On the Validation of Gibbs Algorithms: Training Datasets, Test Datasets and their Aggregation

Samir M. Perlaza, Iñaki Esnaola, Gaetan Bisson, H Vincent Poor

On the Validation of Gibbs Algorithms: Training Datasets, Test Datasets and their Aggregation

Samir M. Perlaza*†‡, Iñaki Esnaola†§, Gaetan Bisson‡, and H. Vincent Poor†
*INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France.
†ECE Dept. Princeton University, Princeton, 08544 NJ, USA.
‡GAATI, Université de la Polynésie Française, Faaa, French Polynesia.
§ACSE Dept., University of Sheffield, Sheffield, United Kingdom.

***Abstract*—The dependence on training data of the Gibbs algorithm (GA) is analytically characterized. By adopting the expected empirical risk as the performance metric, the sensitivity of the GA is obtained in closed form. In this case, sensitivity is the performance difference with respect to an arbitrary alternative algorithm. This description enables the development of explicit expressions involving the training errors and test errors of GAs trained with different datasets. Using these tools, dataset aggregation is studied and different figures of merit to evaluate the generalization capabilities of GAs are introduced. For particular sizes of such datasets and parameters of the GAs, a connection between Jeffrey's divergence, training and test errors is established.**

I. Introduction

The Gibbs algorithm (GA) randomly selects a model by sampling the Gibbs probability measure, which is the unique solution to the empirical risk minimization (ERM) problem with relative entropy regularization (ERM-RER) [1]. The input of the GA is twofold. It requires a number of labeled patterns (datasets); and a prior on the set of models in the form of a σ-measure, e.g., the Lebesgue measure, the counting measure, or a probability measure. One of the main features of the GA is that it does not require an assumption on the statistical properties of the datasets [2]–[4]. Nonetheless, the generalization capabilities of the Gibbs algorithm are often characterized by the generalization error, for which statistical assumptions on the datasets must be considered, e.g., training, and unseen datasets are identically distributed. When the prior on the set of models is a probability measure, a closed-form expression for the generalization error is presented in [5], while upper bounds have been derived in [6]–[27], and references therein.

In a more general setting, when the prior on the set of models is a σ-measure, the generalization capabilities of the GA have been studied in [1], [28], and [29], using the sensitivity of the empirical risk to deviations from the Gibbs probability measure to another probability measure. This method does not require any statistical assumptions on the datasets and is chosen as the workhorse of the present analysis.

This work is supported by the Inria Exploratory Action – Information and Decision Making (AEx IDEM) and in part by a grant from the C3.ai Digital Transformation Institute.

The main motivation of this work is to break away from the implicit assumption in existing literature that all training datasets are drawn from the same probability measure and thus, can be aggregated to improve the generalization capabilities of a given GA. In practical settings, training data might be acquired from multiple sources that might be subject to different impairments during data acquisition, data storage and data transmission. For instance, consider a GA trained upon a particular dataset and assume that a new dataset from a different source is made available. Hence, the following questions arise concerning the generalization capabilities of such a GA: Would such a GA generalize over the new dataset? Should the new dataset be aggregated to the previous dataset to build a new GA in the aim of improving generalization? How does the GA trained upon the existing dataset compare in terms of generalization with respect to a new GA trained upon the new dataset? The answers to such questions are far from trivial. One of the main challenges to answer such questions stems from the fact that the probability measures generating each of those datasets are unknown and potentially different due to a variety of impairments.

This paper introduces a closed-form expression for the difference of the expected empirical risk on a given dataset induced by a GA trained upon this dataset and the one induced by an alternative algorithm (another probability measure). This quantity was coined *sensitivity of the GA algorithm* in [28] and is shown to be central to tackling the questions above. This is in part due to the fact that it allows studying the generalization capabilities of GAs based on actual datasets, which disengages from the assumption that both training and unseen data follow the same probability distribution. More specifically, by studying the sensitivity, closed-form expressions for the difference between training error and test error can be obtained. These expressions lead to a clearer understanding of the roles of the size of datasets chosen for training and testing, as well as the parameters of the GAs. As a byproduct, the difference between the expected empirical risk on the aggregation of two datasets induced by two GAs trained upon the constituent datasets is characterized. Similarly, the difference between the expected empirical risk on one of the constituent datasets induced by two GAs trained upon the aggregated dataset and

the constituent dataset is also characterized. These explicit expressions allow comparing two GAs trained upon different datasets, which is relevant under learning paradigms such as federated learning [30].

II. Problem Formulation

Let $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$, with $\mathcal{M} \subseteq \mathbb{R}^d$ and $d \in \mathbb{N}$, be sets of *models*, *patterns*, and *labels*, respectively. A pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is referred to as a *labeled pattern* or as a *data point*. Given n data points, with $n \in \mathbb{N}$, denoted by (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), a dataset is represented by the tuple $\big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n$.

Let the function $f : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ be such that the label y assigned to the pattern x according to the model $\boldsymbol{\theta} \in \mathcal{M}$ is

$$y = f(\boldsymbol{\theta}, x). \tag{1}$$

Let also the function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty] \tag{2}$$

be such that given a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the risk induced by a model $\boldsymbol{\theta} \in \mathcal{M}$ is $\ell(f(\boldsymbol{\theta}, x), y)$. In the following, the risk function ℓ is assumed to be nonnegative and for all $y \in \mathcal{Y}$, $\ell(y, y) = 0$.

Given a dataset

$$\boldsymbol{z} = \big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n, \tag{3}$$

the *empirical risk* induced by the model $\boldsymbol{\theta}$, with respect to the dataset $\boldsymbol{z}$ in (3), is determined by the function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to [0, +\infty]$, which satisfies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{\theta}, x_i), y_i). \tag{4}$$

Using this notation, the ERM problem consists of the following optimization problem:

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{5}$$

Let the set of solutions to the ERM problem in (5) be denoted by $\mathcal{T}(\boldsymbol{z}) \triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})$. Note that if the set $\mathcal{M}$ is finite, the ERM problem in (5) always possesses a solution, and thus, $|\mathcal{T}(\boldsymbol{z})| > 0$. Nonetheless, in general, the ERM problem might not necessarily possess a solution. Hence, for some cases, it might be observed that $|\mathcal{T}(\boldsymbol{z})| = 0$.

A. Notation

The *relative entropy* is defined below as the extension to σ-finite measures of the relative entropy usually defined for probability measures.

Definition 1 (Relative Entropy): Given two σ-finite measures P and Q on the same measurable space, such that Q is absolutely continuous with respect to P, the relative entropy of Q with respect to P is

$$D(Q\|P) \triangleq \int \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \log\Big(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\Big) \mathrm{d}P(x), \tag{6}$$

where the function $\frac{\mathrm{d}Q}{\mathrm{d}P}$ is the Radon-Nikodym derivative of Q with respect to P.

Given a measurable space $(\Omega, \mathscr{F})$, the set of all σ-finite measures on $(\Omega, \mathscr{F})$ is denoted by $\triangle(\Omega, \mathscr{F})$. Given a σ-measure $Q \in \triangle(\Omega, \mathscr{F})$, the subset of $\triangle(\Omega, \mathscr{F})$ including all σ-finite measures absolutely continuous with Q is denoted by $\triangle_Q(\Omega, \mathscr{F})$. Given a subset $\mathcal{A}$ of $\mathbb{R}^d$, the Borel σ-field on $\mathcal{A}$ is denoted by $\mathscr{B}(\mathcal{A})$.

B. The ERM-RER Problem

The *expected empirical risk* is defined as follows.

Definition 2 (Expected Empirical Risk): Let P be a probability measure in $\Delta(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. The expected empirical risk with respect to the dataset $\boldsymbol{z}$ in (3) induced by the measure P is

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P(\boldsymbol{\theta}), \tag{7}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is in (4).

The following lemma follows immediately from the properties of the Lebesgue integral.

Lemma 1: Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ and two probability measures P_1 and P_2 over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, for all $\alpha \in [0, 1]$, the function $\mathsf{R}_{\boldsymbol{z}}$ in (7) satisfies

$$\mathsf{R}_{\boldsymbol{z}}(\alpha P_1 + (1-\alpha)P_2) = \alpha \mathsf{R}_{\boldsymbol{z}}(P_1) + (1-\alpha)\mathsf{R}_{\boldsymbol{z}}(P_2). \tag{8}$$

The ERM-RER problem is parametrized by a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a positive real, which are referred to as the *reference measure* and the *regularization factor*, respectively. Let Q be a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and let $\lambda > 0$ be a positive real. The ERM-RER problem, with parameters Q and λ, consists in the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P\|Q), \tag{9}$$

where the dataset $\boldsymbol{z}$ is in (3); and the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7). For the ease of presentation, the parameters of the ERM-RER problem in (9) are chosen such that

$$Q(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = +\infty\}) = 0. \tag{10}$$

The case in which the regularization is $D(Q\|P)$ (instead of $D(P\|Q)$) in (9) is left out of the scope of this work. The interested reader is referred to [31].

C. The Solution to the ERM-RER Problem

The solution to the ERM-RER problem in (9) is presented by the following lemma.

Lemma 2 (Theorem 2.1 in [28]): Given a σ-finite measure Q and a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be such that for all $t \in \mathbb{R}$,

$$K_{Q,\boldsymbol{z}}(t) = \log\Big(\int \exp(t\, \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \mathrm{d}Q(\boldsymbol{\theta})\Big), \tag{11}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (4). Let also the set $\mathcal{K}_{Q,\boldsymbol{z}} \subset (0,+\infty)$ be

$$\mathcal{K}_{Q,\boldsymbol{z}} \triangleq \Big\{ s>0 :\ K_{Q,\boldsymbol{z}}\Big(-\frac{1}{s}\Big) < +\infty \Big\}. \tag{12}$$

Then, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the solution to the ERM-RER problem in (9) is a unique measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, whose Radon-Nikodym derivative with respect to Q satisfies that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\Big(-K_{Q,\boldsymbol{z}}\Big(-\frac{1}{\lambda}\Big) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\Big). \tag{13}$$

Among the numerous properties of the solution to the ERM-RER problem in (9), the following property is particularly useful in the remainder of this work.

Lemma 3: Given a σ-finite measure Q over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, and given a dataset $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (12), the following holds:

$$\mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) + \lambda D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\Big) = -\lambda K_{Q,\boldsymbol{z}}\Big(-\frac{1}{\lambda}\Big), \tag{14}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7); the function $K_{Q,\boldsymbol{z}}$ is defined in (11); and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the solution to the ERM-RER problem in (9).

Proof: The proof is presented in [29]. ∎

III. Sensitivity of the ERM-RER Solution

The sensitivity of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}$ to deviations from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ towards an alternative probability measure P is defined as follows.

Definition 3 (Sensitivity [28]): Given a σ-finite measure Q and a positive real $\lambda > 0$, let $\mathsf{S}_{Q,\lambda} : (\mathcal{X}\times\mathcal{Y})^n \times \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to (-\infty, +\infty]$ be a function such that

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise,} \end{cases} \tag{15}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7) and the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the solution to the ERM-RER problem in (9). The sensitivity of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}$ when the measure changes from $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to P is $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$.

The following theorem introduces an exact expression for the sensitivity in Definition 3.

Theorem 1: Given a σ-finite measure Q over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a probability measure $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that for all datasets $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$ and for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (12),

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \lambda\Big(D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\Big) + D\Big(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - D(P\|Q)\Big),$$

where the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the solution to the ERM-RER problem in (9).

Proof: The proof uses the fact that, under the assumption in (10), the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (13) is mutually absolutely continuous with respect to the σ-finite measure Q; see for instance [1]. Hence, the probability measure P is absolutely continuous with respect to $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, as a consequence of the assumption that P is absolutely continuous with respect to Q.

The proof follows by noticing that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\log\left(\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) = \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \tag{16}$$

$$= -\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\Big) \tag{17}$$

$$= K_{Q,\boldsymbol{z}}\Big(-\frac{1}{\lambda}\Big) + \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\Big), \tag{18}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $K_{Q,\boldsymbol{z}}$ are defined in (4) and in (11), respectively; and the equality in (18) follows from Lemma 2. Hence, the relative entropy $D\Big(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ satisfies

$$D\Big(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta})$$

$$= K_{Q,\boldsymbol{z}}\Big(-\frac{1}{\lambda}\Big) + \int\Big(\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\Big)\Big)\mathrm{d}P(\boldsymbol{\theta}) \tag{19}$$

$$= K_{Q,\boldsymbol{z}}\Big(-\frac{1}{\lambda}\Big) + \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}(P) + D(P\|Q) \tag{20}$$

$$= -D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) + \frac{1}{\lambda}\Big(\mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)\Big) + D(P\|Q), \tag{21}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7), the equality in (19) follows from (18), and the equality in (21) follows from Lemma 3. Finally, the proof is completed by re-arranging the terms in (21). ∎

IV. Validation of Gibbs Algorithms

Consider the dataset $\boldsymbol{z}_0 \in (\mathcal{X}\times\mathcal{Y})^{n_0}$ that aggregates dataset $\boldsymbol{z}_1 \in (\mathcal{X}\times\mathcal{Y})^{n_1}$ and dataset $\boldsymbol{z}_2 \in (\mathcal{X}\times\mathcal{Y})^{n_2}$ as constituents. That is, $\boldsymbol{z}_0 = (\boldsymbol{z}_1, \boldsymbol{z}_2)$, with $n_0 = n_1 + n_2$. Datasets $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are referred to as *constituent datasets*, whereas, the dataset $\boldsymbol{z}_0$ is referred to as the *aggregated dataset*. For all $i \in \{0,1,2\}$, the empirical risk function in (4) and the expected empirical risk function in (7) over dataset $\boldsymbol{z}_i$ are denoted by $\mathsf{L}_{\boldsymbol{z}_i}$ and $\mathsf{R}_{\boldsymbol{z}_i}$, respectively. Such functions exhibit the following property.

Lemma 4: The empirical risk functions $\mathsf{L}_{\boldsymbol{z}_0}$, $\mathsf{L}_{\boldsymbol{z}_1}$, and $\mathsf{L}_{\boldsymbol{z}_2}$, defined in (4) satisfy for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\mathsf{L}_{\boldsymbol{z}_0}(\boldsymbol{\theta}) = \frac{n_1}{n_0}\mathsf{L}_{\boldsymbol{z}_1}(\boldsymbol{\theta}) + \frac{n_2}{n_0}\mathsf{L}_{\boldsymbol{z}_2}(\boldsymbol{\theta}). \tag{22}$$

Moreover, the expected empirical risk functions $\mathsf{R}_{\boldsymbol{z}_0}$, $\mathsf{R}_{\boldsymbol{z}_1}$, and $\mathsf{R}_{\boldsymbol{z}_2}$, defined in (7), satisfy for all σ-finite measures $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$,

$$\mathsf{R}_{\boldsymbol{z}_0}(P) = \frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}(P) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}(P). \tag{23}$$

Proof: The proof is presented in [29]. ■

For all $i \in \{0,1,2\}$, let $Q_i \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $\lambda_i \in \mathcal{K}_{Q_i,\boldsymbol{z}_i}$, with $\mathcal{K}_{Q_i,\boldsymbol{z}_i}$ in (12), be the σ-finite measure acting as the reference measure and regularization factor for the learning task with dataset i, respectively. Each dataset induces a different ERM-RER problem formulation of the form

$$\min_{P \in \triangle_{Q_i}(\mathcal{M},\mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}_i}(P) + \lambda_i D(P\|Q_i), \tag{24}$$

where $\mathsf{R}_{\boldsymbol{z}_i}$ is the expected empirical risk defined in (7). For all $i \in \{0,1,2\}$, the solution to the ERM-RER problem in (24) is the probability measure denoted by $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$. In particular, from Lemma 2, it holds that the probability measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q_i$,

$$\frac{\mathrm{d}P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}}{\mathrm{d}Q_i}(\boldsymbol{\theta}) = \exp\Big(-K_{Q_i,\boldsymbol{z}_i}\Big(-\frac{1}{\lambda_i}\Big) - \frac{1}{\lambda_i}\mathsf{L}_{\boldsymbol{z}_i}(\boldsymbol{\theta})\Big). \tag{25}$$

For all $i \in \{0,1,2\}$, the probability measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ in (25) represents a GA trained upon the dataset $\boldsymbol{z}_i$ with parameters (Q_i, λ_i). In the following, such an algorithm is denoted by GA_i and the dataset $\boldsymbol{z}_i$ is often referred to as the *training dataset* of GA_i. The dataset $\boldsymbol{z}_j$, with $j \in \{0,1,2\} \setminus \{i\}$, which might contain datapoints that are not in $\boldsymbol{z}_i$, is referred to as the *test dataset* for GA_i.

A. Gibbs Algorithms Trained on Constituent Datasets

The expected empirical risk induced by GA_i on the training dataset $\boldsymbol{z}_i$ is the *training expected empirical risk*, which is denoted by $\mathsf{R}_{\boldsymbol{z}_i}\big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\big)$ and often referred to as the *training error* [32]. Alternatively, the expected empirical risk induced by GA_i on the test dataset $\boldsymbol{z}_j$ is the *test expected empirical risk*, which is denoted by $\mathsf{R}_{\boldsymbol{z}_j}\big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\big)$ and often referred to as the *test error* [32]. The following theorem provides explicit expressions involving the training and test errors of GA_1 and GA_2.

Theorem 2: Assume that the σ-finite measures Q_1 and Q_2 in (24) are mutually absolutely continuous. Then, for all $i \in \{1,2\}$ and $j \in \{1,2\} \setminus \{i\}$,

$$\mathsf{R}_{\boldsymbol{z}_i}\big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\big) - \mathsf{R}_{\boldsymbol{z}_i}\big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\big) = \lambda_i\Big(D\big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_i\big) + D\big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\|P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\big) - D\big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\|Q_i\big)\Big), \tag{26}$$

where the function $\mathsf{R}_{\boldsymbol{z}_i}$ is defined in (7) and the measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ satisfies (25).

Proof: The proof is immediate from Theorem 1 by noticing that for all $i \in \{1,2\}$ and for all $j \in \{1,2\} \setminus \{i\}$, the differences $\mathsf{R}_{\boldsymbol{z}_i}\big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\big) - \mathsf{R}_{\boldsymbol{z}_i}\big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\big)$ can be written in terms of the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\big(\boldsymbol{z}_i, P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\big)$. ■

A reasonable figure of merit to compare two machine learning algorithms trained upon two different training datasets is the difference between the expected empirical risk they induce upon the aggregation of their training datasets. The following theorem provides an explicit expression for this figure of merit for the case of the algorithms GA_1 and GA_2.

Theorem 3: Assume that the σ-finite measures Q_1 and Q_2 in (24) are mutually absolutely continuous. Then,

$$\begin{aligned}\mathsf{R}_{\boldsymbol{z}_0}\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\big) - \mathsf{R}_{\boldsymbol{z}_0}\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\big) &= \frac{n_1}{n_0}\lambda_1\Big(D\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q_1\big) \\ &+ D\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\big) - D\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q_1\big)\Big) \\ &- \frac{n_2}{n_0}\lambda_2\Big(D\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q_2\big) + D\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\big) \\ &- D\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q_2\big)\Big),\end{aligned} \tag{27}$$

where the function $\mathsf{R}_{\boldsymbol{z}_0}$ is defined in (7) and, for all $i \in \{1,2\}$, the measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ satisfies (25).

Proof: The proof uses the following argument:

$$\begin{aligned}&\mathsf{R}_{\boldsymbol{z}_0}\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\big) - \mathsf{R}_{\boldsymbol{z}_0}\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\big) \\ &= \frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\big) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\big) \\ &\quad - \Big(\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\big) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\big)\Big) \qquad (28) \\ &= \frac{n_1}{n_0}\mathsf{S}_{Q_1,\lambda_1}\big(\boldsymbol{z}_1, P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\big) - \frac{n_2}{n_0}\mathsf{S}_{Q_2,\lambda_2}\big(\boldsymbol{z}_2, P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\big), \qquad (29)\end{aligned}$$

where the equality in (28) follows from Lemma 4; and the equality in (29) follows from Definition 3. The proof is completed by Theorem 1. ■

B. Averaging Gibbs Measures

In practical scenarios, building GAs on the aggregated dataset might be difficult or impossible due to limited computational power or due to the fact that dataset aggregation at one location is not allowed due to privacy constraints. In these cases, a common practice is to average the output of machine learning algorithms trained on constituent datasets, e.g., federated learning [30]. In this case, a figure of merit to validate such an approach is to study the difference of the expected empirical risk induced on the aggregated dataset by GA_0 and a convex combination of GA_1 and GA_2. The following theorem provides an explicit expression for this quantity.

Theorem 4: Assume that the σ-finite measures Q_0, Q_1 and Q_2 in (24) are pair-wise mutually absolutely continuous. Then, for all $\alpha \in [0,1]$,

$$\begin{aligned}&\mathsf{R}_{\boldsymbol{z}_0}\big(\alpha P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1} + (1-\alpha)P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\big) - \mathsf{R}_{\boldsymbol{z}_0}\big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\big) \\ &= \lambda_0\Big(D\big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|Q_0\big) \\ &\quad + \alpha\Big(D\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\big) - D\big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q_0\big)\Big) \\ &\quad + (1-\alpha)\Big(D\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\big) - D\big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q_0\big)\Big)\Big),\end{aligned} \tag{30}$$

where the function $\mathsf{R}_{\boldsymbol{z}_0}$ is defined in (7) and, for all $i \in \{1,2\}$, the measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ satisfies (25).

Proof: The proof uses the following argument:

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}_0}\Big(\alpha P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}+(1-\alpha)P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)-\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)\\
=&\alpha\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)+(1-\alpha)\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)\\
&-\alpha\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)-(1-\alpha)\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) \qquad (31)\\
=&\alpha\Big(\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)-\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)\Big)\\
&+(1-\alpha)\Big(\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)-\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)\Big) \qquad (32)\\
=&\alpha\mathsf{S}_{Q_0,\lambda_0}\Big(\boldsymbol{z}_0,P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)+(1-\alpha)\mathsf{S}_{Q_0,\lambda_0}\Big(\boldsymbol{z}_0,P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big), \qquad (33)
\end{aligned}$$

where the equality in (31) follows from Lemma 1, and the equality in (33) follows from Definition 3. The proof is completed by Theorem 1. ∎

The following corollary of Theorem 4 is obtained by subtracting the equality in (30) with $\alpha = 1$ from the equality in (30) with $\alpha = 0$.

Corollary 1: Assume that the σ-finite measures Q_0, Q_1 and Q_2 in (24) are pair-wise mutually absolutely continuous. Then, for all $i \in \{0,1,2\}$, the probability measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ in (25) satisfies

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)-\mathsf{R}_{\boldsymbol{z}_0}\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)\\
=&\lambda_0\left(D\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)-D\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q_0\Big)\right)\\
&-\lambda_0\left(D\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)-D\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q_0\Big)\right), \qquad (34)
\end{aligned}$$

where the function $\mathsf{R}_{\boldsymbol{z}_0}$ is defined in (7).

Corollary 1 is an alternative to Theorem 3 involving the GA trained upon the aggregated dataset, i.e., GA_0.

C. Gibbs Algorithms Trained on Aggregated Datasets

Training a GA upon the aggregation of datasets does not necessarily imply lower expected empirical risk on the constituent datasets. As argued before, datasets might be obtained up to different levels of fidelity. Hence, a validation method for GA_0 is based on the expected empirical risk induced by GA_0 on a constituent dataset $\boldsymbol{z}_i$, with $i \in \{1,2\}$, which is denoted by $\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)$. A pertinent figure of merit is the difference $\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)-\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)$. The following theorem provides an explicit expression for such quantity.

Theorem 5: Assume that the σ-finite measures Q_0, Q_1 and Q_2 in (24) are pair-wise mutually absolutely continuous. Then, for all $i \in \{0,1,2\}$,

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)-\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)=\lambda_i\Big(D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_i\Big)\\
&+D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)-D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|Q_i\Big)\Big), \qquad (35)
\end{aligned}$$

where, the function $\mathsf{R}_{\boldsymbol{z}_i}$ is defined in (7) and the measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ satisfies (25).

Proof: The proof is immediate from Theorem 1 by noticing that for all $i \in \{1,2\}$, the differences $\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)-\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)$ can be written in terms of the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\Big(\boldsymbol{z}_i,P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big)$. ∎

D. Special Cases

Consider a given σ-finite measure Q and assume that for all $i \in \{0,1,2\}$ and for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$, $Q(\mathcal{A}) = Q_i(\mathcal{A})$. Assume also that the parameters λ_0, λ_1, and λ_2 in (24) satisfy $\lambda_1 = \frac{n_0}{n_1}\lambda_0$ and $\lambda_2 = \frac{n_0}{n_2}\lambda_0$. These assumptions are referred to as the case of *homogeneous priors* with measure Q, and the case of *proportional regularization*, respectively. The term "proportional" stems from the fact that the regularization factor decreases proportionally to the size of the data set in the optimization problem in (24). Under these assumptions, the following corollary of Theorem 2 unveils an interesting connection with the Jeffrey's divergence [33].

Corollary 2: Consider the case of homogeneous priors with a σ-finite measure Q and proportional regularization with parameter λ_0. Then, for all $i \in \{1,2\}$, the probability measure $P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ in (25), satisfies

$$\begin{aligned}
&\Big(\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)-\frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)\Big)\\
&+\Big(\frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)-\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)\Big)\\
&=\lambda_0\left(D\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)+D\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)\right). \qquad (36)
\end{aligned}$$

Note that $D\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)+D\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)$ is the Jeffrey's divergence between the measures $P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ and $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}$. For all $i \in \{1,2\}$ and $j \in \{1,2\}\backslash\{i\}$, the difference $\frac{n_j}{n_0}\mathsf{R}_{\boldsymbol{z}_j}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)-\frac{n_i}{n_0}\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)$ is reminiscent of a *validation* [32, Section 11.2]. This follows from noticing that $\mathsf{R}_{\boldsymbol{z}_j}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)$ is the testing error of GA_i over the test dataset $\boldsymbol{z}_j$, while $\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)$ is the training error of GA_i.

In (36), it holds that $D\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)$ and $D\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)$ are both nonnegative, which leads to the following corollary of Theorem 2.

Corollary 3: Consider the case of homogeneous priors with a σ-finite measure Q and proportional regularization. Then, for all $i \in \{1,2\}$, the probability measure $P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ in (25), satisfies

$$\begin{aligned}
&\Big(\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)+\frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)\Big)\\
\geqslant&\Big(\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)+\frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)\Big). \qquad (37)
\end{aligned}$$

Corollary 3 highlights that the weighted-sum of the test errors induced by GA_1 and GA_2 is not smaller than the weighted sum of their training errors when the weights are proportional to the sizes of the datasets.

REFERENCES

[1] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with generalized relative entropy regularization," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9454, Feb. 2022.

[2] O. Catoni, *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, 1st ed. New York, NY, USA: Springer Science & Business Media, 2004, vol. 1851.

[3] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, Aug. 2016.

[4] P. Alquier, J. Ridgway, and N. Chopin, "On the properties of variational approximations of Gibbs posteriors," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, Dec. 2016.

[5] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," *Advances in Neural information Processing Systems*, vol. 34, pp. 8106–8118, Dec. 2021.

[6] T. Zhang, "From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, vol. 34, no. 5, pp. 2180–2210, 2006.

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

[8] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 1475–1479.

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

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

[11] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 582–586.

[12] 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, Jan. 2019.

[13] 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.

[14] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," *Advances in Neural information Processing Systems*, pp. 7245–7254, Dec. 2018.

[15] A. T. Lopez and V. Jog, "Generalization error bounds using Wasserstein distances," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Guangzhou, China, Nov. 2018, pp. 1–5.

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

[17] 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*, pp. 16 457–16 467, Dec. 2020.

[18] 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*, pp. 9925–9935, Dec. 2018.

[19] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via Wasserstein distance," *Advances in Neural information Processing Systems*, pp. 19 109–19 121, Dec. 2021.

[20] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and maximal leakage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, 2021.

[21] G. Aminian, L. Toni, and M. R. Rodrigues, "Jensen-Shannon information based characterization of the generalization error of learning algorithms," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Kanazawa, Japan, Oct. 2021, pp. 1–5.

[22] G. Aminian, Y. Bu, L. Toni, M. R. D. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," ArXiv Preprint 2210.09864, 2022.

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

[24] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proceedings of Tenth Annual Conference on Computational Learning Theory*, July 1997, pp. 2–9.

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

[26] M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor, "PAC-Bayes unleashed: Generalisation bounds with unbounded losses," *Entropy*, vol. 23, no. 10, Oct. 2021.

[27] B. Guedj and L. Pujol, "Still no free lunches: The price to pay for tighter PAC-Bayes bounds," *Entropy*, vol. 23, no. 11, Nov. 2021.

[28] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization: Optimality and sensitivity," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, Jul. 2022.

[29] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "Sensitivity of the Gibbs algorithm to data aggregation in supervised machine learning," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Research Report RR-9474, Jun. 2022.

[30] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, Florida, Apr. 2017, pp. 1273–1282.

[31] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Analysis of the relative entropy asymmetry in the regularization of empirical risk minimization," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.

[32] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2014.

[33] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.