Analysis of the Relative Entropy Asymmetry in the Regularization of Empirical Risk Minimization

Francisco Daunas, Iñaki Esnaola, Samir M Perlaza, H. Vincent Poor

Analysis of the Relative Entropy Asymmetry in the Regularization of Empirical Risk Minimization

Francisco Daunas*†, Iñaki Esnaola*‡, Samir M. Perlaza†‡§, and H. Vincent Poor‡
*ACSE Dept. University of Sheffield, Sheffield, United Kingdom. {jdaunastorres1, esnaola}@sheffield.ac.uk
†INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France. samir.perlaza@inria.fr
‡ECE Dept. Princeton University, Princeton, 08544 NJ, USA. poor@princeton.edu
§GAATI, Université de la Polynésie Française, Faaa, French Polynesia.

***Abstract*—The effect of the relative entropy asymmetry is analyzed in the empirical risk minimization with relative entropy regularization (ERM-RER) problem. A novel regularization is introduced, coined Type-II regularization, that allows for solutions to the ERM-RER problem with a support that extends outside the support of the reference measure. The solution to the new ERM-RER Type-II problem is analytically characterized in terms of the Radon-Nikodym derivative of the reference measure with respect to the solution. The analysis of the solution unveils the following properties of relative entropy when it acts as a regularizer in the ERM-RER problem: i) relative entropy forces the support of the Type-II solution to collapse into the support of the reference measure, which introduces a strong inductive bias that dominates the evidence provided by the training data; ii) Type-II regularization is equivalent to classical relative entropy regularization with an appropriate transformation of the empirical risk function. Closed-form expressions of the expected empirical risk as a function of the regularization parameters are provided.**

***Index Terms*—Empirical risk minimization; relative entropy; regularization; reference measure; inductive bias**

I. Introduction

Empirical risk minimization (ERM) is a central tool in supervised machine learning that enables the characterization, among others, of sample complexity and probably approximately correct (PAC) learning in a wide range of settings [1]. The application of ERM in the study of theoretical guarantees spans related disciplines such as machine learning [2], information theory [3], [4] and statistics [5], [6]. Classical problems such as classification [7], [8], pattern recognition [9], [10], regression [11], [12], and density estimation [9], [13] can be posed as special cases of the ERM problem [13], [14]. Unfortunately, ERM is prone to training data memorization, a phenomenon also known as overfitting [15]–[17]. For that reason, regularization is used to bound the sensitivity of the solution model to training data and provide generalization guarantees [18]–[20]. Regularization establishes a preference over the models by encoding features of interest that conform to prior knowledge.

In different statistical learning frameworks, such as Bayesian learning [21], [22] and PAC learning [23]–[25], the prior knowledge over the set of models can be described by a reference probability measure. Nonetheless, more general references can be adapted as proved in [26] for the case of σ-finite measures. In either case, the solution to the ERM problem can be cast as a probability distribution over all the candidate models. A common regularizer is the relative entropy of the solution with respect to the reference over the set of models [13], [27]–[29]. The resulting problem formulation, termed ERM with relative entropy regularization (ERM-RER) has been extensively studied and its unique solution is the Gibbs probability measure, for which the most salient properties are well understood [26]–[31]. Despite the many merits of the ERM-RER formulation, it has some significant limitations. Firstly, the definition of the relative entropy in terms of the Radon-Nikodym derivative of the solution with respect to the reference probability measure, sets a hard barrier to the exploration of models outside the support of the reference. These models are not given any consideration by the resulting Gibbs probability measure regardless of the evidence provided by the training dataset. Secondly, the choice of relative entropy over the alternatives often follows arguments based on upper bounds on the performance, which are hard to obtain and are not always informative when evaluated in practical settings [32]–[34]. For these reasons, exploring the asymmetry of the relative entropy is of particular interest to advancing the understanding of entropy regularization and its role in generalization.

Interestingly, there is no literature discussing the asymmetry of relative entropy in the context of ERM regularization. Hence, the issue of regularizing the ERM problem with the relative entropy of the reference with respect to the solution is an open problem. To differentiate between the two cases, we denote by Type-I the use of the relative entropy of the solution with respect to the reference; and by Type-II the use of the relative entropy of the reference with respect to the solution. This paper presents the solution to the Type-II ERM-RER problem and establishes a link to the Type-I ERM-RER problem via a transformation of the risk that can be cast as a tunable loss function [35]–[37].

The remainder of the paper is organized as follows. Section II presents the standard ERM problem. Section III describes the Type-I regularization. The main contribution of the paper is the solution to the Type-II ERM-RER presented in Section IV. Section V studies the equivalence between Type-I

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

and Type-II regularization. The conclusions are summarized in Section VI.

II. Empirical Risk Minimization Problem

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

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

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty), \tag{1}$$

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 $\ell(f(\boldsymbol{\theta}, x), y)$. The risk function ℓ is assumed to be nonnegative and satisfy $\ell(y, y) = 0$ for all $y \in \mathcal{Y}$. Nonetheless, there might exist other models $\boldsymbol{\theta} \in \mathcal{M}$ such that $\ell(f(\boldsymbol{\theta}, x'), y') = 0$ for the labelled data point (x', y'), revealing the need for a large number of labeled patterns for model selection.

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

$$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n, \tag{2}$$

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

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

The ERM problem is given by the optimization problem

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

and the set of solutions to the problem is denoted by

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

Note that if the set $\mathcal{M}$ is finite, the ERM problem in (4) has a solution, and therefore, it holds that $|\mathcal{T}(\boldsymbol{z})| > 0$. Nevertheless, in general, the ERM problem does not always have a solution; that is, there exist choices of the loss function ℓ and the dataset $\boldsymbol{z}$ that yield $|\mathcal{T}(\boldsymbol{z})| = 0$.

A. Statistical Learning

The Bayesian and PAC frameworks in [24] and [22] solve the problem by constructing probability measures $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ conditioned on the dataset $\boldsymbol{z}$, from which models are randomly sampled. In this context, finding probability measures that are minimizers of the ERM problem in (4) over the set $\triangle(\mathcal{M}, \mathscr{F})$ of all probability measures that can be defined on the measurable space $(\mathcal{M}, \mathscr{F})$, requires a metric that enables assessing the goodness of the probability measure. A common metric is the notion of expected empirical risk.

Definition 1 (Expected Empirical Risk): Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}, \mathscr{F}) \to [0, +\infty)$ be such that for all probability measures $Q \in \triangle(\mathcal{M}, \mathscr{F})$,

$$\mathsf{R}_{\boldsymbol{z}}(Q) \triangleq \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{6}$$

where the dataset $\boldsymbol{z}$ is defined in (2); and the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3).

The expected empirical risk is an important performance indicator of learning algorithms. However, it only gives an indication of the risk induced over the training dataset, while the performance of the ERM solutions is characterized by their generalization capability and sensitivity [26], [28], [29], [31]. In the following, we review the Type-I relative entropy regularization that serves as the basis for the analysis of the regularization asymmetry.

III. The Type-I ERM-RER Problem

The Type-I ERM-RER problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M}, \mathscr{F})$ and a positive real λ, where the measure Q is the *reference measure* and λ is the *regularization factor*. The Type-I ERM-RER problem, with parameters Q and λ, consists of the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(P \| Q), \tag{7}$$

where the dataset $\boldsymbol{z}$ is defined in (3), the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (6), and the optimization domain is

$$\triangle_Q(\mathcal{M}, \mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M}, \mathscr{F}) : P \ll Q\}, \tag{8}$$

where the notation $P \ll Q$ stands for P being absolutely continuous with respect to Q.

The solution to the Type-I ERM-RER problem in (7) is the Gibbs probability measure [26]–[28], which is presented by the following lemma.

Lemma 1 ([26, Lemma 1]): Given a probability measure $Q \in \triangle(\mathcal{M}, \mathscr{F})$ and a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ be such that for all $t \in \mathbb{R}$,

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

where the dataset $\boldsymbol{z}$ is defined in (2). Let also the set $\mathcal{K}_{Q,\boldsymbol{z}} \subseteq \mathbb{R}$ 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 $\boldsymbol{\theta} \in \operatorname{supp} Q$ and for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the solution of the Type-I ERM-RER problem in (7), is the unique probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle_Q(\mathcal{M}, \mathscr{F})$, whose Radon-Nikodym derivative with respect to Q satisfies that

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

IV. The Type-II ERM-RER Problem

The Type-II ERM-RER problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M}, \mathscr{F})$ and a positive real λ. The measure Q is referred to as the *reference measure* and λ as the *regularization factor*. Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (2), the Type-II ERM-RER problem, with parameters Q and λ, consists of the following optimization problem:

$$\min_{P \in \bigtriangledown_Q(\mathcal{M}, \mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P), \tag{12}$$

where $\boldsymbol{z}$ is defined in (3), the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (6), and the optimization domain is

$$\bigtriangledown_Q(\mathcal{M}, \mathscr{F}) \triangleq \{P \in \triangle(\mathcal{M}, \mathscr{F}) : Q \ll P\}. \tag{13}$$

A. The Solution to the Type-II ERM-RER Problem

The asymmetry of the relative entropy poses a distinct challenge when tackling the optimization problem given in (12). The approach that leads to the solution of the Type-I in (7) needs to be adapted to accommodate the challenges posed by the absolute continuity requirement in (13). The solution of the Type-II ERM-RER problem in (12) is presented in the following theorem.

Theorem 1: Given a measure $Q \in \triangle(\mathcal{M}, \mathscr{F})$ and a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $\bar{K}_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ be such that for all $t \in (0, \infty)$ it holds that

$$\bar{K}_{Q,\boldsymbol{z}}(t) = \beta, \tag{14}$$

where

$$\int \frac{t}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{15}$$

with $\mathsf{L}_{\boldsymbol{z}}$ being the function defined in (3). The function $\bar{K}_{Q,\boldsymbol{z}}$ in (14) is well defined for a subset of $(0, \infty)$, which is denoted by $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$, and satisfies

$$\bar{\mathcal{K}}_{Q,\boldsymbol{z}} \triangleq \left\{ t \in (0, \infty) : \int \frac{t}{\bar{K}_{Q,\boldsymbol{z}}(t) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta}) = 1 \right\}. \tag{16}$$

Then, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ and for all $\lambda \in \bar{\mathcal{K}}_{Q,\boldsymbol{z}}$, the solution to the optimization problem in (12) is the unique probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, whose Radon-Nikodym derivative with respect to the probability measure Q satisfies

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{17}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $\bar{K}_{Q,\boldsymbol{z}}$ are defined in (3) and (14), respectively.

Proof: The proof is divided into two parts. In the first part, an ancillary optimization problem is solved in a subset of the optimization domain of the Type-II ERM-RER problem. In the second part, it is shown that the solution obtained in this subset is, in fact, the solution of the Type-II ERM-RER problem.

The first part is as follows. Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (2), the ancillary optimization problem is given by:

$$\min_{P \in \bigcirc_Q(\mathcal{M}, \mathscr{F})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P), \tag{18}$$

where the optimization domain is

$$\bigcirc_Q(\mathcal{M}, \mathscr{F}) \triangleq \bigtriangledown_Q(\mathcal{M}, \mathscr{F}) \cap \triangle_Q(\mathcal{M}, \mathscr{F}), \tag{19}$$

and the sets $\triangle_Q(\mathcal{M}, \mathscr{F})$ and $\bigtriangledown_Q(\mathcal{M}, \mathscr{F})$ are respectively defined in (8) and (13). The solution to the ancillary optimization problem in (18) is presented by the following lemma.

Lemma 2: For all $\lambda \in \bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ with $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (14), the solution to the optimization problem in (18) is the unique probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (17).

Proof: From the fact that, for all $P \in \bigcirc_Q(\mathcal{M}, \mathscr{F})$, the measure Q is mutually absolute continuous with respect to P, the ancillary optimization problem in (18) can be written as follows:

$$\min_{P \in \bigcirc_Q(\mathcal{M}, \mathscr{F})} \Bigg[\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}Q(\boldsymbol{\theta}) \Bigg], \tag{20}$$

$$\text{s.t.} \quad \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{21}$$

The Lagrangian of the optimization problem in (20) can be constructed in terms of a function in the set $\mathscr{M}$ of nonnegative measurable functions with respect to the measurable spaces $\bigcirc_Q(\mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{F})$. Let $L : \mathscr{M} \times \mathbb{R} \to \mathbb{R}$ be the Langragian

$$L\left(\frac{\mathrm{d}P}{\mathrm{d}Q}, \beta \right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) - \lambda \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) + \beta \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) - 1 \right) \right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{22}$$

where β is a real value that acts as a Lagrange multiplier due to (21). The Gateaux differential [38] of the functional L in (22) at $\left(\frac{\mathrm{d}P}{\mathrm{d}Q}, \beta \right) \in \mathscr{M} \times \mathbb{R}$ in the direction of $h \in \mathscr{M}$ is

$$\partial L\left(\frac{\mathrm{d}P}{\mathrm{d}Q}, \beta; h \right) = \int h(\boldsymbol{\theta}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \lambda \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{23}$$

The relevance of the Gateaux differential in (23) stems from [38, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (22) to have a stationary point at $\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta \right) \in \mathcal{M} \times \mathbb{R}$ is that for all functions $h \in \mathscr{M}$, the following holds:

$$\partial L\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta; h \right) = 0. \tag{24}$$

From the fact that h is nonnegative, for all $\boldsymbol{\theta} \in \mathcal{M}$ it follows that

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \lambda \left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} + \beta = 0, \tag{25}$$

and thus,

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{26}$$

where the function $\bar{K}_{Q,\boldsymbol{z}}$ is defined in (14).

Finally, note that the objective function in (20) is the sum of two terms. The first one, i.e., $\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})$, is linear in $\frac{\mathrm{d}P}{\mathrm{d}Q}$. The second, i.e., $-\int \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\Big)\,\mathrm{d}Q(\boldsymbol{\theta})$, is strictly convex with $\frac{\mathrm{d}P}{\mathrm{d}Q}$. Hence, given that $\lambda > 0$, the sum of both terms is strictly convex with $\frac{\mathrm{d}P}{\mathrm{d}Q}$. This implies the uniqueness of $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. ■

This completes the first part of the proof of Theorem 1. The second part rests in the following lemma.

Lemma 3: For all $\lambda \in \bar{\mathcal{K}}_{Q,\boldsymbol{z}}$, with $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (16), it holds that

$$\min_{P\in\bigtriangledown_Q\setminus\bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda\mathsf{D}(Q\|P) > \min_{P\in\bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda\mathsf{D}(Q\|P). \tag{27}$$

Proof: The proof is presented in [39]. ■

More specifically, Lemma 3 conveys the fact that the relative entropy regularization penalty for considering models outside of the support is always greater than the reduction in the expected empirical risk induced by including these models. This includes the case in which the set $\mathcal{T}(\boldsymbol{z})$ in (5) lies outside of the support of Q.

From (19), it holds that

$$\bigcirc_Q(\mathcal{M},\mathscr{F}) \subseteq \ \bigtriangledown_Q(\mathcal{M},\mathscr{F}). \tag{28}$$

Hence, from (28), it follows that

$$\min_{P\in\bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P)+\lambda\mathsf{D}(Q\|P) \leq \min_{P\in\bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P)+\lambda\mathsf{D}(Q\|P). \tag{29}$$

From Lemma 3, it holds that

$$\min_{P\in\bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P)+\lambda\mathsf{D}(Q\|P) \geq \min_{P\in\bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P)+\lambda\mathsf{D}(Q\|P). \tag{30}$$

Thus, the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (17) is the solution of the optimization problem in (12), which completes the proof of Theorem 1. ■

B. Properties of the Solution

The properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (14) and the set $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (16) can be studied using the following mathematical objects. Given a positive real δ and the dataset $\boldsymbol{z}$ in (2), consider the set

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \triangleq \{\boldsymbol{\theta}\in\mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leq \delta\}, \tag{31}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3) and $\delta \in [0,\infty)$. Consider also the nonnegative real

$$\delta^{\star}_{Q,\boldsymbol{z}} \triangleq \inf\{\delta\in[0,\infty) : Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0\}, \tag{32}$$

with Q in (12). Let also $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ be the following level set of the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3):

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \triangleq \big\{\boldsymbol{\theta}\in\mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\big\}. \tag{33}$$

The following lemma introduces the properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (14).

Lemma 4: The function $\bar{K}_{Q,\boldsymbol{z}}$ in (14), for fixed Q and $\boldsymbol{z}$, is strictly increasing, continuous, and differentiable infinitely many times.

Proof: The proof is presented in [39]. ■

Note that from Lemma 4, the value $\bar{K}_{Q,\boldsymbol{z}}(\lambda)$ in (14) increases as the regularization factor λ increases, which is consistent with the notion that it acts as a scaling factor in (17). This highlights its dependence with the dataset $\boldsymbol{z}$ in (2) and the reference measure Q in (12).

Similarly, the set $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (16) also depends on the dataset $\boldsymbol{z}$ in (2) and the probability measure Q in (12). The following lemma presents the properties of the set $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (16).

Lemma 5: The set $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (16) is either the empty set or the set

$$\bar{\mathcal{K}}_{Q,\boldsymbol{z}} = (0,\infty). \tag{34a}$$

Moreover, for all $\lambda \in \bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ it holds that

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) \in \big(-\delta^{\star}_{Q,\boldsymbol{z}},\infty\big), \tag{34b}$$

with $\bar{K}_{Q,\boldsymbol{z}}$ defined in (14) and $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32).

Proof: The proof is presented in [39]. ■

Lemma 6 below shows that the expected empirical risk induced by the Type-II ERM-RER solution can be computed in terms of the regularization factor λ and the function $\bar{K}_{Q,\boldsymbol{z}}$ defined in (16). The relation of the expected empirical risk induced by $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (17) is presented by the following lemma.

Lemma 6: For all $\lambda \in \bar{\mathcal{K}}_{Q,\boldsymbol{z}}$, with $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (16), it holds that

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{35}$$

where the functions $\mathsf{R}_{\boldsymbol{z}}$ and $\bar{K}_{Q,\boldsymbol{z}}$ are respectively defined in (6) and (14); and the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is defined in (17).

Proof: The proof is presented in [39]. ■

The equality in (35) provides an upper bound to the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$. The following corollary of Lemma 6 formalizes this observation.

Corollary 7: For all $\lambda \in \bar{\mathcal{K}}_{Q,\boldsymbol{z}}$, with $\bar{\mathcal{K}}_{Q,\boldsymbol{z}}$ in (16), it holds that

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) < \lambda + \delta^{\star}_{Q,\boldsymbol{z}}, \tag{36}$$

where $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the probability measure in (17) and $\delta^{\star}_{Q,\boldsymbol{z}}$ is defined in (32).

The upper bound presented in Corollary 7 is useful as it gives operational meaning to the regularization factor. Indeed, this bound shows that the regularization factor governs the expected empirical risk increase with respect to the infimum of the empirical risk over the support.

C. Discussion on Regularization Properties

The Type-II relative entropy regularizer for the ERM problem in (12) allows for an exploratory minimization, *i.e.* models outside the support of the reference measure are given consideration. However, Theorem 1 shows that the support of the

probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (17) collapses into the support of the reference. A parallel can be established between Type-I and Type-II, as in both cases the support of the solution is the support of the reference measure. In a nutshell, the use of relative entropy regularization inadvertently forces the solution to coincide with the support of the reference regardless of the training data.

V. Interplay Between the Relative Entropy Asymmetry and the Risk

This section presents a connection between the Type-I ERM-RER in (7) and Type-II ERM-RER problems in (12). The log empirical risk is the function $\mathsf{V}_{\boldsymbol{z},\lambda} : \mathcal{M} \to \mathbb{R}$, which satisfies

$$\mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta}) \triangleq \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big), \tag{37}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $\bar{K}_{Q,\boldsymbol{z}}$ are defined in (3) and (14), respectively. For the case in which $\bar{\mathcal{K}}_{Q,\boldsymbol{z}} \neq \emptyset$, replacing the empirical risk in (4) by the notion of log empirical risk in (37) leads to the *expected log empirical risk*, as shown hereunder.

Definition 2 (Expected Log Empirical Risk): Given a dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, let the function $\bar{\mathsf{R}}_{\boldsymbol{z}} : \triangle(\mathcal{M}, \mathscr{F}) \to \mathbb{R}$ be such that for all probability measures $P \in \triangle(\mathcal{M}, \mathscr{F})$ and for all $\lambda \in (0, +\infty)$ it holds that

$$\bar{\mathsf{R}}_{\boldsymbol{z},\lambda}(P) \triangleq \int \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta}) \,\mathrm{d}P(\boldsymbol{\theta}), \tag{38}$$

where the function $\mathsf{V}_{\boldsymbol{z},\lambda}$ is defined in (37).

By considering the expected log empirical risk, an alternative formulation of the Type-I ERM-RER problem is presented. This formulation, also parametrized by Q and λ, consists in the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M},\mathscr{F})} \quad \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}(P) + \mathsf{D}(P\|Q). \tag{39}$$

Using the elements above, the main result of this section is presented in the following theorem.

Theorem 2: The solution to the optimization problem in (39) is the unique probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (17).

Proof: Denote by $\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the solution to the optimization problem in (39). Then, from Lemma 1, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\begin{aligned}
\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \frac{\exp(-\mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta}))}{\int \exp(-\mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\nu}))\,\mathrm{d}Q(\boldsymbol{\nu})} && \text{(40a)}\\
&= \frac{\exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big)}{\int \exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big)\mathrm{d}Q(\boldsymbol{\nu})} && \text{(40b)}\\
&= \frac{\Big(\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\mathrm{d}Q(\boldsymbol{\nu})\Big)^{-1}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} && \text{(40c)}\\
&= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && \text{(40d)}\\
&= \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), && \text{(40e)}
\end{aligned}$$

where equality (40b) follows from the definition of log empirical risk in (37); equality (40d) follows from (14) and (15); and equality (40e) follows from Theorem 1, which completes the proof. ∎

Theorem 2 establishes an equivalence between Type-I and Type-II regularization. It is shown therein that the direction of the relative entropy regularizer can be switched by appropriately transforming the risk function as shown in (37). Indeed, solving the Type-I ERM-RER problem with the expected log empirical risk defined in (38) yields the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that is the solution to the Type-II ERM-RER problem. In view of this, it is not surprising that the support for the probability measure that is the solution to the Type-II ERM-RER collapses into the support of the reference measure. In fact, the mutual absolute continuity between the solution and the reference probability measures is a consequence of the relative entropy regularization, regardless of its direction. Type-I regularization forces the support of the solution to include all the models in the support of the reference measure; on the other hand, Type-II regularization constrains the models in the support of the solution to the models in the support of the reference measure.

VI. Final Remarks

This work has introduced the Type-II ERM-RER problem and has presented its solution through Theorem 1. The solution highlights that regardless of the direction in which relative entropy is used as a regularizer, the models that are considered by the solution are necessarily in the support of the reference measure. In that sense, the restriction over the models introduced by the reference measure cannot be bypassed by the training data when relative entropy is used as the regularizer. We have shown that this is a consequence of the equivalence that can be established between Type-I and Type-II regularization. Remarkably, the direction of the relative entropy regularizer can be switched by a logarithmic transformation of the risk. The mutual absolute continuity of both Type-I and Type-II ERM-RER solutions relative to the reference measure can be understood in the light of the equivalence between both types of regularization. The analytical results have also enabled us to provide an operationally meaningful characterization of the expected empirical risk induced by the Type-II solution in terms of the regularization parameters. This is turn reduces the computational burden of bounding the expected empirical risk. Moreover, the insight provided by the bounds on the expected empirical risk can be distilled into guidelines for the selection of the regularization parameter.

References

[1] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in Neural Information Processing Systems*, vol. 4, pp. 831–838, Jan. 1992.

[2] V. N. Vapnik and A. Y. Chervonenkis, "On a perceptron class," *Avtomatika i Telemkhanika*, vol. 25, no. 1, pp. 112–120, Feb. 1964.

[3] M. R. Rodrigues and Y. C. Eldar, *Information-theoretic Methods in Data Science*, 1st ed. Cambridge, UK: Cambridge University Press, 2021.

[4] M. Mezard and A. Montanari, *Information, Physics, and Computation*, 1st ed. New York, NY, USA: Oxford University Press, 2009.

[5] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, 1st ed. New York, NY, USA: Cambridge University Press, 2019.

[6] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, 1st ed. New York, NY, USA: Cambridge University Press, 2018.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *Journal of the ACM (JACM)*, vol. 36, no. 4, pp. 929–965, Oct. 1989.

[8] I. Guyon, V. Vapnik, B. Boser, L. Bottou, and S. A. Solla, "Structural risk minimization for character recognition," *Advances in Neural Information Processing Systems*, vol. 4, Dec. 1991.

[9] G. Lugosi and K. Zeger, "Nonparametric estimation via empirical risk minimization," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 677–687, May 1995.

[10] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 525–536, Mar. 1998.

[11] V. Vapnik and L. Bottou, "Local algorithms for pattern recognition and dependencies estimation," *Neural Computation*, vol. 5, no. 6, pp. 893–909, Nov. 1993.

[12] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik, "Model complexity control for regression using VC generalization bounds," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1075–1089, Sep. 1999.

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

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

[15] A. Krzyzak, T. Linder, and C. Lugosi, "Nonparametric estimation and classification using radial basis function nets and empirical risk minimization," *IEEE Transactions on Neural Networks*, vol. 7, no. 2, pp. 475–487, Mar. 1996.

[16] W. Deng, Q. Zheng, and L. Chen, "Regularized extreme learning machine," in *Proceedings of the IEEE Symposium on Computational Intelligence in Data Mining (CIDM)*, Nashville, TN, USA, Apr. 2009, pp. 389–395.

[17] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, Aug. 2017, pp. 233–242.

[18] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, no. 1, pp. 499–526, Mar. 2002.

[19] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Measures of complexity: Festschrift for Alexey Chervonenkis*, vol. 16, no. 2, pp. 11–30, Oct. 2015.

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

[21] C. P. Robert, *The Bayesian Choice: From Decision-theoretic Foundations to Computational Implementation*, 1st ed. New York, NY, USA: Springer, 2007.

[22] D. A. McAllester, "Some PAC-Bayesian theorems," in *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, Madison, WI, USA, Jul. 1998, pp. 230–234.

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

[24] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proceedings of the 10th Annual Conference on Computational Learning Theory (COLT)*, Nashville, TN, USA, Jul. 1997, pp. 2–9.

[25] D. Cullina, A. N. Bhagoji, and P. Mittal, "PAC-learning in the presence of adversaries," *Advances in Neural Information Processing Systems*, vol. 31, no. 1, pp. 1–12, Dec. 2018.

[26] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization: Optimality and sensitivity," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, Jul. 2022, pp. 684–689.

[27] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Cambridge, UK, Sep. 2016, pp. 26–30.

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

[29] B. Zou, L. Li, and Z. Xu, "The generalization performance of ERM algorithm with strongly mixing observations," *Machine Learning*, vol. 75, no. 3, pp. 275–295, Feb. 2009.

[30] G. Aminian, Y. Bu, L. Toni, M. R. D. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," arXiv preprint arXiv:2210.09864, Oct. 2022.

[31] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "On the validation of Gibbs algorithms: Training datasets, test datasets and their aggregation," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.

[32] X. Wang and Q. He, "Enhancing generalization capability of SVM classifiers with feature weight adjustment," in *Proceedings of the Knowledge-Based Intelligent Information and Engineering Systems: 8th International Conference (KES)*, Wellington, New Zealand, Sep. 2004, pp. 1037–1043.

[33] Q. Lin, Z. Lu, and L. Xiao, "An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization," arXiv preprint arXiv:1407.1296, Jul. 2014.

[34] X. Yang and D. Li, "Estimation of the empirical risk-return relation: A generalized-risk-in-mean model," *Journal of Time Series Analysis*, vol. 43, no. 6, pp. 938–963, May 2022.

[35] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "A tunable measure for information leakage," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 701–705.

[36] T. Sypherd, M. Diaz, L. Sankar, and P. Kairouz, "A tunable loss function for binary classification," in *Proceedings of the IEEE international symposium on information theory (ISIT)*, Paris, France, Jul. 2019, pp. 2479–2483.

[37] G. R. Kurri, T. Sypherd, and L. Sankar, "Realizing GANs via a tunable loss function," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, virtual conference, 2021, pp. 1–6.

[38] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: Wiley, 1997.

[39] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Empirical risk minimization with relative entropy regularization type-II," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9508, May. 2023.