# Empirical Risk Minimization with Relative Entropy Regularization Type-II

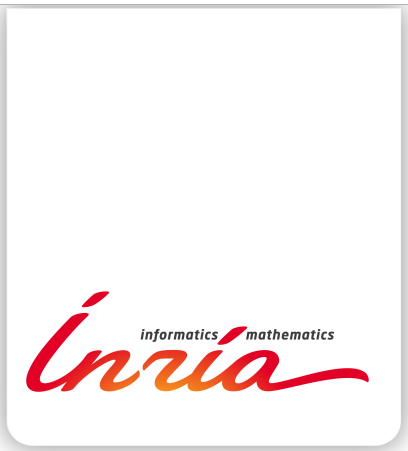
Francisco Daunas, Iñaki Esnaola, Samir M. Perlaza, H. Vincent Poor


# Empirical Risk Minimization with Relative Entropy Regularization Type-II

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

**RESEARCH REPORT**
**N° 9508**
May 2023
Project-Team NEO

ISSN 0249-6399 ISRN INRIA/RR--9508--FR+ENG




# Empirical Risk Minimization with Relative Entropy Regularization Type-II

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

Project-Team NEO

Research Report n° 9508 — version 2 — initial version May 2023 — revised version February 2024 — 59 pages

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

**Key-words:** Supervised Learning, Empirical Risk Minimization; Relative Entropy; Regularization; Gibbs Measure; Inductive Bias; Sensitivity; and Generalization.

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.

**RESEARCH CENTRE**
**SOPHIA ANTIPOLIS – MÉDITERRANÉE**

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

# Minimisation du Risque Empirique avec Régularisation par l'Entropie Relative Tapez-II

**Résumé :** L'effet de l'asymétrie de l'entropie relative est analysé dans le problème de la minimisation du risque empirique avec régularisation par entropie relative (MRE-RER). Une nouvelle régularisation est introduite, appelée régularisation de Type-II, qui permet la recherche des solutions au problème de la MRE-RER avec un support qui s'étend en dehors du support de la mesure de référence. La solution au nouveau problème de la Type-II MRE-RER est caractérisée analytiquement en termes de la dérivée de Radon-Nikodym de la mesure de référence par rapport à la solution. L'analyse de la solution dévoile certaines propriétés de l'entropie relative lorsqu'elle agit comme régularisateur du problème de la MRE-RER : (a) l'entropie relative force le support de la solution de Type-II à s'étaler sur tout le support de la mesure de référence, ce qui introduit un fort biais inductif qui domine l'évidence fournie par les données d'entraînement ; (b) La régularisation de Type-II est équivalente à la régularisation d'entropie relative classique avec une transformation appropriée de la fonction du risque empirique. Enfin, une expression sous forme explicite de la valeur espérée du risque empirique en fonction des paramètres de régularisation est présentée.

**Mots-clés :** Apprentissage Supervisé; Minimisation du Risque Empirique; Entropie Relative; Régularisation; Algorithme de Gibbs; Mesure de Gibbs; Sensitivité; et Généralisation

# Contents

# 1 Introduction

Empirical risk minimization (ERM) is a central tool in supervised machine learning. Among other uses, it enables the characterization of sample complexity and probably approximately correct (PAC) learning in a wide range of settings [1]. The application of ERM in the study of theoretical guarantees spans related disciplines such as machine learning [2], information theory [3,4] and statistics [5,6]. Classical problems such as classification [7,8], pattern recognition [9,10], regression [11,12], and density estimation [9,13] can be posed as special cases of the ERM problem [13,14]. Unfortunately, ERM is prone to training data memorization, a phenomenon also known as overfitting [15–17]. For that reason, ERM is often regularized in order to provide generalization guarantees [18–21]. Regularization establishes a preference over the models by encoding features of interest that conform to prior knowledge. In different statistical learning frameworks, such as Bayesian learning [22,23] and PAC learning [24–26], the prior knowledge over the set of models can be described by a reference probability measure. More general references can be adapted as proved in [27, 28] for the case of $\sigma$-finite measures. Prior knowledge on the set of datasets can also be represented by probability measures, e.g., the worst-case data-generating probability measure introduced in [29]. In either case, the solution to the regularized ERM problem can be cast as a probability distribution over the set of models.

A common regularizer of the ERM problem is the relative entropy of the optimization probability measure with respect to a given reference measure over the set of models [13, 30–32]. The resulting problem formulation, termed ERM with relative entropy regularization (ERM-RER) has been extensively studied for both the case in which the reference measure is a probability measure [30–33] and the case in which it is a $\sigma$-finite measure [27, 28, 34]. While in both cases the solution is unique and corresponds to a Gibbs probability measure, the existence of the solution is ensured only in the case in which the reference measure is a probability measure [28]. Despite the many merits of the ERM-RER formulation, it has some significant limitations. Firstly, absolute continuity of the optimization measure with respect to the reference measure is required for the existence of the corresponding Radon-Nikodym derivative, which is used by the relative entropy regularization. This absolute continuity sets an insurmountable barrier to the exploration of models outside the support of the reference measure. More specifically, models outside the support of the reference measure exhibit zero probability with respect to the Gibbs probability measure solution to ERM-RER. More importantly, this holds regardless of the evidence provided by the training dataset. Secondly, the choice of relative entropy over alternative divergences often follows arguments based on the simplicity of obtaining generalization guarantees in the form of bounds [18]. Nonetheless, such bounds are often hard to calculate and are not always informative when evaluated in practical settings [35–37].

From all the above, exploring the asymmetry of the relative entropy is of particular interest to advancing the understanding of entropy regularization and

its role in generalization. The problem of ERM with a general $f$-divergence regularization has been explored in [38] and [39] in the case of a finite countable set of models, and recently extended to uncountable sets of models in [40]. Nonetheless, the authors in [38–40] constrain the optimization domains to sets of measures that are mutually absolutely continuous with respect to the reference probability. The use of the relative entropy of the optimization measure with respect to the reference in ERM-RER is termed Type-I ERM-RER. Alternatively, the use of the relative entropy of the reference measure with respect to the optimization measure, introduced here, is termed Type-II ERM-RER. Interestingly, the existing results in [38–40], which lead to special cases of the Type-I and Type-II ERM-RER problems by assuming that $f(x) = -x\log(x)$ and $f(x) = -\log(x)$, respectively, do not study the impact of the asymmetry of relative entropy in the context of ERM regularization.

This paper presents the solution to Type-II ERM-RER optimization problem using a new method of proof. In particular, mutual absolute continuity of the measures in the optimization domain with respect to the reference measure is not imposed. Nonetheless, such a mutual absolute continuity is exhibited by the solution as a consequence of the structure of the problem. The main properties of the solution are highlighted and an equivalence between Type-I and a Type-II ERM-RER problems is presented by replacing the empirical risk in the Type-I problem by another function, which can be cast as a tunable loss function as in [41–43].

The remainder of the paper is organized as follows. Section 2 presents the standard ERM problem. Section 3 describes the Type-I regularization. The main contribution of this paper, which is the solution to the Type-II ERM-RER problem, is presented in Section 4. Section 7 studies the equivalence between Type-I and Type-II regularization. The conclusions are summarized in Section 8.

## 2 Empirical Risk Minimization

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

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

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

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

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

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

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

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

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

Let the set of solutions to the ERM problem in (4) be denoted by

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

Note that if the set $\mathcal{M}$ is finite, the ERM problem in (4) always possesses a solution, and thus, $|\mathcal{T}(\boldsymbol{z})| > 0$. Nonetheless, in general, the ERM problem does not necessarily possess a solution, *i.e.*, it might happen that $|\mathcal{T}(\boldsymbol{z})| = 0$.

The PAC and Bayesian frameworks, c.f. in [23] and [25], solve the problem in (4) by constructing probability measures conditioned on the dataset $\boldsymbol{z}$, from which models are randomly sampled. In this context, finding probability measures that are minimizers of the ERM problem in (4) over the set of all probability measures that can be defined on the measurable space $(\mathcal{M}, \mathscr{F})$, which is denoted by $\triangle(\mathcal{M})$, requires a metric that enables assessing the goodness of the probability measure. From this perspective, the underlying assumption in the reminder of this work is that the functions $f$ and $\ell$ in (3) are such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $g_{x,y} : \mathcal{M} \to [0, \infty]$, such that $g_{x,y}(\boldsymbol{\theta}) = \ell(f_{\boldsymbol{\theta}}(x), y)$, is measurable with respect to the Borel measurable spaces $(\mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, where $\mathscr{F}$ and $\mathscr{B}(\mathbb{R})$ are respectively the Borel $\sigma$-fields on $\mathcal{M}$ and $\mathbb{R}$. Under these assumptions, a common metric is the notion of expected empirical risk.

**Definition 2.1** (Expected Empirical Risk)**.** *Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in* (1)*, for all probability measures $P \in \triangle(\mathcal{M})$ let the functional $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}) \to [0, \infty)$ be such that*

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

*where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in* (3)*.*

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

# 3 The Type-I ERM-RER Problem

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

*measure* and $\lambda$ as the *regularization factor*. The Type-I ERM-RER problem, with parameters $Q$ and $\lambda$, is given by the following optimization problem:

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

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

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

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

The solution to the Type-I ERM-RER problem in (7) is the Gibbs probability measure [27, 30, 31], which is presented in the following lemma.

**Lemma 3.1** ( [28, Theorem 3.1]). *The solution of the Type-I ERM-RER problem in* (7)*, denoted by* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle(\mathcal{M})$*, is unique, always exists, and satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right), \tag{9}$$

*where the function* $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ *is such that*

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp(t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right). \tag{10}$$

# 4 The Type-II ERM-RER Problem

The Type-II ERM-RER problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M})$ and a real $\lambda \in (0, \infty)$. As in Type-I ERM-RER problem, the measure $Q$ is referred to as the *reference measure* and $\lambda$ as the *regularization factor*. Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (1), the Type-II ERM-RER problem, with parameters $Q$ and $\lambda$, consists of the following optimization problem:

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

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

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

The difference between Type-I and Type-II ERM-RER problems lies on the regularization. While the former uses the relative entropy $\mathsf{D}(P\|Q)$, the latter uses $\mathsf{D}(Q\|P)$. This translates into different optimization domains due to the asymmetry of the relative entropy. More specifically, in the Type-I ERM-RER problem, the optimization domain is the set of probability measures on the Borel measurable space $(\mathcal{M}, \mathscr{F})$ that are absolutely continuous with the reference

measure $Q$. That is, the set $\triangle_Q(\mathcal{M})$ in (8). Alternatively, in the Type-II ERM-RER problem, the optimization domain consists of probability measures defined on the Borel measurable space $(\mathcal{M}, \mathscr{F})$, with the additional condition that the reference measure $Q$ must be absolutely continuous with respect to them. This corresponds to the set denoted as $\bigtriangledown_Q(\mathcal{M})$ in Equation (12). From this perspective, the techniques used in [28] for solving the Type-I ERM-RER no longer hold. As shown in the next section, a new technique is used for solving the Type-II ERM-RER.

The problem in (11) exhibits a trivial solution when the functional $\mathsf{R}_{\boldsymbol{z}}$ is such that for all $P \in \bigtriangledown_Q(\mathcal{M})$, it holds that $\mathsf{R}_{\boldsymbol{z}}(P) = c$, for some $c \in [0, \infty)$. In such a case, the solution is unique and equal to the probability measure $Q$, independently of the parameter $\lambda$. In order to avoid this trivial case, which arises under particular conditions over the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) and the probability measure $Q$, the notion of separability of the empirical risk function with respect to the measure $Q$ is borrowed from [28]. A separable empirical risk function with respect to a given probability measure $P$ is defined as follows.

**Definition 4.1** (Definition 4.1 in [28])**.** *The empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *is said to be separable with respect to the probability measure* $P \in \triangle(\mathcal{M})$*, if there exist a positive real* $c > 0$ *and two subsets* $\mathcal{A}$ *and* $\mathcal{B}$ *of* $\mathcal{M}$ *that are nonnegligible with respect to* $P$*, and for all* $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$*,*

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

A nonseparable empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) with respect to a measure $P$ is a constant almost surely with respect to the measure $P$. More specifically, there exists a real $a \geq 0$, such that

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

When the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is nonseparable with respect to all measures in $P \in \bigtriangledown_Q(\mathcal{M})$, the trivial case described above is observed. The notion of separable empirical risk functions would play a central role in the study of the optimization problem in (11).

## 4.1 The Solution to the Type-II ERM-RER Problem

The solution of the Type-II ERM-RER problem in (11) is presented in the following theorem, where the measure $Q$ is a parameter of the optimization problem in (11) and the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3).

**Theorem 4.1.** *If there exists a real* $\beta$ *such that*

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

*and*

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

*then the solution to the optimization problem in* (11)*, denoted by* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle(\mathcal{M})$*, is unique and for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*, it satisfies*

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

Before introducing the proof of Theorem 4.1, two important results are presented. The first result consists in the solution to the optimization problem in (11) when the optimization domain is restricted to

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

where the sets $\triangle_Q(\mathcal{M})$ and $\bigtriangledown_Q(\mathcal{M})$ are defined in (8) and (12), respectively. Such an ancillary problem can be formulated as follows:

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

The solution to such an ancillary problem is described by the following lemma.

**Lemma 4.1.** *The solution to the optimization problem in* (18) *is unique and identical to the probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16)*.*

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

The second result consists of comparing the optimal values resulting from the optimizations in (11) and (18). The following lemma formalizes this result.

**Lemma 4.2.** *The optimization problems in* (11) *and* (18) *satisfy*

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

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

Lemma 4.2 unveils the fact that the objective function in (11) when evaluated at measures whose support extends beyond the support of $Q$ is larger than such an objective function evaluated at measures whose support is identical to the reference measure. This includes the case in which the set $\mathcal{T}(\boldsymbol{z})$ in (5) lies outside the support of $Q$.

Using these results, the proof of Theorem 4.1 is as follows.

*Proof of Theorem 4.1:* The proof follows by observing that from (17), it holds that

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

Hence, from (20), it follows that

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

From the inequalities in (19) and (21), it follows that

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

Thus, the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) is the solution of the optimization problem in (11), which completes the proof of Theorem 4.1. ■

Lemma 4.2 implies that the solution to the optimization problem in (11) is in the set $\bigcirc_Q(\mathcal{M})$ in (16). A consequence of this observation is the following corollary.

**Corollary 4.3.** *The probability measures $Q$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (16) *are mutually absolutely continuous.*

Corollary 4.3 also follows from Theorem 4.1 by observing that the solution to the Type-II ERM-RER problem in (11) is expressed in terms of its Radon-Nikodym with respect to $Q$, which implies the absolute continuity of $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to $Q$. The absolute continuity of the measure $Q$ with respect to $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ follows from the optimization domain of the Type-II ERM-RER problem. From this perspective, Corollary 4.3 conveys the fact that there does not exist a dataset that can overcome the inductive bias induced by the reference measure $Q$. That is, set of models outside the support of $Q$ exhibit zero probability measure with respect to the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

This observation is important as at a first glance, the Type-II relative entropy regularization for the ERM problem in (11) does not restrict the solution to be absolutely continuous with respect to the reference measure $Q$. However, Theorem 4.1 shows that the support of the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) collapses into the support of the reference. A parallel can be established between Type-I and Type-II cases, as in both cases the support of the solution is the support of the reference measure. In a nutshell, the use of relative entropy regularization inadvertently forces the solution to coincide with the support of the reference regardless of the training data.

## 4.2 Properties of the Solution

This section introduces a function referred to as the *normalization function* and its properties. This function as well as its properties are central for studying the solution to the ERM-RER problem in (11).

### 4.2.1 The Normalization Function

Let the function

$$\bar{K}_{Q,\boldsymbol{z}} : (0,\infty) \to \mathcal{A}, \tag{23a}$$

with $\mathcal{A} \subseteq \mathbb{R}$, and $Q$ and $\boldsymbol{z}$ in (11), be such that for all $t \in (0,\infty)$

$$\bar{K}_{Q,\boldsymbol{z}}(t) = \alpha, \tag{23b}$$

where $\alpha$ satisfies

$$\int \frac{t}{\alpha + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{24}$$

and the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3). The function $\bar{K}_{Q,\boldsymbol{z}}$ is referred to as the *normalization function*. This is essentially due to the observation that $\beta$ and $\lambda$ in (16) satisfy

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) = \beta, \tag{25}$$

which guarantees that the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle(\mathcal{M})$ in (16) is a probability measure. Hence, the Radon-Nikodym derivative in (16) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$

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

The properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) are studied using the notion of Rashamon sets [44]. Given a real $\delta \in [0, \infty)$, consider the Rashamon set

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leq \delta\}. \tag{27}$$

Consider also the nonnegative real number

$$\delta^{\star}_{Q,\boldsymbol{z}} \triangleq \inf\{\delta \in [0, \infty) : Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0\}. \tag{28}$$

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

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}. \tag{29}$$

Using the objects defined above, the following lemma introduces one of the main properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23).

**Lemma 4.4.** *The function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is strictly increasing and continuous.*

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

As a consequence of Lemma 4.4, the continuity of function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) and equality (24), it follows that for all $\alpha \in \mathcal{A}$, with $\mathcal{A}$ in (23a), the functional inverse $\bar{K}^{-1}_{Q,\boldsymbol{z}} : \mathcal{A} \to (0, \infty)$ is given by

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\alpha) = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \alpha} \, \mathrm{d}Q(\boldsymbol{\theta})}. \tag{30}$$

Note that the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (30) allows defining the codomain of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b), which is formalized by the following lemma.

**Lemma 4.5.** *The set $\mathcal{A}$ in* (23a) *satisfies*

$$\big(-\delta^\star_{Q,\boldsymbol{z}},\infty\big) \subseteq \mathcal{A} \subseteq \big[-\delta^\star_{Q,\boldsymbol{z}},\infty\big). \tag{31}$$

*Moreover, the set $\mathcal{A}$ is identical to $\big[-\delta^\star_{Q,\boldsymbol{z}},\infty\big)$ if and only if*

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}}\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{32}$$

*with $\delta^\star_{Q,\boldsymbol{z}}$ defined in* (28).

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

Observe that for all finite sets of models $\mathcal{M}$, the subset $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ as defined in (29) satisfies $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) > 0$. This implies that the integral in (32) is not finite. The formalization of this observation is presented in the following corollary.

**Corollary 4.6.** *If the set $\mathcal{M}$ is finite, then the set $\mathcal{A}$ in* (23a) *is identical to* $\big(-\delta^\star_{Q,\boldsymbol{z}},\infty\big)$.

An additional remark, is that from Lemma 4.4 and Lemma 4.5, if the set $\mathcal{A}$ in (23a) is equal to the closed set $\big[\delta^\star_{Q,\boldsymbol{z}},\infty\big)$ in (31), there exists a minimum regularization factor $\lambda^\star > 0$, where $\lambda^\star = \bar{K}^{-1}_{Q,\boldsymbol{z}}\big(-\delta^\star_{Q,\boldsymbol{z}}\big)$, with $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ defined in (30). This implies that the mapping of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is such that $\bar{K}_{Q,\boldsymbol{z}} : [\lambda^\star,\infty) \to \big[-\delta^\star_{Q,\boldsymbol{z}},\infty\big)$. This minimum regularization implies that the existence of a solution to the optimization problem in (11) only exists for regularization factors $\lambda \geq \lambda^\star$.

Based on Lemma 4.5, the existence of the minimum regularization factor depends on whether the integral in (32) is finite. The finiteness of the integral is conditioned by the reference measure $Q$; the functions $\ell$ and $f$ in (3); and the dataset $\boldsymbol{z}$ in (1). The subsequent examples illustrate situations where the set $\mathcal{A}$ is the open set $\big(\delta^\star_{Q,\boldsymbol{z}},\infty\big)$ and closed set $\big[\delta^\star_{Q,\boldsymbol{z}},\infty\big)$ in (31).

**Example 4.1.** *Consider the Type-II ERM-RER problem in* (11) *and assume that: (a) $\mathcal{M} = \mathcal{X} = \mathcal{Y} = [0,\infty)$; (b) $\boldsymbol{z} = (1,0)$ and (c) $Q \ll \mu$, with $\mu$ the Lebesgue measure, such that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu}(\boldsymbol{\theta}) = 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}). \tag{33}$$

*Let also the function $f : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ be*

$$f(\boldsymbol{\theta}, x) = x\boldsymbol{\theta}, \tag{34}$$

*and the loss function $\ell$ in* (2) *be*

$$\ell(f(\boldsymbol{\theta}, x), y) = (x\boldsymbol{\theta} - y)^2, \tag{35}$$

*which implies*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = (x\boldsymbol{\theta} - y)^2, \tag{36}$$

*with the function* $\mathsf{L}_{\boldsymbol{z}}$ *defined in* (3)*. Under the current assumptions the objective of this example is to show that* $\mathcal{A} = \left[\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$*. For this purpose, it is sufficient to show that the inequality in* (32) *holds. From Theorem 4.1, it follows that* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}\mu}(\boldsymbol{\theta}) = \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}), \tag{37}$$

*with* $\beta$ *satisfying* (15)*. Thus,*

$$\begin{aligned}
\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) &= \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}) \,\mathrm{d}\mu(\boldsymbol{\theta}) && \text{(38a)}\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(x\boldsymbol{\theta} - y)^2 - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}\boldsymbol{\theta} && \text{(38b)}\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}^2 - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}\boldsymbol{\theta} && \text{(38c)}\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}^2} \,\mathrm{d}\boldsymbol{\theta} && \text{(38d)}\\
&= \int_0^\infty 4 \exp(-2\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} && \text{(38e)}\\
&= 2, && \text{(38f)}
\end{aligned}$$

*where equality* (38a) *follows from equality* (33)*; equality* (38c) *follows from the assumption that* $(x, y) = (1, 0)$*; and equality* (38d) *follows from the fact that* $\delta^\star_{Q,\boldsymbol{z}} = 0$*. Finally, the function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (23) *satisfies* $\bar{K}_{Q,\boldsymbol{z}}\left(\frac{1}{2}\right) = 0$*, which implies* $\delta^\star_{Q,\boldsymbol{z}} = 0 \in \mathcal{A}$*.*

**Example 4.2.** *Consider Example 4.1 with* $\boldsymbol{z} = (1, 1)$*. Under the current assumptions, the objective of this example is to show that* $\mathcal{A} = \left(\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$*. For this purpose, it is sufficient to show that the inequality in* (32) *does not hold:*

$$\begin{aligned}
\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) &= \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}) \,\mathrm{d}\mu(\boldsymbol{\theta}) && \text{(39a)}\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(x\boldsymbol{\theta} - y)^2 - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}\boldsymbol{\theta} && \text{(39b)}\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \,\mathrm{d}\boldsymbol{\theta} && \text{(39c)}\\
&= \infty. && \text{(39d)}
\end{aligned}$$

*where equality* (39a) *follows from equality* (33)*; equality* (39b) *follows from the assumption that* $(x, y) = (1, 0)$*; equality* (39c) *follows from the fact that* $\delta^\star_{Q,\boldsymbol{z}} = 0$*; and the equality* (39d) *follows from an algebraic development shown in Appendix T. Finally, the function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (23) *is undefined at zero, which implies* $\delta^\star_{Q,\boldsymbol{z}} = 0 \notin \mathcal{A}$*.*

The aforementioned examples demonstrate that even if the reference measure $Q$ and functions $\ell$ and $f$ in (3) are fixed, the set $\mathcal{A}$ might be either $\left[\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$ or $\left(\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$ depending on the dataset $\boldsymbol{z}$. This observation underscores that the existence of the minimum regularization factor $\lambda^\star$ is coupled on the specific choices of $Q$, $\ell$, $f$, and $\boldsymbol{z}$. Finally, from Lemma 4.5 it is implied that if the set $\mathcal{M}$ is finite, then the set $\mathcal{A}$ is the open set $\left(-\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$ in (31), as shown by the following example.

**Example 4.3.** *Consider the Type-II ERM-RER problem in* (11) *and assume that:* (*a*) $\mathcal{B}$ *is a proper subset of* $\mathcal{M}$*, and* (*b*) *the probability measure* $Q$ *satisfies*

$$Q(\mathcal{B}) = \epsilon, \qquad \textit{and} \tag{40a}$$

$$Q(\mathcal{M} \setminus \mathcal{B}) = 1 - \epsilon, \tag{40b}$$

*with* $\epsilon > 0$*. Let the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *be*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \textit{if} \quad \boldsymbol{\theta} \in \mathcal{B} \\ c & \textit{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{B}, \end{cases} \tag{41}$$

*with* $c > 0$*.*

*Under the current assumptions, the objective of this example is to show that for all* $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$*, it holds that* $\mathcal{A} = \left(\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$*. For this purpose, it is sufficient to show that for all* $c \in (0, \infty)$*, the function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (23b) *is strictly greater than* $-\delta^\star_{Q,\boldsymbol{z}}$*. From the current assumptions it follows that* $\delta^\star_{Q,\boldsymbol{z}} = 0$ *and that for all* $\lambda \in (0, +\infty)$ *the function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (23b) *satisfies*

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})}. \tag{42}$$

*The proof of equality* (42) *is presented in appendix S. From equality* (42)*, it holds that*

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})} \tag{43a}$$

$$> -\frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2} \tag{43b}$$

$$= -\frac{(c-\lambda)}{2} + \left|\frac{c-\lambda}{2}\right| \tag{43c}$$

$$\geq 0 \tag{43d}$$

$$= -\delta^\star_{Q,\boldsymbol{z}}, \tag{43e}$$

*which proves that for all* $c \in (0, \infty)$*,* $\mathcal{A} = \left(-\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$*.*

Another immediate consequence of Lemma 4.4 is the following corollary.

**Corollary 4.7.** *If the real value* $\delta^\star_{Q,\boldsymbol{z}} = 0$*, with* $\delta^\star_{Q,\boldsymbol{z}}$ *in* (28)*, then the function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (23b) *is strictly positive.*

The following lemma characterizes the limit of $\bar{K}_{Q,\boldsymbol{z}}$ as $\lambda$ approaches zero from the right, under the assumption that the set $\mathcal{A}$ is $\left(-\delta^\star_{Q,\boldsymbol{z}},\infty\right)$.

**Lemma 4.8.** *If the set $\mathcal{A}$ is $\left(-\delta^\star_{Q,\boldsymbol{z}},\infty\right)$, the function $\bar{K}_{Q,\boldsymbol{z}}$ in* (23b) *satisfies*

$$\lim_{\lambda\to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\delta^\star_{Q,\boldsymbol{z}}, \tag{44}$$

*where $\delta^\star_{Q,\boldsymbol{z}}$ is defined in* (28).

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

Studying cases in which the choice of the reference measure $Q$, empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ and dataset $\boldsymbol{z}$ induce the set $\mathcal{A}$ to be equal $\left(-\delta^\star_{Q,\boldsymbol{z}},\infty\right)$ is relevant in the context of learning algorithm, from the fact that if $\mathcal{M}$ is countable the set $\mathcal{A}$ is equal $\left(-\delta^\star_{Q,\boldsymbol{z}},\infty\right)$. Furthermore, as the number of data points in datasets increases, choosing a prior $Q$ such that (32) is satisfied becomes less likely. Hence, in the paper, it will be assumed that $Q$, $\mathsf{L}_{\boldsymbol{z}}$ and $\boldsymbol{z}$ induce $\mathcal{A} = \left(-\delta^\star_{Q,\boldsymbol{z}},\infty\right)$ unless otherwise stated.

### 4.2.2 Bounds on the Radon-Nikodym Derivative

Note that from Theorem 4.1 models resulting in lower empirical risks correspond to greater values of the Radon-Nikodym derivative. The following corollary formalizes this observation.

**Lemma 4.9.** *For all $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \in (\operatorname{supp} Q)^2$, such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) \leq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)$, with $\mathsf{L}_{\boldsymbol{z}}$ in* (3)*, the Radon-Nikodym derivative in* (16) *satisfies*

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2) \leq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1), \tag{45}$$

*with equality if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) = \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)$.*

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

The Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) is always finite and strictly positive. This observation is formalized in the following lemma.

**Lemma 4.10.** *The Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in* (16) *satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$*

$$0 < \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leq \frac{\lambda}{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} < \infty, \tag{46}$$

*where the equality holds if and only if $\boldsymbol{\theta} \in \mathcal{L}^\star_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$.*

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

### 4.2.3 Asymptotes of the Radon-Nikodym Derivative

In the asymptotic regime, when the regularization factor grows to infinity, *i.e.*, $\lambda \to \infty$, the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ becomes identical to the reference measure $Q$, as described in the following lemma.

**Lemma 4.11.** *The Radon-Nikodym derivative* $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (16) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\lim_{\lambda\to\infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 1. \tag{47}$$

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

Lemma 4.11 unveils a similarity between Type-I and Type-II regularization as the Type-I measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (9) exhibits a similar behavior [28].

In the asymptotic regime, when the regularization factor decreases to zero from the right, *i.e.*, $\lambda \to 0^+$, the Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) exhibits the following behavior.

**Lemma 4.12.** *If* $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) > 0$*, with the set* $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ *in* (29)*, then the Radon-Nikodym derivative* $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (16) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q\Big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\Big)} \mathbb{1}_{\{\boldsymbol{\theta}\in\mathcal{L}^\star_{Q,\boldsymbol{z}}\}}. \tag{48}$$

*Alternatively, if* $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) = 0$*, then for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*, it holds that*

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \begin{cases} \infty & \text{if } \boldsymbol{\theta} \in \mathcal{L}^\star_{Q,\boldsymbol{z}} \\ 0 & \text{otherwise.} \end{cases} \tag{49}$$

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

Lemma 4.12 highlights that in the asymptotic regime, when the regularization factor decreases to zero from the right, *i.e.*, $\lambda \to 0^+$, the value $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$ does not depend on the exact model $\boldsymbol{\theta}$ but rather on whether $\boldsymbol{\theta} \in \operatorname{supp} Q \cap \mathcal{L}^\star_{Q,\boldsymbol{z}}$. In the case in which $\boldsymbol{\theta} \in \operatorname{supp} Q \cap \mathcal{L}^\star_{Q,\boldsymbol{z}}$, it holds that $\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) > 0$. Otherwise, $\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0$. In the special case in which $\delta^\star_{Q,\boldsymbol{z}} = 0$, with $\delta^\star_{Q,\boldsymbol{z}}$ in (28), the set $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ satisfies $\mathcal{L}^\star_{Q,\boldsymbol{z}} = \mathcal{T}(\boldsymbol{z})$, where $\mathcal{T}(\boldsymbol{z})$ is defined in (5). This implies a concentration of probability over $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$, which establishes a connection with the ERM problem without regularization in (4).

Furthermore, in the asymptotic regime, when the regularization factor decreases to zero from the right, the solutions to the Type-I and Type-II ERM-RER problems exhibit the same asymptotic behavior, as shown in [28]. This aligns with

the observation that as $\lambda$ decreases, the optimization problems in (7) and (11) exhibit a weaker relative entropy constraint. A stronger result follows from Lemma 4.12 and is presented in the following lemma.

**Lemma 4.13.** *The measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16) *and the set* $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ *in* (29) *satisfy*

$$\lim_{\lambda \to 0^+} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 1. \tag{50}$$

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

Lemma 4.13 shows that indeed when the regularization factor approaches zero from the right, the probability of the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16), in the asymptotic regime, concentrates in the set of models that induce the minimum empirical risk in $\operatorname{supp} Q$.

# 5 The Expected Empirical Risk

This section focuses on the expected empirical risk induced by the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16). Some important properties of the functional $\mathsf{R}_{\boldsymbol{z}}$ in (6) and the value $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ are introduced.

The expected empirical risk can be calculated in terms of the regularization parameter $\lambda$ and the function $\bar{K}_{Q,\boldsymbol{z}}$ defined in (23b), as shown by the following lemma.

**Lemma 5.1.** *The probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16) *satisfies*

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{51}$$

*where the functional* $\mathsf{R}_{\boldsymbol{z}}$*, the function* $\bar{K}_{Q,\boldsymbol{z}}$ *and the measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *are defined in* (6)*,* (23b)*, and* (26)*, respectively.*

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

Lemma 5.1 highlights that the function $r : (0, \infty) \to [0, \infty)$ such that $r(\lambda) = \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$, with $Q$ and $\boldsymbol{z}$ fixed, inherits all properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b). The following lemma formalizes this observation.

**Lemma 5.2.** *The expected empirical risk* $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$*, with the functional* $\mathsf{R}_{\boldsymbol{z}}$ *in* (6) *and the measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16)*, is nondecreasing with respect to* $\lambda$*. Moreover, it is strictly increasing if and only if the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *is separable with respect to the probability measure* $Q$*.*

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

The following lemma highlights a connection existing between the expected empirical risks $\mathsf{R}_{\boldsymbol{z}}(Q)$ and $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$; and the relative entropy $\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$.

**Lemma 5.3.** *The functional* $\mathsf{R}_{\boldsymbol{z}}$ *defined in* (6) *and the measures* $Q$ *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16) *satisfy*

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geq \lambda\Big(\exp\Big(\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)\Big) - 1\Big). \tag{52}$$

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

Note that $\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geq 0$, which leads to the observation that

$$\Big(\exp\Big(\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)\Big) - 1\Big) \geq 0, \tag{53}$$

in (52). Hence, from Lemma 5.3, it follows that the solution to the Type-II ERM-RER problem induces an expected empirical risk that is smaller than the one induced by reference measure $Q$. This is formalized by the following corollary.

**Corollary 5.4.** *The probability measures* $Q$ *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16) *satisfy*

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \leq \mathsf{R}_{\boldsymbol{z}}(Q), \tag{54}$$

*where the functional* $\mathsf{R}_{\boldsymbol{z}}$ *is defined in* (6) *and equality holds if and only if the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *is nonseparable.*

The following lemma presents a lower bound and an upper bound on the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ in which the regularization parameter plays a central role.

**Lemma 5.5.** *The probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16) *satisfies*

$$\delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) < \lambda + \delta^{\star}_{Q,\boldsymbol{z}}, \tag{55}$$

*where the functional* $\mathsf{R}_{\boldsymbol{z}}$ *is defined in* (6) *and* $\delta^{\star}_{Q,\boldsymbol{z}}$ *is defined in* (28)*. Moreover, equality holds if and only if the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *is nonseparable.*

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

The bounds presented in Lemma 5.5 highlight that the regularization parameter $\lambda$ in (11) governs the increase of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ with respect to its minimum, i.e, $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28). Moreover, the lower bound is tight for the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) in the asymptotic regime when $\lambda$ decreases to zero from right, as shown hereunder.

**Lemma 5.6.** *The probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16) *satisfies*

$$\lim_{\lambda\to 0^{+}} \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \delta^{\star}_{Q,\boldsymbol{z}}, \tag{56}$$

*where* $\delta^{\star}_{Q,\boldsymbol{z}}$ *is defined in* (28) *and the functional* $\mathsf{R}_{\boldsymbol{z}}$ *is defined in* (6)*.*

*Proof:* From Lemma 5.1, it holds that

$$\lim_{\lambda\to 0^+} \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lim_{\lambda\to 0^+} \lambda - \lim_{\lambda\to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda) \tag{57a}$$

$$= -\lim_{\lambda\to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda) \tag{57b}$$

$$= \delta^\star_{Q,\boldsymbol{z}}, \tag{57c}$$

where equality (57c) follows from Lemma 4.8. This completes the proof. ■

Finally, note that the functional $\mathsf{R}_{\boldsymbol{z}}$ in (6) is nonnegative. This observation together with Lemma 5.1 lead to a new property for the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b), which is stated by the following corollary

**Corollary 5.7.** *The function $\bar{K}_{Q,\boldsymbol{z}}$ in* (23b) *satisfies, for all $t \in (0,\infty)$,*

$$\bar{K}_{Q,\boldsymbol{z}}(t) \leq t. \tag{58}$$

# 6 $(\delta,\epsilon)$-Optimality

This section presents a PAC guarantee of optimality for models sampled from the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) within the setting of the Type-II ERM-RER problem in (11). Such guarantee is presented using the notion of $(\delta,\epsilon)$-optimality introduced in [28].

**Definition 6.1** ($(\delta,\epsilon)$-Optimality)**.** *Given a pair of positive reals $(\delta,\epsilon)$, with $\epsilon < 1$, the probability measure $P \in \triangle(\mathcal{M})$ is said to be $(\delta,\epsilon)$-optimal if the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ in* (27) *satisfies*

$$P(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 1-\epsilon. \tag{59}$$

The $(\delta,\epsilon)$-optimality guarantee ensures that with probability at least $1-\epsilon$, sampling models from $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) yields models that induce empirical risks not greater than $\delta$. The following theorem presents an $(\delta,\epsilon)$-optimality guarantee for the Type-II ERM-RER solution.

**Theorem 6.1.** *For all $(\delta,\epsilon) \in \left(\delta^\star_{Q,\boldsymbol{z}},\infty\right)\times(0,1)$, with $\delta^\star_{Q,\boldsymbol{z}}$ in* (28)*, there always exists a $\lambda \in (0,\infty)$, such that the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (16) *is $(\delta,\epsilon)$-optimal.*

*Proof:* The proof is presented in Appendix O. ■

# 7 Interplay Between the Relative Entropy Asymmetry and the Empirical Risk

This section presents a connection between the Type-I ERM-RER in (7) and Type-II ERM-RER problems in (11) established via a transformation of the

empirical risk function. The transformation used is termed *log-empirical risk* and is defined hereunder.

**Definition 7.1** (Log-Empirical Risk)**.** *Consider the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *and let the function* $\mathsf{V}_{\boldsymbol{z},\lambda} : \mathcal{M} \to \mathbb{R}$*, with* $Q$ *and* $\lambda$ *in* (11)*, be such that*

$$\mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta}) \triangleq \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big), \tag{60}$$

*with the function* $\bar{K}_{Q,\boldsymbol{z}}$ *defined in* (23b)*. The value* $\mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta})$ *is said to be the log-empirical risk induced by the model* $\boldsymbol{\theta} \in \mathcal{M}$*.*

The notion of log-empirical risk in (60) leads to the *expected log-empirical risk*, which is denoted as follows.

**Definition 7.2** (Expected Log-Empirical Risk)**.** *Consider the log-empirical risk function* $\mathsf{V}_{\boldsymbol{z},\lambda}$ *in* (60) *and let the functional* $\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda} : \triangle(\mathcal{M}) \to \mathbb{R}$ *be such that for all probability measures* $P \in \triangle(\mathcal{M})$*, it holds that*

$$\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(P) \triangleq \int \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}). \tag{61}$$

*The value* $\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(P)$ *is the expected log-empirical risk induced by the measure* $P$*.*

## 7.1 The Connection between Type-I and Type-II

Using the objects defined above, a new Type-I ERM-RER problem is presented:

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(P) + \mathsf{D}(P\|Q), \tag{62}$$

with $\lambda$ and $Q$ being problems of the Type-I and Type-II ERM-RER problems in (7) and (11). The main result of this section is presented in the following theorem.

**Theorem 7.1.** *The solution to the optimization problem in* (62) *is unique and identical to the probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (16)*.*

*Proof:* Denote by $\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the solution to the optimization problem in (62). From Lemma 3.1, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp(-\mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta}))}{\int \exp(-\mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\nu}))\,\mathrm{d}Q(\boldsymbol{\nu})} \tag{63a}$$

$$= \frac{\exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big)}{\int \exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big)\mathrm{d}Q(\boldsymbol{\nu})} \tag{63b}$$

$$= \frac{\Big(\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\mathrm{d}Q(\boldsymbol{\nu})\Big)^{-1}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{63c}$$

$$= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{63d}$$

$$= \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{63e}$$

where the equality in (63b) follows from the definition of log-empirical risk in (60); the equality in (63d) follows from (23b) and (15b); and the equality in (63e) follows from Theorem 4.1, which completes the proof. ■

Theorem 7.1 establishes an equivalence between the regularization of Type-I and Type-II. More specifically, Theorem 7.1 highlights that by transforming the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (60) into the log-empirical risk $\mathsf{V}_{\boldsymbol{z},\lambda}$ in (60), the Type-II ERM-RER problem in (11) can be solved by solving the Type-I ERM-RER problem in (62). In view of this, it is interesting to note that Type-I regularization forces the support of the solution to be included into the support of the reference measure as a consequence of the optimization domain $\triangle_Q(\mathcal{M})$. Hence, $\operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \subseteq \operatorname{supp} Q$. Moreover, as a consequence of the solution, all the models in the support of the reference measure are in the support of the solution. That is, $\operatorname{supp} Q \subseteq \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Alternatively, Type-II regularization forces the support of the solution to be included into the support of the reference measure as a consequence of the optimization domain $\bigtriangledown_Q(\mathcal{M})$. Hence, $\operatorname{supp} Q \subseteq \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Moreover, as a consequence of the solution, all the models in the support of the solution are in the support of the reference measure. That is, $\operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \subseteq \operatorname{supp} Q$. Hence, $\operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} = \operatorname{supp} Q$, regardless of whether the regularization is of Type-I or Type-II.

Finally, the Type-I - Type-II relation can be used to establish an equality involving the relative entropies $\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ and $\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)$; and the expected log-empirical risks $\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ and $\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(Q)$, as shown hereunder.

**Lemma 7.1.** *The functional* $\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}$ *in* (61) *and the probability measures* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *and* $Q$ *in* (16) *satisfy*

$$\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) = \log(\lambda), \tag{64}$$

*and*

$$\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(Q) - \mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \log(\lambda). \tag{65}$$

*Proof:* The proof is presented in Appendix P. ■

## 7.2 Sensitivity of the Log-Empirical Risk

The sensitivity of the expected empirical risk, as presented in [28, Definition 10.1], is defined as follows.

**Definition 7.3** (Sensitivity of the Expected Empirical Risk)**.** *Consider the functional* $\mathsf{R}_{\boldsymbol{z}}$ *defined in* (6) *and let* $\mathsf{S}_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \triangle_Q(\mathcal{M}) \to \mathbb{R}$ *be a functional such that*

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big), \tag{66}$$

*where the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *is in* (9)*. The sensitivity of the expected empirical risk due to a deviation from* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *to* $P$ *is* $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$*.*

Following the same idea, the sensitivity of the expected log-empirical risk is defined as follows.

**Definition 7.4** (Sensitivity of the Expected Log-Empirical Risk)**.** *Consider the functional* $\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}$ *in* (61) *and let* $\bar{\mathsf{S}}_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \bigtriangledown_Q(\mathcal{M}) \to \mathbb{R}$ *be a functional such that*

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P) = \bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(P) - \bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big), \tag{67}$$

*where the probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *is in* (16)*. The sensitivity of the expected log-empirical risk due to a deviation from* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *to* $P$ *is* $\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P)$*.*

The sensitivity of the expected log-empirical risk due to a deviation from $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to $P$ exhibits the following closed-form expression.

**Lemma 7.2.** *The sensitivity* $\bar{\mathsf{S}}_{Q,\lambda}$ *in* (67) *satisfies for all probability measures* $P \in \bigcirc_Q(\mathcal{M})$ *that*

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P) = \mathsf{D}\Big(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \mathsf{D}\left(P\|Q\right) + \mathsf{D}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big), \tag{68}$$

*where the probability measures* $Q$ *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *are defined in* (16)*.*

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

An interesting interpretation of Lemma 7.2 follows from rewriting (68) using the objective function of the Type-I ERM-RER problem in (62) as follows:

$$\begin{aligned}\mathsf{D}\Big(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(P) + \mathsf{D}\left(P\|Q\right) - \bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \\ - \mathsf{D}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big).\end{aligned} \tag{69}$$

That is, the relative entropy $\mathsf{D}\Big(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ represents the variation of the objective function of the Type-I ERM-RER problem in (62) due to a deviation from the solution $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to an alternative probability measure $P$.

In Lemma 7.2, when $P$ is chosen to be identical to the reference measure $Q$, it follows that

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, Q) = \mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) + \mathsf{D}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big), \tag{70}$$

where the right-hand side is a Jeffrey's divergence, also known as the symmetrized Kullback-Liebler divergence, between the measures $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $Q$. Furthermore, by observing that $\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) \geq 0$, and $\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq 0$, Lemma 7.2 leads to the following corollary.

**Corollary 7.3.** *The probability measures $Q$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (16) *satisfy*

$$\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leq \bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}(Q), \tag{71}$$

*where the functional $\bar{\mathsf{R}}_{\boldsymbol{z},Q,\lambda}$ is defined in* (61).

## 7.3 Type-I and Type-II Optimal Measures

The probability measures $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (9) and in (16), respectively, exhibit the following property.

**Lemma 7.4.** *The probability measures $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (9) *and in* (16)*, respectively, satisfy*

$$\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) = \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{72}$$

*where the function $K_{Q,\boldsymbol{z}}$ is defined in* (10).

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

Finally, two important properties of the Type-I and Type-II optimal measures are presented. The first one quantifies the variation of the expected log-empirical risk due to a deviation from the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (9) to the probability measure $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) via the sensitivity $\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$. The second result quantifies the variation of the expected empirical risk due to a deviation from the probability measure $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (9) via the sensitivity $\mathsf{S}_{Q,\alpha}\left(\boldsymbol{z}, \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$.

**Corollary 7.5.** *The probability measures $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (9) *and in* (16)*, respectively, satisfy*

$$\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \mathsf{D}\left(P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \left(\log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) \tag{73}$$

*and*

$$\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{74}$$

*where the functionals $\mathsf{S}_{Q,\lambda}$ and $\bar{\mathsf{S}}_{Q,\alpha}$ are respectively defined in* (66) *and in* (67)*; and the function $K_{Q,\boldsymbol{z}}$ is defined in* (10).

# 8 Final Remarks

This work has introduced the Type-II ERM-RER problem and has presented its solution through Theorem 4.1. The solution highlights that regardless of the direction in which relative entropy is used as a regularizer, the models that are considered by the solution are necessarily in the support of the reference measure. In that sense, the restriction over the models introduced by the reference measure cannot be bypassed by the training data when relative entropy is used as the regularizer. This limitation has been shown to be a consequence of the equivalence that can be established between Type-I and Type-II regularization. Remarkably, the direction of the relative entropy regularizer can be switched by a logarithmic transformation of the risk. The mutual absolute continuity of both Type-I and Type-II ERM-RER solutions relative to the reference measure can be understood in light of the equivalence between both types of regularization. The analytical results have also enabled us to provide an operationally meaningful characterization of the expected empirical risk induced by the Type-II solution in terms of the regularization parameters. This, in turn, reduces the computational burden of bounding the expected empirical risk. Moreover, the insight provided by the bounds on the expected empirical risk can be distilled into guidelines for selecting the regularization parameter.

# A Proof of Lemma 4.1

*Proof:* The optimization problem in (11) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure $P$ with respect to the measure $Q$, which yields:

$$\min_{P\in\bigcirc_Q(\mathcal{M})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) + \lambda\int \frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{\theta})\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{\theta})\right)\ \mathrm{d}P(\boldsymbol{\theta}), \tag{75a}$$

$$\text{s.t.} \quad \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{75b}$$

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

$$\int |\hat{g}(\boldsymbol{\theta})|\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{76}$$

Hence, the optimization problem of interest is:

$$\min_{g\in\mathscr{M}} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - \lambda\int \log(g(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{77a}$$

$$\text{s.t.} \int g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{77b}$$

Let the Lagrangian of the optimization problem in (77) be $L : \mathscr{M} \times \mathbb{R} \to \mathbb{R}$ such that

$$L(g,\beta) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - \lambda\int \log(g(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta}) + \beta\left(\int g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - 1\right) \tag{78}$$

$$= \int \Big(g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta) - \lambda\log(g(\boldsymbol{\theta}))\Big)\ \mathrm{d}Q(\boldsymbol{\theta}) - \ \beta, \tag{79}$$

where $\beta$ is a real that acts as a Lagrange multiplier due to the constraint (77b). Let $\hat{g} : \mathcal{M} \to \mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional $L$ in (78) at $(g, \beta) \in \mathscr{M} \times \mathbb{R}$ in the direction of $\hat{g}$, if it exists, is

$$\partial L(g,\beta;\hat{g}) \triangleq \frac{\mathrm{d}}{\mathrm{d}\gamma}L(g + \gamma\hat{g}, \beta)\bigg|_{\gamma=0}. \tag{80}$$

The proof continues under the assumption that the function $g$ and $\hat{g}$ are such that the Gateaux differential in (80) exists. Under such an assumption, let

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

$$\begin{aligned} r(\gamma) &= L(g + \gamma\hat{g}, \beta) && (81)\\ &= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \\ &\quad + \beta\left(\int (g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) - 1\right) && (82)\\ &= -\beta + \int g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\, \mathrm{d}Q(\boldsymbol{\theta}) + \gamma \int \hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\, \mathrm{d}Q(\boldsymbol{\theta}) \\ &\quad + \lambda \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}), && (83) \end{aligned}$$

where the last equality is simply an algebraic re-arrangement of terms. From the assumption that the function $g$ and $\hat{g}$ are such that the Gateauz differential in (80) exists, it follows that the function $r$ in (81) is differentiable at zero. Note that the first two terms in (83) are independent of $\gamma$; the third term is linear with $\gamma$; 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 $\epsilon$ arbitrarily small, satisfies

$$\hat{r}(\gamma) = -\lambda \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}). \tag{84}$$

Under the same assumption, it follows that the function $\hat{r}$ in (84) is differentiable at zero. That is, the limit

$$\lim_{\delta \to 0} \frac{1}{\delta}(\hat{r}(\gamma + \delta) - \hat{r}(\gamma)) \tag{85}$$

exists for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon$ arbitrarily small. The proof of the existence of such a limit relies on the fact that log is strictly concave and continuous. This implies that $-\log$ 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

$$|\log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})) - \log(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta}))| \leq c\, |\hat{g}(\boldsymbol{\theta})\delta|, \tag{86}$$

with $\delta > 0$, for some constant $c$ positive and finite, which implies that

$$\left| \frac{\log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})) - \log(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta}))}{\delta} \right| \leq c\, |\hat{g}(\boldsymbol{\theta})|. \tag{87}$$

Using these arguments, the limit in (85) satisfies for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, that

$$\begin{aligned} \lim_{\delta \to 0} \frac{1}{\delta}(\hat{r}(\gamma + \delta) - \hat{r}(\gamma)) = \lambda \lim_{\delta \to 0} \frac{1}{\delta} \Bigg( & \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \\ & - \int \log(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \Bigg) \end{aligned} \tag{88}$$

$$
\begin{aligned}
&= \lambda \lim_{\delta \to 0} \int \frac{\log(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))}{\delta} \\
&\quad - \frac{\log(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta}))}{\delta} \, \mathrm{d}Q(\boldsymbol{\theta}) && (89)\\
&= \lambda \int \frac{\mathrm{d}}{\mathrm{d}\gamma} \log(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) && (90)\\
&< \infty, && (91)
\end{aligned}
$$

where both the equality in (90) and the inequality in (91) follow from noticing that the conditions for the dominated convergence theorem hold [45, Theorem 1.6.9], namely:

- For all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$, the inequality in (87) holds;
- The function $\hat{g}$ in (87) satisfies the inequality in (76); and
- For all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, it holds that

$$
\begin{aligned}
&\lim_{\delta \to 0} \frac{\log(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - \log(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))}{\delta} \\
&\quad = \frac{\mathrm{d}}{\mathrm{d}\gamma} \log(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) && (92)\\
&\quad = \frac{\hat{g}(\boldsymbol{\theta})}{(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))}, && (93)
\end{aligned}
$$

  which follows from the fact that log is differentiable.

Hence, the derivative of the real function $r$ in (83) is

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) &= \int \hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta) \, \mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \frac{\hat{g}(\boldsymbol{\theta})}{(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))} \, \mathrm{d}Q(\boldsymbol{\theta}) && (94)\\
&= \int \hat{g}(\boldsymbol{\theta}) \left( \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))} \right) \mathrm{d}Q(\boldsymbol{\theta}). && (95)
\end{aligned}
$$

From (80) and (95), it follows that

$$
\partial L(g, \beta; \hat{g}) = \int \hat{g}(\boldsymbol{\theta}) \left( \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{g(\boldsymbol{\theta})} \right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{96}
$$

The relevance of the Gateaux differential in (96) stems from [46, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional $L$ in (78) 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 $\hat{g} \in \mathscr{M}$,

$$
\partial L \left( \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta; \hat{g} \right) = 0. \tag{97}
$$

From (96) and (97), it follows that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ must satisfy for all functions $\hat{g}$ in $\mathscr{M}$ that

$$\int \hat{g}(\boldsymbol{\theta}) \left( \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \lambda \left( \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \right) \mathrm{d}Q(\boldsymbol{\theta}) = 0. \tag{98}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \lambda \left( \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} = 0, \tag{99}$$

and thus,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{100}$$

where $\beta$ is chosen to satisfy (77b) and guarantee that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \in (0, \infty)$. That is,

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \frac{\lambda}{t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \right\}, \text{ and} \tag{101}$$

$$1 = \int \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta}). \tag{102}$$

which is an assumption of the theorem.

The proof continues by verifying that the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that satisfies (100) is the unique solution to the optimization problem in (75). Such verification is done by showing that the objective function in (75) is strictly convex with the optimization variable. Let $P_1$ and $P_2$ be two different probability measures in $(\mathcal{M}, \mathscr{F})$ and let $\alpha$ be in $(0, 1)$. Hence,

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}}(\alpha P_1 + (1-\alpha)P_2) + \lambda \mathsf{D}(\alpha P_1 + (1-\alpha)P_2 \| Q) && (103)\\
&= \mathsf{R}_{\boldsymbol{z}}(\alpha P_1) + \mathsf{R}_{\boldsymbol{z}}((1-\alpha)P_2) + \lambda \mathsf{D}(\alpha P_1 + (1-\alpha)P_2 \| Q) && (104)\\
&> \mathsf{R}_{\boldsymbol{z}}(\alpha P_1) + \mathsf{R}_{\boldsymbol{z}}((1-\alpha)P_2) + \lambda \mathsf{D}(\alpha P_1 \| Q) + \lambda \mathsf{D}((1-\alpha)P_2 \| Q) && (105)\\
& && (106)
\end{aligned}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (6). The equality above follows from the properties of the Lebesgue integral, while the inequality follows from [28, Theorem 2.2., page ]. This proves that the solution is unique due to the strict concavity of the objective function, which completes the proof. ■

# B Proof of Lemma 4.2

*Proof:* Given a probability measure $V \in \bigtriangledown_Q(\mathcal{M})$, with $\bigtriangledown_Q(\mathcal{M})$ in (12), let $V_0$ and $V_1$ be two probability measures on the measurable space $(\mathcal{M}, \mathscr{F})$ such that for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$V_0(\mathcal{A}) = \frac{V(\mathcal{A} \setminus \operatorname{supp} Q)}{V(\mathcal{M} \setminus \operatorname{supp} Q)}, \tag{107a}$$

and

$$V_1(\mathcal{A}) = \frac{V(\mathcal{A} \cap \operatorname{supp} Q)}{V(\mathcal{M} \cap \operatorname{supp} Q)}. \tag{107b}$$

Let the real value $\alpha$ be

$$\alpha = V(\mathcal{M} \cap \operatorname{supp} Q). \tag{108}$$

Hence, for all $\mathcal{A} \in \mathscr{F}$ the measure $V$ satisfies that

$$V(\mathcal{A}) = (1-\alpha)V_0(\mathcal{A}) + \alpha V_1(\mathcal{A}). \tag{109}$$

Moreover, from (109) it holds that: *i*) If $V(\mathcal{A}) = 0$, then $V_0(\mathcal{A}) = 0$, which implies that $V_0$ is absolutely continuous with respect to $V$; *ii*) If $V(\mathcal{A}) = 0$, then $V_1(\mathcal{A}) = 0$, which implies that $V_1$ is absolutely continuous with respect to $V$. Furthermore, from the definition of $\bigtriangledown_Q(\mathcal{M})$ in (12) the probability measure $V$ is absolutely continuous with respect to $Q$. Hence, for all $\mathcal{A} \in \mathscr{F}$, it follows that

$$\begin{aligned}
Q(\mathcal{A}) &= \int_{\mathcal{A}} \mathrm{d}Q(\boldsymbol{\theta}) && (110)\\
&= \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V(\boldsymbol{\theta}) && (111)\\
&= \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}((1-\alpha)V_0 + \alpha V_1)(\boldsymbol{\theta}) && (112)\\
&= (1-\alpha)\int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V_0(\boldsymbol{\theta}) + \alpha \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V_1(\boldsymbol{\theta}) && (113)\\
&= \int_{\mathcal{A}} \alpha \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V_1(\boldsymbol{\theta}). && (114)
\end{aligned}$$

Hence, from (114) and the Radon-Nikodym Theorem in [45, Theorem 2.2.1, page 65] the probability measure $Q$ is absolutely continuous with respect to $V_1$. This implies that for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$Q(\mathcal{A}) = \int \frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\, \mathrm{d}V_1(\boldsymbol{\theta}), \tag{115}$$

where, for all $\boldsymbol{\theta} \in \operatorname{supp} V$,

$$\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta}) = \alpha \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta}). \tag{116}$$

From (116), the following holds:

$$\begin{aligned}
\mathsf{D}(Q\|V) &= \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(117a)}\\
&= \int \log\left(\frac{1}{\alpha}\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(117b)}\\
&= \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\\
&\quad - \int \log(\alpha)\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(117c)}\\
&= \mathsf{D}(Q\|V_1) - \log(\alpha). && \text{(117d)}
\end{aligned}$$

From (117) it follows that

$$\begin{aligned}
\mathsf{R}_{\boldsymbol{z}}(V) + \lambda\mathsf{D}(Q\|V) &= \mathsf{R}_{\boldsymbol{z}}((1-\alpha)V_0 + \alpha V_1) + \lambda\mathsf{D}(Q\|V_1) - \lambda\log(\alpha) && \text{(118a)}\\
&= (1-\alpha)\mathsf{R}_{\boldsymbol{z}}(V_0) + \alpha\mathsf{R}_{\boldsymbol{z}}(V_1) + \lambda\mathsf{D}(Q\|V_1)\\
&\quad - \lambda\log(\alpha) && \text{(118b)}\\
&\geq \alpha\mathsf{R}_{\boldsymbol{z}}(V_1) + \lambda\mathsf{D}(Q\|V_1) - \lambda\log(\alpha), && \text{(118c)}
\end{aligned}$$

with equality if and only if $\alpha = 1$, which implies that for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$\begin{aligned}
V(\mathcal{A}) &= V_1(\mathcal{A}) && \text{(119a)}\\
&= V(\mathcal{A} \cap \operatorname{supp} Q), && \text{(119b)}
\end{aligned}$$

where the equality in (119b) follows from (107b). This implies that the equality in (118c) holds if and only if

$$\operatorname{supp} Q = \operatorname{supp} V, \tag{120}$$

which implies that the equality in (118c) holds if and only if the measure $V$ is mutually absolutely continuous with respect to the reference measure $Q$. Finally, the above leads to

$$\min_{P \in \bigtriangledown_Q(\mathcal{M}) \setminus \bigcirc_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda\mathsf{D}(Q\|P) > \min_{P \in \bigtriangledown_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda\mathsf{D}(Q\|P), \tag{121}$$

which completes the proof. ■

# C Proof of Lemma 4.4

*Proof:* The properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b), for which an explicit expression is unknown, are proven by studying the functional inverse and the continuous inverse theorem [47, Theorem 5.6]. Hence, the proof is divided into two parts. The first part introduces the functional inverse of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) and properties. The second part proves the continuity of the function

$\bar{K}_{Q,\boldsymbol{z}}$ in (23b) is proved using its functional inverse via the continuous inverse theorem [47, Theorem 5.6]

The first part is as follows. For $\bar{K}_{Q,\boldsymbol{z}}$ defined in (23), assume that $t \in (0,\infty)$ and $\gamma \in \mathcal{A}$, with $\mathcal{A}$ defined in (23a) satisfy that

$$\bar{K}_{Q,\boldsymbol{z}}(t) = \gamma, \tag{122}$$

which implies that

$$1 = \int \frac{\mathrm{d}\bar{P}^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{123a}$$

$$= \int \frac{t}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma}\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{123b}$$

Let the function $\bar{K}^{-1}_{Q,\boldsymbol{z}} : \mathcal{A} \to (0,\infty)$ be the functional inverse of $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) given by

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma) = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\gamma}\,\mathrm{d}Q(\boldsymbol{\theta})}, \tag{124}$$

which follows from the constraint in (123) and the equality in (122).

From (123) and the fact that $t \in (0,\infty)$, it holds that

$$0 < \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma}\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{125}$$

which implies that for all $\gamma \in \mathcal{A}$, with $\mathcal{A}$ in (23a), the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (124) satisfies

$$\infty > \bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma) > 0. \tag{126}$$

Note that from (125), for all $(\gamma_1,\gamma_2) \in \mathcal{A}^2$, such that $\gamma_1 < \gamma_2$, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ it holds that

$$\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_1} > \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_2}, \tag{127}$$

which implies that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_1}\,\mathrm{d}Q(\boldsymbol{\theta}) > \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_2}\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{128}$$

Therefore, from equality (124) and inequality (128) for all $(\gamma_1,\gamma_2) \in \mathcal{A}^2$, such that $\gamma_1 < \gamma_2$, the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (124) satisfies

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma_1) < \bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma_2), \tag{129}$$

which implies that the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (124) is strictly increasing. Furthermore, from inequality (129) and the fact that $\bar{P}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,t)}$ is unique (see Theorem 4.1), it follows that $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ is bijective, which completes the proof of the first part.

The second part is as follows. The continuity of the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (124) is shown by considering an arbitrary $\epsilon > 0$ and a pair $(\gamma_1, \gamma_2) \in \mathcal{A}^2$ under the condition

$$\epsilon > \left|\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_2) - \bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_1)\right|. \tag{130}$$

From the fact that $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ is bijective and strictly increasing, let the pair $\gamma_1$, and $\gamma_2$ satisfy $\gamma_1 < \gamma_2$, without loss of generality. Thus, it holds that

$$\left|\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_2) - \bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_1)\right| = \bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_2) - \bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_1). \tag{131}$$

Substituting equality (131) into (130) yields

$$\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_1) + \epsilon > \bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_2). \tag{132}$$

From the fact that $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ is bijective, evaluating inequality (132) with $\bar{K}_{Q,\boldsymbol{z}}$ in (23a) leads to

$$\begin{aligned} \bar{K}_{Q,\boldsymbol{z}}\Big(\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_1) + \epsilon\Big) &> \bar{K}_{Q,\boldsymbol{z}}\Big(\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_2)\Big) && \text{(133a)} \\ &= \gamma_2. && \text{(133b)} \end{aligned}$$

Subtracting $\gamma_1$ from both sides in (133), results in

$$\begin{aligned} \bar{K}_{Q,\boldsymbol{z}}\Big(\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_1) + \epsilon\Big) - \gamma_1 &> \gamma_2 - \gamma_1 && \text{(134a)} \\ &= |\gamma_2 - \gamma_1|, && \text{(134b)} \end{aligned}$$

where equality (134b) follows from the condition that $\gamma_2 > \gamma_1$. Thus, from (134) it follows that for all $\epsilon > 0$, there exist a $\delta > 0$ that satisfies

$$\delta = \bar{K}_{Q,\boldsymbol{z}}\Big(\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma_1) + \epsilon\Big) - \gamma_1, \tag{135}$$

which implies the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (124) is continuous for all $\gamma \in \mathcal{A}$. From (135) the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ is continuous and strictly increasing. Hence, from the continuous inverse theorem in [47, Theorem 5.6], the inverse of the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (124), which is given by $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is continuous and strictly increasing, which completes the proof. ■

# D Proof of Lemma 4.5

*Proof:* The proof shows by exhaustion that the codomain of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is an interval of $\mathbb{R}$ by evaluating $\beta$ in (25). Three cases are considered: *i)* $\beta < -\delta_{Q,\boldsymbol{z}}^{\star}$; *ii)* $\beta > -\delta_{Q,\boldsymbol{z}}^{\star}$; and *iii)* $\beta = -\delta_{Q,\boldsymbol{z}}^{\star}$, with $\delta_{Q,\boldsymbol{z}}^{\star}$ in (28).

In the first case, from the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (28), it follows from (15b) that if $\beta < -\delta^\star_{Q,\boldsymbol{z}}$, then for all $\boldsymbol{\theta} \in \{\boldsymbol{\theta} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < -\beta\}$, it holds that

$$\frac{\mathrm{d}\bar{P}^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \tag{136a}$$

$$< 0, \tag{136b}$$

which contradicts the fact that the function $\frac{\mathrm{d}\bar{P}^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) is nonnegative [45, Corollary 2.2.2]. The above implies that for all $t \in (0,\infty)$, the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) satisfies that $\bar{K}_{Q,\boldsymbol{z}}(t) \notin \big(-\infty, -\delta^\star_{Q,\boldsymbol{z}}\big)$.

In the second case, if $\beta > -\delta^\star_{Q,\boldsymbol{z}}$, it holds that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\,\mathrm{d}Q(\boldsymbol{\theta}) \leq \int \frac{1}{\delta^\star_{Q,\boldsymbol{z}} + \beta}\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{137a}$$

$$= \frac{1}{\delta^\star_{Q,\boldsymbol{z}} + \beta}. \tag{137b}$$

Moreover, from (137) for all $\beta > -\delta^\star_{Q,\boldsymbol{z}}$, there exists a $\lambda \in (0,\infty)$ such that the constraint in (24) holds, which implies $\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) \in (0,\infty)$.

Finally, under the assumption that $\beta = -\delta^\star_{Q,\boldsymbol{z}}$ and $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ defined in (29), two cases are considered: (*a*) $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) > 0$; and (*b*) $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) = 0$. In case (*a*), if $\beta = -\delta^\star_{Q,\boldsymbol{z}}$ and $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) > 0$, then the integral in the denominator of (30) is undefined, contradicting (24). In the alternative case (*b*), if $\beta = -\delta^\star_{Q,\boldsymbol{z}}$ and $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) = 0$, then the integral in (30) is either

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}}\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{138a}$$

which implies that $-\delta^\star_{Q,\boldsymbol{z}} \in \mathcal{A}$, with $\mathcal{A}$ defined in (23a), or the integral is

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}}\,\mathrm{d}Q(\boldsymbol{\theta}) = \infty, \tag{138b}$$

which implies that, $-\delta^\star_{Q,\boldsymbol{z}} \notin \mathcal{A}$. Note that the integral is always positive from the fact that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > \delta^\star_{Q,\boldsymbol{z}}$. Moreover, the integral in (138) is never zero as it would be a contradiction with (2). Hence, from (136), (137) and (138) the set $\mathcal{A}$ in (23a) is either the open set $\big(-\delta^\star_{Q,\boldsymbol{z}}, \infty\big)$ or the closed set $\big[-\delta^\star_{Q,\boldsymbol{z}}, \infty\big)$, which completes the proof.

■

# E Proof of Lemma 4.8

*Proof:* From (15b), for all $\lambda \in (0,\infty)$, it follows that

$$\lambda = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \beta}\,\mathrm{d}Q(\boldsymbol{\nu})} \tag{139a}$$

$$= \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu})}, \tag{139b}$$

where equality (139a) follows from (15b), and equality (139b) follows from the nominal function in (23b). From the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), for all $\boldsymbol{\nu} \in \operatorname{supp} Q$, the empirical risk satisfies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \geq \delta^{\star}_{Q,\boldsymbol{z}}. \tag{140}$$

Taking the limit of (139) as $\lambda$ approaches zero from the right satisfies

$$\lim_{\lambda \to 0^+} \lambda = \lim_{\lambda \to 0^+} \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu})} \tag{141a}$$

$$\geq \lim_{\lambda \to 0^+} \frac{1}{\int \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}}+\bar{K}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu})} \tag{141b}$$

$$= \lim_{\lambda \to 0^+} \frac{1}{\frac{1}{\delta^{\star}_{Q,\boldsymbol{z}}+\bar{K}(\lambda)}} \tag{141c}$$

$$= \lim_{\lambda \to 0^+} \left(\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}(\lambda)\right) \tag{141d}$$

$$= \delta^{\star}_{Q,\boldsymbol{z}} + \lim_{\lambda \to 0^+} \bar{K}(\lambda), \tag{141e}$$

where inequality (141b) follows from (140). Hence, from (141) it holds that

$$0 \geq \delta^{\star}_{Q,\boldsymbol{z}} + \lim_{\lambda \to 0^+} \bar{K}(\lambda). \tag{142}$$

From (15a) and Lemma 4.4, for all $\lambda \in (0, \infty)$, it holds that

$$0 \leq \delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{143}$$

Taking the limit when $\lambda$ approaches zero from the right in (143) yields

$$0 \leq \delta^{\star}_{Q,\boldsymbol{z}} + \lim_{\lambda \to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{144}$$

Then, from (142) and (144) it follows that

$$\lim_{\lambda \to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\delta^{\star}_{Q,\boldsymbol{z}}, \tag{145}$$

which completes the proof. ■

# F Proof of Lemma 4.9

*Proof:* For all $\boldsymbol{\theta}_1 \in \operatorname{supp} Q$ and for all $\boldsymbol{\theta}_2 \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, it follows that

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) \geq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2), \tag{146}$$

and thus, for all $\lambda \in (0, \infty)$, it holds that

$$\frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)} \leq \frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)}, \tag{147a}$$

which implies

$$\frac{\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)\right)^{-1}}{\int\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)^{-1}\mathrm{d}Q(\boldsymbol{\nu})}\leq\frac{\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)\right)^{-1}}{\int\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)^{-1}\mathrm{d}Q(\boldsymbol{\nu})}. \tag{148}$$

Hence, under the assumption that $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\cap\operatorname{supp}Q\neq\emptyset$, for all $\boldsymbol{\theta}_1\in\operatorname{supp}Q$ and for all $\boldsymbol{\theta}_2\in\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\cap\operatorname{supp}Q$, it holds that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1)\leq\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2), \tag{149}$$

with equality if and only if $\boldsymbol{\theta}_1\in\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\cap\operatorname{supp}Q$, which completes the proof. ■

# G Proof of Lemma 4.10

*Proof:* From Lemma 4.9, it follows that for all $\lambda\in(0,\infty)$, for all $\boldsymbol{\theta}\in\operatorname{supp}Q$, and for all $\boldsymbol{\phi}\in\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\cap\operatorname{supp}Q$, it holds that

$$\begin{aligned}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})&\leq\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\phi}) &\text{(150a)}\\ &=\frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} &\text{(150b)}\\ &=\frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}}+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} &\text{(150c)}\\ &<\infty, &\text{(150d)}\end{aligned}$$

where equality (150a) follows from (16); equality (150b) follows from the fact that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi})\geq\delta^{\star}_{Q,\boldsymbol{z}}$; and the equality in (150d) follows from the fact that for all $\lambda>0$, the function $\bar{K}_{Q,\boldsymbol{z}}(\lambda)<\infty$. From the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28) and $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (29) equality in (150a) holds if and only if $\boldsymbol{\theta}\in\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\cap\operatorname{supp}Q$. This completes the proof of finiteness.

For the proof of positivity, observe that from Lemma 4.4 for all $\lambda\in(0,\infty)$, it holds that

$$-\delta^{\star}_{Q,\boldsymbol{z}}<\bar{K}_{Q,\boldsymbol{z}}(\lambda)<\infty, \tag{151}$$

which implies

$$0<\delta^{\star}_{Q,\boldsymbol{z}}+\bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{152}$$

From (151), it follows that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \tag{153a}$$

$$= \frac{\frac{1}{\int \frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})}}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \tag{153b}$$

$$> 0, \tag{153c}$$

which completes the proof. ∎

# H Proof of Lemma 4.11

*Proof:* From Theorem 4.1, the Radon-Nikodym derivative of the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to $Q$, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, satisfies that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \tag{154a}$$

$$= \frac{\int \frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \tag{154b}$$

$$= \left(\frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\int \left(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)^{-1}\mathrm{d}Q(\boldsymbol{\nu})}\right)^{-1} \tag{154c}$$

$$= \frac{\exp\left(-\log\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)\right)}{\int \exp\left(\log\left(\frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\right)\right)\mathrm{d}Q(\boldsymbol{\nu})} \tag{154d}$$

From Lemma 4.4, the case in which $\lambda \to \infty$ implies $\bar{K}_{Q,\boldsymbol{z}}(\lambda) \to \infty$, and from (154), it follows that

$$\lim_{\lambda\to\infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \lim_{\lambda\to\infty} \frac{\exp\left(-\log\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)\right)}{\int \exp\left(\log\left(\frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\right)\right)\mathrm{d}Q(\boldsymbol{\nu})} \tag{155}$$

$$= \frac{1}{\int \mathrm{d}Q(\boldsymbol{\nu})} \tag{156}$$

$$= 1, \tag{157}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3). This completes the proof. ∎

# I Proof of Lemma 4.12

*Proof:* Assume that $t \in (0,\infty)$ and $\gamma \in \mathbb{R}$ satisfy that

$$1 = \int \frac{t}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\gamma}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{158}$$

which implies

$$\bar{K}_{Q,\boldsymbol{z}}(t) = \gamma. \tag{159}$$

Let the function $\bar{K}_{Q,\boldsymbol{z}}^{-1} : \mathbb{R} \to (0,\infty)$ be the functional inverse of $\bar{K}_{Q,\boldsymbol{z}}$ in (159). Thus,

$$\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma) = t, \tag{160}$$

with $\gamma$ and $t$ in (158). From (158), the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (160) satisfies

$$\bar{K}_{Q,\boldsymbol{z}}^{-1}(\gamma) = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\gamma}\,\mathrm{d}Q(\boldsymbol{\theta})}. \tag{161}$$

From Theorem 4.1, for all $\lambda \in (0,\infty)$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{162a}$$

$$= \frac{\bar{K}_{Q,\boldsymbol{z}}^{-1}\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda)\big)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{162b}$$

$$= \frac{\frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{162c}$$

$$= \left(\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1}. \tag{162d}$$

Given $\boldsymbol{\theta} \in \operatorname{supp} Q$, consider the partition of the set $\mathcal{M}$ formed by the sets $\mathcal{A}_0$, $\mathcal{A}_1$, and $\mathcal{A}_2$, which satisfy the following:

$$\mathcal{A}_0 = \big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\big\}, \tag{163a}$$

$$\mathcal{A}_1 = \big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}}\big\}, \tag{163b}$$

$$\mathcal{A}_2 = \big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^{\star}_{Q,\boldsymbol{z}}\big\}, \tag{163c}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ defined in (28). Using the sets $\mathcal{A}_0$, $\mathcal{A}_1$, and $\mathcal{A}_2$ in (162), for all $\lambda \in (0,\infty)$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ the following holds

$$\begin{aligned}
&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \left(\int_{\mathcal{A}_0} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu}) + \int_{\mathcal{A}_1} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right. \\
&\quad \left. + \int_{\mathcal{A}_2} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(164a)} \\
&= \left(\int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu}) + \int_{\mathcal{A}_1} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(164b)} \\
&= \left(\int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu}) + \int_{\mathcal{A}_1} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(164c)}
\end{aligned}$$

where equality (164b) follows from the fact that the $\mathcal{A}_2$ has measure zero with respect to $Q$ and that the set $\mathcal{A}_0 = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (29); and equality

(164b) follows from the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (28). Following this observation, the rest of the proof is divided into three parts. The first part evaluates $\lim_{\lambda\to\infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\right\}$. The second part considers the case in which $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star_{Q,\boldsymbol{z}}\right\}$. The third part considers the remaining case.

The first part is as follows. Consider that $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\right\}$ and note that $\mathcal{A}_0 = \mathcal{L}^\star_{Q,\boldsymbol{z}}$. From (164), for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\right\}$ it holds that

$$\begin{aligned}
\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \left(\int_{\mathcal{L}^\star_{Q,\boldsymbol{z}}} \frac{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right. \\
&\qquad \left. + \int_{\mathcal{A}_1} \frac{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(165a)} \\
&= \left(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) + \int_{\mathcal{A}_1} \frac{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1}. && \text{(165b)}
\end{aligned}$$

The equality in (165) implies that for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\right\}$ the limit satisfies

$$\begin{aligned}
&\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \left(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_1} \frac{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(166a)} \\
&= \left(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) + \lim_{\lambda\to 0^+} \big(\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)\big)\right. \\
&\qquad \left. \lim_{\lambda\to 0^+} \int_{\mathcal{A}_1} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(166b)} \\
&= \left(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) + \left(\delta^\star_{Q,\boldsymbol{z}} + \lim_{\lambda\to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda)\right)\right. \\
&\qquad \left. \int_{\mathcal{A}_1} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \lim_{\lambda\to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(166c)} \\
&= \left(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) + 0 \int_{\mathcal{A}_1} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(166d)} \\
&= \big(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big)\big)^{-1}, && \text{(166e)}
\end{aligned}$$

where equality (166d) follows from Lemma 4.8.

The second part is as follows. Consider that $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star_{Q,\boldsymbol{z}}\right\}$. Hence, for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star_{Q,\boldsymbol{z}}\right\}$, there exists a real value $\epsilon > 0$ such

that the function $\mathsf{L}_{\boldsymbol{z}}$ satisfies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}} = \epsilon. \tag{167}$$

From (164), for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}}\right\}$ it holds that

$$\begin{aligned}
\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \left( \int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}} \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu}) \right. \\
&\quad \left. + \int_{\mathcal{A}_1} \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} && \text{(168a)} \\
&= \left( \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) \right. \\
&\quad \left. + \int_{\mathcal{A}_1} \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1}. && \text{(168b)}
\end{aligned}$$

The equality in (168) implies that for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}$, the limit satisfies

$$\begin{aligned}
\lim_{\lambda \to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \left( \lim_{\lambda \to 0^+} \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) \right. \\
&\quad \left. + \lim_{\lambda \to 0^+} \int_{\mathcal{A}_1} \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} && \text{(169a)} \\
&= \left( \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon + \lim_{\lambda \to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\delta^{\star}_{Q,\boldsymbol{z}} + \lim_{\lambda \to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda)} Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) \right. \\
&\quad + \lim_{\lambda \to 0^+} \big(\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)\big) \\
&\quad \left. \lim_{\lambda \to 0^+} \int_{\mathcal{A}_1} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} && \text{(169b)} \\
&= \left( \infty \; Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) + \epsilon \int_{\mathcal{A}_1} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} && \text{(169c)}
\end{aligned}$$

The proof continuous by considering the cases in which $Q(\mathcal{A}_0) > 0$ and $Q(\mathcal{A}_0) = 0$ for the limit in (169). Under the assumption that $Q(\mathcal{A}_0) > 0$, from (169b) the limit satisfies

$$\lim_{\lambda \to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{170a}$$

Under the assumption that $Q(\mathcal{A}_0) = 0$, from the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), if $Q(\mathcal{A}_0) = 0$, then it is implied that there is at least an infinite countable sequence of models $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots$ in $\operatorname{supp} Q$ that satisfies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2) > \cdots, \tag{171}$$

such that

$$\lim_{i\to\infty} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_i) = \delta^\star_{Q,\boldsymbol{z}}. \tag{172}$$

From (172) the function inside the integral in (169c) is unbounded over the set $\mathcal{A}_1$. Therefore,

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{0+(\infty)\epsilon} \tag{173a}$$

$$= 0. \tag{173b}$$

The third part of the proof follows by noticing that the set $\left\{\boldsymbol{\nu}\in\operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}(\boldsymbol{\nu})<\delta^\star_{Q,\boldsymbol{z}}}\right\}$ has measure zero with respect to $Q$ and thus, for all $\boldsymbol{\theta}\in\left\{\boldsymbol{\nu}\in\operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}(\boldsymbol{\nu})<\delta^\star_{Q,\boldsymbol{z}}}\right\}$, the value $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$ is unchanged. Hence, it is assumed that for all $\boldsymbol{\theta}\in\left\{\boldsymbol{\nu}\in\operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}(\boldsymbol{\nu})<\delta^\star_{Q,\boldsymbol{z}}}\right\}$, it holds that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{174}$$

This completes the third part and the entire proof.

■

# J Proof of Lemma 4.13

*Proof:* Consider the following partition of the set $\mathcal{M}$ formed by the sets

$$\mathcal{A}_0 = \left\{\boldsymbol{\theta}\in\mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^\star_{Q,\boldsymbol{z}}\right\}, \tag{175a}$$

$$\mathcal{A}_1 = \left\{\boldsymbol{\theta}\in\mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > \delta^\star_{Q,\boldsymbol{z}}\right\}, \tag{175b}$$

$$\mathcal{A}_2 = \left\{\boldsymbol{\theta}\in\mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^\star_{Q,\boldsymbol{z}}\right\}, \tag{175c}$$

with $\delta^\star_{Q,\boldsymbol{z}}$ in (28) and the function $\mathsf{L}_{\boldsymbol{z}}$ in (3). Note that $\mathcal{A}_0 = \mathcal{L}^\star_{Q,\boldsymbol{z}}$, with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (29).

For all $\lambda\in(0,\infty)$, it holds that

$$1 = \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{176a}$$

$$= \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) \tag{176b}$$

$$= \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_1} \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{176c}$$

where equality (176b) follows from the fact that $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) = 0$, which follows from the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (28) and the fact that the probability measure

$\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is mutually absolutely continuous with respect to the reference measure $Q$. The above implies that

$$\lim_{\lambda\to 0^+}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0)+\int_{\mathcal{A}_1}\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right)$$

$$=\lim_{\lambda\to 0^+}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0)+\lim_{\lambda\to 0^+}\int_{\mathcal{A}_1}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{177a}$$

$$=\lim_{\lambda\to 0^+}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0)+\int_{\mathcal{A}_1}\lim_{\lambda\to 0^+}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{177b}$$

$$=\lim_{\lambda\to 0^+}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0), \tag{177c}$$

$$=1, \tag{177d}$$

where, the equality in (177b) follows from the dominated convergence theorem [45, Theorem 1.6.9 page 50], given that from Lemma 4.10 for all $\lambda\in\bar{K}_{Q,\boldsymbol{z}}$, the Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is positive and finite; and the equality in (177c) follows from the fact that for all $\boldsymbol{\theta}\in\mathcal{A}_1$, it holds that $\lim_{\lambda\to 0^+}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=0$ from Lemma 4.12.

Hence, it holds that

$$\lim_{\lambda\to 0^+}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big)=1, \tag{178}$$

which completes the proof. ■

# K Proof of Lemma 5.1

*Proof:* From Lemma 4.1 and the fact that the measures $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $Q$ are mutually absolutely continuous, it holds that for all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})=\frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}, \tag{179}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $\bar{K}_{Q,\boldsymbol{z}}$ are in (3) and (23b), respectively. From (179), it follows that for all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$0=\lambda\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})-\bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{180}$$

Integrating both sides of (180) with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ yields

$$0=\int\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})-\lambda\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1}+\bar{K}_{Q,\boldsymbol{z}}(\lambda)\right)\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{181a}$$

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

$$+ \int \bar{K}_{Q,\boldsymbol{z}}(\lambda) \, \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{181b}$$

$$= \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \lambda \int \mathrm{d}Q(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda) \tag{181c}$$

$$= \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \lambda + \bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{181d}$$

From (181d), it holds that

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{182}$$

which completes the proof. ■

# L Proof of Lemma 5.2

*Proof:* The proof is divided into four parts. The first part characterizes the functional inverse of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b). The second part provides necessary conditions for the differentiation lemma [48, Theorem 6.28, page 160] to hold, which is used in the third and fourth parts of the proof. The third part presents the first derivative of the functional inverse and shows that its derivative is strictly positive. The forth part shows the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ decreases with lambdas decreasing.

The first part is as follows. Under the assumption that $t \in (0, \infty)$ and $\gamma \in \mathbb{R}$ satisfy that

$$1 = \int \frac{\mathrm{d}\bar{P}^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{183a}$$

$$= \int \frac{t}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma} \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{183b}$$

which implies that

$$\bar{K}_{Q,\boldsymbol{z}}(t) = \gamma, \tag{184}$$

where $\bar{K}_{Q,\boldsymbol{z}}$ is defined in (23b). From Lemma 4.4, let the functional $\bar{K}^{-1}_{Q,\boldsymbol{z}} : \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right) \to (0, \infty)$ be the functional inverse of $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) given by

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta})}, \tag{185}$$

which follows from the constraint in (15b) and completes the first part of the proof.

The second part is as follows. From Lemma 4.4, for all $\lambda \in (0, \infty)$, it holds that

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) > -\delta^{\star}_{Q,\boldsymbol{z}}. \tag{186}$$

Then, for the third case, if $\gamma > -\delta^{\star}_{Q,\boldsymbol{z}}$, it follows that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma} \, \mathrm{d}Q(\boldsymbol{\theta}) \leq \int \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}} + \gamma} \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{187a}$$

$$= \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}} + \gamma}. \tag{187b}$$

From the equality in (187b) and the fact that $Q$ is a probability measure, it follows that the integral on the right-hand side in (187a) is bounded. Hence, from the dominated convergence theorem [45, Theorem 1.6.9, page 50] the left-hand side of (187a) is finite for all $\beta \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. Furthermore, for all $\boldsymbol{\theta} \in \mathcal{M}$ the partial derivative of $\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}$ in (187a) with respect to $\beta$ yields

$$\frac{\partial}{\partial \beta}\left(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\right) = -\frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}, \tag{188}$$

which exists for all $\beta \in (-\delta^{\star}, \infty)$. Then, from the differentiation lemma [48, Theorem 6.28, page 160], the interchange of the integral with the derivative on the right-hand side of (187a) is possible. Hence,

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta}) = \int \frac{\partial}{\partial \beta} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{189}$$

which completes the second part of the proof.

The third part is as follows.

For all $\beta \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, the derivative of the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (185) satisfies:

$$\bar{K}^{-1(1)}_{Q,\boldsymbol{z}}(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta})\right)^{-1} \tag{190a}$$

$$= -\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta})\right)^{-2} \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)\right)^{-1} \tag{190b}$$

$$= -\left(\frac{1}{\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta})}\right)^{2} \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)\right)^{-1} \tag{190c}$$

$$= -\frac{\frac{\mathrm{d}}{\mathrm{d}\beta}\left(\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)\right)^{-1}}{\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta})\right)^{2}} \tag{190d}$$

$$= -\frac{\int -\frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2} \, \mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta})\right)^{2}} \tag{190e}$$

$$= \frac{\int \frac{1}{(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^2}\,\mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2}. \tag{190f}$$

where the equality in (190b) follows from (189); and the equality in (190d) follows from (188).

Note that from Jensen's inequality, it follows that

$$\left(\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2 \leq \int \frac{1}{(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^2}\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{191}$$

Then, from (190f) and (191), for all $\beta \in \left(\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$ it holds that

$$\bar{K}^{-1(1)}_{Q,\boldsymbol{z}}(\beta) \geq 1. \tag{192}$$

which completes the third part of the proof.

The fourth part is as follows. Let the real values $(\lambda_1, \lambda_2) \in (0,\infty)^2$ be such that $\lambda_2 > \lambda_1$, which implies from Lemma 4.4 that $\bar{K}_{Q,\boldsymbol{z}}(\lambda_2) > \bar{K}_{Q,\boldsymbol{z}}(\lambda_1)$. Then, from Lemma 5.1 it follows that

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda_2 - \lambda_1 + \bar{K}_{Q,\boldsymbol{z}}(\lambda_1) - \bar{K}_{Q,\boldsymbol{z}}(\lambda_2) \tag{193a}$$

$$= \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_2) - \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_1) + \beta_1 - \beta_2, \tag{193b}$$

where equality (193b) follows from substituting (185) into (193a). Note that (192) implies that

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_2) - \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_1) \geq \beta_2 - \beta_1. \tag{194}$$

Thus, from (193b) and (194) it follows that

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq 0, \tag{195}$$

Furthermore, from Lemma 4.4 and $\mathsf{L}_{\boldsymbol{z}}$ in (3), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ the fraction $(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^{-1}$ is strictly convex. Hence, equality in (192) and in (195) holds if and oln if $\mathsf{L}_{\boldsymbol{z}}$ is nonseparable with respect to $Q$, which completes the proof.

■

# M Proof of Lemma 5.3

*Proof:* From Theorem 4.1 and the definition of the Type-II relative entropy, it holds that

$$\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{196a}$$

$$\leq \log\left(\int \frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\right) \tag{196b}$$

$$= \log\left(\int \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\,\mathrm{d}Q(\boldsymbol{\theta})\right) \tag{196c}$$

$$= \log\left(\frac{1}{\lambda}\int \bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\right) \tag{196d}$$

$$= \log\left(\frac{1}{\lambda}\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{R}_{\boldsymbol{z}}(Q)\big)\right), \tag{196e}$$

where inequality (196b) follows from Jensen's Inequality. From (196e), it follows that

$$\mathsf{R}_{\boldsymbol{z}}(Q) \geq \exp\Big(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\Big)\lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{196f}$$

Hence, the difference between the expected empirical risk of the probability measures $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $Q$, from Lemma 5.1 and (196f), satisfies that

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geq \lambda\Big(\exp\Big(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\Big)-1\Big), \tag{197}$$

which completes the proof. ■

# N Proof of Lemma 5.5

*Proof:* From Lemma 4.4 and Lemma 5.1, it holds that

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) < \lambda + \delta^{\star}_{Q,\boldsymbol{z}}. \tag{198}$$

Similarly, from the definition of the Rashmon set in (27) and $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), it follows that

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geq \delta^{\star}_{Q,\boldsymbol{z}}. \tag{199}$$

The proof continues by determining the conditions for which the equality in (199) holds. Assume the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) (see Definition 4.1). Then, there exists a real value $\epsilon > 0$ and two nonnegligible sets $\mathcal{A}$ and $\mathcal{B}$ with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16), such that

$$\mathcal{A} = \big\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\big\}, \text{and} \tag{200a}$$

$$\mathcal{B} = \big\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geq \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\big\}. \tag{200b}$$

Under the assumption that $\mathsf{L}_{\boldsymbol{z}}$ is separable, the expected empirical risk satisfies

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \int_{\mathcal{A}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})$$

$$
\begin{aligned}
&\quad + \int_{\mathcal{B}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && \text{(201a)}\\
&\geq \int_{\mathcal{A}} \delta^{\star}_{Q,\boldsymbol{z}} \, \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \\
&\quad + \int_{\mathcal{B}} \big(\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\big) \, \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && \text{(201b)}\\
&= \delta^{\star}_{Q,\boldsymbol{z}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) \\
&\quad + \big(\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\big) \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}), && \text{(201c)}\\
&= \delta^{\star}_{Q,\boldsymbol{z}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) \\
&\quad \big(\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\big)\Big(1 - \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\Big) && \text{(201d)}\\
&= \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\Big(1 - \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\Big) && \text{(201e)}\\
&> \delta^{\star}_{Q,\boldsymbol{z}}, && \text{(201f)}
\end{aligned}
$$

where inequality (201b) follows from the fact that $\delta^{\star}_{Q,\boldsymbol{z}}$ and $\big(\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\big)$ are the minimum empirical risk for the set $\mathcal{A}$ and $\mathcal{B}$ respectively; and inequality (201f) follows from the fact that the sets $\mathcal{A}$ and $\mathcal{B}$ are nonnegligible with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which implies $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) < 1$. This proves that for separable empirical risk functions $\mathsf{L}_{\boldsymbol{z}}$ the inequality in (199) is strict.

Considering the case in which the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (3) is not separable with respect to $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16). Then, for all $\boldsymbol{\theta} \in \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, the empirical risk satisfies $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}$, which implies $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \delta^{\star}_{Q,\boldsymbol{z}}$. Hence, equality in (199) holds for all nonseparable empirical risk functions $\mathsf{L}_{\boldsymbol{z}}$, which completes the proof. ■

# O Proof of Theorem 6.1

*Proof:* Let $\delta$ be a real in $\big(\delta^{\star}_{Q,\boldsymbol{z}}, \infty\big)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28). Let also $\gamma \in (0, \infty)$ satisfy the following equality:

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \leq \delta. \tag{202}$$

Note that from Lemma 4.4 the function $\bar{K}_{Q,\boldsymbol{z}}$ is continuous and from Lemma 4.4 the set $(0, \infty)$ is convex. Moreover, from Lemma 5.6, it follows that such $\gamma$ in (202) always exists. From (27), for all $\delta \in \big(\delta^{\star}_{Q,\boldsymbol{z}}, \infty\big)$, it holds that

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \supseteq \mathcal{L}^{\star}_{Q,\boldsymbol{z}}, \tag{203}$$

and thus,

$$\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) \geq \bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big), \tag{204}$$

with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ defined in (29). Let
*lambda* be a positive real such that $\lambda \leq \gamma$, and

$$\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 1-\epsilon. \tag{205}$$

The existence of such a positive real $\lambda$ that satisfies (205) follows from Lemma 4.13. From Theorem 4.1 and (203), it follows that

$$\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) \geq \bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) \tag{206a}$$

Hence, from equality (206a) it holds that

$$1-\epsilon < \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) \tag{207}$$

$$\leq \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)). \tag{208}$$

The equality in (208) implies that the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is $(\delta,\epsilon)$-optimal (Definition 6.1), which completes the proof. ■

# P Proof of Lemma 7.1

*Proof:* From Theorem 4.1, it follows that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) = \log\left(\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) \tag{209a}$$

$$= \log(\lambda) - \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big) \tag{209b}$$

$$= \log(\lambda) - \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta}), \tag{209c}$$

where the function $\mathsf{V}_{\boldsymbol{z},\lambda}$ is defined in (60). Thus,

$$\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) = \int \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{210a}$$

$$= \log(\lambda) - \int \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{210b}$$

$$= \log(\lambda) - \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{210c}$$

where the functional $\bar{\mathsf{R}}_{\boldsymbol{z},\lambda}$ is defined in (61). Hence, it follows from (210c) that

$$\log(\lambda) = \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right), \tag{211}$$

which completes the proof for (64).

From (209), it follows that

$$\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = -\int \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{212a}$$

$$= -\log(\lambda) + \int \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{212b}$$

$$= -\log(\lambda) + \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}(Q), \tag{212c}$$

Hence, it follows from (210c) that

$$\log(\lambda) = \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}(Q) - \mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{213}$$

which completes the proof for (65). ∎

# Q Proof of Lemma 7.2

*Proof:* The proof uses the fact that the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) is mutually absolutely continuous with the probability measure $Q$ in Theorem 4.1. Hence, the probability measure $P$ is mutually absolutely continuous with $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, as a consequence of the assumption that $P \in \bigcirc_Q(\mathcal{M})$.

The proof follows by noticing that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\log\left(\frac{\mathrm{d}P}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) = \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \tag{214a}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \tag{214b}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log\left(\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) \tag{214c}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log(\lambda) + \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big) \tag{214d}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log(\lambda) + \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta}), \tag{214e}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$, $\bar{K}_{Q,\boldsymbol{z}}$ and $\mathsf{V}_{\boldsymbol{z},\lambda}$ are defined in (3), (23b) and in (60), respectively; and the equality in (214c) follows from (16). Hence, the relative entropy $\mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ satisfies,

$$\mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{215a}$$

$$= \int \left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log(\lambda) + \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{215b}$$

$$= \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) - \log(\lambda)$$

$$+\int \mathsf{V}_{\boldsymbol{z},\lambda}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \tag{215c}$$

$$= \mathsf{D}\left(P\|Q\right) - \log(\lambda) + \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}(P) \tag{215d}$$

$$= \mathsf{D}\left(P\|Q\right) - \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \mathsf{D}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) + \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}(P), \tag{215e}$$

where equality (215b) follows from (214); and equality (215e) follows from Lemma 7.1. Thus, from (215), it follows that

$$\bar{\mathsf{R}}_{\boldsymbol{z},\lambda}(P) - \bar{\mathsf{R}}_{\boldsymbol{z},\lambda}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \mathsf{D}\Big(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \mathsf{D}\left(P\|Q\right) + \mathsf{D}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big), \tag{216}$$

which completes the proof. ■

# R Proof of Lemma 7.4

*Proof:* The proof is presented in two parts. First, the sensitivity $\mathsf{S}_{Q,\lambda}$ ( [49, Definition 3.1 page 9]) is evaluated with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16). Second, the log sensitive defined $\bar{\mathsf{S}}_{Q,\lambda}$ in (67) is evaluated with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (9).

For the first part, from Lemma 7.1, for all $\alpha \in (0,\infty)$, it holds that

$$\mathsf{D}\Big(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) = -\bar{\mathsf{R}}_{\boldsymbol{z},\alpha}\Big(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) + \log(\alpha), \tag{217}$$

where the functional $\bar{\mathsf{R}}_{\boldsymbol{z},\alpha}$ is defined in (61).

Similarly, from [49, Lemma 2.2, page 8], for all $\lambda \in (0,\infty)$, it holds that

$$\mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = -\lambda\Big(\mathsf{D}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) + K_{Q,\boldsymbol{z}}\Big(-\frac{1}{\lambda}\Big)\Big), \tag{218}$$

with the functional $\mathsf{R}_{\boldsymbol{z}}$ defined in (6).

From [49, Theorem 3.1 page 9] the sensitivity of the probability measure $\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies

$$\mathsf{S}_{Q,\lambda}\Big(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$$

$$= \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \tag{219a}$$

$$= \lambda\Big(\mathsf{D}\Big(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \mathsf{D}\Big(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big) + \mathsf{D}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big)\Big). \tag{219b}$$

Plugging (217) and (218) into (219) yields

$$\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \bar{\mathsf{R}}_{\boldsymbol{z},\alpha}\Big(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$$

$$= \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \log(\alpha) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{220}$$

which completes the first part of the proof.

For the second part, from Lemma 7.1, for all $\alpha \in (0,\infty)$, it holds that

$$\bar{\mathsf{R}}_{\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = -\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) + \log(\alpha), \tag{221}$$

where the functional $\bar{\mathsf{R}}_{\boldsymbol{z},\alpha}$ is defined in (61).

Similarly, from [49, Lemma 2.2, page 8], for all $\lambda \in (0,\infty)$, it holds that

$$\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) = \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{222}$$

with the functional $\mathsf{R}_{\boldsymbol{z}}$ defined in (6).

From Lemma 7.2 the logarithmic sensitivity of the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is

$$\begin{aligned}&\bar{\mathsf{R}}_{\boldsymbol{z},\alpha}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{R}}_{\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\ &\quad= \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) + \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right).\end{aligned} \tag{223}$$

Plugging (221) and (222) into (223) yields

$$\begin{aligned}&\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{R}}_{\boldsymbol{z},\alpha}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\ &\quad= -\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \left(\log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right).\end{aligned} \tag{224}$$

The proof proceeds by subtracting (224) from (220), resulting in

$$\begin{aligned}&\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\ &\qquad= \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\ &\qquad\quad+2\left(\log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right),\end{aligned} \tag{225}$$

where the functions $\mathsf{S}_{Q,\lambda}$ and $\bar{\mathsf{S}}_{Q,\alpha}$ are respectively defined in [49, Definition 8] and (67). From [34, Theorem 1] and Lemma 7.4, it follows that

$$\begin{aligned}&\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\ &\quad= 2\left(\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right)\right) + \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\ &\qquad-\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right).\end{aligned} \tag{226}$$

Substituting (226) into (225) yields

$$\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) = \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{227}$$

which completes the proof. ■

# S Example 4.3

*Proof.* The Lagrangian multiplier $\beta$ for the optimization problem in (11) satisfies

$$\int \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \, \mathrm{d}Q(\boldsymbol{\nu}) = 1. \tag{228}$$

From the empirical risk function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to \mathbb{R}_0^+$ in (41), which is a simple function, and the probability measure $Q$ in (40a), it holds that

$$\begin{aligned}
\int \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \, \mathrm{d}Q(\boldsymbol{\nu}) =& \lambda \left( \frac{1}{\beta + c_0} Q(\mathcal{T}(\boldsymbol{z})) + \frac{1}{\beta + c_1} Q(\mathcal{M} \setminus \mathcal{T}(\boldsymbol{z})) \right) && \text{(229a)} \\
=& \lambda \left( \frac{1}{\beta + c_0} Q(\mathcal{T}(\boldsymbol{z})) + \frac{1}{\beta + c_1} (1 - Q(\mathcal{T}(\boldsymbol{z}))) \right) && \text{(229b)} \\
=& \lambda \left( \frac{(\beta + c_1) Q(\mathcal{T}(\boldsymbol{z})) + (\beta + c_0)(1 - Q(\mathcal{T}(\boldsymbol{z})))}{\beta^2 + \beta(c_0 + c_1) + c_0 c_1} \right) && \text{(229c)} \\
=& \lambda \left( \frac{(c_1 - c_0) Q(\mathcal{T}(\boldsymbol{z})) + \beta + c_0}{\beta^2 + \beta(c_0 + c_1) + c_0 c_1} \right). && \text{(229d)}
\end{aligned}$$

From (228) and (229d), it follows that

$$\begin{aligned}
0 =& \beta^2 + \beta(c_0 + c_1) + c_0 c_1 - \lambda((c_1 - c_0) Q(\mathcal{T}(\boldsymbol{z})) + \beta + c_0) && \text{(230a)} \\
=& \beta^2 + \beta(c_0 + c_1 - \lambda) + c_0 c_1 - \lambda c_0 - \lambda(c_1 - c_0) Q(\mathcal{T}(\boldsymbol{z})). && \text{(230b)}
\end{aligned}$$

From (230b) and the fact that $c_0 = 0$ in equation (41), it holds that

$$0 = \beta^2 + \beta(c_1 - \lambda) - \lambda c_1 Q(\mathcal{T}(\boldsymbol{z})). \tag{230c}$$

Observe that the equality in (230c) is a quadratic polynomial that has two roots $r_1$ and $r_2$. Hence, (230c) in terms of $r_1$ and $r_2$ satisfies

$$\begin{aligned}
0 =& \beta^2 - (r_1 + r_2)\beta + r_1 r_2 && \text{(231a)} \\
=& (\beta - r_1)(\beta - r_2), && \text{(231b)}
\end{aligned}$$

where the roots $r_1$ and $r_2$ are given by the quadratic formula such that

$$r_1 = -\frac{(c_1 - \lambda)}{2} - \sqrt{\left( \frac{c_1 - \lambda}{2} \right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}, \tag{232a}$$

and

$$r_2 = -\frac{(c_1 - \lambda)}{2} + \sqrt{\left( \frac{c_1 - \lambda}{2} \right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}. \tag{232b}$$

The proof continues by verifying that the roots in (232a) and (232b) are real and there is only one positive root for all $\lambda \in (0, +\infty)$ and for all $Q(\mathcal{T}(\boldsymbol{z})) \in [0, 1)$.

Note that for all $c_1 \in (0, \infty)$ and for all $\lambda \in [0, +\infty)$, it holds that

$$-\frac{c_1 - \lambda}{2} \leq \left|\frac{c_1 - \lambda}{2}\right| \tag{233}$$

$$= \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2} \tag{234}$$

$$\leq \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}. \tag{235}$$

Observe that for all $Q(\mathcal{T}(\boldsymbol{z})) \in [0, 1)$, $c_1 \in (0, \infty)$ and $\lambda \in [0, \infty)$ the expressions $\left(\frac{c_1 - \lambda}{2}\right)^2$ and $\lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))$ are always positive. Thus, the square roots in (232a) and (232b) are real, which implies that $r_1$ and $r_2$ are real. From (235), for all $\lambda \in [0, +\infty)$ and for all $Q(\mathcal{T}(\boldsymbol{z})) \in [0, 1)$, it holds

$$r_1 < 0; \tag{236a}$$

and following the same arguments

$$r_2 > 0. \tag{236b}$$

Hence, the solution for the Lagrange Multiplier $\beta$ that satisfies (228) given the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (41) and the probability measure $Q$ in (40a) is

$$\beta = -\frac{(c_1 - \lambda)}{2} + \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}, \tag{237}$$

which completes the proof. □

# T Example 4.2

*Proof.* Consider the Type-II ERM-RER problem in (11) and assume that: (*a*) $\lambda = 0.5$; (*b*) $\mathcal{M} = \mathcal{X} = \mathcal{Y} = [0, \infty)$; (*c*) $\boldsymbol{z} = ((1, 1))$ and (*d*) $Q \ll \mu$, with $\mu$ the Lebesgue measure, such that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu}(\boldsymbol{\theta}) = 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}). \tag{238}$$

Let also the function $f : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ be

$$f(\boldsymbol{\theta}, x) = x\boldsymbol{\theta}, \tag{239}$$

and the loss function $\ell$ in (2) be

$$\ell(f(\boldsymbol{\theta}, x), y) = (x\boldsymbol{\theta} - y)^2. \tag{240}$$

Hence, from the fact that $\boldsymbol{z}$ is a single data point and the definition of $\mathsf{L}_{\boldsymbol{z}}$ in (3), it follows that

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \ell(f(\boldsymbol{\theta}, x), y) \tag{241a}$$
$$= (x\boldsymbol{\theta} - y)^2. \tag{241b}$$

Based on the previous assumptions regarding the functions $\ell$ and $f$, as well as the sets $\mathcal{M}$, $\mathcal{X}$, $\mathcal{Y}$, and the measure $Q$, the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (28) implies that $\delta^\star_{Q,\boldsymbol{z}} = 0$.

From Theorem 4.1, the solution to the Type-II ERM-RER problem in (11) is the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that is mutually absolutely continuous with respect to $Q$, which implies that under the above assumptions $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to the Lebesgue measure $\mu$ such that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}\mu}(\boldsymbol{\theta}) = \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}). \tag{242}$$

Hence, it follows that

$$\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{M}) = \int_{\mathcal{M}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{243a}$$
$$= \int_{\mathcal{M}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}\mu}(\boldsymbol{\theta})\, \mathrm{d}\mu(\boldsymbol{\theta}) \tag{243b}$$
$$= \int_0^\infty \frac{\lambda 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(x\boldsymbol{\theta} - y)^2 + \beta}\, \mathrm{d}\boldsymbol{\theta} \tag{243c}$$
$$= 4\lambda \int_0^\infty \frac{\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2}\, \mathrm{d}\boldsymbol{\theta} \tag{243d}$$
$$\tag{243e}$$

The integral in (243d) can be rewritten as

$$\int_0^\infty \frac{\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2}\, \mathrm{d}\boldsymbol{\theta}$$
$$= \int_0^\infty \left( \frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \frac{1}{2}\exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1} + \frac{\frac{1}{2}\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \frac{1}{2}\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \right) \mathrm{d}\boldsymbol{\theta} \tag{244}$$
$$= \frac{1}{2}\int_0^\infty \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1}\, \mathrm{d}\boldsymbol{\theta}$$
$$+ \frac{1}{2}\int_0^\infty \frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2}\, \mathrm{d}\boldsymbol{\theta}. \tag{245}$$

Using integration by parts on the second integral in (245) yields

$$\int_0^\infty \frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2}\, \mathrm{d}\boldsymbol{\theta}$$

$$= -\frac{(\boldsymbol{\theta}+1)\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)} - \int_0^\infty \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1}\,\mathrm{d}\boldsymbol{\theta}. \tag{246}$$

Plugging (246) into (245) yields

$$\int_0^\infty \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\boldsymbol{\theta} = \left[-\frac{(\boldsymbol{\theta}+1)\exp(-2\boldsymbol{\theta})}{2(\boldsymbol{\theta}-1)}\right]_0^\infty. \tag{247}$$

From (247), it follows that

$$4\lambda\int_0^\infty \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\boldsymbol{\theta} = 4\lambda\left[-\frac{(\boldsymbol{\theta}+1)\exp(-2\boldsymbol{\theta})}{2(\boldsymbol{\theta}-1)}\right]_0^\infty \tag{248a}$$

$$= 4\lambda\left[-\frac{(\boldsymbol{\theta}+1)\exp(-2\boldsymbol{\theta})}{2(\boldsymbol{\theta}-1)}\right]_0^1 + $$

$$+4\lambda\left[-\frac{(\boldsymbol{\theta}+1)\exp(-2\boldsymbol{\theta})}{2(\boldsymbol{\theta}-1)}\right]_1^\infty \tag{248b}$$

$$\geq 4\lambda\left[-\frac{(\boldsymbol{\theta}+1)\exp(-2\boldsymbol{\theta})}{2(\boldsymbol{\theta}-1)}\right]_0^1 \tag{248c}$$

$$= \lim_{a\to 1^-} 4\lambda\left[-\frac{(\boldsymbol{\theta}+1)\exp(-2\boldsymbol{\theta})}{2(\boldsymbol{\theta}-1)}\right]_0^a \tag{248d}$$

$$= \lim_{a\to 1^-} -4\lambda\frac{(a+1)\exp(-2a)}{2(a-1)}$$

$$-4\lambda\frac{(0+1)\exp(-0)}{2(0-1)} \tag{248e}$$

$$= \lim_{a\to 1^-} -4\lambda\frac{(a+1)\exp(-2a)}{2(a-1)} - 2\lambda \tag{248f}$$

$$= \infty, \tag{248g}$$

which completes the proof. □

# References


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

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

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

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

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

[6] R. Vershynin, *High-dimensional probability: An Introduction with Applications in Data Science*, 1st ed. New York, NY, USA: Cambridge University Press, 2018.

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

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

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

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

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

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

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

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

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

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

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

[18] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky, "Generalization bounds: Perspectives from information theory and PAC-Bayes," arXiv preprint arXiv:2309.04381, Sep. 2023.

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

[20] 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.

[21] 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.

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

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] ——, "Empirical risk minimization with relative entropy regularization," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9454, Feb. 2022.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

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

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

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

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

[39] 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.

[40] 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.

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

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

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

[44] H. Hsu and F. Calmon, "Rashomon capacity: A metric for predictive multiplicity in classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 988–29 000, Dec. 2022.

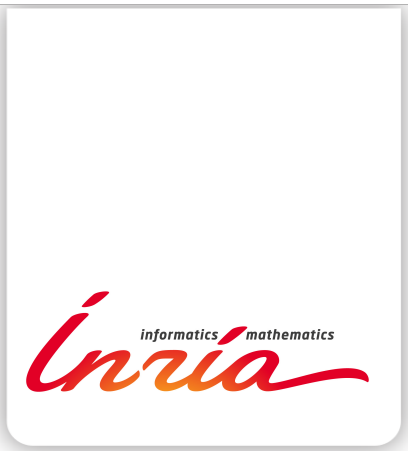
[45] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA, USA: Academic Press, 2000.

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

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

[48] A. Klenke, *Probability Theory: A Comprehensive Course*, 3rd ed. New York, NY, USA: Springer, 2020.

[49] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "Sensitivity of the Gibbs algorithm to data aggregation in supervised machine learning," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9474, Jun. 2022.




**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