Title Journal

Author 1, Author 2, Author 3, and Author 4

Abstract

Abstract of the Journal

Index Terms

Supervised Learning, Empirical Risk Minimization; Relative Entropy; Regularization; Gibbs Measure; Inductive Bias; Gibbs Algorithm; Sensitivity; and Generalization.

I. Introduction

EMPIRICAL risk minimization (ERM) is a central tool in supervised machine learning. Among other uses, it enables the characterization of sample complexity and probably approximately correct (PAC) learning in a wide range of settings [**?**]. The application of ERM in the study of theoretical guarantees spans related disciplines such as machine learning [**?**], information theory [**?**], [**?**] and statistics [**?**], [**?**]. Classical problems such as classification [**?**], [**?**], pattern recognition [**?**], [**?**], regression [**?**], [**?**], and density estimation [**?**], [**?**] can be posed as special cases of the ERM problem [**?**], [**?**]. Unfortunately, ERM is prone to training data memorization, a phenomenon also known as overfitting [**?**], [**?**], [**?**]. For that reason, ERM is often regularized in order to provide generalization guarantees [**?**], [**?**], [**?**], [**?**]. 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 [**?**], [**?**] and PAC learning [**?**], [**?**], [**?**], the prior knowledge over the set of models can be described by a reference probability measure. More general references can be adapted as proved in [**?**], [**?**] for the case of σ-finite measures. Prior knowledge of the set of datasets can also be represented by probability measures, e.g., the worst-case data-generating probability measure introduced in [**?**]. In either case, the solution to the regularized ERM problem can be cast as a probability distribution over the set of models.

A common regularizer of the ERM problem is the relative entropy of the optimization probability measure with respect to a given reference measure over the set of models [**?**], [**?**], [**?**], [**?**]. The resulting problem formulation, termed ERM with relative entropy regularization (ERM-RER) has been extensively studied for both the case in which the reference measure is a probability measure [**?**], [**?**], [**?**], [**?**] and the case in which it is a σ-finite measure [**?**], [**?**], [**?**]. While in both cases the solution is unique and corresponds to a Gibbs probability measure, the existence of the solution is ensured only in the case in which the reference measure is a probability measure [**?**]. Despite the many merits of the ERM-RER formulation, it has some significant limitations. Firstly, the absolute continuity of the optimization measure with respect to the reference measure is required for the existence of the corresponding Radon-Nikodym derivative, which is used by the relative entropy regularization. This absolute continuity sets an insurmountable barrier to the exploration of models outside the support of the reference measure. More specifically, models outside the support of the reference measure exhibit zero probability with respect to the Gibbs probability measure solution to ERM-RER, regardless of the evidence provided by the training dataset. Secondly, the choice of relative entropy over alternative divergences often follows arguments based on the simplicity of obtaining generalization guarantees in the form of bounds [**?**]. Nonetheless, such bounds are often hard to calculate and are not always informative when evaluated in practical settings [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], [**?**].

In view of these, exploring the asymmetry of relative entropy is of particular interest to advancing the understanding of entropy regularization in the context of ERM and its role in generalization. Additionally, examining the asymmetry opens novel pathways to overcome some of the constraints imposed by relative entropy regularization. The problem of ERM with a general f-divergence regularization has been explored in [**?**] and [**?**] in the case of a finite countable set of models, and recently extended to uncountable sets of models in [**?**] and [**?**]. The authors in [**?**], [**?**], [**?**], [**?**] constrain the optimization domains to sets of measures that are mutually absolutely continuous with respect to the reference probability measure. The use of the relative entropy of the optimization measure with respect to the reference measure as a regularizer in the ERM-RER is termed Type-I ERM-RER. Alternatively, the use of the relative entropy of the reference measure with respect to the optimization

Author 1 is with the Department of Automatic Control & Systems Engineering, The University of Sheffield, Sheffield S1 3JD, U.K.; and also with INRIA, Centre Inria d'Université Côte d'Azur, 06902 Sophia Antipolis, France (e-mail: jdaunastorres1@sheffield.ac.uk).

Author 2 is with the Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield S1 3JD, U.K.; and also with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: esnaola@sheffield.ac.uk).

Author 3 is with INRIA, Centre Inria d'Université Côte d'Azur, 06902 Sophia Antipolis, France; also with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA; and also with the GAATI Mathematics Laboratory, University of French Polynesia, 98702 Faaa, French Polynesia (e-mail: samir.perlaza@inria.fr).

Author 4 is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

This paper was presented in part at the IEEE International Symposium on Information Theory (ISIT), Taipei, Taiwan, Jun., 2023 [**?**]; and appears as an INRIA Technical Report in [**?**].

measure is termed Type-II ERM-RER. Interestingly, the existing results in [**?**], [**?**], [**?**], which lead to special cases of the Type-I and Type-II ERM-RER problems by assuming that $f(x) = -x\log(x)$ and $f(x) = -\log(x)$, respectively, do not study the impact of the asymmetry of relative entropy. Another observation that motivates studying the asymmetry of relative entropy in ERM-RER is that numerical analyses of the Type-II ERM-RER, presented in Section **??**, suggest that it achieves better generalization capabilities compared to Type-I ERM-RER, while maintaining a similar expected empirical risk.

This paper presents the solution to Type-II ERM-RER optimization problem using a new method of proof. In particular, mutual absolute continuity between the measures involved is not imposed. Nonetheless, mutual absolute continuity is exhibited by the solution as a consequence of the structure of the problem. The key properties of the solution are highlighted, and an equivalence between the Type-I and Type-II ERM-RER problems is presented. This equivalence is achieved by replacing the empirical risk in the Type-I ERM-RER problem with another function, which can be interpreted as a tunable loss function, as described in [**?**], [**?**], [**?**]. The remainder of the paper is organized as follows. Section **??** presents the ERM-RER problem and its two variations: Type-I and Type-II. The main contribution of this paper, which is the solution to the Type-II ERM-RER problem, is presented in Section **??**. This section also presents key properties of the solution. Section **??** uses these properties to characterize the expected empirical risk. Section **??** studies the equivalence between Type-I and Type-II ERM-RER problems. This work is concluded by Section **??**, with some final remarks.

The mathematical notation used throughout the paper is as follows: Given a measurable space $(\mathcal{M}, \mathscr{F})$, $\triangle(\mathcal{M})$ is used to represent the set of probability measures that can be defined over $(\mathcal{M}, \mathscr{F})$. Often, when the sigma-algebra $\mathscr{F}$ is fixed, it is hidden to ease notation. Given a probability measure $Q \in \triangle(\mathcal{M})$, the subset $\triangle_Q(\mathcal{M})$ of $\triangle(\mathcal{M})$ contains all probability measures that are absolutely continuous with respect to the measure Q. Similarly, the subset $\bigtriangledown_Q(\mathcal{M})$ of $\triangle(\mathcal{M})$ contains all probability measures $P \in \triangle(\mathcal{M})$ such that the probability measure Q is absolutely continuous with respect to P. Finally, the subset $\bigcirc_Q(\mathcal{M})$ of $\triangle(\mathcal{M})$ contains all probability measures that are mutually absolutely continuous with respect to the measure Q.

II. Empirical Risk Minimization

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

$$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n. \tag{1}$$

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

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

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

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

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

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

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

Let the set of solutions to the ERM problem in (**??**) be 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 (**??**) always possesses a solution, and thus, $|\mathcal{T}(\boldsymbol{z})| > 0$. Nonetheless, in general, the ERM problem does not necessarily possess a solution, *i.e.*, it might happen that $|\mathcal{T}(\boldsymbol{z})| = 0$.

The PAC and Bayesian frameworks, as discussed in [**?**] and [**?**], address the problem in (**??**) by constructing probability measures, 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 (**??**) over the set of all probability measures that can be defined on the measurable space $(\mathcal{M}, \mathscr{F})$, which is denoted by $\triangle(\mathcal{M})$, requires a metric that enables assessing the goodness of the probability measure. From this perspective, the underlying assumption in the remainder of this work is that the functions f and ℓ in (**??**) are such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $g_{x,y} : \mathcal{M} \to [0, \infty)$, such that $g_{x,y}(\boldsymbol{\theta}) = \ell(f(\boldsymbol{\theta}, x), y)$, is measurable with respect to the Borel measurable spaces $(\mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, where $\mathscr{F}$ and $\mathscr{B}(\mathbb{R})$ are respectively the Borel σ-fields on $\mathcal{M}$ and $\mathbb{R}$. Under these assumptions, a common metric is the notion of expected empirical risk.

Definition 1 (Expected Empirical Risk): Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (**??**), let the functional $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}) \to [0, \infty)$ be such that

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

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

In the following section, the Type-I relative entropy regularization is reviewed as it serves as the basis for the analysis of the regularization asymmetry.

A. The Type-I ERM-RER Problem

The Type-I ERM-RER problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M})$ and a real $\lambda \in (0, \infty)$. The measure Q is referred to as the *reference measure* and λ as the *regularization factor*. The Type-I ERM-RER problem, with parameters Q and λ, is given by the following optimization problem:

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

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (**??**), and the optimization domain is

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

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

The solution to the Type-I ERM-RER problem in (**??**) is the Gibbs probability measure reported in [**?**], [**?**] and [**?**]. In order to introduce such a measure, consider the function $K_{Q,\boldsymbol{z}} : (0, \infty) \to \mathbb{R}$ that satisfies for all $t \in \mathbb{R}$,

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

with $\mathsf{L}_{\boldsymbol{z}}$ in (**??**). Using this notation, the solution to the Type-I ERM-RER problem in (**??**) is presented by the following lemma.

Lemma 1 ([?, Theorem 3]): The solution to the optimization problem in (**??**) is a unique probability measure, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda} \right) - \frac{1}{\lambda} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \right), \tag{10}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (**??**) and the function $K_{Q,\boldsymbol{z}}$ is defined in (**??**).

B. The Type-II ERM-RER Problem

The Type-II ERM-RER problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M})$ and a real $\lambda \in (0, \infty)$. As in the Type-I ERM-RER problem, 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 (**??**), the Type-II ERM-RER problem, with parameters Q and λ, consists of the following optimization problem:

$$\min_{P \in \triangledown_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q \| P), \tag{11}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (**??**), and the optimization domain is

$$\triangledown_Q(\mathcal{M}) \triangleq \{P \in \triangle(\mathcal{M}) : Q \ll P\}. \tag{12}$$

The difference between Type-I and Type-II ERM-RER problems lies on the regularization. While the former uses the relative entropy $\mathsf{D}(P\|Q)$, the latter uses $\mathsf{D}(Q\|P)$. This translates into different optimization domains due to the asymmetry of the relative entropy. More specifically, in the Type-I ERM-RER problem, the optimization domain is the set of probability measures on the Borel measurable space $(\mathcal{M}, \mathscr{F})$ that are absolutely continuous with the reference measure Q. That is, the set $\triangle_Q(\mathcal{M})$ in (**??**). Alternatively, in the Type-II ERM-RER problem, the optimization domain consists of probability measures defined on the Borel measurable space $(\mathcal{M}, \mathscr{F})$, with the additional condition that the reference measure Q must be absolutely continuous with respect to them. This corresponds to the set denoted as $\triangledown_Q(\mathcal{M})$ in (**??**). From this perspective, the techniques used in [**?**] for solving the Type-I ERM-RER no longer hold. As shown in the next section, a new technique is used for solving the Type-II ERM-RER.

The problems in (**??**) and (**??**) exhibit a trivial solutions when the functional $\mathsf{R}_{\boldsymbol{z}}$ is such that for all $P \in \triangle_Q(\mathcal{M})$ or $P \in \triangledown_Q(\mathcal{M})$, respectively, it holds that $\mathsf{R}_{\boldsymbol{z}}(P) = c$, for some $c \in [0, \infty)$. In such a case, the solution is unique and equal to the probability measure Q, independently of the parameter λ. In order to avoid this trivial case, the notion of separability of the empirical risk function with respect to the measure Q is borrowed from [**?**]. A separable empirical risk function with respect to a given probability measure P is defined as follows.

Definition 2 (Definition 5 in [?]): The empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (**??**) is said to be separable with respect to the probability measure $P \in \triangle(\mathcal{M})$, if there exist a positive real $c > 0$ and two subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{M}$ that are nonnegligible with respect to P, and for all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) < c < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) < \infty. \tag{13}$$

A nonseparable empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (**??**) with respect to a measure P is a constant almost surely with respect to the measure P. More specifically, there exists a real $a \geq 0$, such that

$$P(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = a\}) = 1. \tag{14}$$

When the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (**??**) is nonseparable with respect to all measures in $P \in \bigtriangledown_Q(\mathcal{M})$, the trivial case described above is observed. The notion of separable empirical risk functions would play a central role in the study of the optimization problem in (**??**).

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

The solution of the Type-II ERM-RER problem in (**??**) is presented in the following theorem.

Theorem 1: If there exists a real β such that

$$\beta \in \{t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\}, \tag{15a}$$

and

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

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (**??**), and λ and Q the parameters of the optimization problem in (**??**), then, the solution to such a problem, denoted by $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle(\mathcal{M})$, is unique and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it satisfies

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}. \tag{16}$$

Before introducing the proof of Theorem **??**, two important results are presented. The first result consists in the solution to the optimization problem in (**??**) when the optimization domain is restricted to

$$\bigcirc_Q(\mathcal{M}) \triangleq \bigtriangledown_Q(\mathcal{M}) \cap \triangle_Q(\mathcal{M}), \tag{17}$$

where the sets $\triangle_Q(\mathcal{M})$ and $\bigtriangledown_Q(\mathcal{M})$ are defined in (**??**) and (**??**), respectively. Such an ancillary problem can be formulated as follows:

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

The solution to the problem in (**??**) is described by the following lemma.

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

Proof: The proof is presented in Appendix **??**. ∎

The second result consists of comparing the optimal values resulting from the optimization problems in (**??**) and (**??**), as shown hereunder.

Lemma 3: The optimization problems in (**??**) and (**??**) satisfy

$$\min_{P \in \bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P) \geq \min_{P \in \bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P). \tag{19}$$

Proof: The proof is presented in Appendix **??**. ∎

Lemma **??** unveils the fact that the objective function in (**??**) when evaluated at measures whose support extends beyond the support of Q is larger than such an objective function evaluated at measures whose support is identical to the reference measure. This includes the case in which the set $\mathcal{T}(\boldsymbol{z})$ in (**??**) lies outside the support of Q. Using these results, the proof of Theorem **??** is as follows.

Proof of Theorem **??***:* The proof follows by observing that from (**??**), it holds that

$$\bigcirc_Q(\mathcal{M}) \subseteq \bigtriangledown_Q(\mathcal{M}). \tag{20}$$

Hence, from (**??**), it follows that

$$\min_{P \in \bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P) \leq \min_{P \in \bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P). \tag{21}$$

From the inequalities in (**??**) and (**??**), it follows that

$$\min_{P\in\bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P)+\lambda\mathsf{D}(Q\|P)=\min_{P\in\bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P)+\lambda\mathsf{D}(Q\|P). \tag{22}$$

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

Lemma **??** implies that the solution to the optimization problem in (**??**) is in the set $\bigcirc_Q(\mathcal{M})$ in (**??**). A consequence of this observation is the following corollary.

Corollary 4: The probability measures Q and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (**??**) are mutually absolutely continuous.

Corollary **??** also follows from Theorem **??** by observing that the solution to the Type-II ERM-RER problem in (**??**) is expressed in terms of its Radon-Nikodym derivative with respect to Q, which implies the absolute continuity of $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to Q. The absolute continuity of the measure Q with respect to $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ follows from the optimization domain of the Type-II ERM-RER problem. From this perspective, Corollary **??** conveys the fact that there does not exist a dataset that can overcome the inductive bias induced by the reference measure Q. That is, sets of models outside the support of Q exhibit zero probability measure with respect to the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

This observation is important as, at first glance, the Type-II relative entropy regularization for the ERM problem in (**??**) does not restrict the solution to be absolutely continuous with respect to the reference measure Q. However, Theorem **??** shows that the support of the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (**??**) collapses into the support of the reference. A parallel can be established between Type-I and Type-II cases, 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.

IV. Final Remarks

This work has introduced the Type-II ERM-RER problem and has presented its solution through Theorem **??**. The solution highlights that regardless of whether Type-I or Type-II regularization is used in ERM problems, the models that are considered by the resulting solution are necessarily in the support of the reference measure. In this 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. This limitation has been shown to be a consequence of the equivalence that can be established between Type-I and Type-II regularization. These analytical results lead to providing an operationally meaningful characterization of the expected empirical risk induced by the Type-II solution in terms of the regularization parameters.

V. Acknowledgments

This work is supported by the University of Sheffield ACSE PGR scholarships, the Inria Exploratory Action – Information and Decision Making (AEx IDEM), the European Commission through the H2020-MSCA-RISE-2019 project TESTBED2 under grant agreement no. 872172, and in part by a grant from the C3.ai Digital Transformation Institute.

Appendix A
Proof of Lemma **??**

Proof: The optimization problem in (**??**) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the measure Q, which yields:

$$\min_{P\in\bigcirc_Q(\mathcal{M})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})-\lambda\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q(\boldsymbol{\theta}), \tag{23a}$$

$$\text{s.t.}\quad \int\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})=1. \tag{23b}$$

The remainder of the proof focuses on the problem in which the optimization is over the function $\frac{\mathrm{d}P}{\mathrm{d}Q}:\mathcal{M}\to\mathbb{R}$, instead of optimizing the measure P. This is due to the fact that for all $P\in\bigcirc_Q(\mathcal{M})$, the Radon-Nikodym derivate $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is unique up to sets of zero measure with respect to Q. Let $\mathscr{M}$ be the set of measurable functions $\mathcal{M}\to\mathbb{R}$ with respect to the measurable spaces $(\mathcal{M},\mathscr{F})$ and $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ that are absolutely integrable with respect to Q. That is, for all $\hat{g}\in\mathscr{M}$, it holds that

$$\int|\hat{g}(\boldsymbol{\theta})|\,\mathrm{d}Q(\boldsymbol{\theta})<\infty. \tag{24}$$

Hence, the optimization problem of interest is:

$$\min_{g\in\mathscr{M}}\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})-\lambda\int\log(g(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{25a}$$

$$\text{s.t.}\int g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})=1. \tag{25b}$$

Let the Lagrangian of the optimization problem in (**??**) be $L : \mathscr{M} \times \mathbb{R} \to \mathbb{R}$ such that

$$L(g,\beta)=\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \log(g(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta}) + \ \beta\Bigg(\int g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - 1\Bigg) \tag{26}$$

$$=\int \Big(g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta) - \lambda \log(g(\boldsymbol{\theta}))\Big)\ \mathrm{d}Q(\boldsymbol{\theta}) - \ \beta, \tag{27}$$

where β is a real that acts as a Lagrange multiplier due to the constraint (**??**). Let $\hat{g} : \mathcal{M} \to \mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (**??**) at $(g, \beta) \in \mathscr{M} \times \mathbb{R}$ in the direction of $\hat{g}$ is

$$\partial L(g,\beta;\hat{g}) \triangleq \frac{\mathrm{d}}{\mathrm{d}\gamma} r(g + \gamma\hat{g}, \beta)\Big|_{\gamma=0}, \tag{28}$$

where the function $r : \mathbb{R} \to \mathbb{R}$ is such that for all $\gamma \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small,

$$r(\gamma)=\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta}) - \ \lambda \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta}) + \ \beta\Bigg(\int (g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) - 1\Bigg) \tag{29a}$$

$$=\gamma \int \hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\,\mathrm{d}Q(\boldsymbol{\theta}) + \ \lambda \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta}) + \int g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\,\mathrm{d}Q(\boldsymbol{\theta}) - \beta. \tag{29b}$$

Note that the first term in (**??**) is linear with respect to γ; the second term can be written using the function $\hat{r} : \mathbb{R} \to \mathbb{R}$ in (**??**) such that

$$\hat{r}(\gamma) = \int -\log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}); \tag{30}$$

and the remaining terms are independent of γ.

Hence, based on the fact that the function $\hat{r}$ in (**??**) is differentiable at zero (Lemma **??**), so is the function r in (**??**), which implies that the Gateaux differential of $\partial L(g, \beta; \hat{g})$ in (**??**) exists.

The derivative of the real function r in (**??**) is

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) = \int \hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\,\mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \frac{\hat{g}(\boldsymbol{\theta})}{(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}\ \mathrm{d}Q(\boldsymbol{\theta}) \tag{31}$$

$$=\int \hat{g}(\boldsymbol{\theta})\Bigg(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}\Bigg)\mathrm{d}Q(\boldsymbol{\theta}). \tag{32}$$

From (**??**) and (**??**), it follows that

$$\partial L(g,\beta;\hat{g})=\int \hat{g}(\boldsymbol{\theta})\Bigg(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{g(\boldsymbol{\theta})}\Bigg)\mathrm{d}Q(\boldsymbol{\theta}). \tag{33}$$

The relevance of the Gateaux differential in (**??**) stems from [**?**, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (**??**) to have a stationary point at $\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta\right) \in \mathcal{M} \times [0, \infty)$ is that for all functions $\hat{g} \in \mathscr{M}$,

$$\partial L\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta; \hat{g}\right) = 0. \tag{34}$$

From (**??**) and (**??**), it follows that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ must satisfy for all functions $\hat{g}$ in $\mathscr{M}$ that

$$\int \hat{g}(\boldsymbol{\theta})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \lambda\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1}\right)\mathrm{d}Q(\boldsymbol{\theta}) = 0. \tag{35}$$

This implies that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \lambda\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1} = 0, \tag{36}$$

and thus,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{37}$$

where β is chosen to satisfy (**??**) and guarantee that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \in (0,\infty)$. That is,

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \frac{\lambda}{t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \right\}, \text{ and} \tag{38}$$

$$1 = \int \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta}). \tag{39}$$

which is an assumption of the theorem.

The proof continues by verifying that the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that satisfies (**??**) is the unique solution to the optimization problem in (**??**). Such verification is done by showing that the objective function in (**??**) is strictly convex with the optimization variable. Let P_1 and P_2 be two different probability measures in $(\mathcal{M}, \mathscr{F})$ and let α be in $(0,1)$. Hence,

$$\begin{aligned} \mathsf{R}_{\boldsymbol{z}}(\alpha P_1 + (1-\alpha)P_2) + \lambda \mathsf{D}(\alpha P_1 + (1-\alpha)P_2 \| Q) &= \mathsf{R}_{\boldsymbol{z}}(\alpha P_1) + \mathsf{R}_{\boldsymbol{z}}((1-\alpha)P_2) + \lambda \mathsf{D}(\alpha P_1 + (1-\alpha)P_2 \| Q) \quad (40) \\ &> \alpha \mathsf{R}_{\boldsymbol{z}}(P_1) + (1-\alpha)\mathsf{R}_{\boldsymbol{z}}(P_2) \\ &\quad + \lambda(\alpha \mathsf{D}(P_1 \| Q) + (1-\alpha)\mathsf{D}(P_2 \| Q)) \quad (41) \end{aligned}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (**??**). The equality above follows from the properties of the Lebesgue integral, while the inequality follows from [**?**, Theorem 2]. This proves that the solution is unique due to the strict concavity of the objective function, which completes the proof. ∎

This appendix concludes by presenting Lemma **??** used in the proof of Lemma **??**

Lemma 5: Let $\mathscr{M}$ be the set of measurable functions $h : \mathcal{M} \to \mathbb{R}$, with respect to the measurable space $(\mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Let $\mathscr{S}$ be the subset of $\mathscr{M}$ including all nonnegative functions that are absolutely integrable with respect to a probability measure Q. That is, for all $h \in \mathscr{S}$, it holds that

$$\int |h(\boldsymbol{\theta})| \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{42}$$

Let the function $\hat{r} : \mathbb{R} \to \mathbb{R}$ be such that

$$\hat{r}(\alpha) = \int -\log(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{43}$$

for some functions g and h in $\mathscr{S}$ and $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small. The function $\hat{r}$ in (**??**) is differentiable at zero.

Proof: The objective is to prove that the function $\hat{r}$ in (**??**) is differentiable at zero, which boils down to proving that the limit

$$\lim_{\delta \to 0} \frac{1}{\delta}(\hat{r}(\alpha + \delta) - \hat{r}(\alpha)) \tag{44}$$

exists for all $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small. Let the function $f : (0,\infty) \to \mathbb{R}$ be a function such that

$$f(x) = -\log(x). \tag{45}$$

Note that the function $\hat{r}$ can be written in terms of f as follows

$$\hat{r}(\alpha) = \int f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{46}$$

The proof of the existence of such limit in (**??**) relies on the fact that the function f in (**??**) is strictly convex and differentiable, which implies that f is also Lipschitz continuous. Hence, it follows that

$$|f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))| \leq c \, |h(\boldsymbol{\theta})| \, |\delta|, \tag{47}$$

for some positive and finite constant c, which implies that

$$\frac{|f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))|}{|\delta|} \leq c \, |h(\boldsymbol{\theta})|, \tag{48}$$

and thus, given that $h \in \mathscr{S}$, it holds that

$$\int \frac{|f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))|}{|\delta|} \, \mathrm{d}Q(\boldsymbol{\theta}) \leq \infty. \tag{49}$$

This allows using the dominated convergence theorem as follows. From the fact that the function f is differentiable, let $\dot{f} : \mathbb{R} \to \mathbb{R}$ be the first derivative of f. The limit in (**??**) satisfies for $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small,

$$\lim_{\delta \to 0} \frac{1}{\delta}(\hat{r}(\alpha + \delta) - \hat{r}(\alpha)) = \lim_{\delta \to 0} \frac{1}{\delta}\left(\int f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) - \int f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \right) \tag{50a}$$

$$= \lim_{\delta \to 0} \int \frac{1}{\delta}(f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})))\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{50b}$$

$$= \int \lim_{\delta \to 0} \frac{1}{\delta}(f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})))\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{50c}$$

$$= \int \dot{f}(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{50d}$$

$$< \infty, \tag{50e}$$

where the equalities in (**??**) and (**??**) follow from the dominated convergence theorem [**?**, Theorem 1.6.9]. From (**??**), it follows that the function $\hat{r}$ in (**??**) is differentiable at zero. This completes the proof. ■