Empirical Risk Minimization with Generalized Relative Entropy Regularization

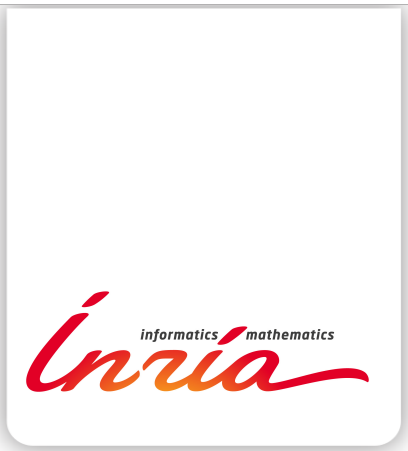
Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, Stefano Rini

Empirical Risk Minimization with Generalized Relative Entropy Regularization

Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, and Stefano Rini

RESEARCH REPORT
N° 9454
February 2022
Project-Team NEO

ISSN 0249-6399 ISRN INRIA/RR--9454--FR+ENG



Empirical Risk Minimization with Generalized Relative Entropy Regularization

Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola,
Alain Jean-Marie, and Stefano Rini

Project-Team NEO

Research Report n° 9454 — version 3 — initial version February 2022 —
revised version July 2022 — 76 pages

Abstract: The empirical risk minimization (ERM) problem with relative entropy regularization (ERM-RER) is investigated under the assumption that the reference measure is a σ-finite measure instead of a probability measure. This assumption leads to a generalization of the ERM-RER (g-ERM-RER) problem that allows for a larger degree of flexibility in the incorporation of prior knowledge over the set of models. The solution of the g-ERM-RER problem is shown to be a unique probability measure mutually absolutely continuous with the reference measure and to exhibit a probably-approximately-correct (PAC) guarantee for the ERM problem. For a given dataset, the empirical risk is shown to be a sub-Gaussian random variable when the models are sampled from the solution to the g-ERM-RER problem. Finally, the sensitivity of the expected empirical risk to deviations from the solution of the g-ERM-RER problem is studied. In particular, the expectation of the absolute value of sensitivity is shown to be upper bounded, up to a constant factor, by the square root of the lautum information between the models and the datasets.

Key-words: Supervised Learning, PAC-Learning, Regularization, Relative Entropy, Empirical Risk Minimization, Maximum Entropy Principle, Sub-Gaussian Random Variables, Bayesian Learning.

Samir M. Perlaza and Alain Jean-Marie are with INRIA at Sophia Antipolis. Gaetan Bisson and Samir M. Perlaza are with the GAATI Laboratory, University of French Polynesia. Iñaki Esnaola is with the Department of Automatic Control and Systems Engineering, University of Sheffield. Iñaki Esnaola and Samir M. Perlaza are also with the Department of Electrical and Computer Engineering, Princeton University. Stefano Rini is with the Department of Electrical and Computer Engineering, National Yang-ming Chao-tung University (NYCU).

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 Généralisée

Résumé : Le problème de minimisation du risque empirique (ERM) avec régularisation par l'entropie relative (ERM-RER) est étudié en considérant que la mesure de référence est une mesure σ-finie au lieu d'une mesure de probabilité. La solution de l'ERM-RER s'avère être une mesure de probabilité unique et son expression explicite est présentée en termes de l'ensemble de données donné et du coefficient de régularisation. Pour un ensemble de données fixe, les ensembles négligeables et la concentration de la mesure optimale (solution à l'ERM-RER) sont caractérisés afin de mettre en évidence l'influence du choix de la mesure de référence et du coefficient de régularisation. Les propriétés de la fonction génératrice de cumulants du risque empirique induite par la mesure optimale sont également étudiées. En utilisant ces propriétés, le risque empirique induit par la mesure optimale s'avère être une variable aléatoire sous-gaussienne. La sensibilité de l'espérance du risque empirique aux déviations de la solution du problème ERM-RER est étudiée. Ensuite, la sensibilité est utilisée pour fournir des bornes supérieures et inférieures sur l'espérance du risque empirique. De plus, il est montré que l'espérance de la sensibilité est majorée, à un facteur constant près, par la racine carrée de l'information lautum entre les modèles et l'ensemble de données.

Mots-clés : Apprentissage Supervisé, Apprentissage PAC, Régularisation, Entropie Relative, Minimisation du Risque Empirique, Principe d'Entropie Maximale, Variables Aléatoires sous-Gaussiennes, Apprentissage Bayésien.

Contents

1 Introduction

In supervised machine learning, the problem of empirical risk minimization (ERM) with relative entropy regularization (ERM-RER) has been the workhorse for building probability measures on the set of models, without any additional assumption on the statistical description of the datasets [1–3]. Instead of additional statistical assumptions on the datasets, which are typical in Bayesian methods [4], relative entropy regularization requires a reference probability measure, which is external to the ERM problem. Often, such reference measure represents prior knowledge and is chosen to assign high probability to the models that induce low empirical risks. The ERM-RER problem is known to possess a unique solution, which is a Gibbs probability measure. Such Gibbs probability measure has been studied using information theoretic notions in [5–10]; statistical physics in [1]; PAC (Probably Approximatively Correct)-Bayesian learning theory in [11–14]; and proved to be of particular interest in classification problems in [15, 16] and supervised learning with energy-based models [17].

In this report, the ERM-RER is generalized to incorporate a σ-finite measure with arbitrary support as the reference measure. Such problem is referred to as the generalized ERM-RER (g-ERM-RER) problem. The flexibility introduced by the g-ERM-RER becomes particularly relevant for the case in which priors are available in the form of probability distributions that can be evaluated up to some normalizing factor [18], or cannot be represented by probability distributions, e.g., equal preferences among elements of infinite countable sets; or equal preferences among the elements of the real numbers.

For some specific choices of the reference measure, the g-ERM-RER boils down to particular cases of special interest: (*i*) the information-risk minimization problem [8]; (*ii*) the ERM with differential entropy regularization; and (*iii*) the ERM with discrete entropy regularization [19]. Hence, the proposed formulation yields a unified mathematical framework that comprises a large class of problems. The g-ERM-RER is shown to possess a unique solution, which is mutually absolutely continuous with the reference measure. Such a solution is recognized as a Gibbs probability measure despite the fact that its partition function is defined with respect to a σ-finite reference measure. Given a dataset, it is shown that the empirical risk observed when models are sampled from the g-ERM-RER optimal probability measure is a sub-Gaussian random variable that exhibits a PAC guarantee for the ERM problem without regularization.

The sensitivity of the expected empirical risk to deviations from the solution of the g-ERM-RER problem is studied. The sensitivity is defined as the difference between two quantities: (*a*) The expectation of the empirical risk with respect to the solution to the mesure that is solution to the g-ERM-RER problem; and (*b*) the expectation of the empirical risk with respect to an alternative measure. The absolute value of the sensitivity is shown to be upper bounded by a term that is proportional to the squared-root of the relative entropy of the alternative measure with respect to the g-ERM-RER-optimal measure. More interestingly,

in a special case, the expectation of the absolute value of the sensitivity with respect to the probability distribution of the data sets is shown to be bounded by a term that is proportional to the squared-root of the lautum information [31] between the models and the datasets. This bound is reminiscent to the result in [10] in which, under certain conditions, the generalization gap of certain machine learning algorithms is upper bounded by a term that is proportional to the squared-root of the mutual information between the models and the datasets.

The reminder of this report is organized as follows. Section 2 introduces three optimization problems: the ERM, the ERM-RER, and the g-ERM-RER. Section 3 presents the solution to the g-ERM-RER problem and introduces its main properties. Section 4 studies the properties of the log-partition function of the g-ERM-RER-optimal probability measure. Section 5 and Section 6 study the properties of the expectation and variance of the empirical risk when the models are sampled from the g-ERM-RER-optimal probability measure. Section 7 describes the monotonic concentration of the g-ERM-RER-optimal probability measure when the regularization factor tends to zero. Section 8 and Section 9 respectively show that the empirical risk when the models are sampled from the g-ERM-RER-optimal probability measure is a sub-Gaussian random variable and exhibits a PAC guarantee with respect to the problem without regularization. Finally, Section 10 studies the sensitivity of the empirical risk function with respect to deviations from the g-ERM-RER optimal measure and shows connections with the lautum information. Section 11 ends this work with conclusions and a discussion on the results.

2 Empirical Risk Minimization (ERM)

Consider three sets $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$, with $\mathcal{M} \subseteq \mathbb{R}^d$ and $d \in \mathbb{N}$. Consider also a function $f : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ such that, for some $\boldsymbol{\theta}^\star \in \mathcal{M}$, there exist two random variables X and Y that satisfy,

$$Y = f(\boldsymbol{\theta}^\star, X). \tag{1}$$

The random variables X and Y jointly form the probability space

$$(\mathcal{X} \times \mathcal{Y}, \mathscr{F}(\mathcal{X} \times \mathcal{Y}), P_{XY}), \tag{2}$$

where $\mathscr{F}(\mathcal{X} \times \mathcal{Y})$ is a σ-algebra on the set $\mathcal{X} \times \mathcal{Y}$, which is assumed to be fixed in this analysis. The elements of the sets $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$ are often referred to as *models*, *patterns*, and *labels*, respectively. A pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ forms a *labeled pattern* or *data point* under the following condition.

Definition 2.1 (Data Point)**.** *The pair (x, y) is said to be a data point if $(x, y) \in \operatorname{supp} P_{XY}$.*

Several data points form a dataset.

Definition 2.2 (Dataset)**.** *Given n data points, with $n \in \mathbb{N}$, denoted by (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), a dataset is represented by the tuple $\big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \big) \in (\mathcal{X} \times \mathcal{Y})^n$.*

The model $\boldsymbol{\theta}^\star$ in (1) is the *ground truth model* and is assumed to be unknown. Given a dataset, the objective is to obtain a model $\boldsymbol{\theta} \in \mathcal{M}$, such that, for all patterns $x \in \mathcal{X}$, the assigned label $f(\boldsymbol{\theta}, x)$ minimizes a notion of *loss* or *risk*. Let the function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty), \tag{3}$$

be such that given a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the cost, loss or risk induced by choosing the model $\boldsymbol{\theta} \in \mathcal{M}$ is $\ell\left(f(\boldsymbol{\theta}, x), y\right)$. Often, the function ℓ is referred to as the *cost function*, *loss function* or *risk function*. In the following, it is assumed that the function ℓ satisfies that, for all $y \in \mathcal{Y}$, the loss $\ell(y, y) = 0$, which implies that correct labelling induces zero cost. Note that there might exist several models $\boldsymbol{\theta} \in \mathcal{M} \setminus \{\boldsymbol{\theta}^\star\}$ such that $\ell\left(f(\boldsymbol{\theta}, x), y\right) = 0$, which reveals the need of a large number of labeled patterns for model selection.

The *empirical risk* induced by the model $\boldsymbol{\theta}$, with respect to a dataset

$$\boldsymbol{z} = \big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \big) \in (\mathcal{X} \times \mathcal{Y})^n, \tag{4}$$

with $n \in \mathbb{N}$, is determined by the function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to [0, +\infty)$, which satisfies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(f(\boldsymbol{\theta}, x_i), y_i\right). \tag{5}$$

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

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}), \tag{6}$$

whose solutions form the set denoted by

$$\mathcal{T}(\boldsymbol{z}) \triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{7}$$

The ground truth model $\boldsymbol{\theta}^\star$ in (1) is one of the solutions to the ERM problem in (6). That is, the model $\boldsymbol{\theta}^\star$ in (1) satisfies that $\boldsymbol{\theta}^\star \in \mathcal{T}(\boldsymbol{z})$ and $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}^\star) = 0$. Hence, the ERM problem in (6) is well posed.

2.1 Main Assumptions

The *generalized relative entropy* is defined below as the extension to σ-finite measures of the relative entropy usually defined for probability measures.

Definition 2.3 (Generalized Relative Entropy)**.** *Given two σ-finite measures P and Q on the same measurable space, such that Q is absolutely continuous with respect to P, the relative entropy of Q with respect to P is*

$$D(Q\|P) = \int \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right) \mathrm{d}P(x), \tag{8}$$

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

In the following, given a measurable space $(\Omega, \mathscr{F})$, the notation $\triangle(\Omega, \mathscr{F})$ is used to represent the set of σ-finite measures that can be defined over such a measurable space. Given a measure $Q \in \triangle(\Omega, \mathscr{F})$, the subset $\triangle_Q(\Omega, \mathscr{F})$ of $\triangle(\Omega, \mathscr{F})$ contains all measures that are absolutely continuous with respect to the measure Q. Given a set $\mathcal{A} \subset \mathbb{R}^d$, with $d \in \mathbb{N}$, the Borel σ-field over $\mathcal{A}$ is denoted by $\mathscr{B}(\mathcal{A})$.

A fundamental assumption in this work is that the function $\bar{\ell} : \mathcal{M} \times \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$, such that for all $(\boldsymbol{\theta}, x, y) \in \mathcal{M} \times \mathcal{X} \times \mathcal{Y}$,

$$\bar{\ell}(\boldsymbol{\theta}, x, y) = \ell(f(\boldsymbol{\theta}, x), y), \tag{9}$$

where the functions f and ℓ are those in (1) and (3), is Borel measurable with respect to the measure space $\big(\mathcal{M} \times \mathcal{X} \times \mathcal{Y}, \mathscr{B}(\mathcal{M}) \times \mathscr{F}(\mathcal{X} \times \mathcal{Y})\big)$, with $\mathscr{F}(\mathcal{X} \times \mathcal{Y})$ the σ-field in (2).

2.2 ERM with Relative Entropy Regularization

When models are chosen by *sampling* from a probability measure over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, one of the performance metrics is the *expected empirical risk*, which is introduced hereunder.

Definition 2.4 (Expected Empirical Risk)**.** *Given a dataset* $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$*, let the function* $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0, +\infty)$ *be such that for all* σ*-finite measures* $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$*, it holds that*

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}), \tag{10}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ *is in* (5)*. Then, when* P *is a probability measure, the expected empirical risk induced by* P *is* $\mathsf{R}_{\boldsymbol{z}}(P)$*.*

The g-ERM-RER problem is parametrized by a σ-finite measure in $\triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a positive real, which are referred to as the *reference measure* and the *regularization factor*, respectively. Let $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ be a σ-finite measure and let λ be a positive real. The g-ERM-RER problem, with parameters Q and λ, consists of the following optimization problem:

$$\begin{aligned} \min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \quad & \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P \| Q), \\ \text{s. t.} \quad & \int \mathrm{d}P(\boldsymbol{\theta}) = 1, \end{aligned} \tag{11}$$

where the dataset $\boldsymbol{z}$ is in (4); and the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (10).

2.3 Special Cases of the G-ERM-RER Problem

Particular choices of the set $\mathcal{M}$ and the reference measure Q lead to special cases of the g-ERM-RER problem in (11). Three cases are of particular interest:

(a) The set $\mathcal{M}$ is the set $\mathbb{R}^d$, with $d \in \mathbb{N}$, and Q is the Lebesgue measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$; ($b$) The set $\mathcal{M} \subset \mathbb{R}^d$ is countable and the measure Q is a counting measure; and (c) The set $\mathcal{M}$ and the measure Q form a probability measure space $(\mathcal{M}, \mathscr{B}(\mathcal{M}), Q)$.

In the former, the g-ERM-RER in (11) satisfies the following

$$
\begin{aligned}
&\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \lambda D(P \| Q) \\
&\qquad \text{s. t.} \qquad \int \mathrm{d}P(\boldsymbol{\theta}) = 1 \\
= &\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) + \lambda \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \\
&\qquad \text{s. t.} \qquad \int \mathrm{d}P(\boldsymbol{\theta}) = 1 \qquad (12) \\
= &\min_{g} \int_{\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, g(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} + \lambda \int_{\mathcal{M}} g(\boldsymbol{\theta}) \log(g(\boldsymbol{\theta})) \, \mathrm{d}\boldsymbol{\theta} \qquad (13) \\
= &\min_{g} \int_{\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, g(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} - \lambda H(g), \qquad (14)
\end{aligned}
$$

where the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ in (12) is a probability density function (p.d.f.) denoted by g, which implies that the optimization domain in (13) is the set of p.d.f.s on $\mathbb{R}^d$. In (14), the notation $H(g)$ represents the differential entropy of the p.d.f. g, c.f., Chapter 8 in [20]. In this particular case, denote the p.d.f. solution of the problem in (12) by $g^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and thus, for all $\boldsymbol{\theta} \in \mathcal{M}$, it follows that

$$
g^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\displaystyle\int_{\mathbb{R}^d} \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}\boldsymbol{\nu}}, \qquad (15)
$$

which is the Gibbs measure with respect to the Lebesgue measure, with parameter λ and energy function $\mathsf{L}_{\boldsymbol{z}}$ in (5).

In case (b), a similar analysis would show that the g-ERM-RER problem in (11) boils down to the ERM with discrete entropy regularization (ERM-DisER).

More specifically,

$$\begin{aligned}
& \min_{P\in\triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P\left(\boldsymbol{\theta}\right) + \lambda D\left(P\|Q\right) \\
& \qquad \text{s. t.} \qquad \int \mathrm{d}P\left(\boldsymbol{\theta}\right) = 1 \\
= & \min_{P\in\triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) + \lambda \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \\
& \qquad \text{s. t.} \qquad \int \mathrm{d}P\left(\boldsymbol{\theta}\right) = 1 \qquad (16) \\
= & \min_{p} \qquad \sum_{\boldsymbol{\theta}\in\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) + \lambda \sum_{\boldsymbol{\theta}\in\mathcal{M}} p(\boldsymbol{\theta}) \log\left(p(\boldsymbol{\theta})\right) \qquad (17) \\
= & \min_{p} \qquad \sum_{\boldsymbol{\theta}\in\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) - \lambda H\left(p\right), \qquad (18)
\end{aligned}$$

where, the Radon-Nikodym derivative $\frac{\mathrm{d}Q}{\mathrm{d}P}$ in (16) is a probability mass function (p.m.f.) denoted by p. This implies that the optimization domain in (17) is the set of p.m.f.s on $\mathcal{M}$; and the entropy $H\left(p\right)$ is that of the p.m.f. p, c.f., Chapter 2 in [20]. In this particular case, denote the p.m.f. solution of the problem in (18) by $p^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, and thus, for all $\boldsymbol{\theta}\in\mathcal{M}$, it follows that

$$p^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\sum\limits_{\boldsymbol{\nu}\in\mathcal{M}} \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)}{\lambda}\right)}, \qquad (19)$$

which is the Gibbs measure with respect to a counting measure, with parameter λ and energy function $\mathsf{L}_{\boldsymbol{z}}$ in (5).

Both, the ERM-DifER and ERM-DisER problems are closely related to those typically arising while using Jayne's maximum entropy principle [21, 22] for classification problems such as those in [15, 16, 23].

Finally, case (c) is a classical optimization problem known as the *information-risk minimization* (IRM) problem in information theory, see for instance [8]. In this case, the g-ERM-RER problem in (11) is known to possess a unique solution consisting in a Gibbs probability measure, see for instance [10, 24] and [25]. Denote such unique solution by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle_Q\left(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right)\right)$, hence, the Radon-Nikodym derivative of $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to Q is

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\lambda}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right)}. \qquad (20)$$

The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (20) is a Gibbs measure with respect to the probability measure Q, parameter λ, and energy function $\mathsf{L}_{\boldsymbol{z}}$ in (5), c.f., [26].

3 The Solution to the G-ERM-RER Problem

The solution to the g-ERM-RER problem in (11) is presented in terms of two objects. First, the function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that for all $t \in \mathbb{R}$,

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\right), \tag{21}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (5). Second, the set $\mathcal{K}_{Q,\boldsymbol{z}} \subset (0, +\infty)$, which is defined by

$$\mathcal{K}_{Q,\boldsymbol{z}} \triangleq \left\{ s \in (0,+\infty) :\ K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) < +\infty \right\}. \tag{22}$$

The notation for the function $K_{Q,\boldsymbol{z}}$ and the set $\mathcal{K}_{Q,\boldsymbol{z}}$ are chosen such that their parametrization by (or dependence on) the dataset $\boldsymbol{z}$ in (4) and the σ-finite measure Q in (11) are highlighted. The following lemma describes the set $\mathcal{K}_{Q,\boldsymbol{z}}$.

Lemma 3.1. *The set $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (22) *is either the empty set or a convex set that satisfies*

$$(0, b) \subset \mathcal{K}_{Q,\boldsymbol{z}}, \tag{23}$$

for some $b \in (0, +\infty]$.

Proof: The proof is presented in Appendix A. ∎

In the special case in which the σ-finite measure Q in (11) is a probability measure, the set $\mathcal{K}_{Q,\boldsymbol{z}}$ is the set formed by all positive reals, as shown by the following lemma.

Lemma 3.2. *Assume that the measure Q in* (11) *is a probability measure. Then, the set $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (22) *satisfies*

$$\mathcal{K}_{Q,\boldsymbol{z}} = (0, +\infty). \tag{24}$$

Proof: The proof is presented in Appendix B. ∎

Using this notation, the solution to the g-ERM-RER problem in (11) is presented by the following theorem.

Theorem 3.1. *The solution to the optimization problem in* (11) *is a unique measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, whose Radon-Nikodym derivative with respect to the σ-finite measure Q satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right), \tag{25}$$

where, the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (5) *and the function $K_{Q,\boldsymbol{z}}$ is defined in* (21).

Proof: The proof is presented in Appendix C. ∎

In Theorem 3.1, when the σ-finite measure Q is a probability measure, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and the function $K_{Q,\boldsymbol{z}}$ are often referred to as a Gibbs measure and a *log-partition function*, see for instance, [27, Section 7.3.1]. In order not to disrupt with the current nomenclature, in the following, independently of whether Q is a probability measure, the function $K_{Q,\boldsymbol{z}}$ and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ are respectively recognized as a log-partition function and a Gibbs measure with parameters Q, λ, and energy function $\mathsf{L}_{\boldsymbol{z}}$.

When the set $\mathcal{M}$ is discrete and the σ-finite measure Q in (11) is the counting measure, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) is equivalent to the p.m.f. $p^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (19). Alternatively, when $\mathcal{M} \subseteq \mathbb{R}^d$ and the σ-finite measure Q in (11) is the Lebesgue measure, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) is the p.d.f. $g^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (15). When Q is a probability measure, the expression in (20) and (25) are identical.

The following lemma shows that the maximum of the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) is achieved by the models that are at the intersection of the support of the reference measure Q and the set of solutions to the optimization problem in (6), that is, the models in the set $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$, with $\mathcal{T}(\boldsymbol{z})$ in (7).

Lemma 3.3. *For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, and for all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in (\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q)^2$, with $\mathcal{T}(\boldsymbol{z})$ in (7), the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) satisfies that*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leqslant \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1) = \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2). \tag{26}$$

Proof: The proof is presented in Appendix D. ∎

When the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) is either the p.d.f. $g^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (15) or the p.m.f. $p^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (19), Lemma 3.4 shows that the elements of the set $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$ in (7) are the modes of the corresponding p.d.f. or p.m.f.

The following lemma shows other bounds on the Radon-Nikodym derivative in (25).

Lemma 3.4. *For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) satisfies that*

$$0 \leqslant \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) < +\infty. \tag{27}$$

The equality $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=0$ *holds if and only if* $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})=+\infty$.

Proof: The proof is presented in Appendix E. ∎

3.1 Asymptotic Regimes

The following lemma describes the asymptotic behavior of the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) when $\lambda\to+\infty$.

Lemma 3.5. *Let the measure Q in* (25) *be a probability measure. Then, for all $\boldsymbol{\theta}\in\operatorname{supp}Q$, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in* (25) *satisfies*

$$\lim_{\lambda\to+\infty}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=1. \tag{28}$$

Proof: From Theorem 3.1, it follows that for all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$\lim_{\lambda\to+\infty}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=\lim_{\lambda\to+\infty}\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})}=\frac{1}{\int\mathrm{d}Q(\boldsymbol{\nu})}=1, \tag{29}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (5). This completes the proof. ∎

Lemma 3.5 unveils the fact that, when Q is a probability measure, for all measurable sets $\mathcal{A}\subseteq\operatorname{supp}Q$,

$$\lim_{\lambda\to+\infty}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})=Q(\mathcal{A}). \tag{30}$$

This is consistent with the fact that when λ tends to infinity, the optimization problem in (11) boils down to exclusively minimizing the relative entropy. Such minimum is zero and is observed when both probability measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q are identical. Alternatively, from Lemma 3.1, it follows that, when Q is not a probability measure, the set $\mathcal{K}_{Q,\boldsymbol{z}}$ might be an interval of the form $(0,b)$, with $b<\infty$. Hence, in such a case, the analysis in which λ tends to infinity is void.

The limit of the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25), when λ tends to zero from the right, can be studied using the following set

$$\mathcal{L}_{\boldsymbol{z}}(\delta)=\{\boldsymbol{\theta}\in\mathcal{M}:\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\leqslant\delta\}, \tag{31}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (5) and $\delta\in[0,+\infty)$. In particular consider the nonnegative real

$$\delta^{\star}_{Q,\boldsymbol{z}}\triangleq\inf\{\delta\in[0,+\infty):Q(\mathcal{L}_{\boldsymbol{z}}(\delta))>0\}. \tag{32}$$

Let also $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ be the following level set of the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5):

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}}=\left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}. \tag{33}$$

Using this notation, the limit of the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25), when λ tends to zero from the right, is described by the following lemma.

Lemma 3.6. *If $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) > 0$, with the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (33) and Q the σ-finite measure in (25), then for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) satisfies*

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)=\frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)}\mathbb{1}_{\left\{\boldsymbol{\theta}\in\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right\}}. \tag{34}$$

Alternatively, if $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$. Then, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) = \begin{cases} +\infty & \text{if} \quad \boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

Proof: The proof is presented in Appendix F. ∎

Consider that $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) > 0$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (33). Under this assumption, from Lemma 3.6, it holds that the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ asymptotically concentrates on the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ when λ tends to zero from the right. The same observation holds for the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. More specifically, note that for all measurable sets $\mathcal{A} \subseteq \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$, it holds that

$$\begin{aligned}
\lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) &= \lim_{\lambda\to 0^+} \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) && (36)\\
&= \lim_{\lambda\to 0^+} \int \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)\mathbb{1}_{\{\boldsymbol{\theta}\in\mathcal{A}\}}\mathrm{d}Q\left(\boldsymbol{\theta}\right) && (37)\\
&= \int \lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)\mathbb{1}_{\{\boldsymbol{\theta}\in\mathcal{A}\}}\mathrm{d}Q\left(\boldsymbol{\theta}\right) && (38)\\
&= \int \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)}\mathbb{1}_{\left\{\boldsymbol{\theta}\in\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right\}}\mathbb{1}_{\{\boldsymbol{\theta}\in\mathcal{A}\}}\mathrm{d}Q\left(\boldsymbol{\theta}\right) && (39)\\
&= \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} \int \mathbb{1}_{\{\boldsymbol{\theta}\in\mathcal{A}\}}\mathrm{d}Q\left(\boldsymbol{\theta}\right) && (40)\\
&= \frac{Q\left(\mathcal{A}\right)}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)}, && (41)
\end{aligned}$$

where the equality in (38) follows from Lemma 3.4 and the dominated convergence theorem [28, Theorem 2.6.9]; and the equality in (39) follows from

Lemma 3.6. In the particular case in which $\mathcal{A} = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (41), it holds that $\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \right) = 1$, which verifies the asymptotic concentration of the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ on the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$. Another interesting observation is that the equality in (41) implies a uniform distribution of the probability among the elements of the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in the limit when λ tends to zero from the right. This becomes more evident in the case in which the set $\mathcal{M}$ is finite and Q is the counting measure. In such a case, the asymptotic probability of each of the elements in $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ when λ tends to zero from the right is $\frac{1}{|\mathcal{L}^{\star}_{Q,\boldsymbol{z}}|}$.

Alternatively, consider that $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (33). Under this assumption, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to the measure Q, and thus, $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$. This is typically the case in which $\mathcal{M} = \mathbb{R}^d$, the measure Q is absolutely continuous with the Lebesgue measure, and the solution to the ERM problem in (6) has a unique solution, i.e., $|\mathcal{T}(\boldsymbol{z})| = 1$.

The following lemma shows that independently of whether the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ is negligible with respect to the measure Q, the limit when λ tends to zero from the right of $P^{(Q,\lambda)}_{\boldsymbol{\Theta}\|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)$ is equal to one.

Lemma 3.7. *The measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (25) *and the set* $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ *in* (33) *satisfy,*

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \right) = 1. \tag{42}$$

Proof: The proof is presented in Appendix G. ∎

In Lemma 3.7, in the case in which $\delta^{\star}_{Q,\boldsymbol{z}} = 0$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32), it holds that $\mathcal{L}^{\star}_{Q,\boldsymbol{z}} = \mathcal{T}(\boldsymbol{z})$ and thus, when λ tends to zero from the right, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ asymptotically concentrates on the set of solutions to the ERM problem in (6). Alternatively, in the case in which $\delta^{\star}_{Q,\boldsymbol{z}} > 0$, it holds that $\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \mathcal{T}(\boldsymbol{z}) = \emptyset$. Hence, when λ tends to zero from the right, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ asymptotically concentrates on a set that does not contain the set of solutions to the ERM problem in (6).

3.2 Coherent and Consistent Measures

Given a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the support of the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is of particular interest. In this regard, note that for all sets $\mathcal{C} \in \mathscr{B}(\mathcal{M})$, it follows from Theorem 3.1 that if $Q(\mathcal{C}) = 0$, then $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = 0$. That is, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with the measure Q. The following lemma shows that the converse is also true, under a certain condition.

Lemma 3.8. *For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), under the assumption that*

$$Q\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = +\infty\}\right) = 0, \tag{43}$$

the σ-finite measure Q and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) are mutually absolutely continuous.

Proof: The proof is presented in Appendix H. ■

The relevance of Lemma 3.8 is that it proves that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the collection of negligible sets with respect to the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) is identical to the collection of negligible sets with respect to the measure Q in (11), under the assumption in (43). Under such assumption, for all subsets $\mathcal{C} \in \mathscr{B}\left(\mathcal{M}\right)$,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{C}\right) > 0 \quad \text{ if and only if } \quad Q\left(\mathcal{C}\right) > 0. \tag{44}$$

The assumption in (43) is trivially true when the function ℓ in (3) is finite.

At the light of Lemma 3.8, a class of reference measures of particular importance in the following sections is that of *coherent* measures.

Definition 3.1 (Coherent Measures)**.** *The σ-finite measure Q in (11) is said to be coherent if, for all $\delta \in (0, +\infty)$, it holds that*

$$Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0, \tag{45}$$

where the set $\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)$ is defined in (31).

The following lemma highlights the relevance of coherence measures.

Lemma 3.9. *The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) satisfies for all $\delta \in (0, +\infty)$ that*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0, \tag{46}$$

with $\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)$ in (31), if and only if the σ-finite measure Q in (11) is coherent.

Proof: The proof is presented in Appendix I. ■

Reference measures that are coherent also exhibit an interesting property in the asymptotic regime when λ tends to zero from the right, as shown by the following lemma.

Lemma 3.10. *The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) and the set $\mathcal{T}\left(\boldsymbol{z}\right)$ in (7) satisfy*

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{T}\left(\boldsymbol{z}\right)\right) = 1, \tag{47}$$

if and only if the σ-finite measure Q in (11) is coherent.

Proof: When the σ-finite measure Q in (11) is coherent, then $\delta^\star_{Q,\boldsymbol{z}} = 0$, with $\delta^\star_{Q,\boldsymbol{z}}$ in (32) and $\mathcal{L}^\star_{Q,\boldsymbol{z}} = \mathcal{T}(\boldsymbol{z})$, with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (33). Thus, from Lemma 3.7, the equality in (47) holds. Alternatively, when the measure Q in (11) is noncoherent, then $\delta^\star_{Q,\boldsymbol{z}} > 0$ and thus, $\mathcal{L}^\star_{Q,\boldsymbol{z}} \cap \mathcal{T}(\boldsymbol{z}) = \emptyset$, which implies

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{T}(\boldsymbol{z}))=0, \tag{48}$$

and completes the proof. ■

Another reference measure Q of particular interest is the one in which for a given $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the set $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (33) satisfies $Q\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) > 0$.

Definition 3.2 (Consistent Measure)**.** *The σ-finite measure Q in* (11) *is said to be consistent if $Q\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) > 0$, with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in* (33)*.*

The relevance of consistent measures is highlighted by the following lemma.

Lemma 3.11. *For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (22)*, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (25) *and the set $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in* (33) *satisfy*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right)>0, \tag{49}$$

if and only if the σ-finite measure Q in (11) *is consistent.*

Proof: When Q is nonconsistent, it holds that $Q\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) = 0$ and thus, from the fact that the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) is absolutely continuous with Q, it holds that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) = 0$. When Q is consistent, it holds that $Q\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) > 0$. Moreover, for all $\boldsymbol{\theta} \in \mathcal{L}^\star_{Q,\boldsymbol{z}}$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < +\infty$ and thus, from Lemma 3.4, it follows that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) > 0$. Hence,

$$\begin{aligned} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) &= \int_{\mathcal{L}^\star_{Q,\boldsymbol{z}}} \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (50)\\ &= \int_{\mathcal{L}^\star_{Q,\boldsymbol{z}}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && (51)\\ &> 0, && (52) \end{aligned}$$

which completes the proof. ■

The case in which the reference measure Q is simultaneously coherent and consistent exhibits a particular property, as shown by the following corollary of Lemma 3.11.

Corollary 3.1. *For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (22)*, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (25) *and the set $\mathcal{T}(\boldsymbol{z})$ in* (7) *satisfy*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{T}(\boldsymbol{z}))>0, \tag{53}$$

if and only if the σ-finite measure Q in (11) *is coherent and consistent.*

3.3 Negligible Sets

Lemma 3.8 shows that, under the assumption in (43), the collection of negligible sets with respect to the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) is identical to the collection of negligible sets with respect to the σ-measure Q. The following lemma shows that the negligible sets with respect to the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) are invariant with respect to the choice of $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$.

Lemma 3.12. *For all $(\alpha, \beta) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), assume that the measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfy (25) with $\lambda = \alpha$ and $\lambda = \beta$, respectively. Then, $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ are mutually absolutely continuous.*

Proof: The proof is presented in Appendix J. ∎

4 The Log-Partition Function

The function $K_{Q,\boldsymbol{z}}$ in (25), often referred to as the log-partition function, exhibits a number of properties, which are relevant in the following sections.

Lemma 4.1. *The function $K_{Q,\boldsymbol{z}}$ in (21) is continuous and differentiable infinitely many times in $(-\infty, 0)$.*

Proof: The proof is presented in Appendix K. ∎

More specific properties for the function $K_{Q,\boldsymbol{z}}$ in (21) can be stated for the case in which the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable with respect to the measure Q in (25).

Definition 4.1 (Separable Empirical Risk Function)**.** *The empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is said to be separable with respect to the σ-finite measure Q in (25), 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 Q, such that for all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$,*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) < c < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) < +\infty. \tag{54}$$

In a nutshell, a nonseparable empirical risk function is a constant almost surely with respect to the measure Q. More specifically, there exists a real $a \geqslant 0$, such that

$$Q(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = a\}) = 1. \tag{55}$$

From this perspective, nonseparable empirical risk functions exhibit little practical interest. This follows from observing that models sampled from a nonseparable probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) induce the same empirical risk.

The definition of separability in Definition 4.1 and Lemma 3.8 lead to the following lemma.

Lemma 4.2. *The empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (5) *is separable with respect to the* σ*-finite measure* Q *in* (25) *if and only if it is separable with respect to the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (25).

Proof: Consider first that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to the σ-finite measure Q. Hence, there exist a positive real $c>0$ and two subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{M}$ that are nonnegligible with respect to Q, such that for all $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\mathcal{A}\times\mathcal{B}$ the inequality in (54) holds. Hence, from (54) the following inequalities hold for all $\lambda\in\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22),

$$-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_1\right)>-\frac{c}{\lambda}>-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_2\right)>-\infty, \tag{56}$$

$$\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_1\right)\right)>\exp\left(-\frac{c}{\lambda}\right)>\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_2\right)\right)>0, \tag{57}$$

and finally,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}_1\right)>\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)-\frac{c}{\lambda}\right)>\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}_2\right)>0. \tag{58}$$

Using the inequality in (58) and the facts that $Q\left(\mathcal{A}\right)>0$ and $Q\left(\mathcal{B}\right)>0$, the following holds

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right)=\int_{\mathcal{A}}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right)>0, \tag{59}$$

and

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{B}\right)=\int_{\mathcal{B}}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right)>0. \tag{60}$$

which implies that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

Consider now that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Hence, there exist a positive real $c>0$ and two subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{M}$ that are nonnegligible with respect to $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, such that for all $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\mathcal{A}\times\mathcal{B}$ the inequality in (54) holds. More specifically,

$$0<P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right)=\int_{\mathcal{A}}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) \tag{61}$$

and

$$0<P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{B}\right)=\int_{\mathcal{B}}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right). \tag{62}$$

From Lemma 3.4 and the inequality in (54), it follows that for all pairs $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\mathcal{A}\times\mathcal{B}$, $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}_1\right)>0$ and $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}_2\right)>0$. Hence, from (61) and (62),

respectively, it follows that $Q(\mathcal{A}) > 0$ and $Q(\mathcal{B}) > 0$, which implies that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to the σ-finite measure Q. This completes the proof. ■

The following lemma presents a general property of the function $K_{Q,\boldsymbol{z}}$ in (21) for the case of separable empirical risk functions.

Lemma 4.3. *The function $K_{Q,\boldsymbol{z}}$ in (21) is convex. The function $K_{Q,\boldsymbol{z}}$ in (21) is strictly convex if and only if the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable with respect to the σ-finite measure Q in (11).*

Proof: The proof is presented in Appendix L. ■

Let the m-th derivative of the function $K_{Q,\boldsymbol{z}}$ in (21) be denoted by $K^{(m)}_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$, with $m \in \mathbb{N}$. Hence, for all $s \in \mathcal{K}_{Q,\boldsymbol{z}}$,

$$K^{(m)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) \triangleq \frac{\mathrm{d}^m}{\mathrm{d}t^m} K_{Q,\boldsymbol{z}}(t)\Big|_{t=-\frac{1}{s}}. \tag{63}$$

The following lemma provides explicit expressions for the first, second and third derivatives of the function $K_{Q,\boldsymbol{z}}$ in (21).

Lemma 4.4. *The first, second and third derivatives of the function $K_{Q,\boldsymbol{z}}$ in (21), denoted respectively by $K^{(1)}_{Q,\boldsymbol{z}}$, $K^{(2)}_{Q,\boldsymbol{z}}$, and $K^{(3)}_{Q,\boldsymbol{z}}$, satisfy for all $\lambda \in \mathrm{int}\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22),*

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{64}$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)^2 \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \text{ and} \tag{65}$$

$$K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)^3 \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{66}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (5) and the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies (25).

Proof: The proof is presented in Appendix M. ■

From Lemma 4.4, it follows that if $\boldsymbol{\Theta}$ is the random vector that induces the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) becomes the random variable

$$W \triangleq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\Theta}), \tag{67}$$

whose mean, variance, and third cumulant are $K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ in (64), $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ in (65), and $K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ in (66), respectively.

5 Expectation of the Empirical Risk

The mean of the random variable W in (67) is equivalent to the expectation of the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), which is equal to $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$, with the function $\mathsf{R}_{\boldsymbol{z}}$ in (10). Often, $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is referred to as the g-ERM-RER-optimal expected empirical risk to emphasize that this is the expected value of the empirical risk when models are sampled from the solution of the g-ERM-RER problem in (11). The following corollary of Lemma 4.4 formalizes this observation.

Corollary 5.1. *For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) verifies that*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{68}$$

where the functions $\mathsf{R}_{\boldsymbol{z}}$ and $K^{(1)}_{Q,\boldsymbol{z}}$ are defined in (10) and (64), respectively.

The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (68) exhibits the following property.

Theorem 5.1. *The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (68) is nondecreasing with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22). Moreover, the function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable with respect to the measure Q if and only if $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is strictly increasing with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$.*

Proof: The proof is presented in Appendix N. ∎

A question that arises from Theorem 5.1 is whether the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (68) can be made arbitrarily close to $\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}^{\star}\right) = 0$, with $\boldsymbol{\theta}^{\star}$ in (1), by making λ arbitrarily small. The following lemma shows that there exist cases in which the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is bounded away from zero, even for arbitrarily small values of λ.

Lemma 5.1. *For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (68) satisfies,*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geqslant \delta^{\star}_{Q,\boldsymbol{z}}, \tag{69}$$

where $\delta^{\star}_{Q,\boldsymbol{z}}$ is defined in (32). Moreover, the function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable with respect to the measure Q in (11) if and only if the inequality in (69) is strict.

Proof: The proof is presented in Appendix O. ∎

In the asymptotic regime when λ tends to zero, the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (68) is equal to $\delta^{\star}_{Q,\boldsymbol{z}}$, as shown by the following lemma.

Theorem 5.2. *The expected empirical risk* $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ *in* (68) *satisfies,*

$$\lim_{\lambda\to 0^+}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\delta^\star_{Q,\boldsymbol{z}}, \tag{70}$$

where $\delta^\star_{Q,\boldsymbol{z}}$ *is defined in* (32)*.*

Proof: The proof is presented in Appendix P. ∎

6 Variance of the Empirical Risk

The monotonicity of the expectation of the random variable W in (67), stated by Theorem 5.1, is not a property exhibited by the variance nor the third cumulant. This section highlights this observation via the following example.

Example 6.1. *Consider the g-ERM-RER problem in* (11)*, under the assumption that* Q *is a probability measure and the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (5) *is such that for all* $\boldsymbol{\theta}\in\mathcal{M}$*,*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})=\begin{cases}0 & \text{if} \quad \boldsymbol{\theta}\in\mathcal{A}\\ 1 & \text{if} \quad \boldsymbol{\theta}\in\mathcal{M}\setminus\mathcal{A},\end{cases} \tag{71}$$

where the sets $\mathcal{A}\subset\mathcal{M}$ *and* $\mathcal{M}\setminus\mathcal{A}$ *are nonnegligible with respect to the reference probability measure* Q*. In this case, the function* $K_{Q,\boldsymbol{z}}$ *in* (21) *satisfies for all* $\lambda>0$*,*

$$K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)=\log\left(Q(\mathcal{A})+\exp\left(-\frac{1}{\lambda}\right)(1-Q(\mathcal{A}))\right). \tag{72}$$

The derivatives $K^{(1)}_{Q,\boldsymbol{z}}$*,* $K^{(2)}_{Q,\boldsymbol{z}}$*, and* $K^{(3)}_{Q,\boldsymbol{z}}$ *in* (63) *of the function* $K_{Q,\boldsymbol{z}}$ *in* (72) *satisfy:*

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)=\frac{\exp\left(-\frac{1}{\lambda}\right)(1-Q(\mathcal{A}))}{Q(\mathcal{A})+\exp\left(-\frac{1}{\lambda}\right)(1-Q(\mathcal{A}))}; \tag{73}$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)=\frac{Q(\mathcal{A})(1-Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{\left(Q(\mathcal{A})+\exp\left(-\frac{1}{\lambda}\right)(1-Q(\mathcal{A}))\right)^2};\ \text{and} \tag{74}$$

$$K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)=\frac{Q(\mathcal{A})(1-Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)\left(Q(\mathcal{A})-(1-Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)\right)}{\left(Q(\mathcal{A})+\exp\left(-\frac{1}{\lambda}\right)(1-Q(\mathcal{A}))\right)^3}. \tag{75}$$

Note that $K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)>0$ *if and only if*

$$Q(\mathcal{A})-(1-Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)>0. \tag{76}$$

Assume that $Q(\mathcal{A})\geqslant\frac{1}{2}$*. Thus, it holds that for all* $\lambda>0$*, the inequality in* (76) *is always satisfied. This follows from observing that for all* $\lambda>0$*,*

$$\exp\left(-\frac{1}{\lambda}\right)<1\leqslant\frac{Q(\mathcal{A})}{1-Q(\mathcal{A})}. \tag{77}$$

Hence, if $Q(\mathcal{A}) \geqslant \frac{1}{2}$*, for all decreasing sequences of positive reals* $\lambda_1 > \lambda_2 > \ldots > 0$*, it holds that*

$$\frac{1}{4} \geqslant K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) > K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) > \ldots > 0. \tag{78}$$

Alternatively, assume that $Q(\mathcal{A}) < \frac{1}{2}$*. In this case, the inequality in* (76) *is satisfied if and only if*

$$\lambda < \left(\log\left(\frac{1-Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1}. \tag{79}$$

Hence, if $Q(\mathcal{A}) < \frac{1}{2}$*, then for all decreasing sequences of positive reals*

$$\left(\log\left(\frac{1-Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1} > \lambda_1 > \lambda_2 > \ldots > 0,$$

it holds that

$$\frac{1}{4} > K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) > K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) > \ldots > 0. \tag{80}$$

Moreover, for all decreasing sequences of positive reals

$$\lambda_1 > \lambda_2 > \ldots > \left(\log\left(\frac{1-Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1},$$

it holds that

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) < K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) < \ldots < \frac{1}{4}. \tag{81}$$

The upperbound by $\frac{1}{4}$ *in* (78)*,* (80) *and* (81) *follows by noticing that the value* $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ *is maximized when* $\lambda = \left(\log\left(\frac{1-Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1}$ *and* $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{1}{4}$*.*

Example 6.1 provides important insights on the choice of the reference measure Q. Note for instance that when the reference measure assigns a probability to the set of models $\mathcal{T}(\boldsymbol{z})$ that is greater than or equal to the probability of suboptimal models $\mathcal{M} \setminus \mathcal{T}(\boldsymbol{z})$, i.e., $Q(\mathcal{T}(\boldsymbol{z})) \geqslant \frac{1}{2}$, the variance is strictly decreasing to zero when λ decreases. See for instance, Figure 1 and Figure 2. That is, when the reference measure assigns higher probability to the set of solutions to the ERM problem in (6), the variance is monotone with respect to the parameter λ.

Alternatively, when the reference measure assigns a probability to the set $\mathcal{T}(\boldsymbol{z})$ that is smaller than the probability of the set $\mathcal{M} \setminus \mathcal{T}(\boldsymbol{z})$, i.e., $Q(\mathcal{T}(\boldsymbol{z})) < \frac{1}{2}$, there exists a critical point for λ at $\left(\log\left(\frac{1-Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1}$. See for instance, Figure 3. More importantly, such a critical point can be arbitrarily close to zero depending on the value $Q(\mathcal{A})$. The variance strictly decreases when λ

decreases beyond the value $\left(\log\left(\frac{1-Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1}$. Otherwise, reducing λ above the value $\left(\log\left(\frac{1-Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1}$ increases the variance.

In general, these observations suggest that reference measures Q that allocate small measures to the sets containing the set $\mathcal{T}(\boldsymbol{z})$ might require reducing the value λ beyond a small threshold in order to observe small values of $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$, which is the variance of the random variable W, in (67). These observations are central to understanding the concentration of probability that occurs when λ decreases, as discussed in the following section.

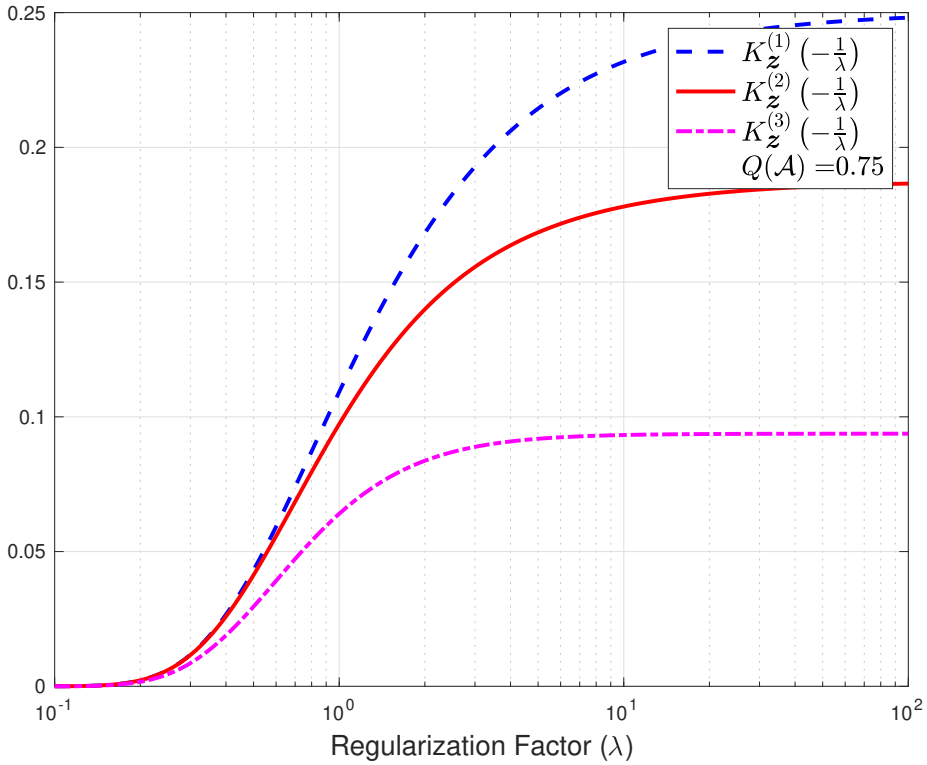


Figure 1: Mean $K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$, variance $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$, and third central moment $K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ of the empirical risk in Example 6.1, with $Q(\mathcal{A}) = \frac{3}{4}$

7 Concentration of Probability

Given a positive real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), consider the following set,

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}, \tag{82}$$

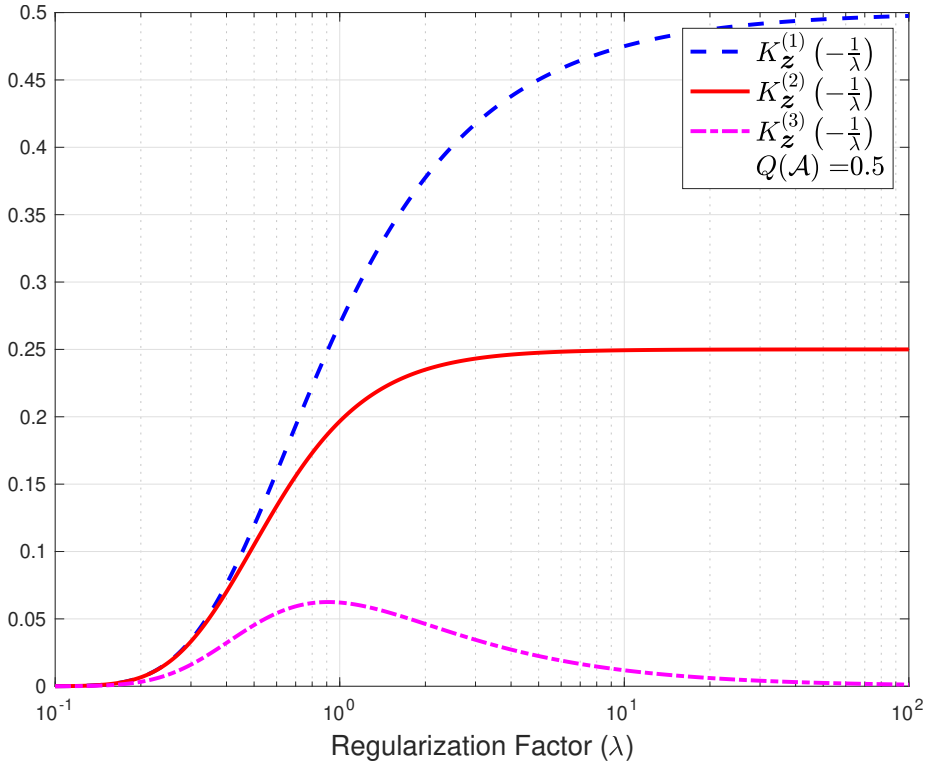


Figure 2: Mean $K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$, variance $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$, and third central moment $K_{Q,\boldsymbol{z}}^{(3)}\left(-\frac{1}{\lambda}\right)$ of the empirical risk in Example 6.1, with $Q\left(\mathcal{A}\right) = \frac{1}{2}$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined by (5); the function $\mathsf{R}_{\boldsymbol{z}}$ is defined by (10); and the probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ is in (25). This section introduces two main results. First, in Theorem 7.1, it is shown that when λ tends to zero, the set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ forms a monotonic sequence of sets that decreases to the set

$$\mathcal{N}_{Q,\boldsymbol{z}}^{\star} \triangleq \mathcal{L}_{\boldsymbol{z}}\left(\delta_{Q,\boldsymbol{z}}^{\star}\right), \tag{83}$$

where, $\delta_{Q,\boldsymbol{z}}^{\star}$ is defined in (32); and the set $\mathcal{L}_{\boldsymbol{z}}(\cdot)$ is defined in (31). Second, in Theorem 7.2, it is shown that the sequence formed by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda))$ when λ tends to zero is increasing and monotone. More importantly, in Theorem 7.3, it is shown that the limit of such sequence is equal to one. These observations justify referring to the set $\mathcal{N}_{Q,\boldsymbol{z}}^{\star}$ as the *limit set*. This section ends by showing that the probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ concentrates on a specific subset $\mathcal{L}_{Q,\boldsymbol{z}}^{\star}$ in (33) of the set $\mathcal{N}_{Q,\boldsymbol{z}}^{\star}$. At the light of this observation, the set $\mathcal{L}_{Q,\boldsymbol{z}}^{\star}$ is referred to as the *nonnegligible limit set*. Finally, it is shown that when the σ-finite measure Q in (25) is coherent, the sets $\mathcal{N}_{Q,\boldsymbol{z}}^{\star}$ and $\mathcal{L}_{Q,\boldsymbol{z}}^{\star}$ are identical.

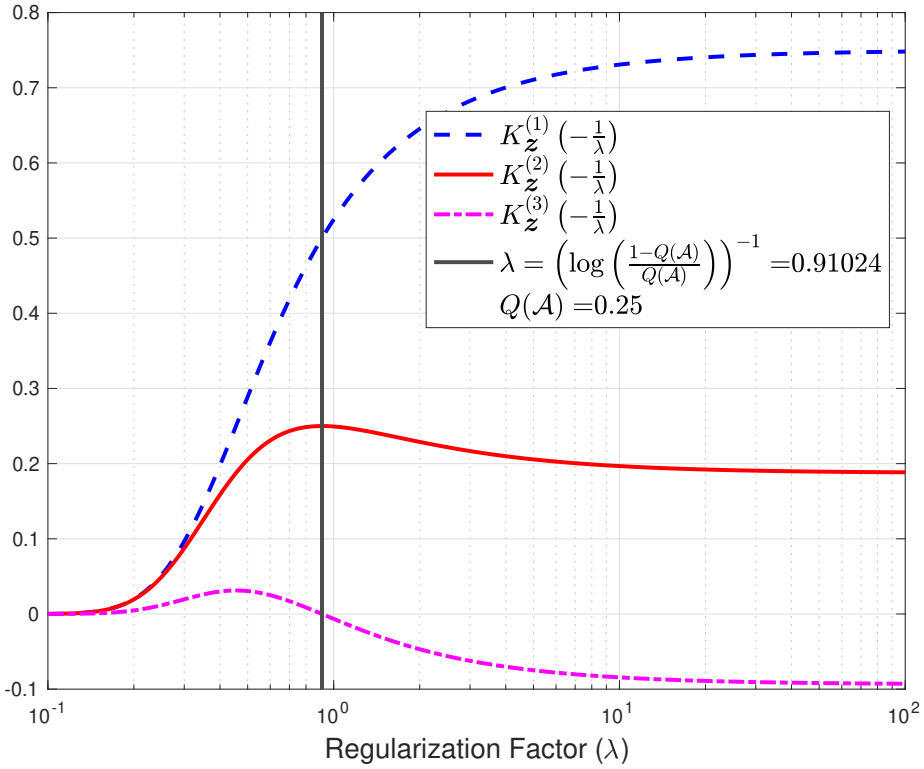


Figure 3: Mean $K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$, variance $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$, and third central moment $K_{Q,\boldsymbol{z}}^{(3)}\left(-\frac{1}{\lambda}\right)$ of the empirical risk in Example 6.1, with $Q\left(\mathcal{A}\right) = \frac{1}{4}$

7.1 The Limit Set

The set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ in (82), with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), contains all the models that induce an empirical risk that is smaller than or equal to $\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)$, i.e., the g-ERM-RER-optimal expected empirical risk in (68). This observation unveils the existence of a relation between the set $\mathcal{N}_{Q,\boldsymbol{z}}^{\star}$ in (83) and the set $\mathcal{T}\left(\boldsymbol{z}\right)$ in (7), as shown by the following lemma.

Lemma 7.1. *The set $\mathcal{N}_{Q,\boldsymbol{z}}^{\star}$ in (83) satisfies*

$$\mathcal{T}\left(\boldsymbol{z}\right) \subseteq \mathcal{N}_{Q,\boldsymbol{z}}^{\star}, \tag{84}$$

where the set $\mathcal{T}\left(\boldsymbol{z}\right)$ is in (7). Moreover,

$$\mathcal{T}\left(\boldsymbol{z}\right) = \mathcal{N}_{Q,\boldsymbol{z}}^{\star}, \tag{85}$$

if and only if the reference measure Q in (25) is coherent.

Proof: The proof of the inclusion in (84) follows from observing that for all $\boldsymbol{\theta} \in$

$\mathcal{T}(\boldsymbol{z})$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = 0 \leqslant \delta^{\star}_{Q,\boldsymbol{z}}$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32). Hence, $\boldsymbol{\theta} \in \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$. This completes the proof of the inclusion in (84).

The proof of the second statement is presented in two parts. In the first part, it is proved that if (85) holds, then the measure Q is coherent. The second part proves the converse.

The proof of the first part is as follows. Under the assumption that $\mathcal{T}(\boldsymbol{z}) = \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ holds, it follows that $\mathcal{T}(\boldsymbol{z}) \supseteq \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$, which implies that $\delta^{\star}_{Q,\boldsymbol{z}} = 0$, and thus, for all $\delta \in (0, +\infty)$, it holds that $Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0$. This verifies that the measure Q is coherent and completes the proof of the first part.

The proof of the second part is as follows. Under the assumption that the measure Q is coherent, it follows that $\delta^{\star}_{Q,\boldsymbol{z}} = 0$. Hence, $\mathcal{T}(\boldsymbol{z}) \supseteq \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$, which together with the inclusion in (84) leads to the equality in (85), which completes the proof of the second part. ■

Lemma 7.1 shows that the limit set $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ in (83) is not empty. This follows from the fact that the set $\mathcal{T}(\boldsymbol{z})$, which is not empty, is a subset of $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$. This observation turns out to be particularly important at the light of the fact that when λ decreases, the set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ decreases to the set $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$. This observation is formalized by following theorem.

Theorem 7.1. *For all $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (22) *and $\lambda_1 > \lambda_2$, the sets $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)$ and $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$ in* (82) *satisfy*

$$\mathcal{M} \supseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \supseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \supseteq \mathcal{N}^{\star}_{Q,\boldsymbol{z}}, \tag{86}$$

with $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ the set defined in (83)*. Moreover, if the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in* (5) *is continuous on $\mathcal{M}$ and separable with respect to the measure Q in* (11)*, then,*

$$\mathcal{M} \supset \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \supset \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \supset \mathcal{N}^{\star}_{Q,\boldsymbol{z}}. \tag{87}$$

Proof: The proof is presented in Appendix Q. ■

An interesting observation is that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), only a subset of $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ might exhibit nonzero probability with respect to the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25). Consider for instance that the measure Q in (25) is non-coherent (Definition 3.1). That is, $\delta^{\star}_{Q,\boldsymbol{z}} > 0$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32), and thus, for all $\gamma \in \left[0, \delta^{\star}_{Q,\boldsymbol{z}}\right)$, it holds that $Q(\mathcal{L}_{\boldsymbol{z}}(\gamma)) = 0$, with the set $\mathcal{L}_{\boldsymbol{z}}(\cdot)$ in (31). Then, for all $\lambda \in \left\{\alpha \in \mathcal{K}_{Q,\boldsymbol{z}} : \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) > \delta^{\star}_{Q,\boldsymbol{z}}\right\}$ and for all $\gamma < \delta^{\star}_{Q,\boldsymbol{z}}$, it holds that $\mathcal{L}_{\boldsymbol{z}}(\gamma) \subseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$, while verifying that $Q(\mathcal{L}_{\boldsymbol{z}}(\gamma)) = 0$, which implies that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\gamma)) = 0$ (Lemma 3.8). Hence, in this case, the set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ possesses a subset $\mathcal{L}_{\boldsymbol{z}}(\gamma)$ that is negligible with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. These observations lead to the analysis of the asymptotic concentration of probability in the following section.

7.2 The Nonnegligible Limit Set

The first step in the analysis of the asymptotic concentration of the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) is to show that the probability $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)\right)$ increases when λ tends to zero, as shown by the following theorem.

Theorem 7.2. *For all $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22) and $\lambda_1 > \lambda_2$, assume that the measures $P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfy (25) with $\lambda = \lambda_1$ and $\lambda = \lambda_2$, respectively. Then, the set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$ in (82) satisfies*

$$0 < P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) \leqslant P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)). \tag{88}$$

Moreover, the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to the σ-finite measure Q, with Q in (25), if and only if for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\lambda_1 > \lambda_2$, it holds that

$$0 < P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)). \tag{89}$$

Proof: The proof is presented in Appendix R. ∎

The following lemma highlights a case in which a stronger concentration of probability is observed.

Lemma 7.2. *Let the function $\mathsf{L}_{\boldsymbol{z}}$ in (5) be separable with respect to the σ-finite measure Q in (25), and consider two positive reals $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22) and $\lambda_1 > \lambda_2$. Assume that*

$$Q\Big(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \cap \left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)\right)^{\mathsf{c}}\Big) = 0. \tag{90}$$

Then, two measures $P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that respectively satisfy (25) with $\lambda = \lambda_1$ and $\lambda = \lambda_2$ verify that

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{91}$$

where, the set $\mathcal{N}_{Q,\boldsymbol{z}}(\cdot)$ is defined in (82).

Proof: The proof is presented in Appendix S. ∎

The following example shows the relevance of Lemma 7.2 in the case in which the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is a simple function and separable with respect to the σ-finite measure Q in (25).

Example 7.1. *Consider Example 6.1. Note that, for all $\lambda > 0$,*

$$0 < \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < 1, \tag{92}$$

where $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is the g-ERM-RER-optimal expected empirical risk in (68). The equality in (92) implies that given two reals λ_1 and λ_2 such that $\lambda_1 > \lambda_2 >$

0*, it holds that,*

$$\begin{aligned}\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)\cap(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2))^{\mathsf{c}}=&\left\{\boldsymbol{\nu}\in\mathcal{M}:\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)<\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\leqslant\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}\\=&\emptyset,\end{aligned}\tag{93}$$

and moreover, $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)=\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$*. Finally, from Lemma 7.2,*

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1))<P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)).\tag{94}$$

Finally, the main result of this section is presented by the following theorem.

Theorem 7.3. *The probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (25)*, with* Q *being a* σ*-finite measure on* $(\mathcal{M},\mathscr{B}(\mathcal{M}))$*, satisfies*

$$\lim_{\lambda\to0^+}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)\right)=1,\tag{95}$$

where, the set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ *is defined in* (82)*.*

Proof: The proof is presented in Appendix T. ∎

Note that Theorem 7.3 and Lemma 3.7 lead to the following conclusion

$$\lim_{\lambda\to0^+}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)\setminus\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)=0,\tag{96}$$

which follows from the fact that $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\subset\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (33). This justifies referring to the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ as the nonnegligible limit set.

8 Sub-Gaussianity of the Empirical Risk

Let λ be a real in $\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), and consider the transport of the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) from $(\mathcal{M},\mathscr{B}(\mathcal{M}))$ to $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ through the function $\mathsf{L}_{\boldsymbol{z}}$ in (5). Denote the resulting probability measure in $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ by $P^{(Q,\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$ in $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. That is, for all $\mathcal{A}\in\mathscr{B}(\mathbb{R})$,

$$P^{(Q,\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})=P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathsf{L}^{-1}_{\boldsymbol{z}}(\mathcal{A})\right),\tag{97}$$

where the term $\mathsf{L}^{-1}_{\boldsymbol{z}}(\mathcal{A})$ represents the set

$$\mathsf{L}^{-1}_{\boldsymbol{z}}(\mathcal{A})\triangleq\{\boldsymbol{\nu}\in\mathcal{M}:\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\in\mathcal{A}\}.\tag{98}$$

Note that the random variable W in (67) induces the probability measure $P^{(Q,\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$ in $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. The objective of this section is to prove that the random variable W is a sub-Gaussian random variable. For this purpose, note that the cumulant generating function induced by the measure $P^{(Q,\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$, denoted by $J_{\boldsymbol{z},Q,\lambda}$:

$\mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, satisfies for all $t \in \mathbb{R}$,

$$J_{\boldsymbol{z},Q,\lambda}(t) = \log\left(\int \exp\left(tw\right) \mathrm{d}P^{(Q,\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}(w)\right) \tag{99}$$

$$= \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{u})\right). \tag{100}$$

For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the following lemma provides an expression for $J_{\boldsymbol{z},Q,\lambda}(t)$ in terms of the cumulant generating function $K_{Q,\boldsymbol{z}}$ in (21), for all $t \in (-\infty, \frac{1}{\lambda})$.

Lemma 8.1. *Given a real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (22)*, the cumulant generating function $J_{\boldsymbol{z},Q,\lambda}$ in* (100)*, verifies the following equality for all $t \in \left(-\infty, \frac{1}{\lambda}\right)$,*

$$J_{\boldsymbol{z},Q,\lambda}(t) = K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty, \tag{101}$$

with the function $K_{Q,\boldsymbol{z}}$ in (21)*.*

Proof: The proof is presented in Appendix U. ∎

Denote by $J^{(m)}_{\boldsymbol{z},Q,\lambda} : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, with $m \in \mathbb{N}$, the m-th derivative of the function $J_{\boldsymbol{z},Q,\lambda}$ in (100). That is, for all $s \in \left(-\infty, \frac{1}{\lambda}\right)$,

$$J^{(m)}_{\boldsymbol{z},Q,\lambda}(s) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} J_{\boldsymbol{z},Q,\lambda}(t)\Big|_{t=s}. \tag{102}$$

From Lemma 8.1, it follows that for all $m \in \mathbb{N}$, and for all $\alpha \in \left(-\infty, \frac{1}{\lambda}\right)$, the following holds,

$$J^{(m)}_{\boldsymbol{z},Q,\lambda}(\alpha) = K^{(m)}_{Q,\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right), \tag{103}$$

where the function $K^{(m)}_{Q,\boldsymbol{z}}$ denotes the m-th derivative of the function $K_{Q,\boldsymbol{z}}$ in (21). See for instance, Lemma 4.4. The equality in (103) establishes a relation between the cumulant generating function $J_{\boldsymbol{z},Q,\lambda}$ and the function $K_{Q,\boldsymbol{z}}$. The following lemma leverages these observations and presents the main result of this section.

Theorem 8.1. *The cumulant generating function $J_{\boldsymbol{z},Q,\lambda}$ in* (100) *verifies the following inequality for all $\alpha \in \left(-\infty, \frac{1}{\lambda}\right)$,*

$$J_{\boldsymbol{z},Q,\lambda}(\alpha) \leqslant \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{2}\alpha^2 B^2_{\boldsymbol{z}}, \tag{104}$$

wheret the constant $B_{\boldsymbol{z}} > 0$ satisfies

$$B^2_{\boldsymbol{z}} = \sup_{\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}} K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right), \tag{105}$$

with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22)*; and the functions $K^{(1)}_{Q,\boldsymbol{z}}$ and $K^{(2)}_{Q,\boldsymbol{z}}$ are respectively defined in* (64) *and* (65)*.*

Proof: The function $K_{Q,\boldsymbol{z}}$ in (21) is differentiable infinitely many times over the interior of the set $\mathcal{K}_{Q,\boldsymbol{z}}$ (Lemma 4.1). Thus, from the Taylor-Lagrange theorem, c.f., [29, Theorem 2.5.4], it follows that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and for all $\alpha \in \left(-\infty, \frac{1}{\lambda}\right)$, there exists a real $\xi \in \left(-\infty, \frac{1}{\lambda}\right)$ such that

$$K_{Q,\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right) = K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{\alpha^2}{2} K^{(2)}_{Q,\boldsymbol{z}}\left(\xi\right). \tag{106}$$

From (106) and Lemma 8.1, it holds that

$$J_{\boldsymbol{z},Q,\lambda}(\alpha) = \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{\alpha^2}{2} K^{(2)}_{Q,\boldsymbol{z}}\left(\xi\right). \tag{107}$$

Finally, the inequality in (104) follows from the maximization of the function $K^{(2)}_{Q,\boldsymbol{z}}$ on the set $\mathcal{K}_{Q,\boldsymbol{z}}$, which completes the proof. ■

The main implication of Theorem 8.1 is that the random variable W in (67) is a sub-Gaussian random variable with parameter $B_{\boldsymbol{z}}$ in (105).

9 (δ, ϵ)-Optimality

This section introduces a PAC guarantee of optimality for the models that are sampled from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) with respect to the ERM problem in (6).

Definition 9.1 ((δ, ϵ)-Optimality)**.** *Given a pair of positive reals (δ, ϵ), with $\epsilon < 1$, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) is said to be (δ, ϵ)-optimal, if the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ in (31) satisfies*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}(\delta)\right) > 1 - \epsilon. \tag{108}$$

For all $\delta > 0$, it holds that $\mathcal{T}(\boldsymbol{z}) \subset \mathcal{L}_{\boldsymbol{z}}(\delta)$, with the sets $\mathcal{T}(\boldsymbol{z})$ and $\mathcal{L}_{\boldsymbol{z}}$ in (7) and (31), respectively. Hence, from Definition 9.1, it follows that if the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) is (δ, ϵ)-optimal, then it assigns a probability that is always greater than $1 - \epsilon$ to the set that contains models that induce an empirical risk that is smaller than δ. From this perspective, particular interest is given to the smallest δ and ϵ for which $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ, ϵ)-optimal.

The main result of this section is presented by the following theorem.

Theorem 9.1. *For all $(\delta, \epsilon) \in (\delta^{\star}_{Q,\boldsymbol{z}}, +\infty) \times (0, 1)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32), there exists a real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), such that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ, ϵ)-optimal.*

Proof of Theorem 9.1: Let δ be a real in $\left(\delta^{\star}_{Q,\boldsymbol{z}}, +\infty\right)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32). Let also $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ satisfy the following equality:

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \leqslant \delta. \tag{109}$$

Note that from Lemma 4.1, it follows that the function $K^{(1)}_{Q,\boldsymbol{z}}$ is continuous. Moreover, from Theorem 5.2, it follows that such a λ in (109) always exists. From (31) and (82), it holds that

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \subseteq \mathcal{L}_{\boldsymbol{z}}\left(\delta\right), \tag{110}$$

and thus,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) \geqslant P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)\right). \tag{111}$$

Let γ be a positive real such that $\gamma \leqslant \lambda$ and

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) > 1-\epsilon. \tag{112}$$

The existence of such a positive real γ follows from Theorem 7.3. Hence, from (112), it holds that,

$$1-\epsilon < P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) \tag{113}$$

$$\leqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right), \tag{114}$$

where the inequality in (114) follows from the fact that $\mathcal{N}_{Q,\boldsymbol{z}}(\gamma) \subseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \subseteq \mathcal{L}_{\boldsymbol{z}}\left(\delta\right)$. Finally, the inequality in (114) implies that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ,ϵ)-optimal (Definition 9.1). This completes the proof. ■

A stronger optimality claim can be stated when the reference measure is coherent.

Theorem 9.2. *For all $(\delta,\epsilon) \in (0,+\infty)\times(0,1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), such that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ,ϵ)-optimal if and only if the reference measure Q is coherent.*

Proof of Theorem 9.2: The proof is divided into two parts. The first part shows that if for all $(\delta,\epsilon) \in (0,+\infty)\times(0,1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), such that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) is (δ,ϵ)-optimal, then, the measure Q is coherent. The second part deals with the converse.

The first part is as follows. If for all $(\delta,\epsilon) \in (0,+\infty)\times(0,1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, such that

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 1-\epsilon, \tag{115}$$

then, it follows from Lemma 3.8 that for all $\delta > 0$,

$$Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0, \tag{116}$$

which implies that the measure Q is coherent. This completes the first part of the proof.

The second part of the proof is as follows. Under the assumption that the measure Q is coherent, it follows that $\delta^{\star}_{Q,\boldsymbol{z}} = 0$. Then, from Theorem 9.1, it follows that for all $(\delta,\epsilon) \in (0,+\infty)\times(0,1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), such that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ,ϵ)-optimal. This completes the second part of the proof. ■

10 Sensitivity

This section studies the sensitivity of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}$ (Definition 2.4) to deviations from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) towards an alternative probability measure P. Deviations from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ towards an alternative probability measure P over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ might arise due to several reasons. For instance, if new datasets become available, a new g-ERM-RER problem can be formulated using a larger dataset obtained by aggregating the old and the new datasets [30]. Similarly, the parameters Q (the reference measure) and λ (the regularization factor) in (11) might be changed based on side-information leading to new g-ERM-RER problems and thus, to new probability measures. Other techniques different from g-ERM-RER might also be used to obtain a probability measure over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, e.g., Bayesian methods. Within this context, the sensitivity is a performance metric defined as follows.

Definition 10.1 (Sensitivity)**.** *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a positive real $\lambda > 0$, let $\mathsf{S}_{Q,\lambda} : (\mathcal{X} \times \mathcal{Y})^n \times \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to (-\infty, +\infty]$ be a function such that for all datasets $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ and for probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that*

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise,} \end{cases} \tag{117}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is defined in (10) *and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is in* (25)*. The sensitivity of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}$ due to a deviation from $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to P is $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$.*

10.1 Dataset-Dependent Bounds

In the aim of characterizing the sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$ in (117), consider the following lemma.

Lemma 10.1. *Given two probability measures P and Q over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, with Q absolutely continuous with P, the following holds for all $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$,*

$$\begin{aligned} &\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}(P) \\ &\leqslant \inf_{t \in (-\infty, 0)} \left(\frac{D(Q\|P) + \log\left(\int \exp\left(t\ (\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{R}_{\boldsymbol{z}}(P))\right) \mathrm{d}P(\boldsymbol{\theta})\right)}{t} \right), \end{aligned} \tag{118}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $\mathsf{R}_{\boldsymbol{z}}$ is defined in (5) *and in* (10)*, respectively.*

Proof: The proof is presented in Appendix V. ∎

Lemma 10.1 together with Theorem 8.1 lead to an upper bound on the absolute value of the sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$ in (117).

Theorem 10.1. *The function* $\mathsf{S}_{Q,\lambda}$ *in* (117) *satisfies for all probability measures* $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$,

$$|\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)| \leqslant \sqrt{2B_{Q,\boldsymbol{z}}^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)}, \tag{119}$$

where the constant $B_{Q,\boldsymbol{z}}$ *is defined in* (105).

Proof: The proof is presented in Appendix W. ∎

Theorem 10.1 establishes an upper and a lower bound on the increase and decrease of the expected empirical risk that can be obtained by deviating from the optimal solution of the g-ERM-RER in (11). More specifically, note that for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that,

$$\mathsf{R}_{\boldsymbol{z}}(P) \geqslant \mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) - \sqrt{2B_{\boldsymbol{z}}^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)} \text{ and} \tag{120a}$$

$$\mathsf{R}_{\boldsymbol{z}}(P) \leqslant \mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) + \sqrt{2B_{\boldsymbol{z}}^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)}. \tag{120b}$$

The following theorem highlights the fact that the measure that minimizes the expected empirical risk subject to a constraint in the relative entropy with respect to the g-ERM-RER-optimal measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ in (25) is also the solution to an g-ERM-RER problem with parameters Q and ω, for some specific $\omega > 0$.

Theorem 10.2. *Given a* σ*-finite measure* $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$*, a dataset* $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$*, and a nonnegative real* $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$*, with* $\mathcal{K}_{Q,\boldsymbol{z}}$ *in* (22)*, consider the following optimization problem*

$$\min_{P \in \triangle_{P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}}(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \quad \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P(\boldsymbol{\theta}), \tag{121a}$$

$$\text{subject to:} \quad D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) \leqslant c, \text{ and} \tag{121b}$$

$$\int \mathrm{d}P(\boldsymbol{\theta}) = 1, \tag{121c}$$

with, c *a given nonnegative constant;* $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ *the probability measure in* (25)*; and* $\mathsf{L}_{\boldsymbol{z}}$ *the function in* (5)*. Then, the solution to the optimization problem in* (121) *is a probability measure* $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\omega)}$ *satisfying for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\omega)}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\omega}\right) - \frac{1}{\omega}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right), \tag{122}$$

with $\omega \in (0, \lambda]$ *such that*

$$D\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\omega)} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) = c. \tag{123}$$

Proof: The proof is presented in Appendix X. ∎

10.2 Dataset-Independent Bounds

Consider a probability measure, denoted by $P_{\boldsymbol{Z}} \in \triangle\Big((\mathcal{X} \times \mathcal{Y})^n, (\mathscr{F}(\mathcal{X} \times \mathcal{Y}))^n \Big)$, such that for all $\mathcal{A} \in \big(\mathscr{F}(\mathcal{X} \times \mathcal{Y})\big)^n$ of the form $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n$ with $\mathcal{A}_i \in \mathscr{F}(\mathcal{X} \times \mathcal{Y})$ and $i \in \{1, 2, \ldots, n\}$, it holds that

$$P_{\boldsymbol{Z}}(\mathcal{A}) = \prod_{t=1}^{n} P_{XY}(\mathcal{A}_t), \tag{124}$$

where the probability measure P_{XY} is defined in (2). More specifically, $P_{\boldsymbol{Z}}(\mathcal{A})$ is the probability measure induced by a random variable

$$\boldsymbol{Z} = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)),$$

in which the n random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are independent and identically distributed according to P_{XY}.

Let the set $\mathcal{K}_Q$, with $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, be

$$\mathcal{K}_Q = \bigcap_{\boldsymbol{z} \in \operatorname{supp} P_{\boldsymbol{Z}}} \mathcal{K}_{Q,\boldsymbol{z}}, \tag{125}$$

where the set $\mathcal{K}_{Q,\boldsymbol{z}}$ is defined in (21) and the probability measure $P_{\boldsymbol{Z}}$ is defined in (124). The set $\mathcal{K}_Q$ in (125) can be empty for some choices of the σ-finite measure Q and empirical loss function $\mathsf{L}_{\boldsymbol{z}}$ in (5). Nonetheless, from Lemma 3.2, it follows that when Q is a probability measure, then,

$$\mathcal{K}_Q = (0, +\infty). \tag{126}$$

Using this notation, the following corollary of Theorem 10.1 provides an upper bound on the expectation of the sensitivity with respect to the probability measure $P_{\boldsymbol{Z}}$ in (124).

Corollary 10.1. *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, for all $\lambda \in \mathcal{K}_Q$, with $\mathcal{K}_Q$ in* (125)*, and for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that*

$$\int |\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)| \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \leqslant \int \sqrt{2B_{Q,\boldsymbol{z}}^2 D\left(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{127}$$

where $B_{Q,\boldsymbol{z}}$ is defined in (105)*; the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is in* (25)*; and the probability measure $P_{\boldsymbol{Z}}$ is defined in* (124)*.*

In the following theorem, the expectation of the sensitivity with respect to the measure $P_{\boldsymbol{Z}}$ in (124) is shown to have an upper bound that can be expressed in terms of the lautum information between the models and the data sets.

Theorem 10.3. *Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, for all $\lambda \in \mathcal{K}_Q$, with $\mathcal{K}_Q$ in* (125)*, it holds that*

$$\int \left| \mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right) \right| \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \leqslant \sqrt{2B_Q^2 \int D\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}^{(Q,\lambda)}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u})}, \tag{128}$$

where the probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ is the solution to the g-ERM-RER problem in (11)*; the probability measure $P_{\boldsymbol{Z}}$ is defined in* (124)*; the probability measure $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ is such that for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$,*

$$P_{\boldsymbol{\Theta}}^{(Q,\lambda)}(\mathcal{A}) = \int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\mathcal{A}) \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}); \tag{129}$$

and the constant B_Q satisfies

$$B_Q^2 = \sup_{\boldsymbol{z} \in \operatorname{supp} P_Z} B_{Q,\boldsymbol{z}}^2, \tag{130}$$

with $B_{Q,\boldsymbol{z}}$ defined in (105)*.*

Proof: The proof follows from Corollary 10.1. In particular, from (127), for all probability measures P over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ absolutely continuous with Q, it holds that

$$\int |\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)| \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \leqslant \int \sqrt{2B_{Q,\boldsymbol{z}}^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \tag{131}$$

$$\leqslant \int \sqrt{2B_Q^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \tag{132}$$

$$\leqslant \sqrt{2B_Q^2 \int D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})}, \tag{133}$$

where the inequality in (132) follows from (130); and the inequality in (133) follows from Jensen's inequality [28, Theorem 6.3.5]. This completes the proof. ■

Given a σ-finite measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and a positive real $\lambda \in \mathcal{K}_Q$, with $\mathcal{K}_Q$ in (125), let $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$ be the random variables that jointly induce a probability measure $P_{\boldsymbol{Z}\boldsymbol{\Theta}}^{(Q,\lambda)}$ with marginals $P_{\boldsymbol{Z}}$ in (124) and $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ in (129). Under these assumptions, the right-hand side in (128) can be written in terms of the lautum information [31] between the random variables $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$, which is denoted by $\mathtt{L}(\boldsymbol{Z}; \boldsymbol{\Theta})$. More specifically, note that

$$\mathtt{L}(\boldsymbol{Z}; \boldsymbol{\Theta}) = \int D\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}). \tag{134}$$

In a nutshell, it can be concluded that the expectation of $\left|\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right)\right|$ with respect to the measure $P_{\boldsymbol{Z}}$ in (124) is upper bounded up to a constant factor by the square root of the lautum information between the random variables $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$.

11 Discussion and Final Remarks

The classical ERM-RER problem in (11) has been studied under the assumption that the reference measure Q is a σ-finite measure, instead of a probability measure, which leads to a more general problem coined as the g-ERM-RER problem. In particular, it has been highlighted that the g-ERM-RER problem provides a larger flexibility for including prior knowledge on the models. Special cases of the g-ERM-RER problem include the ERM problem with (discrete or differential) entropy regularization and the information-risk minimization problem. The solution to the g-ERM-RER problem has been shown to exist and to be unique. Interestingly, the empirical risk observed when models are sampled from the g-ERM-RER-optimal probability measure is a sub-Gaussian random variable that exhibits a PAC guarantee for the ERM problem. That is, for some positive δ and ϵ, it is shown that there always exist some parameters for the g-ERM-RER-optimal measure such that the set of models that induce an empirical loss smaller than δ exhibit a probability that is not smaller that $1-\epsilon$. Interestingly, none of these results relies on statistical assumptions on the datasets. Finally, the sensitivity of the expected empirical risk to deviations from the g-ERM-RER-optimal measure has also been studied. In particular, an upper bound on the expectation of the sensitivity with respect to the dataset is presented. In some particular cases, connections between the sensitivity and the lautum information between the models and the data sets have been stablished.

Appendices

A Proof of Lemma 3.1

The proof is divided into two parts. The first part develops under the assumption that the set $\mathcal{K}_{Q,\boldsymbol{z}} \subset (0,+\infty)$ is not empty. The second part considers the opposite assumption.

The first part is as follows. Under the assumption that the set $\mathcal{K}_{Q,\boldsymbol{z}}$ is not empty, there always exists a real $b \in \mathcal{K}_{Q,\boldsymbol{z}}$, such that $K_{Q,\boldsymbol{z}}\left(-\frac{1}{b}\right) < +\infty$. Note that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) = \frac{1}{t^2} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \geqslant 0, \tag{135}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (5). Thus, from (21), it follows that $K_{Q,\boldsymbol{z}}\left(-\frac{1}{b}\right)$ is nondecreasing with b. This implies that $(0, b] \subset \mathcal{K}_{Q,\boldsymbol{z}}$. This proves the convexity of $\mathcal{K}_{Q,\boldsymbol{z}}$.

Let $b^\star \in (0,+\infty]$ be

$$b^\star = \sup \mathcal{K}_{Q,\boldsymbol{z}}. \tag{136}$$

Hence, if $b^\star = +\infty$, it follows from (22) that

$$\mathcal{K}_{Q,\boldsymbol{z}} = (0,+\infty). \tag{137}$$

Alternatively, if $b^\star < +\infty$, it holds that

$$(0, b^\star) \subset \mathcal{K}_{Q,\boldsymbol{z}}. \tag{138}$$

This completes the first part.

The second part is trivial. Under the assumption that the set $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22) is empty, there is nothing to prove.

This completes the proof.

B Proof of Lemma 3.2

Note that for all $\boldsymbol{\theta} \in \mathcal{M}$ and for all for all $t > 0$, it follows that

$$\exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \leqslant 1, \tag{139}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (5). Thus,

$$K_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right) = \log\left(\int \exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right) \tag{140}$$

$$\leqslant \log\left(\int \mathrm{d}Q(\boldsymbol{\theta})\right) \tag{141}$$

$$\leqslant 0, \tag{142}$$

which implies that $(0,+\infty) \subseteq \mathcal{K}_{Q,\boldsymbol{z}}$. Thus, from (22), it holds that $\mathcal{K}_{Q,\boldsymbol{z}} = (0,+\infty)$, which completes the proof.

C Proof of Theorem 3.1

The objective function in the optimization problem in (11) can be written as follows:

$$\min_{P\in\triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) + \lambda\int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{143a}$$

$$\text{s. t.} \int \mathrm{d}P(\boldsymbol{\theta}) = 1. \tag{143b}$$

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

Let $\mathscr{M}$ be the set of nonnegative measurable functions with respect to the measurable spaces $(\mathcal{M},\mathscr{B}(\mathcal{M}))$ and $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. The Lagrangian of the optimization problem in (143) can be constructed in terms of a function in $\mathscr{M}$, instead of a measure in $\triangle(\mathcal{M},\mathscr{B}(\mathcal{M}))$. Let such Lagrangian be $L:\mathscr{M}\times[0,+\infty)\to\mathbb{R}$ of the form

$$\begin{aligned} L\left(\frac{\mathrm{d}P}{\mathrm{d}Q},\beta\right) = &\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu})\,\mathrm{d}Q(\boldsymbol{\nu}) + \lambda\int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu})\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu}) \\ &+\beta\left(\int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu})\,\mathrm{d}P(\boldsymbol{\nu}) - 1\right), \end{aligned} \tag{144}$$

where β is a positive real that acts as a Lagrangian multiplier due to the constraint (143b).

Let $g:\mathbb{R}^k\to\mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (144) at $\left(\frac{\mathrm{d}P}{\mathrm{d}Q},\beta\right)\in\mathscr{M}\times[0,+\infty)$ in the direction of g is

$$\partial L\left(\frac{\mathrm{d}P}{\mathrm{d}Q},\beta;g\right) \triangleq \left.\frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha)\right|_{\alpha=0}, \tag{145}$$

where the real function $r:\mathbb{R}\to\mathbb{R}$ is such that for all $\alpha\in\mathbb{R}$,

$$\begin{aligned} r(\alpha) = &\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu}) + \alpha g(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu}) \\ &+\beta\left(\int\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu}) + \alpha g(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu}) - 1\right) \\ &+\lambda\int\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu}) + \alpha g(\boldsymbol{\nu})\right)\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu}) + \alpha g(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu}). \end{aligned} \tag{146}$$

Note that the derivative of the real function r in (146) is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha) = &\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,g(\boldsymbol{\nu})\,\mathrm{d}Q(\boldsymbol{\nu}) + \beta\int g(\boldsymbol{\nu})\,\mathrm{d}Q(\boldsymbol{\nu}) \\ &+\lambda\int g(\boldsymbol{\nu})\left(1+\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\nu}) + \alpha g(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu}). \end{aligned} \tag{147}$$

From (145) and (147), it follows that

$$\partial L\left(\frac{\mathrm{d}P}{\mathrm{d}Q},\beta;g\right)=\int g\left(\boldsymbol{\nu}\right)\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\lambda\left(1+\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)\right)\right)+\beta\right)\mathrm{d}Q\left(\boldsymbol{\nu}\right). \tag{148}$$

The relevance of the Gateaux differential in (148) stems from [32, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (144) to have a minimum at $\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q},\beta\right)\in\mathscr{M}\times[0,+\infty)$ is that for all functions $g\in\mathscr{M}$,

$$\partial L\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q};g\right)=0. \tag{149}$$

From (149), it follows that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ must satisfy for all functions g in $\mathscr{M}$ that

$$\int g\left(\boldsymbol{\nu}\right)\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\lambda\left(1+\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)\right)\right)+\beta\right)\mathrm{d}Q\left(\boldsymbol{\nu}\right)=0, \tag{150}$$

which implies that for all $\boldsymbol{\nu}\in\mathcal{M}$,

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\lambda\left(1+\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)\right)\right)+\beta=0, \tag{151}$$

and thus,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)=\exp\left(-\frac{\beta+\lambda}{\lambda}\right)\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)}{\lambda}\right), \tag{152}$$

with β chosen to satisfy (143b). That is,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)=\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\lambda}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right)} \tag{153}$$

$$=\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right). \tag{154}$$

The proof continues by verifying that the objective function in (143) is strictly convex, and thus, the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that satisfies (153) is the unique minimizer. More specifically, note that the objective function in (143) is the sum of two terms. The first one, i.e., $\displaystyle\int\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\frac{\mathrm{d}P}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)\mathrm{d}Q\left(\boldsymbol{\nu}\right)$, is linear in $\frac{\mathrm{d}P}{\mathrm{d}Q}$. The second, i.e., $\displaystyle\int\frac{\mathrm{d}P}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\left(\boldsymbol{\nu}\right)\right)\mathrm{d}Q\left(\boldsymbol{\nu}\right)$, is strictly convex with $\frac{\mathrm{d}P}{\mathrm{d}Q}$. Hence, given that $\lambda>0$, the sum of both terms is strictly convex with $\frac{\mathrm{d}P}{\mathrm{d}Q}$. This implies the uniqueness of $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ and completes the proof.

D Proof of Lemma 3.3

For all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \mathcal{T}(\boldsymbol{z}) \times \mathcal{T}(\boldsymbol{z})$, it follows that

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}), \tag{155}$$

and thus, for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), it holds that

$$\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \leqslant \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) = \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu})}{\lambda}\right), \tag{156}$$

which implies

$$\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\displaystyle\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\alpha})} \leqslant \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\displaystyle\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\alpha})} \tag{157}$$

$$= \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu})}{\lambda}\right)}{\displaystyle\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\alpha})}. \tag{158}$$

Hence, under the assumption that $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q \neq \emptyset$, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ and for all $(\boldsymbol{\mu}, \boldsymbol{\nu}) \in (\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q)^2$, it holds that

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leqslant \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\mu}) = \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}). \tag{159}$$

which completes the proof.

E Proof of Lemma 3.4

From Lemma 3.3, it follows that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, and for all $\boldsymbol{\mu} \in \mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$, it holds that

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leqslant \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\mu}) \tag{160}$$

$$= \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu})\right) \tag{161}$$

$$= \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) \tag{162}$$

$$< +\infty, \tag{163}$$

where the equality in (161) follows from (25); the equality in (162) follows from the fact that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = 0$; and the equality in (163) follows from the fact that $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$. This completes the proof of finiteness.

The proof of positivity follows from observing that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, it holds that $K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty$, and thus, $\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) > 0$. Moreover, for

all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant +\infty$, which implies that $-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geqslant -\infty$, and thus, $\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \geqslant 0$, with equality if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = +\infty$. These two observations put together yield

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \tag{164}$$

$$= \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \tag{165}$$

$$\geqslant 0, \tag{166}$$

with equality if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = +\infty$. This completes the proof.

F Proof of Lemma 3.6

From Theorem 3.1, it follows that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\displaystyle\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})} \tag{167}$$

$$= \left(\exp\left(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} \tag{168}$$

$$= \left(\int \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1}. \tag{169}$$

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

$$\mathcal{A}_0(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = 0\}, \tag{170a}$$

$$\mathcal{A}_1(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < 0\}, \text{ and} \tag{170b}$$

$$\mathcal{A}_2(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > 0\}. \tag{170c}$$

Using the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (169), the following holds for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \left(\int_{\mathcal{A}_0(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu}) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu}) + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} \tag{171}$$

$$= \left(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu}) + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1}. \tag{172}$$

Note that the sets

$$\left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{173}$$

$$\left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{174}$$

$$\left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{175}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32), form a partition of the set $\operatorname{supp} Q$. Following this observation, the rest of the proof is divided into three parts. The first part evaluates $\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$, 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 $\left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\} = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$. Hence, the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (170) satisfy the following:

$$\mathcal{A}_0(\boldsymbol{\theta}) = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}, \tag{176a}$$

$$\mathcal{A}_1(\boldsymbol{\theta}) = \left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{176b}$$

$$\mathcal{A}_2(\boldsymbol{\theta}) = \left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}. \tag{176c}$$

From the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32), it follows that $Q(\mathcal{A}_2(\boldsymbol{\theta})) = 0$. Plugging the equalities in (176) in (172) yields for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \left(Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} \tag{177}$$

The equality in (177) implies that for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}$,

$$\begin{aligned}\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \Bigg(Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) \\ &\quad + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1} \qquad (178)\\ &= \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} \qquad (179)\end{aligned}$$

where the equality in (179) follows from verifying that the dominated convergence theorem [28, Theorem 2.6.9] holds. That is,
(a) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that $\exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) < 1$; and
(b) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that

$$\lim_{\lambda\to 0^+} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) = 0. \tag{180}$$

This completes the first part of the proof.

The second part is as follows. For all $\delta > \delta^{\star}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) = \delta\right\}$, the sets $\mathcal{A}_0\left(\boldsymbol{\theta}\right)$, $\mathcal{A}_1\left(\boldsymbol{\theta}\right)$, and $\mathcal{A}_2\left(\boldsymbol{\theta}\right)$ in (170) satisfy the following:

$$\mathcal{A}_0\left(\boldsymbol{\theta}\right)=\left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\mu}\right) = \delta\right\}, \tag{181a}$$

$$\mathcal{A}_1\left(\boldsymbol{\theta}\right)=\left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\mu}\right) > \delta\right\}, \text{ and} \tag{181b}$$

$$\mathcal{A}_2\left(\boldsymbol{\theta}\right)=\left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\mu}\right) < \delta\right\}. \tag{181c}$$

Consider the sets

$$\mathcal{A}_{2,1}\left(\boldsymbol{\theta}\right)\triangleq\left\{\boldsymbol{\mu} \in \mathcal{A}_2\left(\boldsymbol{\theta}\right) : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\mu}\right) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{182}$$

$$\mathcal{A}_{2,2}\left(\boldsymbol{\theta}\right)\triangleq\left\{\boldsymbol{\mu} \in \mathcal{A}_2\left(\boldsymbol{\theta}\right) : \delta^{\star}_{Q,\boldsymbol{z}} \leqslant \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\mu}\right) < \delta\right\}, \tag{183}$$

and note that $\mathcal{A}_{2,1}\left(\boldsymbol{\theta}\right)$ and $\mathcal{A}_{2,2}\left(\boldsymbol{\theta}\right)$ form a partition of $\mathcal{A}_2\left(\boldsymbol{\theta}\right)$. Moreover, from the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32), it holds that

$$Q\left(\mathcal{A}_{2,1}\left(\boldsymbol{\theta}\right)\right) = 0. \tag{184}$$

Hence, plugging the equalities in (181) and (184) in (172) yields, for all $\delta > \delta^{\star}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) = \delta\}$,

$$\begin{aligned}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)=&\Bigg(Q\left(\mathcal{A}_0\left(\boldsymbol{\theta}\right)\right) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right)\right) \mathrm{d}Q\left(\boldsymbol{\nu}\right)\\ &+ \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right)\right) \mathrm{d}Q\left(\boldsymbol{\nu}\right)\Bigg)^{-1}.\end{aligned} \tag{185}$$

The equality in (185) implies that for all $\delta > \delta^{\star}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M}\colon \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) = \delta\right\}$,

$$\begin{aligned}\lim_{\lambda\to 0^+}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right)=&\Bigg(Q\left(\mathcal{A}_0\left(\boldsymbol{\theta}\right)\right)\\ &+ \lim_{\lambda\to 0^+}\int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right)\right) \mathrm{d}Q\left(\boldsymbol{\nu}\right)\\ &+ \lim_{\lambda\to 0^+}\int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right)\right) \mathrm{d}Q\left(\boldsymbol{\nu}\right)\Bigg)^{-1} &(186)\\ =&\Bigg(Q\left(\mathcal{A}_0\left(\boldsymbol{\theta}\right)\right)\\ &+ \lim_{\lambda\to 0^+}\int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right)\right) \mathrm{d}Q\left(\boldsymbol{\nu}\right)\Bigg)^{-1} &(187)\\ =&\left(Q\left(\mathcal{A}_0\left(\boldsymbol{\theta}\right)\right) + \infty\right)^{-1} &(188)\\ =&0, &(189)\end{aligned}$$

where the equality in (187) follows by verifying that the dominated convergence theorem [28, Theorem 2.6.9] holds. That is,
(*a*) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that $\exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) < 1$; and
(*b*) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that

$$\lim_{\lambda \to 0^+} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) = 0. \tag{190}$$

This completes the second part.

The third part of the proof follows by noticing that the set $\left\{\boldsymbol{\nu} \in \operatorname{supp} Q \colon \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}$ is a negligible set 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 $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$ is immaterial. Hence, it is arbitrarily 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

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{191}$$

This completes the third part of the proof.

Finally, from (179), (189), and (191), it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\lim_{\lambda \to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} \mathbb{1}_{\left\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right\}}, \tag{192}$$

which completes the proof.

G Proof of Lemma 3.7

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

$$\mathcal{A}_0 \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{193a}$$
$$\mathcal{A}_1 \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{193b}$$
$$\mathcal{A}_2 \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{193c}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32) and the function $\mathsf{L}_{\boldsymbol{z}}$ in (5). Note that $\mathcal{A}_0 = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (33).

For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the following holds,

$$1 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{194}$$
$$= P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{195}$$
$$= P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_2} \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{196}$$

where, the equality in (195) follows from noticing that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) = 0$, which follows from the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (32) and the fact that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with the measure Q.

The above implies that

$$1 = \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \lim_{\lambda\to 0^+}\int_{\mathcal{A}_2} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{197}$$

$$= \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_2} \lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{198}$$

$$= \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0), \tag{199}$$

where, the equality in (198) follows from the dominated convergence theorem [28, Theorem 1.6.9], given that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, the Randon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is positive and finite (Lemma 3.4); and the inequality in (199) holds from the fact that for all $\boldsymbol{\theta} \in \mathcal{A}_2$, it holds that $\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0$ (Lemma 3.6). Hence, it finally holds that

$$\lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) = 1, \tag{200}$$

which completes the proof.

H Proof of Lemma 3.8

For all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), and for all $\mathcal{C} \in \mathscr{B}(\mathcal{M})$,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = \int_{\mathcal{C}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{201}$$

and thus, if $Q(\mathcal{C}) = 0$, then

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = 0, \tag{202}$$

which implies the absolute continuity of $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to Q.

Alternatively, given a set $\mathcal{C} \in \mathscr{B}(\mathcal{M})$ and a real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, assume now that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = 0$. Hence, it follows that

$$0 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) \tag{203}$$

$$= \int_{\mathcal{C}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{204}$$

From Lemma 3.4, and the assumption $Q\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = +\infty\}\right) = 0$, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) > 0, \tag{205}$$

which implies that

$$\int_{\mathcal{C}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right) = 0, \tag{206}$$

if and only if $Q\left(\mathcal{C}\right) = 0$. This verifies the absolute continuity of Q with respect to $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, and completes the proof.

I Proof of Lemma 3.9

The proof is presented in two parts. The first part shows that if for all $\delta \in (0, +\infty)$, the inequality in (46) holds, then, Q is coherent. The second part shows that if there exists a $\delta \in (0, +\infty)$ such that

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) = 0, \tag{207}$$

then Q is noncoherent.

The first part is as follows. Note that for all $\delta \in (0, +\infty)$ and for all $\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{z}}\left(\delta\right) \cap \operatorname{supp} Q$, it holds from Lemma 3.4 that

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) > 0. \tag{208}$$

Hence, if for all $\delta \in (0, +\infty)$, the inequality in (46) holds, then

$$0 < P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) \tag{209}$$

$$= \int_{\mathcal{L}_{\boldsymbol{z}}(\delta)} \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{210}$$

$$= \int_{\mathcal{L}_{\boldsymbol{z}}(\delta)} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right), \tag{211}$$

which, together with (208), implies that for all $\delta \in (0, +\infty)$, $Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0$. Hence, Q is coherent.

The second part is as follows. Assume that there exists a $\delta \in (0, +\infty)$, for which it holds that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) = 0$. Hence, the following holds for such δ,

$$0 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) \tag{212}$$

$$= \int_{\mathcal{L}_{\boldsymbol{z}}(\delta)} \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{213}$$

$$= \int_{\mathcal{L}_{\boldsymbol{z}}(\delta)} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right). \tag{214}$$

From Lemma 3.4, it holds that for all $\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{z}}(\delta) \cap \operatorname{supp} Q$, the inequality in (208) holds, which implies that $Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) = 0$. That is, Q is noncoherent. This completes the proof.

J Proof of Lemma 3.12

Consider the function $g : \mathcal{M} \to [0, +\infty)$,

$$g(\theta) = \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \left(\frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1}, \tag{215}$$

and note that from Lemma 3.4, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q \setminus \left\{ \boldsymbol{\nu} \in \mathcal{M} \colon \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = +\infty \right\}$, $g(\boldsymbol{\theta}) > 0$ and for all $\boldsymbol{\theta} \in \{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = +\infty \}$, $g(\boldsymbol{\theta}) = 0$, which follows from the assumption $0 \cdot \frac{1}{0} = 0$.

Consider a measure P on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, such that for all sets $\mathcal{A} \in \mathscr{B}(\mathcal{M})$,

$$P(\mathcal{A}) = \int_{\mathcal{A}} g(\boldsymbol{\theta}) \mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{216}$$

and note that if $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) = 0$, then $P(\mathcal{A}) = 0$. This implies that P is absolutely continuous with $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})$. Moreover, from (216), it follows that

$$\begin{aligned}
P(\mathcal{A}) &= \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \left(\frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (217)\\
&= \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \left(\frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) && (218)\\
&= \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) && (219)\\
&= \int_{\mathcal{A}} \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (220)\\
&= P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}). && (221)
\end{aligned}$$

This proves that $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. The proof that $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ follows the same argument. This completes the proof.

K Proof of Lemma 4.1

Consider the transport of the σ-finite measure Q in (21) from the measure space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ to the measure space $([0, +\infty), \mathscr{B}([0, +\infty)))$ through the

function $\mathsf{L}_{\boldsymbol{z}}$ in (5). Denote the resulting measure in $([0,+\infty),\mathscr{B}([0,+\infty)))$ by P. More specifically, for all $\mathcal{A}\in\mathscr{B}([0,+\infty))$, it holds that $P(\mathcal{A})=Q(\{\boldsymbol{\theta}\in\mathcal{M}:\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\in\mathcal{A}\})$. Hence, the function $K_{Q,\boldsymbol{z}}$ in (21) can be written for all $t\in\mathbb{R}$ in terms of the measure P as follows

$$K_{Q,\boldsymbol{z}}(t)=\log\left(\int\exp(t\,v)\,\mathrm{d}P(v)\right). \tag{222}$$

Denote by ϕ the Laplace transform of the measure P. That is, for all $t\in(0,+\infty)$,

$$\phi(t)=\int\exp(t\,v)\,\mathrm{d}P(v)=\exp\left(K_{Q,\boldsymbol{z}}(-t)\right). \tag{223}$$

From [33, Theorem 1a (page 439)], it follows that the function ϕ has derivatives of all orders in $(0,+\infty)$, and thus, so does the function $K_{Q,\boldsymbol{z}}$ in $(-\infty,0)$. This implies the continuity of $K_{Q,\boldsymbol{z}}$ in $(-\infty,0)$, and completes the proof.

L Proof of Lemma 4.3

Let $(\gamma_1,\gamma_2)\in\mathbb{R}^2$, with $\gamma_1\neq\gamma_2$ and $\alpha\in[0,1]$ be fixed. Assume that $K_{Q,\boldsymbol{z}}(\gamma_1)<+\infty$ and $K_{Q,\boldsymbol{z}}(\gamma_2)<+\infty$. Then, for all $\alpha\in(0,1)$, the following holds

$$\begin{aligned}
&\alpha K_{Q,\boldsymbol{z}}(\gamma_1)+(1-\alpha)K_{Q,\boldsymbol{z}}(\gamma_2)\\
&=\alpha\log\left(\int\exp(\gamma_1\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right)\\
&\quad+(1-\alpha)\log\left(\int\exp(\gamma_2\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right) &(224)\\
&=\log\left(\left(\int\exp(\gamma_1\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{\alpha}\right)\\
&\quad+\log\left(\left(\int\exp(\gamma_2\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{(1-\alpha)}\right) &(225)\\
&=\log\left(\left(\int\exp(\gamma_1\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{\alpha}\left(\int\exp(\gamma_2\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{(1-\alpha)}\right) &(226)\\
&=\log\left(\left(\int\exp(\gamma_1\alpha\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^{p}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{\frac{1}{p}}\right.\\
&\qquad\left.\left(\int\exp(\gamma_2(1-\alpha)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^{q}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{\frac{1}{q}}\right) &(227)\\
&\geqslant\log\left(\int\exp(\gamma_1\alpha\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\exp(\gamma_2(1-\alpha)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right) &(228)\\
&=\log\left(\int\exp\Big(\left(\gamma_1\alpha+\gamma_2(1-\alpha)\right)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\Big)\mathrm{d}Q(\boldsymbol{\theta})\right) &(229)\\
&=K_{Q,\boldsymbol{z}}(\gamma_1\alpha+\gamma_2(1-\alpha)), &(230)
\end{aligned}$$

where the inequality in (227) follows with $\alpha=\frac{1}{p}$ and $1-\alpha=\frac{1}{q}$; the inequality in (228) follows from Hölder's inequality. Hence, equality in (228) holds if and

only if there exist two constants β_1 and β_2, not simultaneously equal to zero, such that the set

$$\mathcal{A} \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \beta_1 \exp\left(\gamma_1 \alpha \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^p = \beta_2 \exp\left(\gamma_2 (1-\alpha) \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^q\} \tag{231}$$

$$= \{\boldsymbol{\theta} \in \mathcal{M} : \beta_1 \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) = \beta_2 \exp\left(\gamma_2 \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\} \tag{232}$$

$$= \left\{\boldsymbol{\theta} \in \mathcal{M} : \exp\left(\left(\gamma_1 - \gamma_2\right) \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) = \frac{\beta_2}{\beta_1}\right\} \tag{233}$$

$$= \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \frac{\log \frac{\beta_2}{\beta_1}}{\left(\gamma_1 - \gamma_2\right)}\right\}, \tag{234}$$

satisfies

$$Q\left(\mathcal{A}\right) = 1. \tag{235}$$

That is, strict inequality in (228) holds if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to the σ-finite measure Q. When $\alpha = 0$ or $\alpha = 1$, the proof is trivial. This completes the proof.

M Proof of Lemma 4.4

For all $s \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the equality in (63) implies the following,

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) = \frac{\mathrm{d}}{\mathrm{d}t} \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right)\Bigg|_{t=-\frac{1}{s}} \tag{236}$$

$$= \frac{1}{\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{v}\right)\right) \mathrm{d}Q(\boldsymbol{v})} \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\Bigg|_{t=-\frac{1}{s}} \tag{237}$$

$$= \frac{1}{\int \exp\left(-\frac{1}{s}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{v}\right)\right) \mathrm{d}Q(\boldsymbol{v})} \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \exp\left(-\frac{1}{s}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{238}$$

$$= \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right) \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \exp\left(-\frac{1}{s}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{239}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) - \frac{1}{s}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{240}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{241}$$

where the equality in (237) holds from the dominated convergence theorem [28]; the equality in (239) follows from (21); and the equality in (241) follows from (25).

For all $s \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the equalities in (63) and (240) imply

that

$$K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right)=\frac{\mathrm{d}}{\mathrm{d}t}\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(t\right)+t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})\Big|_{t=-\frac{1}{s}} \tag{242}$$

$$=\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\left(-K_{Q,\boldsymbol{z}}^{(1)}\left(t\right)+\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(t\right)+t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})\Big|_{t=-\frac{1}{s}} \tag{243}$$

$$=\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\left(-K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)+\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)-\frac{1}{s}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})$$

$$=\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\left(-K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)+\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,s)}\left(\boldsymbol{\theta}\right) \tag{244}$$

$$=-K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,s)}\left(\boldsymbol{\theta}\right)+\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^{2}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,s)}\left(\boldsymbol{\theta}\right) \tag{245}$$

$$=-\left(K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)^{2}+\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^{2}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,s)}\left(\boldsymbol{\theta}\right) \tag{246}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)^{2}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,s)}\left(\boldsymbol{\theta}\right), \tag{247}$$

where the equality in (243) follows from the dominated convergence theorem [28]; the equality in (244) is due to a change of measure through the Radon-Nikodym derivative in (25); and the equality in (246) follows from (241).

For all $s\in\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), the equalities in (63) and (246) imply that

$$
\begin{aligned}
&K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\\
=&\frac{\mathrm{d}}{\mathrm{d}t}\left(\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \mathrm{d}P^{\left(Q,-\frac{1}{t}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - \left(K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)\right)^2\right)\Bigg|_{t=-\frac{1}{s}} && (248)\\
=&\frac{\mathrm{d}}{\mathrm{d}t}\Bigg(\int \Big(\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \exp\left(-K_{Q,\boldsymbol{z}}\left(t\right)+t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\Big)\mathrm{d}Q\left(\boldsymbol{\theta}\right)\\
&-\left(K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)\right)^2\Bigg)\Bigg|_{t=-\frac{1}{s}} && (249)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \left(\frac{\mathrm{d}}{\mathrm{d}t}\exp\left(-K_{Q,\boldsymbol{z}}\left(t\right)+t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\Bigg|_{t=-\frac{1}{s}}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right)\\
&-2K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)K^{(2)}_{Q,\boldsymbol{z}}\left(t\right)\Bigg|_{t=-\frac{1}{s}} && (250)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \left(\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(t\right)+t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\Bigg|_{t=-\frac{1}{s}}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right)\\
&-2K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)K^{(2)}_{Q,\boldsymbol{z}}\left(t\right)\Bigg|_{t=-\frac{1}{s}} && (251)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)-\frac{1}{s}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right)\\
&-2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (252)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)\mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\\
&-2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (253)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\\
&-K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - 2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (254)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\\
&-K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\left(K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)+\left(K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)^2\right)\\
&-2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (255)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)^3 - 3K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (256)\\
=&\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right), && (257)
\end{aligned}
$$

where the equality in (249) follows from (25); the equality in (250) follows from the dominated convergence theorem [28]; the equality in (253) follows from (25); and the equality in (255) follows from (246). This completes the proof.

N Proof of Theorem 5.1

The proof is based on the analysis of the derivative of $K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ with respect to λ in $\operatorname{int}\mathcal{K}_{Q,\boldsymbol{z}}$. That is,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)=\frac{1}{\lambda^2}K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \tag{258}$$

$$\geqslant 0, \tag{259}$$

where the equality in (259) follows from Lemma 4.4. From Lemma 5.1, the inequality in (259) implies that the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ in (21) is nondecreasing with respect to λ. That is, given two reals λ_1 and λ_2, with $\lambda_1 > \lambda_2 > 0$, it holds that

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) \geqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) = \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{260}$$

The rest of the proof consists in showing that for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$, the function $K^{(2)}_{Q,\boldsymbol{z}}$ in (63) satisfies $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > 0$ if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable. For doing so, a handful of preliminary results are described in the following subsection. The proof of Theorem 5.1 resumes in Subsection N.2

N.1 Preliminaries

Given a positive real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), consider a partition of $\mathcal{M}$ formed by the sets $\mathcal{R}_0(\lambda)$, $\mathcal{R}_1(\lambda)$ and $\mathcal{R}_2(\lambda)$, such that

$$\mathcal{R}_0(\lambda) \triangleq \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}, \tag{261a}$$

$$\mathcal{R}_1(\lambda) \triangleq \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}, \text{ and} \tag{261b}$$

$$\mathcal{R}_2(\lambda) \triangleq \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}, \tag{261c}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is in (10) and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is in (25). Note that the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is the g-ERM-RER-optimal expected empirical risk in (68).

The sets in (261) exhibit several properties that are central for proving the main results of this section.

Lemma N.1. *The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), satisfies*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\lambda)\right) > 0, \tag{262}$$

if and only if

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\lambda)\right) > 0, \tag{263}$$

where the sets $\mathcal{R}_1(\alpha)$ *and* $\mathcal{R}_2(\alpha)$ *are in* (261b) *and* (261c)*, respectively.*

Proof: The proof is divided into two parts. In the first part, given a real $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$, it is proven that if the set $\mathcal{R}_1\left(\alpha\right)$ is nonnegligible with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, then the set $\mathcal{R}_2\left(\alpha\right)$ is nonnegligible with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. The second part of the proof consists in proving that, given a real $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$, if the set $\mathcal{R}_2\left(\alpha\right)$ is nonnegligible with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, then the set $\mathcal{R}_1\left(\alpha\right)$ is nonnegligible with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

The first part is proved by contradiction. Assume that set $\mathcal{R}_2\left(\alpha\right)$ is negligible with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Hence, from Lemma 4.4, it holds that

$$\begin{aligned}
K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) =& \int_{\mathcal{R}_0(\alpha)} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) + \int_{\mathcal{R}_1(\alpha)} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \\
&+ \int_{\mathcal{R}_2(\alpha)} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) && (264)\\
=& \int_{\mathcal{R}_0(\alpha)} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) + \int_{\mathcal{R}_1(\alpha)} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) && (265)\\
=& K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0(\alpha)\right) + \int_{\mathcal{R}_1(\alpha)} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) && (266)\\
<& K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0(\alpha)\right) + K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right) && (267)\\
=& K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) \left(P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0(\alpha)\right) + P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right)\right) && (268)\\
=& K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right), && (269)
\end{aligned}$$

which is a contradiction.

The second part of the proof follows the same arguments as in the first part. Assume that the set $\mathcal{R}_1\left(\alpha\right)$ is negligible with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Hence, from

Lemma 4.4, it holds that

$$
\begin{aligned}
K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)=&\int_{\mathcal{R}_0(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\boldsymbol{\nu}\right)+\int_{\mathcal{R}_1(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\boldsymbol{\nu}\right)\\
&+\int_{\mathcal{R}_2(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\boldsymbol{\nu}\right) &(270)\\
=&\int_{\mathcal{R}_0(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\boldsymbol{\nu}\right)+\int_{\mathcal{R}_2(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\boldsymbol{\nu}\right) &(271)\\
=&K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\mathcal{R}_0(\alpha)\right)+\int_{\mathcal{R}_2(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\boldsymbol{\nu}\right) &(272)\\
>&K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\mathcal{R}_0(\alpha)\right)\\
&+K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\mathcal{R}_2(\alpha)\right) &(273)\\
=&K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\mathcal{R}_0(\alpha)\right)+P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\alpha)}\left(\mathcal{R}_2(\alpha)\right)\right) &(274)\\
=&K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right), &(275)
\end{aligned}
$$

which is also a contradiction. This completes the proof. ■

A more general result can be immediately obtained by combining Lemma 3.12 and Lemma N.1.

Lemma N.2. *For all $\alpha\in\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (22)*, the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ in* (25)*, satisfies*

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\mathcal{R}_1(\alpha)\right)>0, \tag{276}$$

if and only if

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\mathcal{R}_2(\alpha)\right)>0, \tag{277}$$

where the sets $\mathcal{R}_1(\alpha)$ and $\mathcal{R}_2(\alpha)$ are in (261b) *and* (261c)*, respectively.*

N.2 The proof

The rest of the proof of Theorem 5.1 is divided into two parts. In the first part, it is shown that if for all $\alpha\in\mathcal{K}_{Q,\boldsymbol{z}}$, $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right)>0$, then the function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable. The second part of the proof, consists in showing that if the function $\mathsf{L}_{\boldsymbol{z}}$ is separable, then, for all $\alpha\in\mathcal{K}_{Q,\boldsymbol{z}}$, $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right)>0$.

The first part is as follows. From Lemma 4.4, it holds that for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$,

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{278}$$

$$= \int_{\mathcal{R}_0(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{279}$$

$$+ \int_{\mathcal{R}_1(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{280}$$

$$+ \int_{\mathcal{R}_2(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right), \tag{281}$$

where the sets $\mathcal{R}_0(\alpha)$, $\mathcal{R}_1(\alpha)$, and $\mathcal{R}_2(\alpha)$ are respectively defined in (261). Hence,

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) = \int_{\mathcal{R}_1(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$$

$$+ \int_{\mathcal{R}_2(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{282}$$

$$\leqslant \left(0 - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right)$$

$$+ \left(\sup_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right). \tag{283}$$

Under the assumption that for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$ the function $K^{(2)}_{Q,\boldsymbol{z}}$ in (63) satisfies $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > 0$, it follows from (283) that

$$0 < \left(0 - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right)$$

$$+ \left(\sup_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right). \tag{284}$$

Note that if

$$P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right) > 0, \tag{285}$$

then, $0 \neq K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)$. Moreover, if

$$P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right) > 0, \tag{286}$$

then, $\sup_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \neq K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)$. Therefore, the inequality in (284) implies that at least one of the following claims is true:

(a) $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right) > 0$; and

(b) $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right) > 0$.

Nonetheless, from Lemma N.1, it follows that both claims **(a)** and **(b)** hold simultaneously. Hence, the sets $\mathcal{R}_1(\alpha)$ and $\mathcal{R}_2(\alpha)$ are both nonnegligible with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and moreover, it holds that for all $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{R}_1(\alpha) \times \mathcal{R}_2(\alpha)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2)\,. \tag{287}$$

This proves that under the assumption that for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$, $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > 0$, the function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. From Lemma 4.2, it holds that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to Q. This completes the first part of the proof.

The second part of the proof is simpler. Assume that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. That is, for all $\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}$, there exist a positive real $c_\gamma > 0$; and two subsets $\mathcal{A}(\gamma)$ and $\mathcal{B}(\gamma)$ of $\mathcal{M}$ that are nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) and verify that for all $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{A}(\gamma) \times \mathcal{B}(\gamma)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > c_\gamma > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2)\,. \tag{288}$$

From Lemma 4.4, it holds that

$$\begin{aligned}
K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right) &= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (289)\\
&= \int_{\mathcal{A}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (290)\\
&\quad + \int_{\mathcal{B}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (291)\\
&\quad + \int_{\mathcal{M}\setminus(\mathcal{A}(\gamma)\cup\mathcal{B}(\gamma))} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (292)\\
&> 0, && (293)
\end{aligned}$$

where the inequality (293) follows from the following facts. First, if $c_\gamma < K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)$, with c_γ in (288), then for all $\boldsymbol{\nu} \in \mathcal{B}(\gamma)$, it holds that

$$\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 > \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2, \tag{294}$$

and thus,

$$\int_{\mathcal{B}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) > \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}(\gamma)) \tag{295}$$

$$> 0. \tag{296}$$

Second, if $c_\gamma \geqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)$ then for all $\boldsymbol{\nu} \in \mathcal{A}(\gamma)$, it holds that

$$\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 > \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2, \tag{297}$$

and thus,

$$\begin{aligned}
&\int_{\mathcal{A}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \\
&> \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\left(\gamma\right)\right) && (298)\\
&> 0. && (299)
\end{aligned}$$

Hence, under the assumption that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable, it holds that for all $\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}$, $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right) > 0$. This completes the proof.

O Proof of Lemma 5.1

Consider the partition of the set $\mathcal{M}$ formed by the sets following sets $\mathcal{A}_0$, $\mathcal{A}_1$, and $\mathcal{A}_2$ in (193). From (64), for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (22), it holds that,

$$\begin{aligned}
K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) =& \int_{\mathcal{A}_0} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (300)\\
&+ \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (301)\\
=& \int_{\mathcal{A}_0} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (302)\\
=& \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (303)\\
\geqslant& \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) + \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) && (304)\\
=& \delta^\star_{Q,\boldsymbol{z}}, && (305)
\end{aligned}$$

where, the equality in (302) follows by noticing that $Q\left(\mathcal{A}_1\right) = 0$, which implies that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) = 0$ (Lemma 3.8); the equality in (303) follows from noticing that $\mathcal{A}_0 = \mathcal{L}^\star_{Q,\boldsymbol{z}}$, with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (33); and the equality in (304) follows from (193c).

This completes the proof.

P Proof of Theorem 5.2

From (303) in the proof of Lemma 5.1, it holds that

$$\lim_{\lambda\to 0^+} K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)=\lim_{\lambda\to 0^+}\delta^\star_{Q,\boldsymbol{z}}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}})+\lim_{\lambda\to 0^+}\int_{\mathcal{A}_2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{306}$$

$$=\lim_{\lambda\to 0^+}\delta^\star_{Q,\boldsymbol{z}}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}})+\lim_{\lambda\to 0^+}\int_{\mathcal{A}_2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{307}$$

$$=\lim_{\lambda\to 0^+}\delta^\star_{Q,\boldsymbol{z}}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}})+\int_{\mathcal{A}_2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\lim_{\lambda\to 0^+}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{308}$$

$$=\delta^\star_{Q,\boldsymbol{z}}\lim_{\lambda\to 0^+}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) \tag{309}$$

$$=\delta^\star_{Q,\boldsymbol{z}}, \tag{310}$$

where, the equality in (308) follows from noticing two facts: (*a*) For all $\lambda\in\mathcal{K}_{Q,\boldsymbol{z}}$, the Randon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is positive and finite (Lemma 3.4); and (*b*) For all $\boldsymbol{\theta}\in\mathcal{A}_2$, it holds that $\lim_{\lambda\to 0^+}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=0$. Hence, the dominated convergence theorem [28, Theorem 1.6.9] holds. The inequality in (309) follows from Lemma 3.7 This completes the proof.

Q Proof of Theorem 7.1

From Theorem 5.1, it follows that for all $(\lambda_1,\lambda_2)\in\mathcal{K}_{Q,\boldsymbol{z}}\times\mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1>\lambda_2$,

$$\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})\,\frac{\mathrm{d}P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\alpha})\,\mathrm{d}Q(\boldsymbol{\alpha})\geqslant\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})\,\frac{\mathrm{d}P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\alpha})\,\mathrm{d}Q(\boldsymbol{\alpha}), \tag{311}$$

which implies the following inclusions:

$$\mathcal{R}_1(\lambda_2)\subseteq\mathcal{R}_1(\lambda_1),\ \text{and} \tag{312a}$$

$$\mathcal{R}_2(\lambda_1)\subseteq\mathcal{R}_2(\lambda_2), \tag{312b}$$

with the sets $\mathcal{R}_1(\cdot)$ and $\mathcal{R}_2(\cdot)$ in (261). From (82), it holds that for all $i\in\{1,2\}$,

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_i)=\mathcal{R}_2(\lambda_i)^{\mathsf{c}}, \tag{313}$$

where the complement is with respect to $\mathcal{M}$. Thus, the inclusion in (312b) and the equality in (313) yields,

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)\supseteq\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2). \tag{314}$$

The inclusion $\mathcal{M} \supseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)$ follows from (82). Alternatively, the inclusion $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \supseteq \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$, follows from Lemma 5.1 and from observing that for all $\boldsymbol{\nu} \in \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$,

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geqslant \delta^{\star}_{Q,\boldsymbol{z}} \tag{315}$$

$$\geqslant \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right), \tag{316}$$

which implies that $\boldsymbol{\nu} \in \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$. This completes the proof of (86).

The proof of (87) is as follows. From the mean value theorem [28] and the assumption that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is continuous on $\mathcal{M}$, it follows that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, there always exists a model $\boldsymbol{\theta} \in \mathcal{M}$, such that

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\alpha}\right), \tag{317}$$

which implies that $\mathcal{R}_0\left(\lambda\right)$ is not empty, and as a consequence, $\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda\right) = \mathcal{R}_0\left(\lambda\right) \cup \mathcal{R}_1\left(\lambda\right)$ is not empty. Hence, for all $\boldsymbol{\theta} \in \mathcal{R}_0\left(\lambda_1\right)$ it holds that $\boldsymbol{\theta} \notin \mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)$. This proves that the elements of $\mathcal{R}_0\left(\lambda_1\right)$ are in $\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right)$ but not in $\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)$. This, together with (314), verifies that

$$\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right) \supset \mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right). \tag{318}$$

The strict inclusion $\mathcal{M} \supset \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)$ is proved by contradiction. Assume that there exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ such that $\mathcal{M} = \mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$. Then, $\mathcal{R}_2\left(\lambda\right) = \emptyset$ and thus, $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2\left(\lambda\right)\right) = 0$, which together with Lemma N.1, implies that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1\left(\lambda\right)\right) = 0$ and consequently,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0\left(\lambda\right)\right) = 1. \tag{319}$$

This contradicts the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable (Definition 4.1). Hence, $\mathcal{M} \supset \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)$.

Finally, the strict inclusion $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \supset \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ is proved by contradiction. Assume that there exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ such that $\mathcal{N}^{\star}_{Q,\boldsymbol{z}} = \mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$. That is,

$$\left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \leqslant \delta^{\star}_{Q,\boldsymbol{z}}\right\} = \mathcal{N}^{\star}_{Q,\boldsymbol{z}} \tag{320}$$

$$= \mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \tag{321}$$

$$= \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \leqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right\}. \tag{322}$$

Hence, three cases might arise:
(a) there exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, such that $\delta^{\star}_{Q,\boldsymbol{z}} < K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ and it holds that

$$\left\{\boldsymbol{\nu} \in \mathcal{M} : \delta^{\star}_{Q,\boldsymbol{z}} < \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \leqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right\} = \emptyset;$$

(b) there exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, such that $\delta^{\star}_{Q,\boldsymbol{z}} > K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ and it holds that

$$\left\{\boldsymbol{\nu} \in \mathcal{M} : K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \leqslant \delta^{\star}_{Q,\boldsymbol{z}}\right\} = \emptyset;$$

or (c) there exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, such that $\delta^{\star}_{Q,\boldsymbol{z}} = K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$.

The cases (a) and (b) are absurd. Hence, the proof is complete only by considering the case (c). In the case (c), it holds that,

$$\mathcal{R}_1\left(\lambda\right) = \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{323}$$

and from the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (32), it holds that

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1\left(\lambda\right)\right) = 0. \tag{324}$$

From Lemma N.1 and (324), it follows that,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2\left(\lambda\right)\right) = 0. \tag{325}$$

Finally, by noticing that

$$1 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0\left(\lambda\right)\right) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1\left(\lambda\right)\right) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2\left(\lambda\right)\right) \tag{326}$$

$$= P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0\left(\lambda\right)\right), \tag{327}$$

reveals a contradiction to the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ (and thus, separable with respect to Q by Lemma 4.2). This completes the proof of (87).

R Proof of Theorem 7.2

The proof of (88) is based on the analysis of the derivative of $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right)$ with respect to λ, for some fixed set $\mathcal{A} \subseteq \mathcal{M}$. More specifically, given a $\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}$, it holds that

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) = \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\alpha}\right) \mathrm{d}Q\left(\boldsymbol{\alpha}\right), \tag{328}$$

and from the fundamental theorem of calculus [34], it follows that for all $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) - P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) = \int_{\lambda_2}^{\lambda_1} \frac{\mathrm{d}}{\mathrm{d}\gamma} P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) \mathrm{d}\gamma \tag{329}$$

$$= \int_{\lambda_2}^{\lambda_1} \frac{\mathrm{d}}{\mathrm{d}\gamma} \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\alpha}\right) \mathrm{d}Q\left(\boldsymbol{\alpha}\right) \mathrm{d}\gamma \tag{330}$$

$$= \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{A}} \frac{\mathrm{d}}{\mathrm{d}\gamma} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\alpha}\right) \mathrm{d}Q\left(\boldsymbol{\alpha}\right) \mathrm{d}\gamma, \tag{331}$$

where the equality in (330) follows from (328); and the equality in (331) holds from Lemma 3.4 and the dominated convergence theorem [28].

For all $\boldsymbol{\theta} \in \operatorname{supp} Q$, the following holds,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})} \tag{332}$$

$$= \frac{\frac{1}{\lambda^2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})} - \frac{\frac{1}{\lambda^2}\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\alpha})}{\left(\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})\right)^2} \tag{333}$$

$$= \frac{1}{\lambda^2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) - \frac{1}{\lambda^2}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\mathrm{d}Q(\boldsymbol{\nu}) \tag{334}$$

$$= \frac{1}{\lambda^2}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right). \tag{335}$$

Plugging (335) into (331) yields,

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) - P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})$$

$$= \int_{\lambda_2}^{\lambda_1}\int_{\mathcal{A}}\frac{1}{\gamma^2}\frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\alpha})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\alpha})\,\mathrm{d}\gamma \tag{336}$$

$$= \int_{\lambda_2}^{\lambda_1}\int_{\mathcal{A}}\frac{1}{\gamma^2}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\,\mathrm{d}\gamma. \tag{337}$$

Note that for all $\boldsymbol{\alpha} \in \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$, it holds that for all $\gamma \in (\lambda_2, \lambda_1)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant 0, \tag{338}$$

and thus,

$$\int_{\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)}\frac{1}{\gamma^2}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \leqslant 0. \tag{339}$$

The equalities in (337) and (339), with $\mathcal{A} = \mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$, imply that

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) - P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) \leqslant 0. \tag{340}$$

The inequality $0 < P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2))$ in (88) is proved by contradiction. Assume that for some $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ it holds that $0 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2))$. Then, $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\lambda_2)) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\lambda_2)) = 0$, which implies that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\lambda_2)) = 1$, which is a contradiction. See for instance, Lemma N.2. This completes the proof of (88).

The proof of (89) is divided into two parts. The first part shows that if for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{341}$$

then the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to Q. The second part of the proof shows that if the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to Q, then, for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$, the inequality in (341) holds.

The first part is as follows. In the proof of Theorem 7.1 it is shown (see (337)) that for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$\begin{aligned} &P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\right) - P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\right) \\ =&\int_{\lambda_2}^{\lambda_1} \int_{\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) - \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) \mathrm{d}\gamma \end{aligned} \tag{342}$$

Assume that for a given pair $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\lambda_1 > \lambda_2$, the inequality in (341) holds. Then, from (342),

$$\begin{aligned} 0 &> \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) - \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) \mathrm{d}\gamma \\ &\geqslant \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{R}_1(\lambda_2)} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) - \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) \mathrm{d}\gamma \end{aligned} \tag{343}$$

where the equality in (343) follows from noticing that $\mathcal{R}_0\left(\lambda_2\right)$ and $\mathcal{R}_1\left(\lambda_2\right)$ form a partition of $\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)$, with the sets $\mathcal{R}_0\left(\lambda_2\right)$, $\mathcal{R}_1\left(\lambda_2\right)$ and $\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)$ defined in (261a), (261b), and (82), respectively.

The inequality in (343) implies that the set $\mathcal{R}_1\left(\lambda_2\right)$ is nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, for some $\gamma \in (\lambda_2, \lambda_1)$. Hence, from Lemma N.2, it follows that both sets $\mathcal{R}_1\left(\lambda_2\right)$ and $\mathcal{R}_2\left(\lambda_2\right)$ are nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

From the above arguments, it has been proved that given a pair $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$, if

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{344}$$

then there always exists a positive $\gamma \in (\lambda_1, \lambda_2)$ such that the sets $\mathcal{R}_1\left(\lambda_2\right)$ and $\mathcal{R}_2\left(\lambda_2\right)$ are not negligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Moreover, such sets $\mathcal{R}_1\left(\lambda_2\right)$ and $\mathcal{R}_2\left(\lambda_2\right)$ satisfy for all $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{R}_2\left(\lambda\right) \times \mathcal{R}_1\left(\lambda\right)$,

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}_1\right) > K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) > \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}_2\right), \tag{345}$$

which together with Definition 4.2 verify that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ (and thus, with respect to Q by Lemma 4.2). This ends the first part of the proof.

The second part of the proof is under the assumption that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable with respect to Q (and thus, with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ by Lemma 4.2). That is, from Definition 4.2, for all $\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}$, there exist a positive real $c_\gamma > 0$ and two subsets $\mathcal{A}(\gamma)$ and $\mathcal{B}(\gamma)$ of $\mathcal{M}$ nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) that verify that for all $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{A}(\gamma) \times \mathcal{B}(\gamma)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > c_\gamma > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2). \tag{346}$$

In the proof of Theorem 7.1, c.f. (337), it has been proved that given a pair $(\alpha_1, \alpha_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\alpha_1 > \gamma > \alpha_2$, it holds that for all subsets $\mathcal{A}$ of $\mathcal{M}$,

$$\begin{aligned}
&P^{(Q,\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) - P^{(Q,\alpha_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) \\
&= \int_{\alpha_2}^{\alpha_1} \int_{\mathcal{A}} \frac{1}{\lambda^2} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\alpha}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P(\boldsymbol{\alpha}) \, \mathrm{d}\lambda \qquad (347) \\
&= \int_{\alpha_2}^{\alpha_1} \int_{\mathcal{A}} \frac{1}{\lambda^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\lambda. \qquad (348)
\end{aligned}$$

Hence, two cases are studied. The first case considers that

$$c_\gamma < K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right), \tag{349}$$

with c_γ in (346). The second case considers that

$$c_\gamma \geqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right). \tag{350}$$

In the first case, it follows from (82) that

$$\mathcal{B}(\gamma) \subset \mathcal{N}_{Q,\boldsymbol{z}}(\gamma), \tag{351}$$

which implies that

$$\begin{aligned}
P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) &\geqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}(\gamma)) \qquad (352) \\
&> 0, \qquad (353)
\end{aligned}$$

where, the inequality in (353) follows from the fact that $\mathcal{B}(\gamma)$ is nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. This implies that the set $\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)$ is not negligible with respect $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Moreover, from (82) and (351), it follows that for all $\boldsymbol{\alpha} \in \mathcal{N}_{Q,\boldsymbol{z}}(\gamma)$ and for all $\lambda \in (\gamma, \alpha_1)$,

$$\begin{aligned}
\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) &< \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - c_\gamma \qquad (354) \\
&< 0, \qquad (355)
\end{aligned}$$

where the inequality in (354) follows from (349); and the inequality in (355) follows from (346). Thus,

$$\int_{\gamma}^{\alpha_1} \int_{\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)} \frac{1}{\lambda^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\lambda < 0 \tag{356}$$

which implies, from (348), that

$$P^{(Q,\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) - P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) < 0. \tag{357}$$

Assume now that $c_\gamma \geqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)$. Hence, the following holds

$$A(\gamma) \subseteq \mathcal{R}_2(\gamma), \tag{358}$$

which implies that

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\gamma)) \geqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}(\gamma)) \tag{359}$$

$$> 0, \tag{360}$$

where the inequality in (360) follows from the fact that $\mathcal{A}(\gamma)$ is nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. This implies that the set $\mathcal{R}_2(\gamma)$ is not negligible with respect $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. From Lemma N.1, it follows that both $\mathcal{R}_1(\gamma)$ and $\mathcal{R}_2(\gamma)$ are nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Using this result, the following holds,

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) \geqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\gamma)) \tag{361}$$

$$> 0, \tag{362}$$

which proves the set $\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)$ is nonnegligible with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

From (82) and Theorem 5.1, it follows that for all $\boldsymbol{\alpha} \in \mathcal{N}_{Q,\boldsymbol{z}}(\gamma)$ and for all $\lambda \in (\gamma, \alpha_1)$,

$$0 \geqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\underline{\boldsymbol{X}}=\boldsymbol{z}=\boldsymbol{y}}(\boldsymbol{\nu}) \tag{363}$$

$$> \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\underline{\boldsymbol{X}}=\boldsymbol{z}=\boldsymbol{y}}(\boldsymbol{\nu}). \tag{364}$$

Thus,

$$\int_{\gamma}^{\alpha_1} \int_{\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)} \frac{1}{\lambda^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\lambda < 0 \tag{365}$$

which implies, from (348), that

$$P^{(Q,\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) - P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) < 0. \tag{366}$$

The inequality $0 < P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2))$ in (89) has already been proved while proving (88), and thus, this completes the proof of (89).

S Proof of Lemma 7.2

The proof is based on the following two observations. First, note that $(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2))^{\mathsf{c}} = \mathcal{R}_2(\lambda_2)$, with the set $\mathcal{R}_2(\cdot)$ defined in (261c). Second, note that

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) = \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \cup (\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \cap \mathcal{R}_2(\lambda_2)), \tag{367}$$

and the fact that the sets $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$ and $(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \cap \mathcal{R}_2(\lambda_2))$ are disjoint. Hence, for all $i \in \{1,2\}$,

$$P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)) = P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \cup (\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \cap \mathcal{R}_2(\lambda_2))\Big) \tag{368}$$

$$= P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)\Big) + P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \cap \mathcal{R}_2(\lambda_2)\Big) \tag{369}$$

$$= P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)\Big), \tag{370}$$

where the equality in (369) follows from Lemma 3.8 and the equality in (90).

Finally, under the assumption that the empirical function $\mathsf{L}_{\boldsymbol{z}}$ in (5) is separable, it holds from Theorem 7.2 that

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)). \tag{371}$$

Plugging (370) into (371), with $i = 1$, yields,

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{372}$$

and this completes the proof.

T Proof of Theorem 7.3

$$\lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)) = \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}) + \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \setminus \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) \tag{373}$$

$$= \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}) \tag{374}$$

$$= 1 \tag{375}$$

where, the equalities in (374) follows Lemma 3.7. This completes the proof.

U Proof of Lemma 8.1

The cumulant generating function $J_{\boldsymbol{z},Q,\lambda}$ in (100) induced by the measure $P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$ in (97) evaluated at t, with $t \leqslant \frac{1}{\lambda}$, is

$$
\begin{aligned}
J_{\boldsymbol{z},Q,\lambda}(t) =& \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{u}\right)\mathrm{d}Q\left(\boldsymbol{u}\right)\right) && (376)\\
=& \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\mathrm{d}Q\left(\boldsymbol{u}\right)\right) && (377)\\
=& \log\left(\int \exp\left(\left(t-\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\mathrm{d}Q\left(\boldsymbol{u}\right)\right) && (378)\\
=& \log\left(\int \exp\left(\left(t-\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)-K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)+K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)\right.\right.\\
& \left.\left.-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\mathrm{d}Q\left(\boldsymbol{u}\right)\right) && (379)\\
=& K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\\
& +\log\left(\int \exp\left(\left(t-\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)-K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)\right)\mathrm{d}Q\left(\boldsymbol{u}\right)\right) && (380)\\
=& K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\\
& +\log\left(\int \frac{\mathrm{d}P^{\left(Q,-\frac{1}{t-\frac{1}{\lambda}}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{u}\right)\mathrm{d}Q\left(\boldsymbol{u}\right)\right) && (381)\\
=& K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), && (382)
\end{aligned}
$$

where the equality in (377) follows from Theorem 3.1; and the equality in (381) follows from the fact that $-\frac{1}{t-\frac{1}{\lambda}} > 0$ for all $t < \frac{1}{\lambda}$. This completes the proof.

V Proof of Lemma 10.1

From [35, Corollary 4.15, Page 100], it follows that the probability measures P and Q in $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ satisfy the following equality:

$$
D\left(Q\|P\right) = \sup_{f}\int f\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) - \log\int \exp\left(f\left(\boldsymbol{\theta}\right)\right)\mathrm{d}P\left(\boldsymbol{\theta}\right), \tag{383}
$$

where the supremum is over the space of all measurable functions f with respect to $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, such that $\int \exp\left(f\left(\boldsymbol{\theta}\right)\right)\mathrm{d}P\left(\boldsymbol{\theta}\right) < \infty$. Hence, for

all $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$ and for all $t \in (-\infty, 0)$, it follows that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5) satisfies that

$$\begin{aligned} D\left(Q\|P\right) \geqslant & \int t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right) - \log \int \exp\left(t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}P\left(\boldsymbol{\theta}\right) && (384) \\ \geqslant & \int t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right) \\ & - \log \int \exp\left(t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) + t\mathsf{R}_{\boldsymbol{z}}\left(P\right) - t\mathsf{R}_{\boldsymbol{z}}\left(P\right)\right) \mathrm{d}P\left(\boldsymbol{\theta}\right) && (385) \\ = & \int t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right) - t\mathsf{R}_{\boldsymbol{z}}\left(P\right) \\ & - \log \int \exp\left(t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - t\mathsf{R}_{\boldsymbol{z}}\left(P\right)\right) \mathrm{d}P\left(\boldsymbol{\theta}\right) && (386) \\ = & t\mathsf{R}_{\boldsymbol{z}}\left(Q\right) - t\mathsf{R}_{\boldsymbol{z}}\left(P\right) - \log \int \exp\left(t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - t\mathsf{R}_{\boldsymbol{z}}\left(P\right)\right) \mathrm{d}P\left(\boldsymbol{\theta}\right), && (387) \end{aligned}$$

which leads to

$$\mathsf{R}_{\boldsymbol{z}}\left(Q\right) - \mathsf{R}_{\boldsymbol{z}}\left(P\right) \leqslant \frac{D\left(Q\|P\right) + \log \int \exp\left(t\left(\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right) - \mathsf{R}_{\boldsymbol{z}}\left(P\right)\right)\right) \mathrm{d}P\left(\boldsymbol{\theta}\right)}{t}. \quad (388)$$

Given that t can be chosen arbitrarily in $(-\infty, 0)$, it holds that

$$\mathsf{R}_{\boldsymbol{z}}\left(Q\right) - \mathsf{R}_{\boldsymbol{z}}\left(P\right) \leqslant \inf_{t \in (-\infty, 0)} \frac{D\left(Q\|P\right) + \log \int \exp\left(t\left(\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right) - \mathsf{R}_{\boldsymbol{z}}\left(P\right)\right)\right) \mathrm{d}P\left(\boldsymbol{\theta}\right)}{t} \quad (389)$$

which completes the proof.

W Proof of Theorem 10.1

From Lemma 10.1, it holds that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), satisfies

$$\begin{aligned} \mathsf{R}_{\boldsymbol{z}}\left(Q\right) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant & \inf_{t \in (-\infty, 0)} \Bigg(\frac{D\left(Q\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}{t} \\ & + \frac{\log\left(\int \exp\left(t\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\tfrac{1}{\lambda}\right)\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right)}{t} \Bigg), \end{aligned} \quad (390)$$

where the function $K^{(1)}_{Q,\boldsymbol{z}}$ is defined in (64) and satisfies (68). Moreover, for all $t \in (-\infty, 0)$,

$$\begin{aligned} & \log\left(\int \exp\left(t\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right) \\ = & \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right) - tK^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) && (391) \\ = & J_{\boldsymbol{z},Q,\lambda}(t) - tK^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) && (392) \\ \leqslant & \frac{1}{2}t^2 B^2_{\boldsymbol{z}}, && (393) \end{aligned}$$

where the inequality in (392) follows from (100); the inequality in (393) follows from Theorem 8.1; and the constant $B_{\boldsymbol{z}}$ is defined in (105).

Plugging (393) into (390) yields for all $t \in (-\infty, 0)$,

$$\mathsf{R}_{\boldsymbol{z}}\left(Q\right) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant \inf_{t\in(-\infty,0)} \frac{D\left(Q\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \frac{1}{2}t^2 B^2_{\boldsymbol{z}}}{t} : \tag{394}$$

Let the $c \in \mathbb{R}$ be defined as follows:

$$c \triangleq \mathsf{R}_{\boldsymbol{z}}\left(Q\right) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{395}$$

Hence, from (394), it follows that for all $t \in (-\infty, 0)$,

$$c\,t - \frac{1}{2}t^2 B^2_{\boldsymbol{z}} \leqslant D\left(Q\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{396}$$

The rest of the proof consists in finding an explicit expression for the absolute value of c. To this aim, consider the function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$\phi(\alpha) = \frac{1}{2}\alpha^2 B^2_{\boldsymbol{z}}, \tag{397}$$

and note that ϕ is a positive and strictly convex function with $\phi(0) = 0$. Let the Legendre-Fenchel transform of ϕ be the function $\phi^* : \mathbb{R} \to \mathbb{R}$, and thus for all $x \in \mathbb{R}$,

$$\phi^*(x) = \max_{t\in(-\infty,0)} xt - \phi(t). \tag{398}$$

In particular, note that

$$\phi^*(c) \leqslant D\left(Q\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{399}$$

Note that for all $x \in \mathbb{R}$ and for all $t \in (-\infty, 0)$, the function $\phi^\star$ in (398) satisfies

$$x\,t - \frac{1}{2}t^2 B^2_{\boldsymbol{z}} \leqslant \phi^\star(x) = x\alpha^\star(x) - \phi\left(\alpha^\star(x)\right), \tag{400}$$

where the term $\alpha^\star(x)$ represents the unique solution in α within the interval $(-\infty, 0)$ to

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(x\alpha - \phi\left(\alpha\right)\right) = x - \alpha B^2_{\boldsymbol{z}} = 0. \tag{401}$$

That is,

$$\alpha^\star(x) = \frac{x}{B^2_{\boldsymbol{z}}}. \tag{402}$$

Plugging (402) into (400) yields,

$$\phi^\star(x) = \frac{x^2}{2B^2_{\boldsymbol{z}}}. \tag{403}$$

Hence, from (399) and (400), given c in (395) for all $t \in (-\infty, 0)$,

$$c\,t - \frac{1}{2}t^2 B_{\boldsymbol{z}}^2 \leqslant \phi^\star(c) \leqslant D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{404}$$

and thus,

$$\frac{c^2}{2B_{\boldsymbol{z}}^2} \leqslant D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{405}$$

This implies that

$$c \leqslant \sqrt{2B_{\boldsymbol{z}}^2 D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)} \tag{406}$$

and

$$c \geqslant -\sqrt{2B_{\boldsymbol{z}}^2 D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}, \tag{407}$$

which leads to

$$\left| \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right| \leqslant \sqrt{2B_{\boldsymbol{z}}^2 D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}, \tag{408}$$

and completes the proof.

X Proof of Theorem 10.2

The optimization problem in (121) can be re-written as follows:

$$\min_{P \in \triangle_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}), \tag{409a}$$

$$\text{subject to:} \int \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu}) \log\left(\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant c, \tag{409b}$$

$$\int \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) = 1. \tag{409c}$$

Let $\mathscr{M}$ be the set of nonnegative measurable functions with respect to the measurable spaces $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. The Lagrangian of the optimization problem in (409) can be constructed in terms of a function in $\mathscr{M}$, instead of a measure over the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. Let such Lagrangian $L : \mathscr{M} \times [0, +\infty)^2 \to \mathbb{R}$ be of the form

$$\begin{aligned} L(g, \alpha, \beta) = & \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, g(\boldsymbol{\nu})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \\ & + \alpha \left(\int g(\boldsymbol{\nu}) \log\left(g(\boldsymbol{\nu})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) - c \right) \\ & + \beta \left(\int g(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) - 1 \right), \end{aligned} \tag{410}$$

where g is a notation to represent the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}$; the reals α and β are both nonnegative and act as Lagrangian multipliers due to the constraint (409b) and (409c), respectively.

Let $h : \mathbb{R}^k \to \mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (410) at $(g, \alpha, \beta) \in \mathscr{M} \times [0, +\infty)^2$ in the direction of h is

$$\partial L\left(g, \alpha, \beta; h\right) \triangleq \left.\frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma)\right|_{\gamma=0}, \tag{411}$$

where the real function $r : \mathbb{R} \to \mathbb{R}$ is such that for all $\gamma \in \mathbb{R}$,

$$\begin{aligned} r(\gamma) = & \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\left(g\left(\boldsymbol{\nu}\right)+\gamma h\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) + \\ & \alpha\left(\int\left(g\left(\boldsymbol{\nu}\right)+\gamma h\left(\boldsymbol{\nu}\right)\right) \log\left(g\left(\boldsymbol{\nu}\right)+\gamma h\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) - c\right) \\ & +\beta\left(\int\left(g\left(\boldsymbol{\nu}\right)+\gamma h\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) - 1\right), \end{aligned} \tag{412}$$

Note that the derivative of the real function r in (412) is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) = & \int \mathsf{L}_{\boldsymbol{z}} h\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \\ & +\alpha \int h\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) + \alpha \int h\left(\boldsymbol{\nu}\right) \log\left(g\left(\boldsymbol{\nu}\right)+\gamma h\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \\ & +\beta \int h\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right). \end{aligned} \tag{413}$$

From (411) and (413), it follows that

$$\partial L\left(g, \alpha, \beta; h\right) = \int h\left(\boldsymbol{\nu}\right)\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) + \alpha\left(1+\log g\left(\boldsymbol{\nu}\right)\right)+\beta\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right). \tag{414}$$

From [32, Theorem 1, page 178], it holds that a necessary condition for the functional L in (410) to have a minimum at $(g, \alpha, \beta) \in \mathscr{M} \times [0, +\infty)^2$ is that for all functions $h \in C\left(\mathcal{M}\right)$

$$\partial L\left(g, \alpha, \beta; h\right) = 0, \tag{415}$$

which implies that for all $\boldsymbol{\nu} \in \mathcal{M}$,

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) + \alpha\left(1+\log g\left(\boldsymbol{\nu}\right)\right)+\beta = 0. \tag{416}$$

Thus,

$$g\left(\boldsymbol{\nu}\right) = \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)}{\alpha}\right) \exp\left(-\frac{\beta+\alpha}{\alpha}\right), \tag{417}$$

where α and β are chosen to satisfy their corresponding constraints. Denote by $P^\star$ the solution of the optimization problem in (121). Hence, from (417), it follows that

$$\frac{\mathrm{d}P^\star}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\alpha}\right)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{418}$$

where α is chosen to satisfy

$$D\left(P^\star \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = c. \tag{419}$$

From Lemma 3.8, it follows that the probability measure $P^\star$ and the σ-finite measure Q satisfy,

$$\frac{\mathrm{d}P^\star}{\mathrm{d}Q}(\boldsymbol{\nu}) = \frac{\mathrm{d}P^\star}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu})\,\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \tag{420}$$

$$= \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\alpha}\right)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})}\right) \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta})}\right) \tag{421}$$

$$= \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\alpha}\right)\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\alpha})}\mathrm{d}Q(\boldsymbol{\theta})}\right) \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta})}\right) \tag{422}$$

$$= \frac{\exp\left(-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)}{\int \exp\left(-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\theta})}, \tag{423}$$

which implies that $P^\star$ is a Gibbs probability measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, with reference measure Q, regularization parameter $\frac{1}{\frac{1}{\alpha}+\frac{1}{\lambda}}$, and energy function $\mathsf{L}_{\boldsymbol{z}}$. That is, for all $\boldsymbol{\nu} \in \operatorname{supp} Q$,

$$P^\star(\boldsymbol{v}) = P^{\left(Q,\frac{\alpha\lambda}{\alpha+\lambda}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}), \tag{424}$$

where α is chosen to satisfy (419). Let the positive real ω be $\omega \triangleq \frac{\alpha\lambda}{\alpha+\lambda}$ and note that $\omega \in (0, \lambda]$ and satisfies $D\left(P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = c$. The proof ends by verifying that the objective function in (410) is strictly convex, and thus, the measure $P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the unique minimizer. This completes the proof.

References

[1] O. Catoni, *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, 1st ed. New York, NY, USA: Springer Science & Business Media, 2004, vol. 1851.

[2] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, Aug. 2016.

[3] P. Alquier, J. Ridgway, and N. Chopin, "On the properties of variational approximations of Gibbs posteriors," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, 2016.

[4] C. P. Robert, *The Bayesian choice: From decision-theoretic foundations to computational implementation*, 1st ed. New York, NY: Springer, 2007.

[5] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," in *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, vol. 4, New Orleans, LA, USA, Dec. 2021, pp. 831–838.

[6] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *Information Theory Workshop*, Cambridge, United Kingdom, Sep. 2016, pp. 26–30.

[7] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, Jan. 2019.

[8] T. Zhang, "Information-theoretic upper and lower bounds for statistical estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, Apr. 2006.

[9] A. R. Asadi and E. Abbe, "Chaining meets chain rule: Multilevel entropic regularization and training of neural networks." *J. Mach. Learn. Res.*, vol. 21, pp. 139–1, 2020.

[10] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. of the Thirty-first Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2017.

[11] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proceedings of the tenth annual conference on Computational learning theory*, 1997, pp. 2–9.

[12] D. A. McAllester, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[13] M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor, "PAC-Bayes unleashed: Generalisation bounds with unbounded losses," *Entropy*, vol. 23, no. 10, 2021.

[14] B. Guedj and L. Pujol, "Still no free lunches: The price to pay for tighter PAC-Bayes bounds," *Entropy*, vol. 23, no. 11, 2021.

[15] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," *Neural Information Processing Systems*, 1999.

[16] J. Zhu and E. P. Xing, "Maximum entropy discrimination Markov networks," *Journal of Machine Learning Research*, vol. 10, no. 11, 2009.

[17] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," in *Predicting structured data*, 1st ed., G. BakIr, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. Vishwanathan, Eds. New York, NY.: The MIT Press, 2007, ch. 10, pp. 191–241.

[18] C. P. Robert and G. Casella, *Monte Carlo statistical methods*, 2nd ed. New York, NY, USA: Springer, 2004.

[19] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*, 1st ed. New York, NY: Wiley, 1989.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

[21] E. T. Jaynes, "Information theory and statistical mechanics I," *Physical Review Journals*, vol. 106, pp. 620–630, May 1957.

[22] ——, "Information theory and statistical mechanics II," *Physical Review Journals*, vol. 108, pp. 171–190, Oct. 1957.

[23] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 721–741, 1984.

[24] O. Catoni, *PAC-Bayesian supervised classification: The thermodynamics of statistical learning*, 1st ed. Beachwood, OH, USA: Institute of Mathematical Statistics Lecture Notes - Monograph Series, 2007, vol. 56.

[25] B. Guedj, "A primer on PAC-Bayesian learning." in *Tutorials of the International Conference on Machine Learning (ICML)*, Jun. 2019.

[26] J. W. Gibbs, *Elementary principles in statistical mechanics*, 1st ed. New Haven, NJ: Yale University Press, 1902.

[27] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY: Springer-Verlag, 2009.

[28] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA: Harcourt/Academic Press, 1999.

[29] W. F. Trench, *Introduction to Real Analysis*, 1st ed. Hoboken, NJ: Prentice Hall/Pearson Education, 2003.

[30] S. M. Perlaza, I. Esnaola, and H. V. Poor, "Sensitivity of the Gibbs algorithm to data aggregation in supervised machine learning," Inria, Centre de Recherche de Sophia Antipolis Méditérranée, Sophia Antipolis, Tech. Rep. RR-9474, Jun. 2022.

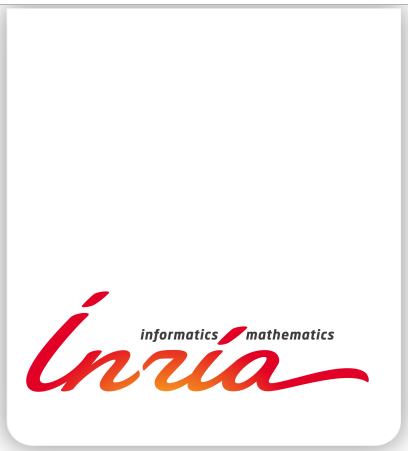
[31] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

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

[33] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. New York, NY: Jhon Wiley & Sons, 1971, vol. II.

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

[35] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, 1st ed. Oxford, UK: Oxford University Press, 2013.



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