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
\[
\Lambda_X(\lambda) \triangleq \log \mathbb{E}[e^{\lambda (X - \mathbb{E}[X])}].
\] (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 $\psi$ defined on the interval $[0, b)$, where $0 < b \leq \infty$, its Legendre dual $\psi^*$ is defined as
\[
\psi^*(x) \triangleq \sup_{\lambda \in [0, b)} (\lambda x - \psi(\lambda)).
\] (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 $\psi^*(x)$ defined above is a non-negative convex and non-decreasing function on $[0, \infty)$ with $\psi^*(0) = 0$. Moreover, its inverse function $\psi^{*-1}(y) = \inf\{x \geq 0 : \psi^*(x) \geq y\}$ is concave, and can be written as
\[
\psi^{*-1}(y) = \inf_{\lambda \in [0, b)} \left( \frac{y + \psi(\lambda)}{\lambda} \right), \quad b > 0.
\] (29)

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

- **Sub-Gaussian:** A random variable $X$ is $\sigma$-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,
\[
\psi^{*-1}(y) = \sqrt{2\sigma^2 y}.
\]
We start with the following two Lemmas:

**Lemma 1.** Let $\psi(x) = \frac{\sigma^2 x^2}{2}$ be a proxy for the population risk. Then,
\[
\mathbb{E}[\psi^{w}(y)] = \psi^{*^{-1}}(y) = \begin{cases} 
\sqrt{2\sigma^2 y}, & \text{if } y \leq \frac{\sigma^2}{2b}; \\
by + \frac{\sigma^2}{2b}, & \text{otherwise}.
\end{cases}
\]

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

Sub-Gamma condition is 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., $\sigma$-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.,
\[
\mathbb{E}[\psi^{w}(y)] = \psi^{*^{-1}}(y) = \begin{cases} 
\sqrt{2\sigma^2 y}, & \text{if } y \leq \frac{\sigma^2}{2b}; \\
by + \frac{\sigma^2}{2b}, & \text{otherwise}.
\end{cases}
\]

**Proof.**
\[
\mathbb{E}_P[w, S] [J_P(W, \mu) - J_E(W, S)] = \alpha \cdot \mathbb{F}(P_W|S, P_S).
\]

\[
\mathbb{E}_P[w, S] [J_P(W, \mu) - J_E(W, S)] = \mathbb{E}_P[w, S] \mathbb{E}_P[z] \left[ \frac{1}{n} \sum_{i=1}^{n} \ell(W, Z_i) - \frac{1}{n} \sum_{i=1}^{n} \ell(W, Z_i) \right] 
+ \mathbb{E}_P[w, S] [g(W) + \mathbb{E}_P[h(S)]] - \mathbb{E}_P[w, S] [g(W) + h(S)]
\]
\[
= \alpha \cdot \mathbb{E}_P[w, S] [L_P(W, \mu) - L_E(W, S)]
= \alpha \cdot \mathbb{F}(P_W|S, P_S).
\]

**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
\[
\mathbb{E}_P[w, S] [J_P(W, \mu) - J_E(W, S)] = I_{\text{SKL}}(W; S).
\]

**Proof.**
\[
I(W; S) + L(W; S)
= \mathbb{E}_P[w, S] \log \frac{P_W|S(W|S)}{P_W(W)} + \mathbb{E}_P[w, S] \log \frac{P_W(W)}{P_W|S(W|S)}
= \mathbb{E}_P[w, S] \log \frac{P_W|S(W|S)}{P_W(W|S)} - \mathbb{E}_P[w, S] \log \frac{P_W|S(W|S)}{P_W(W|S)}
= \mathbb{E}_P[w, S] [-P_S \log P_W|S(W|S) + \log P_W|S(W|S)]
= \mathbb{E}_P[w, S] [J_P(W, \mu) - J_E(W, S)].
\]

**Theorem 1. (restated)** For $(\alpha, \pi(w), L_E(w, s))$-Gibbs algorithm,
\[
P^\alpha_{W|S}(w|s) = \frac{\pi(w) \alpha^{L_E(w, s)}}{V(s, \alpha)}, \quad \alpha > 0,
\]
its expected generalization error is given by
\[
\mathbb{F}(P^\alpha_{W|S}(w|s), P_S) = \frac{I_{\text{SKL}}(W; S)}{\alpha}.
\]
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 symmetrized KL divergence. As shown in [49], symmetrized KL divergence is an $f$-divergence. In general, the product-of-marginal distribution does not minimize $D(Q_W \otimes Q_S \parallel P_W)$, and lautum information satisfies the following variational characterization

$$L(W; S) = \inf_{Q_S} D(P_W \otimes P_S \parallel P_W \parallel Q_S).$$

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

**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

$$J_E(w, s) = -\log P_{W|S}(w|s) = \alpha L_E(w, s) - \log \pi(w) + \log V(s, \alpha),$$

then we have:

$$I_{SKL}(W; S) = \mathbb{E}_{P_{WS}}[J_P(W, P_S) - J_E(W, S)] = \alpha \cdot \text{gem}(P_{WS}^\alpha, P_S).$$

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:

$$\text{gem}(P_{WS}^\alpha, P_S) \geq \frac{TV^2(P_{WS}, P_W \otimes P_S)}{\alpha},$$

where

$$TV(P_{WS}, P_W \otimes P_S) \triangleq \int \int |P_{WS}(w, s) - P_W(w)P_S(s)| dw ds$$

denotes total variation distance.

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

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

Note that the lower bound in Corollary 2 is bounded in $[0, \frac{1}{\alpha}]$.

**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 $S \leftrightarrow W \leftrightarrow W'$,

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

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 $P_W$ 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

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

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 $\inf_{Q_W} D(P_S \otimes Q_W \parallel P_{W,S})$, but it is not true as shown in [49]. In general, the product-of-marginal distribution does not minimize $D(Q_W \otimes Q_S \parallel P_{W,S})$, and lautum information satisfies the following variational characterization

$L(W; S) = \inf_{Q_S} D(P_W \otimes P_S \parallel P_{W|S} \otimes Q_S).$
Thus, individual sample symmetrized KL information cannot be used to characterize the behavior of the generalization error of Gibbs algorithm using individual terms $I_{SKL}(W; Z_i)$. 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 $I_{SKL}(W; Z_i)$. To see this, we have the following example to show that the joint symmetrized KL information $I_{SKL}(W; S)$ can be either larger or smaller than the sum of individual terms $I_{SKL}(W; Z_i)$.

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

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

(42)

It can be verified that $Z_1$ and $Z_2$ are mutually independent Bernoulli random variable with $p = \frac{1}{4}$, 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

$$
I_{SKL}(W; Z_1) = I_{SKL}(W; Z_2) = 0.4200, \quad I_{SKL}(W; Z_1, Z_2) = 0.7329,
$$

$I_{SKL}(W; Z_1) + I_{SKL}(W; Z_2) > I_{SKL}(W; Z_1, Z_2)$.

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

$$
I_{SKL}(W; Z_1) = I_{SKL}(W; Z_2) = 0.1255, \quad I_{SKL}(W; Z_1, Z_2) = 0.2741,
$$

$I_{SKL}(W; Z_1) + I_{SKL}(W; Z_2) < I_{SKL}(W; Z_1, Z_2)$.

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

$$
W = \frac{\sigma_1^2}{\sigma_0^2} \mu_0 + \frac{\sigma_1^2}{\sigma_0^2} \sum_{i=1}^n Z_i + N, \quad \text{with} \quad \sigma_1^2 = \frac{\sigma_0^2 \sigma^2}{n \sigma_0^2 + \sigma^2}
$$

(43)
where $N \sim \mathcal{N}(0, \sigma_Z^2 I_d)$ is independent from the training samples $S = \{Z_i\}_{i=1}^n$. Thus,

$$\mathbb{E}[P_W | S, P_S] = \mathbb{E}_{P,w,s}[L_P(W, \mu) - L_E(W, S)]$$

$$= \mathbb{E}_{P,w,s} \left[ \mathbb{E}_{P_Z} \left[ \|W - \tilde{Z}\|^2 \right] - \frac{1}{n} \sum_{i=1}^n \|W - Z_i\|^2 \right]$$

$$= \mathbb{E}_{P,w, z_i \circ P_Z} \left[ 2\|W - \tilde{Z} - Z_i\| \langle Z_i - \tilde{Z} \rangle \right]$$

$$= \mathbb{E} \left[ \frac{2\sigma^2}{\sigma_0^2} \mathbb{E} \left[ Z_i^\top (Z_i - \tilde{Z}) \right] \right]$$

$$= \frac{2d\sigma^2 \sigma_Z^2}{\sigma^2} = \frac{2d\sigma^2 \sigma_Z^2}{n\sigma_0^2 + \sigma^2}, \quad \text{(44)}$$

where $\tilde{Z} \sim \mathcal{N}(\mu, \sigma_Z^2 I_d)$ 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 $Z_i - \tilde{Z}$ 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

$$Y = AX + N_G, \quad (45)$$

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

$$I(X; Y) = \frac{1}{2} \text{tr} \left( \Sigma_{N_G}^{-1} A \Sigma A^\top \right) - D(P_Y \| P_{N_G}), \quad (46)$$

$$L(X; Y) = \frac{1}{2} \text{tr} \left( \Sigma_{N_G}^{-1} A \Sigma A^\top \right) + D(P_Y \| P_{N_G}). \quad (47)$$

In our example, the output $W$ can be written as

$$W = \frac{\sigma^2}{\sigma_0^2} \mu_0 + \frac{\sigma^2}{\sigma_0^2} \sum_{i=1}^n Z_i + N = \frac{\sigma^2}{\sigma_0^2} \sum_{i=1}^n (Z_i - \mu) + \frac{\sigma^2}{\sigma_0^2} \mu_0 + \frac{n\sigma^2}{\sigma_0^2} \mu + N, \quad (48)$$

where $N \sim \mathcal{N}(0, \sigma^2 I_d)$. Setting $P_{N_G} \sim \mathcal{N}(\frac{\sigma^2}{\sigma_0^2} \mu_0 + \frac{n\sigma^2}{\sigma_0^2} \mu, \sigma^2 I_d)$ and $\Sigma = \sigma^2 I_{nd}$ in Lemma 4 gives

$$\text{tr} \left( \Sigma_{N_G}^{-1} A \Sigma A^\top \right) = \text{tr} \left( \frac{\sigma^2}{\sigma_0^2} A A^\top \right), \quad (49)$$

and noticing that $AA^\top = \frac{n\sigma^4}{\sigma_0^4} I_d$ 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 $P_Z \sim \mathcal{N}(\mu, \sigma_Z^2 I_d)$. 


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

$$\begin{align*}
gen(P_{W|S}, P_S) & \leq \frac{1}{n} \sum_{i=1}^{n} \psi_-^{-1}(I(W; Z_i)), \quad \text{(50)} \\
-\gen(P_{W|S}, P_S) & \leq \frac{1}{n} \sum_{i=1}^{n} \psi_+^{-1}(I(W; Z_i)). \quad \text{(51)}
\end{align*}$$

We need to compute the mutual information between each individual sample and the output hypothesis $I(W; Z_i)$, and the CGF of $\ell(\tilde{W}, \tilde{Z})$, where $\tilde{W}, \tilde{Z}$ 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:

$$\begin{align*}
\text{Cov}[Z_i, W] = \begin{pmatrix} \sigma_Z^2 I_d & \frac{\sigma_Z^2}{\sigma_W^2} \sigma_I^2 I_d \\ \frac{\sigma_Z^2}{\sigma_W^2} \sigma_I^2 I_d & \left(\frac{n\sigma_I^4}{\sigma_Z^4} + \sigma_I^2\right) I_d \end{pmatrix},
\end{align*}$$

then, we have

$$\begin{align*}
I(W; Z_i) & = \frac{d}{2} \log \left(\frac{n\sigma_I^4}{\sigma_Z^4} + \sigma_I^2 + \sigma_Z^2\right) \\
& = \frac{d}{2} \log \left(1 + \frac{\sigma_Z^2 \sigma_I^2}{(n-1)\sigma_I^2 \sigma_Z^2 + \sigma_I^4}\right) \\
& = \frac{d}{2} \log \left(1 + \frac{\sigma_Z^2 \sigma_I^2}{(n-1)\sigma_Z^2 \sigma_I^2 + n\sigma_I^2 \sigma_Z^2 + \sigma_I^4}\right), \quad \text{(53)}
\end{align*}$$

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

$$W \sim \mathcal{N}\left(\frac{\sigma_I^2}{\sigma_Z^2} \mu_0 + \frac{n\sigma_I^2}{\sigma_Z^2} \mu, \left(\frac{n\sigma_I^4}{\sigma_Z^4} + \sigma_I^2\right) I_d\right), \quad \text{(54)}$$

it can be shown that $\ell(\tilde{W}, \tilde{Z}) = \|\tilde{Z} - \tilde{W}\|^2$ is a scaled non-central chi-square distribution with $d$ degrees of freedom, where the scaling factor $\sigma_Z^2 \triangleq \left(\frac{n\sigma_I^4}{\sigma_Z^4} + 1\right)\sigma_Z^2 + \sigma_I^2$ and its non-centrality parameter $\eta \triangleq \frac{\sigma_Z^2}{\sigma_Z^2 + \sigma_I^2} \|\mu_0 - \mu\|^2$.

Note that the expectation of chi-square distribution with non-centrality parameter $\eta$ 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 $\ell(\tilde{W}, \tilde{Z})$ is given by

$$\begin{align*}
\Lambda_{\ell(\tilde{W}, \tilde{Z})}(\lambda) & = -(d\sigma_Z^2 + \eta)\lambda + \frac{\eta \lambda}{1-2\sigma_I^2 \lambda} - \frac{d}{2} \log(1-2\sigma_I^2 \lambda), \quad \text{(55)}
\end{align*}$$

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

$$\begin{align*}
\Lambda_{\ell(\tilde{W}, \tilde{Z})}(\lambda) & = -d\sigma_Z^2 \lambda - \frac{d}{2} \log(1-2\sigma_I^2 \lambda) + \frac{2\sigma_I^2 \eta \lambda^2}{1-2\sigma_I^2 \lambda} \\
& = \frac{d}{2} (-u - \log(1-u)) + \frac{2\sigma_I^2 \eta \lambda^2}{1-2\sigma_I^2 \lambda}, \quad \text{(56)}
\end{align*}$$

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

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

$$\frac{2\sigma_I^2 \eta \lambda^2}{1-2\sigma_I^2 \lambda} \leq 2\sigma_I^2 \eta \lambda^2, \quad \lambda < 0. \quad \text{(58)}$$
We have the following upper bound on the CGF of $\ell(\tilde{W}, Z)$:

$$\Lambda_{\ell(\tilde{W}, Z)}(\lambda) \leq (d_0^2 + 2\sigma^2_2 \eta)\lambda^2, \quad \lambda < 0,$$

which means that $\ell(\tilde{W}, Z)$ is $\sqrt{d_0^2 + 2\sigma^2_2 \eta}$-sub-Gaussian for $\lambda < 0$. Combining the results in (53), Lemma 5 gives the following bound

$$\gen(P_{W|S}, P_S) \leq \sqrt{\frac{d\sigma^2_2 + 2d\sigma^2_2 \eta}{2}} \log(1 + \frac{\sigma^2_0^2 \sigma^2_2}{(n-1)\sigma^2_0^2 \sigma^2_2 + n\sigma^2_0^2 \sigma^2 + \sigma^2}).$$

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

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 $\exists C_E \in \mathbb{R}_+^+$ such that $\frac{L(W, S)}{I(W, S)} \geq C_E$, and we further assume:

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

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

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

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

$$\gen(P_{W|S}^\alpha, P_S) \leq \psi^{-1}\left(\frac{I(W; S)}{n}\right).$$

By Theorem 1 and the assumption on $C_E$, we have

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

Therefore,

$$\frac{(1 + C_E)I(W; S)}{\alpha} \leq \psi^{-1}\left(\frac{I(W; S)}{n}\right). \quad (65)$$

Consider the function $F(u) \triangleq \psi^{-1}\left(\frac{u}{n}\right) - \frac{(1 + C_E)u}{\alpha}$, 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 $\psi^{-1}(\cdot)$ is non-decreasing, we have

$$\gen(P_{W|S}^\alpha, P_S) \leq \psi^{-1}\left(\frac{\kappa}{n}\right) = \frac{(1 + C_E)\kappa}{\alpha}. \quad \square$$

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 $\sigma$-sub-Gaussian assumption.

**Theorem 2. (repeated)** 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 $\sigma$-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

$$0 \leq \gen(P_{W|S}^\alpha, P_S) \leq \frac{2\sigma^2_0 \alpha}{(1 + C_E)n}.$$
We first consider the sub-Exponential case.

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

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

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

\[
\mathbb{E}(P^\alpha_{W|S}, P_S) \leq \frac{(1 + C_E)\kappa}{\alpha} = \frac{2\sigma^2\alpha}{n(1 + C_E)}.
\]

**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 \( (\sigma^2, b) \)-sub-Exponential on the left-tail \(^*\) under distribution \( P_Z \) for all \( w \in W \). We further assume \( C_E \leq \frac{I(W;S)}{I(W;S)} \) for some \( C_E \geq 0 \). Then, for the \((\alpha, \pi(w), L_E(w, s))\)-Gibbs algorithm, we have

\[
\mathbb{E}(P^\alpha_{W|S}, P_S) \leq \begin{cases} 
\frac{2\sigma^2\alpha}{(1 + C_E)n}, & \text{if } n \geq \frac{2bI(W;S)}{\sigma^2} \\
\frac{\sigma^2}{2b} \left(\frac{ab}{n(1 + C_E) - ab} + 1\right), & \text{if } \left\lceil \frac{ab}{1 + C_E} \right\rceil < n < \frac{2bI(W;S)}{\sigma^2}.
\end{cases}
\]

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

\[
\psi^{-1}(y) = \begin{cases} 
\sqrt{2\sigma^2y}, & \text{if } y \leq \frac{\sigma^2}{2b} \\
y + \frac{\sigma^2}{2b}, & \text{otherwise}.
\end{cases}
\]

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

\[
\frac{(1 + C_E)I(W;S)}{\alpha} \leq \sqrt{\frac{2\sigma^2}{n} I(W;S)},
\]

then the following upper bound holds,

\[
I(W;S) \leq \frac{2\sigma^2\alpha^2}{(1 + C_E)^2n},
\]

which gives

\[
\mathbb{E}(P^\alpha_{W|S}, P_S) \leq \frac{2\sigma^2\alpha}{n(1 + C_E)}.
\]

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

\[
\frac{I(W;S)(1 + C_E)}{\alpha} \leq \frac{bI(W;S)}{n} + \frac{\sigma^2}{2b},
\]

then the following upper bound holds when \( n > \frac{ab}{1 + C_E} \),

\[
I(W;S) \leq \frac{\alpha n \sigma^2}{2b(n(1 + C_E) - ab)},
\]

which gives

\[
\mathbb{E}(P^\alpha_{W|S}, P_S) \leq \frac{\sigma^2}{2b} \left(\frac{ab}{n(1 + C_E) - ab} + 1\right).
\]

\(^*\)A random variable \( X \) is \( (\sigma^2, b) \)-sub-Exponential on the left-tail if \( \log \mathbb{E}[e^{\lambda(X - EX)}] \leq \frac{\sigma^2\lambda^2}{2}, \quad -\frac{b}{\sigma^2} \leq \lambda \leq 0.\)
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 \(O(1/n)\) 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 \(\ell(w, Z) = \Gamma(\sigma_z^2, c_z)\)-sub-Gamma on the left-tail \(^*\) under distribution \(P_Z\) for all \(w \in W\). We further assume \(C_E \leq \frac{c_E}{n} \) for some \(C_E \geq 0\). Then, for the \((\alpha, \pi(w), L_E(w, s))\)-Gibbs algorithm, if \(n > \frac{c_E}{C_E} \), we have

\[
\text{gen}(P_{W|S}^n, P_S) \leq \frac{2\sigma_n^2\alpha}{(1 + C_E)n - \alpha c_s} \left(1 + \frac{\alpha c_s}{(1 + C_E)n - \alpha c_s}\right). \tag{73}
\]

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

\[
\frac{(1 + C_E)I(W; S)}{\alpha} \leq \sqrt{\frac{2\sigma_n^2I(W; S)}{n}} + \frac{I(W; S)}{n}. \tag{74}
\]

Then the following upper bound holds when \(n > \frac{C_E}{C_E} \),

\[
I(W; S) \leq \left(\frac{\alpha}{(1 + C_E)n - \alpha c_s}\right)^2 2n\sigma_n^2, \tag{75}
\]

which gives

\[
\text{gen}(P_{W|S}^n, P_S) \leq \frac{2\sigma_n^2\alpha(1 + C_E)n}{(1 + C_E)n - \alpha c_s}. \tag{76}
\]

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 \(O(1/n)\) 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 \(P_{W'}^{\alpha, L_P}\). The following lemma provides an operational interpretation of the symmetrized KL divergence between the Gibbs posterior \(P_{W|S}^\alpha\) and the prior distribution \(P_{W'}^{\alpha, L_P}\).

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

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

where \(\mathbb{E}_{\Delta(P_{W|S}^\alpha \| P_{W'}^{\alpha, L_P})} [f(W)] = \mathbb{E}_{P_{W|S}^\alpha} [f(W)] - \mathbb{E}_{P_{W'}^{\alpha, L_P}} [f(W)]\).

\(^*\)A random variable \(X\) is \(\Gamma(\sigma_x^2, c_x)\)-sub-Gamma on the left-tail if \(\log \mathbb{E}[e^{\lambda(X - EX)}] \leq \frac{\lambda^2 \sigma_x^2}{2(1 - e^{\lambda})}\), for \(\frac{1}{c_x} < \lambda < 0\).
where the last step follows from the sub-Gaussian assumption. Since the above inequality holds for
(a) follows by the fact that partition functions

\[ D_{\text{SKL}}(P^\alpha_{W|S=s} \parallel P^\alpha_{W|L_p}) = \mathbb{E}_{P^\alpha_{W|S=s}} \left[ \log \frac{P^\alpha_{W|S=s}}{P^\alpha_{W|L_p}} \right] - \mathbb{E}_{P^\alpha_{W|L_p}} \left[ \log \frac{P^\alpha_{W|S=s}}{P^\alpha_{W|L_p}} \right] \]

\[ \leq \alpha \mathbb{E} \Delta(p^\alpha_{W|S=s}, p^\alpha_{W|L_p}) \left[ \log (e^{-\alpha(L_E(W,s) - L_p(W,p,s)))} \right] \]

\[ = \alpha \mathbb{E} \Delta(p^\alpha_{W|S=s}, p^\alpha_{W|L_p}) \left[ L_p(W, p_S) - L_E(W,s) \right], \] (77)

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

**Theorem 3.** *(revisited)* 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 \( \sigma \)-sub-Gaussian under data-generating distribution \( P_Z \) for all \( w \in W \). 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 \( P_S \).

\[ \left| \mathbb{E}_{P^\alpha_{W|S=s}} \left[ L_p(W, P_S) - L_E(W,s) \right] \right| \leq \frac{2\sigma^2\alpha}{(1 + C_P(s))n} + \epsilon^2 \]

\[ + 2 \sqrt{\frac{\sigma^2\alpha}{(1 + C_P(s))n} \left( \sqrt{2\sigma^2 D(P_{Z'}||P_Z)} + \epsilon \right)}, \]

where \( \epsilon \triangleq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} \), and \( C_P(s) \leq \frac{D(p^\alpha_{W|S=s})}{D(p^\alpha_{W|S=s}||p^\alpha_{W|L_p})} \) for some \( C_P(s) \geq 0 \).

**Proof.** Using Lemma 6, we have

\[ D_{\text{SKL}}(P^\alpha_{W|S} \parallel P^\alpha_{W|L_p}) = \alpha \left( \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z'})] - \mathbb{E}_{P^\alpha_{W|S=s}}[L_E(W,s)] \right) \]

\[ - \alpha \left( \mathbb{E}_{P^\alpha_{W|L_p}}[L_p(W, P_{Z'})] - \mathbb{E}_{P^\alpha_{W|L_p}}[L_E(W,s)] \right) \]

\[ \leq \alpha \left| \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z'})] - \mathbb{E}_{P^\alpha_{W|S=s}}[L_E(W,s)] \right| \]

\[ + \alpha \left| \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z'})] - \mathbb{E}_{P^\alpha_{W|S=s}}[L_E(W,s)] \right| \]

\[ \leq \alpha \left| \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z'})] - \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z})] \right| \]

\[ + \alpha \left| \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z})] - \mathbb{E}_{P^\alpha_{W|S=s}}[L_E(W,s)] \right| \]

\[ + \alpha \left| \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z})] - \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z})] \right| \]

\[ + \alpha \left| \mathbb{E}_{P^\alpha_{W|S=s}}[L_p(W, P_{Z})] - \mathbb{E}_{P^\alpha_{W|S=s}}[L_E(W,s)] \right| \] (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 \( \lambda \in \mathbb{R} \),

\[ D(P_{Z'}||P_Z) \geq \mathbb{E}_{P_{Z'}}[\lambda(\ell(w, Z'))] - \log \mathbb{E}_{P_Z} [e^{\lambda(\ell(w, Z))}] \]

\[ \geq \lambda(L_p(w, P_{Z'}) - L_p(w, P_{Z})) - \frac{\lambda^2 \sigma^2}{2}, \] (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

\[ |L_p(w, P_{Z'}) - L_p(w, P_{Z})| \leq \sqrt{2\sigma^2 D(P_{Z'}||P_Z)}, \] for all \( w \in W \). (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 $f(w, Z)$ is $\sigma$-sub-Gaussian under $P_Z$ for all $w \in W$, the following bound holds with probability $1 - \delta$,

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

(81)

If we choose $P_{W|S}^\alpha$ as the posterior distribution and $P_{W}^{\alpha, L_p'}$ as the prior distribution, we have

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

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

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

(83)

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

$$
D_{SKL}(P_{W|S}^{\alpha}\|P_{W}^{\alpha, L_p'}) \leq \alpha \sqrt{\frac{2\sigma^2 (D(P_{W|S}^{\alpha}\|P_{W}^{\alpha, L_p'}) + \log(1/\delta))}{n}} + \alpha \sqrt{\frac{2\sigma^2 (\log(1/\delta))}{n}} + 2\alpha \sqrt{2\sigma^2 D(P_Z\|P_Z)}.
$$

(84)

Then, using the assumption that $(1 + C_P(s))D(P_{W|S=s}^{\alpha}\|P_{W}^{\alpha, L_p'}) \leq D_{SKL}(P_{W|S}^{\alpha}\|P_{W}^{\alpha, L_p'})$, we have

$$
(1 + C_P(s))D(P_{W|S=s}^{\alpha}\|P_{W}^{\alpha, L_p'}) \leq \alpha \sqrt{\frac{2\sigma^2 (D(P_{W|S}^{\alpha}\|P_{W}^{\alpha, L_p'}) + \log(1/\delta))}{n}} + \alpha \sqrt{\frac{2\sigma^2 (\log(1/\delta))}{n}} + 2\alpha \sqrt{2\sigma^2 D(P_Z\|P_Z)}.
$$

(85)

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

$$
D(P_{W|S=s}^{\alpha}\|P_{W}^{\alpha, L_p'}) - \sqrt{\frac{2\alpha'^2 \sigma^2 (\log(1/\delta))}{n}} - \sqrt{8\alpha'^2 \sigma^2 D(P_Z\|P_Z)} \leq \frac{2\alpha'^2 \sigma^2 (D(P_{W|S}^{\alpha}\|P_{W}^{\alpha, L_p'}) + \log(1/\delta))}{n}.
$$

(86)

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

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

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

$$
D^2(P_{W|S=s}^{\alpha}\|P_{W}^{\alpha, L_p'}) - 2D(P_{W|S=s}^{\alpha}\|P_{W}^{\alpha, L_p'})(A + B) + B^2 + 2(A - C)B \leq 0.
$$

(87)

Solving the above inequality gives:

$$
0 \leq D(P_{W|S=s}^{\alpha}\|P_{W}^{\alpha, L_p'}) \leq \sqrt{A^2 + 2BC} + A + B.
$$

(88)
As $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for positive $x, y$ and $A \geq C$, we have

$$D(P_{W|S=s}^\alpha\|P_W^\alpha) \leq 2A + B + \sqrt{2BC} \leq 2A + B + \sqrt{2AB} \leq (\sqrt{2A} + \sqrt{B})^2.$$  \hspace{1cm} (89)

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

$$\left| \mathbb{E}_{P_W|S=s} [L_P(W, \mu) - L_E(W, s)] \right| \leq \sqrt{\frac{2\sigma^2(\sqrt{2A} + \sqrt{B})^2 + 2\sigma^2 \log(1/\delta)}{n}}$$

$$\leq \sqrt{\frac{4\sigma^2A}{n} + \frac{2\sigma^2B}{n} + \frac{2\sigma^2 \log(1/\delta)}{n}}$$

$$\leq \frac{2\alpha\sigma^2}{(1 + C_P(s))n} + \frac{\sqrt{2\sigma^2 \log(1/\delta)}}{n}$$

$$+ 2\sqrt{\frac{\alpha\sigma^2}{(1 + C_P(s))n} \left( \frac{\sqrt{2\sigma^2 \log(1/\delta)}}{n} + \sqrt{2\alpha^2 D(P_Z^\alpha\|P_Z)} \right)}.$$  \hspace{1cm} (90)

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 $(\infty, \pi(w), L_E(w, s))$-Gibbs algorithm is

$$\mathbb{E}_{W,S} \left[ \frac{1}{2} W^T H^*(S) W \right]$$

$$+ \mathbb{E}_{W,S} \left[ (W^*(S) - \mathbb{E}[W^*(S)])^T (H^*(S)W^*(S) - \mathbb{E}[H^*(S)W^*(S)]) \right],$$

where $\mathbb{E}_{W,S} [f(W, S)] \equiv \mathbb{E}_{W \otimes S} [f(W, S)] - \mathbb{E}_{W,S} [f(W, S)]$.

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

$$H^*(S) = \nabla_w^2 L_E(w, s)_{|w=W^*(S)}$$

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

$$P_{W|S}^\alpha \rightarrow \mathcal{N}(W^*(S), \frac{1}{\alpha} H^*(S)^{-1})$$

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

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

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

$$I(W; S) + L(W; S)$$

$$= \mathbb{E}_{P_{W,S}} [\log P_{W|S}^\alpha] - \mathbb{E}_{P_W \otimes P_S} [\log P_W^\alpha]$$

$$= \mathbb{E}_{P_{W,S}} \left[ -\frac{\alpha}{2} (W - W^*(S))^T H^*(S)(W - W^*(S)) \right]$$

$$+ \mathbb{E}_{P_W \otimes P_S} \left[ \frac{\alpha}{2} (W - W^*(S))^T H^*(S)(W - W^*(S)) \right]$$

$$= \mathbb{E}_{P_W \otimes P_S} \left[ \frac{\alpha}{2} W^T H^*(S) W \right] - \mathbb{E}_{P_{W,S}} \frac{\alpha}{2} W^T H^*(S) W$$

$$+ \mathbb{E}_{P_S \otimes P_W} \left[ \frac{\alpha}{2} \left( \text{tr}(H^*(S)(W^*(S)W^*(S)^T - WW^*(S)^T - W^*(S)W^T)) \right) \right]$$

$$- \mathbb{E}_{P_S \otimes P_{W|S}} \left[ \frac{\alpha}{2} \left( \text{tr}(H^*(S)(W^*(S)W^*(S)^T - WW^*(S)^T - W^*(S)W^T)) \right) \right].$$  \hspace{1cm} (91)
Note that $\mathbb{E}_{\mathcal{P}_u}[W] = \mathbb{E}_{\mathcal{P}_S}[W^*(S)]$ and $\mathbb{E}_{\mathcal{P}_W|S}[W] = W^*(S)$, we have

$$\gen(P_{W|S}^{\infty}, \mu) = \frac{I(W; S) + L(W; S)}{\alpha}$$

$$= \mathbb{E}_{\mathcal{P}_{W\otimes S}} \left[ \frac{1}{2} W^T H^*(S)W \right] - \mathbb{E}_{\mathcal{P}_{W,S}} \left[ \frac{1}{2} W^T H^*(S)W \right]$$

$$+ \mathbb{E}_{\mathcal{P}_S} \left[ \frac{1}{2} \left( \text{tr}(H^*(S)(-\mathbb{E}[W^*(S)]W^*(S)^\top - W^*(S)\mathbb{E}[W^*(S)]^\top)) \right) \right]$$

$$- \mathbb{E}_{\mathcal{P}_S} \left[ \frac{1}{2} \left( \text{tr}(H^*(S)(-W^*(S)W^*(S)^\top - W^*(S)W^*(S)^\top)) \right) \right]$$

$$= \mathbb{E}_{\mathcal{P}_{W\otimes S}} \left[ \frac{1}{2} W^T H^*(S)W \right] - \mathbb{E}_{\mathcal{P}_{W,S}} \left[ \frac{1}{2} W^T H^*(S)W \right]$$

$$+ \mathbb{E}_{\mathcal{P}_S} \left[ (W^*(S) - \mathbb{E}[W^*(S)])^\top (H^*(S)W^*(S) - \mathbb{E}[H^*(S)W^*(S)]) \right].$$

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

$$\gen(P_{W|S}^{\infty}, \mathcal{P}_S) \leq \frac{1}{M} \sum_{u=1}^{M} \left[ \mathbb{E}_{\Delta_{W_u,S}} \left[ \frac{1}{2} W_u^\top H_u^*(S)W_u \right] + \mathbb{E}_{\mathcal{P}_S} \left[ (W_u^*(S) - \mathbb{E}[W_u^*(S)])^\top H_u(W_u^*(S) - \mathbb{E}[W_u^*(S)]) \right] \right].$$

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

$$P_{W|S}^\alpha \rightarrow \sum_{u=1}^{M} \frac{1}{\Sigma_{u=1}^{M} \pi(W_u^*(S))} \pi(W_u^*(S))N(W_u^*(S), \frac{1}{\alpha} H_u^*(S)^{-1}),$$

as long as $H_u^*(S) \triangleq \nabla^2_{\mathcal{P}_u}L_E(w, S) |_{w=W_u^*(S)}$ 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 $P_{W|S}^\alpha$. 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

$$I(S; W|U) = I(S; W|U) + I(S; U) = I(S; W, U) = I(S; W) + I(S; U|W) \geq I(S; W),$$

and for lautum information

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

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,

$$\gen(P_{W|S}^{\infty}, \mu)$$

$$= \lim_{\alpha \rightarrow \infty} \frac{I(S; W) + L(S; W)}{\alpha}$$

$$\leq \lim_{\alpha \rightarrow \infty} \frac{I(S; W|U) + L(S; W|U)}{\alpha}$$

$$= \mathbb{E}_U \left[ \mathbb{E}_{\mathcal{P}_{W|U}\otimes S} \left[ \frac{1}{2} W^T H(w_u^*(S), S)W \right] - \mathbb{E}_U \left[ \mathbb{E}_{\mathcal{P}_{W,S|U}} \left[ \frac{1}{2} W^T H(w_u^*(S), S)W \right] \right] \right]$$

$$+ \mathbb{E}_U \left[ \mathbb{E}_{\mathcal{P}_S} \left[ (w_u^*(S) - \mathbb{E}[w_u^*(S)])^\top (H(w_u^*(S), S)w_u^*(S) - \mathbb{E}[H(w_u^*(S), S)w_u^*(S)]) \right] \right].$$

27
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. $f(z|\mathbf{w}) \neq f(z|\mathbf{w}')$ for $\mathbf{w} \neq \mathbf{w}'$.
2. $\mathcal{W}$ is an open subset of $\mathbb{R}^d$.
3. The function $\log f(z|\mathbf{w})$ is three times continuously differentiable with respect to $\mathbf{w}$.
4. There exist functions $F_1(z) : Z \to \mathbb{R}$, $F_2(z) : Z \to \mathbb{R}$ and $M(z) : Z \to \mathbb{R}$, such that
   \[ \mathbb{E}_{Z \sim f(z|\mathbf{w})} |M(Z)| < \infty, \]
   and the following inequalities hold for any $\mathbf{w} \in \mathcal{W}$,
   \[ \left| \frac{\partial \log f(z|\mathbf{w})}{\partial w_i} \right| < F_1(z), \quad \left| \frac{\partial^2 \log f(z|\mathbf{w})}{\partial w_i \partial w_j} \right| < F_1(z), \]
   \[ \left| \frac{\partial^3 \log f(z|\mathbf{w})}{\partial w_i \partial w_j \partial w_k} \right| < M(z), \quad i, j, k = 1, 2, \ldots, d. \]
5. The following inequality holds for an arbitrary $\mathbf{w} \in \mathcal{W}$,
   \[ 0 < \mathbb{E}_{Z \sim f(z|\mathbf{w})} \left[ \frac{\partial \log f(z|\mathbf{w})}{\partial w_i} \frac{\partial \log f(z|\mathbf{w})}{\partial w_j} \right] < \infty, \quad i, j = 1, 2, \ldots, d. \]

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

\[ P_{\mathcal{W}|S}(\mathbf{w}|z^n) = \frac{\pi(\mathbf{w}) \prod_{i=1}^n f(z_i|\mathbf{w})}{V(z^n)}, \quad \text{with} \quad V(z^n) = \int \pi(\mathbf{w}) \prod_{i=1}^n f(z_i|\mathbf{w}) d\mathbf{w}, \quad (97) \]

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

We further assume that the parametric family $\{f(z_i|\mathbf{w}) \in \mathcal{W}\}$ and prior $\pi(\mathbf{w})$ 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 $\pi(\mathbf{w})$ is continuous and $\pi(\mathbf{w}) > 0$ for all $\mathbf{w} \in \mathcal{W}$.

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 $P_{\mathcal{W}|S}$ in (97) by

\[ \mathcal{N}(\hat{\mathbf{w}}_{\text{ML}}, \frac{1}{n} J(\mathbf{w}^*)^{-1}), \quad \text{where} \quad \hat{\mathbf{w}}_{\text{ML}} \triangleq \arg \max_{\mathbf{w} \in \mathcal{W}} \sum_{i=1}^n \log f(Z_i|\mathbf{w}), \quad (98) \]

denotes the MLE and

\[ J(\mathbf{w}) \triangleq \mathbb{E}_Z [ -\nabla_w^2 \log f(Z|\mathbf{w}) ] \quad \text{with} \quad \mathbf{w}^* \triangleq \arg \min_{\mathbf{w} \in \mathcal{W}} D(P_Z || f(\cdot|\mathbf{w})). \]
The asymptotic Normality of the MLE states that the distribution of $\hat{W}_{ML}$ will converge to
\[
N(\mathbf{w}^*, \frac{1}{n} J(\mathbf{w}^*)^{-1} I(\mathbf{w}^*) J(\mathbf{w}^*)^{-1}) \quad \text{with} \quad I(\mathbf{w}) \triangleq \mathbb{E}_Z [\nabla_{\mathbf{w}} \log f(Z|\mathbf{w}) \nabla_{\mathbf{w}} \log f(Z|\mathbf{w})^\top]
\]
as $n \to \infty$. Thus, the marginal distribution $P_W$ can be approximated by a Gaussian distribution regardless the choice of prior $\pi(\mathbf{w})$.

Then, the symmetrized KL information can be computed using Lemma 4. By Theorem 1, we have
\[
\text{gen}(P_{W|S}, P_Z) = \frac{I_{SKL}(S; W)}{n} = \frac{\text{tr}(I(\mathbf{w}^*) J(\mathbf{w}^*)^{-1})}{n}.
\] (99)

When the true model is in the parametric family $P_Z = f(\cdot|\mathbf{w}^*)$, we have $I(\mathbf{w}^*) = J(\mathbf{w}^*)$, which gives the Fisher information matrix and $\text{gen}(P_{W|S}, P_Z) = \frac{d}{n}$. 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 $\alpha$. To see this, we first note that the derivative of $P_{W|S}$ with respect to $\alpha$ is given by
\[
\frac{dP_{W|S}^\alpha(w|s)}{d\alpha} = P_{W|S}^\alpha(w|s) \left( \mathbb{E}_{P_{W|S}^\alpha} [L_E(w, S)] - L_E(w, S) \right).
\] (100)

Then, we can compute the derivative of the empirical risk with respect to $\alpha$ as follows:
\[
\begin{align*}
\frac{d\mathbb{E}_{P_{W,S}} [L_E(W, S)]}{d\alpha} &= \mathbb{E}_{P_S} \left[ \frac{d\mathbb{E}_{P_{W|S}^\alpha} [L_E(W, S)]}{d\alpha} \right] \\
&= \mathbb{E}_{P_S} \left[ \int_W L_E(w, S) \frac{dP_{W|S}^\alpha(w|S)}{d\alpha} dw \right] \\
&= \mathbb{E}_{P_S} \left[ \int_W P_{W|S}^\alpha(w|s) \left( L_E(w, S) \mathbb{E}_{P_{W|S}^\alpha} [L_E(w, S)] - L_E^2(w, S) \right) dw \right] \\
&= \mathbb{E}_{P_S} \left[ \mathbb{E}_{P_{W|S}^\alpha} [L_E(w, S)] - \mathbb{E}_{P_{W|S}^\alpha} [L_E^2(w, S)] \right] \\
&= -\mathbb{E}_{P_S} \left[ \text{Var}_{P_{W|S}^\alpha} [L_E(W, S)] \right] \leq 0
\end{align*}
\] (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 $\alpha$ 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
\[
\text{gen}(P_{W|S}^\alpha, P_S) = \frac{I_{SKL}(W; S)}{\alpha} - \lambda \mathbb{E}_{W,S} [R(W, S)],
\]
where $\mathbb{E}_{W,S} [R(W, S)] = \mathbb{E}_{P_W \otimes P_S} [R(W, S)] - \mathbb{E}_{P_W,S} [R(W, S)]$.

**Proof.** For $(\alpha, \pi(w), L_E(w, s) + \lambda R(w, s))$-Gibbs algorithm, we have
\[
I_{SKL}(W; S) = \mathbb{E}_{P_{W,S}} [\log(P_{W|S}^\alpha)] - \mathbb{E}_{P_W \otimes P_S} [\log(P_{W|S}^\alpha)] \\
= \alpha \left( \mathbb{E}_{P_W \otimes P_S} [L_E(W, S)] - \mathbb{E}_{P_{W,S}} [L_E(W, S)] \right) \\
+ \alpha \lambda \left( \mathbb{E}_{P_W \otimes P_S} [R(W, S)] - \mathbb{E}_{P_{W,S}} [R(W, S)] \right) \\
= \alpha \text{gen}(P_{W|S}^\alpha, P_S) + \alpha \lambda \mathbb{E}_{W,S} [R(W, S)].
\]

\[\square\]
Proposition 4. (restated) Suppose that we adopt the $\ell_2$-regularizer $R(w, s) = \|w - T(s)\|_2^2$, where $T(\cdot)$ is an arbitrary deterministic function $T : \mathbb{Z}^n \rightarrow \mathcal{W}$. Then, the expected generalization error of $(\alpha, \pi(w), L_E(w, s) + \lambda R(w, s))$-Gibbs algorithm is

$$\tilde{\mathbb{E}}(P_{W|S}^\alpha, P_S) = \frac{I_{\text{SKL}}(W; S)}{\alpha} - \lambda \text{tr}(\text{Cov}[W, T(S)]),$$

where $\text{Cov}[W, T(S)]$ 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 $R(w, s) = \|w - T(s)\|_2^2$,

$$\begin{align*}
    \mathbb{E}_{P_W \otimes P_S}[R(W, S)] &= \mathbb{E}_{P_W \otimes P_S}[\|W - T(S)\|_2^2] - \mathbb{E}_{P_W \otimes P_S}[\|W - T(S)\|_2^2] \\
    &= \mathbb{E}_{P_W \otimes P_S}[W^T T(S)] - \mathbb{E}_{P_W \otimes P_S}[W^T T(S)] \\
    &= \text{tr}(\text{Cov}(W, T(S))) \quad \Box
\end{align*}$$

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:

$$\frac{I_{\text{SKL}}(W; S)}{\alpha} - \lambda \psi^{-1}(I(W; S)) \leq \tilde{\mathbb{E}}(P_{W|S}^\alpha, P_S) \leq \frac{I_{\text{SKL}}(W; S)}{\alpha} + \lambda \psi^{-1}(I(W; S)) \quad (102)$$

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

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

which means that

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

The final results (102) follows directly from (104) and Proposition 3. \hfill \Box

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 $\sigma$-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:

$$\frac{I_{\text{SKL}}(W; S)}{\alpha} - \lambda \sqrt{2\sigma^2} I(W; S) \leq \tilde{\mathbb{E}}(P_{W|S}^\alpha, P_S) \leq \frac{I_{\text{SKL}}(W; S)}{\alpha} + \lambda \sqrt{2\sigma^2} I(W; S) \quad (105)$$

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

By assuming $\sigma$-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 sub-Gaussian under data-generating distribution $P_Z$ for all $w \in \mathcal{W}$. We further assume $C_E = \frac{I(W; S)}{\tilde{I}(W; S)}$ 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:

$$\tilde{\mathbb{E}}(P_{W|S}^\alpha, P_S) \leq \begin{cases} 
\frac{2\sigma^2 \alpha}{(1 + C_E)} \left( \frac{1}{n} - \frac{\lambda}{\sqrt{n}} \right), & \text{if } 0 \leq \lambda \leq \frac{1}{C_E} \\
\frac{2\sigma^2 \alpha}{(1 + C_E)} \left( \frac{1}{n} + \frac{\lambda}{\sqrt{n}} \right), & \text{otherwise.}
\end{cases} \quad (106)$$
Proof. Using Proposition 5 and [71, Theorem 1], we have

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

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

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

Solving (107) gives

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

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

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

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