Empirical Risk Minimization with f -Divergence Regularization in Statistical Learning

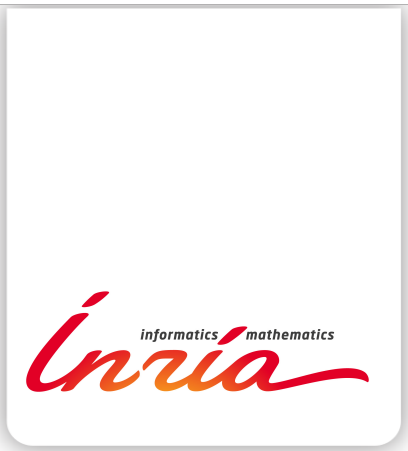
Jose Francisco Daunas Torres, Iñaki Esnaola, Samir M. Perlaza, H. Vincent Poor

Empirical Risk Minimization with f-Divergence Regularization in Statistical Learning

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

RESEARCH REPORT
N° 9521
October 2023
Project-Team NEO

ISSN 0249-6399 ISRN INRIA/RR--9521--FR+ENG



Empirical Risk Minimization with f-Divergence Regularization in Statistical Learning

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

Project-Team NEO

Research Report n° 9521 — October 2023 — 19 pages

Abstract: This report presents the solution to the empirical risk minimization with f-divergence regularization, under mild conditions on f. Under such conditions, the optimal measure is shown to be unique and to always exist. The solution is presented as a closed-form expression of the Radon-Nikodym derivative of the optimal probability measure with respect to the reference measure. Examples for particular choices of the function f are presented. For some choices, existing results are obtained as special cases of the main result. These include the unique solutions to the empirical risk minimization with relative entropy regularization (Type-I and Type-II).

Key-words: Empirical Risk Minimization; f-divergence; Regularization; Inductive Bias; Statistical Learning.

Francisco Daunas and Iñaki Esnaola are with the Department of Automatic Control and Systems Engineering at the University of Sheffield, Sheffield, UK. Francisco Daunas is also a member of the NEO Team at INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis.
Samir M. Perlaza is with INRIA,Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France and with the GAATI Laboratory of the Université de la Polynésie Française, Faaa, French Polynesia.
H. Vincent Poor, Iñaki Esnaola, and Samir M. Perlaza are with the ECE Department at Princeton University, Princeton,NJ.
This work is supported by the University of Sheffield ACSE PGR scholarships, the Inria Exploratory Action – Information and Decision Making (AEx IDEM), and in part by a grant from the C3.ai Digital Transformation Institute.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Minimisation du Risque Empirique avec Régularisation par f-Divergences dans l'Apprentissage Statistique

Résumé : Ce rapport présente la solution au problème de minimisation du risque empirique avec régularisation par une f-divergence dans des faibles contraintes sur f. Dans ces conditions, la mesure optimale s'avère unique et existe toujours. La solution est présentée comme une expression sous forme fermée de la dérivée de Radon-Nikodym de la mesure de probabilité optimale par rapport à la mesure de référence. Des exemples de choix particuliers de la fonction f sont présentés. Pour certains choix, des résultats existants sont obtenus comme cas particuliers du résultat principal. Il s'agit notamment des solutions uniques aux problèmes de minimisation du risque empirique avec régularisation par entropie relative (Type-I et Type-II).

Mots-clés : Minimisation du Risque Empirique, f-divergence, Régularisation, apprentissage statistique.

Contents

1 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 labelled patterns available in the training dataset [1, 2]. In simpler terms, 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 resulting optimal model [6–8]. To remediate this phenomenon, the solution of the ERM shall exhibit a small sensitivity to variations in the training dataset, which is often obtained via regularization [9–14].

In statistical learning theory, the ERM problem consists in the minimization of the expected empirical risk over a subset of all probability measures that can be defined upon the set of models. In this case, regularization is often obtained by adding (up to a constant factor) a *statistical distance* from the optimization measure to a *reference measure* to the expected empirical risk (w.r.t. the optimization measure). 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. A key observation is that such a reference measure often represents prior knowledge and/or desired features on the solution.

A typical example of a statistical distance is an f-divergence. The notion of f-divergence was introduced in [15] and further studied in [16, 17]. A popular f-divergence is the relative entropy [18], which due to its asymmetry, leads to two different problems when it is used for regularizing the ERM problem. Those problems are known as Type-I and Type-II ERM with relative entropy regularization (ERM-RER). In the Type-I ERM-RER, the regularizer is the relative entropy of the optimization measure with respect to (w.r.t.) the reference measure [12]. Alternatively, in the Type-II ERM-RER, the regularizer is the relative entropy of the reference measure w.r.t. the optimization measure [13]. The Type-I ERM-RER problem exhibits a unique solution, which is a Gibbs probability measure. In the case in which the reference measure is a probability measure, the existence of the solution is always ensured. Alternatively, in the case in which it is a σ-finite measure, the existence of a solution is subject to strict conditions. The Type-II ERM-RER problem also exhibits a unique solution when the reference measure is a probability measure. Such a solution is a probability measure that exhibits properties that are analogous to those of the Gibbs probability measure, as pointed out in [13]. Nonetheless, despite these similarities, Type-I ERM-RER appears to be the most popular regularized ERM problem in statistical learning theory. See for instance, [19–30] and references therein.

Optimization problems with f-divergence regularization has been explored in [31,32] and [33]. Nonetheless, the problem of ERM with f-divergence regularization (ERM-fDR), with a general f function, has not been studied before. Only

the special cases, $f(x) = -x\log(x)$ (Type-I ERM-RER) and $f(x) = -\log(x)$ (Type-II ERM-RER), mentioned above have attracted particular attention. This work presents the solution to the ERM-fDR problem using a method of proof that differs from those in [31, 32] and [33] and goes along the lines of the methods in [12, 13] and [14], which rely on the notion of the Gateaux derivative [34].

The remainder of this report is organized as follows. Section 2 presents the standard ERM problem and introduces the notation. The ERM-fDR problem and its solution, which is the main contribution of this work, is presented in Section 3. Section 4 presents examples for specific choices of the function f. Finally, Section 5 draws the main conclusions and finalizes this report.

2 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*. Given the data points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) with $n \in \mathbb{N}$, a *dataset* is represented by the tuple $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$.

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

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

the objective is to obtain a model $\boldsymbol{\theta} \in \mathcal{M}$, such that, for all $i \in \{1, 2, \ldots, n\}$, the label assigned to patern x_i, which is $h(\boldsymbol{\theta}, x_i)$, is "close" to the label y_i specified by the training dataset. This notion of "closeness" is formalized by the function

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

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

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

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

The ERM problem w.r.t. the dataset $\boldsymbol{z}$ in (1) consists of the optimization problem:

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

The set of solutions to such a problem is denoted by

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

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

3 The ERM with f-Divergence Regularization

3.1 Preliminaries

For the ease of notation, the expected empirical risk with respect to a given measure is expressed via the following functional $\mathsf{R}_{\boldsymbol{z}}$, defined hereunder.

Definition 3.1 (Expected Empirical Risk)**.** *The expectation of the empirical risk $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})$ in (3), when $\boldsymbol{\theta}$ is sampled from a probability measure $P \in \triangle(\mathcal{M})$, is determined by the functional $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}) \to [0, +\infty)$, such that*

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

Similarly, a particular notation is used for the f-divergences, as shown in the following definition.

Definition 3.2 (f-divergence [17])**.** *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and $f(0) = \lim_{x \to 0^+} f(x)$. Let P and Q be two probability measures on the same measurable space, with P absolutely continuous with Q. The f-divergence of P w.r.t. Q, denoted by $\mathsf{D}_f(P\|Q)$, is*

$$\mathsf{D}_f(P\|Q) \triangleq \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{7}$$

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

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

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

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

3.2 Problem Formulation

The ERM-fDR problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M})$, a positive real λ, and an f-divergence (Definition 3.2). The measure Q is referred to as the *reference measure*, λ as the *regularization factor* and f as the

regularization function. Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (1), the ERM-fDR problem, with parameters Q , λ and f, consists of the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}_f(P\|Q). \tag{9}$$

3.3 Solution to the ERM-fDR

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

Theorem 3.1. *Let the function f in* (9) *be strictly convex and differentiable. Denote by $\dot{f} : \mathbb{R} \to \mathbb{R}$ and $\dot{f}^{-1} : \mathbb{R} \to (0, \infty)$ the derivative and the reciprocal of the function f, respectively, and assume that $\dot{f}^{-1}$ is strictly positive. Then, the optimization problem in* (9) *always possesses a unique solution, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, whose Radon-Nikodym derivative with respect to the probability measure Q satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$*

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

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3)*; and β is a real chosen such that*

$$\int \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{11}$$

Proof: The optimization problem in (9) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P w.r.t. the measure Q, which yields:

$$\min_{P \in \triangle_Q(\mathcal{M})} \left[\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) + \lambda \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) \right] \tag{12a}$$

$$\text{s.t.} \quad \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{12b}$$

The remainder of the proof focuses on the problem in which the optimization is over the function $\frac{\mathrm{d}P}{\mathrm{d}Q} : \mathcal{M} \to [0, \infty)$, which represents the Radon-Nikodym derivate of P w.r.t. Q. Hence, instead of optimizing the measure P, the optimization is over the function $\frac{\mathrm{d}P}{\mathrm{d}Q}$. This is due to the fact that for all $P \in \triangle_Q(\mathcal{M})$, the Radon-Nikodym derivate $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is unique up to sets of zero measure w.r.t. 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

$$\int |\hat{g}(\boldsymbol{\theta})|\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{13}$$

Hence, the optimization problem of interest is:

$$\min_{g\in\mathscr{M}}\left[\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})+\lambda\int f(g(\boldsymbol{\theta}))\mathrm{d}Q(\boldsymbol{\theta})\right] \tag{14a}$$

$$\text{s.t.}\int g(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta})=1. \tag{14b}$$

Let the Lagrangian of the optimization problem in (14) be $L:\mathscr{M}\times\mathbb{R}\to\mathbb{R}$ such that

$$L(g,\beta)=\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})+\lambda\int f(g(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})+\beta\left(\int g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})-1\right), \tag{15}$$

$$=-\beta+\int\Big(g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)+\lambda f(g(\boldsymbol{\theta}))\Big)\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{16}$$

where β is a real that acts as a Lagrange multiplier due to the constraint (14b).

Let $\hat{g}:\mathcal{M}\to\mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (15) at $(g,\beta)\in\mathscr{M}\times\mathbb{R}$ in the direction of $\hat{g}$, if it exists, is

$$\partial L(g,\beta;\hat{g})\triangleq\frac{\mathrm{d}}{\mathrm{d}\gamma}L(g+\gamma\hat{g},\beta)\bigg|_{\gamma=0}. \tag{17}$$

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

$$r(\gamma)=L(g+\gamma\hat{g},\beta) \tag{18}$$

$$=\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})+\lambda\int f(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})$$
$$+\beta\left(\int(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\mathrm{d}Q(\boldsymbol{\theta})-1\right) \tag{19}$$

$$=-\beta+\int g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)\mathrm{d}Q(\boldsymbol{\theta})+\gamma\int\hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)\mathrm{d}Q(\boldsymbol{\theta})$$
$$+\lambda\int f(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{20}$$

Note that the first two terms in (20) are independent of γ; the third term is linear with γ; and the fourth term can be written using the function $\hat{r}:\mathbb{R}\to\mathbb{R}$ such that for all $\gamma\in(-\epsilon,\epsilon)$, with ϵ arbitrarily small, it holds that

$$\hat{r}(\gamma)=\lambda\int f(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{21}$$

Hence, if the function $\hat{r}$ in (21) is differentiable at zero, so is the function r in (18), which implies that the Gateaux differential of the functional ∂L in (17) exists. The objective is to prove that the function $\hat{r}$ is differentiable at zero, which boils down to proving that the limit

$$\lim_{\delta\to0}\frac{1}{\delta}(\hat{r}(\gamma+\delta)-\hat{r}(\gamma)) \tag{22}$$

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 from the assumption that f is strictly convex, it follows that f is continuous. This implies that f is also 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

$$|f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))| \leqslant c\,|\hat{g}(\boldsymbol{\theta})|\,|\delta|, \tag{23}$$

for some constant c positive and finite, which implies that

$$\left|\frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}{\delta}\right| \leqslant c\,|\hat{g}(\boldsymbol{\theta})|. \tag{24}$$

From the assumption that f is differentiable, let $\dot{f} : (0, \infty) \to \mathbb{R}$ be the first derivative of f. Using these arguments, the limit in (22) satisfies for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, that

$$\begin{aligned}
&\lim_{\delta \to 0} \frac{1}{\delta}(\hat{r}(\gamma + \delta) - \hat{r}(\gamma)) \\
&= \lambda \lim_{\delta \to 0} \frac{1}{\delta}\left(\int f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) - \int f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right) && (25)\\
&= \lambda \lim_{\delta \to 0} \int \frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}{\delta}\,\mathrm{d}Q(\boldsymbol{\theta}) && (26)\\
&= \lambda \int \dot{f}(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\hat{g}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && (27)\\
&< \infty, && (28)
\end{aligned}$$

where both the equality in (27) and the inequality in (28) follow from noticing that the conditions for the dominated convergence theorem hold [35, Theorem 1.6.9], namely:

- For all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$, the inequality in (24) holds;
- The function $\hat{g}$ in (24) satisfies the inequality in (13); and
- For all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, it holds that

$$\begin{aligned}
\lim_{\delta \to 0} \frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}{\delta} &= \frac{\mathrm{d}}{\mathrm{d}\gamma} f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})) && (29)\\
&= \dot{f}(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\hat{g}(\boldsymbol{\theta}) && (30)
\end{aligned}$$

  which follows from the fact that f is assumed to be differentiable.

From (28), it follows that the function $\hat{r}$ in (21) is differentiable in the interval $(-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small. This implies that the function r in (18) is differentiable within the same interval. The derivative of the real function r in (18) at $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, is

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

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

Hence, the Gateaux differential $\partial L(g, \beta; \hat{g})$ in (17) exists and satisfies

$$\partial L(g, \beta; \hat{g}) = \frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) \Big|_{\gamma=0} \tag{33}$$

$$= \int \hat{g}(\boldsymbol{\theta}) \Big(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta + \lambda \dot{f}(g(\boldsymbol{\theta})) \Big) \mathrm{d}Q(\boldsymbol{\theta}). \tag{34}$$

The relevance of the Gateaux differential in (33) stems from [36, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (15) 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 $\hat{g} \in \mathscr{M}$,

$$\partial L \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta; \hat{g} \right) = 0. \tag{35}$$

From (35), 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

$$\int \hat{g}(\boldsymbol{\theta}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta + \lambda \dot{f} \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \right) \mathrm{d}Q(\boldsymbol{\theta}) = 0. \tag{36}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta + \lambda \dot{f} \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) = 0, \tag{37}$$

and thus,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1} \left(\frac{-\beta - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right), \tag{38}$$

where β is chosen to satisfy (11).

Now, the objective is to show the existence of such a β. For this purpose, 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 [37, Theorem 5.6.9]. Moreover, from the assumption that f is convex and differentiable, it holds that $\dot{f}$ is continuous [38, Proposition 5.44]. This implies that $\dot{f}^{-1}$ is continuous. These elements are used hereunder to study the function $k : \mathbb{R} \to (0, \infty)$ such that

$$k(t) = \int \dot{f}^{-1} \left(\frac{-t - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{39}$$

and prove that there always exists a $\beta \in \mathbb{R}$ such that $k(\beta) = 1$. The first step is to prove that the function k in (39) is continuous. This is proved by

showing that k always exhibits a limit. Note that from the fact that $\dot{f}^{-1}$ is strictly increasing, it holds that for all $\beta \in \mathbb{R}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\dot{f}^{-1}\left(\frac{-\beta - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \leqslant \dot{f}^{-1}\left(\frac{-\beta}{\lambda}\right), \tag{40}$$

where equality holds if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = 0$. Now, from the fact that $\dot{f}^{-1}$ is continuous it follows that for all $a \in \mathbb{R}$

$$\lim_{t \to a} \dot{f}^{-1}\left(\frac{-t - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) = \dot{f}^{-1}\left(\frac{-a - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right). \tag{41}$$

Hence, from the the dominated convergence theorem [35, Theorem 1.6.9], it holds that

$$\begin{aligned}
\lim_{t \to a} k(t) &= \lim_{t \to a} \int \dot{f}^{-1}\left(\frac{-\beta - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) && (42)\\
&= \int \left(\lim_{t \to a} \dot{f}^{-1}\left(\frac{-\beta - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) && (43)\\
&= \int \dot{f}^{-1}\left(\frac{-a - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) && (44)\\
&= k(a), && (45)
\end{aligned}$$

which proves that the function k in (39) is continuous. Note also that such a function k is strictly decreasing, from the fact that $\dot{f}^{-1}$ is strictly increasing. The proof continues by showing that there always exists a pair $(t_1, t_2) \in \mathbb{R}^2$, such that $k(t_1) < 1 \leqslant k(t_2)$. Consider the set $\mathcal{A}_{t_1} = \left\{\boldsymbol{\theta} \in \mathcal{M} : \dot{f}(1) \leqslant \left(\frac{-t_1 - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right\} = \left\{\boldsymbol{\theta} \in \mathcal{M} : \dot{f}^{-1}\left(\frac{-t_1 - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \geqslant 1\right\}$ and let $\underline{\mathsf{L}}_{\boldsymbol{z}}$ and $\bar{\mathsf{L}}_{\boldsymbol{z}}$ be two reals such that

$$\begin{aligned}
\underline{\mathsf{L}}_{\boldsymbol{z}} &= \inf\{t \in \mathbb{R} : t = \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \operatorname{supp} Q\}, \text{ and} && (46)\\
\bar{\mathsf{L}}_{\boldsymbol{z}} &= \sup\{t \in \mathbb{R} : t = \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \operatorname{supp} Q\}, && (47)
\end{aligned}$$

and note that $0 \leqslant \underline{\mathsf{L}}_{\boldsymbol{z}} \leqslant \bar{\mathsf{L}}_{\boldsymbol{z}} < +\infty$.

Assume that $1 < k(t_1)$. Hence,

$$\begin{aligned}
1 &< k(t_1) && (48)\\
&= \int_{\mathcal{A}_{t_1}} \dot{f}^{-1}\left(-\frac{t_1 + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M} \setminus \mathcal{A}_{t_1}} \dot{f}^{-1}\left(-\frac{t_1 + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) && (49)\\
&\leqslant \int_{\mathcal{A}_{t_1}} \dot{f}^{-1}\left(-\frac{t_1 + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M} \setminus \mathcal{A}_{t_1}} \mathrm{d}Q(\boldsymbol{\theta}) && (50)\\
&\leqslant \int_{\mathcal{A}_{t_1}} \dot{f}^{-1}\left(-\frac{t_1 + \underline{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M} \setminus \mathcal{A}_{t_1}} \mathrm{d}Q(\boldsymbol{\theta}) && (51)\\
&\leqslant Q(\mathcal{A}_{t_1}) \dot{f}^{-1}\left(-\frac{t_1 + \underline{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}\right) + 1 - Q(\mathcal{A}_{t_1}), && (52)
\end{aligned}$$

which implies that t_1 must satify $\dot{f}^{-1}\left(-\frac{t_1+\underline{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}\right) \geqslant 1$. This is the same as requiring that

$$\dot{f}(1) < -\frac{t_1+\underline{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}, \tag{53}$$

as a consequence of the function $\dot{f}$ monotonically increasing.

Assume now that $1 \geqslant k(t_2)$. Hence,

$$\begin{aligned}
1 \geqslant & k(t_2) && (54)\\
= & \int_{\mathcal{A}_{t_2}} \dot{f}^{-1}\left(-\frac{t_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M}\setminus\mathcal{A}_{t_2}} \dot{f}^{-1}\left(-\frac{t_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) && (55)\\
\geqslant & \int_{\mathcal{A}_{t_2}} \mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M}\setminus\mathcal{A}_{t_2}} \dot{f}^{-1}\left(-\frac{t_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) && (56)\\
\geqslant & \int_{\mathcal{A}_{t_2}} \mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M}\setminus\mathcal{A}_{t_2}} \dot{f}^{-1}\left(-\frac{t_2+\bar{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) && (57)\\
= & Q(\mathcal{A}_{t_2}) + (1-Q(\mathcal{A}_{t_2}))\dot{f}^{-1}\left(-\frac{t_2+\bar{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}\right), && (58)
\end{aligned}$$

which implies that t_2 must satify $\dot{f}^{-1}\left(-\frac{t_2+\bar{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}\right) \leqslant 1$. This is the same as requiring that

$$\dot{f}(1) \geqslant -\frac{t_2+\bar{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}, \tag{59}$$

as a consequence of the function $\dot{f}$ monotonically increasing. It has already been established that $\dot{f}$ is continuous and thus, there always exist two reals t_1 and t_2 that satisfy that

$$-\frac{t_2+\bar{\mathsf{L}}_{\boldsymbol{z}}}{\lambda} < \dot{f}(1) \leqslant -\frac{t_1+\underline{\mathsf{L}}_{\boldsymbol{z}}}{\lambda}, \tag{60}$$

which implies that $k(t_1) < 1 \leqslant k(t_2)$. Hence, given that the function k is continuous, strictly decreasing, and there always exists two reals t_1 and t_2 such that $k(t_1) < 1 \leqslant k(t_2)$, it follows from the intermediate-value theorem [39, Theorem 4.23] that there always exists a unique real t such that $k(t) = 1$.

Finally, note that the objective function in (14) is the sum of two terms. The first one, i.e., $\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta})$, is linear with g. The second, i.e., $\int f(g(\boldsymbol{\theta}))\mathrm{d}Q(\boldsymbol{\theta})$, 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 $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. ■

4 Examples

This section describes some examples for particular choices of the function f in Theorem 3.1.

4.1 Kullback-Leibler Divergence

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that

$$f(x) = x \log(x), \tag{61}$$

whose derivative satisfies

$$\dot{f}(x) = 1 + \log(x). \tag{62}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is the relative entropy of P with respect to Q. From equation (62) and Theorem 3.1 it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$

$$1 + \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) = -\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}, \tag{63}$$

which implies

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-\frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \tag{64}$$

$$= \frac{\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)}{\int \exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\nu)\right)\mathrm{d}Q(\boldsymbol{\nu})}, \tag{65}$$

which is the result independently by several authors in [12, 25, 27, 29, 40], and references therein.

4.2 Reverse Relative Entropy Divergence

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that

$$f(x) = -\log(x), \tag{66}$$

whose derivative satisfies

$$\dot{f}(x) = -\frac{1}{x}. \tag{67}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is the relative entropy of Q with respect to P. From equation (67) and Theorem 3.1 it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$

$$-\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1} = -\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}, \tag{68}$$

which implies

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{69}$$

which is the result reported in [14, 41].

4.3 Jeffrey's Divergence

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that

$$f(x) = x \log(x) - \log(x), \tag{70}$$

such that

$$\dot{f}(x) = \log(x) + 1 - \frac{1}{x}. \tag{71}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is Jeffrey's divergence between P and Q. From equation (71) and Theorem 3.1 it holds that,

$$\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + 1 - \frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) = -\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}, \tag{72}$$

which implies

$$0 = -\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) + \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + 1 + \frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \tag{73}$$

$$= -1 + \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \tag{74}$$

$$= \exp\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) - 1 \tag{75}$$

$$= \left(\exp\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) - 1\right)\frac{\exp\left(\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\exp\left(\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)} \tag{76}$$

$$= \exp\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) - \exp\left(\frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right). \tag{77}$$

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

$$\exp\left(\frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)$$

$$= \exp\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)$$
$$\left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right). \tag{78}$$

Hence, from the equality in (77), it holds that

$$W_0\left(\exp\left(\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right)$$
$$= \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}, \tag{79}$$

which in terms of the Lambert function yields

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$$
$$= \exp\left(W_0\left(\exp\left(\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right) - \frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right). \tag{80}$$

4.4 Hellinger Divergence

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that

$$f(x) = \left(1-\sqrt{x}\right)^2, \tag{81}$$

which implies

$$\dot{f}(x) = 1 - \frac{1}{\sqrt{x}}. \tag{82}$$

From equation (82) and Theorem 3.1 it holds that,

$$1 - \frac{1}{\sqrt{\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})}} = -\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}. \tag{83}$$

Hence, from (83) it follows that

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \left(\frac{\lambda}{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right)^2. \tag{84}$$

4.5 Jennsen-Shannon Divergence

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that

$$f(x) = x \log\left(\frac{2x}{x+1}\right) + \log\left(\frac{2}{x+1}\right), \tag{85}$$

which implies

$$\dot{f}(x) = \log\left(\frac{2x}{x+1}\right). \tag{86}$$

From equation (86) and Theorem 3.1 it holds that,

$$\log\left(\frac{2\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})}{\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) + 1}\right) = -\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}. \tag{87}$$

Hence, from (87) it follows that

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{2 - \exp\left(-\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)} \tag{88}$$

$$= \frac{1}{2\exp\left(\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) - 1}. \tag{89}$$

5 Conclusions

This work has presented the solution to the ERM-fDR 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 always exists. This result allows obtaining closed-form expressions for ERM problems with regularizations by relative entropy ($f(x) = x\log(x)$), reverse relative entropy ($f(x) = -\log(x)$); Jeffrey's divergence ($f(x) = x\log(x) - \log(x)$); Hellinger divergence ($f(x) = (1-\sqrt{x})^2$); Jensen-Shannon divergence ($f(x) = x\log\left(\frac{2x}{x+1}\right) + \log\left(\frac{2}{x+1}\right)$); among others. The result is limited in the sense that popular regularizations such as the one by total variation ($f(x) = |x-1|$) cannot be studied via the main result. This is because in this case, f is not strictly convex.

References

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

[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 relative entropy regularization," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9454, Feb. 2022.

[13] 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-9575, May. 2023.

[14] 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.

[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," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

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

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "Sensitivity of the Gibbs algorithm to data aggregation in supervised machine learning," IN-

RIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9474, Jun. 2022.

[29] 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.

[30] F. Futami and T. Iwata, "Information-theoretic analysis of test data sensitivity in uncertainty," arXiv preprint arXiv:2307.12456, Jul. 2023.

[31] M. Teboulle, "Entropic proximal mappings with applications to nonlinear programming," *Mathematics of Operations Research*, vol. 17, no. 3, pp. 670–690, Aug. 1992.

[32] 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.

[33] 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.

[34] 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.

[35] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA, USA: Academic Press, 2000.

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

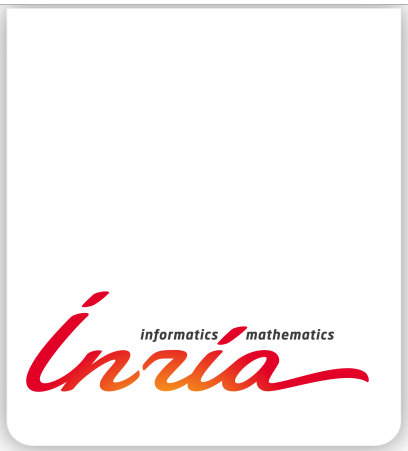
[37] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 2nd ed. New York, NY, USA: Wiley New York, 2000.

[38] J. Douchet, *Analyse : Recueil d'Exercices et Aide-Mémoire*, 3rd ed. Lausanne, Switzerland: PPUR, 2010, vol. 1.

[39] W. Rudin, *Principles of mathematical analysis*, 1st ed. New York, NY, USA: McGraw-Hill Book Company, Inc., 1953.

[40] 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.

[41] 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.



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

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