Information-theoretic Analysis of the Gibbs Algorithm: An Individual Sample Approach

Youheng Zhu
School of Computer Science and Technology
Huazhong University of Science and Technology
Wuhan, China
email: youhengzhu@hust.edu.cn

Yuheng Bu
ECE department
University of Florida
Gainesville, USA
email: buyuheng@ufl.edu

2024 IEEE Information Theory Workshop (ITW) | 979-8-3503-4893-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ITW61385.2024.10806933

***Abstract*—Recent progress has shown that the generalization error of the Gibbs algorithm can be exactly characterized using the symmetrized KL information between the learned hypothesis and the entire training dataset. However, evaluating such a characterization is cumbersome, as it involves a high-dimensional information measure. In this paper, we address this issue by considering individual sample information measures within the Gibbs algorithm. Our main contribution lies in establishing the asymptotic equivalence between the sum of symmetrized KL information between the output hypothesis and individual samples and that between the hypothesis and the entire dataset. We prove this by providing explicit expressions for the gap between these measures in the non-asymptotic regime. Additionally, we characterize the asymptotic behavior of various information measures in the context of the Gibbs algorithm, leading to tighter generalization error bounds. An illustrative example is provided to verify our theoretical results, demonstrating our analysis holds in broader settings.**

I. Introduction

One of the most important research topics in statistical learning theory is to capture the generalization behavior of the learning algorithms to avoid overfitting. Recently, [1], [2] proposed an information-theoretic approach to bound generalization error, where a learning algorithm is modeled as a randomized channel that takes the training dataset as input and outputs the learned hypothesis. In this setting, different information measures can be used to derive various non-trivial generalization error bounds, which capture all components in supervised learning, including the data-generating distribution, hypothesis class, and the learning algorithm itself. In comparison, traditional approaches such as VC-dimension [3], algorithmic stability [4], algorithmic robustness [5], and PAC-Bayesian bounds [6] cannot exploit all the aspects that affect the generalization performance.

After the seminal work [2], several approaches [7]–[15] have been proposed to refine information-theoretic generalization error bounds. Among them, a significant advancement is presented in [16], where the individual sample mutual information bound is introduced. By focusing on information measures involving individual samples, this bound is not only tighter but also simplifies the empirical estimation process, for instance, by using neural estimators like MINE [17]. In contrast, the mutual information-based bound in [2] depends on the mutual information between the hypothesis and the entire training dataset, making it nearly impossible to estimate with a large sample size n.

This paper explores a similar individual sample approach in the context of a specific learning algorithm, the Gibbs algorithm (formally defined in (7)). Such an algorithm can be interpreted as a randomized variant of the standard empirical risk minimization algorithm with mutual information as regularization, and it has other important connections to SGLD [18] and PAC-Bayesian bound [19]. More importantly, it has been shown in [20]–[22] that the generalization error of the Gibbs algorithm can be characterized exactly using the symmetrized KL information between the hypothesis and the entire dataset. Just like the mutual information-based bound, this exact characterization also suffers from the same drawback in practical evaluation due to its high dimensionality.

To address such an issue, in this paper, we study the individual sample information measures within the Gibbs algorithm and their counterparts involving the entire dataset. Our main contribution is an equivalency between the sum of symmetrized KL information between the output hypothesis and individual samples and that between the hypothesis and the entire dataset in the asymptotic regime $n \to \infty$. We also present other interesting properties of information measures in the individual sample context for both non-asymptotic and asymptotic regimes. In particular:

- In Section III, we provide an explicit expression of the gap between the sum of symmetrized KL information w.r.t individual samples and that w.r.t the entire dataset in the non-asymptotic regime.
- In Section IV, we precisely characterize the asymptotic behavior of different information measures for the Gibbs algorithm in terms of both convergence rate and constant factor. We then present our main theorem and additional results derived using similar techniques, which leads to a tighter bound on the generalization error.
- In Section V, an illustrative mean estimation example is provided to verify all theoretical results and demonstrate that our findings can hold in a more general setting.

979-8-3503-4893-4/24/$31.00 ©2024 IEEE

II. Preliminaries

In this section, we first introduce some background about information measures and the Gibbs algorithm.

A. Relevant Information Measures

Given two probability measures P and Q defined on the same probability space $(\Omega, \mathcal{F})$, the symmetrized Kullback-Leibler (KL) divergence is defined as

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

which symmetrizes the standard KL divergence $D(P\|Q)$. When $P \ll Q \ll P$ where $\ll$ denotes absolute continuity between measures, symmetrized KL divergence can be written as

$$D_{\mathrm{SKL}}(P\|Q) = \mathbb{E}_Q\left[\frac{dP}{dQ}\log\frac{dP}{dQ} - \log\frac{dP}{dQ}\right]. \tag{2}$$

It is natural to see that symmetrized KL divergence also belongs to the f-divergence family [23].

For two random variables X and Y, their mutual information is the KL divergence between their joint distribution and the product of the marginal distributions, i.e., $I(X;Y) \triangleq D(P_{X,Y}\|P_X \otimes P_Y)$. Similarly, we can define the symmetrized KL information as

$$\begin{aligned} I_{\mathrm{SKL}}(X;Y) &\triangleq D_{\mathrm{SKL}}(P_{X,Y}\|P_X \otimes P_Y) \\ &= I(X;Y) + L(X;Y), \end{aligned} \tag{3}$$

where $L(X;Y) \triangleq D(P_X \otimes P_Y \| P_{X,Y})$ represents lautum information [24].

B. Generalization Error in Supervised Learning

We denote $\mathcal{W}$ as the hypothesis class and $\mathcal{Z}$ as the instance space. A training dataset $S = \{Z_i\}_{i=1}^n \in \mathcal{S}$ with $Z_i \in \mathcal{Z}$ consists n samples drawn i.i.d from the data-generating distribution μ. A loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_0^+$ is used to measure the performance of a hypothesis on a sample Z. Therefore, we define the empirical and population risks associated with a given hypothesis w by

$$L_e(w,s) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i), \tag{4}$$

$$L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w,Z)], \tag{5}$$

respectively. In statistical learning, a learning algorithm can be modeled as a randomized mapping from the training set S onto a hypothesis $W \in \mathcal{W}$ according to the conditional distribution $P_{W|S}$. We define the expected generalization error quantifying the degree of over-fitting as

$$\begin{aligned} \mathrm{gen}(P_{W|S}, P_S) &\triangleq \mathbb{E}_{P_{W,S}}[L_\mu(W) - L_e(W,S)] \\ &= \mathbb{E}_{P_W \otimes P_S}[L_e(W,S)] - \mathbb{E}_{P_{W,S}}[L_e(W,S)], \end{aligned} \tag{6}$$

where the joint distribution $P_{W,S} = P_{W|S} \otimes P_S = P_{W|S} \otimes \mu^n$.

Following the framework proposed in [21], we focus on a specific learning algorithm $P_{W|S}$, i.e., the Gibbs algorithm (or Gibbs posterior [25]), which is defined as

$$P_{W|S}^{[n]}(w|s) \triangleq \frac{\pi(w)e^{-\gamma L_e(w,s)}}{V_{L_e}(s,\gamma)}. \tag{7}$$

Here, γ is the inverse temperature, $\pi(w)$ is an arbitrarily chosen prior distribution over $\mathcal{W}$, and $V_{L_e}(s,\gamma) \triangleq \int_{\mathcal{W}} \pi(w)e^{-\gamma L_e(w,s)}dw$ is the partition function that normalizes the distribution.

As shown in [20], [21], an important property of the Gibbs algorithm is that its generalization error can be exactly characterized using the symmetrized KL information:

$$\mathrm{gen}(P_{W|S}, P_S) = I_{\mathrm{SKL}}(W;S)/\gamma. \tag{8}$$

C. Other Notations

We will adopt the following notations to express the asymptotic scaling of quantities with n: $f(n) = O(g(n))$ represents that there exists a constant c s.t. $|f(n)| \le cg(n)$; $f(n) = \Theta(g(n))$ when there exist two constants $c_1 > 0$, $c_2 > 0$ s.t. $c_1 g(n) \le f(n) \le c_2 g(n)$; $f(n) = o(g(n))$ when $\lim_{n\to\infty}(f(n)/g(n)) = 0$; and $f(n) \sim g(n)$ when $\lim_{n\to\infty}(f(n)/g(n)) = 1$.

To simplify notation, we denote a probability measure or its corresponding probability density function by P_W when there is no ambiguity. We use $P_{W|Z^n}$ to represent the conditional probability density function, with the capital W, Z representing that it is also a random variable.

Throughout the paper, we will consider the Gibbs algorithm with a fixed inverse temperature γ and study its asymptotic behavior as the number of training samples $n \to \infty$. It is convenient for us to define the Gibbs algorithm using the population risk, i.e.,

$$P_W^\infty(w) \triangleq \frac{\pi(w)e^{-\gamma L_\mu(w)}}{\int_{\mathcal{W}} \pi(w)e^{-\gamma L_\mu(w)}dw}, \tag{9}$$

and the expectation of any measurable function $f(\cdot)$ under P_W^∞ is denoted as

$$\mathbb{E}_W^\infty[f(W)] \triangleq \int_{\mathcal{W}} P_W^\infty(w)f(w)dw. \tag{10}$$

III. Non-asymptotic Results

Motivated by the idea of using individual sample information measures proposed in [16], we first investigate the connection between the joint symmetrized KL information and its individual sample counterpart for the Gibbs algorithm.

The following theorem states that the difference between these two information measures can be characterized using the Jensen gap.

Theorem 1. *For joint distribution $P_{W,S}$ induced by the Gibbs algorithm, we have*

$$\begin{aligned} &\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S) \\ &= \sum_{i=1}^{n}\Big(\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W,Z_i)] - \mathbb{E}_{P_W\otimes P_{Z_i}}[J_i^{[n]}(W,Z_i)]\Big), \end{aligned} \tag{11}$$

where the Jensen gap $J_i^{[n]}(w, z_i)$ is defined as

$$J_i^{[n]}(w, z_i) \triangleq \log \int_{\mathcal{Z}^{n-1}} P^{[n]}_{W|S}(w|z_i, z^{-i}) d\mu^{n-1}(z^{-i}) \quad (12)$$
$$- \int_{\mathcal{Z}^{n-1}} \log \left(P^{[n]}_{W|S}(w|z_i, z^{-i})\right) d\mu^{n-1}(z^{-i}),$$

with $z^{-i} \triangleq \{z_1, \cdots, z_{i-1}, z_{i+1}, \cdots, z_n\}$.

All the proofs of the results presented in the paper can be found in [26]. We note that this theorem holds whenever the samples S are drawn independently but not necessarily identically generated from the distribution μ.

Remark 1. *As the* log *function is concave, the Jensen gap $J_i^{[n]}(w, z_i)$ is always non-negative. However, the RHS of* (11) *can be either negative or positive. An example showing that $I_{\mathrm{SKL}}(W; S)$ can be either larger or smaller than $\sum_{i=1}^n I_{\mathrm{SKL}}(W; Z_i)$ can be found in [20, Example 1].*

It is worth mentioning that the Jensen gap $J_i^{[n]}(w, z_i)$ in Theorem 1 has its own operational meaning by making the connection to the worst-case data-generating distribution introduced in [27]. A detailed discussion can be found in [26]. Other than this, interpreting this Jensen gap directly through finite sample analysis is challenging, prompting us to delve into the asymptotic regime in the next section.

IV. Asymptotic Results

In this section, we provide an asymptotic analysis of different information measures with i.i.d samples, e.g., the joint symmetrized KL information and its individual sample counterpart for the Gibbs algorithm.

A. Asymptotics of Individual Sample Information Measures

We start by rigorously defining the limiting probability space $(\mathcal{W} \times \mathcal{Z}^\infty, \mathcal{F}^\infty, P_W^\infty \otimes P_{Z^\infty})$ in the following definition.

Definition 1. *As the training data were i.i.d sampled from data distribution μ, there exists a filtered probability space $(\mathcal{Z}^\infty, \mathcal{F}_{Z^\infty}, \{\mathcal{F}^{[n]}_{Z^n}\}, P_{Z^\infty})$ where*

$$\mathcal{F}^{[n]}_{Z^n} = \sigma(Z_1, Z_2, \ldots, Z_n), \;\; \mathcal{F}_{Z^\infty} = \sigma\bigg(\bigcup_n \mathcal{F}^{[n]}_{Z^n}\bigg). \quad (13)$$

We define a probability space $(\mathcal{W}, \mathcal{B}, P_W^\infty)$ and the following product probability space

$$(\mathcal{W} \times \mathcal{Z}^\infty, \mathcal{F}^\infty, \{\mathcal{F}^{[n]}_{W,Z^n}\}, P_W^\infty \otimes P_{Z^\infty})$$
$$\triangleq (\mathcal{W}, \mathcal{B}, P_W^\infty) \times (\mathcal{Z}^\infty, \mathcal{F}_{Z^\infty}, \{\mathcal{F}^{[n]}_{Z^n}\}, P_{Z^\infty}). \quad (14)$$

For every sub-σ-algebra $\mathcal{F}^{[n]}_{Z^n}$, $P^{[n]}_{W,Z^n}$ is the probability measure induced by the Gibbs algorithm and the distribution of the dataset with size n, and $P^{[n]}_{W,Z_i}$ is the marginalization of $P^{[n]}_{W,Z^n}$.

Now, we are ready to study the asymptotic behavior of the Gibbs algorithm. We start by presenting the following two lemmas that capture the limit of the joint distribution $P^{[n]}_{W,Z^n}$.

Lemma 1. *For non-negative loss $\ell(w, z) \geq 0$, we have*

$$\lim_{n \to \infty} \left(\frac{dP^{[n]}_{W,Z^n}}{dP_W^\infty \otimes P_{Z^\infty}} \right) = 1 \;\; a.s. \quad (15)$$

Lemma 2. *For non-negative loss $\ell(w, z) \geq 0$ and any individual sample Z_i, we have*

$$\lim_{n \to \infty} \left(\frac{dP^{[n]}_{W,Z_i}}{dP_W^\infty \otimes P_{Z^\infty}} \right) = 1 \;\; a.s. \quad (16)$$

These two lemmas rigorously confirm the intuition that as $n \to \infty$, the asymptotic joint distribution $P^{[n]}_{W,Z^n}$ will converge to a product measure $P_W^\infty \otimes P_{Z^\infty}$, i.e., the learned hypothesis W depends solely on the data distribution μ and is independent of the dataset S. It is worth mentioning that this result is widely applicable, as it only requires the loss function to be non-negative or lower-bounded.

Corollary 1. *If we further assume that the loss function is bounded, i.e., $\ell(w, z) \in [0, C]$, we have that $\big(\frac{dP^{[n]}_{W,Z^n}}{dP_W^\infty \otimes P_{Z^\infty}}\big)$ and $\big(\frac{dP^{[n]}_{W,Z_i}}{dP_W^\infty \otimes P_{Z^\infty}}\big)$ are both uniformly bounded. Furthermore, $\lim_{n \to \infty} \big(\frac{dP^{[n]}_{W,Z_i}}{dP_W^\infty \otimes P_{Z^\infty}}\big) = 1$ almost surely.*

In the following, we will focus on the bounded loss function case. We already know that $dP_W^{[n]} \otimes P_{Z_i} / dP^{[n]}_{W,Z_i}$ converges to 1 as $n \to \infty$, and the following lemma characterizes the exact rate of such convergence.

Lemma 3. *If the loss function $\ell(w, z)$ is bounded, we have*

$$\lim_{n \to \infty} n \cdot \left(1 - \frac{dP_W^{[n]} \otimes P_{Z_i}}{dP^{[n]}_{W,Z_i}}\right) \quad (17)$$
$$= -\gamma[\ell(W, Z_i) - L_\mu(W)] + \mathbb{E}_W^\infty[\gamma(\ell(W, Z_i) - L_\mu(W))].$$

Additionally, $n \cdot \left(1 - \frac{dP_W^{[n]} \otimes P_{Z_i}}{dP^{[n]}_{W,Z_i}}\right)$ is uniformly bounded.

Built upon this Lemma, we provide the following theorem that characterizes the convergence rate of $I_{\mathrm{SKL}}(W; Z_i)$ with a tight constant factor as $n \to \infty$.

Theorem 2. *If the loss function $\ell(w, z)$ is bounded, we have*

$$I_{\mathrm{SKL}}(W; Z_i) \sim \frac{\gamma^2}{n^2} \mathbb{E}_\mu \bigg[\mathbb{E}_W^\infty \big[(\ell(W, Z) - L_\mu(W))^2\big]$$
$$- \mathbb{E}_W^\infty \big[(\ell(W, Z) - L_\mu(W))\big]^2 \bigg]. \quad (18)$$

The constant on the right-hand side of (18) can be interpreted as the variance of $\ell(W, Z) - L_\mu(W)$, which is always non-negative. It is also strictly positive unless $\ell(w, z)$ is a constant for every fixed w.

The proof of Lemma 3 and Theorem 2 mainly use the strong law of large numbers and the dominated convergence theorem, and more details can be found in [26]. As shown in the following corollaries, the same technique can also be applied to other information measures, specifically, the χ^2 information.

Corollary 2. *The χ^2 information has the similar rate if the loss function $\ell(w,z)$ is bounded, i.e.,*

$$I_{\chi^2}(W;Z_i) = \Theta\Big(\frac{1}{n^2}\Big), \tag{19}$$

furthermore,

$$I_{\chi^2}(W;Z_i) \sim I_{\mathrm{SKL}}(W;Z_i). \tag{20}$$

Corollary 3. *If the loss function $\ell(w,z)$ is bounded, we have*

$$I(W;Z_i) = O\Big(\frac{1}{n^2}\Big). \tag{21}$$

Remark 2. *Corollary 3 is directly obtained using Theorem 2 and the fact that $I(W;Z_i) \leq I_{\mathrm{SKL}}(W;Z_i)$. However, if we directly use the bounding technique used in Theorem 2 for mutual information, it yields a weaker conclusion $I(W;Z_i) = o\big(\frac{1}{n}\big)$. One possible reason is that mutual information corresponds to f-divergence with $f(x) = x\log x$, which is not consistently positive for all $x > 0$. On the other hand, $f(x) = x\log x - \log x$ for symmetrized KL information, which is always non-negative. Therefore, swapping the expectation and the limit in the proof of Theorem 2 will have a smaller impact on the analysis, leading to a more refined characterization. The same argument applies to χ^2 divergence as well.*

B. Asymptotics of the Gap

In this subsection, we focus on the gap between the sum of $I_{\mathrm{SKL}}(W;Z_i)$ and $I_{\mathrm{SKL}}(W;S)$ in the asymptotic regime. Our goal is to prove that this gap converges to zero faster than $I_{\mathrm{SKL}}(W;S)$, i.e., the generalization error itself. We begin by presenting two lemmas that capture the asymptotic behaviors of the Jensen gap $J_i^{[n]}(w,z_i)$ defined in Thorem 1.

Lemma 4. *If the loss function $\ell(w,z)$ is bounded, there exists a sequence of functions $\hat{J}^{[n]}(w)$ independent of z_i such that*

$$\lim_{n\to\infty} n\cdot(\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i)) = 0. \tag{22}$$

Furthermore, $n\cdot(\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i))$ is uniformly bounded.

This result shows that despite $J_i^{[n]}(w,z_i)$ is a function of both w and z_i, the influence of z_i is relatively negligible when n goes to infinity.

Lemma 5. *If the loss function $\ell(w,z)$ is bounded, the $\hat{J}^{[n]}(w)$ introduced in Lemma 4 satisfying $n\cdot\hat{J}^{[n]}(w)$ is uniformly bounded. Furthermore, $\lim_{n\to\infty} n\cdot\hat{J}^{[n]}(w)$ exists.*

Equipped with these technical lemmas, we present the main theorem of the paper, which shows that the gap between $I_{\mathrm{SKL}}(W;S)$ and the sum of $I_{\mathrm{SKL}}(W;Z_i)$ converges to zero faster than $\frac{1}{n}$.

Theorem 3. *If the loss function $\ell(w,z)$ is bounded, we have*

$$\sum_{i=1}^n I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S) = o\Big(\frac{1}{n}\Big). \tag{23}$$

It is worth pointing out that the key for proving Theorem 3 is Lemma 4, which indicates that the effect of terms involving variable z_i is order-wise small compared to the remaining terms. Utilizing this result, we apply the dominated convergence theorem on $n\cdot\big(\sum_{i=1}^n I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S)\big)$. More proof details can be found in [26].

From Theorem 3, we can immediately obtain the following corollary.

Corollary 4. *If the loss function $\ell(w,z)$ is bounded, we have*

$$I_{\mathrm{SKL}}(W;S) = \Theta\Big(\frac{1}{n}\Big). \tag{24}$$

More specifically,

$$I_{\mathrm{SKL}}(W;S) \sim \sum_{i=1}^n I_{\mathrm{SKL}}(W;Z_i). \tag{25}$$

The result in Corollary 4 aligns with the conclusion in [28], which states that for a Gibbs algorithm, when the loss function $\ell \in [0,1]$,

$$|\mathrm{gen}(P_{W|S}^{[n]}, P_S)| \leq \frac{\gamma}{2n}. \tag{26}$$

Remark 3. *As a sanity check, we look at a simple coin-tossing example, where $w \in \{0,1\}$ and $z \in \{0,1\}$, $\ell(w,z) = \mathbb{1}_{w=z}$ is a zero-one loss, and $\pi(w)$ is uniform over $\{0,1\}$. From Corollary 4, the convergence behavior of $I_{\mathrm{SKL}}(W;S)$ and thus $\mathrm{gen}(P_{W|S}^{[n]}, P_S)$ can be calculated as*

$$\lim_{n\to\infty} n\cdot\mathrm{gen}(P_{W|S}^{[n]}, P_S) = \frac{\gamma}{4}, \tag{27}$$

which indicates that the $\gamma/2n$ bound in (26) is not tight.

From Corollary 4, it is easy to see that $I(W;S) \leq I_{\mathrm{SKL}}(W;S) = \Theta\big(\frac{1}{n}\big)$, indicating $I(W;S) = O\big(\frac{1}{n}\big)$. With the same argument, the lautum information also satisfies that $L(W;S) = O\big(\frac{1}{n}\big)$. However, it is not clear which quantity contributes more to the generalization error of the Gibbs algorithm. The following theorem answers the question by proving that the two information measures equal each other asymptotically.

Theorem 4. *If the loss function $\ell(w,z)$ is bounded, we have*

$$\lim_{n\to\infty} n\cdot I(W;S) = \lim_{n\to\infty} n\cdot L(W;S) = \frac{1}{2}\lim_{n\to\infty} n\cdot I_{\mathrm{SKL}}(W;S).$$

In other words, $I(W;S) \sim L(W;S) = \Theta\big(\frac{1}{n}\big)$.

The proof technique of Theorem 4 differs from those used in Lemma 3 and Theorem 2. Here, our idea is to sandwich the target quantity between an upper bound and a lower bound that differ only in the third or higher-order terms. We then prove that the two bounds converge to the same value, characterized by the second-order term. Instead of using the dominated convergence theorem as in Lemma 3 and Theorem 2, we directly analyze the integral of the second-order terms for any n before taking the limit. In this process, the independence of the samples plays a crucial role, ensuring that all interaction terms are zero.

Using Theorem 4, we can provide an alternative proof for Corollary 3 by applying the Proposition 2 of [16], i.e.,

$$I(W;S) \geq \sum_{i=1}^{n} I(W;Z_i). \tag{28}$$

We provide the following result to showcase Theorem 4, which tightens the existing generalization error bound for the Gibbs algorithm.

Theorem 5. *For Gibbs algorithm with bounded loss function* $\ell(w,z) \in [a,b]$*,* $\forall \delta > 0$*, there exist an* $N \in \mathbb{N}^+$ *such that* $\forall n > N$*,*

$$0 \leq \mathrm{gen}(P_{W|S}^{[n]}, P_S) \leq \frac{(b-a)^2\gamma}{(4-\delta)n}. \tag{29}$$

This theorem provides a tighter bound compared with (26). Revisiting the coin-tossing example, it is interesting to see that this bound is asymptotically tight in this circumstance.

C. Comparison with Asymptotics of Model Capacity

We would like to compare our result with the asymptotic model capacity studied in [29], [30]. Different from our setting, they considered data $Y^n \in \mathcal{Y}^n$ that are drawn i.i.d from $P_{Y|X}(y|x)$, where $X \in \mathcal{X} \subset \mathbb{R}$ denotes the model parameter from certain model family. For such a Bayesian setting, the model parameter X is modeled using a prior distribution $P_X(x)$. If the model $P_{Y|X}$ is sufficiently smooth in X, then

$$I(X;Y^n) = \frac{1}{2}\log\frac{n}{2\pi e} - D(P_X\|P_X^*) - \log\int_{\mathcal{X}}\sqrt{J(x')}dx' + o(1), \tag{30}$$

where

$$J(x) = \mathbb{E}_{P_{Y|X=x}}\left[\left(\frac{\partial}{\partial x}\log P_{Y|X}(Y|x)\right)^2\right], \tag{31}$$

and P_X^* denotes the least informative prior, i.e., Jeffery's prior [31]. The mutual information is maximized when $P_X = P_X^*$, i.e., P_X^* is the capacity achieving distribution. We can see the growing rate of mutual information is $I(X;Y^n) = O(\log n)$, which is different from the asymptotic result $I(W;S) = \Theta(\frac{1}{n})$ in Theorem 4.

The difference between our setting and the model capacity setting is two-fold: 1) we considered i.i.d samples from data distribution μ, while the model capacity setting considers samples conditionally independent generated from $P_{Y|X}$; 2) In our setting, the channel $P_{W|Z^n}$ is the Gibbs algorithm defined using a bounded loss function $\ell(w,z)$, so that conditional on W, the samples are not independent to each other anymore. However, the posterior $P_{X|Y^n}$ in the model capacity setting is induced by the prior P_X and the likelihood $P_{Y|X}$ given the conditional independent structure among the samples.

Note that the assumption $\ell(W,Z_i)$ being bounded is a sufficient condition for our previous results and is adopted to circumvent the technical difficulty of exchanging the order of integration and Taylor expansion as well as the order of integration and limits.

V. Example

In this section, we will consider an example beyond the bounded loss function. It can be shown that most of our conclusions still hold under this setting.

Estimating the mean. Let $S = \{Z_i\}_{i=1}^n$ be the training set, where Z_i is a d dimensional vector sampled i.i.d. from $\mu = \mathcal{N}(0_d, (\frac{1}{\sqrt{2\beta}})^2\boldsymbol{I}_d)$. We consider the problem of learning the means of the distribution μ. For simplicity, we consider $d = 1$. We adopt square error $\ell(w,Z) \triangleq \|w - Z\|_2$ as the loss function and choose our prior distribution to be $\pi(w) = \frac{1}{\sqrt{\pi}}\exp(-w^2)$.

In this simple example, we can calculate the joint symmetrized KL information between S and W as

$$\gamma\mathrm{gen}(P_{W|S},\mu) = I_{\mathrm{SKL}}(W;S) = \frac{\gamma^2}{n\beta(1+\gamma)}, \tag{32}$$

and the individual sample symmetrized KL information between W and Z_i

$$I_{\mathrm{SKL}}(W;Z_i) = \frac{\gamma^2}{n^2\beta(1+\gamma)+\gamma^2(n-1)}. \tag{33}$$

From (32) and (33), we get

$$\begin{aligned}\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S) &= \Theta\left(\frac{1}{n^2}\right) \\ &= o(I_{\mathrm{SKL}}(W;S)),\end{aligned} \tag{34}$$

which shows that Theorem 3 still holds, despite the fact we are not considering a bounded loss function.

We further investigate the Jensen gap term $J_i(w,z_i)$, since as stated previously, the key to prove Theorem 3 is that the effect of variable z_i is order-wise smaller than that of w. In estimating the mean problem, we can calculate that

$$\begin{aligned}J_i^{[n]}(w,z_i) =& w^2\Theta\left(\frac{1}{n}\right) + wz_1\Theta\left(\frac{1}{n^2}\right) \\ &+ z_1^2\Theta\left(\frac{1}{n^3}\right) + \Theta\left(\frac{1}{n^2}\right)\end{aligned} \tag{35}$$

where all terms represented by big O notation were uniformly small. We can see that the contribution of z_i term is $\Theta\big(\frac{1}{n^2}\big)$, which is indeed order-wise smaller than those terms not containing z_i.

Finally, we provide a similar analysis of mutual information.

$$\begin{aligned}I(W;Z_i) &= \frac{1}{2}\log\left(1 + \frac{\gamma^2}{n^2(1+\gamma)\beta + (n-1)\gamma^2}\right) \\ &= \Theta\left(\frac{1}{n^2}\right),\end{aligned} \tag{36}$$

and

$$\begin{aligned}I(W;S) &= \frac{1}{2}\log\left(1 + \frac{\gamma^2}{n\beta(1+\gamma)}\right) \\ &\sim \frac{1}{2}\cdot\frac{\gamma^2}{n\beta(1+\gamma)} \\ &= \frac{1}{2}I_{\mathrm{SKL}}(W;S),\end{aligned} \tag{37}$$

which corresponds to our result in Theorem 4.

References

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

[2] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Advances in neural information processing systems*, vol. 30, 2017.

[3] V. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[4] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[5] H. Xu and S. Mannor, "Robustness and generalization," *Machine learning*, vol. 86, pp. 391–423, 2012.

[6] D. A. McAllester, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[7] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[8] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 824–839, 2020.

[9] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, "Conditioning and processing: Techniques to improve information-theoretic generalization bounds," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16457–16467, 2020.

[10] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9925–9935, 2020.

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

[12] M. Haghifam, G. K. Dziugaite, S. Moran, and D. Roy, "Towards a unified information-theoretic framework for generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26370–26381, 2021.

[13] H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan, "Information-theoretic generalization bounds for black-box learning algorithms," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24670–24682, 2021.

[14] G. Aminian, Y. Bu, G. W. Wornell, and M. R. Rodrigues, "Tighter expected generalization error bounds via convexity of information measures," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 2481–2486, IEEE, 2022.

[15] R. Zhou, C. Tian, and T. Liu, "Stochastic chaining and strengthened information-theoretic generalization bounds," *Journal of the Franklin Institute*, vol. 360, no. 6, pp. 4114–4134, 2023.

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

[17] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, pp. 531–540, PMLR, 10–15 Jul 2018.

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

[19] N. Thiemann, C. Igel, O. Wintenberger, and Y. Seldin, "A strongly quasiconvex PAC-Bayesian bound," in *International Conference on Algorithmic Learning Theory*, pp. 466–492, PMLR, 2017.

[20] G. Aminian, Y. Bu, L. Toni, M. R. Rodrigues, and G. W. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8106–8118, 2021.

[21] G. Aminian, Y. Bu, L. Toni, M. R. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," *IEEE Transactions on Information Theory*, 2023.

[22] H. Chen, G. W. Wornell, and Y. Bu, "Gibbs-based information criteria and the over-parameterized regime," in *International Conference on Artificial Intelligence and Statistics*, pp. 4501–4509, PMLR, 2024.

[23] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

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

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

[26] Y. Zhu and Y. Bu, "Information-theoretic analysis of the gibbs algorithm: An individual sample approach," *arXiv preprint arXiv:2410.12623*, 2024.

[27] X. Zou, S. M. Perlaza, I. Esnaola, E. Altman, and H. V. Poor, "The worst-case data-generating probability measure in statistical learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 175–189, 2024.

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

[29] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information theory*, vol. 30, no. 4, pp. 629–636, 1984.

[30] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 3, pp. 453–471, 1990.

[31] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[32] D. R. Brillinger, "A note on the rate of convergence of a mean," *Biometrika*, vol. 49, no. 3/4, pp. 574–576, 1962.

[33] H. P. Rosenthal, "On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables," *Israel Journal of Mathematics*, vol. 8, pp. 273–303, 1970.