On the Generalization Error of Meta Learning for the Gibbs Algorithm

Yuheng Bu*, Harsha Vardhan Tetali*, Gholamali Aminian†, Miguel Rodrigues‡ and Gregory Wornell§
*University of Flroida, †The Alan Turing Institute, ‡University College London, §Massachusetts Institute of Technology
Email: {buyuheng, vardhanh71}@ufl.edu, gaminian@turing.ac.uk, m.rodrigues@ucl.ac.uk, gww@mit.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206566

***Abstract*—We analyze the generalization ability of joint-training meta learning algorithms via the Gibbs algorithm. Our exact characterization of the expected meta generalization error for the meta Gibbs algorithm is based on symmetrized KL information, which measures the dependence between all meta-training datasets and the output parameters, including task-specific and meta parameters. Additionally, we derive an exact characterization of the meta generalization error for the super-task Gibbs algorithm, in terms of conditional symmetrized KL information within the super-sample and super-task framework introduced in [1] and [2], respectively. Our results also enable us to provide novel distribution-free generalization error upper bounds for these Gibbs algorithms applicable to meta learning.**

I. Introduction

In meta learning problems,[1] we have access to multiple related tasks generated from a task environment, and our goal is to capture the shared information among all tasks and construct a model that can generalize to new tasks drawn from the same environment. State-of-the-art meta learning algorithms—such as [3]—have been successfully used in a wide range of applications, including object detection, data mining, few-shot learning, continual learning, and natural language processing [4]–[8].

Various analyses have been pursued to explain the success of meta learning. For example, [9] introduces the task environment concept in meta learning, and derives generalization upper bounds via uniform convergence. Other techniques, such as PAC-Bayesian and information-theoretic approaches, have been adopted to construct generalization error bounds, demonstrating both environment and task-level dependencies in the generalization behavior of meta learning. High probability PAC-Bayesian bounds have been proposed in [10]–[14]. Inspired by [15]–[17], information-theoretic upper bounds on the expected generalization error of meta learning are developed in [18], and later refined in [19], which bounds the meta generalization error using mutual information for both joint-training[2] and alternate-training[3] algorithms. More recently, [2] develops upper bounds on the meta generalization error in terms of evaluated conditional mutual information via a super-task framework, which extends the super-sample approach in [1]. However, it is important to appreciate that such upper bounds may not fully capture the generalization ability of a meta learning algorithm, as the tightness of the bounds is subject to the limitations of the bounding technique.

In contrast to such approaches, we develop exact characterizations of the generalization errors for joint-training meta learning algorithms via the Gibbs algorithm. We model the empirical meta risk minimization algorithm proposed by [19] via a meta Gibbs algorithm. We also consider a super-task Gibbs algorithm inspired by the super-task framework in [2].

Our main contributions of this work are as follows:

- We provide an exact characterization of the meta generalization error for the meta Gibbs algorithm in terms of symmetrized KL information.
- We provide an exact characterization of the meta generalization error for the Gibbs algorithm in super-task framework [2] using conditional symmetrized KL information.
- Using our exact characterizations of the meta generalization error, we provide distribution-free upper bounds, which expose the convergence rate of the meta generalization error of the joint-training Gibbs algorithms in terms of the number of samples and tasks.

II. Preliminaries

Our exact characterizations involve various information measures. If P and Q are probability measures over space $\mathcal{X}$, and P is absolutely continuous with respect to Q, the Kullback-Leibler (KL) divergence between P and Q is given by $D(P\|Q) \triangleq \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$. If Q is also absolutely continuous with respect to P, the symmetrized KL divergence (i.e., Jeffrey's divergence [20]) is

$$D_{\mathrm{SKL}}(P\|Q) \triangleq D(P\|Q) + D(Q\|P). \tag{1}$$

The mutual information between random variables X and Y is the KL divergence between the joint distribution and product-of-marginal distribution $I(X;Y) \triangleq D(P_{X,Y}\|P_X \otimes P_Y)$. Swapping the role of $P_{X,Y}$ and $P_X \otimes P_Y$ in mutual information, we obtain the lautum information introduced by [21],

$$L(X;Y) \triangleq D(P_X \otimes P_Y\|P_{X,Y}). \tag{2}$$

The symmetrized KL information [22] between X and Y is

$$I_{\mathrm{SKL}}(X;Y) \triangleq D_{\mathrm{SKL}}(P_{X,Y}\|P_X \otimes P_Y) = I(X;Y) + L(X;Y).$$

The conditional mutual information between two random variables X and Y conditioned on Z is the KL divergence between $P_{X,Y|Z}$ and $P_{X|Z} \otimes P_{Y|Z}$ averaged over P_Z,

$$I(X;Y|Z) \triangleq \mathbb{E}_{P_Z}[D(P_{X,Y|Z=z}\|P_{Y|Z=z} \otimes P_{X|Z=z})].$$

[1]a.k.a. lifelong learning or learning to learn

[2]Meta and task-specific parameters are updated within the same dataset.

[3]Meta parameters and task-specific parameters are updated within two different datasets.

Similarly, we can define the conditional lautum information $L(X;Y|Z)$ and the conditional symmetrized KL information

$$I_{\mathrm{SKL}}(X;Y|Z) \triangleq I(X;Y|Z) + L(X;Y|Z). \tag{3}$$

The $(\gamma, \pi(y), f(y,x))$-Gibbs distribution (a.k.a. Gibbs posterior [23]), which was first proposed by [24] in statistical mechanics and further investigated by [25] in information theory, is defined as:

$$P^{\gamma}_{Y|X}(y|x) \triangleq \frac{\pi(y)\, e^{-\gamma f(y,x)}}{V_f(x,\gamma)}, \quad \gamma \geq 0, \tag{4}$$

where γ is the inverse temperature, $\pi(y)$ is a prior distribution on $\mathcal{Y}$, $f(y,x)$ is energy function, and

$$V_f(x,\gamma) \triangleq \int \pi(y) e^{-\gamma f(y,x)} \mathrm{d}y \tag{5}$$

is the partition function.

III. Background and Related Work

Motivations for Gibbs Algorithm: In supervised learning, the Gibbs algorithm can be viewed as a *randomized* empirical risk minimization (ERM) algorithm. In addition, the Stochastic Gradient Langevin Dynamics (SGLD) algorithm is known to converge to the Gibbs algorithm [26]. The Gibbs algorithm can also be interpreted as the solution to the KL-divergence-regularized ERM problem [16], [27], [28]. For more detailed discussions of the Gibbs algorithm, see, e.g., [29].

Gibbs Algorithm and Generalization Error: An exact characterization of the generalization error of the Gibbs algorithm in terms of symmetrized KL information is provided in [29] for supervised learning. The authors also provide a generalization error upper bound with the rate of $\mathcal{O}(1/n)$ under the sub-Gaussian assumption, where n is the number of training samples. An information-theoretic upper bound with a similar $\mathcal{O}(1/n)$ rate is provided by [30] for the Gibbs algorithm with bounded loss function, and PAC-Bayesian bounds using a variational approximation of Gibbs posteriors are studied by [31]. Both [32], [33] focus on bounding the excess risk of the Gibbs algorithm in supervised learning. The generalization errors of the Gibbs algorithm in transfer learning and semi-supervised learning settings have been analyzed in [34] and [35], respectively.

Other Analysis of Meta Learning: Besides the information-theoretic approach to analyze generalization error, there are other analyses of meta learning. For example, the uniform convergence analysis of meta learning is first conducted in [9], and [36] adopts the tool of algorithmic stability. Distribution-dependent lower bounds on the meta learning algorithms are provided in [37].

IV. Meta Generalization Error of the Meta Gibbs Algorithm

A. Problem Formulation

In meta learning, we aim to learn a model from multiple meta-training tasks that generalize to an unseen new task. Following [9], [19], we assume that all tasks are generated from a common environment τ with a meta distribution P_τ over the probability measures defined on $\mathcal{Z}$ as the space of data samples. We denote m different meta-training tasks i.i.d. drawn from the meta distribution as $M_i \sim P_\tau$, $i \in [m]$. Without loss of generality, we assume that there are n training samples $D_{M_i} = \{Z_j^{M_i}\}_{j=1}^n$ for each meta-training task M_i, which are generated (not necessarily i.i.d.) from the source distribution $P_{D_{M_i}}$.

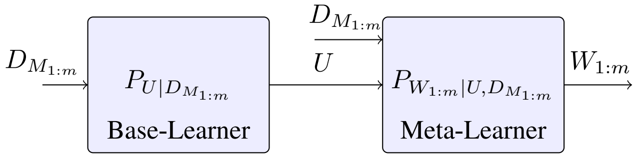

Fig. 1. Joint-training meta learning algorithm

As all tasks, including the unseen test task, are generated from the same meta distribution P_τ, we can use a meta parameter $U \in \mathcal{U}$ to capture the shared knowledge among all tasks and $W_{1:m} = (W_1, \cdots, W_m)$ to denote the task specific-parameters. Here, we adopt a similar formulation as in the two-stage transfer learning considered by [34], where the performance of (U, W_i) is measured by a non-negative loss function $\ell : \mathcal{U} \times \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_0^+$. Thus, we define the following individual empirical risk for a single meta-training task M_i

$$L_E(U, W_i, D_{M_i}) \triangleq \frac{1}{n} \sum_{j=1}^{n} \ell(U, W_i, Z_j^{M_i}), \tag{6}$$

and the joint empirical risk for all meta-training tasks

$$L_E(U, W_{1:m}, D_{M_{1:m}}) \triangleq \frac{1}{m} \sum_{i=1}^{m} L_E(U, W_i, D_{M_i}). \tag{7}$$

A meta learning algorithm, shown in Figure 1, can be decomposed into two components, i.e., a *meta-learner* and a *base-learner*. The meta-learner maps all the dataset of training tasks to a random meta parameter $P_{U|D_{M_{1:m}}}$, and the base-learner maps the meta parameter and dataset of each task to specific parameters, i.e., $P_{W_{1:m}|U,D_{M_{1:m}}} = \prod_{i=1}^n P_{W_i|U,D_{M_i}}$.

We focus on the joint-training meta learning algorithm defined in [19]. In a joint-training algorithm, the training dataset $D_{M_{1:m}}$ are used to obtain all the task-specific parameters $W_{1:m}$ and meta parameter U jointly, which gives the following definition of *empirical meta risk* for meta parameter U,

$$L_E(U, D_{M_{1:m}}) \triangleq \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{P_{W_i|U,D_{M_i}}}[L_E(U, W_i, D_{M_i})]. \tag{8}$$

To evaluate the quality of the meta parameter U, an unseen test task T is drawn from the environment τ with distribution P_τ. We now define the *population meta risk* as follows,

$$L_P(U, \tau) \triangleq \mathbb{E}_{P_\tau}\big[\mathbb{E}_{P_{D_T}}[\mathbb{E}_{P_{W_T|U,D_T}}[L_P(U, W_T, P_{D_T})]]\big], \tag{9}$$

where D_T contains n samples drawn from the test task T, and $L_P(U, W, P_D) = \mathbb{E}_{P_D}[L_E(U, W, D)]$ denotes the standard population risk.

Finally, the expected *meta generalization error* that quantifies the generalizability of meta learning is

$$\begin{aligned}&\overline{\mathrm{gen}}(P_{W_{1:m}|U,D_{M_{1:m}}}, P_{U|D_{M_{1:m}}}, \tau)\\&\quad\triangleq \mathbb{E}_{P_\tau}\big[\mathbb{E}_{P_{U,D_{M_{1:m}}}}[L_P(U,\tau) - L_E(U, D_{M_{1:m}})]\big].\end{aligned} \tag{10}$$

To understand the generalization error in meta learning, we consider the following meta Gibbs algorithm, i.e., $(\gamma, \pi(u, w_{1:m}), L_E(u, w_{1:m}, d_{M_{1:m}}))$-Gibbs algorithm,

$$\begin{aligned}&P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}}(w_{1:m}, u|d_{M_{1:m}})\\&\quad= \frac{\pi(u, w_{1:m})e^{-\gamma L_E(u,w_{1:m},d_{M_{1:m}})}}{V(d_{M_{1:m}},\gamma)}.\end{aligned} \tag{11}$$

Note that this meta Gibbs algorithm is defined by learning U and $W_{1:m}$ jointly. Due to the structure in the joint empirical risk $L_E(U, W_{1:m}, D_{M_{1:m}})$, it can be verified that the induced base-learner satisfies the condition $P_{W_{1:m}|U,D_{M_{1:m}}} = \prod_{i=1}^n P_{W_i|U,D_{M_i}}$, i.e., W_i only depends on D_{M_i} conditioning on the meta-parameter U.

B. Characterization of Expected Meta Generalization Error

The following theorem provides an exact characterization of the expected meta generalization error of the meta Gibbs algorithm using symmetrized KL information. Due to space limitations, all the detailed proofs are provided in [38].

Theorem 1: For the meta Gibbs algorithm in (11), the expected meta generalization error is

$$\overline{\mathrm{gen}}(P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}}, \tau) = \frac{\mathbb{E}_{P_\tau}[I_{\mathrm{SKL}}(U, W_{1:m}; D_{M_{1:m}})]}{\gamma}.$$

Theorem 1 only assumes that the meta-training tasks $P_{D_{M_i}}$ are i.i.d generated from P_τ, and it holds even when the n samples in $D_{M_i} = \{Z_j^{M_i}\}_{j=1}^n$ are not i.i.d.

Some basic properties of the expected meta generalization error can be proved directly from the properties of symmetrized KL information.

a) Non-negativity: The non-negativity of the expected meta generalization error, i.e., $\overline{\mathrm{gen}}(P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}}, \tau) \geq 0$, follows from the non-negativity of $I_{\mathrm{SKL}}(U, W_{1:m}; D_{M_{1:m}})$.

b) Concavity: It is shown in [22] that the symmetrized KL information $I_{\mathrm{SKL}}(X;Y)$ is a concave function of P_X for fixed $P_{Y|X}$. Thus, we have

$$\begin{aligned}&\mathbb{E}_{P_\tau}[I_{\mathrm{SKL}}(P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}}, P_{D_{M_{1:m}}})]\\&\quad\leq I_{\mathrm{SKL}}(P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}}, \mathbb{E}_{P_\tau}[P_{D_{M_{1:m}}}]).\end{aligned} \tag{12}$$

Note that $\mathbb{E}_{P_\tau}[P_{D_{M_{1:m}}}]$ can be viewed as the mixture of all task distributions $P_{D_{M_i}}$ from the environment τ averaged with P_τ. From Theorem 1, an operational interpretation of this inequality is that for fixed meta Gibbs algorithm $P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}}$, the meta generalization error will increase if we mix the datasets from different meta-training tasks, compared to treating different meta-training tasks separately.

To deepen our understanding of the meta Gibbs algorithm, we apply the expansion of lautum information in [21, Eq. (52)] and chain rule of mutual information to Theorem 1,

$$\begin{aligned}&I_{\mathrm{SKL}}(U, W_{1:m}; D_{M_{1:m}})\\&\quad= I_{\mathrm{SKL}}(U; D_{M_{1:m}}) + I(W_{1:m}; D_{M_{1:m}}|U)\\&\qquad+ D(P_{W_{1:m}|U}\|P_{W_{1:m}|U,M_{1:m}}|P_U P_{M_{1:m}}).\end{aligned} \tag{13}$$

Here, the first $I_{\mathrm{SKL}}(U; D_{M_{1:m}})$ term reflects the generalization error caused by learning the shared meta parameter U, and the remaining conditional information and divergence terms correspond to the generalization error in task-specific parameters.

C. Example: Mean Estimation

We now generalize the mean estimation problem considered in [29], [34] to the meta-learning setting, where the symmetrized KL information can be computed easily.

Consider the problem of estimating the mean $\boldsymbol{\mu} \in \mathbb{R}^d$ of the test task using samples from m different meta-training tasks $D_{M_{1:m}} = \{\{Z_j^{M_i}\}_{j=1}^n\}_{i=1}^m$, and $D_T = \{Z_j^T\}_{j=1}^n$, where each task has n i.i.d. samples. We assume that the samples from the meta-training and test tasks satisfying $\mathbb{E}[Z^{M_i}] = \boldsymbol{\mu}_{M_i}$ and $\mathrm{cov}[Z^{M_i}] = \sigma_Z^2\boldsymbol{I}_d$, $\forall i \in [m]$, and $\mathbb{E}[Z^T] = \boldsymbol{\mu}_T$ and $\mathrm{cov}[Z^T] = \sigma_Z^2\boldsymbol{I}_d$, respectively. Thus, the environment τ will generate tasks with different mean $\boldsymbol{\mu}_{M_i} \sim \mathcal{N}(0, \sigma_\tau^2\boldsymbol{I}_d)$, but the covariance matrices of all tasks are the same. We adopt the following regularized mean-squared loss $\ell(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}) = \alpha\|\boldsymbol{z} - \boldsymbol{w}\|_2^2 + (1-\alpha)\|\boldsymbol{u} - \boldsymbol{w}\|_2^2$, for $\boldsymbol{w},\boldsymbol{u},\boldsymbol{z} \in \mathbb{R}^d$, $\alpha \in [0,1]$, and assume uniform distribution over the entire space (improper prior) $\pi(\boldsymbol{w})$ to simplify the computation.

For this setting, the $(\gamma, \pi(\boldsymbol{u}, \boldsymbol{w}_{1:m}), L_E(\boldsymbol{u}, \boldsymbol{w}_i, d_{M_i}))$-Gibbs algorithm is given by the Gaussian posterior distribution, $P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}} \sim \mathcal{N}(\boldsymbol{\mu}_{W_{1:m},U}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_{W_{1:m},U} \in \mathbb{R}^{(m+1)d}$, and

$$\boldsymbol{\mu}_{W_i} = \alpha\bar{Z}^{M_i} + (1-\alpha)\bar{Z}^{M_{1:m}}, \quad \boldsymbol{\mu}_U = \bar{Z}^{M_{1:m}}. \tag{14}$$

Here, the notations

$$\bar{Z}^{M_i} \triangleq \frac{1}{n}\sum_{j=1}^n Z_j^{M_i}, \quad \bar{Z}^{M_{1:m}} \triangleq \frac{1}{m}\sum_{i=1}^m \bar{Z}^{M_i}, \tag{15}$$

are sample means of each meta-training task and the sample mean across all training tasks, respectively. Moreover, the covariance matrix has the following structure,

$$\boldsymbol{\Sigma}^{-1} = \frac{2\gamma}{m}\begin{bmatrix} \boldsymbol{I}_d & \cdots & 0 & (\alpha-1)\boldsymbol{I}_d \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \boldsymbol{I}_d & (\alpha-1)\boldsymbol{I}_d \\ (\alpha-1)\boldsymbol{I}_d & \cdots & (\alpha-1)\boldsymbol{I}_d & m(1-\alpha)\boldsymbol{I}_d \end{bmatrix},$$

which demonstrates the conditional independence between W_i and W_j given U for any $i \neq j$.

Since $P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}}$ is Gaussian, the symmetrized KL information does not depend on the distribution $P_Z^{M_i}$ as long as $\mathrm{cov}[Z^{M_i}] = \sigma_Z^2\boldsymbol{I}_d$, i.e.,

$$I_{\mathrm{SKL}}(U, W_{1:m}; D_{M_{1:m}}) = \frac{2\gamma\alpha((m-1)\alpha+1)d\sigma_Z^2}{mn}. \tag{16}$$

From Theorem 1, the expected meta generalization error of this algorithm can be computed exactly as:

$$\overline{\text{gen}}(P^{\gamma}_{W_{1:m},U|D_{M_{1:m}}},\tau) = \frac{2\alpha^2 d\sigma_Z^2}{n} + \frac{2\alpha(1-\alpha)d\sigma_Z^2}{mn}, \quad (17)$$

which gives a rate of $\mathcal{O}(\frac{d}{mn}+\frac{d}{n})$.

When $\alpha = 1$, the loss function $\ell(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}) = \|\boldsymbol{z}-\boldsymbol{w}\|_2^2$ does not depend on the meta parameter $\boldsymbol{u}$ anymore, which suggests no interaction between different meta-training tasks, and U can be set arbitrarily. Thus, the meta generalization error in (17) reduces to $\frac{2d\sigma_Z^2}{n}$, which is precisely the generalization error of the ERM algorithm with n i.i.d samples from P_Z^T in supervised learning setting (see, [29]).

When $\alpha = 0$, the loss function $\ell(\boldsymbol{w},\boldsymbol{u},\boldsymbol{z}) = \|\boldsymbol{u}-\boldsymbol{w}\|_2^2$ does not depend on any samples. In this case, the meta generalization error in (17) is 0.

For general $\alpha \in (0,1)$, it can be verified that the meta generalization error is always smaller than $\frac{2d\sigma_Z^2}{n}$, i.e., the generalization error of ERM in supervised learning.

Remark 1 (Effect of P_τ): As shown in (17), the meta generalization error of this mean estimation problem does not depend on the meta distribution P_τ, where the variance σ_τ^2 captures the diversity of the means $\boldsymbol{\mu}_{M_i}$ for different meta-training tasks. One reason is that the effect of the means is canceled out in meta generalization error by subtracting the empirical meta risk from the population meta risk. Although different σ_τ^2 do not change meta generalization errors in this example, a large σ_τ^2 implies less similarity between different tasks, and it will lead to large population meta risks. Another reason is that we set sample variance σ_Z^2 to be the same across all tasks. When environment τ generates tasks with different sample variances, meta generalization error will depend on P_τ.

V. Meta Generalization Error of the Super-Task Gibbs Algorithm

In this section, we analyze the super-task framework for meta-learning introduced in [2] from the perspective of the Gibbs algorithm, and we offer the exact characterization of the meta generalization error. All proofs are provided in [38].

A. Notation

We adopt the notation used in [2] for this section. The matrix $\boldsymbol{Z} \in \mathcal{Z}^{n\times 4m}$ represents the entire dataset, where we divide the columns of the matrix into $2m$ groups. Each group consists of a pair of columns, where the first and second columns are the first group, the third and fourth form the second group, and so on. Each group contains $2n$ samples i.i.d. generated from the same meta task drawn from the meta distribution P_τ. The columns in each group are labeled, with the first column labeled 0 and the second column labeled 1. We introduce the notation $\boldsymbol{Z}_{j,l} \in \mathcal{Z}^{2m}$, where $j \in [n]$ and $l \in \{0,1\}$, as a row vector formed by the j-th element in the column labeled by l in each of the $2m$ groups.

To differentiate between different meta tasks, we further label these $2m$ tasks with (i,k) for $i \in [m]$ and $k \in \{0,1\}$. In addition, we use super-scripts to choose the (i,k)-th meta-task among the $2m$ tasks. Thus, $\boldsymbol{Z}_{j,l}^{i,k} \in \mathcal{Z}$ is the (i,k)-th element of the vector $\boldsymbol{Z}_{j,l}$. As shown in Fig. 2, we use superscripts to select among tasks and subscripts to select among samples.

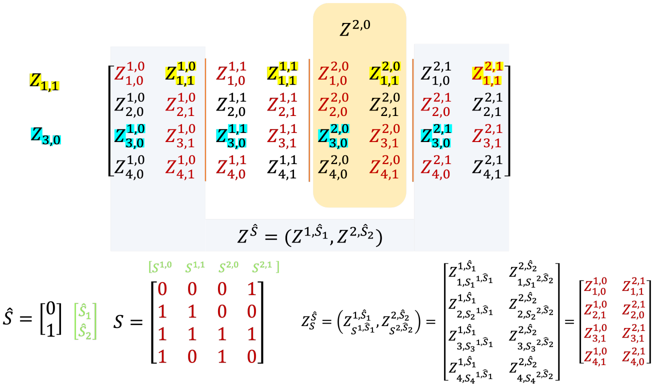

Fig. 2. A graphical representation of the notation system. We chose $m = 2$, i.e., 4 meta tasks, and $n = 4$, i.e., 8 data samples per task.

We define a meta-training task membership vector $\hat{S} \in \{0,1\}^m$, where each element $\hat{S}_i$ is i.i.d. drawn from Bern$(1/2)$. The meta-training tasks are selected according to the elements in $\{(i,\hat{S}_i) : i \in [m]\}$ and the meta test tasks are selected according to $\{(i,-\hat{S}_i) : i \in [m]\}$, where $-\hat{S}_i \triangleq 1-\hat{S}_i$. Within each meta task, we have $2n$ data samples, and we randomly select half of them as the training samples and the remaining as the test samples using a randomly generated matrix $S \in \{0,1\}^{n\times 2m}$, where the elements $S_j^{i,k}$ are drawn from Bern$(1/2)$ for $i \in [m]$, $k \in 0,1$, and $j \in [n]$. Each column of S is a binary vector of length n that indicates which sample is selected as the training data. Our complete meta-training dataset is formed by $\{\boldsymbol{Z}^{i,\hat{S}_i}_{j,S_j^{i,\hat{S}_i}}\}_{i,j=1}^{m,n}$.

B. Characterization of Expected Meta Generalization Error

For given membership variables S and $\hat{S}$, we can rewrite the individual empirical risk for task $(i,\hat{S}_i)$ under this super-task framework as

$$L_E(U, W^{i,\hat{S}_i}, \boldsymbol{Z}_S^{i,\hat{S}_i}) \triangleq \frac{1}{n}\sum_{j=1}^{n} \ell(U, W^{i,\hat{S}_i}, \boldsymbol{Z}^{i,\hat{S}_i}_{j,S_j^{i,\hat{S}_i}}), \quad (18)$$

and the joint empirical risk for all meta-training tasks as

$$L_E(U, W^{\hat{S}}, \boldsymbol{Z}_S^{\hat{S}}) \triangleq \frac{1}{k}\sum_{i=1}^{k} L_E(U, W^{i,\hat{S}_i}, \boldsymbol{Z}_S^{i,\hat{S}_i}). \quad (19)$$

We can consider the super-task Gibbs algorithm for meta-training tasks using the joint empirical risk, i.e., $(\gamma, \pi(u, w^{\hat{S}}), L_E(u, w^{\hat{S}}, \boldsymbol{Z}_S^{\hat{S}}))$-Gibbs algorithm

$$P^{\gamma}_{W^{\hat{S}},U|S,\hat{S},\boldsymbol{Z}}(w^{\hat{S}},u) = \frac{\pi(u,w^{\hat{S}})e^{-\gamma L_E(u,w^{\hat{S}},\boldsymbol{Z}_S^{\hat{S}})}}{V_1(\boldsymbol{Z}_S^{\hat{S}},\gamma)}, \quad (20)$$

for those meta test tasks, the task specific weights $W^{-\hat{S}}$ are obtained by $(\gamma, \pi(w^{-\hat{S}}), L_E(u, w^{-\hat{S}}, \boldsymbol{Z}_S^{-\hat{S}}))$-Gibbs algorithm for a given $U = u$,

$$P^{\gamma}_{W^{-\hat{S}}|U,S,\hat{S},\boldsymbol{Z}}(w^{-\hat{S}}) = \frac{\pi(w^{-\hat{S}})e^{-\gamma L_E(u,w^{-\hat{S}},\boldsymbol{Z}_S^{-\hat{S}})}}{V_2(U,\boldsymbol{Z}_S^{-\hat{S}},\gamma)}.$$

Inspired by [2], we define the following four different types of losses using the membership variables S and $\hat{S}$.

$$\widehat{L} = \mathbb{E}\Big[L_E\big(U, W^{\hat{S}}, \boldsymbol{Z}_S^{\hat{S}}\big)\Big],\ \bar{L} = \mathbb{E}\Big[L_E\big(U, W^{\hat{S}}, \boldsymbol{Z}_{-S}^{\hat{S}}\big)\Big],$$
$$\widetilde{L} = \mathbb{E}\Big[L_E\big(U, W^{-\hat{S}}, \boldsymbol{Z}_S^{-\hat{S}}\big)\Big],\ L_P = \mathbb{E}\Big[L_E\big(U, W^{-\hat{S}}, \boldsymbol{Z}_{-S}^{-\hat{S}}\big)\Big],$$

where $\widehat{L}$ is the expected empirical meta risk evaluated on meta-training tasks, L_P is the population meta risk evaluated on unseen tasks. The remaining two losses are the expected auxiliary test loss $\bar{L}$, which is the loss on test data for training tasks, and the expected auxiliary training loss $\widetilde{L}$, which is the loss on training data for test tasks.

The expected meta generalization error in super-task setting is $\overline{\text{gen}}(P_{W^{\hat{S}},U|S,\hat{S},\boldsymbol{Z}}, \tau) \triangleq \mathbb{E}_{P_\tau}[L_P - \widehat{L}]$.

The following theorem characterizes the meta generalization error using conditional symmetrized KL information by decomposing it into these four different types of losses.

Theorem 2: For the super-task Gibbs algorithm defined in (20), it can be shown

1) $\big(L_P + \bar{L} + \widetilde{L} + \widehat{L}\big) - \gamma\widehat{L} = \dfrac{4}{\gamma} I_{SKL}\big(U, W^{\hat{S}}; S, \hat{S}|\boldsymbol{Z}\big)$,
2) $\big(\bar{L} - \widehat{L}\big) = \dfrac{2}{\gamma} I_{SKL}\big(U, W^{\hat{S}}; S|\hat{S}, \boldsymbol{Z}\big)$,
3) $\big(\widetilde{L} - \widehat{L}\big) = \dfrac{2}{\gamma} I_{SKL}\big(U, W^{\hat{S}}; \hat{S}|S, \boldsymbol{Z}\big)$,
4) $\big(L_P - \widetilde{L}\big) = \dfrac{2}{\gamma} I_{SKL}\big(W^{-\hat{S}}; S|U, \hat{S}, \boldsymbol{Z}\big)$,

and the meta generalization error is given by

$$\overline{\text{gen}}(P_{W^{\hat{S}},U|S,\hat{S},\boldsymbol{Z}}, \tau) = \frac{2}{\gamma}\mathbb{E}_{P_\tau}\Big[I_{SKL}\big(W^{-\hat{S}}; S|U, \hat{S}, \boldsymbol{Z}\big) + I_{SKL}\big(U, W^{\hat{S}}; \hat{S}|S, \boldsymbol{Z}\big)\Big]. \tag{21}$$

As shown in Theorem 2, the meta generalization error can be decomposed into two symmetrized KL information terms $I_{SKL}\big(U, W^{\hat{S}}; \hat{S}|S, \boldsymbol{Z}\big)$ and $I_{SKL}\big(W^{-\hat{S}}; S|U, \hat{S}, \boldsymbol{Z}\big)$, which represents $L_P - \widetilde{L}$ and $\widetilde{L} - \widehat{L}$, respectively.

VI. Distribution-free Upper Bound

In this section, we present distribution-free upper bounds for the meta Gibbs algorithm and super-task Gibbs algorithm. These bounds characterize the relationship between the meta generalization error and the number of tasks m and the number of samples per task n. It can be utilized in situations where direct computation of symmetrized KL information is challenging. All detailed proofs are provided in [38].

In the following Theorem, we provide the distribution-free upper bound on meta Gibbs algorithm by combining Theorem 1 and [19, Theorem 5.1].

Theorem 3: Suppose that the meta target training samples $D_{M_i} = \{Z_j^{M_i}\}_{j=1}^n$ are i.i.d generated from the distribution $P_Z^{M_i}$, and the non-negative loss function $\ell(u, w, Z)$ is σ_{meta}-sub-Gaussian under distribution $Z \sim P_Z^{M_i}$ and $M_i \sim P_\tau$ for all $u \in \mathcal{U}$ and $w \in \mathcal{W}$. If we further assume $C_{\text{meta}} \leq \frac{L(U,W_{1:m};D_{M_{1:m}})}{I(U,W_{1:m};D_{M_{1:m}})}$ for some $C_{\text{meta}} \geq 0$, then for the meta Gibbs algorithm in (11), we have

$$\overline{\text{gen}}(P^\gamma_{W_{1:m},U|D_{M_{1:m}}}, \tau) \leq \frac{2\sigma^2_{\text{meta}}\gamma}{(1 + C_{\text{meta}})mn}. \tag{22}$$

As shown in [16], the sub-Gaussian condition in Theorem 3 holds for all bounded loss functions.

Remark 2: In comparison to the meta generalization upper bounds of the general meta learning algorithm in [2], [19] that scale as $\mathcal{O}(\frac{1}{\sqrt{mn}})$, we prove that the meta generalization error of meta Gibbs algorithm has a faster convergence rate $\mathcal{O}(\frac{1}{mn})$.

Remark 3: It can be verified easily that the loss function $\ell(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{z})$ considered in the mean estimation example in Sec. IV-C is not bounded and does not satisfy the sub-Gaussian assumption in Theorem 3, which results in a rate of $\mathcal{O}(\frac{1}{mn} + \frac{1}{n})$ instead of the faster rate $\mathcal{O}(\frac{1}{mn})$.

The following distribution-free upper bound on the super-task Gibbs algorithm can be obtained by combining Theorem 2 and [2, Corollary 1].

Theorem 4: If the non-negative loss function is bounded, i.e., $\ell(u, w, z) \in [0, 1]$, then for the super-task Gibbs algorithm defined in (20), we have

$$\overline{\text{gen}}(P_{W^{\hat{S}},U|S,\hat{S},\boldsymbol{Z}}, \tau) \leq \frac{\gamma}{m} + \frac{\gamma}{n}. \tag{23}$$

Remark 4: Compared with the bound in Theorem 3, the rate we obtained using super task framework is $\mathcal{O}(\frac{1}{m} + \frac{1}{n})$, which is sub-optimal. We believe this is due to the triangle inequality $L_P - \widehat{L} \leq |L_P - \widetilde{L}| + |\widetilde{L} - \widehat{L}|$ used by the two-step method in [2, Theorem 1], where a similar sub-optimal bound using this approach is obtained in [2, Corollary 6]. Although Theorem 2 adopts a similar decomposition involving $\widetilde{L}$, our characterization of the meta generalization error is exact.

VII. Conclusion and Future Works

We characterize the meta generalization error for the joint-training approach via the meta Gibbs algorithm in terms of symmetrized KL information and the super-task Gibbs algorithm in terms of conditional symmetrized KL information, respectively. We also develop distribution-free upper bounds, which yield better estimates of the convergence rate compared to those available in the existing literature. In future work, we plan to extend our framework to the alternate-training approach. This will include applying asymptotic analysis—similar to [39]—and provide an exact characterization in the asymptotic regime in which $\gamma \to \infty$.

Acknowledgements

Harsha Vardhan Tetali is supported by NSF EECS-1839704 and NSF CISE-1747783. Gholamali Aminian is supported by the UKRI Prosperity Partnership Scheme (FAIR) under the EPSRC Grant EP/V056883/1. M. R. D. Rodrigues and Gholamali Aminian are also supported by the Alan Turing Institute. This work has also been supported in part by the MIT-IBM Watson AI Lab under Agreement No. W1771646, AFRL under Cooperative Agreement No. FA8750-19-2-1000, NSF under Grant No. CCF-1816209.

References

[1] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Conference on Learning Theory*. PMLR, 2020, pp. 3437–3452.

[2] F. Hellström and G. Durisi, "Evaluated CMI bounds for meta learning: Tightness and expressiveness," in *Advances in Neural Information Processing Systems*, 2022.

[3] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[4] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020.

[5] P. Brazdil, C. G. Carrier, C. Soares, and R. Vilalta, *Metalearning: Applications to data mining*. Springer Science & Business Media, 2008.

[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021.

[7] A. Obamuyide and A. Vlachos, "Model-agnostic meta-learning for relation classification with limited supervision," in *Association for Computational Linguistics*, 2019.

[8] J. Harrison, A. Sharma, C. Finn, and M. Pavone, "Continuous meta-learning without tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 571–17 581, 2020.

[9] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, 2000.

[10] A. Pentina and C. Lampert, "A PAC-Bayesian bound for lifelong learning," in *International Conference on Machine Learning*. PMLR, 2014, pp. 991–999.

[11] R. Amit and R. Meir, "Meta-learning by adjusting priors based on extended PAC-Bayes theory," in *International Conference on Machine Learning*. PMLR, 2018, pp. 205–214.

[12] T. Liu, J. Lu, Z. Yan, and G. Zhang, "PAC-Bayes bounds for meta-learning with data-dependent prior," *arXiv preprint arXiv:2102.03748*, 2021.

[13] J. Rothfuss, V. Fortuin, M. Josifoski, and A. Krause, "PACOH: Bayes-optimal meta-learning with PAC-guarantees," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9116–9126.

[14] A. Rezazadeh, "A general framework for PAC-Bayes bounds for meta-learning," *arXiv preprint arXiv:2206.05454*, 2022.

[15] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, 2019.

[16] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, pp. 2524–2533.

[17] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020.

[18] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," *Entropy*, vol. 23, no. 1, p. 126, 2021.

[19] Q. Chen, C. Shui, and M. Marchand, "Generalization bounds for meta-learning: An information-theoretic analysis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 878–25 890, 2021.

[20] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[21] D. P. Palomar and S. Verdú, "Lautum information," *IEEE transactions on information theory*, vol. 54, no. 3, pp. 964–975, 2008.

[22] G. Aminian, H. Arjmandi, A. Gohari, M. Nasiri-Kenari, and U. Mitra, "Capacity of diffusion-based molecular communication networks over LTI-Poisson channels," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 1, no. 2, pp. 188–201, 2015.

[23] O. Catoni, "PAC-Bayesian supervised classification: the thermodynamics of statistical learning," *arXiv preprint arXiv:0712.0248*, 2007.

[24] J. W. Gibbs, "Elementary principles of statistical mechanics," *Compare*, vol. 289, p. 314, 1902.

[25] E. T. Jaynes, "Information theory and statistical mechanics," *Physical review*, vol. 106, no. 4, p. 620, 1957.

[26] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis," in *Conference on Learning Theory*. PMLR, 2017, pp. 1674–1703.

[27] T. Zhang, "Information-theoretic upper and lower bounds for statistical estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[28] T. Zhang *et al.*, "From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, vol. 34, no. 5, pp. 2180–2210, 2006.

[29] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[30] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *2016 IEEE Information Theory Workshop (ITW)*. IEEE, 2016, pp. 26–30.

[31] 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, 2016.

[32] A. R. Asadi and E. Abbe, "Chaining meets chain rule: Multilevel entropic regularization and training of neural networks," *Journal of Machine Learning Research*, vol. 21, no. 139, pp. 1–32, 2020.

[33] I. Kuzborskij, N. Cesa-Bianchi, and C. Szepesvári, "Distribution-dependent analysis of Gibbs-ERM principle," in *Conference on Learning Theory*. PMLR, 2019, pp. 2028–2054.

[34] Y. Bu, G. Aminian, L. Toni, G. W. Wornell, and M. Rodrigues, "Characterizing and understanding the generalization error of transfer learning with Gibbs algorithm," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8673–8699.

[35] H. He, G. Aminian, Y. Bu, M. Rodrigues, and V. Y. Tan, "How does pseudo-labeling affect the generalization error of the semi-supervised gibbs algorithm?" in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 8494–8520.

[36] A. Maurer and T. Jaakkola, "Algorithmic stability and meta-learning." *Journal of Machine Learning Research*, vol. 6, no. 6, 2005.

[37] M. Konobeev, I. Kuzborskij, and C. Szepesvári, "A distribution-dependent analysis of meta learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5697–5706.

[38] Y. Bu, H. V. Tetali, G. Aminian, M. Rodrigues, and G. W. Wornell, "On the generalization error of meta learning for the Gibbs algorithm," *arXiv preprint arXiv: 2304.14332*, 2023.

[39] G. Aminian, Y. Bu, L. Toni, M. R. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," *arXiv preprint arXiv:2210.09864*, 2022.