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 Florida, †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

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-training2 and alternate-training3 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 X, and P is absolutely continuous with respect to Q, the Kullback-Leibler (KL) divergence between P and Q is given by

$$D(P\|Q) = \int \log \frac{dP}{dQ} dP.$$

If Q is also absolutely continuous with respect to P, the symmetrized KL divergence (i.e., Jeffrey’s divergence [20]) is

$$D_{SKL}(P\|Q) = \frac{1}{2} [D(P\|Q) + D(Q\|P)].$$

The mutual information between random variables X and Y is the KL divergence between the joint distribution and product-of-marginal distribution

$$I(X;Y) = D(P_{X,Y}\|P_X \otimes P_Y).$$

Swapping the role of P_{X,Y} and P_X \otimes P_Y in mutual information, we obtain the lautum information introduced by [21],

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

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

$$I_{SKL}(X;Y) = \frac{1}{2} D_{SKL}(P_{X,Y}\|P_X \otimes P_Y) = I(X;Y) + L(X;Y).$$

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 \otimes P_{Y|Z} averaged over P_Z,

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

1a.k.a. lifelong learning or learning to learn
2Meta and task-specific parameters are updated within the same dataset.
3Meta parameters and task-specific parameters are updated within two different datasets.
Similarly, we can define the conditional lautum information $I(X;Y|Z)$ and the conditional symmetrized KL information
\begin{equation}
I_{SKL}(X;Y|Z) \triangleq I(X;Y|Z) + L(X;Y|Z).
\end{equation}

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:
\begin{equation}
P_{\gamma|x}(y|x) \triangleq \frac{\pi(y) e^{-\gamma f(y,x)}}{V_f(x,\gamma)}, \quad \gamma > 0,
\end{equation}
where $\gamma$ is the inverse temperature, $\pi(y)$ is a prior distribution on $Y$, $f(y,x)$ is energy function, and
\begin{equation}
V_f(x,\gamma) \triangleq \int \pi(y)e^{-\gamma f(y,x)}dy
\end{equation}
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-risk minimization (ERM) algorithm. In addition, the Stochastic Gradient Langevin Dynamics (SGLD) algorithm is known to converge to the Gibbs algorithm [26].

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 $\tau$ with a meta distribution $P_{\tau}$ over the probability measures defined on $Z$ 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 $D_{M_i} = \{Z_{i,j}^{M_i}\}_{j=1}^n$ for each meta-training task $M_i$, which are generated (not necessarily i.i.d.) from the source distribution $P_{DM_i}$.

As all tasks, including the unseen test task, are generated from the same meta distribution $P_{\tau}$, we can use a meta parameter $U \in U$ to capture the shared knowledge among all tasks and $W_{1:m} = (W_1, \ldots, W_m)$ 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 $\ell : U \times W \times Z \rightarrow \mathbb{R}_+^+$. Thus, we define the following individual empirical risk for a single meta-training task $M_i$
\begin{equation}
L_E(U,W_i,D_{M_i}) \triangleq \frac{1}{n} \sum_{j=1}^n \ell(U,W_i,Z^{M_i}_j),
\end{equation}
and the joint empirical risk for all meta-training tasks
\begin{equation}
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}).
\end{equation}

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 $U \in U$, while the base-learner maps the meta parameter and dataset of each task to specific parameters, i.e., $P_{W_{1:m}|U,D_{M_{1:m}}} = \prod_{i=1}^m P_{W_i|U,D_{M_i}}$.

We focus on the joint-training meta learning algorithm defined in [19]. In a joint-training algorithm, the training dataset $D_{M_{1:m}}$ is 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 $U$,
\begin{equation}
L_E(U,D_{M_{1:m}}) \triangleq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{P_{W_{1:m}|U,D_{M_i}}} [L_E(U,W_i,D_{M_i})].
\end{equation}

To evaluate the quality of the meta parameter $U$, an unseen test task $T$ is drawn from the environment $\tau$ with distribution $P_{\tau}$. We now define the population meta risk as follows,
\begin{equation}
L_P(U,\tau) \triangleq \mathbb{E}_{P_{\tau}}[\mathbb{E}_{P_{D_T}}[\mathbb{E}_{P_{W_{1:m}|U,D_T}}[L_P(U,W_T,P_{D_T})]]],
\end{equation}
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
\[
\mathbb{E}[P_{W_1}(U, D_{M_1}; P_{U}|D_{M_1}, \tau)] = \mathbb{E}_{P_{\tau}}[\mathbb{E}_{P_{\tau,D_{M_1}}}[P_{U}(U, \tau) - L_{E}(U, D_{M_1})]].
\]
To understand the generalization error in meta learning, we consider the following meta Gibbs algorithm, i.e., \(P_{W_1}(U|D_{M_1})\)-Gibbs algorithm,
\[
P_{W_1}(U|D_{M_1}) = \pi(U|D_{M_1})e^{-\gamma L_{E}(U, D_{M_1})}.
\]
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 \(L_{E}(U, W_{1:m}, D_{M_1})\), it can be verified that the induced base-learner satisfies the condition \(P_{W_1}(U|D_{M_1}) = \prod_{i=1}^{m} P_{W_i}(U|D_{M_1})\), 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
\[
\mathbb{E}[P_{W_1,m}(U|D_{M_1})] = \mathbb{E}_{P_{\tau}}[I_{SKL}(U, W_{1:m}; D_{M_1})].
\]

Theorem 1 only assumes that the meta-training tasks \(P_{D_{M_i}}\) are i.i.d. generated from \(P_{\tau}\), and it holds even when the \(n\) samples in \(D_{M_i} = \{Z_{M_i}^{m, j}\}_{j=1}^{n}\), \(m\) are 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., \(\mathbb{E}[P_{W_1,m}(U|D_{M_1})] \geq 0\), follows from the non-negativity of \(I_{SKL}(U, W_{1:m}; D_{M_1})\).

b) **Concavity:** It is shown in [22] that the symmetrized KL information \(I_{SKL}(X; Y)\) is a concave function of \(P_{X}\) for fixed \(P_{Y|X}\). Thus, we have
\[
\mathbb{E}_{P_{\tau}}[I_{SKL}(P_{W_1,m}|U|D_{M_1}; P_{D_{M_1}})] \leq I_{SKL}(P_{W_1,m}|U|D_{M_1}; \mathbb{E}_{P_{\tau}}[P_{D_{M_1}}]).
\]

To deepen our understanding of the meta Gibbs algorithm, we apply the expansion of lautom information in [21, Eq. (52)] and chain rule of mutual information to Theorem 1,
\[
I_{SKL}(U, W_{1:m}; D_{M_1}) = I_{SKL}(U; D_{M_1}) + I(W_{1:m}; D_{M_1}|U) + D(P_{W_1,m}|U) I_{P_{U}}(P_{W_1,m}|D_{M_1}).
\]
Here, the first \(I_{SKL}(U; D_{M_1})\) 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 \(\mu \in \mathbb{R}^d\) of the test task using samples from \(m\) different meta-training tasks \(D_{M_1} = \{\{Z_{M_1}^{m, j}\}_{j=1}^{n}\}_{i=1}^{m}\), and \(D_{Y} = \{Z_{Y}^{m, j}\}_{j=1}^{n}\), where each task has \(n\) i.i.d. samples. We assume that the samples from the meta-training and test tasks satisfy \(E[Z_{M_1}^{m, j}] = \mu_{M_1}\) and \(cov[Z_{M_1}^{m, j}] = \sigma_{Z}^{2} I_{d}\), respectively. Thus, the environment \(\tau\) will generate tasks with different mean \(\mu_{M_1} \sim N(0, \sigma_{Z}^{2} I_{d})\), but the covariance matrices of all tasks are the same. We adopt the following regularized mean-squared loss \(\ell(w, u, z) = \alpha \|z - w\|^2 + (1 - \alpha)\|u - w\|^2_{2}\) for \(w, u, z \in \mathbb{R}^d\), \(\alpha \in [0, 1]\), and assume uniform distribution over the entire space (improper prior) \(\pi(w)\) to simplify the computation.

For this setting, the following properties of 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., \(\mathbb{E}[P_{W_1,m}(U|D_{M_1})] \geq 0\), follows from the non-negativity of \(I_{SKL}(U, W_{1:m}; D_{M_1})\).

b) **Concavity:** It is shown in [22] that the symmetrized KL information \(I_{SKL}(X; Y)\) is a concave function of \(P_{X}\) for fixed \(P_{Y|X}\). Thus, we have
\[
\mathbb{E}_{P_{\tau}}[I_{SKL}(P_{W_1,m}|U|D_{M_1}; P_{D_{M_1}})] \leq I_{SKL}(P_{W_1,m}|U|D_{M_1}; \mathbb{E}_{P_{\tau}}[P_{D_{M_1}}]).
\]

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 \(L_{E}(U, W_{1:m}, D_{M_1})\), it can be verified that the induced base-learner satisfies the condition \(P_{W_1,m}(U|D_{M_1}) = \prod_{i=1}^{m} P_{W_i}(U|D_{M_1})\), i.e., \(W_i\) only depends on \(D_{M_i}\) conditioning on the meta-parameter \(U\).

Theorem 1 only assumes that the meta-training tasks \(P_{D_{M_i}}\) are i.i.d. generated from \(P_{\tau}\), and it holds even when the \(n\) samples in \(D_{M_i} = \{Z_{M_i}^{m, j}\}_{j=1}^{n}\), \(m\) are 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., \(\mathbb{E}[P_{W_1,m}(U|D_{M_1})] \geq 0\), follows from the non-negativity of \(I_{SKL}(U, W_{1:m}; D_{M_1})\).

b) **Concavity:** It is shown in [22] that the symmetrized KL information \(I_{SKL}(X; Y)\) is a concave function of \(P_{X}\) for fixed \(P_{Y|X}\). Thus, we have
\[
\mathbb{E}_{P_{\tau}}[I_{SKL}(P_{W_1,m}|U|D_{M_1}; P_{D_{M_1}})] \leq I_{SKL}(P_{W_1,m}|U|D_{M_1}; \mathbb{E}_{P_{\tau}}[P_{D_{M_1}}]).
\]
From Theorem 1, the expected meta generalization error of this algorithm can be computed exactly as:

$$
\mathbb{E}\left[ P^{\gamma}_{W, m_1, m_2, m_3} \right] = \frac{2\alpha^2 d \sigma^2}{n} + \frac{2\alpha (1 - \alpha) d \sigma^2}{m n},$$

(17)

which gives a rate of $O\left( \frac{d}{mn} + \frac{d^3}{n^2} \right)$.

When $\alpha = 0$, the loss function $\ell(w, u, z) = \|z - w\|_2^2$ does not depend on the meta parameter $u$ 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 $\frac{2d \sigma^2}{n}$, which is precisely the generalization error of the ERM algorithm with $n$ i.i.d samples from $P^\gamma_Z$ in supervised learning setting (see, [29]).

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

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

**Remark 1 (Effect of $P^\gamma$):** As shown in (17), the meta generalization error of this mean estimation problem does not depend on the meta distribution $P^\gamma$, where the variance $\sigma^2$ captures the diversity of the means $\mu_i$ 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 $\sigma^2$ do not change meta generalization errors in this example, a large $\sigma^2$ implies less similarity between different tasks, and it will lead to large population meta risks. Another reason is that we set sample variance $\sigma^2$ to be the same across all tasks. When environment $\tau$ generates tasks with different sample variances, meta generalization error will depend on $P^\gamma$.

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 $Z \in \mathbb{Z}^{n \times 2m}$ 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 $P^\gamma$. The columns in each group are labeled, with the first column labeled 0 and the second column labeled 1. We introduce the notation $Z_{i,l} \in \mathbb{Z}^{2m}$, 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, $Z_{i,l}^{\gamma, j,k} \in \mathbb{Z}$ is the $(i, k)$-th element of the vector $Z_{j,k}$. As shown in Fig. 2, we use superscripts to select among tasks and subscripts to select among samples.

We define a meta-training task membership vector $S \in \{0, 1\}^n$, where each element $S_i$ is i.i.d. drawn from Bern(1/2). The meta-training tasks are selected according to the elements in $\{[S_i, S_j] : i \in [m]\}$ and the meta test tasks are selected according to $\{[i, -S_i] : i \in [m]\}$, where $S_i = 1 - S_i$. 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 $S_i^{j,k}$ 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 $\{Z_{i, [S_i, S_j]}^{\gamma, j,k} \}_{i,j=1}^{m}$.  

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

$$
L_E(U, W_i, \hat{S}_i, Z_S^{i, \hat{S}_i}) \triangleq \frac{1}{n} \sum_{j=1}^{n} \ell(U, W_i, \hat{S}_i, Z_S^{i, \hat{S}_i}),
$$

(18)

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

$$
L_E(U, W, Z_S^{i, \hat{S}_i}) \triangleq \frac{1}{k} \sum_{i=1}^{k} L_E(U, W, Z_S^{i, \hat{S}_i}).
$$

(19)

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

$$
P^{\gamma}_{W, S, U|S, Z_S^{i, \hat{S}_i}}(u, w) = \frac{\pi(u, w) e^{-\gamma L_E(u, w, Z_S^{i, \hat{S}_i})}}{V_i(Z_S^{i, \hat{S}_i}, \gamma)},
$$

(20)

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

$$
P^{\gamma}_{W, S|U, S, Z_S^{i, \gamma}}(w) = \frac{\pi(w, S) e^{-\gamma L_E(u, w, S, Z_S^{i, \gamma})}}{V_i(U, Z_S^{i, \gamma}, \gamma)}.
$$

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

\[
\bar{L} = \mathbb{E}\left[ L_E(U, W^S, Z_S^S) \right], \quad \bar{\tilde{L}} = \mathbb{E}\left[ L_E(U, W^\tilde{S}, Z_{\tilde{S}}^S) \right],
\]

where \( \bar{L} \) 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 \( \tilde{L} \), which is the loss on test data for training tasks, and the expected auxiliary training loss \( \bar{\tilde{L}} \), which is the loss on training data for test tasks.

The expected meta generalization error in super-task setting is given by

\[
\mathbb{E}\left[ P_{W^S, U|S, S, \tilde{S}, Z}^\gamma \right] \triangleq E_{D_P}[L_P - \bar{L}].
\]

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. \( (L_P + \bar{L} + \tilde{L} + \bar{\tilde{L}}) - \gamma \bar{L} = \frac{1}{\gamma} I_{\text{SKL}}(U, W^S; S, \tilde{S}| Z), \)
2. \( (\bar{L} - \tilde{L}) = \frac{2}{\gamma} I_{\text{SKL}}(U, W^S; S| \tilde{S}, Z), \)
3. \( (\bar{L} - \tilde{L}) = \frac{2}{\gamma} I_{\text{SKL}}(U, W^S; S| \tilde{S}, Z), \)
4. \( (L_P - \bar{\tilde{L}}) = \frac{2}{\gamma} I_{\text{SKL}}(W^\tilde{S}; S| U, \tilde{S}, Z), \)

and the meta generalization error is given by

\[
\mathbb{E}\left[ P_{W^S, U|S, S, \tilde{S}, Z}^\gamma \right] \triangleq \frac{2}{\gamma} E_{D_P} \left[ I_{\text{SKL}}(W^\tilde{S}; S| U, \tilde{S}, Z) + I_{\text{SKL}}(U, W^S; S| S, Z) \right].
\] \hspace{1cm} (21)

As shown in Theorem 2, the meta generalization error can be decomposed into two symmetrized KL information terms \( I_{\text{SKL}}(U, W^S; S| S, Z) \) and \( I_{\text{SKL}}(W^\tilde{S}; S| U, \tilde{S}, Z) \), which represents \( L_P - \bar{\tilde{L}} \) and \( \bar{L} - \tilde{L} \), 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 \( D_{\text{Mt}} = \{Z_j^M\}_{j=1}^n \) are i.i.d generated from the distribution \( P_{Z_M^j} \), and the non-negative loss function \( \ell(u, w, z) \) is \( \sigma_{\text{meta}} \)-sub-Gaussian under distribution \( Z \sim P_Z^M \) and \( M_j \sim P_T \) for all \( u \in U \) and \( w \in W \). If we further assume \( C_{\text{meta}} \leq L(U, W^1, D_{\text{Mt}1}) \) for some \( C_{\text{meta}} \geq 0 \), then for the meta Gibbs algorithm in (11), we have

\[
\mathbb{E}\left[ P_{W^1, U|D_{\text{Mt}1}, \tau}^\gamma \right] \leq \frac{2\sigma_{\text{meta}}^2 \gamma}{(1 + C_{\text{meta}}) m n}.
\] \hspace{1cm} (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 \( O(\frac{1}{mn}) \), we prove that the meta generalization error of meta Gibbs algorithm has a faster convergence rate \( O(\frac{1}{mn}) \).

**Remark 3**: It can be verified easily that the loss function \( \ell(u, w, z) \) 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 \( O(\frac{1}{mn} + \frac{1}{n}) \) instead of the faster rate \( O(\frac{1}{mn}) \).

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

\[
\mathbb{E}\left[ P_{W^S, U|S, S, \tilde{S}, Z}^\gamma \right] \leq \frac{\gamma}{m} + \frac{\gamma}{n}.
\] \hspace{1cm} (23)

**Remark 4**: Compared with the bound in Theorem 3, the rate we obtained using super task framework is \( O(\frac{1}{m} + \frac{1}{n}) \), which is sub-optimal. We believe this is due to the triangle inequality \( L_P - \bar{\tilde{L}} \leq |L_P - \bar{L}| + |\bar{L} - \tilde{L}| \) 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 \( \bar{L} \), 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


