Empirical Risk Minimization with Relative Entropy Regularization: Optimality and Sensitivity Analysis

Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, Stefano Rini

Empirical Risk Minimization with Relative Entropy Regularization: Optimality and Sensitivity Analysis

Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini

***Abstract*—The optimality and sensitivity of the empirical risk minimization problem with relative entropy regularization (ERM-RER) are investigated for the case in which the reference is a σ-finite measure instead of a probability measure. This generalization allows for a larger degree of flexibility in the incorporation of prior knowledge over the set of models. In this setting, the interplay of the regularization parameter, the reference measure, the risk function, and the empirical risk induced by the solution of the ERM-RER problem is characterized. This characterization yields necessary and sufficient conditions for the existence of regularization parameters that achieve arbitrarily small empirical risk with arbitrarily high probability. Additionally, the sensitivity of the expected empirical risk to deviations from the solution of the ERM-RER problem is studied. Dataset-dependent and dataset-independent upper bounds on the absolute value of the sensitivity are presented. In a special case, it is shown that the expectation (with respect to the datasets) of the absolute value of the sensitivity is upper bounded, up to a constant factor, by the square root of the lautum information between the models and the datasets.**

I. Introduction

The problem of empirical risk minimization (ERM) [1], which is strongly related to M-estimation [2], minimum contrast estimation [3], and sample average approximation [4], appears in numerous central problems in machine learning [5]. Among the most popular methods for solving the ERM problem are those based on the gradient. The stochastic gradient descent algorithm [6] and its variants fall within this class of methods. See, for instance, the literature reviews in [7], [8] and [9]. Other methods for solving the ERM problem are based on constructing probability measures over the mesurable space formed by the set of models. In these methods, the model is sampled from a particular probability distribution and thus, the figure of merit is the expectation of the empirical risk with respect to such probability distribution. Methods such as Bayesian methods [10] and PAC-Bayesian methods [11], [12] are typical examples. When a prior on the models is available, a typical choice is to regularize the original ERM problem by the relative entropy with respect to the prior. This problem is know as the ERM with relative entropy regularization (ERM-RER) and has been widely studied in the context of statistical physics [13], [14] and information theory [15]–[19]. Often, the prior is in the form of a probability distribution. Nonetheless, it has been shown that the ERM-RER problem can be generalized to the case in which priors are in the form of σ-finite measures [20]. In particular, when the prior is, for instance, the Lebesgue measure or a counting measure, the ERM-REM problem boils down to the well known problems of ERM with differential and discrete entropy regularization, respectively. See, for instance [21]–[23]. Interestingly, even in the case in which priors are σ-finite measures, the ERM-RER problem is shown to have a unique solution [20].

The contribution of this paper is twofold. First, it introduces a notion of optimality in probability for the ERM-RER problem, which is reminiscent to a probably approximately correct (PAC) guarantee [24]. More specifically, a solution to the ERM-RER problem is said to be (δ, ϵ)-optimal if the set of models that induce empirical risks smaller than δ exhibits a probability higher than $1 - \epsilon$. Necessary and sufficient conditions on the parameters of the ERM-RER problem for observing a (δ, ϵ)-optimal solution are presented. Second, the sensitivity of the ERM-RER problem is studied for a given dataset. The sensitivity is defined as the difference between two quantities: (a) the expectation of the empirical risk for a given dataset with respect to a given probability measure P on the set of models; and (b) The expectation of the empirical risk for such dataset with respect to the measure that is the solution to the ERM-RER problem. In particular, an upper bound on the absolute value of the sensitivity is presented. Such upper-bound is expressed in terms of the square root of the relative entropy of P with respect to the probability measure solution to the ERM-RER problem. As a byproduct, it is shown that the expectation of the absolute value of the sensitivity with respect to the probability distribution of the datasets is bounded. In a special case, it is shown that such bound is, up to a constant term, the square root of the lautum information [25] between the models and the datasets. This result is analogous to the results in [26]–[28], in which, under certain conditions, the generalization gap is upper bounded by a term that is proportional to the square root of the mutual information between the models and the datasets.

Samir M. Perlaza, and Alain Jean-Marie are with INRIA, 2004 Route des Lucioles, 06902 Sophia Antipolis, France. ({samir.perlaza, alain.jean-marie}@inria.fr)

Gaetan Bisson and Samir M. Perlaza are with the Laboratoire de Mathématiques GAATI, Université de la Polynésie Française, BP 6570, 98702 Faaa, French Polynesia. (bisson@gaati.org)

Iñaki Esnaola is with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, United Kingdom. (esnaola@sheffield.ac.uk)

Stefano Rini is with the Department of Electrical and Computer Engineering, National Chao Tung University, Hsinchu, Taiwan.

Samir M. Perlaza and Iñaki Esnaola are also with the Department of Electrical and Computer Engineering, Princeton University, Princeton, 08544 NJ, USA.

This work was supported in part by the INRIA Exploratory Action "Information and Decision Making (IDEM)"; and in part by the Agence Nationale de la Recherche under grant number ANR-20-CE40-0013.

The paper is organized as follows. Section II introduces the ERM and the ERM-RER problem. The solution to the ERM-RER, introduced in [20], is discussed. Section III introduces the notions of (δ, ϵ)-optimality, coherent measures, and consistent measures. Necessary and sufficient conditions for guaranteeing that the solution to the ERM-RER problem is (δ, ϵ)-optimal are presented. Section IV introduces the notion of sensitivity and presents upper-bounds on the absolute value of the sensitivity of the ERM-RER problem. These upper-bounds can be divided into two classes, dataset-dependent bounds and dataset-independent bounds. The former hold for a given dataset, whereas the latter hold for the expectation of the absolute value of the sensitivity with respect to the probability distribution of the data. Section V concludes the paper.

II. Problem Formulation

A. Empirical Risk Minimization

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,

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

The random variables X and Y jointly form the probability space:

$$(\mathcal{X} \times \mathcal{Y}, \mathscr{F}(\mathcal{X} \times \mathcal{Y}), P_{XY}), \tag{2}$$

where $\mathscr{F}(\mathcal{X} \times \mathcal{Y})$ is a σ-algebra on the set $\mathcal{X} \times \mathcal{Y}$, which is assumed to be fixed in this analysis. The elements of the sets $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$ are often referred to as *models*, *patterns* and *labels*, respectively. A pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is referred to as a *labeled pattern* or *data point* under the following condition.

Definition 1 (Data Point)**.** *The pair (x, y) is said to be a data point if $(x, y) \in \operatorname{supp} P_{XY}$.*

Several data points form a dataset.

Definition 2 (Dataset)**.** *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 $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$.*

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 $x \in \mathcal{X}$, the assigned label $f(\boldsymbol{\theta}, x)$ minimizes a notion of *loss* or *risk*. Let the function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty) \tag{3}$$

be such that given a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss or risk induced by choosing the model $\boldsymbol{\theta} \in \mathcal{M}$ is $\ell(f(\boldsymbol{\theta}, x), y)$. 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 $y \in \mathcal{Y}$, the loss $\ell(y, y) = 0$, which implies that correct labelling induces zero cost. Note that there might exist several models $\boldsymbol{\theta} \in \mathcal{M} \setminus \{\boldsymbol{\theta}^\star\}$ such that $\ell(f(\boldsymbol{\theta}, x), y) = 0$, which reveals the need of a large number of labeled patterns for model selection.

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

$$\boldsymbol{z} = \big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n, \tag{4a}$$

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

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

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

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

whose solutions form the set denoted by

$$\mathcal{T}(\boldsymbol{z}) \triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{4d}$$

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 $\boldsymbol{\theta}^\star \in \mathcal{T}(\boldsymbol{z})$ and $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}^\star) = 0$. Hence, the ERM problem in (4) is well posed.

B. Notation and Assumptions

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

Definition 3 (Generalized 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.

In the following, given a measurable space $(\Omega, \mathscr{F})$, the notation $\triangle(\Omega, \mathscr{F})$ is used to represent the set of σ-finite measures that can be defined over such a measurable space. Given a measure $Q \in \triangle(\Omega, \mathscr{F})$, the subset $\triangle_Q(\Omega, \mathscr{F})$ contains all measures that are absolutely continuous with respect to the measure Q. Given a set $\mathcal{A} \subset \mathbb{R}^d$, with $d \in \mathbb{N}$, the Borel σ-field over $\mathcal{A}$ is denoted by $\mathscr{B}(\mathcal{A})$.

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

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

where the functions f and ℓ are those in (1) and (3), is Borel measurable with respect to the measure space $\big(\mathcal{M} \times \mathcal{X} \times \mathcal{Y}, \mathscr{B}(\mathcal{M}) \times \mathscr{F}(\mathcal{X} \times \mathcal{Y})\big)$.

C. Generalized Relative Entropy Regularization

Under the assumptions above, when models are chosen by *sampling* from a probability measure over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, one of the performance metrics is the *expected empirical risk*, which is introduced hereunder.

Definition 4 (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 (4b)*. Then, 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 in $\triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a positive real, which are referred to as the *reference measure* and the *regularization factor*, respectively. Let $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ be a σ-finite measure and let λ be a positive real. The ERM-RER problem, with parameters Q and λ, consists of the following optimization problem:

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

$$\text{s.t.} \int \mathrm{d}P(\boldsymbol{\theta}) = 1. \tag{8b}$$

where the dataset $\boldsymbol{z}$ is in (4a); and the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (7).

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

Lemma 1 (Theorem 2.1 in [20])**.** *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ 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 (4b)*. Let also the set $\mathcal{K}_{Q,\boldsymbol{z}} \subset \mathbb{R}$ be*

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

Then, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the solution to the ERM-RER problem in (8)*, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, is a unique probability measure whose Radon-Nikodym derivative with respect to Q satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

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

III. (δ, ϵ)-Optimality

This section introduces the notion of (δ, ϵ)-optimality. In particular, the focus is on the conditions on the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (4b) and the parameters Q and λ of the ERM-RER problem in (8) for observing expected empirical risks that are arbitrarily small with arbitrarily high probability.

Definition 5. *Given a pair $(\delta, \epsilon) \in [0, +\infty) \times (0, 1)$, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (11)*, which is the solution to the ERM-RER problem in* (8)*, is said to be (δ, ϵ)-optimal, if the set*

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}, \tag{12}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ in (4b)*, satisfies*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 1 - \epsilon. \tag{13}$$

For all $\delta > 0$, it holds that $\mathcal{T}(\boldsymbol{z}) \subset \mathcal{L}_{\boldsymbol{z}}(\delta)$, with the sets $\mathcal{T}(\boldsymbol{z})$ and $\mathcal{L}_{\boldsymbol{z}}$ in (4d) and (12), respectively. Hence, from Definition 5, it follows that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ assigns probability $(1 - \epsilon)$ to a set that contains the models that induce an empirical risk smaller than or equal to δ. In view of this, it is interesting to identify the conditions on the parameters Q and λ for which the solutions to the ERM-RER problem in (8) are (δ, ϵ)-optimal. In the following, such conditions are stated using the following definitions.

Definition 6. *The σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ in* (8) *is said to be coherent if, for all $\delta > 0$, it holds that*

$$Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0, \tag{14}$$

where the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ is defined in (12)*.*

In the case in which the σ-finite measure Q in (8) is coherent, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11) satisfies for all $\delta > 0$,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0. \tag{15}$$

This is a consequence of the fact that the measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q are mutually absolutely continuous [20, Lemma 2.6]. On the other hand, when the measure Q is noncoherent, it follows that $Q(\mathcal{T}(\boldsymbol{z})) = 0$, which implies $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{T}(\boldsymbol{z})) = 0$.

Definition 7. *The σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ in* (8) *is said to be consistent, if the set*

$$\mathcal{L}^{\star}_{\boldsymbol{z}} \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}\} \tag{16}$$

satisfies $Q(\mathcal{L}^{\star}_{\boldsymbol{z}}) > 0$, where the function $\mathsf{L}_{\boldsymbol{z}}$ is in (4b)*; and*

$$\delta^{\star} \triangleq \inf\{\delta \in [0, +\infty) : Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0\}. \tag{17}$$

Note that when Q is coherent, $\delta^{\star} = 0$, and thus, $\mathcal{L}^{\star}_{\boldsymbol{z}} = \mathcal{T}(\boldsymbol{z})$, with $\mathcal{T}(\boldsymbol{z})$ in (4d). Moreover, if Q is coherent and consistent, then $Q(\mathcal{T}(\boldsymbol{z})) > 0$. Using the elements above, the main result of this section is presented by the following theorem.

Theorem 1. *If the σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ in* (8) *is consistent, then for all $(\delta, \epsilon) \in (\delta^{\star}, +\infty) \times (0, 1)$, with $\delta^{\star}$ in* (17)*, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (10)*, such that the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (11) *is (δ, ϵ)-optimal.*

Proof: The proof is presented in [20, Theorem 3.1]. ■

A stronger optimality claim can be obtained when the reference measure is consistent and coherent, as shown by the following corollary of Theorem 1.

Corollary 1. *If the σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ in* (8) *is coherent and consistent, then, for all $(\delta, \epsilon) \in (0, +\infty) \times (0, 1)$, there exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (10)*, such that the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (11) *is (δ, ϵ)-optimal.*

IV. Sensitivity

A performance metric to evaluate the deviations of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}$ (Definition 4) from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11) towards an alternative probability measure P is the sensitivity, which was introduced in [20]. Deviations from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ towards an alternative probability measure P over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ might arise due to several reasons. For instance, if new datasets become available, a new ERM-RER problem can be formulated using a larger dataset obtained by aggregating the old and the new datasets [29]. Similarly, the parameters Q (the reference measure) and λ (the regularization factor) in (8) might be changed based on side-information leading to new ERM-RER problems and thus, to new probability measures. Other techniques different from ERM-RER might also be used to obtain a probability measure over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, e.g., Bayesian methods. Within this context, the sensitivity is a performance metric defined as follows.

Definition 8 (Sensitivity)**.** *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ 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 for all datasets $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ and for probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that*

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise,} \end{cases} \tag{18}$$

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 $\mathsf{R}_{\boldsymbol{z}}$ due to a deviation from $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to P is $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$.*

A. Dataset-Dependent Bounds

The following theorem introduces an upper bound on the sensitivity.

Theorem 2. *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, it holds that, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (10)*, and for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$,*

$$|\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)| \leqslant \sqrt{2B^2_{Q,\boldsymbol{z}} D\left(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}, \tag{19}$$

where the function $\mathsf{S}_{Q,\lambda}$ is defined in (18)*; and the constant $B_{Q,\boldsymbol{z}} \in [0, +\infty)$ is*

$$B^2_{Q,\boldsymbol{z}} = \sup_{\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}} K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right), \tag{20}$$

with $K^{(2)}_{Q,\boldsymbol{z}}$ being the second derivative of the function $K_{Q,\boldsymbol{z}}$ in (9)*.*

Proof: The proof is presented in [20, Appendix V]. ■

In Theorem 2, the second derivative of the function $K_{Q,\boldsymbol{z}}$ in (9) plays a central role. The function $K_{Q,\boldsymbol{z}}$ is continuous and differentiable infinitely many times in $(-\infty, 0)$ [20, Lemma 2.8]. Moreover, from [20, Lemma 2.10], it follows that if $\boldsymbol{\Theta}$ is the random vector that induces the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11), with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (4b) becomes the real random variable $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\Theta})$ whose mean, variance, and third cumulant are respectively $K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$, $K^{(2)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$, and $K^{(3)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$.

Theorem 2 establishes an upper and a lower bound on the increase and decrease of the expected empirical risk that can be obtained by deviating from the optimal solution of the ERM-RER in (8). More specifically, note that for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that,

$$\mathsf{R}_{\boldsymbol{z}}(P) \geqslant \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \sqrt{2B^2_{Q,\boldsymbol{z}} D\left(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}, \text{ and} \tag{21}$$

$$\mathsf{R}_{\boldsymbol{z}}(P) \leqslant \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \sqrt{2B^2_{Q,\boldsymbol{z}} D\left(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}. \tag{22}$$

The following theorem highlights the fact that the measure that minimizes the expected empirical risk subject to a constraint in the relative entropy with respect to the ERM-RER optimal measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (11) is also the solution to an ERM-RER problem with parameters Q and ω, for some specific $\omega > 0$.

Theorem 3. *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, and a nonnegative real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (10)*, consider the following optimization problem*

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P(\boldsymbol{\theta}), \tag{23a}$$

$$\text{subject to:} \quad D\left(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant c, \tag{23b}$$

with, c denoting a nonnegative constant; $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the probability measure in (8)*; and $\mathsf{L}_{\boldsymbol{z}}$ the function in* (4b)*. Then, the solution to the optimization problem in* (23) *is a probability measure $P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfying for all $\boldsymbol{\theta} \in \operatorname{supp} P$,*

$$\frac{\mathrm{d}P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\omega}\right) - \frac{1}{\omega}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right), \tag{24}$$

with $\omega \in (0, \lambda]$ such that

$$D\left(P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = c. \tag{25}$$

Proof: The proof is presented in [20, Appendix W]. ■

B. Dataset-Independent Bounds

Consider a probability measure, denoted by $P_{\boldsymbol{Z}} \in \triangle\Big((\mathcal{X} \times \mathcal{Y})^n, (\mathscr{F}(\mathcal{X} \times \mathcal{Y}))^n\Big)$, such that for all $\mathcal{A} \in$

$\left(\mathscr{F}\left(\mathcal{X}\times\mathcal{Y}\right)\right)^n$ of the form $\mathcal{A}=\mathcal{A}_1\times\mathcal{A}_2\times\ldots\times\mathcal{A}_n$ with $\mathcal{A}_i\in\mathscr{F}\left(\mathcal{X}\times\mathcal{Y}\right)$ and $i\in\{1,2,\ldots,n\}$, it holds that

$$P_{\boldsymbol{Z}}\left(\mathcal{A}\right)=\prod_{t=1}^{n}P_{XY}\left(\mathcal{A}_t\right), \tag{26}$$

where the probability measure P_{XY} is defined in (2). More specifically, $P_{\boldsymbol{Z}}\left(\mathcal{A}\right)$ is the probability measure induced by a random variable $\boldsymbol{Z}=\left(\left(X_1,Y_1\right),\left(X_2,Y_2\right),\ldots,\left(X_n,Y_n\right)\right)$, in which the n random variables $\left(X_1,Y_1\right),\left(X_2,Y_2\right),\ldots,\left(X_n,Y_n\right)$ are independent and identically distributed according to P_{XY}.

Let the set $\mathcal{K}_Q$, with $Q\in\triangle\left(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right)\right)$, be

$$\mathcal{K}_Q=\bigcap_{\boldsymbol{z}\in\operatorname{supp}P_{\boldsymbol{Z}}}\mathcal{K}_{Q,\boldsymbol{z}}, \tag{27}$$

where the set $\mathcal{K}_{Q,\boldsymbol{z}}$ is defined in (9) and the probability measure $P_{\boldsymbol{Z}}$ is defined in (26). The set $\mathcal{K}_Q$ in (27) can be empty for some choices of the σ-finite measure Q and empirical loss function $\mathsf{L}_{\boldsymbol{z}}$ in (4b). Nonetheless, from [20, Lemma 2.2], it follows that when Q is a probability measure, then,

$$\mathcal{K}_Q=\left(0,+\infty\right). \tag{28}$$

Using this notation, the following corollary of Theorem 2 provides an upper bound on the expectation of the sensitivity with respect to the probability measure $P_{\boldsymbol{Z}}$ in (26).

Corollary 2. *Given a σ-finite measure $Q\in\triangle\left(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right)\right)$, for all $\lambda\in\mathcal{K}_Q$, with $\mathcal{K}_Q$ in* (27)*, and for all probability measures $P\in\triangle_Q\left(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right)\right)$, it holds that*

$$\int\left|\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z},P\right)\right|\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})\leqslant\int\sqrt{2B_{Q,\boldsymbol{z}}^2D\left(P\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)}\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{29}$$

where $B_{Q,\boldsymbol{z}}$ is defined in (20)*; the probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ is the solution to the ERM-RER problem in* (8)*; and the probability measure $P_{\boldsymbol{Z}}$ is defined in* (26)*.*

In the following theorem, the expectation of the sensitivity with respect to the measure $P_{\boldsymbol{Z}}$ in (26) is shown to have an upper bound that can be expressed in terms of the lautum information between the models and the data sets.

Theorem 4. *Given a σ-finite measure $Q\in\triangle\left(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right)\right)$, for all $\lambda\in\mathcal{K}_Q$, with $\mathcal{K}_Q$ in* (27)*, it holds that*

$$\begin{aligned}&\int\left|\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z},P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right)\right|\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})\\&\leqslant\sqrt{2B_Q^2\int D\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}^{(Q,\lambda)}\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u})},\end{aligned} \tag{30}$$

where the probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ is the solution to the ERM-RER problem in (8)*; the probability measure $P_{\boldsymbol{Z}}$ is defined in* (26)*; the probability measure $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ is such that for all $\mathcal{A}\in\mathscr{B}\left(\mathcal{M}\right)$,*

$$P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\left(\mathcal{A}\right)=\int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\mathcal{A}\right)\mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right); \tag{31}$$

and the constant B_Q satisfies

$$B_Q^2=\sup_{\boldsymbol{z}\in\operatorname{supp}P_Z}B_{Q,\boldsymbol{z}}^2, \tag{32}$$

with $B_{Q,\boldsymbol{z}}$ defined in (20)*.*

Proof: The proof is presented in [20, Theorem 3.5]. ■

Given a σ-finite measure $Q\in\triangle\left(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right)\right)$ and a positive real $\lambda\in\mathcal{K}_Q$, with $\mathcal{K}_Q$ in (27), let $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$ be the random variables that jointly induce a probability measure $P_{\boldsymbol{Z}\boldsymbol{\Theta}}^{(Q,\lambda)}$ with marginals $P_{\boldsymbol{Z}}$ in (26) and $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ in (31). Under these assumptions, the right-hand side in (30) can be written in terms of the lautum information [25] between the random variables $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$, which is denoted by $\mathtt{L}\left(\boldsymbol{Z};\boldsymbol{\Theta}\right)$. More specifically, note that

$$\mathtt{L}\left(\boldsymbol{Z};\boldsymbol{\Theta}\right)=\int D\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}). \tag{33}$$

In a nutshell, it can be concluded that the expectation of $\left|\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z},P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right)\right|$ with respect to the measure $P_{\boldsymbol{Z}}$ in (26) is upper bounded by the lautum information between the random variables $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$, which represent the datasets and the models, respectively.

V. Final Remarks

This work focuses on a special case of the ERM problem in which the random variables X and Y in (1) are such that Y is deterministic given the ground truth model $\boldsymbol{\theta}^\star$ and a realization of the random variable X. That is, all data points in the dataset $\boldsymbol{z}$ in (4a) are pairs of patterns correctly labeled. A more practical case is that in which the random variables X and Y satisfy $Y=f\left(\boldsymbol{\theta}^\star,X\right)+W$, for some random variable W, which represents an additive noise. In this case, data points in the dataset $\boldsymbol{z}$ in (4a) are pairs of patterns and labels that are not necessarily correct (in the sense of (1)). Nonetheless, the results presented in this paper can be extended to such a case.

References

[1] V. Vapnik, "Principles of risk minimization for learning theory," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 4, Denver, CO, USA, Dec. 1991, pp. 831–838.

[2] L. A. Stefanski and D. D. Boos, "The calculus of M-estimation," *The American Statistician*, vol. 56, no. 1, pp. 29–38, Feb. 2002.

[3] L. Birge and P. Massart, "Rates of convergence for minimum contrast estimators," *Probability Theory and Related Fields*, vol. 97, pp. 113–150, Mar. 1993.

[4] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

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

[6] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[7] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, Sept. 1999.

[8] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[9] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 102–113, May 2020.

[10] C. P. Robert, *The Bayesian choice: From decision-theoretic foundations to computational implementation*, 1st ed. New York, NY: Springer, 2007.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 31, Montréal, Canada, 2018.

[18] A. Asadi and E. Abbe, "Chaining meets chain rule: Multilevel entropic regularization and training of neural networks," *Journal on Machine Learning Research*, vol. 21, pp. 1–32, Jan. 2020.

[19] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 4, New Orleans, LA, USA, Dec. 2021, pp. 831–838.

[20] 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.

[21] E. T. Jaynes, "Information theory and statistical mechanics I," *Physical Review Journals*, vol. 106, pp. 620–630, May 1957.

[22] ——, "Information theory and statistical mechanics II," *Physical Review Journals*, vol. 108, pp. 171–190, Oct. 1957.

[23] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 1999.

[24] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[25] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

[26] I. M. Alabdulmohsin, "Algorithmic stability and uniform generalization," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 28, Montréal, Canada, Dec. 2015.

[27] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 51, Cadiz, Spain, May 2016, pp. 1232–1240.

[28] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, Dec. 2017.

[29] S. M. Perlaza, I. Esnaola, and H. V. Poor, "Sensitivity of the Gibbs algorithm to data aggregation in supervised machine learning," Inria, Centre de Recherche de Sophia Antipolis Méditérranée, Sophia Antipolis, Tech. Rep. RR-9474, Jun. 2022.