A Dual Optimization View to Empirical Risk Minimization with f-Divergence Regularization

Francisco Daunas*†, Iñaki Esnaola *‡, and Samir M. Perlaza†‡¶,
*Sch. of Electrical and Electronic Engineering, University of Sheffield, Sheffield, U.K.
†INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France.
‡ECE Dept., Princeton University, Princeton, 08544 NJ, USA.
¶GAATI, Université de la Polynésie Française, Faaa, French Polynesia.

***Abstract*—The dual formulation of empirical risk minimization with f-divergence regularization (ERM-fDR) is introduced. The solution of the dual optimization problem to the ERM-fDR is connected to the notion of normalization function introduced as an implicit function. This dual approach leverages the Legendre-Fenchel transform and the implicit function theorem to provide a nonlinear ODE expression to the normalization function. Furthermore, the nonlinear ODE expression and its properties provide a computationally efficient method to calculate the normalization function of the ERM-fDR solution under a mild condition.**

***Index Terms*—empirical risk minimization; f-divergence regularization, statistical learning, normalization function.**

I. Introduction

Empirical risk minimization (ERM) [1]–[6] is often posed as an optimization problem regularized by a *statistical distance* between the probability measure to be optimized and a given reference measure [7]–[13]. A well-studied case for such statistical distance is the relative entropy, which leads to the celebrated ERM with relative entropy regularization (ERM-RER) studied, for instance, in [14]–[21]. Other works address the general case for f-divergences, known as ERM with f-divergence regularization (ERM-fDR) [22]–[25]. The discrete case of ERM-fDR is explored in [22] and [23], while more general settings are covered in [24] and [25]. Recent results have shown that f-divergence regularization can improve the robustness of learning algorithms in the context of distributionally robust optimization (DRO) [26]–[29].

However, a main challenge in both the theory and practice of ERM-fDR is that the solution is known only up to a normalization factor [22]–[24]. For many f-divergences, a closed-form expression for the normalization factor is not known. This hinders sampling methods like Markov Chain Monte Carlo (MCMC) [30], where it appears in the likelihood ratio, which is needed for valid transition probabilities. In the case of rejection sampling [31], the absence of such a constant prevents defining an efficient proposal distribution. Even when the normalization factor has a known expression, such as the log-partition function in the case of relative entropy regularization [32], it can be intractable for general loss functions. More broadly, in ERM-fDR problems, computing this normalization factor is especially difficult because it requires evaluating the empirical risk across all models in the support of the reference measure, a task that is $\#P$-hard [33]–[35].

This paper addresses this challenge by formalizing the concept of the normalization function introduced in [25] and deriving the dual optimization problem for ERM-fDR. The benefits of studying the dual optimization problem and its properties are twofold. First, insights for computing the normalization factor are obtained, and second, a characterization of the normalization function for ERM-fDR problems is presented. This characterization is obtained by leveraging tools from convex analysis, notably the Legendre-Fenchel transform [36], [37], and the implicit function theorem [38].

This work is supported in part by a University of Sheffield ACSE PGR scholarship; University of Sheffield PGR Publication Scholarship scheme; the European Commission through the H2020-MSCA-RISE-2019 project 872172; the French National Agency for Research (ANR) through the Project ANR-21-CE25-0013 and the project ANR-22-PEFT-0010 of the France 2030 program PEPR Réseaux du Futur; and the Agence de l'innovation de défense (AID) through the project UK-FR 2024352.

II. Preliminaries

Let Ω be an arbitrary subset of $\mathbb{R}^d$, with $d \in \mathbb{N}$, and let $\mathscr{B}(\Omega)$ denote the Borel σ-field on Ω. The set of probability measures that can be defined upon the measurable space $(\Omega, \mathscr{B}(\Omega))$ is denoted by $\triangle(\Omega)$. Given a probability measure $Q \in \triangle(\Omega)$ the set exclusively containing the probability measures in $\triangle(\Omega)$ that are absolutely continuous with respect to Q is denoted by $\triangle_Q(\Omega)$. That is, $\triangle_Q(\Omega) \triangleq \{P \in \triangle(\Omega) : P \ll Q\}$, where the notation $P \ll Q$ stands for the measure P being absolutely continuous with respect to the measure Q. The Radon-Nikodym derivative of the measure P with respect to Q is denoted by $\frac{\mathrm{d}P}{\mathrm{d}Q} : \Omega \rightarrow [0, \infty)$.

Using this notation, an f-divergence is defined as follows.

Definition 1 (f-divergence [39]): 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 $P \ll Q$. The f-divergence of P with respect to Q, denoted by $\mathsf{D}_f(P\|Q)$, is

$$\mathsf{D}_f(P\|Q) \triangleq \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega)\right) \mathrm{d}Q(\omega). \quad (1)$$

In the case in which the function f in (1) is continuous and differentiable, the derivative of the function f is denoted by $\dot{f} : (0, +\infty) \rightarrow \mathbb{R}$. If the inverse of the function $\dot{f}$ exists, it is denoted by

$$\dot{f}^{-1} : \mathbb{R} \rightarrow (0, +\infty). \quad (2)$$

III. The Learning 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 $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$. 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

$$\boldsymbol{z} = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n, \quad (3)$$

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 the pattern x_i, which is $h(\boldsymbol{\theta}, x_i)$, is "close" to the label y_i. This notion of "closeness" is formalized by the function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$, such that the loss or risk induced by choosing the model $\boldsymbol{\theta} \in \mathcal{M}$ with respect to the labeled 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 to 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 (3) is determined by the function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to [0, +\infty)$, which satisfies

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

The expectation of the empirical risk $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})$ in (4), 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

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}). \quad (5)$$

The ERM-fDR problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M})$, a positive real λ, and a function $f : [0, \infty) \to \mathbb{R}$ that satisfies the conditions in Definition 1. The measure Q is referred to as the *reference measure* and λ as the *regularization factor*.

Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (3), the ERM-fDR problem, with parameters Q, λ and f, is given by the following optimization problem

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}_f(P \| Q), \quad (6)$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (5). The set of solutions to (6) is the singleton $\{Q\}$ in the case in which for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = c$, for some $c > 0$. This distinction is mathematically significant but can be ignored in practice, as it arises only when $\mathsf{R}_{\boldsymbol{z}}(P)$ in (5) is constant for all measures P. In order to avoid the above case, the notion of separable empirical risk functions [18, Definition 5] is adopted.

The solution to the ERM-fDR problems in (6) was first presented in [25, Theorem 1] under the following assumptions:

(a) The function f is strictly convex and twice differentiable;

(b) There exists a β such that

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \dot{f}^{-1}\left(-\frac{t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right) \right\}, \quad (7a)$$

and

$$\int \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right) \mathrm{d}Q(\boldsymbol{\theta}) = 1, \quad (7b)$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (4); and

(c) The function $\mathsf{L}_{\boldsymbol{z}}$ in (4) is separable with respect to the probability measure Q.

Theorem 1 ([25, Theorem 1]): Under Assumptions (a) and (b), the solution to the optimization problem in (6), denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle_Q(\mathcal{M})$, is unique, and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right). \quad (8)$$

The equality in (8) can be written in terms of the *normalization function*, introduced in [25] and defined hereunder.

Definition 2 (Normalization Function): The normalization function of the problem in (6), denoted by

$$N_{Q,\boldsymbol{z}} : \mathcal{A}_{Q,\boldsymbol{z}} \to \mathcal{B}_{Q,\boldsymbol{z}}, \quad (9a)$$

with $\mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0, \infty)$ and $\mathcal{B}_{Q,\boldsymbol{z}} \subseteq \mathbb{R}$, is such that for all $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$,

$$\int \dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right) \mathrm{d}Q(\boldsymbol{\theta}) = 1. \quad (9b)$$

The set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (9a) contains all the regularization factors λ for which Assumption (b) is satisfied. More specifically, it contains the regularization factors λ for which the problem in (6) has a solution. Furthermore, the equality in (9b) justifies referring to the function $N_{Q,\boldsymbol{z}}$ as the *normalization function*, as it ensures that the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (8) is a probability measure.

This section ends by highlighting that the probability measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q in (8) are mutually absolutely continuous [25, Corollary 1]. This is important, as it ensures the Radon-Nikodym derivative of Q with respect to $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is well defined, a property that is used repeatedly throughout the proofs.

IV. The ERM-fDR Dual Problem

The duality principle [37, Chapter 5] enables the analysis of the optimization problem in (6) by studying an alternative form, known as the dual problem. In this section, this dual problem is derived using the Legendre-Fenchel transform [37], which is defined below.

Definition 3 (Legendre-Fenchel transform [37]): Consider a function $f : \mathcal{I}_1 \to \mathcal{I}_2$, with $\mathcal{I}_i \subseteq \mathbb{R}$, with $i \in \{1, 2\}$. The Legendre-Fenchel transform of the function f, denoted by $f^* : \mathcal{J} \to \mathbb{R}$, is

$$f^*(t) = \sup_{x \in \mathcal{I}_1} (tx - f(x)), \quad (10)$$

with

$$\mathcal{J} = \{t \in \mathbb{R} : f^*(t) < \infty\}. \quad (11)$$

Using this notation, consider the following problem

$$\min_{\beta \in \mathbb{R}} \quad \lambda \int f^*\Big(-\frac{1}{\lambda}(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \Big) \mathrm{d}Q(\boldsymbol{\theta}) + \beta, \quad (12)$$

where the real λ, the measure Q and the function f are those in (6); and the functions $\mathsf{L}_{\boldsymbol{z}}$ and f^* are defined in (4) and (10), respectively. The following theorem introduces the solution to the problem in (12).

Theorem 2: Under Assumptions (a) and (b), the solution to the optimization problem in (12) is $N_{Q,\boldsymbol{z}}(\lambda)$, where the function $N_{Q,\boldsymbol{z}}$ is defined in (9).

Proof: Let $G : \mathbb{R} \to \mathbb{R}$ be a function such that

$$G(\beta) = \lambda \int \Big(-\frac{1}{\lambda}(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\, \mathrm{d}Q(\boldsymbol{\theta}) + \beta, \quad (13)$$

which is the objective function of the optimization problem in (12). Note that from Assumption (a) and Definition 3, the function G in (13) is a strictly convex function, which satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\beta}G(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}\Big(\lambda\int f^*\Big(-\frac{1}{\lambda}(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\, \mathrm{d}Q(\boldsymbol{\theta})+\beta\Big) \quad (14)$$

$$= -\int \dot{f}^*\Big(-\frac{1}{\lambda}(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\, \mathrm{d}Q(\boldsymbol{\theta}) + 1, \quad (15)$$

where $\dot{f}^*$ is the derivative of the function f^* in (12). Let the solution to the optimization problem in (13) be denoted by $\widehat{\beta} \in \mathbb{R}$ and note that the derivative of the function G evaluated at $\widehat{\beta}$ is equal to zero, that is

$$\int \dot{f}^*\Big(-\frac{1}{\lambda}(\widehat{\beta} + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\, \mathrm{d}Q(\boldsymbol{\theta}) = 1. \quad (16)$$

From [36, Corollary 23.5.1] and Assumption (a), the following equality holds for all $t \in \mathcal{J}$, with $\mathcal{J}$ in (11),

$$\frac{\mathrm{d}}{\mathrm{d}t}f^*(t) = \dot{f}^*(t) = \dot{f}^{-1}(t), \quad (17)$$

where the functions $\dot{f}^{-1}$ and $\dot{f}^*$ are defined in (2) and (15), respectively. From (16) and (17), it follows that

$$\int \dot{f}^{-1}\Big(-\frac{1}{\lambda}(\widehat{\beta} + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\, \mathrm{d}Q(\boldsymbol{\theta}) = 1, \quad (18)$$

which combined with (9b) and Assumption (b) yields

$$N_{Q,\boldsymbol{z}}(\lambda) = \widehat{\beta}. \quad (19)$$

Note that $\hat{\beta}$ is unique as it is the minimum of a strictly convex function. This observation completes the proof. ■

The following lemma establishes that the problem in (12) is the dual problem to the ERM-fDR problem in (6) and characterizes the difference between their optimal values, which is often referred to as duality gap [40, Section 8.3].

Lemma 1: Under Assumptions (a) and (b), the optimization problem in (12) is the dual problem to the ERM-fDR problem in (6). Moreover, the duality gap is zero.

Proof: Under Assumption (a) and [37, Section 3.3.2], it can be verified that for all $t \in \mathcal{J}$, with $\mathcal{J}$ in (11), the function f^* in (10) satisfies

$$f^*(t) = t\dot{f}^*(t) - f(\dot{f}^*(t)), \quad (20)$$

where the function $\dot{f}^*$ is the same as in (17). From Assumption (a) and (17), the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (8) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^*\Big(-\frac{N_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big), \quad (21)$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $N_{Q,\boldsymbol{z}}$ are defined in (4) and (9), respectively. Then, from (20) and (21), with some algebraic manipulations, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

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

$$= -\lambda f^*\Big(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) - N_{Q,\boldsymbol{z}}(\lambda). \quad (22)$$

Taking the expectation in both sides of (22) with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (8) yields

$$\mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) + \lambda \mathsf{D}_f\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\Big)$$

$$= -\lambda \int f^*\Big(-\frac{1}{\lambda}(N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\, \mathrm{d}Q(\boldsymbol{\theta}) - N_{Q,\boldsymbol{z}}(\lambda). \quad (23)$$

Using Theorem 1 and Theorem 2 in the left-hand and right-hand sides of (23), respectively, yields

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

$$= \max_{\beta \in \mathbb{R}} -\lambda \int f^*\Big(-\frac{1}{\lambda}(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\, \mathrm{d}Q(\boldsymbol{\theta}) - \beta. \quad (24)$$

The claim that the optimization problem in (12) is the dual to the ERM-fDR problem in (6) follows from (24) and [41, Theorem 1, Section 8.4]. The zero duality gap is established by the equality in (24), which completes the proof. ■

The zero duality gap in Lemma 1 ensures strong duality, which implies that the optimal value of the dual variables recovers the primal optimal measure. However, solving the dual problem in (12) remains equally challenging, as the solution must satisfy Assumption (b) in Theorem 1. Thus, unless the normalization function $N_{Q,\boldsymbol{z}}$ in (9) is explicitly characterized, computing $N_{Q,\boldsymbol{z}}(\lambda)$ involves a search over the real line to obtain the value that satisfies (18).

V. Analysis of the Normalization Function

A. Characterization and Properties

The purpose of this section is to characterize the function $N_{Q,\boldsymbol{z}}$ and the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{B}_{Q,\boldsymbol{z}}$ in (9). Given a real $\delta \in [0, \infty)$, consider the Rashomon set $\mathcal{L}_{\boldsymbol{z}}(\delta)$, which is defined as follows $\mathcal{L}_{\boldsymbol{z}}(\delta) \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leq \delta\}$. Consider also the real numbers $\delta^\star_{Q,\boldsymbol{z}}$ and $\lambda^\star_{Q,\boldsymbol{z}}$ defined as follows

$$\delta^\star_{Q,\boldsymbol{z}} \triangleq \inf\{\delta \in [0,\infty) : Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0\}, \quad (25)$$

and

$$\lambda^\star_{Q,\boldsymbol{z}} \triangleq \inf \mathcal{A}_{Q,\boldsymbol{z}}. \quad (26)$$

Using this notation, the following theorem introduces one of the main properties of the function $N_{Q,\boldsymbol{z}}$.

Theorem 3: The function $N_{Q,\boldsymbol{z}}$ in (9) is strictly increasing and continuous within the interior of $\mathcal{A}_{Q,\boldsymbol{z}}$ in (9a). Furthermore, for all $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$,

$$N_{Q,\boldsymbol{z}}(\lambda) = \lambda\frac{\mathrm{d}}{\mathrm{d}\lambda}N_{Q,\boldsymbol{z}}(\lambda) - \mathsf{R}_{\boldsymbol{z}}(P_N), \quad (27)$$

where the probability measure $P_N \in \triangle_Q(\mathcal{M})$ satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P_N}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{\ddot{f}\Big(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\Big)}\Bigg(\int \frac{1}{\ddot{f}\Big(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\Big)}\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1}. \quad (28)$$

B. Proof of Theorem 3

The proof is divided into two parts. The first part leverages the properties of the function f under Assumption (a) in Theorem 1 to prove that the normalization function $N_{Q,\boldsymbol{z}}: \mathcal{A}_{Q,\boldsymbol{z}} \to \mathcal{B}_{Q,\boldsymbol{z}}$ in (9) is strictly increasing. The second part proves the continuity of the function $N_{Q,\boldsymbol{z}}$ in (9).

The first part is as follows. Given a pair $(a,b) \in \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{B}_{Q,\boldsymbol{z}}$, with $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{B}_{Q,\boldsymbol{z}}$ in (9), assume that

$$N_{Q,\boldsymbol{z}}(a) = b. \tag{29}$$

This implies that

$$1=\int \frac{\mathrm{d}P^{(Q,b)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) =\int \dot{f}^{-1}\Big(-\frac{1}{a}(b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\mathrm{d}Q(\boldsymbol{\theta}). \tag{30}$$

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 $\mathcal{B}_{Q,\boldsymbol{z}}$ [42, Theorem 5.6.9]. Moreover, from the assumption that f is strictly convex and differentiable, it holds that $\dot{f}$ is continuous [43, Proposition 5.44]. This implies that $\dot{f}^{-1}$ is continuous. From [44, Lemma A.3], the function $\dot{f}^{-1}$ is strictly increasing such that for all $b \in \mathcal{B}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\dot{f}^{-1}\Big(-\frac{1}{a}(b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big) \leq \dot{f}^{-1}\Big(-\frac{1}{a}\big(b+\delta^{\star}_{Q,\boldsymbol{z}}\big)\Big) < \infty, \tag{31}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ defined in (25) and equality holds if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}$. Then, from (31) and $\mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0,\infty)$, which implies $a > 0$, it follows that

$$\int \dot{f}^{-1}\Big(-\frac{1}{a}(b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big)\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{32}$$

Note that from [44, Lemma A.5] and the assumption in (30) for all $(a_1,a_2) \in \mathcal{A}^2_{Q,\boldsymbol{z}}$, such that $a_1 < a < a_2$, it holds that

$$\int \dot{f}^{-1}\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b}{a_1}\Big)\mathrm{d}Q(\boldsymbol{\theta}) < 1 < \int \dot{f}^{-1}\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b}{a_2}\Big)\mathrm{d}Q(\boldsymbol{\theta}). \tag{33}$$

Then, for $N_{Q,\boldsymbol{z}}(a_i)$, with $i \in \{1,2\}$, to satisfy,

$$\int \dot{f}^{-1}\left(-\frac{1}{a_i}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(a_i))\right)\mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{34}$$

under the assumption that $a_1 < a < a_2$, it must hold that $N_{Q,\boldsymbol{z}}(a_1)$ and $N_{Q,\boldsymbol{z}}(a_2)$ satisfy

$$N_{Q,\boldsymbol{z}}(a_1) < b < N_{Q,\boldsymbol{z}}(a_2). \tag{35}$$

This implies that the function $N_{Q,\boldsymbol{z}}$ in (9) is strictly increasing. Similarly, from [44, Lemma A.5] and the assumption in (30) for all $(b_1,b_2) \in \mathcal{B}^2_{Q,\boldsymbol{z}}$, such that $b_1 < b < b_2$, it holds that

$$\int \dot{f}^{-1}\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b_1}{a}\Big)\mathrm{d}Q(\boldsymbol{\theta}) > 1 > \int \dot{f}^{-1}\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b_2}{a}\Big)\mathrm{d}Q(\boldsymbol{\theta}). \tag{36}$$

Then, for the a_i, with $i \in \{1,2\}$ to satisfy,

$$\int \dot{f}^{-1}\left(-\frac{1}{a_i}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + b_i)\right)\mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{37}$$

under the assumption that $b_1 < b < b_2$, it holds that a_1 and a_2 satisfy

$$a_1 < a < a_2, \tag{38}$$

which implies that the function $N_{Q,\boldsymbol{z}}$ in (9) is strictly increasing. Furthermore, from (35) and (38) the function $N_{Q,\boldsymbol{z}}$ maps one to one the elements of $\mathcal{A}_{Q,\boldsymbol{z}}$ into $\mathcal{B}_{Q,\boldsymbol{z}}$, which implies it is bijective. Thus, the inverse $N^{-1}_{Q,\boldsymbol{z}}: \mathcal{B}_{Q,\boldsymbol{z}} \to \mathcal{A}_{Q,\boldsymbol{z}}$ is well-defined. This completes the proof of the first part.

In the second part, the objective is to prove the continuity of the function $N_{Q,\boldsymbol{z}}$. To do so, an auxiliary function is introduced and proven to be continuous. Under the assumptions (a), (b) and (c) from Theorem 1, the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{B}_{Q,\boldsymbol{z}}$ in (9) are non-empty and the real values: $\bar{a} = \sup \mathcal{A}_{Q,\boldsymbol{z}}$; $\underline{a} = \inf \mathcal{A}_{Q,\boldsymbol{z}}$; $\bar{b} = \sup \mathcal{B}_{Q,\boldsymbol{z}}$; and $\underline{b} = \inf \mathcal{B}_{Q,\boldsymbol{z}}$, are such that

$$\mathcal{A} = (\underline{a},\bar{a}) \subseteq (0,\infty), \text{ and } \mathcal{B} = \big(\underline{b},\bar{b}\big) \subseteq \mathbb{R}. \tag{39}$$

Let the function $F: \mathcal{A}\times\mathcal{B} \to (0,\infty)$ be

$$F(a,b)=\int \dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) - 1. \tag{40}$$

The first step is to prove that the function F in (40) is continuous on the sets $\mathcal{A}$ and $\mathcal{B}$ defined in (39), respectively. This is proved by showing that F always exhibits a limit in $\mathcal{A}$ and $\mathcal{B}$, and that limit is the same as the function evaluated at that value. Then, for all $(a,b) \in \mathcal{A}\times\mathcal{B}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, the function $\dot{f}^{-1}$ satisfies the bound in (31). Hence, from [36, Corollary 24.5.1] the function $\dot{f}^{-1}$ is continuous, such that for all $(a,b) \in \mathcal{A}\times\mathcal{B}$, it holds that

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

Note that the function F in (40) is bounded as a consequence of (31), and thus, from the *dominated convergence theorem* [45, Theorem 1.6.9], the following limits exist and satisfy

$$\lim_{b\to\beta} F(a,b) = F(\beta,a), \tag{42}$$

and

$$\lim_{a\to\lambda} F(a,b) = F(\lambda,b), \tag{43}$$

which proves that the function F in (40) is continuous in $\mathcal{A}$ and $\mathcal{B}$. The proof continues by noting that from the definition of $\mathcal{A}$ and $\mathcal{B}$ in (39) there exists at least one point $(\lambda,\beta) \in \mathcal{A}\times\mathcal{B}$, such that $(\lambda,\beta) \in \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{B}_{Q,\boldsymbol{z}}$, which implies that

$$F(\lambda,\beta) = 0. \tag{44}$$

Note that from (42) and (43) the function F is continuous and thus the partial derivative of F satisfy

$$\frac{\partial}{\partial a}F(a,b)=\int \frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a^2}\frac{1}{\ddot{f}\Big(\dot{f}^{-1}\Big(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\Big)\Big)}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{45}$$

where (45) follows from [44, Lemma A.2]; and

$$\frac{\partial}{\partial b}F(a,b)=\int -\frac{1}{a}\frac{1}{\ddot{f}\Big(\dot{f}^{-1}\Big(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\Big)\Big)}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{46}$$

where (46) follows from [44, Lemma A.2]. Then, from the *implicit function theorem* presented in [38, Theorem 4], the function $N_{Q,\boldsymbol{z}}$ exists and is unique in the open interval $\mathcal{A}$

with $\mathcal{A}$ in (39) and for all $a \in \mathcal{A}$ satisfies that

$$N_{Q,\boldsymbol{z}}(a) = b, \tag{47}$$

such that $F(a, N_{Q,\boldsymbol{z}}(a)) = 0$, which completes the proof of continuity for the normalization function $N_{Q,\boldsymbol{z}}$. Additionally, from [38, Theorem 4] it follows that

$$\frac{\mathrm{d}}{\mathrm{d}a}N_{Q,\boldsymbol{z}}(a) = -\left(\frac{\partial}{\partial b}F(a, N_{Q,\boldsymbol{z}}(a))\right)^{-1}\frac{\partial}{\partial a}F(a, N_{Q,\boldsymbol{z}}(a)), \tag{48}$$

$$= \frac{N_{Q,\boldsymbol{z}}(a)}{a} + \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a} g_a(\boldsymbol{\theta}) \,\mathrm{d}Q(\boldsymbol{\theta}), \tag{49}$$

with the function $g_a : \mathcal{M} \to \mathbb{R}$, such that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$g_a(\boldsymbol{\theta}) = \frac{\mathrm{d}P_N}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{50}$$

with $\frac{\mathrm{d}P_N}{\mathrm{d}Q}$ defined in (28). Note that from the assumption that f is strictly convex and twice differentiable, the derivative $\dot{f}$ is increasing, and the second derivative $\ddot{f}$ is positive for all $\boldsymbol{\theta} \in \operatorname{supp} Q$. Also, the denominator of the fraction is the integral of the reciprocal of $\ddot{f}\left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\right)$ with respect to the measure Q. This term serves as a normalization constant ensuring that the resulting function is a proper probability density such that $\int g_a(\boldsymbol{\theta}) \,\mathrm{d}Q(\boldsymbol{\theta}) = 1$. Therefore, the function g_a in (50) can be interpreted as the Radon-Nikodym derivative of a new probability measure $P^{(a)}$, parametrized by the regularization factor a with respect to Q. Specifically, if we define a measure $P^{(a)}$ such that for any set $\mathcal{A} \in \mathscr{F}_{\mathcal{M}}$,

$$P^{(a)}(\mathcal{A}) = \int_{\mathcal{A}} g_a(\boldsymbol{\theta}) \,\mathrm{d}Q(\boldsymbol{\theta}). \tag{51}$$

From (49) and (51), it follows that

$$N_{Q,\boldsymbol{z}}(a) = a\frac{\mathrm{d}}{\mathrm{d}a}N_{Q,\boldsymbol{z}}(a) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(a)}\right), \tag{52}$$

with $\mathsf{R}_{\boldsymbol{z}}$ defined in (5). This completes the proof of the derivative of the normalization function. ■

The continuity and monotonicity exhibited by the function $N_{Q,\boldsymbol{z}}$ allow the following characterization of the set $\mathcal{A}_{Q,\boldsymbol{z}}$.

Lemma 2: The set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (9a) is either empty or an interval that satisfies $\left(\lambda^\star_{Q,\boldsymbol{z}}, \infty\right) \subseteq \mathcal{A}_{Q,\boldsymbol{z}} \subseteq \left[\lambda^\star_{Q,\boldsymbol{z}}, \infty\right)$, with $\lambda^\star_{Q,\boldsymbol{z}}$ in (26) and satisfies $\lambda^\star_{Q,\boldsymbol{z}} \geq 0$.

Proof: The proof is presented in [44, Appendix B.1]. ■

Lemma 2 highlights two facts. First, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is a convex subset of positive reals. Second, if there exists a solution to the optimization problem in (6) for some $\lambda > 0$, then there exists a solution to such a problem when λ is replaced by $\bar{\lambda} \in (\lambda, \infty)$.

C. Discussion of Results

A major bottleneck in ERM-fDR solutions stems from the computational challenges associated with: (i) evaluating expectations with respect to the prior Q, and (ii) determining the value of $N_{Q,\boldsymbol{z}}(\lambda)$ in (8) for implementing the resulting algorithms. This subsection discusses the significance of the continuity and monotonicity properties of the normalization function $N_{Q,\boldsymbol{z}}$ in (9) and their practical implications for addressing challenge (ii), under the assumption that expectations with respect to the prior Q are computationally tractable. To begin, note that the function $N_{Q,\boldsymbol{z}}$ induces a bijection between the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{B}_{Q,\boldsymbol{z}}$, as a result of the monotonicity and continuity established in Theorem 3. For the cases in which no explicit expression for the normalization function is available, this bijection allows the ERM-fDR solution to be computed directly by selecting a normalization factor and using its inverse to determine the corresponding regularization parameter λ. For example, this approach is used in [17, Eq (26)] for the reverse relative entropy. More importantly, in cases in which neither the normalization function $N_{Q,\boldsymbol{z}}$ nor its inverse can be explicitly defined, as is the case for several f-divergences (see [25]); the monotonicity and continuity of $N_{Q,\boldsymbol{z}}$ allow replacing an exhaustive search over all real values with a root-finding algorithm to approximate the solution, ensuring the existence of the root in (44) and convergence to it. Thus, such an algorithm reduces the number of evaluations of the expectation with respect to Q to find such an approximation. The following Algorithm makes use of the above properties of the normalization function $N_{Q,\boldsymbol{z}}$.

Algorithm 1 Algorithm for $N_{Q,\boldsymbol{z}}(\lambda)$ via Root-finding

Input: $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to [0, \infty)$, Q, $f : (0, \infty) \to \mathbb{R}$, $\lambda > 0$, $\delta^\star_{Q,\boldsymbol{z}}$, tolerance $\epsilon > 0$, max iterations $N_{\max}$

Output: β satisfying $\left|\int \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1\right| \leq \epsilon$

Initialisation:

$b_{\text{low}} \leftarrow \delta^\star_{Q,\boldsymbol{z}} - \lambda\dot{f}(0)$, $b_{\text{high}} \leftarrow \lambda$, $b \leftarrow \frac{1}{2}(b_{\text{low}} + b_{\text{high}})$

$n \leftarrow 0$, $I \leftarrow \int \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta})$

Root-Finding Process:

while $|I - 1| > \epsilon$ **and** $n < N_{\max}$ **do**

  if $I > 1$ **then**

    $b_{\text{high}} \leftarrow b$

  else

    $b_{\text{low}} \leftarrow b$

  end if

  $b \leftarrow \frac{1}{2}(b_{\text{low}} + b_{\text{high}})$

  $I \leftarrow \int \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta})$, $n \leftarrow n + 1$

end while; **return** b

VI. Conclusions

This paper establishes a connection between the ERM-fDR solution and the Legendre-Fenchel transform via the dual formulation in (12). By analyzing the relationship between the primal and dual problems in (6) and (12), it is shown that the duality gap is zero. The identity arising from strong duality is leveraged to prove the continuity and monotonicity of the normalization function, enabling application of the implicit function theorem. This, in turn, yields a nonlinear ODE characterizing the normalization function. These properties are then used to design an algorithm that approximates the normalization function for a tractable reference measure and fixed regularization factor. This improves over searching across the real line when an explicit characterization is unavailable.

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 and A. Chervonenkis, "Theory of pattern recognition," vol. 1, no. 1, Oct. 1974.

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

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

[5] 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.

[6] 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.

[7] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and maximal leakage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, May 2021.

[8] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 577–581.

[9] A. S. Dalalyan and A. G. Karagulyan, "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient," arXiv preprint arXiv:1710.00095, Feb. 2024.

[10] W. Li, N. Klein, B. Gifford, E. Sklute, C. Legett, and S. Clegg, "Regularization via f-divergence: An application to multi-oxide spectroscopic analysis," arXiv preprint arXiv:2502.03755, Feb. 2025.

[11] F. Peng, M. Zhang, and M. Tang, "An information-theoretic analysis for federated learning under concept drift," arXiv preprint arXiv:2506.21036, Jun. 2025.

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

[13] Z. Huang and S. Becker, "Stochastic gradient Langevin dynamics with variance reduction," arXiv preprint arXiv:2102.06759, Jul. 2021.

[14] 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.

[15] 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.

[16] 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.

[17] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Asymmetry of the relative entropy in the regularization of empirical risk minimization," *IEEE Transactions on Information Theory*, vol. 71, no. 8, pp. 6198–6226, Aug. 2025.

[18] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 5122 – 5161, Jul. 2024.

[19] 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.

[20] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, Jan. 2020.

[21] X. Zou, S. M. Perlaza, I. Esnaola, E. Altman, and H. V. Poor, "The worst-case data-generating probability measure in statistical learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, no. 1, pp. 175 – 189, Apr. 2024.

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

[23] 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.

[24] 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.

[25] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Equivalence of empirical risk minimization to regularization on the family of f-divergences," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, Jul. 2024.

[26] J. Wei and Y. Liu, "When optimizing f-divergence is robust with label noise," arXiv preprint arXiv:2011.03687, Aug. 2021.

[27] Z. Liu, B. P. Van Parys, and H. Lam, "Smoothed f-divergence distributionally robust optimization," arXiv preprint arXiv:2306.14041, Oct. 2023.

[28] H. Namkoong and J. C. Duchi, "Stochastic gradient methods for distributionally robust optimization with f-divergences," *Advances in Neural Information Processing Systems*, vol. 26, pp. 1476–1485, Dec. 2016.

[29] Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, vol. 1, no. 2, Nov. 2013.

[30] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, May 1970.

[31] W. R. Gilks and P. Wild, "Adaptive rejection sampling for Gibbs sampling," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 2, pp. 337–348, Jun. 1992.

[32] 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.

[33] A. Bulatov and M. Grohe, "The complexity of partition functions," *Theoretical Computer Science*, vol. 348, no. 2, pp. 148–186, Sep. 2005.

[34] A. A. Bulatov, "The complexity of the counting constraint satisfaction problem," *Journal of the ACM (JACM)*, vol. 60, no. 5, pp. 1–41, Sep. 2014.

[35] C. McQuillan, "The computational complexity of approximation of partition functions," PhD thesis, University of Liverpool, Liverpool, UK, Jun. 2013.

[36] R. T. Rockafellar, *Conjugate Convex Functions in Optimal Control and the Calculus of Variations*, 2nd ed. Princeton, NJ, USA: Princeton University Press, 1970.

[37] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*, 1st ed. Cambridge, UK: Cambridge University Press, 2004.

[38] O. de Oliveira, "The Implicit and Inverse Function Theorems: Easy Proofs," *Real Analysis Exchange*, vol. 39, no. 1, pp. 207 – 218, 2013.

[39] 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.

[40] D. G. Luenberger, "Observing the state of a linear system," *IEEE Transactions on Military Electronics*, vol. 8, no. 2, pp. 74–80, Apr. 1964.

[41] ——, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: Wiley, 1997.

[42] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 3rd ed. New York, NY, USA: Wiley New York, 2000.

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

[44] F. Daunas, I. Esnaola, and S. M. Perlaza, "A dual optimization view to empirical risk minimization with f-divergence regularization," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9594, Jul. 2025.

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