Towards Optimal Inverse Temperature in the Gibbs Algorithm

Yuheng Bu
Department of ECE, University of Florida, Email: buyuheng@ufl.edu

***Abstract*—This paper explores the problem of selecting optimal hyperparameters in the Gibbs algorithm to minimize the population risk, specifically focusing on the inverse temperature. The inverse temperature is a hyperparameter that controls the tradeoff between data fitting and generalization. We present a characterization of the derivative of the *population risk* with respect to the inverse temperature, expressed in terms of the covariance between empirical risk and test loss. This characterization enables us to identify the optimal inverse temperature that minimizes population risk. Additionally, we provide two illustrative examples—a mean estimate and linear regression—to validate our analytical findings. Notably, our analysis reveals that the optimal inverse temperature exhibits different behaviors in two different regimes based on data quality and prior distribution. These insights contribute to our understanding of linear regression and more general machine learning models.**

I. Introduction

Understanding how a learning algorithm generalizes to unseen data is a fundamental problem in statistical learning theory. Various approaches have been developed including VC dimension-based approach [1], algorithmic stability-based approaches [2], PAC-Bayesian bounding approach [3]. However, these classical techniques are shown to be loose in the context of deep learning [4] because they cannot fully capture all the aspects of a learning problem, including model class, learning algorithm, loss function, and data-generating distribution.

Recently, [5], [6] proposed to use the mutual information between the training data and the learned model weights to bound generalization error, which captures all components of learning problems. Since then, multiple approaches [7]–[18] have been proposed to refine information-theoretic generalization error bounds. However, recent papers [19], [20] revealed inherent limitations in existing information-theoretic bounds, preventing them from achieving optimal rates for some well-studied problems.

In addition, simply bounding the generalization error does not complete the story. As the generalization error is the difference between the population risk and empirical risk, achieving zero generalization error does not guarantee a good population performance if the empirical risk (training loss) is large. Therefore, we need a method to capture the tradeoff between generalization error and empirical risk accurately.

Due to the inherent sophistication of learning problems, accurately characterizing such a tradeoff of *any* learning algorithm using the existing information-theoretic approach is challenging. In this paper, we shift the focus from *arbitrary* learning algorithms to a *specific* one, the Gibbs algorithm. The Gibbs algorithm (formally defined in (6)) can be interpreted as a randomized variant of the standard empirical risk minimization algorithm with regularization. Consequently, it possesses the flexibility to approximate the behavior of many commonly used algorithms in practice. As shown in [21], [22], one benefit of the Gibbs algorithm is that the aforementioned fundamental quantities, i.e., expected generalization error and the empirical risk, can be characterized exactly.

Given such exact characterizations, we explore the problem of how to choose the optimal hyper-parameters in the Gibbs algorithm that minimizes the population risk to guide the practical algorithm design. This paper focuses on the inverse temperature, which controls the tradeoff between fitting the training data and generalization. Note that there have been other works aiming to highlight the Gibbs algorithm (referred to as the power posterior in their context) in terms of its robustness in handling model misspecification [23], [24], which further motivates the problem considered here. Our main contributions to this work are as follows:

- We provide an exact characterization of the derivative for the *population risk* with respect to inverse temperature in terms of the covariance between the empirical risk and test loss, which captures the optimal inverse temperature that minimizes the population risk.
- We further provide two simple examples to validate the analysis, i.e., mean estimate and linear regression. In the mean estimate example, our analysis shows that the optimal inverse temperature has different forms in two regimes depending on the quality of the data and the prior distribution. Such a result provides valuable insights that contribute to our understanding of linear regression and more general machine learning models.

II. Background and Related Works

In this section, we first introduce some background about supervised learning and the Gibbs algorithm.

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 natural ones, and all the information measure units are nats. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

A. Generalization Error in Supervised Learning

Let $S = \{Z_i\}_{i=1}^n \in \mathcal{S}$ be the training set, where each $Z_i = \{X_i, Y_i\}$ is defined on the same alphabet $\mathcal{Z}$. Note that

979-8-3503-8284-6/24/$31.00 ©2024 IEEE

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 space of probability distributions over $\mathcal{W}$ and $\mathcal{S}$ by $\mathcal{P}(\mathcal{W})$ and $\mathcal{P}(\mathcal{S})$, respectively. We also 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 define the empirical and population risks associated with a given hypothesis w via

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

$$L_P(w, P_S) \triangleq \mathbb{E}_{P_S}[L_E(w, S)], \tag{2}$$

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

$$\overline{\text{gen}}(P_{W|S}, P_S) \triangleq \mathbb{E}_{P_{W,S}}[L_P(W, P_S) - L_E(W, S)], \tag{3}$$

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

As P_S is unknown in practice, we need to estimate the population risk using the test loss, which is the empirical risk evaluated on an independent test dataset $S' = \{Z_i'\}_{i=1}^n \sim P_S$,

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

where we assume S' also contains n samples for simplicity. Thus, we can also define the following expected generalization error on fixed training and testing data as

$$\overline{\text{gen}}(P_{W|S}, s, s') \triangleq \mathbb{E}_{P_{W|S=s}}[L_E(W, s') - L_E(W, s)], \tag{5}$$

where the randomness only comes from the learning algorithm. If we further take expectation with respect to S and S', it reduces back to the expected generalization error in (3).

B. Motivation of Gibbs Algorithm

In this paper, we focus on the Gibbs algorithm (or Gibbs posterior [25]), first proposed by [26] in statistical mechanics and further investigated by [27] in information theory.

The Gibbs algorithm arises when conditional KL-divergence is used as a regularizer to penalize over-fitting in the empirical risk minimization (ERM) algorithm [22]. It was first shown in [5], [6] that under the sub-Gaussian assumption, the expected generalization error of *any* learning algorithm $P_{W|S}$ could be bounded using the mutual information between the input training data S and the learned model weights W. This result motivates regularizing the mutual information in the standard ERM to improve the generalization performance.

As computing $I(S; W)$ requires the knowledge of P_W, [6], [28] propose the following *information risk minimization* (IRM) problem, which replaces $I(S; W)$ with an upper bound $D(P_{W|S} \| \pi | P_S)$, i.e.,

$$P_{W|S}^{\gamma} \triangleq \arg\min_{P_{W|S}} \Big(\mathbb{E}_{P_{W,S}}[L_E(W, S)] + \frac{1}{\gamma} D(P_{W|S} \| \pi | P_S) \Big).$$

Here, $\pi \in \mathcal{P}(\mathcal{W})$ is an arbitrary prior distribution, and the inverse temperature $\gamma \geq 0$ controls the regularization term and balances between data fitting and generalization. When $\gamma \to \infty$, IRM reduces back to the standard ERM algorithm; when $\gamma \to 0$, the learning algorithm ignores the training data, and $P_{W|S}(w|s) = \pi(w)$.

It is shown in [6], [28], the solution of this information risk minimization is the following $(\gamma, \pi(w), L_E(w, s))$-Gibbs algorithm, which is defined as:

$$P_{W|S}^{\gamma}(w|s) = \frac{\pi(w) e^{-\gamma L_E(w,s)}}{V(\gamma, s)}, \tag{6}$$

where

$$V(\gamma, s) \triangleq \int \pi(w) e^{-\gamma L_E(w,s)} dw \tag{7}$$

is the partition function.

In practice, one way to implement the Gibbs algorithm is to use the Stochastic Gradient Langevin Dynamics (SGLD) algorithm [29], which can be viewed as the discrete version of the continuous-time Langevin diffusion, or the noisy variant of Stochastic Gradient Descent (SGD),

$$W_{k+1} = W_k - \eta \nabla_w L_E(W_k, S) + \sqrt{\frac{2\eta}{\gamma}} \zeta_k, \tag{8}$$

for $k = 0, 1, \cdots$, where ζ_k is a standard Gaussian vector, $\eta > 0$ is the step size and γ controls the variance of the noise. In [30], it is proved that under certain conditions, the conditional distribution $P_{W_k|S}$ induced by the SGLD algorithm is close to the Gibbs distribution in the Wasserstein distance for sufficiently large k.

Another approach to sample from the Gibbs posterior is the Metropolis-adjusted Langevin algorithm (MALA) [31]. MALA and SGLD are first-order sampling methods since they have similar gradient update formulas as in (8), guaranteeing that both algorithms converge to the Gibbs distribution. MALA differs from the SGLD by introducing an additional Metropolis-adjusted step, which provides a faster convergence rate, as shown in [31]–[33].

C. Existing Results

In this subsection, we review some existing results on the information-theoretic characterization of the generalization error and empirical risk for the Gibbs algorithm.

1) Generalization error of Gibbs algorithm: It has been shown in [21], [34] that the expected generalization error for the Gibbs algorithm can be characterized using the symmetrized Kullback-Leibler information, i.e.,

$$\overline{\text{gen}}(P_{W|S}^{\gamma}, P_S) = \frac{I_{\text{SKL}}(W; S)}{\gamma}, \quad \text{with} \tag{9}$$

$$I_{\text{SKL}}(W; S) \triangleq D(P_{W,S} \| P_W \otimes P_S) + D(P_W \otimes P_S \| P_{W,S}).$$

Similar to mutual information $I(W;S)$ and KL divergence, $I_{\mathrm{SKL}}(W;S)$ is induced from the symmetrized KL divergence (or Jeffrey's divergence [27]), which is also an f-divergence [35]. Such a result highlights the role the symmetrized KL information plays in learning theory, and it holds for *non-i.i.d* training samples S with finite n.

However, generalization error is just a part of the story, as setting $\gamma \to 0$ will lead to zero generalization error, but the algorithm ignores the training data. Thus, to understand the tradeoff between generalization and data fitting, we need to characterize the empirical risk of the Gibbs algorithm.

2) Empirical risk of Gibbs algorithm: We notice that the partition function $V(\gamma, s)$ defined in (7) can be viewed as the moment-generating function of the empirical risk $-L_E(W,s)$ under $\pi(W)$. Therefore, the first and second-order moments of $L_E(W,s)$ can be obtained using the derivative of the cumulant generating function, as shown in the following lemma.

Lemma 1 ([22, Lemma 17]): The log-partition function $\log V(\gamma, s)$ is convex and differentiable infinitely many times with respect to γ in the interior of $\{\gamma \geq 0 : \log V(\gamma, s) < \infty\}$. In particular, the first and second derivatives satisfy

$$\frac{\partial \log V(\gamma,s)}{\partial \gamma} = -\mathbb{E}_\gamma[L_E(W,s)], \tag{10}$$

$$\frac{\partial^2 \log V(\gamma,s)}{\partial \gamma^2} = \mathrm{Var}_\gamma[L_E(W,s)], \tag{11}$$

where $\mathbb{E}_\gamma[\,\cdot\,] \triangleq \mathbb{E}_{P^\gamma_{W|S=s}}[\,\cdot\,]$, and

$$\mathrm{Var}_\gamma[L_E(W,s)] \triangleq \mathbb{E}_\gamma[L_E(W,s)^2] - \mathbb{E}_\gamma[L_E(W,s)]^2 \tag{12}$$

denote the expectation and variance under the Gibbs algorithm $P^\gamma_{W|S=s}$, respectively.

From Lemma 1, we can conclude that the expected empirical risk of the Gibbs algorithm is a non-increasing function of the inverse temperature γ. To see this, note that

$$\frac{\partial}{\partial \gamma}\mathbb{E}_\gamma[L_E(W,s)] = -\mathrm{Var}_\gamma[L_E(W,s)] \leq 0.$$

To understand the tradeoff between empirical risk and generalization error of the Gibbs algorithm, in the following, we will mainly focus on the population risk and generalization error of the Gibbs algorithm. Moreover, we adopt the following notations for the expected test loss

$$L_P(\gamma, s, s') \triangleq \mathbb{E}_\gamma[L_E(W,s')], \tag{13}$$

and expected generalization error

$$\overline{\mathrm{gen}}(\gamma, s, s') \triangleq \mathbb{E}_\gamma[L_E(W,s') - L_E(W,s)]. \tag{14}$$

on fixed training data s and testing data s' for the Gibbs algorithm, where the expectation is over the distribution $P^\gamma_{W|S=s}$.

III. Main results

In this section, we aim to understand the monotonicity of the population risk and the generalization error for the Gibbs algorithm as we increase the inverse temperature γ. Such an analysis could guide us in selecting the optimal inverse temperature γ that minimizes the population risk when we use the Gibbs algorithm in practice.

A. Expected Test Loss

The following theorem captures the first-order derivative of the expected test loss $L_P(\gamma, s, s')$.

Theorem 1: For $\gamma \in \{\gamma \geq 0 : \log V(\gamma, s) < \infty\}$, the first order-derivative of the expected test loss for fixed training and testing data with respect to γ is given by

$$\frac{\partial}{\partial \gamma} L_P(\gamma, s, s') = -\mathrm{Cov}_\gamma[L_E(W,s'), L_E(W,s)], \tag{15}$$

where

$$\begin{aligned}&\mathrm{Cov}_\gamma[L_E(W,s'), L_E(W,s)] \\ &\triangleq \mathbb{E}_\gamma[L_E(W,s)L_E(W,s')] - \mathbb{E}_\gamma[L_E(W,s)]\mathbb{E}_\gamma[L_E(W,s')].\end{aligned} \tag{16}$$

Proof: As Gibbs posterior is differentiable with respect to γ within the set $\{\gamma \geq 0 : \log V(\gamma, s) < \infty\}$, we can swap the order of partial derivative and integral, then

$$\begin{aligned}\frac{\partial}{\partial \gamma} L_P(\gamma, s, s') &= \frac{\partial}{\partial \gamma}\int L_E(w,s')\frac{\pi(w)e^{-\gamma L_E(w,s)}}{V(\gamma,s)}dw \\ &= \int L_E(w,s')\pi(w)\frac{\partial}{\partial \gamma}\frac{e^{-\gamma L_E(w,s)}}{V(\gamma,s)}dw \\ &\overset{(a)}{=} \int L_E(w,s')\pi(w)\frac{e^{-\gamma L_E(w,s)}}{V(\gamma,s)} \\ &\qquad \cdot \big(\mathbb{E}_\gamma[L_E(W,s)] - L_E(w,s)\big)dw \\ &= \mathbb{E}_\gamma\big[L_E(W,s')\big(\mathbb{E}_\gamma[L_E(W,s)] - L_E(W,s)\big)\big] \\ &= -\mathrm{Cov}_\gamma[L_E(W,s'), L_E(W,s)].\end{aligned} \tag{17}$$

Here, equality (a) follows from the fact that

$$\begin{aligned}&\frac{\partial}{\partial \gamma}\frac{e^{-\gamma L_E(w,s)}}{V(\gamma,s)} \\ &= \frac{-e^{-\gamma L_E(w,s)}L_E(w,s)V(\gamma,s) - V'(\gamma,s)e^{-\gamma L_E(w,s)}}{V^2(\gamma,s)} \\ &= \frac{e^{-\gamma L_E(w,s)}}{V(\gamma,s)}\Big(-\frac{V'(\gamma,s)}{V(\gamma,s)} - L_E(w,s)\Big) \\ &= \frac{e^{-\gamma L_E(w,s)}}{V(\gamma,s)}\Big(-\frac{\partial}{\partial \gamma}\log V(\gamma,s) - L_E(w,s)\Big).\end{aligned} \tag{18}$$

Then, the result follows by Lemma 1, which shows that the first-order derivative of $-\log V(\gamma, s)$ is $\mathbb{E}_\gamma[L_E(W,s)]$. ∎

Unlike the expected empirical risk, the covariance term $\mathrm{Cov}_\gamma[L_E(W,s'), L_E(W,s)]$ can be either positive or negative, and it is hard to conclude the monotonicity of the expected test loss $L_P(\gamma, s, s')$. Thus, to find the optimal inverse temperature γ^* that minimizes the population risk, we also need to check the second-order derivative of $L_P(\gamma^*, s, s')$ under the first-order condition $\mathrm{Cov}_{\gamma^*}[L_E(W,s'), L_E(W,s)] = 0$.

We have the following theorem that characterizes the second-order derivative of the expected test loss, given that the first-order condition is satisfied. The proof is omitted due to the space limit.

Theorem 2: For $\gamma \in \{\gamma \geq 0 : \log V(\gamma, s) < \infty\}$, under the condition that $\mathrm{Cov}_{\gamma^*}[L_E(W,s'), L_E(W,s)] = 0$, the second

order-derivative of the expected test loss for fixed training and testing data with respect to γ is given by

$$\frac{\partial^2}{\partial\gamma^2}L_P(\gamma,s,s')\Big|_{\gamma=\gamma^*}=\mathrm{Cov}_{\gamma^*}[L_E(W,s'),L_E(W,s)^2].$$

From Theorem 1 and 2, a condition of the optimal inverse temperature γ^* that achieves the local minimum of $L_P(\gamma,s,s')$ is given by,

$$\mathrm{Cov}_{\gamma^*}[L_E(W,s'),L_E(W,s)]=0, \tag{19}$$
$$\mathrm{Cov}_{\gamma^*}[L_E(W,s'),L_E(W,s)^2]>0. \tag{20}$$

In addition, it is easy to see that by taking the expectation over both P_S and $P_{S'}$, the optimal inverse temperature γ^* that achieves the local minimum of $L_P(P^\gamma_{W|S},P_S)$ should satisfy the following two conditions,

$$\mathbb{E}_{P_SP_{S'}}\big[\mathrm{Cov}_{\gamma^*}[L_E(W,S'),L_E(W,S)]\big]=0, \tag{21}$$
$$\mathbb{E}_{P_SP_{S'}}\big[\mathrm{Cov}_{\gamma^*}[L_E(W,s'),L_E(W,s)^2]\big]>0. \tag{22}$$

We note that the optimal γ^* given by the aforementioned conditions may not exist. As shown in our example in Section IV-A, it implies that the optimal $\gamma^*\to\infty$.

B. Expected Generalization Error

As generalization error is the difference between population risk and empirical risk, we obtain the following corollary that characterizes the derivative of the expected generalization error with respect to γ by combining our Theorem 1 with Lemma 1.

Corollary 1: For $\gamma\in\{\gamma\geq 0:\log V(\gamma,s)<\infty\}$, the first order-derivative of the expected generalization error for fixed training and testing data with respect to γ is given by

$$\frac{\partial}{\partial\gamma}\overline{\mathrm{gen}}(\gamma,s,s')$$
$$=\mathrm{Var}_\gamma(L_E(W,s))-\mathrm{Cov}_\gamma[L_E(W,s'),L_E(W,s)].$$

In addition, we have

$$\frac{\partial}{\partial\gamma}\overline{\mathrm{gen}}(P^\gamma_{W|S},P_S)$$
$$=\mathbb{E}_{P_SP_{S'}}\big[\mathrm{Var}_\gamma(L_E(W,S))-\mathrm{Cov}_\gamma[L_E(W,S'),L_E(W,S)]\big].$$

Unlike the empirical risk, which is always non-increasing with respect to γ, we cannot show that the generalization error is non-decreasing, i.e., $\frac{\partial}{\partial\gamma}\overline{\mathrm{gen}}(P_{W|S},P_S)\geq 0$. This is because Cauchy-Schwarz Inequality only guarantees that

$$\big|\mathrm{Cov}_\gamma[L_E(W,s'),L_E(W,s)]\big|$$
$$\leq\sqrt{\mathrm{Var}_\gamma(L_E(W,s))\mathrm{Var}_\gamma(L_E(W,s'))}. \tag{23}$$

However, we cannot compare $\mathrm{Var}_\gamma(L_E(W,s))$ and $\mathrm{Var}_\gamma(L_E(W,s'))$ in general cases.

It is easy to see that the generalization error reaches zero as $\gamma\to 0$. Moreover, [21] provides a bound of order $\mathcal{O}\big(\frac{\gamma}{n}\big)$ by simply combining the I_{SKL} characterization with the mutual information-based generalization error bound, which may hint that generalization error is always increasing with γ. However, in Section IV-B, we will illustrate an example to elucidate how the generalization error initially rises from zero and subsequently decreases as γ increases.

IV. Illustrative Examples

In this section, we provide two illustrative examples to deepen our understanding of the theoretical results for optimal inverse temperature: the mean estimation and linear regression.

A. Mean Estimation Example

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 $S=\{\boldsymbol{z}_i\}_{i=1}^n$. We do not assume Gaussian distribution for the data, but the covariance matrix of Z satisfies $\Sigma_Z=\sigma_Z^2\boldsymbol{I}_d$ 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 $\pi(\boldsymbol{w})=\mathcal{N}(\boldsymbol{\mu}_0,\sigma_0^2\boldsymbol{I}_d)$. Then the $(\gamma,\mathcal{N}(\boldsymbol{\mu}_0,\sigma_0^2\boldsymbol{I}_d),L_E(\boldsymbol{w},s))$-Gibbs algorithm is given by the following Gaussian posterior distribution as shown in [21],

$$P^\gamma_{W|S}(\boldsymbol{w}|\boldsymbol{z}^n)\sim\mathcal{N}\Big(\alpha\boldsymbol{\mu}_0+(1-\alpha)\bar{\boldsymbol{z}},\alpha\sigma_0^2\boldsymbol{I}_d\Big), \tag{24a}$$

with

$$\alpha\triangleq\frac{1}{2\sigma_0^2\gamma+1},\quad \bar{\boldsymbol{z}}\triangleq\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}_i. \tag{24b}$$

One way to find the optimal γ^* is to evaluate the population risk by decomposing it into generalization error and empirical risk. The generalization error can be obtained by computing $I_{\mathrm{SKL}}(W;S)$ for Gaussian channel directly, which has the decay rate of $\mathcal{O}(1/n)$ as shown in [21]. By lemma 1, we can also compute the expected empirical risk using $\log V(\gamma,s)$, which follows the non-central chi-squared distribution under P_S. Thus, we obtain the following exact characterization of the population risk, i.e.,

$$L_P(P^\gamma_{W|S},P_S) \tag{25}$$
$$=\underbrace{\frac{4d\sigma_0^2\sigma_Z^2\gamma}{n(1+2\sigma_0^2\gamma)}}_{\text{generalization error}}+\underbrace{\frac{\|\boldsymbol{\mu}_0-\boldsymbol{\mu}\|_2^2+d\sigma_z^2/n}{(1+2\sigma_0^2\gamma)^2}+\frac{d\sigma_0^2}{1+2\sigma_0^2\gamma}+\frac{n-1}{n}d\sigma_Z^2}_{\text{empirical risk}}.$$

Then, optimizing over γ will identify the optimal inverse temperature.

Another approach is to directly evaluate the derivative of $L_P(\gamma,s,s')$ by computing the covariance term in Theorem 1. In particular, for an independent test dataset $S'=\{\boldsymbol{z}_i'\}_{i=1}^n$, we can show that

$$\frac{\partial}{\partial\gamma}L_P(\gamma,s,s')$$
$$=-\mathrm{Cov}_\gamma[L_E(W,s'),L_E(W,s)]$$
$$=\frac{-2\sigma_0^2}{(2\gamma\sigma_0^2+1)^3}\Big(2(\|\boldsymbol{\mu}_0\|^2-\boldsymbol{\mu}_0^\top(\bar{\boldsymbol{z}}+\bar{\boldsymbol{z}}')+\bar{\boldsymbol{z}}'^\top\bar{\boldsymbol{z}})$$
$$+4\gamma\sigma_0^2\boldsymbol{\mu}_0^\top(\bar{\boldsymbol{z}}-\bar{\boldsymbol{z}}')-4\gamma\sigma_0^2\|\bar{\boldsymbol{z}}\|^2+4\gamma\sigma_0^2\bar{\boldsymbol{z}}^\top\bar{\boldsymbol{z}}'$$
$$+2d\gamma\sigma_0^4+d\sigma_0^2\Big), \tag{26}$$

which only depends on S and S' through the average across the samples. Taking expectations over the training and test data set, we obtain

$$\frac{\partial}{\partial\gamma}L_P(P^{\gamma}_{W|S},P_S) \tag{27}$$
$$=\frac{2\sigma_0^2}{(2\gamma\sigma_0^2+1)^3}\Big(2d\sigma_0^2(2\frac{\sigma_Z^2}{n}-\sigma_0^2)\gamma-2\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2-d\sigma_0^2\Big),$$

which is equivalent to taking derivative with γ in (25).

Thus, the optimal inverse temperature of γ that minimizes the population risk depends on other parameters of the problem in a non-trivial manner, i.e.,

$$\gamma^*=\begin{cases}+\infty, & \text{if } \frac{\sigma_Z^2}{n}\in[0,\frac{\sigma_0^2}{2}),\ \text{(high-SNR)}\\ \frac{\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2+d\sigma_0^2/2}{d(2\sigma_Z^2/n-\sigma_0^2)\sigma_0^2}, & \text{if } \frac{\sigma_Z^2}{n}\in[\frac{\sigma_0^2}{2},\infty).\ \text{(low-SNR)}\end{cases} \tag{28}$$

Here, the term $\frac{\sigma_Z^2}{n}$ only depends on the training data S, which can be interpreted as the normalized noise of the samples, and σ_0^2 and $\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2$ captures the confidence and the bias of the prior knowledge. Therefore, we can connect the two cases here with the concept of signal-to-noise ratio (SNR).

The optimal Gibbs algorithm has different forms in the following two regimes, i.e., 1) high-SNR regime, where the quality of training samples surpasses prior knowledge, the optimal algorithm involves discarding the prior distribution and employing the standard ERM algorithm; 2) low-SNR regime, where we should incorporate knowledge from both the training samples and prior distribution, and the optimal γ depends on $\frac{\sigma_Z^2}{n}$, σ_0^2 and $\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|^2$. In the extreme case where $\boldsymbol{\mu}_0=\boldsymbol{\mu}$ and $\sigma_0^2=0$, the optimal $\gamma^*=0$, indicating that we should discard the samples and solely rely on the prior knowledge.

Although the above result is obtained within the mean estimation example, the optimal inverse temperature in two different regimes provides insights to help us understand more general machine learning models.

B. Simulation for Linear Regression

In this example, we consider a simple linear regression problem, where the training data $S=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n$, with $\mathcal{X}=\mathbb{R}^d$ and $\mathcal{Y}=\mathbb{R}$. Specifically, the data is generated using the true weights $W^*\in\mathbb{R}^d$ with additive noise, i.e.,

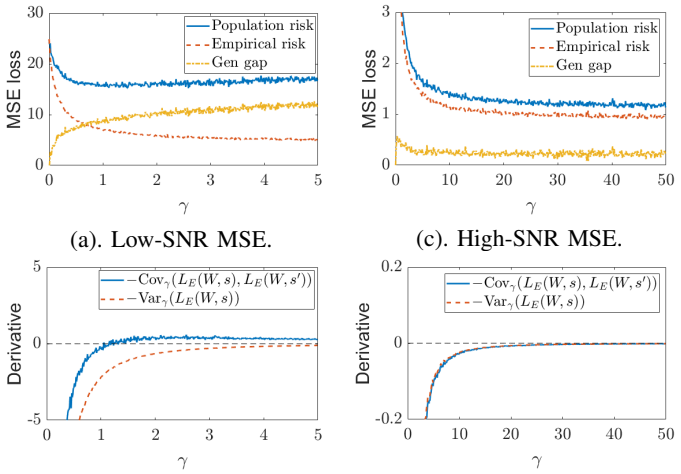
$$Y_i=X_i\cdot W^*+\varepsilon_i,\quad \varepsilon\sim\mathcal{N}(0,\sigma_\varepsilon^2). \tag{29}$$

We adopt the mean-squared loss $\ell(\boldsymbol{w},\boldsymbol{z})=(y-\boldsymbol{x}\cdot\boldsymbol{w})^2$, and assume a zero-mean Gaussian prior for the weights $\pi(\boldsymbol{w})=\mathcal{N}(0,\sigma_0^2\boldsymbol{I}_d)$. Then the $(\gamma,\mathcal{N}(0,\sigma_0^2\boldsymbol{I}_d),L_E(\boldsymbol{w},s))$-Gibbs algorithm is given by the following Gaussian posterior distribution

$$P^{\gamma}_{W|S}(\boldsymbol{w}|S)\sim\mathcal{N}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top\boldsymbol{Y},\frac{n}{2\gamma}\boldsymbol{\Sigma}^{-1}\Big), \tag{30}$$

with $\boldsymbol{\Sigma}\triangleq\boldsymbol{X}^\top\boldsymbol{X}+\frac{n}{2\sigma_0^2\gamma}\boldsymbol{I}_d$, and $\boldsymbol{X}\in\mathbb{R}^{n\times d},\boldsymbol{Y}\in\mathbb{R}^n$ are the matrix form of the training data.

To avoid the computation involving multivariate non-central chi-squared distribution, we directly perform simulation to verify our theoretical results. Specifically, we use the following parameters: $\boldsymbol{\Sigma}_X=\boldsymbol{I}_d$, $d=5$, $\sigma_0^2=2$. Motivated by (28), we further consider the following two cases: 1) low SNR regime, where the number of training samples $n=10$, and $\sigma_\varepsilon^2=3$; and 2) high SNR regime, with $n=100$, and $\sigma_\varepsilon^2=1$.



(a). Low-SNR MSE. (c). High-SNR MSE.

(b). Low-SNR Covariance. (d). High-SNR Covariance.

Fig. 1. Simulation results for the linear regression example.

We plot the average empirical risk, population risk, and generalization error over 500 runs for different values of γ on Figure 1 (top) using the Gaussian Gibbs posterior in (30). We further estimate the covariance between the test loss and training loss and the variance for the training loss, respectively, and plot them in the bottom of Figure 1.

As shown in Figure 1 (a) and (b), in this low SNR regime, the optimal γ^* that minimizes the population risk is roughly at 1, which is also reflected by the zero point of the estimated covariance. In this case, the generalization error is increasing as we increase the inverse temperature γ.

However, for the high SNR regime in Figure 1 (c) and (d), the population risk is always decreasing, and the estimated covariance converges to zero but never reaches zero in the considered range of γ, implying that the optimal $\gamma^*\to\infty$. More interestingly, the generalization error in this case is not monotone. It first jumps to 0.4 from zero and slowly decreases to 0.2, which further demonstrates the looseness of the previous $\mathcal{O}\big(\frac{\gamma}{n}\big)$ bound in [21].

V. Conclusion

Our investigation into the exact characterization of the Gibbs algorithm in this paper has paved the way for addressing the problem of selecting optimal hyperparameters, specifically focusing on the inverse temperature, to minimize population risk. We hope the optimal inverse temperature in high-SNR and low-SNR regimes provides insights to inform practical algorithm design. While we have concentrated on the inverse temperature in this paper, it is important to acknowledge that other design choices in the Gibbs algorithm, such as the model family, loss function, and the prior distribution π, remain areas of future exploration and research.

REFERENCES

[1] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

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

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

[4] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations*, 2016.

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

[6] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

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

[8] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via wasserstein distance," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 577–581.

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

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

[11] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, "Conditioning and processing: Techniques to improve information-theoretic generalization bounds," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 457–16 467, 2020.

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

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

[14] M. Haghifam, G. K. Dziugaite, S. Moran, and D. Roy, "Towards a unified information-theoretic framework for generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 370–26 381, 2021.

[15] H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan, "Information-theoretic generalization bounds for black-box learning algorithms," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 670–24 682, 2021.

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

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

[18] Y. Chu and M. Raginsky, "A unified framework for information-theoretic generalization bounds," *arXiv preprint arXiv:2305.11042*, 2023.

[19] M. Haghifam, B. Rodríguez-Gálvez, R. Thobaben, M. Skoglund, D. M. Roy, and G. K. Dziugaite, "Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization," in *International Conference on Algorithmic Learning Theory*. PMLR, 2023, pp. 663–706.

[20] R. Livni, "Information theoretic lower bounds for information theoretic upper bounds," *arXiv preprint arXiv:2302.04925*, 2023.

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

[22] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, 2024.

[23] M. A. Medina, J. L. M. Olea, C. Rush, and A. Velez, "On the robustness to misspecification of α-posteriors and their variational approximations," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 6579–6629, 2022.

[24] R. Ray, M. A. Medina, and C. Rush, "Asymptotics for power posterior mean estimation," in *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2023, pp. 1–8.

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

[26] J. W. Gibbs, "Elementary principles of statistical mechanics," *Compare*, vol. 289, p. 314, 1902.

[27] E. T. Jaynes, "Information theory and statistical mechanics," *Physical review*, vol. 106, no. 4, p. 620, 1957.

[28] T. Zhang, "Information-theoretic upper and lower bounds for statistical estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[29] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 681–688.

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

[31] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu, "Log-concave sampling: Metropolis-hastings algorithms are fast!" in *Conference on learning theory*. PMLR, 2018, pp. 793–797.

[32] O. Mangoubi and N. K. Vishnoi, "Nonconvex sampling with the metropolis-adjusted langevin algorithm," in *Conference on Learning Theory*. PMLR, 2019, pp. 2259–2293.

[33] D. Holzmüller and F. Bach, "Convergence rates for non-log-concave sampling and log-partition estimation," *arXiv preprint arXiv:2303.03237*, 2023.

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

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