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

***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.

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

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

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

See Appendix A for the proof. 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 [26]. A detailed discussion can be found in Appendix A. 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 $(\mathcal{Z}^{\infty}, \mathcal{F}_{Z^{\infty}}, \{\mathcal{F}_{Z^n}^{[n]}\}, P_{Z^{\infty}})$ where*

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

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

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

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

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 $P_{W,Z^n}^{[n]}$.

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

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

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

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

These two lemmas rigorously confirm the intuition that as $n \to \infty$, the asymptotic joint distribution $P_{W,Z^n}^{[n]}$ 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 $\left(\frac{dP_{W,Z^n}^{[n]}}{dP_W^{\infty} \otimes P_{Z^{\infty}}}\right)$ and $\left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^{\infty} \otimes P_{Z^{\infty}}}\right)$ are both uniformly bounded. Furthermore, $\lim_{n\to\infty}\left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^{\infty} \otimes P_{Z^{\infty}}}\right) = 1$ almost surely.*

In the following, we will focus on the bounded loss function case. We already know that $dP_W^{[n]} \otimes P_{Z_i} / dP_{W,Z_i}^{[n]}$ 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*

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

Additionally, $n \cdot \left(1 - \frac{dP_W^{[n]} \otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right)$ 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*

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

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 Appendix B. 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.,*

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

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*

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

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 $I(W;Z_i) = o\left(\frac{1}{n}\right)$. 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*

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

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 $n \cdot \left(\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S)\right)$. More proof details can be found in Appendix B.

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

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

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

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 [27], 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 $I(W;S) \leq I_{\mathrm{SKL}}(W;S) = \Theta\left(\frac{1}{n}\right)$, indicating $I(W;S) = O\left(\frac{1}{n}\right)$. With the same argument, the lautum information also satisfies that $L(W;S) = O\left(\frac{1}{n}\right)$. 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, $I(W;S) \sim L(W;S) = \Theta\left(\frac{1}{n}\right)$.

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 [28], [29]. 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 [30]. 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. A more detailed elaboration of the example can be found in Appendix C.

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 $\Theta(\frac{1}{n^2})$, 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] 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.

[27] 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.

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

[29] 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.

[30] 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.

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

[32] 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.

Appendix A
Non-asymptotic Results

Notation 1. *For simplicity, we use*

$$\mathbb{E}_e^{[n]}[f(W)] \triangleq \frac{\int_{\mathcal{W}} \pi(w) e^{-\gamma \frac{1}{n}\sum_{i=1}^n \ell(w,z_i)}(f(w))dw}{\int_{\mathcal{W}} \pi(w) e^{-\gamma \frac{1}{n}\sum_{i=1}^n \ell(w,z_i)}dw}. \tag{38}$$

A. Proof of Theorem 1

Proof. First from the definition, we have

$$I_{\mathrm{SKL}}(W;Z_i) = \mathbb{E}_{P_{W,Z_i}}[\log(P_{W|Z_i})] - \mathbb{E}_{P_W\otimes P_{Z_i}}[\log(P_{W|Z_i})]. \tag{39}$$

Using z^{-i} to represent $\{z_j\}_{j\neq i}$, the inner term $\log(P_{W|Z_i})$ can be represented by:

$$\log(P_{W|Z_i}) = \log\left(\int_{\mathcal{Z}^{n-1}} P_{W|Z_i,z^{-i}} d\mu^{n-1}(z^{-i})\right). \tag{40}$$

Notice that this holds true only when Z_is are sampled independently.

Our aim is to prove

$$\sum_{i=1}^n I_{\mathrm{SKL}}(W;Z_i) = \sum_{i=1}^n (\mathbb{E}_{P_{W,Z_i}}[J_i^{[n]}(W,Z_i)] - \mathbb{E}_{P_W\otimes P_{Z_i}}[J_i^{[n]}(W,Z_i)]) + I_{\mathrm{SKL}}(W;S). \tag{41}$$

Then from (12) (39) and (40), the equation to prove can be rewritten as

$$\begin{aligned} I_{\mathrm{SKL}}(W;S) = &\sum_{i=0}^n \Bigg[\mathbb{E}_{P_{W,Z_i}}\left[\int_{\mathcal{Z}^{n-1}} \log(P_{W|Z_i,z^{-i}}) d\mu^{n-1}(z^{-i})\right] - \\ &\mathbb{E}_{P_W\otimes P_{Z_i}}\left[\int_{\mathcal{Z}^{n-1}} \log(P_{W|Z_i,z^{-i}}) d\mu^{n-1}(z^{-i})\right]\Bigg]. \end{aligned} \tag{42}$$

Using the Gibbs posterior, we have

$$P_{W|Z_i,z^{-i}} = \frac{\pi(W)e^{-\gamma L_e(W,Z_i,z^{-i})}}{V_{L_e}(Z_i,z^{-i},\gamma)}. \tag{43}$$

With this, we can now calculate the right hand side of equation (42)

$$\begin{aligned} &\sum_{i=0}^n \left[\mathbb{E}_{P_{W,Z_i}}\left[\int_{\mathcal{Z}^{n-1}} \log(P_{W|Z_i,z^{-i}}) d\mu^{n-1}(z^{-i})\right] - \mathbb{E}_{P_W\otimes P_{Z_i}}\left[\int_{\mathcal{Z}^{n-1}} \log(P_{W|Z_i,z^{-i}}) d\mu^{n-1}(z^{-i})\right]\right] \\ &= \sum_{i=0}^n \left\{\left[\int_{\mathcal{Z}^{n-1}} \mathbb{E}_{P_{W,Z_i}}[\log(P_{W|Z_i,z^{-i}})] d\mu^{n-1}(z^{-i})\right] - \left[\int_{\mathcal{Z}^{n-1}} \mathbb{E}_{P_W\otimes P_{Z_i}}[\log(P_{W|Z_i,z^{-i}})] d\mu^{n-1}(z^{-i})\right]\right\} \\ &= \sum_{i=0}^n \int_{\mathcal{Z}^{n-1}} \left[\mathbb{E}_{P_{W,Z_i}}[\log(P_{W|Z_i,z^{-i}})] - \mathbb{E}_{P_W\otimes P_{Z_i}}[\log(P_{W|Z_i,z^{-i}})]\right] d\mu^{n-1}(z^{-i}) \\ &= \sum_{i=0}^n \int_{\mathcal{Z}^{n-1}} \gamma \left[\mathbb{E}_{P_W\otimes P_{Z_i}}[L_e(W,Z_i,z^{-i})] - \mathbb{E}_{P_{W,Z_i}}[L_e(W,Z_i,z^{-i})]\right] d\mu^{n-1}(z^{-i}) \\ &= \sum_{i=0}^n \int_{\mathcal{Z}^{n-1}} \frac{\gamma}{n} \left[\mathbb{E}_{P_W\otimes P_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)]\right] d\mu^{n-1}(z^{-i}) \\ &= \sum_{i=0}^n \frac{\gamma}{n} \left[\mathbb{E}_{P_W\otimes P_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)]\right] \\ &= \sum_{i=0}^n \frac{\gamma}{n} \left[\mathbb{E}_{P_W\otimes P_S}[\ell(W,Z_i)] - \mathbb{E}_{P_{W,S}}[\ell(W,Z_i)]\right] \\ &= \gamma \left[\mathbb{E}_{P_W\otimes P_S}[L_e(W,S)] - \mathbb{E}_{P_{W,S}}[L_e(W,S)]\right] \\ &= \gamma \cdot \mathrm{gen}(P_{W|S}^\gamma, P_S). \end{aligned} \tag{44}$$

Using the fact that for Gibbs algorithm, we have

$$I_{\mathrm{SKL}}(W;S) = \gamma \cdot \mathrm{gen}(P_{W|S}^\gamma, P_S). \tag{45}$$

Combining (44) and (45), we get (42), and that completes our proof. □

B. Discussion on Jensen's gap and WCDG distribution

For a given loss function $\ell(\theta, s)$ and a given measure of reference P_0, the WCDG distribution $P^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}$ is a measure defined as

$$\frac{dP^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}}{dP_0} = \exp\left(\frac{1}{\beta}\ell(\theta, s) - \log\int \exp\left(\frac{1}{\beta}\ell(\theta, s)\right)dP_0(s)\right). \tag{46}$$

Note that this $\ell(\theta, s)$ could be different from what we introduced in the preliminaries section. Therefore, each β corresponds to a WCDG distribution, and if we assign

$$\alpha = D(P^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}\|P_0), \tag{47}$$

then it can be proved that $P^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}$ is the solution to the following optimization problem

$$\max_{P\ll P_0}\int \ell(\theta, s)dP(s) \quad s.t. \quad D(P\|P_0) \leq \alpha \tag{48}$$

where θ, P_0 is fixed. In other words, $P^{(P_0,\beta)}_{\hat{S}|\Theta=\theta}$ represents the worst case distribution such that the expected loss is the highest while the distribution itself is not "too far" from a reference distribution P_0. [26] used calculus of variation with constraint to generate the result for this optimization problem by seeing it as a functional of dP/dP_0, which one-to-one corresponds to a measure P $s.t.$ $P \ll P_0$.

Under this setting, and we assign $\theta = w, z_i$, and $s = z^{-i}$, let

$$\ell(\theta, s) = L_e(w, z_i, z^{-i}) - \frac{1}{\gamma}\log V_{L_e}(z_i, z^{-i}, \gamma), \tag{49}$$

then we have

$$\mathbb{E}_{P_W\otimes P_{Z_i}}[J_i^{[n]}(W, Z_i)] = D(P_W \otimes P_S\|P^{(\mu^{n-1},\frac{1}{\gamma})}_{\hat{Z}^{n-1},Z_i,W}) \tag{50}$$

where $P^{(\mu^{n-1},\frac{1}{\gamma})}_{\hat{Z}^{n-1},Z_i,W} = P^{(\mu^{n-1},\frac{1}{\gamma})}_{\hat{Z}^{n-1}|Z_i,W}P_{Z_i,W}$, and $P^{(\mu^{n-1},\frac{1}{\gamma})}_{\hat{Z}^{n-1}|Z_i,W}$ indicates the worst case data generating distribution of Z^{-i} that maximize the loss $\ell(\theta, s)$ given in (49), with a fixed hypothesis W and one sample Z_i.

This also results in an upper bound for the generalization error

$$\begin{aligned}\sum_{i=1}^{n}\left(I_{\text{SKL}}(W; Z_i) + D(P_W \otimes P_S\|P^{(\mu^{n-1},\frac{1}{\gamma})}_{\hat{Z}^{n-1},Z_i,W})\right) &\geq I_{\text{SKL}}(W; S) \qquad (51)\\ &= \gamma\text{gen}(P^{\gamma}_{W|S}, P_S). \qquad (52)\end{aligned}$$

Note that this bound is not order-wise optimal as it omitted another non-negative Jensen term.

Appendix B
Asymptotic Results

A. Proof of Lemma 1

Proof. It is easy to show that

$$\begin{aligned}\lim_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z^n}}{dP^{\infty}_W\otimes P_{Z^\infty}}\right) &= \lim_{n\to\infty}\left(\frac{dP^{[n]}_{W,Z^n}}{dP^{\infty}_W\otimes \mu^n}\right)\\ &= \lim_{n\to\infty}\frac{\pi(W)e^{-\gamma\frac{1}{n}\sum_1^n \ell(W,Z_i)}}{\int_{\mathcal{W}}\pi(w)e^{-\gamma\frac{1}{n}\sum_1^n \ell(w,Z_i)}dw}\cdot\frac{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}dw}{\pi(W)e^{-\gamma L_\mu(W)}}\\ &= \frac{\pi(W)\lim_{n\to\infty}e^{-\gamma\frac{1}{n}\sum_1^n \ell(W,Z_i)}}{\int_{\mathcal{W}}\pi(w)\lim_{n\to\infty}e^{-\gamma\frac{1}{n}\sum_1^n \ell(w,Z_i)}dw}\cdot\frac{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}dw}{\pi(W)e^{-\gamma L_\mu(W)}}\\ &= \frac{\pi(W)e^{-\gamma\lim_{n\to\infty}\frac{1}{n}\sum_1^n \ell(W,Z_i)}}{\int_{\mathcal{W}}\pi(w)e^{-\gamma\lim_{n\to\infty}\frac{1}{n}\sum_1^n \ell(w,Z_i)}dw}\cdot\frac{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}dw}{\pi(W)e^{-\gamma L_\mu(W)}}\\ &= \frac{\pi(W)e^{-\gamma L_\mu(W)}}{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}dw}\cdot\frac{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}dw}{\pi(W)e^{-\gamma L_\mu(W)}}\\ &= 1 \quad a.s. \end{aligned} \tag{53}$$

where we used the strong law of large numbers to show that for any fixed W,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{1}^{n} \ell(W, Z_i) = L_\mu(W) \tag{54}$$

almost surely with regard to P_{Z^∞}, and thus holds almost surely in $P_W^\infty \otimes P_{Z^\infty}$ using Fubini's theorem. We also used the dominated convergence theorem to exchange the integral and limit. □

B. Proof of Lemma 2

Proof. Using similar techniques as in the proof of the previous Lemma B-A

$$\begin{aligned}
\varliminf_{n\to\infty} \left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^\infty \otimes P_{Z^\infty}}\right) &= \varliminf_{n\to\infty} \left(\frac{P_{W|Z_i}^{[n]}}{P_W^\infty}\right) \\
&= \varliminf_{n\to\infty} \left(\int_{\mathcal{Z}^{n-1}} P_{W|Z_i,z^{-i}}^{[n]} d\mu^{n-1}(z^{-i})\right) \cdot \frac{d\mathrm{Leb}}{dP_W^\infty} \\
&\geq \left(\int_{\mathrm{Z}^\infty} \varliminf_{n\to\infty} P_{W|Z_i,z^{-i}}^{[n]} dP_{Z^\infty}(z^\infty)\right) \cdot \frac{d\mathrm{Leb}}{dP_W^\infty} \\
&= \frac{\pi(W)e^{-\gamma L_\mu(W)}}{\int_{\mathcal{W}} \pi(w)e^{-\gamma L_\mu(w)} \cdot dw} \cdot \frac{\int_{\mathcal{W}} \pi(w)e^{-\gamma L_\mu(w)} \cdot dw}{\pi(W)e^{-\gamma L_\mu(W)}} \\
&= 1
\end{aligned} \tag{55}$$

where the inequality is obtained using Fatou's lemma. Here, we use Leb to denote the Lebesgue measure on $\mathcal{W} \subset \mathbb{R}^d$. By saying that, we acknowledge that $\pi(w)$ is a distribution with regard to Lebesgue measure on $\mathcal{W}$. Of course, the proof still works for other measures even when Lebesgue measure doesn't exist on $\mathcal{W}$, namely, when $\mathcal{W}$ is an abstract measure space (for example a general function space).

From the Radon-Nikodym theorem, $dP_{W,Z_i}^{[n]}/dP_W^\infty \otimes P_{Z^\infty}$ is a measurable function on $\mathcal{F}_{W,Z^n}^{[n]}$, and thus measurable on $\mathcal{F}^\infty$. Consequently, $\varliminf_{n\to\infty} \left(dP_{W,Z_i}^{[n]}/dP_W^\infty \otimes P_{Z^\infty}\right)$ is also measurable on $\mathcal{F}^\infty$. We consider the set Ω_δ as

$$\Omega_\delta = \left\{\omega \in \mathcal{W} \times \mathcal{Z}^\infty \,\middle|\, \varliminf_{n\to\infty} \left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^\infty \otimes P_{Z^\infty}}\right)(\omega) > 1\right\}, \tag{56}$$

and the set is obviously measurable. Suppose that $P_W^\infty \otimes P_{Z^\infty}(\Omega_\delta) > 0$, then

$$1 = \varliminf_{n\to\infty} \int_{\Omega^\infty} \left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^\infty \otimes P_{Z^\infty}}\right) dP_W^\infty \otimes P_{Z^\infty} \tag{57}$$

$$\geq \int_{\Omega^\infty} \varliminf_{n\to\infty} \left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^\infty \otimes P_{Z^\infty}}\right) dP_W^\infty \otimes P_{Z^\infty} > 1 \tag{58}$$

generates contradiction, indicating that $P_W^\infty \otimes P_{Z^\infty}(\Omega_\delta) = 0$, which proves the lemma. □

C. Proof of Corollary 1

Proof. This is a direct result of the dominated convergence theorem. □

D. Proof of Lemma 3

Proof. First, we have

$$\begin{aligned}
n \cdot \left(1 - \frac{dP_W^{[n]} \otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right) &= -\frac{n \cdot \int_{Z^n} (P_{W|z^n} - P_{W|Z_i,z^{-i}}) d\mu^n(z^n)}{\int_{Z^n} (P_{W|Z_i,z^{-i}}) d\mu^n(z^n)} \\
&= -\frac{\int_{Z^\infty} n(P_{W|z^n} - P_{W|Z_i,z^{-i}}) dP_{Z^\infty}(z^\infty)}{\int_{Z^\infty} (P_{W|Z_i,z^{-i}}) dP_{Z^\infty}(z^\infty)}.
\end{aligned} \tag{59}$$

For $\frac{f(t,x)}{(Af)(x)}$ where $A: g(t) \mapsto y$ is a linear functional. We expand it into

$$\begin{aligned}\frac{f(t,x)}{(Af)(x)} =& (f(t,x_0) + f_x(t,x_0)(x-x_0) + \frac{1}{2}f_{xx}(t,\xi)(x-x_0)^2) \cdot \Bigg(\frac{1}{(Af(t,x_0))} \\ &- \frac{1}{(Af(t,x_0))^2} A(f_x(t,x_0)(x-x_0) + \frac{1}{2}f_{xx}(t,\xi)(x-x_0)^2) \\ &+ \frac{1}{\xi_1^3}\big(A(f_x(t,x_0)(x-x_0) + \frac{1}{2}f_{xx}(t,\xi)(x-x_0)^2)\big)^2 \Bigg). \end{aligned} \tag{60}$$

where we used the Lagrange remainder, and $\xi \in (x_0, x)$, $\xi_1 \in (Af(t,x_0), Af(t,x))$. Applying this technique where we choose the linear operator as

$$A: f(w) \mapsto \int_{\mathcal{W}} \pi(w) f(w) dw. \tag{61}$$

We have

$$\begin{aligned} & n(P_{W|z^n} - P_{W|Z_i,z^{-i}}) \\ =& \frac{\pi(W)\big(e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(W,z_j)} - e^{-\gamma\frac{1}{n}\sum_{j=1,j\neq i}^n \ell(W,z_j) - \gamma\frac{1}{n}\ell(W,Z_i)}\big)}{\int_{\mathcal{W}} \pi(w) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(w,z_j)} \cdot dw - \int_{\mathcal{W}} \pi(w)\big(e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(W,z_j)} - e^{-\gamma\frac{1}{n}\sum_{j=1,j\neq i}^n \ell(w,z_j) - \gamma\frac{1}{n}\ell(w,Z_i)}\big)dw} \\ =& \Bigg\{ \frac{\pi(W) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(W,z_j)} (\frac{\gamma}{n}(\ell(W,Z_i) - \ell(W,z)) \cdot n)}{\int_{\mathcal{W}} \pi(w) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(w,z_j)} \cdot dw} \\ & - \frac{\pi(W) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(W,z_j)}}{[\int_{\mathcal{W}} \pi(w) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(w,z_j)} \cdot dw]^2} \int_{\mathcal{W}} \pi(w) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(w,z_j)} (\frac{\gamma}{n}(\ell(w,Z_i) - \ell(w,z)) \cdot n) \cdot dw \Bigg\} \\ & + \left[\frac{\pi(W) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(W,z_j)}}{\int_{\mathcal{W}} \pi(w) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(w,z_j)} \cdot dw} \right] R(W, Z_i, z^{-i}, n^{-1}) \\ =& \left[\frac{\pi(W) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(W,z_j)}}{\int_{\mathcal{W}} \pi(w) e^{-\gamma\frac{1}{n}\sum_{j=1}^n \ell(w,z_j)} \cdot dw} \right] \Bigg(\gamma(\ell(W,Z_i) - \ell(W,z)) - \mathbb{E}_e^{[n]}[\ell(W,Z_i) - \ell(W,z)] \\ & + R(W, Z_i, z^{-i}, n^{-1}) \Bigg) \end{aligned} \tag{62}$$

$$\tag{63}$$

where $R(W, Z_i, z^{-i}, n^{-1})$ is the remainder of the Taylor's expansion. It is easy to see that

$$R(W, Z_i, z^{-i}, n^{-1}) = \Theta\left(\frac{1}{n}\right). \tag{64}$$

Using the fact that the energy function is bounded, it is not difficult to show that $R(W, Z_i, z^{-i}, n^{-1})$ is uniformly small. In other words, $n \cdot R(W, Z_i, z^{-i}, n^{-1})$ is uniformly bounded, satisfying $\forall n \in \mathbb{N}, \forall W, \forall Z_i, \forall z^{-i}$, there exists a constant C, such that

$$R(W, Z_i, z^{-i}, n^{-1}) \leq \frac{C}{n}. \tag{65}$$

Then using that the energy function is bounded, together with (62) and (65), we get that $\frac{n(P_{W|z^n} - P_{W|Z_i,z^{-i}})}{\pi(W)}$ is uniformly bounded, satisfying $\forall n \in \mathbb{N}, \forall W, \forall Z_i, \forall z^{-i}$, there exists a constant C_1, such that

$$\frac{n(P_{W|z^n} - P_{W|Z_i,z^{-i}})}{\pi(W)} \leq C_1. \tag{66}$$

Note that here $\pi(W)$ may not be bounded, but it is not essential since $\pi(W)$ is a probability distribution. Afterwards, we will prove that the denominator in (59) is lower-bounded with any given W, Z_i.

$$\begin{aligned} \left| \int_{Z^\infty} (P_{W|Z_i,z^{-i}}) dP_{Z^\infty}(z^\infty) \right| &= \int_{Z^\infty} (P_{W|Z_i,z^{-i}}) dP_{Z^\infty}(z^\infty) \\ &= \int_{Z^\infty} \frac{\pi(W) e^{-\gamma\frac{1}{n}\sum_1^n \ell(W,Z_i)}}{\int_{\mathcal{W}} \pi(w) e^{-\gamma\frac{1}{n}\sum_1^n \ell(w,z_j)} dw} dP_{Z^\infty}(z^\infty) \\ &\geq \frac{\pi(W) e^{-\gamma \sup \ell(w,z)}}{e^{-\gamma \inf \ell(w,z)}}. \end{aligned} \tag{67}$$

With (66) (67), we can bound (59)

$$
\begin{aligned}
\left|n\cdot\left(1-\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right)\right| &= \frac{\left|\int_{Z^\infty} n(P_{W|z^n}-P_{W|Z_i,z^{-i}})dP_{Z^\infty}(z^\infty)\right|}{\left|\int_{Z^\infty}(P_{W|Z_i,z^{-i}})dP_{Z^\infty}(z^\infty)\right|}\\
&\le \frac{\pi(W)\int_{Z^\infty}\left|\frac{n(P_{W|z^n}-P_{W|Z_i,z^{-i}})}{\pi(W)}\right|dP_{Z^\infty}(z^\infty)}{\frac{\pi(W)e^{-\gamma\sup\ell(w,z)}}{e^{-\gamma\inf\ell(w,z)}}}\\
&\le \frac{e^{-\gamma\inf\ell(w,z)}\cdot C_1}{e^{-\gamma\sup\ell(w,z)}},
\end{aligned}\tag{68}
$$

and that proves the uniformly bounded conclusion.

With all the bounded condition, we can now apply the dominated convergence theorem and the strong law of large numbers to get (17)

$$
\begin{aligned}
\lim_{n\to\infty} n\cdot\left(1-\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right) &= -\frac{\lim\limits_{n\to\infty}\int_{Z^\infty} n(P_{W|z^n}-P_{W|Z_i,z^{-i}})dP_{Z^\infty}(z^\infty)}{\lim\limits_{n\to\infty}\int_{Z^\infty}(P_{W|Z_i,z^{-i}})dP_{Z^\infty}(z^\infty)}\\
&= -\frac{\int_{Z^\infty}\lim\limits_{n\to\infty} n(P_{W|z^n}-P_{W|Z_i,z^{-i}})dP_{Z^\infty}(z^\infty)}{\int_{Z^\infty}\lim\limits_{n\to\infty}(P_{W|Z_i,z^{-i}})dP_{Z^\infty}(z^\infty)}\\
&= \left[\frac{\pi(W)e^{-\gamma L_\mu(W)}}{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}\cdot dw}\right]\Bigg(-(\gamma(\ell(W,Z_i)-L_\mu(W)))\\
&\quad+\mathbb{E}_W^\infty(\gamma(\ell(W,Z_i)-L_\mu(W)))\Bigg)\cdot\frac{1}{\left[\frac{\pi(W)e^{-\gamma L_\mu(W)}}{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}\cdot dw}\right]}\\
&= (-\gamma(\ell(W,Z_i)-L_\mu(W)))+\mathbb{E}_W^\infty(\gamma(\ell(W,Z_i)-L_\mu(W))).
\end{aligned}\tag{69}
$$

And that finishes our proof. □

E. Proof of Theorem 2

Proof. From the definition

$$
I_{\mathrm{SKL}}(W;Z_i)=\int_{W,Z^\infty}\left(\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\log\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}-\log\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right)dP_{W,Z^n}^{[n]}.\tag{70}
$$

Let

$$
x=1-\frac{dP_W^{[n]}\otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}},\tag{71}
$$

then the term inside the integration becomes

$$
(1-x)\log(1-x)-\log(1-x)=x^2+\frac{1}{2(1-\xi)^2}x^3,\tag{72}
$$

where $\xi\in(0,x)$.

Using Lemma 3, $|x|\le\frac{C_1}{n}$ for some constant C_1, then for sufficiently large n, say $n>2C_1$, we easily get

$$
\frac{1}{(1-\xi)^2}<4.\tag{73}
$$

Consider that

$$\begin{aligned}\left|n^2\big((1-x)\log(1-x)-\log(1-x)\big)\right| &= n^2\left|x^2+\frac{1}{2(1-\xi)^2}x^3\right| \\ &\le (nx)^2+n^2\frac{1}{2(1-\xi)^2}|x|^3 \\ &\le (nx)^2+2|nx|^3\cdot\frac{1}{n} \\ &\le C_1^2+2C_1^3\cdot\frac{1}{n} \end{aligned} \tag{74}$$

is uniformly bounded for sufficiently large n. From Corollary 1 we know $\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_W\otimes P_{Z^\infty}}$ is also uniformly bounded, therefore we can exchange the order of integration and limits using the dominated convergence theorem, which yields

$$\begin{aligned}\lim_{n\to\infty} n^2 I_{\mathrm{SKL}}(W;Z_i) &= \int \lim_{n\to\infty} n^2\big((1-x)\log(1-x)-\log(1-x)\big)\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_W\otimes P_{Z^\infty}}dP^{\infty}_W\otimes P_{Z^\infty}(w,z^\infty) \\ &= \int \lim_{n\to\infty} n^2\left(x^2+\frac{1}{2(1-\xi)^2}x^3\right)\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_W\otimes P_{Z^\infty}}dP^{\infty}_W\otimes P_{Z^\infty}(w,z^\infty) \\ &= \int \lim_{n\to\infty} n^2x^2\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_W\otimes P_{Z^\infty}}dP^{\infty}_W\otimes P_{Z^\infty}(w,z^\infty) \\ &= \int \Big((\gamma(\ell(w,z_i)-L_\mu(w)))-\mathbb{E}^{\infty}_W[\gamma(\ell(W,z_i)-L_\mu(W))]\Big)^2 dP^{\infty}_W\otimes P_{Z^\infty}(w,z^\infty) \qquad (75) \\ &< +\infty \qquad (76)\end{aligned}$$

where the last equality (75) is obtained using Corollary 1 and Lemma 3. Therefore,

$$\begin{aligned}\lim_{n\to\infty} n^2\cdot I_{\mathrm{SKL}}(W;Z_i) &= \mathbb{E}_{P^{\infty}_W\otimes\mu}\left[\Big((\gamma(\ell(W,Z)-L_\mu(W)))-\mathbb{E}^{\infty}_W[\gamma(\ell(W,Z)-L_\mu(W))]\Big)^2\right] \qquad (77) \\ &= \gamma^2\mathbb{E}_\mu\Big[\mathbb{E}^{\infty}_W[(\ell(W,Z)-L_\mu(W))^2]-\mathbb{E}^{\infty}_W[(\ell(W,Z)-L_\mu(W))]^2\Big]. \qquad (78)\end{aligned}$$

□

F. Proof of Corollary 3

Proof. This conclusion is directly obtained from $I(W;Z_i)\le I_{\mathrm{SKL}}(W;Z_i)$. □

G. Proof of Remark 2

Proof. Using Lemma 3, we start by showing

$$\lim_{n\to\infty} n\log\frac{dP^{[n]}_{W,Z_i}}{dP^{[n]}_W\otimes P_{Z_i}} = \Big(-(\gamma(\ell(W,Z_i)-L_\mu(W)))+\mathbb{E}^{\infty}_W[\gamma(\ell(W,Z_i)-L_\mu(W))]\Big). \tag{79}$$

Using similar techniques as in Theorem 2, we again have the terms inside the integration being uniformly bounded for sufficiently large n. Therefore, using the dominated convergence theorem, we have

$$\begin{aligned}\lim_{n\to\infty} nI(W;Z_i) &= \int_{W,Z^\infty}\lim_{n\to\infty} n\log\left(\frac{dP^{[n]}_{W,Z_i}}{dP^{[n]}_W\otimes P_{Z_i}}\right)\cdot\left(\frac{dP^{[n]}_{W,Z_i}}{dP^{\infty}_W\otimes P_{Z^\infty}}\right)dP^{\infty}_W\otimes P_{Z^\infty}(w,z^\infty) \\ &= -\int_{W,Z_i}\Big((\gamma(\ell(w,z_i)-L_\mu(w)))-\mathbb{E}^{\infty}_W[\gamma(\ell(W,z_i)-L_\mu(W))]\Big)dP^{\infty}_W\otimes P_{Z_i}(w,z_i) \\ &= 0, \end{aligned} \tag{80}$$

which completes our proof. □

H. Proof of Lemma 4

Proof. Without the lost of generality, we assume $\pi(w)$ is also bounded to simplify our proof. Note that this assumption can surely be taken off without affecting the conclusion.

We first define

$$\Delta_{z_i}(w) \triangleq P_{W|Z^n}(w|z_i, z^{-i}) - P_{W|Z^n}(w|z^n) \tag{81}$$
$$= P_{w|z_i,z^{-i}} - P_{w|z^n}. \tag{82}$$

And therefore as proved in Lemma 3

$$\Delta_{z_i}(w, z^n) = \left[\frac{\pi(w)e^{-\gamma\frac{1}{n}\sum_{j=1}^{n}\ell(w,z_j)}}{\int_{\mathcal{W}}\pi(w)e^{-\gamma\frac{1}{n}\sum_{j=1}^{n}\ell(w,z_j)}\cdot dw}\right]\cdot\Bigg(-(\gamma(\ell(w,z_i)-\ell(w,z))) + \mathbb{E}_e^{[n]}[\gamma(\ell(W,z_i) - \ell(W,z))]\Bigg)\cdot\frac{1}{n} + o\left(\frac{1}{n}\right). \tag{83}$$

Note that if ℓ is only lower-bounded, the expansion still rings true almost surely. Only that this time, the $o\left(\frac{1}{n}\right)$ will no longer be uniformly small.

Then we consider

$$\begin{aligned} n\cdot J_i^{[n]}(w, z_i) &= n\cdot\left[\log\left(\int_{\mathcal{Z}^{n-1}} P_{w|z_i,z^{-i}}d\mu^{n-1}(z^{-i})\right) - \int_{\mathcal{Z}^{n-1}}\log(P_{w|z_i,z^{-i}})d\mu^{n-1}(z^{-i})\right] \\ &= n\cdot\Bigg[\log\left(\int_{\mathrm{Z^n}} P_{w|z^n} + \Delta_{z_i}(w,z^n)\ d\mu^n(z^n)\right) \\ &\quad - \int_{\mathrm{Z^n}}\log(P_{w|z^n} + \Delta_{z_i}(w,z^n))d\mu^n(z^n)\Bigg]. \end{aligned} \tag{84}$$

For any $r \in [r_1, r_2] \subset \mathbb{R}$ where $r_1 > 0, r_2 < +\infty$, let $0 < m_1 \leq r + x \leq m_2 < +\infty$, it's easy to prove that there exists a real number $a < 0$, such that

$$\begin{aligned} \log(r+x) &\leq \log(r) + \frac{x}{r}, \\ \log(r+x) &\geq \log(r) + \frac{x}{r} + ax^2. \end{aligned} \tag{85}$$

Now, we assign

$$\hat{J}^{[n]}(w) = \log\left(\int_{\mathrm{Z^n}} P_{w|z^n}d\mu^n(z^n)\right) - \int_{\mathrm{Z^n}}\log P_{w|z^n}d\mu^n(z^n). \tag{86}$$

From the proof of Lemma 3, we know that $n\cdot\Delta_{z_i}(w,z^n)$ is uniformly bounded. It is also easy to see that $P_{w|z^n} \in [r_1, r_2]$ where $r_1 > 0, r_2 < +\infty$, and $0 < r_1 \leq P_{w|z^n} + \Delta_{z_i}(w,z^n) \leq r_2 < +\infty$. Combining (84) (85) and (86) we get

$$\begin{aligned} n\cdot(\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i)) \leq &\left[\int_{\mathrm{Z^n}}\frac{n\Delta_{z_i}(w,z^n)}{P_{w|z^n}}d\mu^n(z^n) - \frac{\int_{\mathrm{Z^n}} n\Delta_{z_i}(w,z^n)d\mu^n(z^n)}{\int_{\mathrm{Z^n}} P_{w|z^n}d\mu^n(z^n)}\right] \\ &- \frac{a}{n}\left(\int_{\mathrm{Z^n}} n\cdot\Delta_{z_i}(w,z^n)d\mu^n(z^n)\right)^2, \end{aligned} \tag{87}$$

$$\begin{aligned} n\cdot(\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i)) \geq &\left[\int_{\mathrm{Z^n}}\frac{n\Delta_{z_i}(w,z^n)}{P_{w|z^n}}d\mu^n(z^n) - \frac{\int_{\mathrm{Z^n}} n\Delta_{z_i}(w,z^n)d\mu^n(z^n)}{\int_{\mathrm{Z^n}} P_{w|z^n}d\mu^n(z^n)}\right] \\ &+ \frac{a}{n}\int_{\mathrm{Z^n}}\left(n\cdot\Delta_{z_i}(w,z^n)\right)^2 d\mu^n(z^n), \end{aligned} \tag{88}$$

from which we get $n\cdot(\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i))$ is uniformly bounded. Then we apply the squeeze theorem, and yield

$$\lim_{n\to\infty} n\cdot(\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i)) = \lim_{n\to\infty}\left[\int_{\mathrm{Z^n}}\frac{n\Delta_{z_i}(w,z^n)}{P_{w|z^n}}d\mu^n(z^n) - \frac{\int_{\mathrm{Z^n}} n\Delta_{z_i}(w,z^n)d\mu^n(z^n)}{\int_{\mathrm{Z^n}} P_{w|z^n}d\mu^n(z^n)}\right]. \tag{89}$$

Eventually combining the result that $n \cdot \Delta_{z_i}(w, z^n)$ is uniformly bounded and (83), we complete our proof by showing

$$\lim_{n\to\infty} n \cdot (\hat{J}^{[n]}(w) - J_i^{[n]}(w, z_i)) = \Bigg[\int_{Z^\infty} \frac{\lim_{n\to\infty} n\Delta_{z_i}(w, z^n)}{\lim_{n\to\infty} P_{w|z^n}} dP_{Z^\infty}(z^\infty) - \frac{\int_{Z^\infty} \lim_{n\to\infty} n\Delta_{z_i}(w, z^n) dP_{Z^\infty}(z^\infty)}{\int_{Z^\infty} \lim_{n\to\infty} P_{w|z^n} dP_{Z^\infty}(z^\infty)} \Bigg] \tag{90}$$

$$= 0 \tag{91}$$

where we used the strong low of large numbers and the dominated convergence theorem. □

I. Proof of Lemma 5

Proof. We define

$$\delta(w, z^n) \triangleq \frac{1}{n} \sum_{i=0}^{n} \ell(w, z_i) - L_\mu(w). \tag{92}$$

We first prove that

$$\int_{Z^n} \mathbb{E}_W^\infty \left[\sqrt{n} \cdot |\delta(W, z^n)|\right]^2 d\mu^n(z^n), \tag{93}$$

$$\int_{Z^n} \mathbb{E}_W^\infty \left[\sqrt{n} \cdot |\delta(W, z^n)|\right]^4 d\mu^n(z^n) \tag{94}$$

are both bounded. For (93), using Cauchy's inequality

$$\begin{aligned}
\int_{Z^n} \mathbb{E}_W^\infty \left[\sqrt{n} \cdot |\delta(W, z^n)|\right]^2 d\mu^n(z^n) &\le \int_{Z^n} \mathbb{E}_W^\infty \left[n \cdot \delta(W, z^n)^2\right] d\mu^n(z^n) \\
&= \mathbb{E}_W^\infty \Big[\int_{Z^n} n \cdot \delta(W, z^n)^2 d\mu^n(z^n)\Big] \\
&= \mathbb{E}_W^\infty \left[\mathbb{E}_{\mu^n}[n \cdot \delta(W, Z^n)^2]\right] \\
&= \mathbb{E}_W^\infty \Bigg[\mathbb{E}_{\mu^n}[\frac{1}{n} \cdot \sum_{i=1}^{n} (\ell(W, Z_i) - L_\mu(W))^2] + \\
&\underbrace{\mathbb{E}_{\mu^n}[\frac{1}{n} \sum_{i\neq j} (\ell(W, Z_i) - L_\mu(W))(\ell(W, Z_j) - L_\mu(W))] \Bigg]}_{=0} \\
&= \mathbb{E}_W^\infty \Bigg[\mathbb{E}_\mu[(\ell(W, Z) - L_\mu(W))^2]\Bigg]
\end{aligned} \tag{95}$$

and is therefore bounded.

Similarly for (94), we again use Cauchy's inequality and yield

$$\begin{aligned}
\int_{Z^n} \mathbb{E}_W^\infty \left[\sqrt{n} \cdot |\delta(W, z^n)|\right]^4 d\mu^n(z^n) &\le \int_{Z^n} \mathbb{E}_W^\infty \left[n^2 \cdot \delta(W, z^n)^4\right] d\mu^n(z^n) \\
&= \mathbb{E}_W^\infty \Big[\int_{Z^n} n^2 \cdot \delta(W, z^n)^4 d\mu^n(z^n)\Big] \\
&= \mathbb{E}_W^\infty \left[\mathbb{E}_{\mu^n}[n^2 \cdot \delta(W, Z^n)^4]\right] \\
&= \mathbb{E}_W^\infty \Bigg[\frac{1}{n} \mathbb{E}_\mu[(\ell(W, Z) - L_\mu(W))^4] + \\
&\frac{3(n-1)}{n} \mathbb{E}_\mu[(\ell(W, Z) - L_\mu(W))^2]^2 \Bigg]
\end{aligned} \tag{96}$$

and is therefore bounded as well. And as will be discussed in Annotation 1, this result holds true for all power k besides 2 and 4.

Notice that $\hat{J}^{[n]}(w)$ can also be written as

$$\hat{J}^{[n]}(w) = \log \int_{Z^n} \frac{P_{w|z^n}}{P_w^\infty} d\mu^n(z^n) - \int_{Z^n} \log \frac{P_{w|z^n}}{P_w^\infty} d\mu^n(z^n). \tag{97}$$

Using Taylor's theorem with Lagrange form remainder (similar to (60)), we have the following expansion

$$\begin{aligned}\frac{P_{w|z^n}}{P_w^\infty} &= \frac{\pi(w)e^{-\gamma(L_\mu(w)+\delta(w,z^n))}}{\int_{\mathcal{W}}\pi(w)e^{-\gamma(L_\mu(w)+\delta(w,z^n))}\cdot dw}\cdot\frac{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}\cdot dw}{\pi(w)e^{-\gamma L_\mu(w)}}\\ &= \Big(1+R_1(w,\delta)\delta(w,z^n)\Big)\cdot\Big(1-R_2(w,\delta)\cdot\mathbb{E}_W^\infty[R_1(W,\delta)\delta(W,z^n)]\Big)\end{aligned} \tag{98}$$

where $R_1(w,\delta), R_2(w,\delta)$ are both bounded. More specifically

$$R_1(w,\delta) = -\gamma e^{-\gamma\xi_1}\quad \xi_1\in(0,\delta), \tag{99}$$

$$R_2(w,\delta) = \left(\frac{\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}\cdot dw}{\xi_2}\right)^2, \tag{100}$$

where $\xi_2\in(\int_{\mathcal{W}}\pi(w)e^{-\gamma L_\mu(w)}\cdot dw, \int_{\mathcal{W}}\pi(w)e^{-\gamma(L_\mu(w)+\delta(w,z^n))}\cdot dw)$. For simplicity, we respectively use M_1 and M_2 to each represent a upper-bound for $|R_1|$ and $|R_2|$.

We employ the same technique used in the proof of Lemma 4 to bound $\log(r+x)$, where $r=1$ in this circumstance. With this, we are able to dominate $n\cdot\hat{J}^{[n]}(w)$ using the expanded form of $\delta(w,z^n)$. This time for some $a<0$

$$\begin{aligned}n\cdot\hat{J}^{[n]}(w) \le& -a\int_{\mathcal{Z}^n} n\cdot\Big(R_1(w,\delta)\delta(w,z^n)-R_2(w,\delta)\mathbb{E}_W^\infty[R_1(W,\delta)\delta(W,z^n)]\\ &-R_1(w,\delta)R_2(w,\delta)\delta(w,z^n)\mathbb{E}_W^\infty[R_1(W,\delta)\delta(W,z^n)]\Big)^2 d\mu^n(z^n)\\ \le& |a|\Big\{M_1^2\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^2 d\mu^n(z^n) + M_1^2M_2^2\int_{\mathcal{Z}^n}\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(w,z^n)|]^2 d\mu^n(z^n)\\ &+2M_1^2M_2\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|\cdot\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]d\mu^n(z^n)\\ &+\frac{1}{\sqrt{n}}\cdot 2M_1^3M_2\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^2\cdot\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]d\mu^n(z^n)\\ &+\frac{1}{\sqrt{n}}\cdot 2M_1^3M_2^2\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|\cdot\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]^2 d\mu^n(z^n)\\ &+\frac{1}{n}\cdot M_1^4M_2^2\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^2\cdot\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]^2 d\mu^n(z^n)\Big\}\\ \le& |a|\Big\{M_1^2\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^2 d\mu^n(z^n) + M_1^2M_2^2\int_{\mathcal{Z}^n}\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(w,z^n)|]^2 d\mu^n(z^n)\\ &+2M_1^2M_2\Big(\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^2 d\mu^n(z^n)\Big)^{\frac{1}{2}}\Big(\int_{\mathcal{Z}^n}\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]^2 d\mu^n(z^n)\Big)^{\frac{1}{2}}\\ &+\frac{1}{\sqrt{n}}\cdot 2M_1^3M_2\Big(\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^4 d\mu^n(z^n)\Big)^{\frac{1}{2}}\Big(\int_{\mathcal{Z}^n}\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]^2 d\mu^n(z^n)\Big)^{\frac{1}{2}}\\ &+\frac{1}{\sqrt{n}}\cdot 2M_1^3M_2^2\Big(\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^2 d\mu^n(z^n)\Big)^{\frac{1}{2}}\Big(\int_{\mathcal{Z}^n}\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]^4 d\mu^n(z^n)\Big)^{\frac{1}{2}}\\ &+\frac{1}{n}\cdot M_1^4M_2^2\Big(\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^4 d\mu^n(z^n)\Big)^{\frac{1}{2}}\Big(\int_{\mathcal{Z}^n}\mathbb{E}_W^\infty[\sqrt{n}\cdot|\delta(W,z^n)|]^4 d\mu^n(z^n)\Big)^{\frac{1}{2}}\Big\}.\end{aligned} \tag{101}$$

Let

$$\mathrm{Mmt}_2(\ell(w,Z)) = \mathrm{Var}(\ell(w,Z)) = \int_{\mathrm{Z}}(\ell(w,z)-L_\mu(w))^2 dP(z), \tag{102}$$

$$\mathrm{Mmt}_4(\ell(w,Z)) = \int_{\mathrm{Z}}(\ell(w,z)-L_\mu(w))^4 dP(z), \tag{103}$$

then it can be calculated that

$$\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^2 d\mu^n(z^n) = \mathrm{Var}(\ell(w,Z)), \tag{104}$$

$$\int_{\mathcal{Z}^n}|\sqrt{n}\delta(w,z^n)|^4 d\mu^n(z^n) = \frac{1}{n}\mathrm{Mmt}_4(\ell(w,Z)) + \frac{3(n-1)}{n}\mathrm{Var}(\ell(w,Z))^2. \tag{105}$$

With the bounded energy function, (102) (103) (104) and (105) are all uniformly bounded, thus for some $C_1 > 0, C_2 > 0$,

$$\int_{\mathcal{Z}^n} |\sqrt{n}\delta(w, z^n)|^2 d\mu^n(z^n) \leq C_1, \tag{106}$$

$$\int_{\mathcal{Z}^n} |\sqrt{n}\delta(w, z^n)|^4 d\mu^n(z^n) \leq C_2. \tag{107}$$

Then with condition (93) and (94), we assign upper-bounds C_3 and C_4 for them respectively, getting

$$\int_{Z^n} \mathbb{E}_W^\infty\big[\sqrt{n} \cdot |\delta(W, z^n)|\big]^2 d\mu^n(z^n) \leq C_3, \tag{108}$$

$$\int_{Z^n} \mathbb{E}_W^\infty\big[\sqrt{n} \cdot |\delta(W, z^n)|\big]^4 d\mu^n(z^n) \leq C_4. \tag{109}$$

Now, (101) becomes

$$\begin{aligned} n \cdot \hat{J}^{[n]}(w) \leq |a| \bigg(& M_1^2 C_1 + M_1^2 M_2^2 C_3 + 2M_1^2 M_2 C_1^{\frac{1}{2}} C_3^{\frac{1}{2}} + \frac{1}{\sqrt{n}} \cdot 2M_1^3 M_2 C_2^{\frac{1}{2}} C_3^{\frac{1}{2}} \\ & + \frac{1}{\sqrt{n}} \cdot 2M_1^3 M_2^2 C_1^{\frac{1}{2}} C_4^{\frac{1}{2}} + \frac{1}{n} \cdot M_1^4 M_2^2 C_2^{\frac{1}{2}} C_4^{\frac{1}{2}} \bigg), \end{aligned} \tag{110}$$

and is therefore dominated by a constant, which surely proves that $n \cdot \hat{J}^{[n]}(w)$ is uniformly bounded. From this we also see that the terms with order $k \geq 3$ (in other words, containing $\hat{J}^{[n]}(w)^k$) are controlled by a factor $(1/n)^{(k/2-1)}$, thus those terms converges to zero with $n \to \infty$. This observation is used many times in later proofs.

Then, we prove that $\lim_{n\to\infty} n \cdot \hat{J}^{[n]}(w)$ exists. Similar to (98), we have

$$\begin{aligned} \frac{P_{w|z^n}}{P_w^\infty} &= \frac{\pi(w) e^{-\gamma(L_\mu(w) + \delta(w, z^n))}}{\int_{\mathcal{W}} \pi(w) e^{-\gamma(L_\mu(w)+\delta(w,z^n))} \cdot dw} \cdot \frac{\int_{\mathcal{W}} \pi(w) e^{-\gamma L_\mu(w)} \cdot dw}{\pi(w) e^{-\gamma L_\mu(w)}} \\ &= \bigg(1 - \gamma\delta(w, z^n) + R_1(w, \delta)\delta(w, z^n)^2\bigg) \cdot \bigg(1 - \mathbb{E}_W^\infty[-\gamma\delta(W, z^n) + R_1(W, \delta)\delta(W, z^n)^2] \\ &\quad + R_2(w, \delta) \cdot \mathbb{E}_W^\infty[-\gamma\delta(W, z^n) + R_1(W, \delta)\delta(W, z^n)^2]^2\bigg) \end{aligned} \tag{111}$$

where $R_1(w, \delta) = \frac{1}{2}\gamma^2 e^{-\gamma\xi_1}$, $\xi_1 \in (0, \delta)$, $R_2(w, \delta) = \left(\frac{\int_{\mathcal{W}} \pi(w) e^{-\gamma L_\mu(w)} \cdot dw}{\xi_2}\right)^3$, and $\xi_2 \in (\int_{\mathcal{W}} \pi(w) e^{-\gamma L_\mu(w)} \cdot dw, \int_{\mathcal{W}} \pi(w) e^{-\gamma(L_\mu(w)+\delta(w,z^n))} \cdot dw)$.

Again similar to (85), we have for some $a < 0$

$$\log(1 + x) \leq x - \frac{1}{2}x^2 + \frac{1}{3}x^3, \tag{112}$$

$$\log(1 + x) \geq x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + ax^4, \tag{113}$$

with $0 < m \leq 1 + x \leq M < +\infty$. Very similar to (101), combining (97), (111), (112) and (113), we yield an upper-bound and a lower-bound of $n \cdot \hat{J}^{[n]}(w)$. It is easy to see that the terms with order equal to or greater than 3 converges to zero. Therefore, using the squeeze theorem, the upper-bound and lower-bound converges to the same value characterized by terms with order 1 and 2, and we eventually get

$$\begin{aligned} \lim_{n\to\infty} n \cdot \hat{J}^{[n]}(w) = & \underbrace{\lim_{n\to\infty} -\frac{1}{2}\bigg[\int_{Z^n} -\gamma\sqrt{n}\delta(w, z^n) d\mu^n(z^n) + \int_{Z^n} \gamma\mathbb{E}_W^\infty[\sqrt{n}\delta(W, z^n)] d\mu^n(z^n)\bigg]^2}_{=0} \\ & + \int_{Z^n} \frac{1}{2}\bigg(-\gamma\sqrt{n}\delta(w, z^n) + \gamma\mathbb{E}_W^\infty[\sqrt{n}\delta(W, z^n)]\bigg)^2 d\mu^n(z^n) \\ & = \lim_{n\to\infty} \int_{Z^n} \frac{1}{2}\bigg(-\gamma\sqrt{n}\delta(w, z^n) + \gamma\mathbb{E}_W^\infty[\sqrt{n}\delta(W, z^n)]\bigg)^2 d\mu^n(z^n) \\ & = \frac{1}{2}\gamma^2\mathbb{E}_\mu[(\ell(w, Z) - L_\mu(w))^2] - \gamma^2\mathbb{E}_\mu\big[(\ell(w, Z) - L_\mu(w))\mathbb{E}_W^\infty[\ell(W, Z) - L_\mu(W)]\big] \\ & + \frac{1}{2}\gamma^2\mathbb{E}_\mu\big[\mathbb{E}_W^\infty[\ell(W, Z) - L_\mu(W)]^2\big], \end{aligned} \tag{114}$$

which is a bounded function of w. That proves the lemma. □

Annotation 1. *According to [31], for a series of i.i.d sampled X with zero mean, when the k^{th}-moment of X exists ,*

$$\mathbb{E}\left[\left(\frac{|X_1+X_2+\cdots+X_n|}{\sqrt{n}}\right)^k\right] < C_k \tag{115}$$

for some constant C_k depending only on the distribution of X.

Following its proof, it is easy to see that C_k only depends on $\mathbb{E}[|X|^k]$ and C_{k-1}. The idea of the proof is to use mathematical induction. Knowing that when $k=2$, the conclusion holds for a $C_2=\mathbb{E}[X^2]$. Then if the conclusion holds for k with C_k, it is able to prove that there exists a C_{k+1} which only depends on C_k and $\mathbb{E}[|X|^{k+1}]$ for the inequality to hold for all n.

Therefore, for $\forall k>0$, the k^{th}-moment of $\ell(w,Z)$, $\mathbb{E}[|\ell(w,Z)|^k]$ surly exists, and is uniformly bounded since $\ell(w,z)$ is a bounded function. We then have the C_k introduced above is uniform over w for $\ell(w,Z)-L_\mu(w)$, in other words,

$$\exists M_k < +\infty, \forall n, \forall w \ \mathbb{E}_{\mu^n}[(\sqrt{n}|\delta(w,Z^n)|)^k] < M_k. \tag{116}$$

And similar to our proof in (95) *and* (96)*,*

$$\exists M_k < +\infty, \forall n, \ \mathbb{E}_{\mu^n}[(\sqrt{n}\mathbb{E}_W^\infty[|\delta(W,Z^n)|])^k] < M_k. \tag{117}$$

The result plays an important role in guaranteeing the third or higher order terms converges to zero.

For a more general setting, consider $2<p<\infty$ and independent random variables $X_i \in L^p$ with zero mean that are not necessarily identical. Then by [32], we have

$$\left(\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^p\right]\right)^{\frac{1}{p}} \le K_p \max\left\{\left(\sum_{i=1}^n \mathbb{E}[|X_i|^p]\right)^{\frac{1}{p}}, \left(\sum_{i=1}^n \mathbb{E}[|X_i|^2]\right)^{\frac{1}{2}}\right\} \tag{118}$$

and

$$\left(\mathbb{E}\left[\left|\sum_{i=1}^n X_i\right|^p\right]\right)^{\frac{1}{p}} \ge \frac{1}{2} \max\left\{\left(\sum_{i=1}^n \mathbb{E}[|X_i|^p]\right)^{\frac{1}{p}}, \left(\sum_{i=1}^n \mathbb{E}[|X_i|^2]\right)^{\frac{1}{2}}\right\}, \tag{119}$$

where K_p is a constant only depending on p. Note that this theorem can be seen as a version of Khintchine inequality.

It provide an estimate of the moment of the sum of independent random variables in a broader setting. Letting all the random variables be identically distributed, and the result recovers to (115)*, and can surely yield* (116) *and* (117)*.*

J. Proof of Theorem 3

Proof. We define

$$K_0(w) \triangleq \lim_{n\to\infty} n \cdot \hat{J}^{[n]}(w). \tag{120}$$

The existence of the limit is obtained using Lemma 5. Using Lemma 4 and Lemma 5, we get

$$\lim_{n\to\infty} n \cdot J_i^{[n]}(w,z_i) = K_0(w) \tag{121}$$

and

$$0 \le n \cdot J_i^{[n]}(w,z_i) = n \cdot \hat{J}^{[n]}(w) - n \cdot (\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i)) \tag{122}$$

$$\le |n \cdot J_i^{[n]}(w,z_i)| + |n \cdot (\hat{J}^{[n]}(w) - J_i^{[n]}(w,z_i))| \tag{123}$$

which is uniformly bounded.

Using Lemma 3, we know that $n \cdot \left(1 - \frac{dP_W^{[n]} \otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right)$ is uniformly bounded. Therefore, together with Corollary 1, we have

$$\begin{aligned}
&\lim_{n\to\infty} n \cdot \left(\sum_{i=1}^n I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S)\right) \\
&= \lim_{n\to\infty} \int_{W,Z^\infty} n \cdot J_i^{[n]}(w,z_i) \cdot n \cdot \left(1 - \frac{dP_W^{[n]} \otimes P_{Z_i}}{dP_{W,Z_i}^{[n]}}\right)\left(\frac{dP_{W,Z_i}^{[n]}}{dP_W^\infty \otimes P_{Z^\infty}}\right) dP_W^\infty \otimes P_{Z^\infty}(w,z^\infty) \\
&= \int_{W,Z_i} K_0(w)\Big(-(\gamma(\ell(w,z_i)-L_\mu(w))) + \mathbb{E}_W^\infty[\gamma(\ell(W,z_i)-L_\mu(W))]\Big) dP_W^\infty \otimes P_{Z_i}(w,z_i) \\
&= 0,
\end{aligned} \tag{124}$$

which proves the theorem. □

K. Proof of Theorem 4

Proof. The proof of this theorem utilizes the very same techniques in the proof of Lemma 5. Similar to our previous expansion (98) and (111), we can write

$$\begin{aligned}\frac{P_{w|z^n}}{P_w^\infty} &= \frac{\pi(w)e^{-\gamma(L_\mu(w)+\delta(w,z^n))}}{\int_{\mathcal{W}} \pi(w)e^{-\gamma(L_\mu(w)+\delta(w,z^n))}\cdot dw} \cdot \frac{\int_{\mathcal{W}} \pi(w)e^{-\gamma L_\mu(w)}\cdot dw}{\pi(w)e^{-\gamma L_\mu(w)}} \\ &= \left(1-\gamma\delta(w,z^n)+\frac{1}{2}\gamma^2\delta(w,z^n)^2+R_1(w,\delta)\delta(w,z^n)^3\right)\cdot\Bigg(1-\mathbb{E}_W^\infty[-\gamma\delta(W,z^n)+\frac{1}{2}\gamma^2\delta(W,z^n)^2 \\ &\quad + R_1(W,\delta)\delta(W,z^n)^3]+\mathbb{E}_W^\infty[-\gamma\delta(W,z^n)+\frac{1}{2}\gamma^2\delta(W,z^n)^2+R_1(W,\delta)\delta(W,z^n)^3]^2 \\ &\quad + R_2(w,\delta)\cdot\mathbb{E}_W^\infty[-\gamma\delta(W,z^n)+\frac{1}{2}\gamma^2\delta(W,z^n)^2+R_1(W,\delta)\delta(W,z^n)^3]^3\Bigg).\end{aligned} \tag{125}$$

We once again assign $x = 1-\frac{P_{w|z^n}}{P_w^\infty}$, then similar to (112) and (113), for some $a>0$,

$$-\log(1-x) \geq x+\frac{1}{2}x^2+\frac{1}{3}x^3, \tag{126}$$

$$-\log(1-x) \leq x+\frac{1}{2}x^2+\frac{1}{3}x^3+ax^4. \tag{127}$$

Thus for lautum information, we can calculate the value for any given n as

$$n\cdot L(W;S) = n\cdot\int_{\mathcal{W},\mathcal{Z}^n} -\log(1-x)dP_W^{[n]}\otimes\mu^n(w,z^n). \tag{128}$$

Therefore,

$$n\cdot L(W;S) \leq n\cdot\int_{\mathcal{W},\mathcal{Z}^n}(x+\frac{1}{2}x^2+\frac{1}{3}x^3+ax^4)dP_W^{[n]}\otimes\mu^n(w,z^n), \tag{129}$$

$$n\cdot L(W;S) \geq n\cdot\int_{\mathcal{W},\mathcal{Z}^n}(x+\frac{1}{2}x^2+\frac{1}{3}x^3)dP_W^{[n]}\otimes\mu^n(w,z^n), \tag{130}$$

It is natural to then combine (125) (129) and (130). Once again, using the squeeze theorem, the terms with order higher or equal to 3 converges to zero (as proved in Lemma 5 and Annotation 1), and the two sides of the inequalities converges to the same value as shown below

$$\begin{aligned}&\lim_{n\to\infty} n\cdot L(W;S) \\ =& \lim_{n\to\infty} n\cdot\int_{\mathcal{W},\mathcal{Z}^n}\Bigg[\gamma\Big(-\delta(w,z^n)+\mathbb{E}_W^\infty[\delta(W,z^n)]\Big)+ \\ &\gamma^2\Big(-\delta(w,z^n)\mathbb{E}_W^\infty[\delta(W,z^n)]+\frac{1}{2}\delta(w,z^n)^2-\frac{1}{2}\mathbb{E}_W^\infty[\delta(W,z^n)^2]+\mathbb{E}_W^\infty[\delta(W,z^n)]^2+ \\ &\frac{1}{2}\big(\mathbb{E}_W^\infty[\delta(W,z^n)]-\delta(w,z^n)\big)^2\Big)\Bigg]dP_W^{[n]}\otimes\mu^n(w,z^n) \\ =& \lim_{n\to\infty} n\cdot\int_{\mathcal{W},\mathcal{Z}^n}\gamma^2\Big(-\delta(w,z^n)\mathbb{E}_W^\infty[\delta(W,z^n)]+\frac{1}{2}\delta(w,z^n)^2-\frac{1}{2}\mathbb{E}_W^\infty[\delta(W,z^n)^2]+ \\ &\mathbb{E}_W^\infty[\delta(W,z^n)]^2+\frac{1}{2}\big(\mathbb{E}_W^\infty[\delta(W,z^n)]-\delta(w,z^n)\big)^2\Big)dP_W^{[n]}\otimes\mu^n(w,z^n) \\ =& \lim_{n\to\infty}\gamma^2 n\cdot\int_{\mathcal{W},\mathcal{Z}^n}\big(\mathbb{E}_W^\infty[\delta(W,z^n)]-\delta(w,z^n)\big)^2-\frac{1}{2}\big(\mathbb{E}_W^\infty[\delta(W,z^n)^2]-\mathbb{E}_W^\infty[\delta(W,z^n)]^2\big)dP_W^{[n]}\otimes\mu^n(w,z^n) \\ =& \gamma^2\lim_{n\to\infty}\Big\{\mathbb{E}_\mu\big[\mathbb{E}_W^\infty[(\ell(W,Z)-L_\mu(W))]^2\big]+\mathbb{E}_{P_W^{[n]}\otimes\mu}\big[(\ell(W,Z)-L_\mu(W))^2\big]- \\ &2\mathbb{E}_\mu\big[\mathbb{E}_W^\infty[(\ell(W,Z)-L_\mu(W))]\mathbb{E}_{P_W^{[n]}}[(\ell(W,Z)-L_\mu(W))]\big]- \\ &\frac{1}{2}\mathbb{E}_\mu\big[\mathbb{E}_W^\infty[(\ell(W,Z)-L_\mu(W))^2]\big]+\frac{1}{2}\mathbb{E}_\mu\big[\mathbb{E}_W^\infty[(\ell(W,Z)-L_\mu(W))]^2\big]\Big\} \\ =& \frac{\gamma^2}{2}\mathbb{E}_\mu\Big[\mathbb{E}_W^\infty[(\ell(W,Z)-L_\mu(W))^2]-\mathbb{E}_W^\infty[(\ell(W,Z)-L_\mu(W))]^2\Big]\end{aligned} \tag{131}$$

where the final step uses the fact that $P_W^{[n]} \to P_W^\infty$. Comparing (131) with (77) yields

$$\lim_{n\to\infty} n \cdot L(W;S) = \frac{1}{2} \lim_{n\to\infty} n \cdot I_{\mathrm{SKL}}(W;S). \tag{132}$$

Eventually using $I_{\mathrm{SKL}}(W;S) = I(W;S) + L(W;S)$, we get

$$\lim_{n\to\infty} n \cdot I(W;S) = \lim_{n\to\infty} n \cdot L(W;S) = \frac{1}{2} \lim_{n\to\infty} n \cdot I_{\mathrm{SKL}}(W;S). \tag{133}$$

□

L. Proof of Theorem 5

Lemma 6 (Theorem 2 [20]). *Suppose the non-negative loss function $\ell(w, Z)$ is σ-sub-Gaussian on the left-tail under distribution μ for all $w \in \mathcal{W}$. If we further assume $C_E \le \frac{L(W;S)}{I(W;S)}$ for some $C_E \ge 0$, then for the $(\gamma, \pi(w), L_E(w,s))$-Gibbs algorithm, we have*

$$0 \le \mathrm{gen}(P_{W|S}^\gamma, P_S) \le \frac{2\sigma^2\gamma}{(1+C_E)n} \tag{134}$$

Proof. Using Theorem 4, we get

$$\lim_{n\to\infty} \frac{L(W;S)}{I(W;S)} = 1 \tag{135}$$

Therefore, for $\frac{\delta}{2} > 0$, there exists an $N \in \mathbb{N}^+$ such that for all $n > N$,

$$\frac{L(W;S)}{I(W;S)} > 1 - \frac{\delta}{2} \tag{136}$$

Using the fact that a bounded random variable $\ell \in [a,b]$ is $(b-a)/2$-sub-Gaussian, we apply Lemma 6 and gets the result. □

APPENDIX C
MEAN ESTIMATION EXAMPLE

A. Preparation: SKL-divergence between Gaussians

Suppose X, Y follows a multi-Gaussian distribution: $[X,Y] \sim \mathcal{N}(m, \Sigma_{[X,Y]})$, where $\Sigma_{[X,Y]}$ can be written as:

$$\Sigma_{[X,Y]} = \begin{pmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{pmatrix} \tag{137}$$

The independent $\tilde{X}, \tilde{Y} \sim P_X P_Y$ also follows a multi-gaussian distribution, that is $\tilde{X}, \tilde{Y} \sim \mathcal{N}(m, \Sigma_{[\tilde{X},\tilde{Y}]})$, and:

$$\Sigma_{[\tilde{X},\tilde{Y}]} = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix} \tag{138}$$

then:

$$\begin{aligned} I_{\mathrm{SKL}}(X;Y) &= \frac{1}{2} \log \frac{|\Sigma_X||\Sigma_Y|}{|\Sigma_{[X,Y]}|} + \frac{1}{2} \left(\log \frac{|\Sigma_{[X,Y]}|}{|\Sigma_X||\Sigma_Y|} + \mathrm{tr}(\Sigma_{[X,Y]}^{-1} \Sigma_{[\tilde{X},\tilde{Y}]} - I) \right) &(139) \\ &= \frac{1}{2} \mathrm{tr}(\Sigma_{[X,Y]}^{-1} \Sigma_{[\tilde{X},\tilde{Y}]} - I) &(140) \end{aligned}$$

B. Estimating the mean

$S = \{Z_i\}_{i=1}^n$ is the training set, where Z_i is a d dimentional vector sampled i.i.d. from $\mathcal{N}(0_d, (\frac{1}{\sqrt{2\beta}})^2 I_d)$. We consider the problem of learning the means of the dataset. For simplisity, we consider $d = 1$. Loss function is the square error $\ell(w, Z) \triangleq \|w - Z\|_2$. We further choose our prior distribution $\pi(w) = \frac{1}{\sqrt{\pi}} \exp(-w^2)$

Under such settings, it can be calculated that

$$P_{W,S} \sim \mathcal{N}(\{0\}^{n+1}, \Sigma) \tag{141}$$

where Σ is an $n+1$ dimensional matrix as follow

$$\Sigma = \frac{1}{2}\begin{pmatrix} \frac{n\gamma\beta+n\beta+\gamma^2}{n(1+\gamma)^2\beta} & \frac{\gamma}{n\beta(1+\gamma)} & \frac{\gamma}{n\beta(1+\gamma)} & \frac{\gamma}{n\beta(1+\gamma)} & \cdots \\ \frac{\gamma}{n\beta(1+\gamma)} & \frac{1}{\beta} & 0 & 0 & \cdots \\ \frac{\gamma}{n\beta(1+\gamma)} & 0 & \frac{1}{\beta} & 0 & \cdots \\ \frac{\gamma}{n\beta(1+\gamma)} & 0 & 0 & \frac{1}{\beta} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{142}$$

$$\Sigma^{-1} = 2\begin{pmatrix} \gamma+1 & -\frac{\gamma}{n} & -\frac{\gamma}{n} & -\frac{\gamma}{n} & \cdots \\ -\frac{\gamma}{n} & \frac{\gamma^2}{n^2(\gamma+1)}+\beta & \frac{\gamma^2}{n^2(\gamma+1)} & \frac{\gamma^2}{n^2(\gamma+1)} & \cdots \\ -\frac{\gamma}{n} & \frac{\gamma^2}{n^2(\gamma+1)} & \frac{\gamma^2}{n^2(\gamma+1)}+\beta & \frac{\gamma^2}{n^2(\gamma+1)} & \cdots \\ -\frac{\gamma}{n} & \frac{\gamma^2}{n^2(\gamma+1)} & \frac{\gamma^2}{n^2(\gamma+1)} & \frac{\gamma^2}{n^2(\gamma+1)}+\beta & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{143}$$

With the joint distribution, we are able to calculate the symmetrized KL information of S and W using (139), which is the characterization of generalization error

$$\begin{aligned} \gamma \mathrm{gen}(\mu, P_{W|S}) &= I_{\mathrm{SKL}}(W;S) \\ &= \frac{1}{2}\mathrm{tr}(\Sigma_{[W,S]}^{-1}\Sigma_{[\tilde{W},\tilde{S}]} - I) \\ &= \frac{\gamma^2}{n\beta(1+\gamma)} \end{aligned} \tag{144}$$

Next, we consider the symmetrized KL information of Z_i and W. Similarly, we get

$$\begin{aligned} I_{\mathrm{SKL}}(W;Z_i) &= \frac{1}{2}\mathrm{tr}(\Sigma_{[W,Z_i]}^{-1}\Sigma_{[\tilde{W},\tilde{Z}_i]} - I) \\ &= \frac{1}{2}\mathrm{tr}\left(\begin{pmatrix} \frac{n\gamma\beta+n\beta+\gamma^2}{n(1+\gamma)^2\beta} & \frac{\gamma}{n\beta(1+\gamma)} \\ \frac{\gamma}{n\beta(1+\gamma)} & \frac{1}{\beta} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \frac{n\gamma\beta+n\beta+\gamma^2}{n(1+\gamma)^2\beta} & 0 \\ 0 & \frac{1}{\beta} \end{pmatrix} - I\right) \\ &= \frac{\gamma^2}{n^2\beta(1+\gamma)+\gamma^2(n-1)} \end{aligned} \tag{145}$$

From the result of (144) and (145), we get

$$\sum_{i=1}^{n} I_{\mathrm{SKL}}(W;Z_i) - I_{\mathrm{SKL}}(W;S) = \Theta\left(\frac{1}{n^2}\right) = o(I_{\mathrm{SKL}}(W;S)) \tag{146}$$

We can see that (145) and (146) corresponds to Theorem 2 and Theorem 3, despite 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

$$\begin{aligned} J_i^{[n]}(w,z_i) =& \log\left(\sqrt{\frac{n^2(1+\gamma)^2\beta}{\pi(n^2\beta(1+\gamma)+(n-1)\gamma^2)}}\right) - \log\left(\sqrt{\frac{1+\gamma}{\pi}}\right) + \frac{n-1}{n^2}\cdot\frac{\gamma^2}{1+\gamma}\cdot\frac{1}{2\beta} \\ &+ w^2(1+\gamma)\left[1-\frac{n^2(1+\gamma)^2\beta}{n^2\beta(1+\gamma)+(n-1)\gamma^2}\right] - \frac{2\gamma}{n}wz_i\left[1-\frac{n^2(1+\gamma)^2\beta}{n^2\beta(1+\gamma)+(n-1)\gamma^2}\right] \\ &+ \frac{\gamma^2}{n^2(1+\gamma)}z_i^2\left[1-\frac{n^2(1+\gamma)^2\beta}{n^2\beta(1+\gamma)+(n-1)\gamma^2}\right] \end{aligned} \tag{147}$$

$$= w^2\Theta\left(\frac{1}{n}\right) + wz_1\Theta\left(\frac{1}{n^2}\right) + z_1^2\Theta\left(\frac{1}{n^3}\right) + \Theta\left(\frac{1}{n^2}\right). \tag{148}$$

We can see that the contribution of z_i term is $\Theta\left(\frac{1}{n^2}\right)$ which is indeed 1 order smaller than those terms not containing z_i.

Finally, we provide similar analysis on mutual information.

$$I(W;Z_i) = \frac{1}{2}\log\left(1+\frac{\gamma^2}{n^2(1+\gamma)\beta+(n-1)\gamma^2}\right) \tag{149}$$

$$= \Theta\left(\frac{1}{n^2}\right), \tag{150}$$

and

$$I(W;S) = \frac{1}{2}\log\left(1+\frac{\gamma^2}{n\beta(1+\gamma)}\right) \tag{151}$$

$$\sim \frac{1}{2}\cdot\frac{\gamma^2}{n\beta(1+\gamma)} \tag{152}$$

$$= \frac{1}{2}I_{\mathrm{SKL}}(W;S), \tag{153}$$

which corresponds to our result in Theorem 4.