Empirical Risk Minimization with Relative Entropy Regularization

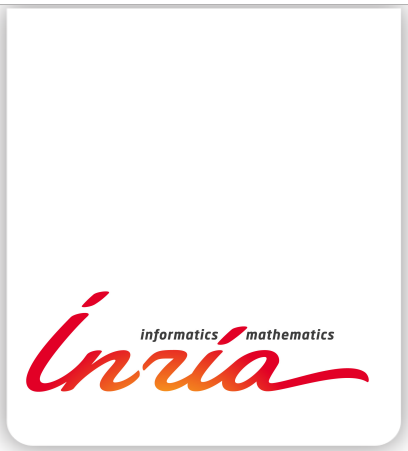
Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, Stefano Rini

Empirical Risk Minimization with 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 Relative Entropy Regularization

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

Project-Team NEO

Research Report n° 9454 — version 6 — initial version February 2022 —
revised version November 2023 — 100 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, and not necessarily a probability measure. Under this assumption, which leads to a generalization of the ERM-RER problem allowing a larger degree of flexibility for incorporating prior knowledge, numerous relevant properties are stated. Among these properties, the solution to this problem, if it exists, is shown to be a unique probability measure, often mutually absolutely continuous with the reference measure. Such a solution exhibits a probably-approximately-correct guarantee for the ERM problem independently of whether the latter possesses a solution. For a fixed dataset, the empirical risk is shown to be a sub-Gaussian random variable when the models are sampled from the solution to the ERM-RER problem. The generalization capabilities of the solution to the ERM-RER problem (the Gibbs algorithm) are studied via the sensitivity of the expected empirical risk to deviations from such a solution towards alternative probability measures. Finally, an interesting connection between sensitivity, generalization error, and lautum information is established.

Key-words: Supervised Learning, PAC-Learning, Regularization, Relative Entropy, Empirical Risk Minimization, Gibbs Measure, Gibbs Algorithm, Generalization, and Sensitivity.

Samir M. Perlaza and Alain Jean-Marie are with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France. Gaetan Bisson and Samir M. Perlaza are with the GAATI Laboratory at the Université de la Polynésie Française, Faaa 98702, French Polynesia. Iñaki Esnaola is with the ACSE Dept. at The University of Sheffield, Sheffield S1 3JD, UK. Stefano Rini is with the ECE Dept. at the National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan 30010, ROC. Samir M. Perlaza and Iñaki Esnaola are also with the ECE Dept. at Princeton University, Princeton N.J. 08544, USA. This work has been presented in part at the IEEE International Symposium on Information Theory (ISIT) in [1].

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

Résumé : Le problème de minimisation du risque empirique (ERM) avec régularisation d'entropie relative (ERM-RER) est étudié sous l'hypothèse que la mesure de référence est une mesure σ-finie, et pas nécessairement une mesure de probabilité. Sous cette hypothèse, qui conduit à une généralisation du problème ERM-RER permettant une plus grande flexibilité pour l'incorporation des connaissances antérieures, de nombreuses propriétés pertinentes sont énoncées. Parmi ces propriétés, la solution à ce problème, si elle existe, se révèle être une mesure de probabilité unique, souvent mutuellement absolument continue avec la mesure de référence. Une telle solution présente une garantie probablement-approximativement-correcte pour le problème ERM indépendamment du fait que ce dernier possède ou non une solution. Pour un ensemble de données fixe, le risque empirique s'avère être une variable aléatoire sous-gaussienne lorsque les modèles sont échantillonnés à partir de la solution au problème ERM-RER. Les capacités de généralisation de la solution au problème ERM-RER (l'algorithme de Gibbs) sont étudiées via la sensibilité de la valeur espérée du risque empirique aux déviations d'une telle solution vers des mesures de probabilité alternatives. Enfin, un lien intéressant entre la sensibilité, l'erreur de généralisation et l'information lautum est établi.

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

Contents

1 Introduction

In statistical 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. See for instance [2–4] and [5]. Instead of additional statistical assumptions on the datasets, which are typical in Bayesian methods [6], relative entropy regularization requires a reference probability measure on the set of models, which is external to the ERM problem. Often, such a reference measure represents prior knowledge or side information and is chosen for guiding the search of models towards those inducing low empirical risks with high probability over seen and unseen datasets. From this perspective, the reference measure can be seen as an additional degree of freedom to improve the generalization capabilities of machine learning algorithms based on ERM-RER, e.g, Gibbs algorithms [4, 7–14] and [15]. This new degree of freedom is one of the main motivations for regularizing the ERM problem using relative entropy, or more generally, any f-divergence regularization, as discussed in [16, 17] and [18]. Beyond probability measures, as shown in this paper, the reference measure can be any σ-finite measure with arbitrary support. The flexibility introduced by this generalization 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, cf. [19], or cannot be represented by probability distributions, e.g., equal preferences among elements of infinite countable sets. For some specific choices of σ-finite reference measures, the ERM-RER boils down to particular cases of special interest: (*i*) the information-risk minimization problem presented in [20]; (*ii*) the ERM with differential entropy regularization (ERM-DiffER); and (*iii*) the ERM with discrete entropy regularization (ERM-DisER). See for instance [21] and references therein. From this perspective, the proposed ERM-RER formulation yields a unified mathematical framework that comprises a large class of problems.

When the reference measure is a probability measure, the solution to the ERM-RER problem is known to be unique and correspond to a Gibbs probability measure. Such a Gibbs probability measure has been studied using measure theoretic and information theoretic notions in [8, 20, 22–29]; statistical physics in [2]; PAC (Probably Approximatively Correct)-Bayesian learning theory in [30–33]; and proved to be of particular interest in classification problems in [4, 11, 17, 34–36] and [37]. In the general case in which the reference is a σ-finite measure, a solution to the ERM-RER problem does not always exist. Nonetheless, if it exists, it is shown to be a unique Gibbs probability despite the fact that its partition function is defined with respect to a σ-finite measure. The condition for the existence is mild and is always satisfied when the reference measure is a probability measure, as highlighted above. Interestingly, such a solution is mutually absolutely continuous with the reference measure in most practical cases. Interestingly, most of the properties known for the classical ERM-RER problem are shown to hold in the most general case. For instance,

the empirical risk observed when models are sampled from the ERM-RER-optimal probability measure is a sub-Gaussian random variable that exhibits a PAC guarantee for the ERM problem without regularization.

When the solution to the ERM-RER problem is used to sample models to label unseen patterns, the process is known as the Gibbs algorithm. One of the traditional performance metrics to evaluate the generalization capabilities of the Gibbs algorithm is the generalization error. When the reference measure is a probability measure, a closed-form expression for the generalization error of the Gibbs algorithm is presented in [8], while upper bounds have been derived in [15, 20, 27–33, 38–51], and references therein. In this work, a new performance metric coined *sensitivity*, which quantifies the variations of the expected empirical risk due to deviations from the solution of the ERM-RER problem is introduced. The sensitivity is defined as the difference between two quantities: (a) The expectation of the empirical risk with respect to the solution to the 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 ERM-RER-optimal measure. Such bound allows providing lower and upper bounds on the expected empirical risk after a deviation from the ERM-RER-optimal measure towards an alternative probability measure. More interestingly, the expectation (with respect to the probability distribution of the datasets) of the sensitivity to deviations to a specific measure is shown to be equal to the generalization error of the Gibbs algorithm. Using this result, the closed-form expression for the generalization error of the Gibbs algorithm presented in [8] is shown to hold even in the case in which the reference measure is a σ-finite measure. Moreover, the generalization error is shown to be upper bounded by a term that is proportional to the squared-root of the lautum information between the models and the datasets, cf. [52]. This bound is reminiscent of the result in [29, Theorem 1] in which a similar bound is presented using the mutual information instead of the lautum information. While [29, Theorem 1] follows immediately from the variational representation of relative entropy, c.f., [53, Lemma 4.18 (Transportation Lemma)], the new result follows from the fact that the empirical risk when models are sampled from the ERM-RER-optimal probability measure is a sub-Gaussian random variable. Interestingly, the new upper-bound does not require any of the conditions in [29, Theorem 1].

The remainder of this work is organized as follows. Section 2 introduces two optimization problems: the ERM and the ERM-RER. The asymmetry of the relative entropy is analyzed in the context of the ERM-RER and two variants, coined Type-I and Type-II, are distinguished. The former considers the case in which the regularization is the relative entropy of the optimization measure with respect to the reference measure. The latter considers a regularization by the relative entropy of the reference measure with respect to the optimization measure. Section 3 presents the solution to the ERM-RER problem in the general case and introduces its main properties. Section 4 introduces two new classes

of reference measures and the solution of the ERM-RER problem is shown to exhibit different properties for each class. This section ends by studying the ERM-RER problem in the special case in which the reference measure is a Gibbs probability measure. This special case exhibits a solution that is identical to the solution to an ERM-RER problem whose reference measure is the same used to build the above mentioned Gibbs measure. Section 5 studies the properties of the log-partition function of the ERM-RER-optimal probability measure. The first, second, and third cumulants of the empirical risk when the models are sampled from the ERM-RER-optimal measure and the reference measure are respectively characterized. Section 6 and Section 7 study the properties of the expectation and variance of the empirical risk when the models are sampled from the ERM-RER-optimal probability measure. These mean and variance are compared with the mean and variance of the empirical risk when models are sampled from the reference measure. Section 8 introduces several explicit expressions for the cumulant generating function of the empirical risk when the models are sampled from the ERM-RER-optimal measure. Using these equivalent expressions, it is shown that empirical risk is a sub-Gaussian random variable when models are sampled from the ERM-RER-optimal measure. Section 9 describes the monotonic concentration of the ERM-RER-optimal probability measure when the regularization factor tends to zero. Section 10 show that the empirical risk when the models are sampled from the ERM-RER-optimal probability measure exhibits a PAC-type guarantee with respect to the ERM problem without regularization. Finally, Section 11 studies the sensitivity of the expected empirical risk with respect to deviations from the ERM-RER-optimal measure to alternative measures and shows connections with the generalization error and the lautum information. Section 12 ends this work with conclusions and a discussion on the results.

2 Empirical Risk Minimization (ERM)

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

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

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

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

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

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

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

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

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

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

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

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

2.1 Notation and Main Assumptions

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 $(\Omega, \mathscr{F})$. Given a measure $Q \in \triangle(\Omega, \mathscr{F})$, the subset $\triangle_Q(\Omega, \mathscr{F})$ of $\triangle(\Omega, \mathscr{F})$ contains all σ-finite measures that are absolutely continuous with respect to the measure Q. Alternatively, the subset $\bigtriangledown_Q(\Omega, \mathscr{F})$ of $\triangle(\Omega, \mathscr{F})$ contains all probability measures P such that Q is absolutely continuous with respect to P. Given a set $\mathcal{A} \subset \mathbb{R}^d$, the Borel σ-field over $\mathcal{A}$ is denoted by $\mathscr{B}(\mathcal{A})$.

The main assumption adopted in this work is that the function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is measurable with respect to the Borel measurable spaces $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $([0, +\infty], \mathscr{B}([0, +\infty]))$.

2.2 Relative Entropy Extended to σ-Finite Measures

In this work, the *relative entropy*, which is usually defined for probability measures, is extended to σ-finite measures.

Definition 2.1 (Generalized Relative Entropy)**.** *Given two σ-finite measures P and Q on the same measurable space, such that P is absolutely continuous with respect to Q, the relative entropy of P with respect to Q is*

$$D(P\|Q) = \int \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \mathrm{d}Q(x), \tag{6}$$

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

The relative entropy exhibits a property often referred to as the *information inequality* [54, Theorem 2.6.3] in the case of probability measures on $(\Omega, \mathscr{F})$, with Ω a countable set. The following theorem explores this property in a more general scenario.

Theorem 2.1. *If P and Q are both probability measures on a general measurable space $(\Omega, \mathscr{F})$, then,*

$$D\left(P\|Q\right)\geqslant 0, \tag{7}$$

with equality if and only if P and Q are identical.

Proof: Consider the function $f : [0, \infty) \to \mathbb{R}$ such that for all $x \in (0, +\infty)$, $f(x) = x\log(x)$ and $f(0) = 0$. Note that f is strictly convex. If P and Q are both probability measures on the measurable space $(\Omega, \mathscr{F})$, the following holds:

$$\begin{aligned} D\left(P\|Q\right) &= \int \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \mathrm{d}Q(x) && (8)\\ &= \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \mathrm{d}Q(x) && (9)\\ &\geqslant f\left(\int \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \mathrm{d}Q(x)\right) && (10)\\ &= f\left(1\right) && (11)\\ &= 0, && (12) \end{aligned}$$

where the inequality (11) follows from Jensen's inequality [55, Section 6.3.5]. Equality in (11) holds if and only if for all $x \in \operatorname{supp} Q$, $\frac{\mathrm{d}P}{\mathrm{d}Q}(x) = 1$, which implies that both P and Q are identical. This completes proof. ■

If Q is not a probability measure, then it might be observed that $D\left(P\|Q\right) < 0$. Consider for instance the case in which P is a zero-mean Gaussian probability measure with variance σ^2 and Q is the Lebesgue measure on $(\mathbb{R}, \mathscr{B}\left(\mathbb{R}\right))$. Hence, the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Gaussian probability density function such that for all $x \in \mathbb{R}$,

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right). \tag{13}$$

Under this assumption, the relative entropy of P with respect to Q is the negative of the differential entropy of P. That is,

$$D\left(P\|Q\right) = -\frac{1}{2}\log\left(2\pi\epsilon\sigma^2\right), \tag{14}$$

with ϵ being Néper's constant. See for instance [54, Example 8.1.2]. Hence, $D\left(P\|Q\right)$ is negative for all $\sigma^2 \in \left(\frac{1}{2\pi\epsilon}, +\infty\right)$ and nonnegative for all $\sigma^2 \in \left(0, \frac{1}{2\pi\epsilon}\right]$. Finally, note also that

$$\begin{aligned} \lim_{\sigma^2 \to 0} D\left(P\|Q\right) &= +\infty, \text{ and} && (15)\\ \lim_{\sigma^2 \to +\infty} D\left(P\|Q\right) &= -\infty. && (16) \end{aligned}$$

A central observation from (14) is that the equality $D\left(P\|Q\right) = 0$ does not necessarily imply that P and Q are identical measures. For instance, when

$\sigma^2 = \frac{1}{2\pi\epsilon}$ in (15), it holds that $D(P\|Q) = 0$, while P is a Gaussian probability measure and Q is the Lebesgue measure.

The following property, known for the case of probability measures as the *joint-convexity of the relative entropy*, is extended by the following theorem.

Theorem 2.2. *Let P_1 and P_2 be two probability measures and Q_1 and Q_2 be two σ-finite measures, all on the same measurable space. For all $i \in \{1, 2\}$, let P_i be absolutely continuous with respect to Q_i. Then, for all $\lambda \in [0, 1]$,*

$$\begin{aligned} &D(\lambda P_1 + (1-\lambda)P_2 \| \lambda Q_1 + (1-\lambda)Q_2) \\ \leqslant &\lambda D(P_1\|Q_1) + (1-\lambda) D(P_2\|Q_2). \end{aligned} \tag{17}$$

Equality in (17) *holds if and only if $P_1 = P_2$ and $Q_1 = Q_2$.*

Proof: The proof is presented in Appendix A. ∎

2.3 ERM with Relative Entropy Regularization

Given a dataset, the *expected empirical risk* induced by a probability measure $P \in \Delta(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ is defined as follows.

Definition 2.2 (Expected Empirical Risk)**.** *Let P be a probability measure in $\Delta(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. The expected empirical risk with respect to the dataset $\boldsymbol{z}$ in* (1) *induced by the measure P is*

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}), \tag{18}$$

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

The 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 ERM-RER problem, with parameters Q and λ, consists of the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(P\|Q), \tag{19a}$$

$$\text{s. t.} \quad \int \mathrm{d}P(\boldsymbol{\theta}) = 1, \tag{19b}$$

where the dataset $\boldsymbol{z}$ is in (1), and the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (18).

2.4 Type-I and Type-II Relative Entropy Regularization

The optimization problem in (19) is coined Type-I ERM-RER in [56] in the aim of distinguishing it from the optimization problem

$$\min_{P \in \triangledown_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda D(Q\|P), \tag{20a}$$

$$\text{s. t.} \quad \int \mathrm{d}P(\boldsymbol{\theta}) = 1, \tag{20b}$$

which is coined Type-II ERM-RER.

The Type-II ERM-RER problem in (20), when Q is a probability measure, exhibits a solution that is identical to the solution to the following Type-I ERM-RER problem [56, Theorem 1]:

$$\min_{P\in\triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \int \log(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\mathrm{d}P(\boldsymbol{\nu})+D(P\|Q), \tag{21a}$$

$$\text{s. t.} \qquad \int \mathrm{d}P\left(\boldsymbol{\theta}\right)=1, \tag{21b}$$

where β is a constant chosen to satisfy

$$\int \frac{\lambda}{\beta+\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)}\mathrm{d}Q(\boldsymbol{\nu})=1. \tag{21c}$$

Essentially, by appropriately transforming the objective function, an equivalence can be established between Type-I and Type-II ERM-RER problems. Hence, without loss of generality, the remainder of this work focuses exclusively on Type-I ERM-RER, which is simply referred to as ERM-RER.

3 The Solution to the ERM-RER Problem

The solution to the ERM-RER problem in (19) 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 $t\in\mathbb{R}$,

$$K_{Q,\boldsymbol{z}}\left(t\right)=\log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})\right), \tag{22}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (3). Second, the set $\mathcal{K}_{Q,\boldsymbol{z}}\subset(0,+\infty)$, which is defined by

$$\mathcal{K}_{Q,\boldsymbol{z}}\triangleq\left\{s\in(0,+\infty):\ K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)<+\infty\right\}. \tag{23}$$

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 (1) and the σ-finite measure Q in (19) are highlighted.

The following lemma describes the set $\mathcal{K}_{Q,\boldsymbol{z}}$.

Lemma 3.1. *The set $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (23) *is a convex subset of* $\mathbb{R}$. *If the measure Q in* (19) *is a probability measure, then, the set $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (23) *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 ERM-RER problem in (19) is presented by the following theorem.

Theorem 3.1. *If $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (23)*, the solution to the optimization problem in* (19) *is a unique probability measure, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which 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 (3) *and the function $K_{Q,\boldsymbol{z}}$ is defined in* (22)*.*

Proof: The proof is presented in Appendix C. ■

Contrary to the ERM problem in (4), which does not necessarily possess a solution, the ERM-RER problem in (19) always possess a solution when Q is a probability measure. This is essentially because the set $\mathcal{K}_{Q,\boldsymbol{z}}$ is the set of all positive reals (Lemma 3.1), and thus, the condition $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ is always verified. On the contrary, when Q is a σ-finite measure, the solution to the ERM-RER problem in (19) depends on whether $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$. If the solution exists, it is $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), which is a unique probability measure referred to as the Gibbs measure [57]. The function $K_{Q,\boldsymbol{z}}$ is often referred to as the *log-partition function*, see for instance, [58, Section 7.3.1].

The following lemma shows that the Radon-Nikodym derivative in (25) is both nonnegative and finite.

Lemma 3.2. *The Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in* (25) *satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ that*

$$0 \leqslant \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) < +\infty, \tag{26}$$

where 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 D. ■

An immediate consequence of Lemma 3.2 is the equality

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = +\infty\}\right) = 0.$$

Theorem 3.1 shows that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to the measure Q. The following lemma shows that the converse is also true if and only if the set of models that lead to an infinite empirical risk exhibit zero measure with respect to the reference measure Q.

Lemma 3.3. *The σ-finite measure Q and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (25) *are mutually absolutely continuous if and only if*

$$Q\left(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = +\infty\}\right) = 0. \tag{27}$$

Proof: The proof is presented in Appendix E. ■

The relevance of Lemma 3.3 is that it shows that if $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, 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 (19), under the assumption in (27). Such an assumption is trivially true when the function ℓ in (2) is bounded.

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 λ.

Lemma 3.4. *For all $(\alpha, \beta) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), assume that the probability measures $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ 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 F. ■

Particular assumptions on the set $\mathcal{M}$ and the reference measure Q lead to well-known instances of the ERM-RER problem in (19), as discussed hereunder.

3.1 Examples

Three examples are of particular interest: (*a*) The set $\mathcal{M} \subset \mathbb{R}^d$ is countable and the measure Q is the counting measure in $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, which leads to the ERM-DisER problem; (*b*) The set $\mathcal{M}$ is an uncountable subset of $\mathbb{R}^d$, and Q is the Lebesgue measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, which leads to the ERM-DiffER problem; and (*c*) The set $\mathcal{M}$ and the measure Q form a Borel probability measure space $(\mathcal{M}, \mathscr{B}(\mathcal{M}), Q)$, which leads to the information-risk minimization problem.

3.1.1 ERM with Discrete Entropy Regularization

When the set $\mathcal{M} \subset \mathbb{R}^d$ is countable and the σ-finite measure Q in (19) is the counting measure in $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, given a probability measure P on the same measurable space, the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is a probability mass function, denoted by p. Thus, the relative entropy $D(P\|Q)$ is equivalent to the negative of the discrete entropy induced by p [54, Chapter 2], denoted by $H(p)$. In this case, the ERM-RER in (19) can be re-written as the following ERM-DisER problem:

$$\min_{p} \sum_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, p(\boldsymbol{\theta}) - \lambda H(p), \tag{28}$$

where the optimization domain in (28) is the set of probability mass functions that can be defined over the measure space $\triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. In this special case, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) whose probability mass function is the

solution to the ERM-DisER problem in (28) satisfies

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\sum\limits_{\boldsymbol{\nu}\in\mathcal{M}} \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}, \tag{29}$$

which describes the discrete Gibbs probability measure on $\triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, with temperature parameter λ, and energy function $\mathsf{L}_{\boldsymbol{z}}$ in (3).

3.1.2 ERM with Differential Entropy Regularization

When $\mathcal{M} \subseteq \mathbb{R}^d$ is uncountable and the σ-finite measure Q in (19) is the Lebesgue measure in $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, for all probability measures $P \in \Delta_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is a probability density function, denoted by g. Thus, the relative entropy $D(P\|Q)$ is equivalent to the negative of the differential entropy induced by g [54, Chapter 8], denoted by $h(g)$. In this special case, the ERM-RER in (19) can be re-written as the following ERM-DiffER problem:

$$\min_{g} \int_{\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, g(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} - \lambda h(g), \tag{30}$$

where the optimization domain in (30) is the set of probability density functions that can be defined over the measure space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) whose probability density function is the solution to the ERM-RER problem in (30) satisfies

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int_{\mathcal{M}} \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}\boldsymbol{\nu}}, \tag{31}$$

which describes the absolutely continuous Gibbs probability measure with temperature parameter λ and energy function $\mathsf{L}_{\boldsymbol{z}}$ in (3).

Both, the ERM-DiffER and ERM-DisER problems are closely related to those typically arising while using Jayne's maximum entropy principle [59, 60] for classification problems such as those in [34–36], and [61].

3.1.3 Information-Risk Minimization

When Q is a probability measure, the ERM-RER in (19) is equivalent to the *information-risk minimization* (IRM) problem in [20]. The IRM problem in (19) is known to possess a unique solution equal to the Gibbs probability measure in (25), as independently shown in [20, 29, 57, 62, 63] and [64].

3.2 Bounds on the Radon-Nikodym Derivative

The Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) is bigger for models inducing smaller empirical risks, as shown by the following corollary of Theorem 3.1.

Corollary 3.1. *The Radon-Nikodym derivative* $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (25) *satisfies for all* $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \operatorname{supp} Q \times \operatorname{supp} Q$*, with* $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) \leqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)$*, that*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1) \leqslant \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2), \tag{32}$$

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

The intuition that follows from corollary 3.1 is that under the assumption that the ERM problem in (4) possesses a solution in the support of the reference measure, i.e., $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$ is not empty, with $\mathcal{T}(\boldsymbol{z})$ in (5), the maximum of the function $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) is achieved by the models in $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$. 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 probability mass function in (29) or the probability density function in (31), Corollary 3.1 shows that the elements of the set $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$ are the *modes* of the corresponding probability density function or probability mass function.

3.3 Asymptotes of the Radon-Nikodym Derivative

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 the regulariation factor increases, i.e., $\lambda \to +\infty$ and the reference measure Q is a probability measure.

Lemma 3.5. *Let the measure* Q *in* (19) *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*

$$\lim_{\lambda \to +\infty} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 1. \tag{33}$$

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

$$\lim_{\lambda \to +\infty} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \lim_{\lambda \to +\infty} \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\displaystyle\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\nu})} \tag{34}$$

$$= \frac{1}{\displaystyle\int \mathrm{d}Q(\boldsymbol{\nu})} \tag{35}$$

$$= 1, \tag{36}$$

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

Lemma 3.5 unveils the fact that, when Q is a probability measure, in the limit when $\lambda \to +\infty$, both probability measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q are identical. This is consistent with the fact that when λ tends to infinity, the optimization problem

in (19) 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 (Theorem 2.1). Such intuition breaks when the reference measure is a σ-finite measure, but not a probability measure. In such a case, the relative entropy term in (19) might be negative and a minimum might not exist. See for instance, the case of the relative entropy between a Gaussian measure and the Lebesgue measure in (14), which satisfies (16).

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

$$\mathcal{L}_{\boldsymbol{z}}\left(\delta\right) \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \leqslant \delta\right\}, \tag{37}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3) and $\delta \in [0, +\infty)$. In particular consider the nonnegative real

$$\delta^{\star}_{Q,\boldsymbol{z}} \triangleq \inf\left\{\delta \in [0, +\infty) : Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0\right\}. \tag{38}$$

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

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \triangleq \left\{\boldsymbol{\theta} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}. \tag{39}$$

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* (39) *and Q the σ-finite measure in* (19)*, 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*

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) = \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} \mathbb{1}_{\left\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right\}}. \tag{40}$$

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

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) = \begin{cases} +\infty & \text{if } \boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \\ 0 & \text{otherwise.} \end{cases} \tag{41}$$

Proof: The proof is presented in Appendix G. ∎

Consider that $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) > 0$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (39). Under this assumption, from Lemma 3.6, it holds that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ asymptotically concentrates on the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ when λ tends to zero from the right. More specifically,

note that for all measurable sets $\mathcal{A} \subseteq \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$, it holds that

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) = \lim_{\lambda \to 0^+} \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{42}$$

$$= \int \lim_{\lambda \to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{A}\}} \mathrm{d}Q(\boldsymbol{\theta}) \tag{43}$$

$$= \int \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\}} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{A}\}} \mathrm{d}Q(\boldsymbol{\theta}) \tag{44}$$

$$= \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} \int \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{A}\}} \mathrm{d}Q(\boldsymbol{\theta}) \tag{45}$$

$$= \frac{Q(\mathcal{A})}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)}, \tag{46}$$

where the equality in (43) follows from Lemma 3.2 and the dominated convergence theorem [55, Theorem 2.6.9]. The equality in (44) follows from Lemma 3.6. In the particular case in which $\mathcal{A} = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (46), 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 Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (25) is a constant among the elements of the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$. This can be assimilated to 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, as previously highlighted in [22–24] and [25]. 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 $\frac{1}{|\mathcal{L}^{\star}_{Q,\boldsymbol{z}}|}$.

Consider now that $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (39). Under this assumption, in the asymptotic regime when $\lambda \to 0$, the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is not a probability measure but either the trivial measure or the infinite measure. This is typically the case in which $\mathcal{M} = \mathbb{R}^d$, the measure Q is absolutely continuous with respect to the Lebesgue measure, and the solution to the ERM problem in (4) has a unique solution on the support of Q, i.e., $\mathcal{L}^{\star}_{Q,\boldsymbol{z}} = \mathcal{T}(\boldsymbol{z})$ and $|\mathcal{T}(\boldsymbol{z})| = 1$, which implies $Q(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}) = 0$.

An interesting question, which is left out of the scope of this paper, is the rate at which $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ converges to such limiting measure. The interested reader is referred to [22, 25], and references therein.

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* (39) *satisfy,*

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

Proof: The proof is presented in Appendix H. ■

Note that if the ERM problem in (4) possesses at least one solution and such solution is within the support of the measure Q, i.e., $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q \neq \emptyset$, then, when λ tends to zero from the right, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ asymptotically concentrates on the solution (or the set of solutions within the support of Q) to the ERM problem in (4). Alternatively, in the case in which $\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \mathcal{T}(\boldsymbol{z}) = \emptyset$, 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 (4). This observation leads to the introduction to two new classes of reference measures, namely, *coherent* and *consistent* measures, in the following section.

4 Reference Measures

This section introduces two classes of reference measures, namely *coherent* and *consistent* measures, and discusses the special case of Gibbs reference measures.

4.1 Coherent and Consistent Reference Measures

A class of reference measures of particular importance to establish connections between the set of solutions to the ERM problem in (4) and the solution to the ERM-RER problem in (19) is that of *coherent* measures. Let $\rho^{\star} \geqslant 0$ be the infimum of the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (3). That is,

$$\rho^{\star} \triangleq \inf\{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{M}\}. \tag{48}$$

Using this notation, coherent measures are defined as follows.

Definition 4.1 (Coherent Measures)**.** *The* σ*-finite measure* Q *in* (19) *is said to be coherent if, for all* $\delta \in (\rho^{\star}, +\infty)$*, with* $\rho^{\star}$ *in* (48)*, it holds that*

$$Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0, \tag{49}$$

where the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ *is defined in* (37)*.*

When the reference measure Q in the EMR-RER problem in (19) is a coherent measure, it holds that for all $\delta > \rho^{\star}$, the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ in (37) exhibits positive probability with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25). The following lemma highlights this observation.

Lemma 4.1. *The probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (25) *satisfies for all* $\delta \in (\rho^\star, +\infty)$*, with* $\rho^\star$ *in* (48)*, that*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right)>0, \tag{50}$$

with $\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)$ *in* (37)*, if and only if the* σ*-finite measure* Q *in* (19) *is coherent.*

Proof: The proof is presented in Appendix I. ■

Under the assumption that the ERM problem in (4) possesses a solution, it holds that

$$\min_{\theta\in\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)=\inf\{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) : \boldsymbol{\theta}\in\mathcal{M}\}. \tag{51}$$

Hence, when the σ-finite measure Q in (19) is coherent, then

$$\delta^\star_{Q,\boldsymbol{z}} = \rho^\star, \tag{52}$$

with $\delta^\star_{Q,\boldsymbol{z}}$ in (38) and $\rho^\star$ in (48), which implies that

$$\mathcal{L}^\star_{Q,\boldsymbol{z}} \subseteq \mathcal{T}\left(\boldsymbol{z}\right), \tag{53}$$

with $\mathcal{T}\left(\boldsymbol{z}\right)$ in (5) and $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (39). This observation, together with Lemma 3.7, leads to the following result.

Lemma 4.2. *Assume that the ERM problem in* (4) *possesses a solution. Then, the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (25) *and the sets* $\mathcal{T}\left(\boldsymbol{z}\right)$ *in* (5) *and* $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ *in* (39) *satisfy*

$$\lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\cap\mathcal{T}\left(\boldsymbol{z}\right)\right)=1, \tag{54}$$

if and only if the σ*-finite measure* Q *in* (19) *is coherent.*

Proof: The proof follows by observing that if Q is a coherent measure and the ERM problem in (4) possesses a solution, the inclusion in (53) holds. Thus, from Lemma 3.7, the equality in (54) holds. Alternatively, when the measure Q in (19) is noncoherent, then $\delta^\star_{Q,\boldsymbol{z}} > \rho^\star$, which implies that $\mathcal{L}^\star_{Q,\boldsymbol{z}}\cap\mathcal{T}\left(\boldsymbol{z}\right) = \emptyset$. Hence, from Lemma 3.7, it follows that

$$\lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\cap\mathcal{T}\left(\boldsymbol{z}\right)\right)=0, \tag{55}$$

and completes the proof. ■

The relevance of coherent measures in ERM-RER problems is well highlighted by Lemma 4.2. Essentially, when the ERM problem in (4) possesses at least one solution, the concentration of the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) on the set (or a subset) of solutions to the ERM problem in (4) occurs asymptotically when λ tends to zero from the right, if only if the reference measure Q in (19) is coherent. Nonetheless, such asymptotic concentration is not a guarantee that for strictly positive values of λ in (19), the set $\mathcal{T}\left(\boldsymbol{z}\right)$ in (5) and the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) satisfy $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{T}\left(\boldsymbol{z}\right)\right) > 0$. In order to ensure this, another class of reference measures, known as *consistent measures*, is introduced.

Definition 4.2 (Consistent Measure)**.** *The σ-finite measure Q in* (19) *is said to be consistent if $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) > 0$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in* (39)*.*

Note that every consistent measure is not necessarily coherent. For instance, if Q is consistent but $\delta^{\star}_{Q,\boldsymbol{z}} > \rho^{\star}$, with $\rho^{\star}$ in (48) and $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38), then, for all $\delta \in (\rho^{\star}, \delta^{\star}_{Q,\boldsymbol{z}})$, it follows that $Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) = 0$, and thus, Q is not coherent. Alternatively, every coherent measure is not necessarily consistent. For instance, if $\left|\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right| = 1$ and Q is coherent and absolutely continuous with respect to the Lebesgue measure, it follows that $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$, and thus, Q is not consistent.

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

Lemma 4.3. *The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (25) *and the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in* (39) *satisfy*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)>0, \tag{56}$$

if and only if the σ-finite measure Q in (19) *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 respect to Q, it holds that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$. 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 $\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) < +\infty$ and thus, from Lemma 3.2, it follows that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) > 0$. Hence,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = \int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{57}$$

$$= \int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right) > 0, \tag{58}$$

which completes the proof. ■

The following lemma highlights a central property of consistent measures when the ERM problem in (4) possesses a solution.

Lemma 4.4. *Assume that the ERM problem in* (4) *possesses a solution in the support of Q. The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in* (25) *and the sets $\mathcal{T}\left(\boldsymbol{z}\right)$ in* (5) *and $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in* (39) *satisfy*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \mathcal{T}\left(\boldsymbol{z}\right)\right)>0, \tag{59}$$

if and only if the σ-finite measure Q in (19) *is consistent.*

Proof: The proof follows from Lemma 4.3 by noticing that when the ERM problem in (4) possesses a solution in the support of Q, the inclusion in (53) holds. ■

The distinction between coherent and consistent measures becomes more evident under certain conditions. Consider the case in which $\mathcal{M}$ is finite. In this case, if the solution to the ERM problem in (4) is in the support of the σ-finite measure Q, then Q is both coherent and consistent. This is essentially because all measurable singletons (models) in $\operatorname{supp} Q$ exhibit positive measure with respect to Q. Alternatively, if the solution to the ERM problem in (4) is not in the support of Q, then Q is consistent but not coherent. Consider the case in which $\mathcal{M}$ is the set $\mathbb{R}^d$; the loss function ℓ in (2) is continuous; and the ERM problem in (4) admits a unique solution. In this case, any probability measure Q absolutely continuous with respect to the Lebesgue measure is a coherent measure, but it is not a consistent measure. Alternatively, if the set of solutions to the ERM problem in (4) exhibits positive Lebesgue measure, then, the measure Q is both coherent and consistent.

4.2 Gibbs Reference Measures

In model selection, a natural idea is to proceed by successive approximations in the seek of lower computation complexity. From this perspective, one might wonder whether the solution to a current instance of an ERM-RER problem might serve as reference measure for the next instance. In this section, it is shown that this yields no benefit. Composing two successive ERM-REM problems boils down to a unique ERM-RER problem with the initial reference measure and a particular regularization factor. Under the assumption that $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), the problem of interest is:

$$\min_{P \in \triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}(P) + \alpha D\left(P \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{60a}$$

$$\text{s. t.} \qquad \int \mathrm{d}P(\boldsymbol{\theta}) = 1, \tag{60b}$$

where $\alpha > 0$; the reference measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which satisfies (25), is the solution of the ERM-RER problem in (19); and the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (18). From Theorem 3.1, the solution to the ERM-RER problem in (60), which is denoted by $P^{\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\alpha\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, satisfies for all $\theta \in \operatorname{supp} Q$ that

$$\frac{\mathrm{d}P^{\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\alpha\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) = \exp\left(-K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) - \frac{1}{\alpha}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right). \tag{61}$$

The log-partition functions $K_{Q,\boldsymbol{z}}$ in (22) and $K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}$ in (61) are strongly related, as shown by the following lemma.

Lemma 4.5. *The functions $K_{Q,\boldsymbol{z}}$ in (22) and $K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}$ in (61) satisfy for all $t \in \mathbb{R}$,*

$$K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(t) = K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right). \tag{62}$$

Moreover, for all $t \leqslant 0$,

$$K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(t) \leqslant 0. \tag{63}$$

Proof: The proof of (62) relies on the fact that for all

$$t \in \left\{ \nu \in \mathbb{R} : K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(\nu) < \infty \right\},$$

the function $K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}$ in (61) satisfies

$$\begin{aligned}
K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(t) &= \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) && (64)\\
&= \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) \right) && (65)\\
&= \log\left(\int \exp\left(\left(t - \frac{1}{\lambda}\right) \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) \right) && (66)\\
&= \log\left(\int \exp\left(\left(t - \frac{1}{\lambda}\right) \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) && (67)\\
&= K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), && (68)
\end{aligned}$$

where the equality in (66) follows from (25). Moreover, from Lemma 5.2, it follows that the function $K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}$ is continuous and nondecresing. Let $s^\star \in \mathbb{R} \cup \{+\infty\}$ be defined by

$$s^\star \triangleq \sup \left\{ \nu \in \mathbb{R} : K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(\nu) < \infty \right\}. \tag{69}$$

If $s^\star = +\infty$, then for all $t \in \mathbb{R}$, $K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(t) < +\infty$, and the proof of (62) is completed.

Alternatively, if $s^\star < +\infty$, it follows that for all $t > s^\star$, $K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(t) = +\infty$, which implies that $K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) = +\infty$, as the function $K_{Q,\boldsymbol{z}}$ is also continuous (Lemma 5.2) and $K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < \infty$ (due to the choice of λ). Hence, in this case, the equality in (62) is of the form $+\infty = +\infty$. This completes the proof of (62).

The proof of (63) follows by noticing that for all $t \leqslant 0$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \leqslant 1$. Hence,

$$\begin{aligned}
K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}(t) &= \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) && (70)\\
&\leqslant \log\left(\int \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) && (71)\\
&= 0, && (72)
\end{aligned}$$

which completes the proof. ∎

The following lemma establishes that the solution to the ERM-RER problem in (60) is identical to the solution to another ERM-RER problem of the form

$$\min_{P\in\triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}\left(P\right)+\left(\frac{1}{\frac{1}{\alpha}+\frac{1}{\lambda}}\right)D\left(P\|Q\right), \tag{73a}$$

$$\text{s. t.} \qquad \int \mathrm{d}P\left(\boldsymbol{\theta}\right)=1, \tag{73b}$$

with $\lambda\in\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), and whose solution, denoted by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{\alpha}}\right)}$, satisfies for all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{\alpha}}\right)}}{\mathrm{d}Q}(\boldsymbol{\theta})=\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}-\frac{1}{\alpha}\right)-\left(\frac{1}{\lambda}+\frac{1}{\alpha}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right). \tag{74}$$

The formal statement is as follows.

Lemma 4.6. *Let $\alpha\in(0,+\infty)$ and $\lambda\in\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23). Then, the probability measures $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\alpha\right)}$ in (61) and $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{\alpha}}\right)}$ in (74) are identical.*

Proof: For all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\alpha\right)}}{\mathrm{d}Q}(\boldsymbol{\theta})=\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\alpha\right)}}{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}}(\boldsymbol{\theta})\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) \tag{75}$$

$$=\exp\left(-K_{P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \tag{76}$$

$$=\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}-\frac{1}{\lambda}\right)-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \tag{77}$$

$$=\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{\alpha}}\right)}}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{78}$$

where the equality in (75) follows from the fact that the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\alpha\right)}$ is absolutely continuous with respect to $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ and $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ is absolutely continuous with respect to the measure Q; the equality in (76) follows from Lemma 4.5; and the equality in (78) follows from Theorem 3.1.

For all measurable subsets $\mathcal{A}$ of $\mathcal{M}$, the following holds:

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\alpha\right)}(\mathcal{A}) = \int_{\mathcal{A}} \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\alpha\right)}}{\mathrm{d}Q}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) \tag{79}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{\alpha}}\right)}}{\mathrm{d}Q}\mathrm{d}Q(\boldsymbol{\theta}) \tag{80}$$

$$= \int_{\mathcal{A}} \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{\alpha}}\right)}(\boldsymbol{\theta}) \tag{81}$$

$$= P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{\alpha}}\right)}(\mathcal{A}), \tag{82}$$

where the equality in (80) follows from (78). This completes the proof. ■

The following theorem establishes a relation between the solutions to the following optimization problems

$$\min_{P\in\triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \quad \mathsf{R}_{\boldsymbol{z}}\left(P\right), \tag{83a}$$

$$\text{s. t.} \quad D\left(P\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) \leqslant c, \text{ and} \tag{83b}$$

$$\int \mathrm{d}P(\boldsymbol{\theta}) = 1, \tag{83c}$$

and

$$\min_{P\in\triangle_Q(\mathcal{M},\mathscr{B}(\mathcal{M}))} \quad \mathsf{R}_{\boldsymbol{z}}\left(P\right) + \omega D\left(P\|Q\right), \tag{84a}$$

$$\text{s. t.} \quad \int \mathrm{d}P\left(\boldsymbol{\theta}\right) = 1, \tag{84b}$$

with $c > 0$ and $\omega \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), two constants; $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ the probability measure in (25); and $\mathsf{R}_{\boldsymbol{z}}$ the functional in (18).

From Theorem 3.1, the solution to the ERM-RER problem in (84), which is denoted by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\omega)}$, satisfies for all $\theta \in \operatorname{supp} Q$ that

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\omega)}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\omega}\right) - \frac{1}{\omega}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right), \tag{85}$$

where the function $K_{Q,\boldsymbol{z}}$ is in (22).

The following theorem formalizes the relation between both optimization problems.

Theorem 4.1. *Assume that c and ω in* (83) *and* (84) *satisfy*

$$D\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\omega)}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) = c, \tag{86}$$

with $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *and* $P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *being the probability measures in* (25) *and* (85)*, respectively. Then, the solution to the optimization problem in* (83) *is the probability measure* $P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$*.*

Proof: The proof is presented in Appendix J. ■

5 The Log-Partition Function

This section introduces some properties of the log-partition function $K_{Q,\boldsymbol{z}}$ in (22) using the notion of *separable* empirical risk functions.

5.1 Separable Empirical Risk Functions

Separable empirical risk functions are defined with respect to a measure $P \in \triangle(\mathcal{M})$.

Definition 5.1 (Separable Empirical Risk Function)**.** *The empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *is said to be separable with respect to a* σ*-finite measure* $P \in \triangle(\mathcal{M})$*, if there exist a positive real* $c > 0$ *and two subsets* $\mathcal{A}$ *and* $\mathcal{B}$ *of* $\mathcal{M}$ *that are nonnegligible with respect to* P*, and for all* $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$*,*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) < c < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) < +\infty. \tag{87}$$

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

$$Q(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = a\}) = 1. \tag{88}$$

From this perspective, nonseparable empirical risk functions exhibit little practical interest for model selection.

The definition of separability in Definition 5.1 and Lemma 3.3 lead to the following lemma.

Lemma 5.1. *The empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *is separable with respect to the* σ*-finite measure* Q *in* (19) *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 (87) holds. Hence, from (87) the following inequalities hold:

$$-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) > -\frac{c}{\lambda} > -\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) > -\infty, \text{ and} \tag{89}$$

$$\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)\right) > \exp\left(-\frac{c}{\lambda}\right) > \exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)\right) > 0. \tag{90}$$

This implies that

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1) > \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)-\frac{c}{\lambda}\right) \tag{91}$$

$$> \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2) \tag{92}$$

$$> 0. \tag{93}$$

Using the inequality in (91) and the facts that $Q(\mathcal{A}) > 0$ and $Q(\mathcal{B}) > 0$, the following holds

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) = \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) > 0, \tag{94}$$

and

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}) = \int_{\mathcal{B}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) > 0. \tag{95}$$

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 (87) holds. More specifically, $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) > 0$ and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}) > 0$. From Lemma 3.2 and the inequality in (87), it follows that for all pairs $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$, $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1) > 0$ and $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2) > 0$. Hence, from the fact that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) > 0$ and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}) > 0$, 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. ■

Lemma 5.1 shows that separable empirical risk functions, and only these functions, lead to ERM-RER-optimal probability measures from which models are sampled with different probabilities. For the case of nonseparable empirical risk functions, all models are sampled from the ERM-RER-optimal probability measure with the same probability.

5.2 Properties of the Log-Partition Function

The log-partition function $K_{Q,\boldsymbol{z}}$ in (22) is a nondecreasing continuous convex function as shown by the following lemmas.

Lemma 5.2. *The function $K_{Q,\boldsymbol{z}}$ in (22) is nondecreasing and differentiable infinitely many times in the interior of $\{t \in \mathbb{R} : K_{Q,\boldsymbol{z}}(t) < +\infty\}$.*

Proof: The proof is presented in Appendix K. ■

Lemma 5.3. *The function $K_{Q,\boldsymbol{z}}$ in (22) is convex in $\{t \in \mathbb{R} : K_{Q,\boldsymbol{z}}(t) < +\infty\}$. Moreover, it is strictly convex if and only if the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable with respect to the σ-finite measure Q in (19).*

Proof: The proof is presented in Appendix L. ■

In Lemma 5.2, it has been established that the log-partition function $K_{Q,\boldsymbol{z}}$ in (22) is differentiable infinitely many times in the interval

$$\{t \in \mathbb{R} : K_{Q,\boldsymbol{z}}(t) < +\infty\}.$$

Let the m-th derivative of the function $K_{Q,\boldsymbol{z}}$ in (22) be denoted by $K_{Q,\boldsymbol{z}}^{(m)} : \mathbb{R} \to \mathbb{R}$, with $m \in \mathbb{N}$. Hence, for all $s \in \mathcal{K}_{Q,\boldsymbol{z}}$,

$$K_{Q,\boldsymbol{z}}^{(m)}\left(-\frac{1}{s}\right) \triangleq \frac{\mathrm{d}^m}{\mathrm{d}t^m} K_{Q,\boldsymbol{z}}(t)\Big|_{t=-\frac{1}{s}}. \tag{96}$$

The following lemma provides explicit expressions for the first, second and third derivatives of the function $K_{Q,\boldsymbol{z}}$ in (22).

Lemma 5.4. *The first, second and third derivatives of the function $K_{Q,\boldsymbol{z}}$ in (22), denoted respectively by $K_{Q,\boldsymbol{z}}^{(1)}$, $K_{Q,\boldsymbol{z}}^{(2)}$, and $K_{Q,\boldsymbol{z}}^{(3)}$, satisfy for all $\lambda \in \mathrm{int}\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23),*

$$K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\boldsymbol{\theta}), \tag{97}$$

$$K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)\right)^2 \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\boldsymbol{\theta}), \tag{98}$$

$$K_{Q,\boldsymbol{z}}^{(3)}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)\right)^3 \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\boldsymbol{\theta}), \tag{99}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3) and the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ satisfies (25).

Proof: The proof is presented in Appendix M. ■

From Lemma 5.4, it follows that if $\boldsymbol{\Theta} \sim P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$, with $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ in (25), the random variable

$$W \triangleq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\Theta}), \tag{100}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ in (3), possesses a mean, variance, and third cumulant that are equivalent to $K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$ in (97), $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$ in (98), and $K_{Q,\boldsymbol{z}}^{(3)}\left(-\frac{1}{\lambda}\right)$ in (99), respectively.

Note that if there exists a $\delta > 0$ such that the log-partition function $K_{Q,\boldsymbol{z}}$ is differentiable within the open interval $(-\delta, \delta)$ and Q in (19) is a probability measure, the function $K_{Q,\boldsymbol{z}}$ is the cumulant generating function of the random variable

$$V \triangleq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\Theta}), \text{ with } \boldsymbol{\Theta} \sim Q. \tag{101}$$

The following lemma leverages this observation.

Lemma 5.5. *Assume that Q in (19) is a probability measure and that there exists real $\delta > 0$ such that the log-partition function $K_{Q,\boldsymbol{z}}$ in (22) is differentiable within $(-\delta, \delta)$. Then, the first, second and third derivatives of $K_{Q,\boldsymbol{z}}$, denoted respectively by $K^{(1)}_{Q,\boldsymbol{z}}$, $K^{(2)}_{Q,\boldsymbol{z}}$, and $K^{(3)}_{Q,\boldsymbol{z}}$, satisfy*

$$K^{(1)}_{Q,\boldsymbol{z}}(0) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{102}$$

$$K^{(2)}_{Q,\boldsymbol{z}}(0) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \right)^2 \mathrm{d}Q(\boldsymbol{\theta}), \tag{103}$$

$$K^{(3)}_{Q,\boldsymbol{z}}(0) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \right)^3 \mathrm{d}Q(\boldsymbol{\theta}), \tag{104}$$

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

Proof: The proof follows along the same arguments of the proof of Lemma 5.4. ■

The mean, variance, and third cumulant of the random variable V in (101) are $K^{(1)}_{Q,\boldsymbol{z}}(0)$ in (102), $K^{(2)}_{Q,\boldsymbol{z}}(0)$ in (103), and $K^{(3)}_{Q,\boldsymbol{z}}(0)$ in (104), respectively.

6 Expectation of the Empirical Risk

The mean of the random variable W in (100) 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 functional $\mathsf{R}_{\boldsymbol{z}}$ in (18). Often, $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is referred to as the 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 ERM-RER problem in (19). The following corollary of Lemma 5.4 formalizes this observation.

Corollary 6.1. *The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) verifies that*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{105}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ and the function $K^{(1)}_{Q,\boldsymbol{z}}$ are defined in (18) and (97), respectively.

The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (105) exhibits the following property.

Theorem 6.1. *The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (105) is nondecreasing with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23). Moreover, $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is strictly increasing with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable with respect to the measure Q.*

Proof: The proof is presented in Appendix N. ■

The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (105) has been shown to be non-decreasing with λ in [8, Appendix E.4] for the special case in which Q is a probability measure.

A question that arises from Theorem 6.1 is whether the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (105) can be made arbitrarily close to $\delta^{\star}_{Q,\boldsymbol{z}}$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38), by making λ arbitrarily small. The following lemma shows that the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is often bounded away from $\delta^{\star}_{Q,\boldsymbol{z}}$, even for arbitrarily small values of λ.

Lemma 6.1. *The expected empirical risk* $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ *in* (105) *satisfies,*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geqslant \delta^{\star}_{Q,\boldsymbol{z}}, \tag{106}$$

where $\delta^{\star}_{Q,\boldsymbol{z}}$ *is defined in* (38)*. Moreover, the inequality in* (106) *is strict if and only if the function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *is separable with respect to the measure* Q *in* (19)*.*

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 (105) is equal to $\delta^{\star}_{Q,\boldsymbol{z}}$, as shown by the following lemma.

Theorem 6.2. *The expected empirical risk* $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ *in* (105) *satisfies,*

$$\lim_{\lambda\to 0^{+}} \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=\delta^{\star}_{Q,\boldsymbol{z}}, \tag{107}$$

where $\delta^{\star}_{Q,\boldsymbol{z}}$ *is defined in* (38)*.*

Proof: The proof is presented in Appendix P. ■

The following lemma determines the value of the objective function of the ERM-RER problem in (19) when it is evaluated at its solution. This result appeared first in [10, Lemma 3].

Lemma 6.2 (Lemma 3 in [10])**.** *The probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (25) *and the* σ*-finite measure* Q *in* (19) *satisfy*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\lambda D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)=-\lambda K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right). \tag{108}$$

Moreover, if the condition in (27) *holds, then,*

$$\mathsf{R}_{\boldsymbol{z}}\left(Q\right)-\lambda D\left(Q\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=-\lambda K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{109}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ *is defined in* (18)*; and the function* $K_{Q,\boldsymbol{z}}$ *is defined in* (22)*.*

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

$$\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) = -K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}), \tag{110}$$

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

$$D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) = \int \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{111}$$

$$= -K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{112}$$

$$= -K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{113}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (18). This completes the proof of (108).

From Lemma 3.3 and (110), it follows that

$$D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = -\int \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{114}$$

$$= K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{\lambda}\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{115}$$

$$= K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}(Q), \tag{116}$$

which completes the proof of (109). ∎

The following corollary of Lemma 6.2 characterizes the difference between the expected values of the random variables W and V in (100) and (101), respectively.

Corollary 6.2. *If measures Q and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) are both probability measures, then,*

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda\left(D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right)\right). \tag{117}$$

The right-hand side of (117) is a symmetrized Kullback-Liebler divergence, also known as Jeffrey's divergence [65], between the measures Q and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. More importantly, when Q is a probability measure, it follows that $D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) \geqslant 0$ and $D\left(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geqslant 0$, which leads to the following corollary from Lemma 6.2.

Corollary 6.3. *If the σ-finite measure Q in (19) is a probability measure, then, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) satisfies*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant \mathsf{R}_{\boldsymbol{z}}(Q), \tag{118}$$

where, the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (18).

7 Variance of the Empirical Risk

In Lemma 5.2, it has be established that if there exists a $\delta > 0$ such that the log-partition function $K_{Q,\boldsymbol{z}}$ in (22) is finite within the open interval $(-\delta, \delta)$ the log-partition function $K_{Q,\boldsymbol{z}}$ is differentiable infinitely many times within the interval $(-\infty, \delta)$. This together with the *mean value theorem* [66, Theorem 5.10] lead to the following characterization of the differences of the values $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right)$ and $K^{(2)}_{Q,\boldsymbol{z}}(0)$, with $t > 0$.

Lemma 7.1. *If the measure Q in* (19) *is a probability measure and there exists a $\delta > 0$ such that the function $K_{Q,\boldsymbol{z}}$ in* (22) *is differentiable within the open interval $(-\delta, \delta)$, then for all $t > 0$,*

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right) - K^{(2)}_{Q,\boldsymbol{z}}(0) = -\frac{1}{t} K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\beta}\right) < +\infty, \tag{119}$$

for some $\beta \in (t, +\infty)$, where the functions $K^{(2)}_{Q,\boldsymbol{z}}$, and $K^{(3)}_{Q,\boldsymbol{z}}$ are defined in (96).

Proof: The proof is an immediate consequence of Lemma 5.2 and the mean value theorem [66, Theorem 5.10]. ■

The relevance of Lemma 7.1 lies on the fact that $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ and $K^{(2)}_{Q,\boldsymbol{z}}(0)$ are the variances of the random variables W in (100) and V in (101). See Lemma 5.4 and Lemma 5.5. Under the assumptions of Lemma 7.1, it follows that the function $K^{(3)}_{Q,\boldsymbol{z}}$ is continuous in $(-\infty, \delta)$, where $\delta > 0$. Hence, for all $t > 0$, the function $K^{(3)}_{Q,\boldsymbol{z}}$ achieves a maximum and a minimum within the interval $\left[-\frac{1}{t}, 0\right]$. Such extrema allow providing lower and upper bounds on the variance $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ of the random variable W in terms of the variance $K^{(2)}_{Q,\boldsymbol{z}}(0)$ of the random variable V, as shown hereunder.

Corollary 7.1. *If the measure Q in* (19) *is a probability measure and there exists a $\delta > 0$ such that the function $K_{Q,\boldsymbol{z}}$ in* (22) *is differentiable within the open interval $(-\delta, \delta)$, then for all $t > 0$,*

$$K^{(2)}_{Q,\boldsymbol{z}}(0) - \frac{1}{t} c_2 \leqslant K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right) \leqslant K^{(2)}_{Q,\boldsymbol{z}}(0) - \frac{1}{t} c_1, \tag{120}$$

where,

$$c_1 = \min_{s \in \left[-\frac{1}{t}, 0\right]} K^{(3)}_{Q,\boldsymbol{z}}(s) \quad \textit{and} \tag{121}$$

$$c_2 = \max_{s \in \left[-\frac{1}{t}, 0\right]} K^{(3)}_{Q,\boldsymbol{z}}(s), \tag{122}$$

and the functions $K^{(2)}_{Q,\boldsymbol{z}}$, and $K^{(3)}_{Q,\boldsymbol{z}}$ are defined in (96).

The inequality in (120) reveals that under the assumptions of Corollary 7.1, in the asymptotic regime when $t \to +\infty$, the variances of the random variables W

in (100) and V in (101) are identical. Additionally, unlike the means $K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ and $K^{(1)}_{Q,\boldsymbol{z}}(0)$ of the random variables W and V, which satisfy $K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \leqslant K^{(1)}_{Q,\boldsymbol{z}}(0)$ (Corollary 6.3), their variances $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ and $K^{(2)}_{Q,\boldsymbol{z}}(0)$ might satisfy $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < K^{(2)}_{Q,\boldsymbol{z}}(0)$ or $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \geqslant K^{(2)}_{Q,\boldsymbol{z}}(0)$ depending on whether the function $K^{(3)}_{Q,\boldsymbol{z}}$ is positive or negative within the interval $[-\frac{1}{\lambda}, 0]$. Using this observation the values $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right)$, with $t > 0$, and $K^{(2)}_{Q,\boldsymbol{z}}(0)$ can be compared as follows.

Lemma 7.2. *Assume that the measure Q in (19) is a probability measure and there exists a $\delta > 0$ such that the function $K_{Q,\boldsymbol{z}}$ in (22) is differentiable within the open interval $(-\delta, \delta)$. Hence, the following holds for all $t > 0$:*

- *If for all $s > t$, $K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) < 0$, then*

$$K^{(2)}_{Q,\boldsymbol{z}}(0) < K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right) < +\infty; \tag{123}$$

- *If for all $s > t$, $K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) > 0$, then*

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right) < K^{(2)}_{Q,\boldsymbol{z}}(0) < +\infty; \tag{124}$$

- *If for some $s > t$, $K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) = 0$, then there exists two positive reals c_1 and c_2 such that*

$$c_1 \leqslant \min\{K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right), K^{(2)}_{Q,\boldsymbol{z}}(0)\} \tag{125}$$

$$\leqslant \max\{K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right), K^{(2)}_{Q,\boldsymbol{z}}(0)\} \tag{126}$$

$$\leqslant c_2. \tag{127}$$

Proof: The proofs of the inequalities in (123) and (124) are immediate consequences of Lemma 7.1. The inequalities in (125) and (127) follow from the fact that the function $K^{(2)}_{Q,\boldsymbol{z}}$, which is continuous, exhibits critical points at $-\frac{1}{s}$, with s satisfying $K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) = 0$. Some of such critical points might be local extrema of the function $K^{(2)}_{Q,\boldsymbol{z}}$, either local minima or local maxima. Hence, the inequalities (125) and (127) follow by choosing c_1 as the smallest minimum of the function $K^{(2)}_{Q,\boldsymbol{z}}$ within the interval $\left[-\frac{1}{t}, 0\right]$; and c_2 as the biggest maximum of the function $K^{(2)}_{Q,\boldsymbol{z}}$ within the interval $\left[-\frac{1}{t}, 0\right]$. If none of such critical points is a local extremum, then, (125) and (127) hold with equality. ■

Lemma 7.1 and Lemma 7.2 show that the monotonicity of the expectation of the random variable W in (100), stated by Theorem 6.1, is not a property exhibited by the variance nor the third cumulant. The following example highlights this observation.

Example 7.1. *Consider the ERM-RER problem in* (19)*, under the assumption that Q is a probability measure and the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in* (3) *is such that for all $\boldsymbol{\theta} \in \mathcal{M}$,*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{A} \\ 1 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{A}, \end{cases} \tag{128}$$

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 (22) *satisfies for all $\lambda > 0$,*

$$K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \log\left(Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))\right). \tag{129}$$

The derivatives $K^{(1)}_{Q,\boldsymbol{z}}$, $K^{(2)}_{Q,\boldsymbol{z}}$, and $K^{(3)}_{Q,\boldsymbol{z}}$ in (96) *of the function $K_{Q,\boldsymbol{z}}$ in* (129) *satisfy for all $\lambda > 0$,*

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{\exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))}{Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))}; \tag{130}$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{Q(\mathcal{A})(1 - Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{\left(Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))\right)^2}; \text{ and} \tag{131}$$

$$K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\left(\frac{Q(\mathcal{A}) - (1 - Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{Q(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - Q(\mathcal{A}))}\right). \tag{132}$$

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

$$Q(\mathcal{A}) - (1 - Q(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right) > 0. \tag{133}$$

Assume that $Q(\mathcal{A}) \geqslant \frac{1}{2}$. Thus, it holds that for all $\lambda > 0$, the inequality in (133) *is always satisfied. This follows from observing that for all $\lambda > 0$,*

$$\exp\left(-\frac{1}{\lambda}\right) < 1 \leqslant \frac{Q(\mathcal{A})}{1 - Q(\mathcal{A})}. \tag{134}$$

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

$$\frac{1}{4} \geqslant K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) > K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) > \ldots > 0. \tag{135}$$

Alternatively, assume that $Q(\mathcal{A}) < \frac{1}{2}$. In this case, the inequality in (133) *is satisfied if and only if*

$$\lambda < \left(\log\left(\frac{1 - Q(\mathcal{A})}{Q(\mathcal{A})}\right)\right)^{-1}. \tag{136}$$

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

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

it holds that

$$\frac{1}{4} > K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) > K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) > \ldots > 0. \tag{137}$$

Moreover, for all decreasing sequences of positive reals

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

it holds that

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) < K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) < \ldots < \frac{1}{4}. \tag{138}$$

The upperbound by $\frac{1}{4}$ *in* (135), (137) *and* (138) *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 7.1 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 $\mathcal{T}\left(\boldsymbol{z}\right)$ in (5) that is greater than or equal to the probability of suboptimal models $\mathcal{M} \setminus \mathcal{T}\left(\boldsymbol{z}\right)$, i.e., $Q\left(\mathcal{T}\left(\boldsymbol{z}\right)\right) \geqslant \frac{1}{2}$, 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 (4), the variance is monotone with respect to the parameter λ.

Alternatively, when the reference measure assigns a probability to the set $\mathcal{T}\left(\boldsymbol{z}\right)$ that is smaller than the probability of the set $\mathcal{M} \setminus \mathcal{T}\left(\boldsymbol{z}\right)$, i.e., $Q\left(\mathcal{T}\left(\boldsymbol{z}\right)\right) < \frac{1}{2}$, 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 $Q\left(\mathcal{A}\right)$. 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 $\mathcal{T}\left(\boldsymbol{z}\right)$ 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 (100). These observations are central to understanding the concentration of probability that occurs when λ decreases to zero, as discussed in Section 9.

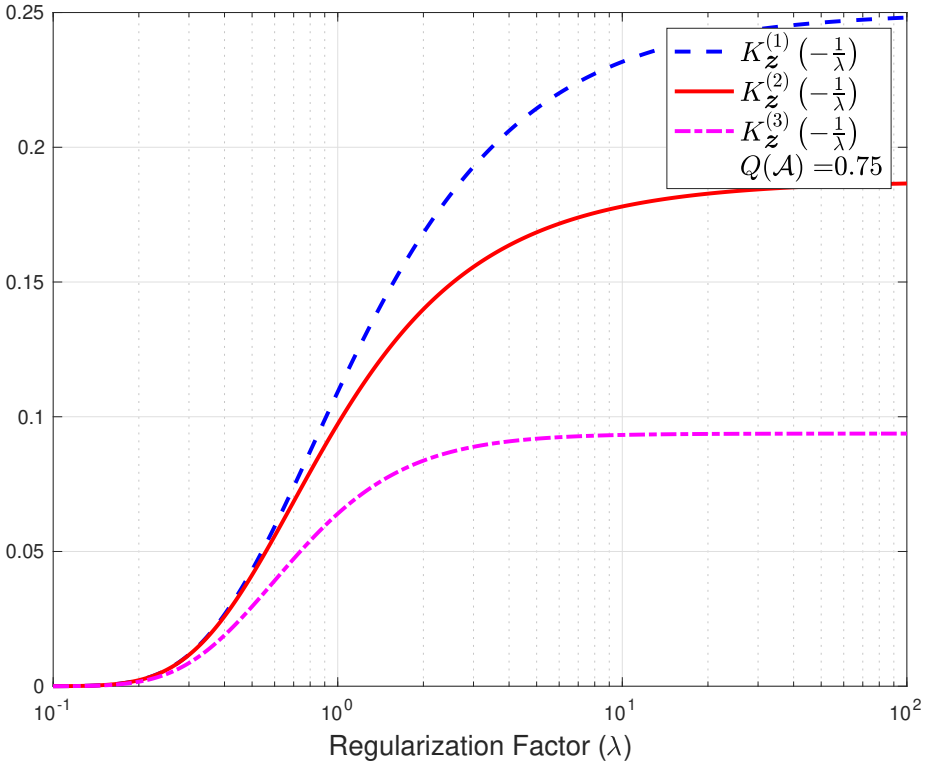


Figure 1: Mean $K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$, variance $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$, and third central moment $K_{Q,\boldsymbol{z}}^{(3)}\left(-\frac{1}{\lambda}\right)$ of the empirical risk in Example 7.1, with $Q\left(\mathcal{A}\right)=\frac{3}{4}$

8 Cumulant Generating Function of the Empirical Risk

Consider the transport of the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ in (25) from $(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right))$ to $([0,+\infty],\mathscr{B}\left([0,+\infty]\right))$ through the function $\mathsf{L}_{\boldsymbol{z}}$ in (3). Denote the resulting probability measure in $([0,+\infty],\mathscr{B}\left([0,+\infty]\right))$ by $P_{W|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$. That is, for all $\mathcal{A}\in\mathscr{B}\left([0,+\infty]\right)$,

$$P_{W|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\mathcal{A}\right)=P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\mathsf{L}_{\boldsymbol{z}}^{-1}\left(\mathcal{A}\right)\right), \tag{139}$$

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

$$\mathsf{L}_{\boldsymbol{z}}^{-1}\left(\mathcal{A}\right)\triangleq\{\boldsymbol{\nu}\in\mathcal{M}:\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\in\mathcal{A}\}. \tag{140}$$

Note that the random variable W in (100) induces the probability measure $P_{W|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ in $([0,+\infty],\mathscr{B}\left([0,+\infty]\right))$. The objective of this section is to study the properties of the cumulant generating function of the probability measure $P_{W|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$, denoted by $J_{\boldsymbol{z},Q,\lambda}:\mathbb{R}\to\mathbb{R}\cup\{+\infty\}$, which satisfies for all $t\in$

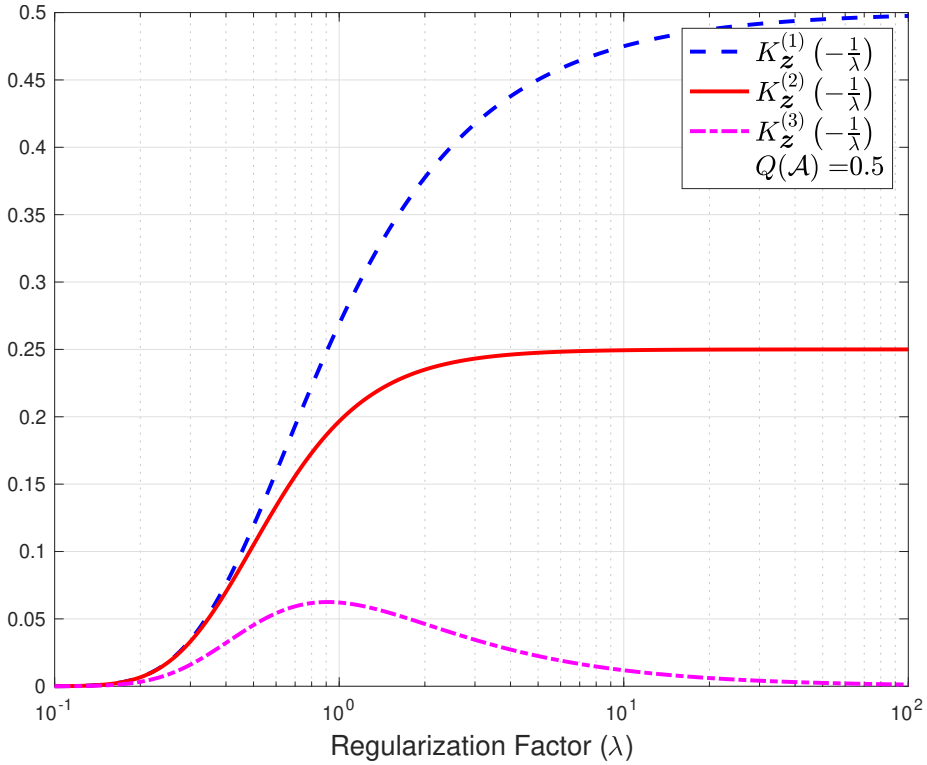


Figure 2: Mean $K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$, variance $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$, and third central moment $K_{Q,\boldsymbol{z}}^{(3)}\left(-\frac{1}{\lambda}\right)$ of the empirical risk in Example 7.1, with $Q\left(\mathcal{A}\right) = \frac{1}{2}$

$\mathbb{R}$,

$$J_{\boldsymbol{z},Q,\lambda}(t) = \log\left(\int \exp\left(tw\right) \mathrm{d}P_{W|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(w)\right) \tag{141}$$

$$= \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\boldsymbol{\theta})\right), \tag{142}$$

where the equality in (142) follows from [55, Theorem 1.6.12].

The following lemma provides an expression for $J_{\boldsymbol{z},Q,\lambda}$ in terms of the log-partition function $K_{Q,\boldsymbol{z}}$ in (22).

Lemma 8.1. *If $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in* (23)*, then, the function $J_{\boldsymbol{z},Q,\lambda}$ in* (141)*,*

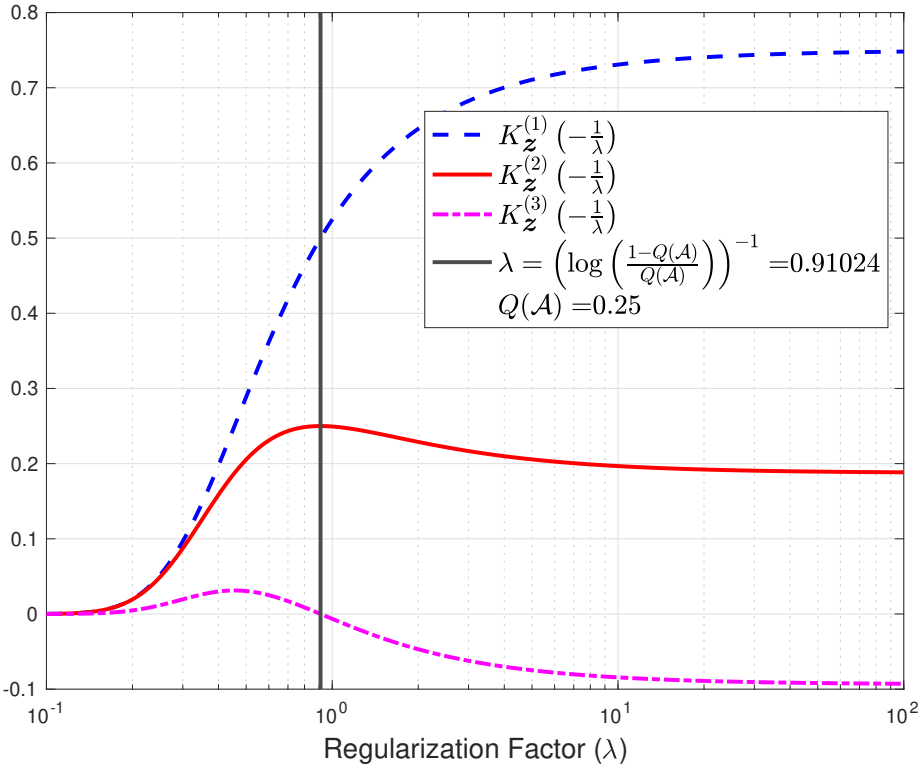


Figure 3: Mean $K_{Q,\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$, variance $K_{Q,\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$, and third central moment $K_{Q,\boldsymbol{z}}^{(3)}\left(-\frac{1}{\lambda}\right)$ of the empirical risk in Example 7.1, with $Q\left(\mathcal{A}\right) = \frac{1}{4}$

verifies for all $t \in \mathbb{R}$,

$$J_{\boldsymbol{z},Q,\lambda}(t) = K_{P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\boldsymbol{z}}\left(t\right) \tag{143}$$

$$= K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \tag{144}$$

$$= \sum_{m=1}^{+\infty} \frac{t^m}{m!} K_{Q,\boldsymbol{z}}^{(m)}\left(-\frac{1}{\lambda}\right), \tag{145}$$

with the function $K_{Q,\boldsymbol{z}}$ *in* (22) *and the function* $K_{Q,\boldsymbol{z}}^{(m)}$ *in* (96).

Proof: The proof of (143) follows immediately from (22) and (142). The proof of (144) follows from Lemma 4.5. Finally, the proof of (145) follows by observing that a Taylor expansion of the function $K_{Q,\boldsymbol{z}}$ in (22) at the point $-\frac{1}{\lambda}$, yields for all $t \in \{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}(\nu) < +\infty\}$,

$$K_{Q,\boldsymbol{z}}\left(t\right) = K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \sum_{s=1}^{+\infty} \frac{K_{Q,\boldsymbol{z}}^{(s)}\left(-\frac{1}{\lambda}\right)}{s!} \left(t + \frac{1}{\lambda}\right)^s. \tag{146}$$

Choosing $\alpha \in \left\{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}(\nu - \frac{1}{\lambda}) < +\infty\right\}$ such that $t = \alpha - \frac{1}{\lambda}$ in (146) yields

$$K_{Q,\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right) = K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \sum_{s=1}^{+\infty} \frac{\alpha^s}{s!} K_{Q,\boldsymbol{z}}^{(s)}\left(-\frac{1}{\lambda}\right), \tag{147}$$

which implies that for all $t \in \left\{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}(\nu - \frac{1}{\lambda}) < +\infty\right\}$,

$$K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \sum_{s=1}^{+\infty} \frac{t^s}{s!} K_{Q,\boldsymbol{z}}^{(s)}\left(-\frac{1}{\lambda}\right). \tag{148}$$

Let $s^\star \in \mathbb{R} \cup \{+\infty\}$ be defined by

$$s^\star \triangleq \sup\left\{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}\left(\nu - \frac{1}{\lambda}\right) < \infty\right\}. \tag{149}$$

If $s^\star = +\infty$, then for all $t \in \mathbb{R}$, $K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty$, and thus,

$$+\infty > J_{\boldsymbol{z},Q,\lambda}(t) = K_{P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)},\boldsymbol{z}}(t) \tag{150}$$

$$= K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \tag{151}$$

$$= \sum_{m=1}^{+\infty} \frac{t^m}{m!} K_{Q,\boldsymbol{z}}^{(m)}\left(-\frac{1}{\lambda}\right). \tag{152}$$

Alternatively, if $s^\star < +\infty$, it follows that for all $t > s^\star$, $K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) = +\infty$. From the fact that the function $K_{Q,\boldsymbol{z}}$ is continuous (Lemma 5.2) and $K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < \infty$ (due to the fact that $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ in (23)), it follows that

$$+\infty = J_{\boldsymbol{z},Q,\lambda}(t) = K_{Q,\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \tag{153}$$

$$= \sum_{m=1}^{+\infty} \frac{t^m}{m!} K_{Q,\boldsymbol{z}}^{(m)}\left(-\frac{1}{\lambda}\right), \tag{154}$$

which implies that $\sum_{m=1}^{+\infty} \frac{t^m}{m!} K_{Q,\boldsymbol{z}}^{(m)}\left(-\frac{1}{\lambda}\right) = +\infty$. Hence, in this case, the equality in (145) is of the form $+\infty = +\infty$. This completes the proof. ■

Alternative expressions for $J_{\boldsymbol{z},Q,\lambda}$ in (141) are provided hereunder.

Lemma 8.2. *If $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), then, the function $J_{\boldsymbol{z},Q,\lambda}$ in (141), verifies for all $t \in (0, +\infty)$,*

$$J_{\boldsymbol{z},Q,\lambda}\left(-\frac{1}{t}\right) = -\frac{1}{t}\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{t}}\right)}\right) - D\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{t}}\right)} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) \tag{155}$$

$$= -\frac{1}{t}\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) + D\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{t}}\right)}\right) \tag{156}$$

$$\leqslant 0, \tag{157}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ *is in* (18)*; the function* $K_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}$ *is in* (61)*; and the probability measures* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *and* $P^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{t}}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *are respectively in* (25) *and* (74).

Proof: The proof of (155) follows from (108) in Lemma 6.2 by observing that for all $t \in (0, +\infty)$,

$$-tK_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}\left(-\frac{1}{t}\right)=\mathsf{R}_{\boldsymbol{z}}\left(P^{\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},t\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+tD\left(P^{\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},t\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \quad (158)$$

$$=\mathsf{R}_{\boldsymbol{z}}\left(P^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{t}}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+tD\left(P^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{t}}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \quad (159)$$

where the equality in (159) follows from Lemma 4.6. The proof of (156) follows from (109) in Lemma 6.2 by observing that for all $t \in (0, +\infty)$,

$$-tK_{P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},\boldsymbol{z}}\left(-\frac{1}{t}\right)=\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-tD\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}},t\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \quad (160)$$

$$=\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-tD\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{\left(Q,\frac{1}{\frac{1}{\lambda}+\frac{1}{t}}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \quad (161)$$

where the equality in (161) follows from Lemma 4.6, which completes the proof. ■

From Lemma 5.2 and Lemma 8.1, it follows that the function $J_{\boldsymbol{z},Q,\lambda}$ in (141) is increasing and differentiable infinitely many times in the interior of

$$\left\{t \in \mathbb{R} : K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right) < +\infty\right\}.$$

Moreover, note that $\left(-\infty, \frac{1}{\lambda}\right] \subset \left\{t \in \mathbb{R} : K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right) < +\infty\right\}$. 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 (141). That is, for all $s \in \mathbb{R}$,

$$J^{(m)}_{\boldsymbol{z},Q,\lambda}(s) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} J_{\boldsymbol{z},Q,\lambda}(t)\Big|_{t=s}. \quad (162)$$

From Lemma 8.1, it follows that for all $m \in \mathbb{N}$, and for all $\alpha \in \mathbb{R}$, the following holds,

$$J^{(m)}_{\boldsymbol{z},Q,\lambda}(\alpha) = K^{(m)}_{Q,\boldsymbol{z}}\left(\alpha-\frac{1}{\lambda}\right), \quad (163)$$

where the function $K^{(m)}_{Q,\boldsymbol{z}}$ denotes the m-th derivative of the function $K_{Q,\boldsymbol{z}}$ in (22). See for instance, Lemma 5.4. The equality in (163) establishes a relation between the cumulant generating function $J_{\boldsymbol{z},Q,\lambda}$ and the function $K_{Q,\boldsymbol{z}}$. This observation becomes an alternative proof to Lemma 5.4.

The following theorem presents the relation between the cumulant generating function $J_{\boldsymbol{z},Q,\lambda}$ and the functions $K^{(1)}_{Q,\boldsymbol{z}}$ and $K^{(2)}_{Q,\boldsymbol{z}}$ in (97) and (98).

Theorem 8.1. *For all $\alpha \in \mathbb{R}$, the function $J_{\boldsymbol{z},Q,\lambda}$ in* (141) *verifies the following equality*

$$J_{\boldsymbol{z},Q,\lambda}(\alpha) = \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{2}\alpha^2 K^{(2)}_{Q,\boldsymbol{z}}(\xi) \tag{164}$$

with

$$\xi \in \left(\min\left\{-\frac{1}{\lambda}, \alpha - \frac{1}{\lambda}\right\}, \max\left\{-\frac{1}{\lambda}, \alpha - \frac{1}{\lambda}\right\}\right), \tag{165}$$

where the functions $K^{(1)}_{Q,\boldsymbol{z}}$ and $K^{(2)}_{Q,\boldsymbol{z}}$ are defined in (97) *and* (98)*, respectively.*

Proof: From Lemma 5.2, it follows that the function $K_{Q,\boldsymbol{z}}$ is differentiable infinitely many times in the interior of $\{t \in \mathbb{R} : K_{Q,\boldsymbol{z}}(t) < +\infty\}$. Then, a Taylor expansion of the function $K_{Q,\boldsymbol{z}}$ in (22) at the point $-\frac{1}{\lambda}$ yields for all $t \in \{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}(\nu) < +\infty\}$,

$$K_{Q,\boldsymbol{z}}(t) = K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \sum_{s=1}^{+\infty} \frac{1}{s!}\left(t + \frac{1}{\lambda}\right)^s K^{(s)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right). \tag{166}$$

Choosing $t = \alpha - \frac{1}{\lambda}$, with $\alpha \in \left\{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}(\nu - \frac{1}{\lambda}) < +\infty\right\}$ in (166), it holds from the Taylor-Lagrange theorem [67, Theorem 2.5.4] that

$$K_{Q,\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right) = K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{2}\alpha^2 K^{(2)}_{Q,\boldsymbol{z}}(\xi), \tag{167}$$

where $\xi \in \big(\min\{-\frac{1}{\lambda}, \alpha - \frac{1}{\lambda}\}, \max\{-\frac{1}{\lambda}, \alpha - \frac{1}{\lambda}\},\big)$.

Let $s^\star \in \mathbb{R} \cup \{+\infty\}$ be defined by

$$s^\star \triangleq \sup\left\{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}\left(\nu - \frac{1}{\lambda}\right) < \infty\right\}. \tag{168}$$

If $s^\star = +\infty$, then for all $\alpha \in \mathbb{R}$, $K_{Q,\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty$, and thus, the proof is completed by noticing that from Lemma 8.1, it holds that $J_{\boldsymbol{z},Q,\lambda}(\alpha) = K_{Q,\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$.

Alternatively, if $s^\star < +\infty$, it follows that for all $\alpha > s^\star$, $K_{Q,\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right) = +\infty$. From from Lemma 8.1, it holds that $J_{\boldsymbol{z},Q,\lambda}(\alpha) = +\infty$, which implies that $+\infty = \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{\alpha^2}{2} K^{(2)}_{Q,\boldsymbol{z}}(\xi)$, and thus, $K^{(2)}_{Q,\boldsymbol{z}}(\xi)$ is infinite. Hence, in this case, the equality in (164) is of the form $+\infty \leqslant +\infty$. This completes the proof. ■

In (164), the parameter ξ depends on α, as shown in (165). To highlight this dependence, in the following, the parameter ξ is denoted by ξ_α. Using this notation, the focus is now on the term $K^{(2)}_{Q,\boldsymbol{z}}(\xi_\alpha)$, when $\alpha \in \{t \in \mathbb{R} : J_{\boldsymbol{z},Q,\lambda}(t) < +\infty\}$.

Theorem 8.2. *The function $J_{\boldsymbol{z},Q,\lambda}$ in* (141) *verifies the following inequality, for all $\alpha \in \{t \in \mathbb{R} : J_{\boldsymbol{z},Q,\lambda}(t) < +\infty\}$,*

$$J_{\boldsymbol{z},Q,\lambda}(\alpha) \leqslant \alpha K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{2}\alpha^2 \beta^2_{Q,\boldsymbol{z}} \tag{169}$$

where $\beta_{Q,\boldsymbol{z}}$ is finite, and satisfies

$$\beta_{Q,\boldsymbol{z}}=\sup\left\{\sqrt{K^{(2)}_{Q,\boldsymbol{z}}(\alpha)}:\alpha\in\left(-\infty,b-\frac{1}{\lambda}\right)\right\}, \tag{170}$$

with

$$b\triangleq\sup\left\{t\in\mathbb{R}:J_{\boldsymbol{z},Q,\lambda}(t)<+\infty\right\}, \tag{171}$$

and the functions $K^{(1)}_{Q,\boldsymbol{z}}$ and $K^{(2)}_{Q,\boldsymbol{z}}$ defined in (97) *and* (98)*, respectively.*

Proof: The proof of the inequality in (169) is trivial from Theorem 8.1 and the choice of $\beta_{Q,\boldsymbol{z}}$ in (170). Hence, the remainder of the proof focuses on proving that $\beta_{Q,\boldsymbol{z}}<+\infty$. From Lemma 5.2 and Lemma 8.1, it holds that

$$\left\{t\in\mathbb{R}:J_{\boldsymbol{z},Q,\lambda}(t)<+\infty\right\}=\left\{t\in\mathbb{R}:K_{Q,\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)<+\infty\right\},$$

which implies that the set $\left\{t\in\mathbb{R}:J_{\boldsymbol{z},Q,\lambda}(t)<+\infty\right\}$ is an interval of the form $(-\infty,b)$, with b in (171). This follows from the fact that the function $K_{Q,\boldsymbol{z}}$ is continuous and nondecreasing (Lemma 5.2) and the fact that

$$\lim_{t\to b}J_{\boldsymbol{z},Q,\lambda}(t)=+\infty. \tag{172}$$

For all $\alpha\in(-\infty,b)$, the function $K^{(2)}_{Q,\boldsymbol{z}}$ is continuous (Lemma 5.2). Hence, for all $t\in\left(\min\left\{-\frac{1}{\lambda},\alpha-\frac{1}{\lambda}\right\},\max\left\{-\frac{1}{\lambda},\alpha-\frac{1}{\lambda}\right\}\right)\subset(-\infty,b)$, the value $K^{(2)}_{Q,\boldsymbol{z}}(t)$ is finite. Moreover, the values $K^{(2)}_{Q,\boldsymbol{z}}\left(\min\left\{-\frac{1}{\lambda},\alpha-\frac{1}{\lambda}\right\}\right)$ and $K^{(2)}_{Q,\boldsymbol{z}}\left(\max\left\{-\frac{1}{\lambda},\alpha-\frac{1}{\lambda}\right\}\right)$ are both finite. This implies that the function $K^{(2)}_{Q,\boldsymbol{z}}$ achieves a minimum and maximum within the closed interval $\left[\min\left\{-\frac{1}{\lambda},\alpha-\frac{1}{\lambda}\right\},\max\left\{-\frac{1}{\lambda},\alpha-\frac{1}{\lambda}\right\}\right]$. Thus, the corresponding term $K^{(2)}_{Q,\boldsymbol{z}}(\xi_\alpha)$ is finite.

In the asymptotic regime, when $\alpha\to-\infty$, the following holds:

$$\lim_{\alpha\to-\infty}\xi_\alpha\in\left(-\infty,-\frac{1}{\lambda}\right). \tag{173}$$

The function $K^{(2)}_{Q,\boldsymbol{z}}$ is continuous in $\left(-\infty,-\frac{1}{\lambda}\right)$, as a consequence of the inclusion $\left(-\infty,-\frac{1}{\lambda}\right)\subset(-\infty,b)$, and thus, for all $t\in\left(-\infty,-\frac{1}{\lambda}\right)$, $K^{(2)}_{Q,\boldsymbol{z}}(t)<+\infty$. Moreover, from the assumption that $\lambda\in\mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), it holds that

$$\lim_{t\to-\frac{1}{\lambda}}K^{(2)}_{Q,\boldsymbol{z}}(t)=K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)<+\infty. \tag{174}$$

Alternatively,

$$\lim_{t\to-\infty} K^{(2)}_{Q,\boldsymbol{z}}(t) = \lim_{t\to 0^+} K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right) \tag{175}$$

$$= \lim_{t\to 0^+} \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{176}$$

$$= \lim_{t\to 0^+} \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \mathrm{d}P^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) - \lim_{t\to 0^+} \left(K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right)\right)^2, \tag{177}$$

$$= \lim_{t\to 0^+} \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \mathrm{d}P^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) - \left(\delta^\star_{Q,\boldsymbol{z}}\right)^2 \tag{178}$$

$$= \lim_{t\to 0^+} \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \frac{\mathrm{d}P^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) - \left(\delta^\star_{Q,\boldsymbol{z}}\right)^2 \tag{179}$$

$$= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \left(\lim_{t\to 0^+} \frac{\mathrm{d}P^{(Q,t)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) - \left(\delta^\star_{Q,\boldsymbol{z}}\right)^2 \tag{180}$$

$$= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \left(\frac{1}{Q\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right)} \mathbb{1}_{\left\{\boldsymbol{\theta}\in\mathcal{L}^\star_{Q,\boldsymbol{z}}\right\}}\right) \mathrm{d}Q(\boldsymbol{\theta}) - \left(\delta^\star_{Q,\boldsymbol{z}}\right)^2 \tag{181}$$

$$= \frac{1}{Q\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right)} \int_{\mathcal{L}^\star_{Q,\boldsymbol{z}}} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \mathrm{d}Q(\boldsymbol{\theta}) - \left(\delta^\star_{Q,\boldsymbol{z}}\right)^2 \tag{182}$$

$$= \left(\delta^\star_{Q,\boldsymbol{z}}\right)^2 - \left(\delta^\star_{Q,\boldsymbol{z}}\right)^2 = 0, \tag{183}$$

where the equality in (176) follows from Lemma 5.4; the equality in (178) follows from Theorem 6.2, with $\delta^\star_{Q,\boldsymbol{z}}$ in (38); the equality in (180) follows from the dominated convergence theorem [55, Theorem 1.6.9]; the equality in (181) follows from Lemma 3.6; and the equality in (183) follows from the definition of the set $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (39).

Hence, from (173), (174), and (183), it follows that

$$\lim_{\alpha\to-\infty} K^{(2)}_{Q,\boldsymbol{z}}(\xi_\alpha) \in \left[0, \max_{c\in\left(-\infty,-\frac{1}{\lambda}\right]} K^{(2)}_{Q,\boldsymbol{z}}(c)\right], \tag{184}$$

where the maximum exists and is finite.

On the other hand, in the asymptotic regime, when $\alpha \to b^-$, two cases are considered: (i) $b \geqslant 0$; and (ii) $b < 0$. In the first case, the following holds from (165):

$$\lim_{\alpha\to b^-} \xi_\alpha \in \left[-\frac{1}{\lambda}, b - \frac{1}{\lambda}\right). \tag{185}$$

The function $K^{(2)}_{Q,\boldsymbol{z}}$ is continuous in $\left(-\frac{1}{\lambda}, b-\frac{1}{\lambda}\right)$, as a consequence of the inclusion $\left(-\frac{1}{\lambda}, b-\frac{1}{\lambda}\right) \subset (-\infty, b)$, and thus, for all $t \in \left(-\frac{1}{\lambda}, b-\frac{1}{\lambda}\right)$, $K^{(2)}_{Q,\boldsymbol{z}}(t) < +\infty$.

Moreover,

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)<+\infty, \text{ and} \tag{186}$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(b-\frac{1}{\lambda}\right)<+\infty. \tag{187}$$

This implies that when $b \geqslant 0$,

$$\lim_{\alpha\to b^-} K^{(2)}_{Q,\boldsymbol{z}}(\xi_\alpha) \in \left[0, \max_{c\in\left[-\frac{1}{\lambda}, b-\frac{1}{\lambda}\right]} K^{(2)}_{Q,\boldsymbol{z}}(c)\right], \tag{188}$$

where the maximum exists and is finite. Finally, In the second case, the following holds from (165):

$$\lim_{\alpha\to b^-} \xi_\alpha \in \left(b-\frac{1}{\lambda}, -\frac{1}{\lambda}\right). \tag{189}$$

The function $K^{(2)}_{Q,\boldsymbol{z}}$ is continuous in $\left(b-\frac{1}{\lambda}, -\frac{1}{\lambda}\right)$, as a consequence of the inclusion $\left(b-\frac{1}{\lambda}, -\frac{1}{\lambda}\right) \subset (-\infty, b)$, and thus, for all $t \in \left(b-\frac{1}{\lambda}, -\frac{1}{\lambda}\right)$, $K^{(2)}_{Q,\boldsymbol{z}}(t) < +\infty$. Moreover,

$$K^{(2)}_{Q,\boldsymbol{z}}\left(b-\frac{1}{\lambda}\right)<+\infty, \text{ and} \tag{190}$$

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)<+\infty. \tag{191}$$

This implies that when $b < 0$,

$$\lim_{\alpha\to b^-} K^{(2)}_{Q,\boldsymbol{z}}(\xi_\alpha) \in \left[0, \max_{c\in\left(b-\frac{1}{\lambda}, -\frac{1}{\lambda}\right]} K^{(2)}_{Q,\boldsymbol{z}}(c)\right], \tag{192}$$

where the maximum exists and is finite. From all the above, it holds that for all $\alpha \in \{t \in \mathbb{R} : J_{\boldsymbol{z},Q,\lambda}(t) < +\infty\}$, the value $K^{(2)}_{Q,\boldsymbol{z}}(\xi_\alpha)$ is finite, and this completes the proof. ■

The main implication of Theorem 8.2 is that the random variable W in (100) is a sub-Gaussian random variable with sub-Gaussianity parameter $\beta_{Q,\boldsymbol{z}}$ in (170) [53, Section 2.3]. This follows by noticing that the function $J_{\boldsymbol{z},Q,\lambda}$ in (142) is the cumulant generating function of the random variable W. Hence, whenever it is finite, it is upper bounded as shown in Theorem 8.2. The following corollary of Theorem 8.2 highlights this observation.

Corollary 8.1. *The random variable W in* (100) *is a sub-Gaussian random variable with sub-Gaussianity parameter $\beta_{Q,\boldsymbol{z}}$ in* (170)*.*

The relevance of Corollary 8.1 is that it highlights the fact that when the models are sampled from the ERM-RER optimal measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), the empirical risk with respect to the dataset $\boldsymbol{z}$ is a sub-Gaussian random variable with sub-Gaussianity parameter $\beta_{Q,\boldsymbol{z}}$ in (170).

9 Concentration of Probability

Consider the following set,

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \triangleq \left\{ \boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \right\}, \tag{193}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined by (3); the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined by (18); and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is in (25). This section introduces two results. First, in Theorem 9.1, it is shown that when λ tends to zero, the set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ forms an indexed family of sets that is monotonic and decreases to the set

$$\mathcal{N}^{\star}_{Q,\boldsymbol{z}} \triangleq \mathcal{L}_{\boldsymbol{z}}\left(\delta^{\star}_{Q,\boldsymbol{z}}\right), \tag{194}$$

where $\delta^{\star}_{Q,\boldsymbol{z}}$ is defined in (38); and the set $\mathcal{L}_{\boldsymbol{z}}(\delta^{\star}_{Q,\boldsymbol{z}})$ is defined in (37). Second, in Theorem 9.2, it is shown that the probability $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda))$ strictly increases when λ tends to zero. More importantly, in Theorem 9.3, it is shown that the limit of the probability $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda))$, when $\lambda \to 0$, is equal to one. These observations justify referring to the set $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ as the *limit set*. These observations are complementary to those stated in Section 3.2 and Section 3.3. This section ends by showing that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ concentrates on a specific subset $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (39) of the set $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$. At the light of this observation, the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ is referred to as the *nonnegligible limit set*. Finally, it is shown that when the σ-finite measure Q in (19) is coherent, the sets $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ and $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ are identical.

9.1 The Limit Set

The set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ in (193), with $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), contains all the models that induce an empirical risk that is smaller than or equal to $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$, i.e., the ERM-RER-optimal expected empirical risk in (105). This observation unveils the existence of a relation between the set $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ in (194) and the set $\mathcal{T}(\boldsymbol{z})$ in (5), as shown by the following lemma.

Lemma 9.1. *The set $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ in (194) satisfies*

$$\mathcal{T}(\boldsymbol{z}) \subseteq \mathcal{N}^{\star}_{Q,\boldsymbol{z}}, \tag{195}$$

where the set $\mathcal{T}(\boldsymbol{z})$ is in (5). Moreover,

$$\mathcal{T}(\boldsymbol{z}) = \mathcal{N}^{\star}_{Q,\boldsymbol{z}}, \tag{196}$$

if and only if (a) the ERM problem in (4) possesses a solution; and (b) the reference measure Q in (19) is coherent.

Proof: If the set $\mathcal{T}(\boldsymbol{z})$ in (5) is empty, the inclusion in (195) is trivially true. Assume that $|\mathcal{T}(\boldsymbol{z})| > 0$. Hence, the proof of the inclusion in (195) follows from observing that for all $\boldsymbol{\theta} \in \mathcal{T}(\boldsymbol{z})$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \rho^{\star} \leqslant \delta^{\star}_{Q,\boldsymbol{z}}$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38) and $\rho^{\star}$ in (48). Hence, $\boldsymbol{\theta} \in \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$. This completes the proof of the inclusion in (195).

The proof of the equality in (196) is presented in two parts. In the first part, it is proved that if (196) holds, then the ERM problem in (4) possesses a solution and 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 $\delta^{\star}_{Q,\boldsymbol{z}} = \rho^{\star}$, with $\rho^{\star}$ in (48), which implies that the ERM problem in (4) possesses a solution. Moreover, for all $\delta \in (\rho^{\star}, +\infty)$, it holds that $Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0$, which 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 ERM problem in (4) possesses a solution and the measure Q is coherent, it follows that $\delta^{\star}_{Q,\boldsymbol{z}} = \rho^{\star}$. Hence, $\mathcal{T}(\boldsymbol{z}) = \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$, which completes the proof of the second part. ■

The following theorem highlights that the set $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ is decreasing with λ.

Theorem 9.1. *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* (23) *and $\lambda_1 > \lambda_2$, the sets $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)$ and $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$ in* (193) *satisfy*

$$\mathcal{M} \supseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \supseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \supseteq \mathcal{N}^{\star}_{Q,\boldsymbol{z}}, \tag{197}$$

with $\mathcal{N}^{\star}_{Q,\boldsymbol{z}}$ being the set defined in (194)*. Moreover, if the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in* (3) *is continuous on $\mathcal{M}$ and separable with respect to the measure Q in* (19)*, then,*

$$\mathcal{M} \supset \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \supset \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \supset \mathcal{N}^{\star}_{Q,\boldsymbol{z}}. \tag{198}$$

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 (23), 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 (19) is noncoherent (Definition 4.1). That is, $\delta^{\star}_{Q,\boldsymbol{z}} > \rho^{\star}$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38) and $\rho^{\star}$ in (48). Thus, for all $\gamma \in \left(\rho^{\star}, \delta^{\star}_{Q,\boldsymbol{z}}\right)$, it holds that $Q(\mathcal{L}_{\boldsymbol{z}}(\gamma)) = 0$, with the set $\mathcal{L}_{\boldsymbol{z}}(\cdot)$ in (37). From Lemma 3.3, this implies that for all $\gamma \in \left(\rho^{\star}, \delta^{\star}_{Q,\boldsymbol{z}}\right)$, the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) satisfies $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\gamma)) = 0$, while verifying that $\mathcal{L}_{\boldsymbol{z}}(\gamma) \subseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$. These observations lead to the analysis of the asymptotic concentration of probability in the following section.

9.2 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 $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda))$ increases when λ tends to zero, as shown by the following theorem.

Theorem 9.2. *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* (23) *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* (193) *satisfies*

$$0 < P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) \leqslant P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{199}$$

where strict inequality holds if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ *is separable with respect to the* σ*-finite measure* Q*.*

Proof: The proof is presented in Appendix R. ■

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

Lemma 9.2. *Let the function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (3) *be separable with respect to the* σ*-finite measure* Q *in* (19)*. Let also* $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$*, with* $\mathcal{K}_{Q,\boldsymbol{z}}$ *in* (23)*, be two positive reals such that* $\lambda_1 > \lambda_2$ *and*

$$Q\Big(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right) \cap \left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\right)^{\mathsf{c}}\Big) = 0, \tag{200}$$

with the complement with respect to the set of models $\mathcal{M}$*. 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*

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{201}$$

where, the set $\mathcal{N}_{Q,\boldsymbol{z}}\left(\cdot\right)$ *is defined in* (193)*.*

Proof: The proof is presented in Appendix S. ■

The following example shows the relevance of Lemma 9.2 in the case in which the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is a simple function and separable with respect to the σ-finite measure Q in (19).

Example 9.1. *Consider Example 7.1. Note that, for all* $\lambda > 0$*,*

$$0 < \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < 1, \tag{202}$$

where $\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ *is the ERM-RER-optimal expected empirical risk in* (105)*. The equality in* (202) *implies that given two reals* λ_1 *and* λ_2 *such that* $\lambda_1 > \lambda_2 > 0$*, it holds that,*

$$\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right) \cap \left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\right)^{\mathsf{c}} = \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \leqslant \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\} \tag{203}$$

$$= \emptyset, \tag{204}$$

and moreover, $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) = \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$*. Finally, from Lemma 9.2,*

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)). \tag{205}$$

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

Theorem 9.3. *The probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (25) *satisfies*

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda\right)\right) = 1, \tag{206}$$

where, the set $\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda\right)$ *is defined in* (193)*.*

Proof: The proof follows immediately from Lemma 3.7 and by noticing that for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), the sets $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (39) and $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$ in (193) satisfies $\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \subseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$. ■

Note that Theorem 9.3 and Lemma 3.7 lead to the following conclusion

$$\lim_{\lambda \to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \setminus \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \right) = 0, \tag{207}$$

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 (39). This justifies referring to the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ as the nonnegligible limit set.

10 (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 (4). Such guarantee is defined as follows.

Definition 10.1 ((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 (37) satisfies*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \left(\mathcal{L}_{\boldsymbol{z}}(\delta) \right) > 1 - \epsilon. \tag{208}$$

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 a 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 10.1. *For all $(\delta,\epsilon) \in (\delta^{\star}_{Q,\boldsymbol{z}}, +\infty) \times (0,1)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38), there exists a real $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), such that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ,ϵ)-optimal.*

Proof: Let δ be a real in $\left(\delta^{\star}_{Q,\boldsymbol{z}}, +\infty\right)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38). Let also $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ satisfy the following equality:

$$K^{(1)}_{Q,\boldsymbol{z}} \left(-\frac{1}{\lambda} \right) \leqslant \delta. \tag{209}$$

Note that from Lemma 5.2, it follows that the function $K^{(1)}_{Q,\boldsymbol{z}}$ is continuous. Moreover, from Theorem 6.2, it follows that such a λ in (209) always exists. From (37) and (193), it holds that

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \subseteq \mathcal{L}_{\boldsymbol{z}}(\delta), \tag{210}$$

and thus,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) \geqslant P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda)\right). \tag{211}$$

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

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) > 1-\epsilon. \tag{212}$$

The existence of such a positive real γ follows from Theorem 9.3. Hence, from (212), it holds that,

$$1-\epsilon < P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) \tag{213}$$

$$\leqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right), \tag{214}$$

where the inequality in (214) 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 (214) implies that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ,ϵ)-optimal (Definition 10.1). This completes the proof. ■

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

Theorem 10.2. *For all $(\delta,\epsilon) \in (\rho^\star,+\infty)\times(0,1)$, with $\rho^\star$ in (48), there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), 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: The proof is divided into two parts. The first part shows that if for all $(\delta,\epsilon) \in (\rho^\star,+\infty)\times(0,1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), 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. Let $\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}$ be such that

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 1-\epsilon, \tag{215}$$

then, for all measurable subsets $\mathcal{A}$ of $\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)$, it holds that

$$\begin{aligned} 1-\epsilon &< P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) \\ &= \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\mathrm{d}Q(\boldsymbol{\nu}) + \int_{\mathcal{L}_{\boldsymbol{z}}(\delta)\setminus\mathcal{A}} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\mathrm{d}Q(\boldsymbol{\nu}), \end{aligned} \tag{216}$$

which, together with Lemma 3.2, implies that there exists at least one measurable subset $\mathcal{A}$ for which $Q\left(\mathcal{A}\right) > 0$, and thus,

$$Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > Q\left(\mathcal{A}\right) > 0, \tag{217}$$

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 $\delta^\star_{Q,\boldsymbol{z}} = \rho^\star$. Then, from Theorem 10.1, it

follows that for all $(\delta, \epsilon) \in (\delta^{\star}_{Q,\boldsymbol{z}}, +\infty) \times (0,1)$, there always exists a $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), 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. ■

11 Sensitivity and Generalization

This section introduces the notion of sensitivity and establishes its connections with the notion of generalization error of the Gibbs algorithm, cf. [8].

11.1 Sensitivity

The sensitivity of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}$ in (18) to deviations from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) towards an alternative probability measure $P \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ is introduced as a novel metric to evaluate the generalization capabilities of the ERM-RER-optimal measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Deviations from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ towards an alternative probability measure P would allow comparing the ERM-RER-optimal measure with alternative measures (or algorithms). For instance, if new datasets become available, a new ERM-RER problem can be formulated using a larger dataset obtained by aggregating the old and the new datasets, cf. [10] and [68]. Intuitively, the ERM-RER-optimal measure obtained after the aggregation of datasets might exhibit better generalization capabilities, see for instance [10]. This analysis is the motivation of the sensitivity, which is defined as follows.

Definition 11.1 (Sensitivity)**.** *Given the σ-finite measure Q and the positive real $\lambda > 0$ in* (19)*, 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 functional such that*

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \begin{cases} \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) & \text{if } \lambda \in \mathcal{K}_{Q,\boldsymbol{z}} \\ +\infty & \text{otherwise,} \end{cases} \tag{218}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (18) *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 $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$.*

Recently, the following exact expression for the sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$ in (218) was introduced in [10].

Theorem 11.1 (Theorem 1 in [10])**.** *The sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$ in* (218) *satisfies*

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

where the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is in (25)*.*

The following theorem introduces an upper bound on the absolute value of the sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$ in (218), which requires the calculation of only one of the relative entropies in Theorem 11.1.

Theorem 11.2. *For all $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, the sensitivity $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$ in (218) satisfies*

$$|\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)| \leqslant \sqrt{2\beta_{Q,\boldsymbol{z}}^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)}, \tag{220}$$

where the constant $\beta_{Q,\boldsymbol{z}}$ is defined in (170).

Proof: The proof is presented in Appendix T. ■

Note that equality holds in (220) in the trivial case in which the empirical risk function is not separable with respect to Q (Definition 5.1). In such case, for all $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = 0$ and $\beta_{Q,\boldsymbol{z}} = 0$.

Theorem 11.2 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 ERM-RER problem in (19). More specifically, note that for all probability measures $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that,

$$\mathsf{R}_{\boldsymbol{z}}(P) \geqslant \mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) - \sqrt{2\beta_{Q,\boldsymbol{z}}^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)} \text{ and} \tag{221}$$

$$\mathsf{R}_{\boldsymbol{z}}(P) \leqslant \mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right) + \sqrt{2\beta_{Q,\boldsymbol{z}}^2 D\left(P \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)}. \tag{222}$$

11.2 Generalization Error

This section unveils the interesting connection between the notion of sensitivity and the notion of generalization error of the Gibbs algorithm, cf. [8]. The Generalization error is defined under the assumption that datasets are sampled from a probability measure

$$P_{\boldsymbol{Z}} \in \triangle\left((\mathcal{X} \times \mathcal{Y})^n, \mathscr{F}\right), \tag{223}$$

where $\mathscr{F}$ denotes a given σ-field on the set $(\mathcal{X} \times \mathcal{Y})^n$. For such a probability measure $P_{\boldsymbol{Z}}$ in (223), let the set $\mathcal{K}_{Q,P_{\boldsymbol{Z}}} \subset \mathbb{R}$ be

$$\mathcal{K}_{Q,P_{\boldsymbol{Z}}} = \bigcap_{\boldsymbol{z} \in \operatorname{supp} P_{\boldsymbol{Z}}} \mathcal{K}_{Q,\boldsymbol{z}}, \tag{224}$$

where the σ-finite measure Q is in (19). The set $\mathcal{K}_{Q,P_{\boldsymbol{Z}}}$ in (224) can be empty for some choices of the σ-finite measure Q. Nonetheless, from Lemma 3.1, it follows that if Q is a probability measure, then,

$$\mathcal{K}_{Q,P_{\boldsymbol{Z}}} = (0, +\infty). \tag{225}$$

Under the assumption that datasets are sampled from $P_{\boldsymbol{Z}}$ in (223), the generalization error of the Gibbs algorithm with parameters Q and λ, is defined as the

expectation with respect to the product measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}} \cdot P_{\boldsymbol{Z}}$, with $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in (25), of difference between: (*a*) the *population risk* due to a model $\boldsymbol{\theta} \in \mathcal{M}$,

$$\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right) \tag{226}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (3); and (*b*) The empirical risk induced by the model $\boldsymbol{\theta}$ with respect to a training dataset $\boldsymbol{z}$, that is, $\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)$. More specifically, the generalization error of the Gibbs algorithm with parameters Q and λ is

$$\begin{aligned}
&\int\int \left(\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right) - \mathsf{L}_{\boldsymbol{\nu}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\left(\boldsymbol{\theta}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \\
=&\int \left(\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\right) \\
&- \int \mathsf{L}_{\boldsymbol{\nu}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\left(\boldsymbol{\theta}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \qquad (227) \\
=&\int \left(\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right) \\
&- \int \mathsf{L}_{\boldsymbol{\nu}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\left(\boldsymbol{\theta}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \qquad (228) \\
=&\int \left(\mathsf{R}_{\boldsymbol{\nu}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right) - \mathsf{R}_{\boldsymbol{\nu}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right)\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), \qquad (229)
\end{aligned}$$

where the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}}$ satisfies for all sets $\mathcal{A} \in \mathscr{B}\left(\mathcal{M}\right)$,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\left(\mathcal{A}\right) = \int P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\left(\mathcal{A}\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{\nu}\right), \tag{230}$$

and the functional $\mathsf{R}_{\boldsymbol{\nu}}$ is defined in (18).

The following theorem establishes a connection between sensitivity and generalization error in the particular case in which Q in (19) is a probability measure.

Theorem 11.3. *Under the assumption that datasets are sampled from $P_{\boldsymbol{Z}}$ in* (223)*, the generalization error of the Gibbs algorithm with parameters Q (a probability measure) and $\lambda > 0$, is*

$$\int \mathsf{S}_{Q,\lambda}\left(\boldsymbol{\nu}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), \tag{231}$$

where the functional $\mathsf{S}_{Q,\lambda}$ is in (218)*; and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}}$ is in* (230)*.*

Proof: The proof uses the fact that under the assumption that Q is a probability measure, for all $\boldsymbol{\nu} \in \operatorname{supp} P_{\boldsymbol{Z}}$, it follows from Lemma 3.1 that $\mathcal{K}_{Q,\boldsymbol{\nu}} = (0, +\infty)$. This implies that for all $\boldsymbol{z} \in \operatorname{supp} P_{\boldsymbol{Z}}$ and for all $\lambda > 0$, the ERM-RER problem in (19), always possesses as solution the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25). Thus,

the measure $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ in (230) is well defined. Moreover, $\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right) = \mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right) - \mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)$ and the integral in (229) is also well defined, which completes the proof. ■

Theorem 11.3 provides an interesting viewpoint of the generalization error. For instance, the probability measure $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ in (223) can be understood as the barycenter of a subset of $\triangle\left(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right)\right)$ containing the solutions to ERM-RER problems of the form in (19), with $\boldsymbol{z} \in \operatorname{supp} P_{\boldsymbol{Z}}$ in (223). Hence, the generalization error of the Gibbs algorithm is the expectation (with respect to $P_{\boldsymbol{Z}}$) of the sensitivity of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}}$ in (18) to variations from the ERM-RER-optimal measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}$ towards the barycenter, i.e., the measure $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$.

The following definition extends the notion of generalization error to Gibbs algorithms obtained by assuming that the reference measure Q in (19) is a σ-finite measure. This definition also exploits the relation between the notions of sensitivity and generalization error introduced by Theorem 11.3.

Definition 11.2 (Generalization Error of the Gibbs Algorithm)**.** *Given a σ-finite measure $Q \in \triangle\left(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right)\right)$ and a real $\lambda > 0$, let the functional $\mathsf{G}_{Q,\lambda} : \triangle\left(\left(\mathcal{X} \times \mathcal{Y}\right)^n, \mathscr{F}\right) \to (-\infty, +\infty]$ be such that*

$$\mathsf{G}_{Q,\lambda}(P_{\boldsymbol{Z}}) = \begin{cases} \displaystyle\int \mathsf{S}_{Q,\lambda}\left(\boldsymbol{\nu}, P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) & \text{if } \lambda \in \mathcal{K}_{Q,P_{\boldsymbol{Z}}} \\ +\infty & \text{otherwise,} \end{cases} \tag{232}$$

where the functional $\mathsf{S}_{Q,\lambda}$ is in (218)*; the set $\mathcal{K}_{Q,P_{\boldsymbol{Z}}}$ is in* (224)*; and the probability measure $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ is in* (230)*. The generalization error induced by the Gibbs algorithm with parameters Q and λ under the assumption that datasets are sampled from the probability measure $P_{\boldsymbol{Z}}$, is $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$.*

The main difficulty for extending the notion of generalization error to Gibbs algorithms obtained under the assumption that the reference measure is not a probability measure, but a σ-finite measure, is that the integrals in (229) and (230) might not be well defined. This is essentially due to the fact that, while the ERM-RER problem in (19) always possesses a solution when Q is a probability measure, the existence of a solution when Q is not a probability measure is subject to the condition that for all $\boldsymbol{z} \in \operatorname{supp} P_{\boldsymbol{Z}}$, $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23). This leads to the condition that $\lambda \in \mathcal{K}_{Q,P_{\boldsymbol{Z}}}$, with the set $\mathcal{K}_{Q,P_{\boldsymbol{Z}}}$ in (224). When such a condition is not met, the definition of sensitivity is void.

The following theorem provides a closed-form expression for the generalization error of the Gibbs algorithm in the general case in which the reference measure Q in (19) is a σ-finite measure.

Theorem 11.4. *If $\lambda \in \mathcal{K}_{Q,P_{\boldsymbol{Z}}}$, with $\mathcal{K}_{Q,P_{\boldsymbol{Z}}}$ in* (224)*, the generalization error*

$\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ *in* (232) *satisfies*

$$\begin{aligned}&\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)\\&=\lambda\left(\int D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\Big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})+\int D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\Big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})\right),\end{aligned}\tag{233}$$

where for all $\boldsymbol{z}\in\operatorname{supp}P_{\boldsymbol{Z}}$*, the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *is in* (25)*; and the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}}$ *is defined in* (230)*.*

Proof: The proof is presented in Appendix U. ∎

The terms $\int D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})$ and $\int D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})$ in the right-hand side of (233) are respectively the mutual and the lautum information [52] induced by a joint probability measure $P_{\boldsymbol{\Theta},\boldsymbol{Z}}$ whose marginals are $P_{\boldsymbol{Z}}$ in (223) and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}}$ in (230). When the reference measure Q in (19) is a probability measure, Theorem 11.4 reduces to [8, Theorem 1]. Interestingly, independently of whether the reference measure Q in (19) is a probability measure, or whether the n data points in the datasets are independent and identically distributed, the generalization error $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ in (232) is always a factor of the sum of the mutual and lautum information induced by the joint probability measure $P_{\boldsymbol{\Theta},\boldsymbol{Z}}$ mentioned above.

Theorem 11.4 also provides an alternative interpretation of the generalization error $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ in (232). Note that by writing one of the factors in the right-hand side of (233) as

$$\int\Big(D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\Big)+D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\Big)\Big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}),$$

it becomes clear that $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ is the expectation with respect to $P_{\boldsymbol{Z}}$ of the symmetrized Kullback-Leibler divergence, also known as Jeffrey's divergence [65], of the probability measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}}$. That is, the solution to the ERM-RER problem in (19) and the barycenter induced by $P_{\boldsymbol{Z}}$.

The following theorem provides an upper-bound on the generalization error of the Gibbs algorithm only in terms of the lautum information induced by such a joint probability measure $P_{\boldsymbol{\Theta},\boldsymbol{Z}}$.

Theorem 11.5. *The generalization error* $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ *in* (232) *satisfies for all* $\lambda\in\mathcal{K}_{Q,P_{\boldsymbol{Z}}}$*,*

$$0\leqslant\mathsf{G}_{Q,\lambda}(P_{\boldsymbol{Z}})\leqslant\sqrt{2\sigma_Q^2\int D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\Big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})},\tag{234}$$

where for all $\boldsymbol{z}\in\operatorname{supp}P_{\boldsymbol{Z}}$*, the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *is in* (25)*; the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}}$ *is defined in* (230)*; and*

$$\sigma_Q=\sup\left\{\beta_{Q,\boldsymbol{z}}:\boldsymbol{z}\in(\mathcal{X}\times\mathcal{Y})^n\right\},\tag{235}$$

with $\beta_{Q,\boldsymbol{z}}$ *in* (165)*.*

Proof: The proof of the inequality $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right) \geqslant 0$ follows from observing that for all $\boldsymbol{\nu} \in (\mathcal{X} \times \mathcal{Y})^n$, the terms $D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right)$ and $D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right)$ in (233) are nonnegative (Theorem 2.1). The proof of the remaining inequality follows from (232) and the following inequalities:

$$\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right) = \left| \int \mathsf{S}_{Q,\lambda}\left(\boldsymbol{\nu}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \right| \tag{236}$$

$$\leqslant \int \left| \mathsf{S}_{Q,\lambda}\left(\boldsymbol{\nu}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right) \right| \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \tag{237}$$

$$\leqslant \int \sqrt{2\beta_{Q,\boldsymbol{\nu}} D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), \tag{238}$$

$$\leqslant \int \sqrt{2\sigma_Q^2 D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \tag{239}$$

$$\leqslant \sqrt{2\sigma_Q^2 \int D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})}, \tag{240}$$

where the equality in (236) follows from (232); the inequality in (237) follows from [55, Theorem 1.5.9(c)]; the inequality in (238) follows from Theorem 11.2; the inequality in (239) follows from (235); and the inequality in (240) follows from Jensen's inequality [55, Section 6.3.5]. This completes the proof. ■

In a nutshell, the generalization error $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ in (232) is upper bounded up to a constant factor by the square root of the lautum information induced by the joint probability measure $P_{\boldsymbol{\Theta},\boldsymbol{Z}}$ mentioned above. Theorem 11.5 is reminiscent of [29, Theorem 1], which provides a similar upper-bound on $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ using the mutual information instead of the lautum information induced by the joint probability measure $P_{\boldsymbol{\Theta},\boldsymbol{Z}}$. The interest in Theorem 11.5 for the specific case of the Gibbs algorithm, lies on the fact that it holds under milder conditions than those in [29, Theorem 1]. For instance, no additional conditions on the loss function ℓ in (2) concerning sub-Gaussianity are assumed. Moreover, the probability measure $P_{\boldsymbol{Z}}$ from which datasets are sampled is not necessarily a product measure.

12 Conclusions and Final Remarks

The classical ERM-RER problem in (19) 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 that includes the ERM problem with (discrete or differential) entropy regularization and the information-risk minimization problem. While in the case in which the reference measure is a probability measure the solution to the ERM-RER problem always exists, in this general case, the existence of a solution is subject to a condition that depends on the loss function, the reference measure, the regularization factor, and the training dataset. When a solution exists, it has been proved that it is unique. Additionally, if it exists, such a solution and the reference measure are mutually

absolutely continuous in most of the practical cases of interest. Interestingly, the empirical risk observed when models are sampled from the 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 ERM-RER problem such that the set of models that induce an empirical risk smaller than δ exhibits a probability that is not smaller that $1-\epsilon$. Interestingly, none of these results relies on statistical assumptions on the datasets.

The sensitivity of the expected empirical risk to deviations from the ERM-RER-optimal measure to alternative measures is introduced as a new performance metric to evaluate the generalization capabilities of the Gibbs algorithm. In particular, an upper bound on the absolute value of the sensitivity, which depends on the training dataset, is presented. This bound is formed by a constant factor and the square root of the relative entropy of the alternative measure (the deviation) with respect to the ERM-RER solution. Finally, it is shown that the expectation of the sensitivity (with respect to the datasets) to deviations towards a particular measure is equivalent to the generalization error of the Gibbs algorithm. Equipped with this observation, the generalization error is shown to be in the most general case, up to a constant factor, the sum of the mutual and lautum information between the models and the datasets, which was a result known exclusively for the case in which the reference is a probability measure, cf. [8]. From this perspective, it is argued that the study of the generalization capabilities of the Gibbs algorithm based on generalization error is a significantly narrow view. This is essentially because it is looking at an expectation of the sensitivity to deviations to a particular measure, i.e., the barycenter of the set of ERM-RER solutions induced by a prior on the datasets. A broader view is offered by the study of the sensitivity to deviations towards other measures, i.e., ERM-RER-optimal measures obtained with different training data sets. This approach has lead already to a few initial results in [10] that highlight the connections to sensitivity, training error, and test error. Nonetheless, the study of the sensitivity in the aim of describing the generalization capabilities of learning algorithms remains by now as an open problem.

Appendices

A Proof of Theorem 2.2

Consider the function $f : [0, +\infty) \to \mathbb{R}$ such that

$$f(x) = \begin{cases} x \log(x) & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0, \end{cases} \tag{241}$$

and note that it is strictly convex. From the assumption that for all $i \in \{1,2\}$, P_i and Q_i are both measures on the same measurable space $(\Omega, \mathscr{F})$, with P_i absolutely continuous with respect to Q_i, let $g : \Omega \to [0,\infty)$ be the function

$$g(x) = \frac{\mathrm{d}\left(\lambda P_1 + (1-\lambda)P_2\right)}{\mathrm{d}\left(\lambda Q_1 + (1-\lambda)Q_2\right)}(x), \tag{242}$$

where $\frac{\mathrm{d}(\lambda P_1+(1-\lambda)P_2)}{\mathrm{d}(\lambda Q_1+(1-\lambda)Q_2)}$ is the Radon-Nikodym derivative of the measure $\lambda P_1 + (1-\lambda)P_2$ with respect to $\lambda Q_1 + (1-\lambda)Q_2$. Using this notation, for all $\lambda \in (0,1)$,

$$\begin{aligned}
&D\left(\lambda P_1 + (1-\lambda)P_2 \| \lambda Q_1 + (1-\lambda)Q_2\right) \\
&-\lambda D\left(P_1 \| Q_1\right) + (1-\lambda) D\left(P_2 \| Q_2\right) \qquad (243)\\
=&\int \log(g(x)) \mathrm{d}(\lambda P_1 + (1-\lambda)P_2)(x) \\
&-\lambda \int \log\Big(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)\Big)\mathrm{d}P_1(x) - (1-\lambda)\int \log\Big(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)\Big)\mathrm{d}P_2(x) \\
=&\lambda \int \log\left(g\left(x\right)\right)\mathrm{d}P_1(x) + (1-\lambda)\int \log\left(g\left(x\right)\right)\mathrm{d}P_2(x) \\
&-\lambda \int \log\Big(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)\Big)\mathrm{d}P_1(x) - (1-\lambda)\int \log\Big(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)\Big)\mathrm{d}P_2(x) \\
=&\lambda \int \log\left(\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)\right)^{-1} g(x)\right)\mathrm{d}P_1(x) \\
&+(1-\lambda)\int \log\left(\left(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)\right)^{-1} g(x)\right)\mathrm{d}P_2(x) \\
=&\lambda \int \frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x) \log\left(\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)\right)^{-1} g(x)\right)\mathrm{d}Q_1(x) \\
&+(1-\lambda)\int \frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x) \log\left(\left(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)\right)^{-1} g(x)\right)\mathrm{d}Q_2(x) \\
=&\lambda \int \frac{g(x)\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)}{g(x)} \log\left(\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)\right)^{-1} g(x)\right)\mathrm{d}Q_1(x) \\
&+(1-\lambda)\int \frac{g(x)\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)}{g(x)} \log\left(\left(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)\right)^{-1} g(x)\right)\mathrm{d}Q_2(x) \\
=&-\lambda \int g(x) f\Big(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)(g(x))^{-1}\Big)\mathrm{d}Q_1(x) - (1-\lambda)\int g(x) f\Big(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)(g(x))^{-1}\Big)\mathrm{d}Q_2(x), \qquad (244)
\end{aligned}$$

where the function f is defined in (241). Let β_1 and β_2 be the following constants:

$$\beta_1 \triangleq \int g(\nu)\mathrm{d}Q_1(\nu) \text{ and } \beta_2 \triangleq \int g(\nu)\mathrm{d}Q_2(\nu). \tag{245}$$

From (244) and (245), it follows that for all $\lambda \in (0,1)$,

$$
\begin{aligned}
& D\left(\lambda P_1+(1-\lambda)P_2 \| \lambda Q_1+(1-\lambda)Q_2\right)-\lambda D\left(P_1 \| Q_1\right)+(1-\lambda) D\left(P_2 \| Q_2\right) \\
=&-\lambda\beta_1 \int \frac{g(x)}{\beta_1} f\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)(g(x))^{-1}\right) \mathrm{d}Q_1(x) \\
&-(1-\lambda)\beta_2 \int \frac{g(x)}{\beta_2} f\left(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)(g(x))^{-1}\right) \mathrm{d}Q_2(x) && (246)\\
\leqslant&-\lambda\beta_1 f\left(\int \frac{g(x)}{\beta_1}\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x)(g(x))^{-1}\mathrm{d}Q_1(x)\right) && (247)\\
&-(1-\lambda)\beta_2 f\left(\int \frac{g(x)}{\beta_2}\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x)(g(x))^{-1}\mathrm{d}Q_2(x)\right) \\
=&-\lambda\beta_1 f\left(\frac{1}{\beta_1}\int \mathrm{d}P_1(x)\right)-(1-\lambda)\beta_2 f\left(\frac{1}{\beta_2}\int \mathrm{d}P_2(x)\right) && (248)\\
=&-\lambda\beta_1 f\left(\frac{1}{\beta_1}\right)-(1-\lambda)\beta_2 f\left(\frac{1}{\beta_2}\right) && (249)\\
\leqslant&-f\left(\lambda\beta_1\frac{1}{\beta_1}+(1-\lambda)\beta_2\frac{1}{\beta_2}\right) && (250)\\
=&-f(1) && (251)\\
=&0, && (252)
\end{aligned}
$$

where the inequalities in (247) and (250) follow from Jensen's inequality [55, Section 6.3.5] and the fact that the function f in (244) is strictly concave. Note that from (245), in (247), for all $i \in \{1,2\}$, $\int \frac{g(x)}{\beta_i}\mathrm{d}Q_i(x) = 1$; while in (250),

$$
\begin{aligned}
\lambda\beta_1+(1-\lambda)\beta_2 =& \int g(\nu)\mathrm{d}(\lambda Q_1+(1-\lambda)Q_2)(\nu) && (253)\\
=& \int \mathrm{d}\left(\lambda P_1+(1-\lambda)P_2\right)(\nu) && (254)\\
=& \lambda\int \mathrm{d}P_1(\nu)+(1-\lambda)\int \mathrm{d}P_2(\nu) && (255)\\
=& 1. && (256)
\end{aligned}
$$

Given the strict convexity of the function f in (241), equality in (247) and (250) hold if and only if $P_1 = P_2$ and $Q_1 = Q_2$. This completes the proof.

B Proof of Lemma 3.1

The proof is divided into two parts. The first part is as follows. Under the assumption that the set $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23) is empty, there is nothing to prove. Alternatively, 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 $\boldsymbol{\theta} \in \mathcal{M}$,

$$
\frac{\mathrm{d}}{\mathrm{d}t}\exp\left(-\frac{1}{t}\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) = \frac{1}{t^2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\exp\left(-\frac{1}{t}\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \geqslant 0, \tag{257}
$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (3). Thus, from (22), it follows that $K_{Q,\boldsymbol{z}}\left(-\frac{1}{b}\right)$ is nondecreasing with b. This implies that $(0, b] \subseteq \mathcal{K}_{Q,\boldsymbol{z}}$.

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

$$b^\star = \sup \mathcal{K}_{Q,\boldsymbol{z}}. \tag{258}$$

Hence, if $b^\star = +\infty$, it follows from (23) that

$$\mathcal{K}_{Q,\boldsymbol{z}} = (0, +\infty). \tag{259}$$

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

$$(0, b^\star) \subseteq \mathcal{K}_{Q,\boldsymbol{z}} \subseteq (0, b^\star]. \tag{260}$$

In either case, it follows that $\mathcal{K}_{Q,\boldsymbol{z}}$ is a convex set. This completes the first part of the proof.

The second part of the proof is under the assumption that Q is a probability measure. Under this assumption, for all $\boldsymbol{\theta} \in \mathcal{M}$ and for all for all $t > 0$, it follows that

$$\exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \leqslant 1, \tag{261}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (3). Thus,

$$K_{Q,\boldsymbol{z}}\left(-\frac{1}{t}\right) = \log\left(\int \exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right) \tag{262}$$

$$\leqslant \log\left(\int \mathrm{d}Q(\boldsymbol{\theta})\right) \tag{263}$$

$$= 0, \tag{264}$$

which implies that $(0, +\infty) \subseteq \mathcal{K}_{Q,\boldsymbol{z}}$. Thus, if Q is a probability measure, from (23), it holds that $\mathcal{K}_{Q,\boldsymbol{z}} = (0, +\infty)$, which completes the proof.

C Proof of Theorem 3.1

The optimization problem in (19) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the measure Q, denoted by $\frac{\mathrm{d}P}{\mathrm{d}Q} : \mathcal{M} \to [0, \infty)$, which yields:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) + \lambda \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{265a}$$

$$\text{s. t.} \qquad \int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \mathrm{d}Q\left(\boldsymbol{\theta}\right) = 1. \tag{265b}$$

The remainder of the proof focuses on the problem in which the optimization is over the function $\frac{\mathrm{d}P}{\mathrm{d}Q}$ instead of the measure P. This is due to the fact that for all $P \in \triangle_Q\left(\mathcal{M}\right)$, the Radon-Nikodym derivate $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is unique up to sets of zero measure with respect to the measure Q. Let $\mathscr{M}$ be the set of measurable

functions $\mathcal{M} \to \mathbb{R}$ with respect to the measurable spaces $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ that are absolutely integrable with respect to Q. That is, for all $\hat{g} \in \mathscr{M}$, it holds that

$$\int |\hat{g}(\boldsymbol{\theta})| \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{266}$$

Hence, the optimization problem of interest is:

$$\min_{g \in \mathscr{M}} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) + \lambda \int g(\boldsymbol{\theta}) \log(g(\boldsymbol{\theta})) \mathrm{d}Q(\boldsymbol{\theta}) \tag{267a}$$

$$\text{s. t.} \int g(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{267b}$$

Let the Lagrangian of the optimization problem in (267) be the functional $L : \mathscr{M} \times \mathbb{R} \to \mathbb{R}$ such that

$$\begin{aligned} L(g, \beta) = & \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) g(\boldsymbol{\nu}) \mathrm{d}Q(\boldsymbol{\nu}) + \lambda \int g(\boldsymbol{\nu}) \log(g(\boldsymbol{\nu})) \mathrm{d}Q(\boldsymbol{\nu}) \\ & + \beta \left(\int g(\boldsymbol{\nu}) \, \mathrm{d}Q(\boldsymbol{\nu}) - 1 \right), \end{aligned} \tag{268}$$

where β is a real that acts as a Lagrangian multiplier due to the constraint (267b). Let $\hat{g} : \mathcal{M} \to \mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (268) at $(g, \beta) \in \mathscr{M} \times \mathbb{R}$ in the direction of $\hat{g}$, if it exists, is

$$\partial L(g, \beta; \hat{g}) \triangleq \frac{\mathrm{d}}{\mathrm{d}\gamma} L(g + \gamma \hat{g}, \beta) \bigg|_{\gamma = 0}. \tag{269}$$

The proof continues under the assumption that the functions g and $\hat{g}$ are such that the Gateaux differential in (269) exists. Under such an assumption, let the function $r : \mathbb{R} \to \mathbb{R}$ satisfy for all $\alpha \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small, that

$$\begin{aligned} r(\alpha) = & \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \left(g(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu}) \right) \mathrm{d}Q(\boldsymbol{\nu}) \\ & + \beta \left(\int \left(g(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu}) \right) \mathrm{d}Q(\boldsymbol{\nu}) - 1 \right) \\ & + \lambda \int (\hat{g}(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu})) \log(g(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu})) \mathrm{d}Q(\boldsymbol{\nu}) \end{aligned} \tag{270}$$

$$\begin{aligned} = & \int g(\boldsymbol{\nu}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \beta \right) \mathrm{d}Q(\boldsymbol{\nu}) - \beta \\ & + \alpha \left(\int \hat{g}(\boldsymbol{\nu}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \beta \right) \mathrm{d}Q(\boldsymbol{\nu}) \right) \\ & + \lambda \int (g(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu})) \log(g(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu})) \mathrm{d}Q(\boldsymbol{\nu}), \end{aligned} \tag{271}$$

where the last equality is simply an algebraic re-arrangement of terms. From the assumption that the functions g and $\hat{g}$ are such that the Gateaux differential

in (269) exists, it follows that the function r in (271) is differentiable at zero. Note that the first two terms in (271) are independent of α; the third term is linear with α; and the fourth term can be written using the function $\hat{r} : \mathbb{R} \to \mathbb{R}$ such that for all $\alpha \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small, satisfies

$$\hat{r}(\alpha) = \lambda \int (g(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu})) \log(g(\boldsymbol{\nu}) + \alpha \hat{g}(\boldsymbol{\nu})) \mathrm{d}Q(\boldsymbol{\nu}) \tag{272}$$

$$= \lambda \int f\left(g\left(\boldsymbol{\nu}\right) + \alpha \hat{g}\left(\boldsymbol{\nu}\right)\right) \mathrm{d}Q\left(\boldsymbol{\nu}\right), \tag{273}$$

where $f : (0, +\infty) \to \mathbb{R}$ is such that $f(t) = t \log(t)$. Under the same assumption, it follows that the function $\hat{r}$ in (272) is differentiable at zero. That is, the limit

$$\lim_{\delta \to 0} \frac{1}{\delta} (\hat{r}(\gamma + \delta) - \hat{r}(\gamma)) \tag{274}$$

exists for all $\gamma \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small. Note that the function f in (273) is continuous and differentiable (with finite derivate) in $(0, +\infty)$. Thus, the function f is also Lipschitz continuous. This implies that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, and for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, it holds that

$$|f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))| \leqslant c |\hat{g}(\boldsymbol{\theta})| |\delta|, \tag{275}$$

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

$$\left| \frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))}{\delta} \right| \leqslant c |\hat{g}(\boldsymbol{\theta})|. \tag{276}$$

Using these arguments, the limit in (274) satisfies for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, that

$$\begin{aligned}
&\lim_{\delta \to 0} \frac{1}{\delta} (\hat{r}(\gamma + \delta) - \hat{r}(\gamma)) \\
&= \lambda \lim_{\delta \to 0} \int \frac{f(g(\boldsymbol{\theta}) + (\gamma + \delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))}{\delta} \mathrm{d}Q(\boldsymbol{\theta}) \\
&= \lambda \int \dot{f}(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta})) \hat{g}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) && (277) \\
&< \infty, && (278)
\end{aligned}$$

where the function $\dot{f} : (0, +\infty) \to \mathbb{R}$ is the derivative of f. That is, $\dot{f}(t) = 1 + \log(t)$. The equality in (277) and the inequality in (278) follow from noticing that the conditions for the dominated convergence theorem hold [55, Theorem 1.6.9], namely:

- For all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$, the inequality in (276) holds;
- The function $\hat{g}$ in (276) satisfies the inequality in (266); and

- For all $\boldsymbol{\theta} \in \operatorname{supp} Q$ and for all $\gamma \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small, it holds that

$$\lim_{\delta \to 0} \frac{f(g(\boldsymbol{\theta}) + (\gamma+\delta)\hat{g}(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}{\delta} = \frac{\mathrm{d}}{\mathrm{d}\gamma} f(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta})) \quad (279)$$

$$= \dot{f}(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\hat{g}(\boldsymbol{\theta}). \quad (280)$$

Hence, the derivative of the real function r in (271) is

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} r(\alpha) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\hat{g}(\boldsymbol{\nu})\mathrm{d}Q(\boldsymbol{\nu}) + \beta \int \hat{g}(\boldsymbol{\nu})\mathrm{d}Q(\boldsymbol{\nu}) + \lambda \int \hat{g}(\boldsymbol{\nu})(1+\log(g(\boldsymbol{\nu})+\alpha\hat{g}(\boldsymbol{\nu})))\mathrm{d}Q(\boldsymbol{\nu}). \quad (281)$$

From (269) and (281), it follows that

$$\partial L\,(g, \beta; \hat{g}) = \int \hat{g}(\boldsymbol{\nu})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \lambda(1+\log(g(\boldsymbol{\nu}))) + \beta)\mathrm{d}Q(\boldsymbol{\nu}). \quad (282)$$

The relevance of the Gateaux differential in (282) stems from [69, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (268) to have a minimum at $\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta\right) \in \mathscr{M} \times \mathbb{R}$ is that for all functions $\hat{g} \in \mathscr{M}$,

$$\partial L\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta; \hat{g}\right) = 0. \quad (283)$$

From (283), it follows that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ must satisfy for all functions $\hat{g}$ in $\mathscr{M}$ that

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

which implies that for all $\boldsymbol{\nu} \in \operatorname{supp} Q$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \lambda\left(1 + \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\right)\right) + \beta = 0, \quad (284)$$

and thus,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) = \exp\left(-\frac{\beta+\lambda}{\lambda}\right)\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right), \quad (285)$$

with β chosen to satisfy (265b). That is,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta})} \quad (286)$$

$$= \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right). \quad (287)$$

The proof continues by verifying that the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that satisfies (286) is the unique solution to the ERM-RER problem in (19). Such verification is done by showing that the objective function in (19) is strictly convex with the optimization variable. Let P_1 and P_2 be two different probability measures in $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and let α be in $(0,1)$. Hence,

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

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (18). The equality above follows from the properties of the Lebesgue integral, while the inequality follows from Theorem 2.2. This proves that the solution is unique due to the strict concavity of the objective function, which completes the proof.

D Proof of Lemma 3.2

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

$$\begin{aligned}
\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) =& \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) && (288)\\
\leqslant& \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) && (289)\\
<& +\infty, && (290)
\end{aligned}$$

where the inequality in (289) follows from the fact that the function $\mathsf{L}_{\boldsymbol{z}}$ is non-negative; and the equality in (290) 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 $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$ and thus, $K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty$, and thus, $\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) > 0$. Moreover, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, 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

$$\begin{aligned}
\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) =& \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) && (291)\\
=& \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) && (292)\\
\geqslant& 0, && (293)
\end{aligned}$$

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

E Proof of Lemma 3.3

The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) satisfies for all $\mathcal{C} \in \mathscr{B}(\mathcal{M})$,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = \int_{\mathcal{C}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{294}$$

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

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = 0, \tag{295}$$

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

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

$$0 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) \tag{296}$$

$$= \int_{\mathcal{C}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{297}$$

From Lemma 3.2, and the assumption $Q(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = +\infty\}) = 0$, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) > 0, \tag{298}$$

which implies that

$$\int_{\mathcal{C}} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) = 0, \tag{299}$$

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

F Proof of Lemma 3.4

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

$$g(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \left(\frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1}, \tag{300}$$

and note that for all $\boldsymbol{\theta} \in \operatorname{supp} Q \setminus \left\{\boldsymbol{\nu} \in \mathcal{M}\colon \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = +\infty\right\}$, $g(\boldsymbol{\theta}) > 0$. Alternatively, 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 $\mathcal{A} \in \mathscr{B}(\mathcal{M})$,

$$P(\mathcal{A}) = \int_{\mathcal{A}} g(\boldsymbol{\theta}) \mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{301}$$

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 respect to $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})$. Moreover, from (301), it follows that

$$P(\mathcal{A}) = \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \left(\frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{302}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \left(\frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \frac{\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) \tag{303}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{304}$$

$$= \int_{\mathcal{A}} \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{305}$$

$$= P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}), \tag{306}$$

which implies that the probability measures P in (301) and $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ are identical. Thus, $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. The proof that $P^{(Q,\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ follows the same argument. This completes the proof.

G Proof of Lemma 3.6

From Theorem 3.1, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\nu})} \tag{307}$$

$$= \left(\exp\left(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} \tag{308}$$

$$= \left(\int \exp\left(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\right) \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1}. \tag{309}$$

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:

$$\mathcal{A}_0(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = 0\}, \tag{310a}$$

$$\mathcal{A}_1(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < 0\}, \text{ and} \tag{310b}$$

$$\mathcal{A}_2(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > 0\}. \tag{310c}$$

Using the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (309), the following holds for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \Bigg(\int_{\mathcal{A}_0(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big)\mathrm{d}Q(\boldsymbol{\nu}) \\ &\quad + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big)\mathrm{d}Q(\boldsymbol{\nu}) \\ &\quad + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big)\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1} \qquad (311) \\ &= \Bigg(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big)\mathrm{d}Q(\boldsymbol{\nu}) \\ &\quad + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big)\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1}. \qquad (312)\end{aligned}$$

Note that the sets

$$\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\}, \qquad (313)$$
$$\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star_{Q,\boldsymbol{z}}\}, \text{ and} \qquad (314)$$
$$\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^\star_{Q,\boldsymbol{z}}\}, \qquad (315)$$

with $\delta^\star_{Q,\boldsymbol{z}}$ in (38), 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 $\boldsymbol{\theta} \in \big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\big\}$. The second part considers the case in which $\boldsymbol{\theta} \in \big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star_{Q,\boldsymbol{z}}\big\}$. The third part considers the remaining case.

The first part is as follows. Consider that $\boldsymbol{\theta} \in \big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\big\}$ and note that $\big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\big\} = \mathcal{L}^\star_{Q,\boldsymbol{z}}$. Hence, the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (310) satisfy the following:

$$\mathcal{A}_0(\boldsymbol{\theta}) = \mathcal{L}^\star_{Q,\boldsymbol{z}}, \qquad (316a)$$
$$\mathcal{A}_1(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta^\star_{Q,\boldsymbol{z}}\}, \text{ and} \qquad (316b)$$
$$\mathcal{A}_2(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star_{Q,\boldsymbol{z}}\}. \qquad (316c)$$

From the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (38), it follows that $Q(\mathcal{A}_2(\boldsymbol{\theta})) = 0$. Plugging the equalities in (316) in (312) yields for all $\boldsymbol{\theta} \in \big\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\big\}$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \Bigg(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big)\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1}. \qquad (317)$$

The equality in (317) implies that for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}$,

$$\lim_{\lambda \to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \Bigg(\lim_{\lambda \to 0^+} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big) \mathrm{d}Q(\boldsymbol{\nu}) + \tag{318}$$

$$+ Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) \Bigg)^{-1} \tag{319}$$

$$= \begin{cases} +\infty & \text{if } Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0 \\ \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} & \text{otherwise.} \end{cases} \tag{320}$$

where the equality in (320) follows from verifying that the dominated convergence theorem [55, 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

$$\lim_{\lambda \to 0^+} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) = 0. \tag{321}$$

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 $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta\right\}$, the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (310) satisfy the following:

$$\mathcal{A}_0(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = \delta\}, \tag{322a}$$

$$\mathcal{A}_1(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta\}, \text{ and} \tag{322b}$$

$$\mathcal{A}_2(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta\}. \tag{322c}$$

Consider the sets

$$\mathcal{A}_{2,1}(\boldsymbol{\theta}) \triangleq \left\{\boldsymbol{\mu} \in \mathcal{A}_2(\boldsymbol{\theta}) : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{323}$$

$$\mathcal{A}_{2,2}(\boldsymbol{\theta}) \triangleq \left\{\boldsymbol{\mu} \in \mathcal{A}_2(\boldsymbol{\theta}) : \delta^{\star}_{Q,\boldsymbol{z}} \leqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta\right\}, \tag{324}$$

and note that $\mathcal{A}_{2,1}(\boldsymbol{\theta})$ and $\mathcal{A}_{2,2}(\boldsymbol{\theta})$ form a partition of $\mathcal{A}_2(\boldsymbol{\theta})$. Moreover, from the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38), it holds that

$$Q\left(\mathcal{A}_{2,1}(\boldsymbol{\theta})\right) = 0. \tag{325}$$

Hence, plugging the equalities in (322) and (325) in (312) yields, for all $\delta > \delta^{\star}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta\}$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \Bigg(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\Big(\frac{1}{\lambda}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}))\Big) \mathrm{d}Q(\boldsymbol{\nu}) \tag{326}$$

$$+ \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}Q(\boldsymbol{\nu}) \Bigg)^{-1}. \tag{327}$$

The equality in (327) implies that for all $\delta > \delta^{\star}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \mathcal{M}\colon \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta\right\}$,

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \Bigg(\lim_{\lambda\to 0^+} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}Q(\boldsymbol{\nu}) \tag{328}$$

$$+ \lim_{\lambda\to 0^+} \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}Q(\boldsymbol{\nu}) \tag{329}$$

$$+ Q\left(\mathcal{A}_0(\boldsymbol{\theta})\right) \Bigg)^{-1} \tag{330}$$

$$= \Bigg(\lim_{\lambda\to 0^+} \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}Q(\boldsymbol{\nu})$$

$$+ Q\left(\mathcal{A}_0(\boldsymbol{\theta})\right) \Bigg)^{-1} \tag{331}$$

$$= \left(Q\left(\mathcal{A}_0(\boldsymbol{\theta})\right) + \infty\right)^{-1} \tag{332}$$

$$= 0, \tag{333}$$

where the equality in (331) follows by verifying that the dominated convergence theorem [55, 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

$$\lim_{\lambda\to 0^+} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) = 0. \tag{334}$$

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

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{335}$$

This completes the third part and completes the proof.

H Proof of Lemma 3.7

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

$$\mathcal{A}_0 \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{336a}$$

$$\mathcal{A}_1 \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{336b}$$

$$\mathcal{A}_2 \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{336c}$$

with $\delta^\star_{Q,\boldsymbol{z}}$ in (38) and the function $\mathsf{L}_{\boldsymbol{z}}$ in (3). Note that $\mathcal{A}_0 = \mathcal{L}^\star_{Q,\boldsymbol{z}}$, with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (39) and

$$1 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{337}$$

$$= P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{338}$$

$$= P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_2} \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{339}$$

where, the equality in (338) 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 (38) and the fact that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to the measure Q.

The above implies that

$$1 = \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_2} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) \tag{340}$$

$$= \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_2} \lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) \tag{341}$$

$$= \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0), \tag{342}$$

where, the equality in (341) follows from the dominated convergence theorem [55, Theorem 1.6.9], given that the Randon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is positive and finite (Lemma 3.2); and the inequality in (342) 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

$$\lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) = 1, \tag{343}$$

which completes the proof.

I Proof of Lemma 4.1

The proof is presented in two parts. The first part shows that if for all $\delta \in (\rho^\star, +\infty)$, the inequality in (50) holds, then, Q is coherent. The second part shows that if Q is not coherent, then there exists a $\delta \in (\rho^\star, +\infty)$ such that

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) = 0. \tag{344}$$

The first part is as follows. Note that for all $\delta \in (\rho^\star, +\infty)$ and for all $\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{z}}(\delta) \cap \operatorname{supp} Q$, it holds from Lemma 3.2 that

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) > 0. \tag{345}$$

Hence, if for all $\delta \in (\rho^\star, +\infty)$, the inequality in (50) holds, then

$$0 < P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) \tag{346}$$

$$= \int_{\mathcal{L}_{\boldsymbol{z}}(\delta)} \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{347}$$

$$= \int_{\mathcal{L}_{\boldsymbol{z}}(\delta)} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\theta}\right) \mathrm{d}Q\left(\boldsymbol{\theta}\right), \tag{348}$$

which, together with (345), implies that for all $\delta \in (\rho^\star, +\infty)$, $Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0$. Hence, Q is coherent.

The second part is as follows. Assume that Q is not coherent. Then, there exists a $\delta \in (\rho^\star, +\infty)$ such that $Q\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) = 0$. Hence, from the fact that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to Q, it follows that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) = 0$. This completes the proof.

J Proof of Theorem 4.1

The optimization problem in (83) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, denoted by $\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}} : \mathcal{M} \to [0, +\infty)$, which yields:

$$\min_{P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right), \tag{349a}$$

subject to:

$$\int \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\left(\boldsymbol{\nu}\right) \log\left(\frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \leqslant c, \text{ and} \tag{349b}$$

$$\int \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = 1. \tag{349c}$$

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

$$\int \left|\hat{g}(\boldsymbol{\theta})\right| \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) < \infty. \tag{350}$$

Hence, the optimization problem of interest is:

$$\min_{g\in\mathscr{M}} \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) g\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \tag{351a}$$

$$\text{s.t:} \quad \int g\left(\boldsymbol{\nu}\right) \log\left(g\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \leqslant c, \text{ and} \tag{351b}$$

$$\int g\left(\boldsymbol{\theta}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) = 1. \tag{351c}$$

The Lagrangian of the optimization problem in (351) is a functional $L : \mathscr{M} \times [0,+\infty)^2 \to \mathbb{R}$ of the form

$$\begin{aligned} L\left(g,\alpha,\beta\right) = & \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) g\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \\ & + \alpha \left(\int g\left(\boldsymbol{\nu}\right) \log\left(g\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) - c \right) \\ & + \beta \left(\int g\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) - 1 \right), \end{aligned} \tag{352}$$

where the reals α and β are both nonnegative and act as Lagrangian multipliers due to the constraints (351b) and (351c), respectively.

Let $h : \mathcal{M} \to \mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (352) at $(g,\alpha,\beta) \in \mathscr{M} \times [0,+\infty)^2$ in the direction of h, if it exists, is

$$\partial L\left(g,\alpha,\beta;h\right) \triangleq \frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma)\Big|_{\gamma=0}, \tag{353}$$

where the real function $r : \mathbb{R} \to \mathbb{R}$ is such that for all $\gamma \in (-\epsilon,\epsilon)$, with some $\epsilon > 0$, satisfies

$$\begin{aligned} r(\gamma) = & \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \left(g\left(\boldsymbol{\nu}\right) + \gamma h\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \\ & + \alpha \left(\int \left(g(\boldsymbol{\nu}) + \gamma h(\boldsymbol{\nu})\right) \log\left(g(\boldsymbol{\nu}) + \gamma h(\boldsymbol{\nu})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) - c \right) \\ & + \beta \left(\int \left(g\left(\boldsymbol{\nu}\right) + \gamma h\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) - 1 \right). \end{aligned} \tag{354}$$

The proof continues under the assumption that the functions g and h are such that the Gateaux differential in (353) exists. That is, the function r in (354) is differentiable in $(-\epsilon,\epsilon)$, with some $\epsilon > 0$. Using the same arguments as in the proof of Theorem 3.1, it follows that the derivative of the real function r in (354) is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) = & \int \mathsf{L}_{\boldsymbol{z}} h\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) + \alpha \int h\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \\ & + \alpha \int h\left(\boldsymbol{\nu}\right) \log\left(g\left(\boldsymbol{\nu}\right) + \gamma h\left(\boldsymbol{\nu}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \\ & + \beta \int h\left(\boldsymbol{\nu}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right). \end{aligned} \tag{355}$$

From (353) and (355), it follows that

$$\partial L\left(g,\alpha,\beta;h\right)=\int h\left(\boldsymbol{\nu}\right)\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\alpha\left(1+\log g\left(\boldsymbol{\nu}\right)\right)+\beta\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\boldsymbol{\nu}\right). \quad (356)$$

From [69, Theorem 1, page 217], it holds that a necessary condition for the functional L in (352) to have a minimum at $(g,\alpha,\beta)\in\mathscr{M}\times[0,+\infty)^2$ is that for all functions $h\in\mathscr{M}$,

$$\partial L\left(g,\alpha,\beta;h\right)=0, \quad (357)$$

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

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\alpha\left(1+\log g\left(\boldsymbol{\nu}\right)\right)+\beta=0. \quad (358)$$

Thus,

$$g\left(\boldsymbol{\nu}\right)=\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)}{\alpha}\right)\exp\left(-\frac{\beta+\alpha}{\alpha}\right), \quad (359)$$

where α and β are chosen to satisfy their corresponding constraints with equality. Denote by $P^{\star}$ the solution of the optimization problem in (83). Hence, from (359), it follows that

$$\frac{\mathrm{d}P^{\star}}{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}}\left(\boldsymbol{\nu}\right)=\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)}{\alpha}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\alpha}\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\boldsymbol{\theta}\right)}, \quad (360)$$

where α is chosen to satisfy

$$D\left(P^{\star}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)=c. \quad (361)$$

From Lemma 3.3, it follows that the probability measure $P^{\star}$ and the σ-finite

measure Q satisfy,

$$\frac{\mathrm{d}P^\star}{\mathrm{d}Q}(\boldsymbol{\nu}) = \frac{\mathrm{d}P^\star}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu}) \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \tag{362}$$

$$= \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\alpha}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})} \right) \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta})} \right) \tag{363}$$

$$= \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\alpha}\right)\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\alpha})} \mathrm{d}Q(\boldsymbol{\theta})} \right) \left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta})} \right) \tag{364}$$

$$= \frac{\exp\left(-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)}{\int \exp\left(-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\theta})}, \tag{365}$$

which implies that $P^\star$ is a Gibbs probability measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, with energy function $\mathsf{L}_{\boldsymbol{z}}$, reference measure Q, and regularization parameter $\frac{1}{\frac{1}{\alpha}+\frac{1}{\lambda}}$, where α is chosen to satisfy (361). 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 (352) is strictly convex, and thus, the measure $P^{(Q,\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the unique minimizer. This completes the proof.

K Proof of Lemma 5.2

Note that for all $(\lambda_1, \lambda_2) \in \{x \in \mathbb{R} : K_{Q,\boldsymbol{z}}(x) < +\infty\}^2$, such that $\lambda_1 > \lambda_2$, it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, the inequality $\exp(\lambda_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \leqslant \exp(\lambda_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))$ holds. This implies that $K_{Q,\boldsymbol{z}}(\lambda_2) \leqslant K_{Q,\boldsymbol{z}}(\lambda_1) < +\infty$, which proves that the function is nondecreasing.

The proof of continuity of the function $K_{Q,\boldsymbol{z}}$ follows from observing that for all

$\alpha \in \{x \in \mathbb{R} : K_{Q,\boldsymbol{z}}(x) < +\infty\}$, it holds that

$$\begin{aligned}
\lim_{t\to\alpha} K_{Q,\boldsymbol{z}}(t) &= \lim_{t\to\alpha} \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right) && (366)\\
&= \log\left(\lim_{t\to\alpha} \int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right) && (367)\\
&= \log\left(\int \lim_{t\to\alpha} \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right) && (368)\\
&= \log\left(\int \exp\left(\alpha\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right) && (369)\\
&= K_{Q,\boldsymbol{z}}(\alpha), && (370)
\end{aligned}$$

where (367) and (369) follow from the fact that both the logarithmic and exponential functions are continuous; and the equality in (368) follows from the monotone convergence theorem [55, Theorem 1.6.2]. This shows that the function $K_{Q,\boldsymbol{z}}$ is continuous in $\{x \in \mathbb{R} : K_{Q,\boldsymbol{z}}(x) < +\infty\}$.

The proof of differentiability follows by considering the transport of the σ-finite measure Q in (22) 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 (3). 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 $K_{Q,\boldsymbol{z}}$ satisfies for all $t \in \{\nu \in \mathbb{R} : K_{Q,\boldsymbol{z}}(\nu) < +\infty\}$,

$$\begin{aligned}
K_{Q,\boldsymbol{z}}(t) &= \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \mathrm{d}Q(\boldsymbol{\theta})\right) && (371)\\
&= \log\left(\int \exp\left(t\, w\right) \mathrm{d}P(w)\right), && (372)
\end{aligned}$$

where the equality (372) follows from [55, Theorem 1.6.12]. Denote by ϕ the Laplace transform of the measure P. That is, for all $t \in \{x \in \mathbb{R} : K_{Q,\boldsymbol{z}}(x) < +\infty\}$,

$$\phi(t) = \int \exp\left(t\, v\right) \mathrm{d}P(v). \tag{373}$$

Hence, $\phi(t) = \exp\left(K_{Q,\boldsymbol{z}}(t)\right)$. From [70, Theorem 1$a$ (page 439)], it follows that the function ϕ has derivatives of all orders in $\{x \in \mathbb{R} : K_{Q,\boldsymbol{z}}(x) < +\infty\}$, and thus, so does the function $K_{Q,\boldsymbol{z}}$ in the interior of $\{x \in \mathbb{R} : K_{Q,\boldsymbol{z}}(x) < +\infty\}$. This completes the proof.

L Proof of Lemma 5.3

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

$$
\begin{aligned}
&\alpha K_{Q,\boldsymbol{z}}(\gamma_1) + (1-\alpha) K_{Q,\boldsymbol{z}}(\gamma_2) \\
=&\alpha \log\left(\int \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\right) \\
&+(1-\alpha)\log\left(\int \exp\left(\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\right) && (374)\\
=&\log\left(\left(\int \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\right)^{\alpha}\right) \\
&+\log\left(\left(\int \exp\left(\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\right)^{(1-\alpha)}\right) && (375)\\
=&\log\Bigg(\left(\int \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\right)^{\alpha} \\
&\left(\int \exp(\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \mathrm{d}Q(\boldsymbol{\theta})\right)^{(1-\alpha)}\Bigg) && (376)\\
=&\log\Bigg(\left(\int \exp\left(\gamma_1 \alpha \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^{p} \mathrm{d}Q(\boldsymbol{\theta})\right)^{\frac{1}{p}} \\
&\left(\int \exp\left(\gamma_2 (1-\alpha) \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^{q} \mathrm{d}Q(\boldsymbol{\theta})\right)^{\frac{1}{q}}\Bigg) && (377)\\
\geqslant&\log\left(\int \exp(\gamma_1 \alpha \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \exp(\gamma_2 (1-\alpha) \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \mathrm{d}Q(\boldsymbol{\theta})\right) && (378)\\
=&\log\left(\int \exp\Big(\left(\gamma_1 \alpha + \gamma_2 (1-\alpha)\right) \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\Big) \mathrm{d}Q(\boldsymbol{\theta})\right) && (379)\\
=&K_{Q,\boldsymbol{z}}\left(\gamma_1 \alpha + \gamma_2 (1-\alpha)\right), && (380)
\end{aligned}
$$

where the inequality in (377) follows with $\alpha \triangleq \frac{1}{p}$ and $1-\alpha \triangleq \frac{1}{q}$; the inequality in (378) follows from Hölder's inequality. Hence, equality in (378) holds if and only if there exist two constants β_1 and β_2, not simultaneously equal to zero, such that the set

$$
\begin{aligned}
\mathcal{A} &\triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \beta_1 \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) = \beta_2 \exp\left(\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)\right\} \\
&= \left\{\boldsymbol{\theta} \in \mathcal{M} : \exp\left(\left(\gamma_1 - \gamma_2\right) \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) = \frac{\beta_2}{\beta_1}\right\} && (381)\\
&= \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{\log \frac{\beta_2}{\beta_1}}{\left(\gamma_1 - \gamma_2\right)}\right\}, && (382)
\end{aligned}
$$

satisfies $Q(\mathcal{A}) = 1$. That is, strict inequality in (378) 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 5.4

For all $s \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), the equality in (96) implies the following,

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})\right)\bigg|_{t=-\frac{1}{s}} \tag{383}$$

$$= \int \frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)}{\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{v}\right)\right)\mathrm{d}Q(\boldsymbol{v})}\mathrm{d}Q(\boldsymbol{\theta})\bigg|_{t=-\frac{1}{s}} \tag{384}$$

$$= \int \frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\exp\left(-\frac{1}{s}\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)}{\int \exp\left(-\frac{1}{s}\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{v}\right)\right)\mathrm{d}Q(\boldsymbol{v})}\mathrm{d}Q(\boldsymbol{\theta}) \tag{385}$$

$$= \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\exp\left(-\frac{1}{s}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{386}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) - \frac{1}{s}\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{387}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{388}$$

where the equality in (384) holds from the dominated convergence theorem [55, Theorem 1.6.9]; the equality in (386) follows from (22); and the equality in (388) follows from (25).

For all $s \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), the equalities in (96) and (387) imply that

$$K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\exp\left(-K_{Q,\boldsymbol{z}}\left(t\right) + t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})\bigg|_{t=-\frac{1}{s}} \tag{389}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\left(-K^{(1)}_{Q,\boldsymbol{z}}\left(t\right) + \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \exp\left(-K_{Q,\boldsymbol{z}}\left(t\right) + t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})\bigg|_{t=-\frac{1}{s}} \tag{390}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\left(-K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) + \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) - \frac{1}{s}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{391}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\left(-K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) + \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{392}$$

$$= -K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2\mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{393}$$

$$= -\left(K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)^2 + \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{394}$$

$$= \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)^2 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right), \tag{395}$$

where the equality in (390) follows from the dominated convergence theorem [55, Theorem 1.6.9]; the equality in (392) is due to a change of measure through the Radon-Nikodym derivative in (25); and the equality in (394) follows from (388).

For all $s \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), the equalities in (96) and (394) imply that

$$K^{(3)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \mathrm{d}P^{\left(Q,-\frac{1}{t}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - \left(K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)\right)^2\right)\Bigg|_{t=-\frac{1}{s}} \tag{396}$$

$$= \frac{\mathrm{d}}{\mathrm{d}t}\Bigg(\int \Big(\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \exp\left(-K_{Q,\boldsymbol{z}}\left(t\right) + t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\Big)\mathrm{d}Q\left(\boldsymbol{\theta}\right) - \left(K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)\right)^2\Bigg)\Bigg|_{t=-\frac{1}{s}} \tag{397}$$

$$= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \left(\frac{\mathrm{d}}{\mathrm{d}t}\exp(-K_{Q,\boldsymbol{z}}(t) + t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big|_{t=-\frac{1}{s}}\right)\mathrm{d}Q(\boldsymbol{\theta}) - 2K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)K^{(2)}_{Q,\boldsymbol{z}}\left(t\right)\Big|_{t=-\frac{1}{s}} \tag{398}$$

$$= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2 \Big(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}(t)\Big)\exp(-K_{Q,\boldsymbol{z}}(t) + t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\Big|_{t=-\frac{1}{s}}\mathrm{d}Q(\boldsymbol{\theta}) - 2K^{(1)}_{Q,\boldsymbol{z}}\left(t\right)K^{(2)}_{Q,\boldsymbol{z}}\left(t\right)\Big|_{t=-\frac{1}{s}} \tag{399}$$

$$= \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right) \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) - \frac{1}{s}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) - 2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) \tag{400}$$

$$= \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)\mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - 2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) \tag{401}$$

$$
\begin{aligned}
&= \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \\
&\quad -K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^2 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \\
&\quad -2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (402) \\
&= \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \\
&\quad -K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) \left(K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) + \left(K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)^2\right) \\
&\quad -2K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (403) \\
&= \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)^3 && (404) \\
&\quad -3K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right) && (405) \\
&= \int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{s}\right)\right)^3 \mathrm{d}P^{(Q,s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right), && (406)
\end{aligned}
$$

where the equality in (397) follows from (25); and the equality in (398) follows from the dominated convergence theorem [55, Theorem 1.6.9]; the equality in (401) follows from (25); and the equality in (403) follows from (394).

This completes the proof.

N Proof of Theorem 6.1

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 $\mathrm{int}\mathcal{K}_{Q,\boldsymbol{z}}$. This is due to Corollary 6.1. For instance, note that

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) && (407) \\
&= \frac{1}{\lambda^2}K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) && (408) \\
&\geqslant 0, && (409)
\end{aligned}
$$

where the equality in (408) follows from Lemma 5.4. The inequality in (409) 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 (105) is nondecreasing with respect to λ. 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 (96) satisfies $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > 0$ if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable. For doing so, a handful of preliminary results are described in the following subsection. The proof of Theorem 6.1 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 (23), 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

$$\mathcal{R}_0(\lambda) \triangleq \left\{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) = \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}, \tag{410a}$$

$$\mathcal{R}_1(\lambda) \triangleq \left\{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) < \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}, \text{ and} \tag{410b}$$

$$\mathcal{R}_2(\lambda) \triangleq \left\{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) > \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right\}, \tag{410c}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is in (18) and the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is in (25). The sets in (410) 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*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\lambda)\right) > 0, \tag{411}$$

if and only if

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\lambda)\right) > 0, \tag{412}$$

where the sets $\mathcal{R}_1(\cdot)$ and $\mathcal{R}_2(\cdot)$ are in (410b) *and* (410c)*, 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 proves the converse.

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 5.4, 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) && (413)\\
&+ \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) && (414)\\
=& \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) && (415)\\
=& 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) && (416)\\
&+ \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) && (417)\\
<& 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) && (418)\\
=& 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) && (419)\\
=& K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right), && (420)
\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 5.4, 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) && (421)\\
&+ \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) && (422)\\
=& \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}_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) && (423)\\
=& 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}_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) && (424)\\
>& 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}_2(\alpha)\right) && (425)\\
=& 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}_2(\alpha)\right)\right) && (426)\\
=& K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right), && (427)
\end{aligned}
$$

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

A more general result can be immediately obtained by combining Lemma 3.4 and Lemma N.1.

Lemma N.2. *For all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), satisfies*

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right) > 0, \tag{428}$$

if and only if

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right) > 0, \tag{429}$$

where the sets $\mathcal{R}_1(\alpha)$ and $\mathcal{R}_2(\alpha)$ are in (410b) and (410c), respectively.

N.2 The proof

The rest of the proof of Theorem 6.1 is divided into two parts. In the first part, it is shown that if for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$, $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > 0$, then the function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable. The second part of the proof, consists in showing that if the function $\mathsf{L}_{\boldsymbol{z}}$ is separable, then, for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$, $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > 0$.

The first part is as follows. From Lemma 5.4, it holds that for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$,

$$\begin{aligned}
K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) &= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (430)\\
&= \int_{\mathcal{R}_0(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (431)\\
&\quad + \int_{\mathcal{R}_1(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (432)\\
&\quad + \int_{\mathcal{R}_2(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), && (433)
\end{aligned}$$

where the sets $\mathcal{R}_0(\alpha)$, $\mathcal{R}_1(\alpha)$, and $\mathcal{R}_2(\alpha)$ are respectively defined in (410). Hence,

$$\begin{aligned}
K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) &= \int_{\mathcal{R}_1(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (434)\\
&\quad + \int_{\mathcal{R}_2(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). && (435)
\end{aligned}$$

Under the assumption that for all $\alpha \in \mathcal{K}_{Q,\boldsymbol{z}}$ the function $K^{(2)}_{Q,\boldsymbol{z}}$ in (96) satisfies $K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > 0$, it follows 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 $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{R}_1(\alpha) \times \mathcal{R}_2(\alpha)$,

$$+\infty > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2), \tag{436}$$

where $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) < +\infty$ follows from the fact that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = +\infty\}) = 0$ (Lemma 3.2). 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 (3) is separable with respect to $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. From Lemma 5.1, 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 (3) 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 $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{A}(\gamma) \times \mathcal{B}(\gamma)$,

$$+\infty > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > c_\gamma > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2). \tag{437}$$

From Lemma 5.4, it holds that

$$\begin{aligned}
K^{(2)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right) &= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (438)\\
&= \int_{\mathcal{A}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (439)\\
&+ \int_{\mathcal{B}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (440)\\
&+ \int_{\mathcal{M}\setminus(\mathcal{A}(\gamma)\cup\mathcal{B}(\gamma))} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (441)\\
&> 0, && (442)
\end{aligned}$$

where the inequality (442) follows from the following facts. First, if $c_\gamma < K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)$, with c_γ in (437), then for all $\boldsymbol{\nu} \in \mathcal{B}(\gamma)$, it holds that $K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right) > c_\gamma > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})$, and thus,

$$\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 > \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2, \tag{443}$$

which implies,

$$\begin{aligned}
\int_{\mathcal{B}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) &> \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}(\gamma)) && (444)\\
&> 0. && (445)
\end{aligned}$$

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 $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > c_\gamma \geqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)$, and thus,

$$\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 > \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2, \tag{446}$$

which implies,

$$\int_{\mathcal{A}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) > \left(c_\gamma - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)\right)^2 P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}(\gamma)) \tag{447}$$

$$> 0. \tag{448}$$

Hence, under the assumption that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) 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 6.1

Consider the partition of the set $\mathcal{M}$ formed by the sets $\mathcal{A}_0$, $\mathcal{A}_1$, and $\mathcal{A}_2$ in (336). From (97), for all $\lambda \in \mathcal{K}_{Q,\boldsymbol{z}}$, with $\mathcal{K}_{Q,\boldsymbol{z}}$ in (23), it holds that,

$$K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int_{\mathcal{A}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{449}$$

$$+ \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{450}$$

$$= \int_{\mathcal{A}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{451}$$

$$= \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{452}$$

$$\geqslant \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) + \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{453}$$

$$= \delta^\star_{Q,\boldsymbol{z}}, \tag{454}$$

where the equality in (451) follows by noticing that $Q(\mathcal{A}_1) = 0$, which implies that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) = 0$ (Lemma 3.3); the equality in (452) follows from noticing that $\mathcal{A}_0 = \mathcal{L}^\star_{Q,\boldsymbol{z}}$, with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (39); and the equality in (453) follows from (336c). This completes the proof.

P Proof of Theorem 6.2

From (452) in the proof of Lemma 6.1, it holds that

$$\lim_{\lambda\to 0^+} K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \lim_{\lambda\to 0^+} \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{455}$$

$$= \lim_{\lambda\to 0^+} \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) \tag{456}$$

$$= \lim_{\lambda\to 0^+} \delta^\star_{Q,\boldsymbol{z}} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) \tag{457}$$

$$= \delta^\star_{Q,\boldsymbol{z}} \lim_{\lambda\to 0^+} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{Q,\boldsymbol{z}}) \tag{458}$$

$$= \delta^\star_{Q,\boldsymbol{z}}, \tag{459}$$

where, the equality in (457) 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 3.2); and (*b*) For all $\boldsymbol{\theta} \in \mathcal{A}_2$, it holds that $\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0$ (Lemma 3.6). Hence, the dominated convergence theorem [55, Theorem 1.6.9] holds. The inequality in (458) follows from Lemma 3.7. This completes the proof.

Q Proof of Theorem 9.1

From Theorem 6.1, it follows that for all $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

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

which implies the following inclusions:

$$\mathcal{R}_1(\lambda_2) \subseteq \mathcal{R}_1(\lambda_1), \text{ and} \tag{460a}$$

$$\mathcal{R}_2(\lambda_1) \subseteq \mathcal{R}_2(\lambda_2), \tag{460b}$$

with the sets $\mathcal{R}_1(\cdot)$ and $\mathcal{R}_2(\cdot)$ in (410). From (193), it holds that for all $i \in \{1, 2\}$,

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_i) = \mathcal{R}_2(\lambda_i)^{\mathsf{c}}, \tag{461}$$

where the complement is with respect to $\mathcal{M}$. Thus, the inclusion in (460b) and the equality in (461) yields,

$$\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1) \supseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2). \tag{462}$$

The inclusion $\mathcal{M} \supseteq \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)$ follows from (193). Alternatively, the inclusion $\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2) \supseteq \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$, follows from Lemma 6.1 and from observing that for all $\boldsymbol{\nu} \in \mathcal{N}^{\star}_{Q,\boldsymbol{z}}$,

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geqslant \delta^{\star}_{Q,\boldsymbol{z}} = \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right), \tag{463}$$

which implies that $\boldsymbol{\nu} \in \mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)$. This completes the proof of (197).

The proof of (198) is as follows. From the intermediate value theorem [66, Theorem 4.23] and the assumption that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) 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

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) = \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\alpha}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\alpha}\right), \tag{464}$$

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 (462), verifies that

$$\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right) \supset \mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right). \tag{465}$$

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,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0\left(\lambda\right)\right) = 1. \tag{466}$$

This contradicts the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable (Definition 5.1). 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,

$$\left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \leqslant \delta^{\star}_{Q,\boldsymbol{z}}\right\} = \mathcal{N}^{\star}_{Q,\boldsymbol{z}} \tag{467}$$

$$= \mathcal{N}_{Q,\boldsymbol{z}}(\lambda) \tag{468}$$

$$= \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \leqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right\}. \tag{469}$$

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,

$$\mathcal{R}_1\left(\lambda\right) = \left\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \tag{470}$$

and from the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (38), it holds that

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1\left(\lambda\right)\right) = 0. \tag{471}$$

From Lemma N.1 and (471), it follows that,

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2\left(\lambda\right)\right) = 0. \tag{472}$$

Finally, by noticing that

$$1 = P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0\left(\lambda\right)\right) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1\left(\lambda\right)\right) + P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2\left(\lambda\right)\right) \tag{473}$$

$$= P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0\left(\lambda\right)\right), \tag{474}$$

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 5.1). This completes the proof of (198).

R Proof of Theorem 9.2

The proof of (199) 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 $\mathcal{A} \subseteq \mathscr{B}\left(\mathcal{M}\right)$. More specifically, given a $\gamma \in \mathcal{K}_{Q,\boldsymbol{z}}$, it holds that

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) = \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\alpha}\right) \mathrm{d}Q\left(\boldsymbol{\alpha}\right), \tag{475}$$

and from the fundamental theorem of calculus [66, Theorem 6.21], it follows that for all $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) - P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) = \int_{\lambda_2}^{\lambda_1} \frac{\mathrm{d}}{\mathrm{d}\gamma} P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{A}\right) \mathrm{d}\gamma \tag{476}$$

$$= \int_{\lambda_2}^{\lambda_1} \frac{\mathrm{d}}{\mathrm{d}\gamma} \int_{\mathcal{A}} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\alpha}\right) \mathrm{d}Q\left(\boldsymbol{\alpha}\right) \mathrm{d}\gamma \tag{477}$$

$$= \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{A}} \frac{\mathrm{d}}{\mathrm{d}\gamma} \frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}\left(\boldsymbol{\alpha}\right) \mathrm{d}Q\left(\boldsymbol{\alpha}\right) \mathrm{d}\gamma, \tag{478}$$

where the equality in (477) follows from (475); and the equality in (478) holds from Lemma 3.2 and the dominated convergence theorem [55, Theorem 1.6.9].

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

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})} \tag{479}$$

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

$$- \frac{\frac{1}{\lambda^2}\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\alpha})}{\left(\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})\right)^2} \tag{480}$$

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

$$- \frac{1}{\lambda^2}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\mathrm{d}Q(\boldsymbol{\nu}) \tag{481}$$

$$= \frac{1}{\lambda^2}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right). \tag{482}$$

Plugging (482) into (478) yields,

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

$$= \int_{\lambda_2}^{\lambda_1}\int_{\mathcal{A}}\frac{1}{\gamma^2}\frac{\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\alpha})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\alpha})\,\mathrm{d}\gamma \tag{483}$$

$$= \int_{\lambda_2}^{\lambda_1}\int_{\mathcal{A}}\frac{1}{\gamma^2}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\mathrm{d}\gamma. \tag{484}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant 0, \tag{485}$$

and thus,

$$\int_{\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)}\frac{1}{\gamma^2}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \leqslant 0. \tag{486}$$

The equalities in (484) and (486), with $\mathcal{A} = \mathcal{N}_{Q,\boldsymbol{z}}(\lambda)$, imply that

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) - P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) \leqslant 0. \tag{487}$$

The inequality $0 < P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2))$ in (199) 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 (199).

The proof of strict inequality in (199) 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 $\lambda_1 > \lambda_2$,

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{488}$$

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 (488) holds.

The first part is as follows. In the proof of Theorem 9.1 it is shown (see (484)) that for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{Q,\boldsymbol{z}} \times \mathcal{K}_{Q,\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$\begin{aligned} &P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\right) - P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\right) \\ &= \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)} \frac{1}{\gamma^2}\Big(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - K^{(1)}_{Q,\boldsymbol{z}}\Big(-\frac{1}{\gamma}\Big)\Big) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\mathrm{d}\gamma. \end{aligned} \tag{489}$$

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 (488) holds. Then, from (489),

$$0 > \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)} \frac{1}{\gamma^2}\Big(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - K^{(1)}_{Q,\boldsymbol{z}}\Big(-\frac{1}{\gamma}\Big)\Big) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\mathrm{d}\gamma \tag{490}$$

$$= \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{R}_1(\lambda_2)} \frac{1}{\gamma^2}\Big(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - K^{(1)}_{Q,\boldsymbol{z}}\Big(-\frac{1}{\gamma}\Big)\Big) \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\mathrm{d}\gamma, \tag{491}$$

where the equality in (491) 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 (410a), (410b), and (193), respectively.

The inequality in (491) 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 arguments above, 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

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{492}$$

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 $\mathcal{R}_2(\lambda_2)$ satisfy for all $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{R}_2(\lambda) \times \mathcal{R}_1(\lambda)$,

$$+\infty > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2), \tag{493}$$

which together with Definition 5.1 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 5.1). 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 (3) is separable with respect to Q (and thus, with respect to $P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ by Lemma 5.1). That is, from Definition 5.1, 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 $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{A}(\gamma) \times \mathcal{B}(\gamma)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > c_\gamma > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2). \tag{494}$$

In the proof of Theorem 9.1, cf. (484), 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 $\mathcal{M}$,

$$\begin{aligned}
&P^{(Q,\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) - P^{(Q,\alpha_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) \\
&= \int_{\alpha_2}^{\alpha_1}\int_{\mathcal{A}} \frac{1}{\lambda^2} \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\alpha}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) \mathrm{d}P(\boldsymbol{\alpha})\mathrm{d}\lambda \\
&= \int_{\alpha_2}^{\alpha_1}\int_{\mathcal{A}} \frac{1}{\lambda^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\mathrm{d}\lambda.
\end{aligned} \tag{495}$$

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

$$c_\gamma < K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right), \tag{496}$$

with c_γ in (494). The second case considers that

$$c_\gamma \geqslant K^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\gamma}\right). \tag{497}$$

In the first case, it follows from (193) that

$$\mathcal{B}(\gamma) \subset \mathcal{N}_{Q,\boldsymbol{z}}(\gamma), \tag{498}$$

which implies that

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) \geqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}(\gamma)) \tag{499}$$

$$> 0, \tag{500}$$

where, the inequality in (500) 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 (193) and (498), it follows that for all $\boldsymbol{\alpha} \in \mathcal{N}_{Q,\boldsymbol{z}}(\gamma)$ and for all $\lambda \in (\gamma, \alpha_1)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - c_\gamma \tag{501}$$

$$< 0, \tag{502}$$

where the inequality in (501) follows from (496); and the inequality in (502) follows from (494). Thus,

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

which implies, from (495), that

$$P^{(Q,\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) - P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) < 0. \tag{503}$$

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

$$A(\gamma) \subseteq \mathcal{R}_2(\gamma), \tag{504}$$

which implies that

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\gamma)) \geqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}(\gamma)) \tag{505}$$

$$> 0, \tag{506}$$

where the inequality in (506) 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,

$$P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) \geqslant P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\gamma)) \tag{507}$$

$$> 0, \tag{508}$$

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 (193) and Theorem 6.1, it follows that for all $\boldsymbol{\alpha} \in \mathcal{N}_{Q,\boldsymbol{z}}(\gamma)$ and for all $\lambda \in (\gamma, \alpha_1)$,

$$0 \geqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \tag{509}$$

$$> \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}). \tag{510}$$

Thus,

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

which implies, from (495), that

$$P^{(Q,\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) - P^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\gamma)) < 0. \tag{511}$$

This completes the proof.

S Proof of Lemma 9.2

The proof is based on the following two observations. First, note that $\left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\right)^{\mathsf{c}} = \mathcal{R}_2\left(\lambda_2\right)$, with the set $\mathcal{R}_2\left(\cdot\right)$ defined in (410c). Second, note that

$$\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right)=\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right) \cup \left(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right) \cap \mathcal{R}_2\left(\lambda_2\right)\right), \tag{512}$$

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

$$\begin{aligned} P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1))&=P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)\cup(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)\cap\mathcal{R}_2(\lambda_2))\Big) && (513)\\ &=P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\Big) \\ &\quad +P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_1\right)\cap\mathcal{R}_2\left(\lambda_2\right)\Big) && (514)\\ &=P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}_{Q,\boldsymbol{z}}\left(\lambda_2\right)\Big), && (515)\end{aligned}$$

where the equality in (514) follows from Lemma 3.3 and the equality in (200).

Finally, under the assumption that the empirical function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable, it holds from Theorem 9.2 that

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)). \tag{516}$$

Plugging (515) into (516), with $i=1$, yields,

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_1)) < P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}_{Q,\boldsymbol{z}}(\lambda_2)), \tag{517}$$

and this completes the proof.

T Proof of Theorem 11.2

Consider the following lemma.

Lemma T.1. *Given two probability measures P_1 and P_2 over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, with P_2 absolutely continuous with respect to P_1, the following holds for all $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$,*

$$\mathsf{R}_{\boldsymbol{z}}\left(P_2\right) - \mathsf{R}_{\boldsymbol{z}}\left(P_1\right) \leqslant \inf_{t<0}\left(\frac{D(P_2\|P_1)+\log\big(\int\exp(t\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})-\mathsf{R}_{\boldsymbol{z}}(P_1)\right))\mathrm{d}P_1(\boldsymbol{\theta})\big)}{t}\right), \tag{518}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ and the functional $\mathsf{R}_{\boldsymbol{z}}$ are defined in (3) *and in* (18)*, respectively.*

Proof: From [53, Corollary 4.15, Page 100], it follows that the probability measures P_1 and P_2 in $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ satisfy the following equality:

$$D(P_2\|P_1)=\sup_f \int f(\boldsymbol{\theta})\mathrm{d}P_2(\boldsymbol{\theta})-\log\int \exp(f(\boldsymbol{\theta}))\mathrm{d}P_1(\boldsymbol{\theta}), \tag{519}$$

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 $\int \exp(f(\boldsymbol{\theta}))\,\mathrm{d}P_1(\boldsymbol{\theta}) < \infty$. Hence, for all $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$ and for all $t \in (-\infty, 0)$, it follows that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) satisfies that

$$\begin{aligned}
D(P_2\|P_1) &\geqslant \int t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P_2(\boldsymbol{\theta}) - \log\int \exp(t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}P_1(\boldsymbol{\theta}) && (520)\\
&\geqslant \int t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P_2(\boldsymbol{\theta}) \\
&\quad - \log\int \exp(t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + t\mathsf{R}_{\boldsymbol{z}}(P_1) - t\mathsf{R}_{\boldsymbol{z}}(P_1))\,\mathrm{d}P_1(\boldsymbol{\theta}) && (521)\\
&= \int t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P_2(\boldsymbol{\theta}) - t\mathsf{R}_{\boldsymbol{z}}(P_1) \\
&\quad - \log\int \exp(t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - t\mathsf{R}_{\boldsymbol{z}}(P_1))\,\mathrm{d}P_1(\boldsymbol{\theta}) && (522)\\
&= t\mathsf{R}_{\boldsymbol{z}}(P_2) - t\mathsf{R}_{\boldsymbol{z}}(P_1) - \log\int \exp(t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - t\mathsf{R}_{\boldsymbol{z}}(P_1))\mathrm{d}P_1(\boldsymbol{\theta}), && (523)
\end{aligned}$$

which leads to

$$\mathsf{R}_{\boldsymbol{z}}(P_2) - \mathsf{R}_{\boldsymbol{z}}(P_1) \leqslant \frac{D(P_2\|P_1) + \log\int \exp(t(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{R}_{\boldsymbol{z}}(P_1)))\mathrm{d}P_1(\boldsymbol{\theta})}{t}. \tag{524}$$

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

$$\mathsf{R}_{\boldsymbol{z}}(P_2) - \mathsf{R}_{\boldsymbol{z}}(P_1) \leqslant \inf_{t\in(-\infty,0)} \frac{D(P_2\|P_1) + \log\int \exp(t(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{R}_{\boldsymbol{z}}(P_1)))\mathrm{d}P_1(\boldsymbol{\theta})}{t}, \tag{525}$$

which completes the proof. ∎

From Lemma T.1, it holds that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (25), satisfies for all $P \in \triangle_Q(\mathcal{M}, \mathscr{B}(\mathcal{M}))$,

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\leqslant \inf_{t\in(-\infty,0)} \left(\frac{D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}{t} + \frac{\log\left(\int \exp\left(t\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{Q,\boldsymbol{z}}\left(\frac{1}{\lambda}\right)\right)\right)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right)}{t} \right), && (526)
\end{aligned}$$

where the function $K^{(1)}_{Q,\boldsymbol{z}}$ is defined in (97) and satisfies (105). Moreover, for

all $t \in (-\infty, 0)$,

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

$$= \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right) - tK^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \tag{527}$$

$$= J_{\boldsymbol{z},Q,\lambda}(t) - tK^{(1)}_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \tag{528}$$

$$\leqslant \frac{1}{2}t^2\beta^2_{Q,\boldsymbol{z}}, \tag{529}$$

where the equality in (528) follows from (142); the inequality in (529) follows from Theorem 8.1; and the constant $\beta_{Q,\boldsymbol{z}}$ is defined in (165).

Plugging (529) into (526) yields for all $t \in (-\infty, 0)$,

$$\mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant \inf_{t\in(-\infty,0)} \frac{D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \frac{1}{2}t^2\beta^2_{Q,\boldsymbol{z}}}{t}. \tag{530}$$

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

$$c \triangleq \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{531}$$

Hence, from (530), it follows that for all $t \in (-\infty, 0)$,

$$c\,t - \frac{1}{2}t^2\beta^2_{Q,\boldsymbol{z}} \leqslant D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{532}$$

The rest of the proof consists in finding an explicit expression for the absolute value of c in (532). To this aim, consider the function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$\phi(\alpha) = \frac{1}{2}\alpha^2\beta^2_{Q,\boldsymbol{z}}, \tag{533}$$

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 $x \in \mathbb{R}$,

$$\phi^*(x) = \max_{t\in(-\infty,0)} xt - \phi(t). \tag{534}$$

In particular, note that

$$\phi^*(c) \leqslant D\left(P\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{535}$$

Note that for all $x \in \mathbb{R}$ and for all $t \in (-\infty, 0)$, the function $\phi^\star$ in (534) satisfies

$$x\,t - \frac{1}{2}t^2\beta^2_{Q,\boldsymbol{z}} \leqslant \phi^\star(x) = x\alpha^\star(x) - \phi\left(\alpha^\star(x)\right), \tag{536}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(x\alpha - \phi\left(\alpha\right)\right) = x - \alpha\beta^2_{Q,\boldsymbol{z}} = 0. \tag{537}$$

That is,

$$\alpha^{\star}(x)=\frac{x}{\beta_{Q,\boldsymbol{z}}^2}. \tag{538}$$

Plugging (538) into (536) yields,

$$\phi^{\star}(x)=\frac{x^2}{2\beta_{Q,\boldsymbol{z}}^2}. \tag{539}$$

Hence, from (535) and (536), given c in (531) for all $t \in (-\infty, 0)$,

$$c\,t-\frac{1}{2}t^2\beta_{Q,\boldsymbol{z}}^2 \leqslant \phi^{\star}(c) \leqslant D\left(P\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right), \tag{540}$$

and thus,

$$\frac{c^2}{2\beta_{Q,\boldsymbol{z}}^2} \leqslant D\left(P\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right). \tag{541}$$

This implies that

$$c \leqslant \sqrt{2\beta_{Q,\boldsymbol{z}}^2 D\left(P\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)} \tag{542}$$

and

$$c \geqslant -\sqrt{2\beta_{Q,\boldsymbol{z}}^2 D\left(P\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)}, \tag{543}$$

which leads to

$$\left|\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}(\boldsymbol{\theta})\right| \leqslant \sqrt{2\beta_{Q,\boldsymbol{z}}^2 D\left(P\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\right)}, \tag{544}$$

and completes the proof.

U Proof of Theorem 11.4

Under the condition that $\lambda \in \mathcal{K}_{Q,P_{\boldsymbol{Z}}}$, from Theorem 11.1 and Definition 11.2, it follows that the generalization error $\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)$ in (232) satisfies

$$\begin{aligned}
&\mathsf{G}_{Q,\lambda}\left(P_{\boldsymbol{Z}}\right)\\
&=\lambda\int\left(D\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}\|Q\right)+D\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}\right)-D\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\|Q\right)\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), &(545)\\
&=\lambda\left(\int D\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}\|P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})+\int D\left(P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})\right), &(546)
\end{aligned}$$

where the equality in (546) follows from the fact that

$$
\begin{aligned}
&\int \Big(D\big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| Q\big) - D\big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| Q\big) \Big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \\
&= \int D\big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| Q\big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) - D\big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| Q\big) && (547) \\
&= \int \left(\int \log \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}(\boldsymbol{\theta}) \right) P_{\boldsymbol{Z}}(\boldsymbol{\nu}) - D\big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| Q\big) && (548) \\
&= \int \left(\int \log \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \\
&\quad - \int \log \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) && (549) \\
&= \int \left(\int \log \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \\
&\quad - \int \left(\int \log \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) && (550) \\
&= \int \left(\int \left(\log \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \right. \right. \\
&\quad \left. \left. + \log \left(\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}(\boldsymbol{\theta}) \right) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) && (551) \\
&= \int \left(\int \log \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) && (552) \\
&= \int D\big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}). && (553)
\end{aligned}
$$

The equality in (550) follows from (230); and the equality in (551) follows from the fact that the measures Q and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}$, with $\boldsymbol{\nu} \in \operatorname{supp} P_{\boldsymbol{Z}}$, are mutually absolutely continuous. This completes the proof.

References

[1] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization: Optimality and sensitivity," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, Jul. 2022, pp. 684–689.

[2] 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.

[3] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, Aug. 2016.

[4] P. Alquier, J. Ridgway, and N. Chopin, "On the properties of variational approximations of Gibbs posteriors," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, Dec. 2016.

[5] H. P. Young, *Strategic Learning and its Limits*, 1st ed. Oxford, UK: Oxford University Press, 2004.

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

[7] G. Aminian, L. Toni, and M. R. Rodrigues, "Information-theoretic bounds on the moments of the generalization error of learning algorithms," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Melbourne, Australia, Jul. 2021, pp. 682–687.

[8] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8106–8118, Dec. 2021.

[9] W. Jiang and M. A. Tanner, "Gibbs posterior for variable selection in high-dimensional classification and data mining," *The Annals of Statistics*, vol. 36, no. 5, pp. 2207–2231, Oct. 2008.

[10] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "On the validation of Gibbs algorithms: Training datasets, test datasets and their aggregation," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.

[11] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, "The worst-case data-generating probability measure," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9515, Aug. 2023.

[12] Y. Bu, G. Aminian, L. Toni, G. W. Wornell, and M. Rodrigues, "Characterizing and understanding the generalization error of transfer learning with

Gibbs algorithm," in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Virtual Conference, Mar. 2022, pp. 8673–8699.

[13] H. He, G. Aminian, Y. Bu, M. Rodrigues, and V. Y. Tan, "How does pseudo-labeling affect the generalization error of the semi-supervised Gibbs algorithm?" in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Valencia, Spain, Apr. 2023, pp. 8494–8520.

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

[15] G. Aminian, L. Toni, and M. R. Rodrigues, "Jensen-Shannon information based characterization of the generalization error of learning algorithms," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Kanazawa, Japan, Oct. 2021.

[16] S. Masiha, A. Gohari, and M. H. Yassaee, "f-divergences and their applications in lossy compression and bounding generalization error," *IEEE Transactions on Information Theory*, to appear, arXiv preprint arXiv:2206.11042.

[17] P. Alquier, "Non-exponentially weighted aggregation: regret bounds for unbounded loss functions," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, Jul. 2021, pp. 207–218.

[18] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Empirical risk minimization with f-divergence regularization in statistical learning," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9521, Oct. 2023.

[19] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*, 2nd ed. New York, NY, USA: Springer, 2004.

[20] 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.

[21] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*, 1st ed. New York, NY, USA: Wiley, 1989.

[22] V. De Bortoli and A. Desolneux, "On quantitative Laplace-type convergence results for some exponential probability measures, with two applications," *arXiv preprint arXiv:2110.12922*, 2021.

[23] K. B. Athreya and C.-R. Hwang, "Gibbs measures asymptotics," *Sankhyā: The Indian Journal of Statistics, Series A*, vol. 72, no. 1, pp. 191–207, 2010.

[24] C.-R. Hwang, "Laplace's method revisited: Weak convergence of probability measures," *The Annals of Probability*, vol. 8, no. 6, pp. 1177–1182, 1980.

[25] M. Hasenpflug, D. Rudolf, and B. Sprungk, "Wasserstein convergence rates of increasingly concentrating probability measures," *arXiv preprint arXiv:2207.08551*, 2022.

[26] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Cambridge, UK, Sep. 2016, pp. 26–30.

[27] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, Jan. 2019.

[28] 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, Jun. 2020.

[29] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–10, Dec. 2017.

[30] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proceedings of the 10th Annual Conference on Computational Learning Theory (COLT)*, Nashville, TN, USA, Jul. 1997, pp. 2–9.

[31] D. A. McAllester, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, Apr. 2003.

[32] M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor, "PAC-Bayes unleashed: Generalisation bounds with unbounded losses," *Entropy*, vol. 23, no. 10, pp. 1–20, Oct. 2021.

[33] B. Guedj and L. Pujol, "Still no free lunches: The price to pay for tighter PAC-Bayes bounds," *Entropy*, vol. 23, no. 11, Nov. 2021.

[34] S. Mazuelas, Y. Shen, and A. Pérez, "Generalized maximum entropy for supervised classification," *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2530–2550, 2022.

[35] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," *Neural Information Processing Systems*, vol. 12, pp. 470–476, Dec. 1999.

[36] J. Zhu and E. P. Xing, "Maximum entropy discrimination Markov networks," *Journal of Machine Learning Research*, vol. 10, no. 11, pp. 2531–2569, Nov. 2009.

[37] Y. Le Cun, 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, USA: The MIT Press, 2007, ch. 10, pp. 191–241.

[38] L. P. Barnes, A. Dytso, and H. V. Poor, "Improved information-theoretic generalization bounds for distributed, federated, and iterative learning," *Entropy*, vol. 24, no. 9, 2022.

[39] T. Zhang, "From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, vol. 34, no. 5, pp. 2180–2210, Oct. 2006.

[40] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 1475–1479.

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

[42] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 582–586.

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

[44] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," *Advances in Neural Information Processing Systems*, pp. 7245–7254, Dec. 2018.

[45] A. T. Lopez and V. Jog, "Generalization error bounds using Wasserstein distances," in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Guangzhou, China, Nov. 2018, pp. 1–5.

[46] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, "Conditioning and processing: Techniques to improve information-theoretic generalization bounds," *Advances in Neural Information Processing Systems*, pp. 16 457–16 467, Dec. 2020.

[47] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms," *Advances in Neural Information Processing Systems*, pp. 9925–9935, Dec. 2018.

[48] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via Wasserstein distance," *Advances*

in Neural Information Processing Systems, vol. 34, pp. 19 109–19 121, Dec. 2021.

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

[50] G. Aminian, Y. Bu, L. Toni, M. R. D. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," *IEEE Transactions on Information Theory*, to appear, arXiv preprint arXiv:2210.09864.

[51] G. Aminian, Y. Bu, G. W. Wornell, and M. R. Rodrigues, "Tighter expected generalization error bounds via convexity of information measures," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aalto, Finland, Jun. 2022, pp. 2481–2486.

[52] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

[53] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Non-asymptotic Theory of Independence*, 1st ed. Oxford, UK: Oxford University Press, 2013.

[54] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

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

[56] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Analysis of the relative entropy asymmetry in the regularization of empirical risk minimization," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.

[57] H.-O. Georgii, *Gibbs measures and phase transitions*, 2nd ed. New York, NY, USA: De Gruyter, 2011.

[58] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.

[59] E. T. Jaynes, "Information theory and statistical mechanics I," *Physical Review Journals*, vol. 106, no. 4, pp. 620–630, May 1957.

[60] ——, "Information theory and statistical mechanics II," *Physical Review Journals*, vol. 108, no. 2, pp. 171–190, Oct. 1957.

[61] 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, Nov. 1984.

[62] J. W. Gibbs, *Elementary principles in statistical mechanics*, 1st ed. New Haven, NJ, USA: Yale University Press, 1902.

[63] 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.

[64] B. Guedj, "A primer on PAC-Bayesian learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, Long Beach, CA, USA, May 2019, pp. 2931—2940.

[65] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, Sep. 1946.

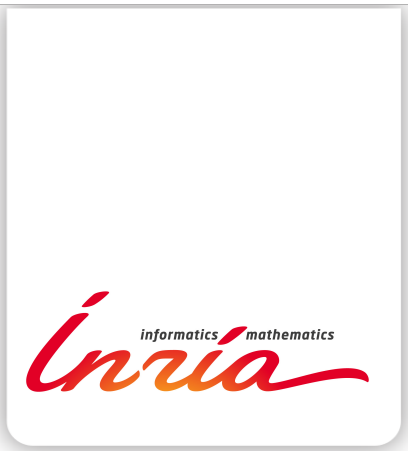
[66] W. Rudin, *Principles of mathematical analysis*, 1st ed. New York, NY, USA: McGraw-Hill Book Company, Inc., 1953.

[67] W. F. Trench, *Introduction to Real Analysis*, 1st ed. Hoboken, NJ, USA: Prentice Hall Pearson Education, 2003.

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

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

[70] W. Feller, *An Introduction to Probability Theory and Its Applications II*, 2nd ed. New York, NY, USA: Jhon Wiley & Sons, 1971.



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