Information-theoretic Analysis of Bayesian Test Data Sensitivity

Futoshi Futami
Osaka University / RIKEN AIP

Tomoharu Iwata
NTT Corporation

Abstract

Bayesian inference is often used to quantify uncertainty. Several recent analyses have rigorously decomposed uncertainty in prediction by Bayesian inference into two types: the inherent randomness in the data generation process and the variability due to lack of data respectively. Existing studies have analyzed these uncertainties from an information-theoretic perspective, assuming the model is well-specified and treating the model parameters as latent variables. However, such information-theoretic uncertainty analysis fails to account for a widely believed property of uncertainty known as sensitivity between test and training data. This means that if the test data is similar to the training data in some sense, the uncertainty will be smaller. In this study, we study such sensitivity using a new decomposition of uncertainty. Our analysis successfully defines such sensitivity using information-theoretic quantities. Furthermore, we extend the existing analysis of Bayesian meta-learning and show the novel sensitivities among tasks for the first time.

1 INTRODUCTION

Evaluating uncertainty in predictions of machine learning algorithms has become increasingly important. Such information is used in the detection of domain shifts (Ovadia et al., 2019), adversarial attacks (Ye and Zhu, 2018), Bayesian optimization (Hernández-Lobato et al., 2014), and reinforcement learning (Janz et al., 2019). The Bayesian inference is widely used in such applications since it represents uncertainties through a posterior distribution updated from the prior distribution using training data (Bishop, 2006).

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

Recently, Xu and Raginsky (2022) have rigorously decomposed the uncertainty in the prediction of Bayesian inference into two types; one is **aleatoric uncertainty**, which is caused by the noise inherent in the data-generating process. The other is called **epistemic uncertainty**, caused by the lack of data. The key idea of their analysis is that assuming the model is well-specified, model parameters are treated as latent variables and marginalized as is done in Bayesian inference. Thus, they called the setting **Bayesian learning**.

Under the Bayesian learning setting, they clarified that the aleatoric uncertainty corresponds to the Bayes risk since the noise in the data-generating process is closely related to the fundamental difficulty of learning. As we introduce in Sec. 2, they showed that the epistemic uncertainty can be regarded as the excess risk under the posterior predictive distribution since such excess risk corresponds to the "loss due to lack of data" when the model is well specified. Furthermore, they clarified that the excess risk is closely related to **conditional mutual information (CMI)**, see Eq. (4) for details. The CMI satisfies the desirable property of epistemic uncertainty, such as monotonically decreasing with the training dataset size. These settings have recently been extended to Bayesian meta-learning settings, where we assume a hyperprior distribution on prior distributions (Theresa Jose et al., 2022).

A limitation of these existing Bayesian learning analyses is that they cannot account for the widely believed geometric property of epistemic uncertainty: If a given test data point is similar to the training data in some sense, the uncertainty at such a test data point should be small because there is enough information to make a prediction. On the other hand, if the test data is not very similar to the training data, the uncertainty should be large. This property is called **sensitivity** between test and training data points. Linear models and Gaussian processes exhibit this property (Bishop, 2006) because the variance of the posterior predictive distribution depends explicitly on the distance between test and training data under a given feature map. However, existing analyses in Bayesian learning failed to explain such sensitivity because they analyzed the CMI

only focusing on the training data without considering its relation to the test data.

In this paper, we continue the uncertainty analysis under Bayesian learning and aim to analyze the sensitivity between the test and training data points. To achieve this, we first present the novel decomposition of the CMI in Theorem 1. Using this, we formally define the sensitivity between the test and training data using an information-theoretic quantity. Then we provide the theoretical and numerical analyses of this quantity. We also apply our analysis to the meta-learning setting similarly to Theresa Jose et al. (2022) and determined the sensitivity between the meta-training and meta-test tasks in Theorem 6. To the best of our knowledge, the sensitivity among tasks is presented for the first time.

Another contribution of this work is a new information-theoretic upper bound of the CMI, which includes the interaction between training data points in Corollary 1. Our new bound is tighter than the existing bound proposed by Xu and Raginsky (2022). Finally, we present a new exact characterization of the generalization error using our novel decomposition in Theorem 4 and show a new connection to the existing information-theoretic generalization error bounds under the frequentist setting (Aminian et al., 2021).

2 PRELIMINARIES

Here, we review the setting of **Bayesian learning** used by Xu and Raginsky (2022) and its extension to the meta-learning setting proposed by Theresa Jose et al. (2022). Capital letters such as X represent random variables, whereas lowercase letters such as x represent deterministic values.

2.1 Bayesian Learning

We consider a supervised setting and denote input–output pairs by $Z = (X, Y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Learners can access N training data, $Z^N := (Z_1, \dots, Z_N)$ with $Z_n := (X_n, Y_n)$, which are independent and identically distributed (i.i.d.) samples from some underlying distribution. The goal of supervised learning is to use Z^N to predict the test target variable Y given the test input X, independently drawn from the same distribution as the training data. For this purpose, we consider a parametric generative model. We assume that the underlying distribution belongs to a model class $\{p(z|w) : w \in \mathcal{W}\}$ with model parameter w in the set $\mathcal{W}$. As discussed in Theresa Jose et al. (2022) and Hafez-Kolahi et al. (2021), this implies that the model is well-specified. We also assume that $p(Z|W) = p(Y|X, W)p(X)$ for simplicity. This means that the input data X is independent of the model parameter.

In **Bayesian learning**, model parameters are treated as latent random variables following a prior distribution $p(w)$. Conditioned on $W = w$, the data are generated by $p(Z|W = w)$. Thus, the joint distribution of the training data Z^N, the test data Z^*, and the model parameter W is given by

$$p(W, Z^N, Z^*) \coloneqq p(W)p(Z|W)^N p(Z^*|W). \quad (1)$$

Since Z^N are i.i.d. samples, we can express $p(Z|W)^N = p(Z^N|W)$. See Sec. 6 for the intuition of this joint model. We express the Bayesian posterior as $p(W|Z^N)$ and the posterior predictive distribution as $p(Y^*|X^*, Z^N) := \mathbb{E}_{p(W|Z^N)}p(Y^*|X^*, W)$.

Next, we introduce action a and loss function l to measure the performance of supervised learning. We define $\mathcal{A}$ as an action space and the loss function as $l : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. The loss of action $a \in \mathcal{A}$ and the target variable y are written as $l(y, a)$; for example, the log loss is given as $l(y, q) = -\ln q(y)$, where q is the probability density of Y and $\mathcal{A}$ is the set of all probability densities on Y. The squared loss is given as $l(y, a) = |y-a|^2$, where $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Our goal is to infer the decision rule $\psi : \mathcal{X} \times \mathcal{Z}^N \to \mathcal{A}$ that minimizes the expected loss $\mathbb{E}_{p(Y^*, X^*, Z^N)}[l(Y^*, \psi(X^*, Z^N))]$ among all decision rules. Following the previous work by Xu and Raginsky (2022), we define the infimum of the expected loss as the **Bayesian risk**:

$$R_l(Y^*|X^*, Z^N) \coloneqq \inf_{\psi:\mathcal{X}\times\mathcal{Z}^N\to\mathcal{A}} \mathbb{E}_{p(Y^*, X^*, Z^N)}[l(Y^*, \psi(X^*, Z^N))].$$

For example, when the log loss is used, $R_{\log}(Y^*|X^*, Z^N) = H[Y^*|X^*, Z^N]$, where $H[Y^*|X^*, Z^N]$ is the entropy of the posterior predictive distribution defined as $H[Y^*|X^*, Z^N] =$

$$\mathbb{E}_{p(Z^N)p(X^*)}\mathbb{E}_{p(Y^*|X^*, Z^N)}[-\log p(Y^*|X^*, Z^N)].$$

Thus, the Bayesian risk equals the test error under a posterior predictive distribution.

Next, we define a fundamental limit of learning as $\phi : \mathcal{X} \times \mathcal{W} \to \mathcal{A}$, which takes the true parameter W instead of the training dataset Z^N. Then, the corresponding risk is given as

$$R_l(Y^*|X^*, W) \coloneqq \inf_{\phi:\mathcal{X}\times\mathcal{W}\to\mathcal{A}} \mathbb{E}_{p(Y^*, X^*, W)}[l(Y^*, \phi(X^*, W))]. \quad (2)$$

We cannot improve this risk by increasing the number of training data. Thus, $R_l(Y^*|X^*, W)$ can be regarded as the **aleatoric uncertainty** since it expresses the fundamental difficulty of learning. In other words, this risk implies the inherent presence of randomness in the

data-generating mechanism. When the log loss is used, $R_l(Y^*|X^*, W)$ corresponds to the conditional entropy

$$R_{\log}(Y^*|X^*, W) = \mathbb{E}_{p(X^*)p(W)} H[Y^*|X^*, W] \\ = \mathbb{E}_{p(Y^*, X^*, W)}[-\log p(Y^*|X^*, W)].$$

Finally, we define the difference between the Bayesian risk and the fundamental limit of learning as the **minimum excess risk (MER)**:

$$\mathrm{MER}_l(Y^*|X^*, Z^N) := R_l(Y^*|X^*, Z^N) - R_l(Y^*|X^*, W). \tag{3}$$

This corresponds to the **epistemic uncertainty** since it is defined as the difference between the Bayesian risk and the fundamental limit of learning. Thus, MER implies the loss due to insufficient training data under the well-specified model assumption (Xu and Raginsky, 2022; Hafez-Kolahi et al., 2021). When the log loss is used, MER is given as

$$\mathrm{MER}_{\log}(Y^*|X^*, Z^N) = I(W; Y^*|X^*, Z^N), \tag{4}$$

where $I(W; Y^*|X^*, Z^N)$ is **the conditional mutual information (CMI)**. Other than the log loss, if the loss function satisfies the σ^2 sub-Gaussian property conditioned on $(X^*, Z^N) = (x^*, z^N)$, Xu and Raginsky (2022) showed that

$$\mathrm{MER}_l(Y^*|X^*, Z^N) \leq \sqrt{2\sigma^2 I(W; Y^*|X^*, Z^N)}. \tag{5}$$

Thus, MER is upper-bounded by the square root of the CMI. Thus, understanding the CMI is crucial to understand MER_l and epistemic uncertainty. For this reason, we mainly analyze the CMI in this paper. CMI has some desirable properties for understanding epistemic uncertainty. Xu and Raginsky (2022) proved that CMI is greater than 0 and it decreases as we increase N. Moreover, they showed that CMI can be upper-bounded by the mutual information (MI) as follows:

lmm 1 (Xu and Raginsky (2022))**.** *Under the joint distribution of Eq.* (1)*, we obtain*

$$I(W; Y^*|X^*, Z^N) \leq \frac{1}{N} I(W; Z^N). \tag{6}$$

In many practical settings, $I(W; Z^N)$ is upper-bounded by $\mathcal{O}(\log N)$; thus, the CMI is bounded by $\mathcal{O}(\log N/N)$. Therefore, it converges to 0 as $\mathcal{O}(\log N/N)$ for the log loss and $\mathcal{O}(\sqrt{\log N/N})$ for sub-Gaussian loss functions. It has been discussed that $I(W; Z^N)$ captures the sensitivity of the learned parameter and training dataset and is closely connected to the generalization error bound (Xu and Raginsky, 2017).

2.2 Bayesian Meta-learning

Uncertainty also plays an important role in meta-learning. In traditional Bayesian inference, the prior distribution is selected on the basis of prior knowledge about the task. In Bayesian meta-learning, the prior distribution is automatically inferred by observing related tasks. We model the statistical relationship between different tasks using a hierarchical Bayesian model with a global latent variable U in the set $\mathcal{U}$. To understand the uncertainty captured by this hierarchical structure, Theresa Jose et al. (2022) extended the analysis of Bayesian learning.

We observe M related tasks and aim to infer a suitable prior distribution for a new unknown task. Each meta-training dataset has N data points drawn i.i.d from $p(Z|W = w_m)$, where w_m is the task-specific parameter. We express the m-th meta-training dataset as $Z^{N,(m)} = (Z_1^{(m)}, \ldots, Z_N^{(m)})$. We assume that the parameter W_m is drawn i.i.d from the shared prior $p(W|U)$ parametrized by the global latent variable U. We assume the hyperprior distribution $p(U)$ on U. We express meta-training dataset as $Z^{NM} = (Z^{N,(1)}, \ldots, Z^{N,(M)})$. We express model parameters of the meta-training dataset as $W^M = (W_1, \cdots, W_M)$. Finally, we have a new unknown task called the meta-test task generated from the meta-test task parameter W. We can use the meta-test training data $Z^N = (Z_1, \cdots, Z_N)$ and the meta-test input data X^* at the test stage.

The above setting is summarized as the joint distribution below:

$$p(U, W^M, Z^{NM}, W, Z^N, Z) := \\ p(U) \underbrace{\Big(p(W|U)p(Z^N|W)\Big)^M}_{\text{meta-training}} \underbrace{p(W|U)p(Z^N|W)p(Z^*|W)}_{\text{meta-testing}}. \tag{7}$$

Here, we omit the index for the meta-training dataset for simplicity. Under this setting, we consider the decision rules and excess risk in the same way as in Sec. 2.1. We define the Bayesian meta-risk as

$$R_l(Y|X^*, Z^N, Z^{NM}) := \inf_{\psi_{\mathrm{meta}}: \mathcal{X} \times \mathcal{Z}^{NM} \times \mathcal{Z}^N \to \mathcal{A}} \\ \mathbb{E}_{p(Z^{NM}, Z^N, Z^*)}[l(Y^*, \psi_{\mathrm{meta}}(X^*, Z^{NM}, Z^N))]. \tag{8}$$

We also define the fundamental limit of learning in meta-learning as

$$R_l(Y^*|X^*, W, U) := \\ \inf_{\phi_{\mathrm{meta}}: \mathcal{X} \times \mathcal{W} \times \mathcal{U} \to \mathcal{A}} \mathbb{E}_{p(U,W,Z)}[l(Y^*, \phi_{\mathrm{meta}}(X^*, W, U))]. \tag{9}$$

We define the minimum excess meta-risk (MEMR) as

$$\begin{aligned}&\mathrm{MEMR}_l(Y^*|X^*, Z^N, Z^{NM})\\ &:= R_l(Y^*|X^*, Z^N, Z^{NM}) - R_l(Y^*|X^*, W, U). \quad (10)\end{aligned}$$

Theresa Jose et al. (2022) showed that the MEMR of the log loss equals to the CMI:

$$\mathrm{MEMR}_{\log}(Y^*|X^*, Z^N, Z^{NM}) = I(W; Y^*|X^*, Z^N, Z^{NM}), \quad (11)$$

and derived the upper-bound of $\mathrm{MEMR}_{\log}$, which is similar to Eq. (6), as

$$I(W; Y^*|X^*, Z^N, Z^{NM}) \leq \frac{I(U; Z^{NM})}{NM} + \frac{I(W; Z^N|U)}{N}. \quad (12)$$

We can see that the CMI is also upper-bounded by the MI that captures the sensitivities of the learned meta-test task parameter, hyperparameter, and meta-training dataset.

3 EXACT CHARACTERIZATION OF CMI

As we pointed out in Sec. 1, the analysis of MER_l introduced in Sec. 2 cannot explain the sensitivity in uncertainty between the test and training data. The limitation of the information-theoretic analysis of Lemma 1 is that the decomposition of the CMI focuses only on the training dataset and not on the test data point.

Here, we show our novel CMI decomposition and present the information-theoretic quantity of the sensitivity. First, we consider the Bayesian learning setting introduced in Sec. 2.1. All the proofs are shown in the Supplementary material.

3.1 Decomposition of CMI Using Sensitivity

The following theorem is our first main result, which decomposes the CMI as the sum of the MI and the sensitivities of the test data and each training data point.

thm 1. *Under the joint distribution of Eq.* (1)*, we have*

$$\begin{aligned}I(W; Y^*|X^*, Z^N) = &\frac{1}{N} I(W; Z^N) - \frac{1}{N}\sum_{n=1}^{N} I(Z^*, Z_n|Z^{N\setminus n})\\ &- \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}), \quad (13)\end{aligned}$$

where $Z^{N\setminus n} := (Z_1, \ldots, Z_{n-1}, Z_{n+1}, \ldots, Z_N)$ *and* $Z^{n-1} := (Z_1, \ldots, Z_{n-1})$.

Different from the bound in Lemma 1, the CMI is decomposed into three terms connected with equality, not inequality. The first term on the right-hand side of Eq. (13) is the MI between the learned parameter and the training dataset. The second and third terms correspond to the binary relation about how much information each data point has to predict other data points. The second term $I(Z^*, Z_n|Z^{N\setminus n})$ represents the information-theoretic quantity of the sensitivity of the test and training data points. This term indicates how useful the training data point Z_n is to predict the test data point Z. If the training data point Z_n has more information about the test data, then the uncertainty at Z decreases. If Z_n is almost independent of Z, then the mutual information becomes 0, which means that the uncertainty increases.

From this observation, we introduce the definition of test data sensitivity as follows.

dfn 1. *The sensitivity of the test data and the single training data point is defined as*

$$I_n \coloneqq I(Z^*, Z_n|Z^{N\setminus n}). \quad (14)$$

For simplicity, we also express $I_{n+1,n} := I(Z_{n+1}, Z_n|Z^{n-1})$, which appears in Eq. (13).

We transform I_n into a more intuitive expression as

$$\begin{aligned}I_n &= H(Z^*|Z^{N\setminus n}) - H(Z^*|Z^N) \quad (15)\\ &= \mathbb{E}_{p(W,Z^N,Z^*)} \ln \frac{\mathbb{E}_{p(W|Z^{N\setminus n})} p(Z^*, Z_n|W)}{p(Z^*|Z^{N\setminus n}) p(Z_n|Z^{N\setminus n})}. \quad (16)\end{aligned}$$

Eq. (15) is useful for explicitly calculating the sensitivity for some models shown in Sec. 3.2. Eq. (16) states that the joint posterior predictive distribution $\mathbb{E}_{p(W|Z^{N\setminus n})} p(Z^*, Z_n|W)$ differs from the single-point posterior predictive distribution. The joint predictive distribution has recently attracted attention in decision problems (Rosenfeld et al., 2020; Osband et al., 2021; Wen et al., 2021). Thus, our theoretical results suggest new insights into the connection between decision problems, joint predictive distribution, and uncertainty. However, this is outside the scope of this study, and we leave it to future work to explore this connection.

Finally, from Theorem 1, we obtain the new information-theoretic bound for the CMI as follows.

col 1. *Under the joint distribution of Eq.* (1)*, we obtain*

$$\begin{aligned}I(W; Y^*|X^*, Z^N) \leq &\frac{1}{N} I(W; Z^N)\\ &- \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}). \quad (17)\end{aligned}$$

This bound is tighter than that of Lemma 1 owing to the second term on the right-hand side. In Sec. 7, we numerically compare this bound with that of Lemma 1.

3.2 Linear Regression Model

As we discussed in Sec. 1, we can explicitly evaluate the sensitivity in linear models. In this section, we use a linear regression model to explicitly discuss the relationship between such traditional sensitivity and the results in Sec. 3.1.

The likelihood of the model is given as the Gaussian distribution with the mean $w^\top \phi(x)$ and the variance $\beta^{-1} \in \mathbb{R}^+$. We express it as $p(Y^*|X^* = x^*, w) = \mathcal{N}(w^\top \phi(x^*), \beta^{-1})$, where $\mathcal{Y} = \mathbb{R}$ and $\phi(x) := (\phi_1(x), \ldots, \phi_d(x))^\top \in \mathbb{R}^d$ is a d-dimensional feature vector of input x and each $\phi_i : \mathcal{X} \to \mathbb{R}$. We assume a prior distribution $p(w) = \mathcal{N}(0, \alpha^{-1} I_d)$ with some positive constant $\alpha > 0$. We define a design matrix as $\Phi = (\phi(x_1), \ldots, \phi(x_N))^\top \in \mathbb{R}^{N \times d}$. We also define a target vector as $\mathbf{y} = (y_1, \ldots, y_N)^\top$. Then, a posterior distribution is given by $p(w|z^N) = \mathcal{N}(m_N, S_N)$, where $m_N = \beta S_N \Phi^\top \mathbf{y}$ and $S_N^{-1} := \alpha I_d + \beta \Phi^\top \Phi$. We also have a posterior predictive distribution as $p(Y^*|X^* = x^*, z^N) := \mathcal{N}(m_N^\top \phi(x^*), \sigma_N^2(x^*))$, where $\sigma_N^2(x^*) := \beta^{-1} + \phi(x^*)^\top S_N \phi(x^*)$.

Since the posterior predictive distribution is given as the Gaussian distribution, its entropy is calculated on the basis of its variance. Thus, $I(W; Y^*|X^*, Z^N) = \mathbb{E}_{p(X^{N+1})} \log \sigma_N^2(X^*)/2 + \text{Const}$, and the interplay between $\phi(x)$ and S_N characterizes the sensitivity of the test and training data points. Similar arguments still hold for Gaussian process models, where the inner products of the feature maps are replaced with kernel functions.

We can explicitly calculate the sensitivity I_n using Eq. (15). Then, we obtain

$$I_n = \mathbb{E}_{p(X^{N+1})} \frac{1}{2} (\ln \sigma_{N\setminus n}^2(X^*) - \ln \sigma_N^2(X^*)). \quad (18)$$

We can simplify this as follows.

thm 2. *For linear models, the sensitivity I_n satisfies the following relation:*

$$\mathbb{E}_{p(X^{N+1})} \frac{(\phi(X^*)^\top S_N \phi(X_n))^2}{2\omega(X_n)(\alpha^{-1}\phi(X^*)^\top \phi(X^*) + \beta^{-1})} \le I_n \le \mathbb{E}_{p(X^{N+1})} \frac{(\phi(X^*)^\top S_N \phi(X_n))^2}{2\omega(X_n)(\beta^{-1} + \phi(X^*)^\top S_N \phi(X^*)},$$

where $\omega(x) := \beta^{-1} - \phi(x)^\top S_N \phi(x)$.

This bound implies that the posterior covariance matrix $S_N := (\alpha I_d + \beta \Phi^\top \Phi)^{-1}$ can be seen as a metric for measuring the similarity between the training data x_n and the test data x. In the Supplementary material, we numerically evaluated this bound.

Combined with Theorem 1, we obtain

$$\mathrm{MER}_{\log}(Y^*|X^*, Z^N) \le \frac{1}{N} I(W; Z^N) - \mathbb{E}_{p(X^{N+1})} \frac{1}{N} \sum_{n=1}^N \frac{(\phi(X^*)^\top S_N \phi(X_n))^2}{2\omega(X_n)(\alpha^{-1}\phi(X^*)^\top \phi(X^*) + \beta^{-1})}.$$

This suggests that the test error becomes small if the given test and training data points are similar under the feature map with the metric S_N. This relation explicitly validates our intuition about the geometric property of the sensitivity introduced in Sec. 1.

3.3 Asymptotic Behavior

Here, we discuss the asymptotic behavior of sensitivity. Using the asymptotic expansion of Bayesian inference introduced in Watanabe (2018), we obtain the following relation:

thm 3. *Assume that $p(z|w)$ has a relatively finite variance, that is, for any pair of $w_0, w \in \mathcal{W}$, there exists a positive constant c_0 such that*

$$c_0 \mathbb{E}_{p(Z|w_0)} (\ln p(Z|w_0) - \ln p(Z|w))^2 \le \mathbb{E}_{p(Z|w_0)} [\ln p(Z|w_0) - \ln p(Z|w)]. \quad (19)$$

Then, we obtain $I_n = I(Z^*, Z_n|Z^{N\setminus n}) = o\left(\frac{1}{N}\right)$, where $o(\frac{1}{N})$ is little o.

The relatively finite variance assumption in Eq. (19) is satisfied in many widely used models. For example, generalized linear models, including the linear and logistic regression models, satisfy this condition. See the Supplementary material and Watanabe (2018) for other examples.

Combined with Theorem 1, since $\frac{1}{N} I(W; Z^N) = O\left(\frac{1}{N}\right)$, we obtain

$$I(W; Y^*|X^*, Z^N) \le \underbrace{\frac{1}{N} I(W; Z^N)}_{=O(\frac{1}{N})} - \underbrace{\frac{1}{N} \sum_{n=1}^N I(Z^*, Z_n|Z^{N\setminus n})}_{=o(\frac{1}{N})}. \quad (20)$$

Thus, since the order of the sensitivity term is $o(1/N)$, it is much smaller than the MI, which is $O(1/N)$. Finally, using Theorem 3, we obtain the following relation:

$$\frac{1}{N} \sum_{n'=1}^{N-1} \sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}) = O\left(\frac{1}{N}\right). \quad (21)$$

4 EXACT CHARACTERIZATION OF GENERALIZATION ERROR

Information theoretic quantities play an important role in understanding generalization as discussed in Xu and

Raginsky (2017). Here, we discuss the relationship between the sensitivity and the generalization using Theorem 1.

4.1 Relation to Generalization Error

First, we show that the statement of Lemma 1 is closely related to the generalization error analysis.

lmm 2. *When a log loss is used, Eq.* (6) *is equivalent to the following inequality;*

$$R_{\log}(Y^*|X^*,Z^N) \leq -\mathbb{E}_{p(Z^N)}\mathbb{E}_{p(W|Z^N)}\frac{1}{N}\sum_{n=1}^{N}\ln p(Z_n|W) + \frac{1}{N}\mathrm{KL}(p(W|Z^N)|p(W)). \tag{22}$$

The left-hand side of Eq. (22) corresponds to the test error, and the right-hand side is the training error plus the regularization term. Thus, Lemma 1 (Eq. (6)) is closely related to the generalization error.

With this observation, using our Theorem 1, we can incorporate the sensitivity of the test and training data to the generalization error as follows.

thm 4. *Under the joint distribution of Eq.* (1) *with a log loss, we obtain*

$$R_{\log}(Y^*|X^*,Z^N) = -\mathbb{E}_{p(Z^N)}\mathbb{E}_{p(W|Z^N)}\frac{1}{N}\sum_{n=1}^{N}\ln p(Y_n|X_n,W) + \frac{1}{N}\mathrm{KL}(p(W|Z^N)|p(W)) - \frac{1}{N}\sum_{n=1}^{N}I_n - \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1}I_{n+1,n}.$$

This theorem states that if training data x_n has sufficient information to predict x, I_n becomes large, leading to a smaller test error. Thus, this relation formalizes our intuition that we can predict a test data point, which is similar to the training data in some sense, better than the test data, which are completely different from the training data. In this way, the geometric properties of the sensitivity and generalization are unified explicitly in Bayesian learning.

Another interesting point is that, unlike Lemma 2, this theorem is the identity, not the inequality. Thus, we can precisely characterize the relationship between the test and training errors. We will discuss the relation between our result and the recently proposed exact characterization of the generalization error (Aminian et al., 2021) in Sec. 4.2.

4.2 Relationship between the Sensitivity and the Gibbs Test Error

So far, we focused on the Bayesian risk, $R_l(Y^*|X^*,Z^N)$ as the test error, and it is based on the Bayesian posterior predictive distribution. On the other hand, in many generalization error analyses, we often use the Gibbs test error defined as

$$R_{\log}^{\mathrm{Gibbs}}(Y^*|X^*,Z^N) := \mathbb{E}_{p(W)p(Z^N|W)p(\tilde{W}|Z^N)}[-\mathbb{E}_{p(Z^*|W)}\log p(Y^*|X^*,\tilde{W})].$$

Here, we express the learned parameter as $\tilde{W}$, which follows the Bayesian posterior distribution $p(\tilde{W}|Z^N)$. By comparing with $R_{\log}(Y^*|X^*,Z^N)$, which uses the posterior predictive distribution, we obtain

$$R_{\log}(Y^*|X^*,Z^N) \leq R_{\log}^{\mathrm{Gibbs}}(Y^*|X^*,Z^N), \tag{23}$$

where we used the Jensen inequality. This relation is general since we only use the convexity of the log loss.

We further explore the relationship between the Gibbs test error $R_{\log}^{\mathrm{Gibbs}}(Y^*|X^*,Z^N)$ and the Bayesian risk $R_{\log}(Y^*|X^*,Z^N)$ using the Lautum information (LI), which was used by Aminian et al. (2021). First, we present the exact characterization of the generalization error of the Gibbs test error.

thm 5. *Under the joint distribution of Eq.* (1) *with a log loss, we obtain*

$$R_{\log}^{\mathrm{Gibbs}}(Y^*|X^*,Z^N) = -\mathbb{E}_{p(W)p(Z^N|W)}\frac{\sum_{n=1}^{N}\ln p(Y_n|X_n,W)}{N} + \frac{LI(\tilde{W};Z^N|W)}{N},$$

where LI is the Lautum information defined as

$$LI(\tilde{W};Z^N|W) = \mathbb{E}_{p(W)p(\tilde{Z}^N|W)p(\tilde{W}|\tilde{Z}^N)p(Z^N|W)}\log\frac{p(Z^N|W)p(\tilde{W}|W)}{p(Z^N,\tilde{W}|W)} = \mathrm{KL}(p(\tilde{W}|W)p(Z^N|W)|p(\tilde{W},Z^N|W)). \tag{24}$$

Note that the LI can also be regarded as the reverse KL divergence (Aminian et al., 2021). This result is similar to the exact characterization of generalization error under the **frequentist setting** used by Aminian et al. (2021).

From Theorems 4 and 5, we can evaluate the difference between the Gibbs test error and Bayesian risk.

col 2. *Under the joint distribution of Eq.* (1)*, we obtain*

$$R_{\log}^{\mathrm{Gibbs}}(Y^*|X^*,Z^N) - R_{\log}(Y^*|X^*,Z^N) = \frac{LI(\tilde{W};Z^N|W)}{N} + \sum_{n=1}^{N}\frac{I_n}{N} + \sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1}\frac{I_{n+1,n}}{N} - \frac{I(W;Z^N)}{N}. \tag{25}$$

From the Jensen inequality of Eq. (23), if the posterior $p(\tilde{W}|Z^N)$ is a point mass, the Jensen gap vanishes. From this relation, as the sensitivity term I_n increases,

the Jensen gap becomes large. The Jensen gap has been studied in relation to the model misspecification under the frequentist setting (Grünwald, 2012; Grünwald and Van Ommen, 2017). Since our setting is Bayesian learning, it is difficult to directly compare our Eq. (25) with previously reported results of the frequentist setting. We leave it to future work to clarify how the existing analysis of the Jensen gap under model misspecification is translated into our setting.

5 EXACT CHARACTERIZATION OF CMI IN META-LEARNING

In this section, we extend our information-theoretic analysis of the sensitivity to a Bayesian meta-learning setting. The following is our main result:

thm 6. *Under the joint distribution of Eq.* (7)*, we obtain*

$$
\begin{aligned}
&I(W;Y^*|X^*,Z^N,Z^{NM}) \\
&= \frac{1}{N} I(W;Z^N|U) + \frac{1}{NM} I(U;Z^{NM}) && (26)\\
&- \frac{1}{NM}\sum_{m=1}^{M} I(Z^N, Z^{N,(m)}|Z^{N(M\backslash m)}) && (27)\\
&- \frac{1}{NM}\sum_{m'=1}^{M-1}\sum_{m=m'}^{M-1} I(Z^{N,(m+1)}, Z^{N,(m)}|Z^{N(m-1)}) && (28)\\
&- \frac{1}{N}\sum_{n=1}^{N} I(Z^*, Z_n|Z^{N\backslash n}, Z^{NM}) && (29)\\
&- \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}, Z^{NM}), && (30)
\end{aligned}
$$

where $Z^{Nm} := (Z^{N,(1)},\cdots,Z^{N,(m)})$, $Z^{N(m-1)} := (Z^{N,(1)},\cdots,Z^{N,(m-1)})$, *and* $Z^{N(M\backslash m)} := (Z^{N,(1)},\ldots,Z^{N,(m-1)},Z^{N,(m+1)},\ldots,Z^{N,(M)})$.

This theorem rigorously quantifies how the different levels of sensitivities (Eqs. (27) to (30)) contribute to the CMI, which corresponds to the epistemic uncertainty. We explain each sensitivity below. First, Eq. (27) represents the sensitivity between the test and training tasks since it quantifies how useful the m-th training task is to predict the meta-test task. To the best of our knowledge, our study is the first to quantify task sensitivity theoretically. Eq. (29) quantifies the sensitivities of the meta-test training data and meta-test test data points similarly to Theorem 1.

The information from the relevant tasks is captured by the hyper-posterior distribution $p(U|Z^{NM})$, and the information from meta-test training data is incorporated into the posterior distribution $p(W|Z^N,U)$. These correspond to Eq. (26), which also appears in the existing MEMR bound. Such relations are also summarized as the posterior predictive distribution:

$$
\begin{aligned}
&p(Y^*|X^*,Z^N,Z^{NM}) \\
&= \mathbb{E}_{p(W|Z^N,Z^{NM})} p(Y^*|X^*,W) \\
&= \mathbb{E}_{p(U|Z^{NM})p(W|Z^N,U)} p(Y^*|X^*,W).
\end{aligned}
$$

The sensitivity between the meta-test and meta-training tasks of Eq. (27) is given as

$$
\begin{aligned}
&I(Z^N, Z^{N,m}|Z^{N(M\backslash m)}) \\
&= H(Z^N|Z^{N(M\backslash m)}) - H(Z^N|Z^{NM}).
\end{aligned}
$$

We can evaluate the above by evaluating the hyper-posterior distributions $p(U|Z^{N(M\backslash m)})$ and $p(U|Z^{NM})$.

We can obtain the improved information-theoretic upper bound about MEMR, which improves Eq. (12), in the same way as Corollary 1. We numerically evaluate these information-theoretic quantities in Sec. 7.

6 RELATED WORK

First, we point out that the joint model in Eq. (1) often appears in the Bayesian decision theory (Robert et al., 2007) in statistics. This model evaluates the *average* performance of the risk function over a prior distribution and leads to the minimax rate analysis of the parameter estimation. The lower bound for the decision rule in the minimum excess risk has recently been reported (Hafez-Kolahi et al., 2021) using the rate-distortion theory (Cover and Thomas, 2006). Moreover, the joint model in Eq. (1) is used in Bayesian experimental design, in which the stochastic dependencies of data and parameters are introduced to incorporate uncertainty. We also remark that this model is closely related to Bayesian online learning, and we show the regret analysis in the Supplementary material.

The sensitivity between test and training data points has been an important property theoretically and practically (Bishop, 2006; Murphy, 2013). Linear models and Gaussian processes have extensively been studied to analyze the sensitivity since their posterior predictive distribution is expressed analytically (Fiedler et al., 2021; Lederer et al., 2019). Our result extends this relationship to general probabilistic models using the information-theoretic quantity for the first time. Such a relationship provides an important contribution in practice since some recent studies, such as Angelopoulos et al. (2021); Liu et al. (2020); Tian et al. (2021), and He et al. (2020), explicitly introduced the sensitivity property into deep neural networks to enhance the uncertainty quantification performance. Similarly to sensitivity, the CMI is widely used as the objective function in Bayesian experimental designs (Foster

et al., 2019). Thus, understanding the sensitivity of the CMI will lead to the analysis of such applications. In meta-learning tasks, information-theoretic quantities are widely used (Titsias et al., 2021; Chen et al., 2022) to quantify the similarity of tasks. In these applications, evaluating exact information-theoretic quantities is difficult for many practical models. Thus, various approximation methods have been proposed, including variational inference (Bishop, 2006). We leave it to future work to explore how the approximation quality affects the sensitivity in uncertainty.

The information-theoretic analysis has recently received attention in the generalization error analysis (Xu and Raginsky, 2017; Pensia et al., 2018). In such generalization error analysis, including the PAC-Bayesian theory (Sheth and Khardon, 2017; Alquier, 2021), the data generating distribution may not be well-specified, and model parameters are not treated as latent variables. Compared with previous studies, we can specify the correct model families in the Bayesian learning settings. The CMI of Theorem 3 in Aminian et al. (2023) is defined as the information between the super-sample index U and W conditioned on super-sample S. On the other hand, our CMI is defined as the information between the test data and W conditioned on training data, which is slightly different. However, we found that Theorem 1 in Aminian et al. (2023) corresponds to our Theorem 5, and both evaluate the generalization error using Lautum information.

Theorem 1 in Bu et al. (2023) closely aligns with our Theorem 5. While our Theorem 5 evaluates the generalization error in a non-meta-learning setting using Lautum information, Bu et al. (2023) demonstrates that similar results can also be obtained in a meta-learning setting. This suggests that we can extend our Theorem 5 to a meta-learning setting. However, such a specific derivation would significantly expand beyond the scope of our current study, so we leave it a future research.

Finally, we remark that the relatively finite variance assumption in Theorem 3 is equivalent to the Generalized Bernstein condition Watanabe (2018) and, thanks to this assumption, the fast learning rate $\mathcal{O}(1/N)$ can be achieved.

7 NUMERICAL EXPERIMENTS

We show the numerical evaluation of the sensitivities in Theorems 1 and 6. Detailed experimental settings and additional numerical results are shown in the Supplementary material.

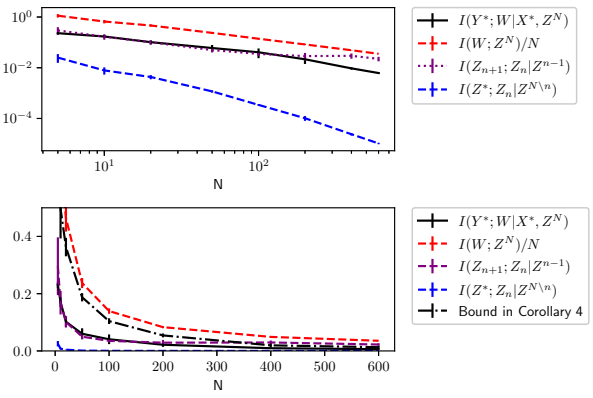

Figure 1: Information-theoretic quantities appearing in Theorem 1 and Corollary 1. Enlarged figures are provided in the Supplementary material.

7.1 Bayesian Learning Setting

First, using linear regression models introduced in Sec. 3.2, we numerically evaluated information-theoretic quantities appearing in Theorem 1, changing the training data size N. We can calculate all the information-theoretic quantities analytically. In the main paper, we only show the results of Gaussian basis functions as ϕ, whose dimension is set to 10.

The results are shown in Fig. 1, where we plot the CMI ($I(W;Y^*|X^*,Z^N)$), MI ($I(W;Z^N)/N$), and the sum of the test data sensitivity ($\sum_n I(Z^*;Z_n|Z^{(N\setminus n)})/N$) and the sum of the training data sensitivity ($\frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1};Z_n|Z^{(n-1)})$). In the figure legend, we omit the summation with respect to n and n' for clarity. In the left panel of this figure, we plot them in the log scale, and we can see that all the terms converge linearly in the plot. This is consistent with Eq. (20), which describes the asymptotic order of each quantity. Note that the sensitivity term I_n converges faster than the other terms, as indicated by the asymptotic analysis in Theorem 3. In the right panel of this figure, in addition to the CMI, MI, and sensitivities, we plot our proposed bound in Corollary 1. Our bound is tighter than the existing bound, which corresponds to $I(W;Z^N)/N$ owing to the sensitivity given as $I(Z_{n+1};Z_n|Z^{(n-1)})$. In the Supplementary material, we numerically evaluated the upper and lower bounds of I_n in Theorem 1.

7.2 Bayesian Meta-learning Setting

Next, we numerically evaluated the theoretical findings of meta-learning settings in Theorem 6. For this purpose, we put a hyperprior on the parameters of the linear regression model. We consider that

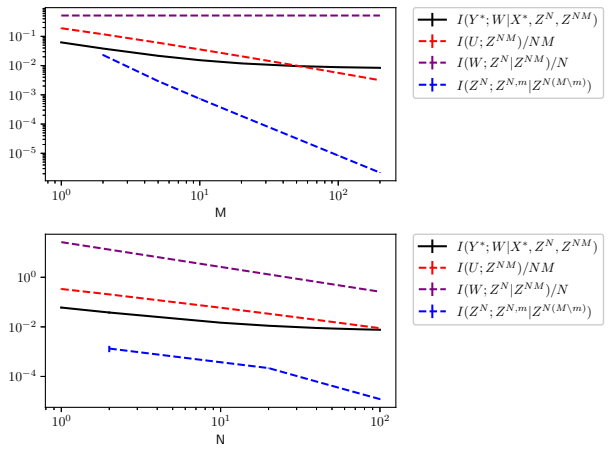

Figure 2: Information-theoretic quantities appearing in Theorem 6. The left panel shows the results under different M (the number of tasks) at fixed N (the number of training datasets). The right panel shows the results under different N at fixed M. Here, we omit the summation with respect to n, n', m, and m' for clarity.

$p(W|U) = \mathcal{N}(U, \alpha^{-1}I_d)$ and $p(U) = \mathcal{N}(0, \gamma^{-1}I_d)$. Under these settings, we can analytically calculate the posterior distributions $p(U|Z^{NM}), p(W|Z^N, U)$, and $p(W;Y^*|X^*, Z^N, Z^{NM})$. Thus, we can analytically evaluate the information-theoretic quantities in Theorem 6, see the Supplementary material for details.

The result is shown in Fig. 2. In the left panel of this figure, fixing $N = 50$, we plot the information-theoretic quantities with increasing M. We can see that MEMR (black line) decreases as we increase the number of meta-training datasets. We can also see that the MI between the hyperparameter and meta-training datasets (red dot line) also decreases. Finally, we can see that the sensitivity between the meta-test and meta-training tasks decreases faster than other information-theoretic quantities. In the right panel of this figure, fixing $M = 20$, we plot the information-theoretic quantities with increasing N. By increasing N, we find that all the quantities decrease as we expected.

8 CONCLUSION

In this work, we showed the novel decomposition of the CMI and then provided the information-theoretic quantity of the sensitivity between the test and training data points. Our analysis rigorously characterizes the uncertainty's widely believed sensitivity property for the first time. Our analysis is also extended to the meta-learning setting and showed the sensitivity between tasks for the first time. It will be interesting to analyze the sensitivity under model misspecification and approximation in future work.

Acknowledgements

FF was supported by JSPS KAKENHI Grant Number JP23K16948. FF was supported by JST, PRESTO Grant Number JPMJPR22C8, Japan.

References

Alquier, P. (2021). User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*.

Aminian, G., Bu, Y., Toni, L., Rodrigues, M., and Wornell, G. (2021). An exact characterization of the generalization error for the gibbs algorithm. In *Advances in Neural Information Processing Systems*.

Aminian, G., Bu, Y., Toni, L., Rodrigues, M. R., and Wornell, G. W. (2023). Information-theoretic characterizations of generalization error for the gibbs algorithm. *IEEE Transactions on Information Theory*.

Angelopoulos, A. N., Bates, S., Jordan, M., and Malik, J. (2021). Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bu, Y., Tetali, H. V., Aminian, G., Rodrigues, M., and Wornell, G. (2023). On the generalization error of meta learning for the gibbs algorithm. *arXiv preprint arXiv:2304.14332*.

Chen, Y., Zhang, S., and Kian Hsiang Low, B. (2022). Near-optimal task selection for meta-learning with mutual information and online variational bayesian unlearning. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9091–9113. PMLR.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Fiedler, C., Scherer, C. W., and Trimpe, S. (2021). Practical and rigorous uncertainty bounds for gaussian process regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7439–7447.

Foster, A., Jankowiak, M., Bingham, E., Horsfall, P., Teh, Y. W., Rainforth, T., and Goodman, N. (2019). Variational bayesian optimal experimental design.

Advances in Neural Information Processing Systems, 32.

Grünwald, P. (2012). The safe bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer.

Grünwald, P. and Van Ommen, T. (2017). Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Hafez-Kolahi, H., Moniri, B., Kasaei, S., and Baghshah, M. S. (2021). Rate-distortion analysis of minimum excess risk in Bayesian learning. In *International Conference on Machine Learning*, pages 3998–4007. PMLR.

He, B., Lakshminarayanan, B., and Teh, Y. W. (2020). Bayesian deep ensembles via the neural tangent kernel. *Advances in neural information processing systems*, 33:1010–1022.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27.

Janz, D., Hron, J., Mazur, P., Hofmann, K., Hernández-Lobato, J. M., and Tschiatschek, S. (2019). Successor uncertainties: exploration and uncertainty in temporal difference learning. *Advances in Neural Information Processing Systems*, 32.

Lederer, A., Umlauft, J., and Hirche, S. (2019). Posterior variance analysis of gaussian processes with application to average learning curves.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. (2020). Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512.

Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press.

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Hao, B., Ibrahimi, M., Lawson, D., Lu, X., O'Donoghue, B., and Roy, B. V. (2021). Evaluating predictive distributions: Does bayesian deep learning work?

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*.

Pensia, A., Jog, V., and Loh, P.-L. (2018). Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE.

Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.

Rosenfeld, N., Hilgard, A., Ravindranath, S. S., and Parkes, D. C. (2020). From predictions to decisions: Using lookahead regularization. *Advances in Neural Information Processing Systems*, 33:4115–4126.

Sheth, R. and Khardon, R. (2017). Excess risk bounds for the bayes risk using variational inference in latent gaussian models. In *Advances in Neural Information Processing Systems*.

Theresa Jose, S., Park, S., and Simeone, O. (2022). Information-theoretic analysis of epistemic uncertainty in bayesian meta-learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR.

Tian, J., Yung, D., Hsu, Y.-C., and Kira, Z. (2021). A geometric perspective towards neural calibration via sensitivity decomposition. *Advances in Neural Information Processing Systems*, 34.

Titsias, M. K., Ruiz, F. J. R., Nikoloutsopoulos, S., and Galashov, A. (2021). Information theoretic meta-learning with Gaussian processes. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research. PMLR.

Watanabe, S. (2018). *Mathematical theory of Bayesian statistics*. Chapman and Hall/CRC.

Wen, Z., Osband, I., Qin, C., Lu, X., Ibrahimi, M., Dwaracherla, V., Asghari, M., and Van Roy, B. (2021). From predictions to decisions: The importance of joint predictive distributions.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30.

Xu, A. and Raginsky, M. (2022). Minimum excess risk in bayesian learning. *IEEE Transactions on Information Theory*.

Ye, N. and Zhu, Z. (2018). Bayesian adversarial learning. In *Advances in Neural Information Processing Systems*.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, see Sec.2.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, see Sec.3.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, see the Supplementary material.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, see each theorem.]
   (b) Complete proofs of all theoretical results. [Yes, see the Supplementary material.]
   (c) Clear explanations of any assumptions. [Yes, see Sec.2, 3, 4.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes, see the Supplementary material.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, see the Supplementary material.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, see the Supplementary material.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, see the Supplementary material.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the the Supplementary material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

In Supplementary material, we express the test data as $Z = (Y, X)$ and drop the superscript $*$ for simplicity.

9 Proof of Theorem 1

Proof. First, we show the following lemma.

lmm 3. *Under the same setting as Theorem 1, for any $k \in (1, N]$, we have*

$$H[Z_{k+1}|Z^k] = H[Z_{k+1}|Z^{k-1}] - I(Z_{k+1}; Z_k|Z^{k-1}). \tag{31}$$

We can derive this by directly calculating the definition of entropy and conditional mutual information. This lemma states how much uncertainty at Z_{k+1} reduces by adding a data point Z_k. From this relation, we get the relation about the conditional mutual information as follows,

lmm 4. *Under the same setting as Theorem 1, for any $k \in (1, N]$, we have*

$$I(W; Z_{k+1}|Z^k) = I(W; Z_k|Z^{k-1}) - I(Z_{k+1}; Z_k|Z^{k-1}). \tag{32}$$

Proof. From the definition of conditional mutual information, we have

$$\begin{aligned}
I(W; Z_k|Z^{k-1}) &= H[Z_k|Z^{k-1}] - H[Z_k|W, Z^{k-1}] \\
&= H[Z_{k+1}|Z^{k-1}] - H[Z_k|W] \\
&= H[Z_{k+1}|Z^k] + I(Z_{k+1}, Z_k|Z^{k-1}) - H[Z_{k+1}|W] \\
&= H[Z_{k+1}|Z^k] + I(Z_{k+1}, Z_k|Z^{k-1}) - H[Z_{k+1}|W, Z^K] \\
&= I[W; Z_{k+1}|Z^k] + I(Z_{k+1}, Z_k|Z^{k-1}),
\end{aligned}$$

where the second equality is because the joint distribution of (Z^{k-1}, Z_k) is equivalent to the joint distribution of (Z^{k-1}, Z_{k+1}), and Z^{k-1} is independent of Z_k given W. The third inequality is because of Lemma 3. □

This relation explains how much information about W we obtain by adding a data point Z_k into the training dataset.

Finally, by using the chain rule of the mutual information and applying Lemma 4 recursively, we obtain

$$\begin{aligned}
I(W; Z^N) &= \sum_{n=1}^{N} I(W; Z_n|Z^{N-1}) \\
&= NI(W; Z_N|Z^{N-1}) + \sum_{n'=1}^{N-1} \sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{N-1})
\end{aligned} \tag{33}$$

We transform the first term. By definition, we have

$$\begin{aligned}
I(W; Z_N|Z^{N-1}) &= H[Z_N|Z^{N-1}] - H[Z_N|W] \\
&= \mathbb{E}_{p(W)p(Z^N|W)}[-\ln p(Z^N) + \ln p(Z^{N-1})] - H[Z_n|W].
\end{aligned}$$

Here $Z^{N-1} = (Z_1, \cdots, Z_{N-1})$. Let us consider the distribution of $Z^{N\backslash n} = (Z_1, \ldots, Z_{n-1}, Z_{n+1}, \ldots, Z_N)$. We express the marginal distribution of this as $p(Z^{N-1}) = \mathbb{E}_{p(W)}p(Z^{N-1}|W)$ and $p(Z^{N\backslash n}) = \mathbb{E}_{p(W)}p(Z^{N\backslash n}|W)$. Note that $p(Z^N|W)$ and $p(Z^{N\backslash n}|W)$ are equivalent since all the data is drawn i.i.d. Thus, from the chain rule of the KL divergence $\mathbb{E}_{p(W)p(Z^N|W)}[\ln p(Z^{N-1}) - \ln p(Z^{N\backslash n}|W)] = 0$. Thus we have

$$\begin{aligned}
I(W; Z_N|Z^{N-1}) &= H[Z_N|Z^{N-1}] - H[Z_N|W] \\
&= \mathbb{E}_{p(W)p(Z^N|W)}[-\ln p(Z^N) + \ln p(Z^{N-1})] - H[Z_n|W] \\
&= \mathbb{E}_{p(W)p(Z^N|W)}[-\ln p(Z^N) + \ln p(Z^{N\backslash n})] - H[Z_n|W] = I(W; Z_n|Z^{N\backslash n}).
\end{aligned}$$

Then we substitute this to Eq. (33)

$$I(W;Z^N) = NI(W;Z_N|Z^{N-1}) + \sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1},Z_n|Z^{N-1})$$
$$= \sum_{n=1}^{N} I(W;Z_n|Z^{N\setminus n}) + \sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1},Z_n|Z^{N-1}). \tag{34}$$

Finally, we apply Lemma 4 to each $I(W;Z_n|Z^{N\setminus n})$. Then, for each $n \in (1,N]$ we have

$$I(W;Z_{N+1}|Z^N) = I(W;Z_n|Z^{N-1}) - I(Z_{N+1};Z_n|Z^{N-1}). \tag{35}$$

This is because the training dataset $\{Z_n\}_{n=1}^N$ are i.i.d., and thus, we can permute the index of the training data points in Lemma 4. By setting $Z_{N+1} := Z$, which is the test data point, we obtain

$$I(W;Z^N) = \sum_{n=1}^{N} I(W;Z_n|Z^{N\setminus n}) + \sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1},Z_n|Z^{N-1})$$
$$= NI(W;Z|Z^N) + \sum_{n=1}^{N} I(Z,Z_n|Z^{N\setminus n}) + \sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1},Z_n|Z^{N-1}). \tag{36}$$

Finally, we rearrange this equation and divide both hand sides by N; we get the result. □

10 Proof of Theorem 2 in Linear Regression Model

Here we show the results of Bayesian linear regression again. A Bayesian linear regression model is given as $p(Y|x,w) = \mathcal{N}(w^\top\phi(x),\beta^{-1})$ where $\phi(x) := (\phi_1(x),\dots,\phi_d(x))^\top \in \mathbb{R}^d$ is a d-dimensional feature vector of the input x and each $\phi_i : \mathbb{R}^d \to \mathbb{R}$. We assume a prior distribution over W as $P(W) = \mathcal{N}(0,\alpha^{-1}I_d)$. We define a design matrix as $\Phi = (\phi(x_1),\dots,\phi(x_N))^\top \in \mathbb{R}^{N\times d}$. We also define a target vector as $\mathbf{Y} = (Y_1,\dots,Y_N)^\top$. Then a posterior distribution is given by $P(W|Z^N) = \mathcal{N}(m_N,S_N)$ where $m_N = \beta S_N\Phi^\top\mathbf{Y}$ and $S_N^{-1} := \alpha I_d + \beta\Phi^\top\Phi$. We also have a posterior predictive distribution as $P(Y|X,Z^N) := \mathcal{N}(m_N^\top\phi(x),\sigma_N^2(x))$ where $\sigma_N^2(x) := \beta^{-1} + \phi(X)^\top S_N\phi(X)$. Thus, the entropy of the posterior predictive distribution is given as

$$H(Z|Z^N) = H(X) + H(Y|X,Z^N) = H(X) + \mathbb{E}_{p(X)}\frac{1}{2}\ln\sigma_N^2(X) + \frac{1}{2}d(\ln(2\pi)+1). \tag{37}$$

Thus, we can calculate I_n as follows

$$I(Z,Z_n|Z^{N\setminus n}) = \mathbb{E}_{p(X)}\frac{1}{2}(\ln\sigma_{N\setminus n}^2(X) - \ln\sigma_N^2(X)). \tag{38}$$

Note that since $\sigma_N^2(X) \le \sigma_{N\setminus n}^2$ holds (Bishop, 2006), $I(Z,Z_n|Z^{N\setminus n})$ is always positive. Then we use the relation

$$\frac{x-y}{x+y} \le \frac{1}{2}(\ln x - \ln y) \le \frac{1}{2}\frac{x-y}{y}. \tag{39}$$

Using these relations, we have

$$\frac{1}{2}\frac{\sigma_{N\setminus n}^2(X)-\sigma_N^2(X)}{\alpha^{-1}\phi(x)^\top\phi(x)+\beta^{-1}} \le \frac{\sigma_{N\setminus n}^2(X)-\sigma_N^2(X)}{\sigma_N^2(X)+\sigma_{N\setminus n}^2(X)} \le \frac{1}{2}(\ln\sigma_{N\setminus n}^2(X) - \ln\sigma_N^2(X)) \le \frac{1}{2}\frac{\sigma_{N\setminus n}^2(X)-\sigma_N^2(X)}{\sigma_N^2(X)}. \tag{40}$$

Note that $S_N^{-1} := \alpha I_d + \beta\Phi^\top\Phi$ and $\Phi^\top\Phi = \sum_n \phi(X_n)\phi(X_n)^\top$. Thus, we have $S_N^{-1} = S_{N-1}^{-1} + \beta\phi(X_N)\phi(X_N)^\top$. From this relation, by using the Woodbury formula, we have

$$S_{N-1} = S_N + S_N\phi(\beta^{-1} - \phi(x_N)^\top S_N\phi(x_N))^{-1}\phi^\top S_N. \tag{41}$$

Then by definition, we have

$$\sigma_{N\setminus n}^2(X) - \sigma_N^2(X) = \phi(x)^\top(S_{N\setminus n} - S_N)\phi(x)). \tag{42}$$

Combining these results, we have

$$\sigma^2_{N\backslash n}(X) - \sigma^2_N(X) = \frac{(\phi(x)^\top S_N \phi(x_n))^2}{\beta^{-1} - \phi(x_n)^\top S_N \phi(x_n)}. \tag{43}$$

Note that $\beta^{-1} - \phi(x_n)^\top S_N \phi(x_n)$ is always positive. To confirm this, first, we regard this as the shur complement of the matrix

$$M = \begin{pmatrix} S_N^{-1} & \phi(x_n) \\ \phi(x_n)^\top & \beta^{-1} \end{pmatrix} \tag{44}$$

and use the fact that $\det(M) = \det(S_N^{-1})\det(\beta^{-1} - \phi(x_n)^\top S_N \phi(x_n))$. We can confirm that M is positive definite because we can show that $v^\top M v > 0$, where $v \in \mathbb{R}^{N+1}$ is an arbitrary vector, using the fact that S_N is positive definite. Then $\beta^{-1} - \phi(x_n)^\top S_N \phi(x_n)$ must be positive definite since M and S_N are positive definite. Thus, we have

$$\frac{(\phi(x)^\top S_N \phi(x_n))^2}{\beta^{-1} - \phi(x_n)^\top S_N \phi(x_n)} \geq (\phi(x)^\top S_N \phi(x_n))^2. \tag{45}$$

In conclusion, we have

$$\frac{1}{N}\sum_{n=1}^N I(Z, Z_n | Z^{N\backslash n}) \geq \mathbb{E}_{p(X)} \frac{1}{2N} \frac{1}{\alpha^{-1}\phi(x)^\top \phi(x) + \beta^{-1}} \sum_{n=1}^N \frac{(\phi(x)^\top S_N \phi(x_n))^2}{\omega(x_n)}, \tag{46}$$

where $\omega(x_n) := \beta^{-1} - \phi(x_n)^\top S_N \phi(x_n)$.

11 Proof of the Asymptotic Result in Theorem 3

We introduce the following definition,

$$\mathcal{G}_N(\alpha) := -\mathbb{E}_{W,Z^N,Z} \log \mathbb{E}_{\tilde{W}|Z^N} p(Z|W)^\alpha, \tag{47}$$

where $p(Z|W)^\alpha$ is the power of $\alpha \in \mathbb{R}$ of $p(Z|W)^\alpha$. We only consider the α such that $|\mathcal{G}_N(\alpha)| < \infty$. We also simplify the expression of the expectation since the distribution with which we take the expectation is clear. When $\alpha = 1$, this is the Bayesian risk in the Bayesian learning setting. We also define

$$\mathcal{T}_N(\alpha) := -\mathbb{E}_{W,Z^N} \frac{1}{N}\sum_{n=1}^N \log \mathbb{E}_{\tilde{W}|Z^N} p(Z_n|W)^\alpha. \tag{48}$$

Compared to $\mathcal{G}_N(\alpha)$, this corresponds to the training error.

We then have the following relation

$$\begin{aligned}
\mathcal{G}_{N-1}(\alpha = 1) &= -\mathbb{E}_{W,Z^{N-1},Z_N} \log \mathbb{E}_{\tilde{W}|Z^{N-1}} p(Z_N|W) \\
&= \mathbb{E}_{W,Z^{N-1},Z_N}\left[\log \frac{1}{\mathbb{E}_W \prod_{n=1}^N P_{Z_n|W}} \mathbb{E}_W\left[p(Z_N|W)^{-1}\prod_{n=1}^N p(Z_n|W)\right]\right] \\
&= \mathbb{E}_{W,Z^{N-1},Z_N} \log \mathbb{E}_{\tilde{W}|Z^N} p(Z_N|W)^{-1} \\
&= \mathbb{E}_{W,Z^{N-1},Z_N}\left[\frac{1}{N}\sum_{n=1}^N \log \mathbb{E}_{\tilde{W}|Z^N} p(Z_n|W)^{-1}\right] \\
&= -\mathcal{T}_N(\alpha = -1).
\end{aligned} \tag{49}$$

Recall that

$$I(W; Z|Z^N) = I(W; Z_N|Z^{N-1}) - I(Z, Z_N|Z^{N-1}), \tag{50}$$

then we have

$$I(Z, Z_N|Z^{N-1}) = \mathcal{G}_{N-1}(\alpha = 1) - \mathcal{G}_N(\alpha = 1), \tag{51}$$

thus we have

$$I(Z, Z_N|Z^{N-1}) = -\mathcal{T}_N(\alpha = -1) - \mathcal{G}_N(\alpha = 1). \tag{52}$$

Next we consider the Taylor expansion of $\mathcal{T}_N(\alpha)$ and $\mathcal{G}_N(\alpha)$ about α. This expansion is studied asymptotically in Watanabe (2018) rigorously. Using Theorem 3 in Watanabe (2018), we have

$$I(Z, Z_N|Z^{N-1}) = o\left(\frac{1}{N}\right), \tag{53}$$

where $o(\frac{1}{N})$ is little o.

Finally, from Theorem 1, we have

$$\frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}) = -I(W; Y|X, Z^N) + \frac{1}{N}I(W; Z^N) - \frac{1}{N}\sum_n I_n \leq \frac{1}{N}I(W; Z^N). \tag{54}$$

Thus $\frac{1}{N}I(W; Z^N) = O(1/N)$, we obtain Eq. (20).

Remark 1. *The finite variance assumption is satisfied for regular models. We say that a statistical model is regular if its map $w \to p(z|w)$ is one-to-one, which implies $p(z|w_1) = p(z|w_2) \Rightarrow w_1 = w_2$ and its Fisher information matrix is positive definite for arbitrary $w \in \mathcal{W}$. Intuitively, this means that when we have enough samples, the posterior distribution of a regular model converges to the Gaussian distribution.*

12 Proofs of Lemma 2 and Theorem 4

Proof. The conditional mutual information is rewritten as

$$\begin{aligned}
&I(W; Y|X, Z^N) \\
&= E_{W,Z^N,Z}\left[-\ln \mathbb{E}_{W|Z^N} p(Y|X, W) + \ln p(Y|X, W)\right] \\
&= E_{W,Z^N,Z}\left[-\ln \mathbb{E}_{W|Z^N} p(Y|X, W)\right] - H[Y|X, W].
\end{aligned} \tag{55}$$

On the other hand, the mutual information is rewritten as

$$\begin{aligned}
&I(W; Z^N) \\
&= \mathbb{E}_{Z^N,W}[-\ln p(Z^N) + \ln p(Z^N|W)] \\
&= -\mathbb{E}_{Z^N} \ln p(Z^N) - H[Z^N|W] \\
&= -\mathbb{E}_{Z^N}\mathbb{E}_{W|Z^N} \ln p(Z^N|W) + \mathbb{E}_{Z^N}\mathrm{KL}(p(W|Z^N)|p(W)) - H[Z^N|W].
\end{aligned} \tag{56}$$

This implies

$$-\mathbb{E}_{Z^N} \ln p(Z^N) = I(W; Z^N) + H[Z^N|W]. \tag{57}$$

Note that $H[Z^N|W] = NH[Z|W]$ since the test and training data points are i.i.d. By combining these results, we get Lemma 2. Moreover, combining these relations with Theorem 1, we get Theorem 4. □

12.1 Relation to Bayesian Regret

Here, we discuss how the Bayesian learning setting is related to Bayesian regret problems. Bayesian inference has been utilized for the sequential decision-making problem since we can incorporate the information from past observations into the posterior distribution. Here, we discuss how to utilize our developed theories for the sequential decision problem. We assume that, at each round n, we are given data X_n and predict Y_n using Z^{n-1}.

When using an optimal decision, we suffer from loss $R_l(Y_n|X_n, Z^{n-1}) - R_l(Y_n|X_n, W) = \mathrm{MER}_l(Y_n|X_n, Z^{n-1})$ at round n. We define an average cumulative regret as

$$\mathrm{Reg}_l(N) := \frac{1}{N}\sum_{n=1}^{N} \mathrm{MER}_l(Y_n|X_n, Z^{n-1}). \tag{58}$$

Under this definition, we obtain the following relation between regret and MER:

col 3.

$$\mathrm{MER}_{\log}(Y|X, Z^N) = \mathrm{Reg}_{\log}(N) - \frac{1}{N}\sum_{n=1}^{N} I_n - \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I_{n+1,n}. \tag{59}$$

Proof. Note that

$$-\ln p(Z_n|Z^{N-1}) = -\ln \mathbb{E}_{W|Z^{N-1}} p(Z_n|W) = -\ln p(Z^n) + \ln p(Z^{N-1}). \tag{60}$$

Thus, we have

$$\mathbb{E}_{Z^n}[-\ln p(Z_n|Z^{N-1})] = I(W; Z^n) - I(W; Z^{N-1}) + H. \tag{61}$$

where H is the entropy $H[Z|W]$. Here we omit the data point index from this entropy since all the entropy $H[Z_n|W]$ are equivalent since each Z_n are i.i.d. By adding from $n = 1$ to N, we can reformulate a regret as

$$E_{Z^N}\sum_{n=1}^{N}[-\ln p(Z_n|Z^{N-1})] = I(W; Z^N) + NH. \tag{62}$$

This concludes the proof. □

13 Proof of Theorem 5

We express the learned parameters as $\tilde{W}$, which follows the Bayesian posterior distribution $p(\tilde{W}|Z^N)$. Note that we have the Markov chain $W - Z^N - \tilde{W}$. Recall that $p(W|Z^N) := \frac{p(Z^N,W)}{p(Z^N)}$. Then we have

$$\begin{aligned} L(\tilde{W}; Z^N|W) &= \mathbb{E}_{p(W)p(\tilde{Z}^N|W)}\mathbb{E}_{p(\tilde{W}|\tilde{Z}^N)p(Z^N|W)}\left[\log \frac{p(Z^N|W)p(\tilde{W}|W)}{p(Z^N, \tilde{W}|W)}\right] \\ &= \mathbb{E}_{p(W)p(\tilde{Z}^N|W)}\mathbb{E}_{p(\tilde{W}|\tilde{Z}^N)p(Z^N|W)}\left[\log \frac{p(Z^N|W)}{p(Z^N|\tilde{W})}\right] \\ &= NR_{\log}^{\mathrm{Gibbs}} - NH[Y|X, W]. \end{aligned} \tag{63}$$

14 Proofs of Meta-learning in Theorem 6

The proof of the sensitivity in meta-learning is almost identical to that of Theorem 1. First, using Theorem 1 we can decompose the MEMR as follows.

$$\begin{aligned} &I(Y; W|X, Z^N, Z^{NM}) \\ &= \frac{I(W; Z^N|Z^{NM})}{N} - \frac{1}{N}\sum_{n=1}^{N} I(Z, Z_n|Z^{N\setminus n}, Z^{NM}) - \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}, Z^{NM}). \end{aligned} \tag{64}$$

Here we applied Theorem 1 to the meta-test test data and meta-test training dataset. Then we have

$$\begin{aligned} &I(Y; W|X, Z^N, Z^{NM}) \\ &= \frac{I(W, U; Z^N|Z^{NM})}{N} - \frac{1}{N}\sum_{n=1}^{N} I(Z, Z_n|Z^{N\setminus n}, Z^{NM}) - \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}, Z^{NM}) \\ &= \frac{I(W; Z^N|U)}{N} + \frac{I(U; Z^N|Z^{NM})}{N} - \frac{1}{N}\sum_{n=1}^{N} I(Z, Z_n|Z^{N\setminus n}, Z^{NM}) - \frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1} I(Z_{n+1}, Z_n|Z^{n-1}, Z^{NM}), \end{aligned} \tag{65}$$

where we used the fact that $I(U;Z^N|Z^{NM})=0$, which results from the Markov chain $(U,Z^{NM})-W-Z^N$. Next, we use the following two lemmas, the modified versions of Lemma 3 and 4, in the proof of Theorem 1.

lmm 5. *For any $m\in(1,2,\dots,M]$, we have*

$$H[Z^{N,(m+1)}|Z^{Nm}]=H[Z^{N,(m+1)}|Z^{N(m-1)}]-I(Z^{N,(m+1)};Z^{N,(m)}|Z^{N(m-1)}), \tag{66}$$

where $Z^{Nm}:=(Z^{N,(1)},\cdots,Z^{N,(m)})$ and $Z^{N(m-1)}:=(Z^{N,(1)},\cdots,Z^{N,(m-1)})$.

We can derive this by direct calculation. We also have the following lemma.

lmm 6. *For any $m\in(1,2,\dots,M]$*

$$I(U;Z^{N,(m+1)}|Z^{Nm})=I(U;Z^{N,(m)}|Z^{N(m-1)})-I(Z^{N,(m+1)};Z^{N,(m)}|Z^{N(m-1)}). \tag{67}$$

We can prove this almost in the same way as Lemma 4. Then, similarly to Eq. (33) in Theorem 1, we have

$$\begin{aligned}
I(U;Z^{NM})&=\sum_{m=1}^{M}I(U;Z^{N,(m)}|Z^{N(m-1)})\\
&=MI(W;Z^{N,(M)}|Z^{N(M-1)})+\sum_{m'=1}^{M-1}\sum_{m=m'}^{M-1}I(Z^{N,(m+1)},Z^{N,(m)}|Z^{N(m-1)})\\
&=MI(W;Z^{N}|Z^{NM})+\sum_{m=1}^{M}I(Z^{N},Z^{N,(m)}|Z^{N(M\backslash m)})+\sum_{m'=1}^{M-1}\sum_{m=m'}^{M-1}I(Z^{N,(m+1)},Z^{N,(m)}|Z^{N(m-1)}),
\end{aligned} \tag{68}$$

where $Z^{N(M\backslash m)}:=(Z^{N,(1)},\dots,Z^{N,(m-1)},Z^{N,(m+1)},\dots,Z^{N,(M)})$. Combining these results, we have

$$\begin{aligned}
&I(Y;W|X,Z^N,Z^{NM})\\
&=\frac{I(W;Z^N|U)}{N}+\frac{1}{NM}I(U;Z^{NM}) &(69)\\
&-\frac{1}{NM}\sum_{m=1}^{M}I(Z^{N},Z^{N,(m)}|Z^{N(M\backslash m)})-\frac{1}{NM}\sum_{m'=1}^{M-1}\sum_{m=m'}^{M-1}I(Z^{N,(m+1)},Z^{N,(m)}|Z^{N(m-1)}) &(70)\\
&-\frac{1}{N}\sum_{n=1}^{N}I(Z,Z_n|Z^{N\backslash n},Z^{NM})-\frac{1}{N}\sum_{n'=1}^{N-1}\sum_{n=n'}^{N-1}I(Z_{n+1},Z_n|Z^{n-1},Z^{NM}). &(71)
\end{aligned}$$

15 Numerical Experiments

Here we show a detailed explanation of the numerical experiments in the main paper and additional results. We conducted on all the numerical experiments with a CPU machine (Intel(R) Core(TM) i9-9980HK CPU)

15.1 Experiments in Bayesian Learning Setting

Here we explain the settings in Sec. 7.1. In Figure 3, we show the enlarged version of Figure 1 in the main paper.

We generated input $x\ \sim\mathcal{N}(0,\mathbf{1}_{d'})$. We used $d'=1$ in the paper. We then set $\alpha=\beta=1$ in all the experiments. As the feature map, we used the Gaussian feature map, defined as $\phi(x)=[e^{-\frac{1}{2}\|x-\mu_1\|^2},\cdots,e^{-\frac{1}{2}\|x-\mu_d\|^2}]$. We set $d=10$ and μ_1 to μ_d evenly in the interval $[-2,2]$.

Here we show additional numerical results, the upper and lower bounds of I_n derived in Theorem 2. The result is shown in Fig. 4. We can see that the upper bound is very accurate, while the lower bound is relatively loose.

Next, we show the results when using a polynomial feature map, $\phi(x)=[x,x^2,\cdots,x^5]$. We show the results in Fig. 5. We can see that the information-theoretic quantities behave similarly to when we use the Gaussian basis function for the feature map.

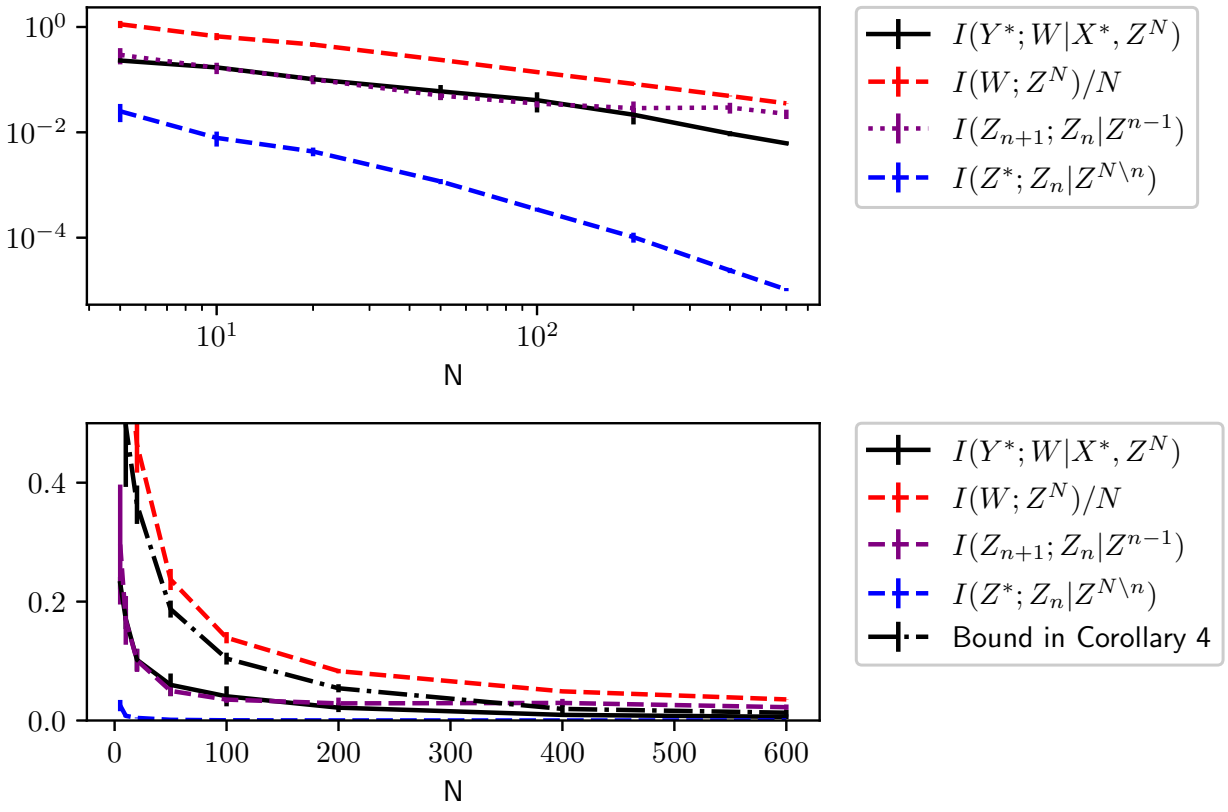

Figure 3: Information-theoretic quantities appearing in Theorem 1 and Corollary 1.

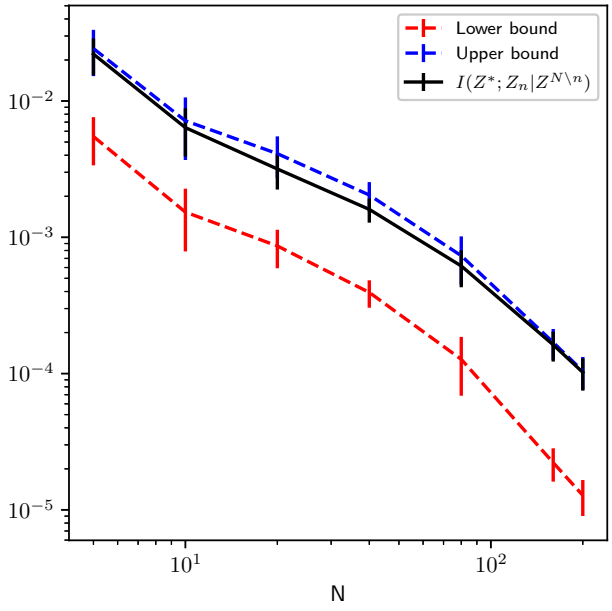

Figure 4: The upper and lower bounds of the sensitivity derived in Theorem 2.

15.2 Experiments in Bayesian Meta-learning Setting

Here we introduce the meta-learning settings of our numerical experiments. In Figure 6, we show the enlarged version of Figure 2 in the main paper.

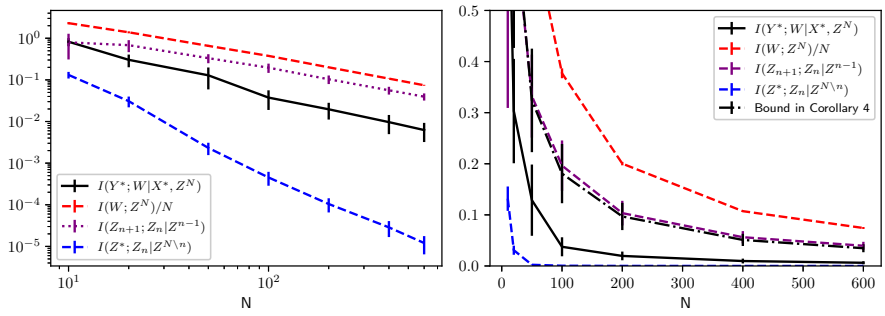

Figure 5: The information-theoretic quantities appeared in Theorem 1 and Corollary 1 for polynomial basis function.

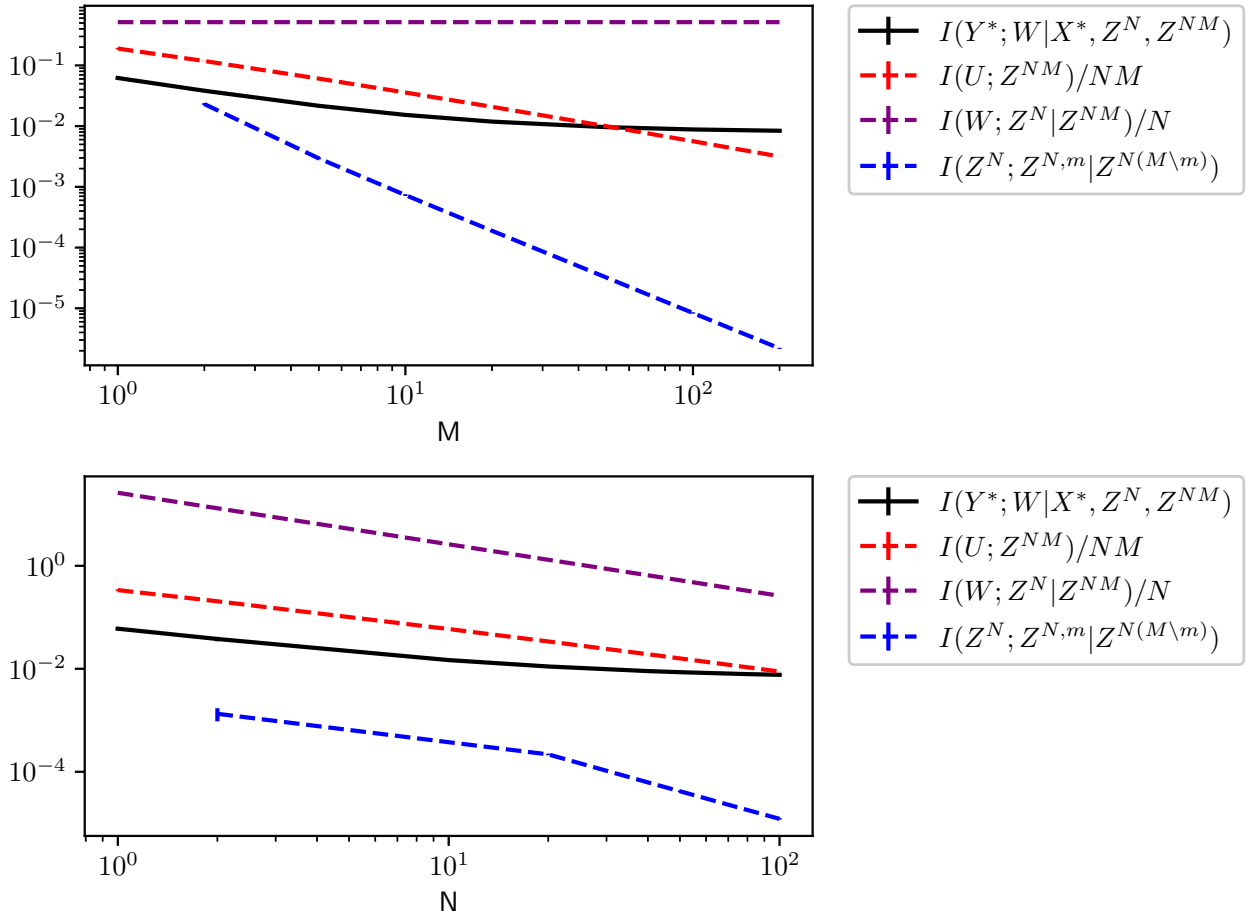

Figure 6: Information-theoretic quantities appearing in Theorem 6. The left panel shows the results under different M (the number of tasks) at fixed N (the number of training datasets). The right panel shows the results under different N at fixed M. Here, we omit the summation with respect to n, n', m, and m' for clarity.

For the prediction stage, we have the meta-test posterior predictive distribution as follows;

$$p(Y;W|X,Z^N,Z^{NM}) = \int p(Y|X,W)p(W|Z^N,Z^{NM})dW$$
$$= \int p(Y|X,W)p(W|Z^N,U)p(U|Z^{NM})dWdU. \quad (72)$$

Note that from the meta-training dataset, we can learn the posterior of the hyper-prior $p(U)$ as $p(U|Z^{NM})$. We can learn the posterior of the parameter of the meta-task as $p(W|Z^N, U)$.

We are focusing on the linear regression model, and we put on a Gaussian prior on W and Gaussian hyper-prior on U as $p(W|U) = \mathcal{N}(U, \alpha^{-1}\mathbf{1})$ and $p(U) = \mathcal{N}(0, \gamma^{-1}\mathbf{1})$ where $\mathbf{1}$ is the identity matrix. Thus, we can analytically calculate all the related posterior and predictive distributions. Thus, we can analytically calculate the CMI and MI in the information-theoretic bound.

We first present the explicit form of $p(W|Z^N, U)$. Recall the definitions. A linear regression model is given as, $p(Y|X, W) = \mathcal{N}(W^\top \phi(x), \beta^{-1})$ where $\phi_t(x) := (\phi_1(x), \dots, \phi_d(x))^\top \in \mathbb{R}^d$ is a d-dimensional feature vector of the input x of meta-test training dataset. We define a design matrix as $\Phi_t = (\phi(x_1), \dots, \phi(x_N))^\top \in \mathbb{R}^{N\times d}$. We also define a target vector as $\mathbf{y} = (y_1, \dots, y_N)^\top$.

lmm 7. *The posterior distribution of the meta-task parameter is given as Gaussian distribution,* $p(W|Z^N, U) = \mathcal{N}(m_t, S_t)$ *where*

$$m_t := S_t(\beta\Phi_t^\top \mathbf{y} + \alpha U), \tag{73}$$

$$S_t^{-1} := \alpha I_d + \beta\Phi_t^\top \Phi_t. \tag{74}$$

Proof. The joint distribution $p(U, Z^N, W)$ is Gaussian; thus, $p(W|Z^N, U)$ is Gaussian distribution. So we only need to calculate $\log p(U, Z^N, W)$ and focus on the quadratic and linear terms of W since those coefficients characterize the mean and variance of the Gaussian distribution. Then we have

$$-\log p(U, Z^N, W) = W^\top(\alpha I_d + \beta\Phi^\top\Phi)W + 2W^\top(\beta\Phi^\top \mathbf{y} + \alpha U) + \text{Const.} \tag{75}$$

By re-arranging this, we obtain the result. □

Next, we present the posterior distribution of hyperparameter U. We introduce the definition of the design matrix of the m-th meta-training dataset $\phi_m(x) := (\phi_1(x), \dots, \phi_d(x))^\top \in \mathbb{R}^d$ and a design matrix as $\Phi_m = (\phi_m(x_1), \dots, \phi_m(x_N))^\top \in \mathbb{R}^{N\times d}$. We use the same feature map for the meta-training and meta-testing stages. Under this setting, we have the following result.

lmm 8. *The posterior distribution of hyper-parameter is given as the Gaussian distribution,* $p(U|Z^{NM}) = \mathcal{N}(m_u, S_u)$ *where*

$$m_u := S_u \sum_{m=1}^{M} \mathbf{y}_m^\top s_m \Phi_m, \tag{76}$$

$$S_u^{-1} := \sum_{m=1}^{M} (\gamma\mathbf{1} + \Phi_m^\top s_m \Phi_m), \tag{77}$$

$$s_m^{-1} := \beta^{-1}\mathbf{1} + \alpha^{-1}\Phi_m\Phi_m^\top. \tag{78}$$

Proof. We consider the joint distribution for the meta-training as $p(U, Z^N M, W^M) = p(Z^{NM}|W^M)p(W^M|U)p(U)$. First, we can easily integrate out W from the joint distribution. Then, for the m-th task, $\int p(Z^{N,m}|W_m)p(W_m|U = u)dW_m$ is given as the Gaussian distribution $\mathcal{N}(\Phi_m u, \beta^{-1}\mathbf{1} + \alpha^{-1}\Phi_m\Phi_m^\top)$. Since the joint distribution $p(U, Z^{NM})$ is the Gaussian, $p(U|Z^{NM})$ is also the Gaussian distribution. Thus, we only need to calculate $\log p(U, Z^{NM})$ and focus on the quadratic and linear terms of U. Then we have

$$-\log p(U, Z^{NM}) = U^\top(\gamma + \sum_{m=1}^{M} \Phi_m^\top s_m \Phi_m)U + 2U^\top(\sum_{m=1}^{M} \Phi^\top s_m \mathbf{y}_m) + \text{Const}, \tag{79}$$

where $s_m^{-1} := \beta^{-1}\mathbf{1} + \alpha^{-1}\Phi_m\Phi_m^\top$ and $\mathbf{y}_m$ is the target variables of the m-th task. Thus, we obtain the result. □

Finally, we calculate the meta-test predictive distribution. This is also given as the Gaussian distribution.

lmm 9. *The posterior predictive distribution of the meta-test is given as the Gaussian distribution,* $p(Y; W|X, Z^N, Z^{NM}) = \mathcal{N}(m_f, S_f)$ *where*

$$m_f = \beta\mathbf{y}^\top\Phi_t S_t\phi(x) + \alpha m_u^\top S_t\phi(x), \tag{80}$$

$$S_f = \beta^{-1} + \phi(x)^\top S_t\phi(x)) + (\alpha S_t\phi(x))^\top S_u(\alpha S_t\phi(x)). \tag{81}$$

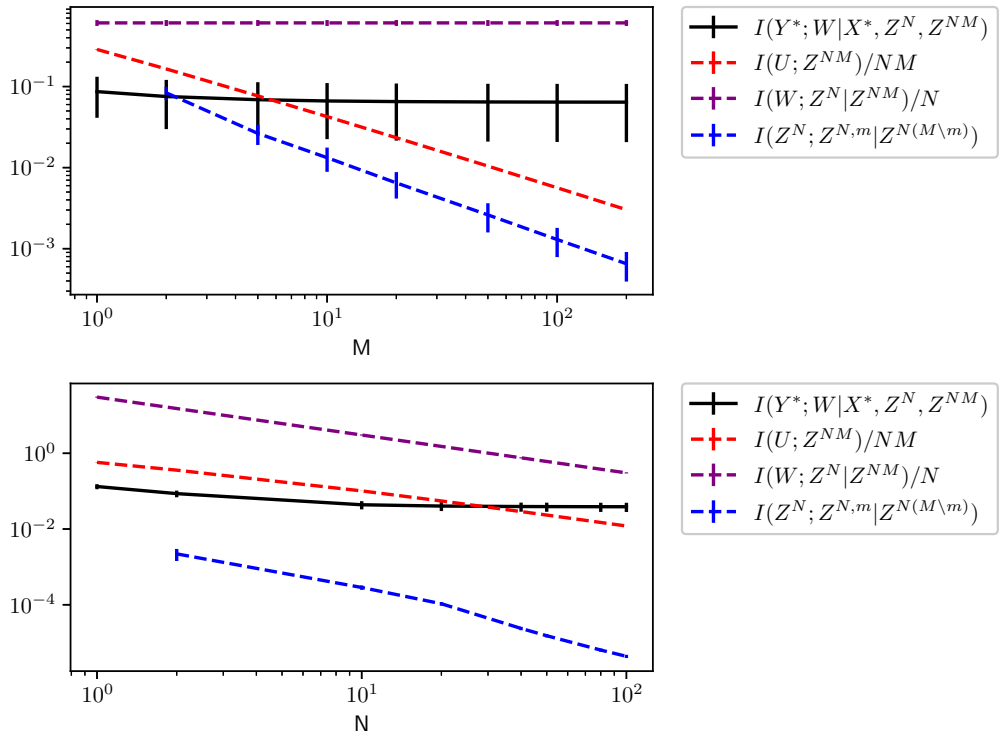

Figure 7: The information-theoretic quantities appeared in Theorem 6 for polynomial basis function.

Proof. We first calculate $\int p(Y|X,W)p(W|Z^N,U)p(U|Z^{NM})dW$. This is given as the Gaussian distribution $\mathcal{N}(m_t^\top \phi(x), \beta^{-1} + \phi(x)^\top S_t \phi(x))$, where we defined m_t and S_t in the above. Then applying the expectation formula of the Gaussian distribution, we have

$$m_f = \mathbb{E}_U[m_t^\top \phi(x)] = \mathbb{E}_U[(S_t(\beta \Phi_t^\top \mathbf{y} + \alpha U))^\top \phi(x)] = \beta \mathbf{y}^\top \Phi_t S_t \phi(x) + \alpha m_u^\top S_t \phi(x), \tag{82}$$

$$S_f = \beta^{-1} + \phi(x)^\top S_t \phi(x)) + (\alpha S_t \phi(x))^\top S_u (\alpha S_t \phi(x)). \tag{83}$$

□

As for the experimental settings, we set $\gamma = 1$. All the other settings are the same as the not meta-learning settings.

In the main paper, we showed the results of the Gaussian feature map. Here we show additional experimental results about the polynomial feature map. In Fig. 7, we show the results when the polynomial feature map, $\phi(x) = [x, x^2, \cdots, x^5]$ is used. We can see that the information-theoretic quantities behave similarly to when we use the Gaussian basis function for the feature map.