Asymmetry of the Relative Entropy in the Regularization of Empirical Risk Minimization

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

***Abstract*—The effect of relative entropy asymmetry is analyzed in the context of empirical risk minimization (ERM) with relative entropy regularization (ERM-RER). Two regularizations are considered: (a) the relative entropy of the measure to be optimized with respect to a reference measure (Type-I ERM-RER); and (b) the relative entropy of the reference measure with respect to the measure to be optimized (Type-II ERM-RER). The main result is the characterization of the solution to the Type-II ERM-RER problem and its key properties. By comparing the well-understood Type-I ERM-RER with Type-II ERM-RER, the effects of entropy asymmetry are highlighted. The analysis shows that in both cases, regularization by relative entropy forces the support of the solution to collapse into the support of the reference measure, introducing a strong inductive bias that negates the evidence provided by the training data. Finally, it is shown that Type-II regularization is equivalent to Type-I regularization with an appropriate transformation of the empirical risk function.**

***Index Terms*—Empirical risk minimization; relative entropy regularization; reference measure; inductive bias**

I. Introduction

EMPIRICAL risk minimization (ERM) is a central tool in supervised machine learning. Among other uses, it enables the characterization of sample complexity and probably approximately correct (PAC) learning in a wide range of settings [2]. The application of ERM in the study of theoretical guarantees spans related disciplines such as machine learning [3], information theory [4], [5] and statistics [6], [7]. Classical problems such as classification [8], [9], pattern recognition [10], [11], regression [12], [13], and density estimation [10], [14] can be posed as special cases of the ERM problem [14], [15]. Unfortunately, ERM is prone to training data memorization, a phenomenon also known as overfitting [16]–[18]. For that reason, ERM is often regularized in order to provide generalization guarantees [19]–[22]. Regularization establishes a preference over the models by encoding features of interest that conform to prior knowledge. In different statistical learning frameworks, such as Bayesian learning [23], [24] and PAC learning [25]–[27], the prior knowledge over the set of models can be described by a reference probability measure. More general references can be adopted as proved in [28], [29] for the case of σ-finite measures. In either case, the solution to the regularized ERM problem can be cast as a probability distribution over the set of models. Prior knowledge of the set of datasets can also be represented by probability measures, e.g., the worst-case data-generating probability measure introduced in [30].

A. Motivation

A common regularizer of the ERM problem is the relative entropy of the optimization probability measure with respect to a given reference measure over the set of models [14], [31]–[33]. The resulting problem formulation, termed ERM with relative entropy regularization (ERM-RER) has been extensively studied for both the case in which the reference measure is a probability measure [31]–[34] and the case in which it is a σ-finite measure [28], [29], [35]. While in both cases the solution is unique and corresponds to a Gibbs probability measure, the existence of the solution is guaranteed only in the case in which the reference measure is a probability measure [29]. Despite the many merits of the ERM-RER formulation, it has some significant limitations. Firstly, the absolute continuity of the optimization measure with respect to the reference measure is required for the existence of the corresponding Radon-Nikodym derivative, which is used by the relative entropy regularization. This absolute continuity sets an insurmountable barrier to the exploration of models outside the support of the reference measure. More specifically, models outside the support of the reference measure exhibit zero probability with respect to the Gibbs probability measure solution to ERM-RER, regardless of the evidence provided by the training dataset. Furthermore, selecting priors with full support often leads to computationally expensive partition functions [29], [31], [32] in high-dimensional spaces. While priors with large supports ensure the inclusion of high-performing models, they also assign non-zero probability to models with poor empirical risk. Secondly, the choice of relative entropy over alternative divergences often follows arguments based on the simplicity

This work was presented in part at the International Symposium on Information Theory (ISIT) 2023 in [1]. This work is supported in part by a University of Sheffield ACSE PGR scholarship; the European Commission through the H2020-MSCA-RISE-2019 project 872172; the French National Agency for Research (ANR) through the Project ANR-21-CE25-0013 and the project ANR-22-PEFT-0010 of the France 2030 program PEPR Réseaux du Futur; the Agence de l'innovation de défense (AID) through the project UK-FR 2024352; and in part by the U.S National Science Foundation under Grant ECCS-2335876

F. Daunas is with the School of Electrical and Electronic Engineering, University of Sheffield, Sheffield S1 3JD, U.K.; and also with INRIA, Centre Inria d'Université Côte d'Azur, 06902 Sophia Antipolis, France (e-mail: jdaunastorres1@sheffield.ac.uk).

I. Esnaola is with the School of Electrical and Electronic Engineering, University of Sheffield, Sheffield S1 3JD, U.K.; and also with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: esnaola@sheffield.ac.uk).

Samir M. Perlaza is with INRIA, Centre Inria d'Université Côte d'Azur, 06902 Sophia Antipolis, France; also with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA; and also with the GAATI Mathematics Laboratory, University of French Polynesia, 98702 Faaa, French Polynesia (e-mail: samir.perlaza@inria.fr).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

of obtaining generalization guarantees in the form of bounds [19] or even closed form expressions, see [29] and [36]. Nonetheless, such bounds and closed form expressions are often hard to calculate and are not always informative when evaluated in practical settings [1], [29], [30], [37]–[41]. The problem of ERM with a general f-divergence regularization has been explored in [42] and [43] in the case of a finite countable set of models, and recently extended to uncountable sets of models in [44] and [45]. The authors in [42]–[45] constrain the optimization domains to sets of measures that are mutually absolutely continuous with respect to the reference probability measure. In view of these, exploring the asymmetry of relative entropy is of particular interest to advancing the understanding of entropy regularization in the context of ERM and its role in generalization. Additionally, examining the asymmetry opens novel pathways to overcome some of the constraints imposed by relative entropy regularization, such as, the ability to select models outside the support of the prior.

The use of the relative entropy of the optimization measure with respect to the reference measure as a regularizer in the ERM-RER is termed Type-I ERM-RER. Alternatively, the use of the relative entropy of the reference measure with respect to the optimization measure is termed Type-II ERM-RER. Interestingly, the results in [42]–[44], which lead to special cases of the Type-I and Type-II ERM-RER problems by assuming that $f(x) = -x\log(x)$ and $f(x) = -\log(x)$, respectively, do not study the impact of the asymmetry of relative entropy. Another observation that motivates studying the asymmetry of relative entropy in ERM-RER is that numerical analyses of the Type-II ERM-RER, presented in Section VII, suggest that Type-II regularization exhibits a markedly different relationship between test error and training error when compared to that of Type-I regularization. While, the generalization capabilities of Type-I are better in the simulations carried out for this work, the performance of the Type-II regularization is comparable and displays promising properties that warrant further research.

B. Contributions

This paper presents the solution to Type-II ERM-RER optimization problem using a new method of proof. In particular, mutual absolute continuity between the measures involved is not imposed. Surprisingly, mutual absolute continuity is exhibited by the solution as a consequence of the structure of the problem. The key properties of the solution are highlighted, and an equivalence between the Type-I and Type-II ERM-RER problems is presented. This equivalence is achieved by replacing the empirical risk in the Type-I ERM-RER problem with another function, which can be interpreted as a tunable loss function, as described in [46]–[48]. The remainder of the paper is organized as follows. Section II presents the ERM-RER problem and its two variations: Type-I and Type-II. The main contribution of this paper, which is the solution to the Type-II ERM-RER problem, is presented in Section III. This section also presents key properties of the solution. Section VI uses these properties to characterize the expected empirical risk. Section VII studies the equivalence between Type-I and Type-II ERM-RER problems. This work is concluded by Section VIII, with some final remarks.

II. Empirical Risk Minimization

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

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

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

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

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

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

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

Using this notation, 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 does not necessarily possess a solution, *i.e.*, it might happen that $|\mathcal{T}(\boldsymbol{z})| = 0$.

The PAC and Bayesian frameworks, as discussed in [24] and [26], address the problem in (4) by constructing probability measures, conditioned on the dataset $\boldsymbol{z}$, from which models are randomly sampled. In this context, the focus is on probability measures that assign high probability to minimizers of the ERM problem in (4). Such probability measures are defined on the measurable space $(\mathcal{M}, \mathscr{F})$, which is denoted by $\triangle(\mathcal{M})$. From this perspective, the underlying assumption in the remainder of this work is that the functions f and ℓ in (3) are such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the function $g_{x,y} : \mathcal{M} \to [0, \infty)$, such that $g_{x,y}(\boldsymbol{\theta}) = \ell(f(\boldsymbol{\theta}, x), y)$, is measurable with respect to the Borel measurable spaces $(\mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, where $\mathscr{F}$ and $\mathscr{B}(\mathbb{R})$ are, respectively, Borel σ-fields on $\mathcal{M}$ and $\mathbb{R}$. Under these assumptions, a common metric is the expected empirical risk.

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

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

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

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

A. The Type-I ERM-RER Problem

The Type-I ERM-RER problem is parameterized by a probability measure $Q \in \triangle(\mathcal{M})$ and a real $\lambda \in (0, \infty)$. The measure Q is referred to as the *reference measure* and λ as the *regularization factor*. The Type-I ERM-RER problem, with parameters Q and λ, is given by the following optimization problem:

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

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

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

with the notation $P \ll Q$ standing for P being absolutely continuous with respect to Q.

The solution to the Type-I ERM-RER problem in (7) is the Gibbs probability measure reported in [28], [31] and [32]. In order to introduce such a measure, consider the function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ that satisfies for all $t \in \mathbb{R}$,

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp(t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right), \tag{9}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (3). Using this notation, the solution to the Type-I ERM-RER problem in (7) is presented by the following lemma.

Lemma 1 ([29, Theorem 3]): The solution to the optimization problem in (7) is a unique probability measure, denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

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

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3) and the function $K_{Q,\boldsymbol{z}}$ is defined in (9).

B. The Type-II ERM-RER Problem

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

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

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

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

The difference between Type-I and Type-II ERM-RER problems lies on the regularization. While the former uses the relative entropy $\mathsf{D}(P\|Q)$, the latter uses $\mathsf{D}(Q\|P)$. This translates into different optimization domains due to the asymmetry of the relative entropy. More specifically, in the Type-I ERM-RER problem, the optimization domain is the set of probability measures on the Borel measurable space $(\mathcal{M}, \mathscr{F})$ that are absolutely continuous with the reference measure Q. That is, the set $\triangle_Q(\mathcal{M})$ in (8). Alternatively, in the Type-II ERM-RER problem, the optimization domain consists of probability measures defined on the Borel measurable space $(\mathcal{M}, \mathscr{F})$, with the additional condition that the reference measure Q must be absolutely continuous with respect to them. This corresponds to the set denoted as $\triangledown_Q(\mathcal{M})$ in (12). From this perspective, the techniques used in [29] for solving the Type-I ERM-RER no longer hold. In the next section, a new technique is developed for solving the Type-II ERM-RER.

The problems in (7) and (11) exhibit trivial solutions when the functional $\mathsf{R}_{\boldsymbol{z}}$ is such that for all $P \in \triangle_Q(\mathcal{M})$ or $P \in \triangledown_Q(\mathcal{M})$, respectively, it holds that $\mathsf{R}_{\boldsymbol{z}}(P) = c$, for some $c \in [0, \infty)$. In such a case, the solution is unique and equal to the probability measure Q, independently of the parameter λ. In order to avoid this trivial case, the notion of separability of the empirical risk function with respect to the measure Q is borrowed from [29]. A separable empirical risk function with respect to a given probability measure P is defined as follows.

Definition 2 (Definition 5 in [29]): The empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is said to be separable with respect to the probability measure $P \in \triangle(\mathcal{M})$, if there exist a positive real $c > 0$ and two subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{M}$ that are nonnegligible with respect to P, and for all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$,

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

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

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

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

III. The Solution to the Type-II ERM-RER Problem

The solution to the Type-II ERM-RER problem in (11) is presented in the following theorem.

Theorem 1: If there exists a real β such that

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

and

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

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (3), and λ and Q the parameters of the optimization problem in (11), then, the solution to such a problem, denoted by $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle(\mathcal{M})$, is unique and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it satisfies

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

Before introducing the proof of Theorem 1, two important results are presented. The first result provides the solution to the optimization problem in (11) when the optimization domain is restricted to

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

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

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

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (6). The solution to the problem in (18) is described by the following lemma.

Lemma 2: The solution to the optimization problem in (18) is unique and identical to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16).

Proof: The proof is presented in Appendix B. ■

The second result consists of comparing the optimal values resulting from the optimization problems in (11) and (18), as shown hereunder.

Lemma 3: The optimization problems in (11) and (18) satisfy

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

Proof: The proof is presented in Appendix C. ■

Lemma 3 unveils the fact that the objective function in (11) when evaluated at measures whose support extends beyond the support of Q is larger than such an objective function evaluated at measures whose support is identical to the reference measure. This includes the case in which the set $\mathcal{T}(\boldsymbol{z})$ in (5) lies outside the support of Q. Using these results, the proof of Theorem 1 is as follows.

Proof of Theorem 1: The proof follows by observing that from (17), it holds that

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

Hence, from (20), it follows that

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

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

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

Thus, the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) is the solution of the optimization problem in (11), which completes the proof of Theorem 1. ■

Lemma 3 implies that the solution to the optimization problem in (11) is in the set $\bigcirc_Q(\mathcal{M})$ in (16). A consequence of this observation is the following corollary.

Corollary 4: The probability measures Q and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) are mutually absolutely continuous.

Corollary 4 also follows from Theorem 1 by observing that the solution to the Type-II ERM-RER problem in (11) is expressed in terms of its Radon-Nikodym derivative with respect to Q, which implies the absolute continuity of $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to Q. The absolute continuity of the measure Q with respect to $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ follows from the optimization domain of the Type-II ERM-RER problem. From this perspective, Corollary 4 conveys the fact that there does not exist a dataset that can overcome the inductive bias induced by the reference measure Q. That is, sets of models outside the support of Q exhibit zero probability measure with respect to the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

This observation is important as, at first glance, the Type-II relative entropy regularization for the ERM problem in (11) does not restrict the solution to be absolutely continuous with respect to the reference measure Q. However, Theorem 1 shows that the support of the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) collapses into the support of the reference. A parallel can be established between Type-I and Type-II cases, as in both cases, the support of the solution is the support of the reference measure. In a nutshell, the use of relative entropy regularization inadvertently forces the solution to coincide with the support of the reference regardless of the training data.

IV. The Normalization Function

Let the set $\mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0,\infty)$ and $\mathcal{C}_{Q,\boldsymbol{z}} \subset \mathbb{R}$, with Q and $\boldsymbol{z}$ in (11), be such that if $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$, then there exists a $\beta \in \mathcal{C}_{Q,\boldsymbol{z}}$ that satisfies the inclusion in (15a) and (15b). From Theorem 1, specifically from the uniqueness of the solution to (11), it follows that for all $(\lambda,\beta) \in \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{C}_{Q,\boldsymbol{z}}$ and for all $\alpha \in \mathbb{R}$, with $\alpha \neq \beta$, it holds that $(\lambda,\alpha) \notin \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{C}_{Q,\boldsymbol{z}}$. This observation allows establishing a bijection between these two sets. Let such a bijection be represented by the function

$$\bar{K}_{Q,\boldsymbol{z}} : \mathcal{A}_{Q,\boldsymbol{z}} \rightarrow \mathcal{C}_{Q,\boldsymbol{z}}, \tag{23a}$$

which satisfies

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

with λ and β satisfying (15). The function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is referred to as the *normalization function*. This is due to the observation that the Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) can be re-written for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, as

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{24}$$

which together with (15b), implies that the function $\bar{K}_{Q,\boldsymbol{z}}$ ensures that $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) is a probability measure.

The analysis of the normalization function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) relies on the analysis of its functional inverse, denoted by $\bar{K}^{-1}_{Q,\boldsymbol{z}} : \mathcal{C}_{Q,\boldsymbol{z}} \rightarrow \mathcal{A}_{Q,\boldsymbol{z}}$, which can be defined by noticing that

$$1 = \int \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{25a}$$

$$= \int \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{25b}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (3), and λ and Q the parameters of the optimization problem in (11). More specifically, from (23b), it follows that $\lambda = \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)$; and from (25b), it follows that

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta})}. \tag{26}$$

Note that while the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is defined implicitly, its functional inverse $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ is defined explicitly in (26). The existence of the function and inverese follows from the fact that $\bar{K}_{Q,\boldsymbol{z}}$ is a bijection.

The purpose of the remaining of this section is to provide a characterization of the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$. To do so, some mathematical objects are introduced. Given a real $\delta \in [0,\infty)$, consider the Rashomon set [49], $\mathcal{L}_{\boldsymbol{z}}(\delta)$, defined as follows

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

Consider also the real numbers $\delta^{\star}_{Q,\boldsymbol{z}}$ and $\lambda^{\star}_{Q,\boldsymbol{z}}$ defined as follow:

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

and

$$\lambda^{\star}_{Q,\boldsymbol{z}} \triangleq \inf \mathcal{A}_{Q,\boldsymbol{z}}. \tag{29}$$

Let also $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ be the level set of the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) for the value $\delta^{\star}_{Q,\boldsymbol{z}}$. That is,

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}. \tag{30}$$

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

Lemma 5: The function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is strictly increasing and continuous.

Proof: The proof is presented in Appendix D. ■

The following lemma characterizes the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$ in (23a), which are the domain and codomain of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b).

Lemma 6: The set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (23a) is either empty or an interval of the form

$$\mathcal{A}_{Q,\boldsymbol{z}} = \begin{cases} [\lambda_{Q,\boldsymbol{z}},\infty) & \text{if } \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) < \infty \\ (0,\infty) & \text{otherwise,} \end{cases} \tag{31}$$

where

$$\lambda_{Q,\boldsymbol{z}} = \bar{K}_{Q,\boldsymbol{z}}^{-1}\big(-\delta^{\star}_{Q,\boldsymbol{z}}\big), \tag{32}$$

with $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (26), the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3), and $\delta^{\star}_{Q,\boldsymbol{z}}$ is defined in (28). Moreover, the set $\mathcal{C}_{Q,\boldsymbol{z}}$ in (23a) is either empty or an interval of the form

$$\mathcal{C}_{Q,\boldsymbol{z}} = \begin{cases} \big[-\delta^{\star}_{Q,\boldsymbol{z}},\infty\big) & \text{if } \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \\ \big(-\delta^{\star}_{Q,\boldsymbol{z}},\infty\big) & \text{otherwise.} \end{cases} \tag{33}$$

Proof: The proof is presented in Appendix E. ■

Lemma 6 shows that the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$ in (23a) are convex sets (intervals). Appendix T introduces some examples to illustrate particular cases in which the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is open or semi-open. The convexity of $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$ is crucial for analyzing how the choice of λ influences whether the Type-II ERM-RER problem in (11) has a solution. For instance, if $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$, then the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the unique solution to the problem in (11) (Theorem 1). Moreover, as λ increases, the resulting Type-II ERM-RER problem still possesses a solution, which is formalized by the following corollary.

Corollary 7: If the Type-II ERM-RER problem in (11) possesses a solution, then, the following problem

$$\min_{P \in \triangledown_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \alpha \mathsf{D}(Q\|P), \tag{34}$$

with $\alpha \geqslant \lambda$, also possesses a solution.

Additionally, Lemma 6 allows identifying how small λ in (11) can be, such that the Type-II ERM-RER problem in (11) still possesses a solution. The regularization factor λ can be made arbitrarily close to zero in some cases, as shown hereunder.

Corollary 8: If the set $\mathcal{M}$ is finite, then the set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (23a) is $(0,\infty)$.

Corollary 8 follows by noticing that if the set $\mathcal{M}$ is finite, the subset $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (30) satisfies $Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) > 0$. Thus, the integral in (33) is not finite, which follows from the fact that for all $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}} = 0$. Another immediate consequence of Lemma 5 and Lemma 6 is the following corollary.

Corollary 9: If the real value $\delta^{\star}_{Q,\boldsymbol{z}} = 0$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), then the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) is strictly positive.

This section is closed by leveraging Lemma 6 for presenting a key property of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b).

Lemma 10: The function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) satisfies

$$\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\delta^{\star}_{Q,\boldsymbol{z}}, \tag{35}$$

where $\delta^{\star}_{Q,\boldsymbol{z}}$ and $\lambda^{\star}_{Q,\boldsymbol{z}}$ are defined in (28) and (29), respectively.

Proof: The proof is presented in Appendix F. ■

The limit in (35) is determined by the set of models in the support of the prior with the lowest empirical risk determined by the choice of the loss function ℓ and the function f in (3).

V. Properties of the Solution

A. Bounds on the Radon-Nikodym Derivative

Note that from Theorem 1, models resulting in lower empirical risks correspond to greater values of the Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16). The following corollary formalizes this observation.

Lemma 11: For all $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \in (\operatorname{supp} Q)^2$, such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) \leq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)$, with $\mathsf{L}_{\boldsymbol{z}}$ in (3), the Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) satisfies

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2) \leq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1), \tag{36}$$

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

Proof: The proof is presented in Appendix G. ■

The intuition that follows from Lemma 11 is that, under the assumption that the ERM problem in (4) possesses a solution in the support of the reference measure, the maximum of the function $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) is achieved by the models in $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$, provided that it is not empty, with $\mathcal{T}(\boldsymbol{z})$ in (5). Furthermore, the Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is monotonic with respect to the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (3). This property is similar to that of the solution to the Type-I ERM-RER problem in (10), as established in [29, Corollary 1].

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

Lemma 12: The Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$0 < \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leq \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} < \infty, \tag{37}$$

where the function $\bar{K}_{Q,\boldsymbol{z}}$ and the real $\delta^{\star}_{Q,\boldsymbol{z}}$ are defined in (23a) and (28), respectively. The equality holds if and only if $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (30).

Proof: The proof is presented in Appendix H. ■

B. Asymptotes of the Radon-Nikodym Derivative

In the asymptotic regime, when the regularization factor λ in (11) grows to infinity, *i.e.*, $\lambda \to \infty$, the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ becomes identical to the reference measure Q, up to sets of measure zero, as described in the following lemma.

Lemma 13: The Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\lim_{\lambda \to \infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 1. \tag{38}$$

Proof: The proof is presented in Appendix I. ■

Lemma 13 unveils a similarity between Type-I and Type-II regularization as the Type-I measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (10), also exhibits a similar behavior [29, Lemma 7].

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

Lemma 14: If $Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) > 0$, with the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (30), then the Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (16) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big)} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\}}. \tag{39}$$

On the ofther hand, if $Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 0$ and $\lambda^{\star}_{Q,\boldsymbol{z}}$ in (29) satisfies $\lambda^{\star}_{Q,\boldsymbol{z}} = 0$, then for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \begin{cases} \infty & \text{if } \boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

Conversely, if $Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 0$ and $\lambda^{\star}_{Q,\boldsymbol{z}} > 0$, then for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda^{\star}_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}. \tag{41}$$

Proof: The proof is presented in Appendix J. ■

Lemma 14 highlights that in the asymptotic regime when the regularization factor decreases to zero from the right, *i.e.*, $\lambda \to 0^{+}$, the value $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$ does not depend on the exact model $\boldsymbol{\theta}$ but rather on whether $\boldsymbol{\theta} \in \operatorname{supp} Q \cap \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$. In the case in which $\boldsymbol{\theta} \in \operatorname{supp} Q \cap \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, it holds that $\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) > 0$. Otherwise, $\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0$. In the special case in which $\delta^{\star}_{Q,\boldsymbol{z}} = 0$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ satisfies $\mathcal{L}^{\star}_{Q,\boldsymbol{z}} = \mathcal{T}(\boldsymbol{z})$, where $\mathcal{T}(\boldsymbol{z})$ is defined in (5). This implies a concentration of probability over $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$, which establishes a connection with the ERM problem without regularization in (4).

Furthermore, in the asymptotic regime, when the regularization factor decreases to zero from the right, the solutions to the Type-I and Type-II ERM-RER problems exhibit the same asymptotic behavior, as shown in [29, Lemma 6]. This aligns with the observation that as λ decreases, the optimization problems in (7) and (11) exhibit a weaker relative entropy constraint. A stronger result that follows from Lemma 14 is presented in the following lemma.

Lemma 15: If $\lambda^{\star}_{Q,\boldsymbol{z}}$ in (29) satisfies $\lambda^{\star}_{Q,\boldsymbol{z}} = 0$, then the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) and the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (30) satisfy

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

Alternatively, if $\lambda^{\star}_{Q,\boldsymbol{z}} > 0$, then the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) and the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (30) satisfy

$$\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 0. \tag{43}$$

Proof: The proof is presented in Appendix K. ■

Lemma 15 shows that indeed, if $\lambda^{\star}_{Q,\boldsymbol{z}} = 0$ and the regularization factor λ approaches zero from the right, the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) concentrates on a set of models that induce minimum empirical risk, *i.e.*, the set $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (30).

VI. The Expected Empirical Risk

This section focuses on the expected empirical risk induced by the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16). That is, the value $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$, with the functional $\mathsf{R}_{\boldsymbol{z}}$ defined in (6).

The following lemma establishes a relation between $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$, λ, and the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b).

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

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{44}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ and the function $\bar{K}_{Q,\boldsymbol{z}}$ are defined in (6) and (23b), respectively.

Proof: The proof is presented in Appendix L. ■

Lemma 16 characterizes the expected empirical risk of the Type-II ERM-RER solution and establishes a direct connection to the regularization factor λ. For example, from Lemma 10 if there exists a model $\boldsymbol{\theta}^{\star} \in \operatorname{supp} Q$, such that its empirical risk is zero, the function $\bar{K}_{Q,\boldsymbol{z}}$ is nonegative, which implies that λ serves as an explicit upper bound to the expected empirical risk. This provides a clear interpretation: the choice of λ directly controls and bounds the average expected empirical risk. This property gives Type-II ERM-RER a distinct risk management strategy over that of Type-I ERM-RER, mainly lowering the expected empirical risk by appropriately choosing the value of λ. Additionally, Lemma 16 highlights that the the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$, with Q and $\boldsymbol{z}$ fixed, inherits all properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b). The following lemma formalizes this observation.

Lemma 17: The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$, with the functional $\mathsf{R}_{\boldsymbol{z}}$ in (6) and the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16), is continuous and nondecreasing with respect to λ. Moreover, it is strictly increasing if and only if the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable with respect to the probability measure Q.

Proof: The proof is presented in Appendix M. ■

A. Bounds on the Expected Empirical Risk

This section builds on the characterization of the expected empirical risk and its monotonicity with respect to the regularization factor λ in (11) to establish a range of bounds on the expected empirical risk. The following lemma highlights a connection existing between the expected empirical risks $\mathsf{R}_{\boldsymbol{z}}(Q)$ and $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$; and the relative entropy $\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$.

Lemma 18: The functional $\mathsf{R}_{\boldsymbol{z}}$ defined in (6) and the measures Q and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) satisfy

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq \lambda\Big(\exp\Big(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\Big) - 1\Big). \tag{45}$$

Proof: The proof is presented in Appendix N. ■

Note that $\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq 0$ in (45), which leads to the observation that

$$\Big(\exp\Big(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\Big) - 1\Big) \geq 0. \tag{46}$$

Hence, from Lemma 18, it follows that the solution to the Type-II ERM-RER problem induces an expected empirical risk that is smaller than the one induced by reference measure Q. This is formalized by the following corollary.

Corollary 19: The probability measures Q and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) satisfy

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leq \mathsf{R}_{\boldsymbol{z}}(Q), \tag{47}$$

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

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

Lemma 20: The probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) satisfies

$$\delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < \lambda + \delta^{\star}_{Q,\boldsymbol{z}}, \tag{48}$$

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

Proof: The proof is presented in Appendix O. ■

The bounds presented in Lemma 20 highlight that the regularization parameter λ in (11) governs the increase of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ with respect to its minimum, i.e, $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28). Moreover, the lower bound is tight for the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) in the asymptotic regime when λ decreases to $\lambda^{\star}_{Q,\boldsymbol{z}}$ from right, as shown hereunder.

Lemma 21: The probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) satisfies

$$\lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda^{\star}_{Q,\boldsymbol{z}} + \delta^{\star}_{Q,\boldsymbol{z}}, \tag{49}$$

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

Proof: From Lemma 16, it holds that

$$\begin{aligned}
&\lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad= \lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \lambda - \lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) &(50)\\
&\quad= \lambda^{\star}_{Q,\boldsymbol{z}} - \lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) &(51)\\
&\quad= \lambda^{\star}_{Q,\boldsymbol{z}} + \delta^{\star}_{Q,\boldsymbol{z}}, &(52)
\end{aligned}$$

where equality (52) follows from Lemma 10. This completes the proof. ■

Finally, note that the functional $\mathsf{R}_{\boldsymbol{z}}$ in (6) is nonnegative. This observation together with Lemma 16 lead to a new property for the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b), which is stated by the following corollary

Corollary 22: The function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) satisfies, for all $t > \lambda^{\star}_{Q,\boldsymbol{z}}$, with $\lambda^{\star}_{Q,\boldsymbol{z}}$ in (29),

$$\bar{K}_{Q,\boldsymbol{z}}(t) \leq t. \tag{53}$$

B. (δ, ϵ)-Optimality

This section presents a PAC guarantee for models sampled from the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16), with respect to the Type-II ERM-RER problem in (11). Such guarantee is presented using the notion of (δ, ϵ)-optimality introduced in [29, Definition 6].

Definition 3 ([29, Definition 6]): Given a pair of positive reals (δ, ϵ), with $\epsilon < 1$, the probability measure $P \in \triangle(\mathcal{M})$ is said to be (δ, ϵ)-optimal if the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ in (27) satisfies

$$P(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 1 - \epsilon. \tag{54}$$

The following theorem presents a (δ, ϵ)-optimality guarantee for the solution to the Type-II ERM-RER problem in (11).

Theorem 2: Assume that $\lambda^{\star}_{Q,\boldsymbol{z}} = 0$, then for all $(\delta, \epsilon) \in \left(\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right) \times (0, 1)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), there always exists a $\lambda \in (0, \infty)$, such that the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) is (δ, ϵ)-optimal.

Proof: The proof is presented in Appendix P. ■

VII. Interplay Between the Relative Entropy Asymmetry and the Empirical Risk

This section presents a connection between the Type-I ERM-RER in (7) and Type-II ERM-RER in (11) established via a transformation of the empirical risk function. The connection is established by proving the existence of two functions $\mathsf{W}_{Q,\boldsymbol{z},\lambda} : \mathcal{M} \to \mathbb{R}$ and $\mathsf{V}_{Q,\boldsymbol{z},\lambda} : \mathcal{M} \to \mathbb{R}$, such that the solution to the optimization problem in (7) is identical to the solution of the following problem:

$$\min_{P\in\triangledown_{Q}(\mathcal{M})} \int \mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}) + \mathsf{D}(Q\|P), \qquad (55)$$

with λ and Q in (7); and the solution to the optimization problem in (11) is identical to the solution of the following problem:

$$\min_{P\in\triangle_{Q}(\mathcal{M})} \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}) + \mathsf{D}(P\|Q), \qquad (56)$$

with λ and Q in (11). The main result of this section is presented in the following theorem.

Theorem 3: If the problems in (7) and in (11) have solutions, then

$$\min_{P\in\triangledown_{Q}(\mathcal{M})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}) \ + \lambda\mathsf{D}(Q\|P) \ = \ \min_{P\in\triangle_{Q}(\mathcal{M})} \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}) \ + \mathsf{D}(P\|Q), \qquad (57a)$$

where the function $\mathsf{V}_{Q,\boldsymbol{z},\lambda} : \mathcal{M} \to \mathbb{R}$ is defined as

$$\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) = \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big), \qquad (57b)$$

and

$$\min_{P\in\triangle_{Q}(\mathcal{M})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}) \ + \lambda\mathsf{D}(P\|Q) \ = \ \min_{P\in\triangledown_{Q}(\mathcal{M})} \int \mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}) \ + \mathsf{D}(Q\|P), \qquad (58a)$$

where the function $\mathsf{W}_{Q,\boldsymbol{z},\lambda} : \mathcal{M} \to \mathbb{R}$ is such that

$$\mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) = \frac{\lambda}{\exp\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - K_{Q,\boldsymbol{z}}\big(-\frac{1}{\lambda}\big)\Big)} - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \qquad (58b)$$

where the functions $K_{Q,\boldsymbol{z}}$ and $\bar{K}_{Q,\boldsymbol{z}}$ are defined in (9) and (23), respectively.

Proof: Denote by $\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the solution to the optimization problem in (56). From Lemma 1, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\begin{aligned}
&\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \frac{\exp(-\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}))}{\int \exp(-\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\nu}))\, \mathrm{d}Q(\boldsymbol{\nu})} && (59)\\
&= \frac{\exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big)}{\int \exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big) \mathrm{d}Q(\boldsymbol{\nu})} && (60)\\
&= \frac{\Big(\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu})\Big)^{-1}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} && (61)\\
&= \frac{\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} && (62)\\
&= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && (63)\\
&= \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), && (64)
\end{aligned}$$

where (60) follows from (57b); (62) follows from (26); (63) follows from (23b); and (64) follows from Theorem 1. This completes the proof of (58a).

Similarly, denote by $\tilde{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the solution to the optimization problem in (55). From Theorem 1, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\begin{aligned}
&\frac{\mathrm{d}\tilde{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \frac{\lambda}{\mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && (65)\\
&= \frac{\lambda}{\frac{\lambda}{\exp\big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - K_{Q,\boldsymbol{z}}\big(-\frac{1}{\lambda}\big)\big)} - \bar{K}_{Q,\boldsymbol{z}}(\lambda) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && (66)\\
&= \exp\bigg(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - K_{Q,\boldsymbol{z}}\bigg(-\frac{1}{\lambda}\bigg)\bigg) && (67)\\
&= \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), && (68)
\end{aligned}$$

where (65) follows from (24); (66) follows from (58b); and (68) follows from Lemma 1. This completes the proof of (57a). ■

Theorem 3 establishes an equivalence between the regularization of Type-I and Type-II. More specifically, Theorem 3 highlights that by modifying the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (57b) using the function $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ in (57b), the Type-II ERM-RER problem in (11) can be solved by solving the Type-I ERM-RER problem in (56). It is noteworthy that Type-I regularization imposes the support of the solution to be contained within the support of the reference measure, i.e., $\operatorname{supp} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \subseteq \operatorname{supp} Q$. Similarly, Type-II regularization imposes the support of the solution to contain the support of the reference measure, i.e., $\operatorname{supp} Q \subseteq \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Interestingly, these inclusions can be shown to be equalities from Theorem 1 and Lemma 1. That is,

$$\operatorname{supp} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} = \operatorname{supp} Q = \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}. \qquad (69)$$

The remainder of the section focuses on the transformation from Type-I to Type-II. The function $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ in (57b) is referred to as the *log-empirical* risk function. The *expected log-empirical risk* is defined as follows.

Definition 4 (Expected Log-Empirical Risk): Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (1) and the log-empirical risk function $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ in (57b), let the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda} : \triangle(\mathcal{M}) \to \mathbb{R}$ be such that

$$\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) = \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}). \qquad (70)$$

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

Using the *expected log-empirical risk* defined above, the optimization problem in (55) can be rewritten as follows

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) + \mathsf{D}(P\|Q), \tag{71}$$

with λ and Q being parameters of the Type-I and Type-II ERM-RER problems in (7) and (11). The Type-I - Type-II relation in Theorem 3 can be used to establish an equality involving the relative entropies $\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ and $\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)$; and the expected log-empirical risks $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ and $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q)$, as shown hereunder.

Lemma 23: The functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ in (70) and the probability measures $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q in (16) satisfy

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

$$= \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q) - \mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{73}$$

Proof: The proof is presented in Appendix Q. ■

A. Sensitivity of the Log-Empirical Risk

The sensitivity of the expected empirical risk, as presented in [29, Definition 7], is defined as follows.

Definition 5 (Sensitivity of the Expected Empirical Risk): Consider the functional $\mathsf{R}_{\boldsymbol{z}}$ in (6) and let $\mathsf{S}_{Q,\lambda} : (\mathcal{X}\times\mathcal{Y})^n \times \triangle_Q(\mathcal{M}) \to \mathbb{R}$ be a functional such that

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{74}$$

where the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is defined in (10). 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 $\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P)$.

Similarly, the sensitivity of the expected log-empirical risk $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ in (57b) is defined as follows.

Definition 6 (Sensitivity of the Expected Log-Empirical Risk): Consider the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ in (70) and let $\bar{\mathsf{S}}_{Q,\lambda} : (\mathcal{X}\times\mathcal{Y})^n \times \triangledown_Q(\mathcal{M}) \to \mathbb{R}$ be a functional such that

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{75}$$

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

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

Lemma 24: The sensitivity $\bar{\mathsf{S}}_{Q,\lambda}$ in (75) satisfies for all probability measures $P \in \bigcirc_Q(\mathcal{M})$ that

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P) = \mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}(P\|Q) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right), \tag{76}$$

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

Proof: The proof is presented in Appendix R. ■

An interesting interpretation of Lemma 24 follows from rewriting (76) using the objective function of the Type-I ERM-RER problem in (56) as follows:

$$\mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}(P\|Q) - \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right). \tag{77}$$

That is, the relative entropy $\mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ represents the variation of the objective function of the Type-I ERM-RER problem in (56) due to a deviation from the solution $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to an alternative probability measure P.

In Lemma 24, when P is chosen to be identical to the reference measure Q, it follows that

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, Q) = \mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right), \tag{78}$$

where the right-hand side is a Jeffreys divergence [50], also known as the symmetrized Kullback-Leibler divergence between the measures $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q. Furthermore, by observing that $\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) \geq 0$, and $\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq 0$ [29, Theorem 1], Lemma 24 leads to the following corollary.

Corollary 25: The probability measures Q and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) satisfy

$$\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leq \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q), \tag{79}$$

where the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ is defined in (70).

B. Type-I and Type-II Optimal Measures

The solutions to the optimization problems (7) and (11), with regularization factors λ and α, respectively, are $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (10) and in (16). These measures exhibit the following property.

Lemma 26: The probability measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (10) and in (16), respectively, satisfy

$$\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) = \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{80}$$

where the function $K_{Q,\boldsymbol{z}}$ is defined in (9).

Proof: The proof is presented in Appendix S. ■

Lemma 26 characterizes the relative entropy difference of Type-I and Type-II ERM-RER solution with respect to the prior Q. In doing so, it provides an alternative way to evaluate this difference without directly computing the corresponding relative entropies.

Finally, two important properties of the Type-I and Type-II optimal measures are presented by the following corollary of Lemma 26.

Corollary 27: The probability measures $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (10) and in (16), respectively, satisfy

$$\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \mathsf{D}\left(P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \left(\log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) \tag{81}$$

and

$$\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$$
$$= \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \quad (82)$$

where the functionals $\mathsf{S}_{Q,\lambda}$ and $\bar{\mathsf{S}}_{Q,\alpha}$ are respectively defined in (74) and in (75); and the function $K_{Q,\boldsymbol{z}}$ is defined in (9).

The equality in (81) quantifies the variation of the expected log-empirical risk due to a deviation from the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (10) to the probability measure $\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) via the sensitivity $\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$. The equality in (82) quantifies the variation of the expected empirical risk due to a deviation from the probability measure $\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (10) via the sensitivity $\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$.

C. Numerical Comparison of Type-I and Type-II Regularization

In machine learning, the generalization error describes the capacity of a learning algorithm to select models based on training data that performs well with unseen test data. The sensitivity (Definition 5) of the optimization problem in (7) is closely related to the generalization error [29, Theorem 16]. See also [30], [51]. A probability measure $P \in \triangle_Q(\mathcal{M})$, e.g., a machine learning algorithm, that yields larger sensitivity indicates that the learning overfits with respect to the training data, leading to an increase in the generalization error [51]. In this context, algorithms arising from the Type-I and Type-II ERM-RER are used for the classification of two handwritten numbers from the MNIST dataset [52]. The MNIST example is simplified to accommodate a parameterized model in $\mathbb{R}^2$ such that the numerical approximations of the generalization error for different regularization factors are meaningful [36].

The MNIST dataset consists of 60,000 images for training and 10,000 images for testing. Out of the 60,000 training images, 12,183 are labeled as the digits six and seven, while 1,986 out of the 10,000 test images correspond to these digits. Each image is a 28×28 grayscale picture and is represented by a matrix in $[0,1]^{28\times 28}$. To reduce the computational complexity, the pictures are processed following the procedure described in Appendix U. Consider the Type-I ERM-RER problem in (7) and the Type-II ERM-RER problem in (11) and assume that: (i) the set of models is $\mathcal{M} = [-50, 50]^2$; (ii) the set of patterns $\mathcal{X}$ is formed by computing the histogram of gradients (HOG) of the pictures such that $\mathcal{X} \subset \mathbb{R}^{1296}$ of the handwritten six and seven in the MNIST dataset; (iii) the set of labels is $\mathcal{Y} = \{6, 7\}$; (iv) the reference measure Q is chosen to be a uniform probability measure over the set of models; (v) the function f in (3) is defined as

$$f(\boldsymbol{\theta}, x) = \begin{cases} 6 & \text{if } 0 < (x\boldsymbol{W})\boldsymbol{\theta}, \\ 7 & \text{if } 0 > (x\boldsymbol{W})\boldsymbol{\theta}, \end{cases} \quad (83)$$

where the matrix $\boldsymbol{W}$ is defined in (354) in Appendix U; and (vi) the loss function ℓ in (2) satisfies

$$\ell(f(\boldsymbol{\theta}, x), y) = \mathbb{1}_{\{f(\boldsymbol{\theta},x)\neq y\}}. \quad (84)$$

For the simulation, 8,100 data points are uniformly sampled from the 12,183 available training images, forming the dataset $\boldsymbol{z}_1$, referred to as the *training dataset*. Similarly, 1,300 data points are uniformly sampled from the 1,986 available test images, forming the dataset $\boldsymbol{z}_2$, referred to as the *test dataset*. Not all images are used because the simulation is repeated only 100 times, and at each iteration, the datasets $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are uniformly resampled. Figure 5 displays the average (over 100 repetitions) generalization gap for Type-I and Type-II algorithms. The average generalization gap of Type-II suggests it is less prone to overfitting. In contrast, Figure 3 and Figure 4 show that Type-I achieves a lower average training error, which results on a lower average test error. These observations imply that Type-II promotes a more positive conservative learning, reducing the generalization gap by keeping average training and testing error closer, while Type-I achieves lower training error at the cost of a higher avergae generalization gap, indicating greater reliance on the training data.

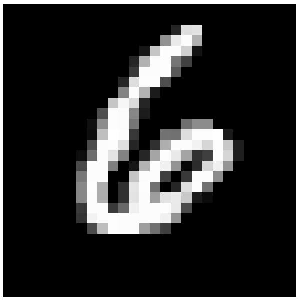

Fig. 1. 28×28 Image of a handwritten 6 from MNIST dataset.

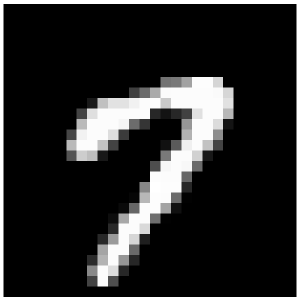

Fig. 2. 28×28 Image of a handwritten 7 from MNIST dataset.

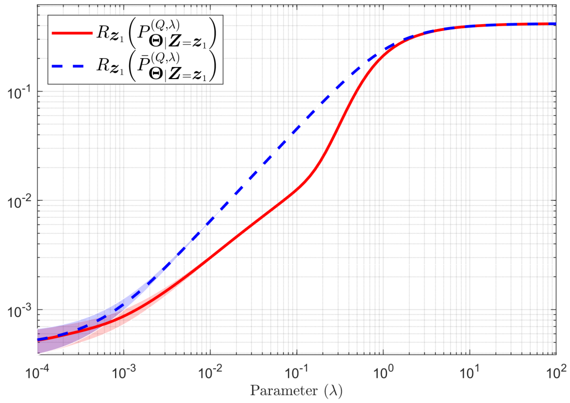

Fig. 3. Average Training Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ and $\mathsf{R}_{\boldsymbol{z}_1}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, with the measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ in (10) and (16), respectively, computed over one hundred different $\boldsymbol{z}_1$ (training dataset) random selections.

Another key observation is that, for certain ranges of the regularization factor (e.g.,$\lambda \in (0.002, 0.09)$ and $\lambda \in (0.3, 0.8)$ for Type-II), where the average test error are comparable, Type-II exhibits lower average generalization gap. This suggests that Type-II can achieve both small test error and generalization error for certain regularization factor, a highly desirable outcome. However, selecting such regularization

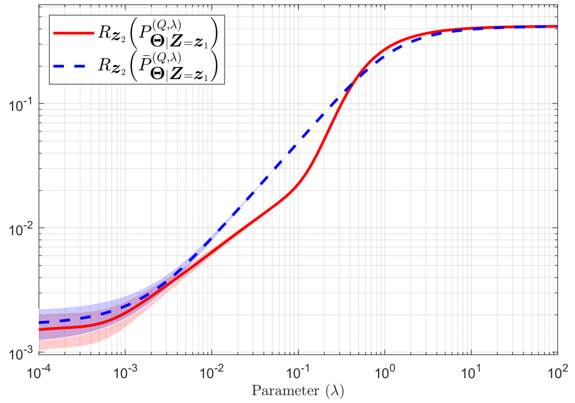

Fig. 4. Average Test Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ and $\mathsf{R}_{\boldsymbol{z}_2}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, with the measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ in (10) and (16), respectively, computed over one hundred different $\boldsymbol{z}_1$ (training dataset) and $\boldsymbol{z}_2$ (test dataset) random selections.

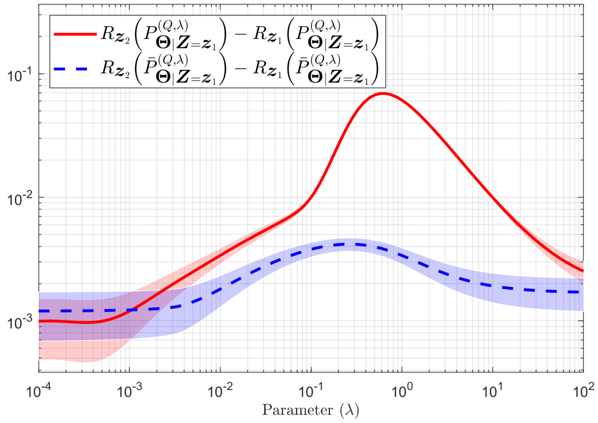

Fig. 5. Average of the differences (generalization gaps) $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ and $\mathsf{R}_{\boldsymbol{z}_2}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_1}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, with the measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ and $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ in (10) and (16), respectively, computed over one hundred different $\boldsymbol{z}_1$ (training dataset) and $\boldsymbol{z}_2$ (test dataset) random selections.

ranges remains an open question for future research under the theoretical framework presented in this paper.

VIII. Final Remarks

This work has introduced the Type-II ERM-RER problem and has presented its solution through Theorem 1. The solution highlights that regardless of whether Type-I or Type-II regularization is used in ERM problems, the models that are considered by the resulting solution are necessarily in the support of the reference measure. In this sense, the restriction over the models introduced by the reference measure cannot be bypassed by the training data when relative entropy is used as the regularizer. This limitation has been shown to be a consequence of the equivalence that can be established between Type-I and Type-II regularization. These analytical results lead to an operationally meaningful characterization of the expected empirical risk induced by the Type-II solution in terms of the regularization parameters. The closed-form expressions for the expected empirical risk induced by Type-I and Type-II errors are used to characterize the sensitivity of the expected empirical risk and the sensitivity of the expected log-empirical risk, in terms of the cumulant generating function and Kullback-Leibler divergence. The analysis of the solution to the optimization problem (11) shows that, under mild assumptions, there always exists a positive real value λ such, with probability $1-\epsilon$, the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ concentrates on the set of models that minimize the empirical risk.

Appendix A
Preliminary

This appendix introduces a preliminary result, which is central in the proof of Lemma 2.

Lemma 28: Let $\mathscr{M}$ be the set of measurable functions $h : \mathcal{M} \rightarrow \mathbb{R}$, with respect to the measurable space $(\mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Let $\mathscr{S}$ be the subset of $\mathscr{M}$ including all non-negative functions that are absolutely integrable with respect to a probability measure Q. That is, for all $h \in \mathscr{S}$, it holds that

$$\int |h(\boldsymbol{\theta})| \,\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{85}$$

Let the function $\hat{r} : \mathbb{R} \rightarrow \mathbb{R}$ be such that

$$\hat{r}(\alpha) = \int -\log(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{86}$$

for some functions g and h in $\mathscr{S}$. The function $\hat{r}$ in (86) is differentiable at zero.

Proof: The objective is to prove that the function $\hat{r}$ in (86) is differentiable at zero, which reduces to proving that the limit

$$\lim_{\delta \rightarrow 0} \frac{1}{\delta}(\hat{r}(\alpha+\delta) - \hat{r}(\alpha)) \tag{87}$$

exists for all $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small. Let the function $f : (0, \infty) \rightarrow \mathbb{R}$ be a function such that

$$f(x) = -\log(x). \tag{88}$$

Note that the function $\hat{r}$ can be written in terms of f as follows:

$$\hat{r}(\alpha) = \int f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{89}$$

The proof of the existence of such limit in (87) relies on the fact that the function f in (88) is strictly convex and differentiable, which implies that f is also Lipschitz continuous. Hence, it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}&|f(g(\boldsymbol{\theta}) + (\alpha+\delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))|\\ &\leq c\,|h(\boldsymbol{\theta})|\,|\delta|,\end{aligned} \tag{90}$$

for some positive and finite constant c, which implies that

$$\begin{aligned}&\frac{|f(g(\boldsymbol{\theta}) + (\alpha+\delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))|}{|\delta|}\\ &\leq c\,|h(\boldsymbol{\theta})|,\end{aligned} \tag{91}$$

and thus, given that $h \in \mathscr{S}$, it holds that

$$\begin{aligned}&\int \frac{|f(g(\boldsymbol{\theta}) + (\alpha+\delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))|}{|\delta|}\,\mathrm{d}Q(\boldsymbol{\theta})\\ &\leq \infty.\end{aligned} \tag{92}$$

This facilitates using the dominated convergence theorem as follows. From the fact that the function f is differentiable, the limit in (87) satisfies for $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small,

$$\begin{aligned}
&\lim_{\delta\to 0}\frac{1}{\delta}(\hat{r}(\alpha+\delta)-\hat{r}(\alpha))\\
&= \lim_{\delta\to 0}\frac{1}{\delta}\bigg(\int f(g(\boldsymbol{\theta})+(\alpha+\delta)h(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\\
&\quad -\int f(g(\boldsymbol{\theta})+\alpha h(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\bigg) && \text{(93a)}\\
&= \lim_{\delta\to 0}\int \frac{1}{\delta}(f(g(\boldsymbol{\theta})+(\alpha+\delta)h(\boldsymbol{\theta}))\\
&\quad -f(g(\boldsymbol{\theta})+\alpha h(\boldsymbol{\theta})))\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(93b)}\\
&\leq \lim_{\delta\to 0}\int \frac{1}{|\delta|}|f(g(\boldsymbol{\theta})+(\alpha+\delta)h(\boldsymbol{\theta}))\\
&\quad -f(g(\boldsymbol{\theta})+\alpha h(\boldsymbol{\theta}))|\mathrm{d}Q(\boldsymbol{\theta}) && \text{(93c)}\\
&= \int \lim_{\delta\to 0}\frac{1}{|\delta|}|f(g(\boldsymbol{\theta})+(\alpha+\delta)h(\boldsymbol{\theta}))\\
&\quad -f(g(\boldsymbol{\theta})+\alpha h(\boldsymbol{\theta}))|\mathrm{d}Q(\boldsymbol{\theta}) && \text{(93d)}\\
&= c\int h(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(93e)}\\
&< \infty, && \text{(93f)}
\end{aligned}$$

where (93c) follows from [53, Theorem 1.5.9(c)]; (93d) follows from the dominated convergence theorem [53, Theorem 1.6.9]; (93e) follows from (91); and (93f) follows from (85). Finally, from (93f), it follows that the function $\hat{r}$ in (86) is differentiable at zero, which completes the proof. ■

Appendix B
Proof of Lemma 2

The optimization problem in (18) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the measure Q, which yields:

$$\begin{aligned}
\min_{P\in\bigcirc_Q(\mathcal{M})} &\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta})\\
&-\lambda\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}), && \text{(94a)}\\
\text{s.t.}\quad &\int\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})=1. && \text{(94b)}
\end{aligned}$$

The remainder of the proof focuses on the problem in which the optimization is over the function $\frac{\mathrm{d}P}{\mathrm{d}Q}:\mathcal{M}\to\mathbb{R}$, instead of the measure P. This is due to the fact that for all $P\in\bigcirc_Q(\mathcal{M})$, the Radon-Nikodym derivate $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is unique up to sets of zero measure with respect to Q [53, Theorem 2.2.1]. Let $\mathscr{S}$ be the set defined in Lemma 28. Using this notation, the optimization problem of interest is:

$$\begin{aligned}
\min_{g\in\mathscr{S}}&\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta})-\lambda\int\log(g(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(95a)}\\
\text{s.t.}&\int g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})=1. && \text{(95b)}
\end{aligned}$$

Let the Lagrangian of the optimization problem in (95) be $L:\mathscr{S}\times\mathbb{R}\to\mathbb{R}$ such that

$$\begin{aligned}
L(g,\beta)=&\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta})-\lambda\int\log(g(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta})\\
&+\beta\left(\int g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta})-1\right) && \text{(96)}\\
=&\int\Big(g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)-\lambda\log(g(\boldsymbol{\theta}))\Big)\mathrm{d}Q(\boldsymbol{\theta})-\ \beta, && \text{(97)}
\end{aligned}$$

where β is a real that acts as a Lagrange multiplier due to the constraint (95b). Let $\hat{g}:\mathcal{M}\to\mathbb{R}$ be a function in $\mathscr{S}$. The Gateaux differential of the functional L in (96) at $(g,\beta)\in\mathscr{S}\times\mathbb{R}$ in the direction of $\hat{g}$ is

$$\partial L(g,\beta;\hat{g})\triangleq\frac{\mathrm{d}}{\mathrm{d}\gamma}L(g+\gamma\hat{g},\beta)\bigg|_{\gamma=0}. \tag{98}$$

Let the function $r:\mathbb{R}\to\mathbb{R}$ be defined for some fixed functions g and $\hat{g}$ and some fixed β such that for all $\gamma\in(-\epsilon,\epsilon)$, with ϵ arbitrarily small,

$$r(\gamma)=L(g+\gamma\hat{g},\beta). \tag{99}$$

The proof follows by showing that the function r in (99) is differentiable at zero, in order to prove the existence of the Gateaux differential in (98) for those functions g and $\hat{g}$ and real β. For doing so, note that

$$\begin{aligned}
r(\gamma)=&\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta})\\
&-\lambda\int\log(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta})\\
&+\beta\left(\int(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})-1\right) && \text{(100a)}\\
=&\gamma\int\hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)\,\mathrm{d}Q(\boldsymbol{\theta})\\
&-\lambda\int\log(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta})\\
&+\int g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)\,\mathrm{d}Q(\boldsymbol{\theta})-\beta, && \text{(100b)}
\end{aligned}$$

where the first term in (100b) is linear with respect to γ and the third term is independent of γ. The second term can be written using the function $\hat{r}:\mathbb{R}\to\mathbb{R}$, which is defined for fixed functions g and $\hat{g}$ as follows

$$\hat{r}(\gamma)=-\int\log(g(\boldsymbol{\theta})+\gamma\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{101}$$

and is verified to be differentiable at zero in Lemma 28. This implies that

$$\begin{aligned}
r(\gamma)=&\gamma\int\hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)\,\mathrm{d}Q(\boldsymbol{\theta})+\lambda\hat{r}(\gamma)\\
&+\int g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta)\,\mathrm{d}Q(\boldsymbol{\theta})-\beta, && \text{(102)}
\end{aligned}$$

which verifies that the function r in (99) is differentiable at zero and more importantly, verifies that the Gateaux differential $\partial L(g,\beta;\hat{g})$ in (98) exists.

The proof proceeds by calculating the Gateaux differential $\partial L(g, \beta; \hat{g})$ in (98), which requires calculating the derivative of the real function r in (99). That is,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) = \int \hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\,\mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \frac{\hat{g}(\boldsymbol{\theta})}{(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))}\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{103}$$

$$= \int \hat{g}(\boldsymbol{\theta}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{(g(\boldsymbol{\theta}) + \gamma \hat{g}(\boldsymbol{\theta}))} \right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{104}$$

From (98) and (104), it follows that

$$\partial L(g, \beta; \hat{g}) = \int \hat{g}(\boldsymbol{\theta}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{g(\boldsymbol{\theta})} \right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{105}$$

The relevance of the Gateaux differential in (105) stems from [54, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (96) to have a stationary point at $\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta \right) \in \mathcal{M} \times [0, \infty)$ is that for all functions $\hat{g} \in \mathscr{S}$,

$$\partial L \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta; \hat{g} \right) = 0. \tag{106}$$

From (105) and (106), 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 $\mathscr{S}$ that

$$\int \hat{g}(\boldsymbol{\theta}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \lambda \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \right) \mathrm{d}Q(\boldsymbol{\theta}) = 0. \tag{107}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \lambda \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} = 0, \tag{108}$$

and thus,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{109}$$

where β is chosen to satisfy (95b) and guarantee that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \in (0, \infty)$. That is,

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \frac{\lambda}{t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \right\}, \text{ and} \tag{110}$$

$$1 = \int \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{111}$$

which is an assumption of the theorem.

The proof continues by verifying that the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that satisfies (109) is the unique solution to the optimization problem in (94). Such verification is done by showing that the objective function in (94) is strictly convex with the optimization variable. Let P_1 and P_2 be two different probability measures in $(\mathcal{M}, \mathscr{F})$ and let α be in $(0,1)$. Hence,

$$\begin{aligned} &\mathsf{R}_{\boldsymbol{z}}(\alpha P_1 + (1-\alpha)P_2) + \lambda \mathsf{D}(\alpha P_1 + (1-\alpha)P_2 \| Q) \\ &= \mathsf{R}_{\boldsymbol{z}}(\alpha P_1) + \mathsf{R}_{\boldsymbol{z}}((1-\alpha)P_2) \\ &\quad + \lambda \mathsf{D}(\alpha P_1 + (1-\alpha)P_2 \| Q) \end{aligned} \tag{112}$$

$$\begin{aligned} &> \alpha \mathsf{R}_{\boldsymbol{z}}(P_1) + (1-\alpha)\mathsf{R}_{\boldsymbol{z}}(P_2) \\ &\quad + \lambda(\alpha \mathsf{D}(P_1 \| Q) + (1-\alpha)\mathsf{D}(P_2 \| Q)), \end{aligned} \tag{113}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (6). The equality above follows from the properties of the Lebesgue integral, while the inequality follows from [29, Theorem 2]. This proves that the solution is unique due to the strict concavity of the objective function, which completes the proof. ■

Appendix C
Proof of Lemma 3

Given a probability measure $V \in \triangledown_Q(\mathcal{M})$, with $\triangledown_Q(\mathcal{M})$ in (12), it follows that

$$\operatorname{supp} Q \subseteq \operatorname{supp} V. \tag{114}$$

Using this observation, let V_0 and V_1 be two probability measures on the measurable space $(\mathcal{M}, \mathscr{F})$ such that for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$V_0(\mathcal{A}) = \frac{V(\mathcal{A} \setminus \operatorname{supp} Q)}{V(\mathcal{M} \setminus \operatorname{supp} Q)}, \tag{115a}$$

and

$$V_1(\mathcal{A}) = \frac{V(\mathcal{A} \cap \operatorname{supp} Q)}{V(\mathcal{M} \cap \operatorname{supp} Q)}. \tag{115b}$$

Let the real value α be

$$\alpha \triangleq V(\mathcal{M} \cap \operatorname{supp} Q) \in (0, 1], \tag{116}$$

which implies that

$$1 - \alpha \triangleq V(\mathcal{M} \setminus \operatorname{supp} Q) \in [0, 1). \tag{117}$$

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

$$V(\mathcal{A}) = (1-\alpha)V_0(\mathcal{A}) + \alpha V_1(\mathcal{A}). \tag{118}$$

Moreover, from the definition of $\triangledown_Q(\mathcal{M})$ in (12), it follows that the probability measure Q is absolutely continuous with respect to V. Hence, for all measurable sets $\mathcal{A}$, it follows that

$$Q(\mathcal{A}) = \int_{\mathcal{A}} \mathrm{d}Q(\boldsymbol{\theta}) \tag{119a}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\,\mathrm{d}V(\boldsymbol{\theta}) \tag{119b}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\,\mathrm{d}((1-\alpha)V_0 + \alpha V_1)(\boldsymbol{\theta}) \tag{119c}$$

$$\begin{aligned} &= (1-\alpha) \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\,\mathrm{d}V_0(\boldsymbol{\theta}) \\ &\quad + \alpha \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\,\mathrm{d}V_1(\boldsymbol{\theta}). \end{aligned} \tag{119d}$$

Note that if $\mathcal{A} \subseteq \operatorname{supp} Q$, it follows that

$$0 < Q(\mathcal{A}) = \int_{\mathcal{A}} \alpha \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\,\mathrm{d}V_1(\boldsymbol{\theta}), \tag{120}$$

which follows from the fact that $V_0(\mathcal{A}) = 0$ under the assumption that $\mathcal{A} \subseteq \operatorname{supp} Q$. Moreover, noting that the measures Q and V_1 are mutually absolutely continuous, it follows from the Radon-Nikodym Theorem [53, Theorem 2.2.1], that the Radon-Nikodym derivative of Q with respect to V_1 exists. Moreover, from (120), it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta}) = \alpha \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta}). \tag{121}$$

Alternatively, for all measurable sets $\mathcal{A}$, with $\mathcal{A} \subseteq \mathcal{M} \backslash \operatorname{supp} Q$, it follows that

$$0 = Q(\mathcal{A}) = \int_{\mathcal{A}} (1-\alpha) \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta}) \, \mathrm{d}V_0(\boldsymbol{\theta}), \tag{122}$$

which follows from the fact that $V_1(\mathcal{A}) = 0$ under the assumption that $\mathcal{A} \subseteq \mathcal{M} \setminus \operatorname{supp} Q$. Hence, it holds that for all $\boldsymbol{\theta} \in \mathcal{M} \backslash \operatorname{supp} Q$, $\frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta}) = 0$. In a nutshell, for all $\boldsymbol{\theta} \in \operatorname{supp} V$,

$$\frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta}) = \begin{cases} \frac{1}{\alpha} \frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta}) & \text{if } \boldsymbol{\theta} \in \operatorname{supp} Q \\ 0 & \text{otherwise.} \end{cases} \tag{123}$$

From (123), the following holds:

$$\begin{aligned} \mathsf{D}(Q \| V) &= \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(124a)} \\ &= \int \log\left(\frac{1}{\alpha} \frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(124b)} \\ &= \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \\ &\quad - \int \log(\alpha) \, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(124c)} \\ &= \mathsf{D}(Q \| V_1) - \log(\alpha). && \text{(124d)} \end{aligned}$$

From (124), it follows that

$$\begin{aligned} &\mathsf{R}_{\boldsymbol{z}}(V) + \lambda \mathsf{D}(Q \| V) \\ &\quad = \mathsf{R}_{\boldsymbol{z}}((1-\alpha)V_0 + \alpha V_1) + \lambda \mathsf{D}(Q \| V_1) \\ &\qquad - \lambda \log(\alpha) && \text{(125a)} \\ &\quad = (1-\alpha)\mathsf{R}_{\boldsymbol{z}}(V_0) + \alpha \mathsf{R}_{\boldsymbol{z}}(V_1) + \lambda \mathsf{D}(Q \| V_1) \\ &\qquad - \lambda \log(\alpha) && \text{(125b)} \\ &\quad \geq \alpha \mathsf{R}_{\boldsymbol{z}}(V_1) + \lambda \mathsf{D}(Q \| V_1), && \text{(125c)} \end{aligned}$$

with equality if and only if $\alpha = 1$, which implies that for all

$$P^\star \in \arg \min_{V \in \triangledown_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(V) + \lambda \mathsf{D}(Q \| V), \tag{126}$$

it holds that $P^\star \in \bigcirc_Q(\mathcal{M})$, which follows from observing that

$$\operatorname{supp} Q = \operatorname{supp} P^\star, \tag{127}$$

which implies (19) and completes the proof ■

Appendix D
Proof of Lemma 5

The proof is divided into two parts. The first part proves the monotonicity of the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (26); while the second part proves the continuity of the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$. The proof is finalized by using the continuous inverse theorem [55, Theorem 5.6.5] to show both the monotonicity and continuity of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23).

The first part is as follows. Let λ and β be two reals that satisfy (23b). Hence, $0 < \lambda < \infty$ and from (25b), it holds that

$$0 < \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{128}$$

which, together with (26), imply

$$\infty > \bar{K}_{Q,\boldsymbol{z}}^{-1}(\beta) > 0. \tag{129}$$

That is, the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (26) is positive and finite. Using this observation, let the reals γ_1 and γ_2 be elements of the set $\mathcal{C}_{Q,\boldsymbol{z}}$, with $\mathcal{C}_{Q,\boldsymbol{z}}$ in (23a) and $\gamma_1 < \gamma_2$. Hence, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_1} > \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_2}, \tag{130}$$

which implies that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_1} \, \mathrm{d}Q(\boldsymbol{\theta}) > \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_2} \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{131}$$

and thus, from (26), it holds that

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

This proves that the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (26) is strictly increasing, and completes the first part of the proof.

In the second part, the objective is to prove the continuity of the function $\bar{K}_{Q,\boldsymbol{z}}^{-1}$. To do so, two auxiliary functions are introduced and proven to be continuous. Then, the fact that $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (26) is the composition of the two auxiliary functions is leveraged to prove its continuity. Let the function $h : (0, \infty) \to (0, \infty)$ be

$$h(x) = \frac{1}{x}. \tag{133}$$

Let also the function $k : \mathcal{C}_{Q,\boldsymbol{z}} \to (0, \infty)$, be such that

$$k(\gamma) {=} \int h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}). \tag{134}$$

The first step is to prove that the function k in (134) is continuous in $\mathcal{C}_{Q,\boldsymbol{z}}$. This is proved by showing that k always exhibits a limit in $\mathcal{C}_{Q,\boldsymbol{z}}$. Note that if $\gamma \in \mathcal{C}_{Q,\boldsymbol{z}}$, with $\mathcal{C}_{Q,\boldsymbol{z}}$ in (23), then from (15a), it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, the inequality $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma > 0$ holds, which implies that $\gamma > -\delta_{Q,\boldsymbol{z}}^\star$, with $\delta_{Q,\boldsymbol{z}}^\star$ in (28). Hence, the proof of continuity of the function k in (134) is restricted to $\left(-\delta_{Q,\boldsymbol{z}}^\star, \infty\right)$.

For two models $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\operatorname{supp} Q$, such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)$, the function h satisfies

$$h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)) {>} h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)). \tag{135}$$

Then, for all $\gamma \in \left(-\delta_{Q,\boldsymbol{z}}^\star, \infty\right)$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$h\left(\gamma + \delta_{Q,\boldsymbol{z}}^\star\right) {\geq} h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})), \tag{136}$$

where equality holds if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta_{Q,\boldsymbol{z}}^\star$. The function h is continuous, and thus, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ and for all $a \in \left(-\delta_{Q,\boldsymbol{z}}^\star, \infty\right)$, it holds that

$$\lim_{\gamma \to a} h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) {=} h(a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})). \tag{137}$$

Hence, from the dominated convergence theorem [53, Theorem 1.6.9], the following limit exists and satisfies

$$\lim_{\gamma \to a} k(b) = \lim_{\gamma \to a} \int h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{138a}$$
$$= \int \left(\lim_{\gamma \to a} h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{138b}$$
$$= \int h(a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{138c}$$
$$= k(a), \tag{138d}$$

where (138c) follows from (137). The equality in (138d) proves that the function k in (134) is continuous in the interval $\left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. Note that from (26) and (134), it holds that

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma) = \frac{1}{k(\gamma)}. \tag{139}$$

Using (139), for all $a \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, it holds that

$$\lim_{\gamma \to a} \bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma) = \lim_{\gamma \to a} \frac{1}{k(\gamma)} \tag{140}$$
$$= \frac{1}{\lim_{\gamma \to a} k(\gamma)} \tag{141}$$
$$= \frac{1}{\int h(a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta})} \tag{142}$$
$$= \frac{1}{\int \frac{1}{a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\, \mathrm{d}Q(\boldsymbol{\theta})} \tag{143}$$
$$= \bar{K}^{-1}_{Q,\boldsymbol{z}}(a), \tag{144}$$

where (141) follows from the continuity of the function h in (133) over the interval $(0, \infty)$; (142) follows from (138d); and (143) follows from (26). Thus, the existence of the limit in (140) implies that the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ is continuous in $\mathcal{C}_{Q,\boldsymbol{z}}$. This completes the second part of the proof.

The proof ends by using the continuous inverse theorem [55, Theorem 5.6.5]. That is, given that the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ is both continuous and strictly increasing, then, so is the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23). This concludes the proof. ■

Appendix E
Proof of Lemma 6

The proof is divided into two parts. In the first part, it is shown that the set $\mathcal{C}_{Q,\boldsymbol{z}}$ is an interval of $\mathbb{R}$. In the second part, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is shown to be also an interval. The first part uses a partition of $\mathbb{R}$ formed by the following sets: $\left(-\infty, -\delta^{\star}_{Q,\boldsymbol{z}}\right)$; $\left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$; and $\{\delta^{\star}_{Q,\boldsymbol{z}}\}$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28). Each of these intervals is studied separately.

Let β be such that $\bar{K}_{Q,\boldsymbol{z}}(\lambda) = \beta$, with λ in (11) and assume that $\beta \in \left(-\infty, -\delta^{\star}_{Q,\boldsymbol{z}}\right)$. Under this assumption, the inclusion in (15a) does not hold. This follows from the fact that, if $\beta < -\delta^{\star}_{Q,\boldsymbol{z}}$, for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < -\beta\right\}$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta < 0$, which contradicts (15a). This implies that

$$\left(-\infty, -\delta^{\star}_{Q,\boldsymbol{z}}\right) \cap \mathcal{C}_{Q,\boldsymbol{z}} = \emptyset. \tag{145}$$

Assume now that $\beta \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. Then, from (28), it can be verified that the constraint in (15a) is satisfied. More specifically, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta > 0$. The proof continuous by showing that (15b) is also verified. For this purpose, note that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\, \mathrm{d}Q(\boldsymbol{\theta}) \leq \int \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}} + \beta}\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{146a}$$
$$= \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}} + \beta} \tag{146b}$$
$$< \infty. \tag{146c}$$

The finiteness of the integral in the left-hand side of (146a) implies that

$$0 < \lambda \tag{147a}$$
$$= \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) \tag{147b}$$
$$= \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\, \mathrm{d}Q(\boldsymbol{\theta})} \tag{147c}$$
$$< \infty, \tag{147d}$$

where (147a) follows from the assumption that $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0, \infty)$; (147b) follows from the fact that $\bar{K}_{Q,\boldsymbol{z}}(\lambda) = \beta$; (147c) follows from (26); and (147d) follows from the inequality in (146c). In a nutshell,

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\, \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \text{ and} \tag{148}$$
$$\lambda < \infty, \tag{149}$$

which, implies that the product

$$\lambda \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\, \mathrm{d}Q(\boldsymbol{\theta}) = \int \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{150a}$$
$$= \int \frac{\frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \beta}\, \mathrm{d}Q(\boldsymbol{\nu})}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{150b}$$
$$= 1, \tag{150c}$$

where (150b) follows from (147c). This verifies (15b), which implies that

$$\left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right) \subseteq \mathcal{C}_{Q,\boldsymbol{z}}. \tag{151}$$

Finally, under the assumption that $\beta = -\delta^{\star}_{Q,\boldsymbol{z}}$, two cases are considered: (a) $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) > 0$; and (b) $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ defined in (30). In case (a), if $\beta = -\delta^{\star}_{Q,\boldsymbol{z}}$ and $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) > 0$, then for all $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, it follows that

$$\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = \infty, \tag{152}$$

which implies that,

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\, \mathrm{d}Q(\boldsymbol{\theta}) = \infty. \tag{153}$$

The equality in (153) implies that the constraint (15b) is not satisfied. Therefore, for case (a) it follows that,

$$-\delta^{\star}_{Q,\boldsymbol{z}} \notin \mathcal{C}_{Q,\boldsymbol{z}}. \tag{154}$$

In the alternative case (b), if $\beta = -\delta^{\star}_{Q,\boldsymbol{z}}$ and $Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0$, then, the integral in (26) is either

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}\, \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{155}$$

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

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) = \infty, \tag{156}$$

which implies that $-\delta^{\star}_{Q,\boldsymbol{z}} \notin \mathcal{C}_{Q,\boldsymbol{z}}$. Hence, from (145), (151), (154), (155), and (156) the set $\mathcal{C}_{Q,\boldsymbol{z}}$ in (23a) is either the open set $\left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$ or the closed set $\left[-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. Note that the equality $\mathcal{C}_{Q,\boldsymbol{z}} = \left[-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$ is observed, if and only if,

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{157}$$

which completes the first part of the proof.

The second part of the proof is as follows. Two cases are considered: *i)* $\mathcal{C}_{Q,\boldsymbol{z}} = \left[-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$; and *ii)* $\mathcal{C}_{Q,\boldsymbol{z}} = \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. In case *i)*, the value $-\delta^{\star}_{Q,\boldsymbol{z}}$ is in the domain of the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (26), that is, the set $\mathcal{C}_{Q,\boldsymbol{z}}$. Given that the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ is strictly increasing, then, $-\delta^{\star}_{Q,\boldsymbol{z}}$ should be mapped to the smallest value in the range of $\bar{K}^{-1}_{Q,\boldsymbol{z}}$, denoted by $\lambda_{Q,\boldsymbol{z}}$. Hence,

$$\begin{aligned} \bar{K}^{-1}_{Q,\boldsymbol{z}}\left(-\delta^{\star}_{Q,\boldsymbol{z}}\right) &= \lambda_{Q,\boldsymbol{z}} \qquad (158)\\ &> 0, \qquad (159) \end{aligned}$$

where (159) follows from the fact that zero is not in the domain of the function $\bar{K}_{Q,\boldsymbol{z}}$, that is, the set $\mathcal{A}_{Q,\boldsymbol{z}}$. Using these elements, it is concluded that the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is the interval $[\lambda_{Q,\boldsymbol{z}}, \infty)$, which ends the analysis of case *i.)*.

In case *ii)*, from Lemma 5, the continuity and strict monotonicity of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) imply that $\mathcal{A}_{Q,\boldsymbol{z}} = \left(\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, with $\lambda^{\star}_{Q,\boldsymbol{z}}$ in (29). The remaining of the proof focuses on showing that $\lambda^{\star}_{Q,\boldsymbol{z}} = 0$ in this case, and thus, $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$. From Lemma 5 and the continuous inverse theorem [55, Theorem 5.6.5], it follows that function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ is strictly increasing and continuous. Hence, using (26), it holds that

$$\begin{aligned} \lim_{\gamma \to -\delta^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma) &= \lim_{\gamma \to -\delta^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma} \mathrm{d}Q(\boldsymbol{\theta})} \qquad (160)\\ &= \frac{1}{\lim_{\gamma \to -\delta^{\star}_{Q,\boldsymbol{z}}{}^{+}} \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma} \mathrm{d}Q(\boldsymbol{\theta})} \qquad (161)\\ &= \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta})} \qquad (162)\\ &= 0, \qquad (163) \end{aligned}$$

where (161) follows from [56, Theorem 4.4], that permits the change of the limit to the reciprocal; and (162) follows from (138); and (163) follows from (156). From Lemma 5 and (163), it follows that in this second case, in which $\mathcal{C}_{Q,\boldsymbol{z}} = \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, it holds that $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$. This completes the proof. ■

Appendix F
Proof of Lemma 10

From Lemma 5, the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is strictly increasing and continuous. Additionally, from Lemma 6, the domain and range of the function $\bar{K}_{Q,\boldsymbol{z}}$, defined by the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$, respectively, are convex intervals. Consequently, combining Lemma 5 and Lemma 6, it follows that

$$\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\delta^{\star}_{Q,\boldsymbol{z}}, \tag{164}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ defined in (28) and $\lambda^{\star}_{Q,\boldsymbol{z}}$ defined in (29). ■

Appendix G
Proof of Lemma 11

For all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in (\operatorname{supp} Q)^2$, such that

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) \leq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2), \tag{165}$$

it follows that

$$\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)} \geq \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)}, \tag{166a}$$

where the function $\bar{K}_{Q,\boldsymbol{z}}$ is defined in (23); and equality holds if and only if (165) holds with equality. The proof is completed by noticing that from (24), the inequality above can be rewritten as

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1) \geq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2), \tag{167}$$

which completes the proof. ■

Appendix H
Proof of Lemma 12

From Lemma 11, it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, and for all $\boldsymbol{\phi} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$, it holds that

$$\begin{aligned} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &\leq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\phi}) \qquad (168a)\\ &= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \qquad (168b)\\ &\leqslant \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \qquad (168c)\\ &< \infty, \qquad (168d) \end{aligned}$$

where (168b) follows from (24); the equality in (168c) follows from the fact that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi}) \geq \delta^{\star}_{Q,\boldsymbol{z}}$; and (168d) follows from the fact that $\bar{K}_{Q,\boldsymbol{z}}(\lambda) < \infty$. Note that equalities in (168a) and (168c) hold if and only if $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$ (Lemma 10). This completes the proof of finiteness.

For the proof of positivity, observe that from Lemma 6, it holds that

$$-\delta^{\star}_{Q,\boldsymbol{z}} < \bar{K}_{Q,\boldsymbol{z}}(\lambda) < \infty, \tag{169}$$

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

$$\begin{aligned} 0 &< \delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda) \qquad (170)\\ &\leqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda). \qquad (171) \end{aligned}$$

Hence, from the fact that $\lambda > 0$, it holds from (24) and (171) that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) > 0, \tag{172a}$$

which completes the proof. ■

APPENDIX I
PROOF OF LEMMA 13

From Theorem 1, the Radon-Nikodym derivative of the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to Q, satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}
&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \qquad (173a)\\
&= \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \frac{1}{\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu})} \qquad (173b)\\
&= \frac{1}{\int \frac{\beta}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) + \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu})}, \qquad (173c)
\end{aligned}$$

where (173b) follows from (26), which implies that

$$\lambda = \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta} \mathrm{d}Q(\boldsymbol{\theta})}, \qquad (174)$$

with the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ being the inverse of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23). Using the function $\bar{K}_{Q,\boldsymbol{z}}$, the equation in (173) can be written in term of λ such that

$$\begin{aligned}
&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \frac{1}{\int \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) + \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu})}. \qquad (175)
\end{aligned}$$

Furthermore, using Lemma 5 and Lemma 6, the following holds from (175),

$$\begin{aligned}
&\lim_{\lambda\to\infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \lim_{\lambda\to\infty} \frac{1}{\int \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) + \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu})} \qquad (176a)\\
&= \frac{1}{\lim_{\beta\to\infty} \int \frac{\beta}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) + \lim_{\beta\to\infty} \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu})}, \qquad (176b)
\end{aligned}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3); and (176b) follows from Theorem 1, which implies that the terms in the denominator are positive and the fact that the function $g(x) = \frac{1}{x}$ is continuous. Recall that from the definition of the function $\mathsf{L}_{\boldsymbol{z}}$ in (3), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, the empirical risk satisfies that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \infty$. Using this fact, the proof continues by evaluating the limits in the denominator, which yields

$$\begin{aligned}
&\lim_{\beta\to\infty} \int \frac{\beta}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) \\
&= \int \lim_{\beta\to\infty} \frac{\beta}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) \qquad (177a)\\
&= \int \mathrm{d}Q(\boldsymbol{\nu}) \qquad (177b)\\
&= 1, \qquad (177c)
\end{aligned}$$

where (177a) follows from the dominated convergence theorem [53, Theorem 1.6.9]; and,

$$\begin{aligned}
&\lim_{\beta\to\infty} \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) \\
&= \int \lim_{\beta\to\infty} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\nu}) \qquad (178a)\\
&= \int 0 \, \mathrm{d}Q(\boldsymbol{\nu}) \qquad (178b)\\
&= 0, \qquad (178c)
\end{aligned}$$

where (178a) also follows from the dominated convergence theorem [53, Theorem 1.6.9]. Substituting (177) and (178) into (176) yields

$$\lim_{\lambda\to\infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 1, \qquad (179)$$

which completes the proof. ■

APPENDIX J
PROOF OF LEMMA 14

From Theorem 1, the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}
\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \qquad (180a)\\
&= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \qquad (180b)\\
&= \frac{\bar{K}^{-1}_{Q,\boldsymbol{z}}\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda)\big)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \qquad (180c)\\
&= \frac{\frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu})}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \qquad (180d)\\
&= \left(\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1}, \qquad (180e)
\end{aligned}$$

where (180b) follows from (23); and (180d) follows from (26) and observing that $\lambda = \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)$. Given $\boldsymbol{\theta} \in \operatorname{supp} Q$, consider the partition of the $\operatorname{supp} Q$ formed by the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$, which satisfy the following:

$$\begin{aligned}
\mathcal{A}_0(\boldsymbol{\theta}) &= \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = 0\}, \qquad (181a)\\
\mathcal{A}_1(\boldsymbol{\theta}) &= \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < 0\}, \text{ and} \qquad (181b)\\
\mathcal{A}_2(\boldsymbol{\theta}) &= \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > 0\}. \qquad (181c)
\end{aligned}$$

Using the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (181), the following holds for all $\boldsymbol{\theta} \in \operatorname{supp} Q$.

$$\begin{aligned}
&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\
&= \Bigg(\int_{\mathcal{A}_0(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \\
&\quad + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \\
&\quad + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \Bigg)^{-1} \qquad (182a)
\end{aligned}$$

$$= \left(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right.$$
$$\left. + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1}. \tag{182b}$$

Consider the following partition of the $\operatorname{supp} Q$:

$$\left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}} \right\}, \tag{183a}$$
$$\left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}} \right\}, \text{ and} \tag{183b}$$
$$\left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^{\star}_{Q,\boldsymbol{z}} \right\}, \tag{183c}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28). The proof is divided into two cases. The first case follows under the assumption that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) = \infty; \tag{184}$$

and the second case follows under the assumption that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{185}$$

From Lemma 6, it follows that in Case 1, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (23) is $(0,\infty)$. Similarly, in Case 2, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is $\left[\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. Hence, Case 1 considers the limit $\lambda \to 0^{+}$, which comprehends the equalities (39) and (40). Case 2 considers the limit $\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}$, which comprehends the equality (41).

A. *Case* 1

This case is divided into three parts. The first part evaluates $\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}} \right\}$. The second part considers the case in which $\boldsymbol{\theta} \in \left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}} \right\}$. The third part considers the remaining case in (183).

1) Part 1*:* The first part is as follows. Consider that $\boldsymbol{\theta} \in \left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}} \right\}$ and note that

$$\left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}} \right\} = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}, \tag{186}$$

with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ defined in (30). Hence, the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (181) satisfy the following:

$$\mathcal{A}_0(\boldsymbol{\theta}) = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}, \tag{187a}$$
$$\mathcal{A}_1(\boldsymbol{\theta}) = \left\{ \boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta^{\star}_{Q,\boldsymbol{z}} \right\}, \text{ and} \tag{187b}$$
$$\mathcal{A}_2(\boldsymbol{\theta}) = \left\{ \boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^{\star}_{Q,\boldsymbol{z}} \right\}. \tag{187c}$$

From the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), it follows that $Q(\mathcal{A}_2(\boldsymbol{\theta})) = 0$. Substituting the equalities in (187) in (182) yields for all $\boldsymbol{\theta} \in \left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}} \right\}$,

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$$
$$= \left(Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1}, \tag{188}$$

which implies that for all $\boldsymbol{\theta} \in \left\{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}} \right\}$,

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$$
$$= \left(Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) + \lim_{\lambda \to 0^{+}} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} \tag{189}$$
$$= \begin{cases} \infty & \text{if } Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0 \\ \frac{1}{Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right)} & \text{otherwise} \end{cases}, \tag{190}$$

where (190) follows from verifying that the dominated convergence theorem [53, Theorem 2.6.9] holds. That is,
(a) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that $\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \leq \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}$.
(b) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that

$$\lim_{\lambda \to 0^{+}} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}$$
$$= \lim_{\lambda \to 0^{+}} \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{191a}$$
$$= \left(\delta^{\star}_{Q,\boldsymbol{z}} + \lim_{\lambda \to 0^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) \right) \lim_{\lambda \to 0^{+}} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{191b}$$
$$= 0, \tag{191c}$$

where (191b) follows from observing that for all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that $\lim_{\lambda \to 0^{+}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda) \neq 0$ and [56, Theorem 4.4]; and (191c) follows from Lemma 10. This completes the first part of Case 1.

2) Part 2*:* For all $\delta > \delta^{\star}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \{ \boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta \}$, the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (181) satisfy the following:

$$\mathcal{A}_0(\boldsymbol{\theta}) = \{ \boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = \delta \}, \tag{192a}$$
$$\mathcal{A}_1(\boldsymbol{\theta}) = \{ \boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta \}, \text{ and} \tag{192b}$$
$$\mathcal{A}_2(\boldsymbol{\theta}) = \{ \boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta \}. \tag{192c}$$

Consider the sets

$$\mathcal{A}_{2,1}(\boldsymbol{\theta}) = \left\{ \boldsymbol{\mu} \in \mathcal{A}_2(\boldsymbol{\theta}) : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^{\star}_{Q,\boldsymbol{z}} \right\}, \text{ and} \tag{193a}$$
$$\mathcal{A}_{2,2}(\boldsymbol{\theta}) = \left\{ \boldsymbol{\mu} \in \mathcal{A}_2(\boldsymbol{\theta}) : \delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta \right\}, \tag{193b}$$

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 (28), it holds that

$$Q(\mathcal{A}_{2,1}(\boldsymbol{\theta})) = 0. \tag{194}$$

Hence, substituting the equalities in (192) and (194) in (182) yields,

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$$
$$= \left(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \lim_{\lambda \to 0^{+}} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right.$$
$$\left. + \lim_{\lambda \to 0^{+}} \int_{\mathcal{A}_2(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} \tag{195a}$$
$$= \left(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \lim_{\lambda \to 0^{+}} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu}) \right.$$

$$+ \lim_{\lambda\to 0^+}\int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1} \tag{195b}$$

$$= \Bigg(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \lim_{\lambda\to 0^+} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\mathrm{d}Q(\boldsymbol{\nu})$$

$$+ \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \lim_{\lambda\to 0^+} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1}, \tag{195c}$$

where (195c) follows by verifying that the dominated convergence theorem [53, Theorem 1.6.9] holds. That is,
(a) For all $\boldsymbol{\nu} \in \mathcal{A}_{2,2}(\boldsymbol{\theta})$, it holds that $\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} \leq \frac{\lambda}{\delta^\star_{Q,\boldsymbol{z}}+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} < \infty$; and
(b) For all $\boldsymbol{\nu} \in \mathcal{A}_{2,2}(\boldsymbol{\theta})$, it holds that

$$\lim_{\lambda\to 0^+} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}$$

$$= \lim_{\lambda\to 0^+} \frac{\delta+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{196a}$$

$$= \left(\delta + \lim_{\lambda\to 0^+}\bar{K}_{Q,\boldsymbol{z}}(\lambda)\right)\lim_{\lambda\to 0^+}\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{196b}$$

$$= \left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right)\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}, \tag{196c}$$

where (196b) follows from observing that for all $\boldsymbol{\nu} \in \mathcal{A}_{2,2}(\boldsymbol{\theta})$, it holds that $\lim_{\lambda\to 0^+}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda) \neq 0$ and [56, Theorem 4.2]; and (196c) follows from Lemma 10. From (196c), it follows that

$$\int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \lim_{\lambda\to 0^+} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\mathrm{d}Q(\boldsymbol{\nu})$$

$$= \left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right)\int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu}). \tag{197}$$

Moreover, from the fact that

$$Q(\mathcal{A}_0(\boldsymbol{\theta})) \leq 1, \tag{198}$$

and the fact that

$$\int_{\mathcal{A}_1(\boldsymbol{\theta})} \lim_{\lambda\to 0^+} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\bar{K}_{Q,\boldsymbol{z}}(\lambda)}\mathrm{d}Q(\boldsymbol{\nu})$$

$$= \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\delta-\delta^\star_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu}) \tag{199}$$

$$< \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\delta-\delta^\star_{Q,\boldsymbol{z}}}{\delta-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu}) \tag{200}$$

$$= Q(\mathcal{A}_1(\boldsymbol{\theta})) \tag{201}$$

$$\leq 1, \tag{202}$$

the following holds from Lemma 6 under the assumptions of Case 1:

$$\infty = \left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right)\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu}) \tag{203}$$

$$= \left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right)\Bigg(\int_{\mathcal{A}_0(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu})$$

$$+ \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu})$$

$$+ \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu})\Bigg) \tag{204}$$

$$= \Bigg(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\delta-\delta^\star_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu})$$

$$+ \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{\delta-\delta^\star_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu})\Bigg). \tag{205}$$

From (198), (202), and (205), it follows that

$$\left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right)\int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})-\delta^\star_{Q,\boldsymbol{z}}}\mathrm{d}Q(\boldsymbol{\nu}) = \infty. \tag{206}$$

Finally, from (195c), (197), and (206), for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star_{Q,\boldsymbol{z}}\right\}$,

$$\lim_{\lambda\to 0^+}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{207}$$

This completes the second part of Case 1.

3) Part 3: The third part of the proof follows by noticing that the set $\left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star_{Q,\boldsymbol{z}}\right\}$ is a negligible set with respect to Q and thus, for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star_{Q,\boldsymbol{z}}\right\}$, the value $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$ is immaterial. Hence, it is arbitrarily assumed that for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star_{Q,\boldsymbol{z}}\right\}$, it holds that

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

which implies that for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star_{Q,\boldsymbol{z}}\right\}$, it holds that

$$\lim_{\lambda\to 0^+}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{209}$$

This completes the third part of Case 1.

Under the assumption that $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) > 0$, from (190), (207), and (209), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\lim_{\lambda\to 0^+}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q\Big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\Big)}\mathbb{1}_{\left\{\boldsymbol{\theta}\in\mathcal{L}^\star_{Q,\boldsymbol{z}}\right\}}, \tag{210}$$

which completes the proof of (39). Alternatively, under the assumption that $Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) = 0$, from (190), (207), and (209), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\lim_{\lambda\to 0^+}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \begin{cases}\infty & \text{if } \boldsymbol{\theta} \in \mathcal{L}^\star_{Q,\boldsymbol{z}}\\ 0 & \text{otherwise}\end{cases}, \tag{211}$$

which completes the proof of (40).

B. Case 2

Under the assumptions of Case 2, namely (185), it holds that

$$Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) = 0. \tag{212}$$

This can be proved by noticing that if $Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) > 0$, then

$$\begin{aligned}
&\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) \\
&= \int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) \\
&\quad + \int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}{}^{\mathsf{c}}} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) &\text{(213a)}\\
&> \int_{\mathcal{L}^{\star}_{Q,\boldsymbol{z}}} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) &\text{(213b)}\\
&= \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}} - \delta^{\star}_{Q,\boldsymbol{z}}} Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) &\text{(213c)}\\
&= \infty, &\text{(213d)}
\end{aligned}$$

which contradicts (212).

The proof of Case 2 is divided into three parts. The first part evaluates $\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \big\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\big\}$. The second part considers the case in which $\boldsymbol{\theta} \in \big\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}}\big\}$. The third part considers the remaining case in (183).

1) Part 1: From (212) it holds that the set the set $\big\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = \delta^{\star}_{Q,\boldsymbol{z}}\big\}$ is a negligible set with respect to Q and thus, for all $\boldsymbol{\theta} \in \big\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = \delta^{\star}_{Q,\boldsymbol{z}}\big\}$, the value $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$ is immaterial.

2) Part 2: For all $\boldsymbol{\theta} \in \big\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}}\big\}$, it holds that

$$\begin{aligned}
\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} &\text{(214a)}\\
&= \frac{\lambda^{\star}_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}, &\text{(214b)}
\end{aligned}$$

where (214b) follows from observing that $\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda) = \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}} \neq 0$ (Lemma 10) and [56, Theorem 4.2].

3) Part 3: The third part of the proof follows by noticing that the set $\big\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^{\star}_{Q,\boldsymbol{z}}\big\}$ is a negligible set with respect to Q and thus, for all $\boldsymbol{\theta} \in \big\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^{\star}_{Q,\boldsymbol{z}}\big\}$, the value $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$ is immaterial. Hence, it is arbitrarily assumed that

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{215}$$

This completes the third part of Case 2.

From (214b), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda^{\star}_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}, \tag{216}$$

which completes the proof of (41). This completes the proof. ■

Appendix K
Proof of Lemma 15

The proof is divided into two cases. The first case follows under the assumption that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) = \infty; \tag{217}$$

and the second case follows under the assumption that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{218}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28) and the function $\mathsf{L}_{\boldsymbol{z}}$ in (3). From Lemma 6, it follows that in Case 1, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (23) is $(0,\infty)$. Similarly, in Case 2, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is $\big[\lambda^{\star}_{Q,\boldsymbol{z}},\infty\big)$. Hence, Case 1 considers the limit $\lambda \to 0^{+}$, which comprehends the equality (42). Case 2 considers the limit $\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}$, which comprehends the equality (43).

A. *Case* 1

The first case is as follows. Consider the following partition of the set $\mathcal{M}$ formed by the sets

$$\begin{aligned}
\mathcal{A}_0 &= \big\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\big\}, &\text{(219a)}\\
\mathcal{A}_1 &= \big\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > \delta^{\star}_{Q,\boldsymbol{z}}\big\}, \text{ and} &\text{(219b)}\\
\mathcal{A}_2 &= \big\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^{\star}_{Q,\boldsymbol{z}}\big\}. &\text{(219c)}
\end{aligned}$$

Note that $\mathcal{A}_0 = \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (30). For all $\lambda \in (0,\infty)$, it holds that

$$\begin{aligned}
1 &= \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) &\text{(220a)}\\
&= \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) &\text{(220b)}\\
&= \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_1} \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), &\text{(220c)}
\end{aligned}$$

where (220b) follows from the fact that $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) = 0$, which follows from the mutual absolute continuity of $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q (Corollary 4). The above implies that

$$\begin{aligned}
&\lim_{\lambda \to 0^{+}} \left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_1} \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) \\
&\quad = \lim_{\lambda \to 0^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) \\
&\qquad + \lim_{\lambda \to 0^{+}} \int_{\mathcal{A}_1} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \,\mathrm{d}Q(\boldsymbol{\theta}) &\text{(221a)}\\
&\quad = \lim_{\lambda \to 0^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) \\
&\qquad + \int_{\mathcal{A}_1} \lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \,\mathrm{d}Q(\boldsymbol{\theta}) &\text{(221b)}\\
&\quad = \lim_{\lambda \to 0^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0), &\text{(221c)}\\
&\quad = 1, &\text{(221d)}
\end{aligned}$$

where, (221b) follows from Lemma 12 and the dominated convergence theorem [53, Theorem 1.6.9 page 50]; and (221c) follows from Lemma 14. Hence, it holds that

$$\lim_{\lambda \to 0^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 1, \tag{222}$$

which completes the proof of (42).

B. *Case* 2

Under the assumptions of Case 2, namely (218), it holds that

$$Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 0, \tag{223}$$

which can be shown using the arguments in (213). Hence, the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies

$$\begin{aligned}
&\lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) \\
&= \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \int_{\mathcal{A}_0} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && (224a)\\
&= \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \int_{\mathcal{A}_0} \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\theta}) && (224b)\\
&= \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \int_{\mathcal{A}_0} \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\theta}) && (224c)\\
&= \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} Q(\mathcal{A}_0) && (224d)\\
&= \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) && (224e)\\
&= \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} 0 && (224f)\\
&= 0, && (224g)
\end{aligned}$$

which completes the proof of (43). This completes the proof. ■

Appendix L
Proof of Lemma 16

From Lemma 2 and Corollary 4, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}
\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) &= \left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1} && (225)\\
&= \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}, && (226)
\end{aligned}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $\bar{K}_{Q,\boldsymbol{z}}$ are in (3) and (23b), respectively. From (226), it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$0 = \lambda \frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{227}$$

Integrating both sides of (227) with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ yields

$$\begin{aligned}
0 &= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \lambda \left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1} \right.\\
&\qquad \left. + \bar{K}_{Q,\boldsymbol{z}}(\lambda) \right) \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (228a)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\\
&\qquad -\lambda \int \left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1} \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\\
&\qquad + \int \bar{K}_{Q,\boldsymbol{z}}(\lambda)\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (228b)\\
&= \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \lambda \int \mathrm{d}Q(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda) && (228c)\\
&= \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \lambda + \bar{K}_{Q,\boldsymbol{z}}(\lambda). && (228d)
\end{aligned}$$

From (228d), it holds that

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{229}$$

which completes the proof. ■

Appendix M
Proof of Lemma 17

The proof of continuity is immediate from Lemma 5 and Lemma 16. The proof of monotonicity is divided into two parts. The first part presents the first derivative of the functional inverse $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (26) and shows that its derivative is strictly positive. The second part shows the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ is decreasing with λ.

The first part is as follows. For all $\boldsymbol{\theta} \in \mathcal{M}$, the partial derivative of $\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}$, with respect to $\beta \in \big(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\big)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (28), is

$$\frac{\partial}{\partial \beta}\left(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\right) = -\frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}. \tag{230}$$

From [57, Theorem 6.28, page 160], the following holds

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta}\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\,\mathrm{d}Q(\boldsymbol{\theta}) &= \int \frac{\partial}{\partial \beta} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\,\mathrm{d}Q(\boldsymbol{\theta}) && (231)\\
&= -\int \frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}\,\mathrm{d}Q(\boldsymbol{\theta}). && (232)
\end{aligned}$$

From Lemma 5, the derivative of the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (26) satisfies:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\beta}\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)\\
&= \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{-1} && (233a)\\
&= -\left(\frac{1}{\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})}\right)^2 \frac{\mathrm{d}}{\mathrm{d}\beta}\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}(\boldsymbol{\theta})}}\,\mathrm{d}Q(\boldsymbol{\theta}) && (233b)\\
&= -\frac{\int -\frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}\,\mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}(\boldsymbol{\theta})}}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2} && (233c)\\
&= \frac{\int \frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}\,\mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}(\boldsymbol{\theta})}}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2}, && (233d)
\end{aligned}$$

where (233c) follows from (232).

Jensen's inequality [58, Theorem 2.6.2] leads to the following inequality:

$$\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2 \leq \int \frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{234}$$

with equality if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ in (3) is nonseparable (Definition 2). Then, from (233d) and (234), for all $\beta \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, it holds that

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta}\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) &= \frac{\int \frac{1}{(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^{2}}\,\mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^{2}} \qquad (235)\\
&\geq 1. \qquad (236)
\end{aligned}
$$

This completes the first part of the proof.

The second part is as follows. Consider the pairs $(\lambda_1, \beta_1) \in \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{C}_{Q,\boldsymbol{z}}$ and $(\lambda_2, \beta_2) \in \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{C}_{Q,\boldsymbol{z}}$, such that $\lambda_2 > \lambda_1$, which implies that $\bar{K}_{Q,\boldsymbol{z}}(\lambda_2) > \bar{K}_{Q,\boldsymbol{z}}(\lambda_1)$ (Lemma 5). Then, from Lemma 16, it follows that

$$
\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\qquad = \lambda_2 - \lambda_1 + \bar{K}_{Q,\boldsymbol{z}}(\lambda_1) - \bar{K}_{Q,\boldsymbol{z}}(\lambda_2) \quad (237a)\\
&\qquad = \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_2) - \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_1) + \beta_1 - \beta_2, \quad (237b)
\end{aligned}
$$

where (237b) follows from substituting (26) into (237a). Note that (236) implies that

$$
\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_2) - \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_1) \geq \beta_2 - \beta_1, \tag{238}
$$

with equality if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ is nonseparable. Thus, from (237b) and (238) it follows that

$$
\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq 0, \tag{239}
$$

with equality if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ is nonseparable. This completes the second part of the proof. ■

Appendix N
Proof of Lemma 18

From Theorem 1 and Corollary 4, it holds that

$$
\begin{aligned}
&\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad = \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\right)\mathrm{d}Q(\boldsymbol{\theta}) \qquad (240)\\
&\quad \leq \log\left(\int \frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\right) \qquad (241)\\
&\quad = \log\left(\int \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\,\mathrm{d}Q(\boldsymbol{\theta})\right) \qquad (242)\\
&\quad = \log\left(\frac{1}{\lambda}\int \bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\right) \qquad (243)\\
&\quad = \log\left(\frac{1}{\lambda}\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{R}_{\boldsymbol{z}}(Q)\right)\right), \qquad (244)
\end{aligned}
$$

where (241) follows from Jensen's inequality [58, Theorem 2.6.2]; (242) follows from (16); and (244) follows from (6). From (244), it follows that

$$
\begin{aligned}
\mathsf{R}_{\boldsymbol{z}}(Q) &\geq \lambda \exp\left(\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right) - \bar{K}_{Q,\boldsymbol{z}}(\lambda) \qquad (245)\\
&= \lambda \exp\left(\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right) + \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \lambda, \qquad (246)
\end{aligned}
$$

where (246) follows from Lemma 16. Hence, the difference between the expected empirical risk of the probability measures $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and Q, from (246), satisfies that

$$
\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq \lambda\left(\exp\left(\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right) - 1\right), \tag{247}
$$

which completes the proof. ■

Appendix O
Proof of Lemma 20

From Lemma 5, Lemma 6, and Lemma 16, it holds that

$$
\begin{aligned}
\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) &= \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda) \qquad (248)\\
&< \lambda + \delta^{\star}_{Q,\boldsymbol{z}}. \qquad (249)
\end{aligned}
$$

Note also that

$$
\begin{aligned}
\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) &= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \qquad (250)\\
&\geq \int \delta^{\star}_{Q,\boldsymbol{z}}\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \qquad (251)\\
&= \delta^{\star}_{Q,\boldsymbol{z}}, \qquad (252)
\end{aligned}
$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (3). The proof continues by determining the conditions for which (252) holds with equality. Assume the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (3) is separable with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) (see Definition 2). Then, there exists a real value $\epsilon > 0$ and two nonnegligible sets $\mathcal{A}$ and $\mathcal{B}$ with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16), such that

$$
\begin{aligned}
\mathcal{A} &= \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\right\}, \text{and} \qquad (253)\\
\mathcal{B} &= \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geq \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\right\}. \qquad (254)
\end{aligned}
$$

Note that the sets $\mathcal{A}$ and $\mathcal{B}$ form a partition of the set $\mathcal{M}$. Hence, the expected empirical risk satisfies

$$
\begin{aligned}
\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) &= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \qquad (255)\\
&= \int_{\mathcal{A}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\\
&\quad + \int_{\mathcal{B}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \qquad (256)\\
&\geq \int_{\mathcal{A}} \delta^{\star}_{Q,\boldsymbol{z}}\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\\
&\quad + \int_{\mathcal{B}} \left(\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\right)\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \qquad (257)\\
&= \delta^{\star}_{Q,\boldsymbol{z}}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\\
&\quad + \left(\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\right)\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}), \qquad (258)\\
&= \delta^{\star}_{Q,\boldsymbol{z}}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\\
&\quad + \left(\delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\right)\left(1 - \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\right) \qquad (259)\\
&= \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\left(1 - \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\right) \qquad (260)\\
&> \delta^{\star}_{Q,\boldsymbol{z}}, \qquad (261)
\end{aligned}
$$

where inequality (257) follows from (28) and (254); (259) follows from the fact that the sets $\mathcal{A}$ and $\mathcal{B}$ form a partition of the set $\mathcal{M}$; and (261) follows from the fact that the sets $\mathcal{A}$ and

$\mathcal{B}$ are nonnegligible with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which implies $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) < 1$. This proves the strict inequality in (251).

Consider the case in which the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (3) is not separable with respect to $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16). Then, for all $\boldsymbol{\theta} \in \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, the empirical risk satisfies $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}$, which implies $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \delta^{\star}_{Q,\boldsymbol{z}}$. Hence, if the function $\mathsf{L}_{\boldsymbol{z}}$ is nonseparable, then (251) holds with equality. Therefore, (251) holds with equality, if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ is nonseparable, which completes the proof. ■

Appendix P
Proof of Theorem 2

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

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leq \delta, \tag{262}$$

where the existence of such a γ is ensured by the continuity of $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ with respect to γ (Lemma 17); and Lemma 20 and Lemma 21. From (27), it holds that

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \supseteq \mathcal{L}^{\star}_{Q,\boldsymbol{z}}, \tag{263}$$

and thus,

$$\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) \geq \bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right), \tag{264}$$

with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ defined in (30). Let λ be a positive real such that $\lambda \leq \gamma$, and

$$\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) > 1-\epsilon. \tag{265}$$

The existence of such a positive real λ follows from Lemma 15. From (263), it follows that

$$\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) \geq \bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right). \tag{266}$$

Hence, from (265) and (266), it holds that

$$1-\epsilon < \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) \tag{267}$$

$$\leq \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)). \tag{268}$$

The equality in (268) implies that the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ,ϵ)-optimal (Definition 3), which completes the proof.

■

Appendix Q
Proof of Lemma 23

The proof is divided into two parts. The first part is as follows, from Theorem 1, it follows that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) = \log\left(\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) \tag{269}$$

$$= \log(\lambda) - \log\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \tag{270}$$

$$= \log(\lambda) - \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}), \tag{271}$$

where the function $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ is defined in (57b). Thus,

$$\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) = \int \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{272}$$

$$= \log(\lambda) - \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{273}$$

$$= \log(\lambda) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{274}$$

where the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ is defined in (70). Hence, it follows from (274) that

$$\log(\lambda) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right), \tag{275}$$

which completes the proof of (72) and concludes the first part.

The second part is as follows. From (271), it follows that

$$\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = -\int \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{276}$$

$$= -\log(\lambda) + \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{277}$$

$$= -\log(\lambda) + \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q). \tag{278}$$

Hence, it follows from (278) that

$$\log(\lambda) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q) - \mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{279}$$

which completes the proof of (73). This completes the proof. ■

Appendix R
Proof of Lemma 24

The proof uses the mutual absolute continuity between $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) and Q (Corollary 4). Hence, a probability measure $P \in \bigcirc_Q(\mathcal{M})$ is mutually absolutely continuous with $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. The proof follows by noticing that for such P and for all $\boldsymbol{\theta} \in \mathcal{M}$, it holds that

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

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \tag{280}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \tag{281}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log\left(\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) \tag{282}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log(\lambda) + \log\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \tag{283}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log(\lambda) + \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}), \tag{284}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$, $\bar{K}_{Q,\boldsymbol{z}}$ and $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ are defined in (3), (23) and in (57b), respectively. The equality (282) follows from (16). Hence, the relative entropy $\mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ satisfies,

$$\begin{aligned}
&\mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) && (285)\\
&\quad=\int\Bigg(\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)-\log(\lambda)\\
&\qquad+\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\Bigg)\mathrm{d}P(\boldsymbol{\theta}) && (286)\\
&\quad=\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta})-\log(\lambda)\\
&\qquad+\int\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) && (287)\\
&\quad=\mathsf{D}(P\|Q)-\log(\lambda)+\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) && (288)\\
&\quad=\mathsf{D}(P\|Q)-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\qquad-\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)+\ \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P), && (289)
\end{aligned}$$

where (286) follows from (284); (288) follows from (70); and (289) follows from Lemma 23. Thus, from (289), it follows that

$$\begin{aligned}
&\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P)-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=\mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{D}(P\|Q)+\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right), && (290)
\end{aligned}$$

which completes the proof. ■

Appendix S
Proof of Lemma 26

From Lemma 23, for all $\alpha\in(0,\infty)$, it holds that

$$\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)=-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\log(\alpha), \quad (291)$$

where the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}$ is defined in (70).

Similarly, from [29, Lemma 20], for all $\lambda\in(0,\infty)$, it holds that

$$\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)=-\left(\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right), \quad (292)$$

with the functional $\mathsf{R}_{\boldsymbol{z}}$ defined in (6). From [29, Theorem 3], the function $\mathsf{S}_{Q,\lambda}$ in [29, Definition 7] satisfies that

$$\begin{aligned}
&\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z},\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) && (293)\\
&\quad=\lambda\Bigg(\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\\
&\qquad-\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\Bigg) && (294)\\
&\quad=\lambda\Bigg(\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\\
&\qquad+\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\log(\alpha)\Bigg) && (295)\\
&\quad=\lambda\Bigg(\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\qquad-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)+\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\log(\alpha)\Bigg), && (296)
\end{aligned}$$

where (295) follows from (291); and (296) follows from (292). Rearranging (296) yields

$$\begin{aligned}
&\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\log(\alpha)-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right). && (297)
\end{aligned}$$

Similarly, from Lemma 24 the function $\bar{\mathsf{S}}_{Q,\alpha}$ in (75) satisfies that

$$\begin{aligned}
&\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z},P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) && (298)\\
&\quad=\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\\
&\qquad+\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) && (299)\\
&\quad=\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\\
&\qquad-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\log(\alpha) && (300)\\
&\quad=\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\qquad+\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\\
&\qquad-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\log(\alpha), && (301)
\end{aligned}$$

where (300) follows from (291); and (301) follows from (292). Rearranging (301) yields

$$\begin{aligned}
&\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=-\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\qquad-\left(\log(\alpha)+K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right). && (302)
\end{aligned}$$

The proof proceeds by subtracting (302) from (297), resulting in

$$\begin{aligned}
&\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z},\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z},P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\qquad+2\left(\log(\alpha)+K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right), && (303)
\end{aligned}$$

where the functions $\mathsf{S}_{Q,\lambda}$ and $\bar{\mathsf{S}}_{Q,\alpha}$ are respectively defined in [29, Definition 7] and (75). From [35, Theorem 1] and Lemma 24, it follows that

$$\begin{aligned}
&\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z},\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z},P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad=\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)-\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\qquad+2\left(\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)-\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\right). && (304)
\end{aligned}$$

Substituting (304) into (303) yields

$$\begin{aligned}
&\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)-\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\\
&\qquad=\log(\alpha)+K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), && (305)
\end{aligned}$$

which completes the proof. ■

Appendix T
Examples

A. Example 1

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

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu}(\boldsymbol{\theta}) = 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}). \tag{306}$$

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

$$f(\boldsymbol{\theta}, x) = x\boldsymbol{\theta}, \tag{307}$$

and the loss function ℓ in (2) be

$$\ell(f(\boldsymbol{\theta}, x), y) = (x\boldsymbol{\theta} - y)^2, \tag{308}$$

which implies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = (x\boldsymbol{\theta} - y)^2, \tag{309}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (3). Furthermore, from assumptions (a), (b), and (309), it follows that there exists $\boldsymbol{\theta}^\star \in \operatorname{supp} Q$ such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}^\star) = 0$, which implies that

$$\delta^\star_{Q,\boldsymbol{z}} = 0. \tag{310}$$

Under the current assumptions, the objective of this example is to show that $\mathcal{C}_{Q,\boldsymbol{z}} = \left[\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$. For this purpose, from Lemma 6, it is sufficient to show that the condition in (31) holds. From Theorem 1, it follows that $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}\mu}(\boldsymbol{\theta}) = \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}), \tag{311}$$

with β satisfying (15). Thus,

$$\begin{aligned}
&\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) \\
&= \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}) \,\mathrm{d}\mu(\boldsymbol{\theta}) && (312a)\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(x\boldsymbol{\theta} - y)^2 - \delta^\star_{Q,\boldsymbol{z}}} \mathrm{d}\boldsymbol{\theta} && (312b)\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}^2 - \delta^\star_{Q,\boldsymbol{z}}} \mathrm{d}\boldsymbol{\theta} && (312c)\\
&= \int_0^\infty \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}^2} \mathrm{d}\boldsymbol{\theta} && (312d)\\
&= \int_0^\infty 4 \exp(-2\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} && (312e)\\
&= 2, && (312f)
\end{aligned}$$

where (312a) follows from (306); (312c) follows from the assumption that $(x, y) = (1, 0)$; and (312d) follows from the fact that $\delta^\star_{Q,\boldsymbol{z}} = 0$. Finally, the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) satisfies $\bar{K}_{Q,\boldsymbol{z}}\left(\frac{1}{2}\right) = 0$, which implies $-\delta^\star_{Q,\boldsymbol{z}} \in \mathcal{C}_{Q,\boldsymbol{z}}$, thus the set $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$.

B. Example 2

Consider Example 1 in Appendix T-A with $\boldsymbol{z} = ((1, 1))$. Note that (310) holds for this example. Under the current assumptions, the objective of this example is to show that $\mathcal{A}_{Q,\boldsymbol{z}} = \left(\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$. For this purpose, from Lemma 6, it is sufficient to show that the condition in (31) does not hold. That is,

$$\begin{aligned}
&\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) \\
&= \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \frac{\mathrm{d}Q}{\mathrm{d}\mu}(\boldsymbol{\theta}) \,\mathrm{d}\mu(\boldsymbol{\theta}) && (313a)\\
&= \int \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \mathrm{d}\mu(\boldsymbol{\theta}) && (313b)\\
&= \int \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(x\boldsymbol{\theta} - y)^2 - \delta^\star_{Q,\boldsymbol{z}}} \mathrm{d}\mu(\boldsymbol{\theta}) && (313c)\\
&= \int \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \mathrm{d}\mu(\boldsymbol{\theta}) && (313d)\\
&= 4 \int \frac{\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \mathrm{d}\mu(\boldsymbol{\theta}) && (313e)\\
&= 4 \int \Bigg(\frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \frac{\exp(-2\boldsymbol{\theta})}{2}}{\boldsymbol{\theta} - 1} \\
&\quad - \frac{-\frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta})}{2} - \frac{\exp(-2\boldsymbol{\theta})}{2}}{(\boldsymbol{\theta} - 1)^2} \Bigg) \mathrm{d}\mu(\boldsymbol{\theta}) && (313f)\\
&= 4 \Bigg(\int \frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \frac{\exp(-2\boldsymbol{\theta})}{2}}{\boldsymbol{\theta} - 1} \mathrm{d}\mu(\boldsymbol{\theta}) \\
&\quad + \int \frac{\frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta})}{2} + \frac{\exp(-2\boldsymbol{\theta})}{2}}{(\boldsymbol{\theta} - 1)^2} \mathrm{d}\mu(\boldsymbol{\theta}) \Bigg) && (313g)\\
&= 4 \Bigg(\frac{1}{2} \int \frac{2\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1} \mathrm{d}\mu(\boldsymbol{\theta}) \\
&\quad + \frac{1}{2} \int \frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \mathrm{d}\mu(\boldsymbol{\theta}) \Bigg), && (313h)
\end{aligned}$$

where (313a) follows from the assumption that $Q \ll \mu$, (313b) follows from (306), (313d) follows from (310) and the assumption that $(x, y) = (1, 1)$. Using integration by parts on the second integral in (313h), let the functions $\phi : \mathcal{M} \to \mathbb{R}$ and $\psi : \mathcal{M} \to \mathbb{R}$ be

$$\begin{aligned}
\phi(\boldsymbol{\theta}) &= \boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta}), \text{ and} && (314a)\\
\psi(\boldsymbol{\theta}) &= -\frac{1}{\boldsymbol{\theta} - 1}. && (314b)
\end{aligned}$$

The derivatives of ϕ and ψ satisfy

$$\begin{aligned}
\frac{\mathrm{d}\phi}{\mathrm{d}\mu}(\boldsymbol{\theta}) &= -2\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) - \exp(-2\boldsymbol{\theta}), \text{ and} && (315a)\\
\frac{\mathrm{d}\psi}{\mathrm{d}\mu}(\boldsymbol{\theta}) &= \frac{1}{(\boldsymbol{\theta} - 1)^2}, && (315b)
\end{aligned}$$

respectively. Note that given a subset $[a, b] \subset \mathcal{M}$ with $a, b \in \mathbb{R}$ such that $a < b$ it holds that,

$$\int_{[a,b]} \frac{\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \mathrm{d}\mu(\boldsymbol{\theta})$$

$$= \int_{[a,b]} \phi(\boldsymbol{\theta}) \frac{\mathrm{d}\psi}{\mathrm{d}\mu}(\boldsymbol{\theta})\mu(\boldsymbol{\theta}) \tag{316a}$$

$$= \Big[\phi(\boldsymbol{\theta})\psi(\boldsymbol{\theta})\Big]_a^b - \int_{[a,b]} \frac{\mathrm{d}\phi}{\mathrm{d}\mu}(\boldsymbol{\theta})\psi(\boldsymbol{\theta}) \,\mathrm{d}\mu(\boldsymbol{\theta}) \tag{316b}$$

$$\begin{aligned} &= \Big[- \frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1}\Big]_a^b \\ &\quad + \int_{[a,b]} \frac{-2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) - \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1} \,\mathrm{d}\mu(\boldsymbol{\theta}), \end{aligned} \tag{316c}$$

where (316c) follows the equalities (314) and (315). Substituting (316c) into (313h) yields

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta})$$

$$= 4\int_{[0,\infty)} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \,\mathrm{d}\mu(\boldsymbol{\theta}) \tag{317a}$$

$$\begin{aligned} &= 4\int_{[0,1]} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \,\mathrm{d}\mu(\boldsymbol{\theta}) \\ &\quad + 4\int_{(1,\infty)} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \,\mathrm{d}\mu(\boldsymbol{\theta}) \end{aligned} \tag{317b}$$

$$\geq 4\int_{[0,1]} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \,\mathrm{d}\mu(\boldsymbol{\theta}) \tag{317c}$$

$$\begin{aligned} &= 2\Bigg(\int_{[0,1]} \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1} \,\mathrm{d}\mu(\boldsymbol{\theta}) \\ &\quad + \int_{[0,1]} \frac{\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \,\mathrm{d}\mu(\boldsymbol{\theta})\Bigg) \end{aligned} \tag{317d}$$

$$\begin{aligned} &= 2\Bigg(\int_{[0,1]} \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1} \,\mathrm{d}\mu(\boldsymbol{\theta}) \\ &\quad + \Big[- \frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1}\Big]_0^1 \\ &\quad - \int_{[0,1]} \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1} \,\mathrm{d}\mu(\boldsymbol{\theta})\Bigg) \end{aligned} \tag{317e}$$

$$= 2\Big[- \frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1}\Big]_0^1 \tag{317f}$$

$$= \infty, \tag{317g}$$

where (317a) follows from the assumption that $\mathcal{M} = [0,\infty)$, (317c) follows from observing that for all $\boldsymbol{\theta} \in [0,\infty)$, it holds that $\frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2} > 0$, in (317d) follows from (313h), and (317e) follows from substituting (316c) into (317d). From (317g), it follows that the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23) is undefined at zero, which implies $\delta^{\star}_{Q,\boldsymbol{z}} \notin \mathcal{A}_{Q,\boldsymbol{z}}$, and this, $\mathcal{A}_{Q,\boldsymbol{z}} = (0,\infty)$.

C. Example 3

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

$$Q(\mathcal{B}) = \epsilon, \qquad \text{and} \tag{318a}$$

$$Q(\mathcal{M} \setminus \mathcal{B}) = 1 - \epsilon, \tag{318b}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{B} \\ c & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{B}, \end{cases} \tag{319}$$

with $c > 0$. Under the current assumptions, the objective of this example is to show that for all $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, it holds that $\mathcal{C}_{Q,\boldsymbol{z}} = \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$ and $\mathcal{A}_{Q,\boldsymbol{z}} = (0,\infty)$. To show this, it is necessary to characterize the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b). Hence, from the fact that the Lagrangian multiplier β for the optimization problem in (11) satisfies

$$\int \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})} \,\mathrm{d}Q(\boldsymbol{\nu}) = 1, \tag{320}$$

which follows from Theorem 1, the empirical risk function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \rightarrow \mathbb{R}_0^+$ in (319), which is a simple function, and the probability measure Q in (318a), it holds that

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

$$= \lambda\left(\frac{1}{\beta + c_0} Q(\mathcal{T}(\boldsymbol{z})) + \frac{1}{\beta + c_1} Q(\mathcal{M} \setminus \mathcal{T}(\boldsymbol{z}))\right) \tag{321a}$$

$$= \lambda\left(\frac{1}{\beta + c_0} Q(\mathcal{T}(\boldsymbol{z})) + \frac{1}{\beta + c_1} (1 - Q(\mathcal{T}(\boldsymbol{z})))\right) \tag{321b}$$

$$= \lambda\left(\frac{(\beta + c_1)Q(\mathcal{T}(\boldsymbol{z})) + (\beta + c_0)(1 - Q(\mathcal{T}(\boldsymbol{z})))}{\beta^2 + \beta(c_0 + c_1) + c_0 c_1}\right) \tag{321c}$$

$$= \lambda\left(\frac{(c_1 - c_0)Q(\mathcal{T}(\boldsymbol{z})) + \beta + c_0}{\beta^2 + \beta(c_0 + c_1) + c_0 c_1}\right). \tag{321d}$$

From (320) and (321d), it follows that

$$0 = \beta^2 + \beta(c_0 + c_1) + c_0 c_1 - \lambda((c_1 - c_0)Q(\mathcal{T}(\boldsymbol{z})) + \beta + c_0) \tag{322a}$$

$$= \beta^2 + \beta(c_0 + c_1 - \lambda) + c_0 c_1 - \lambda c_0 - \lambda(c_1 - c_0)Q(\mathcal{T}(\boldsymbol{z})). \tag{322b}$$

Hence,

$$0 = \beta^2 + \beta(c_1 - \lambda) - \lambda c_1 Q(\mathcal{T}(\boldsymbol{z})). \tag{322c}$$

Observe that the expression in (322c) is a quadratic polynomial that has two roots r_1 and r_2. Hence, (322c) in terms of r_1 and r_2 satisfies

$$0 = \beta^2 - (r_1 + r_2)\beta + r_1 r_2 \tag{323a}$$

$$= (\beta - r_1)(\beta - r_2), \tag{323b}$$

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

$$r_1 = -\frac{(c_1 - \lambda)}{2} - \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}, \tag{324a}$$

and

$$r_2 = -\frac{(c_1 - \lambda)}{2} + \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}. \tag{324b}$$

The proof continues by verifying that the roots in (324a) and (324b) are real and there is only one positive root for all $\lambda \in (0, +\infty)$ and for all $Q(\mathcal{T}(\boldsymbol{z})) \in [0, 1)$.

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

$$-\frac{c_1-\lambda}{2} \leq \left|\frac{c_1-\lambda}{2}\right| \tag{325}$$

$$=\sqrt{\left(\frac{c_1-\lambda}{2}\right)^2} \tag{326}$$

$$\leq\sqrt{\left(\frac{c_1-\lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}. \tag{327}$$

Observe that for all $Q(\mathcal{T}(\boldsymbol{z})) \in [0,1)$, $c_1 \in (0,\infty)$ and $\lambda \in [0,\infty)$ the expressions $\left(\frac{c_1-\lambda}{2}\right)^2$ and $\lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))$ are always positive. Thus, the square roots in (324a) and (324b) are real, which implies that r_1 and r_2 are real. From (324) and (327), for all $\lambda \in [0,+\infty)$ and for all $Q(\mathcal{T}(\boldsymbol{z})) \in [0,1)$, it holds that

$$r_1 <0; \tag{328a}$$

and following the same arguments

$$r_2 >0. \tag{328b}$$

Hence, the solution for the Lagrange Multiplier β that satisfies (320) given the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (319) and the probability measure Q in (318a) is

$$\beta = -\frac{(c_1-\lambda)}{2} + \sqrt{\left(\frac{c_1-\lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}, \tag{329}$$

which implies that the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) under the current assumptions in (318) and (319) satisfies

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})}. \tag{330}$$

From Theorem 1, it follows that $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (16) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})}}. \tag{331}$$

Under the current assumptions, from Lemma 6, it is sufficient to show that for all $c \in (0,\infty)$ and $\lambda \in (0,\infty)$ in (319), the function $\bar{K}_{Q,\boldsymbol{z}}$ in (23b) is strictly greater than $-\delta^{\star}_{Q,\boldsymbol{z}}$. From equality (330), it holds that

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})} \tag{332a}$$

$$> -\frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2} \tag{332b}$$

$$= -\frac{(c-\lambda)}{2} + \left|\frac{c-\lambda}{2}\right| \tag{332c}$$

$$\geq 0 \tag{332d}$$

$$= -\delta^{\star}_{Q,\boldsymbol{z}}, \tag{332e}$$

which proves that for all $c \in (0,\infty)$ and for all $\lambda \in (0,\infty)$, it holds that $\bar{K}_{Q,\boldsymbol{z}}(\lambda) > -\delta^{\star}_{Q,\boldsymbol{z}}$ which implies that $-\delta^{\star}_{Q,\boldsymbol{z}} \notin \mathcal{C}_{Q,\boldsymbol{z}}$ with the set $\mathcal{C}_{Q,\boldsymbol{z}}$ defined in (23) and thus $\mathcal{A}_{Q,\boldsymbol{z}} = (0,\infty)$.

Appendix U
Numerical Simulation

The MNIST dataset consists of 60,000 images for training and 10,000 images for testing. Out of the 60,000 training images, 12,183 are labeled as the digits six or seven, while 1,986 out of the 10,000 test images correspond to these digits. Each image is a 28×28 grayscale picture and is represented by the matrix $I \in [0,1]^{28\times 28}$.

A. Feature Extraction of the Histogram of Oriented Gradients

The grayscale images are processed by calculating their corresponding *histogram of oriented gradients* (HOG) [59]. The HOG for each image is computed through the following steps:

1.) For each pixel location $(i,j) \in \{1,2,...,28\}^2$ in the image, the gradients in the w- and h-directions (*width, height*) are computed using finite differences given by the functions $\mathsf{G}_w : \{1,2,...,28\}^2 \to \mathbb{R}$ and $\mathsf{G}_h : \{1,2,...,28\}^2 \to \mathbb{R}$, which are defined as

$$\mathsf{G}_w(i,j) = \begin{cases} I(i+1,j) - I(i-1,j) & \text{if } i \in \{2,\dots,27\} \\ I(i+1,j) - I(i,j) & \text{if } i = 1 \\ I(i,j) - I(i-1,j) & \text{if } i = 28 \end{cases}, \tag{333}$$

and

$$\mathsf{G}_h(i,j) = \begin{cases} I(i,j+1) - I(i,j-1) & \text{if } j \in \{2,\dots,27\} \\ I(i,j+1) - I(i,j) & \text{if } j = 1 \\ I(i,j) - I(i,j-1) & \text{if } j = 28 \end{cases}, \tag{334}$$

where $I(i,j) \in [0,1]$ represents the pixel intensity at location (i,j).

2.) Given a pixel location $(i,j) \in \{1,2,...,28\}^2$, the magnitude and orientation of a pixel at location (i,j) is given by the functions $\mathsf{M} : \{1,2,...,28\}^2 \to \mathbb{R}$ and $\phi : \{1,2,...,28\}^2 \to \mathbb{R}$, such that

$$\mathsf{M}(i,j) = \sqrt{\mathsf{G}_w(i,j)^2 + \mathsf{G}_h(i,j)^2}, \text{ and} \tag{335}$$

$$\phi(i,j) = \arctan\left(\frac{\mathsf{G}_h(i,j)}{\mathsf{G}_w(i,j)}\right). \tag{336}$$

3.) The matrix I is divided into sub-matrices of size 4×4, such that the number of sub-matrices is 7. These sub-matrices are referred to as *cells* and are denoted, for all $(w,h) \in \{1,\cdots,7\}^2$, by

$$C_{w,h} = \begin{bmatrix} I(a_w,b_h) & \cdots & I(a_w+3,b_h) \\ \vdots & \ddots & \vdots \\ I(a_w,b_h+3) & \cdots & I(a_w+3,b_h+3) \end{bmatrix}, \tag{337}$$

where the real values a_w and b_h are

$$a_w = 4(w-1)+1 \tag{338}$$

$$b_h = 4(h-1)+1. \tag{339}$$

This implies that the matrix I can be represented as

$$I = \begin{bmatrix} C_{1,1} & \cdots & C_{7,1} \\ \vdots & \ddots & \vdots \\ C_{1,7} & \cdots & C_{7,7} \end{bmatrix}. \tag{340}$$

From (337), the set of all pairs (i, j) of pixel coordinates in I that lie within the cell $C_{w,h}$ is given by:

$$\mathcal{A}_{w,h} = \{a_w, a_w + 3\} \times \{b_h, b_h + 3\}, \tag{341}$$

with a_w in (338) and b_h in (339).

4.) For each cell $C_{w,h}$ in (337) the orientations $\phi(i, j)$ in (336) are divided into 9 bins. That is, the n^{th} bin, with $1 \leq n \leq 9$, satisfies that

$$\mathcal{B}^{(n)}_{w,h} = \left\{ \phi(i,j) \in \mathbb{R} : 180\left(\frac{n-1}{9}\right) \leq \phi(i,j) < 180\left(\frac{n}{9}\right) : (i,j) \in \mathcal{A}_{w,h} \right\}. \tag{342}$$

The contribution of each pixel to its corresponding bin is based on its gradient magnitude. That is, the value of the n-th bin from the (w, h)-th cell $C_{w,h}$ in (337) is given by the function $H_{w,h}(n) : \{1, 2, \ldots, 9\} \to \mathbb{R}$, such that

$$H_{w,h}(n) = \sum_{(i,j)\in\mathcal{A}_{w,h}} \mathsf{M}(i,j) \mathbb{1}_{\left\{\phi(i,j)\in\mathcal{B}^{(n)}_{w,h}\right\}}, \tag{343}$$

with M in (335); and $\mathcal{B}^{(n)}_{w,h}$ in (342). Thus, the histogram of gradient orientations of the cell $C_{w,h}$ is represented by the vector $\boldsymbol{H}_{w,h} \in \mathbb{R}^9$, such that

$$\boldsymbol{H}_{w,h} = [H_{w,h}(1), H_{w,h}(2), \cdots H_{w,h}(9)], \tag{344}$$

with the function $H_{w,h}$ in (343).

5.) To account for illumination and contrast variations, the histogram $\boldsymbol{H}_{w,h}$ in (344) is normalized. To normalize the histograms for all cells $C_{w,h}$ in (337), the cells are grouped into sub-matrices formed by 2×2 cells with a *cell overlap* set to 1 pixel, such that the number of sub-matrices is:

$$(7-1) \times (7-1) = 36. \tag{345}$$

These sub-matrices of the matrix I in (340), with $(m, s) \in \left\{1, \cdots \sqrt{36}\right\}^2$ are referred to as *blocks*, and denoted by

$$B_{m,s} = \begin{bmatrix} C_{m,s} & C_{m+1,s} \\ C_{m,s+1} & C_{m+1,s+1} \end{bmatrix}, \tag{346}$$

with the matrix $C_{m,s}$ in (337). From (340) and (346), a block $B_{m,s}$ is a sub-matrix of size 8×8, *i.e.*, $B_{m,s} \in \mathbb{R}^{8\times 8}$. The *size* of a block is given by the ratio of the total number of pixels in a block to the number of pixels in a cell:

$$\frac{8 \times 8}{4 \times 4} = 4, \tag{347a}$$

The normalized histogram of a cell $C_{w,h}$ in a block $B_{m,s}$ is denoted by the vector $\hat{\boldsymbol{H}}^{(m,s)}_{w,h} \in \mathbb{R}^9$. This normalization of is typically done using the ℓ_2-norm, such that

$$\hat{\boldsymbol{H}}^{(m,s)}_{w,h} = \frac{\boldsymbol{H}_{w,h}}{\sqrt{\sum_{(i,j)\in\{m,m+1\}\times\{s,s+1\}} \boldsymbol{H}^2_{i,j} + \epsilon^2}}, \tag{348}$$

where $\boldsymbol{H}_{w,h}$ in (344) is the unnormalized histogram, and the $\epsilon = 10^{-4}$ to avoid division by zero.

6.) For an image with 36 blocks (see (345)), 9 orientation bins, and a size of block 4 (see (347)), the dimension of the HOG feature vector $\hat{\mathbf{x}}$ is:

$$36 \times 4 \times 9 = 1296. \tag{349}$$

The HOG feature vector $\hat{\mathbf{x}} \in \mathbb{R}^{1296}$ is formed by concatenating all the normalized histograms $\hat{\boldsymbol{H}}^{(m,s)}_{w,h}$ such that

$$\begin{aligned} \hat{\mathbf{x}} = \Big[& \hat{\boldsymbol{H}}^{(1,1)}_{1,1}, \hat{\boldsymbol{H}}^{(1,1)}_{1,2}, \hat{\boldsymbol{H}}^{(1,1)}_{2,1}, \hat{\boldsymbol{H}}^{(1,1)}_{2,2}, \\ & \hat{\boldsymbol{H}}^{(2,1)}_{2,1}, \hat{\boldsymbol{H}}^{(2,1)}_{2,2}, \hat{\boldsymbol{H}}^{(2,1)}_{3,1}, \hat{\boldsymbol{H}}^{(2,1)}_{3,2}, \cdots, \\ & \hat{\boldsymbol{H}}^{(6,6)}_{6,6}, \hat{\boldsymbol{H}}^{(6,6)}_{6,7}, \hat{\boldsymbol{H}}^{(6,6)}_{7,6}, \hat{\boldsymbol{H}}^{(6,6)}_{7,7} \Big]^\top. \end{aligned} \tag{350}$$

B. Principal Component Analysis

The final step in the data processing is to reduce the dimensionality of the pattern $\hat{\mathbf{x}}$ in (350) from $\mathbb{R}^{1296}$ to $\mathbb{R}^2$, while ensuring that the important structure of the pattern is preserved. In this simulation, *principal component analysis (PCA)* is used to project the high-dimensional data onto a lower-dimensional subspace. From 60,000 images for training in the MNIST, the HOG of two handwritten numbers (in this simulation 6 and 7) are computed, as mentioned in Appendix U-A. The resulting 12,183 HOG vectors $\hat{\mathbf{x}} \in \mathbb{R}^{1296}$ are reduced to $\mathbb{R}^2$ using PCA as follows:

1.) To reduce the dimensionality, the first step in PCA is to compute the *covariance matrix* of the data. This matrix captures the relationships between the different features (or dimensions) of the data. The covariance matrix is calculated as follows:

$$\mathbf{C} = \frac{1}{n-1} \sum_{i=1}^{n} (\hat{\mathbf{x}}_i - \boldsymbol{\mu})(\hat{\mathbf{x}}_i - \boldsymbol{\mu})^\top, \tag{351}$$

where $n = 12{,}183$, $\mathbf{C} \in \mathbb{R}^{1296\times 1296}$, and $\boldsymbol{\mu}$ is the mean of all the training patterns given by

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{x}}_i. \tag{352}$$

2.) The next step in PCA is to perform an *eigenvalue decomposition* of the covariance matrix $\mathbf{C}$ in (351). The decomposition can be written as:

$$\mathbf{C} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top, \tag{353}$$

where $\mathbf{V} \in \mathbb{R}^{1296\times 1296}$ is a matrix whose columns are the eigenvectors of $\mathbf{C}$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{1296\times 1296}$ is a diagonal matrix containing the corresponding eigenvalues.

3.) Following the computation of the eigenvectors, the dimensionality is reduced from $\mathbb{R}^{1296}$ to $\mathbb{R}^2$ by selecting the two eigenvectors associated with the largest eigenvalues. Denote these top two eigenvectors as $\mathbf{w}_1$ and $\mathbf{w}_2$. These eigenvectors constitute the columns of the projection matrix $\mathbf{W} \in \mathbb{R}^{1296\times 2}$, defined as

$$\mathbf{W} = [\mathbf{w}_1\, \mathbf{w}_2]. \tag{354}$$

4.) Once the projection matrix $\mathbf{W}$ is computed, each high-dimensional pattern $\hat{\mathbf{x}} \in \mathbb{R}^{1296}$ can be projected onto the new $\mathbb{R}^2$ subspace. The projection is performed as follows:

$$\mathbf{x} = \mathbf{W}^{\top}\hat{\mathbf{x}}, \tag{355}$$

with $\hat{\mathbf{x}}$ in (350), $\boldsymbol{W}$ in (354) and $\mathbf{x} \in \mathbb{R}^2$ is the 2-dimensional coordinates of the original pattern $\hat{\mathbf{x}}$ in the reduced-dimensional space.

C. Simulation Dataset

In this simulation, a datapoint is a tuple $(\hat{\mathbf{x}}, y) \in \mathbb{R}^{1296} \times \{6, 7\}$, with $\hat{\mathbf{x}}$ in (350) and y being the label assigned by MNIST to the image I in (340). The label y corresponds to the digit in the image I. Such an image produces the vector $\hat{\mathbf{x}}$, when its HOG features are computed.

References

[1] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[19] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky, "Generalization bounds: Perspectives from information theory and PAC-Bayes," *Foundations and Trends® in Machine Learning*, vol. 18, no. 1, pp. 1–223, Jan. 2025.

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

[21] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Measures of Complexity: Festschrift for Alexey Chervonenkis*, vol. 16, no. 2, pp. 11–30, Oct. 2015.

[22] 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.

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

[24] D. A. McAllester, "Some PAC-Bayesian theorems," in *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, Madison, WI, USA, Jul. 1998, pp. 230–234.

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

[26] 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.

[27] D. Cullina, A. N. Bhagoji, and P. Mittal, "PAC-learning in the presence of adversaries," *Advances in Neural Information Processing Systems*, vol. 31, no. 1, pp. 1–12, Dec. 2018.

[28] 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.

[29] ——, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 5122 – 5161, Jul. 2024.

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

[31] 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.

[32] 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.

[33] B. Zou, L. Li, and Z. Xu, "The generalization performance of ERM algorithm with strongly mixing observations," *Machine Learning*, vol. 75, no. 3, pp. 275–295, Feb. 2009.

[34] 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*, vol. 70, no. 1, pp. 632–655, Nov. 2023.

[35] 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.

[36] S. M. Perlaza and X. Zou, "The generalization error of machine learning algorithms," arXiv preprint 2411.12030, Nov. 2024.

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

[38] Q. Lin, Z. Lu, and L. Xiao, "An accelerated proximal coordinate gradient method and its application to regularized empirical risk minimization," *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2244–2273, Jul. 2015.

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

[40] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, “On the generalization for transfer learning: An information-theoretic analysis,” *IEEE Transactions on Information Theory*, vol. 70, no. 10, pp. 7089–7124, Aug. 2024.

[41] B. Rodríguez Gálvez, “An information-theoretic approach to generalization theory,” PhD thesis, KTH Royal Institute of Technology, Stockholm, Sweden, Jun. 2024.

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

[43] A. Beck and M. Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, Jan. 2003.

[44] 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.

[45] A. R. Esposito and M. Gastpar, “From generalisation error to transportation-cost inequalities and back,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aalto, Finland, Jun. 2022, pp. 294–299.

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

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

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

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

[50] 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.

[51] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, “Generalization analysis of machine learning algorithms via the worst-case data-generating probability measure,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2024, pp. 17 271–17 279.

[52] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Aug. 1998.

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

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

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

[56] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill Book Company, Inc., 1976.

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

[58] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

[59] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, Apr. 2005, pp. 886–893.