Information-Theoretic Analysis of Epistemic Uncertainty in Bayesian Meta-learning

Sharu Theresa Jose   **Sangwoo Park**   **Osvaldo Simeone**

King's Communications, Information Processing and Learning (KCLIP) Lab
Department of Engineering
King's College London
London WC2R 2LS, United Kingdom
sharu.jose@kcl.ac.uk, sangwoo.park@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk

Abstract

The overall predictive uncertainty of a trained predictor can be decomposed into separate contributions due to epistemic and aleatoric uncertainty. Under a Bayesian formulation, assuming a well-specified model, the two contributions can be exactly expressed (for the log-loss) or bounded (for more general losses) in terms of information-theoretic quantities (Xu and Raginsky, 2020). This paper addresses the study of epistemic uncertainty within an information-theoretic framework in the broader setting of Bayesian meta-learning. A general hierarchical Bayesian model is assumed in which hyperparameters determine the per-task priors of the model parameters. Exact characterizations (for the log-loss) and bounds (for more general losses) are derived for the epistemic uncertainty – quantified by the minimum excess meta-risk (MEMR) – of optimal meta-learning rules. This characterization is leveraged to bring insights into the dependence of the epistemic uncertainty on the number of tasks and on the amount of per-task training data. Experiments are presented that use the proposed information-theoretic bounds, evaluated via neural mutual information estimators, to compare the performance of conventional learning and meta-learning as the number of meta-learning tasks increases.
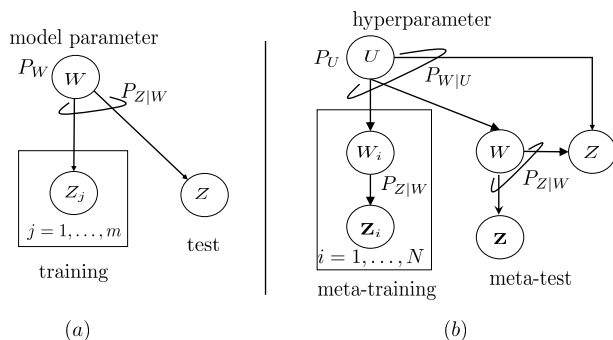
Figure 1: A graphical model representation of the joint distribution of the relevant quantities for : (a) conventional Bayesian learning; and (b) Bayesian meta-learning.

1   Introduction

Bayesian learning and epistemic uncertainty. Bayesian machine learning is well understood to have important advantages in terms of uncertainty quantification, model selection, and out-of-distribution detection (MacKay, 2003; Wilson, 2020). Bayesian learning assumes the probabilistic model illustrated in Figure 1(a), in which the training data $\mathbf{Z}$ is generated in an i.i.d. manner given a model parameter W that is considered to be a random variable endowed with a prior distribution P_W. Assuming that the model is well specified, the overall uncertainty of the optimal predictor for a test target variable Y given input X, when measured by the log-loss, is given by the conditional entropy $H(Y|X, \mathbf{Z})$. This can be decomposed as (Xu and Raginsky, 2020)

$$
\begin{aligned}
&H(Y|X, \mathbf{Z}) \\
&= \underbrace{H(Y|X, W)}_{\text{aleatoric uncertainty}} + \underbrace{I(Y; W|X, \mathbf{Z})}_{\text{epistemic uncertainty}},
\end{aligned} \quad (1)
$$

where the conditional entropy $H(Y|X, W)$ quantifies the **aleatoric uncertainty** in the prediction, while the conditional mutual information (MI) $I(Y; W|X, \mathbf{Z})$ accounts for the **epistemic uncertainty**. The aleatoric uncertainty captures the inherent randomness in the data generation process, and is independent of the amount of the available data; while the epistemic uncertainty, also known as the **minimum excess risk** (**MER**), is caused by limitations in the availability of data, and vanishes as more training data is processed. In this work, we aim at extending this decomposition, and related analysis, from conventional Bayesian learning to Bayesian meta-learning (Kim et al., 2018; Grant et al., 2018; Ravi and Beatson, 2019).

Bayesian meta-learning. In conventional Bayesian learning, the prior P_W of the model parameter is fixed a priori based on knowledge about the problem or tractability. A choice of the prior that matches the data generation mechanism can reduce the amount of data required to meet accuracy requirement. Bayesian meta-learning aims to automatically infer the prior P_W by observing a finite number N of "related" tasks, so that the predictive performance on a new, previously unseen task, in the same class can be improved (Kim et al., 2018; Grant et al., 2018; Ravi and Beatson, 2019). The shared statistical properties of a class of tasks can be modelled using the **hierarchical Bayesian model** show in Figure 1(*b*). In it, a **latent hyperpameter** U determines the prior of all tasks via the conditional distribution as $P_{W|U}$. Assuming a well-specified model, knowing the hyperparameter U hence yields the correct prior $P_{W|U}$ for the new task. The hyperparameter U itself is assumed to be random, and endowed with a **hyperprior** distribution P_U.

Contributions. This work aims at developing, analyzing, and evaluating information-theoretic characterizations of the epistemic uncertainty associated with a meta-learning predictor assuming a well-specified model. The meta-learner has access to limited labelled data from a set of related tasks, known as **meta-learning tasks**, as well as from the new task of interest, known as **meta-test task**. Intuitively, meta-learning data can help reduce the epistemic uncertainty associated with the hyperparameter U, while meta-test task data is important to further reduce the epistemic uncertainty at the level of the per-task model parameter W. With this in mind, the main contributions of this work are as follows:

1. We first develop an exact information-theoretic characterization of the overall epistemic uncertainty for an optimal Bayesian meta-learner for the **log-loss**. The characterization is given in terms of the **minimum excess meta-risk (MEMR)**, which generalize the notion of MER to Bayesian meta-learning. The

bound reveals that, under suitable assumptions, the first contribution to epistemic uncertainty – due to the hyperparameter U – scales as $O(d \log(N)/N)$, where N is the number of meta-training tasks for fixed number of per-task data samples m; while the contribution due to model parameter uncertainty scales as $O(d \log(m)/m)$.

2. We evaluate the derived information-theoretic bounds for the problem of meta-learning the priors of a Bayesian neural network (BNN) by leveraging mutual information neural estimation (MINE) (Belghazi et al., 2018; Mukherjee et al., 2020), and compare the performance of conventional learning and meta-learning as the number of meta-training tasks increses.

Additional material including a generalization to a broader class of loss functions and an information-theoretic comparison of Bayesian learning and meta-learning can be found in the supplementary file.

2   Related Work

Information-theoretic generalization analysis. The works (Russo and Zou, 2016; Xu and Raginsky, 2017) have shown that the **generalization error** of conventional learning algorithms in the frequentist setting can be upper bounded in terms of the MI $I(W; \mathbf{Z})$ between the input training set and the output model parameter. This metric captures the **sensitivity** of the learning mechanism to the input training set. Various refinements of these MI-based bounds have been studied since by Bu et al. (2019); Negrea et al. (2019); Steinke and Zakynthinou (2020), among others.

Moving from the frequentist to Bayesian learning, the recent work in Xu and Raginsky (2020) introduces an information-theoretic analysis of the MER in Bayesian learning. The MER for a general class of loss functions, including log-loss and bounded loss, is shown to be upper bounded via functions of the ratio $I(W; \mathbf{Z})/m$ where m is the number of data samples. Under appropriate regularity assumptions on the model, this upper bound is shown to vanish in the limit as $m \to \infty$.

Generalization analysis of meta-learning. Originating in the work by Schmidhuber (1987) and Thrun and Pratt (1998), meta-learning has been extensively studied in recent years both in terms of algorithm design (Finn et al., 2017; Nichol et al., 2018) and of analytical studies on the **meta-generalization error** within a frequentist setting (Pentina and Lampert, 2014; Amit and Meir, 2018; Rothfuss et al., 2020; Jose and Simeone, 2021a). While the works in Pentina and Lampert (2014); Amit and Meir (2018); Rothfuss et al. (2020) obtain high-probability PAC-Bayesian bounds on the meta-generalization error with respect to the meta-training data, reference (Jose and Simeone, 2021a)

presents information-theoretic bounds on the average meta-generalization error, thereby extending the work of Xu and Raginsky (2017) for conventional learning to meta-learning. Refinements and extensions to these bounds have been studied in Rezazadeh et al. (2020) and Jose and Simeone (2021b).

Bayesian meta-learning. Most activity on Bayesian meta-learning has modelled the hyperparameter U in Fig. 1(b) as deterministic and only captured epistemic uncertainty related to the model parameter W. The approach is akin to empirical Bayes (Grant et al., 2018), and it has been investigated in Finn et al. (2018); Kim et al. (2018); Ravi and Beatson (2019); Gordon et al. (2018b); Nguyen et al. (2020). Notably, reference (Kim et al., 2018) proposed the use of Stein Variational Gradient Descent (SVGD) (Liu and Wang, 2016) to carry out non-parametric variational inference (VI) for W. SVGD is more flexible and can be effective than standard parametric VI methods based on Gaussian distributions (Liu and Zhu, 2018). Fully Bayesian meta-learning methods were derived in Rothfuss et al. (2020); Amit and Meir (2018) from a PAC Bayes perspective by using parametric VI with Gaussian models. In this paper, we are not concerned with introducing new approximate meta-learning algorithms, but rather to evaluate the generalization performance of exact Bayesian meta-learning.

3 Problem Setting

In this section, we first review the setting studied in Xu and Raginsky (2020) for conventional Bayesian learning along with the key definition of Minimum Excess Risk (MER). Then, we generalize the framework to the Bayesian meta-learning setup introduced and analyzed in this paper. Central to our analysis is the Minimum Excess Meta-Risk (MEMR) metric, which extends the MER to meta-learning. We adopt standard notations for information-theoretic quantities such as (conditional) entropy and (conditional) MI as defined in Cover and Thomas (2006).

3.1 Conventional Bayesian Learning

In supervised learning, each data point $Z = (X, Y) \in \mathcal{Z}$ consists of a tuple of input feature vector $X \in \mathcal{X}$ and target variable $Y \in \mathcal{Y}$, which is drawn from an unknown population distribution. The learner observes a training data set $\mathbf{Z} = (Z_1, \ldots, Z_m)$, of m samples, $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, m$, that are generated i.i.d. according to the underlying unknown distribution, and uses it to predict the label of a test feature input X drawn independently from the training set $\mathbf{Z}$ from the same distribution. Considering a **parametric generative model**, we assume that the unknown population distri-

bution belongs to a model class $\mathcal{M} = \{P_{Z|w} : w \in \mathcal{W}\}$ parametrized by a model parameter w in the set $\mathcal{W}$. This implies that the model class is **well-specified** (Knoblauch et al., 2019).

In conventional Bayesian learning, the model parameter W is treated as a latent random vector, and is endowed with a prior distribution P_W. Conditioned on the model parameter W, the data samples are drawn i.i.d. from the model $P_{Z|W}$. Consequently, the joint distribution of model parameter W, training set $\mathbf{Z}$, and test sample $Z = (X, Y)$ is given as the product

$$P_{W, \mathbf{Z}, Z} = P_W \otimes P_{Z|W}^{\otimes m} \otimes P_{Z|W}, \qquad (2)$$

which factorizes according to the Bayesian network illustrated in Figure 1(a).

Let $\mathcal{A}$ denote an action space and $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ denote a loss function. The loss accrued by action $a \in \mathcal{A}$ on target variable $y \in \mathcal{Y}$ is measured by the loss function $\ell(y, a)$. Under the generative model (2), the Bayesian learning problem is to infer a **decision rule**, $\Psi_{\text{base}} : \mathcal{Z}^m \times \mathcal{X} \to \mathcal{A}$, mapping the input training data $\mathbf{Z} \in \mathcal{Z}^m$ and test input $X \in \mathcal{X}$ to an action $a \in \mathcal{A}$, that minimizes the expected loss $\mathbb{E}_{P_{Y, X, \mathbf{Z}}}[\ell(Y, \Psi_{\text{base}}(X, \mathbf{Z}))]$, where $P_{Y, X, \mathbf{Z}}$ is the marginal of (2) over Y, X, and $\mathbf{Z}$.

Definition 3.1 *(Xu and Raginsky, 2020) The* **Bayesian risk** *for a loss function* $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ *is the minimum expected loss across all possible choices of the decision rule, i.e.,*

$$R_\ell(Y|X, \mathbf{Z})$$
$$:= \min_{\Psi_{\text{base}}: \mathcal{Z}^m \times \mathcal{X} \to \mathcal{A}} \mathbb{E}_{P_{Y, X, \mathbf{Z}}} \left[\ell(Y, \Psi_{\text{base}}(\mathbf{Z}, X)) \right]. \quad (3)$$

The Bayesian risk is lower bounded by the expected loss obtained by an ideal decision rule, $\Phi_{\text{base}} : \mathcal{W} \times \mathcal{X} \to \mathcal{A}$, that has access to the true model parameter W generating the test sample $Z = (X, Y) \sim P_{Z|W}$.

Definition 3.2 *The* **genie-aided Bayesian risk** *is defined as*

$$R_\ell(Y|X, W)$$
$$:= \min_{\Phi_{\text{base}}: \mathcal{W} \times \mathcal{X} \to \mathcal{A}} \mathbb{E}_{P_{Y, X, W}} \left[\ell(Y, \Phi_{\text{base}}(W, X)) \right]. \quad (4)$$

The difference between the Bayesian risk (3) and the genie-aided risk (4) is the **minimum excess risk (MER)**,

$$\text{MER}_\ell := R_\ell(Y|X, \mathbf{Z}) - R_\ell(Y|X, W). \qquad (5)$$

The MER defined in (5) satisfies the following properties (Xu and Raginsky, 2020): (i) $\text{MER}_\ell \geq 0$; and (ii) it is non-increasing with respect to the number of

data samples m. Importantly, by (5), the minimum predictive uncertainty, $R_\ell(Y|X, \mathbf{Z})$, can be written as the sum $R_\ell(Y|X, W) + \text{MER}_\ell$ of genie-aided Bayesian risk and MER. The first term quantifies the **aleatoric uncertainty** resulting from the inherent presence of randomness in the data generation process; while the MER quantifies the **epistemic uncertainty** due to availability of insufficient data to identify the model parameter W.

3.2 Bayesian Meta-Learning

In conventional Bayesian learning, the prior distribution P_W on the model parameters is conventionally chosen based on prior knowledge about the problem. In contrast, in Bayesian meta-learning, this selection is *data-driven* and *automated*. Specifically, by observing data from a number N of tasks with shared statistical characteristics, meta-learning aims at inferring a suitable prior P_W, to be used on a new, a priori unknown task. As we detail next, the statistical relationship among different tasks is accounted for via a hierarchical Bayesian model that includes a global latent hyperparameter $U \in \mathcal{U}$ (Gordon et al., 2018a). As illustrated in Figure 1(b), the meta-learner is given data from N **meta-training tasks**. Data for each task i is drawn from the distribution $P_{Z|W=W_i}$ with the task-specific model parameter W_i. In particular, conditioned on model parameter W_i, the training data samples of each ith task, $\mathbf{Z}_i = (Z_1^i, \ldots, Z_m^i)$, are i.i.d. and drawn from the data distribution $P_{Z|W=W_i}$. The model parameter W_i of each task i is drawn from a shared prior distribution $P_{W|U}$, parameterized by a common hyperparameter U. Both the model parameter W_i and hyperparameter U are assumed to be latent random variables, with joint distribution factorizing as $P_U \otimes P_{W|U}$, with P_U denoting the **hyper-prior** distribution. The parameterized prior $P_{W|U}$ is assumed to be the same for all tasks, and the statistical relationship of the observed tasks is captured through the hyperparameter U.

The **meta-training set** $\mathbf{Z}_{1:N} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ includes the data sets from the N meta-training tasks. The goal is to use this data to reduce the expected loss measured on a **meta-test task**. The latter is a priori unknown, and is modelled as being generated by drawing an independent model parameter $W \sim P_{W|U}$ for the given hyperparameter U that is shared with the meta-training data. This model parameter underlies the generation of the **meta-test training data** $\mathbf{Z} \sim P_{Z|W}^{\otimes m}$, and an independently generated **meta-test test data** $Z = (X, Y) \sim P_{Z|W}$.

To summarize, as shown in Figure 1(b), the joint distribution of global parameter U, model parameters

$W_{1:N} = (W_1, \ldots, W_N)$ for the meta-training tasks, meta-training data set $\mathbf{Z}_{1:N}$, training data $\mathbf{Z}$ and test data Z of the meta-test task with model parameter W, is given as

$$P_{U,W_{1:N},\mathbf{Z}_{1:N},W,\mathbf{Z},Z}$$

$$= P_U \otimes \underbrace{\left(P_{W|U} \otimes P_{\mathbf{Z}|W} \right)^{\otimes N}}_{\text{meta-training}} \otimes \underbrace{P_{W|U} \otimes P_{\mathbf{Z}|W} \otimes P_{Z|W}}_{\text{meta-testing}} .$$

$$(6)$$

The **meta-learning decision rule** is defined as a mapping $\Psi_{\text{meta}} : \mathcal{Z}^{Nm} \times \mathcal{Z}^m \times \mathcal{X} \to \mathcal{A}$ from observed meta-training set $\mathbf{Z}_{1:N} \in \mathcal{Z}^{Nm}$, training set $\mathbf{Z} \in \mathcal{Z}^m$ and test feature input $X \in \mathcal{X}$ of the meta-test task to the action space $\mathcal{A}$. In words, the meta-learning rule Ψ_{meta} leverages meta-training data, along with the training data for the meta-test task, to predict the target variable Y of the test sample (X, Y) for the meta-test task. Under the meta-learning generative model in (6), the Bayesian meta-learning problem is to infer a decision rule $\Psi_{\text{meta}}(\mathbf{Z}_{1:N}, \mathbf{Z}, X)$ so as to minimize the expected loss $\mathbb{E}_{P_{\mathbf{Z}_{1:N},\mathbf{Z},X,Y}}[\ell(Y, \Psi_{\text{meta}}(\mathbf{Z}_{1:N}, \mathbf{Z}, X))]$, where $P_{\mathbf{Z}_{1:N},\mathbf{Z},X,Y}$ is the marginal of the joint distribution (6) over $(\mathbf{Z}_{1:N}, \mathbf{Z}, X, Y)$. Accordingly, we have the following definitions.

Definition 3.3 *For a given loss function* $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$*, the **Bayesian meta-risk** is the minimum expected loss across all possible choices of the meta-learning decision rule*

$$R_\ell(Y|X, \mathbf{Z}_{1:N}, \mathbf{Z})$$
$$:= \min_{\Psi_{\text{meta}}:\mathcal{Z}^{Nm} \times \mathcal{Z}^m \times \mathcal{X} \to \mathcal{A}}$$
$$\mathbb{E}_{P_{\mathbf{Z}_{1:N},\mathbf{Z},X,Y}} \left[\ell(Y, \Psi_{\text{meta}}(\mathbf{Z}_{1:N}, \mathbf{Z}, X)) \right]. \quad (7)$$

The Bayesian meta-risk reduces to the conventional Bayesian risk (3) when the per-task prior distribution $P_{W|U}$ does not depend on the hyperparameter U, and no meta-training set is observed i.e.,

$$\mathbf{Z}_{1:N} = \emptyset, \quad \text{and} \quad P_{W|U} = P_W. \quad (8)$$

The Bayesian meta-risk is lower bounded by the expected loss obtained by a genie-aided decision rule, $\Phi_{\text{meta}} : \mathcal{U} \times \mathcal{W} \times \mathcal{X} \to \mathcal{A}$ that has access to the true shared hyperparameter U and the true model parameter W of the test task.

Definition 3.4 *The **genie-aided Bayesian meta-risk** is defined as*

$$R_\ell(Y|X, W, U)$$
$$:= \min_{\Phi_{\text{meta}}:\mathcal{U} \times \mathcal{W} \times \mathcal{X} \to \mathcal{A}} \mathbb{E}_{P_{U,W,X,Y}}[\ell(Y, \Phi_{\text{meta}}(U, W, X))].$$
$$(9)$$

The difference between the Bayesian meta-risk in (7) *and the genie-aided Bayesian meta-risk in* (9) *is the **minimum excess meta risk (MEMR)**, i.e.,*

$$\text{MEMR}_\ell := R_\ell(Y|X, \mathbf{Z}_{1:N}, \mathbf{Z}) - R_\ell(Y|X, W, U). \tag{10}$$

The MEMR reduces to the MER when condition (8) holds.

4 Information-Theoretic Analysis of the MEMR

In this section, we first provide some general properties of the MEMR. Then, we analyze the MEMR with log-loss as the loss function, and obtain information-theoretic upper bounds that explicitly reveal the dependence of the MEMR on the number of meta-training tasks and per-task data samples. The detailed proofs of all the results can be found in the supplementary material.

4.1 Exact Analysis of the MEMR

Generalizing the properties of the MER reviewed in Section 3.1 and proved in Xu and Raginsky (2020), the MEMR can be shown to satisfy the following properties.

Lemma 4.1 *The minimum excess meta-risk* MEMR_ℓ *is non-negative, i.e.,* $\text{MEMR}_\ell \geq 0$ *and it is non-increasing with respect to the number of tasks, N, and to number of data samples per task, m.*

We now evaluate the MEMR explicitly in terms of information-theoretic metrics when the loss function $\ell(\cdot, \cdot)$ is the log-loss. To this end, consider the action space $\mathcal{A}$ to be the space of all probability distributions $q(\cdot)$ on $\mathcal{Y}$. We assume that all necessary measurability conditions are satisfied (Xu and Raginsky, 2020). The log-loss accrued by distribution $q(\cdot)$ on a given target y is defined as $\ell(y, q) = -\log q(y)$.

MER for conventional Bayesian learning. For reference, we first review a result from Xu and Raginsky (2020) that expresses the MER for conventional Bayesian learning in terms of a conditional MI, and bounds it as a function of a scaled MI.

Lemma 4.2 *(Xu and Raginsky, 2020) The minimum excess risk* (5) *for the log-loss satisfies*

$$\text{MER}_{\log} = I(Y; W|X, \mathbf{Z}) = H(Y|X, \mathbf{Z}) - H(Y|X, W) \tag{11}$$

$$\leq \frac{I(W; \mathbf{Z})}{m}. \tag{12}$$

In (11), the conditional entropy $H(Y|X, \mathbf{Z})$ captures the **overall predictive uncertainty** of the target Y when tested on the feature input X using the training data $\mathbf{Z}$, while the term $H(Y|X, W)$ accounts for the **aleatoric uncertainty**. The latter results from the inherent randomness in the observations, which applies even when true model parameter W is known. The difference between the two yield the conditional mutual information $I(Y; W|X, \mathbf{Z})$, which captures the **epistemic uncertainty** in predicting Y. In (12), the MER is upper bounded by the term that depends on the MI $I(W; \mathbf{Z})$ between the model parameter and the training data. This term captures the "**sensitivity**" of the trained model parameter W on the training data $\mathbf{Z}$ (see Sec. 2), in the sense that it quantifies the dependence of the trained model parameter W on $\mathbf{Z}$ (Raginsky et al., 2016).

MEMR for Bayesian meta-learning. Our first main result is the generalization of the information-theoretic characterization (11) to Bayesian meta-learning.

Proposition 4.1 *The minimum excess meta-risk* (10) *for the log-loss is given as*

$$\begin{aligned} \text{MEMR}_{\log} &= I(Y; W|X, \mathbf{Z}, \mathbf{Z}_{1:N}) \\ &= H(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}) - H(Y|X, W). \end{aligned} \tag{13}$$

The proof can be found in the supplementary materials.

In (13), the conditional entropy $H(Y|X, W)$ captures the **aleatoric uncertainty** in predicting Y, which applies even when the true model parameter W is known. In contrast, the term $H(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N})$ captures the average predictive uncertainty of the optimal meta-learning decision rule. As such, the conditional MI $I(Y; W|X, \mathbf{Z}, \mathbf{Z}_{1:N})$ captures the **epistemic uncertainty**. This uncertainty results from the availability of limited meta-training tasks and meta-test training data, which causes the true model parameter W and true hyperparameter U to be inaccurately estimated.

Dependence of MEMR on N and m. We now decouple the contributions of hyperparameter-level and per-task-level uncertainties by developing an information-theoretic upper bound on the MEMR (13). The bound will be used to relate the MEMR to the number of meta-training tasks, N, and to the number of samples of the meta-test training set, m.

Theorem 4.2 *The following upper bounds on the*

MEMR hold under the log-loss,

$$\mathrm{MEMR}_{\log} \leq \frac{I(W, U; \mathbf{Z}|\mathbf{Z}_{1:N})}{m} \tag{14}$$

$$\leq \frac{I(U; \mathbf{Z}_{1:N})}{Nm} + \frac{I(W; \mathbf{Z}|U)}{m} =: \mathrm{MEMR}_{\log}^{\mathrm{UB}}, \tag{15}$$

where the inequality (15) *holds for* $N \geq 1$.

The upper bound $\mathrm{MEMR}_{\log}^{\mathrm{UB}}$ (15) on the MEMR for the log-loss is the sum of two contributions. The first term captures the **sensitivity** of the hyperparameter U on the meta-training set $\mathbf{Z}_{1:N}$. The second term corresponds to the average **sensitivity** of the model parameter W on the meta-test task training data $\mathbf{Z}$ assuming that the hyperparameter U is known. The result (15) shows that the epistemic uncertainty $I(Y; W|X, \mathbf{Z}, \mathbf{Z}_{1:N})$, which applies to the domain of the target variable Y, is upper bounded by the sum of two contributions that pertain the uncertainty levels in the spaces of hyperparameter and model parameter, respectively.

The additive dependence of the upper bound (15) on two mutual information terms, one at the hyperparameter level and other at the per-task model parameter level, bears resemblance to the information-theoretic bounds on the generalization error of frequentist meta-learning problems obtained in Jose and Simeone (2021a,b). However, the two types of bounds are conceptually different. In fact, the generalization error bounds in Jose and Simeone (2021a,b) quantify the error in approximating the meta-population loss with an empirical meta-training loss for an arbitrary stochastic learning algorithm. In contrast, the MEMR metric captures the excess prediction risk obtained by the posterior distribution under the assumption of a well-specified model. As a result, while the information-theoretic upper bounds in Jose and Simeone (2021a,b) can be used to re-derive the regularized training loss objectives in Rothfuss et al. (2020), the upper bound (15) cannot play this role.

A generalization of Theorem 4.2 to any loss function and a **comparison between meta-learning and conventional learning** in terms of predictive accuracy is available in the supplementary materials.

Asymptotic analysis of the MEMR. The first term in (15) is a function of (N, m) and the second is of m, obscuring the scaling of the MEMR with (N, m). To investigate this point, we now study the asymptotic behavior of the above two terms in (15).

Lemma 4.3 *Let* $W \in \mathcal{W}$ *and* $U \in \mathcal{U}$ *be d-dimensional vectors taking values in compact subsets* $\mathcal{W}, \mathcal{U} \subset \mathbb{R}^d$ *re-*

spectively. Assume that the data distribution $P_{Z|W}(\cdot|w)$ *is smooth in* $w \in \mathcal{W}$, *and that the distribution* $P_{\mathbf{Z}|U}(\cdot|u)$ *is smooth in* $u \in \mathcal{U}$. *Then, under additional technical conditions (included in supplementary material), we have that for fixed* m, *as* $N \to \infty$,

$$I(U; \mathbf{Z}_{1:N}) = \frac{d}{2} \log\left(\frac{N}{2\pi e}\right) + H(U)$$
$$+ \mathbb{E}_{P_U}\left[\log |J_{\mathbf{Z}|U}(U)|\right] + o(1), \tag{16}$$

and as $m \to \infty$, *we have*

$$I(W; \mathbf{Z}|U) = \frac{d}{2} \log\left(\frac{m}{2\pi e}\right) + H(W|U)$$
$$+ \mathbb{E}_{P_{W,U}}\left[\log |J_{Z|W}(W)|\right] + o(1), \tag{17}$$

where $H(\cdot)$ *denotes the differential entropy of the argument random variable and* $J_{A|B}(B)$ *is the Fisher information matrix (FIM) about* B *contained in* A *with respect to conditional distribution* $P_{A|B}$, *whose* (j, k)*th entry is*

$$[J_{A|B}(B)]_{j,k} = \left[\frac{\partial^2}{\partial B'_j \partial B'_k} D_{\mathrm{KL}}(P_{A|B} || P_{A|B'})\Big|_{B'=B}\right]. \tag{18}$$

Using Lemma 4.3, it can be seen that the epistemic uncertainty at the hyperparameter level, quantified by the sensitivity $I(U; \mathbf{Z}_{1:N})/Nm$, scales as $O(d \log(N)/N)$ for fixed m; while the epistemic uncertainty at the per-task level, accounted for by the sensitivity $I(W; \mathbf{Z}|U)/m$, scales as $O(d \log(m)/m)$. Therefore, if $N \to \infty$, and m is finite, the MEMR depends solely on the per-task epistemic uncertainty term $I(W; \mathbf{Z}|U)/m$ in (15). That the MEMR does not vanish as $N \to \infty$ is a consequence of the fact that the meta-test task is a priori unknown. As a result, even an infinite amount of meta-training data does not resolve the epistemic uncertainty about the meta-test task (Gordon et al., 2018a; Jose and Simeone, 2021b).

4.2 Note on the Optimality of Bi-Level Meta-Learning

The meta-decision rule maps directly the observed meta-training set $\mathbf{Z}_{1:N}$, the training data $\mathbf{Z}$ of the meta-test task, and test feature input X into a predictive distribution $q(y|X, \mathbf{Z}, \mathbf{Z}_{1:N})$ on the space $\mathcal{Y}$ of target labels. By standard results in Bayesian inference (see e.g., Bishop (2006)), the optimal predictive distribution is hence given by the posterior $P_{Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}}$. To conclude this section and prepare for the next, we recall here that the joint distribution (6) can be factorized as

$$P_{Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}} = \mathbb{E}_{P_{U|\mathbf{Z}, \mathbf{Z}_{1:N}} P_{W|\mathbf{Z}, X, U}}[P_{Y|X, W}]. \tag{19}$$

This factorization reveals that the optimal meta-decision rule can be implemented as a **two-step procedure**, whereby one first obtains the **hyperposterior** distribution $P_{U|\mathbf{Z},\mathbf{Z}_{1:N}}$ using meta-training data $\mathbf{Z}_{1:N}$ and the meta-test task training data $\mathbf{Z}$; and then evaluates the **per-task posterior** distribution $P_{W|\mathbf{Z},X,U}$ to evaluate the ensemble predictor (19).

4.3 Impact of Model Misspecification

The results discussed so far rely on the assumption that the model is well specified, in the sense that the unknown population distribution belongs to a model class $\mathcal{M} = \{P_{Z|w} : w \in \mathcal{W}\}$. This assumption is violated if the true data generating distribution does not belong to the model class. In this subsection, we extend our results to this scenario.

To account for model misspecification, we assume that a task environment distribution Q_T defines a distribution over tasks; and that each task $T_i \sim Q_T$, for $i = 1, \ldots, N$, corresponds to a data distribution $Q_{Z|T_i}$, with its training data generated as $\mathbf{Z}_i \sim Q_{Z|T_i}^{\otimes m}$. The meta-training tasks $T_{1:N}$, the meta-training set $\mathbf{Z}_{1:N}$, the meta-test task T, the test task training data $\mathbf{Z}$, and the test data Z are jointly distributed as

$$
\begin{aligned}
&Q_{T_{1:N},\mathbf{Z}_{1:N},T,\mathbf{Z},Z} \\
&= \left(Q_T \otimes Q_{Z|T}^{\otimes m}\right)^{\otimes N} \otimes Q_T \otimes Q_{Z|T}^{\otimes m} \otimes Q_{Z|T}. \quad (20)
\end{aligned}
$$

Consequently, the observed meta-training data, meta-test training and test data are generated according to the marginal distribution $Q_{Y,X,\mathbf{Z},\mathbf{Z}_{1:N}}$. Crucially, a learner that assumes this distribution computes the optimal ensemble predictor that minimizes the MEMR as $Q_{Y|X,\mathbf{Z},\mathbf{Z}_{1:N}}$, whereas a learner assuming the marginal $P_{Y,X,\mathbf{Z},\mathbf{Z}_{1:N}}$ under the joint distribution (2) computes it as $P_{Y|X,\mathbf{Z},\mathbf{Z}_{1:N}}$. The model is misspecified in the sense that the marginal distributions $Q_{Y,X,\mathbf{Z},\mathbf{Z}_{1:N}}$ and $P_{Y,X,\mathbf{Z},\mathbf{Z}_{1:N}}$ differ.

We now analyze the impact of model misspecification on the Bayesian meta-risk (7) under the log-loss. For a well-specified (WS) model, the optimal ensemble predictor $P_{Y|X,\mathbf{Z},\mathbf{Z}_{1:N}}$ (as in (19)) results in the following meta-risk

$$
R_{\log}^{WS}(Y|X,\mathbf{Z}_{1:N},\mathbf{Z}) = \mathbb{E}_{P_{\mathbf{Z}_{1:N},\mathbf{z},x,Y}}[-\log P_{Y|X,\mathbf{Z},\mathbf{Z}_{1:N}}]. \quad (21)
$$

When the model is misspecified (MS), the ensemble predictor $P_{Y|X,\mathbf{Z},\mathbf{Z}_{1:N}}$ yields the following risk

$$
R_{\log}^{MS}(Y|X,\mathbf{Z}_{1:N},\mathbf{Z}) = \mathbb{E}_{Q_{\mathbf{Z}_{1:N},\mathbf{z},x,Y}}[-\log P_{Y|X,\mathbf{Z},\mathbf{Z}_{1:N}}]. \quad (22)
$$

The average excess risk due to model misspecification can be then quantified as

$$
\Delta(Q,P) = R_{\log}^{MS}(Y|X,\mathbf{Z}_{1:N},\mathbf{Z}) - R_{\log}^{WS}(Y|X,\mathbf{Z}_{1:N},\mathbf{Z}). \quad (23)
$$

The overall minimum excess meta-risk of the ensemble predictor $P_{Y|X,\mathbf{Z},\mathbf{Z}_{1:N}}$ under model misspecification is then given as

$$
\mathrm{MEMR}_{\log}^{MS} = R_{\log}^{MS}(Y|X,\mathbf{Z}_{1:N},\mathbf{Z}) - H(Y|X,W). \quad (24)
$$

This can be decomposed as the sum of uncertainty due to model-misspecification and the epistemic uncertainty in prediction, i.e.,

$$
\mathrm{MEMR}_{\log}^{MS} = \Delta(Q,P) + \mathrm{MEMR}_{\log}, \quad (25)
$$

where the second term, $\mathrm{MEMR}_{\log}$, was studied in the previous sections.

5 Examples

In this section, we describe a regression example based on Bayesian neural networks (Neal, 2012). Additional experiments can be found in supplementary material [1].

In order to ensure that the model is well specified – an underlying assumption of the analysis of generalization for Bayesian learning as studied in this work – we consider a synthetic data set generated as follows. We focus on a regression problem in which the target variable is distributed as $Y = f_W(X) + \xi$, with input $X \sim \mathcal{N}(0,1)$; regression function $f_W(\cdot)$ specified by a neural network parameter vector W; and observation noise $\xi \sim \mathcal{N}(0, 0.1^2)$. The neural network consists of one hidden layer with ReLU activation in the hidden layer and a linear activation in the last layer. The prior distribution of model parameter W is determined by hyperparameter U as $P_{W|U} = \mathcal{N}(U, \sigma_w^2 \mathbb{1})$ with fixed standard deviation σ_w and identity matrix $\mathbb{1}$ with the same dimension as vectors W and U. Lastly, the hyperprior distribution for hyperparameter U is defined as $P_U = \mathcal{N}(U|\mathbf{0}, \mathbb{1})$ with an all-zero mean vector $\mathbf{0}$.

Fig. 2 compares the MEMR under the log-loss in (13) (top) and the information-theoretic upper bound $\mathrm{MEMR}_{\log}^{UB}$ in (15) (bottom) as a function of the increasing number N of tasks for fixed $m = 1$ (left) and of the increasing number of per-task samples m for fixed $N = 1$ (right). Note that when $N = 0$, MEMR corresponds to the minimum excess risk for conventional learning. We use conditional MINE (C-MINE) (Mukherjee et al., 2020) along with Smoothed Mutual
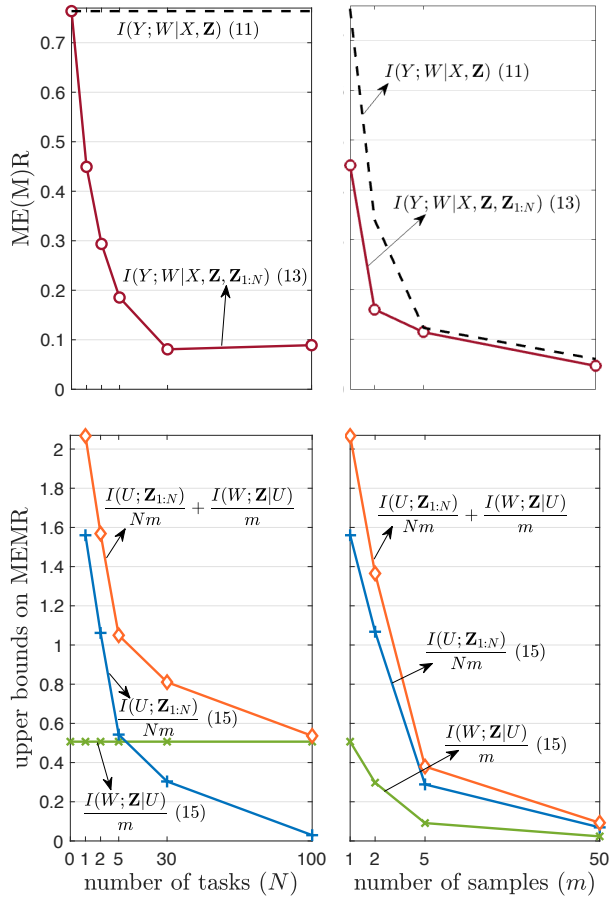
Figure 2: MEMR (top) and information-theoretic upper bounds (bottom); (left) as a function of number of meta-training tasks (N) with fixed number of samples $m = 1$ (right) as a function of number of per-task samples (m) with fixed number of meta-training tasks $N = 1$. Experiments are evaluated with prior variance $\sigma_w^2 = 0.01$ on the model parameters.

Information Lower-bound Estimator (SMILE) (Song and Ermon, 2020) to estimate the mutual information terms. Details for the experiment can be found in supplementary material.

The top part of the figure demonstrates the advantages of meta-learning over conventional learning as the number N of meta-training tasks increases, and it also highlights the different dependence of the MEMR on N and m. In particular, the MEMR does not vanish as N grows larger, whereas having more per-task data, i.e., increasing m, yields a vanishing MEMR. This is due to the fact that, even when one has access to infinitely many meta-training tasks, there is generally still some amount of unresolved uncertainty about the new meta-test task $\mathbf{Z}$ (see Lemma 4.3).

The bottom panels show the the information-theoretic

upper bound $\text{MEMR}_{\log}^{\text{UB}}$ in (15), which separates the contributions to the MEMR due to epistemic uncertainty at the levels of hyperparameters and per-task model parameters. The bound, while numerically loose (see, e.g., Hellström and Durisi (2020) and Wang et al. (2021) for similar results), reproduce well the dependence of the MEMR on N and m. Furthermore, the decomposition into the separate hyparparameter-level and model parameter-level contributions to epistemic uncertainty helps explain the non-vanishing behavior of the MEMR when N, as opposed to m, increases. While the hyperparameter-level sensitivity term decreases and vanishes asymptotically with N, the model parameter sensitivity term, which captures the uncertainty of the model parameter, is not influenced by N and remains constant as N varies. The non-vanishing MEMR can be thus attributed to the residual epistemic uncertainty about the newly encountered meta-test task at the level of model parameters.

The analysis of the two contributions is also useful to assess the relative merits of increasing m or N. The left-bottom panel of Fig. 2 shows, for instance, that as N increases, the contribution due to hyperparameter-level uncertainty becomes less relevant than that of model parameter-level uncertainty. In this regime, further increases in N have limited impact, and is generally preferable to increase m (not shown).

In Figure 3, we increase the prior variance σ_w^2 for the model parameters from $\sigma_w^2 = 0.01$, assumed in the previous figure, to $\sigma_w^2 = 1$. Intuitively, this change affects the amount of information that can be extracted from the hyperparameters U on the model parameters and hence on the target variables. Accordingly, the model parameter sensitivity is seen to dominate the hyperparameter sensitivity, and meta-learning is observed to yield marginal benefits over conventional learning (i.e. when $N = 0$).

6 Conclusion

This paper studies epistemic uncertainty for Bayesian meta-learning from an information-theoretic perspective. We show that this uncertainty can be evaluated exactly (for log-loss) or bounded (for general loss functions) using a conditional MI $I(Y; W | X, \mathbf{Z}, \mathbf{Z}_{1:N})$ involving model parameter, hyperparameter, and data. A novel information-theoretic upper bound on this term is also presented that explicitly shows the dependence of epistemic uncertainty on the number of meta-training tasks, N, and per-task samples, m.

The information-theoretic analysis conducted in this work assume optimal Bayesian inference. Future work may try to alleviate this limitations by considering the impact of approximations due to variational inference.
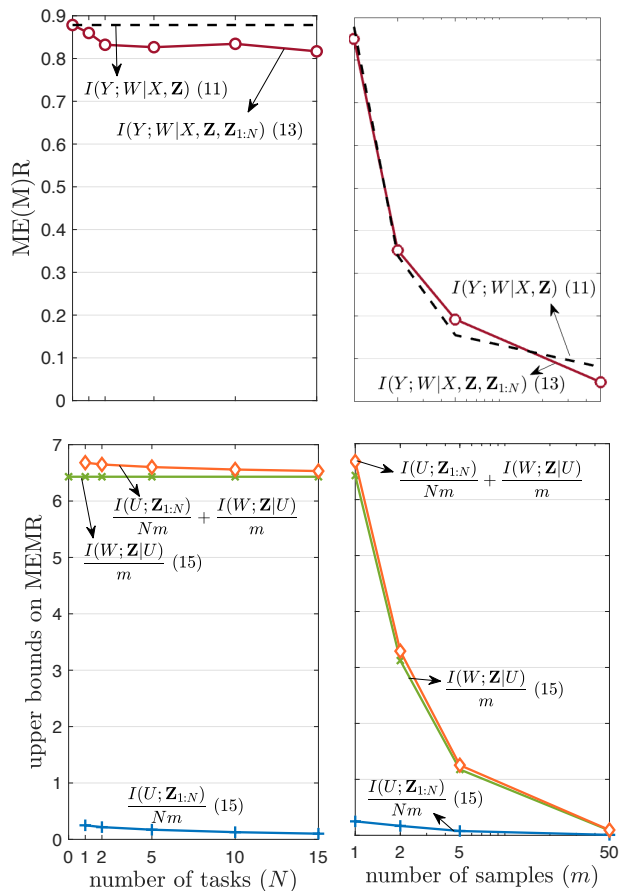
Figure 3: MEMR (top) and information-theoretic upper bounds (bottom); (left) as a function of number of meta-training tasks (N) with fixed number of samples $m = 1$ (right) as a function of number of per-task samples (m) with fixed number of meta-training tasks $N = 1$. Experiments are evaluated with prior variance $\sigma_w^2 = 1$ on the model parameters.

As a final note, as this paper addresses purely theoretical analysis, the results presented have no significant societal impact.

Acknowledgements

References

Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *Proc. of Int. Conf. Machine Learning (ICML)*, pages 205–214, Jul 2018.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, March 2000.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, pages 587–591, July 2019.

Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory, 2nd Edition,*. Wiley-Interscience, July 2006. ISBN 0471241954.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proc. of Int. Conf. Machine Learning-Volume 70*, pages 1126–1135, Aug. 2017.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. *arXiv preprint arXiv:1806.02817*, 2018.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E. Turner. Decision-theoretic meta-learning: Versatile and efficient amortization of few-shot learning. *CoRR*, abs/1805.09921, 2018a. URL http://arxiv.org/abs/1805.09921.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. Meta-learning probabilistic inference for prediction. *arXiv preprint arXiv:1805.09921*, 2018b.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

Fredrik Hellström and Giuseppe Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. *arXiv preprint arXiv:2010.11552*, 2020.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1):126, 2021a.

Sharu Theresa Jose and Osvaldo Simeone. An information-theoretic analysis of the impact of task similarity on meta-learning. *arXiv preprint arXiv:2101.08390*, 2021b.

Taesup Kim, Jaesik Yoon, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. *arXiv preprint arXiv:1806.03836*, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized Variational Inference. *arXiv preprint arXiv:1904.02063*, 2019.

Chang Liu and Jun Zhu. Riemannian stein variational gradient descent for Bayesian inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016.

David JC MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

Andreas Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6:967–994, Jun 2005.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pages 1083–1093. PMLR, 2020.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Proc. of Adv. Neural Inf. Processing Sys. (NIPS)*, pages 11013–11023, Dec 2019.

Cuong Nguyen, Thanh-Toan Do, and Gustavo Carneiro. Uncertainty in model-agnostic meta-learning using variational inference. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3100, 2020.

Alex Nichol, Joshua Achiam, and John Schulman. On First-Order Meta-Learning Algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Anastasia Pentina and Christoph Lampert. A PAC-Bayesian bound for lifelong learning. In *Proc. of Int. Conf. on Machine Learning (ICML)*, pages 991–999, June 2014.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *Proc. of IEEE Inf. Theory Workshop (ITW)*, pages 26–30, Sep. 2016.

Sachin Ravi and Alex Beatson. Amortized Bayesian meta-learning. In *ICLR (Poster)*, 2019.

Arezou Rezazadeh, Sharu Theresa Jose, Giuseppe Durisi, and Osvaldo Simeone. Conditional mutual information-based generalization bound for meta learning. *arXiv preprint arXiv:2010.10886*, 2020.

Jonas Rothfuss, Vincent Fortuin, and Andreas Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. *arXiv preprint arXiv:2002.05551*, 2020.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proc. of Artificial Intelligence and Statistics (AISTATS)*, pages 1232–1240, May 2016.

Jürgen Schmidhuber. *Evolutionary Principles in Self-Referential Learning, or On Learning How to Learn: The Meta-meta-... Hook.* PhD thesis, Technische Universität München, 1987.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. *arXiv preprint arXiv:2001.09122*, 2020.

Sebastian Thrun and Lorien Pratt. Learning to Learn: Introduction and Overview. In *Learning to Learn*, pages 3–17. Springer, 1998.

Hao Wang, Yizhe Huang, Rui Gao, and Flavio P Calmon. Learning while dissipating information: Understanding the generalization capability of sgld. *arXiv preprint arXiv:2102.02976*, 2021.

Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Proc. of Adv. in Neural Inf. Processing Sys. (NIPS)*, pages 2524–2533, Dec. 2017.

Aolin Xu and Maxim Raginsky. Minimum excess risk in Bayesian learning. *arXiv preprint arXiv:2012.14868*, 2020.

Sharu Theresa Jose, Sangwoo Park, Osvaldo Simeone

Information-Theoretic Analysis of Epistemic Uncertainty in Bayesian Meta-learning: Supplementary Material

A Proof of Lemma 4.1

The properties are a direct consequence of the data processing inequality satisfied by the Bayesian risk (Xu and Raginsky, 2020). This states that, given jointly distributed random variables A, B and C, if the Markov chain $A - B - C$ holds, we have the inequality $R_\ell(C|A) \geq R_\ell(C|B)$. Noting that $(\mathbf{Z}_{1:N}, \mathbf{Z}, X) - (U, W, X) - Y$ forms a Markov chain, the non-negativity of the MEMR follows from the data processing inequality. Through the same argument, it can be proved that MEMR is non-increasing with the number of tasks and per-task data samples.

B Proof of Proposition 4.1

Under log-loss, the Bayesian meta-risk is given by,

$$R_{\log}(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}) = \min_{q(\cdot)} \mathbb{E}_{P_{\mathbf{Z}_{1:N}, \mathbf{z}, X} P_{Y|X, \mathbf{z}, \mathbf{z}_{1:N}}} [-\log q(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N})]. \tag{26}$$

From standard results in information theory (Cover and Thomas, 2006), it can be verified that the optimal meta-decision rule $q(\cdot)$ that minimizes the Bayesian meta-risk corresponds to the posterior predictive distribution $P_{Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}}$, whereby we have

$$R_{\log}(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}) = \mathbb{E}_{P_{\mathbf{Z}_{1:N}, \mathbf{z}, X} P_{Y|X, \mathbf{z}, \mathbf{z}_{1:N}}} [-\log P_{Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}}] = H(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}). \tag{27}$$

Similarly, it can be shown that

$$\begin{aligned} R_{\log}(Y|X, W, U) &= \mathbb{E}_{P_{U,W,X} P_{Y|X,U,W}} [-\log P_{Y|X,U,W}] \\ &= H(Y|X, W, U) = H(Y|X, W), \end{aligned} \tag{28}$$

where the last equality follows since $U - W - Z$ forms a Markov chain whereby $P_{Z|W,U} = P_{Z|W}$. Together, we then have that

$$\begin{aligned} \text{MEMR}_{\log} &= H(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}) - H(Y|X, W) \\ &\overset{(a)}{=} H(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}) - H(Y|X, W, \mathbf{Z}, \mathbf{Z}_{1:N}) \\ &= I(Y; W|X, \mathbf{Z}, \mathbf{Z}_{1:N}), \end{aligned} \tag{29} \tag{30}$$

where the equality in (a) follows since conditioned on test input X and model parameter W, the test output Y is independent of $(\mathbf{Z}, \mathbf{Z}_{1:N})$.

C   Proof of Theorem 4.2

To obtain the required bound on MEMR, we note that the following set of relations hold.

$$\text{MEMR}_{\log} = I(W; Y | X, \mathbf{Z}, \mathbf{Z}_{1:N})$$

$$\leq I(W; Z | \mathbf{Z}, \mathbf{Z}_{1:N}) \tag{31}$$

$$\leq \frac{I(W; \mathbf{Z} | \mathbf{Z}_{1:N})}{m} \tag{32}$$

$$= \frac{I(W, U; \mathbf{Z} | \mathbf{Z}_{1:N})}{m} \tag{33}$$

$$= \frac{I(W; \mathbf{Z} | U)}{m} + \frac{I(U; \mathbf{Z} | \mathbf{Z}_{1:N})}{m} \tag{34}$$

$$\leq \frac{I(W; \mathbf{Z} | U)}{m} + \frac{I(U; \mathbf{Z}_{1:N})}{Nm}. \tag{35}$$

Here, (31) follows from the chain rule of mutual information. To prove inequality (32), we use the technique of (Xu and Raginsky, 2020, Proof of Thm. 2) which we now detail here. Towards this, we define $Z^j = (Z_1, \ldots, Z_j)$ and note that $I(W; \mathbf{Z} | \mathbf{Z}_{1:N}) = \sum_{j=1}^{m} I(W; Z_j | Z^{j-1}, \mathbf{Z}_{1:N})$. We then have the following set of relations

$$I(W; Z_{j+1} | Z^j, \mathbf{Z}_{1:N})$$

$$= H(Z_{j+1} | Z^j, \mathbf{Z}_{1:N}) - H(Z_{j+1} | Z^j, W, \mathbf{Z}_{1:N})$$

$$\overset{(a)}{=} H(Z_{j+2} | Z^j, \mathbf{Z}_{1:N}) - H(Z_{j+1} | W)$$

$$\overset{(b)}{\geq} H(Z_{j+2} | Z^{j+1}, \mathbf{Z}_{1:N}) - H(Z_{j+2} | W)$$

$$= H(Z_{j+2} | Z^{j+1}, \mathbf{Z}_{1:N}) - H(Z_{j+2} | \mathbf{Z}_{1:N}, W, Z^{j+1})$$

$$= I(W; Z_{j+2} | \mathbf{Z}_{1:N}, Z^{j+1}), \tag{36}$$

where (a) follows since $(Z_{j+1}, Z^j, \mathbf{Z}_{1:N}) \triangleq (Z_{j+2}, Z^j, \mathbf{Z}_{1:N})$ in distribution and that Z_{j+1} is conditionally independent of $(Z^j, \mathbf{Z}_{1:N})$ given W, and (b) follows since conditioning reduces entropy, and that $(Z_{j+1}, W) \triangleq (Z_{j+2}, W)$ in distribution. Consequently, we get the inequality

$$I(W; \mathbf{Z} | \mathbf{Z}_{1:N}) = \sum_{j=1}^{m} I(W; Z_j | Z^{j-1}, \mathbf{Z}_{1:N})$$

$$\geq m I(W; Z_{m+1} | Z^m, \mathbf{Z}_{1:N})$$

$$= m I(W; Z | \mathbf{Z}, \mathbf{Z}_{1:N}), \tag{37}$$

whereby $I(W; Z | \mathbf{Z}, \mathbf{Z}_{1:N}) \leq I(W; \mathbf{Z} | \mathbf{Z}_{1:N}) / m$.

The equality in (33) follows since $I(U; \mathbf{Z} | W, \mathbf{Z}_{1:N}) = 0$, which results from $(U, \mathbf{Z}_{1:N}) - W - \mathbf{Z}$ forming a Markov chain, whereby $I(W, U; \mathbf{Z} | \mathbf{Z}_{1:N}) = I(W; \mathbf{Z} | \mathbf{Z}_{1:N})$. Finally to see (35), we follow similar steps as in the proof of (32). Denoting $\mathbf{Z}^k = (\mathbf{Z}_1, \ldots, \mathbf{Z}_k)$, we have the mutual information $I(U; \mathbf{Z}_{1:N}) = \sum_{k=1}^{N} I(U; \mathbf{Z}_k | \mathbf{Z}^{k-1})$, each individual component of which can be written as

$$I(U; \mathbf{Z}_{k+1} | \mathbf{Z}^k)$$

$$= H(\mathbf{Z}_{k+1} | \mathbf{Z}^k) - H(\mathbf{Z}_{k+1} | U, \mathbf{Z}^k)$$

$$\overset{(a)}{=} H(\mathbf{Z}_{k+2} | \mathbf{Z}^k) - H(\mathbf{Z}_{k+1} | U) \tag{38}$$

$$\overset{(b)}{\geq} H(\mathbf{Z}_{k+2} | \mathbf{Z}^{k+1}) - H(\mathbf{Z}_{k+2} | U)$$

$$= H(\mathbf{Z}_{k+2} | \mathbf{Z}^{k+1}) - H(\mathbf{Z}_{k+2} | U, \mathbf{Z}^{k+1})$$

$$= I(\mathbf{Z}_{k+2}; U | \mathbf{Z}^{k+1}). \tag{39}$$

Here, the equality in (a) follows since $(\mathbf{Z}_{k+1}, \mathbf{Z}^k) \triangleq (\mathbf{Z}_{k+2}, \mathbf{Z}^k)$ in distribution, and that $\mathbf{Z}_{k+1}$ is conditionally independent of $\mathbf{Z}^k$ given U, and (b) follows since conditioning reduces entropy, and that $(\mathbf{Z}_{k+1}, U) \triangleq (\mathbf{Z}_{k+2}, U)$ in distribution. Consequently, we have that the mutual information $I(U; \mathbf{Z}_{1:N}) \geq N I(U; \mathbf{Z} | \mathbf{Z}_{1:N})$, which results in the inequality in (35).

D Meta-Learning vs Conventional Learning

One of the advantages of the Bayesian viewpoint on meta-learning is that one can obtain general information-theoretic conclusions about the performance comparison of meta-learning and conventional learning. This is in contrast to the frequentist analyses that focus on the meta-generalization error (Baxter, 2000; Maurer, 2005; Pentina and Lampert, 2014), making it difficult to draw general conclusions on this comparison.

To start, it is easy to see that under the assumption $\mathbf{Z}_{1:N} = \emptyset$ and $P_{W|U} = P_W$, the MEMR (defined in (10) in the main text) reduces to the MER (equation (5) in the main text) for conventional Bayesian learning. In fact, we have $\mathrm{MEMR}_{\log} = I(W; Y|X, \mathbf{Z}) = \mathrm{MER}_{\log}$.

Generalizing this observation, the following proposition quantifies the gains of meta-learning with respect to conventional learning under log-loss.

Proposition D.1 *Under log-loss, the* $\mathrm{MEMR}_{\log}$ *for meta-learning and the* $\mathrm{MER}_{\log}$ *for conventional learning are related as*

$$\mathrm{MER}_{\log} - \mathrm{MEMR}_{\log} = I(\mathbf{Z}_{1:N}; Y|X, \mathbf{Z}) \geq 0. \tag{40}$$

Proof: The relation in (40) is obtained as follows.

$$\mathrm{MER}_{\log} - \mathrm{MEMR}_{\log} = I(Y; W|X, \mathbf{Z}) - I(Y; W|X, \mathbf{Z}, \mathbf{Z}_{1:N}) \tag{41}$$
$$= H(Y|X, \mathbf{Z}) - H(Y|X, \mathbf{Z}, \mathbf{Z}_{1:N}) \tag{42}$$
$$= I(Y; \mathbf{Z}_{1:N}|X, \mathbf{Z}). \tag{43}$$

∎

Proposition D.1 shows that under the log-loss, meta-learning yields a lower minimum excess risk than conventional learning. The gain in minimum excess risk is quantified by the conditional MI $I(\mathbf{Z}_{1:N}; Y|X, \mathbf{Z})$, which grows as the meta-training set $\mathbf{Z}_{1:N}$ becomes more informative about the meta-test target variable Y beyond the information already available in the meta-test training set $\mathbf{Z}$ and input X.

E Assumptions for the Convergence Rates of Lemma 4.3

In this section, we specialize the assumptions required for the convergence rate of $I(W; \mathbf{Z})$ for Bayesian learning in Clarke and Barron (1994) to the case of Bayesian meta-learning, where we have two MI terms $I(U; \mathbf{Z}_{1:N})$ and $I(W; \mathbf{Z}|U)$. We first list the assumptions required for the convergence of $I(U; \mathbf{Z}_{1:N})$, and explain how these extend to $I(W; \mathbf{Z}|U)$.

Assumption E.1 *The following assumptions must be satisfied for ensuring the convergence of the mutual information term* $I(U; \mathbf{Z}_{1:N})$ *in Lemma 4.3.*

1. *Let* $U \in \mathcal{U} \subset \mathbb{R}^d$, *and that the density* $P_{\mathbf{Z}|U}$ *exists with respect to Lebesgue measure. Moreover,* $\mathcal{U}$ *has a non-void interior and its boundary has d-dimensional Lebesgue measure 0.*

2. *The density* $P_{\mathbf{Z}|U}(\mathbf{Z}|u)$ *is twice continuously differentiable in u for almost every* $\mathbf{Z}$ *and there exists* $\delta(u)$ *so that for each* $j, k = 1, \ldots, d$,

$$f(u) = \mathbb{E}_{P_{\mathbf{Z}|U=u}} \left[\sup_{u': ||u'-u|| < \delta(u)} \left| \frac{\partial^2}{\partial u'_j \partial u'_k} \log P_{\mathbf{Z}|U}(\mathbf{Z}|u') \right|^2 \right] \tag{44}$$

   is finite and continuous.

3. *For* $j = 1, \ldots, d$,

$$\mathbb{E}_{P_{\mathbf{Z}|U=u}} \left[\left| \frac{\partial}{\partial u_j} \log P_{\mathbf{Z}|U}(\mathbf{Z}|u) \right|^{2+\xi} \right] \tag{45}$$

   is finite and continuous, as a function of u, for some $\xi > 0$.

4. *Fisher information matrix (FIM) and second derivative of relative entropy are equal i.e. for matrices,*

$$[I_{\mathbf{Z}|U}(u)]_{j,k} = \mathbb{E}\left[\frac{\partial}{\partial u_j} \log P_{\mathbf{Z}|U}(\mathbf{Z}|u) \frac{\partial}{\partial u_k} \log P_{\mathbf{Z}|U}(\mathbf{Z}|u)\right] \tag{46}$$

and

$$[J_{\mathbf{Z}|U}(u)]_{j,k} = \left[\frac{\partial^2}{\partial u_j' \partial u_k'} D_{\mathrm{KL}}(P_{\mathbf{Z}|u}||P_{\mathbf{Z}|u'})\Big|_{u'=u}\right], \tag{47}$$

we have $I_{\mathbf{Z}|U}(u) = J_{\mathbf{Z}|U}(u)$ and that the matrix $I_{\mathbf{Z}|U}(u)$ is assumed to be positive definite.

5. *For $u \neq u'$, we have $P_{\mathbf{Z}|U=u} \neq P_{\mathbf{Z}|U=u'}$.*

6. *The hyperprior P_U is assumed continuous and is supported on a compact subset in the interior of $\mathcal{U}$.*

Under Assumption E.1, Theorem 1 of Clarke and Barron (1994) then yields the required asymptotic of the MI $I(U; \mathbf{Z}_{1:N})$ in Lemma 4.3.

To analyze the asymptotic of the MI $I(W; \mathbf{Z}|U)$, we note that $I(W; \mathbf{Z}|U) = \mathbb{E}_{P_U}[I(W; \mathbf{Z}|U = u)]$. Consequently, we specialize Assumption E.1 to ensure convergence of $I(W; \mathbf{Z}|U = u)$ for each $u \in \mathcal{U}$. This can be done by replacing the distribution $P_{\mathbf{Z}|U}$ with $P_{Z|W}$, the hyperprior P_U by the prior $P_{W|U=u}$ for each $u \in \mathcal{U}$, such that the resulting assumptions hold at the level of model parameter. Subsequently, Theorem 1 of Clarke and Barron (1994) ensures that as $m \to \infty$,

$$I(W; \mathbf{Z}|U = u) = \frac{d}{2}\log\left(\frac{m}{2\pi e}\right) + H(W|U = u) + \mathbb{E}_{P_{W|U=u}}\left[\log|J_{Z|W}(W)|\right] + o(1). \tag{48}$$

Taking expectation of (48) with respect to the hyperprior P_U, then yields the asymptotic behaviour of $I(W; \mathbf{Z}|U)$ in Lemma 4.3.

F Information-Theoretic Analysis of the MEMR for General Loss Functions

In this section, we extend the characterization in Theorem 4.2 of the MEMR from the log-loss to general loss functions $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. We specifically show that, under suitable assumptions on the loss function, the MEMR (equation (10) in the main text) can be upper bounded using a concave, non-decreasing, function of the conditional mutual information $I(Y; W|X, \mathbf{Z}, \mathbf{Z}_{1:N})$.

To upper bound the MEMR$_l$, we consider the performance of the following randomized meta-decision rule $\Psi_{\mathrm{meta}}(X, \mathbf{Z}, \mathbf{Z}_{1:N})$. Define as $\Phi^*_{\mathrm{meta}}(X, W, U)$ the optimal genie-aided decision rule that minimizes the Bayesian meta-risk, i.e., $R_\ell(Y|X, W, U) = \mathbb{E}[\ell(Y, \Phi^*_{\mathrm{meta}}(X, W, U))]$. This rule is not directly applicable since the pair (U, W) is not known. Having computed the posterior $P_{W,U|X,\mathbf{Z},\mathbf{Z}_{1:N}}$ (see Section 4.2 of the main text), we draw a sample (U', W') from it. Note that conditioned on $(X, \mathbf{Z}, \mathbf{Z}_{1:N})$, the pairs (U, W) and (U', W') are independent. The meta-decision rule is chosen as $\Phi^*_{\mathrm{meta}}(X, W', U')$, substituting the true pair (U, W) with the sample (U', W'). Consequently, the MEMR can be upper bounded as

$$\mathrm{MEMR}_l \leq \mathbb{E}_{P_{X,\mathbf{Z},\mathbf{Z}_{1:N}} P_{Y,W',U'|X,\mathbf{Z},\mathbf{Z}_{1:N}}}[\ell(Y, \Phi^*_{\mathrm{meta}}(X, W', U'))] - \mathbb{E}_{P_{X,Y,W,U}}[\ell(Y, \Phi^*_{\mathrm{meta}}(X, W, U))]. \tag{49}$$

We now obtain an information-theoretic upper bound on (49) under the following assumption. Towards this, we first define the following zero mean random variable

$$\Delta\ell(Y, W', U'|X, \mathbf{Z}, \mathbf{Z}_{1:N}) = \ell(Y, \Phi^*_{\mathrm{meta}}(X, W', U') - \mathbb{E}_{P_{Y,W',U'|X,\mathbf{Z},\mathbf{Z}_{1:N}}}[\ell(Y, \Phi^*_{\mathrm{meta}}(X, W', U')]).$$

Assumption F.1 *There exists function $\Upsilon(\lambda)$ for $\lambda \in (0, b]$ satisfying $\Upsilon(0) = \Upsilon'(0) = 0$ such that the cumulant generating function (CGF) of $\Delta\ell(Y, W', U'|x, \mathbf{z}, \mathbf{z}_{1:N})$ is upper bounded by $\Upsilon(\lambda)$, i.e., the following inequality holds*

$$\log \mathbb{E}_{P_{Y,W',U'|x,\mathbf{z},\mathbf{z}_{1:N}}}\left[\exp\left(\lambda\Delta\ell(Y, W', U'|x, \mathbf{z}, \mathbf{z}_{1:N})\right)\right] \leq \Upsilon(\lambda) \tag{50}$$

for all $x \in \mathcal{X}$, $\mathbf{z} \in \mathcal{Z}^m$ and $\mathbf{z}_{1:N} \in \mathcal{Z}^{Nm}$.

We also define the Legendre dual of $\Upsilon(\lambda)$ as $\Upsilon^*(x) = \sup_{\lambda \in (0,b]} \lambda x - \Upsilon(\lambda)$. It is a non-negative, convex and a non-decreasing function on $[0, \infty)$ with $\Upsilon^*(0) = 0$ (Boucheron et al., 2013, Lemma 2.4). The inverse of this function, called inverse Legendre dual, is defined as $\Upsilon^{*-1}(y) = \inf_{\lambda \in (0,b]} (y + \Upsilon(\lambda))/\lambda$, and is concave. We now state our result.

Theorem F.1 *Under Assumption F.1, the following bound on MEMR holds*

$$\text{MEMR}_l \leq \Upsilon^{*-1}\Big(I(W; Y | X, \mathbf{Z}, \mathbf{Z}_{1:N}) \Big) \tag{51}$$

$$\leq \Upsilon^{*-1}\Big(\frac{I(U; \mathbf{Z}_{1:N})}{Nm} + \frac{I(W; \mathbf{Z}|U)}{m} \Big). \tag{52}$$

Proof: The proof follows the approach in Xu and Raginsky (2020) and we outline the main steps here. The following set of relations hold:

$$\text{MEMR}_l$$
$$\leq \mathbb{E}_{P_{X,\mathbf{z},\mathbf{z}_{1:N}} P_{Y,W',U'|X,\mathbf{z},\mathbf{z}_{1:N}}} [\ell(Y, \Phi^*_{\text{meta}}(X, W', U'))] - \mathbb{E}_{P_{X,Y,W,U}} [\ell(Y, \Phi^*_{\text{meta}}(X, W, U))]$$
$$= \mathbb{E}_{P_{X,\mathbf{z},\mathbf{z}_{1:N}}} \Big[\mathbb{E}_{P_{Y,W',U'|X,\mathbf{z},\mathbf{z}_{1:N}}} [\ell(Y, \Phi^*_{\text{meta}}(X, W', U'))] - \mathbb{E}_{P_{Y,W,U|X,\mathbf{z},\mathbf{z}_{1:N}}} [\ell(Y, \Phi^*_{\text{meta}}(X, W, U))] \Big]$$
$$\overset{(a)}{\leq} \mathbb{E}_{P_{X,\mathbf{z},\mathbf{z}_{1:N}}} \Big[\Upsilon^{*-1}\Big(D_{\text{KL}}(P_{Y,W,U|X,\mathbf{z},\mathbf{z}_{1:N}} || P_{Y,W',U'|X,\mathbf{z},\mathbf{z}_{1:N}}) \Big) \Big]$$
$$\overset{(b)}{\leq} \Upsilon^{*-1}\Big(\mathbb{E}_{P_{X,\mathbf{z},\mathbf{z}_{1:N}}} \Big[D_{\text{KL}}(P_{Y,W,U|X,\mathbf{z},\mathbf{z}_{1:N}} || P_{Y,W',U'|X,\mathbf{z},\mathbf{z}_{1:N}}) \Big] \Big)$$
$$\overset{(c)}{=} \Upsilon^{*-1}\Big(I(Y; W, U | X, \mathbf{Z}, \mathbf{Z}_{1:N}) \Big)$$
$$= \Upsilon^{*-1}\Big(I(Y; W | X, \mathbf{Z}, \mathbf{Z}_{1:N}) \Big)$$
$$\overset{(d)}{\leq} \Upsilon^{*-1}\Big(\frac{I(U; \mathbf{Z}_{1:N})}{Nm} + \frac{I(W; \mathbf{Z}|U)}{m} \Big). \tag{53}$$

Here, the inequality in (a) follows from Assumption F.1 and using Donsker-Varadhan inequality (see (Xu and Raginsky, 2020, Lemma A.1)). The inequality in (b) follows by using Jensen's inequality on the concave inverse Legendre dual function $\Upsilon^{*-1}(\cdot)$. The equality in (c) follows from the observation that while the distribution $P_{Y,W,U|X,\mathbf{z},\mathbf{z}_{1:N}}$ factorizes as $P_{Y|X,\mathbf{z},\mathbf{z}_{1:N}} \otimes P_{W,U|X,Y,\mathbf{z},\mathbf{z}_{1:N}}$, the distribution $P_{Y,W',U'|X,\mathbf{z},\mathbf{z}_{1:N}}$ is obtained as $P_{Y|X,\mathbf{z},\mathbf{z}_{1:N}} \otimes P_{W,U|X,\mathbf{z},\mathbf{z}_{1:N}}$ with (W', U') conditionally independent of Y given $(X, \mathbf{Z}, \mathbf{Z}_{1:N})$. The last inequality in (d) follows since the inverse Legendre dual is a non-decreasing function. ∎

It can be seen that if the random variable $\Delta\ell(Y, W', U'|x, \mathbf{z}, \mathbf{z}_{1:N})$ is σ^2-sub Gaussian[2] when $(Y, W', U') \sim P_{Y,W',U'|x,\mathbf{z},\mathbf{z}_{1:N}}$ for all $x \in \mathcal{X}$, $\mathbf{z} \in \mathcal{Z}^m$ and $\mathbf{z}_{1:N} \in \mathcal{Z}^{Nm}$, then Assumption F.1 is satisfied with $b = \infty$, $\Upsilon(\lambda) = \lambda^2 \sigma^2 / 2$ and $\Upsilon^{*-1}(y) = \sqrt{2\sigma^2 y}$. We now specialize Theorem F.1 to account for this case.

Corollary F.2 *Assume that $\Delta\ell(Y, W', U'|x, \mathbf{z}, \mathbf{z}_{1:N})$ is σ^2-sub Gaussian for all $x \in \mathcal{X}$, $\mathbf{z} \in \mathcal{Z}^m$ and $\mathbf{z}_{1:N} \in \mathcal{Z}^{Nm}$. Then, the following upper bound on the MEMR holds:*

$$\text{MEMR}_l \leq \sqrt{2\sigma^2 \Big(\frac{I(U; \mathbf{Z}_{1:N})}{Nm} + \frac{I(W; \mathbf{Z}|U)}{m} \Big)}. \tag{54}$$

---

[2]A zero-mean random variable X is said to be σ^2-sub-Gaussian if the cumulant generating function (CGF) satisfies $\log \mathbb{E}_{P_X}[\exp(\lambda(X))] \leq \frac{\lambda^2 \sigma^2}{2}$ for all $\lambda \in \mathbb{R}$.

G Additional Experiments

G.1 Bayesian Sinusoidal Regression

We focus on a sinusoidal regression problem, in which we have $Y = W \sin(X) + \xi$, with amplitude W and observation noise $\xi \sim \mathcal{N}(0, 1)$. The prior distribution of model parameter W is determined via hyperparameter U as $P_{W|U} = \mathcal{N}(W|U, \sigma_w^2)$ for some fixed variance σ_w^2. The hyperprior on the prior-mean U is taken as $P_U = \mathcal{N}(0, 1)$.
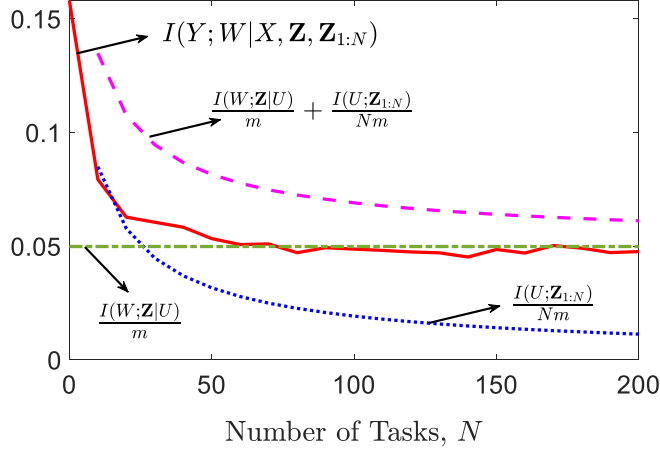

Figure 4: MEMR (13), information-theoretic upper bound (15) and the sensitivity terms $I(W; \mathbf{Z}|U)/m$ and $I(U; \mathbf{Z}_{1:N})/Nm$ as a function of number of meta-training tasks (N) with fixed number of samples $m = 1$ for $\sigma_w^2 = 0.2$.

Figure 4 compares the MEMR under log loss in (13) with the upper bound $\mathrm{MEMR}_{\log}^{\mathrm{UB}}$ in (15), as well as its two component sensitivity terms – model parameter-level sensitivity $I(W; \mathbf{Z}|U)/m$ and hyperparameter-level sensitivity $I(U; \mathbf{Z}_{1:N})/Nm$ – as a function of the increasing number N of tasks for fixed $m = 1$ and $\sigma_w^2 = 0.2$. Note that when $N = 0$, MEMR corresponds to the minimum excess risk for conventional learning. Figure 4 shows that for the problem setting studied, meta-learning using large number N of tasks can yield significantly lower excess risk than conventional learning. In fact, while increasing N decreases both MEMR and the corresponding upper bound at first, they remain non-vanishing in the limit of large number of tasks. This can be explained by looking at the two sensitivity terms - while the hyper-parameter level sensitivity term decreases and vanishes asymptotically, the model parameter sensitivity term, which captures the uncertainty due to limited number m of samples observed about the new, previously unobserved meta-test task, is not influenced by N and remains constant. The non-vanishing MEMR can be thus attributed to the residual epistemic uncertainty about the newly encountered meta-test task.

In Figure 5, we compare the MEMR (13), the upper bound (15) and the two sensitivity terms as a function of increasing number m of per-task samples, for fixed $N = 1$ and $\sigma_w^2 = 0.2$. It can be seen that availability of abundant number of per-task training samples decreases the MEMR, the corresponding upper bound as well as the two sensitivity terms, all of which vanish asymptotically. When large number m (say $m > 75$) of training samples is available, the MEMR is largely determined by the model parameters-level sensitivity. As observed in Figure 4, meta-learning using large N has no impact on model parameter sensitivity, and is thus not very beneficial over conventional learning in this regime. However, when m is small (say $m < 10$), it can be seen that the major contributor to the MEMR is the hyperparameter-level sensitivity term. Consequently, in this regime, meta-learning using large number N of tasks can significantly reduce the hyperparameter sensitivity and thus the MEMR.

In Figure 6, we increase the prior variance to $\sigma_w^2 = 2$ and compare the MEMR, the upper bound and the two sensitivity terms, as a function of the increasing number of meta-training tasks N for fixed $m = 1$. Recall that in Figure 4, we considered the prior $P_{W|U}$ to be more concentrated at the mean or hyperparameter U (by choosing small $\sigma_w^2 = 0.2$). This results in lower model parameter sensitivity (given knowledge of U), while the
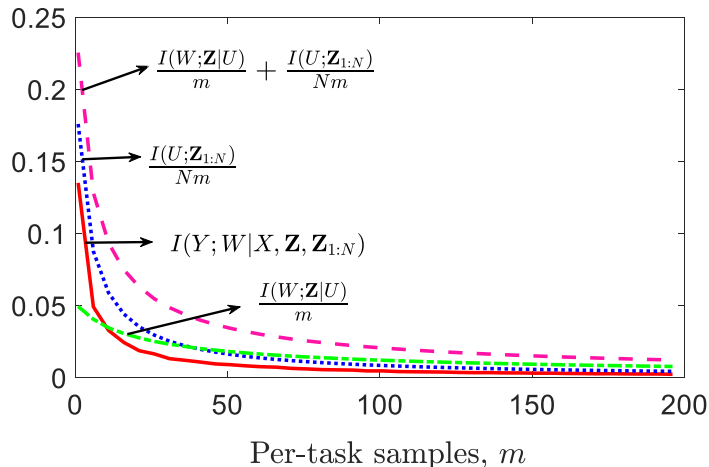
Figure 5: MEMR (13), information-theoretic upper bound (15) and the sensitivity terms $I(W; \mathbf{Z}|U)/m$ and $I(U; \mathbf{Z}_{1:N})/Nm$ as a function of number per-task samples (m) with fixed number of meta-training tasks $N = 1$ for $\sigma_w^2 = 0.2$.
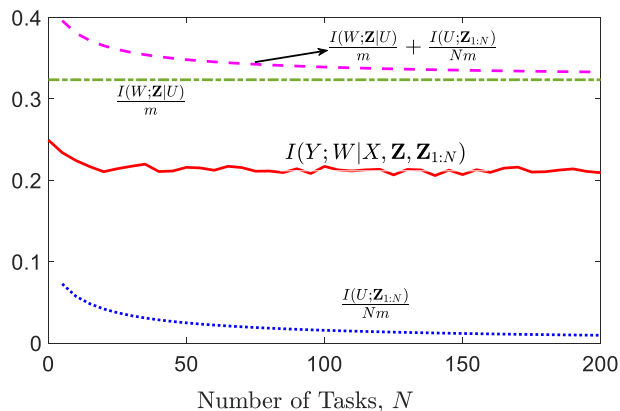

Figure 6: MEMR (13), information-theoretic upper bound (15) and the sensitivity terms $I(W; \mathbf{Z}|U)/m$ and $I(U; \mathbf{Z}_{1:N})/Nm$ as a function of number per-task samples (m) with fixed number of meta-training tasks $N = 1$ for $\sigma_w^2 = 2$.

hyperparameter sensitivity is large and is shown to decrease when meta-learning using large N. In contrast, by assuming a larger prior variance, the model parameter sensitivity outweighs the hyperparameter sensitivity. As such, meta-learning using large number N of tasks brings marginal benefits over conventional learning (i.e. when $N = 0$) as shown by the MEMR curve.

H Experimental Details for Bayesian Neural Network Regression

We detail the essential experimental settings for reproducibility of the results.

Settings	Value
regression architecture	MLP
# hidden layers	1
# hidden units	3
activation	ReLU

Table 1: BNN architecture for Bayesian neural network regression

Settings	Value
classifier architecture for C-MINE	MLP + Sigmoid layer
# hidden layers	3
# hidden units	(64,64,1)
step size	0.001
optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
# epoch	200
batch size	64
regularizer	L2 (0.001)
activation	ReLU

Table 2: classifier details in C-MINE for Bayesian neural network regression

Settings	Value
ratio between dataset used for training C-MINE and computing estimated mutual information	1:1 (use same dataset)
number of samples in the dataset	30000
clipping value τ for SMILE (clip log density ratio estimator for marginal dataset between $-\tau$ and τ)	1.0

Table 3: training details for C-MINE and SMILE for Bayesian neural network regression