Sensitivity of the Gibbs Algorithm to Data Aggregation in Supervised Machine Learning

Samir M Perlaza, Iñaki Esnaola, H Vincent Poor

Sensitivity of the Gibbs Algorithm to Data Aggregation in Supervised Machine Learning

Samir M. Perlaza(1), Iñaki Esnaola(2), and H. Vincent Poor(3).

ISSN 0249-6399 ISRN INRIA/RR-9474--FR+ENG

RESEARCH REPORT
N° 9474
Jun 20, 2022
Project-Team NEO

Sensitivity of the Gibbs Algorithm to Data Aggregation in Supervised Machine Learning

Samir M. Perlaza(1), Iñaki Esnaola(2), and H. Vincent Poor(3).

Project-Team NEO

Research Report n° 9474 — Jun 20, 2022 — 22 pages

Abstract: An explicit expression for the sensitivity of the expected empirical risk (EER) induced by the Gibbs algorithm (GA) is presented in the context of supervised machine learning. The sensitivity is defined as the difference between the EER induced by the GA and the EER induced by an alternative probability measure on the models. When several datasets are available, the sensitivity plays a central role to determine whether or not a lower EER might be observed by aggregating several datasets. Necessary and sufficient conditions for decreasing the EER by dataset aggregation are presented. Such conditions, which are on the GA parameters and the reference measures assumed for each constituent dataset, boils down to the evaluation of the sign of the sum of some relative entropy terms. From this perspective, sensitivity appears as (a) an alternative metric to evaluate the generalization capabilities of the Gibbs algorithm; and (b) a theoretical ground to study the use of several datasets describing the same phenomenon, yet subject to different data acquisition systems, i.e., datasets with different statistical properties.

Key-words: Supervised Learning, Gibbs Algorithm, PAC-Learning, Relative Entropy Regularization, Empirical Risk Minimization, Maximum Entropy Principle, and Bayesian Learning.

(1) Samir M. Perlaza is with INRIA, 2004 Route des Lucioles, 06902 Sophia Antipolis, France (samir.perlaza@inria.fr); with the Mathematics Department (GAATI Laboratory) of the University of French Polynesia, BP 6570, 98702 Faaa, French Polynesia; and with the Electrical and Computer Engineering Department at Princeton University, 08544 Princeton, NJ.

(2) Iñaki Esnaola is with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom; and with the Electrical and Computer Engineering Department at Princeton University, 08544 Princeton, NJ.

(3) H. Vincent Poor is with the Electrical and Computer Engineering Department at Princeton University, 08544 Princeton, NJ.

This work was developed in the context of the Inria Exploratory Action “Information and Decision Making” (IDEM).

Sensibilité de la Valeur Espérée du Risque Empirique Induit par l'Algorithme de Gibbs dans le Problème d'Apprentissage Supervisé

Résumé : Une expression explicite de la sensibilité du risque empirique espéré (REE) induit par l'algorithme de Gibbs (AG) dans le problème de l'apprentissage automatique supervisé est présentée. La sensibilité est définie comme la différence entre l'REE induit par l'AG et l'REE induit par une mesure de probabilité alternative sur les modèles. Lorsque plusieurs ensembles de données sont disponibles, la sensibilité joue un rôle central pour déterminer si un REE plus petit peut-être observé comme résultat de l'aggregation de plusieurs ensembles de données. Les conditions nécessaires et suffisantes pour observer une diminution de l'EER due à l'agrégation des données sont présentées. De telles conditions, qui sont sur les paramètres de l'AG et les mesures de référence supposées pour chaque ensemble de données, se résument à l'évaluation du signe d'une somme de certains termes d'entropie relative. À la lumière de ces résultats, la sensibilité apparaît comme (a) une métrique alternative pour évaluer les capacités de généralisation de l'algorithme de Gibbs; et (b) un cadre théorique pour étudier l'impact de l'utilisation de plusieurs ensembles de données décrivant le même phénomène mais soumis à différents systèmes d'acquisition de données, ce qui implique par exemple, différentes propriétés statistiques pour chaque ensemble de données.

Mots-clés : Apprentissage Supervisé, Apprentissage PAC, Régularisation, Entropie Relative, Minimisation du Risque Empirique, Principe d'Entropie Maximale, et Apprentissage Bayésien.

Contents

1 Introduction

In the context of supervised learning, the Gibbs algorithm labels unseen data by randomly selecting a model sampled from the probability measure that solves the empirical risk minimization (ERM) problem with relative entropy regularization (ERM-RER) [1]. One of the main advantages of the Gibbs algorithm is that it does not require additional assumptions on the statistical description of the datasets [2–4]. Instead, it requires a prior or preference over the set of models in the form of a σ-finite measure, e.g., a probability measure. The regularization term in the ERM-RER is precisely a relative entropy with respect to such prior, which in the case of probability measures, has been shown to govern the generalization capabilities of the Gibbs algorithm [5,6]. The solution to the ERM-RER problem is unique and described by the Gibbs probability measure, which has been extensively studied using information theoretic tools in [5–8]; statistical physics [2]; PAC (Probably Approximatively Correct)-Bayesian learning theory [9–12]; and shown to be central to classification problems [13, 14].

A dataset is constructed by acquiring data from a source that is linked to a phenomenon for which a learning task must be implemented. In practical settings, training data might be acquired from different sources and through different data acquisition systems. This leads to datasets sampled from different probability measures. In such cases, data aggregation techniques that combine several constituent datasets acquired from sources related to the same phenomenon can be useful for tackling poor performance induced by data scarcity from some of the sources. For instance, consider that a dataset is used to implement a Gibbs algorithm and that a new constituent dataset about the same phenomenon is made available. The question addressed in this report is the following: Should this new dataset be aggregated to the previous dataset to form a larger training dataset aiming at improving the performance of the Gibbs algorithm? The main challenge stems from the fact that the probability measures generating each constituent dataset are not known, and therefore, Bayesian characterizations of the aggregate risk are not feasible. Moreover, assuming that both datasets have been generated by the same probability measure is not a prudent premise because differences in dataset acquisition methods often result in varying degrees of fidelity for each constituent dataset.

In this report, the performance of the Gibbs algorithm is analyzed for datasets constructed by aggregating two constituent datasets for which the probability measures from which they have been sampled from are unknown. Interestingly, the sensitivity of the expected empirical risk induced by the Gibbs algorithm quantifies the benefit of data aggregation. It is shown that the sensitivity change induced by data aggregation is a function of the Gibbs probability measures obtained for each constituent dataset and the corresponding reference measures. This analytical characterization of the sensitivity is used to distill necessary and sufficient conditions for improving the Gibbs algorithm performance obtained with the aggregate dataset with respect to the performance obtained with the constituent datasets. Remarkably, the resulting sensitivity expressions are computable without knowledge of the probability measures generating the datasets and can be expressed in terms of the Gibbs probability measures obtained for the constituent datasets and for the aggregate datasets.

2 Problem Formulation

Consider three sets $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$, with $\mathcal{M} \subseteq \mathbb{R}^d$ and $d \in \mathbb{N}$. Let the function $f : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ be such that, for some $\boldsymbol{\theta}^\star \in \mathcal{M}$, there exist two random variables X and Y that satisfy,

$$Y = f(\boldsymbol{\theta}^\star, X). \tag{1}$$

A pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is referred to as a data point. Given n data points, with $n \in \mathbb{N}$, denoted by (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), a dataset is the tuple $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$. The model $\boldsymbol{\theta}^\star$ in (1), which is often referred to as the *ground truth model*, is unknown. Given a dataset, the objective is to obtain a model $\boldsymbol{\theta} \in \mathcal{M}$, such that for all patterns $u \in \mathcal{X}$, the assigned label $f(\boldsymbol{\theta}, u)$ minimizes a notion of *loss* or *risk*. Let 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 loss or risk induced by choosing the model $\boldsymbol{\theta} \in \mathcal{M}$, which assigns the label $f(\boldsymbol{\theta}, x)$ to the pattern x, is $\ell\left(f(\boldsymbol{\theta}, x), y\right)$. Often the function ℓ is referred to as the *loss function* or *risk function*. In the following, it is assumed that the function ℓ satisfies that for all $y \in \mathcal{Y}$, $\ell\left(y, y\right) = 0$, which implies that correct labelling implies zero cost. Note that there might exist several models $\boldsymbol{\theta} \in \mathcal{M} \setminus \{\boldsymbol{\theta}^\star\}$ such that $\ell\left(f(\boldsymbol{\theta}, x), y\right) = 0$, which reveals the need of a large number of labelled patterns for model selection.

The *empirical risk* induced by the model $\boldsymbol{\theta}$, with respect to a dataset

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

with $n \in \mathbb{N}$, is determined by the function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to [0, +\infty)$, which satisfies

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell\left(f(\boldsymbol{\theta}, x_i), y_i\right). \tag{3b}$$

Using this notation, the ERM problem with respect to the data set $\boldsymbol{z}$ is the following optimization problem

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right), \tag{4}$$

whose solutions form the set denoted by $\mathcal{T}\left(\boldsymbol{z}\right) \triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$. The ERM problem in (4) is well posed. Note for instance that the ground truth model $\boldsymbol{\theta}^\star$ in (1) is one of the solutions to the ERM problem in (4). That is, the model $\boldsymbol{\theta}^\star$ in (1) satisfies that $\boldsymbol{\theta}^\star \in \mathcal{T}\left(\boldsymbol{z}\right)$ and $\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}^\star\right) = 0$.

2.1 Generalized Relative Entropy Regularization

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

Definition 2.1 (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*

$$D(Q\|P) = \int \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right) \mathrm{d}P(x), \tag{5}$$

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

A fundamental assumption in this work is that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $\bar{\ell}_{x,y} : \mathcal{M} \to [0, +\infty)$, such that for all $(\boldsymbol{\theta}, x, y) \in \mathcal{M} \times \mathcal{X} \times \mathcal{Y}$,

$$\bar{\ell}_{x,y}(\boldsymbol{\theta}) = \ell\left(f(\boldsymbol{\theta}, x), y\right), \tag{6}$$

where the functions f and ℓ are those in (1) and (2), is Borel measurable with respect to the measure space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. Under this assumption, the *expected empirical risk* is introduced.

Definition 2.2 (Expected Empirical Risk)**.** *Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0, +\infty)$ be such that for all σ-finite measures $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that*

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

where the function $\mathsf{L}_{\boldsymbol{z}}$ is in (3b)*. When P is a probability measure, the expected empirical risk induced by P is $\mathsf{R}_{\boldsymbol{z}}(P)$.*

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:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P\|Q), \tag{8}$$

where the dataset $\boldsymbol{z}$ is in (3a); the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7); and the optimization domain is the set of all probability measures on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ that are absolutely continuous with the measure Q. The notation $\triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ in (8) is used to represent the set of probability measures on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ that are absolutely continuous with the measure Q.

2.2 The Gibbs Algorithm

The solution to the ERM-RER problem in (8) is presented by the following lemma.

Lemma 2.1 (Theorem 2.1 in [1]). *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 $t \in \mathbb{R}$,*

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\right), \tag{9}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3b)*. Let also the set $\mathcal{K}_{Q,\boldsymbol{z}} \subset (0, +\infty)$ be*

$$\mathcal{K}_{Q,\boldsymbol{z}} \triangleq \left\{ s > 0 : \; K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) < +\infty \right\}. \tag{10}$$

Then, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the solution to the ERM-RER problem in (8) *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 $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right). \tag{11}$$

When Q is a probability measure, the ERM-RER problem in (8) is known to possess a unique solution [6,15,16] described by a Gibbs probability measure. In the more general case in which Q is a σ-finite measure, the ERM-RER problem in (8) also possesses a unique solution to which we refer to as a Gibbs probability measure [1], despite the fact that it is defined for a σ-finite measure instead of a probability measure. Similarly, the function $K_{Q,\boldsymbol{z}}$ in (9) is referred to as the *log-partition function*, independently of whether the reference measure Q is a probability measure. This is in order to avoid disrupting with the current nomenclature.

Using Lemma 2.1, the Gibbs algorithm can be described as follows.

Algorithm 1: The Gibbs Algorithm

Parameter: Training Data $\boldsymbol{z}$ in (3a);
Reference Measure Q in (8); and
Regularization Factor λ in (8)

Input: Unseen Pattern $x \in \mathcal{X}$

1 Obtain $\boldsymbol{\theta} \in \mathcal{M}$ by sampling from $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11)

Output: Label $y = f(\boldsymbol{\theta}, x)$, with f in (1)

In the following, the Algorithm 1 is represented by the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11), which justifies that often, such probability measure is referred to as the Gibbs algorithm itself, c.f., [5–7, 17].

The expected empirical risk induced by the Gibbs algorithm $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is denoted by $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$, where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7). Among the numerous properties of the Gibbs algorithm, the following property plays a central role in this work.

Lemma 2.2. *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* (10)*, it holds that*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \lambda D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\Big) = -\lambda K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \text{ and} \tag{12}$$

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \lambda D\Big(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = -\lambda K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{13}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7)*; the function $K_{Q,\boldsymbol{z}}$ is defined in* (9)*; and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the solution to the ERM-RER problem in* (8)*.*

Proof: From Lemma 2.1, it follows that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) = -K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}), \tag{14}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3b). Thus,

$$\begin{aligned} D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\Big) &= \int \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (15)\\ &= -K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (16)\\ &= -K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), && (17) \end{aligned}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7). This completes the proof of (12).

From (14), it follows that

$$\begin{aligned} D\Big(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) &= -\int \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && (18)\\ &= K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{\lambda}\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && (19)\\ &= K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}(Q), && (20) \end{aligned}$$

which completes the proof of (13). ■

Note that Lemma 2.2 states that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (10), it holds that

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P\|Q) = -\lambda K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right). \tag{21}$$

3 Sensitivity

The sensitivity of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}$ in (7) to deviations from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11) towards an alternative probability measure P is defined as follows.

Definition 3.1 (Definition 8 in [8]). *Given a σ-finite measure Q and a positive real $\lambda > 0$, let $\mathsf{S}_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \triangle_P(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to (-\infty, +\infty]$ be a function such that for all datasets $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ and for probability measures $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that*

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise,} \end{cases} \tag{22}$$

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 (8). 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 sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$ in (22) is a means to quantify the change of the expected empirical risk function $\mathsf{R}_{\boldsymbol{z}}$ around the optimal solution of a given ERM-RER problem in (8). That is, it quantifies the loss or gain obtained by using an alternative to the Gibbs algorithm $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, i.e., Algorithm 1. The following theorem introduces an exact expression for the sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$.

Theorem 3.1. *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 (10),*

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \lambda\left(D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\Big) + D\Big(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - D(P\|Q)\right), \tag{23}$$

where the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the solution to the ERM-RER problem in (8).

Proof: The proof uses the fact that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11) is mutually absolutely continuous with the σ-finite measure Q [1, Lemma 2.6]. Hence, the probability measure P is absolutely continuous with $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, as a consequence of the assumption of the lemma that P is absolutely continuous with Q.

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

$$\log\left(\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)=\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \tag{24}$$

$$=-\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)+\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \tag{25}$$

$$=K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)+\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right), \tag{26}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $K_{Q,\boldsymbol{z}}$ are defined in (3b) and in (9), respectively; and the equality in (26) follows from Lemma 2.1. Hence, the relative entropy $D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ satisfies,

$$D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{27}$$

$$=K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)+\frac{1}{\lambda}\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \tag{28}$$

$$+\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{29}$$

$$=K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)+\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}(P)+D(P\|Q) \tag{30}$$

$$=-D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)-\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}(P)+D(P\|Q), \tag{31}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7), and the equality in (31) follows from Lemma 2.2. Finally, from (31), it follows that

$$\mathsf{R}_{\boldsymbol{z}}(P)-\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\lambda\left(D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)-D(P\|Q)\right), \tag{32}$$

which completes the proof. ■

An interesting observation from Theorem 3.1 follows by re-writing (22) in terms of the objective function of the ERM-RER problem in (8), as shown by the following corollary.

Corollary 3.1. *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ 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 (10),*

$$\left(\mathsf{R}_{\boldsymbol{z}}(P)+\lambda D(P\|Q)\right)-\left(\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\lambda D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\right)$$
$$=\lambda D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \tag{33}$$

where the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the solution to the ERM-RER problem in (8).

Corollary 3.1 characterizes the deviation of the objective function of the ERM-RER problem in (8) from its solution, i.e., the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11), to an alternative probability measure $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$.

3.1 Priors and Posteriors

In Theorem 3.1, when P is chosen to be identical to the reference measure Q, it follows that

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda\left(D\Big(Q\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) + D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big)\right), \tag{34}$$

where the right-hand side is a symmetrized Kullback-Liebler divergence, also known as Jeffrey's divergence, between the measures Q and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. More importantly, when Q is a probability measure, it follows that $D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) \geqslant 0$ and $D\Big(Q\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geqslant 0$, which leads to the following corollary from Theorem 3.1.

Corollary 3.2. *Given a probability measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a probability measure $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, 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* (10)*, it holds that*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant \mathsf{R}_{\boldsymbol{z}}(Q), \tag{35}$$

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

Corollary 3.2 reveals the fact that, under the assumption that the reference measure Q is a probability measure, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (10), the expected empirical risk induced by the Gibbs algorithm $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is smaller than or equal to expected empirical risk obtained by randomly sampling models from the reference measure Q. This observation can be interpreted from a Bayesian perspective. Note for instance that the probability measure Q on the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ does not depend upon the available dataset. Hence, Q can be interpreted as a prior probability measure on the models. This also justifies interpreting the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ as a posterior probability measure on the set of models.

3.2 A Geometric Interpretation of Sensitivity

In Theorem 3.1, note that if Q is a probability measure, then it holds that $D(P\|Q) \geqslant 0$, $D\Big(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geqslant 0$, and $D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) \geqslant 0$, and thus, from (22), it holds that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (10),

$$\begin{aligned}
\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) + \lambda D(P\|Q) &= \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P\|Q) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) &(36)\\
&\geqslant \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \lambda D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) &(37)\\
&= \lambda D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big), &(38)
\end{aligned}$$

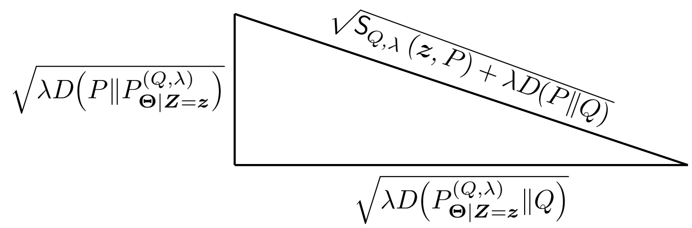

Figure 1: Geometric interpretation of the sensitivity (Definition 3.1).

and

$$\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, P\right) + \lambda D(P\|Q) = \lambda\left(D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) + D\Big(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)\right) \tag{39}$$

$$\geqslant \lambda D\Big(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big), \tag{40}$$

where the inequality in (37), follows from Lemma 2.1.

Hence, under the assumption that the measure Q is a probability measure, the inequalities in (38) and (40) allow writing the equality in (22) as follows:

$$\left(\sqrt{\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, P\right) + \lambda D(P\|Q)}\right)^2 = \left(\sqrt{\lambda D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big)}\right)^2 \tag{41}$$

$$+\left(\sqrt{\lambda D\Big(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)}\right)^2, \tag{42}$$

which implies that a right-angled triangle can be constructed such that the hypotenuse exhibits length $\sqrt{\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, P\right) + \lambda D(P\|Q)}$ and the short sides exhibits lengths $\sqrt{\lambda D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big)}$ and $\sqrt{\lambda D\Big(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)}$, respectively. Figure 1 shows this interpretation of sensitivity.

4 Sensitivity to Dataset Aggregation

4.1 Dataset Aggregation

For all $i \in \{1, 2\}$, let $n_i \in \mathbb{N}$ be the number of labelled patterns in the dataset i and denote by $\boldsymbol{z}_i \in (\mathcal{X} \times \mathcal{Y})^{n_i}$ such dataset. Consider the dataset $\boldsymbol{z}_0 \in (\mathcal{X} \times \mathcal{Y})^{n_0}$ that aggregates dataset 1 and dataset 2 as constituents such that

$$\boldsymbol{z}_0 = (\boldsymbol{z}_1, \boldsymbol{z}_2). \tag{43a}$$

The total number of labelled patterns is given by

$$n_0 = n_1 + n_2, \tag{43b}$$

and for all $i \in \{0, 1, 2\}$, the entries of the datasets are denoted as

$$\boldsymbol{z}_i = \left(\left(x_{i,1}, y_{i,1}\right), \left(x_{i,2}, y_{i,2}\right), \ldots, \left(x_{i,n_i}, y_{i,n_i}\right)\right) \in (\mathcal{X} \times \mathcal{Y})^{n_i}. \tag{43c}$$

For all $i \in \{0,1,2\}$, let $Q_i \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ be a σ-finite measure. Let also $\lambda_i \in \mathcal{K}_{Q_i,\boldsymbol{z}_i}$, with $\mathcal{K}_{Q_i,\boldsymbol{z}_i}$ in (10), be a positive real number. Each dataset induces a different ERM-RER problem formulation of the form

$$\min_{P \in \triangle_{Q_i}(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}_i}(P) + \lambda_i D(P \| Q_i), \tag{44}$$

where $\mathsf{R}_{\boldsymbol{z}_i}$ is the expected empirical risk defined in (7).

The following lemma shows that the empirical risk function and the expected empirical risk function that emerge as a result of aggregating dataset $\boldsymbol{z}_1$ and dataset $\boldsymbol{z}_2$ are convex combinations of the empirical risk functions and expected empirical risk functions induced by the constituent datasets.

Lemma 4.1. *Consider the datasets $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, and $\boldsymbol{z}_2$ with lengths n_0, n_1, and n_2, respectively, that satisfy* (43). *Then, the empirical risk functions $\mathsf{L}_{\boldsymbol{z}_0}$, $\mathsf{L}_{\boldsymbol{z}_1}$, and $\mathsf{L}_{\boldsymbol{z}_2}$, defined in* (3b) *satisfy for all $\boldsymbol{\theta} \in \mathcal{M}$,*

$$\mathsf{L}_{\boldsymbol{z}_0}(\boldsymbol{\theta}) = \frac{n_1}{n_0}\mathsf{L}_{\boldsymbol{z}_1}(\boldsymbol{\theta}) + \frac{n_2}{n_0}\mathsf{L}_{\boldsymbol{z}_2}(\boldsymbol{\theta}). \tag{45}$$

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 $P \in \triangle(M, \mathscr{B}(M))$,*

$$\mathsf{R}_{\boldsymbol{z}_0}(P) = \frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}(P) + \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}(P). \tag{46}$$

Proof: From the definition of the empirical risk function $\mathsf{L}_{\boldsymbol{z}_i}$, with $i \in \{0,1,2\}$, in (3b) and the structure of the datasets $\boldsymbol{z}_0$, $\boldsymbol{z}_1$, and $\boldsymbol{z}_2$ in (43), it follows that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\begin{aligned}
\mathsf{L}_{\boldsymbol{z}_0}(\boldsymbol{\theta}) &\triangleq \frac{1}{n_0}\sum_{t=1}^{n_0} \ell(f(\boldsymbol{\theta}, x_{0,t}), y_{0,t}) && (47)\\
&= \frac{n_1}{n_0}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} \ell(f(\boldsymbol{\theta}, x_{0,t}), y_{0,t})\right) + \frac{n_2}{n_0}\left(\frac{1}{n_2}\sum_{t=n_1+1}^{n_2} \ell(f(\boldsymbol{\theta}, x_{0,t}), y_{0,t})\right) && (48)\\
&= \frac{n_1}{n_0}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} \ell(f(\boldsymbol{\theta}, x_{1,t}), y_{1,t})\right) + \frac{n_2}{n_0}\left(\frac{1}{n_2}\sum_{t=1}^{n_2} \ell(f(\boldsymbol{\theta}, x_{2,t}), y_{2,t})\right) && (49)\\
&= \frac{n_1}{n_0}\mathsf{L}_{\boldsymbol{z}_1} + \frac{n_2}{n_0}\mathsf{L}_{\boldsymbol{z}_2}, && (50)
\end{aligned}$$

where the equality in (49) follows from the concatenation of datasets $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ into $\boldsymbol{z}_0$, as shown in (43a). This completes the proof of the equality in (45). The proof of the equality in (46) follows by integrating with respect to the σ-finite measure P the equality in (45). ∎

Lemma 4.1 highlights that the risk contribution of each dataset is proportional to the corresponding number of labelled patterns supplied to the aggregate dataset.

4.2 Sensitivity Analysis with Constituent Datasets

For all $i \in \{0,1,2\}$, the solution to the ERM-RER problem in (44) is a probability measure denoted by $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$. In particular, from Lemma 2.1, it holds that the probability measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q_i$,

$$\frac{\mathrm{d}P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}}{\mathrm{d}Q_i}(\boldsymbol{\theta})=\exp\left(-K_{Q_i,\boldsymbol{z}_i}\left(-\frac{1}{\lambda_i}\right)-\frac{1}{\lambda_i}\mathsf{L}_{\boldsymbol{z}_i}(\boldsymbol{\theta})\right). \tag{51}$$

The probability measure $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ in (51) defines a Gibbs algorithm, c.f., Algorithm 1, whose training dataset is $\boldsymbol{z}_i$. The following theorem provides expressions for the differences $\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)-\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$, i.e., the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\left(\boldsymbol{z}_i,P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)$.

Theorem 4.1. *Assume that the σ-finite measures Q_1 and Q_2 in* (44) *are mutually absolutely continuous. Hence, for all $i \in \{1,2\}$ and $j \in \{1,2\}\setminus\{i\}$, the solution to the ERM-RER problem in* (44)*, denoted by $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$, satisfies:*

$$\begin{aligned}&\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)-\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)\\ &=\lambda_i\left(D\Big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\|P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)+D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_i\Big)-D\Big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\|Q_i\Big)\right).\end{aligned} \tag{52}$$

where the functions $\mathsf{R}_{\boldsymbol{z}_1}$ and $\mathsf{R}_{\boldsymbol{z}_2}$ are defined in (7).

Proof: The proof of Theorem 4.1 is immediate from Theorem 3.1 by noticing that for all $i \in \{1,2\}$ and for all $j \in \{1,2\}\setminus\{i\}$, the diferences $\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)-\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$ can be written in terms of the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\left(\boldsymbol{z}_i,P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)$. ■

The relevance of the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\left(\boldsymbol{z}_i,P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)$ in Theorem 4.1 is revealed by the interpretation of the terms $\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)$ and $\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$, with $i \in \{1,2\}$ and $j \in \{1,2\}\setminus\{i\}$. The former represents the expected empirical risk induced by the Gibbs algorithm $P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}$ on the unseen dataset $\boldsymbol{z}_i$, whereas the latter represents the expected empirical risk induced by the Gibbs algorithm $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ on the training dataset $\boldsymbol{z}_i$. Hence, the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\left(\boldsymbol{z}_i,P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\right)$ represents the loss or gain of using the Gibbs algorithm $P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}$ or $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ over the data set $\boldsymbol{z}_i$. More specifically, the algorithm $P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}$ would induce a smaller expected empirical risk than the one induced by the algorithm $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ over the data set $\boldsymbol{z}_i$ if and only if

$$D\Big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\|P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)+D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_i\Big)-D\Big(P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_j}\|Q_i\Big)<0. \tag{53}$$

4.3 Sensitivity Analysis with Aggregate Datasets

The focus of this section is on the differences $\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right) - \mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right)$ and $\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right) - \mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$, with $i \in \{1,2\}$, which written in terms of sensitivity respectively yield $\mathsf{S}_{Q_0,\lambda_0}\left(\boldsymbol{z}_0, P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$ and $\mathsf{S}_{Q_i,\lambda_i}\left(\boldsymbol{z}_i, P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right)$.

The following theorem provides explicit expressions for $\mathsf{S}_{Q_0,\lambda_0}\left(\boldsymbol{z}_0, P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$.

Theorem 4.2. *Assume that the σ-finite measures Q_0, Q_1 and Q_2 in* (44) *are (pair-wise) mutually absolutely continuous. Hence, for all $i \in \{0,1,2\}$, the solution to the ERM-RER problem in* (44)*, denoted by $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$, satisfies:*

$$\begin{aligned}&\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right) - \mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right)\\ =&\lambda_0\left(D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) + D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|Q_0\Big) - D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_0\Big)\right),\end{aligned} \quad (54)$$

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

Proof: The proof of Theorem 4.2 is immediate from Theorem 3.1 by noticing that for all $i \in \{1,2\}$, the diference $\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right) - \mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right)$ can be written in terms of the sensitivity $\mathsf{S}_{Q_0,\lambda_0}\left(\boldsymbol{z}_0, P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$. ∎

The aggregate sensitivity $\mathsf{S}_{Q_0,\lambda_0}\left(\boldsymbol{z}_0, P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$ in Theorem 4.2 is reminiscent to the generalization error or population error [17], except that instead of an assumption on the probability measure on the data sets, additional data is required to evaluate the generalization capabilities. More specifically, the difference $\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right) - \mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right)$ represents the gain or loss on expected empirical risk over the aggregated dataset $\boldsymbol{z}_0$ obtained by using the Gibbs algorithm $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ instead of the Gibbs algorithm $P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}$. Remarkably, the Gibbs algorithm obtained with all available data (the aggregate set $\boldsymbol{z}_0$) induces a larger expected empirical risk than the one obtained using only the constituent dataset i, that is, $\mathsf{S}_{Q_0,\lambda_0}\left(\boldsymbol{z}_0, P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right) < 0$, if and only if

$$D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) + D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|Q_0\Big) - D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_0\Big) < 0. \quad (55)$$

The following corollary of Theorem 4.2 is obtained by subtracting the equalities in (54) for $i = 1$ and $i = 2$.

Corollary 4.1. *Assume that the σ-finite measures Q_0, Q_1 and Q_2 in* (44) *are (pair-wise) mutually absolutely continuous. Hence, for all $i \in \{0,1,2\}$, the* *solution to the ERM-RER problem in* (44)*, denoted by* $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$*, satisfies:*

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)\\
=&\lambda_0\Bigg(D\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) - D\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q_0\Big) - D\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) \quad (56)\\
&+D\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q_0\Big)\Bigg), \quad (57)
\end{aligned}$$

where, the functions $\mathsf{R}_{\boldsymbol{z}_1}$ *and* $\mathsf{R}_{\boldsymbol{z}_2}$ *are defined in* (7)*.*

Corollary 4.1 allows the comparison of the Gibbs algorithms $P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ and $P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}$ with respect to the same dataset, i.e., the aggregated data set $\boldsymbol{z}_0$. Note for instance that the Gibbs algorithm $P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ induces a smaller expected empirical risk over the aggregated dataset than the one induced by the Gibbs algorithm $P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}$ if and only if

$$\begin{aligned}
&D\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) - D\Big(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q_0\Big) \quad (58)\\
<&D\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\Big) - D\Big(P^{(Q_2,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q_0\Big). \quad (59)
\end{aligned}$$

Consider now the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\left(\boldsymbol{z}_i, P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right)$ with $i \in \{1,2\}$, which is characterized by the following theorem.

Theorem 4.3. *Assume that the σ-finite measures Q_0, Q_1 and Q_2 in* (44) *are (pair-wise) mutually absolutely continuous. Hence, for all $i \in \{0,1,2\}$, the solution to the ERM-RER problem in* (44)*, denoted by* $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$*, satisfies:*

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right) - \mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)\\
=&\lambda_i\left(D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big) + D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_i\Big) - D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|Q_i\Big)\right), \quad (60)
\end{aligned}$$

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

Proof: The proof of Theorem 4.3 is immediate from Theorem 3.1 by noticing that for all $i \in \{1,2\}$, the diferences $\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right) - \mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$ can be written in terms of the sensitivity $\mathsf{S}_{Q_i,\lambda_i}\left(\boldsymbol{z}_i, P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right)$. ■

More specifically, for all $i \in \{0,1,2\}$, the difference $\mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\right) - \mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right)$ represents the gain or loss on expected empirical risk over the aggregated dataset $\boldsymbol{z}_i$ obtained by using the Gibbs algorithm $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$ instead of the Gibbs algorithm $P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}$.

Interestingly, the Gibbs algorithm that uses all available data induces a smaller expected empirical risk (over the constituent dataset $\boldsymbol{z}_i$) than the one induced by the Gibbs algorithm that uses only the constituent dataset $\boldsymbol{z}_i$ if and only if

$$D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big) + D\Big(P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\|Q_i\Big) - D\Big(P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}\|Q_i\Big) < 0. \tag{61}$$

4.4 Homogeneous Priors and Proportional Regularization

In this section two assumptions are made. Firstly, given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it is assumed that for all $i \in \{0,1,2\}$ and for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$,

$$Q(\mathcal{A}) = Q_i(\mathcal{A}). \tag{62}$$

Secondly, the parameters λ_0, λ_1, and λ_2 in (44) satisfy

$$\lambda_1 = \frac{n_0}{n_1}\lambda_0 \quad \text{and} \quad \lambda_2 = \frac{n_0}{n_2}\lambda_0, \tag{63}$$

with n_0, n_1, and n_2 integers satisfying (43b). These assumptions are referred to as the case of *homogeneous priors* with measure Q, and the case of *proportional regularization* with parameter λ_0, respectively.

Under these assumptions, the following corollary follows from Theorem 4.1 and Lemma 4.1.

Corollary 4.2. *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 solution to the ERM-RER problem in* (44)*, denoted by $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$, satisfies:*

$$\begin{aligned}
&\Big(\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big) - \frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big)\Big)\\
&+\Big(\frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big) - \frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)\Big)\\
&= \lambda_0\left(D\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big) + D\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big)\right),
\end{aligned} \tag{64}$$

where the functions $\mathsf{R}_{\boldsymbol{z}_1}$ and $\mathsf{R}_{\boldsymbol{z}_2}$ are defined in (7)*; and the integers n_0, n_1, and n_2 satisfy* (43b).

For all $i \in \{1,2\}$ and $j \in \{1,2\} \setminus \{i\}$, the Gibbs algorithm $P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$, which is obtained by using the training set $\boldsymbol{z}_i$, induces an expected empirical risk $\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)$ over the training set; and an expected empirical risk $\mathsf{R}_{\boldsymbol{z}_j}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big)$ over the unseen dataset $\boldsymbol{z}_j$. The coefficient $\frac{n_i}{n_0}$ weighs the function $\mathsf{R}_{\boldsymbol{z}_i}$ proportionally to the number of datapoints in the dataset $\boldsymbol{z}_i$. Hence, the difference

$$\frac{n_j}{n_0}\mathsf{R}_{\boldsymbol{z}_j}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big) - \frac{n_i}{n_0}\mathsf{R}_{\boldsymbol{z}_i}\Big(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\Big) \tag{65}$$

is reminiscent to a *validation* [18, Section 11.2] for the special case in which the probability distribution of the data sets is unknown and thus, only available datasets can be used. More specifically, the dataset $\boldsymbol{z}_j$ can be assumed to be a validation dataset for the Gibbs algorithm $P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$. Hence, from this perspective, the most performing Gibbs algorithm is $P^{(Q,\lambda_k)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_k}$, with $k \in \{1,2\}$, such that

$$k = \arg\min_{i \in \{1,2\}} \frac{n_j}{n_0} \mathsf{R}_{\boldsymbol{z}_j}\left(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right) - \frac{n_i}{n_0} \mathsf{R}_{\boldsymbol{z}_i}\left(P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}\right), \text{ with } j \in \{1,2\} \setminus \{i\}. \quad (66)$$

In (64), it holds that $D\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1} \| P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big) \geqslant 0$ and $D\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2} \| P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big) \geqslant 0$, which leads to the following corollary of Theorem 4.1.

Corollary 4.3. *Consider the case of homogeneous priors with a σ-finite measure Q and proportional regularization. Then, for all $i \in \{1,2\}$, the solution to the ERM-RER problem in* (44)*, denoted by $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$, satisfies:*

$$\begin{aligned}
&\Big(\frac{n_1}{n_0} \mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right) + \frac{n_2}{n_0} \mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)\Big) \\
\geqslant &\Big(\frac{n_1}{n_0} \mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) + \frac{n_2}{n_0} \mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)\Big), \qquad (67)
\end{aligned}$$

where, the functions $\mathsf{R}_{\boldsymbol{z}_1}$ and $\mathsf{R}_{\boldsymbol{z}_2}$ are defined in (7)*; and the integers n_0, n_1, and n_2 satisfy* (43b).

Corollary 4.3 highlights that the weighted-sum of the expected empirical risks respectively induced by the Gibbs algorithms $P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ and $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}$ over their unseen datasets is not smaller than the weighted sum of the expected empirical risks respectively induced by those algorithms over their training datasets.

The following theorem provides an alternative expression for the difference $\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)$ in Corollary 4.1 under the assumption of homogeneous priors and proportional regularization.

Theorem 4.4. *Consider the case of homogeneous priors with a σ-finite measure Q and proportional regularization. Hence, for all $i \in \{1,2\}$, the solution to the ERM-RER problem in* (44)*, denoted by $P^{(Q_i,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_i}$, satisfies:*

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right) \\
=&\lambda_0 \Bigg(D\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1} \| P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\Big) + 2D\Big(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1} \| Q\Big) - D\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2} \| P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\Big) \\
&-2D\Big(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2} \| Q\Big) \Bigg), \qquad (68)
\end{aligned}$$

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

Proof: In the case of proportional regularization, the following holds from Theorem 4.1,

$$\frac{n_2}{n_0}\left(\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)\right)$$
$$=\lambda_0\left(D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)+D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q\right)-D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q\right)\right), \text{ and} \quad (69)$$
$$\frac{n_1}{n_0}\left(\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)-\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)\right)$$
$$=\lambda_0\left(D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)+D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q\right)-D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q\right)\right). \quad (70)$$

The substraction of the equality in (70) from the equality in (69) yields

$$\frac{n_2}{n_0}\left(\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)\right)-\frac{n_1}{n_0}\left(\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)-\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)\right)$$
$$=\lambda_0\Bigg(D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)+D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q\right)-D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q\right)$$
$$-D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q\right)+D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q\right)\Bigg). \quad (71)$$

The left-hand side of the equality in (71) satisfies the following equalities:

$$\frac{n_2}{n_0}\left(\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)\right)-\frac{n_1}{n_0}\left(\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)-\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)\right)$$
$$=\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)+\frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-\frac{n_1}{n_0}\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)$$
$$-\frac{n_2}{n_0}\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right) \quad (72)$$
$$=\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right). \quad (73)$$

Plugging (73) into (71) yields,

$$\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-\mathsf{R}_{\boldsymbol{z}_0}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)$$
$$=\lambda_0\Bigg(D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)+D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q\right)-D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q\right) \quad (74)$$
$$-D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)-D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q\right)+D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q\right)\Bigg) \quad (75)$$
$$=\lambda_0\left(D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\right)-D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)\right) \quad (76)$$
$$+2\lambda_0\left(D\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_2}\|Q\right)-D\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\|Q\right)\right). \quad (77)$$

This completes the proof. ■

The advantage of the expression in (68) with respect to the one in (57) is that it does not depend on the probability measure $P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}$. That is, it does not require to implement the Gibbs algorithm $P^{(Q_0,\lambda_0)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_0}$.

5 Conclusions and Final Remarks

Using the notion of sensitivity (Definition 3.1), the exact difference between the expected empirical risks induced by Gibbs algorithms obtained with different training datasets has been presented. More specifically, given two datasets (constituent datasets), three Gibbs algorithms are obtained by using each of the constituent datasets and the aggregate dataset as training data. In this context, explicit expressions for the differences of the following quantities have been obtained:
(a) The expected empirical risks (over a constituent dataset) induced by the Gibbs algorithms respectively trained with constituent datasets (Theorem 4.1);
(b) The expected empirical risks (over the aggregated dataset) induced by the Gibbs algorithm respectively trained with the aggregated dataset and a constituent dataset (Theorem 4.2);
(c) The expected empirical risks (over the aggregated dataset) induced by the Gibbs algorithms respectively trained with constituent datasets (Corollary 4.1); and
(d) The expected empirical risks (over a constituent dataset) induced by the Gibbs algorithm respectively trained with the aggregated dataset and a constituent dataset (Theorem 4.3);
These differences, which correspond to forms of sensitivity, are in terms of the relative entropy of the probability measures associated with the Gibbs algorithms (solutions to the corresponding ERM-RER problems) and the reference measures associated to each dataset. This reveals the impact of the choice of these reference measures on the expected empirical risk induced by the corresponding algorithms.

The sensitivities in (a) - (d) arise as performance metrics that are relevant when a probability measure over the datasets is not available and thus, expectations cannot be performed. This is even more relevant in the case in which different datasets are obtained from unknown and different probability distributions. This justifies that in this work, different references measures are associated to the constituent and aggregate datasets. In some special cases, e.g, homogeneous regularization and proportional regularization, it is shown that the expressions in (a) - (d) are further simplified and the symmetrized Kullback-Liebler divergence appears as one of the driving terms.

Finally, equipped with this analytical insights, necessary and sufficient conditions for data aggregation to improve the performance of the Gibbs algorithms are obtained.

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 de Recherche de Sophia Antipolis Méditérranée, Sophia Antipolis, 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," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, 2016.

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

[6] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. of the Thirty-first Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2017.

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

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

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

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

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

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

[13] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," *Neural Information Processing Systems*, 1999.

[14] J. Zhu and E. P. Xing, "Maximum entropy discrimination Markov networks," *Journal of Machine Learning Research*, vol. 10, no. 11, 2009.

[15] O. Catoni, *PAC-Bayesian supervised classification: The thermodynamics of statistical learning*, 1st ed. Beachwood, OH, USA: Institute of Mathematical Statistics Lecture Notes - Monograph Series, 2007, vol. 56.

[16] B. Guedj, "A primer on PAC-Bayesian learning." in *Tutorials of the International Conference on Machine Learning (ICML)*, Jun. 2019.

[17] 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. 4, pp. 831–838, 2021.

[18] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2014.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau -
Rocquencourt
BP 105 - 78153 Le Chesnay
Cedex
inria.fr