An Exact Characterization of the Generalization Error for the Gibbs Algorithm

Gholamali Aminian[*]
University College London
g.aminian @ucl.ac.uk

Yuheng Bu[*]
Massachusetts Institute of Technology
buyuheng@mit.edu

Laura Toni
University College London
l.toni@ucl.ac.uk

Miguel Rodrigues
University College London
m.rodrigues@ucl.ac.uk

Gregory Wornell
Massachusetts Institute of Technology
gww@mit.edu

Abstract

Various approaches have been developed to upper bound the generalization error of a supervised learning algorithm. However, existing bounds are often loose and lack of guarantees. As a result, they may fail to characterize the exact generalization ability of a learning algorithm. Our main contribution is an exact characterization of the expected generalization error of the well-known Gibbs algorithm (a.k.a. Gibbs posterior) using symmetrized KL information between the input training samples and the output hypothesis. Our result can be applied to tighten existing expected generalization error and PAC-Bayesian bounds. Our approach is versatile, as it also characterizes the generalization error of the Gibbs algorithm with data-dependent regularizer and that of the Gibbs algorithm in the asymptotic regime, where it converges to the empirical risk minimization algorithm. Of particular relevance, our results highlight the role the symmetrized KL information plays in controlling the generalization error of the Gibbs algorithm.

1 Introduction

Evaluating the generalization error of a learning algorithm is one of the most important challenges in statistical learning theory. Various approaches have been developed [55], including VC dimension-based bounds [66], algorithmic stability-based bounds [16], algorithmic robustness-based bounds [72], PAC-Bayesian bounds [45], and information-theoretic bounds [71].

However, upper bounds on generalization error may not entirely capture the generalization ability of a learning algorithm. One apparent reason is the tightness issue, some upper bounds [9] can be far away from the true generalization error or even vacuous when evaluated in practice. More importantly, existing upper bounds do not fully characterize all the aspects that could influence the generalization error of a supervised learning problem. For example, VC dimension-based bounds depend only on the hypothesis class, and algorithmic stability-based bounds only exploit the properties of the learning algorithm. As a consequence, both methods fail to capture the fact that generalization error depends strongly on the interplay between the hypothesis class, learning algorithm, and the

[*] Equal contribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

underlying data-generating distribution, as discussed in [71, 73]. This paper overcomes the above limitations by deriving an exact characterization of the generalization error for a specific supervised learning algorithm, namely the Gibbs algorithm.

1.1 Problem Formulation

Let $S = \{Z_i\}_{i=1}^n$ be the training set, where each Z_i is defined on the same alphabet $\mathcal{Z}$. Note that Z_i is not required to be i.i.d generated from the same data-generating distribution P_Z, and we denote the joint distribution of all the training samples as P_S. We denote the hypotheses by $w \in \mathcal{W}$, where $\mathcal{W}$ is a hypothesis class. The performance of the hypothesis is measured by a non-negative loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_0^+$, and we can define the empirical risk and the population risk associated with a given hypothesis w as

$$L_E(w, s) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i), \quad \text{and} \quad L_P(w, P_S) \triangleq \mathbb{E}_{P_S}[L_E(w, S)], \tag{1}$$

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

$$\overline{\text{gen}}(P_{W|S}, P_S) \triangleq \mathbb{E}_{P_{W,S}}[L_P(W, P_S) - L_E(W, S)], \tag{2}$$

where the expectation is taken over the joint distribution $P_{W,S} = P_{W|S} \otimes P_S$.

In this paper we focus on the generalization error of the Gibbs algorithm (a.k.a. Gibbs posterior [21]). The $(\alpha, \pi(w), f(w, s))$-Gibbs distribution, which was first proposed by [28] in statistical mechanics and further investigated by [35] in information theory, is defined as:

$$P_{W|S}^{\alpha}(w|s) \triangleq \frac{\pi(w)e^{-\alpha f(w,s)}}{V(s, \alpha)}, \quad \alpha \geq 0, \tag{3}$$

where α is the inverse temperature, $\pi(w)$ is an arbitrary prior distribution of W, $f(w, s)$ is energy function, and $V(s, \alpha) \triangleq \int \pi(w)e^{-\alpha f(w,s)}dw$ is the partition function.

If P and Q are probability measures over space $\mathcal{X}$, and P is absolutely continuous with respect to Q, the Kullback-Leibler (KL) divergence between P and Q is given by $D(P\|Q) \triangleq \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP$. If Q is also absolutely continuous with respect to P, then the symmetrized KL divergence (a.k.a. Jeffrey's divergence [36]) is

$$D_{\mathrm{SKL}}(P\|Q) \triangleq D(P\|Q) + D(Q\|P). \tag{4}$$

The mutual information between two random variables X and Y is defined as the KL divergence between the joint distribution and product-of-marginal distribution $I(X; Y) \triangleq D(P_{X,Y}\|P_X \otimes P_Y)$, or equivalently, the conditional KL divergence between $P_{Y|X}$ and P_Y averaged over P_X, $D(P_{Y|X}\|P_Y|P_X) \triangleq \int_{\mathcal{X}} D(P_{Y|X=x}\|P_Y)dP_X(x)$. By swapping the role of $P_{X,Y}$ and $P_X \otimes P_Y$ in mutual information, we get the lautum information introduced by [49], $L(X; Y) \triangleq D(P_X \otimes P_Y\|P_{X,Y})$. Finally, the symmetrized KL information between X and Y is given by [6]:

$$I_{\mathrm{SKL}}(X; Y) \triangleq D_{\mathrm{SKL}}(P_{X,Y}\|P_X \otimes P_Y) = I(X; Y) + L(X; Y). \tag{5}$$

Throughout the paper, upper-case letters denote random variables (e.g., Z), lower-case letters denote the realizations of random variables (e.g., z), and calligraphic letters denote sets (e.g., $\mathcal{Z}$). All the logarithms are the natural ones, and all the information measure units are nats. $\mathcal{N}(\mu, \Sigma)$ denotes the Gaussian distribution with mean μ and covariance matrix Σ.

1.2 Contributions

The core contribution of this paper (see Theorem 1) is an *exact* characterization of the expected generalization error for the Gibbs algorithm in terms of the symmetrized KL information between the input training samples S and the output hypothesis W, as follows:

$$\overline{\text{gen}}(P_{W|S}^{\alpha}, P_S) = \frac{I_{\mathrm{SKL}}(W; S)}{\alpha}.$$

This result highlights the fundamental role of such an information quantity in learning theory that does not appear to have been recognized before. We also discuss some general properties of the symmetrized KL information, which could be used to prove the non-negativity and concavity of the expected generalization error for the Gibbs algorithm.

Building upon this result, we further expand our contributions in various directions:

- In Section 3, we tighten existing expected generalization error bound (see Theorem 2) and PAC-Bayesian bound (see Theorem 3) for Gibbs algorithm under i.i.d and sub-Gaussian assumptions by combining our symmetrized KL information characterization with the existing bounding techniques.
- In Section 4 (Proposition 1 and 2), we show how to use our method to characterize the asymptotic behavior of the generalization error for Gibbs algorithm under large inverse temperature limit $\alpha \to \infty$, where Gibbs algorithm converges to the empirical risk minimization algorithm. Note that existing bounds, such as [39, 52, 71], become vacuous in this regime.
- In Section 5, we characterize the generalization error of the Gibbs algorithm with data-dependent regularizer using symmetrized KL information, which provides some insights on how to reduce the generalization error using regularization.

1.3 Motivations for the Gibbs Algorithm

As we discuss below, the choice of Gibbs algorithm is not arbitrary since it arises naturally in many different applications and is sufficiently general to model many learning algorithms used in practice:

Empirical Risk Minimization: The $(\alpha, \pi(w), L_E(w,s))$-Gibbs algorithm can be viewed as a randomized version of the empirical risk minimization (ERM) algorithm if we specify the energy function $f(w,s) = L_E(w,s)$. As the inverse temperature $\alpha \to \infty$, the prior distribution $\pi(w)$ becomes negligible, and the Gibbs algorithm converges to the standard ERM algorithm.

Information Risk Minimization: The Gibbs algorithm also arises when conditional KL-divergence is used as a regularizer to penalize over-fitting in the information risk minimization framework. In particular, it is shown in [71, 75, 76] that the solution to the following regularized ERM problem

$$P^{\star}_{W|S} = \arg \inf_{P_{W|S}} \left(\mathbb{E}_{P_{W,S}}[L_E(W,S)] + \frac{1}{\alpha} D(P_{W|S} \| \pi(W) | P_S) \right), \tag{6}$$

corresponds to the $(\alpha, \pi(w), L_E(w,s))$-Gibbs distribution. The inverse temperature α controls the regularization term and balances between over-fitting and generalization.

PAC-Bayesian Bound: The following upper bound on population risk from [63] holds with probability at least $1-\delta$ for $0<\delta<1$, and $0<\lambda<2$ under distribution P_S,

$$\mathbb{E}_{P_{W|S=s}}[L_P(W,P_S)] \leq \frac{\mathbb{E}_{P_{W|S=s}}[L_E(W,s)]}{1-\frac{\lambda}{2}} + \frac{D(P_{W|S=s}\|\pi(W)) + \log(\frac{2\sqrt{n}}{\delta})}{\lambda(1-\frac{\lambda}{2})n}. \tag{7}$$

If we fix λ, $\pi(w)$ and optimize over $P_{W|S=s}$, the distribution that minimizes the PAC-Bayes bound in (7) would be the $(n\lambda, \pi(w), L_E(w,s))$-Gibbs distribution. Similar bounds are proposed in [21, Theorem 1.2.1] and [65, Lemma 10], where optimizing over posterior distribution would result in a Gibbs distribution.

SGLD Algorithm: The Stochastic Gradient Langevin Dynamics (SGLD) can be viewed as the discrete version of the continuous-time Langevin diffusion, and it is defined as follows:

$$W_{k+1} = W_k - \beta \nabla L_E(W_k, s) + \sqrt{\frac{2\beta}{\alpha}}\zeta_k, \quad k = 0, 1, \cdots, \tag{8}$$

where ζ_k is a standard Gaussian random vector and $\beta > 0$ is the step size. In [51], it is proved that under some conditions on loss function, the conditional distribution $P_{W_k|S}$ induced by SGLD algorithm is close to $(\alpha, \pi(W_0), L_E(w_k, s))$-Gibbs distribution in 2-Wasserstein distance for sufficiently large k. Under some conditions on the loss function $\ell(w,z)$, [22, 42] shows that in the continuous-time Langevin diffusion, the stationary distribution of hypothesis W is the Gibbs distribution.

1.4 Other Related Work

Information-theoretic generalization error bounds: Recently, [58, 71] propose to use the mutual information between the input training set and the output hypothesis to upper bound the expected generalization error. However, those bounds are known not to be tight, and multiple approaches have been proposed to tighten the mutual information-based bound. [19] provides tighter bounds by considering the individual sample mutual information, [10, 11] propose using chaining mutual information, and [30, 31, 62] advocate the conditioning and processing techniques. Information-theoretic generalization error bounds using other information quantities are also studied, such as, f-divergence [37], α-Rényi divergence and maximal leakage [25, 34], Jensen-Shannon divergence [7] and Wasserstein distance [41, 56, 69]. Using rate-distortion theory, [17, 18, 43] provide information-theoretic generalization error upper bounds for model misspecification and model compression.

PAC-Bayesian generalization error bounds: First proposed by [44, 45, 60], PAC-Bayesian analysis provides high probability bounds on the generalization error in terms of KL divergence between the data-dependent posterior induced by the learning algorithm and a data-free prior that can be chosen arbitrarily. There are multiple ways to generalize the standard PAC-Bayesian bounds, including using different information measures other than the KL divergence [3, 8, 14, 32, 48] and considering data-dependent priors (prior depends on the training data) [5, 13, 21, 23, 24, 53] or distribution-dependent priors (prior depends on data-generating distribution) [20, 40, 50, 54]. In [27], a more general PAC-Bayesian framework is proposed, which provides a high probability bound on the convex function of the expected population and empirical risk with respect to the posterior distribution, whereas in [26] the connection between Bayesian inference and PAC-Bayesian theorem is explored by considering Gibbs posterior and negative log loss function.

Generalization error of Gibbs algorithm: Both information-theoretic and PAC-Bayesian approaches have been used to bound the generalization error of the Gibbs algorithm. An information-theoretic upper bound with a convergence rate of $\mathcal{O}\left(\alpha/n\right)$ is provided in [52] for the Gibbs algorithm with bounded loss function, and PAC-Bayesian bounds using a variational approximation of Gibbs posteriors are studied in [4]. [11, Appendix D] provides an upper bound on the excess risk of the Gibbs algorithm under sub-Gaussian assumption. [39] focuses on the excess risk of the Gibbs algorithm and a similar generalization bound with rate of $\mathcal{O}\left(\alpha/n\right)$ is provided under sub-Gaussian assumption. Although these bounds are tight in terms of the sample complexity n, they become vacuous when the inverse temperature $\alpha \to \infty$, hence are unable to capture the behaviour of the ERM algorithm.

Our work differs from this body of research in the sense that we provide an exact characterization of generalization error of the Gibbs algorithm in terms of the symmetrized KL information. Our work also further leverages this characterization to tighten the existing expected and PAC-Bayesian generalization error bounds in literature.

2 Generalization Error of Gibbs Algorithm

Our main result, which characterizes the exact expected generalization error of the Gibbs algorithm with prior distribution $\pi(w)$, is as follows:

Theorem 1. *For $(\alpha, \pi(w), L_E(w,s))$-Gibbs algorithm,*

$$P^{\alpha}_{W|S}(w|s) = \frac{\pi(w)e^{-\alpha L_E(w,s)}}{V(s,\alpha)}, \quad \alpha > 0, \tag{9}$$

the expected generalization error is given by

$$\overline{gen}(P^{\alpha}_{W|S}, P_S) = \frac{I_{\mathrm{SKL}}(W;S)}{\alpha}. \tag{10}$$

Sketch of Proof: It can be shown that the symmetrized KL information can be written as

$$I_{\mathrm{SKL}}(W;S) = \mathbb{E}_{P_{W,S}}[\log(P^{\alpha}_{W|S})] - \mathbb{E}_{P_W \otimes P_S}[\log(P^{\alpha}_{W|S})]. \tag{11}$$

Just like the generalization error, the above expression is the difference between the expectations of the same function evaluated under the joint distribution and the product-of-marginal distribution. Note that $P_{W,S}$ and $P_W \otimes P_S$ share the same marginal distribution, we have $\mathbb{E}_{P_{W,S}}[\log \pi(W)] =$

$\mathbb{E}_{P_W}[\log \pi(W)]$, and $\mathbb{E}_{P_{W,S}}[\log V(S,\alpha)] = \mathbb{E}_{P_S}[\log V(S,\alpha)]$. Then, combining (9) with (11) completes the proof. More details together with the full proof are provided in Appendix B.1. □

To the best of our knowledge, this is the first exact characterization of the expected generalization error for the Gibbs algorithm. Note that Theorem 1 only assumes that the loss function is non-negative, and it holds even for non-i.i.d training samples.

In Section 2.1, we discuss some general properties of the expected generalization error that can be learned directly from the properties of symmetrized KL information. In Section 2.2, we provide a mean estimation example to show that the symmetrized KL information can be computed exactly for squared loss with Gaussian prior.

2.1 General Properties

By Theorem 1, some basic properties of the expected generalization error, including non-negativity and concavity, can be proved directly from the properties of symmetrized KL information.

The non-negativity of the expected generalization error, i.e., $\overline{\text{gen}}(P^\alpha_{W|S}, P_S) \geq 0$, follows by the non-negativity of the symmetrized KL information. Note that the non-negativity result could also be proved using [39, Appendix A.2] under much more stringent assumptions, including i.i.d samples and a sub-Gaussian loss function.

It is shown in [6] that the symmetrized KL information $I_{\text{SKL}}(X;Y)$ is a concave function of P_X for fixed $P_{Y|X}$, and a convex function of $P_{Y|X}$ for fixed P_X. Thus, we have the following corollary.

Corollary 1. *For a fixed $(\alpha, \pi(w), L_E(w,s))$-Gibbs algorithm $P^\alpha_{W|S}$, the expected generalization error $\overline{gen}(P^\alpha_{W|S}, P_S)$ is a concave function of P_S.*

The concavity of the generalization error for the Gibbs algorithm $P^\alpha_{W|S}$ can be immediately used to explain why training a model by mixing multiple datasets from different domains leads to poor generalization. Suppose that the data-generating distribution is domain-dependent, i.e., there exists a random variable D, such that $D \leftrightarrow S \leftrightarrow W$ holds. Then, $P_S = \mathbb{E}_{P_D}[P_{S|D}]$ can be viewed as the mixture of the data-generating distribution across all domains. From Corollary 1 and Jensen's inequality, we have

$$\overline{\text{gen}}(P^\alpha_{W|S}, P_S) \geq \mathbb{E}_{P_D}\big[\overline{\text{gen}}(P^\alpha_{W|S}, P_{S|D})\big], \tag{12}$$

which shows that the generalization error of Gibbs algorithm achieved with the mixture distribution P_S is larger than the averaged generalization error for each $P_{S|D}$.

More discussions about other properties of symmetrized KL divergence, including data processing inequality (symmetrized KL divergence is an f-divergence), variational representation, chain rule, and their implications in learning problems are provided in Appendix B.2.

2.2 Example: Mean Estimation

We now consider a simple learning problem, where the symmetrized KL information can be computed exactly, to demonstrate the usefulness of Theorem 1. All details are provided in Appendix B.3.

Consider the problem of learning the mean $\boldsymbol{\mu} \in \mathbb{R}^d$ of a random vector Z using n i.i.d training samples $S = \{Z_i\}_{i=1}^n$. We assume that the covariance matrix of Z satisfies $\Sigma_Z = \sigma_Z^2 I_d$ with unknown σ_Z^2. We adopt the mean-squared loss $\ell(\boldsymbol{w}, \boldsymbol{z}) = \|\boldsymbol{z} - \boldsymbol{w}\|_2^2$, and assume a Gaussian prior for the mean $\pi(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 I_d)$. If we set inverse-temperature $\alpha = \frac{n}{2\sigma^2}$, then the $(\frac{n}{2\sigma^2}, \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 I_d), L_E(\boldsymbol{w}, s))$-Gibbs algorithm is given by the following posterior distribution [47],

$$P^\alpha_{W|S}(\boldsymbol{w}|Z^n) \sim \mathcal{N}\Big(\frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{\sigma_1^2}{\sigma^2}\sum_{i=1}^n Z_i, \sigma_1^2 I_d\Big), \quad \text{with} \quad \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}. \tag{13}$$

Since $P^\alpha_{W|S}$ is Gaussian, the mutual information and lautum information are given by

$$I(S;W) = \frac{nd\sigma_0^2\sigma_Z^2}{2(n\sigma_0^2 + \sigma^2)\sigma^2} - D\big(P_W \| \mathcal{N}(\boldsymbol{\mu}_W, \sigma_1^2 I_d)\big), \tag{14}$$

$$L(S;W) = \frac{nd\sigma_0^2\sigma_Z^2}{2(n\sigma_0^2 + \sigma^2)\sigma^2} + D\big(P_W \| \mathcal{N}(\boldsymbol{\mu}_W, \sigma_1^2 I_d)\big), \quad \text{with} \quad \boldsymbol{\mu}_W = \frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{n\sigma_1^2}{\sigma^2}\boldsymbol{\mu}. \tag{15}$$

For additive Gaussian channel $P_{W|S}$, it is well known that Gaussian input distribution (which also gives a Gaussian output distribution P_W) maximizes the mutual information under a second-order moment constraint. As we can see from the above expressions, the opposite is true for lautum information. In addition, symmetrized KL information $I_{\mathrm{SKL}}(W;S)$ is independent of the distribution of P_Z, as long as $\Sigma_Z = \sigma_Z^2 I_d$.

From Theorem 1, the generalization error of this algorithm can be computed exactly as:

$$\overline{\mathrm{gen}}(P_{W|S}^{\alpha}, P_S) = \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} = \frac{2d\sigma_0^2\sigma_Z^2}{n\sigma_0^2 + \sigma^2} = \frac{2d\sigma_0^2\sigma_Z^2}{n(\sigma_0^2 + \frac{1}{2\alpha})}, \tag{16}$$

which has the decay rate of $\mathcal{O}(1/n)$. As a comparison, the individual sample mutual information (ISMI) bound from [19], which is shown to be tighter than the mutual information-based bound in [71, Theorem 1], gives a sub-optimal bound with order $\mathcal{O}(1/\sqrt{n})$, as $n \to \infty$, (see Appendix B.4).

3 Tighter Generalization Error Upper Bounds

In this section, we show that by combining Theorem 1 with existing information-theoretic and PAC-Bayesian approaches, we can provide tighter generalization error upper bounds for the Gibbs algorithm. These bounds quantify how the generalization error of the Gibbs algorithm depends on the number of samples n, and are useful when directly evaluating the symmetrized KL information is hard.

3.1 Expected Generalization Error Upper Bound

The following upper bound on the expected generalization error for the Gibbs algorithm can be obtained by combining our Theorem 1 with the information-theoretic bound proposed in [71] under i.i.d and sub-Gaussian assumptions.

Theorem 2. *(proved in Appendix C.1) Suppose that the training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z, and the non-negative loss function $\ell(w, Z)$ is σ-sub-Gaussian on the left-tail* [*] *under distribution P_Z for all $w \in \mathcal{W}$. If we further assume $C_E \leq \frac{L(W;S)}{I(W;S)}$ for some $C_E \geq 0$, then for the $(\alpha, \pi(w), L_E(w,s))$-Gibbs algorithm, we have*

$$0 \leq \overline{gen}(P_{W|S}^{\alpha}, P_S) \leq \frac{2\sigma^2\alpha}{(1+C_E)n}. \tag{17}$$

Theorem 2 establishes the convergence rate $\mathcal{O}(\alpha/n)$ of the generalization error of Gibbs algorithm with i.i.d training samples, and suggests that a smaller inverse temperature α leads to a tighter upper bound. Note that all the σ-sub-Gaussian loss functions are also σ-sub-Gaussian on the left-tail under the same distribution (loss function in Section 2.2 is sub-Gaussian on the left-tail under P_Z, but not sub-Gaussian). Therefore, our result also applies to any bounded loss function $\ell : \mathcal{W} \times \mathcal{Z} \to [a, b]$, since bounded functions are $(\frac{b-a}{2})$-sub-Gaussian.

Remark 1 (Previous Results)**.** *Using the fact that Gibbs algorithm is differentially private [46] for bounded loss functions $\ell \in [0,1]$, directly applying [71, Theorem 1] gives a sub-optimal bound $|\overline{gen}(P_{W|S}^{\alpha}, P_S)| \leq \sqrt{\frac{\alpha}{n}}$. By further exploring the bounded loss assumption using Hoeffding's lemma, a tighter upper bound $|\overline{gen}(P_{W|S}^{\alpha}, P_S)| \leq \frac{\alpha}{2n}$ is obtained in [52], which has the similar decay rate order of $\mathcal{O}(\alpha/n)$. In [39, Theorem 1], the upper bound $\overline{gen}(P_{W|S}^{\alpha}, P_S) \leq \frac{4\sigma^2\alpha}{n}$ is derived with a different assumption, i.e., $\ell(W,z)$ is σ-sub-Gaussian under Gibbs algorithm $P_{W|S}^{\alpha}$. In Theorem 2, we assume the loss function is σ-sub-Gaussian on left-tail under data-generating distribution P_Z for all $w \in \mathcal{W}$, which is more general as we discussed above. Our upper bound is also improved by a factor of $\frac{1}{2(1+C_E)}$ compared to the result in [39].*

Remark 2 (Choice of C_E)**.** *Since $L(W;S) > 0$ when $I(W;S) > 0$, setting $C_E = 0$ is always valid in Theorem 2, which gives $\overline{gen}(P_{W|S}^{\alpha}, P_S) \leq \frac{2\sigma^2\alpha}{n}$. As shown in [49, Theorem 15], $L(S;W) \geq$*

[*] A random variable X is σ-sub-Gaussian if $\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \frac{\sigma^2\lambda^2}{2}$, $\forall \lambda \in \mathbb{R}$, and X is σ-sub-Gaussian on the left-tail if $\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \frac{\sigma^2\lambda^2}{2}$, $\forall \lambda \leq 0$.

$I(S;W)$ holds for any Gaussian channel $P_{W|S}$. In addition, it is discussed in [49, Example 1], if either the entropy of training S or the hypothesis W is small, $I(S;W)$ would be smaller than $L(S;W)$ (as it is not upper-bounded by the entropy), which implies that the lautum information term is not negligible in general.

We extend Theorem 2 by considering other types of tail behavior including sub-Gamma and sub-Exponential on left-tail in Appendix C.2.

3.2 PAC-Bayesian Upper Bound

As discussed in Section 1.4, the prior distribution used in PAC-Bayesian bounds is different from the prior in Gibbs algorithm, since the former priors can be chosen arbitrarily to tighten the generalization error bound. In this section, we provide a tighter PAC-Bayesian bound based on the symmetrized KL divergence, which is inspired by the distribution-dependent PAC-Bayesian bound proposed in [40] using $(\alpha, \pi(w), L_P(w, P_S))$-Gibbs distribution as the PAC-Bayesian prior.

As the data-generating distribution P_S is unknown in practice, we consider the $(\alpha, \pi(w), L_P(w, P_{S'}))$-Gibbs distribution in the following discussion, where $P_{S'}$ is an arbitrary data-generating distribution. Since $(\alpha, \pi(w), L_P(w, P_{S'}))$-Gibbs distribution is independent of the samples S and only depends on the population risk $L_P(w, P_{S'})$, we can denote it as P_W^{α, L'_P}.

By exploiting the connection between the symmetrized KL divergence $D_{\mathrm{SKL}}(P^\alpha_{W|S=s} \| P_W^{\alpha, L'_P})$ and the KL divergence term $D(P^\alpha_{W|S=s} \| P_W^{\alpha, L'_P})$in the PAC-Bayesian bound from [40], the following PAC-Bayesian bound can be obtained under i.i.d and sub-Gaussian assumptions.

Theorem 3. *(proved in Appendix D) Suppose that the training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z, and the non-negative loss function $\ell(w, Z)$ is σ-sub-Gaussian under data-generating distribution P_Z for all $w \in \mathcal{W}$. If we use the $(\alpha, \pi(w), L_P(w, P_{S'}))$-Gibbs distribution as the PAC-Bayesian prior, where $P_{S'}$ is an arbitrary chosen (and known) distribution, the following upper bound holds for the generalization error of $(\alpha, \pi(w), L_E(w, s))$-Gibbs algorithm with probability at least $1 - 2\delta$, $0 < \delta < \frac{1}{2}$ under distribution P_S,*

$$\left|\mathbb{E}_{P^\alpha_{W|S=s}}[L_P(W, P_S) - L_E(W, s)]\right| \leq \frac{2\sigma^2\alpha}{(1 + C_P(s))n} + 2\sqrt{\frac{\sigma^2\alpha}{(1 + C_P(s))n}} \left(\sqrt[4]{2\sigma^2 D(P_{Z'} \| P_Z)} + \epsilon\right) + \epsilon^2,$$

where $\epsilon \triangleq \sqrt[4]{\frac{2\sigma^2 \log(1/\delta)}{n}}$, and $C_P(s) \leq \frac{D\left(P_W^{\alpha, L'_P} \middle\| P^\alpha_{W|S=s}\right)}{D\left(P^\alpha_{W|S=s} \middle\| P_W^{\alpha, L'_P}\right)}$ for some $C_P(s) \geq 0$.

Remark 3 (Previous Result)**.** *We could recover the distribution-dependent bound in [40, Theorem 6] by setting $P_{Z'} = P_Z$, choosing a bounded loss function in $[0, 1]$ and $C_P(s) = 0$ in our Theorem 3. Note that multiple terms in our upper bound in Theorem 3 are tightened by a factor of $1/(1 + C_P(s))$, and we also consider σ-sub-Gaussian loss functions.*

Remark 4 (Choice of $C_P(s)$)**.** *Since the distribution $P_{Z'}$ can be set arbitrarily, the prior distribution P_W^{α, L'_P} is accessible. Then, we can set $C_P(s) = D(P_W^{\alpha, L'_P} \| P^\alpha_{W|S=s}) / D(P^\alpha_{W|S=s} \| P_W^{\alpha, L'_P})$ to tighten the bound, as it can be computed exactly using the training set.*

4 Asymptotic Behavior of Generalization Error for Gibbs Algorithm

In this section, we consider the asymptotic behavior of the generalization error for Gibbs algorithm as the inverse temperature $\alpha \to \infty$. Note that the upper bounds obtained in the previous section, as well as the ones in the literature, have the order $\mathcal{O}(\frac{\alpha}{n})$, which becomes vacuous in this regime. However, it is known that the Gibbs algorithm will converge to ERM as $\alpha \to \infty$, which has finite generalization error with bounded loss function. To resolve this issue, we provide an exact characterization of the generalization error in this regime using Theorem 1.

It is shown in [12, 33] that the asymptotic behavior of the Gibbs algorithm depends on the number of minimizers for the empirical risk, so we consider the single-well case and multiple-well case separately.

Single-well case: In this case, there exists a unique $W^*(S)$ that minimizes the empirical risk, i.e.,

$$W^*(S) = \arg\min_{w \in \mathcal{W}} L_E(w, S). \tag{18}$$

It is shown in [33] that if $H^*(S) \triangleq \nabla_w^2 L_E(w,S)\big|_{w=W^*(S)}$is not singular, then $P_{W|S}^{\alpha} \to \mathcal{N}(W^*(S), \frac{1}{\alpha}H^*(S)^{-1})$ in distribution. Thus, the symmetrized KL information in Theorem 1 can be evaluated using this Gaussian approximation, which gives the following result.

Proposition 1. *(proved in Appendix E.1) In the single-well case, if the Hessian matrix $H^*(S)$ is not singular, then the generalization error of the $(\infty, \pi(\boldsymbol{w}), L_E(\boldsymbol{w},s))$-Gibbs algorithm is*

$$\overline{gen}(P_{W|S}^{\infty}, P_S) = \mathbb{E}_{\Delta_{W,S}}\big[\frac{1}{2}W^\top H^*(S)W\big] \tag{19}$$
$$+ \mathbb{E}_{P_S}\Big[(W^*(S) - \mathbb{E}[W^*(S)])^\top (H^*(S)W^*(S) - \mathbb{E}[H^*(S)W^*(S)])\Big],$$

where $\mathbb{E}_{\Delta_{W,S}}[f(W,S)] \triangleq \mathbb{E}_{P_W \otimes P_S}[f(W,S)] - \mathbb{E}_{P_{W,S}}[f(W,S)]$.

Proposition 1 shows that the generalization error of the Gibbs algorithm in the limiting regime $\alpha \to \infty$ highly depends on the landscape of the empirical risk function.

As an example, we use Proposition 1 to obtain the generalization error of the maximum likelihood estimates (MLE) in the asymptotic regime $n \to \infty$. More specifically, suppose that we have n i.i.d. training samples generated from the distribution P_Z, and we want to fit the training data with a parametric distribution family $\{f(z_i|\boldsymbol{w})\}_{i=1}^n$, where $\boldsymbol{w} \in \mathcal{W} \subset \mathbb{R}^d$ denotes the parameter. Here, the true data-generating distribution may not belong to the parametric family, i.e., $P_Z \neq f(\cdot|\boldsymbol{w})$ for $\boldsymbol{w} \in \mathcal{W}$. If we use the log-loss $\ell(\boldsymbol{w}, z) = -\log f(z|\boldsymbol{w})$ in the Gibbs algorithm, as $\alpha \to \infty$, it converges to the ERM algorithm, which is equivalent to MLE, i.e.,

$$W^*(S) = \hat{W}_{\mathrm{ML}} \triangleq \arg\max_{\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n \log f(Z_i|\boldsymbol{w}). \tag{20}$$

As $n \to \infty$, under regularization conditions (details in Appendix E.2) which guarantee that $W^*(S)$ is unique, the asymptotic normality of the MLE [64] states that the distribution of $\hat{W}_{\mathrm{ML}}$ converges to

$$\mathcal{N}(\boldsymbol{w}^*, \frac{1}{n}J(\boldsymbol{w}^*)^{-1}\mathcal{I}(\boldsymbol{w}^*)J(\boldsymbol{w}^*)^{-1}), \quad \text{with} \quad \boldsymbol{w}^* \triangleq \arg\min_{\boldsymbol{w}\in\mathcal{W}} D(P_Z \| f(\cdot|\boldsymbol{w})),$$
$$J(\boldsymbol{w}) \triangleq \mathbb{E}_Z\big[-\nabla_{\boldsymbol{w}}^2 \log f(Z|\boldsymbol{w})\big] \quad \text{and} \quad \mathcal{I}(\boldsymbol{w}) \triangleq \mathbb{E}_Z\big[\nabla_{\boldsymbol{w}} \log f(Z|\boldsymbol{w}) \nabla_{\boldsymbol{w}} \log f(Z|\boldsymbol{w})^\top\big].$$

In addition, the Hessian matrix $H^*(S) \to J(\boldsymbol{w}^*)$ as $n \to \infty$, which is independent of the training samples S. Thus, $\mathbb{E}_{\Delta_{W,S}}[\frac{1}{2}W^\top H^*(S)W] = 0$, and Proposition 1 gives

$$\overline{\mathrm{gen}}(P_{W|S}^{\infty}, P_S) = \frac{\mathrm{tr}(\mathcal{I}(\boldsymbol{w}^*)J(\boldsymbol{w}^*)^{-1})}{n}. \tag{21}$$

When the true model is in the parametric family $P_Z = f(\cdot|\boldsymbol{w}^*)$, we have $\mathcal{I}(\boldsymbol{w}^*) = J(\boldsymbol{w}^*)$ and the above expression reduces to $\overline{\mathrm{gen}}(P_{W|S}^{\infty}, P_Z) = \frac{d}{n}$, which corresponds to the well-known Akaike information criterion (AIC) [2] used in MLE model selection.

Multiple-well case: In this case, there exist M distinct $W_u^*(S)$ such that

$$W_u^*(S) \in \arg\min_{w\in\mathcal{W}} L_E(w,S), \quad u \in \{1, \cdots, M\}, \tag{22}$$

where M is a fixed constant, and all the minimizers $W_u^*(S)$ are isolated, meaning that a sufficiently small neighborhood of each $W_u^*(S)$ contains a unique minimum.

In this multiple-well case, it is shown in [12] that the the Gibbs algorithm can be approximated by a Gaussian mixture, as long as $H_u^*(S) \triangleq \nabla_w^2 L_E(w,S)\big|_{w=W_u^*(S)}$ is not singular for all $u \in \{1, \cdots, M\}$. However, there is no closed form for the symmetrized KL information for Gaussian mixtures. Thus, we provide the following upper bound of the generalization error by evaluating Theorem 1 under the assumption that $\pi(W)$ is a uniform distribution over $\mathcal{W}$.

Proposition 2. *(proved in Appendix E.1) If we assume that $\pi(W)$ is a uniform distribution over $\mathcal{W}$, and the Hessian matrices $H_u^*(S)$ are not singular for all $u \in \{1, \cdots, M\}$, then the generalization error of the $(\infty, \pi(\boldsymbol{w}), L_E(\boldsymbol{w},s))$-Gibbs algorithm in the multiple-well case can be bounded as*

$$\overline{gen}(P_{W|S}^{\infty}, P_S) \leq \frac{1}{M}\sum_{u=1}^M \Big[\mathbb{E}_{\Delta_{W_u,S}}\big[\frac{1}{2}W_u^\top H_u^*(S)W_u\big]$$
$$+ \mathbb{E}_{P_S}\Big[(W_u^*(S) - \mathbb{E}[W_u^*(S)])^\top H_u(W_u^*(S) - \mathbb{E}[W_u^*(S)])\Big]\Big]. \tag{23}$$

Comparing with Proposition 1, Proposition 2 shows that the global generalization error in the multiple-well case can be upper bounded by the mean of the generalization errors achieved by each local minimizer.

Remark 5. *In [39], a similar Gaussian approximation technique is used to bound the* excess risk *of Gibbs algorithm in both single-well and multiple-well cases. However, their result is based on a loose generalization error bound with the order $\mathcal{O}(\frac{\alpha}{n})$. Thus, our method can also be used to obtain a tighter characterization of the excess risk for the Gibbs algorithm.*

In Appendix E.3, we consider a slightly different asymptotic regime, where the Gibbs algorithm converges to the Bayesian posterior instead of ERM. A similar result as in (21) can be obtained from Bernstein–von–Mises theorem [38] and the asymptotic normality of the MLE.

5 Regularized Gibbs Algorithm

In this section, we show how regularization will influence the generalization error of the Gibbs algorithm. Our definition of the regularizer is more general compared to the standard data-independent regularizer, as it may also depend on the training samples. There are many applications of such data-dependent regularization in the literature—e.g., data-dependent spectral norm regularization is proposed in [57], ℓ_1 regularizer over data-dependent hypothesis space is studied in [70] and dropout is modeled as data-dependent ℓ_2 regularization in [68].

In the following proposition, we consider the Gibbs algorithm with a regularization term $R : \mathcal{W} \times \mathcal{Z}^n \to \mathbb{R}_0^+$ and characterize the generalization error of this $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm, which is the solution of the following regularized ERM problem:

$$P_{W|S}^{\star} = \arg\inf_{P_{W|S}} \left(\mathbb{E}_{P_{W,S}}[L_E(W,S) + \lambda R(W,S)] + \frac{1}{\alpha} D(P_{W|S} \| \pi(W) | P_S) \right), \tag{24}$$

where $\lambda \geq 0$ controls the regularization term.

Proposition 3. *(proved in Appendix F.1) For $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm, its expected generalization error is given by*

$$\overline{gen}(P_{W|S}^{\alpha}, P_S) = \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} - \lambda \mathbb{E}_{\Delta_{W,S}}[R(W,S)], \tag{25}$$

where $\mathbb{E}_{\Delta_{W,S}}[R(W,S)] = \mathbb{E}_{P_W \otimes P_S}[R(W,S)] - \mathbb{E}_{P_{W,S}}[R(W,S)]$.

Proposition 3 holds for non-i.i.d samples and any non-negative loss function, and it shows that to improve the generalization ability of the Gibbs algorithm, the data-dependent regularizer needs to 1) minimize the symmetrized KL information $I_{\mathrm{SKL}}(W;S)$ and 2) maximize the $\mathbb{E}_{\Delta_{W,S}}[R(W,S)]$ term which corresponds to a “generalization error” defined with the regularization term $R(W,S)$.

Remark 6. *If the regularizer is independent of the data, i.e., $R(w,s) = R(w)$, we have $\mathbb{E}_{\Delta_{W,S}}[R(W,S)] = 0$, and Proposition 3 gives $\overline{gen}(P_{W|S}^{\alpha}, P_S) = \frac{I_{\mathrm{SKL}}(W;S)}{\alpha}$, which implies that the data-independent regularizer needs to improve the generalization ability of learning algorithm by reducing the symmetrized KL information $I_{\mathrm{SKL}}(W;S)$.*

Inspired by the data-dependent regularizer proposed in [61] for support vector machines, we consider a similar data-dependent ℓ_2-regularizer in the following proposition.

Proposition 4. *(proved in Appendix F.1) Suppose that we adopt the ℓ_2-regularizer $R(w,s) = \|w - T(s)\|_2^2$, where $T(\cdot)$ is an arbitrary deterministic function $T : \mathcal{Z}^n \to \mathcal{W}$. Then, the expected generalization error of $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm is*

$$\overline{gen}(P_{W|S}^{\alpha}, P_S) = \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} - \lambda \mathrm{tr}\big(\mathrm{Cov}[W, T(S)]\big), \tag{26}$$

where $\mathrm{Cov}[W, T(S)]$ denotes the covariance matrix between W and $T(S)$.

Our result suggests that to reduce the generalization error with data-dependent ℓ_2-regularizer, the function $T(S)$ should be chosen in a way, such that the term $\mathrm{tr}(\mathrm{Cov}[W, T(S)])$ is maximized. One way is to leave a part of the training set and learn the $T(S)$ function. Note that similar idea has been explored in the development of PAC-Bayesian bound with data-dependent prior [5]. More discussions and results about data-dependent regularizer are provided in Appendix F.2.

6 Conclusion

We provide an exact characterization of the generalization error for the Gibbs algorithm using symmetrized KL information. We demonstrate the power and versatility of our approach in multiple applications, including tightening expected generalization error and PAC-Bayesian bounds, characterizing the behaviors of the Gibbs algorithm with large inverse temperature and the regularized Gibbs algorithm.

This work motivates further investigation of the Gibbs algorithm in a variety of settings, including extending our results to characterize the generalization ability of an over-parameterized Gibbs algorithm, which could potentially provide more understanding of the generalization ability for deep learning.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their valuable feedback, which helped us to improve the paper greatly. Gholamali Aminian is supported by the Royal Society Newton International Fellowship, grant no. NIF\R1 \192656. Yuheng Bu is supported, in part, by NSF under Grant CCF-1717610 and by the MIT-IBM Watson AI Lab.

References

[1] Abou-Moustafa, K., and Szepesvári, C. An exponential Efron-Stein inequality for L_q stable learning rules. In *Algorithmic Learning Theory* (2019), PMLR, pp. 31–63.

[2] Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[3] Alquier, P., and Guedj, B. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning 107*, 5 (2018), 887–902.

[4] Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research 17*, 1 (2016), 8374–8414.

[5] Ambroladze, A., Parrado-Hernández, E., and Shawe-Taylor, J. Tighter PAC-Bayes bounds. *Advances in neural information processing systems 19* (2007), 9.

[6] Aminian, G., Arjmandi, H., Gohari, A., Nasiri-Kenari, M., and Mitra, U. Capacity of diffusion-based molecular communication networks over LTI-Poisson channels. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications 1*, 2 (2015), 188–201.

[7] Aminian, G., Toni, L., and Rodrigues, M. R. Jensen-Shannon information based characterization of the generalization error of learning algorithms. In *2020 IEEE Information Theory Workshop (ITW)* (2020), IEEE.

[8] Aminian, G., Toni, L., and Rodrigues, M. R. Information-theoretic bounds on the moments of the generalization error of learning algorithms. In *IEEE International Symposium on Information Theory (ISIT)* (2021).

[9] Anthony, M., and Bartlett, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[10] Asadi, A., Abbe, E., and Verdú, S. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems* (2018), pp. 7234–7243.

[11] Asadi, A. R., and Abbe, E. Chaining meets chain rule: Multilevel entropic regularization and training of neural networks. *Journal of Machine Learning Research 21*, 139 (2020), 1–32.

[12] Athreya, K., and Hwang, C.-R. Gibbs measures asymptotics. *Sankhya A 72*, 1 (2010), 191–207.

[13] Banerjee, P. K., and Montúfar, G. Information complexity and generalization bounds. In *IEEE International Symposium on Information Theory (ISIT)* (2021).

[14] Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics* (2016), PMLR, pp. 435–444.

[15] Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[16] Bousquet, O., and Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research 2*, Mar (2002), 499–526.

[17] Bu, Y., Gao, W., Zou, S., and Veeravalli, V. V. Information-theoretic understanding of population risk improvement with model compression. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 3300–3307.

[18] Bu, Y., Gao, W., Zou, S., and Veeravalli, V. V. Population risk improvement with model compression: An information-theoretic approach. *Entropy 23*, 10 (2021), 1255.

[19] Bu, Y., Zou, S., and Veeravalli, V. V. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory 1*, 1 (2020), 121–130.

[20] Catoni, O. A PAC-Bayesian approach to adaptive classification. *preprint 840* (2003).

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

[22] Chiang, T.-S., Hwang, C.-R., and Sheu, S. J. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization 25*, 3 (1987), 737–753.

[23] Dziugaite, G. K., and Roy, D. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors. In *International Conference on Machine Learning* (2018), PMLR, pp. 1377–1386.

[24] DZIUGAITE, G. K., AND ROY, D. M. Data-dependent PAC-Bayes priors via differential privacy. In *Advances in Neural Information Processing Systems* (2018), pp. 8440–8450.

[25] ESPOSITO, A. R., GASTPAR, M., AND ISSA, I. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory* (2021).

[26] GERMAIN, P., BACH, F., LACOSTE, A., AND LACOSTE-JULIEN, S. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems* (2016), pp. 1884–1892.

[27] GERMAIN, P., LACASSE, A., LAVIOLETTE, F., MARCH, M., AND ROY, J.-F. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research 16*, 26 (2015), 787–860.

[28] GIBBS, J. W. Elementary principles of statistical mechanics. *Compare 289* (1902), 314.

[29] GUEDJ, B., AND PUJOL, L. Still no free lunches: the price to pay for tighter PAC-Bayes bounds. *arXiv preprint arXiv:1910.04460* (2019).

[30] HAFEZ-KOLAHI, H., GOLGOONI, Z., KASAEI, S., AND SOLEYMANI, M. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems 33* (2020).

[31] HAGHIFAM, M., NEGREA, J., KHISTI, A., ROY, D. M., AND DZIUGAITE, G. K. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems* (2020).

[32] HELLSTRÖM, F., AND DURISI, G. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory* (2020).

[33] HWANG, C.-R. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability* (1980), 1177–1182.

[34] ISSA, I., ESPOSITO, A. R., AND GASTPAR, M. Strengthened information-theoretic bounds on the generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)* (2019), IEEE, pp. 582–586.

[35] JAYNES, E. T. Information theory and statistical mechanics. *Physical review 106*, 4 (1957), 620.

[36] JEFFREYS, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences 186*, 1007 (1946), 453–461.

[37] JIAO, J., HAN, Y., AND WEISSMAN, T. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)* (2017), IEEE, pp. 1475–1479.

[38] KLEIJN, B. J., VAN DER VAART, A. W., ET AL. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics 6* (2012), 354–381.

[39] KUZBORSKIJ, I., CESA-BIANCHI, N., AND SZEPESVÁRI, C. Distribution-dependent analysis of Gibbs-ERM principle. In *Conference on Learning Theory* (2019), PMLR, pp. 2028–2054.

[40] LEVER, G., LAVIOLETTE, F., AND SHAWE-TAYLOR, J. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science 473* (2013), 4–28.

[41] LOPEZ, A. T., AND JOG, V. Generalization error bounds using Wasserstein distances. In *2018 IEEE Information Theory Workshop (ITW)* (2018), IEEE, pp. 1–5.

[42] MARKOWICH, P. A., AND VILLANI, C. On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis. *Mat. Contemp 19* (2000), 1–29.

[43] MASIHA, M. S., GOHARI, A., YASSAEE, M. H., AND AREF, M. R. Learning under distribution mismatch and model misspecification. In *IEEE International Symposium on Information Theory (ISIT)* (2021).

[44] MCALLESTER, D. A. Some PAC-Bayesian theorems. *Machine Learning 37*, 3 (1999), 355–363.

[45] MCALLESTER, D. A. PAC-Bayesian stochastic model selection. *Machine Learning 51*, 1 (2003), 5–21.

[46] McSherry, F., and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* (2007), IEEE, pp. 94–103.

[47] Murphy, K. P. Conjugate Bayesian analysis of the Gaussian distribution. *def 1*, $2\sigma 2$ (2007), 16.

[48] Ohnishi, Y., and Honorio, J. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation. In *International Conference on Artificial Intelligence and Statistics* (2021), PMLR, pp. 1711–1719.

[49] Palomar, D. P., and Verdú, S. Lautum information. *IEEE transactions on information theory 54*, 3 (2008), 964–975.

[50] Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. PAC-Bayes bounds with data dependent priors. *The Journal of Machine Learning Research 13*, 1 (2012), 3507–3531.

[51] Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory* (2017), PMLR, pp. 1674–1703.

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

[53] Rivasplata, O., Kuzborskij, I., Szepesvári, C., and Shawe-Taylor, J. PAC-Bayes analysis beyond the usual bounds. *Advances in Neural Information Processing Systems* (2020).

[54] Rivasplata, O., Parrado-Hernández, E., Shawe-Taylor, J., Sun, S., and Szepesvári, C. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In *NeurIPS* (2018).

[55] Rodrigues, M. R., and Eldar, Y. C. *Information-Theoretic Methods in Data Science*. Cambridge University Press, 2021.

[56] Rodríguez-Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. Tighter expected generalization error bounds via Wasserstein distance. *arXiv preprint arXiv:2101.09315* (2021).

[57] Roth, K., Kilcher, Y., and Hofmann, T. Adversarial training is a form of data-dependent operator norm regularization. In *Advances in Neural Information Processing Systems* (2020), vol. 33, pp. 14973–14985.

[58] Russo, D., and Zou, J. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory 66*, 1 (2019), 302–323.

[59] Sason, I., and Verdú, S. f-divergence inequalities. *IEEE Transactions on Information Theory 62*, 11 (2016), 5973–6006.

[60] Shawe-Taylor, J., and Williamson, R. C. A PAC analysis of a Bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory* (1997), pp. 2–9.

[61] Shivaswamy, P. K., and Jebara, T. Maximum relative margin and data-dependent regularization. *Journal of Machine Learning Research 11*, 2 (2010).

[62] Steinke, T., and Zakynthinou, L. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory* (2020), PMLR, pp. 3437–3452.

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

[64] Van der Vaart, A. W. *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.

[65] van Erven, T. PAC-Bayes mini-tutorial: a continuous union bound. *arXiv preprint arXiv:1405.1580* (2014).

[66] Vapnik, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks 10*, 5 (1999), 988–999.

[67] Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press, 2018.

[68] Wager, S., Wang, S., and Liang, P. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems* (2013), pp. 351–359.

[69] WANG, H., DIAZ, M., SANTOS FILHO, J. C. S., AND CALMON, F. P. An information-theoretic view of generalization via Wasserstein distance. In *2019 IEEE International Symposium on Information Theory (ISIT)* (2019), IEEE, pp. 577–581.

[70] XIAO, Q.-W., ZHOU, D.-X., ET AL. Learning by nonsymmetric kernels with data dependent spaces and l^1-regularizer. *Taiwanese Journal of Mathematics 14*, 5 (2010), 1821–1836.

[71] XU, A., AND RAGINSKY, M. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems* (2017), pp. 2524–2533.

[72] XU, H., AND MANNOR, S. Robustness and generalization. *Machine learning 86*, 3 (2012), 391–423.

[73] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B., AND VINYALS, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM 64*, 3 (2021), 107–115.

[74] ZHANG, H., AND CHEN, S. X. Concentration inequalities for statistical inference. *arXiv preprint arXiv:2011.02258* (2020).

[75] ZHANG, T. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory 52*, 4 (2006), 1307–1321.

[76] ZHANG, T., ET AL. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics 34*, 5 (2006), 2180–2210.

An Exact Characterization of the Generalization Error for the Gibbs Algorithm: Supplementary Material

Gholamali Aminian[*]
University College London
g.aminian @ucl.ac.uk

Yuheng Bu[*]
Massachusetts Institute of Technology
buyuheng@mit.edu

Laura Toni
University College London
l.toni@ucl.ac.uk

Miguel Rodrigues
University College London
m.rodrigues@ucl.ac.uk

Gregory Wornell
Massachusetts Institute of Technology
gww@mit.edu

A Preliminaries

In this section, we introduce the notion of cumulant generating function, which characterizes different tail behaviors of random variables.

Definition 1. *The cumulant generating function (CGF) of a random variable X is defined as*

$$\Lambda_X(\lambda) \triangleq \log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]. \tag{27}$$

Assuming $\Lambda_X(\lambda)$ exists, it can be verified that $\Lambda_X(0) = \Lambda'_X(0) = 0$, and that it is convex.

Definition 2. *For a convex function ψ defined on the interval $[0, b)$, where $0 < b \leq \infty$, its Legendre dual $\psi^\star$ is defined as*

$$\psi^\star(x) \triangleq \sup_{\lambda\in[0,b)} \big(\lambda x - \psi(\lambda)\big). \tag{28}$$

The following lemma characterizes a useful property of the Legendre dual and its inverse function.

Lemma 1. *[15, Lemma 2.4] Assume that $\psi(0) = \psi'(0) = 0$. Then $\psi^\star(x)$ defined above is a non-negative convex and non-decreasing function on $[0,\infty)$ with $\psi^\star(0) = 0$. Moreover, its inverse function $\psi^{\star-1}(y) = \inf\{x \geq 0 : \psi^\star(x) \geq y\}$ is concave, and can be written as*

$$\psi^{\star-1}(y) = \inf_{\lambda\in[0,b)} \Big(\frac{y + \psi(\lambda)}{\lambda}\Big), \quad b > 0. \tag{29}$$

We consider the distributions with the following tail behaviors in the appendices:

- **Sub-Gaussian:** A random variable X is σ-sub-Gaussian, if $\psi(\lambda) = \frac{\sigma^2\lambda^2}{2}$ is an upper bound on $\Lambda_X(\lambda)$, for $\lambda \in \mathbb{R}$. Then by Lemma 1,

$$\psi^{\star-1}(y) = \sqrt{2\sigma^2 y}.$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

- **Sub-Exponential:** A random variable X is (σ_e^2, b)-sub-Exponential, if $\psi(\lambda) = \frac{\sigma_e^2\lambda^2}{2}$ is an upper bound on $\Lambda_X(\lambda)$, for $0 \leq |\lambda| \leq \frac{1}{b}$ and $b > 0$. Using Lemma 1, we have

$$\psi^{\star -1}(y) = \begin{cases} \sqrt{2\sigma_e^2 y}, & \textit{if } y \leq \frac{\sigma_e^2}{2b}; \\ by + \frac{\sigma_e^2}{2b}, & \textit{otherwise.} \end{cases}$$

- **Sub-Gamma:** A random variable X is $\Gamma(\sigma_s^2, c_s)$-sub-Gamma [74], if $\psi(\lambda) = \frac{\lambda^2\sigma_s^2}{2(1-c_s|\lambda|)}$ is an upper bound on $\Lambda_X(\lambda)$, for $0 < |\lambda| < \frac{1}{c_s}$ and $c_s > 0$. Using Lemma 1, we have

$$\psi^{\star -1}(y) = \sqrt{2\sigma_s^2 y} + c_s y.$$

Sub-Exponential condition is a slightly milder compared with sub-Gaussian condition. All the definition above can be generalized by considering only the left ($\lambda < 0$) or right ($\lambda > 0$) tails, e.g., σ-sub-Gaussian in the left tail as in Theorem 2.

B Generalization Error of Gibbs Algorithm

B.1 Theorem 1 Details

We start with the following two Lemmas:

Lemma 2. *We define the following $J_E(w, S)$ function as a proxy for the empirical risk, i.e., $J_E(w, S) \triangleq \frac{\alpha}{n}\sum_{i=1}^n \ell(w, Z_i) + g(w) + h(S)$, where $\alpha \in \mathbb{R}_0^+$, $g : \mathcal{W} \to \mathbb{R}$, $h : \mathcal{Z}^n \to \mathbb{R}$, and the function $J_P(w, \mu) \triangleq \mathbb{E}_{P_S}[J_E(w, S)]$ as a proxy for the population risk. Then,*

$$\mathbb{E}_{P_{W,S}}[J_P(W, \mu) - J_E(W, S)] = \alpha \cdot \overline{\textit{gen}}(P_{W|S}, P_S). \tag{30}$$

Proof.

$$\begin{aligned} &\mathbb{E}_{P_{W,S}}[J_P(W, \mu) - J_E(W, S)] \\ &= \mathbb{E}_{P_{W,S}}\Big[\mathbb{E}_{P_{Z^n}}[\frac{\alpha}{n}\sum_{i=1}^n \ell(W, Z_i)] - \frac{\alpha}{n}\sum_{i=1}^n \ell(W, Z_i)\Big] \\ &\quad + \mathbb{E}_{P_W}\Big[g(W) + \mathbb{E}_{P_S}[h(S)]\Big] - \mathbb{E}_{P_{W,S}}\Big[g(W) + h(S)\Big] \\ &= \alpha \cdot \mathbb{E}_{P_{W,S}}[L_P(W, \mu) - L_E(W, S)] \\ &= \alpha \cdot \overline{\text{gen}}(P_{W|S}, P_S). \end{aligned} \tag{31}$$

□

Lemma 3. *Consider a learning algorithm $P_{W|S}$, if we set the proxy function $J_E(w, z^n) = -\log P_{W|S}(w|s)$, then*

$$\mathbb{E}_{P_{W,S}}[J_P(W, \mu) - J_E(W, S)] = I_{\text{SKL}}(W; S). \tag{32}$$

Proof.

$$\begin{aligned} &I(W; S) + L(W; S) \\ &= \mathbb{E}_{P_{W,S}}\Big[\log \frac{P_{W|S}(W|S)}{P_W(W)}\Big] + \mathbb{E}_{P_W \otimes P_S}\Big[\log \frac{P_W(W)}{P_{W|S}(W|S)}\Big] \\ &= \mathbb{E}_{P_{W,S}}\Big[\log P_{W|S}(W|S)\Big] - \mathbb{E}_{P_W \otimes P_S}\Big[\log P_{W|S}(W|S)\Big] \\ &= \mathbb{E}_{P_{W,S}}[-\mathbb{E}_{P_S}[\log P_{W|S}(W|S)] + \log P_{W|S}(W|S)] \\ &= \mathbb{E}_{P_{W,S}}[J_P(W, \mu) - J_E(W, S)]. \end{aligned} \tag{33}$$

□

Theorem 1. ***(restated)*** *For $(\alpha, \pi(w), L_E(w, s))$-Gibbs algorithm,*

$$P_{W|S}^\alpha(w|s) = \frac{\pi(w)e^{-\alpha L_E(w,s)}}{V(s, \alpha)}, \quad \alpha > 0,$$

its expected generalization error is given by

$$\overline{\textit{gen}}(P_{W|S}^\alpha, P_S) = \frac{I_{\text{SKL}}(W; S)}{\alpha}.$$

Proof. Considering Lemma 2 and Lemma 3, we just need to verify that $J_E(w,s) = -\log P_{W|S}(w|s)$ can be decomposed into $J_E(w,s) = \frac{\alpha}{n}\sum_{i=1}^{n}\ell(w,z_i) + g(w) + h(s)$, for $\alpha > 0$. Note that

$$J_E(w,s) = -\log P^{\alpha}_{W|S}(w|s) = \alpha L_E(w,s) - \log \pi(w) + \log V(s,\alpha), \tag{34}$$

then we have:

$$\begin{aligned} I_{\mathrm{SKL}}(W;S) &= \mathbb{E}_{P_{W,S}}[J_P(W,P_S) - J_E(W,S)] \\ &= \alpha \cdot \overline{\text{gen}}(P^{\alpha}_{W|S}, P_S). \end{aligned} \tag{35}$$

□

Using Theorem 1, we can also derive the following lower bound on the expected generalization error in terms of total variation distance. As a comparison, an *upper* bound on the generalization error of a learning algorithm in terms of total variation distance is provided in [52].

Corollary 2. *For $(\alpha, \pi(w), L_E(w,s))$-Gibbs algorithm, the following lower bound on the generalization error of the Gibbs algorithm holds:*

$$\overline{gen}(P^{\alpha}_{W|S}, P_S) \geq \frac{TV^2(P_{W,S}, P_W \otimes P_S)}{\alpha}, \tag{36}$$

where

$$TV(P_{W,S}, P_W \otimes P_S) \triangleq \int\int \Big|P_{W,S}(w,s) - P_W(w)P_S(s)\Big| dwds \tag{37}$$

denotes total variation distance.

Proof. This can be proved immediately by combining Theorem 1 with the well-known Pinsker's inequality [49],

$$TV(P_{W,S}, P_W \otimes P_S) \leq \sqrt{2\min(I(W;S), L(W;S))}. \tag{38}$$

Note that the lower bound in Corollary 2 is bounded in $[0, \frac{4}{\alpha}]$. □

B.2 General Properties

In this section, we provide more discussions about other properties of the symmetrized KL divergence, including data processing inequality, variational representation, chain rule, and their implications in learning problems.

Data Processing Inequality: As shown in [59], symmetrized KL divergence is an f-divergence. Thus, the data processing inequality holds, i.e., for Markov chain $S \leftrightarrow W \leftrightarrow W'$,

$$I_{\mathrm{SKL}}(S;W) \geq I_{\mathrm{SKL}}(S;W'). \tag{39}$$

Using the data processing inequality for mutual information, [17, 71] show that pre/post-processing improves generalization, since these techniques give tighter mutual information-based generalization error bounds. However, our Theorem 1 only holds for Gibbs algorithm, which cannot characterize the generalization error for all conditional distributions $P_{W'|S}$ induced by the post-processing $P_{W'|W}$ in the Markov chain. Thus, it is hard to conclude that the pre/post-processing will reduce the exact generalization error for Gibbs algorithm by directly applying the data processing inequality.

Variational Representation: It is well-known that the mutual information has the following variational characterization

$$I(W;S) = \inf_{Q_W} D(P_{W|S}\|Q_W|P_S) = \inf_{Q_W,Q_S} D(P_{W,S}\|Q_W \otimes Q_S), \tag{40}$$

which implies that the product-of-marginal distribution minimizes the KL divergence for a given joint distribution. One may think that the counterpart for lautum information would be $\inf_{Q_W} D(P_S \otimes Q_W\|P_{W,S})$, but it is not true as shown in [49]. In general, the product-of-marginal distribution does not minimize $D(Q_W \otimes Q_S\|P_{W,S})$, and lautum information satisfies the following variational characterization

$$L(W;S) = \inf_{Q_S} D(P_W \otimes P_S\|P_{W|S} \otimes Q_S). \tag{41}$$

Thus, the product-of-marginal distribution $P_S \otimes P_W$ does not minimize the symmetrized KL divergence $D_{\mathrm{SKL}}(P_{W,S}\|Q_W \otimes Q_S)$.

Chain Rule: As shown in [17], using the chain rule of mutual information, i.e., $I(W;S) = \sum_{i=1}^{n} I(W;Z_i|Z^{i-1})$ and the fact that $I(W;Z_i|Z^{i-1}) \geq I(W;Z_i)$ for i.i.d. samples, the mutual information based generalization bound can be tightened by considering the individual sample mutual information $I(W;Z_i)$.

However, lautum information does not satisfy the same chain rule as mutual information in general, and it is hard to characterize the generalization error of Gibbs algorithm using individual terms $I_{\mathrm{SKL}}(W;Z_i)$. To see this, we have the following example to show that the joint symmetrized KL information $I_{\mathrm{SKL}}(W;S)$ can be either larger or smaller than the sum of individual terms $I_{\mathrm{SKL}}(W;Z_i)$.

Example 1. *Consider the following joint distribution for binary random variables $W, Z_1, Z_2 \in \{0,1\}$,*

$$P_{W,Z_1,Z_2}(w,z_1,z_2) = \begin{cases} \frac{1}{8}, & \textit{if } (z_1,z_2) = (0,0), \\ \frac{1}{4} - \epsilon, & \textit{if } w = 1, \textit{ and } (z_1,z_2) \neq (0,0), \\ \epsilon, & \textit{otherwise.} \end{cases} \tag{42}$$

It can be verified that Z_1 and Z_2 are mutually independent Bernoulli random variable with $p = \frac{1}{2}$, and the conditional distribution is symmetric in the sense that $P_{W|Z_1,Z_2}(w|0,1) = P_{W|Z_1,Z_2}(w|1,0)$.

Case I: *When $\epsilon = 0.0001$, we can compute the mutual information as*

$$I(W;Z_1) = I(W;Z_2) = 0.0943, \quad I(W;Z_1,Z_2) = 0.2014,$$

which satisfies the bound $I(W;Z_1,Z_2) \geq I(W;Z_1) + I(W;Z_2)$ when $Z_1 \perp Z_2$. However, for lautum information

$$L(W;Z_1) = L(W;Z_2) = 0.3257, \quad L(W;Z_1,Z_2) = 0.5315,$$

$L(W;Z_1) + L(W;Z_2) > L(W;Z_1,Z_2)$, and

$$\begin{aligned} &I_{\mathrm{SKL}}(W;Z_1) = I_{\mathrm{SKL}}(W;Z_2) = 0.4200, \quad I_{\mathrm{SKL}}(W;Z_1,Z_2) = 0.7329, \\ &I_{\mathrm{SKL}}(W;Z_1) + I_{\mathrm{SKL}}(W;Z_2) > I_{\mathrm{SKL}}(W;Z_1,Z_2). \end{aligned}$$

Case II: *When $\epsilon = 0.01$, it can be verified that*

$$\begin{aligned} &I_{\mathrm{SKL}}(W;Z_1) = I_{\mathrm{SKL}}(W;Z_2) = 0.1255, \quad I_{\mathrm{SKL}}(W;Z_1,Z_2) = 0.2741, \\ &I_{\mathrm{SKL}}(W;Z_1) + I_{\mathrm{SKL}}(W;Z_2) < I_{\mathrm{SKL}}(W;Z_1,Z_2). \end{aligned}$$

Thus, individual sample symmetrized KL information cannot be used to characterize the behavior of $I_{\mathrm{SKL}}(W;S)$ in general.

B.3 Example Details: Mean Estimation

B.3.1 Generalization Error

We first evaluate the generalization error of the learning algorithm in (13) directly. Note that the output W can be written as

$$W = \frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{\sigma_1^2}{\sigma^2}\sum_{i=1}^{n} Z_i + N, \quad \text{with} \quad \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{43}$$

where $N \sim \mathcal{N}(0, \sigma_1^2 I_d)$ is independent from the training samples $S = \{Z_i\}_{i=1}^n$. Thus,

$$
\begin{aligned}
&\overline{\text{gen}}(P_{W|S}, P_S)\\
&= \mathbb{E}_{P_{W,S}}[L_P(W,\mu) - L_E(W,S)]\\
&= \mathbb{E}_{P_{W,S}}\Big[\mathbb{E}_{P_{\widetilde{Z}}}[\|W - \widetilde{Z}\|_2^2] - \frac{1}{n}\sum_{i=1}^n \|W - Z_i\|_2^2\Big]\\
&\overset{(a)}{=} \mathbb{E}_{P_{W,Z_i}\otimes P_{\widetilde{Z}}}\Big[(2W - \widetilde{Z} - Z_i)^\top (Z_i - \widetilde{Z})\Big]\\
&= \mathbb{E}\Big[2(\frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{\sigma_1^2}{\sigma^2}\sum_{i=1}^n Z_i + N)^\top (Z_i - \widetilde{Z}) - (Z_i + \widetilde{Z})^\top (Z_i - \widetilde{Z})\Big]\\
&\overset{(b)}{=} \frac{2\sigma_1^2}{\sigma^2}\mathbb{E}\Big[Z_i^\top (Z_i - \widetilde{Z})\Big]\\
&= \frac{2d\sigma_1^2\sigma_Z^2}{\sigma^2} = \frac{2d\sigma_0^2\sigma_Z^2}{n\sigma_0^2 + \sigma^2},
\end{aligned}
\tag{44}
$$

where $\widetilde{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_Z^2 I_d)$ denotes an independent copy of the training sample, (a) follows due to the fact that Z^n are i.i.d, and (b) follows from the fact that $Z_i - \widetilde{Z}$ has zero mean, and it is only correlated with Z_i.

B.3.2 Symmetrized KL Divergence

The following lemma from [49] characterizes the mutual and lautum information for the Gaussian channel.

Lemma 4. *[49, Theorem 14] Consider the following model*

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{N}_{\mathrm{G}}, \tag{45}$$

where $\boldsymbol{X} \in \mathbb{R}^{d_X}$ denotes the input random vector with zero mean (not necessarily Gaussian), $\boldsymbol{A} \in \mathbb{R}^{d_Y \times d_X}$ denotes the linear transformation undergone by the input, $\boldsymbol{Y} \in \mathbb{R}^{d_Y}$ is the output vector, and $\boldsymbol{N}_{\mathrm{G}} \in \mathbb{R}^{d_Y}$ is a Gaussian noise vector independent of $\boldsymbol{X}$. The input and the noise covariance matrices are given by $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{N_{\mathrm{G}}}$. Then, we have

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}_{N_{\mathrm{G}}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big) - D\big(P_{\boldsymbol{Y}}\|P_{N_{\mathrm{G}}}\big), \tag{46}$$

$$L(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}_{N_{\mathrm{G}}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big) + D\big(P_{\boldsymbol{Y}}\|P_{N_{\mathrm{G}}}\big). \tag{47}$$

In our example, the output W can be written as

$$W = \frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{\sigma_1^2}{\sigma^2}\sum_{i=1}^n Z_i + N = \frac{\sigma_1^2}{\sigma^2}\sum_{i=1}^n (Z_i - \boldsymbol{\mu}) + \frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{n\sigma_1^2}{\sigma^2}\boldsymbol{\mu} + N, \tag{48}$$

where $N \sim \mathcal{N}(0, \sigma_1^2 I_d)$. Setting $P_{N_{\mathrm{G}}} \sim \mathcal{N}(\frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{n\sigma_1^2}{\sigma^2}\boldsymbol{\mu}, \sigma_1^2 I_d)$ and $\boldsymbol{\Sigma} = \sigma_Z^2 I_{nd}$ in Lemma 4 gives

$$\mathrm{tr}\big(\boldsymbol{\Sigma}_{N_{\mathrm{G}}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big) = \mathrm{tr}\big(\frac{\sigma_Z^2}{\sigma_1^2}\boldsymbol{A}\boldsymbol{A}^\top\big), \tag{49}$$

and noticing that $\boldsymbol{A}\boldsymbol{A}^\top = \frac{n\sigma_1^4}{\sigma^4} I_d$ completes the proof.

B.4 ISMI Bound

In this subsection, we evaluate the following individual sample mutual information (ISMI) bound from [19, Theorem 2] for the example discussed in Section 2.2 with i.i.d. samples generated from Gaussian distribution $P_Z \sim \mathcal{N}(\boldsymbol{\mu}, \sigma_Z^2 I_d)$.

Lemma 5. *[19, Theorem 2] Suppose $\ell(\widetilde{W},\widetilde{Z})$ satisfies $\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \leq \psi_+(\lambda)$ for $\lambda \in [0, b_+)$, and $\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \leq \psi_-(-\lambda)$ for $\lambda \in (b_-, 0]$ under $P_{\widetilde{Z},\widetilde{W}} = P_Z \otimes P_W$, where $0 < b_+ \leq \infty$ and $-\infty \leq b_- < 0$. Then,*

$$\mathrm{gen}(P_{W|S}, P_S) \leq \frac{1}{n}\sum_{i=1}^{n} \psi_-^{*-1}\big(I(W;Z_i)\big), \tag{50}$$

$$-\mathrm{gen}(P_{W|S}, P_S) \leq \frac{1}{n}\sum_{i=1}^{n} \psi_+^{*-1}\big(I(W;Z_i)\big). \tag{51}$$

We need to compute the mutual information between each individual sample and the output hypothesis $I(W;Z_i)$, and the CGF of $\ell(\widetilde{W},\widetilde{Z})$, where $\widetilde{W}$, $\widetilde{Z}$ are independent copies of W and Z with the same marginal distribution, respectively.

Since W and Z_i are Gaussian, $I(W;Z_i)$ can be computed exactly using covariance matrix:

$$\mathrm{Cov}[Z_i, W] = \begin{pmatrix} \sigma_Z^2 I_d & \frac{\sigma_1^2}{\sigma^2}\sigma_Z^2 I_d \\ \frac{\sigma_1^2}{\sigma^2}\sigma_Z^2 I_d & (\frac{n\sigma_1^4}{\sigma^4}\sigma_Z^2 + \sigma_1^2) I_d \end{pmatrix}, \tag{52}$$

then, we have

$$\begin{aligned} I(W;Z_i) &= \frac{d}{2}\log\frac{\frac{n\sigma_1^4}{\sigma^4}\sigma_Z^2 + \sigma_1^2}{\frac{(n-1)\sigma_1^4}{\sigma^4}\sigma_Z^2 + \sigma_1^2} \\ &= \frac{d}{2}\log\Big(1 + \frac{\sigma_1^2\sigma_Z^2}{(n-1)\sigma_1^2\sigma_Z^2 + \sigma^4}\Big) \\ &= \frac{d}{2}\log\Big(1 + \frac{\sigma_0^2\sigma_Z^2}{(n-1)\sigma_0^2\sigma_Z^2 + n\sigma_0^2\sigma^2 + \sigma^4}\Big), \end{aligned} \tag{53}$$

for $i = 1, \cdots, n$, $n \geq 2$. In addition, since

$$W \sim \mathcal{N}\Big(\frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{n\sigma_1^2}{\sigma^2}\boldsymbol{\mu}, \big(\frac{n\sigma_1^4}{\sigma^4}\sigma_Z^2 + \sigma_1^2\big)I_d\Big), \tag{54}$$

it can be shown that $\ell(\widetilde{W},\widetilde{Z}) = \|\widetilde{Z} - \widetilde{W}\|^2$ is a scaled non-central chi-square distribution with d degrees of freedom, where the scaling factor $\sigma_\ell^2 \triangleq (\frac{n\sigma_1^4}{\sigma^4} + 1)\sigma_Z^2 + \sigma_1^2$ and its non-centrality parameter $\eta \triangleq \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}\|_2^2$.

Note that the expectation of chi-square distribution with non-centrality parameter η and d degrees of freedom is $d + \eta$ and its moment generating function is $\exp(\frac{\eta\lambda}{1-2\lambda})(1-2\lambda)^{-d/2}$. Therefore, the CGF of $\ell(\widetilde{W},\widetilde{Z})$ is given by

$$\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) = -(d\sigma_\ell^2 + \eta)\lambda + \frac{\eta\lambda}{1 - 2\sigma_\ell^2\lambda} - \frac{d}{2}\log(1 - 2\sigma_\ell^2\lambda), \tag{55}$$

for $\lambda \in (-\infty, \frac{1}{2\sigma_\ell^2})$. Since $\mathrm{gen}(P_{W|S}, P_Z) \geq 0$, we only need to consider the case $\lambda < 0$. It can be shown that:

$$\begin{aligned} \Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) &= -d\sigma_\ell^2\lambda - \frac{d}{2}\log(1 - 2\sigma_\ell^2\lambda) + \frac{2\sigma_\ell^2\eta\lambda^2}{1 - 2\sigma_\ell^2\lambda} \\ &= \frac{d}{2}(-u - \log(1-u)) + \frac{2\sigma_\ell^2\eta\lambda^2}{1 - 2\sigma_\ell^2\lambda}, \end{aligned} \tag{56}$$

where $u \triangleq 2\sigma_\ell^2\lambda$. Further note that

$$-u - \log(1-u) \leq \frac{u^2}{2},\ u < 0, \tag{57}$$

$$\frac{2\sigma_\ell^2\eta\lambda^2}{1 - 2\sigma_\ell^2\lambda} \leq 2\sigma_\ell^2\eta\lambda^2,\ \lambda < 0. \tag{58}$$

We have the following upper bound on the CGF of $\ell(\widetilde{W},\widetilde{Z})$:

$$\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \leq (d\sigma_\ell^4 + 2\sigma_\ell^2\eta)\lambda^2, \quad \lambda < 0, \tag{59}$$

which means that $\ell(\widetilde{W},\widetilde{Z})$ is $\sqrt{d\sigma_\ell^4 + 2\sigma_\ell^2\eta}$-sub-Gaussian for $\lambda < 0$. Combining the results in (53), Lemma 5 gives the following bound

$$\overline{\text{gen}}(P_{W|S}, P_S) \leq \sqrt{\frac{d^2\sigma_\ell^4 + 2d\sigma_\ell^2\eta}{2}\log(1 + \frac{\sigma_0^2\sigma_Z^2}{(n-1)\sigma_0^2\sigma_Z^2 + n\sigma_0^2\sigma^2 + \sigma^4})}. \tag{60}$$

If $\sigma^2 = \frac{n}{2\alpha}$ is a constant, i.e., $\alpha = \mathcal{O}(n)$, then as $n \to \infty$, $\sigma_1^2 = \mathcal{O}\big(\frac{1}{n}\big)$ and $\sigma_\ell^2 = \mathcal{O}(1)$, and the above bound is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

C Expected Generalization Error Upper Bound

C.1 Proof of Theorem 2

We prove a slightly more general form of Theorem 2 as follows:

Theorem 4. *Suppose that the training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z and the loss function $\ell(w, Z)$ satisfies $\Lambda_{\ell(w,Z)}(\lambda) \leq \psi(-\lambda)$, for $\lambda \in (-b, 0)$ and $0 < b$ under data-generating distribution P_Z for all $w \in \mathcal{W}$. Let us assume $\exists\, C_E \in \mathbb{R}_0^+$ such that $\frac{L(W;S)}{I(W;S)} \geq C_E$, and we further assume:*

$$\exists\, 0 < \kappa < \infty, \quad s.t. \quad \psi^{\star -1}(\frac{\kappa}{n}) - \frac{(1 + C_E)\kappa}{\alpha} = 0. \tag{61}$$

Then, the following upper bound holds for the expected generalization error of $(\alpha, \pi(w), L_E(w, s))$-Gibbs algorithm:

$$0 \leq \overline{gen}(P_{W|S}^\alpha, P_S) \leq \frac{(1 + C_E)\kappa}{\alpha}. \tag{62}$$

Proof. It is shown in [19, Proposition 2] that the following generalization error bound holds,

$$\overline{\text{gen}}(P_{W|S}^\alpha, P_S) \leq \psi^{\star -1}\Big(\frac{I(W;S)}{n}\Big). \tag{63}$$

By Theorem 1 and the assumption on C_E, we have

$$\overline{\text{gen}}(P_{W|S}^\alpha, P_S) = \frac{I(W;S) + L(W;S)}{\alpha} \geq \frac{(1 + C_E)I(W;S)}{\alpha}. \tag{64}$$

Therefore,

$$\frac{(1 + C_E)I(W;S)}{\alpha} \leq \psi^{\star -1}\Big(\frac{I(W;S)}{n}\Big). \tag{65}$$

Consider the function $F(u) \triangleq \psi^{\star -1}(\frac{u}{n}) - \frac{(1+C_E)u}{\alpha}$, which is concave and satisfies $F(0) = 0$ by Lemma 1. If there exists $0 < \kappa < \infty$, such that $F(\kappa) = 0$, then $F(I(W;S)) \geq 0$ implies that

$$0 \leq I(W;S) \leq \kappa.$$

Since $\psi^{\star -1}(\cdot)$ is non-decreasing, we have

$$\overline{\text{gen}}(P_{W|S}^\alpha, P_S) \leq \psi^{\star -1}\Big(\frac{\kappa}{n}\Big) = \frac{(1 + C_E)\kappa}{\alpha}. \qquad \square$$

In the following, we specify the different forms of $\psi(\lambda)$ function in Theorem 4 to capture different tail behaviors of the loss function. We first consider the σ-sub-Gaussian assumption.

Theorem 2. ***(restated)*** *Suppose that the training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z, and the non-negative loss function $\ell(w, Z)$ is σ-sub-Gaussian on the left-tail under distribution P_Z for all $w \in \mathcal{W}$. We further assume $C_E \leq \frac{L(W;S)}{I(W;S)}$ for some $C_E \geq 0$. Then, for the $(\alpha, \pi(w), L_E(w, s))$-Gibbs algorithm, we have*

$$0 \leq \overline{gen}(P_{W|S}^\alpha, P_S) \leq \frac{2\sigma^2\alpha}{(1 + C_E)n}.$$

Proof. If the loss function is σ-sub-Gaussian on the left-tail we have $\psi^{\star-1}(y) = \sqrt{2\sigma^2 y}$. Using Theorem 4 we have

$$\sqrt{2\sigma^2\frac{\kappa}{n}} - \frac{(1+C_E)\kappa}{\alpha} = 0, \tag{66}$$

and the solution is $\kappa = \frac{2\sigma^2}{n}\frac{\alpha^2}{(1+C_E)^2}$. Therefore,

$$\overline{\text{gen}}(P_{W|S}^{\alpha}, P_S) \leq \frac{(1+C_E)\kappa}{\alpha} = \frac{2\sigma^2\alpha}{n(1+C_E)}. \qquad \square$$

C.2 Other Tail Distributions

In this section, we consider the sub-Exponential and sub-Gamma assumptions for the loss function and it is shown that the rates of convergence in these two cases are the same as that of the sub-Gaussian assumption, i.e., $\mathcal{O}(1/n)$.

We first consider the sub-Exponential case.

Corollary 3. *Suppose that the training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z, and the non-negative loss function $\ell(w, Z)$ is (σ_e^2, b)-sub-Exponential on the left-tail* [*] *under distribution P_Z for all $w \in \mathcal{W}$. We further assume $C_E \leq \frac{L(W;S)}{I(W;S)}$ for some $C_E \geq 0$. Then, for the $(\alpha, \pi(w), L_E(w,s))$-Gibbs algorithm, we have*

$$\overline{gen}(P_{W|S}^{\alpha}, P_S) \leq \begin{cases} \frac{2\sigma_e^2\alpha}{n(1+C_E)}, & \text{if } n \geq \frac{2bI(W;S)}{\sigma_e^2}; \\ \frac{\sigma_e^2}{2b}\Big(\frac{\alpha b}{(n(1+C_E)-\alpha b)} + 1\Big), & \text{if } \lceil \frac{\alpha b}{1+C_E} \rceil < n < \frac{2bI(W;S)}{\sigma_e^2}. \end{cases} \tag{67}$$

Proof. If the loss function is sub-Exponential on the left-tail we have

$$\psi^{\star-1}(y) = \begin{cases} \sqrt{2\sigma_e^2 y}, & \text{if } y \leq \frac{\sigma_e^2}{2b}; \\ by + \frac{\sigma_e^2}{2b}, & \text{otherwise.} \end{cases}$$

If $\frac{I(W;S)}{n} \leq \frac{\sigma_e^2}{2b}$, by Theorem 4, we have

$$\frac{(1+C_E)I(W;S)}{\alpha} \leq \sqrt{2\sigma_e^2\frac{I(W;S)}{n}}, \tag{68}$$

then the following upper bound holds,

$$I(W;S) \leq \frac{2\sigma_e^2\alpha^2}{(1+C_E)^2 n}, \tag{69}$$

which gives

$$\overline{\text{gen}}(P_{W|S}^{\alpha}, P_S) \leq \frac{2\sigma_e^2\alpha}{n(1+C_E)}. \tag{70}$$

If $\frac{I(W;S)}{n} > \frac{\sigma_e^2}{2b}$, we have

$$\frac{I(W;S)(1+C_E)}{\alpha} \leq \frac{bI(W;S)}{n} + \frac{\sigma_e^2}{2b}, \tag{71}$$

then the following upper bound holds when $n > \frac{\alpha b}{1+C_E}$,

$$I(W;S) \leq \frac{\alpha n \sigma_e^2}{2b(n(1+C_E) - \alpha b)}, \tag{72}$$

which gives

$$\overline{\text{gen}}(P_{W|S}^{\alpha}, P_S) \leq \frac{\sigma_e^2}{2b}\Big(\frac{\alpha b}{(n(1+C_E)-\alpha b)} + 1\Big). \qquad \square$$

[*] A random variable X is (σ_e^2, b)-sub-Exponential on the left-tail if $\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \frac{\sigma_e^2\lambda^2}{2}$, $-\frac{1}{b} \leq \lambda \leq 0$.

Note that all the sub-Exponential loss functions are also sub-Exponential on the left-tail under the same distribution (the converse statement is not true).

The authors in [48, 58] also consider the sub-Exponential assumption for general learning algorithms and provide PAC-Bayesian upper bounds. The result in Corollary 3 is an upper bound on the expected generalization error for Gibbs algorithm under sub-Exponential assumption, which establishes the $\mathcal{O}(1/n)$ convergence rate.

Next, we provide an upper bound under sub-Gamma assumption.

Corollary 4. *Suppose that the training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z, and the non-negative loss function $\ell(w, Z)$ is $\Gamma(\sigma_s^2, c_s)$-sub-Gamma on the left-tail*[*] *under distribution P_Z for all $w \in \mathcal{W}$. We further assume $C_E \le \frac{L(W;S)}{I(W;S)}$ for some $C_E \ge 0$. Then, for the $(\alpha, \pi(w), L_E(w, s))$-Gibbs algorithm, if $n > \frac{c_s \alpha}{(1+C_E)}$, we have*

$$\overline{\text{gen}}(P^{\alpha}_{W|S}, P_S) \le \frac{2\sigma_s^2 \alpha}{(1+C_E)n - \alpha c_s}\Big(1 + \frac{\alpha c_s}{(1+C_E)n - \alpha c_s}\Big). \tag{73}$$

Proof. By considering $\psi^{\star -1}(y) = \sqrt{2\sigma_s^2 y} + cy$ in Theorem 4, we have

$$\frac{(1+C_E)I(W;S)}{\alpha} \le \sqrt{2\sigma_s^2 \frac{I(W;S)}{n}} + c_s \frac{I(W;S)}{n}. \tag{74}$$

Then the following upper bound holds when $n > \frac{c_s \alpha}{(1+C_E)}$,

$$I(W;S) \le \Big(\frac{\alpha}{(1+C_E)n - \alpha c_s}\Big)^2 2n\sigma_s^2, \tag{75}$$

which gives

$$\overline{\text{gen}}(P^{\alpha}_{W|S}, P_S) \le \frac{2\sigma_s^2 \alpha (1+C_E)n}{\big((1+C_E)n - \alpha c_s\big)^2}. \qquad \square$$

The sub-Gamma assumption is also considered in [1, 26] and PAC-Bayesian upper bounds are provided. Our Corollary 4 provides an upper bound on the expected generalization error for Gibbs algorithm under sub-Gamma assumption, which establishes the $\mathcal{O}(1/n)$ convergence rate.

D PAC-Bayesian Upper Bound

Since the $(\alpha, \pi(w), L_P(w, P_{S'}))$-Gibbs distribution only depends on the population risk $L_P(w, P_{S'})$ and is independent of the samples S, we can denote it as P_W^{α, L'_P}. The following lemma provides an operational interpretation of the symmetrized KL divergence between the Gibbs posterior $P^{\alpha}_{W|S}$ and the prior distribution P_W^{α, L'_P}.

Lemma 6. *Let us denote the $(\alpha, \pi(w), L_E(w, s))$-Gibbs algorithm as $P^{\alpha}_{W|S}$ and the $(\alpha, \pi(w), L_P(w, P_{S'}))$-Gibbs algorithm as P_W^{α, L'_P}. Then, the following equality holds for these two Gibbs distributions with the same inverse temperature and prior distribution*

$$\mathbb{E}_{\Delta(P^{\alpha}_{W|S=s}, P_W^{\alpha, L'_P})}[L_P(W, P_{S'}) - L_E(W, s)] = \frac{D_{\text{SKL}}(P^{\alpha}_{W|S=s} \| P_W^{\alpha, L'_P})}{\alpha}, \tag{76}$$

where $\mathbb{E}_{\Delta(P^{\alpha}_{W|S=s}, P_W^{\alpha, L'_P})}[f(W)] = \mathbb{E}_{P^{\alpha}_{W|S=s}}[f(W)] - \mathbb{E}_{P_W^{\alpha, L'_P}}[f(W)]$.

[*] A random variable X is $\Gamma(\sigma_s^2, c_s)$-sub-Gamma on the left-tail if $\log \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \le \frac{\lambda^2 \sigma_s^2}{2(1 - c|\lambda|)}$, for $-\frac{1}{c_s} < \lambda < 0$.

Proof.

$$
\begin{aligned}
&D_{\mathrm{SKL}}(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P})\\
&=\mathbb{E}_{P^{\alpha}_{W|S=s}}\left[\log\frac{P^{\alpha}_{W|S=s}}{P_W^{\alpha,L'_P}}\right]-\mathbb{E}_{P_W^{\alpha,L'_P}}\left[\log\frac{P^{\alpha}_{W|S=s}}{P_W^{\alpha,L'_P}}\right]\\
&\overset{(a)}{=}\mathbb{E}_{\Delta(P^{\alpha}_{W|S=s},P_W^{\alpha,L'_P})}\big[\log(e^{-\alpha(L_E(W,s)-L_P(W,P_{S'}))})\big]\\
&=\alpha\,\mathbb{E}_{\Delta(P^{\alpha}_{W|S=s},P_W^{\alpha,L'_P})}\big[L_P(W,P_{S'})-L_E(W,s)\big],
\end{aligned}
\tag{77}
$$

where (a) follows by the fact that partition functions $V(s,\alpha)$ do not depend on W. □

Theorem 3. ***(restated)*** *Suppose that the training samples $S=\{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z, and the non-negative loss function $\ell(w,Z)$ is σ-sub-Gaussian under data-generating distribution P_Z for all $w\in\mathcal{W}$. If we use the $(\alpha,\pi(w),L_P(w,P_{S'}))$-Gibbs distribution as the PAC-Bayesian prior, where $P_{S'}$ is an arbitrary chosen (and known) distribution, the following upper bound holds for the generalization error of $(\alpha,\pi(w),L_E(w,s))$-Gibbs algorithm with probability at least $1-2\delta$, $0<\delta<\frac{1}{2}$ under distribution P_S,*

$$
\begin{aligned}
\Big|\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,P_S)-L_E(W,s)]\Big|\leq&\frac{2\sigma^2\alpha}{(1+C_P(s))n}+\epsilon^2\\
&+2\sqrt{\frac{\sigma^2\alpha}{(1+C_P(s))n}}\left(\sqrt[4]{2\sigma^2D(P_{Z'}\|P_Z)}+\epsilon\right),
\end{aligned}
$$

where $\epsilon\triangleq\sqrt[4]{\frac{2\sigma^2\log(1/\delta)}{n}}$, and $C_P(s)\leq\frac{D\left(P_W^{\alpha,L'_P}\middle\|P^{\alpha}_{W|S=s}\right)}{D\left(P^{\alpha}_{W|S=s}\middle\|P_W^{\alpha,L'_P}\right)}$ for some $C_P(s)\geq 0$.

Proof. Using Lemma 6, we have

$$
\begin{aligned}
D_{\mathrm{SKL}}(P^{\alpha}_{W|S}\|P_W^{\alpha,L'_P})=&\,\alpha(\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,P_{Z'})]-\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_E(W,s)])\\
&-\alpha(\mathbb{E}_{P_W^{\alpha,L'_P}}[L_P(W,P_{Z'})]-\mathbb{E}_{P_W^{\alpha,L'_P}}[L_E(W,s)])\\
\leq&\,\alpha\Big|\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,P_{Z'})]-\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_E(W,s)]\Big|\\
&+\alpha\Big|(\mathbb{E}_{P_W^{\alpha,L'_P}}[L_P(W,P_{Z'})]-\mathbb{E}_{P_W^{\alpha,L'_P}}[L_E(W,s)]\Big|\\
\leq&\,\alpha\Big|\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,P_{Z'})]-\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,P_Z)]\Big|\\
&+\alpha\Big|\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,P_Z)]-\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_E(W,s)]\Big|\\
&+\alpha\Big|\mathbb{E}_{P_W^{\alpha,L'_P}}[L_P(W,P_{Z'})]-\mathbb{E}_{P_W^{\alpha,L'_P}}[L_P(W,P_Z)]\Big|\\
&+\alpha\Big|\mathbb{E}_{P_W^{\alpha,L'_P}}[L_P(W,P_Z)]-\mathbb{E}_{P_W^{\alpha,L'_P}}[L_E(W,s)]\Big|,
\end{aligned}
\tag{78}
$$

and we just need to bound the four terms in the above inequality.

The first and the third term in (78) can be bounded using the Donsker-Varadhan variational characterization of KL divergence, note that for all $\lambda\in\mathbb{R}$,

$$
\begin{aligned}
D(P_{Z'}\|P_Z)&\geq\mathbb{E}_{P_{Z'}}[\lambda\ell(w,Z')]-\log\mathbb{E}_{P_Z}[e^{\lambda\ell(w,Z)}]\\
&\geq\lambda(L_P(w,P_{Z'})-L_P(w,P_Z))-\frac{\lambda^2\sigma^2}{2},
\end{aligned}
\tag{79}
$$

where the last step follows from the sub-Gaussian assumption. Since the above inequality holds for all $\lambda\in\mathbb{R}$, the discriminant must be non-positive, which implies

$$
|L_P(w,P_{Z'})-L_P(w,P_Z)|\leq\sqrt{2\sigma^2D(P_{Z'}\|P_Z)},\quad\text{for all}\quad w\in\mathcal{W}.
\tag{80}
$$

We use the PAC-Bayesian bound in [29, Proposition 3] to bound the second and the fourth term in (78). For any posterior distribution $Q_{W|S=s}$, and prior distribution Q_W, if $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$, the following bound holds with probability $1-\delta$,

$$\left|\mathbb{E}_{Q_{W|S=s}}[L_P(W,P_Z)] - \mathbb{E}_{Q_{W|S=s}}[L_E(W,s)]\right| \leq \sqrt{\frac{2\sigma^2\big(D(Q_{W|S=s}\|Q_W) + \log(1/\delta)\big)}{n}}. \tag{81}$$

If we choose $P^{\alpha}_{W|S}$ as the posterior distribution and P_W^{α,L'_P} as the prior distribution, we have

$$\left|\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,P_Z)] - \mathbb{E}_{P^{\alpha}_{W|S=s}}[L_E(W,s)]\right| \leq \sqrt{\frac{2\sigma^2\left(D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) + \log(1/\delta)\right)}{n}} \tag{82}$$

holds with probability $1-\delta$. If we set $Q_{W|S=s} = Q_W = P_W^{\alpha,L'_P}$, we have

$$\left|\mathbb{E}_{P_W^{\alpha,L'_P}}[L_P(W,P_Z)] - \mathbb{E}_{P_W^{\alpha,L'_P}}[L_E(W,s)]\right| \leq \sqrt{\frac{2\sigma^2\left(\log(1/\delta)\right)}{n}}. \tag{83}$$

Combining the bounds in (80), (82) and (83) with (78), we have

$$\begin{aligned} D_{\mathrm{SKL}}(P^{\alpha}_{W|S}\|P_W^{\alpha,L'_P}) \leq& \alpha\sqrt{\frac{2\sigma^2\left(D(P^{\alpha}_{W|S=s}\|P_{W|S}^{\alpha,L'_P}) + \log(1/\delta)\right)}{n}} \\ &+ \alpha\sqrt{\frac{2\sigma^2\left(\log(1/\delta)\right)}{n}} + 2\alpha\sqrt{2\sigma^2 D(P_{Z'}\|P_Z)}. \end{aligned} \tag{84}$$

Then, using the assumption that $(1+C_P(s))D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) \leq D_{\mathrm{SKL}}(P^{\alpha}_{W|S}\|P_{W|S}^{\alpha,L_P})$, we have

$$\begin{aligned} (1+C_P(s))D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) \leq \alpha&\sqrt{\frac{2\sigma^2\left(D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) + \log(1/\delta)\right)}{n}} \\ &+ \alpha\sqrt{\frac{2\sigma^2\left(\log(1/\delta)\right)}{n}} + 2\alpha\sqrt{2\sigma^2 D(P_{Z'}\|P_Z)}. \end{aligned} \tag{85}$$

Denote $\alpha' \triangleq \frac{\alpha}{(1+C_P(s))}$, then we have

$$\begin{aligned} D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) - \sqrt{\frac{2\alpha'^2\sigma^2\left(\log(1/\delta)\right)}{n}} - \sqrt{8\alpha'^2\sigma^2 D(P_{Z'}\|P_Z)} \\ \leq \sqrt{\frac{2\alpha'^2\sigma^2\left(D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) + \log(1/\delta)\right)}{n}}. \end{aligned} \tag{86}$$

If we have $0 \leq D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) \leq \sqrt{\frac{2\alpha'^2\sigma^2(\log(1/\delta))}{n}} + \sqrt{8\alpha'^2\sigma^2 D(P_{Z'}\|P_Z)}$, then the above inequality holds. Otherwise, we could take square over both sides in (86), and denote

$$A \triangleq C + \sqrt{\frac{2\sigma^2\alpha'^2\log(1/\delta)}{n}}, \quad B \triangleq \sqrt{8\alpha'^2\sigma^2 D(P_{Z'}\|P_Z)},$$

where $C \triangleq \frac{\sigma^2\alpha'^2}{n}$, then we have

$$D^2(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) - 2D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P})(A+B) + B^2 + 2(A-C)B \leq 0. \tag{87}$$

Solving the above inequality gives:

$$0 \leq D(P^{\alpha}_{W|S=s}\|P_W^{\alpha,L'_P}) \leq \sqrt{A^2+2BC} + A + B. \tag{88}$$

As $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for positive x, y and $A \ge C$, we have

$$D(P^{\alpha}_{W|S=s} \| P^{\alpha, L'_P}_{W}) \le 2A + B + \sqrt{2BC} \le 2A + B + \sqrt{2AB} \le (\sqrt{2A} + \sqrt{B})^2. \tag{89}$$

Now using (89) in (82) and applying the inequality $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, we have:

$$\begin{aligned}
&\left|\mathbb{E}_{P^{\alpha}_{W|S=s}}[L_P(W,\mu) - L_E(W,s)]\right| \\
&\le \sqrt{\frac{2\sigma^2(\sqrt{2A}+\sqrt{B})^2 + 2\sigma^2\log(1/\delta)}{n}} \\
&\le \sqrt{\frac{4\sigma^2 A}{n}} + \sqrt{\frac{2\sigma^2 B}{n}} + \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}} \\
&\le \frac{2\alpha\sigma^2}{(1+C_P(s))n} + \sqrt{\frac{2\sigma^2\left(\log(1/\delta)\right)}{n}} \\
&\quad + 2\sqrt{\frac{\alpha\sigma^2}{(1+C_P(s))n}}\left(\sqrt[4]{\frac{2\sigma^2\log(1/\delta)}{n}} + \sqrt[4]{2\sigma^2 D(P_{Z'}\|P_Z)}\right).
\end{aligned}$$

As both (82) and (83) hold with probability at least $1-\delta$, the above inequality holds with probability at least $1-2\delta$ by the union bound [67]. □

E Asymptotic Behavior of Generalization Error for Gibbs Algorithm

E.1 Large Inverse Temperature Details

Proposition 1. ***(restated)*** *In the single-well case, if the Hessian matrix $H^*(S)$ is not singular, then the generalization error of the $(\infty, \pi(\boldsymbol{w}), L_E(\boldsymbol{w}, s))$-Gibbs algorithm is*

$$\begin{aligned}
\overline{gen}(P^{\infty}_{W|S}, P_S) &= \mathbb{E}_{\Delta_{W,S}}\big[\frac{1}{2}W^\top H^*(S) W\big] \\
&\quad + \mathbb{E}_{P_S}\Big[(W^*(S) - \mathbb{E}[W^*(S)])^\top (H^*(S)W^*(S) - \mathbb{E}[H^*(S)W^*(S)])\Big],
\end{aligned}$$

where $\mathbb{E}_{\Delta_{W,S}}[f(W,S)] \triangleq \mathbb{E}_{P_W \otimes P_S}[f(W,S)] - \mathbb{E}_{P_{W,S}}[f(W,S)]$.

Proof. It is shown in [12, 33] that if the following Hessian matrix

$$H^*(S) = \nabla^2_w L_E(w,S)\big|_{w=W^*(S)} \tag{90}$$

is not singular, then as $\alpha \to \infty$

$$P^{\alpha}_{W|S} \to \mathcal{N}(W^*(S), \frac{1}{\alpha}H^*(S)^{-1}) \tag{91}$$

in distribution. Then, the mean of the marginal distribution P_W equals to the mean of $W^*(S)$, i.e.,

$$\mathbb{E}_{P_W}[W] = \mathbb{E}_{P_S}[W^*(S)]. \tag{92}$$

To apply Theorem 1, we evaluate the symmetrized KL information using the Gaussian approximation:

$$\begin{aligned}
&I(W;S) + L(W;S) \\
&= \mathbb{E}_{P_{W,S}}[\log P^{\alpha}_{W|S}] - \mathbb{E}_{P_W \otimes P_S}[\log P^{\alpha}_{W|S}] \\
&= \mathbb{E}_{P_{W,S}}\big[-\frac{\alpha}{2}(W - W^*(S))^\top H^*(S)(W - W^*(S))\big] \\
&\quad + \mathbb{E}_{P_W \otimes P_S}\big[\frac{\alpha}{2}(W - W^*(S))^\top H^*(S)(W - W^*(S))\big] \\
&= \mathbb{E}_{P_W \otimes P_S}\big[\frac{\alpha}{2}W^\top H^*(S) W\big] - \mathbb{E}_{P_{W,S}}\big[\frac{\alpha}{2}W^\top H^*(S) W\big] \\
&\quad + \mathbb{E}_{P_S \otimes P_W}\Big[\frac{\alpha}{2}\Big(\mathrm{tr}\big(H^*(S)(W^*(S)W^*(S)^\top - WW^*(S)^\top - W^*(S)W^\top)\big)\Big)\Big] \\
&\quad - \mathbb{E}_{P_S \otimes P_{W|S}}\Big[\frac{\alpha}{2}\Big(\mathrm{tr}\big(H^*(S)(W^*(S)W^*(S)^\top - WW^*(S)^\top - W^*(S)W^\top)\big)\Big)\Big].
\end{aligned} \tag{93}$$

Note that $\mathbb{E}_{P_W}[W] = \mathbb{E}_{P_S}[W^*(S)]$ and $\mathbb{E}_{P_{W|S}}[W] = W^*(S)$, we have

$$
\begin{aligned}
\overline{\text{gen}}(P^{\infty}_{W|S}, \mu) &= \frac{I(W;S) + L(W;S)}{\alpha} \\
&= \mathbb{E}_{P_W \otimes P_S}\big[\frac{1}{2} W^\top H^*(S) W\big] - \mathbb{E}_{P_{W,S}}\big[\frac{1}{2} W^\top H^*(S) W\big] \\
&\quad + \mathbb{E}_{P_S}\Big[\frac{1}{2}\Big(\text{tr}\big(H^*(S)(-\mathbb{E}[W^*(S)]W^*(S)^\top - W^*(S)\mathbb{E}[W^*(S)]^\top)\big)\Big)\Big] \\
&\quad - \mathbb{E}_{P_S}\Big[\frac{1}{2}\Big(\text{tr}\big(H^*(S)(-W^*(S)W^*(S)^\top - W^*(S)W^*(S)^\top)\big)\Big)\Big] \\
&= \mathbb{E}_{P_W \otimes P_S}\big[\frac{1}{2} W^\top H^*(S) W\big] - \mathbb{E}_{P_{W,S}}\big[\frac{1}{2} W^\top H^*(S) W\big] \\
&\quad + \mathbb{E}_{P_S}\Big[(W^*(S) - \mathbb{E}[W^*(S)])^\top (H^*(S)W^*(S) - \mathbb{E}[H^*(S)W^*(S)])\Big]. \quad \square
\end{aligned}
$$

Proposition 2. ***(restated)*** *If we assume that $\pi(W)$ is a uniform distribution over $\mathcal{W}$, and the Hessian matrices $H^*_u(S)$ are not singular for all $u \in \{1, \cdots, M\}$, then the generalization error of the $(\infty, \pi(\boldsymbol{w}), L_E(\boldsymbol{w}, s))$-Gibbs algorithm in the multiple-well case can be bounded as*

$$
\begin{aligned}
\overline{gen}(P^{\infty}_{W|S}, P_S) \leq \frac{1}{M}\sum_{u=1}^{M} \Big[& \mathbb{E}_{\Delta_{W_u,S}}\big[\frac{1}{2} W_u^\top H^*_u(S) W_u\big] \\
& + \mathbb{E}_{P_S}\Big[(W^*_u(S) - \mathbb{E}[W^*_u(S)])^\top H_u (W^*_u(S) - \mathbb{E}[W^*_u(S)])\Big]\Big].
\end{aligned}
$$

Proof. In this multiple-well case, it is shown in [12] that the Gibbs algorithm can be approximated by the following Gaussian mixture distribution

$$
P^{\alpha}_{W|S} \to \frac{1}{\sum_{u=1}^{M} \pi(W^*_u(S))} \sum_{u=1}^{M} \pi\big(W^*_u(S)\big)\mathcal{N}\big(W^*_u(S), \frac{1}{\alpha} H^*_u(S)^{-1}\big), \tag{94}
$$

as long as $H^*_u(S) \triangleq \nabla^2_w L_E(w, S)\big|_{w = W^*_u(S)}$ is not singular for all $u \in \{1, \cdots, M\}$.

However, there is no closed form for the symmetrized KL information for Gaussian mixtures. Thus, we use Theorem 1 to construct an upper bound of the generalization error.

Consider the latent random variable $U \in \{1, \cdots, M\}$ which denotes the index of the Gaussian component of $P^{\alpha}_{W|S}$. Then, conditioning on U and S, W is a Gaussian random variable. Moreover, since $\pi(W)$ is a uniform prior, U is a discrete uniform distribution $P_U(U = u) = \frac{1}{M}$, and $U \perp S$. Note that for mutual information, we have

$$
I(S;W|U) = I(S;W|U) + I(S;U) = I(S;W,U) = I(S;W) + I(S;U|W) \geq I(S;W), \tag{95}
$$

and for lautum information

$$
L(W;S) \overset{(a)}{\leq} L(W,U;S) \overset{(b)}{=} L(U;S) + L(W;S|U) = L(W;S|U), \tag{96}
$$

where (a) is due to the data processing inequality for any f-divergence, and (b) follows by the fact that the chain rule of lautum information holds when $U \perp S$ as shown in [49].

Thus, we can upper bound $I(S;W)$ and $L(S;W)$ with $I(S;W|U)$ and $L(S;W|U)$, respectively,

$$
\begin{aligned}
&\overline{\text{gen}}(P^{\infty}_{W|S}, \mu) \\
&= \lim_{\alpha \to \infty} \frac{I(S;W) + L(S;W)}{\alpha} \\
&\leq \lim_{\alpha \to \infty} \frac{I(S;W|U) + L(S;W|U)}{\alpha} \\
&= \mathbb{E}_U\Big[\mathbb{E}_{P_{W|U} \otimes P_S}\big[\frac{1}{2} W^\top H(w^*_u(S), S) W\big]\Big] - \mathbb{E}_U\Big[\mathbb{E}_{P_{W,S|U}}\big[\frac{1}{2} W^\top H(w^*_U(S), S) W\big]\Big] \\
&\qquad + \mathbb{E}_U\Big[\mathbb{E}_{P_S}\Big[(w^*_U(S) - \mathbb{E}[w^*_U(S)])^\top (H(w^*_U(S), S) w^*_U(S) - \mathbb{E}[H(w^*_U(S), S) w^*_U(S)])\Big]\Big]. \, \square
\end{aligned}
$$

E.2 Regularity Conditions for MLE

In this section, we present the regularity conditions required by the asymptotic normality [64] of maximum likelihood estimates.

Assumption 1. ***Regularity Conditions for MLE:***

1. *$f(z|\boldsymbol{w}) \neq f(z|\boldsymbol{w}')$ for $\boldsymbol{w} \neq \boldsymbol{w}'$.*
2. *$\mathcal{W}$ is an open subset of $\mathbb{R}^d$.*
3. *The function $\log f(z|\boldsymbol{w})$ is three times continuously differentiable with respect to $\boldsymbol{w}$.*
4. *There exist functions $F_1(z) : \mathcal{Z} \to \mathbb{R}$, $F_2(z) : \mathcal{Z} \to \mathbb{R}$ and $M(z) : \mathcal{Z} \to \mathbb{R}$, such that*
$$\mathbb{E}_{Z\sim f(z|\boldsymbol{w})}[M(Z)] < \infty,$$
and the following inequalities hold for any $\boldsymbol{w} \in \mathcal{W}$,
$$\left|\frac{\partial \log f(z|\boldsymbol{w})}{\partial w_i}\right| < F_1(z), \qquad \left|\frac{\partial^2 \log f(z|\boldsymbol{w})}{\partial w_i \partial w_j}\right| < F_1(z),$$
$$\left|\frac{\partial^3 \log f(z|\boldsymbol{w})}{\partial w_i \partial w_j \partial w_k}\right| < M(z), \qquad i, j, k = 1, 2, \cdots, d.$$
5. *The following inequality holds for an arbitrary $\boldsymbol{w} \in \mathcal{W}$,*
$$0 < \mathbb{E}_{Z\sim f(z|\boldsymbol{w})}\left[\frac{\partial \log f(z|\boldsymbol{w})}{\partial w_i}\frac{\partial \log f(z|\boldsymbol{w})}{\partial w_j}\right] < \infty, \quad i, j = 1, 2, \cdots, d.$$

E.3 Bayesian Learning Algorithm

In this section, we show that the symmetrized KL information can be used to characterize the generalization error of Gibbs algorithm in a different asymptotic regime, i.e., inverse temperature $\alpha = n$, then α and n go to infinity simultaneously. In this regime, the Gibbs algorithm is equivalent to the Bayesian posterior distribution instead of ERM.

Suppose that we have n i.i.d. training samples $S = \{Z_i\}_{i=1}^n$ generated from the distribution P_Z defined on $\mathcal{Z}$, and we want to fit the training data with a parametric distribution family $\{f(z_i|\boldsymbol{w})\}_{i=1}^n$, where $\boldsymbol{w} \in \mathcal{W} \subset \mathbb{R}^d$ denotes the parameter and $\pi(\boldsymbol{w})$ denotes a pre-selected prior distribution. Here, the true data-generating distribution may not belong to the parametric family, i.e., $P_Z \neq f(\cdot|\boldsymbol{w})$ for $\boldsymbol{w} \in \mathcal{W}$. The following Bayesian posterior distribution

$$P_{W|S}(\boldsymbol{w}|z^n) = \frac{\pi(\boldsymbol{w})\prod_i^n f(z_i|\boldsymbol{w})}{V(z^n)}, \quad \text{with} \quad V(z^n) = \int \pi(\boldsymbol{w})\prod_i^n f(z_i|\boldsymbol{w})dw, \tag{97}$$

is equivalent to the $(n, \pi(\boldsymbol{w}), L_E(\boldsymbol{w}, s))$-Gibbs algorithm with log-loss $\ell(\boldsymbol{w}, z) = -\log f(z|\boldsymbol{w})$. Thus, Theorem 1 can be applied directly, and we just need to evaluate $I_{\mathrm{SKL}}(W; S)$.

We further assume that the parametric family $\{f(z|\boldsymbol{w}), \boldsymbol{w} \in \mathcal{W}\}$ and prior $\pi(\boldsymbol{w})$ satisfy all the regularization conditions required for the Bernstein–von-Mises theorem [64] and the asymptotic Normality of the maximum likelihood estimate (MLE), including Assumption 1 and the condition that $\pi(w)$ is continuous and $\pi(w) > 0$ for all $w \in \mathcal{W}$.

In the asymptotic regime $n \to \infty$, Bernstein–von-Mises theorem under model mismatch [38, 64] states that we could approximate the Bayesian posterior distribution $P_{W|S}$ in (97) by

$$\mathcal{N}(\hat{W}_{\mathrm{ML}}, \frac{1}{n}J(\boldsymbol{w}^*)^{-1}), \quad \text{where} \quad \hat{W}_{\mathrm{ML}} \triangleq \underset{\boldsymbol{w}\in\mathcal{W}}{\arg\max}\sum_{i=1}^n \log f(Z_i|\boldsymbol{w}), \tag{98}$$

denotes the MLE and

$$J(\boldsymbol{w}) \triangleq \mathbb{E}_Z\big[-\nabla_{\boldsymbol{w}}^2 \log f(Z|\boldsymbol{w})\big] \quad \text{with} \quad \boldsymbol{w}^* \triangleq \underset{\boldsymbol{w}\in\mathcal{W}}{\arg\min}\, D(P_Z\|f(\cdot|\boldsymbol{w})).$$

The asymptotic Normality of the MLE states that the distribution of $\hat{W}_{\mathrm{ML}}$ will converge to

$$\mathcal{N}(\boldsymbol{w}^*, \frac{1}{n} J(\boldsymbol{w}^*)^{-1}\mathcal{I}(\boldsymbol{w}^*)J(\boldsymbol{w}^*)^{-1}) \quad \text{with} \quad \mathcal{I}(\boldsymbol{w}) \triangleq \mathbb{E}_Z\left[\nabla_{\boldsymbol{w}} \log f(Z|\boldsymbol{w}) \nabla_{\boldsymbol{w}} \log f(Z|\boldsymbol{w})^\top\right]$$

as $n \to \infty$. Thus, the marginal distribution P_W can be approximated by a Gaussian distribution regardless the choice of prior $\pi(\boldsymbol{w})$.

Then, the symmetrized KL information can be computed using Lemma 4. By Theorem 1, we have

$$\overline{\text{gen}}(P_{W|S}, P_Z) = \frac{I_{\text{SKL}}(S;W)}{n} = \frac{\text{tr}(\mathcal{I}(\boldsymbol{w}^*)J(\boldsymbol{w}^*)^{-1})}{n}. \tag{99}$$

When the true model is in the parametric family $P_Z = f(\cdot|\boldsymbol{w}^*)$, we have $\mathcal{I}(\boldsymbol{w}^*) = J(\boldsymbol{w}^*)$, which gives the Fisher information matrix and $\overline{\text{gen}}(P_{W|S}, P_Z) = \frac{d}{n}$. This result suggests that the expected generalization error of MLE and that of the Bayesian posterior distribution are the same under suitable regularity conditions.

E.4 Behavior of Empirical Risk

As an aside, we show that the empirical risk is a decreasing function of the inverse temperature α. To see this, we first note that the derivative of $P^\alpha_{W|S}$ with respect to α is given by

$$\frac{\mathrm{d}P^\alpha_{W|S}(w|s)}{\mathrm{d}\alpha} = P^\alpha_{W|S}(w|s)\left(\mathbb{E}_{P^\alpha_{W|S}}[L_E(w,S)] - L_E(w,S)\right). \tag{100}$$

Then, we can compute the derivative of the empirical risk with respect to α as follows:

$$\begin{aligned}
&\frac{\mathrm{d}\mathbb{E}_{P_{W,S}}[L_E(W,S)]}{\mathrm{d}\alpha} = \mathbb{E}_{P_S}[\frac{\mathrm{d}\mathbb{E}_{P^\alpha_{W|S}}[L_E(W,S)]}{\mathrm{d}\alpha}] \\
&= \mathbb{E}_{P_S}\left[\int_{\mathcal{W}} L_E(w,S)\frac{\mathrm{d}P^\alpha_{W|S}(w|S)}{\mathrm{d}\alpha}dw\right] \\
&= \mathbb{E}_{P_S}\left[\int_{\mathcal{W}} P^\alpha_{W|S}(w|s)\left(L_E(w,S)\mathbb{E}_{P^\alpha_{W|S}}[L_E(w,S)] - L_E^2(w,S)\right)dw\right] \\
&= \mathbb{E}_{P_S}\left[\mathbb{E}^2_{P^\alpha_{W|S}}[L_E(w,S)] - \mathbb{E}_{P^\alpha_{W|S}}[L^2_E(w,S)]\right] \\
&= -\mathbb{E}_{P_S}[\text{Var}_{P^\alpha_{W|S}}[L_E(W,S)]] \leq 0
\end{aligned} \tag{101}$$

When $\alpha = 0$, it can be shown that $(0, \pi(w), L_E(w,s))$-Gibbs algorithm has zero generalization error. However, the empirical risk in this case could be large, since the training samples are not used at all. As $\alpha \to \infty$, the empirical risk is decreasing, but the generalization error could be large. Thus, the inverse temperature α controls the trade-off between the empirical risk and the generalization error.

F Regularized Gibbs Algorithm

F.1 Proofs of Proposition 3 and Proposition 4

Proposition 3. ***(restated)*** *For $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm, its expected generalization error is given by*

$$\overline{gen}(P^\alpha_{W|S}, P_S) = \frac{I_{\text{SKL}}(W;S)}{\alpha} - \lambda\mathbb{E}_{\Delta_{W,S}}[R(W,S)],$$

where $\mathbb{E}_{\Delta_{W,S}}[R(W,S)] = \mathbb{E}_{P_W\otimes P_S}[R(W,S)] - \mathbb{E}_{P_{W,S}}[R(W,S)]$.

Proof. For $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm, we have

$$\begin{aligned}
I_{\text{SKL}}(W;S) &= \mathbb{E}_{P_{W,S}}[\log(P^\alpha_{W|S})] - \mathbb{E}_{P_W\otimes P_S}[\log(P^\alpha_{W|S})] \\
&= \alpha\left(\mathbb{E}_{P_W\otimes P_S}[L_E(W,S)] - \mathbb{E}_{P_{W,S}}[L_E(W,S)]\right) \\
&\quad + \alpha\lambda\left(\mathbb{E}_{P_W\otimes P_S}[R(W,S)] - \mathbb{E}_{P_{W,S}}[R(W,S)]\right) \\
&= \alpha\overline{\text{gen}}(P^\alpha_{W|S}, P_S) + \alpha\lambda\mathbb{E}_{\Delta_{W,S}}[R(W,S)].
\end{aligned}$$

□

Proposition 4. ***(restated)*** *Suppose that we adopt the ℓ_2-regularizer $R(w,s) = \|w - T(s)\|_2^2$, where $T(\cdot)$ is an arbitrary deterministic function $T : \mathcal{Z}^n \to \mathcal{W}$. Then, the expected generalization error of $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm is*

$$\overline{gen}(P_{W|S}^{\alpha}, P_S) = \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} - \lambda \mathrm{tr}\big(\mathrm{Cov}[W, T(S)]\big),$$

where $\mathrm{Cov}[W, T(S)]$ denotes the covariance matrix between W and $T(S)$.

Proof. We just need to compute $\mathbb{E}_{\Delta_{W,S}}[R(W,S)]$ by considering $R(w,s) = \|w - T(s)\|_2^2$,

$$\begin{aligned}
&\mathbb{E}_{P_W \otimes P_S}[R(W,S)] - \mathbb{E}_{P_{W,S}}[R(W,S)] \\
&= \mathbb{E}_{P_W \otimes P_S}\left[\|W - T(S)\|_2^2\right] - \mathbb{E}_{P_{W,S}}\left[\|W - T(S)\|_2^2\right] \\
&= \mathbb{E}_{P_{W,S}}\left[W^T T(S)\right] - \mathbb{E}_{P_W \otimes P_S}\left[W^T T(S)\right] \\
&= \mathrm{tr}(\mathrm{Cov}(W, T(S))).
\end{aligned}$$

□

F.2 Generalization Error Upper Bounds for Regularized Gibbs Algorithm

For general regularization function $R(w,s)$, we can bound the $\mathbb{E}_{\Delta_{W,S}}[R(W,S)]$ term using the mutual information-based generalization error bound in [19, 71].

Proposition 5. *Suppose that the regularizer function $R(w,s)$ satisfies $\Lambda_{R(w,s)}(\lambda) \leq \psi(\lambda)$, for $\lambda \in (-b, b)$ and $b > 0$ under data-generating distribution P_Z for all $w \in \mathcal{W}$. Then the following lower and upper bounds holds for $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm:*

$$\frac{I_{\mathrm{SKL}}(W;S)}{\alpha} - \lambda \psi^{*-1}(I(W;S)) \leq \overline{gen}(P_{W|S}^{\alpha}, P_S) \leq \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} + \lambda \psi^{*-1}(I(W;S)) \tag{102}$$

Proof. Using the decoupling lemma from [19, Theorem 1], we have:

$$|\mathbb{E}_{\Delta_{W,S}}[R(W,S)]| \leq \psi^{*-1}(I(W;S)), \tag{103}$$

which means that

$$-\psi^{*-1}(I(W;S)) \leq \mathbb{E}_{\Delta_{W,S}}[R(W,S)] \leq \psi^{*-1}(I(W;S)). \tag{104}$$

The final results (102) follows directly from (104) and Proposition 3. □

Note that the bounded CGF assumption is on the regularizer function $R(w,s)$. We could consider different assumptions on $\psi(\lambda)$ in Proposition 5 including sub-Gaussian, sub-Exponential and sub-Gamma. We focus on sub-Gaussian assumption for regularizer function in the following result.

Corollary 5. *Suppose that the regularizer function $R(w,s)$ is σ-sub-Gaussian under the distribution P_S for all $w \in \mathcal{W}$. Then the following bounds holds for $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm:*

$$\frac{I_{\mathrm{SKL}}(W;S)}{\alpha} - \lambda\sqrt{2\sigma^2 I(W;S)} \leq \overline{gen}(P_{W|S}^{\alpha}, P_S) \leq \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} + \lambda\sqrt{2\sigma^2 I(W;S)} \tag{105}$$

Proof. Considering $\psi^{*-1}(I(W;S)) = \sqrt{2\sigma^2 I(W;S)}$ in Proposition 5 completes the proof. □

By assuming σ-sub-Gaussianity for both loss function and the regularizer, we provide a generalization error upper bound for regularized Gibbs algorithm in the following proposition.

Proposition 6. *Suppose that the training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d generated from the distribution P_Z, and the non-negative loss function $\ell(w,Z)$ and the regularizer function $R(w,s)$ are σ-sub-Gaussian under data-generating distribution P_Z for all $w \in \mathcal{W}$. We further assume $C_E = \frac{L(W;S)}{I(W;S)}$ for some $C_E \geq 0$. Then the following bounds holds for $(\alpha, \pi(w), L_E(w,s) + \lambda R(w,s))$-Gibbs algorithm:*

$$\overline{gen}(P_{W|S}^{\alpha}, P_S) \leq \begin{cases} \frac{2\sigma^2\alpha}{(1+C_E)}\left(\frac{1}{n} - \frac{\lambda}{\sqrt{n}}\right), & \textit{if} \quad 0 \leq \lambda \leq \frac{1}{\sqrt{n}} \quad \textit{and} \quad I(W;S) \leq \frac{2\sigma^2\alpha^2}{(1+C_E)^2}\left(\frac{1}{\sqrt{n}} - \lambda\right)^2; \\ \frac{2\sigma^2\alpha}{(1+C_E)}\left(\frac{1}{n} + \frac{\lambda}{\sqrt{n}}\right), & \textit{otherwise.} \end{cases} \tag{106}$$

Proof. Using Proposition 5 and [71, Theorem 1], we have

$$\frac{I_{\mathrm{SKL}}(W;S)}{\alpha} - \lambda\sqrt{2\sigma^2 I(W;S)} \leq \min\left(\sqrt{\frac{2\sigma^2 I(W;S)}{n}}, \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} + \lambda\sqrt{2\sigma^2 I(W;S)}\right).$$

If $\sqrt{\frac{2\sigma^2 I(W;S)}{n}} \leq \frac{I_{\mathrm{SKL}}(W;S)}{\alpha} + \lambda\sqrt{2\sigma^2 I(W;S)}$, and using $C_E I(W;S) = L(W;S)$, then we have:

$$\frac{I(W;S)(1+C_E)}{\alpha} - \lambda\sqrt{2\sigma^2 I(W;S)} \leq \sqrt{\frac{2\sigma^2 I(W;S)}{n}}. \tag{107}$$

Solving (107) gives

$$I(W;S) \leq \frac{2\sigma^2\alpha^2}{(1+C_E)^2}\left(\frac{1}{\sqrt{n}} + \lambda\right)^2. \tag{108}$$

If $\frac{I_{\mathrm{SKL}}(W;S)}{\alpha} + \lambda\sqrt{2\sigma^2 I(W;S)} < \sqrt{\frac{2\sigma^2 I(W;S)}{n}}$, and using $C_E I(W;S) = L(W;S)$, then we have:

$$I(W;S) \leq \frac{2\sigma^2\alpha^2}{(1+C_E)^2}\left(\frac{1}{\sqrt{n}} - \lambda\right)^2, \tag{109}$$

for $0 \leq \lambda \leq \frac{1}{\sqrt{n}}$. Combining the (108) and (109) with [71, Theorem 1] completes the proof. □

In Proposition 6, if $0 \leq \lambda \leq \frac{1}{\sqrt{n}}$ and $\frac{I(W;S)(1+C_E)^2}{2\alpha^2\sigma^2} \leq \left(\frac{1}{\sqrt{n}} - \lambda\right)^2$ hold, then the upper bound would be tighter than the upper bound in Theorem 2 with $C_E = \frac{L(W;S)}{I(W;S)}$.