



HAL
open science

Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences

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

► **To cite this version:**

Francisco Daunas, Iñaki Esnaola, Samir M Perlaza, H. Vincent Poor. Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences. Proceedings of the IEEE International Symposium on Information Theory, Jul 2024, Athens, Greece. hal-04431558v1

HAL Id: hal-04431558

<https://hal.science/hal-04431558v1>

Submitted on 1 Feb 2024 (v1), last revised 16 May 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Equivalence of the Empirical Risk Minimization to Regularization on the Family of f -Divergences

Francisco Daunas^{*†}, Iñaki Esnaola^{*‡}, Samir M. Perlaza^{†‡§}, and H. Vincent Poor[‡]

^{*}Dept. of ACSE, 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

[‡]Dept. of ECE, Princeton University, Princeton, 08544 NJ, USA. poor@princeton.edu

[§]GAATI, Université de la Polynésie Française, Faaa, French Polynesia.

Abstract—The solution to empirical risk minimization with f -divergence regularization (ERM- f DR) is presented under mild conditions on f . Under such conditions, the optimal measure is shown to be unique. Examples of the solution for particular choices of the function f are presented. Previously known solutions to common regularization choices are obtained by leveraging the flexibility of the family of f -divergences. These include the unique solutions to empirical risk minimization with relative entropy regularization (Type-I and Type-II). The analysis of the solution unveils the following properties of f -divergences when used in the ERM- f DR problem: *i*) f -divergence regularization forces the support of the solution to coincide with the support of the reference measure, which introduces a strong inductive bias that dominates the evidence provided by the training data; and *ii*) any f -divergence regularization is equivalent to a different f -divergence regularization with an appropriate transformation of the empirical risk function.

Index Terms—Empirical risk minimization; f -divergence regularization, statistical learning.

I. INTRODUCTION

Empirical Risk Minimization (ERM) is a fundamental principle in machine learning. It is a tool for selecting a model from a given set by minimizing the empirical risk, which is the average loss or error induced by such a model on each of the labeled patterns available in the training dataset [1], [2]. In a nutshell, ERM aims to find a model that performs well on a given training dataset. However, ERM is prone to overfitting [3]–[5], which affects the generalization capability of the selected model [6]–[8]. To remediate this phenomenon, the solution of ERM must exhibit a small sensitivity to variations in the training dataset, which is often obtained via regularization [9]–[13].

In statistical learning theory, the ERM problem amounts to the minimization of the expected empirical risk over a subset of all probability measures that can be defined on the set of models. In this case, regularization is often obtained by adding to the expected empirical risk a *statistical distance* from the optimization measure, weighted by a regularization factor. Such a statistical distance is essentially a non-negative measure of dissimilarity between the optimization measure and the reference measure, which might be a σ -finite measure

and not necessarily a probability measure, as shown in [12] and [14]. A key observation is that the reference measure often represents prior knowledge or the inductive bias on the solution.

The notion of f -divergence, introduced in [15], and further studied in [16] and [17], describes a family of hallmark statistical distances. A popular f -divergence is the relative entropy [18], which due to its asymmetry, leads to two different problem formulations known as Type-I and Type-II ERM with relative entropy regularization (ERM-RER) [19]–[21]. Relative entropy regularization also plays a central role in obtaining the worst-case data-generating probability measure introduced in [13] and [22]. The Type-I ERM-RER problem exhibits a unique solution, which is a Gibbs probability measure, independently of whether the reference measure is a probability measure or a σ -finite measure, as shown in [12]. The Type-II ERM-RER problem also has a unique solution when the reference measure is a probability measure. This solution exhibits properties that are analogous to those of the Gibbs probability measure [20]. Type-I ERM-RER appears to be the more popular regularized ERM problem, despite the fact that both types of regularization have distinct advantages. See for instance, [14], [20], [23]–[32] and references therein.

Optimization problems with f -divergence regularization have been explored before in [33] and [34] for the discrete case. In [35], the problem of non-exponentially weighted aggregation is studied. Such a problem involves an ERM with f -divergence regularization (ERM- f DR) identical to the one studied in this work. Nonetheless, the ERM- f DR imposes strong solution existence conditions on the function f , and thus, it holds for a reduced number of f -divergences. This work presents the solution to the ERM- f DR problem using a method of proof that differs from those in [33], [34] and [35] and goes along the lines of the methods in [12], [13] and [20], which rely on the notion of the Gateaux derivative [36] and vector space methods [37].

The method of proof favored in this paper enables the derivation of new results that have not been reported before. Firstly, the permissible values of the regularization factor that guarantee the existence of a solution are analytically characterized. Secondly, the obtained solution holds for a family of f -divergences that is larger than the one in [35]. For instance, the Type-II ERM-RER studied in [21] and the ERM with Jensen-

This work is supported by the University of Sheffield ACSE PGR scholarships, the Inria Exploratory Action – Information and Decision Making (AEX IDEM), H2020 project CYBERNETS under EU Grant agreement 659316, and in part by a grant from the C3.ai Digital Transformation Institute.

Shannon divergence regularization are both special cases of the ERM- f DR problem studied in this paper. These are examples of ERM- f DR problems that are not considered in [35]. More importantly, the new method of proof allows showing that any f -divergence regularization is equivalent to a different f -divergence regularization with an appropriate transformation of the empirical risk function.

II. EMPIRICAL RISK MINIMIZATION PROBLEM

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 *data point*, and a *dataset* is represented by the tuple $((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$.

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

$$z = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n, \quad (1)$$

the objective is to obtain a model $\theta \in \mathcal{M}$, such that, for all $i \in \{1, 2, \dots, n\}$, the label assigned to pattern x_i , which is $h(\theta, x_i)$, is “close” to the label y_i . This notion of “closeness” is formalized by the function

$$\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty), \quad (2)$$

such that the loss or risk induced by choosing the model $\theta \in \mathcal{M}$ with respect to the labeled pattern (x_i, y_i) , with $i \in \{1, 2, \dots, n\}$, is $\ell(h(\theta, x_i), y_i)$. The risk function ℓ is assumed to be nonnegative and to satisfy $\ell(y, y) = 0$, for all $y \in \mathcal{Y}$.

The *empirical risk* induced by a model θ with respect to the dataset z in (1) is determined by the function $L_z : \mathcal{M} \rightarrow [0, +\infty)$, which satisfies

$$L_z(\theta) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(h(\theta, x_i), y_i). \quad (3)$$

The ERM problem with respect to the dataset z in (1) consists of the optimization problem:

$$\min_{\theta \in \mathcal{M}} L_z(\theta). \quad (4)$$

The set of solutions to such a problem is denoted by

$$\mathcal{T}(z) \triangleq \arg \min_{\theta \in \mathcal{M}} L_z(\theta). \quad (5)$$

Note that if the set \mathcal{M} is finite, the ERM problem in (4) has a solution, and therefore, it holds that $|\mathcal{T}(z)| > 0$. 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 z that yield $|\mathcal{T}(z)| = 0$.

III. THE ERM WITH f -DIVERGENCE REGULARIZATION

A. Preliminaries

For ease of notation, the expected empirical risk with respect to a given measure is expressed via the following functional R_z , defined below.

Definition 1 (Expected Empirical Risk): The expectation of the empirical risk $L_z(\theta)$ in (3), when θ is sampled from a probability measure $P \in \Delta(\mathcal{M})$, is determined by the functional $R_z : \Delta(\mathcal{M}) \rightarrow [0, +\infty)$, such that

$$R_z(P) = \int L_z(\theta) dP(\theta). \quad (6)$$

In the following, the family of f -divergences is defined.

Definition 2 (f -divergence [17]): Let $f : [0, \infty) \rightarrow \mathbb{R}$ be a convex function with $f(1) = 0$ and $f(0) \triangleq \lim_{x \rightarrow 0^+} f(x)$. Let P and Q be two probability measures on the same measurable space, with P absolutely continuous with respect to Q . The f -divergence of P with respect to Q , denoted by $D_f(P\|Q)$, is

$$D_f(P\|Q) \triangleq \int f\left(\frac{dP}{dQ}(\theta)\right) dQ(\theta), \quad (7)$$

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

In the case in which the function f is continuous and differentiable, denote by $\dot{f} : [0, +\infty) \rightarrow \mathbb{R}$ and $\dot{f}^{-1} : \mathbb{R} \rightarrow [0, +\infty)$, the derivative of f and the inverse of the function \dot{f} , respectively.

The set of probability measures that can be defined upon the measurable space $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$, with $\mathcal{B}(\mathcal{M})$ being the Borel σ -field on \mathcal{M} , is denoted by $\Delta(\mathcal{M})$. Given a probability measure $Q \in \Delta(\mathcal{M})$ the set containing exclusively the probability measures in $\Delta(\mathcal{M})$ that are absolutely continuous with respect to Q is denoted by $\Delta_Q(\mathcal{M})$. That is,

$$\Delta_Q(\mathcal{M}) \triangleq \{P \in \Delta(\mathcal{M}) : P \ll Q\}, \quad (8)$$

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

B. Problem Formulation

The ERM- f DR problem is parametrized by a probability measure $Q \in \Delta(\mathcal{M})$, a positive real λ , and an f -divergence (Definition 2). The measure Q is referred to as the *reference measure*, λ as the *regularization factor*, and f as the *regularization function*. Given the dataset $z \in (\mathcal{X} \times \mathcal{Y})^n$ in (1), the ERM- f DR problem, with parameters Q , λ and f , consists of the following optimization problem:

$$\min_{P \in \Delta_Q(\mathcal{M})} R_z(P) + \lambda D_f(P\|Q). \quad (9)$$

C. Solution to the ERM- f DR

The solution of the ERM- f DR problem in (9) is presented in the following theorem under the assumption that the function f is strictly convex.

Theorem 1: If the function f in (9) is strictly convex, differentiable and there exists a β in

$$\mathcal{B} = \left\{ t \in \mathbb{R} : \forall \theta \in \text{supp } Q, 0 < \dot{f}^{-1}\left(-\frac{t + L_z(\theta)}{\lambda}\right) \right\}, \quad (10a)$$

such that

$$\int \dot{f}^{-1}\left(-\frac{\beta + L_z(\theta)}{\lambda}\right) dQ(\theta) = 1, \quad (10b)$$

then the solution to the optimization problem in (9), denoted by $P_{\Theta|Z=z}^{(Q,\lambda)} \in \Delta_Q(\mathcal{M})$, is unique, and for all $\theta \in \text{supp } Q$ satisfies

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \dot{f}^{-1}\left(-\frac{\beta + L_z(\theta)}{\lambda}\right). \quad (11)$$

Proof: The proof is presented in Appendix A. \blacksquare Theorem 1 implies that the Radon-Nikodym derivative $\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}$ in (11) is strictly positive. A consequence of this observation is the following corollary.

Corollary 1: The probability measures Q and $P_{\Theta|Z=z}^{(Q,\lambda)}$ in (11) are mutually absolutely continuous.

Corollary 1 reveals that, as is also the case with Type-II regularization, the support of the reference measure Q establishes an inductive bias that cannot be overcome, regardless of the f -divergence choice. That is, the support of the solution is the support of the reference measure. In a nutshell, the use of any strictly convex f -divergence regularization inadvertently forces the solution to coincide with the support of the reference independently of the training data. Remarkably, from [38, Corollary 23.5.1] the function \dot{f}^{-1} is the derivative of the convex conjugate of f .

D. Examples

Under the assumptions in Theorem 1 and assuming that \mathcal{B} in (10a) is not empty, this section presents the solutions for typical choices of the function f .

1) *Kullback-Leibler Divergence:* Let the function $f : (0, +\infty) \rightarrow \mathbb{R}$ be such that $f(x) = x \log(x)$, whose derivative satisfies

$$\dot{f}(x) = 1 + \log(x). \quad (12a)$$

In this case, the resulting f -divergence $D_f(P||Q)$ is the relative entropy of P with respect to Q . From (12a) and Theorem 1, it holds that for all $\theta \in \text{supp } Q$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \exp\left(-\frac{\beta + \lambda + L_z(\theta)}{\lambda}\right) \quad (12b)$$

$$= \frac{\exp(-\frac{1}{\lambda}L_z(\theta))}{\int \exp(-\frac{1}{\lambda}L_z(\nu)) dQ(\nu)}. \quad (12c)$$

This result has been independently reported by several authors in [12], [14], [29], [31], [39], and the references therein.

2) *Reverse Relative Entropy Divergence:* Let the function $f : (0, +\infty) \rightarrow \mathbb{R}$ be such that $f(x) = -\log(x)$, whose derivative satisfies

$$\dot{f}(x) = -\frac{1}{x}. \quad (13a)$$

In this case, the resulting f -divergence $D_f(P||Q)$ is the relative entropy of Q with respect to P . From (13a) and Theorem 1, it holds that for all $\theta \in \text{supp } Q$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \frac{\lambda}{\beta + L_z(\theta)}. \quad (13b)$$

This result has been reported in [13] and [21].

3) *Jeffrey's Divergence:* Let the function $f : (0, +\infty) \rightarrow \mathbb{R}$ be such that $f(x) = x \log(x) - \log(x)$, whose derivative satisfies

$$\dot{f}(x) = \log(x) + 1 - \frac{1}{x}. \quad (14a)$$

In this case, the resulting f -divergence $D_f(P||Q)$ is Jeffrey's divergence between P and Q . From (14a) and Theorem 1, it holds that for all $\theta \in \text{supp } Q$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \exp\left(W_0\left(\exp\left(\frac{\beta + \lambda + L_z(\theta)}{\lambda}\right)\right) - \frac{\beta + \lambda + L_z(\theta)}{\lambda}\right), \quad (14b)$$

where the function $W_0 : [0, \infty) \rightarrow [0, \infty)$ is the Lambert function, which for a function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that $g(x) = x \exp(x)$ satisfies $W_0(g(x)) = x$.

4) *Hellinger Divergence:* Let the function $f : (0, +\infty) \rightarrow \mathbb{R}$ be such that $f(x) = (1 - \sqrt{x})^2$, whose derivative satisfies

$$\dot{f}(x) = 1 - \frac{1}{\sqrt{x}}. \quad (15a)$$

In this case, the resulting f -divergence $D_f(P||Q)$ is Hellinger's divergence of P with respect to Q . From (15a) and Theorem 1, it holds that for all $\theta \in \text{supp } Q$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \left(\frac{\lambda}{\beta + \lambda + L_z(\theta)}\right)^2. \quad (15b)$$

5) *Jensen-Shannon Divergence:* Let the function $f : (0, +\infty) \rightarrow \mathbb{R}$ be such that $f(x) = x \log\left(\frac{2x}{x+1}\right) + \log\left(\frac{2}{x+1}\right)$, whose derivative satisfies

$$\dot{f}(x) = \log\left(\frac{2x}{x+1}\right). \quad (16a)$$

From (16a) and Theorem 1, it holds that for all $\theta \in \text{supp } Q$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \frac{1}{2 \exp\left(\frac{\beta + L_z(\theta)}{\lambda}\right) - 1}. \quad (16b)$$

6) χ^2 *Divergence:* Let the function $f : (0, \infty) \rightarrow \mathbb{R}$ be such that $f(x) = (x - 1)^2$, whose derivative satisfies

$$\dot{f}(x) = 2(x - 1). \quad (17a)$$

In this case, the resulting f -divergence $D_f(P||Q)$ is the Pearson-divergence, also known as, the χ^2 -divergence. From (17a) and Theorem 1, it holds that for all $\theta \in \text{supp } Q$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \frac{2\lambda - \beta - L_z(\theta)}{2\lambda}. \quad (17b)$$

IV. ANALYSIS OF REGULARIZATION FACTOR

This section explores the set of permissible regularization factors, which is defined as the collection of regularization factors for which the optimization problem in (9) admits a solution. The analysis of the set \mathcal{B} , as defined in (10a), forms the basis for comprehending this set and facilitates the introduction of the *normalization function* described below. Let the function

$$N_{Q,z} : \mathcal{A} \rightarrow \mathcal{B}, \quad (18a)$$

where $\mathcal{A} \subseteq (0, \infty)$ represents the set of permissible regularization factors, and \mathcal{B} is defined in (10a). This function is defined such that for all $\gamma \in \mathcal{A}$,

$$N_{Q,z}(\gamma) = t, \quad (18b)$$

where t satisfies

$$\int \hat{f}^{-1}\left(-\frac{t + L_z(\boldsymbol{\theta})}{\gamma}\right) dQ(\boldsymbol{\theta}) = 1, \quad (19)$$

and the function L_z is defined in (3). Combining (18) and (19), it follows that

$$\int \hat{f}^{-1}\left(-\frac{N_{Q,z}(\lambda) + L_z(\boldsymbol{\theta})}{\lambda}\right) dQ(\boldsymbol{\theta}) = 1, \quad (20)$$

which justifies calling the function $N_{Q,z}$ as the *normalization function*. Some of the properties of interest of the function $N_{Q,z}$ in (18) are characterized by

$$t_{Q,z}^* \triangleq \inf \mathcal{B}, \quad (21)$$

under the assumption that $\mathcal{B} \neq \emptyset$. The following lemma introduces relevant properties of the function $N_{Q,z}$ in (18).

Lemma 2: The function $N_{Q,z}$ in (18) is strictly increasing and continuous.

Proof: The proof is presented in Appendix B. ■

Since the function f is strictly convex, then \hat{f}^{-1} is strictly increasing, which in conjunction with Lemma 2, connects the term $t_{Q,z}^*$ to the infimum of the set \mathcal{A} in (18a). More specifically,

$$\lambda^* \triangleq \inf \mathcal{A} = N_{Q,z}(t_{Q,z}^*). \quad (22)$$

Lemma 3: If the set \mathcal{B} in (10a) is not empty, then it satisfies

$$(t_{Q,z}^*, \infty) \subseteq \mathcal{B} \subseteq [t_{Q,z}^*, \infty). \quad (23)$$

Moreover, the set \mathcal{B} is identical to $[t_{Q,z}^*, \infty)$ if and only if

$$\int \hat{f}^{-1}(-t_{Q,z}^* - L_z(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) < \infty, \quad (24)$$

with $t_{Q,z}^*$ defined in (21).

Proof: The proof is presented in Appendix C. ■

For the case in which \mathcal{B} is closed from the left, Lemma 2 and Lemma 3 imply the existence of a minimum regularization factor $\lambda^* > 0$, with $\lambda^* \in \mathcal{A}$. As a result, the solution to the optimization problem in (9) only exists for regularization factors $\lambda \geq \lambda^*$. For the case in which \mathcal{B} is open from the left, the following lemma shows sufficient conditions for observing that $\mathcal{A} = (0, \infty)$.

Lemma 4: If the function \hat{f}^{-1} in (11) is nonnegative and \mathcal{B} is not empty, then \mathcal{B} in (10a) is identical to $(t_{Q,z}^*, \infty)$ and \mathcal{A} in (18a) is identical to $(0, \infty)$, with $t_{Q,z}^*$ defined in (21).

Proof: The proof is presented in Appendix D. ■

Under the assumptions of Lemma 4, the ERM- f DR optimization in (9) exhibits a unique solution for all $\lambda \in (0, \infty)$. This is the case of the *Kullback-Leibler Divergence*, *Jeffrey's Divergence* and *Hellinger Divergence*, which makes them easy to implement regularizers as the constraint for existence in (10a) is always satisfied.

In the case of divergences such as the *Reverse Relative Entropy Divergence*, *Jensen-Shannon Divergence* and χ^2 *Divergence*, the existence of a lower bound on the regularization factor is dependent on the parameters of the ERM- f DR optimization in (9), which complicates their implementation in practical settings. The following examples illustrate this dependence on the parameters by providing cases in which \mathcal{B} in (23) is the open set $(t_{Q,z}^*, \infty)$ and closed set $[t_{Q,z}^*, \infty)$ for the *Reverse Relative Entropy Divergence*.

Example 1: Consider the ERM- f DR problem in (9) for $f(x) = -\log(x)$ and assume that: (a) $\mathcal{M} = \mathcal{X} = \mathcal{Y} = [0, \infty)$; (b) $\mathbf{z} = (1, 0)$ and (c) $Q \ll \mu$, with μ the Lebesgue measure, such that for all $\boldsymbol{\theta} \in \text{supp } Q$,

$$\frac{dQ}{d\mu}(\boldsymbol{\theta}) = 4\theta^2 \exp(-2\boldsymbol{\theta}). \quad (25a)$$

Let also the function $h : \mathcal{M} \times \mathcal{X} \rightarrow \mathcal{Y}$ be

$$h(\boldsymbol{\theta}, x) = x\boldsymbol{\theta}, \quad (25b)$$

and the risk function ℓ in (2) be

$$\ell(h(\boldsymbol{\theta}, x), y) = (x\boldsymbol{\theta} - y)^2, \quad (25c)$$

which implies

$$L_z(\boldsymbol{\theta}) = (x\boldsymbol{\theta} - y)^2, \quad (25d)$$

with the function L_z defined in (3). Under the current assumptions, the objective of this example is to show that $\mathcal{B} = [t_{Q,z}^*, \infty)$. For this purpose, it is sufficient to show that the inequality in (24) holds. From Theorem 1, it follows that $P_{\Theta|Z=z}^{(Q,\lambda)}$ in (11) satisfies for all $\boldsymbol{\theta} \in \text{supp } Q$,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{d\mu}(\boldsymbol{\theta}) = \frac{\lambda}{L_z(\boldsymbol{\theta}) + \beta} 4\theta^2 \exp(-2\boldsymbol{\theta}), \quad (26)$$

with β satisfying (10). Thus,

$$\int \frac{1}{L_z(\boldsymbol{\theta}) + t_{Q,z}^*} dQ(\boldsymbol{\theta}) = \int_0^\infty 4 \exp(-2\boldsymbol{\theta}) d\boldsymbol{\theta} \quad (27a)$$

$$= 2, \quad (27b)$$

where equality (27a) follows from (25), the assumption that $(x, y) = (1, 0)$ and the fact that $t_{Q,z}^* = 0$. Finally, the function $N_{Q,z}$ in (18) satisfies $N_{Q,z}(\frac{1}{2}) = 0$, which implies $t_{Q,z}^* \in \mathcal{B}$, that is, $\mathcal{B} = [0, \infty)$ and $\mathcal{A} = [\frac{1}{2}, \infty)$.

Example 2: Consider Example 1 with $\mathbf{z} = (1, 1)$. Under the current assumptions, the objective of this example is to show

that $\mathcal{B} = (t_{Q,z}^*, \infty)$. For this purpose, it is sufficient to show that the inequality in (24) does not hold:

$$\int \frac{1}{L_z(\theta) - t_{Q,z}^*} dQ(\theta) = \int \frac{4\theta^2 \exp(-2\theta)}{L_z(\theta) + t_{Q,z}^*} d\mu(\theta) \quad (28a)$$

$$= \int_0^\infty \frac{4\theta^2 \exp(-2\theta)}{(\theta - 1)^2} d\theta \quad (28b)$$

$$= \infty. \quad (28c)$$

where equality (28a) follows from equality (25a); equality (28b) follows from the assumption that $(x, y) = (1, 0)$ and the fact that $t_{Q,z}^* = 0$; and the equality (28c) follows from an algebraic development. Finally, the function $N_{Q,z}$ in (18) is undefined at zero, which implies $t_{Q,z}^* \notin \mathcal{B}$, that is, $\mathcal{B} = (0, \infty)$.

These examples illustrate that even if the reference measure Q and functions ℓ and h in (3) are fixed, the set \mathcal{B} might be either $[t_{Q,z}^*, \infty)$ or $(t_{Q,z}^*, \infty)$ depending on the dataset z . This observation underscores that the existence of the minimum regularization factor λ^* is coupled on the specific choices of Q , ℓ , f , and z .

V. EQUIVALENCE OF THE f -REGULARIZATION VIA TRANSFORMATION OF THE EMPIRICAL RISK

This section shows that given two strictly convex and differentiable functions f and g that satisfy the conditions in Definition 2, there exists a function $v : [0, \infty) \rightarrow [0, \infty)$, such that the solution to the optimization problem in (9) is identical to the solution of the following problem:

$$\min_{P \in \Delta_Q(\mathcal{M})} \int v(L_z(\theta)) dP(\theta) + \lambda D_g(P||Q), \quad (29)$$

with λ and Q in (9). The main result of this section is presented in the following theorem.

Theorem 2: Let f and g be two strictly convex and differentiable functions satisfying the conditions in Definition 2. If the problem in (9) possesses a solution, then

$$\begin{aligned} & \min_{P \in \Delta_Q(\mathcal{M})} \int L_z(\theta) dP(\theta) + \lambda D_f(P||Q) \\ &= \min_{P \in \Delta_Q(\mathcal{M})} \int v(L_z(\theta)) dP(\theta) + \lambda D_g(P||Q), \end{aligned} \quad (30)$$

where the function $v : [0, \infty) \rightarrow [0, \infty)$ is such that

$$v(x) = \lambda \dot{g} \left(\dot{f}^{-1} \left(-\frac{N_{Q,z}(\lambda) + x}{\lambda} \right) \right) - N'_{Q,z}(\lambda), \quad (31)$$

where $N_{Q,z}$ and $N'_{Q,z}$ are the normalization functions of the optimization problems in (9) and (29).

Proof: Note that from Theorem 1 the functions f and g are differentiable and strictly convex. Hence, the functional inverse of the derivative is well-defined from the fact that \dot{f} and \dot{g} are strictly increasing and bijective. Denote by $\hat{P}_{\Theta|Z=z}^{(Q,\lambda)}$

the solution to the optimization problem in (29). Then, from Theorem 1, for all $\theta \in \text{supp } Q$, it follows that

$$\frac{d\hat{P}_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta) = \dot{g}^{-1} \left(-\frac{N'_{Q,z}(\lambda) + v(L_z(\theta))}{\lambda} \right) \quad (32a)$$

$$= \dot{g}^{-1} \left(\dot{g} \left(\dot{f}^{-1} \left(-\frac{N_{Q,z}(\lambda) + L_z(\theta)}{\lambda} \right) \right) \right) \quad (32b)$$

$$= \dot{f}^{-1} \left(-\frac{N_{Q,z}(\lambda) + L_z(\theta)}{\lambda} \right) \quad (32c)$$

$$= \frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\theta), \quad (32d)$$

where the equality in (32b) follows from (31), which completes the proof. \blacksquare

Theorem 2 establishes an equivalence between two ERM problems subject to different f -divergence regularizations. Such equivalence can always be established as long as the corresponding divergences are defined by strictly convex and differentiable functions. More importantly, for all strictly convex f functions, the solution to the corresponding ERM with f -divergence regularization is mutually absolutely continuity with respect to the reference measure.

VI. CONCLUSIONS

This work has presented the solution to the ERM- f DR problem under mild conditions on f , namely, (a) strict convexity; and (b) differentiability. Under these conditions, the optimal measure is shown to be unique and sufficient conditions for the existence of the solution are presented. This result unveils the fact that all parameters are involved in guaranteeing the existence of a solution. Remarkably, the f -divergence regularizer can be transformed into a different f -divergence regularizer by a transformation of the empirical risk. The mutual absolute continuity of the ERM- f DR solutions to the reference measure can be understood in light of the equivalence between the regularization. The analytical results have also enabled us to provide insights into choices of f -divergences for algorithm design in statistical machine learning.

REFERENCES

- [1] V. Vapnik and A. Y. Chervonenkis, "On a perceptron class," *Avtomatika i Telemekhanika*, vol. 25, no. 1, pp. 112–120, Feb. 1964.
- [2] V. Vapnik, "Principles of risk minimization for learning theory," *Advances in Neural Information Processing Systems*, vol. 4, pp. 831–838, Jan. 1992.
- [3] 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.
- [4] 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.
- [5] D. Arpit, S. Jastrzebski, 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.
- [6] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematics Doklady*, vol. 4, no. 6, pp. 1035–1038, Dec. 1963.

- [7] A. E. Horel, "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, vol. 58, no. 1, pp. 54–59, Jun. 1962.
- [8] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 2006.
- [9] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, no. 1, pp. 499–526, Mar. 2002.
- [10] 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.
- [11] 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.
- [12] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with generalized relative entropy regularization," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9454, Feb. 2022.
- [13] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, "The worst-case data-generating probability measure," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9515, Aug. 2023.
- [14] 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.
- [15] A. Rényi *et al.*, "On measures of information and entropy," in *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, Berkeley, CA, USA, Jun. 1961, pp. 547–561.
- [16] I. Sason and S. Verdú, " f -divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, Jun. 2016.
- [17] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, no. 1, pp. 299–318, Jun. 1967.
- [18] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
- [19] 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.
- [20] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "Sensitivity of the Gibbs algorithm to data aggregation in supervised machine learning," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9474, Jun. 2022.
- [21] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Analysis of the relative entropy asymmetry in the regularization of empirical risk minimization," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.
- [22] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, "Generalization analysis of machine learning algorithms via the worst-case data-generating probability measure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2024, pp. 1–9.
- [23] C. P. Robert, *The Bayesian Choice: From Decision-theoretic Foundations to Computational Implementation*, 1st ed. New York, NY, USA: Springer, 2007.
- [24] 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.
- [25] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, Nov. 1984.
- [26] 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.
- [27] 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.
- [28] V. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] F. Futami and T. Iwata, "Information-theoretic analysis of test data sensitivity in uncertainty," arXiv preprint arXiv:2307.12456, Jul. 2023.
- [33] M. Teboulle, "Entropic proximal mappings with applications to nonlinear programming," *Mathematics of Operations Research*, vol. 17, no. 3, pp. 670–690, Aug. 1992.
- [34] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, Jan. 2003.
- [35] P. Alquier, "Non-exponentially weighted aggregation: regret bounds for unbounded loss functions," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, Jul. 2021, pp. 207–218.
- [36] R. Gateaux, "Sur les fonctionnelles continues et les fonctionnelles analytiques," *Comptes rendus hebdomadaires des séances de l'Académie des Sciences, Paris*, vol. 157, no. 325-327, p. 65, 1913.
- [37] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: Wiley, 1997.
- [38] R. T. Rockafellar, *Convex Functions in Optimal Control and the Calculus of Variations*, 2nd ed. Princeton, NJ, USA: Princeton University Press, 1970.
- [39] 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.
- [40] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA, USA: Academic Press, 2000.
- [41] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 2nd ed. New York, NY, USA: Wiley New York, 2000.
- [42] J. Douchet, *Analyse : Recueil d'Exercices et Aide-Mémoire*, 3rd ed. Lausanne, Switzerland: PPUR, 2010, vol. 1.
- [43] W. Rudin, *Principles of mathematical analysis*, 1st ed. New York, NY, USA: McGraw-Hill Book Company, Inc., 1953.

APPENDIX A
PROOF OF THEOREM 1

Proof: The optimization problem in (9) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the measure Q , which yields

$$\min_{P \in \Delta_Q(\mathcal{M})} \int \mathbf{L}_z(\boldsymbol{\theta}) \frac{dP}{dQ}(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) + \lambda \int f\left(\frac{dP}{dQ}(\boldsymbol{\theta})\right) dQ(\boldsymbol{\theta}) \quad (33a)$$

$$\text{s.t.} \quad \int \frac{dP}{dQ}(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) = 1. \quad (33b)$$

The remainder of the proof focuses on the problem in which the optimization is over the function $\frac{dP}{dQ} : \mathcal{M} \rightarrow [0, \infty)$, which represents the Radon-Nikodym derivative of P with respect to Q . Hence, instead of optimizing the measure P , the optimization is over the function $\frac{dP}{dQ}$. This is due to the fact that for all $P \in \Delta_Q(\mathcal{M})$, the Radon-Nikodym derivative $\frac{dP}{dQ}$ is unique up to sets of zero measure with respect to Q . Let \mathcal{M} be the set of measurable functions $\mathcal{M} \rightarrow \mathbb{R}$ with respect to the measurable spaces $(\mathcal{M}, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that are absolutely integrable with respect to Q . That is, for all $\hat{g} \in \mathcal{M}$, it holds that

$$\int |\hat{g}(\boldsymbol{\theta})| dQ(\boldsymbol{\theta}) < \infty. \quad (34)$$

Hence, the optimization problem of interest is

$$\min_{g \in \mathcal{M}} \int \mathbf{L}_z(\boldsymbol{\theta}) g(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) + \lambda \int f(g(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) \quad (35a)$$

$$\text{s.t.} \quad \int g(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) = 1. \quad (35b)$$

Let the Lagrangian of the optimization problem in (35) be $L : \mathcal{M} \times \mathbb{R} \rightarrow \mathbb{R}$ such that

$$L(g, \beta) = \int \mathbf{L}_z(\boldsymbol{\theta}) g(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) + \lambda \int f(g(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) + \beta \left(\int g(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) - 1 \right), \quad (36)$$

$$= \int \left(g(\boldsymbol{\theta})(\mathbf{L}_z(\boldsymbol{\theta}) + \beta) + \lambda f(g(\boldsymbol{\theta})) \right) dQ(\boldsymbol{\theta}) - \beta, \quad (37)$$

where β is a real that acts as a Lagrange multiplier due to the constraint (35b). Let $\hat{g} : \mathcal{M} \rightarrow \mathbb{R}$ be a function in \mathcal{M} . The Gateaux differential of the functional L in (36) at $(g, \beta) \in \mathcal{M} \times \mathbb{R}$ in the direction of \hat{g} , if it exists, is

$$\partial L(g, \beta; \hat{g}) \triangleq \frac{d}{d\gamma} L(g + \gamma \hat{g}, \beta) \Big|_{\gamma=0}. \quad (38)$$

The proof continues under the assumption that the function g and \hat{g} are such that the Gateaux differential in (38) exists.

Under such an assumption, let the function $r : \mathbb{R} \rightarrow \mathbb{R}$ satisfy for all $\gamma \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small, that

$$r(\gamma) = L(g + \gamma \hat{g}, \beta) \quad (39)$$

$$= \int \mathbf{L}_z(\boldsymbol{\theta})(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) + \lambda \int f(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) + \beta \left(\int (g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) - 1 \right) \quad (40)$$

$$= \int g(\boldsymbol{\theta})(\mathbf{L}_z(\boldsymbol{\theta}) + \beta) dQ(\boldsymbol{\theta}) + \gamma \int \hat{g}(\boldsymbol{\theta})(\mathbf{L}_z(\boldsymbol{\theta}) + \beta) dQ(\boldsymbol{\theta}) - \beta + \lambda \int f(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}), \quad (41)$$

where the last equality is simply an algebraic re-arrangement of terms. From the assumption that the function g and \hat{g} are such that the Gateaux differential in (38) exists, it follows that the function r in (39) is differentiable at zero. Note that the first two terms in (41) are independent of γ ; the third term is linear with γ ; and the fourth term can be written using the function $\hat{r} : \mathbb{R} \rightarrow \mathbb{R}$ such that for all $\gamma \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small, it holds that

$$\hat{r}(\gamma) = \lambda \int f(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}). \quad (42)$$

Under the same assumption, it follows that the function \hat{r} in (42) is differentiable at zero. That is, the limit

$$\lim_{\delta \rightarrow 0} \frac{1}{\delta} (\hat{r}(\gamma + \delta) - \hat{r}(\gamma)) \quad (43)$$

exists for all $\gamma \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small. The proof of the existence of such a limit relies on the fact that f is strictly convex and continuous. This implies that f is Lipschitz continuous, which implies that for all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, it holds that

$$|f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))| \leq c |\hat{g}(\boldsymbol{\theta})| |\delta|, \quad (44)$$

for some constant c positive and finite, which implies that

$$\left| \frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}{\delta} \right| \leq c |\hat{g}(\boldsymbol{\theta})|. \quad (45)$$

Using these arguments, the limit in (43) satisfies for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, that

$$\begin{aligned} & \lim_{\delta \rightarrow 0} \frac{1}{\delta} (\hat{r}(\gamma + \delta) - \hat{r}(\gamma)) \\ &= \lambda \lim_{\delta \rightarrow 0} \frac{1}{\delta} \left(\int f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) - \int f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) \right) \\ &= \lambda \lim_{\delta \rightarrow 0} \int \left(\frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta}))}{\delta} \right) \end{aligned} \quad (46)$$

$$-\frac{f(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))}{\delta} \Big) dQ(\boldsymbol{\theta}) \quad (47)$$

$$= \lambda \int \dot{f}(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) \hat{g}(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) \quad (48)$$

$$< \infty, \quad (49)$$

where both the equality in (48) and the inequality in (49) follow from noticing that the conditions for the dominated convergence theorem hold [40, Theorem 1.6.9], namely:

- For all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$, the inequality in (45) holds;
- The function \hat{g} in (45) satisfies the inequality in (34); and
- For all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, it holds that

$$\lim_{\delta \rightarrow 0} \frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}{\delta} \\ = \frac{d}{d\gamma} f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})) \quad (50)$$

$$= \dot{f}(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})), \quad (51)$$

which follows from the fact that f is differentiable.

Hence, the derivative of the real function r in (41) is

$$\frac{d}{d\gamma} r(\gamma) \\ = \int \hat{g}(\boldsymbol{\theta})(L_z(\boldsymbol{\theta}) + \beta) dQ(\boldsymbol{\theta}) \\ + \lambda \int \dot{f}(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})) \hat{g}(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) \quad (52)$$

$$= \int \hat{g}(\boldsymbol{\theta})(L_z(\boldsymbol{\theta}) + \beta + \lambda \dot{f}(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))) dQ(\boldsymbol{\theta}). \quad (53)$$

From (38) and (53), it follows that

$$\partial L(g, \beta; \hat{g}) = \frac{d}{d\gamma} r(\gamma) \Big|_{\gamma=0} \quad (54)$$

$$= \int \hat{g}(\boldsymbol{\theta})(L_z(\boldsymbol{\theta}) + \beta + \lambda \dot{f}(g(\boldsymbol{\theta}))) dQ(\boldsymbol{\theta}). \quad (55)$$

The relevance of the Gateaux differential in (54) stems from [37, Theorem 1], which unveils the fact that a necessary condition for the functional L in (36) to have a stationary point at $\left(\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}, \beta\right) \in \mathcal{M} \times [0, \infty)$ is that for all functions $\hat{g} \in \mathcal{M}$,

$$\partial L\left(\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}, \beta; \hat{g}\right) = 0. \quad (56)$$

From (55) and (56), it follows that $\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}$ must satisfy for all functions \hat{g} in \mathcal{M} that

$$\int \hat{g}(\boldsymbol{\theta}) \left(L_z(\boldsymbol{\theta}) + \beta + \lambda \dot{f}\left(\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta})\right) \right) dQ(\boldsymbol{\theta}) = 0. \quad (57)$$

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

$$L_z(\boldsymbol{\theta}) + \beta + \lambda \dot{f}\left(\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta})\right) = 0, \quad (58)$$

and thus,

$$\frac{dP_{\Theta|Z=z}^{(Q,\lambda)}}{dQ}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\beta + L_z(\boldsymbol{\theta})}{\lambda}\right), \quad (59)$$

where β is chosen to satisfy (10b) and guarantee that for all $\boldsymbol{\theta} \in \text{supp } Q$, it holds that $\dot{f}^{-1}\left(-\frac{\beta + L_z(\boldsymbol{\theta})}{\lambda}\right) \in (0, \infty)$. That is,

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \text{supp } Q, 0 < \dot{f}^{-1}\left(-\frac{t + L_z(\boldsymbol{\theta})}{\lambda}\right) \right\}, \quad (60)$$

which is an assumption of the theorem.

Finally, note that the objective function in (35) is the sum of two terms. The first one, *i.e.*, $\int L_z(\boldsymbol{\theta})g(\boldsymbol{\theta}) dQ(\boldsymbol{\theta})$, is linear with g . The second, *i.e.*, $\int -\log(g(\boldsymbol{\theta})) dQ(\boldsymbol{\theta})$, is strictly convex with g from the assumption that f is strictly convex. Hence, given that $\lambda > 0$, the sum of both terms is strictly convex with g . This implies the uniqueness of $P_{\Theta|Z=z}^{(Q,\lambda)}$, which completes the proof. ■

APPENDIX B

PROOF OF LEMMA 2

Proof: The properties of the function $N_{Q,z}$ in (18b) are proven into two parts. The first part uses the properties of the f -divergences regularization to prove it is strictly increasing. The second part proves the continuity of the function $N_{Q,z}$ in (18b).

The first part is as follows. For $N_{Q,z}$ defined in (18), assume that $t \in \mathcal{B}$, and $\gamma \in \mathcal{A}$ satisfy that

$$N_{Q,z}(t) = \gamma, \quad (61)$$

which implies that

$$1 = \int \frac{dP_{\Theta|Z=z}^{(Q,t)}}{dQ}(\boldsymbol{\theta}) dQ(\boldsymbol{\theta}) \quad (62a)$$

$$= \int \dot{f}^{-1}\left(-\frac{t + L_z(\boldsymbol{\theta})}{\gamma}\right) dQ(\boldsymbol{\theta}), \quad (62b)$$

with \mathcal{B} and \mathcal{A} defined in (10a) and (18a), respectively. Note that the inverse \dot{f}^{-1} exists from the fact that f is strictly convex, which implies that \dot{f} is a strictly increasing function. Hence, \dot{f}^{-1} is also a strictly increasing function in \mathcal{B} [41, Theorem 5.6.9]. Moreover, from the assumption that f is strictly convex and differentiable, it holds that \dot{f} is continuous [42, Proposition 5.44]. This implies that \dot{f}^{-1} is continuous.

From (62) and the fact that \dot{f}^{-1} is continuous, if $t \in \mathcal{B}$, then it holds that

$$0 < \int \dot{f}^{-1}(-t - L_z(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) < \infty. \quad (63)$$

From (63) and the fact that \dot{f}^{-1} is strictly increasing, for all $(t_1, t_2) \in \mathcal{B}^2$, such that $t_1 < t_2$, for all $\boldsymbol{\theta} \in \text{supp } Q$, it holds that

$$\dot{f}^{-1}(-t_1 - L_z(\boldsymbol{\theta})) > \dot{f}^{-1}(-t_2 - L_z(\boldsymbol{\theta})), \quad (64)$$

which implies that

$$\int \dot{f}^{-1}(-t_1 - L_z(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}) > \int \dot{f}^{-1}(-t_2 - L_z(\boldsymbol{\theta})) dQ(\boldsymbol{\theta}). \quad (65)$$

Therefore, from equality (62) and inequality (65) for all $(t_1, t_2) \in \mathcal{B}^2$, such that $t_1 < t_2$, the function $N_{Q,z}$ in (18) satisfies

$$N_{Q,z}(\gamma_1) < N_{Q,z}(\gamma_2), \quad (66)$$

which implies that the function $N_{Q,z}$ in (18) is strictly increasing. Furthermore, from inequality (66) and the fact that $P_{\Theta|Z=z}^{(Q,t)}$ is unique (see Theorem 1), it follows that $N_{Q,z}$ is bijective, which completes the proof of the first part.

The second part is as follows. The continuity of the function $N_{Q,z}$ in (18) is shown by considering an arbitrary $\epsilon > 0$ and a pair $(\gamma_1, \gamma_2) \in \mathcal{A}^2$ under the condition

$$\epsilon > |N_{Q,z}(\gamma_2) - N_{Q,z}(\gamma_1)|. \quad (67)$$

From the fact that $N_{Q,z}$ is bijective and strictly increasing, let the pair γ_1 , and γ_2 satisfy $\gamma_1 < \gamma_2$, without loss of generality. Thus, it holds that

$$|N_{Q,z}(\gamma_2) - N_{Q,z}(\gamma_1)| = N_{Q,z}(\gamma_2) - N_{Q,z}(\gamma_1). \quad (68)$$

Substituting equality (68) into (67) yields

$$N_{Q,z}(\gamma_1) + \epsilon > N_{Q,z}(\gamma_2). \quad (69)$$

From the fact that $N_{Q,z}$ is bijective, evaluating inequality (69) with $N_{Q,z}$ in (18a) leads to

$$N_{Q,z}^{-1}(N_{Q,z}(\gamma_1) + \epsilon) > N_{Q,z}^{-1}(N_{Q,z}(\gamma_2)) \quad (70a)$$

$$= \gamma_2. \quad (70b)$$

Subtracting γ_1 from both sides in (70), results in

$$N_{Q,z}^{-1}(N_{Q,z}(\gamma_1) + \epsilon) - \gamma_1 > \gamma_2 - \gamma_1 \quad (71a)$$

$$= |\gamma_2 - \gamma_1|, \quad (71b)$$

where equality (71b) follows from the condition that $\gamma_2 > \gamma_1$. Thus, from (71) it follows that for all $\epsilon > 0$, there exist a $\delta > 0$ that satisfies

$$\delta = N_{Q,z}^{-1}(N_{Q,z}(\gamma_1) + \epsilon) - \gamma_1, \quad (72)$$

which implies the function $N_{Q,z}$ in (18) is continuous for all $\gamma \in \mathcal{A}$. From (72) the function $N_{Q,z}$ is continuous and strictly increasing, which completes the proof. ■

APPENDIX C PROOF OF LEMMA 3

Proof: The proof shows that if \mathcal{B} in (10a) is not empty, then the domain of the function $N_{Q,z}$ in (18) is an interval of \mathbb{R} . The proof is divided into two parts, in the first part it is shown that \mathcal{B} is always open from the right. The second part uses $t_{Q,z}^*$ the set from the left.

The first part is as follows. From Lemma 2, if there exists a $t_1 \in \mathcal{B}$, then it holds

$$0 < \int f^{-1}(-t_1 - L_z(\theta)) dQ(\theta) < \infty. \quad (73)$$

From (73) and the fact that $N_{Q,z}$ is strictly increasing, for all $t \in \{s \in \mathbb{R} : s \geq t_1\}$, it holds that

$$\int f^{-1}(-t_1 - L_z(\theta)) dQ(\theta) > \int f^{-1}(-t - L_z(\theta)) dQ(\theta). \quad (74)$$

From (74) and the fact that f^{-1} is continuous and decreasing, there exists a $\lambda \in (0, \infty)$ such that for all $t \in \{s \in \mathbb{R} : s \geq t_1\}$ constrain in (10b) is satisfied, which implies that $\mathcal{B} \supseteq [t, \infty)$. This completes the first part of the proof.

The second part is as follows. There are two cases considered under the assumption that $t = t_{Q,z}^*$ with $t_{Q,z}^*$ defined in (21). In the first case, if $t_{Q,z}^* \in \mathcal{B}$, then it holds that

$$\int f^{-1}(-t_{Q,z}^* - L_z(\theta)) dQ(\theta) < \infty, \quad (75a)$$

which implies that $\mathcal{B} = [t_{Q,z}^*, \infty)$, with \mathcal{B} defined in (18a). In the second case, if $t_{Q,z}^* \notin \mathcal{B}$, then it holds that

$$\int f^{-1}(-t_{Q,z}^* - L_z(\theta)) dQ(\theta) = \infty, \quad (75b)$$

which implies that $\mathcal{B} = (t_{Q,z}^*, \infty)$, with \mathcal{B} defined in (18a). Note that the integral in (75) is never zero as it would be a contradiction with (2). Hence, from (74) and (75) the set \mathcal{B} in (10b) is either the open set $(t_{Q,z}^*, \infty)$ or the closed set $[t_{Q,z}^*, \infty)$, which completes the proof. ■

APPENDIX D PROOF OF LEMMA 4

Proof: The following proof by contradiction is divided into three parts. In the first part, an auxiliary function is introduced and proven to be continuous. In the second part, a contradiction is shown under the assumption that f^{-1} is nonnegative and the continuity of the auxiliary. Finally, it is shown that for nonnegative f^{-1} , the set of admissible regularization factors is the positive reals.

The first part is as follows. Let the function $k : \mathcal{B} \rightarrow (0, +\infty)$ be such that

$$k(t) = \int f^{-1}\left(\frac{-t - L_z(\theta)}{\lambda}\right) dQ(\theta). \quad (76)$$

The first step is to prove that the function k in (76) is continuous in \mathcal{B} . This is proved by showing that k always exhibits a limit. Note that from the fact that f^{-1} is strictly increasing, it holds that for all $t \in \mathcal{B}$ and for all $\theta \in \text{supp } Q$, it holds that

$$f^{-1}\left(\frac{-t - L_z(\theta)}{\lambda}\right) \leq f^{-1}\left(\frac{-t}{\lambda}\right), \quad (77)$$

where equality holds if and only if $L_z(\theta) = 0$. Now, from the fact that f^{-1} is continuous it follows that for all $a \in \mathcal{B}$, it holds that

$$\lim_{t \rightarrow a} f^{-1}\left(\frac{-t - L_z(\theta)}{\lambda}\right) = f^{-1}\left(\frac{-a - L_z(\theta)}{\lambda}\right). \quad (78)$$

Hence, from the dominated convergence theorem [40, Theorem 1.6.9], the following limit exists and satisfies

$$\lim_{t \rightarrow a} k(t) = \lim_{t \rightarrow a} \int f^{-1} \left(\frac{-t - \mathbf{L}_z(\boldsymbol{\theta})}{\lambda} \right) dQ(\boldsymbol{\theta}) \quad (79)$$

$$= \int \left(\lim_{t \rightarrow a} f^{-1} \left(\frac{-t - \mathbf{L}_z(\boldsymbol{\theta})}{\lambda} \right) \right) dQ(\boldsymbol{\theta}) \quad (80)$$

$$= \int f^{-1} \left(\frac{-a - \mathbf{L}_z(\boldsymbol{\theta})}{\lambda} \right) dQ(\boldsymbol{\theta}) \quad (81)$$

$$= k(a), \quad (82)$$

which proves that the function k in (76) is continuous.

The second part is as follows. Under the assumption that f induces a nonnegative \dot{f}^{-1} and that \mathcal{B} is nonempty, from Lemma 3 the set \mathcal{B} is either $[t_{Q,z}^*, \infty)$ or $(t_{Q,z}^*, \infty)$. Hence, it is sufficient to show that assuming \mathcal{B} is equal to $[t_{Q,z}^*, \infty)$ and \dot{f}^{-1} is nonnegative leads to a contradiction. From Lemma 3 it follows that

$$k(t_{Q,z}^*) < \infty, \quad (83)$$

with the function k in (76) and $t_{Q,z}^*$ defined in (21). Thus, from the fact that k in (76) is continuous, for $t < t_{Q,z}^*$, it follows that $k(t) < \infty$, which implies there exists a set $C = \left\{ \boldsymbol{\theta} \in \text{supp } Q : \dot{f}^{-1} \left(\frac{-t - \mathbf{L}_z(\boldsymbol{\theta})}{\lambda} \right) < 0 \right\}$ such that $t \notin \mathcal{B}$. However, this contradicts the assumption that \dot{f}^{-1} is nonnegative, which implies that $\mathcal{B} = (t_{Q,z}^*, \infty)$. This completes the second part of the proof.

The third part is as follows. From the fact that \mathcal{B} is nonempty, there is a $t \in \mathcal{B}$ and a $\lambda \in (0, \infty)$ such that,

$$1 = \int \dot{f}^{-1} \left(-\frac{t + \mathbf{L}_z}{\lambda} \right) dQ(\boldsymbol{\theta}). \quad (84)$$

From the fact that the function \dot{f}^{-1} is continuous and strictly increasing, for all $t_1 \in (t_{Q,z}^*, t)$ and for all $t_2 \in (t, \infty)$, it holds that

$$\int \dot{f}^{-1} \left(-\frac{t_1 + \mathbf{L}_z}{\lambda} \right) dQ(\boldsymbol{\theta}) > 1 > \int \dot{f}^{-1} \left(-\frac{t_2 + \mathbf{L}_z}{\lambda} \right) dQ(\boldsymbol{\theta}). \quad (85)$$

Under the same argument, for all $\lambda_1 \in (0, \lambda)$ and for all $\lambda_2 \in (\lambda, \infty)$, it holds that

$$\int \dot{f}^{-1} \left(-\frac{t + \mathbf{L}_z}{\lambda_1} \right) dQ(\boldsymbol{\theta}) > 1 > \int \dot{f}^{-1} \left(-\frac{t + \mathbf{L}_z}{\lambda_2} \right) dQ(\boldsymbol{\theta}). \quad (86)$$

Hence, given that the function k in (76) is continuous, strictly decreasing, from (85) then, there always exists two reals t_1 and t_2 in \mathcal{B} such that $k(t_1) < 1 < k(t_2)$, it follows from the intermediate-value theorem [43, Theorem 4.23] that there always exists a unique real $t \in \mathcal{B}$ such that $k(t) = 1$. Furthermore, for all $t \in \mathcal{B}$ there always exists two reals λ_1 and λ_2 in $(0, \infty)$ such that inequality (86) holds, it follows from the intermediate-value theorem [43, Theorem 4.23] that there always exists a unique real $t \in \mathcal{B}$ for all $\lambda \in (0, \infty)$ such that $k(t) = 1$. Finally, from the fact that $N_{Q,z}$ in (18) is continuous and strictly increasing, if $\mathcal{B} = (t_{Q,z}^*, \infty)$ then the set of admissible regularization factors \mathcal{A} in (18a) is identical to $(0, \infty)$, which completes the proof. ■