University of Sheffield

Doctoral Thesis

Empirical Risk Minimization with f-Divergence Regularization for Machine Learning

Author:
Francisco Daunas

Supervisor:
Dr. Iñaki Esnaola

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

School of Electrical and Electronic Engineering

October 29, 2025

Dedicated to my wife and family

Acknowledgements

I am deeply grateful to my supervisors for their unwavering guidance and patience throughout my doctorate. Special mention to Dr. Samir M. Perlaza for all his time, input, and mentoring through my PhD. To my research group colleaguesthank you for enduring my endless measure theory tangents with humor and insight. Finally, to Sandra, my rock: your support and encouragement made this journey possible.

UNIVERSITY OF SHEFFIELD

Abstract

Engineering

School of Electrical and Electronic Engineering

Doctor of Philosophy

Empirical Risk Minimization with f-Divergence Regularization for Machine Learning

by Francisco Daunas

This work contributes to the field of statistical machine learning by providing a theoretical characterization of the role of regularization in supervised learning through the lens of information measures. The asymmetry of the relative entropy is analyzed in the context of its role in the regularization of empirical risk minimization. Building on this insight, a broad family of f-divergences is introduced as potential regularizers for empirical risk minimization. Under mild assumptions, solutions for general f-divergences are derived, and the concept of the normalization function is formally defined. Furthermore, a dual optimization problem associated with empirical risk minimization using f-divergence regularization is explored. By studying the normalization function, it is demonstrated that the duality gap is zero, and insights from the dual formulation are used to derive explicit expressions for the generalization error of general statistical learning algorithms in terms of f-divergence-regularized learning frameworks.

Contents

Acknowledgements **v**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Aim and Objectives . . . 3
1.3 Publications . . . 4
1.4 Contributions . . . 4
1.5 Notation and General Definitions . . . 6

2 Brief Introduction To Statistical Machine Learning **9**
2.1 Machine Learning . . . 9
2.1.1 Supervised Machine Learning . . . 11
2.1.2 Statistical Machine Learning . . . 12
2.2 Empirical Risk Minimization . . . 13
2.3 Information Theory . . . 14
2.3.1 Classical Generalization Bounds in Machine Learning . . . 15
2.3.2 Information Stability Methods . . . 16
2.4 Type-I ERM-RER Problem . . . 17

3 ERM-RER Asymmetry **19**
3.1 Introduction . . . 19
3.2 Type-II ERM-RER Problem . . . 20
3.3 Type-II ERM-RER Solution & Results . . . 21
3.3.1 The Normalization Function . . . 23
3.3.2 Type-II ERM-RER Properties . . . 25
3.3.3 Relative Entropy Asymmetry via Empirical Risk . . . 30
3.4 Numerical Results of Type-I and Type-II Regularization . . . 35
3.5 Conclusion . . . 38

4 ERM with f-Divergence Regularization **39**
4.1 Introduction . . . 39
4.2 ERM-fDR Problem . . . 40
4.3 ERM-fDR Solution & Results . . . 40
4.3.1 ERM-fDR Normalization Function . . . 42
4.3.2 Properties of the ERM-fDR solution . . . 44
4.3.3 Equivalence of the f-Regularization via Transformation of the Empirical Risk . . . 45
4.4 Numerical Comparison of ERM-fDR Regularizations . . . 48
4.5 Conclusions . . . 50

5 ERM-fDR Duality **51**
5.1 Introduction . . . 51

5.2 ERM-fDR Dual Problem . . . 51
5.3 ERM-fDR Dual Solution & Results . . . 52
5.3.1 ERM-fDR Expected Empirical Risk . . . 54
5.4 Exact Characterization of the Generalization Error . . . 54
5.5 Conclusions . . . 56

6 Conclusion **59**
6.1 Contributions . . . 59
6.2 Limitations of the current work . . . 60
6.3 Future Work . . . 61

A Preliminaries **63**
A.1 Canonical form f-Divergence . . . 63
A.2 Asymmetry Minor Results . . . 64
A.3 Canonical form f-Divergence . . . 65

B ERM-RER Type-II **69**
B.1 Proof of Lemma 3.3.1 . . . 69
B.2 Proof of Lemma 3.3.2 . . . 71
B.3 Proof of Lemma 3.3.4 . . . 73
B.4 Proof of Lemma 3.3.5 . . . 75
B.5 Proof of Lemma 3.3.9 . . . 78
B.6 Proof of Lemma 3.3.10 . . . 78
B.7 Proof of Lemma 3.3.11 . . . 79
B.8 Proof of Lemma 3.3.12 . . . 79
B.9 Proof of Lemma 3.3.13 . . . 81
B.9.1 Case 1 . . . 82
Part 1 . . . 82
Part 2 . . . 83
Part 3 . . . 85
B.9.2 Case 2 . . . 86
Part 1 . . . 86
Part 2 . . . 87
Part 3 . . . 87
B.10 Proof of Lemma 3.3.14 . . . 87
B.10.1 Case 1 . . . 88
B.10.2 Case 2 . . . 88
B.11 Proof of Lemma 3.3.15 . . . 89
B.12 Proof of Lemma 3.3.16 . . . 90
B.13 Proof of Lemma 3.3.17 . . . 91
B.14 Proof of Lemma 3.3.19 . . . 92
B.15 Proof of Theorem 3.3.3 . . . 93
B.16 Proof of Lemma 3.3.22 . . . 94
B.17 Proof of Lemma 3.3.23 . . . 95
B.18 Proof of Lemma 3.3.25 . . . 96

C ERM-fDR **99**
C.1 Proof of Theorem 4.3.1 . . . 99
C.2 Proof of Theorem 4.3.2 . . . 109
C.3 Proof of Lemma 4.3.2 . . . 114
C.4 Proof of Lemma 4.3.3 . . . 116

C.5 Proof of Lemma 4.3.4 . . . 119
C.6 Proof of Lemma 4.3.5 . . . 121
C.7 Proof of Lemma 4.3.6 . . . 121
C.8 Proof of Lemma 4.3.7 . . . 122

D Dual ERM-fDR **125**
D.1 Proof of Theorem 5.4.1 . . . 125
D.2 Proof of Lemma 5.3.3 . . . 126
D.3 Proof Theorem 5.4.2 . . . 127
D.4 Proof Theorem 5.4.3 . . . 128
D.5 Proof Remark 1 . . . 128

E Complementary **131**
E.1 Examples . . . 131
E.1.1 Example 1 . . . 131
E.1.2 Example 2 . . . 132
E.1.3 Example 3 . . . 134
E.2 ERM-fDR Table Derivation . . . 136
E.2.1 Relative Entropy . . . 136
E.2.2 Reverse Relative Entropy . . . 137
E.2.3 Jeffreys Divergence . . . 138
E.2.4 Jensen-Shannon Divergence . . . 138
E.2.5 Hellinger Divergence . . . 139
E.2.6 X2 Divergence . . . 139
E.3 Numerical Simulation . . . 140
E.3.1 Features extraction of the Histogram of Oriented Gradients . . . 140
E.3.2 Principal Component Analysis . . . 143
E.3.3 Simulation Dataset . . . 144

List of Figures

3.1 28×28 image of a handwritten six from the MNIST dataset. 36
3.2 28×28 image of a handwritten seven from the MNIST dataset. 36
3.3 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 (2.16) and (3.4), respectively, computed over one hundred different training and test dataset random selections. 37
3.4 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 (2.16) and (3.4), respectively, computed over one hundred different training and test dataset random selections. 37
3.5 Average of the differences $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ 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 (2.16) and (3.4), respectively, computed over one hundred different training and test dataset random selections. 38

4.1 Representation of the empirical risk transformation from the f-divergence induced by $f(t) = -\log(t)$ and the g-divergence induced by $g(t) = t\log(t)$. 48
4.2 Average Training Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, for the f-divergence regularization Type-I ERM-RER, Type-II ERM-RER, Shannon-Jensen and Hellinger in Table 4.1, respectively, computed over one hundred different training and test dataset random selections. 48
4.3 Average Test Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, for the f-divergence regularization Type-I ERM-RER, Type-II ERM-RER, Shannon-Jensen and Hellinger in Table 4.1, respectively, computed over one hundred different training and test dataset random selections. 49
4.4 Average of the differences $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, for the f-divergence regularization Type-I ERM-RER, Type-II ERM-RER, Shannon-Jensen and Hellinger in Table 4.1, respectively, computed over one hundred different training and test dataset random selections. 49

Chapter 1

Introduction

1.1 Motivation

This thesis aims to build a mathematical theory to understand, model, and determine the fundamental limits[1] of the influence that training data has in machine learning (ML) under specific regularization schemes. In this context, training data is understood as a set of samples drawn from an unknown probability distribution, often composed of a pattern and a corresponding label, used to train machine learning models. In general, the process of learning a model with an ML algorithm can be viewed as a decision-making process driven by some statistic, which is a numerical quantity used to describe or infer properties about a larger population from which the samples are drawn, often considered in an infinite hypothesis sample space [1]. In machine learning, an algorithm is a set of rules or procedures that enable a system to learn patterns from data and make predictions or decisions [2]. Therefore, in most training processes, ML algorithms rely on partial information about the stochastic phenomena that characterize the environment in which the learning occurs. That is, the learning process, more often than not, operates with incomplete information, and for that reason, it must account for its information acquisition[2]. For example, the sampled data for the training process might contain information that is distorted, biased, or incomplete, exposing learning algorithms to undesired influences that can steer the decision-making process away from the target task.

The problem tackled in this thesis lies at the intersection of information theory[3] and ML. These subjects have gained attention in recent years due to the integration of information and digital communication technologies in a wide variety of fields. The increase in data storage capacity, the availability of large databases that can be used to train ML algorithms for the discovery of patterns, the rapid increase of computational capabilities due to technological developments such as GPUs, and the relative simplicity of deploying existing algorithms are factors that have also contributed to their popularity in different areas [4, 5]. Ensuring the reliability of ML algorithms plays a central role. Examples of this reliability characteristic needed arise in contexts that involve humans, machines, or both of them, e.g., (*a*) Healthcare diagnosis systems evaluating possible diseases in different patients; (*b*) Bio-metric

[1] A fundamental limit is a performance bound determined by Physics and/or Mathematics, *e.g.* , the fundamental limit of the lowest temperature of a body is zero Kelvin; the fundamental limit of information transmission over an additive white Gaussian noise channel is $\log(1 + \mathrm{SNR})$ bits per channel use, where SNR is the signal to noise ratio.

[2] The process of supplying data, often sampled from an unknown distribution to machine learning algorithms for training, validation, or testing [3].

[3] Information theory is the scientific study of the quantification, storage, and communication of digital information

recognition for identity verification; (*c*) Autonomous driving vehicles in unstructured environments; (*d*) Detection of buried explosive objects; and (*e*) Financial service companies which rely on information to anticipate customer behavior and plan their business strategy.

The wide range of applications in which ML can be used has propelled the technology to the forefront of research and increased the demand in the industry. As ML makes its way into everyday life, guaranteeing the reliability of systems that make use of these algorithms is of prime importance to protect computer systems and networks. Consequently, the accuracy of systems that rely on ML algorithms must be ensured before deployment. However, evaluating the performance of ML algorithms is non-trivial and poses great difficulty both theoretically and in practical terms. As a result, the mathematical challenges involved in developing rigorous theoretical guarantees have naturally led to a stronger reliance on empirical testing to assess robustness and reliability [6]. Although straightforward, empirical approaches inherently constrain the evaluation to case-specific performance analyses and data-driven testing procedures. However, due to the intrinsic limitations of finite datasets and experimental designs, such methods may unavoidably yield performance estimates that do not fully capture the generalization capabilities of the underlying algorithms [7]. Thus, the implementation of ML in safety-critical settings is limited by the ability to provide guarantees based on empirical methods [8]. Furthermore, the current implementation of ML algorithms is required to handle sensitive data privately and prevent negative impacts on safety or fundamental rights [9]. This implies the algorithm must ensure the reliability of the ML models derived from the data employed in the training process, along with being robust and efficient [10]. For this reason, this thesis follows the information-theoretic framework proposed in [11], which follows the derivation of fundamental limits that can be generalized under the umbrella of supervised learning in ML, as shown in [11–13].

As previously mentioned, evaluating the performance of ML algorithms is non-trivial and poses great difficulty. As such, theoretical tools such as the VC dimension and Rademacher complexity enable research to progress our understanding of ML accuracy. For example, [14] introduced the notion of adversarial examples[4] which showed ML algorithms have blind spots that are connected to the data distribution in inconspicuous ways. The potential for overfitting in these learning problems, even though ML algorithms aim for strong performance on new data, requires careful study.

Despite current empirical validation frameworks enabling the discovery and testing of the intrinsic relations between ML algorithms and data, building an understanding of fundamental limits requires the characterization of these relations, which this framework does not provide. Thus, information theory tools can be leveraged to provide operational meaning to the coupling enclosed by the statistical dependencies in training data and the resulting model selection of ML algorithms [15]. Specifically, the methodological framework developed for the study of the encoding, storage, and transmission of data in information theory is a good fit for research on the foundation of ML. This stems from the fact that ML foundations come from probabilistic methods that drive the development of new algorithms [16].

Having established the general motivation of this thesis and the reason why information theory and statistical learning are of particular interest, the context is hereby

[4] An adversarial example in ML is data points to which an imperceptible non-random perturbation is applied in order to maximize the prediction error [14, p. 2]

presented. The field of ML is significantly wide. However, most of the algorithms fall into one of the following categories: (a) Supervised learning, (b) Unsupervised learning, (c) Semi-Supervised Learning, (d) Reinforcement Learning, (e) Ensemble Learning and (f) Instance-based Learning [6, 17]. The scope of this thesis is on the branch of supervised learning problems, as these can be formulated in terms of the classical Empirical Risk Minimization problem, which has been the central framework across several fields. In this setting, learning can be understood as estimating the underlying mapping from patterns $\mathcal{X}$ to labels $\mathcal{Y}$, or equivalently, the conditional distribution $P_{\mathcal{X}|\mathcal{Y}=y}$.

1.2 Aim and Objectives

This thesis follows an exploratory approach, *i.e.* the research first develops comprehension over an issue more thoroughly, before attempting to establish patterns into statistically inferable data. Therefore, the main objective is to theoretically characterize the impact of regularization in supervised ML using information measures. To accomplish this, the objectives are the following:

1. To characterize the role of asymmetry in the empirical risk minimization with relative entropy regularization in supervised ML.

- **1.1** To develop bounds over the expected empirical risk with relative entropy regularization that aids the selection of the regularization parameter.
- **1.2** To determine the benefits of relative entropy regularization $\mathsf{D}(Q\|P)$ over $\mathsf{D}(P\|Q)$ and vice versa.
- **1.3** To generalize relative entropy regularization $\mathsf{D}(Q\|P)$ and $\mathsf{D}(P\|Q)$ to the symmetrized relative entropy.

2. To study general f-divergences as regularizers in supervised ML.

- **2.1** To extend results obtained in the relative entropy analysis to the general regularization based on basic properties of the f-divergences.
- **2.2** To characterize the normalization factor for f-divergences regularized learning algorithms.
- **2.3** To derive a general analytical solution to the empirical risk minimization with f-divergence regularization in the form of a Radon-Nikodym derivative.

3. To characterize the generalization error for f-divergence regularized algorithms.

- **3.1** To solve the dual problem for empirical risk minimization with f-divergence.
- **3.2** To show the duality gap is zero via the Legendre-Fenchel transform.
- **3.3** To characterize the expected empirical risk for f-divergences regularized learning algorithms.
- **3.4** To apply the *method of gaps*[5] to empirical risk minimization with f-divergence regularization.

[5]The method of gaps is a framework to characterize the generalization error of empirical risk minimization with relative entropy regularization introduced in [18]

1.3 Publications

Journal

F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor. Analysis of the Relative Entropy Asymmetry in the Regularization of Empirical Risk Minimization, Submitted to IEEE Transactions on Information Theory, in February 2024.

Conference

F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor. Equivalence of the Empirical Risk Minimization to Regularization on the Family of f-Divergences, In Proc. of the IEEE International Symposium on Information Theory (ISIT), Athens, Greece, Jun., 2024.

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 Proc. of the IEEE International Symposium on Information Theory (ISIT), Taipei, Taiwan, Jun., 2023.

F. Daunas, I. Esnaola, and S. M. Perlaza A Dual Optimization View to Empirical Risk Minimization with f-Divergence Regularization, In Proc. of the IEEE Information Theory Workshop (ITW), Sydney, Australia, Sep., 2025.

Technical Reports

F. Daunas, I. Esnaola, S. M. Perlaza, and G. Aminian. Generalization Error of f-Divergence Stabilized Algorithms via Duality, Arxiv Jun., 2025. (available at arxiv:2502.14544)

F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor. Empirical Risk Minimization with f-Divergence Regularization in Statistical Learning, INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9521, Oct., 2023.

F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor. Empirical Risk Minimization with Relative Entropy Regularization Type-II, INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9508, May., 2023.

1.4 Contributions

This thesis explores how divergence-based regularization on empirical risk minimization affects learning algorithms in statistical learning. It focuses on both relative entropy and its generalization to the broader class of f-divergences. The goal is to understand how these regularizers influence model behavior, generalization, and optimization. The work is organized into three main contributions.

Chapter 4:

The first contribution, *Characterization of Asymmetry in Relative Entropy Regularization*, investigates the role of asymmetry in the relative entropy functional within empirical risk minimization. The asymmetry is analyzed by solving the empirical risk minimization problem regularized with the reverse relative entropy, where the resulting solution is expressed as a function defining a change of measure. Then, by

comparing this formulation with the known solution for the classical relative entropy regularization, the properties and differences between the solutions are examined. This work reveals how the use of reverse relative entropy $\mathsf{D}(Q\|P)$, as opposed to the classical form $\mathsf{D}(P\|Q)$, induces a collapse of the support of the solution onto that of the reference measure, revealing an inductive bias. The work also shows how relative entropy can be made equivalent to reverse relative entropy through appropriate transformations of the empirical risk, highlighting how the choice of regularizer shapes the learning process.

- Solution to the Type-II ERM-RER problem [19, Theorem 1] (see Theorem 3.3.1)
- Collapse of the support, showing inductive bias [19, Lemma 3] (see Corollary 3.3.3)
- (δ, ϵ) Optimality of Type-II ERM-RER solution [19, Theorem 2] (see Theorem 3.3.3)
- Asymmetry via empirical risk transformation [19, Theorem 3] (see Theorem 3.3.4)

Chapter 3:

The second contribution, *Generalization to f-Divergences as Regularizers in the Empirical Risk Minimization*, extends the analysis from relative entropy to general f-divergences. It provides general solutions to the regularized learning problem, expressed through a function determined by the choice of f-divergence, which characterizes the change of measure between the reference and the optimal distributions. Furthermore, it introduces a normalization function that guarantees the resulting solutions define valid probability measures. The properties of this function, such as being continuous and increasing, help in understanding and computing the solutions. This broadens the use of divergence-based regularization in supervised learning and connects theory with practical methods.

- Expanded the set of f-divergences to the ERM-fDR problem for which explicit solutions can be derived [20, Theorem 1] (see Theorem 4.3.1)
- Introduction of the Normalization Function [20] (see Definition 4.3.1)
- Properties of the Normalization Function, [20, Lemma 2] (see Theorem 4.3.2)
- Equivalence of ERM-fDR via empirical risk transformation [20, Theorem 2] (see Theorem 4.3.3)

Chapter 5:

The third contribution, Characterization of Generalization Error for f-Divergence Regularized Algorithms, focuses on how these regularized models perform on unseen data. Solving the dual version of the learning problem and proving that the duality gap is zero offers a solid theoretical basis for generalization. The zero duality gap establishes the equivalence between the Empirical Risk Minimization problem with f-divergence regularization and its dual formulation at the optimal solution. This, in turn, enables the derivation of expressions that characterize the generalization error in terms of the optimal solution. Furthermore, the extension of the method of gaps to f-divergences also gives a new tool to study generalization in a wide range of models. These results help explain how regularization influences model performance beyond the training data.

- Solution to the Dual ERM-fDR problem [21, Theorem 2] (see Theorem 5.3.1)
- Show duality gap is zero between the Dual ERM-fDR and ERM-fDR problems [21, Lemma 1] (see Lemma 5.3.2)
- Characterization of generalization error for general learning algorithm, [22, Theorem 5] (see Theorem 5.4.1)
- Characterization of generalization Error for f-divergence regularized algorithm, [22, Theorem 6] (see Theorem 5.4.3)

Together, these contributions build a clearer and more general understanding of how divergence-based regularization affects learning and generalization in machine learning.

1.5 Notation and General Definitions

In this work, sets are denoted by calligraphic letters. Given a set $\mathcal{M}$, the notation $\mathscr{F}$ represents a sigma-field (σ-field) on $\mathcal{M}$, such that the measurable space on $\mathcal{M}$ is denoted by $(\mathcal{M}, \mathscr{F})$. In the case in which $\mathcal{M} \subset \mathbb{R}^d$, for some $d \in \mathbb{N}$, the Borel σ-field on $\mathcal{M}$ is denoted by $\mathscr{B}(\mathcal{M})$. The set of probability measures that can be defined upon the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ is denoted by $\triangle(\mathcal{M})$. Given a probability measure $Q \in \triangle(\mathcal{M})$ the set exclusively containing the probability measures in $\triangle(\mathcal{M})$ that are absolutely continuous with respect to Q is denoted by $\triangle_Q(\mathcal{M})$, later defined in (2.14). Alternatively, the set exclusively containing the probability measures $P \in \triangle(\mathcal{M})$ for which the reference measure Q is absolutely continuous with respect to P is denoted by $\bigtriangledown_Q(\mathcal{M})$, later defined in (3.2). The Radon-Nikodym derivative of the measure P with respect to Q is denoted by $\frac{\mathrm{d}P}{\mathrm{d}Q} : \mathcal{M} \to [0, \infty)$.

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

Definition 1.5.1 (f-divergence [23])**.** *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and $f(0) \triangleq \lim_{x \to 0^+} f(x)$. Let P and Q be two probability measures on the same measurable space, with P absolutely continuous with Q. The f-divergence of P with respect to Q, denoted by $\mathsf{D}_f(P\|Q)$, is*

$$\mathsf{D}_f(P\|Q) \triangleq \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(s)\right) \mathrm{d}Q(s), \tag{1.1}$$

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

The Kullback-Leibler divergence (or relative entropy) of P with respect to Q, denoted by $\mathsf{D}(P\|Q)$, is obtained from (1.1) when the function f satisfies $f(x) = x \log(x)$. More specifically,

$$\mathsf{D}(P\|Q) = \int \frac{\mathrm{d}P}{\mathrm{d}Q}(s) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(s)\right) \mathrm{d}P(s), \tag{1.2}$$

and the symmetries Kullback-Leibler divergence (or Jeffreys divergence) is denoted by

$$\mathsf{D}_J(P\|Q) = \mathsf{D}(P\|Q) + \mathsf{D}(Q\|P). \tag{1.3}$$

Furthermore, in the case in which the function f in (1.1) is continuous and differentiable, the derivative of the function f is denoted by

$$\dot{f} : (0, +\infty) \to \mathbb{R}. \tag{1.4}$$

If the inverse of the function $\dot{f}$ exists, it is denoted by

$$\dot{f}^{-1} : \mathbb{R} \to (0, +\infty). \tag{1.5}$$

Note that under the assumption that the function f is strictly convex and differentiable, the *inverse function theorem* in [24, Chapter 9, Theorem 1] holds, which guarantees the existence of $\dot{f}^{-1}$. Moreover, the Legendre-Fenchel transform [25] of the function f is defined below. Additionaly, an important assumption is that the function $\dot{f}$ in (1.4) satisfies that

$$\dot{f}(1) = 0. \tag{1.6}$$

This important distinction can be satisfied for all f-divergences without loss of generality as the f-divergence is invariant to afine transofmations. (see Appendix A.1)

Definition 1.5.2 (Legendre-Fenchel transform [25])**.** *Consider a function $f : \mathcal{I} \to \mathbb{R}$, with $\mathcal{I} \subset \mathbb{R}$. The Legendre-Fenchel transform of the function f, denoted by $f^* : \mathcal{J} \to \mathbb{R}$, is*

$$f^*(t) = \sup_{x \in \mathcal{I}} (tx - f(x)), \tag{1.7}$$

with

$$\mathcal{J} = \{t \in \mathbb{R} : f^*(t) < \infty\}. \tag{1.8}$$

Chapter 2

Brief Introduction To Statistical Machine Learning

Machine learning has become central to many modern applications, influencing diverse sectors such as healthcare, finance, and autonomous driving by enabling early disease detection, fraud prevention, and vehicle control. It is essential to understand when and why these algorithms perform reliably in high-stakes scenarios, which necessarily follows from a clear understanding of the fundamentals behind machine learning. At its core, machine learning relies on pattern recognition, which involves analyzing statistical properties to model trends and build models [2, 26]. The objective of designing such a model is to classify or predict outcomes based on data, which can be viewed as estimation and detection problems in general.

2.1 Machine Learning

The field of pattern recognition has become an important area of research because it enables exploiting statistical properties of data to model patterns and trends that are otherwise not obvious. The core idea of pattern recognition is to tune the parameters of models using a set of data, often referred to as the training set [2]. The objective of designing such a model is to classify or predict outcomes based on data.

This task of learning from data can be traced back to works from before the formalization of machine learning (ML). For example, the works on linear regression, introduced by Legendre and Gauss in the early 19th century, paved the way [27] for learning. Linear regression, a statistical method for fitting linear models and remains central to learning, estimation theory, and many other fields. It uses norms such as least squares or L_1 regularization (lasso) to estimate parameters, drawing from principles of estimation theory.

The notion of machine learning was first formalized by [28] and [29] in their work on decision trees, which was later refined by [30]. Decision trees use conditional hypothesis testing, rooted in the works of [31] and [32], to maximize outcomes at decision nodes. Variants like ID3, C4.5, CART, and CHAID have been developed for classification and regression tasks [33–35]. The Min-Max (MM) algorithm, for instance, minimizes the worst-case loss, reflecting early connections between information theory and machine learning. The k-nearest neighbor (k-NN) algorithm, introduced by [36], overcomes the linear limitations of earlier models by creating nonlinear decision boundaries based on distance metrics like Euclidean distance. While simple to implement, k-NN suffers from computational inefficiency as the sample size

grows, though it benefits from the convergence of conditional probabilities with more data [37].

The perceptron, introduced by [38], marked a significant advancement in linear discriminant models. It uses a weight vector to transform inputs into outputs but is limited to linearly separable data [39, 40]. Building on the perceptron, [41] developed the Support Vector Machine (SVM), which leverages kernel functions to enable non-linear classification and scales well for high-dimensional data [42]. SVMs use quadratic programming to find global optima, offering strong generalization guarantees.

Artificial Neural Networks (ANNs), inspired by biological systems [39, 43, 44], represent another significant class of algorithms. Feed-forward neural networks (FNNs), including multi-layer perceptrons (MLPs), use activation functions and gradient descent for training [2, 45]. Unlike SVMs, ANNs are prone to local optima and require careful validation, though they can be more computationally efficient for prediction tasks [46, 47].

From the long history of learning, the algorithms used to solve the task can be grouped according to the type of training data and information that is available to the algorithm. These classes in which an algorithm may fall are:

(a) *Supervised Learning :* Supervised learning is the branch of machine learning concerned with learning a mapping from inputs or patterns $\mathcal{X}$ to outputs or labels $\mathcal{Y}$. Thus, relying on parings or also referred to as labeled data. Its core idea is to infer a predictive function that generalizes well to unseen data by minimizing the expected loss, typically approximated through empirical risk minimization [48]. Classical examples include linear regression, logistic regression, and support vector machines, which formalize this principle under different loss and regularization settings. Supervised learning underpins most practical applications in classification and regression tasks.

(b) *Unsupervised Learning :* Unsupervised learning deals with discovering patterns, structure, or representations from data without labeled outputs. The main objective is to model the underlying distribution or find latent structure within the observed data. Early methods such as k-means clustering [49] and principal component analysis [50, 51] represent foundational approaches, focusing on data compression and grouping. Modern developments extend these ideas in different areas, some rely on semantic segmentation, such as [52], and others through probabilistic modeling and generative methods that capture complex dependencies in high-dimensional data [53].

(c) *Semi-Supervised Learning :* Semi-supervised learning (SSL) lies between supervised and unsupervised learning, leveraging both labeled and unlabeled data to improve predictive performance. The central idea is that unlabeled data can reveal the data manifold or structure, which guides the learning of a more accurate decision function. Foundational works by [54] formalized key methods such as self-training, graph-based regularization, and consistency-based approaches. SSL has proven especially valuable in scenarios where labeled data are scarce or costly to obtain.

(d) *Reinforcement Learning :* Reinforcement learning (RL) studies how an agent learns to make sequential decisions by interacting with an environment to maximize cumulative rewards. The key principle is learning through trial and error, guided by feedback rather than explicit supervision. Early foundations were laid in [55], which built on ideas from control theory and dynamic programming in [56]. Reinforcement

learning has matured into a comprehensive framework that integrates both value-based and policy-based methodologies, with successes in wide areas such as game playing and robotics.

(*e*) *Ensemble Learning :* Ensemble learning aims to improve predictive accuracy and robustness by combining multiple models instead of relying on a single one. The central idea is to use a mixture of different learning algorithms on the same learning task. The aggregation of different predictors reduces the variance, bias, or both. Pioneering approaches include bagging [57], boosting [58], and stacking [59]. These methods have become fundamental components of modern machine learning, often being present in the state-of-the-art predictive systems.

2.1.1 Supervised Machine Learning

The focus of this thesis is on supervised machine learning. As mentioned above, supervised learning is the branch of machine learning concerned with learning a mapping from inputs or features $\mathcal{X}$ to outputs or labels $\mathcal{Y}$. More formally, it operates on a training dataset consisting of inputoutput pairs $\boldsymbol{z} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ drawn from an unknown joint distribution $P_{X,Y}$. Although always assumed, the data set is not a proper set, as this is just a sequence of pairs where repetition might occur, which has significant implications for the learning task compared to having access to the set of learning pairs in the space. The goal is to infer a predictive function $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by θ, that accurately predicts the output y for new, unseen inputs x. This predictive ability is evaluated in terms of the expected loss, or population risk,

$$\mathsf{L}(\boldsymbol{\theta}) = \int \ell(\boldsymbol{\theta}, x, y)\, \mathrm{d}P_{X,Y}(x, y), \tag{2.1}$$

where ℓ is a task-specific loss function, such as the squared error for regression or the cross-entropy loss for classification. However, due to the unknown distribution $P_{X,Y}$, evaluating the population or true risk is not possible. Hence, the need for the empirical risk minimization framework, which is further explained in Section 2.2.

Inferring a model solely from data without assumptions about its structure is, in general infeasible. To make learning possible, one must introduce inductive biases, that is, assumptions that restrict or guide the space of candidate functions (the hypothesis space) from which the algorithm selects a model [60, 61]. These biases can take the form of structural constraints (e.g., linearity), smoothness assumptions, regularization terms, or architectural choices in neural networks. Without such constraints, the hypothesis space would be too large, and the model could fit the training data perfectly while failing to capture the true underlying relationship between inputs and outputs.

These assumptions play a crucial role in the generalization ability of the model, which refers to how well it performs on new data sampled from the same distribution as the training set. A well-generalizing model captures the essential statistical dependencies in the data rather than memorizing individual training examples. Generalization is typically quantified through the generalization error, defined as the difference between the expected and empirical risks,

Definition 2.1.1 (Generalization Error)**.** *Given a dataset* $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$*, let* L *and* $\mathsf{L}\boldsymbol{z}$ *denote the true (population) and empirical risks, respectively. For any probability*

measure $P \in \triangle(\mathcal{M}, \mathscr{F})$, the generalization gap is defined as

$$\mathrm{Gen}(P) = \int (\mathsf{L}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\, \mathrm{d}P(\boldsymbol{\theta}) \tag{2.2}$$

$$= \mathsf{R}(P) - \mathsf{R}_{\boldsymbol{z}}(P). \tag{2.3}$$

Minimizing this error is central to the design and analysis of supervised learning algorithms, ensuring reliable performance beyond the training set and preventing overfitting.

Supervised learning has evolved from early statistical estimation methods to a broad class of algorithms capable of modeling complex, high-dimensional relationships. Early approaches such as linear regression and logistic regression assume simple parametric forms, emphasizing interpretability and analytical tractability. Later developments, including support vector machines (SVMs) and artificial neural networks (ANNs), extended the ERM framework to nonlinear and nonparametric settings through the use of kernel methods and layered representations. Despite these methodological advances, all supervised learning algorithms rely on the same foundational principle: learning from labeled examples under a set of assumptions that make generalization possible.

Supervised learning involves learning a function from a training set of input-output pairs. The goal is to infer the correct model parameters so that the model can accurately map new, unseen inputs to their corresponding outputs. However, inferring the model without assumptions about its structure is infeasible, necessitating the introduction of learning biasesassumptions that guide the model's hypothesis space [60,61]. This leads to the concept of generalization, which refers to a model's ability to perform well on unseen data, ensuring it captures underlying patterns rather than memorizing training examples. It is often measured by the generalization error and is critical for avoiding overfitting and ensuring reliable performance of the model for field deployment.

In summary, supervised learning has evolved from early statistical methods like linear regression to sophisticated algorithms such as SVMs and ANNs. These advancements, rooted in statistical theory and estimation, have enabled the development of models capable of handling complexity, thus requiring a general description of algorithms for their study.

2.1.2 Statistical Machine Learning

Statistical learning originated from the study of estimation and detection, which seeks to infer population properties from sample data [62, p. 157]. Foundational contributions by Fisher introduced key concepts such as consistency, efficiency, and sufficiency, and led to the development of maximum likelihood estimation (MLE) [1,31], a cornerstone of modern statistical inference. Subsequently, Neyman and Pearson formalized hypothesis testing and defined Type-I and Type-II errors, providing a systematic approach to decision-making under uncertainty [63]. Their work distinguished between testing a single hypothesis, as in Fishers framework, and comparing competing hypotheses through controlled error probabilities. Further developments by Wishart established the role of moment-based methods in parameter estimation [64], while later research extended these principles to dynamic systems through the Wiener and Kalman filters, linking statistical inference to signal estimation [65, 66]. Together,

these advances provided the theoretical and methodological foundations for statistical learning, emphasizing inference, uncertainty quantification, and optimal decision-making.

Statistical machine learning builds on these foundations by integrating statistical inference with computational and algorithmic tools to address problems of prediction, generalization, and uncertainty quantification. It leverages probabilistic modeling and hypothesis testing to extract meaningful patterns from complex datasets in fields such as computer vision, speech recognition, and bioinformatics. At its core, it seeks to model uncertainty to guide algorithm design [2], balancing model complexity and generalization through techniques such as regularization and cross-validation. The field maintains close ties to information theory, which quantifies uncertainty and information in data (e.g., through entropy and mutual information), and to Bayesian methods, which combine prior knowledge with observed evidence for inference and decision-making [67]. By bridging statistical theory and computational algorithms, statistical machine learning provides a rigorous framework for developing interpretable and predictive models capable of learning from data under uncertainty.

In summary, statistical machine learning bridges the gap between statistical theory and computational algorithms, allowing the derivation of guarantees that extend beyond the behaviour of an algorithm over training data and into unseen data, which provides tools to quantify and control the generalization error. Through the integration of concepts from statistics, probability, and information theory, statistical machine learning offers a unified and rigorous framework for understanding, analyzing, and improving data-driven models.

2.2 Empirical Risk Minimization

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 [68]. The application of ERM in the study of theoretical guarantees spans related disciplines such as machine learning [69], information theory [70, 71], and statistics [16, 72]. Classical problems such as classification [73, 74], pattern recognition [75, 76], regression [77, 78], and density estimation [48, 75] can be posed as special cases of the ERM problem [48, 79].

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

$$\boldsymbol{z} \triangleq ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n. \tag{2.4}$$

Let the function $h : \mathcal{M} \times \mathcal{X} \to \mathcal{Y}$ be such that the label assigned to a pattern $x \in \mathcal{X}$ according to the model $\boldsymbol{\theta} \in \mathcal{M}$ is $h(\boldsymbol{\theta}, x)$. Then, given the dataset $\boldsymbol{z}$ in (2.4), the objective is to obtain a model $\boldsymbol{\theta} \in \mathcal{M}$, such that, for all $i \in \{1, 2, \ldots, n\}$, the label assigned to pattern x_i, which is $h(\boldsymbol{\theta}, x_i)$, is "close" to the label y_i. This notion of "closeness" is formalized by the function

$$\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty), \tag{2.5}$$

such that the loss or risk induced by choosing the model $\boldsymbol{\theta} \in \mathcal{M}$ with respect to the labeled pattern (x_i, y_i) is $\ell(h(\boldsymbol{\theta}, x_i), y_i)$. The risk function ℓ is assumed to be nonnegative, and for all $y \in \mathcal{Y}$, it satisfies $\ell(y, y) = 0$. Hence, the smaller the risk $\ell(h(\boldsymbol{\theta}, x_i), y_i)$, the closer the labels $h(\boldsymbol{\theta}, x_i)$ and y_i are.

The *empirical risk* induced by a model $\boldsymbol{\theta}$ with respect to the dataset $\boldsymbol{z}$ in (2.4) is determined by the function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to [0, +\infty)$, which satisfies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\boldsymbol{\theta}, x_i), y_i). \tag{2.6}$$

The ERM problem with respect to the dataset $\boldsymbol{z}$ in (2.4) consists of the optimization problem:

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{2.7}$$

The set of solutions to such a problem is denoted by

$$\mathcal{T}(\boldsymbol{z}) \triangleq \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{2.8}$$

Note that if the set $\mathcal{M}$ is finite, the ERM problem in (2.7) has a solution, and therefore, it holds that $|\mathcal{T}(\boldsymbol{z})| > 0$. Nevertheless, in general, the ERM problem does not always have a solution. That is, there exist choices of the loss function ℓ and the dataset $\boldsymbol{z}$ that yield $|\mathcal{T}(\boldsymbol{z})| = 0$.

2.3 Information Theory

Information theory, pioneered by Claude Shannon in [80], provides a mathematical framework for quantifying information, communication, and uncertainty. Although initially developed for communication systems, its principles have profound implications for estimation, detection, and machine learning. At its core, information theory deals with the encoding, transmission, and decoding of information, which can be viewed analogously to the goals of statistical estimation and machine learning: extracting meaningful patterns from data to make informed decisions.

The connection between estimation and information theory is further highlighted by the concept of Fisher information [1], which quantifies the amount of information that an observable random variable carries about an unknown parameter. Fisher information plays a crucial role in determining the efficiency of estimators and is closely related to the Cramér-Rao bound, which provides a lower bound on the variance of unbiased estimators [81]. Furthermore, the Fisher Information is related to the second derivative of the relative entropy, as shown [82]. These ideas are foundational in both statistics and information theory, bridging the gap between parameter estimation and information quantification. Therefore, ML can be thought of as a communication problem where a sequence of data points is observed from an unknown probability distribution, and the goal is to estimate a model or infer aspects of the source distribution, but unlike classical information theory, the design of the underlying model is fixed, that is, the encoder since it describes the process in the learning problem [11, 13, 83, 84].

Information theory plays a central role in modern machine learning. Entropy measures uncertainty and is used in decision trees for optimal splits [85], while mutual

information identifies relevant features [86]. KL divergence quantifies differences between distributions, enabling techniques such as variational autoencoders [87]. The information bottleneck theory explains how neural networks compress data while preserving relevant information [88]. Tools such as the Kalman filter [66] and the Luenberger observer [89] apply information-theoretic principles to minimize uncertainty in the estimation.

2.3.1 Classical Generalization Bounds in Machine Learning

The theoretical foundations of machine learning are grounded in frameworks that establish fundamental limits on the performance of learning algorithms. These frameworks describe the trade-offs between model complexity, sample size, and generalization performance, that is, how well a model trained on finite data performs on unseen data. Three central concepts in this theory are the probably approximately correct (PAC) learning framework, the VapnikChervonenkis (VC) dimension, and Rademacher complexity. Together, these notions characterize when learning is possible, how many samples are required, and how the complexity of a hypothesis class affects generalization.

The PAC framework, introduced by Valiant in [90], formalizes the idea that a learning algorithm can, with high probability, achieve a small generalization error using only a finite number of samples. A problem is said to be PAC-learnable if there exists an algorithm that, for any $\epsilon > 0$ and $\delta > 0$, outputs a hypothesis with error at most ϵ with probability at least $1 - \delta$, using a finite sample size. This formalism provides probabilistic guarantees on performance and establishes that not all problems are learnable with finite data; for instance, hypothesis classes with infinite VC dimension are not PAC-learnable [26]. A key result connecting generalization error in Definition 2.1.1 and information theory is given by the PAC-Bayes bound [91].

Theorem 2.3.1 (PAC-Bayes Bound [91])**.** *For any prior $Q \in \triangle(\mathcal{M}, \mathscr{F})$ and any posterior $P \in \triangle_Q(\mathcal{M}, \mathscr{F})$, with probability at least $1 - \delta$ over the dataset $\boldsymbol{z}$,*

$$\mathsf{R}(P) \leq \mathsf{R}_{\boldsymbol{z}}(P) + \sqrt{\frac{\mathsf{D}(P\|Q) - \log(\delta)}{2n}}. \tag{2.9}$$

This result shows that the generalization error depends on the empirical risk and the divergence between the posterior and prior distributions over hypotheses, reflecting how much the algorithm learns from data. Although general and widely applicable, PAC bounds are often conservative in practice, particularly for complex models such as deep neural networks.

An alternative but related perspective arises from information-theoretic generalization bounds, which express the generalization gap in terms of mutual information between the data and the learned parameters [92].

Theorem 2.3.2 (Mutual Information Bound [92])**.** *If the loss function $\ell(\theta, X, Y)$ is σ-subgaussian under $P_{(X,Y)^n}$, denoted $P_{\boldsymbol{Z}}$, then*

$$\left|\mathrm{Gen}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}\big)\right| \leq \sqrt{\frac{2\sigma^2}{n}, I(\boldsymbol{Z}; \boldsymbol{\Theta})}. \tag{2.10}$$

The VapnikChervonenkis (VC) dimension, introduced in [41], provides a combinatorial measure of the expressive power of a hypothesis class $\mathcal{H}$. It is defined as the largest

number of data points that can be shattered, that is, classified in all possible labelings, by $\mathcal{H}$. A finite VC dimension implies that the class is PAC-learnable [26].

Theorem 2.3.3 (VC Dimension Bound [41]). *Let $\ell : \mathcal{M} \to [0, 1]$ and suppose $\mathcal{H}$ has VC dimension d. Then, with probability at least $1 - \delta$ over an i.i.d. sample $\boldsymbol{z}$, for all $\boldsymbol{\theta} \in \mathcal{M}$,*

$$\mathsf{L}(\boldsymbol{\theta}) \leq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \sqrt{\frac{8}{n} d \log\left(\frac{2n}{d}\right) + \log\left(\frac{4}{\delta}\right)}. \tag{2.11}$$

While fundamental to statistical learning theory, the VC dimension does not account for the specific structure or inductive biases of the model nor the influence of the learning algorithm, which can lead to better generalization in practice. For example, the VC dimension of deep neural networks with ReLU activations is typically infinite, which renders the bounds vacuous [76]. Moreover, it does not capture the structural or inductive properties of models that contribute to their ability to generalize well in practice.

The Rademacher complexity [93–95] refines the VC-based approach by providing a data-dependent measure of model capacity. It quantifies the ability of a hypothesis class to fit random noise, yielding tighter, empirically grounded generalization bounds. However, while more precise in principle, Rademacher complexity is difficult to compute for high-dimensional or non-linear models, and it does not fully explain the generalization behavior observed in deep networks.

The PAC, VC, and Rademacher frameworks collectively highlight the trade-off between model complexity and generalization. Simpler models, characterized by low VC or Rademacher complexity, tend to generalize better but may underfit, while overly complex models risk overfitting. This principle underlies classical model selection and regularization strategies such as weight decay and dropout, which effectively control model capacity [2, 96]. Nevertheless, the generalization performance of deep neural networks often exceeds the predictions derived from these classical theories [97], indicating the need for new complexity measures that account for structure, optimization dynamics, and implicit regularization effects.

2.3.2 Information Stability Methods

Unfortunately, ERM is prone to training data memorization, a phenomenon also known as overfitting [98–100]. For that reason, ERM is often regularized in order to provide generalization guarantees [101–104]. 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 [91, 105] and PAC learning [90, 106, 107], the prior knowledge over the set of models can be described by a reference probability measure. More general references can be adapted as proved in [84, 108] for the case of σ-finite measures. 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 [109]. In either case, the solution to the regularized ERM problem can be cast as a probability distribution over the set of models.

The PAC and Bayesian frameworks, as discussed in [91] and [106], address the problem in (2.7) by constructing probability measures, conditioned on the dataset $\boldsymbol{z}$, from which models are randomly sampled. In this context, finding probability measures that are minimizers of the ERM problem in (2.7) over the set of all probability measures that can be defined on the measurable space $(\mathcal{M}, \mathscr{F})$, which is denoted

by $\triangle(\mathcal{M})$, requires a metric that enables assessing the goodness of the probability measure. From this perspective, the underlying assumption in the remainder of this work is that the functions h and ℓ in (2.6) 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 $g_{x,y}(\boldsymbol{\theta}) = \ell(h(\boldsymbol{\theta}, x), y)$, is measurable with respect to the Borel measurable spaces $(\mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, where $\mathscr{F}$ and $\mathscr{B}(\mathbb{R})$ are respectively the Borel σ-fields on $\mathcal{M}$ and $\mathbb{R}$. Under these assumptions, a common metric is the notion of expected empirical risk.

Definition 2.3.1 (Expected Empirical Risk)**.** *The expectation of the empirical risk* $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})$ *in* (2.6)*, when* $\boldsymbol{\theta}$ *is sampled from a probability measure* $P \in \triangle(\mathcal{M})$*, is determined by the functional* $\mathsf{R}_{\boldsymbol{z}} : \triangle(\mathcal{M}) \to [0, +\infty)$*, such that*

$$\mathsf{R}_{\boldsymbol{z}}(P) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}). \tag{2.12}$$

However, ERM is prone to overfitting [98–100], which affects the generalization capability of the selected model [2, 110, 111]. To remediate this phenomenon, the solution of ERM must exhibit a small sensitivity to variations in the training dataset, which is often obtained via regularization [84, 102–104, 112]. In this case, regularization is achieved by adding to the expected empirical risk a *statistical distance* from a reference measure to the optimization measure. 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 [11, 48, 92, 113]. 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 [11, 92, 113, 114] and the case in which it is a σ-finite measure [84, 108, 115]. While in both cases, the solution is unique and corresponds to a Gibbs probability measure, the existence of the solution is ensured only in the case in which the reference measure is a probability measure [84].

2.4 Type-I ERM-RER Problem

The classical ERM-RER problem formulation uses the relative entropy of the optimization measure with respect to the reference measure. In this work, two distinct problem formulations are used in the study of asymmetry. Thus, the classical ERM-RER will be referred to as Type-I ERM-RER to avoid ambiguities and is formally defined hereunder. The Type-I ERM-RER problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M})$ and a real $\lambda \in (0, \infty)$. The measure Q is referred to as the *reference measure* and λ as the *regularization factor*. The Type-I ERM-RER problem, with parameters Q and λ, is given by the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(P \| Q), \tag{2.13}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (2.12), and the optimization domain is

$$\triangle_Q(\mathcal{M}) \triangleq \{P \in \triangle(\mathcal{M}) : P \ll Q\}, \tag{2.14}$$

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 (2.13) is the Gibbs probability measure reported in [92, 108] and [11]. To introduce such a measure, consider the

function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ that satisfies for all $t \in \mathbb{R}$,

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp(t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})\right), \tag{2.15}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (2.6). Using this notation, the solution to the Type-I ERM-RER problem in (2.13) is presented by the following lemma.

Lemma 2.4.1 ([84, Theorem 3]). *The solution to the optimization problem in* (2.13) *is a unique probability measure, denoted by* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$*, which satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right), \tag{2.16}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ *is defined in* (2.6) *and the function* $K_{Q,\boldsymbol{z}}$ *is defined in* (2.15)*.*

The problem in (2.13) exhibits a trivial solutions when the functional $\mathsf{R}_{\boldsymbol{z}}$ is such that for all $P \in \triangle_Q(\mathcal{M})$. In such a case, the solution is unique and equal to the probability measure Q, independently of the parameter λ. To avoid this trivial case, the notion of separability of the empirical risk function with respect to the measure Q is borrowed from [84]. A separable empirical risk function with respect to a given probability measure P is defined as follows.

Definition 2.4.1 (Definition 5 in [84]). *The empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (2.6) *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* $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$*,*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) < c < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) < \infty. \tag{2.17}$$

A nonseparable empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) 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

$$P(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = a\}) = 1. \tag{2.18}$$

When the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is nonseparable with respect to all measures in $P \in \triangle_Q(\mathcal{M})$, the trivial case described above is observed.

Optimization problems with f-divergence regularization have been explored before in [116] and [117] for the discrete case. In [118], the problem of non-exponentially weighted aggregation is studied, and in the context of ERM, there are works on [20]. Such a problem involves an ERM with f-divergence regularization (ERM-fDR) similar to the one studied in this work.

Chapter 3

ERM-RER Asymmetry

The effect of relative entropy asymmetry is analyzed in the context of empirical risk minimization (ERM) with relative entropy regularization (ERM-RER). The relative entropy regularization of the reference measure with respect to the measure to be optimized is considered, that is, the Type-II ERM-RER, and compared to the Type-I ERM-RER regularization in Section 2.4. 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 solution's support to collapse into the support of the reference measure, introducing a strong inductive bias that can overshadow 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.

3.1 Introduction

Despite the many merits of the Type-I ERM-RER formulation in Section 2.4, 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 in high-dimensional spaces. While such priors ensure the inclusion of high-performing models, the resulting mutual absolute continuity between the classical ERM-RER solution and the chosen prior also assigns 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 of obtaining generalization guarantees in the form of bounds [101]. Nonetheless, such bounds are often hard to calculate and are not always informative when evaluated in practical settings [84, 109, 119–124].

In view of these, exploring the asymmetry of relative entropy is of particular interest in 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 choose models selectively outside the support of the prior. The

problem of ERM with a general f-divergence regularization has been explored in [116] and [117] in the case of a finite countable set of models and recently extended to uncountable sets of models in [118] and [125]. The authors in [116–118,125] constrain the optimization domains to sets of measures that are mutually absolutely continuous with respect to the reference probability measure. 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 existing results in [116–118], 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 3.3.3, 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, the performance of the Type-II regularization is comparable and displays promising properties that warrant further research.

This chapter presents the solution to the Type-II ERM-RER optimization problem through a new method of proof. Unlike the Type-I ERM-RER case, where the optimization is carried out in a space that allows the direct use of Gâteaux derivatives, the Type-II formulation operates in a more general space where such differentiation cannot be applied straightforwardly. In particular, mutual absolute continuity between the measures involved is not imposed. Nonetheless, mutual absolute continuity is exhibited by the solution as a consequence of the structure of the problem. To address this difficulty, the proof proceeds by first formulating and solving an auxiliary optimization problem defined in a space compatible with the classical Type-I analysis. It is then shown that the solution to this auxiliary problem coincides with the solution of the original Type-II problem, thereby establishing the desired result through an indirect but rigorous argument. 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 [126–128]. The remainder of the chapter is organized as follows. Section 3.2 presents the ERM-RER problem and its two variations: Type-I and Type-II. The main contribution of this chapter, which is the solution to the Type-II ERM-RER problem, is presented in Section 3.3. This section also presents key properties of the solution. Section 3.3.14 uses these properties to characterize the expected empirical risk. Section 3.3.3 studies the equivalence between Type-I and Type-II ERM-RER problems. This work is concluded by Section 3.5, with some final remarks.

3.2 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 (2.4), the Type-II ERM-RER problem, with parameters Q

and λ, consists of the following optimization problem:

$$\min_{P \in \bigtriangledown_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P), \tag{3.1}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (2.12), and the optimization domain is

$$\bigtriangledown_Q(\mathcal{M}) \triangleq \{P \in \triangle(\mathcal{M}) : Q \ll P\}, \tag{3.2}$$

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

The difference between Type-I and Type-II ERM-RER problems lies in 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 (2.14). Alternatively, in the Type-II ERM-RER problem, the optimization domain consists of probability measures defined on the Borel measurable space $(\mathcal{M}, \mathscr{F})$, with the additional condition that the reference measure Q must be absolutely continuous with respect to them. This corresponds to the set denoted as $\bigtriangledown_Q(\mathcal{M})$ in (3.2). From this perspective, the techniques used in [84] for solving the Type-I ERM-RER no longer hold. As shown in the next section, a new technique is used for solving the Type-II ERM-RER.

3.3 Type-II ERM-RER Solution & Results

The solution of the Type-II ERM-RER problem in (3.1) is presented in the following theorem.

Theorem 3.3.1. *If there exists a real β such that*

$$\beta \in \{t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\}, \tag{3.3a}$$

and

$$\int \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{3.3b}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (2.6), and λ and Q the parameters of the optimization problem in (3.1), 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

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}. \tag{3.4}$$

Before introducing the proof of Theorem 3.3.1, two important results are presented. The first result consists of the solution to the optimization problem in (3.1) when the optimization domain is restricted to

$$\bigcirc_Q(\mathcal{M}) \triangleq \bigtriangledown_Q(\mathcal{M}) \cap \triangle_Q(\mathcal{M}), \tag{3.5}$$

where the sets $\triangle_Q(\mathcal{M})$ and $\bigtriangledown_Q(\mathcal{M})$ are defined in (2.14) and (3.2), respectively. Such an ancillary problem can be formulated as follows:

$$\min_{P \in \bigcirc_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P). \tag{3.6}$$

The solution to the problem in (3.6) is described by the following lemma.

Lemma 3.3.1. *The solution to the optimization problem in* (3.6) *is unique and identical to the probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4).

Proof: The proof is presented in Appendix B.1. ■

The second result consists of comparing the optimal values resulting from the optimization problems in (3.1) and (3.6), as shown hereunder.

Lemma 3.3.2. *The optimization problems in* (3.1) *and* (3.6) *satisfy*

$$\min_{P \in \bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P) \geq \min_{P \in \bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P). \tag{3.7}$$

Proof: The proof is presented in Appendix B.2. ■

Lemma 3.3.2 unveils the fact that the objective function in (3.1), 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 (2.8) lies outside the support of Q. Using these results, the proof of Theorem 3.3.1 is as follows.

Proof of Theorem 3.3.1: The proof follows by observing that from (3.5), it holds that

$$\bigcirc_Q(\mathcal{M}) \subseteq \bigtriangledown_Q(\mathcal{M}). \tag{3.8}$$

Hence, from (3.8), it follows that

$$\min_{P \in \bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P) \leq \min_{P \in \bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P). \tag{3.9}$$

From the inequalities in (3.7) and (3.9), it follows that

$$\min_{P \in \bigtriangledown_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P) = \min_{P \in \bigcirc_Q} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P). \tag{3.10}$$

Thus, the measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) is the solution of the optimization problem in (3.1), which completes the proof of Theorem 3.3.1. ■

Lemma 3.3.2 implies that the solution to the optimization problem in (3.1) is in the set $\bigcirc_Q(\mathcal{M})$ in (3.4). A consequence of this observation is the following corollary.

Corollary 3.3.3. *The probability measures* Q *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *are mutually absolutely continuous.*

Corollary 3.3.3 also follows from Theorem 3.3.1 by observing that the solution to the Type-II ERM-RER problem in (3.1) 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 3.3.3 conveys the fact that there does not exist a dataset that

can overcome the inductive bias induced by the reference measure Q. That is, 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 (3.1) does not restrict the solution to be absolutely continuous with respect to the reference measure Q. However, Theorem 3.3.1 shows that the support of the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) 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.

3.3.1 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 (3.1), 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 (3.3a) and (3.3b). From Theorem 3.3.1, specifically from the uniqueness of the solution to (3.1), 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

$$\bar{K}_{Q,\boldsymbol{z}} : \mathcal{A}_{Q,\boldsymbol{z}} \to \mathcal{C}_{Q,\boldsymbol{z}}, \tag{3.11a}$$

which satisfies,

$$\bar{K}_{Q,\boldsymbol{z}}(\lambda) = \beta. \tag{3.11b}$$

The function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11) is referred to as the *normalization function*. This is essentially 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 (3.4) can be re-written for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, as

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{3.12}$$

which together with (3.3b), implies that the function $\bar{K}_{Q,\boldsymbol{z}}$ ensures that $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) is a probability measure.

The analysis of the normalization function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11) relies on the analysis of its functional inverse, denoted by $\bar{K}^{-1}_{Q,\boldsymbol{z}} : \mathcal{C}_{Q,\boldsymbol{z}} \to \mathcal{A}_{Q,\boldsymbol{z}}$, which can be defined by noticing that

$$1 = \int \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{3.13a}$$

$$= \int \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{3.13b}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (2.6), and λ and Q the parameters of the optimization problem in (3.1). More specifically, from (3.11b), it follows that $\lambda = \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)$; and from (3.13b), it follows that

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) = \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta})}. \tag{3.14}$$

Note that while the function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11) is implicitly defined, as a closed-form expression is not provided, its functional inverse $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ is explicitly defined in (3.14). Such functional inverse exists from the fact that $\bar{K}_{Q,\boldsymbol{z}}$ is a bijection.

The purpose of the remainder of this section is to provide a characterization of the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$. Some mathematical objects are introduced to characterize these sets. Given a real $\delta \in [0,\infty)$, consider the Rashomon set [129], $\mathcal{L}_{\boldsymbol{z}}(\delta)$, defined as follows

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leq \delta\}. \tag{3.15}$$

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

$$\delta^{\star}_{Q,\boldsymbol{z}} \triangleq \inf\{\delta \in [0,\infty) : Q(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0\}, \tag{3.16}$$

and

$$\lambda^{\star}_{Q,\boldsymbol{z}} \triangleq \inf \mathcal{A}_{Q,\boldsymbol{z}}. \tag{3.17}$$

Let also $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ be the level set of the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) for the value $\delta^{\star}_{Q,\boldsymbol{z}}$. That is,

$$\mathcal{L}^{\star}_{Q,\boldsymbol{z}} \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}\right\}. \tag{3.18}$$

Using the objects defined above, the following lemma introduces one of the main properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11).

Lemma 3.3.4. *The function $\bar{K}_{Q,\boldsymbol{z}}$ in* (3.11) *is strictly increasing and continuous.*

Proof: The proof is presented in Appendix B.3. ■

The following lemma characterizes the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$ in (3.11a), which are the domain and codomain of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11b).

Lemma 3.3.5. *The set $\mathcal{A}_{Q,\boldsymbol{z}}$ in* (3.11a) *is either empty or an interval of the form*

$$\mathcal{A}_{Q,\boldsymbol{z}} = \begin{cases} [\lambda_{Q,\boldsymbol{z}},\infty) & \text{if } \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})-\delta^{\star}_{Q,\boldsymbol{z}}}\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty \\ (0,\infty) & \text{otherwise,} \end{cases} \tag{3.19}$$

where

$$\lambda_{Q,\boldsymbol{z}} = \bar{K}_{Q,\boldsymbol{z}}^{-1}\big(-\delta^{\star}_{Q,\boldsymbol{z}}\big), \tag{3.20}$$

with $\bar{K}_{Q,\boldsymbol{z}}^{-1}$ in (3.14)*, the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in* (2.6)*, and $\delta^{\star}_{Q,\boldsymbol{z}}$ is defined in* (3.16)*. Moreover, the set $\mathcal{C}_{Q,\boldsymbol{z}}$ in* (3.11a) *is either empty or an interval of the form*

$$\mathcal{C}_{Q,\boldsymbol{z}} = \begin{cases} \left[-\delta^{\star}_{Q,\boldsymbol{z}},\infty\right) & \text{if } \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})-\delta^{\star}_{Q,\boldsymbol{z}}}\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \\ \left(-\delta^{\star}_{Q,\boldsymbol{z}},\infty\right) & \text{otherwise.} \end{cases} \tag{3.21}$$

Proof: The proof is presented in Appendix B.4. ■

Lemma 3.3.5 shows that the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{C}_{Q,\boldsymbol{z}}$ in (3.11a) are convex sets (intervals). This convexity is crucial for analyzing how the choice of λ influences whether the Type-II ERM-RER problem in (3.1) 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 (3.1) (Theorem 3.3.1). Moreover, if such λ is increased, the resulting Type-II ERM-RER problem still possesses a solution, which is formalized by the following corollary.

Corollary 3.3.6. *If the Type-II ERM-RER problem in* (3.1) *possesses a solution, then, the following problem*

$$\min_{P\in\nabla_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \alpha \mathsf{D}(Q\|P), \tag{3.22}$$

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

Additionally, Lemma 3.3.5 allows identifying how small λ in (3.1) can be, such that the Type-II ERM-RER problem in (3.1) still possesses a solution. The regularization factor λ can be made arbitrarily close to zero in some cases, as shown hereunder.

Corollary 3.3.7. *If the set* $\mathcal{M}$ *is finite, then the set* $\mathcal{A}_{Q,\boldsymbol{z}}$ *in* (3.11a) *is* $(0,\infty)$.

Corollary 3.3.7 follows by noticing that if the set $\mathcal{M}$ is finite, the subset $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (3.18) satisfies $Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big) > 0$. Thus, the integral in (3.21) 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 3.3.4 and Lemma 3.3.5 is the following corollary.

Corollary 3.3.8. *If the real value* $\delta^{\star}_{Q,\boldsymbol{z}} = 0$*, with* $\delta^{\star}_{Q,\boldsymbol{z}}$ *in* (3.16)*, then the function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (3.11b) *is strictly positive.*

Appendix E.1 introduces some examples to illustrate particular cases in which the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is open or semi-open. This section is closed by leveraging Lemma 3.3.5 for presenting a key property of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11b).

Lemma 3.3.9. *The function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (3.11) *satisfies*

$$\lim_{\lambda\to{\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\delta^{\star}_{Q,\boldsymbol{z}}, \tag{3.23}$$

where $\delta^{\star}_{Q,\boldsymbol{z}}$ *and* $\lambda^{\star}_{Q,\boldsymbol{z}}$ *are defined in* (3.16) *and* (3.17)*, respectively.*

Proof: The proof is presented in Appendix B.5. ∎

The limit in (3.23) is determined by the set of models in the support of the prior with the lowest empirical risk determined by the choice of loss function ℓ and function f in (2.6).

3.3.2 Type-II ERM-RER Properties

Note that from Theorem 3.3.1, 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 (3.4). The following corollary formalizes this observation.

Lemma 3.3.10. *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* (2.6)*, the Radon-Nikodym derivative* $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (3.4) *satisfies*

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2) \leq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1), \tag{3.24}$$

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 B.6. ∎

The intuition that follows from Lemma 3.3.10 is that under the assumption that the ERM problem in (2.7) possesses a solution in the support of the reference measure. That is, the maximum of the function $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (3.4) is achieved by the models

in $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} Q$, provided that it is not empty, where $\mathcal{T}(\boldsymbol{z})$ is defined in (2.8). 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 (2.6). This property is similar to that of the solution to the Type-I ERM-RER problem in (2.16), as established in [84, Corollary 1].

The Radon-Nikodym derivative $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (3.4) is always finite and strictly positive. This observation is formalized in the following lemma.

Lemma 3.3.11. *The Radon-Nikodym derivative* $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (3.4) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$0 < \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leq \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} < \infty, \tag{3.25}$$

where the function $\bar{K}_{Q,\boldsymbol{z}}$ *and the real* $\delta^{\star}_{Q,\boldsymbol{z}}$ *are defined in* (3.11a) *and* (3.16)*, respectively. The equality holds if and only if* $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$*.*

Proof: The proof is presented in Appendix B.7. ∎

Asymptotes of the Radon-Nikodym Derivative

In the asymptotic regime, when the regularization factor λ in (3.1) 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 3.3.12. *The Radon-Nikodym derivative* $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (3.4) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\lim_{\lambda \to \infty} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 1. \tag{3.26}$$

Proof: The proof is presented in Appendix B.8. ∎

Lemma 3.3.12 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 (2.16), also exhibits a similar behavior [84].

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 (3.4) exhibits the following behavior.

Lemma 3.3.13. *If* $Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big) > 0$*, with the set* $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ *in* (3.18)*, then the Radon-Nikodym derivative* $\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (3.4) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big)} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}\}}. \tag{3.27}$$

Alternatively, if $Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big) = 0$ *and* $\lambda^{\star}_{Q,\boldsymbol{z}}$ *in* (3.17) *satisfies* $\lambda^{\star}_{Q,\boldsymbol{z}} = 0$*, then for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*, it holds that*

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \begin{cases} \infty & \text{if } \boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \\ 0 & \text{otherwise.} \end{cases} \tag{3.28}$$

Conversely, if $Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big) = 0$ *and* $\lambda^{\star}_{Q,\boldsymbol{z}}$ *in* (3.17) *satisfies* $\lambda^{\star}_{Q,\boldsymbol{z}} > 0$*, then for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*, it holds that*

$$\lim_{\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda^{\star}_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}. \tag{3.29}$$

Proof: The proof is presented in Appendix B.9. ∎

Lemma 3.3.13 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 (3.16), 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 (2.8). 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 (2.7).

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 [84, Lemma 6]. This aligns with the observation that as λ decreases, the optimization problems in (2.13) and (3.1) exhibit a weaker relative entropy constraint. A stronger result follows from Lemma 3.3.13 and is presented in the following lemma.

Lemma 3.3.14. *If* $\lambda^{\star}_{Q,\boldsymbol{z}}$ *in* (3.17) *satisfies* $\lambda^{\star}_{Q,\boldsymbol{z}} = 0$*, then the measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *and the set* $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ *in* (3.18) *satisfy*

$$\lim_{\lambda \to 0^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 1. \tag{3.30}$$

Alternatively, if $\lambda^{\star}_{Q,\boldsymbol{z}} > 0$*, then the measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *and the set* $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ *in* (3.18) *satisfy*

$$\lim_{\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = 0. \tag{3.31}$$

Proof: The proof is presented in Appendix B.10. ∎

Lemma 3.3.14 shows that indeed when the regularization factor approaches zero from the right, the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) concentrates in the set of models that induce the minimum empirical risk in $\operatorname{supp} Q$.

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 (3.4). 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 (2.12).

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 (3.11b).

Lemma 3.3.15. *The probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *satisfies*

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{3.32}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ *and the function* $\bar{K}_{Q,\boldsymbol{z}}$ *are defined in* (2.12) *and* (3.11b)*, respectively.*

Proof: The proof is presented in Appendix B.11. ∎

Lemma 3.3.15 characterizes the expected empirical risk of the Type-II ERM-RER solution and establishes a direct connection to the regularization factor λ. Unlike Type-I ERM-RER, in Type-II, λ serves as an explicit upper bound to the expected empirical risk if there exists a model $\boldsymbol{\theta}^\star \in \operatorname{supp} Q$, such that its empirical risk is zero. Thus, providing a clear interpretation: the choice of λ directly controls and bounds the average expected empirical risk. This property makes Type-II ERM-RER advantageous for risk management through the explicit selection of λ. Additionally, Lemma 3.3.15 highlights that the function $r : (0,\infty) \to (0,\infty)$ such that $r(\lambda) = \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$, with Q and $\boldsymbol{z}$ fixed, inherits all properties of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11b). The following lemma formalizes this observation.

Lemma 3.3.16. *The expected empirical risk* $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$*, with the functional* $\mathsf{R}_{\boldsymbol{z}}$ *in* (2.12) *and the measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4)*, 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* (2.6) *is separable with respect to the probability measure* Q*.*

Proof: The proof is presented in Appendix B.12. ∎

Another application of Lemma 3.3.15 is the characterization of the distance of the Type-II ERM-RER solution to any probability measure $P \in \triangle_Q(\mathcal{M})$, which is formalized by the following theorem.

Theorem 3.3.2. *The probability* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *satisfies for all* $P \in \triangle_Q(\mathcal{M})$*,*

$$\mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda \int \frac{\mathrm{d}P}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - \lambda. \tag{3.33}$$

Proof: The proof follows from observing that

$$\begin{aligned}
\lambda \int \frac{\mathrm{d}P}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - \lambda &= \lambda \int \frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - \lambda && (3.34)\\
&= \lambda \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\,\mathrm{d}P(\boldsymbol{\theta}) - \lambda && (3.35)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)\,\mathrm{d}P(\boldsymbol{\theta}) - \lambda && (3.36)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) - \big(\lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda)\big) && (3.37)\\
&= \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big), && (3.38)
\end{aligned}$$

where (3.34) follows from Corollary 3.3.3, (3.35) follows from (3.4), and (3.38) follows from Lemma 3.3.15. ∎

The importance of the expression provided in Theorem 3.3.2 is that it can also be used to characterize the generalization error of the Type-II ERM-RER solution when

used with the results presented in [18, Lemma 3]. This will be further developed in general form in Section 5.4.

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 (3.1) 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 $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$; and the relative entropy $\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$.

Lemma 3.3.17. *The functional* $\mathsf{R}_{\boldsymbol{z}}$ *defined in* (2.12) *and the measures* Q *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *satisfy*

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geq \lambda\Big(\exp\Big(\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)\Big) - 1\Big). \tag{3.39}$$

Proof: The proof is presented in Appendix B.13. ∎

Note that $\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \geq 0$ in (3.39), which leads to the observation that

$$\Big(\exp\Big(\mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)\Big) - 1\Big) \geq 0. \tag{3.40}$$

Hence, from Lemma 3.3.17, 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 3.3.18. *The probability measures* Q *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *satisfy*

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \leq \mathsf{R}_{\boldsymbol{z}}(Q), \tag{3.41}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ *is defined in* (2.12)*, and equality holds if and only if the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (2.6) *is nonseparable.*

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

Lemma 3.3.19. *The probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *satisfies*

$$\delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) < \lambda + \delta^{\star}_{Q,\boldsymbol{z}}, \tag{3.42}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ *is defined in* (2.12) *and* $\delta^{\star}_{Q,\boldsymbol{z}}$ *is defined in* (3.16)*. Moreover, equality holds if and only if the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (2.6) *is nonseparable.*

Proof: The proof is presented in Appendix B.14. ∎

The bounds presented in Lemma 3.3.19 highlight that the regularization parameter λ in (3.1) governs the increase of the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ with respect to its minimum, i.e, $\delta^{\star}_{Q,\boldsymbol{z}}$ in (3.16). Moreover, the lower bound is tight for the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) in the asymptotic regime when λ decreases to zero from the right, as shown hereunder.

Lemma 3.3.20. *The probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *satisfies*

$$\lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda^{\star}_{Q,\boldsymbol{z}} + \delta^{\star}_{Q,\boldsymbol{z}}, \tag{3.43}$$

where $\delta^{\star}_{Q,\boldsymbol{z}}$ *is defined in* (3.16) *and the functional* $\mathsf{R}_{\boldsymbol{z}}$ *is defined in* (2.12).

Proof: From Lemma 3.3.15, it holds that

$$\begin{aligned}
&\lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\
&\quad= \lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \lambda - \lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) &(3.44)\\
&\quad= \lambda^{\star}_{Q,\boldsymbol{z}} - \lim_{\lambda\to\lambda^{\star}_{Q,\boldsymbol{z}}{}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) &(3.45)\\
&\quad= \lambda^{\star}_{Q,\boldsymbol{z}} + \delta^{\star}_{Q,\boldsymbol{z}}, &(3.46)
\end{aligned}$$

where equality (3.46) follows from Lemma 3.3.9. This completes the proof. ■

Finally, note that the functional $\mathsf{R}_{\boldsymbol{z}}$ in (2.12) is nonnegative. This observation together with Lemma 3.3.15 lead to a new property for the function $\bar{K}_{Q,\boldsymbol{z}}$ in (3.11b), which is stated by the following corollary

Corollary 3.3.21. *The function* $\bar{K}_{Q,\boldsymbol{z}}$ *in* (3.11b) *satisfies, for all* $t > \lambda^{\star}_{Q,\boldsymbol{z}}$, *with* $\lambda^{\star}_{Q,\boldsymbol{z}}$ *in* (3.17),

$$\bar{K}_{Q,\boldsymbol{z}}(t) \leq t. \tag{3.47}$$

(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 (3.4), with respect to the Type-II ERM-RER problem in (3.1). Such a guarantee is presented using the notion of (δ,ϵ)-optimality introduced in [84, Definition 6].

Definition 3.3.1 ([84, 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* (3.15) *satisfies*

$$P(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 1-\epsilon. \tag{3.48}$$

The (δ,ϵ)-optimality guarantee in Definition 3.3.1 ensures that with probability at least $1-\epsilon$, sampling models from $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) yields models that induce empirical risks not greater than δ. The following theorem presents a (δ,ϵ)-optimality guarantee for the Type-II ERM-RER solution.

Theorem 3.3.3. *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* (3.16), *there always exists a* $\lambda \in (0,\infty)$, *such that the probability measure* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *is* (δ,ϵ)*-optimal.*

Proof: The proof is presented in Appendix B.15. ■

3.3.3 Relative Entropy Asymmetry via Empirical Risk

This section presents a connection between the Type-I ERM-RER in (2.13) and Type-II ERM-RER in (3.1) 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 (2.13) is identical to the solution of the following problem:

$$\min_{P \in \bigtriangledown_Q(\mathcal{M})} \int \mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \mathsf{D}(Q\|P), \tag{3.49}$$

with λ and Q in (2.13); and the solution to the optimization problem in (3.1) is identical to the solution of the following problem:

$$\min_{P \in \bigtriangleup_Q(\mathcal{M})} \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \mathsf{D}(P\|Q), \tag{3.50}$$

with λ and Q in (3.1). The main result of this section is presented in the following theorem.

Theorem 3.3.4. *If the problems in* (2.13) *and in* (3.1) *have solutions, then*

$$\begin{aligned} &\min_{P \in \bigtriangledown_Q(\mathcal{M})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \lambda \mathsf{D}(Q\|P) \\ &\quad = \min_{P \in \bigtriangleup_Q(\mathcal{M})} \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \mathsf{D}(P\|Q), \end{aligned} \tag{3.51a}$$

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

$$\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) = \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big), \tag{3.51b}$$

and

$$\begin{aligned} &\min_{P \in \bigtriangleup_Q(\mathcal{M})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \lambda \mathsf{D}(P\|Q) \\ &\quad = \min_{P \in \bigtriangledown_Q(\mathcal{M})} \int \mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) + \mathsf{D}(Q\|P), \end{aligned} \tag{3.52a}$$

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

$$\mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) = \frac{\lambda}{\exp\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - K_{Q,\boldsymbol{z}}\big(-\frac{1}{\lambda}\big)\Big)} - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{3.52b}$$

where the functions $K_{Q,\boldsymbol{z}}$ *and* $\bar{K}_{Q,\boldsymbol{z}}$ *are defined in* (2.15) *and* (3.11)*, respectively.*

Proof: Denote by $\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the solution to the optimization problem in (3.50). From Lemma 2.4.1, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\begin{aligned} &\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\ &= \frac{\exp(-\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}))}{\int \exp(-\mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\nu})) \, \mathrm{d}Q(\boldsymbol{\nu})} && (3.53) \\ &= \frac{\exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big)}{\int \exp\Big(\log\Big(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\Big)\Big) \mathrm{d}Q(\boldsymbol{\nu})} && (3.54) \\ &= \frac{\Big(\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \mathrm{d}Q(\boldsymbol{\nu})\Big)^{-1}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && (3.55) \end{aligned}$$

$$= \frac{\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{3.56}$$

$$= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{3.57}$$

$$= \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{3.58}$$

where (3.54) follows from (3.51b); (3.56) follows from (3.14); (3.57) follows from (3.11b); and (3.58) follows from Theorem 3.3.1. This completes the proof of (3.52a).

Similarly, denote by $\tilde{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the solution to the optimization problem in (3.49). From Theorem 3.3.1, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

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

$$= \frac{\lambda}{\mathsf{W}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{3.59}$$

$$= \frac{\lambda}{\frac{\lambda}{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)} - \bar{K}_{Q,\boldsymbol{z}}(\lambda) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{3.60}$$

$$= \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right) \tag{3.61}$$

$$= \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{3.62}$$

where (3.59) follows from (3.12); (3.60) follows from (3.52b); and (3.62) follows from Lemma 2.4.1. This completes the proof of (3.51a). ■

Theorem 3.3.4 establishes an equivalence between the regularization of Type-I and Type-II. More specifically, Theorem 3.3.4 highlights that by modifying the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (3.51b) using the function $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ in (3.51b), the Type-II ERM-RER problem in (3.1) can be solved by solving the Type-I ERM-RER problem in (3.50). 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 3.3.1 and Lemma 2.4.1. That is,

$$\operatorname{supp} P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} = \operatorname{supp} Q = \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}. \tag{3.63}$$

The remainder of the section focuses on the transformation from Type-I to Type-II. From the notion of *log-empirical* risk in (3.51b), the *expected log-empirical risk* is defined as follows.

Definition 3.3.2 (Expected Log-Empirical Risk)**.** *Given the dataset* $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ *in* (2.4) *and the log-empirical risk function* $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ *in* (3.51b)*, let the functional* $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda} : \triangle(\mathcal{M}) \to \mathbb{R}$ *be such that*

$$\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) = \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}P(\boldsymbol{\theta}). \tag{3.64}$$

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 (3.49) can be rewritten as follows

$$\min_{P\in\triangle_Q(\mathcal{M})} \quad \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) + \mathsf{D}(P\|Q), \tag{3.65}$$

with λ and Q being parameters of the Type-I and Type-II ERM-RER problems in (2.13) and (3.1). The Type-I - Type-II relation in Theorem 3.3.4 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 3.3.22. *The functional* $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ *in* (3.64) *and the probability measures* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *and* Q *in* (3.4) *satisfy*

$$\log(\lambda) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) \tag{3.66}$$

$$= \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q) - \mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{3.67}$$

Proof: The proof is presented in Appendix B.16. ■

Sensitivity of the Log-Empirical Risk

The sensitivity of the expected empirical risk, as presented in [84, Definition 7], is defined as follows.

Definition 3.3.3 (Sensitivity of the Expected Empirical Risk)**.** *Consider the functional* $\mathsf{R}_{\boldsymbol{z}}$ *in* (2.12) *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*

$$\mathsf{S}_{Q,\lambda}(\boldsymbol{z}, P) = \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{3.68}$$

where the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *is defined in* (2.16)*. 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 (3.51b) is defined as follows.

Definition 3.3.4 (Sensitivity of the Expected Log-Empirical Risk)**.** *Consider the functional* $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ *in* (3.64) *and let* $\bar{\mathsf{S}}_{Q,\lambda} : (\mathcal{X}\times\mathcal{Y})^n \times \bigtriangledown_Q(\mathcal{M}) \to \mathbb{R}$ *be a functional such that*

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{3.69}$$

where the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *is in* (3.4)*. The sensitivity of the expected log-empirical risk due to a deviation from* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *to* P *is* $\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P)$*.*

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

Lemma 3.3.23. *The sensitivity* $\bar{\mathsf{S}}_{Q,\lambda}$ *in* (3.69) *satisfies for all probability measures* $P \in \bigcirc_Q(\mathcal{M})$ *that*

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, P) = \mathsf{D}\left(P \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}(P\|Q) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right), \tag{3.70}$$

where the probability measures Q *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *are defined in* (3.4).

Proof: The proof is presented in Appendix B.17. ■

An interesting interpretation of Lemma 3.3.23 follows from rewriting (3.70) using the objective function of the Type-I ERM-RER problem in (3.50) as follows:

$$\begin{aligned}\mathsf{D}\left(P \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\ + \mathsf{D}(P\|Q) - \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right).\end{aligned} \tag{3.71}$$

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 (3.50) 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 3.3.23, when P is chosen to be identical to the reference measure Q, it follows that

$$\bar{\mathsf{S}}_{Q,\lambda}(\boldsymbol{z}, Q) = \mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right), \tag{3.72}$$

where the right-hand side is a Jeffreys divergence, 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 $\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq 0$ [84, Theorem 1], Lemma 3.3.23 leads to the following corollary.

Corollary 3.3.24. *The probability measures* Q *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (3.4) *satisfy*

$$\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leq \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q), \tag{3.73}$$

where the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ *is defined in* (3.64).

Type-I and Type-II Optimal Measures

The solutions to the optimization problems (2.13) and (3.1), 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 (2.16) and in (3.4). These measures exhibit the following property.

Lemma 3.3.25. *The probability measures* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *and* $\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (2.16) *and in* (3.4)*, respectively, satisfy*

$$\begin{aligned}&\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) \\ &= \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right),\end{aligned} \tag{3.74}$$

where the function $K_{Q,\boldsymbol{z}}$ *is defined in* (2.15).

Proof: The proof is presented in Appendix B.18. ■

Lemma 3.3.25 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 3.3.25.

Corollary 3.3.26. *The probability measures* $P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *and* $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (2.16) *and in* (3.4)*, respectively, satisfy*

$$\begin{aligned}&\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\&= \mathsf{D}\left(P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \left(\log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\end{aligned} \tag{3.75}$$

and

$$\begin{aligned}&\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\\&= \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right),\end{aligned} \tag{3.76}$$

where the functionals $\mathsf{S}_{Q,\lambda}$ *and* $\bar{\mathsf{S}}_{Q,\alpha}$ *are respectively defined in* (3.68) *and in* (3.69)*; and the function* $K_{Q,\boldsymbol{z}}$ *is defined in* (2.15)*.*

The equality in (3.75) 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 (2.16) to the probability measure $\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) via the sensitivity $\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$. The equality in (3.76) 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 (3.4) to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (2.16) via the sensitivity $\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$.

3.4 Numerical Results of Type-I and Type-II Regularization

In machine learning, the generalization error indicates how well a probability measure concentrates on models that correctly assign labels to unseen test data. The sensitivity (Definition 3.3.3) of the optimization problem in (2.13) is closely related to the generalization error [84, Theorem 16]. See also [109, 130, 131]. 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 [130]. 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 [132]. The MNIST example is simplified to accommodate parameterized models in $\mathbb{R}^2$ such that the approximations of the generalization error for different regularization factors are valid [18]. In practice, however, the Type-I and Type-II ERM-RER algorithm only require solving the statistical model once for given λ, $\boldsymbol{z}$, and Q, after which a model can be sampled from the resulting measure. This approach enables training in higher-dimensional spaces. The increase in space dimensionality continues to pose a significant challenge for ERM-RER statistical models. This is primarily due to the computational complexity associated with calculating the normalization function, which necessitates evaluating the empirical risk across all possible models within the support of the reference measure. This task is known to be $\#P$-hard [133].

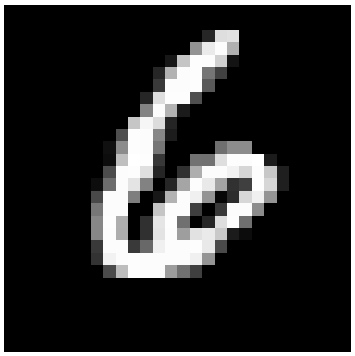


FIGURE 3.1: 28×28 image of a handwritten six from the MNIST dataset.

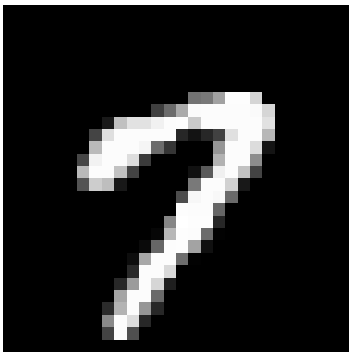


FIGURE 3.2: 28×28 image of a handwritten seven from the MNIST dataset.

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 Appendix E.3 in order to ensure the approximation of the generalization error. Consider the Type-I ERM-RER problem in (2.13) and the Type-II ERM-RER problem in (3.1) 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; (ii) 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 (2.6) is defined as

$$f(\boldsymbol{\theta}, x) = \begin{cases} 6 & \text{if } 0 < (x\boldsymbol{W})\boldsymbol{\theta}, \\ 7 & \text{if } 0 > (x\boldsymbol{W})\boldsymbol{\theta}, \end{cases} \tag{3.77}$$

where the matrix $\boldsymbol{W}$ is defined in (E.112) in Appendix E.3; and (f) the loss function ℓ in (2.5) satisfies

$$\ell(f(\boldsymbol{\theta}, x), y) = \mathbb{1}_{\{f(\boldsymbol{\theta},x)\neq y\}}. \tag{3.78}$$

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 3.5 displays an empirical approximation of the generalization error for Type-I and Type-II algorithms. The smaller generalization error approximation of Type-II suggests it is less prone to overfitting. In contrast, Figure 3.3 and Figure 3.4 show that Type-I achieves a lower average training error, which results in a lower average test error. These observations imply that Type-II promotes more 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 generalization gap, indicating greater reliance on the training data. This asymmetry balances the trade-off between data preference and solution robustness.

Another key observation is that, for certain ranges of the regularization factor (e.g.,$\lambda \in$

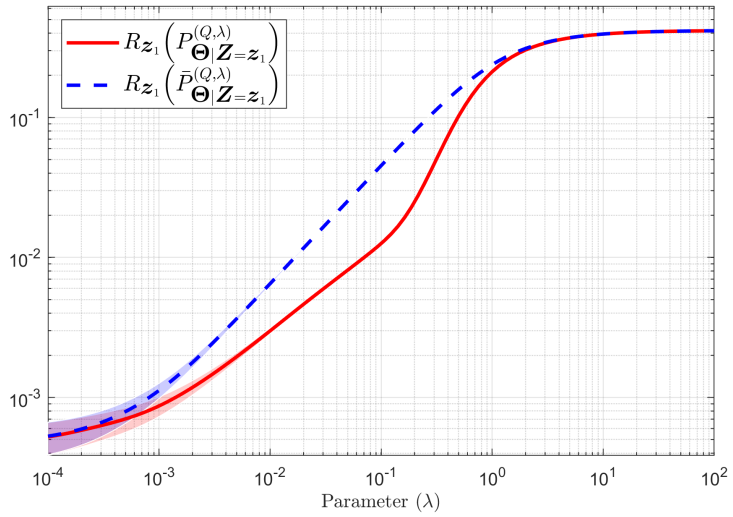


FIGURE 3.3: Average Training Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ and $\mathsf{R}_{\boldsymbol{z}_1}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, 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 (2.16) and (3.4), respectively, computed over one hundred different training and test dataset random selections.

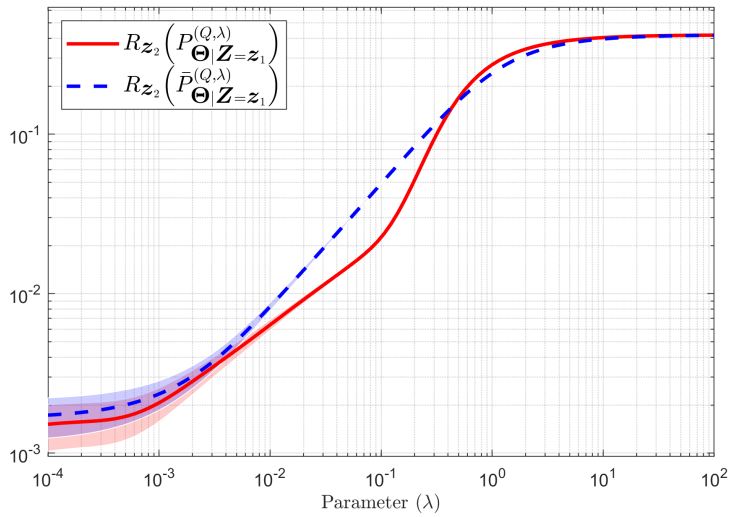


FIGURE 3.4: Average Test Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ and $\mathsf{R}_{\boldsymbol{z}_2}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, 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 (2.16) and (3.4), respectively, computed over one hundred different training and test dataset random selections.

$(0.002, 0.09)$ and $\lambda \in (0.3, 0.8)$ for Type-II), where the average test errors are comparable, Type-II exhibits lower generalization error. This suggests that Type-II can achieve both small test error and generalization error for certain regularization factors, a highly desirable outcome. Overall, the observed behavior of the Type-I and Type-II ERM-RER problems is as expected, where for larger values of λ the expected empirical risk is larger than for smaller, which follows from the intuition of classical empirical risk minimization. Furthermore, the observation of points in which Type-II performs better than Type-I is expected due to the asymmetry. However, selecting such regularization ranges and whether such points exist in all learning cases remains an open question for future research under the theoretical framework presented in this thesis.

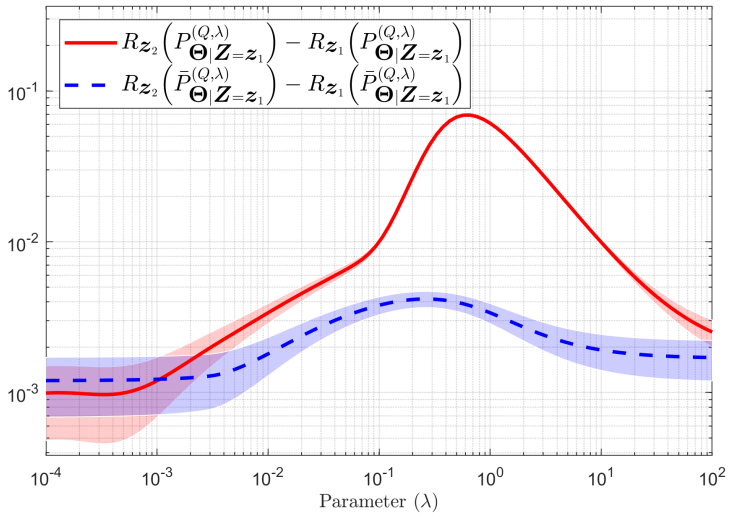


FIGURE 3.5: Average of the differences $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ and $\mathsf{R}_{\boldsymbol{z}_2}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_1}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, 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 (2.16) and (3.4), respectively, computed over one hundred different training and test dataset random selections.

3.5 Conclusion

This work has introduced the Type-II ERM-RER problem and has presented its solution through Theorem 3.3.1. 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 providing 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 (3.1) shows that, under mild assumptions, there always exists a positive real value λ such that, 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.

Chapter 4

ERM with f-Divergence Regularization

The solution to empirical risk minimization with f-divergence regularization (ERM-fDR) is derived under mild conditions on the function f. Under such conditions, the optimal probability measure is shown to be unique. Examples of the solution for particular choices of the function f are presented, and various previously known results are obtained as special cases. These include the unique solutions to the empirical risk minimization with Type-I and Type-II relative entropy regularizations. Further analysis of the solution reveals the following properties of the ERM-fDR problem: *i*) f-divergence regularization forces the support of the solution to match the support of the reference measure, creating a strong inductive bias that can override the evidence provided by the training data; *ii*) the normalization function defined as an implicit function in Section 3.3.1 is extended for the general f-divergence regularization and is explicitly characterized; and *iii*) the solution to an ERM with f-divergence regularization is identical to an ERM with g-divergence regularization, with an appropriate modification of the empirical risk function, under some conditions on the functions f and g.

4.1 Introduction

In statistical learning, the classical empirical risk minimization (ERM) problem [68, 69] is transformed by minimizing the expected empirical risk over a subset of all probability measures defined on the set of models. A regularization term is added to this expected empirical risk, often expressed as a statistical distance between the optimization measure and a reference measure. A well-studied case in the information theory community involves the use of relative entropy [11, 19, 84, 92, 108]. Other works address the general case for f-divergences, known as ERM problems with f-divergence regularization (ERM-fDR), with f-divergences, introduced in [134] and further developed in [135] and [23]. For discrete cases, the ERM-fDR is explored in [116] and [117], while more general settings are covered in [118] and [20]. The method of proof favored in this chapter enables the derivation of new results that have not been reported before. Firstly, the obtained solution holds for a family of f-divergences that is larger than the one in [118]. For instance, the Type-II ERM-RER studied in [119] and the ERM with Jensen-Shannon divergence regularization are both special cases of the ERM-fDR problem studied in this paper. These are examples of ERM-fDR problems that are not considered in [118]. Secondly, the permissible values of the regularization factor that guarantee the existence of a solution are analytically characterized. This is due that the works in [116, 117] and [118] are known only

up to a regularization factor, without further analysis on the value. In this work, analysis of the normalization factor is used to provide insights into possible future paths to tackle the computational bottleneck of such a factor to make practical use of the solution. For example, when using relative entropy as the regularizer, this factor corresponds to the log partition function [115]. This computation for ERM-fDR solutions is particularly challenging due to its dependence on evaluating the empirical risk over all possible models in the support of the reference measure, a task known to be $\#P$-hard [133].

This chapter presents the solution to the ERM-fDR, which is obtained under milder conditions than those in [118] using a method of proof that differs from those in [116, 117] and [118]. Then the solution is used to formalize the concept of the normalization function introduced in [20] and derive an explicit form expression to the normalization function by leveraging the implicit function theorem. Additionally, the new method of proof allows showing that the solution to an ERM with f-divergence regularization is identical to another ERM with g-divergence regularization, with an appropriate transformation of the empirical risk function and some conditions on both functions f and g.

4.2 ERM-fDR Problem

The ERM-fDR problem is parametrized by a probability measure $Q \in \triangle(\mathcal{M})$, a positive real λ, and a function $f : [0, \infty) \to \mathbb{R}$ that satisfies the conditions in Definition 1.5.1. The measure Q is referred to as the *reference measure* and λ as the *regularization factor*.

Given the dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ in (2.4), the ERM-fDR problem, with parameters Q, λ and f, consists of the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}_f(P\|Q), \tag{4.1}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (2.12). The optimization problem in (4.1) is closely related to the following optimization problem:

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \mathsf{R}_{\boldsymbol{z}}(P) \tag{4.2a}$$

$$\text{s.t.} \quad \mathsf{D}_f(P\|Q) \leq \eta, \tag{4.2b}$$

with $\eta \in [0, \infty)$. The optimization problems in (4.1) and (4.2) do not share the same solutions when for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = c$, for some $c > 0$. More specifically, the set of solutions to the problem in (4.2) is $\{P \in \triangle_Q(\mathcal{M}) : \mathsf{D}_f(P\|Q) \leq \eta\}$, while the set of solutions to (4.1) is the singleton $\{Q\}$. This distinction is mathematically significant but can be ignored in practice, as it arises only when $\mathsf{R}_{\boldsymbol{z}}(P)$ in (2.12) is constant for all measures P. In order to avoid the above case, the notion of separable empirical risk functions in Definition 2.4.1 is adopted.

4.3 ERM-fDR Solution & Results

The solutions to the ERM-fDR problems in (4.1) and (4.2) are presented under the following assumptions:

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

(b) There exists a β such that

$$\beta \in \left\{ t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \dot{f}^{-1}\left(-\frac{t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right) \right\}, \tag{4.3a}$$

and

$$\int \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right) \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{4.3b}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (2.6); and

(c) The function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is separable with respect to the probability measure Q.

Under Assumptions (a) and (b), the solution to the optimization problem in (4.1) was first presented in [20, Theorem 1]. Using Assumption (c), the following theorem shows that the problems in (4.1) and (4.2) share the same unique solution.

Theorem 4.3.1. *Under Assumptions (a) and (b), the solution to the optimization problem in (4.1), denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle_Q(\mathcal{M})$, is unique, and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right). \tag{4.4}$$

Moreover, under Assumptions (a), (b), and (c), if λ in (4.1) and η in (4.2) satisfy

$$\mathsf{D}_f\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q \right) = \eta, \tag{4.5}$$

then, the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) is also the unique solution to the optimization problem in (4.2).

Proof: The proof is presented in Appendix C.1. ■

Interestingly, the proof presented in Appendix C.1 for (4.4), is different from the one in [20, Theorem 1]. Also, the condition that β in (4.3b) satisfies Assumption (b) in (4.3a), leads to observing that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) > 0, \tag{4.6}$$

which leads to the following corollary.

Corollary 4.3.1. *Under Assumptions (a) and (b), the probability measures Q and $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) are mutually absolutely continuous.*

Corollary 4.3.1 reveals that the support of the reference measure Q establishes an inductive bias that cannot be overcome, regardless of the choice of f-divergence. That is, the support of the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) is identical to the support of the reference measure Q. In a nutshell, the use of any f-divergence regularization, under Assumptions (a) and (b), inadvertently forces the solution to the optimization problems in (4.1) and (4.2) to coincide with the support of the reference measure Q, independent of the training data. This observation has been already pointed out for particular cases. For instance, when the function f is such that $f(x) = x\log(x)$ in [84]; and $f(x) = -\log(x)$ in [119]. In general, this observation is particularly important for choosing the reference measure Q.

Examples

The following table presents the solutions to the optimization problem in (4.1) for specific choices of the function f. Derivation of the table results are presented in Appendix E.2.

TABLE 4.1: Solution to the ERM-fDR problem for different choices of f.

Name	$f(x)$	$\dot{f}(x)$	$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}\mid\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})$
Relative Entropy	$x\log(x)$	$1+\log(x)$	$\exp\Big(-\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)$
Reverse Relative Entropy	$-\log(x)$	$-\frac{1}{x}$	$\frac{\lambda}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}$
Jeffreys	$x\log(x)-\log(x)$	$\log(x)+1-x^{-1}$	$\exp\Big(W_0\Big(\exp\Big(\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)\Big)-\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)$
Jensen-Shannon	$x\log\Big(\frac{2x}{x+1}\Big)+\log\Big(\frac{2}{x+1}\Big)$	$\log(2x)-\log(x+1)$	$\frac{1}{2\exp\left(\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)-1}$
Hellinger	$(\sqrt{x}-1)^2$	$1-\frac{1}{\sqrt{x}}$	$\Big(\frac{\lambda}{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\Big)^2$
Pearson/ χ^2	x^2-1	$2x$	$\frac{\lambda+\mathsf{R}_{\boldsymbol{z}}(Q)-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{2\lambda}$

It is noted that for the choice of *relative entropy*, the result from Theorem 4.3.1 recovers the results independently reported by several authors in [84, 92, 108, 113, 115], and proved via a large variety of methods. Similarly, for the choice of *reverse relative entropy*, the result recovers the solution to the Type-II ERM-RER in Chapter 3, which has been reported in [119]. Furthermore, the result also provides solutions to unexplored divergences such as the *Jeffreys Divergence* and *Hellinger*, among others.

Note also that for cases such as the χ^2 divergence, it can be shown from Table 4.1 that there exist cases in which Assumption (b) is not met. For instance, if there exists a model $\boldsymbol{\theta}\in\operatorname{supp}Q$ for which $\lambda+\mathsf{R}_{\boldsymbol{z}}(Q)<\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})$, then, for such a model $\boldsymbol{\theta}$, it holds that $\dot{f}^{-1}\Big(-\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)=\frac{\lambda+\mathsf{R}_{\boldsymbol{z}}(Q)-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{2\lambda}<0$, which implies condition (4.3a) does not hold. Alternatively, if $\lambda>\sup\{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})-\mathsf{R}_{\boldsymbol{z}}(Q):\boldsymbol{\theta}\in\operatorname{supp}Q\}$, then $\beta\in\mathcal{B}$, and thus, $P^{(Q,\lambda)}_{\boldsymbol{\Theta}\mid\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) is the solution to the optimization problem in (4.1).

When Assumptions (a) and (b) are not met, the solution to the optimization problems in (4.1) and (4.2) cannot be obtained from Theorem 4.3.1. Moreover, in such a case, nothing can be asserted on whether solutions to these optimization problems exist. In a nutshell, the existence of the solution to the optimization problems in (4.1) and (4.2) for the case in which Assumptions (a) and (b) are not met is still an open problem. Some interesting f-divergences do not meet both Assumptions (a) and (b), which is the case of the total variation divergence.

Total Variation Divergence

Let the function $f:(0,+\infty)\to\mathbb{R}$ be such that $f(x)=|x-1|$. The resulting f-divergence $\mathsf{D}_f(P\|Q)$ is the total variation between P and Q. Nonetheless, the function f is not strictly convex and nondifferentiable at 1. Hence, the solution to the optimization problems in (4.1) and (4.2) cannot be obtained from Theorem 4.3.1.

4.3.1 ERM-fDR Normalization Function

The notion of *normalization function* introduced in Chapter 3 is generalized by providing a definition for the ERM-fDR problem, defined hereunder, with a slight abuse of notation for the definition of the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{B}_{Q,\boldsymbol{z}}$.

Definition 4.3.1 (Normalization Function)**.** *The normalization function of the problem in* (4.1)*, denoted by*

$$N_{Q,\boldsymbol{z}} : \mathcal{A}_{Q,\boldsymbol{z}} \to \mathcal{B}_{Q,\boldsymbol{z}}, \tag{4.7a}$$

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

$$\int \dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{4.7b}$$

The set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7a) contains all the regularization factors for which Assumption (b) is satisfied. More specifically, it contains the regularization factors λ for which the problem in (4.1) has a solution. Furthermore, the equality in (4.7b) justifies referring to the function $N_{Q,\boldsymbol{z}}$ as the *normalization function*, as it ensures that the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) is a probability measure.

The following theorem introduces one of the main properties of the function $N_{Q,\boldsymbol{z}}$ in (4.7), as it ensures that the measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) is a probability measure.

Theorem 4.3.2. *The function* $N_{Q,\boldsymbol{z}}$ *in* (4.7) *is continuous within the interior of* $\mathcal{A}_{Q,\boldsymbol{z}}$ *in* (4.7a) *and for all* $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$*, the* $N_{Q,\boldsymbol{z}}(\lambda)$ *is unique. Furthermore, if the function* f *is twice differentiable, for all* $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}}$*,*

$$N_{Q,\boldsymbol{z}}(\lambda) = \lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} N_{Q,\boldsymbol{z}}(\lambda) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}\right), \tag{4.8}$$

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

$$\frac{\mathrm{d}P^{(\lambda)}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)^{-1}}{\int \left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\right)\right)^{-1} \mathrm{d}Q(\boldsymbol{\nu})}. \tag{4.9}$$

Proof: The proof is presented in Appendix C.2. ∎

The result in Theorem 4.3.2 does not provide a direct computational method for obtaining the solution to the optimization problem (4.1); instead, it functions as an analytical tool for characterizing and controlling the evolution of the regularization path. Its practical significance arises primarily when the choice of λ is treated as a tunable parameter rather than as a fixed hyperparameter.

Lemma 4.3.2. *Under the assumption that* (i) *the function* f *is twice differentiable,* (ii) *the function* $\dot{f}^{-1}$ *is log convex, which is equivalent that for all* $u \in \operatorname{dom} f$ *such that* $\frac{\mathrm{d}}{\mathrm{d}u}\left(u\ddot{f}(u)\right) \geq 0$*, and* (iii) *the function* f *satisfies for all probability measures* $P \in \triangle_Q(\mathcal{M})$*,*

$$\int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \dot{f}\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \geq 0. \tag{4.10}$$

Then, the normalization function $N_{Q,\boldsymbol{z}}$ *in* (4.7) *is monotonicaly decreasing.*

Proof: The proof is presented in Appendix C.3. ∎

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

Lemma 4.3.3. *The set $\mathcal{A}_{Q,\boldsymbol{z}}$ in* (4.7a) *is either empty or an interval that satisfies*

$$\mathcal{A}_{Q,\boldsymbol{z}} = \begin{cases} \left[\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right) & \text{if } \int \dot{f}^{-1}\left(-\frac{t+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda^{\star}_{Q,\boldsymbol{z}}}\right) \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \\ \left(\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right) & \text{otherwise,} \end{cases} \tag{4.11}$$

where $t > \lim_{\lambda \to \lambda^{\star}_{Q,\boldsymbol{z}}+} N_{Q,\boldsymbol{z}}(\lambda)$ and $N_{Q,\boldsymbol{z}}$ is defined in (4.7).

Proof: The proof is presented in Appendix C.4. ■

Lemma 4.3.3 highlights two facts. First, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is a convex subset of positive reals. Second, if there exists a solution to the optimization problem in (4.1) for some $\lambda > 0$, then there exists a solution to such a problem when λ is replaced by $\bar{\lambda} \in (\lambda, \infty)$.

Note that even in cases where the reference measure Q is analytically or numerically computable with respect to the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6), finding the value of $N_{Q,\boldsymbol{z}}(\lambda)$ remains challenging. This is because $N_{Q,\boldsymbol{z}}(\lambda)$ appears inside the function $\dot{f}^{-1}$, which tends to introduce nonlinearity into the equation. The exception is the χ^2divergence, where the equation remains linear. In general, solving nonlinear equations involving high-dimensional integrals is computationally demanding.

However, for the case in which the integrals with respect to Q are computable, the results presented in Theorem 4.3.2, combined with Lemma 4.3.3, provide a pathway to compute the value of the normalization function using bracketing methods. This approach leverages the monotonicity and continuity of the solutions within the set $\mathcal{A}_{Q,\boldsymbol{z}}$, which allows computing the integral in (4.7b), for real values $\underline{\beta}$ and $\bar{\beta}$, such that $\underline{\beta} < N_{Q,\boldsymbol{z}}(\lambda) < \bar{\beta}$. Specifically, in this case, the value $N_{Q,\boldsymbol{z}}(\lambda)$ that satisfies the condition in (4.7b) can be obtained using root-finding algorithms, more precisely via the *bisection method*.

The following lemma presents a case in which the set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7a) can be fully characterized.

Lemma 4.3.4. *If the function $\dot{f}^{-1}$ in* (4.4) *is strictly positive, then the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is identical to* $(0, \infty)$.

Proof: The proof is presented in Appendix C.5. ■

From Table 4.1, the condition of Lemma 4.3.4 is satisfied by the *Kullback-Leibler Divergence*, *Jeffreys Divergence*, and *Hellinger Divergence*. For those cases, it holds that $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$. Alternatively, Table 4.1 shows that the *Reverse Relative Entropy Divergence*, *Jensen-Shannon Divergence*, and χ^2 *Divergence* do not satisfy the condition of Lemma 4.3.4. Thus, the value $\lambda^{\star}_{Q,\boldsymbol{z}}$ in (3.17) depends on Q, $\boldsymbol{z}$ and ℓ, such that in the case in which $\lambda^{\star}_{Q,\boldsymbol{z}} > 0$ implies the existence of an interval $\left(0, \lambda^{\star}_{Q,\boldsymbol{z}}\right)$ or $\left(0, \lambda^{\star}_{Q,\boldsymbol{z}}\right]$, for which Theorem 4.3.1 does not provide insights into the solution to the optimization problem (4.1). Appendix E.1 introduces some examples to illustrate particular cases in which the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is open or semi-open.

4.3.2 Properties of the ERM-fDR solution

This section studies the properties of the solution to the ERM-fDR problem in (4.1) and (4.2). Note that from Theorem 4.3.1 models resulting in lower empirical risks

correspond to greater values of the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (4.4). The following lemma formalizes this observation.

Lemma 4.3.5. *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* (2.6)*, the Radon-Nikodym derivative in* (4.4) *satisfies*

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2) \leq \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1), \tag{4.12}$$

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 C.6. ■

The Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ in (4.4) is always finite and strictly positive. This observation is formalized by the following lemma.

Lemma 4.3.6. *The Radon-Nikodym derivative* $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (4.4) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$

$$0 < \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \leq \dot{f}^{-1}\left(-\frac{\delta^{\star}_{Q,\boldsymbol{z}} + \beta}{\lambda}\right) < \infty, \tag{4.13}$$

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

Proof: The proof is presented in Appendix C.7. ■

The next lemma shows that the expected empirical risk induced by the solution to the optimization problems in (4.1) and (4.2), $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4), monotonically decreases with respect to the regularization factor λ. This observation is formalized by the following lemma.

Lemma 4.3.7. *For all* $(\lambda_1, \lambda_2) \in (0, \infty)^2$*, such that* $\lambda_1 < \lambda_2$*, the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (4.4) *satisfies*

$$\delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leq \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leq \mathsf{R}_{\boldsymbol{z}}(Q), \tag{4.14}$$

where the equality holds if and only if $\mathsf{L}_{\boldsymbol{z}}$ *in* (2.6) *is nonseparable with respect to the probability measure* Q.

Proof: The proof is presented in Appendix C.8. ■

4.3.3 Equivalence of the f-Regularization via Transformation of the Empirical Risk

This section shows that given two strictly convex and differentiable functions f and g that satisfy the conditions in Definition 1.5.1, there exists a function $v : [0, \infty) \to \mathbb{R}$, such that the solution to the optimization problem in (4.1) is identical to the solution of the following problem:

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \int v(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\, \mathrm{d}P(\boldsymbol{\theta}) + \lambda \mathsf{D}_g(P \| Q), \tag{4.15}$$

with λ and Q in (4.1). The main result of this section is presented in the following theorem.

Theorem 4.3.3. *Let* f *and* g *be two strictly convex and differentiable functions satisfying the conditions in Definition 1.5.1. If the problem in* (4.1) *possesses a solution,*

then

$$\begin{aligned}&\arg\min_{P\in\triangle_Q(\mathcal{M})}\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta})+\lambda\mathsf{D}_f(P\|Q)\\&\quad=\arg\min_{P\in\triangle_Q(\mathcal{M})}\int v(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}P(\boldsymbol{\theta})+\lambda\mathsf{D}_g(P\|Q),\end{aligned}\tag{4.16}$$

where the function $v:[0,\infty)\to\mathbb{R}$ *is such that*

$$v(t)=-\lambda\dot{g}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+t}{\lambda}\right)\right)-c,\tag{4.17}$$

where $N_{Q,\boldsymbol{z}}$ *is the normalization function of the optimization problem in* (4.1) *and* c *is a constant such that* $c\in\mathbb{R}$*.*

Proof: Note that from Theorem 4.3.1, the functions f and g are differentiable and strictly convex. Hence, the functional inverse of the derivative is well-defined from the fact that $\dot{f}$ and $\dot{g}$ are strictly increasing and bijective. Denote by $\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ the solution to the optimization problem in (4.15). Then, from Theorem 4.3.1, for all $\boldsymbol{\theta}\in\operatorname{supp}Q$, it follows that

$$\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=\dot{g}^{-1}\left(-\frac{\hat{N}_{Q,\boldsymbol{z}}(\lambda)+v(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))}{\lambda}\right)\tag{4.18}$$

$$=\dot{g}^{-1}\left(\dot{g}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right)+\frac{c-\hat{N}_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\right),\tag{4.19}$$

where $\hat{N}_{Q,\boldsymbol{z}}$ is the normalization function of the optimization problem in (4.15). Then, under the assumption that $\hat{N}_{Q,\boldsymbol{z}}$ satisfies $\hat{N}_{Q,\boldsymbol{z}}(\lambda)>c$, from the monotonicity of $\dot{g}^{-1}$, for all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$\begin{aligned}&\dot{g}^{-1}\left(\dot{g}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right)-\frac{\hat{N}_{Q,\boldsymbol{z}}(\lambda)+c}{\lambda}\right)\\&<\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right).\end{aligned}\tag{4.20}$$

Similarly, under the assumption that $\hat{N}_{Q,\boldsymbol{z}}$ satisfies $\hat{N}_{Q,\boldsymbol{z}}(\lambda)<c$, from the monotonicity of $\dot{g}^{-1}$, for all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$\begin{aligned}&\dot{g}^{-1}\left(\dot{g}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right)-\frac{\hat{N}_{Q,\boldsymbol{z}}(\lambda)+c}{\lambda}\right)\\&>\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right).\end{aligned}\tag{4.21}$$

Note that $\hat{N}_{Q,\boldsymbol{z}}$ needs be such that,

$$1=\int\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}).\tag{4.22}$$

Under the assumption that $\hat{N}_{Q,\boldsymbol{z}}(\lambda)>c$, from (4.20), it follows that

$$\int\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})<\int\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta})\tag{4.23}$$

$$=1.\tag{4.24}$$

Similarly, under the assumption that $\hat{N}_{Q,\boldsymbol{z}}(\lambda) < c$, from (4.21), it follows that

$$\int \frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) > \int \dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{4.25}$$

$$= 1. \tag{4.26}$$

Therefore, the normalization function $\hat{N}_{Q,\boldsymbol{z}}$ satisfies

$$\hat{N}_{Q,\boldsymbol{z}}(\lambda) = c. \tag{4.27}$$

Thus, from (4.19) and (4.27), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ that

$$\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \tag{4.28}$$

$$= \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{4.29}$$

where (4.28) follows from (4.17), which completes the proof. ■

Theorem 4.3.3 establishes an equivalence between two ERM problems subject to different f-divergence regularizations. Such equivalence can always be established as long as the corresponding divergences are defined by strictly convex and differentiable functions. Furthermore, for all strictly convex f functions, the solution to the corresponding ERM with f-divergence regularization is mutually absolutely continuous with respect to the reference measure. Another important observation follows from noting that for any strictly convex function f, the derivative $\dot{f}$ is strictly increasing, and from Lemma 4.3.5 the function $\dot{f}^{-1}$ is strictly decreasing with respect to the empirical risk implies that for a given λ the function v in (4.17) is monotonic.

The following example illustrates the equivalence between two f-divergence regularizations.

Example 4.3.1. *Consider the optimization problems in* (4.1) *and* (4.15) *with* $f(t) = -\log(t)$ *and* $g(t) = t\log(t)$*, respectively. Under the current assumptions, the objective of this example is to demonstrate the equivalence of the solutions to the optimization problems in* (4.1) *and* (4.15)*. The solution to the optimization problem in* (4.1) *is described in Section E.2.1. Denote by* $\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *the solution to the optimization problem in* (4.15)*. From Theorem 4.3.1, it follows that for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{v(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))+\beta}, \tag{4.30}$$

where the function v *is defined in* (4.17) *and under the assumptions of this example satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$*,*

$$v(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) = \frac{\lambda}{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - \log\left(\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\nu})\right)\right)} - \beta. \tag{4.31}$$

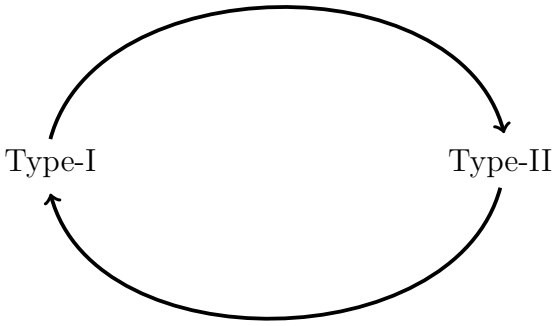
Plugging (4.31) *into* (4.30) *yields*

$$\frac{\mathrm{d}\hat{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)}{\int \exp\left(-\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu})}, \tag{4.32}$$

which is the solution to the optimization problem in (4.1) *presented in Section E.2.1.*

Hence, from the above example, it is possible to transform the empirical risk function to map from problems with regularization Type-I to problems with regularization Type-II and the converse is possible. Note that this is true for all other strictly convex f-divergences as long as the choice of λ belongs to the set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7a) for the optimization problems (4.1) and (4.2) with each of the chosen divergences.

$$v(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) = \lambda \log(\hat{N}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))$$



$$\bar{v}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) = \frac{\lambda}{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} - \log\left(\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\nu})\right)\right)} - \hat{N}_{Q,\boldsymbol{z}}(\lambda)$$

FIGURE 4.1: Representation of the empirical risk transformation from the f-divergence induced by $f(t) = -\log(t)$ and the g-divergence induced by $g(t) = t\log(t)$.

4.4 Numerical Comparison of ERM-fDR Regularizations

This section presents a machine learning example used in Chapter 3 Section 3.4, by computing the results for new choices of f in the ERM-fDR optimization problem. The solution $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}$ to the ERM-fDR optimization problems in (4.1) and (4.2) when

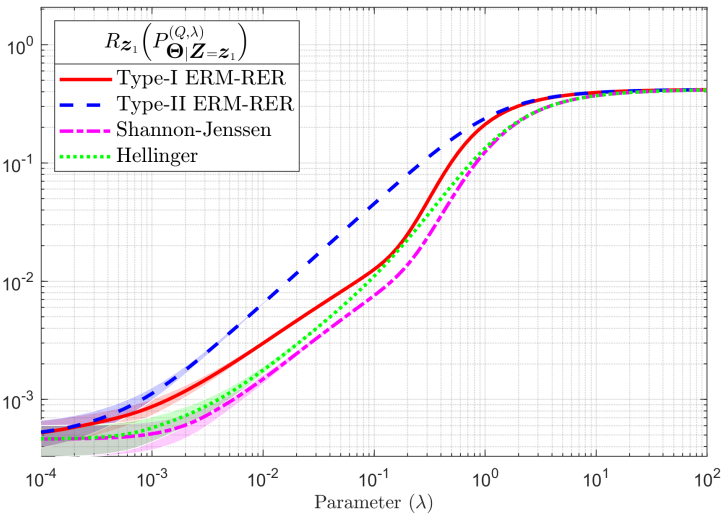


FIGURE 4.2: Average Training Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, for the f-divergence regularization Type-I ERM-RER, Type-II ERM-RER, Shannon-Jensen and Hellinger in Table 4.1, respectively, computed over one hundred different training and test dataset random selections.

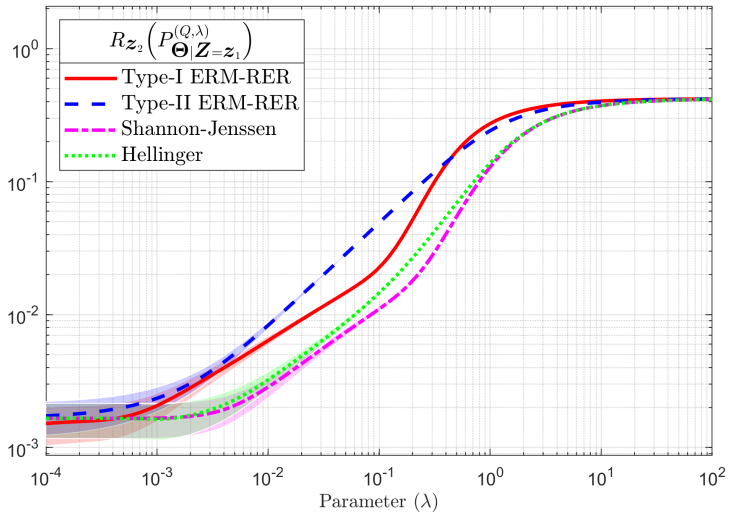


FIGURE 4.3: Average Test Error: average of the expected empirical risks $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, for the f-divergence regularization Type-I ERM-RER, Type-II ERM-RER, Shannon-Jensen and Hellinger in Table 4.1, respectively, computed over one hundred different training and test dataset random selections.

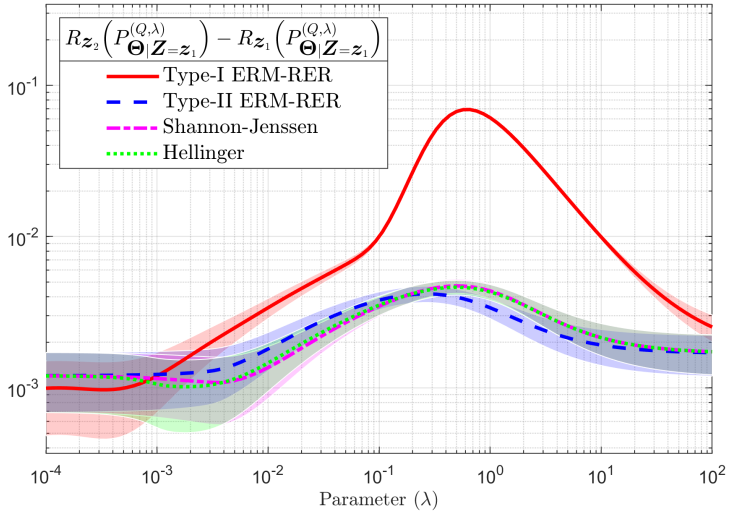


FIGURE 4.4: Average of the differences $\mathsf{R}_{\boldsymbol{z}_2}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right) - \mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$, for the f-divergence regularization Type-I ERM-RER, Type-II ERM-RER, Shannon-Jensen and Hellinger in Table 4.1, respectively, computed over one hundred different training and test dataset random selections.

choosing f to be the reverse relative entropy (Type-II ERM-RER) exhibits the largest expected empirical risk $\mathsf{R}_{\boldsymbol{z}_1}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}_1}\right)$ compared to the ERM-fDR solutions with relative entropy (Type-I ERM-RER), Hellinger and Shannon-Jensen regularization in Table 4.1, as shown in Figure 4.2 and Figure 4.3.

However, the sensitivity analysis in Figure 4.4 reveals that Type-I has a larger sensitivity than the f-divergence regularization Type-II, Shannon-Jensen, and Hellinger in Table 4.1. This suggests that Type-I is more likely to overfit than Type-II, Shannon-Jensen, and Hellinger. Remarkably, for small values of the regularization factor λ, the sensitivity of Type-I is significantly smaller on average than that of Type-II, Shannon-Jensen and Hellinger, despite all solutions yielding comparable expected empirical risk as shown in Figure 4.3. The example above showcases a simple implementation of the

solutions for the ERM-fDR problem with different choices of f, in which sampling from these distributions is achieved through MCMC and rejection sampling [136]. Following on the discussion in Chapter 3 Section 3.4, the tradeoff off asymmetry between the Type-I and Type-II ERM-RER solutions, where Type-I favors overfitting data at the expense of sensitivity, while Type-II leans more towards a conservative approach by underfitting the data at the expense of training error and test error, seems to be balanced for cases in which symmetric regularization functions, such as Hellinger and the Shannon-Jensen divergence are chosen. This particular observation opens the door for future study of f-divergences as potential regularizers that can give tighter guarantees in statistical learning algorithms compared to the Gibbs probability distribution, which results from the Type-I ERM-RER problem.

4.5 Conclusions

This work has presented the solution to the ERM-fDR problem under mild conditions on f, namely, (a) strict convexity and (b) differentiability. Under these conditions, the optimal measure is shown to be unique and sufficient conditions for the existence of the solution are presented. These conditions, jointly with the implicit function theorem, provide an explicit expression for the normalization function. This result unveils the fact that all parameters are involved in guaranteeing the existence of a solution. Remarkably, the f-divergence regularizer can be transformed into a different f-divergence regularizer by a transformation of the empirical risk. The mutual absolute continuity of the ERM-fDR solutions to the reference measure can be understood in light of the equivalence between the regularization. The analytical results have also enabled us to provide insights into choices of f-divergences for algorithm design in statistical machine learning.

Chapter 5

ERM-fDR Duality

A dual formulation of ERM-fDR is introduced and shown to have a duality gap zero with respect to the ERM-fDR solution. This dual approach leverages the Legendre-Fenchel transform to provide an alternative method to compute the expected empirical risk of ERM-fDR solution. Furthermore, the alternative characterization of the expected empirical risk enables explicit characterizations of the generalization error for general algorithms under mild conditions and another for ERM-fDR solutions.

5.1 Introduction

Dual problem formulations play a central role in optimization theory [137] and [25], which can offer both theoretical insights and computational advantages for optimization problems. Transforming the original (primal) problem into its dual counterpart opens the door to making use of properties such as convexity and separability, which are often not readily apparent in the primal formulation. In this context, the ERM-fDR problem benefits from dual formulations due to its strict convexity. By leveraging the techniques presented in [20] – which goes along the lines of the methods in [84, 109, 138]; and relying on the Gâteaux derivative [139] and vector space methods [24], this chapter introduces a dual formulation for the ERM-fDR problem. The proposed dual formulation provides a convenient way to compute the normalization function, while offering operational insights into the characterization of the generalization error for statistical learning algorithms.

This chapter makes the following contributions: A dual optimization problem to the ERM-fDR is introduced in Section 5.2 and its solution is derived, jointly with conditions provided to ensure equivalence between the dual and primal solutions. The dual solution is used to characterize the normalization function via the implicit function theorem, connecting the ERM-fDR solution to the dual through the Legendre-Fenchel transform. Finally, the connection between the Legendre-Fenchel transform and f-divergence regularization is used to explicitly characterize the generalization error for the ERM-fDR problem in Section 5.4.

5.2 ERM-fDR Dual Problem

The duality principle [25, Chapter 5] enables the reformulation of the optimization problem in (4.1) into an alternative form, known as the dual problem. In this section, this dual formulation is derived using the Legendre-Fenchel transform [25], see Definition 1.5.2.

Using this notation, consider the following problem

$$\min_{\beta\in\mathbb{R}} \quad \lambda \int f^*\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + \beta, \tag{5.1}$$

where the real λ, the measure Q and the function f are those in (4.1); and the functions $\mathsf{L}_{\boldsymbol{z}}$ and f^* are defined in (2.6) and (1.7), respectively.

5.3 ERM-fDR Dual Solution & Results

The following theorem introduces the solution to the problem in (5.1).

Theorem 5.3.1. *Under Assumptions* (a) *and* (b) *the solution to the optimization problem in* (5.1) *is* $N_{Q,\boldsymbol{z}}(\lambda)$*, where the function* $N_{Q,\boldsymbol{z}}$ *is defined in* (4.7)*.*

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

$$G(\beta) = \lambda \int f^*\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + \beta, \tag{5.2}$$

which is the objective function of the optimization problem in (5.1). Note that G in (5.2) is a convex function, and thus satisfies:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta} G(\beta) &= \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\lambda \int f^*\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + \beta\right) && (5.3)\\
&= \lambda \int \frac{\mathrm{d}}{\mathrm{d}\beta} f^*\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + 1 && (5.4)\\
&= -\int \dot{f}^*\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) + 1, && (5.5)
\end{aligned}$$

where $\dot{f}^*$ is the derivative of the function f^* in (5.1). Let the solution to the optimization problem in (5.2) be denoted by $\widehat{\beta} \in \mathbb{R}$ and note that the derivative of the function G evaluated at $\widehat{\beta}$ is equal to zero, that is

$$\int \dot{f}^*\left(-\frac{\widehat{\beta} + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{5.6}$$

From [137, Corollary 23.5.1] and Assumption (a), the following equality holds for all $t \in \mathcal{J}$, with $\mathcal{J}$ in (1.8),

$$\frac{\mathrm{d}}{\mathrm{d}t} f^*(t) = \dot{f}^*(t) = \dot{f}^{-1}(t), \tag{5.7}$$

where the functions $\dot{f}^{-1}$ and $\dot{f}^*$ are defined in (1.5) and (5.5), respectively. From (5.6) and (5.7), it follows that

$$\int \dot{f}^{-1}\left(-\frac{\widehat{\beta} + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{5.8}$$

which combined with (4.7b) and Assumption (b) yields

$$N_{Q,\boldsymbol{z}}(\lambda) = \widehat{\beta}, \tag{5.9}$$

and completes the proof. ■

Note that the proof leads to the following corollary.

Corollary 5.3.1. *The Radon-Nikodym derivative* $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ *in* (4.4) *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^*\bigg(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\bigg), \tag{5.10}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$, $N_{Q,\boldsymbol{z}}$ *and* $\dot{f}^*$ *are defined in* (2.6), (4.7) *and* (5.7), *respectively.*

Corollary 5.3.1 connect the results from Theorem 4.3.1 to the optimization problem in (5.1) via Legendre-Fenchel transform in Definition 1.5.2. Using this connection, the following lemma establishes that the problem in (5.1) is the dual problem to the ERM-fDR problem in (4.1) and characterizes the difference between their optimal values, which is often referred to as the duality gap [89, Section 8.3].

Lemma 5.3.2. *Under Assumptions* (*a*) *and* (*b*), *the optimization problem in* (5.1) *is the dual problem to the ERM-*f*DR problem in* (4.1). *Moreover, the duality gap is zero.*

Proof: Under Assumption (*a*) and [25, Section 3.3.2], it can be verified that for all $t \in \mathcal{J}$, with $\mathcal{J}$ in (1.8), the function f^* in (1.7) satisfies

$$f^*(t) = t\dot{f}^*(t) - f(\dot{f}^*(t)), \tag{5.11}$$

where the function $\dot{f}^*$ is the same as in (5.7). Then, from (5.11) and Corollary 5.3.1, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\begin{aligned}
&\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \lambda f\bigg(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\bigg)\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) \\
&= -\lambda f^*\Big(-\frac{N_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) - N_{Q,\boldsymbol{z}}(\lambda).
\end{aligned} \tag{5.12}$$

Taking the expectation in both sides of (5.12) with respect to the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) yields

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) + \lambda \mathsf{D}_f\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\Big) \\
&= -\lambda \int f^*\Big(-\frac{N_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)\,\mathrm{d}Q(\boldsymbol{\theta}) - N_{Q,\boldsymbol{z}}(\lambda).
\end{aligned} \tag{5.13}$$

Using Theorem 4.3.1 and Theorem 5.3.1 in the left-hand and right-hand sides of (5.13), respectively, yields

$$\begin{aligned}
&\min_{P \in \triangle_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}_f(P\|Q) \\
&= \max_{\beta \in \mathbb{R}} -\lambda \int f^*\Big(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\Big)\,\mathrm{d}Q(\boldsymbol{\theta}) - \beta.
\end{aligned} \tag{5.14}$$

The proof that the optimization problem in (5.1) is the dual to the ERM-fDR problem in (4.1) follows from (5.14) and [24, Theorem 1, Section 8.4]. The zero duality gap is established by the equality in (5.14), which completes the proof. ∎

5.3.1 ERM-fDR Expected Empirical Risk

The following theorem connects the Legendre-Fenchel transform in Definition 1.5.2 and the normalization $N_{Q,\boldsymbol{z}}$ function in (4.7), with respect to the solution to the optimization problem in (4.1).

Lemma 5.3.3. *The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) satisfies*

$$\mathsf{R}_{\boldsymbol{z}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)+\lambda\mathsf{D}_f\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\Big)=-\lambda\int f^*\bigg(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\bigg)\,\mathrm{d}Q(\boldsymbol{\theta})-N_{Q,\boldsymbol{z}}(\lambda), \tag{5.15}$$

and

$$\mathsf{R}_{\boldsymbol{z}}(Q)+\lambda\int f\bigg(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\bigg)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})=-\lambda\int f^*\bigg(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\bigg)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})$$
$$-N_{Q,\boldsymbol{z}}(\lambda), \tag{5.16}$$

where f^* is the Legendre-Fenchel transform of f (see Definition 1.5.2), the functions $N_{Q,\boldsymbol{z}}$ is defined in (4.7), and the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (2.12).

Proof: The proof is presented in Appendix D.2. ■

5.4 Exact Characterization of the Generalization Error

Let the functional $\mathsf{G}:(\mathcal{X}\times\mathcal{Y})^n\times\triangle(\mathcal{M})\times\triangle(\mathcal{M})\to\mathbb{R}$ satisfy

$$\mathsf{G}(\boldsymbol{z},P_1,P_2)=\mathsf{R}_{\boldsymbol{z}}(P_1)-\mathsf{R}_{\boldsymbol{z}}(P_2). \tag{5.17}$$

The value $\mathsf{G}(\boldsymbol{z},P_1,P_2)$ in (5.17) represents the variation of the functional $\mathsf{R}_{\boldsymbol{z}}$ in (2.12) when its argument changes from P_2 to P_1. This variation is referred to as an *algorithm driven gap* in [18], which is justified by the fact that P_1 and P_2 can be assimilated to learning algorithms. Using this notion, this section studies the generalization error of machine learning algorithms.

Definition 5.4.1 (Generalization Error [18, Definition 4])**.** *The generalization error induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}\in\triangle(\mathcal{M}|(\mathcal{X}\times\mathcal{Y})^n)$ under the assumption that training and test datasets are independently sampled from a probability measure $P_{\boldsymbol{Z}}\in\triangle((\mathcal{X}\times\mathcal{Y})^n)$, which is denoted by $\bar{\bar{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}\big)$, is*

$$\bar{\bar{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}\big)=\iint\big(\mathsf{R}_{\boldsymbol{u}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big)-\mathsf{R}_{\boldsymbol{z}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big)\big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u})\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}). \tag{5.18}$$

Consider the following assumptions:

(*d*) For all $\boldsymbol{z}\in(\mathcal{X}\times\mathcal{Y})^n$, the probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to the probability measure $P_{\boldsymbol{\Theta}}\in\triangle_Q(\mathcal{M})$, which satisfies for all measurable subsets $\mathcal{C}$ of $\mathcal{M}$

$$P_{\boldsymbol{\Theta}}(\mathcal{C})=\int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}). \tag{5.19}$$

(*e*) The probability measure $P_{\boldsymbol{\Theta}}$ in (5.19) and Q in (4.1) are mutually absolutely continuous.

Under Assumptions (d) and (e), it follows from [18, Lemma 3] that the generalization error $\bar{\bar{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\big)$ in (5.18) satisfies

$$\bar{\bar{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\big) = \int \mathsf{G}\big(\boldsymbol{z}, P_{\boldsymbol{\Theta}}, P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big)\, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \tag{5.20}$$

$$= \int \mathsf{G}\Big(\boldsymbol{z}, P_{\boldsymbol{\Theta}}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \mathsf{G}\Big(\boldsymbol{z}, P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{5.21}$$

where the measures $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P_{\boldsymbol{\Theta}}$ are defined in (4.4) and (5.19), respectively; and the functional G is defined in (5.17). The following theorem presents the main tool in this section.

Theorem 5.4.1. *The probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) satisfies for all $P \in \triangle_Q(\mathcal{M})$,*

$$\mathsf{G}\Big(\boldsymbol{z}, P, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda \int \left(1 - \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{5.22}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$, $N_{Q,\boldsymbol{z}}$, and f^* are defined in (2.6), (4.7) and (1.7), respectively; and the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is defined in (4.4).

Proof: The proof is presented in Appendix D.1. ■

Note that the Theorem 5.4.1 serves two purposes. The first is that it can be used to quantify the information gap between any measure with respect to the solution. This is particularly useful when compared to the prior Q as it becomes a metric on the improvement in expected empirical risk with respect to priors for different choices of f, which can be reduced to other statistical distances, as shown by the following examples.

Relative Entropy: Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that $f(x) = x \log(x)$, the gap in (5.17) of solution of the Type-I ERM-RER in (2.16) with respect to the reference measure Q reduces to

$$\mathsf{G}\Big(\boldsymbol{z}, Q, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda \mathsf{D}_J\Big(Q \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big), \tag{5.23}$$

which recovers the results previously reported by other authors in [115, Lemma 3] and [84, Lemma 20].

Reverse Relative Entropy: Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that $f(x) = -\log(x)$, the gap in (5.17) of solution of the Type-II ERM-RER in (3.4) with respect to the reference measure Q reduces to

$$\mathsf{G}\Big(\boldsymbol{z}, Q, \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \lambda \chi^2\Big(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big). \tag{5.24}$$

The second purpose of Theorem 5.4.1, it that it allows generalizing the method of algorithm-driven gaps introduced in [18]. In particular the choice of $f(x) = x \log(x)$ in Theorem 5.4.1 leads to [115, Theorem 1]. Moreover, using Theorem 5.4.1 in (5.21) leads to the following characterization of $\bar{\bar{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\big)$ in (5.18).

Theorem 5.4.2. *The generalization error* $\bar{\bar{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\big)$ *in* (5.18)*, under Assumptions* (*a*)*,* (*b*)*,* (*d*) *and* (*e*)*, satisfies*

$$\bar{\bar{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\big) = \lambda \int\int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda} \right) \right) \left(\frac{\mathrm{d}P_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) - \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) \right) \mathrm{d}Q(\boldsymbol{\theta})\, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{5.25}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$*,* $N_{Q,\boldsymbol{z}}$*, and* f^* *are defined in* (2.6)*,* (4.7) *and* (1.7)*, respectively; and the probability measure* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *and the real* λ *are those in* (4.4)*.*

Proof: The proof is presented in Appendix D.3. ■

The expression in (5.25) significantly simplifies for some choices of f. See for instance, the case in which $f(x) = x\log(x)$ in [18, Lemma 4]. Another case in which such expression becomes particularly simple is the case in which the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is the solution to the optimization problems (4.1) and (4.2). In such a case, the following holds.

Theorem 5.4.3. *Consider the solution to the optimization problem in* (4.1)*,* $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (4.4)*, and consider also the generalization error* $\bar{\bar{\mathsf{G}}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ *defined as in* (5.18)*. Under Assumptions* (*a*)*,* (*b*)*,* (*d*) *and* (*e*)*, the following holds*

$$\bar{\bar{\mathsf{G}}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\Big) = \lambda \int \left(\int \dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) - \int \dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{5.26}$$

where the function $\dot{f}$ *is defined in* (1.4)*.*

Proof: The proof is presented in Appendix D.4. ■

The case in which $f(x) = x\log(x)$ in Theorem 5.4.2 and Theorem 5.4.3, which corresponds to the Gibbs algorithm, leads to the existing results in [84, Theorem 16] and [104, Theorem 1]. Note that, from both Theorem 5.4.2 and Theorem 5.4.3, the expression on the generalization error connect the population risk directly to the stochastic behavior of the learning algorithm through the distribution of models it induces. In practical terms, whenever the inner integral can be computed exactly or bounded effectively, it becomes possible to derive quantitative guarantees on the expected performance of the learned model across the population. Such expressions enable the development of theoretically grounded generalization bounds, facilitate the study of regularization effects, and provide a systematic route for comparing algorithms under uncertainty. Nevertheless a major bottle neck to the usefulness of this results remains the ability to computed the expressions. However, effective bounding could prove useful.

5.5 Conclusions

This work has established the conditions under which the solution to the ERM-fDR problem in (4.1) also serves as the solution to the optimization problem in (4.2). Furthermore, the solution of the ERM-fDR problem has been connected to the

Legendre-Fenchel transform by establishing the dual problem in (5.1) to the optimization problem in (4.1). This connection between the dual formulation and the Legendre-Fenchel transform provides explicit expressions on the expected empirical risk. Notably, the link to the Legendre-Fenchel transform, under mild assumptions, enables the derivation of an explicit expression for the generalization error of general learning algorithms. Lastly, it offers a separate explicit expression for evaluating the generalization error of algorithms arising as the solution to an ERM-fDR optimization problem.

Chapter 6

Conclusion

This research project has made significant contributions to the fields of statistical learning and machine learning by theoretically characterizing the impact of regularization in supervised learning using information measures. The exploratory approach adopted in this work has enabled a thorough comprehension of the role of regularization, particularly through the lens of relative entropy and general f-divergences, and their implications for empirical risk minimization and generalization error.

6.1 Contributions

The contributions to statistical learning and ML are presented as follows:

The first objective, the *Characterization of Asymmetry in Relative Entropy Regularization*, focused on understanding the role of asymmetry in empirical risk minimization with relative entropy regularization. By solving the Type-II ERM-RER, this work provides insight into the inductive bias of reverse relative entropy and relative entropy, which causes the support of the solution to collapse into the support of the reference measure. Additionally, the analysis of the benefits of using $\mathsf{D}(Q\|P)$ versus $\mathsf{D}(P\|Q)$ and the generalization to symmetrized relative entropy has advanced the understanding of how the choice of the empirical risk function influences learning algorithms, via their equivalence through appropriate empirical risk function transformation. These contributions are particularly valuable for statistical learning as they offer insights into the trade-offs between model complexity and generalization.

The second objective, the *Generalization to f-divergences as Regularizers*, extended the findings from relative entropy regularization to general f-divergences. By deriving bounds over the expected empirical risk and providing a general analytical solution in terms of the f-divergence functional, this research has broadened the applicability of regularization techniques in supervised statistical machine learning. These results also enable the characterization of the normalization function, which is akin to the normalization factor. Thus, enhancing the understanding of the ERM-fDR solution and providing properties that can be leveraged to compute the solution. This contribution opens the door for further research in this venue as a possible tool to bridge statistical learning theory and practical machine learning applications.

The third objective, the *Characterization of Generalization Error for f-divergence Regularized Algorithms*, addressed the generalization error of f-divergence regularized algorithms. By solving the dual problem for empirical risk minimization and demonstrating that the duality gap is zero via the Legendre-Fenchel transform, this work has provided a rigorous theoretical foundation for understanding the generalization properties of regularized models. Furthermore, the extension of the *method

of gaps* to f-divergence regularization represents a novel contribution to the field, offering a new framework for analyzing generalization errors in a broader class of learning algorithms. This advancement is particularly significant for machine learning as it enhances the theoretical understanding of how regularization impacts model performance on unseen data.

In summary, this research has made the following key contributions:

- Developed theoretical bounds and insights into the role of asymmetry in relative entropy regularization, aiding in the selection of regularization parameters.
- Generalized regularization techniques to f-divergences, providing a broader framework for empirical risk minimization.
- Characterized the generalization error for f-divergence regularized algorithms, advancing the understanding of model performance and robustness.

These contributions collectively enhance the theoretical foundations of f-divergence regularization in statistical learning, providing insight into properties of the f-divergence family of regularization. The findings provide valuable tools for researchers and practitioners to provide guarantees on the Generalization error of algorithms and design their algorithms.

6.2 Limitations of the current work

The primary practical limitation of the ERM-fDR framework lies in the computational complexity associated with evaluating and optimizing the f-divergence term. Except in special cases, most notably the relative entropy divergence and χ^2, the normalization condition for the optimal distribution $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ does not admit a closed-form expression, requiring the solution of nonlinear ODEs. When the support of Q contains uncountably many models, the resulting optimization involves iterating over all data points at each step, which introduces significant computational overhead. This difficulty is further amplified in high-dimensional parameter spaces, where numerical integration or Monte Carlo estimation of expectations under Q becomes expensive or untractable.

Moreover, the pointwise transformation $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\Big)$ can be ill-conditioned for certain choices of f, as expresed by the need of verifing Assumption (b). Divergences such as reverse-relative entropy or Hellinger introduce sharp descontinuitis that amplify small numerical errors in the loss or in the estimation of $N_{Q,\boldsymbol{z}}$. As a consequence, implementations often rely on approximate solvers, discretization schemes. Despite its analytical structure, the ERM-fDR formulation therefore remains computationally challenging to scale in modern large-sample or high-dimensional settings.

Another important concern is the sensitivity of the learned distribution $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to both the choice of the f-divergence and the regularization parameter λ. Different divergences impose qualitatively distinct penalization: the relative entropy encourages smooth reweighting relative to Q, while the reverse-relative entropy produce heavy-tailed solutions from the analyisis in Chapter 3. However, current framework while it gives solutions for $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, the solution does not directly allows understand which choices of divergence may lead to desirable bias, and which to undesirable bias. In

practice, the interpretability of λ also varies with f, ℓ and the given data in $\boldsymbol{z}$, complicating interpretability.

6.3 Future Work

The theoretical framework developed in this work opens several avenues for future research that extend beyond its current scope.

A natural direction concerns the computational realization of f-divergence regularized learning. While the theoretical solutions establish existence and optimality conditions, practical implementation requires approximating intractable expectations and normalization terms. Future work could investigate stochastic and variational schemes to approximate these quantities efficiently, as well as explore gradient-based optimization methods that can accurately aproximate the solution.

Additionally, this work can explore the idea of laveraging the generalization expressions from Section 5.4 to find bounds. In particular, having acces to the explicit expresion, even if it is not computationaly tractable, under specific assumed priors, loss functions or other assumptions, computationaly tractable upper and lower bounds can be derivede. Furthermore, by exploring this bounds for different f-divergences it may shed light on desired properties of each f-divergene for specific senarios and guiding selection of regularizers.

Other areas of interest for future work include, exploring the connection between the ERMfDR formulation and Distributionally Robust Optimization (DRO). By recognizing that ERMfDR can be interpreted as a DRO problem, where the learner minimizes the worst-case risk over all distributions within an f-divergence ball around the empirical distribution, one can leverage the extensive algorithmic advances in DRO to make the theoretical framework of ERMfDR practically implementable. This connection provides a principled mechanism to quantify and control distributional shifts, thereby linking information-theoretic regularization to out-of-distribution (OOD) generalization and robust learning. Further research could investigate how different choices of f-divergence influence the robustnessgeneralization trade-off, and how sampling-based estimation methods such as contrastive variational inference or divergence-constrained MCMC can be adapted to approximate or sample from the induced robust solution. Bridging these perspectives would enable the design of practical, information-theoretically grounded DRO algorithms that achieve both robustness to distributional uncertainty and theoretical guarantees on generalization performance.

Overall, these research directions aim to bridge the gap between the theoretical understanding of information-regularized learning and its computational and practical realization, paving the way for the development of robust, interpretable, and theoretically grounded learning algorithms.

Appendix A

Preliminaries

A.1 Canonical form f-Divergence

Lemma A.1.1. *For all differentiable functions $f : [0, \infty) \to \mathbb{R}$ that induce an f-divergence, such that $\dot{f}(1) \neq 0$, there exists a function $\tilde{f} : [0, \infty) \to \mathbb{R}$ that satisfies $\dot{\tilde{f}}(1) = 0$ and that for all probability measures P and Q it holds that*

$$\mathsf{D}_f(P\|Q) = \mathsf{D}_{\tilde{f}}(P\|Q). \tag{A.1}$$

Proof: Let the function $f : [0, \infty) \to \mathbb{R}$ be differentiable and the function $\tilde{f} : [0, \infty) \to \mathbb{R}$ be

$$\tilde{f}(u) = f(u) - \dot{f}(1)(u - 1). \tag{A.2}$$

Note that, from (A.2) it follows that

$$\begin{aligned} \tilde{f}(1) &= f(1) - \dot{f}(1)(0) && \text{(A.3)} \\ &= 0. && \text{(A.4)} \end{aligned}$$

Similarly, the derivative of (A.2) yields,

$$\dot{\tilde{f}}(u) = \dot{f}(u) - \dot{f}(1), \tag{A.5}$$

which implies that

$$\dot{\tilde{f}}(1) = 0. \tag{A.6}$$

The proof continues by showing that the equality in (A.1) holds. From the definition of f-divergences, it follows that

$$\begin{aligned} \mathsf{D}_{\tilde{f}}(P\|Q) &= \int \tilde{f}\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(A.7)} \\ &= \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \dot{f}(1)\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) - 1\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(A.8)} \\ &= \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) - \dot{f}(1) \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) - 1\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(A.9)} \\ &= \mathsf{D}_f(P\|Q) - \dot{f}(1)\left(\int \mathrm{d}P(\boldsymbol{\theta}) - \int \mathrm{d}Q(\boldsymbol{\theta})\right) && \text{(A.10)} \\ &= \mathsf{D}_f(P\|Q), && \text{(A.11)} \end{aligned}$$

where (A.8) follows from (A.2), (A.9) follows from the change of measure, (A.10) follows from the fact that both P and Q are probability measures. This completes

the proof of (A.1). ■

A.2 Asymmetry Minor Results

This appendix concludes by presenting Lemma A.2.1 used in the proof of Lemma 3.3.1

Lemma A.2.1. *Let $\mathscr{M}$ be the set of measurable functions $h : \mathcal{M} \to \mathbb{R}$, 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 nonnegative functions that are absolutely integrable with respect to a probability measure Q. That is, for all $h \in \mathscr{S}$, it holds that*

$$\int |h(\boldsymbol{\theta})| \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{A.12}$$

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

$$\hat{r}(\alpha) = \int -\log(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{A.13}$$

for some functions g and h in $\mathscr{S}$ and $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small. The function $\hat{r}$ in (A.13) *is differentiable at zero.*

Proof: The objective is to prove that the function $\hat{r}$ in (A.13) is differentiable at zero, which reduces to proving that the limit

$$\lim_{\delta \to 0} \frac{1}{\delta} (\hat{r}(\alpha + \delta) - \hat{r}(\alpha)), \tag{A.14}$$

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

$$f(x) = -\log(x). \tag{A.15}$$

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

$$\hat{r}(\alpha) = \int f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{A.16}$$

The proof of the existence of such a limit in (A.14) relies on the fact that the function f in (A.13) is strictly convex and differentiable, which implies that f is also Lipschitz continuous. Hence, it follows that

$$\begin{aligned} &|f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))| \\ &\leq c \, |h(\boldsymbol{\theta})| \, |\delta|, \end{aligned} \tag{A.17}$$

for some positive and finite constant c, which implies that

$$\begin{aligned} &\frac{|f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))|}{|\delta|} \\ &\leq c \, |h(\boldsymbol{\theta})|, \end{aligned} \tag{A.18}$$

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

$$\begin{aligned} &\int \frac{|f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))|}{|\delta|} \, \mathrm{d}Q(\boldsymbol{\theta}) \\ &\leq \infty. \end{aligned} \tag{A.19}$$

This allows using the dominated convergence theorem as follows. From the fact that the function f is differentiable, let $\dot{f} : \mathbb{R} \to \mathbb{R}$ be the first derivative of f. The limit in (A.14) satisfies for $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small,

$$\begin{aligned}
&\lim_{\delta \to 0} \frac{1}{\delta}(\hat{r}(\alpha + \delta) - \hat{r}(\alpha)) \\
&= \lim_{\delta \to 0} \frac{1}{\delta}\left(\int f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \right. \\
&\quad \left. - \int f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \right) && \text{(A.20a)} \\
&= \lim_{\delta \to 0} \int \frac{1}{\delta}(f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) \\
&\quad - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})))\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(A.20b)} \\
&= \int \lim_{\delta \to 0} \frac{1}{\delta}(f(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta})) \\
&\quad - f(g(\boldsymbol{\theta}) + \alpha h(\boldsymbol{\theta})))\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(A.20c)} \\
&= \int \dot{f}(g(\boldsymbol{\theta}) + (\alpha + \delta)h(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(A.20d)} \\
&< \infty, && \text{(A.20e)}
\end{aligned}$$

where the equalities in (A.20c) and (A.20e) follow from the dominated convergence theorem [140, Theorem 1.6.9]. From (A.20e), it follows that the function $\hat{r}$ in (A.13) is differentiable at zero. This completes the proof. ■

A.3 Canonical form f-Divergence

Theorem A.3.1. *Given a probability measure space $(\mathcal{M}, \mathscr{F}(\mathcal{M}), \mu)$ and an open subset $\mathcal{A}$ od $\mathbb{R}$, let the function $f : \mathcal{A} \times \mathcal{M} \to \mathbb{R}$ be measurable with respect to $(\mathcal{A} \times \mathcal{M}, \mathscr{F})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. If for all $\nu \in \mathcal{M}$, the function $f(\cdot, \boldsymbol{\nu}) : \mathcal{A}_{Q,\boldsymbol{z}} \to \mathbb{R}$ is Lipschitz continuous and for some $u \in \mathcal{A}$, $\int f(u, \nu)\, \mathrm{d}\mu(\nu) < \infty$, then*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int f(t, \nu)\, \mathrm{d}\mu(\nu)\bigg|_{t=u} = \int \frac{\mathrm{d}}{\mathrm{d}t} f(t, \nu)\, \mathrm{d}\mu(\nu)\bigg|_{t=u}, \tag{A.21}$$

Proof: Note that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \int f(t, \nu)\, \mathrm{d}\mu(\nu)\bigg|_{t=u} &= \lim_{\delta \to 0} \frac{\int f(u + \delta, \nu)\, \mathrm{d}\mu(\nu) - \int f(u, \nu)\, \mathrm{d}\mu(\nu)}{\delta} && \text{(A.22)} \\
&= \lim_{\delta \to 0} \int \frac{f(u + \delta, \nu) - f(u, \nu)}{\delta}\, \mathrm{d}\mu(\nu), && \text{(A.23)}
\end{aligned}$$

where (A.23) follows from [140, Theorem 1.6.3]. The assumption that for all $\nu \in \mathcal{M}$, the function $f(\cdot, \boldsymbol{\nu})$ is Lipschitz continuous implies that for all $u \in \mathcal{A}$ and some $\delta \in \mathbb{R}$,

$$|f(u + \delta, \nu) - f(u, \nu)| < L\,|\delta|, \tag{A.24}$$

$$\tag{A.25}$$

with $L < \infty$. And thus, dividing the RHS and LHS of (A.24) by $|\delta|$ yields

$$\left| \frac{f(u + \delta, \nu) - f(u, \nu)}{\delta} \right| < L, \tag{A.26}$$

(A.27)

which implies that

$$\int \left| \frac{f(u+\delta,\nu) - f(u,\nu)}{\delta} \right| \mathrm{d}\mu(\nu) < \infty. \tag{A.28}$$

(A.29)

This allows using the dominated convergence theorem [140, Theorem 1.6.9] as follows. From (A.23), the following holds

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \int f(t,\nu)\,\mathrm{d}\mu(\nu)\Big|_{t=u} &= \lim_{\delta\to 0} \int \frac{f(u+\delta,\nu) - f(u,\nu)}{\delta}\,\mathrm{d}\mu(\nu) && \text{(A.30)} \\ &= \int \lim_{\delta\to 0} \frac{f(u+\delta,\nu) - f(u,\nu)}{\delta}\,\mathrm{d}\mu(\nu) && \text{(A.31)} \\ &= \int \frac{\mathrm{d}}{\mathrm{d}t} f(t,\nu)\,\mathrm{d}\mu(\nu)\Big|_{t=u}, && \text{(A.32)} \end{aligned}$$

where (A.31) follows from the dominated convergence theorem [140, Theorem 1.6.9]. This completes the proof. ■

Lemma A.3.1. *Given a strictly convex and differentiable function $f : \mathcal{I} \to \mathbb{R}$, the inverse of the derivative of f, denoted by the function $\dot{f}^{-1} : \mathcal{J} \to \mathcal{I}$, is strictly increasing.*

Proof: From the assumption that the function $f : \mathcal{I} \to \mathbb{R}$ is strictly convex, it follows from the strict convexity definition that the derivative $\dot{f} : \mathcal{I} \to \mathcal{J}$ is strictly increasing. Using the continuous inverse theorem in [141, Theorem 5.6] implies that the function $\dot{f}^{-1} : \mathcal{J} \to \mathcal{I}$ is strictly increasing, which completes the proof. ■

Lemma A.3.2. *Given a strictly convex and twice differentiable function $f : \mathcal{I} \to \mathbb{R}$, and a differentiable function $h : \mathcal{I}^\star \to \mathcal{I}$, for all $x \in \mathcal{I}^\star$ it holds that*

$$\frac{\mathrm{d}}{\mathrm{d}x} \dot{f}^{-1}(h(x)) = \frac{\dot{h}(x)}{\ddot{f}\left(\dot{f}^{-1}(x)\right)}. \tag{A.33}$$

Proof: Let the function $g : \mathcal{I} \to \mathbb{R}$ be defined for all $x \in \mathcal{I}^\star$ by

$$g(x) = \dot{f}^{-1}(h(x)) \tag{A.34}$$

By the definition of the inverse function, it follows that

$$\dot{f}(g(x)) = h(x). \tag{A.35}$$

Differentiating (A.35) with respect to x yields

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} \dot{f}(g(x)) &= \ddot{f}(g(x))\,\dot{g}(x) && \text{(A.36)} \\ &= \dot{h}(x). && \text{(A.37)} \end{aligned}$$

From (A.36) the derivative of the function g in (A.35) is given by

$$\dot{g}(x) = \frac{\dot{h}(x)}{\ddot{f}(g(x))} \tag{A.38}$$

$$= \frac{\dot{h}(x)}{\ddot{f}\Big(\dot{f}^{-1}(h(x))\Big)}, \tag{A.39}$$

where (A.39) follows from (A.34). Hence, from (A.34) and (A.39) it follows from

$$\frac{\mathrm{d}}{\mathrm{d}x}\dot{f}^{-1}(h(x)) = \frac{\dot{h}(x)}{\ddot{f}\Big(\dot{f}^{-1}(h(x))\Big)}. \tag{A.40}$$

This completes the proof. ■

Lemma A.3.3. *The* Legendre-Fenchel *transform of strictly convex and differentiable function f, satisfies for all $t \in \mathcal{J}$, with $\mathcal{J}$ in* (1.8)*,*

$$f^*(t) = t\dot{f}^*(t) - f(\dot{f}^*(t)). \tag{A.41}$$

Proof: From the *Legendre-Fenchel* transform in Definition 1.5.2 it holds that for all $t \in \mathcal{J}$, with $\mathcal{J}$ in (1.8),

$$f^*(t) = \sup_{x\in\mathcal{I}}(tz - f(x)). \tag{A.42}$$

For any $z \in \mathcal{I}$, setting $x = z$ yields

$$f^*(y) \geq yx - f(x), \tag{A.43}$$

which rearranges to the Fenchel inequality,

$$f(x) + f^*(y) \geq xy, \tag{A.44}$$

where equality in (A.44) holds if and only if,

$$f^*(y) = yx - f(x). \tag{A.45}$$

By definition of f^* in (A.42), the equality (A.45) implies that x achieves the supremum in $f^*(y)$, which will be denoted by x_y. In other words, x_y is the maximizing argument

$$x_y = \arg\max_{x\in\mathcal{I}} xy - f(x). \tag{A.46}$$

Note that under Assumption (a), the solution to the maximization problem

$$\max_{x\in\mathcal{I}} xy - f(x), \tag{A.47}$$

is unique, and satisfies

$$\frac{\mathrm{d}}{\mathrm{d}x}(xy - f(x)) = y - \dot{f}(x) \tag{A.48}$$

$$= 0. \tag{A.49}$$

From (A.49), the maximizing argument x_y in (A.46) satisfies

$$x_y = \dot{f}^{-1}(y) \tag{A.50}$$

$$= \dot{f}^*(y), \tag{A.51}$$

where (A.51) follows from [137, Corollary 23.5.1]. Hence, from (A.51), the *Legendre-Fenchel* transform of function the f, under Assumption (a), satisfies for all $t \in \mathcal{J}$,

$$f^*(t) = t\dot{f}^*(t) - f(\dot{f}^*(t)), \tag{A.52}$$

which completes the proof. ■

Appendix B

ERM-RER Type-II

B.1 Proof of Lemma 3.3.1

The optimization problem in (3.6) can be rewritten in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the measure Q, which yields:

$$\min_{P\in\bigcirc_Q(\mathcal{M})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q(\boldsymbol{\theta}), \tag{B.1a}$$

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

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

$$\int |\hat{g}(\boldsymbol{\theta})|\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{B.2}$$

Hence, the optimization problem of interest is:

$$\min_{g\in\mathscr{M}} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \log(g(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{B.3a}$$

$$\text{s.t.} \int g(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{B.3b}$$

Let the Lagrangian of the optimization problem in (B.3) be $L : \mathscr{M} \times \mathbb{R} \to \mathbb{R}$ such that

$$L(g,\beta) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - \lambda \int \log(g(\boldsymbol{\theta}))\ \mathrm{d}Q(\boldsymbol{\theta}) + \beta\left(\int g(\boldsymbol{\theta})\ \mathrm{d}Q(\boldsymbol{\theta}) - 1\right) \tag{B.4}$$

$$= \int \Big(g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta) - \lambda \log(g(\boldsymbol{\theta}))\Big)\ \mathrm{d}Q(\boldsymbol{\theta}) - \beta, \tag{B.5}$$

where β is a real that acts as a Lagrange multiplier due to the constraint (B.3b). Let $\hat{g} : \mathcal{M} \to \mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (B.4) at $(g, \beta) \in \mathscr{M} \times \mathbb{R}$ in the direction of $\hat{g}$ is

$$\partial L(g, \beta; \hat{g}) \triangleq \frac{\mathrm{d}}{\mathrm{d}\gamma} r(g + \gamma\hat{g}, \beta)\bigg|_{\gamma=0}, \tag{B.6}$$

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

$$\begin{aligned} r(\gamma) = & \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \\ & - \lambda \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \\ & + \beta\left(\int (g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) - 1\right) \qquad \text{(B.7a)} \\ = & \gamma \int \hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\, \mathrm{d}Q(\boldsymbol{\theta}) \\ & + \lambda \int \log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}) \\ & + \int g(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\, \mathrm{d}Q(\boldsymbol{\theta}) - \beta. \qquad \text{(B.7b)} \end{aligned}$$

Note that the first term in (B.7b) is linear with respect to γ; the second term can be written using the function $\hat{r} : \mathbb{R} \to \mathbb{R}$ in (A.13) such that

$$\hat{r}(\gamma) = \int -\log(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta}); \tag{B.8}$$

and the remaining terms are independent of γ.

Hence, based on the fact that the function $\hat{r}$ in (B.8) is differentiable at zero (Lemma A.2.1), so is the function r in (B.7), which implies that the Gateaux differential of $\partial L(g, \beta; \hat{g})$ in (B.6) exists. The derivative of the real function r in (B.7) is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) = & \int \hat{g}(\boldsymbol{\theta})(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)\, \mathrm{d}Q(\boldsymbol{\theta}) \\ & - \lambda \int \frac{\hat{g}(\boldsymbol{\theta})}{(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}\, \mathrm{d}Q(\boldsymbol{\theta}) \qquad \text{(B.9)} \\ = & \int \hat{g}(\boldsymbol{\theta})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{(g(\boldsymbol{\theta}) + \gamma\hat{g}(\boldsymbol{\theta}))}\right)\mathrm{d}Q(\boldsymbol{\theta}). \qquad \text{(B.10)} \end{aligned}$$

From (B.6) and (B.10), it follows that

$$\partial L(g, \beta; \hat{g}) = \int \hat{g}(\boldsymbol{\theta})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta - \frac{\lambda}{g(\boldsymbol{\theta})}\right)\mathrm{d}Q(\boldsymbol{\theta}). \tag{B.11}$$

The relevance of the Gateaux differential in (B.11) stems from [24, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (B.4) to have a stationary point at $\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta\right) \in \mathcal{M} \times [0, \infty)$ is that for all functions $\hat{g} \in \mathscr{M}$,

$$\partial L\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}, \beta; \hat{g}\right) = 0. \tag{B.12}$$

From (B.11) and (B.12), it follows that $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ must satisfy for all functions $\hat{g}$ in $\mathscr{M}$ that

$$\int \hat{g}(\boldsymbol{\theta})\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta-\lambda\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1}\right)\mathrm{d}Q(\boldsymbol{\theta})=0. \tag{B.13}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta-\lambda\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1}=0, \tag{B.14}$$

and thus,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=\frac{\lambda}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}, \tag{B.15}$$

where β is chosen to satisfy (B.3b) 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,

$$\beta\in\left\{t\in\mathbb{R}:\forall\boldsymbol{\theta}\in\operatorname{supp}Q, 0<\frac{\lambda}{t+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right\},\ \text{and} \tag{B.16}$$

$$1=\int\frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{B.17}$$

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 (B.15) is the unique solution to the optimization problem in (B.1). Such verification is done by showing that the objective function in (B.1) 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,

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}}(\alpha P_1+(1-\alpha)P_2)+\lambda\mathsf{D}(\alpha P_1+(1-\alpha)P_2\|Q)\\
&=\mathsf{R}_{\boldsymbol{z}}(\alpha P_1)+\mathsf{R}_{\boldsymbol{z}}((1-\alpha)P_2)\\
&\quad+\lambda\mathsf{D}(\alpha P_1+(1-\alpha)P_2\|Q) &\text{(B.18)}\\
&>\alpha\mathsf{R}_{\boldsymbol{z}}(P_1)+(1-\alpha)\mathsf{R}_{\boldsymbol{z}}(P_2)\\
&\quad+\lambda(\alpha\mathsf{D}(P_1\|Q)+(1-\alpha)\mathsf{D}(P_2\|Q)) &\text{(B.19)}
\end{aligned}$$

where the functional $\mathsf{R}_{\boldsymbol{z}}$ is defined in (2.12). The equality above follows from the properties of the Lebesgue integral, while the inequality follows from [84, Theorem 2]. This proves that the solution is unique due to the strict concavity of the objective function, which completes the proof. ■

B.2 Proof of Lemma 3.3.2

Given a probability measure $V\in\bigtriangledown_Q(\mathcal{M})$, with $\bigtriangledown_Q(\mathcal{M})$ in (3.2), 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

$$V_0(\mathcal{A})=\frac{V(\mathcal{A}\setminus\operatorname{supp}Q)}{V(\mathcal{M}\setminus\operatorname{supp}Q)}, \tag{B.20a}$$

and

$$V_1(\mathcal{A}) = \frac{V(\mathcal{A} \cap \operatorname{supp} Q)}{V(\mathcal{M} \cap \operatorname{supp} Q)}. \tag{B.20b}$$

Let the real value α be

$$\alpha = V(\mathcal{M} \cap \operatorname{supp} Q). \tag{B.21}$$

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

$$V(\mathcal{A}) = (1-\alpha)V_0(\mathcal{A}) + \alpha V_1(\mathcal{A}). \tag{B.22}$$

Moreover, from (B.22) it holds that: *i*) If $V(\mathcal{A}) = 0$, then $V_0(\mathcal{A}) = 0$, which implies that V_0 is absolutely continuous with respect to V; *ii*) If $V(\mathcal{A}) = 0$, then $V_1(\mathcal{A}) = 0$, which implies that V_1 is absolutely continuous with respect to V. Furthermore, from the definition of $\bigtriangledown_Q(\mathcal{M})$ in (3.2), the probability measure Q is absolutely continuous with respect to V. Hence, for all $\mathcal{A} \in \mathscr{F}$, it follows that

$$\begin{aligned}
Q(\mathcal{A}) &= \int_{\mathcal{A}} \mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.23a)}\\
&= \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V(\boldsymbol{\theta}) && \text{(B.23b)}\\
&= \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}((1-\alpha)V_0 + \alpha V_1)(\boldsymbol{\theta}) && \text{(B.23c)}\\
&= (1-\alpha) \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V_0(\boldsymbol{\theta})\\
&\quad + \alpha \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V_1(\boldsymbol{\theta}) && \text{(B.23d)}\\
&= \int_{\mathcal{A}} \alpha \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\, \mathrm{d}V_1(\boldsymbol{\theta}). && \text{(B.23e)}
\end{aligned}$$

Hence, from (B.23e) and the Radon-Nikodym Theorem in [140, Theorem 2.2.1, page 65] the probability measure Q is absolutely continuous with respect to V_1. This implies that for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$Q(\mathcal{A}) = \int \frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\, \mathrm{d}V_1(\boldsymbol{\theta}), \tag{B.24}$$

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

$$\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta}) = \alpha \frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta}). \tag{B.25}$$

From (B.25), the following holds:

$$\begin{aligned}
\mathsf{D}(Q\|V) &= \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}V}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.26a)}\\
&= \int \log\left(\frac{1}{\alpha}\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.26b)}\\
&= \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}V_1}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\\
&\quad - \int \log(\alpha)\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.26c)}\\
&= \mathsf{D}(Q\|V_1) - \log(\alpha). && \text{(B.26d)}
\end{aligned}$$

From (B.26) it follows that

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}}(V) + \lambda \mathsf{D}(Q\|V) \\
&\quad = \mathsf{R}_{\boldsymbol{z}}((1-\alpha)V_0 + \alpha V_1) + \lambda \mathsf{D}(Q\|V_1) \\
&\qquad -\lambda \log(\alpha) && \text{(B.27a)} \\
&\quad = (1-\alpha)\mathsf{R}_{\boldsymbol{z}}(V_0) + \alpha \mathsf{R}_{\boldsymbol{z}}(V_1) + \lambda \mathsf{D}(Q\|V_1) \\
&\qquad - \lambda \log(\alpha) && \text{(B.27b)} \\
&\quad \geq \alpha \mathsf{R}_{\boldsymbol{z}}(V_1) + \lambda \mathsf{D}(Q\|V_1) - \lambda \log(\alpha), && \text{(B.27c)}
\end{aligned}$$

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

$$\begin{aligned}
V(\mathcal{A}) &= V_1(\mathcal{A}) && \text{(B.28a)} \\
&= V(\mathcal{A} \cap \operatorname{supp} Q), && \text{(B.28b)}
\end{aligned}$$

where (B.28b) follows from (B.20b). This implies that (B.27c) holds if and only if

$$\operatorname{supp} Q = \operatorname{supp} V, \tag{B.29}$$

which implies that (B.27c) holds if and only if the measure V is mutually absolutely continuous with respect to the reference measure Q. Finally, the above leads to

$$\begin{aligned}
&\min_{P \in \nabla_Q(\mathcal{M}) \setminus \bigcirc_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P) \\
&> \min_{P \in \nabla_Q(\mathcal{M})} \mathsf{R}_{\boldsymbol{z}}(P) + \lambda \mathsf{D}(Q\|P), && \text{(B.30)}
\end{aligned}$$

which completes the proof. ■

B.3 Proof of Lemma 3.3.4

The proof is divided into two parts. The first part proves the monotonicity of the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (3.14); while the second part proves the continuity of the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$. The proof is finalized by using the continuous inverse theorem [141, Theorem 5.6.5] to show both the monotonicity and continuity of the function $\bar{K}_{Q,\boldsymbol{z}}$ in (4.7).

The first part is as follows. Let λ and β be two reals that satisfy (3.11b). Hence, $0 < \lambda < \infty$ and from (C.108), it holds that

$$0 < \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{B.31}$$

which, together with (3.14), imply

$$\infty > \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) > 0. \tag{B.32}$$

That is, the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (3.14) 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 (4.7a) and $\gamma_1 < \gamma_2$. Hence, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_1} > \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \gamma_2}, \tag{B.33}$$

which implies that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\gamma_1}\,\mathrm{d}Q(\boldsymbol{\theta}) > \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\gamma_2}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{B.34}$$

and thus, from (3.14), it holds that

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma_1) < \bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma_2). \tag{B.35}$$

This proves that the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (3.14) 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 $\bar{K}^{-1}_{Q,\boldsymbol{z}}$. To do so, two auxiliary functions are introduced and proven to be continuous. Then, the fact that $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (3.14) is the composition of the two auxiliary functions is leveraged to prove its continuity. Let the function $h : (0,\infty) \to (0,\infty)$ be

$$h(x) = \frac{1}{x}. \tag{B.36}$$

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

$$k(\gamma) = \int h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{B.37}$$

The first step is to prove that the function k in (B.37) 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 (4.7), then from (3.3a), 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 $\gamma > -\delta^{\star}_{Q,\boldsymbol{z}}$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (3.16). Hence, the proof of continuity of the function k in (B.37) is restricted to $\left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$.

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

$$h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)) > h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)). \tag{B.38}$$

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

$$h\big(\gamma + \delta^{\star}_{Q,\boldsymbol{z}}\big) \geq h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})), \tag{B.39}$$

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

$$\lim_{\gamma \to a} h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) = h(a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})). \tag{B.40}$$

Hence, from the dominated convergence theorem [140, Theorem 1.6.9], the following limit exists and satisfies

$$\lim_{\gamma \to a} k(b) = \lim_{\gamma \to a} \int h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{B.41a}$$

$$= \int \left(\lim_{\gamma \to a} h(\gamma + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})) \right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{B.41b}$$

$$= \int h(a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{B.41c}$$

$$= k(a), \tag{B.41d}$$

where (B.41c) follows from (B.40). The equality in (B.41d) proves that the function k in (B.37) is continuous in the interval $\left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. Note that from (3.14) and (B.37), it holds that

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma) = \frac{1}{k(\gamma)}. \tag{B.42}$$

Using (B.42), for all $a \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, it holds that

$$\lim_{\gamma \to a} \bar{K}^{-1}_{Q,\boldsymbol{z}}(\gamma) = \lim_{\gamma \to a} \frac{1}{k(\gamma)} \tag{B.43}$$

$$= \frac{1}{\lim\limits_{\gamma \to a} k(\gamma)} \tag{B.44}$$

$$= \frac{1}{\int h(a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\, \mathrm{d}Q(\boldsymbol{\theta})} \tag{B.45}$$

$$= \frac{1}{\int \frac{1}{a + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\, \mathrm{d}Q(\boldsymbol{\theta})} \tag{B.46}$$

$$= \bar{K}^{-1}_{Q,\boldsymbol{z}}(a), \tag{B.47}$$

where (B.44) follows from the continuity of the function h in (B.36) over the interval $(0, \infty)$; (B.45) follows from (B.41d); and (B.46) follows from (B.36). Thus, the existence of the limit in (B.43) 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 [141, 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 (4.7). This concludes the proof. ■

B.4 Proof of Lemma 3.3.5

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 (3.16). Each of these intervals is studied separately.

Let β be such that $\bar{K}_{Q,\boldsymbol{z}}(\lambda) = \beta$, with λ in (3.1) and assume that $\beta \in \left(-\infty, -\delta^{\star}_{Q,\boldsymbol{z}}\right)$. Under this assumption, the inclusion in (3.3a) 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 (3.3a). This implies that

$$\left(-\infty, -\delta^{\star}_{Q,\boldsymbol{z}}\right) \cap \mathcal{C}_{Q,\boldsymbol{z}} = \emptyset. \tag{B.48}$$

Assume now that $\beta \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$. Then, from (3.16), it can be verified that the constraint in (3.3a) is satisfied. More speciffically, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta > 0$. The proof continues by showing that (3.3b) is also verified. For this

purpose, note that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}) \leq \int \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}}+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{B.49a}$$

$$= \frac{1}{\delta^{\star}_{Q,\boldsymbol{z}}+\beta} \tag{B.49b}$$

$$< \infty. \tag{B.49c}$$

The finiteness of the integral in the left-hand side of (B.49a) implies that

$$0 < \lambda \tag{B.50a}$$

$$= \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) \tag{B.50b}$$

$$= \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta})} \tag{B.50c}$$

$$< \infty, \tag{B.50d}$$

where (B.50a) follows from the assumption that $\lambda \in \mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0,\infty)$; (B.50b) follows from the fact that $\bar{K}_{Q,\boldsymbol{z}}(\lambda) = \beta$; (B.50c) follows from (3.14); and (B.50d) follows from the inequality in (B.49c). In a nutshell,

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}) < \infty, \text{ and} \tag{B.51}$$

$$\lambda < \infty, \tag{B.52}$$

which implies that the product

$$\lambda \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}) = \int \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{B.53a}$$

$$= \int \frac{\frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}\,\mathrm{d}Q(\boldsymbol{\nu})}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{B.53b}$$

$$= 1, \tag{B.53c}$$

where (B.53b) follows from (B.50c). This verifies (3.3b), which implies that

$$\big(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\big) \subseteq \mathcal{C}_{Q,\boldsymbol{z}}. \tag{B.54}$$

Finally, under the assumption that $\beta = -\delta^{\star}_{Q,\boldsymbol{z}}$, two cases are considered: (a) $Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big) > 0$; and (b) $Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big) = 0$, with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ defined in (3.18). In case (a), if $\beta = -\delta^{\star}_{Q,\boldsymbol{z}}$ and $Q\Big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\Big) > 0$, then for all $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}}$, it follows that

$$\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} Q\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) = \infty, \tag{B.55}$$

which implies that,

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\,\mathrm{d}Q(\boldsymbol{\theta}) = \infty. \tag{B.56}$$

The equality in (B.56) implies that the constraint (3.3b) is not satisfied. Therefore, for case (a) it follows that,

$$-\delta^{\star}_{Q,z} \notin \mathcal{C}_{Q,z}. \tag{B.57}$$

In the alternative case (b), if $\beta = -\delta^{\star}_{Q,z}$ and $Q\Big(\mathcal{L}^{\star}_{Q,z}\Big) = 0$, then, the integral in (3.14) is either

$$\int \frac{1}{\mathsf{L}_{z}(\boldsymbol{\theta}) - \delta^{\star}_{Q,z}} \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{B.58}$$

which implies that $-\delta^{\star}_{Q,z} \in \mathcal{C}_{Q,z}$, with $\mathcal{C}_{Q,z}$ defined in (4.7a), or the integral is

$$\int \frac{1}{\mathsf{L}_{z}(\boldsymbol{\theta}) - \delta^{\star}_{Q,z}} \, \mathrm{d}Q(\boldsymbol{\theta}) = \infty, \tag{B.59}$$

which implies that $-\delta^{\star}_{Q,z} \notin \mathcal{C}_{Q,z}$. Hence, from (B.48), (B.54), (B.57), (B.58), and (B.59) the set $\mathcal{C}_{Q,z}$ in (4.7a) is either the open set $\Big(-\delta^{\star}_{Q,z}, \infty\Big)$ or the closed set $\Big[-\delta^{\star}_{Q,z}, \infty\Big)$. Note that the equality $\mathcal{C}_{Q,z} = \Big[-\delta^{\star}_{Q,z}, \infty\Big)$ is observed, if and only if,

$$\int \frac{1}{\mathsf{L}_{z}(\boldsymbol{\theta}) - \delta^{\star}_{Q,z}} \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{B.60}$$

which completes the first part of the proof.

The second part of the proof is as follows. Two cases are considered: *i)* $\mathcal{C}_{Q,z} = \Big[-\delta^{\star}_{Q,z}, \infty\Big)$; and *ii)* $\mathcal{C}_{Q,z} = \Big(-\delta^{\star}_{Q,z}, \infty\Big)$. In case $i)$, the value $-\delta^{\star}_{Q,z}$ is in the domain of the function $\bar{K}^{-1}_{Q,z}$ in (3.14), that is, the set $\mathcal{C}_{Q,z}$. Given that the function $\bar{K}^{-1}_{Q,z}$ is strictly increasing, then, $-\delta^{\star}_{Q,z}$ should be mapped to the smallest value in the range of $\bar{K}^{-1}_{Q,z}$, denoted by $\lambda_{Q,z}$. Hence,

$$\bar{K}^{-1}_{Q,z}\big(-\delta^{\star}_{Q,z}\big) = \lambda_{Q,z} \tag{B.61}$$

$$> 0, \tag{B.62}$$

where (B.62) follows from the fact that zero is not in the domain of the function $\bar{K}_{Q,z}$, that is, the set $\mathcal{A}_{Q,z}$. Using these elements, it is concluded that the set $\mathcal{A}_{Q,z}$ is the interval $[\lambda_{Q,z}, \infty)$, which ends the analysis of case $i)$.

In the second case, from Lemma 3.3.4, the continuity and strict monotonicity of the function $\bar{K}_{Q,z}$ in (4.7) imply that $\mathcal{A}_{Q,z} = \Big(\lambda^{\star}_{Q,z}, \infty\Big)$, with $\lambda^{\star}_{Q,z}$ in (3.17). The remaining of the proof focuses on showing that $\lambda^{\star}_{Q,z} = 0$ in this case, and thus, $\mathcal{A}_{Q,z} = (0, \infty)$. From Lemma 3.3.4 and the continuous inverse theorem [141, Theorem 5.6.5], it follows that function $\bar{K}^{-1}_{Q,z}$ is strictly increasing and continuous. Hence, using (3.14), it holds that

$$\lim_{\gamma \to -\delta^{\star}_{Q,z}{}^{+}} \bar{K}^{-1}_{Q,z}(\gamma) = \lim_{\gamma \to -\delta^{\star}_{Q,z}{}^{+}} \frac{1}{\int \frac{1}{\mathsf{L}_{z}(\boldsymbol{\theta}) + \gamma} \, \mathrm{d}Q(\boldsymbol{\theta})} \tag{B.63}$$

$$= \frac{1}{\lim\limits_{\gamma \to -\delta^{\star}_{Q,z}{}^{+}} \int \frac{1}{\mathsf{L}_{z}(\boldsymbol{\theta}) + \gamma} \, \mathrm{d}Q(\boldsymbol{\theta})} \tag{B.64}$$

$$= \frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta})} \tag{B.65}$$

$$= 0, \tag{B.66}$$

where (B.64) follows from [142, Theorem 4.4], that permits the change of the limit to the reciprocal; and (B.65) follows from (B.41); and (B.66) follows from (B.59). From Lemma 3.3.4 and (B.66), 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. ■

B.5 Proof of Lemma 3.3.9

From Lemma 3.3.4, the function $\bar{K}_{Q,\boldsymbol{z}}$ in (4.7) is strictly increasing and continuous. Additionally, from Lemma 3.3.5, 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 3.3.4 and Lemma 3.3.5, it follows that

$$\lim_{\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} \bar{K}_{Q,\boldsymbol{z}}(\lambda) = -\delta^{\star}_{Q,\boldsymbol{z}}, \tag{B.67}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ defined in (3.16) and $\lambda^{\star}_{Q,\boldsymbol{z}}$ defined in (3.17). ■

B.6 Proof of Lemma 3.3.10

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) \geq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2), \tag{B.68}$$

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

$$\frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)} \leq \frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)}, \tag{B.69a}$$

which implies

$$\frac{\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1)\right)^{-1}}{\int \left(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)^{-1} \mathrm{d}Q(\boldsymbol{\nu})} \leq \frac{\left(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)\right)^{-1}}{\int \left(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)^{-1} \mathrm{d}Q(\boldsymbol{\nu})}. \tag{B.70}$$

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

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1) \leq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2), \tag{B.71}$$

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

B.7 Proof of Lemma 3.3.11

From Lemma 3.3.10, it follows that for all $\lambda \in (0, \infty)$, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, and for all $\boldsymbol{\phi} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$, it holds that

$$\begin{aligned}
\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &\leq \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\phi}) && \text{(B.72a)}\\
&= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && \text{(B.72b)}\\
&= \frac{\lambda}{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && \text{(B.72c)}\\
&< \infty, && \text{(B.72d)}
\end{aligned}$$

where (B.72a) follows from (3.4); the equality in (B.72b) follows from the fact that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi}) \geq \delta^{\star}_{Q,\boldsymbol{z}}$; and (B.72d) follows from the fact that for all $\lambda > 0$, the function $\bar{K}_{Q,\boldsymbol{z}}(\lambda) < \infty$. From the definition of $\delta^{\star}_{Q,\boldsymbol{z}}$ in (3.16) and $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ in (3.18) equality in (B.72a) holds if and only if $\boldsymbol{\theta} \in \mathcal{L}^{\star}_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$. This completes the proof of finiteness.

For the proof of positivity, observe that from Lemma 3.3.4 for all $\lambda > 0$, it holds that

$$-\delta^{\star}_{Q,\boldsymbol{z}} < \bar{K}_{Q,\boldsymbol{z}}(\lambda) < \infty, \tag{B.73}$$

which implies

$$0 < \delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{B.74}$$

From (3.4) and (B.74), it follows that

$$\begin{aligned}
\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} && \text{(B.75a)}\\
&= \frac{\frac{1}{\int \frac{1}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})}}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} && \text{(B.75b)}\\
&> 0, && \text{(B.75c)}
\end{aligned}$$

which completes the proof. ■

B.8 Proof of Lemma 3.3.12

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 $\boldsymbol{\theta} \in \operatorname{supp} Q$ that

$$\begin{aligned}
&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\\
&= \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} && \text{(B.76a)}\\
&= \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \frac{1}{\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})} && \text{(B.76b)}\\
&= \frac{1}{\int \frac{\beta}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu}) + \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})}. && \text{(B.76c)}
\end{aligned}$$

Note that from the function $\bar{K}_{Q,\boldsymbol{z}}$ in (4.7) the equation in (B.76) can be written in term of λ such that

$$\begin{aligned}&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\\&= \frac{1}{\int \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})+\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})}.\end{aligned} \tag{B.77}$$

Furthermore, by Lemma 3.3.5, the case where λ increases is equivalent to evaluating the case where β increases, and from (B.77), it follows that

$$\begin{aligned}&\lim_{\lambda\to\infty}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\\&=\lim_{\lambda\to\infty}\frac{1}{\int\frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\mathrm{d}Q(\boldsymbol{\nu})+\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})}&\text{(B.78a)}\\&=\frac{1}{\lim_{\beta\to\infty}\int\frac{\beta}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\mathrm{d}Q(\boldsymbol{\nu})+\lim_{\beta\to\infty}\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\mathrm{d}Q(\boldsymbol{\nu})}.&\text{(B.78b)}\end{aligned}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (3); and (B.78b) follows from Theorem 3.3.1, which implies that the terms in the denominator are positive and the fact that the function $g(x)=\frac{1}{x}$ is continuous over the positive reals. Recall that from the definition of the function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) 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

$$\begin{aligned}&\lim_{\beta\to\infty}\int\frac{\beta}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})\\&=\int\lim_{\beta\to\infty}\frac{\beta}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})&\text{(B.79a)}\\&=\int\mathrm{d}Q(\boldsymbol{\nu})&\text{(B.79b)}\\&=1,&\text{(B.79c)}\end{aligned}$$

where (B.79a) follows from the dominated convergence theorem [140, Theorem 1.6.9]; and,

$$\begin{aligned}&\lim_{\beta\to\infty}\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})\\&=\int\lim_{\beta\to\infty}\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\,\mathrm{d}Q(\boldsymbol{\nu})&\text{(B.80a)}\\&=\int 0\,\mathrm{d}Q(\boldsymbol{\nu})&\text{(B.80b)}\\&=0,&\text{(B.80c)}\end{aligned}$$

where (B.80a) also follows from the dominated convergence theorem [140, Theorem 1.6.9]; Substituting (B.79) and (B.80) into (B.78) yields

$$\lim_{\lambda\to\infty}\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})=1, \tag{B.81}$$

which completes the proof. ■

B.9 Proof of Lemma 3.3.13

From Theorem 3.3.1, the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}
\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} && \text{(B.82a)}\\
&= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && \text{(B.82b)}\\
&= \frac{\bar{K}^{-1}_{Q,\boldsymbol{z}}\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda)\big)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && \text{(B.82c)}\\
&= \frac{\frac{1}{\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} && \text{(B.82d)}\\
&= \left(\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1}, && \text{(B.82e)}
\end{aligned}$$

where (B.82b) follows from (4.7); and (B.82d) follows from (3.14). 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:

$$\begin{aligned}
\mathcal{A}_0(\boldsymbol{\theta}) &= \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = 0\}, && \text{(B.83a)}\\
\mathcal{A}_1(\boldsymbol{\theta}) &= \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < 0\}, \text{ and} && \text{(B.83b)}\\
\mathcal{A}_2(\boldsymbol{\theta}) &= \{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > 0\}. && \text{(B.83c)}
\end{aligned}$$

Using the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (B.83), the following holds for all $\boldsymbol{\theta} \in \operatorname{supp} Q$.

$$\begin{aligned}
&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\\
&= \left(\int_{\mathcal{A}_0(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right.\\
&\quad + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\\
&\quad \left. + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1} && \text{(B.84a)}\\
&= \left(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right.\\
&\quad \left. + \int_{\mathcal{A}_2(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1}. && \text{(B.84b)}
\end{aligned}$$

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

$$\begin{aligned}
&\big\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^{\star}_{Q,\boldsymbol{z}}\big\}, && \text{(B.85a)}\\
&\big\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^{\star}_{Q,\boldsymbol{z}}\big\}, \text{ and} && \text{(B.85b)}\\
&\big\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^{\star}_{Q,\boldsymbol{z}}\big\}, && \text{(B.85c)}
\end{aligned}$$

with $\delta^\star_{Q,\boldsymbol{z}}$ in (3.16). The proof is divided into two cases. The first case follows under the assumption that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) = \infty; \tag{B.86}$$

and the second case follows under the assumption that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{B.87}$$

From Lemma 3.3.5, it follows that in Case 1, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7) 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 (3.27) and (3.28). Case 2 considers the limit $\lambda \to {\lambda^\star_{Q,\boldsymbol{z}}}^+$, which comprehends the equality (3.29).

B.9.1 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 (B.85).

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

$$\left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\right\} = \mathcal{L}^\star_{Q,\boldsymbol{z}}, \tag{B.88}$$

with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ defined in (3.18). Hence, the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (B.83) satisfy the following:

$$\mathcal{A}_0(\boldsymbol{\theta}) = \mathcal{L}^\star_{Q,\boldsymbol{z}}, \tag{B.89a}$$
$$\mathcal{A}_1(\boldsymbol{\theta}) = \left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta^\star_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{B.89b}$$
$$\mathcal{A}_2(\boldsymbol{\theta}) = \left\{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star_{Q,\boldsymbol{z}}\right\}. \tag{B.89c}$$

From the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (3.16), it follows that $Q(\mathcal{A}_2(\boldsymbol{\theta})) = 0$. Substituting the equalities in (B.89) in (B.84) yields for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star_{Q,\boldsymbol{z}}\right\}$,

$$\begin{aligned}&\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\\ &= \left(Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu})\right)^{-1},\end{aligned} \tag{B.90}$$

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

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

$$= \left(Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) + \lim_{\lambda \to 0^+} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \, \mathrm{d}Q(\boldsymbol{\nu}) \right)^{-1} \tag{B.91}$$

$$= \begin{cases} \infty & \text{if } Q\left(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\right) = 0 \\ \frac{1}{Q(\mathcal{L}^{\star}_{Q,\boldsymbol{z}})} & \text{otherwise} \end{cases}, \tag{B.92}$$

where (B.92) follows from verifying that the dominated convergence theorem [140, 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

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

$$= \lim_{\lambda \to 0^+} \frac{\delta^{\star}_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{B.93a}$$

$$= \left(\delta^{\star}_{Q,\boldsymbol{z}} + \lim_{\lambda \to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda) \right) \lim_{\lambda \to 0^+} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{B.93b}$$

$$= 0, \tag{B.93c}$$

where (B.93b) 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 [142, Theorem 4.2]; and (B.93c) follows from Lemma 3.3.9. This completes the first part of Case 1.

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 (B.83) satisfy the following:

$$\mathcal{A}_0(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = \delta\}, \tag{B.94a}$$

$$\mathcal{A}_1(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta\}, \text{ and} \tag{B.94b}$$

$$\mathcal{A}_2(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta\}. \tag{B.94c}$$

Consider the sets

$$\mathcal{A}_{2,1}(\boldsymbol{\theta}) = \left\{\boldsymbol{\mu} \in \mathcal{A}_2(\boldsymbol{\theta}) : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^{\star}_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{B.95a}$$

$$\mathcal{A}_{2,2}(\boldsymbol{\theta}) = \left\{\boldsymbol{\mu} \in \mathcal{A}_2(\boldsymbol{\theta}) : \delta^{\star}_{Q,\boldsymbol{z}} \leq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta\right\}, \tag{B.95b}$$

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 (3.16), it holds that

$$Q(\mathcal{A}_{2,1}(\boldsymbol{\theta})) = 0. \tag{B.96}$$

Hence, substituting the equalities in (B.94) and (B.96) in (B.84) yields,

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

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

$$+ \lim_{\lambda\to 0^+} \int_{\mathcal{A}_2(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1} \tag{B.97a}$$

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

$$+ \lim_{\lambda\to 0^+} \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1} \tag{B.97b}$$

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

$$+ \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \lim_{\lambda\to 0^+} \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \,\mathrm{d}Q(\boldsymbol{\nu})\Bigg)^{-1}, \tag{B.97c}$$

where (B.97c) follows by verifying that the dominated convergence theorem [140, 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

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

$$= \lim_{\lambda\to 0^+} \frac{\delta + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{B.98a}$$

$$= \Big(\delta + \lim_{\lambda\to 0^+} \bar{K}_{Q,\boldsymbol{z}}(\lambda)\Big) \lim_{\lambda\to 0^+} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{B.98b}$$

$$= \big(\delta - \delta^\star_{Q,\boldsymbol{z}}\big) \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}}, \tag{B.98c}$$

where (B.98b) 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 [142, Theorem 4.2]; and (B.98c) follows from Lemma 3.3.9. From (B.98c), it follows that

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

$$= \big(\delta - \delta^\star_{Q,\boldsymbol{z}}\big) \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}). \tag{B.99}$$

Moreover, from the fact that

$$Q(\mathcal{A}_0(\boldsymbol{\theta})) \leq 1, \tag{B.100}$$

and the fact that

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

$$= \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\delta - \delta^\star_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \tag{B.101}$$

$$< \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\delta - \delta^\star_{Q,\boldsymbol{z}}}{\delta - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \tag{B.102}$$

$$= Q(\mathcal{A}_1(\boldsymbol{\theta})) \tag{B.103}$$

$$\leq 1, \tag{B.104}$$

the following holds from Lemma 3.3.5 under the assumptions of Case 1:

$$\infty = \left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right) \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \tag{B.105}$$

$$\begin{aligned} &= \left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right) \Bigg(\int_{\mathcal{A}_0(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \\ &\quad + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \\ &\quad + \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \Bigg) \end{aligned} \tag{B.106}$$

$$\begin{aligned} &= \Bigg(Q(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \frac{\delta - \delta^\star_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \\ &\quad + \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{\delta - \delta^\star_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) \Bigg). \end{aligned} \tag{B.107}$$

From (B.100), (B.104), and (B.107), it follows that

$$\left(\delta - \delta^\star_{Q,\boldsymbol{z}}\right) \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\nu}) = \infty. \tag{B.108}$$

Finally, from (B.97c), (B.99), and (B.108), for all $\boldsymbol{\theta} \in \left\{\boldsymbol{\nu} \in \operatorname{supp} Q : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star_{Q,\boldsymbol{z}}\right\}$,

$$\lim_{\lambda \to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{B.109}$$

This completes the second part of Case 1.

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

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0, \tag{B.110}$$

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

$$\lim_{\lambda \to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{B.111}$$

This completes the third part of Case 1.

Under the assumption that $Q\Big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\Big) > 0$, from (B.92), (B.109), and (B.111), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{Q\Big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\Big)} \mathbb{1}_{\{\boldsymbol{\theta}\in\mathcal{L}^\star_{Q,\boldsymbol{z}}\}}, \tag{B.112}$$

which completes the proof of (3.27). Alternatively, under the assumption that $Q\Big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\Big) = 0$, from (B.92), (B.109), and (B.111), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \begin{cases} \infty & \text{if } \boldsymbol{\theta} \in \mathcal{L}^\star_{Q,\boldsymbol{z}} \\ 0 & \text{otherwise} \end{cases}, \tag{B.113}$$

which completes the proof of (3.28).

B.9.2 Case 2

Under the assumptions of Case 2, namely (B.87), it holds that

$$Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) = 0. \tag{B.114}$$

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

$$\begin{aligned}
&\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) \\
&= \int_{\mathcal{L}^\star_{Q,\boldsymbol{z}}} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) \\
&\quad + \int_{\mathcal{L}^\star_{Q,\boldsymbol{z}}{}^{\mathsf{c}}} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.115a)} \\
&> \int_{\mathcal{L}^\star_{Q,\boldsymbol{z}}} \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^\star_{Q,\boldsymbol{z}}} \,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.115b)} \\
&= \frac{1}{\delta^\star_{Q,\boldsymbol{z}} - \delta^\star_{Q,\boldsymbol{z}}} Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) && \text{(B.115c)} \\
&= \infty, && \text{(B.115d)}
\end{aligned}$$

which contradicts (B.114).

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 (B.85).

Part 1

From (B.114) 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.

Part 2

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

$$\lim_{\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \lim_{\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} \tag{B.116a}$$

$$= \frac{\lambda^{\star}_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}, \tag{B.116b}$$

where (B.116b) 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 3.3.9) and [142, Theorem 4.2].

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

$$\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = 0. \tag{B.117}$$

This completes the third part of Case 2.

From (B.116b), for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it follows that

$$\lim_{\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda^{\star}_{Q,\boldsymbol{z}}}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}, \tag{B.118}$$

which completes the proof of (3.29). This completes the proof. ■

B.10 Proof of Lemma 3.3.14

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

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) = \infty; \tag{B.119}$$

and the second case follows under the assumption that

$$\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \mathrm{d}Q(\boldsymbol{\theta}) < \infty, \tag{B.120}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (3.16) and the function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6). From Lemma 3.3.5, it follows that in Case 1, the set $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7) 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 equality (3.30). Case 2 considers the limit $\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}$, which comprehends the equality (3.31).

B.10.1 Case 1

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

$$\mathcal{A}_0 = \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^\star_{Q,\boldsymbol{z}}\right\}, \tag{B.121a}$$
$$\mathcal{A}_1 = \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > \delta^\star_{Q,\boldsymbol{z}}\right\}, \text{ and} \tag{B.121b}$$
$$\mathcal{A}_2 = \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^\star_{Q,\boldsymbol{z}}\right\}. \tag{B.121c}$$

Note that $\mathcal{A}_0 = \mathcal{L}^\star_{Q,\boldsymbol{z}}$, with $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (3.18). For all $\lambda \in (0, \infty)$, it holds that

$$1 = \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{B.122a}$$
$$= \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) \tag{B.122b}$$
$$= \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_1} \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{B.122c}$$

where (B.122b) 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 4.3.1). The above implies that

$$\begin{aligned}
&\lim_{\lambda\to 0^+}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_1} \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right) \\
&\quad= \lim_{\lambda\to 0^+} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) \\
&\qquad+ \lim_{\lambda\to 0^+} \int_{\mathcal{A}_1} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.123a)} \\
&\quad= \lim_{\lambda\to 0^+} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) \\
&\qquad+ \int_{\mathcal{A}_1} \lim_{\lambda\to 0^+} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.123b)} \\
&\quad= \lim_{\lambda\to 0^+} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0), && \text{(B.123c)} \\
&\quad= 1, && \text{(B.123d)}
\end{aligned}$$

where, (B.123b) follows from Lemma 3.3.11 and the dominated convergence theorem [140, Theorem 1.6.9 page 50]; and (B.123c) follows from Lemma 3.3.13. Hence, it holds that

$$\lim_{\lambda\to 0^+} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) = 1, \tag{B.124}$$

which completes the proof of (3.30).

B.10.2 Case 2

Under the assumptions of Case 2, namely (B.120), it holds that

$$Q\left(\mathcal{L}^\star_{Q,\boldsymbol{z}}\right) = 0, \tag{B.125}$$

which can be shown using the arguments in (B.115). Hence, the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies

$$\begin{aligned}
&\lim_{\lambda\to\lambda^\star_{Q,\boldsymbol{z}}{}^+} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) \\
&\quad= \lim_{\lambda\to\lambda^\star_{Q,\boldsymbol{z}}{}^+} \int_{\mathcal{A}_0} \frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.126a)}\\
&\quad= \lim_{\lambda\to\lambda^\star_{Q,\boldsymbol{z}}{}^+} \int_{\mathcal{A}_0} \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.126b)}\\
&\quad= \lim_{\lambda\to\lambda^\star_{Q,\boldsymbol{z}}{}^+} \int_{\mathcal{A}_0} \frac{\lambda}{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)}\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.126c)}\\
&\quad= \lim_{\lambda\to\lambda^\star_{Q,\boldsymbol{z}}{}^+} \frac{\lambda}{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} Q(\mathcal{A}_0) && \text{(B.126d)}\\
&\quad= \lim_{\lambda\to\lambda^\star_{Q,\boldsymbol{z}}{}^+} \frac{\lambda}{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} Q\big(\mathcal{L}^\star_{Q,\boldsymbol{z}}\big) && \text{(B.126e)}\\
&\quad= \lim_{\lambda\to\lambda^\star_{Q,\boldsymbol{z}}{}^+} \frac{\lambda}{\delta^\star_{Q,\boldsymbol{z}} + \bar{K}_{Q,\boldsymbol{z}}(\lambda)} 0 && \text{(B.126f)}\\
&\quad= 0, && \text{(B.126g)}
\end{aligned}$$

which completes the proof of (3.31). This completes the proof. ■

B.11 Proof of Lemma 3.3.15

From Lemma 3.3.1 and Corollary 4.3.1, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned}
\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) &= \left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1} && \text{(B.127)}\\
&= \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}, && \text{(B.128)}
\end{aligned}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$ and $\bar{K}_{Q,\boldsymbol{z}}$ are in (2.6) and (3.11b), respectively. From (B.128), it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$0 = \lambda \frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \bar{K}_{Q,\boldsymbol{z}}(\lambda). \tag{B.129}$$

Integrating both sides of (B.129) with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ yields

$$\begin{aligned}
0 = \int \Bigg(&\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \lambda \left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)^{-1} \\
&+ \bar{K}_{Q,\boldsymbol{z}}(\lambda) \Bigg) \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && \text{(B.130a)}\\
= \int &\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})
\end{aligned}$$

$$
\begin{aligned}
&-\lambda \int \left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)^{-1} \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \\
&+ \int \bar{K}_{Q,\boldsymbol{z}}(\lambda)\, \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && \text{(B.130b)} \\
&= \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \lambda \int \mathrm{d}Q(\boldsymbol{\theta}) + \bar{K}_{Q,\boldsymbol{z}}(\lambda) && \text{(B.130c)} \\
&= \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \lambda + \bar{K}_{Q,\boldsymbol{z}}(\lambda). && \text{(B.130d)}
\end{aligned}
$$

From (B.130d), it holds that

$$
\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda), \tag{B.131}
$$

which completes the proof. ■

B.12 Proof of Lemma 3.3.16

The proof of continuity is immediate from Lemma 3.3.4 and Lemma 3.3.15. 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 (3.14) and shows that its derivative is strictly positive. The second part shows the expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ decreases with lambdas decreasing.

The first part is as follows. For all $\boldsymbol{\theta} \in \mathcal{M}$, the partial derivative of $\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}$, with respect to $\beta \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (3.16), is

$$
\frac{\partial}{\partial \beta}\left(\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\right) = -\frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}. \tag{B.132}
$$

From [143, Theorem 6.28, page 160], the following holds

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta}\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\, \mathrm{d}Q(\boldsymbol{\theta}) &= \int \frac{\partial}{\partial \beta}\frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.133)} \\
&= -\int \frac{1}{\left(\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}\, \mathrm{d}Q(\boldsymbol{\theta}). && \text{(B.134)}
\end{aligned}
$$

From Lemma 3.3.4, the derivative of the function $\bar{K}^{-1}_{Q,\boldsymbol{z}}$ in (3.14) satisfies:

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\beta}\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) \\
&= \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\, \mathrm{d}Q(\boldsymbol{\theta})\right)^{-1} && \text{(B.135a)} \\
&= -\left(\frac{1}{\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}(\boldsymbol{\theta})}}\, \mathrm{d}Q(\boldsymbol{\theta})}\right)^2 \frac{\mathrm{d}}{\mathrm{d}\beta}\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}(\boldsymbol{\theta})}}\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.135b)} \\
&= -\frac{\int -\frac{1}{\left(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2}\, \mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}(\boldsymbol{\theta})}}\, \mathrm{d}Q(\boldsymbol{\theta})\right)^2} && \text{(B.135c)}
\end{aligned}
$$

$$= \frac{\int \frac{1}{(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^2}\,\mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2}, \tag{B.135d}$$

where (B.135c) follows from (B.134).

Note that from Jensen's inequality [86, Theorem 2.6.2], it follows that

$$\left(\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2 \leq \int \frac{1}{(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^2}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{B.136}$$

with equality if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is noseparable (Definition 2.4.1). Then, from (B.135d) and (B.136), for all $\beta \in \left(-\delta^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta) = \frac{\int \frac{1}{(\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))^2}\,\mathrm{d}Q(\boldsymbol{\theta})}{\left(\int \frac{1}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\,\mathrm{d}Q(\boldsymbol{\theta})\right)^2} \tag{B.137}$$

$$\geq 1. \tag{B.138}$$

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 3.3.4). Then, from Lemma 3.3.15, it follows that

$$\begin{aligned}
&\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad = \lambda_2 - \lambda_1 + \bar{K}_{Q,\boldsymbol{z}}(\lambda_1) - \bar{K}_{Q,\boldsymbol{z}}(\lambda_2) && \text{(B.139a)} \\
&\qquad = \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_2) - \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_1) + \beta_1 - \beta_2, && \text{(B.139b)}
\end{aligned}$$

where (B.139b) follows from substituting (3.14) into (B.139a). Note that (B.138) implies that

$$\bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_2) - \bar{K}^{-1}_{Q,\boldsymbol{z}}(\beta_1) \geq \beta_2 - \beta_1, \tag{B.140}$$

with equality if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ is noseparable. Thus, from (B.139b) and (B.140) it follows that

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq 0, \tag{B.141}$$

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

B.13 Proof of Lemma 3.3.17

From Theorem 3.3.1 and Corollary 4.3.1, it holds that

$$\begin{aligned}
&\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad = \int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\right)\mathrm{d}Q(\boldsymbol{\theta}) && \text{(B.142)}
\end{aligned}$$

$$\leq \log\left(\int \frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\right) \tag{B.143}$$

$$= \log\left(\int \frac{\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\,\mathrm{d}Q(\boldsymbol{\theta})\right) \tag{B.144}$$

$$= \log\left(\frac{1}{\lambda}\int \bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\right) \tag{B.145}$$

$$= \log\left(\frac{1}{\lambda}\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda)+\mathsf{R}_{\boldsymbol{z}}(Q)\big)\right), \tag{B.146}$$

where (B.143) follows from Jensen's inequality [86, Theorem 2.6.2]; (B.144) follows from (3.4); and (B.146) follows from (2.12). From (B.146), it follows that

$$\mathsf{R}_{\boldsymbol{z}}(Q) \geq \lambda \exp\left(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right) - \bar{K}_{Q,\boldsymbol{z}}(\lambda) \tag{B.147}$$

$$= \lambda \exp\left(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right) + \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \lambda, \tag{B.148}$$

where (B.148) follows from Lemma 3.3.15. 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 (B.148), satisfies that

$$\mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geq \lambda\left(\exp\left(\mathsf{D}\left(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right) - 1\right), \tag{B.149}$$

which completes the proof. ■

B.14 Proof of Lemma 3.3.19

From Lemma 3.3.4 and Lemma 3.3.15, it holds that

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \lambda - \bar{K}_{Q,\boldsymbol{z}}(\lambda) \tag{B.150}$$

$$< \lambda + \delta^{\star}_{Q,\boldsymbol{z}}. \tag{B.151}$$

Note also that

$$\mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{B.152}$$

$$\geq \int \delta^{\star}_{Q,\boldsymbol{z}}\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{B.153}$$

$$= \delta^{\star}_{Q,\boldsymbol{z}}, \tag{B.154}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) and $\delta^{\star}_{Q,\boldsymbol{z}}$ in (3.16). The proof continues by determining the conditions for which (B.154) holds with equality. Assume the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is separable with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) (see Definition 2.4.1). Then, there exists a real value $\epsilon > 0$ and two nonnegligible sets $\mathcal{A}$ and $\mathcal{B}$ with respect to the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4), such that

$$\mathcal{A} = \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\right\}, \text{and} \tag{B.155}$$

$$\mathcal{B} = \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geq \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\right\}. \tag{B.156}$$

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

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{B.157}$$

$$= \int_{\mathcal{A}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{B}} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{B.158}$$

$$\geq \int_{\mathcal{A}} \delta^{\star}_{Q,\boldsymbol{z}}\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{B}} \big(\delta^{\star}_{Q,\boldsymbol{z}}+\epsilon\big)\,\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{B.159}$$

$$= \delta^{\star}_{Q,\boldsymbol{z}}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) + \big(\delta^{\star}_{Q,\boldsymbol{z}}+\epsilon\big)\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}), \tag{B.160}$$

$$= \delta^{\star}_{Q,\boldsymbol{z}}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) + \big(\delta^{\star}_{Q,\boldsymbol{z}}+\epsilon\big)\Big(1-\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\Big) \tag{B.161}$$

$$= \delta^{\star}_{Q,\boldsymbol{z}} + \epsilon\Big(1-\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\Big) \tag{B.162}$$

$$> \delta^{\star}_{Q,\boldsymbol{z}}, \tag{B.163}$$

where inequality (B.159) follows from (3.16) and (B.156); (B.161) follows from the fact that the sets $\mathcal{A}$ and $\mathcal{B}$ form a partition of the set $\mathcal{M}$; and (B.163) 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 (B.153).

Consider the case in which the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is not separable with respect to $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4). Then, for all $\boldsymbol{\theta} \in \operatorname{supp} \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, the empirical risk satisfies $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^{\star}_{Q,\boldsymbol{z}}$, which implies $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) = \delta^{\star}_{Q,\boldsymbol{z}}$. Hence, if the function $\mathsf{L}_{\boldsymbol{z}}$ is noseparable, then (B.153) holds with equality. Therefore, (B.153) holds with equality, if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ is noseparable, which completes the proof. ■

B.15 Proof of Theorem 3.3.3

Let δ be a real in $\Big(\delta^{\star}_{Q,\boldsymbol{z}}, \infty\Big)$, with $\delta^{\star}_{Q,\boldsymbol{z}}$ in (3.16). Let also $\gamma \in (0,\infty)$ satisfy the following equality:

$$\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) \leq \delta, \tag{B.164}$$

where the existence of such a γ is ensured by the continuity of $\mathsf{R}_{\boldsymbol{z}}\Big(\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ with respect to γ (Lemma 3.3.16); and Lemma 3.3.20 and Lemma 3.3.19. From (3.15), it holds that

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \supseteq \mathcal{L}^{\star}_{Q,\boldsymbol{z}}, \tag{B.165}$$

and thus,

$$\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) \geq \bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big), \tag{B.166}$$

with $\mathcal{L}^{\star}_{Q,\boldsymbol{z}}$ defined in (3.18). Let λ be a positive real such that $\lambda \leq \gamma$, and

$$\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) > 1 - \epsilon. \tag{B.167}$$

The existence of such a positive real λ follows from Lemma 3.3.14. From (B.165), it follows that

$$\bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) \geq \bar{P}^{(Q,\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big). \tag{B.168}$$

Hence, from (B.167) and (B.168), it holds that

$$1 - \epsilon < \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\big(\mathcal{L}^{\star}_{Q,\boldsymbol{z}}\big) \tag{B.169}$$

$$\leq \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)). \tag{B.170}$$

The equality in (B.170) implies that the probability measure $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is (δ,ϵ)-optimal (Definition 3.3.1), which completes the proof.

■

B.16 Proof of Lemma 3.3.22

The proof is divided into two parts. The first part is as follows, from Theorem 3.3.1, it follows that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) = \log\left(\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) \tag{B.171}$$

$$= \log(\lambda) - \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big) \tag{B.172}$$

$$= \log(\lambda) - \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}), \tag{B.173}$$

where the function $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ is defined in (3.51b). Thus,

$$\mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) = \int \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{B.174}$$

$$= \log(\lambda) - \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\, \mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{B.175}$$

$$= \log(\lambda) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{B.176}$$

where the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}$ is defined in (3.64). Hence, it follows from (B.176) that

$$\log(\lambda) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right), \tag{B.177}$$

which completes the proof of (3.66) and concludes the first part.

The second part is as follows. From (B.173), it follows that

$$\mathsf{D}\left(Q \| \bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = -\int \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{B.178}$$

$$= -\log(\lambda) + \int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{B.179}$$

$$= -\log(\lambda) + \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q). \tag{B.180}$$

Hence, it follows from (B.180) that

$$\log(\lambda) = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(Q) - \mathsf{D}\Big(Q\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big), \tag{B.181}$$

which completes the proof of (3.67). This completes the proof. ■

B.17 Proof of Lemma 3.3.23

The proof uses the mutual absolute continuity between $\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (3.4) and Q (Corollary 4.3.1). 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

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

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\frac{\mathrm{d}Q}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \tag{B.182}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log\left(\frac{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \tag{B.183}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log\left(\frac{\lambda}{\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right) \tag{B.184}$$

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

$$\quad + \log\big(\bar{K}_{Q,\boldsymbol{z}}(\lambda) + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\big) \tag{B.185}$$

$$= \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) - \log(\lambda) + \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta}), \tag{B.186}$$

where the functions $\mathsf{L}_{\boldsymbol{z}}$, $\bar{K}_{Q,\boldsymbol{z}}$ and $\mathsf{V}_{Q,\boldsymbol{z},\lambda}$ are defined in (2.6), (4.7) and in (3.51b), respectively. The equality (B.184) follows from (3.4). Hence, the relative entropy $\mathsf{D}\Big(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big)$ satisfies,

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

$$= \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{B.187}$$

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

$$\quad \left. + \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{B.188}$$

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

$$+\int \mathsf{V}_{Q,\boldsymbol{z},\lambda}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \tag{B.189}$$

$$= \mathsf{D}\left(P\|Q\right) - \log(\lambda) + \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) \tag{B.190}$$

$$= \mathsf{D}\left(P\|Q\right) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) + \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P), \tag{B.191}$$

where (B.188) follows from (B.186); (B.190) follows from (3.64); and (B.191) follows from Lemma 3.3.22. Thus, from (B.191), it follows that

$$\bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}(P) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\lambda}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \mathsf{D}\left(P\|\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P\|Q\right) + \mathsf{D}\left(\bar{P}^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right), \tag{B.192}$$

which completes the proof. ■

B.18 Proof of Lemma 3.3.25

From Lemma 3.3.22, for all $\alpha \in (0,\infty)$, it holds that

$$\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) = -\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \log(\alpha), \tag{B.193}$$

where the functional $\bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}$ is defined in (3.64).

Similarly, from [84, Lemma 20], for all $\lambda \in (0,\infty)$, it holds that

$$\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) = -\left(\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right), \tag{B.194}$$

with the functional $\mathsf{R}_{\boldsymbol{z}}$ defined in (2.12). From [84, Theorem 3], the function $\mathsf{S}_{Q,\lambda}$ in [84, Definition 7] satisfies that

$$\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = \mathsf{R}_{\boldsymbol{z}}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \tag{B.195}$$

$$= \lambda\left(\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right)\right) \tag{B.196}$$

$$= \lambda\left(\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) + \bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \log(\alpha)\right) \tag{B.197}$$

$$= \lambda\left(\mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \log(\alpha)\right), \tag{B.198}$$

where (B.197) follows from (B.193); and (B.198) follows from (B.194). Rearranging (B.198) yields

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

$$= \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \log(\alpha) - K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right). \tag{B.199}$$

Similarly, from Lemma 3.3.23 the function $\bar{\mathsf{S}}_{Q,\alpha}$ in (3.69) satisfies that

$$\begin{aligned}
&\bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\quad = \bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) && \text{(B.200)} \\
&\quad = \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) \\
&\qquad + \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) && \text{(B.201)} \\
&\quad = \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) \\
&\qquad - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \log(\alpha) && \text{(B.202)} \\
&\quad = \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad + \frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \\
&\qquad - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \log(\alpha), && \text{(B.203)}
\end{aligned}$$

where (B.202) follows from (B.193); and (B.203) follows from (B.194). Rearranging (B.203) yields

$$\begin{aligned}
&\frac{1}{\lambda}\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{R}}_{Q,\boldsymbol{z},\alpha}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad = -\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad\quad - \left(\log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right). && \text{(B.204)}
\end{aligned}$$

The proof proceeds by subtracting (B.204) from (B.199), resulting in

$$\begin{aligned}
&\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad = \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad\quad + 2\left(\log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right), && \text{(B.205)}
\end{aligned}$$

where the functions $\mathsf{S}_{Q,\lambda}$ and $\bar{\mathsf{S}}_{Q,\alpha}$ are respectively defined in [84, Definition 7] and (3.69). From [115, Theorem 1] and Lemma 3.3.23, it follows that

$$\begin{aligned}
&\frac{1}{\lambda}\mathsf{S}_{Q,\lambda}\left(\boldsymbol{z}, \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \bar{\mathsf{S}}_{Q,\alpha}\left(\boldsymbol{z}, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad = \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| \bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \\
&\qquad\quad + 2\Big(\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right)\Big). && \text{(B.206)}
\end{aligned}$$

Substituting (B.206) into (B.205) yields

$$\begin{aligned}
&\mathsf{D}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) - \mathsf{D}\left(\bar{P}^{(Q,\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) \\
&\qquad = \log(\alpha) + K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), && \text{(B.207)}
\end{aligned}$$

which completes the proof. ■

Appendix C

ERM-fDR

C.1 Proof of Theorem 4.3.1

Proof: The optimization problems in (4.1) and (4.2) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the reference measure Q, denoted by $\frac{\mathrm{d}P}{\mathrm{d}Q} : \mathcal{M} \to [0, \infty)$, which yields:

$$\min_{P \in \triangle_Q(\mathcal{M})} \quad \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{C.1a}$$

$$\text{s.t.} \quad \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \leq \eta \tag{C.1b}$$

$$\int \frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{C.1c}$$

The remainder of the proof focuses on the problem in which the optimization is over the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ instead of the measures P. This is due to the fact that for all $P \in \triangle_Q(\mathcal{M})$, the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is unique up to sets of measure zero with respect to Q. The first part is as follows. Let $\mathscr{M}$ be the set of measurable functions $\mathcal{M} \to \mathbb{R}$ 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 nonnegative functions that are absolutely integrable with respect to Q. That is, for all $\hat{g} \in \mathscr{S}$, it holds that

$$\int |\hat{g}(\boldsymbol{\theta})| \, \mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{C.2}$$

Note that the set $\mathscr{M}$ forms a real vector space and the set $\mathscr{S}$ is a convex subset of $\mathscr{M}$. Note also that the constraints (C.1b) and (C.1c) are satisfied by the probability measure Q, which also satisfies $Q \in \triangle_Q(\mathcal{M})$. Hence, the constraints do not induce an empty feasible set. Finally, note that without loss of generality, the minimization in (C.1) can be written as a minimization problem of the form:

$$\min_{g \in \mathscr{S}} \quad \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{C.3a}$$

$$\text{s.t.} \quad \frac{1}{\eta} \int f(g(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) \leq 1 \tag{C.3b}$$

$$\int g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{C.3c}$$

where the expressions $\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta})$ and $\int g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta})$ are linear with g; the expression $\frac{1}{\eta} \int f(g(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta})$ is convex with g. The proof continues by assuming that the problem in (C.3) possesses a solution, which is denoted by $g^\star \in \mathscr{S}$. Let

$\mu_0 \in [0, \infty)$ be

$$\mu_0 \triangleq \min_{g \in \mathscr{S}} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{C.4a}$$

$$\text{s.t. } \frac{1}{\eta} \int f(g(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) \leq 1 \tag{C.4b}$$

$$\int g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) = 1 \tag{C.4c}$$

$$= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.4d}$$

From [24, Theorem 1, Section 8.3], it holds that there exists two tuples (a_1, b_1) and (a_2, b_2) in $\mathbb{R}^2$ such that

$$\mu_0 = \min_{g \in \mathscr{S}} \left\{ \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) + \frac{a_1}{\eta} \int f(g(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_1 + a_2 \int g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_2 \right\}, \tag{C.5a}$$

and moreover,

$$0 = \frac{a_1}{\eta} \int f(g^\star(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_1, \text{ and} \tag{C.5b}$$

$$0 = a_2 \int g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_2. \tag{C.5c}$$

Hence, the proof continues by solving the ancillary optimization problem in (C.5), which allows the reformulation of the optimization problem in an unconstrained dual problem. This reformulation is possible as the tuples (a_1, b_1) and (a_2, b_2) are such that equalities (C.5b) and (C.5c) are satisfied, by definition.

Let the function $L : \mathscr{S} \to \mathbb{R}$ be such that

$$L(g) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) + \frac{a_1}{\eta} \int f(g(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_1 + a_2 \int g(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_2. \tag{C.6}$$

Let $\hat{g} : \mathcal{M} \to \mathbb{R}$ be a function in $\mathscr{S}$. The Gateaux differential of the functional L in (C.6) at $(g, \beta) \in \mathscr{S} \times \mathbb{R}$ in the direction of $\hat{g}$ is

$$\partial L(g; \hat{g}) \triangleq \frac{\mathrm{d}}{\mathrm{d}\alpha} L(g + \alpha \hat{g}, \beta) \bigg|_{\alpha = 0}. \tag{C.7}$$

Let the function $r : \mathbb{R} \to \mathbb{R}$ be defined for some fixed functions g and $\hat{g}$ and some fixed a_1, b_1, a_2 and b_2 such that for all $\alpha \in (-\epsilon, \epsilon)$, with ϵ arbitrarily small,

$$r(\alpha) = L(g + \alpha \hat{g}). \tag{C.8}$$

The proof follows by showing that the function r in (C.8) is differentiable at zero, to prove the existence of the Gateaux differential in (C.7) for those functions g and $\hat{g}$ and reals a_1, b_1, a_2, and b_2. To do so, note that

$$\begin{aligned} r(\alpha) = & \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) (g(\boldsymbol{\theta}) + \alpha \hat{g}(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) \\ & + \frac{a_1}{\eta} \int f(g(\boldsymbol{\theta}) + \alpha \hat{g}(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_1 \\ & + a_2 \int (g(\boldsymbol{\theta}) + \alpha \hat{g}(\boldsymbol{\theta})) \, \mathrm{d}Q(\boldsymbol{\theta}) + b_2 \end{aligned} \tag{C.9}$$

which can be rewritten as follows,

$$\begin{aligned} r(\alpha)=&\alpha \int \hat{g}(\boldsymbol{\theta})(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \\ &+\frac{a_1}{\eta}\int f(g(\boldsymbol{\theta})+\alpha\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \\ &+\int g(\boldsymbol{\theta})(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\;\mathrm{d}Q(\boldsymbol{\theta})+b_1+b_2. \end{aligned} \tag{C.10}$$

Note that the first term in (C.10) is linear with α; the second term can be written using the function $\hat{r}:\mathbb{R}\to\mathbb{R}$ such that for all $\alpha\in(-\epsilon,\epsilon)$, with ϵ arbitrarily small, it holds that

$$\hat{r}(\alpha)=\lambda\int f(g(\boldsymbol{\theta})+\alpha\hat{g}(\boldsymbol{\theta}))\;\mathrm{d}Q(\boldsymbol{\theta}); \tag{C.11}$$

and the remaining terms are independent of α. Hence, based on the fact that the function $\hat{r}$ in (C.11) is differentiable at zero (see Lemma A.2.1 in Appendix A.3), so is the function r in (C.10), which implies that the Gâteaux differential of $\partial L(g,\hat{g})$ in (C.7) exists.

The proof proceeds by calculating the Gateaux differential $\partial L(g,\hat{g})$ in (C.7), which requires calculating the derivative of the real function r in (C.10). That is,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\Bigg(\alpha\int\hat{g}(\boldsymbol{\theta})(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})+\frac{a_1}{\eta}\int f(g(\boldsymbol{\theta})+\alpha\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \\ &\quad+\int g(\boldsymbol{\theta})(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\;\mathrm{d}Q(\boldsymbol{\theta})+b_1+b_2\Bigg) && \text{(C.12)} \\ &= \int\hat{g}(\boldsymbol{\theta})(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})+\frac{a_1}{\eta}\int\frac{\mathrm{d}}{\mathrm{d}\alpha}f(g(\boldsymbol{\theta})+\alpha\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.13)} \\ &= \int\hat{g}(\boldsymbol{\theta})(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})+\frac{a_1}{\eta}\int\hat{g}(\boldsymbol{\theta})\dot{f}(g(\boldsymbol{\theta})+\alpha\hat{g}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.14)} \end{aligned}$$

where (C.13) follows from Theorem A.3.1. From equations (C.7) and (C.14), it follows that

$$\begin{aligned} \partial L(g;\hat{g}) &= \int\hat{g}(\boldsymbol{\theta})(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})+\frac{a_1}{\eta}\int\hat{g}(\boldsymbol{\theta})\dot{f}(g(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.15a)} \\ &= \int\hat{g}(\boldsymbol{\theta})\left(a_2+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\frac{a_1}{\eta}\dot{f}(g(\boldsymbol{\theta}))\right)\mathrm{d}Q(\boldsymbol{\theta}). && \text{(C.15b)} \end{aligned}$$

From [24, Theorem 1, Chapter 7], a necessary condition to use for the functional L in (C.6) to have a minimum at $g^\star$ is that for all functions $\hat{g}\in\mathscr{S}$,

$$\partial L(g^\star;\hat{g})=0. \tag{C.16}$$

From (C.15b) and (C.16), it follows that $g^\star$ must satisfy for all $\boldsymbol{\theta}\in\operatorname{supp}Q$, that

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\frac{a_1}{\eta}\dot{f}(g^\star(\boldsymbol{\theta}))+a_2=0. \tag{C.17}$$

Assuming that

$$a_1\neq 0, \tag{C.18}$$

From (C.17), it follows that

$$g^\star(\boldsymbol{\theta}) = \dot{f}^{-1}\bigg(-\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2)\bigg), \tag{C.19}$$

where the values a_1 and a_2 satisfy (C.5b) and (C.5c) and (C.18).

The remainder of the proof focuses on determining the values of a_1, a_2, b_1, and b_2, which must also be such that $g^\star$ in (C.19) satisfies the constraints (C.3b) and (C.3c) under the assumption that $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is separable. For instance, from constraints (C.3b) and (C.5c), it follows that

$$a_2 = -b_2. \tag{C.20}$$

From (C.20), the constraint in (C.5c) implies that the choice of a_2 satisfies

$$1 = \int g^\star(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}). \tag{C.21}$$

Similarly, the function $g^\star$ in (C.19) is the Radon-Nikodym derivative with respect to Q of the solution $P^\star \in \triangle(\mathcal{M})$ to the problem in (C.1). Hence, (C.5b) can be written as follows

$$\frac{a_1}{\eta}\mathsf{D}_f(P^\star\|Q) + b_1 = 0, \tag{C.22}$$

which implies

$$b_1 = -\frac{a_1}{\eta}\mathsf{D}_f(P^\star\|Q). \tag{C.23}$$

Note that if $a_1 < 0$, given two models $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\mathcal{M}$, such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)$, it holds that

$$-\frac{a_1}{\eta}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) + a_2) < -\frac{a_1}{\eta}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) + a_2). \tag{C.24}$$

From Lemma A.3.1 in Appendix A.3, the function $\dot{f}^{-1}$ in (C.19) is strictly increasing, and thus,

$$\dot{f}^{-1}\bigg(-\frac{a_1}{\eta}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) + a_2)\bigg) < \dot{f}^{-1}\bigg(-\frac{a_1}{\eta}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) + a_2)\bigg). \tag{C.25}$$

Hence, from (C.19) and (C.25), it follows that

$$g^\star(\boldsymbol{\theta}_1) < g^\star(\boldsymbol{\theta}_2). \tag{C.26}$$

The proof continues by showing that (C.26) implies that the expectation of $g^\star \in \mathscr{S}$ with respect to Q induces an expected empirical risk larger than, the function $g^\star$ is nonnegative. Let the real value $k \in \mathbb{R}$ be

$$k = -\frac{a_1}{\eta}\dot{f}(1) - a_2, \tag{C.27}$$

and consider the partition of the set $\mathcal{M}$ formed by the sets $\mathcal{M}_0$ and $\mathcal{M}_1$, which satisfy the following:

$$\mathcal{M}_0 = \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < k\}, \tag{C.28}$$

$$\mathcal{M}_1 = \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geq k\}. \tag{C.29}$$

Note that from (C.27) for all $\boldsymbol{\theta} \in \mathcal{M}$, such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = k$, it follows

$$g^\star(\boldsymbol{\theta}) = 1. \tag{C.30}$$

From (C.26) and (C.30), it implies that for all $\boldsymbol{\theta} \in \mathcal{M}_0$,

$$g^\star(\boldsymbol{\theta}) < 1, \tag{C.31}$$

and for all $\boldsymbol{\theta} \in \mathcal{M}_1$,

$$g^\star(\boldsymbol{\theta}) \geq 1, \tag{C.32}$$

Let $P^\star$ denote the probability measures defined by the pairs (a_1, a_2), such that

$$P^\star(\mathcal{M}) = \int_{\mathcal{M}_0} g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M}_1} g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{C.33}$$

where $g^\star$ is defined in (C.19). Under the assumption $a_1 < 0$, the measures $P^\star$ over the set $\mathcal{M}_0$ in (C.28) satisfies

$$\begin{aligned} P^\star(\mathcal{M}_0) &= \int_{\mathcal{M}_0} g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.34a)} \\ &< \int_{\mathcal{M}_0} \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.34b)} \\ &= Q(\mathcal{M}_0), && \text{(C.34c)} \end{aligned}$$

where (C.34b) follows from (C.31). Hence, from (C.31) and (C.34) it follows that

$$\int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) < \int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.35}$$

Similarly, under the assumption $a_1 < 0$, the measures $P^\star$ over the set $\mathcal{M}_1$ in (C.29) satisfies

$$\begin{aligned} P^\star(\mathcal{M}_1) &= \int_{\mathcal{M}_1} g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.36a)} \\ &\geq \int_{\mathcal{M}_1} \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.36b)} \\ &= Q(\mathcal{M}_1). && \text{(C.36c)} \end{aligned}$$

where (C.36b) follows from (C.32). Hence, from (C.32) and (C.36) it follows that

$$\int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g^\star(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \geq \int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.37}$$

Note that from (C.28) and (C.35), it holds that

$$\begin{aligned} 0 &< \int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) - \int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P^\star(\boldsymbol{\theta}) && \text{(C.38a)} \\ &< k(Q(\mathcal{M}_0) - P^\star(\mathcal{M}_0)). && \text{(C.38b)} \end{aligned}$$

Similarly, from (C.29) and (C.37), it holds that

$$\int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^\star(\boldsymbol{\theta}) - \int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) > k(P^\star(\mathcal{M}_1) - Q(\mathcal{M}_1)) \tag{C.39a}$$

$$= k((1 - P^\star(\mathcal{M}_0)) - (1 - Q(\mathcal{M}_0))) \tag{C.39b}$$

$$= k(Q(\mathcal{M}_0) - P^\star(\mathcal{M}_0)) \tag{C.39c}$$

Plugging, (C.38) into (C.39) yields

$$\int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - \int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^\star(\boldsymbol{\theta}) < \int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^\star(\boldsymbol{\theta}) - \int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{C.40}$$

which can be rearranged into,

$$\int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) + \int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) < \int_{\mathcal{M}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^\star(\boldsymbol{\theta}) + \int_{\mathcal{M}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^\star(\boldsymbol{\theta}). \tag{C.41}$$

Thus, the expected empirical risk of the measures $P^\star$ under the assumption $a_1 < 0$ implies that

$$\mathsf{R}_{\boldsymbol{z}}(Q) < \mathsf{R}_{\boldsymbol{z}}(P^\star). \tag{C.42}$$

Note that from Definition 1.5.1, for all $P \in \triangle_Q(\mathcal{M})$, it follows that

$$\mathsf{D}_f(Q\|Q) \leq \mathsf{D}_f(P\|Q), \tag{C.43}$$

with equality, if and only if $P = Q$. Hence, (C.42) and (C.43) imply

$$\mathsf{R}_{\boldsymbol{z}}(Q) + \lambda \mathsf{D}_f(Q\|Q) < \mathsf{R}_{\boldsymbol{z}}(P^\star) + \lambda \mathsf{D}_f(P^\star\|Q), \tag{C.44}$$

which is a contradiction. Thus, the focus in the remainder of the proof is on the case in which

$$a_1 > 0, \tag{C.45}$$

which implies that models $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in $\operatorname{supp} Q$ that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2)$, satisfy

$$g^\star(\boldsymbol{\theta}_1) > g^\star(\boldsymbol{\theta}_2). \tag{C.46}$$

Given the pairs (a_1, a_2) and $(\hat{a}_1, \hat{a}_2)$ in $\mathbb{R}^2$ such that each pair satisfies the constraints in (C.5b) and (C.5c), from (C.19) there exist a solution for each pair given by

$$g^\star(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2)\right), \tag{C.47}$$

and

$$\hat{g}^\star(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\right), \tag{C.48}$$

where the functions $g^\star$ and $\hat{g}^\star$ are the Radon-Nikodym derivative of the solutions $P^\star$ and $\hat{P}^\star$ with respect to Q for each pair (a_1, a_2) and $(\hat{a}_1, \hat{a}_2)$, respectively. Under the assumption that $a_1 < \hat{a}_1$, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$-\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2) < -\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2), \tag{C.49}$$

which from Lemma A.3.1 in Appendix A.3 implies that that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\dot{f}^{-1}\left(-\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2)\right) < \dot{f}^{-1}\left(-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2)\right). \tag{C.50}$$

From (C.50), it holds that

$$1 = \int \dot{f}^{-1}\left(-\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2)\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{C.51}$$

$$< \int \dot{f}^{-1}\left(-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2)\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.52}$$

Then, for the pair $(\hat{a}_1, \hat{a}_2)$ to satisfy,

$$\int \dot{f}^{-1}\left(-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\right) \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{C.53}$$

under the assumption that $a_1 < \hat{a}_1$, the value $\hat{a}_2$ must satisfy $a_2 < \hat{a}_2$. Using the fact that $0 < a_1 < \hat{a}_1$ and $a_2 < \hat{a}_2$, let the real value $c \in \mathbb{R}$ be

$$c = \frac{a_2\hat{a}_1 - \hat{a}_2 a_1}{a_1 - \hat{a}_1}, \tag{C.54}$$

and consider the partition of the set $\mathcal{M}$ formed by the sets $\mathcal{A}_0$, $\mathcal{A}_1$ and $\mathcal{A}_2$, which satisfy the following:

$$\mathcal{A}_0 \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = c\}, \tag{C.55}$$

$$\mathcal{A}_1 \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < c\}, \tag{C.56}$$

$$\mathcal{A}_2 \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > c\}. \tag{C.57}$$

Note that for all $\boldsymbol{\theta} \in \mathcal{A}_0$, the pair $(\hat{a}_1, \hat{a}_2)$ satisfies

$$-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2) = -\frac{\eta}{\hat{a}_1}(c + \hat{a}_2) \tag{C.58}$$

$$= -\frac{\eta}{\hat{a}_1}\left(\frac{a_2\hat{a}_1 - \hat{a}_2 a_1}{a_1 - \hat{a}_1} + \frac{\hat{a}_2(a_1 - \hat{a}_1)}{a_1 - \hat{a}_1}\right) \tag{C.59}$$

$$= -\frac{\eta}{\hat{a}_1}\left(\frac{a_2\hat{a}_1 - \hat{a}_2\hat{a}_1}{a_1 - \hat{a}_1}\right) \tag{C.60}$$

$$= -\eta\left(\frac{a_2 - \hat{a}_2}{a_1 - \hat{a}_1}\right). \tag{C.61}$$

Similarly, for all $\boldsymbol{\theta} \in \mathcal{A}_0$, the pair (a_1, a_2) satisfies

$$-\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2) = -\frac{\eta}{a_1}(c + a_2) \tag{C.62}$$

$$= -\frac{\eta}{a_1}\left(\frac{a_2\hat{a}_1 - \hat{a}_2 a_1}{a_1 - \hat{a}_1} + \frac{a_2(a_1 - \hat{a}_1)}{a_1 - \hat{a}_1}\right) \tag{C.63}$$

$$= -\frac{\eta}{a_1}\left(\frac{a_2 a_1 - \hat{a}_2 a_1}{a_1 - \hat{a}_1}\right) \tag{C.64}$$

$$= -\eta\left(\frac{a_2 - \hat{a}_2}{a_1 - \hat{a}_1}\right). \tag{C.65}$$

Hence, from (C.61) and (C.65) for all $\boldsymbol{\theta} \in \mathcal{A}_0$, it holds that

$$-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2) = -\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2). \tag{C.66}$$

Then, from (C.19) and (C.66), it follows that for all $\boldsymbol{\theta} \in \mathcal{A}_0$,

$$\begin{aligned} g^\star(\boldsymbol{\theta}) &= \dot{f}^{-1}\left(-\frac{\eta}{a_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2)\right) && \text{(C.67a)} \\ &= \dot{f}^{-1}\left(-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\right) && \text{(C.67b)} \\ &= \hat{g}^\star(\boldsymbol{\theta}). && \text{(C.67c)} \end{aligned}$$

Furthermore, from the fact that $\dot{f}^{-1}$ in (C.19) is strictly increasing (see Lemma A.3.1 in Appendix A.3.1), for all $\boldsymbol{\theta} \in \mathcal{A}_1$, it holds that

$$g^\star(\boldsymbol{\theta}) > \hat{g}^\star(\boldsymbol{\theta}), \tag{C.68}$$

and for all $\boldsymbol{\theta} \in \mathcal{A}_2$, it holds that

$$g^\star(\boldsymbol{\theta}) < \hat{g}^\star(\boldsymbol{\theta}). \tag{C.69}$$

Let $P^\star$ and $\hat{P}^\star$ denote the probability measures defined by the pairs (a_1, a_2) and $(\hat{a}_1, \hat{a}_2)$, respectively. From (C.47) and (C.48), it follows that

$$P^\star(\mathcal{A}_1) = \int_{\mathcal{A}_1} g^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}), \tag{C.70}$$

and

$$\hat{P}^\star(\mathcal{A}_1) = \int_{\mathcal{A}_1} \hat{g}^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.71}$$

From (C.70) and (C.71), the measures $P^\star$ and $\hat{P}^\star$ over the set $\mathcal{A}_1$ in (C.56) satisfy

$$\begin{aligned} P^\star(\mathcal{A}_1) &= \int_{\mathcal{A}_1} g^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.72a)} \\ &> \int_{\mathcal{A}_1} \hat{g}^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.72b)} \\ &= \hat{P}^\star(\mathcal{A}_1), && \text{(C.72c)} \end{aligned}$$

where (C.72b) follows from (C.68). Hence, from (C.46) and (C.72) it follows that

$$\int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) > \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \hat{g}^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.73}$$

Similarly, from (C.70) and (C.71), the measures $P^\star$ and $\hat{P}^\star$ over the set $\mathcal{A}_2$ in (C.57) satisfy

$$\begin{aligned} P^\star(\mathcal{A}_2) &= \int_{\mathcal{A}_2} g^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.74a)} \\ &< \int_{\mathcal{A}_2} \hat{g}^\star(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.74b)} \\ &= \hat{P}^\star(\mathcal{A}_2). && \text{(C.74c)} \end{aligned}$$

where (C.74b) follows from (C.69). Hence, from (C.46) and (C.74) it follows that

$$\int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) g^\star(\boldsymbol{\theta}) \,\mathrm{d}Q(\boldsymbol{\theta}) < \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \hat{g}^\star(\boldsymbol{\theta}) \,\mathrm{d}Q(\boldsymbol{\theta}). \tag{C.75}$$

Note that from (C.56) and (C.73), it holds that

$$\begin{aligned} 0 &< \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta}) - \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) && \text{(C.76a)} \\ &< c(P^\star(\mathcal{A}_1) - \hat{P}^\star(\mathcal{A}_1)). && \text{(C.76b)} \end{aligned}$$

Similarly, from (C.57) and (C.75), it holds that

$$\begin{aligned} \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) - \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta}) &> c(\hat{P}^\star(\mathcal{A}_2) - P^\star(\mathcal{A}_2)) && \text{(C.77a)} \\ &= c((1 - \hat{P}^\star(\mathcal{A}_1)) - (1 - P^\star(\mathcal{A}_1))) && \text{(C.77b)} \\ &= c(P^\star(\mathcal{A}_1) - \hat{P}^\star(\mathcal{A}_1)) && \text{(C.77c)} \end{aligned}$$

Plugging, (C.76) into (C.77) yields

$$\int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta}) - \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) < \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) - \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta}), \tag{C.78}$$

which can be rearranged into,

$$\int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta}) < \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) + \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}). \tag{C.79}$$

Thus, the expected empirical risk of the measures $P^\star$ and $\hat{P}^\star$ satisfy

$$\mathsf{R}_{\boldsymbol{z}}(P^\star) < \mathsf{R}_{\boldsymbol{z}}(\hat{P}^\star). \tag{C.80}$$

Observe that from (C.80) and the assumption of $a_1 < \hat{a}_1$, it follows that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}a_1} \mathsf{R}_{\boldsymbol{z}}(P^\star) &= \frac{\mathrm{d}}{\mathrm{d}a_1} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta}) && \text{(C.81a)} \\ &= \lim_{\hat{a}_1 \to a_1} \frac{\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \,\mathrm{d}P^\star(\boldsymbol{\theta})}{\hat{a}_1 - a_1} && \text{(C.81b)} \\ &> 0, && \text{(C.81c)} \end{aligned}$$

where (C.81c) follows from (C.80). The proof continues by showing the induced f-divergence by the function f is increasing with respect to a_1, which is equivalent to showing that the derivative of D_f with respect to a_1 is always positive. In order to do so, note that from the Assumption (a) in Theorem 4.3.1, that is the strict convexity it holds from Jensen inequality

$$\begin{aligned} &\mathsf{D}_f(P^\star \| Q) - \mathsf{D}_f(\hat{P}^\star \| Q) \\ &= \int f(g^\star(\boldsymbol{\theta})) \,\mathrm{d}Q(\boldsymbol{\theta}) - \int f(\hat{g}^\star(\boldsymbol{\theta})) \,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.82a)} \\ &= \int f(g^\star(\boldsymbol{\theta})) - f(\hat{g}^\star(\boldsymbol{\theta})) \,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.82b)} \\ &> \int \dot{f}(\hat{g}^\star(\boldsymbol{\theta}))(g^\star(\boldsymbol{\theta}) - \hat{g}^\star(\boldsymbol{\theta})) \,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.82c)} \\ &= \int \dot{f}\left(\dot{f}^{-1}\left(-\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\right)\right)(g^\star(\boldsymbol{\theta}) - \hat{g}^\star(\boldsymbol{\theta})) \,\mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.82d)} \end{aligned}$$

$$= \int -\frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)(g^\star(\boldsymbol{\theta}) - \hat{g}^\star(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.82e}$$

$$= \int \frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\hat{g}^\star(\boldsymbol{\theta}) - \frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)g^\star(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.82f}$$

$$= \int \frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\hat{g}^\star(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - \int \frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)g^\star(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.82g}$$

$$= \int \frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) - \int \frac{\eta}{\hat{a}_1}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \hat{a}_2)\,\mathrm{d}P^\star(\boldsymbol{\theta}) \tag{C.82h}$$

$$= \frac{\eta}{\hat{a}_1}\left(\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) + \hat{a}_2 - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^\star(\boldsymbol{\theta}) - \hat{a}_2\right) \tag{C.82i}$$

$$= \frac{\eta}{\hat{a}_1}\left(\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}\hat{P}^\star(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^\star(\boldsymbol{\theta})\right) \tag{C.82j}$$

$$= \frac{\eta}{\hat{a}_1}(\mathsf{R}_{\boldsymbol{z}}(\hat{P}^\star) - \mathsf{R}_{\boldsymbol{z}}(P^\star)) \tag{C.82k}$$

$$> 0, \tag{C.82l}$$

where (C.82c) follows from first-order contidon (see [25, Secction 3.1.3]; (C.82d) follows from (C.48); (C.82h) follows from (C.70) and (C.71); and (C.82l) follows from (C.80) and the facts that $0 < \eta, \hat{a}_1$. Observe that from (C.82) and the assumption of $a_1 < \hat{a}_1$, it follows that

$$\frac{\mathrm{d}}{\mathrm{d}a_1}\mathsf{D}_f(P^\star\|Q) = \frac{\mathrm{d}}{\mathrm{d}a_1}\int f(g^\star(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.83}$$

$$= \lim_{\hat{a}_1 \to a_1} \frac{\mathsf{D}_f(P^\star\|Q) - \mathsf{D}_f(\hat{P}^\star\|Q)}{a_1 - \hat{a}_1} \tag{C.84}$$

$$< 0, \tag{C.85}$$

where (C.85) follows from (C.82l) and that $a_1 - \hat{a}_1 < 0$. Hence, from (C.81) the terms $\int g^\star(\boldsymbol{\theta})\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})$ in (C.1b) is strictly increasing with a_1 and from (C.85) the term $\int f(g^\star(\boldsymbol{\theta}))\,\mathrm{d}Q(\boldsymbol{\theta})$ in (C.1c) is strictly decreasing with a_1. This implies that $a_1 > 0$ shall be chosen such that

$$\mathsf{D}_f(P^\star\|Q) = \eta, \tag{C.86}$$

and justify the uniqueness of the solution.

For the case in which the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is nonseparable (see Definition 2.4.1), the objective function in (4.2) is a constant. Hence, the optimization problems in (4.1) and (4.2) do not share the same solutions when for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = c$, for some $c > 0$ and $\mathsf{L}_{\boldsymbol{z}}$ in (2.6). More specifically, the set of solutions to (4.1) is the singleton $\{Q\}$, while the set of solutions to the problem in (4.2) is $\{P \in \triangle_Q(\mathcal{M}) : \mathsf{D}_f(P\|Q) \leq \eta\}$. Thus, the problem is ill-posted and justifies the need for Assumption (c) in Theorem 4.3.1.

In a nutshell, choosing a real value $\lambda = \frac{a_1}{\eta}$, the real a_2 to satisfy

$$a_2 \in \left\{t \in \mathbb{R} : \forall \boldsymbol{\theta} \in \operatorname{supp} Q, 0 < \dot{f}^{-1}\left(-\frac{t + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right\}, \tag{C.87}$$

and denoting the solution $P^\star$ as $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, it holds that $g^\star$ in (C.19) can be written as $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$, and thus, for all $(\boldsymbol{\theta}) \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + a_2}{\lambda}\right), \tag{C.88}$$

where λ is such that $\mathsf{D}_f\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| Q\right) = \eta$. This completes the proof. ■

C.2 Proof of Theorem 4.3.2

Proof: The proof is divided into three parts. The first part uses the properties of the f-divergence regularization to prove the properties of an ancillary function. The second part proves that the normalization function $N_{Q,\boldsymbol{z}} : \mathcal{A}_{Q,\boldsymbol{z}} \to \mathcal{B}_{Q,\boldsymbol{z}}$, with $N_{Q,\boldsymbol{z}}$ implicitly defined in (4.7), has a unique solution and is continuous by using the implicit function theorem on the ancillary function properties. The third part uses the implicit function theorem result to derive an implicit expression for the normalization function $N_{Q,\boldsymbol{z}}$.

Under the assumptions (a), (b) and (c) from Theorem 4.3.1, the sets $\mathcal{A}_{Q,\boldsymbol{z}}$ and $\mathcal{B}_{Q,\boldsymbol{z}}$ in (4.7) are non-empty such that

$$\bar{a} = \sup \mathcal{A}_{Q,\boldsymbol{z}}, \tag{C.89}$$
$$\underline{a} = \inf \mathcal{A}_{Q,\boldsymbol{z}}, \tag{C.90}$$
$$\bar{b} = \sup \mathcal{B}_{Q,\boldsymbol{z}}, \text{ and} \tag{C.91}$$
$$\underline{b} = \inf \mathcal{B}_{Q,\boldsymbol{z}}, \tag{C.92}$$

such that

$$\mathcal{A} = (\underline{a}, \bar{a}) \subseteq (0, \infty), \text{ and} \tag{C.93a}$$
$$\mathcal{B} = \left(\underline{b}, \bar{b}\right) \subseteq \mathbb{R}. \tag{C.93b}$$

Let the function $F : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ be

$$F(a,b) = \int \dot{f}^{-1}\left(-\frac{b + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1. \tag{C.94}$$

The first part is as follows. Given the function F in (C.94), the continuity of F over the sets $\mathcal{A}$ and $\mathcal{B}$, defined in (C.93), is established by showing that F exhibits a limit at every point in $\mathcal{A}$ and $\mathcal{B}$, respectively. Then, for all $(a,b) \in \mathcal{A} \times \mathcal{B}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\dot{f}^{-1}\left(\frac{-b - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) \leq \dot{f}^{-1}\left(-\frac{b + \delta^\star_{Q,\boldsymbol{z}}}{a}\right) < \infty, \tag{C.95}$$

where equality holds if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \delta^\star_{Q,\boldsymbol{z}}$. Now, from [137, Corollary 24.5.1] the function $\dot{f}^{-1}$ is continuous, such that for all $b \in \mathcal{B}$, it holds that

$$\lim_{b \to \beta} \dot{f}^{-1}\left(\frac{-b - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) = \dot{f}^{-1}\left(\frac{-\beta - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right). \tag{C.96}$$

Hence, from the dominated convergence theorem [140, Theorem 1.6.9], the following limit exists and satisfies

$$\lim_{b\to\beta} F(a,b) = \lim_{b\to\beta} \int \dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.97}$$

$$= \int \left(\lim_{b\to\beta} \dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.98}$$

$$= \int \dot{f}^{-1}\left(-\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.99}$$

$$= F(\beta, a), \tag{C.100}$$

which proves that the function F in (C.94) is continuous in $\mathcal{B}$. Similarly, from [137, Corollary 24.5.1] the function $\dot{f}^{-1}$ is continuous, such that for all $a \in \mathcal{A}$, it holds that

$$\lim_{a\to\lambda} \dot{f}^{-1}\left(\frac{-b-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) = \dot{f}^{-1}\left(\frac{-b-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right). \tag{C.101}$$

Hence, from the dominated convergence theorem [140, Theorem 1.6.9], the following limit exists and satisfies

$$\lim_{a\to\lambda} F(a,b) = \lim_{a\to\lambda} \int \dot{f}^{-1}\left(\frac{-b-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.102}$$

$$= \int \left(\lim_{a\to\lambda} \dot{f}^{-1}\left(\frac{-b-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.103}$$

$$= \int \dot{f}^{-1}\left(\frac{-b-\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.104}$$

$$= F(\lambda, b), \tag{C.105}$$

which proves that the function F in (C.94) is continuous in $\mathcal{A}\times\mathcal{B}$. Given a pair $(a,b) \in \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{B}_{Q,\boldsymbol{z}}$, with $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7), assume that

$$N_{Q,\boldsymbol{z}}(a) = b. \tag{C.106}$$

This implies that

$$0 = \int \frac{\mathrm{d}P^{(Q,b)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.107}$$

$$= \int \dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.108}$$

$$= F(a,b). \tag{C.109}$$

Note that the inverse $\dot{f}^{-1}$ exists from the fact that f is strictly convex, which implies that $\dot{f}$ is a strictly increasing function. Hence, $\dot{f}^{-1}$ is also a strictly increasing function in $\mathcal{B}_{Q,\boldsymbol{z}}$ [141, Theorem 5.6.9]. Moreover, from the assumption that f is strictly convex and differentiable, it holds that $\dot{f}$ is continuous [144, Proposition 5.44]. This implies that $\dot{f}^{-1}$ is continuous. From Lemma A.3.1 the function $\dot{f}^{-1}$ is strictly increasing such that for all $b \in \mathcal{B}_{Q,\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right) \leq \dot{f}^{-1}\left(-\frac{b+\delta^{\star}_{Q,\boldsymbol{z}}}{a}\right), \tag{C.110}$$

with $\delta^{\star}_{Q,\boldsymbol{z}}$ defined in (3.16). Then, from (C.110) it follows that

$$\int \dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) < \int \dot{f}^{-1}\left(-\frac{b+\delta^{\star}_{Q,\boldsymbol{z}}}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.111}$$

$$= \dot{f}^{-1}\left(-\frac{b+\delta^{\star}_{Q,\boldsymbol{z}}}{a}\right) \tag{C.112}$$

$$< \infty, \tag{C.113}$$

where (C.113) follows from $\mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0,\infty)$, which implies $a > 0$. For all $(b_1, b_2) \in \mathcal{B}^2_{Q,\boldsymbol{z}}$, such that $b_1 < b < b_2$, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$-\frac{1}{a}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b_1) > -\frac{1}{a}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b) > -\frac{1}{a}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b_2), \tag{C.114}$$

which from Lemma A.3.1 implies that

$$\dot{f}^{-1}\left(-\frac{1}{a}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b_1)\right) > \dot{f}^{-1}\left(-\frac{1}{a}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b)\right) > \dot{f}^{-1}\left(-\frac{1}{a}(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+b_2)\right) \tag{C.115}$$

From (C.115), it holds that

$$F(a,b_1) > 0 > F(a,b_2), \tag{C.116}$$

which implies that the function F in (C.94) is strictly monotonic with respect to b.

The second part is as follows. From the definnition of $\mathcal{A}$ and $\mathcal{B}$ in (C.93) there exists atleast one point $(\lambda,\beta) \in \mathcal{A}\times\mathcal{B}$, such that

$$(\lambda,\beta) \in \mathcal{A}_{Q,\boldsymbol{z}} \times \mathcal{B}_{Q,\boldsymbol{z}}, \tag{C.117}$$

which implies that

$$F(\lambda,\beta) = \int \dot{f}^{-1}\left(-\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.118}$$

$$= \int \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - 1 \tag{C.119}$$

$$= \int \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) - 1 \tag{C.120}$$

$$= 0. \tag{C.121}$$

Note that from (C.97) and (C.102) the function F is continouse and thus the partial derivative of F satisfy

$$\frac{\partial}{\partial a}F(a,b) = \frac{\partial}{\partial a}\left(\int \dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) - 1\right) \tag{C.122}$$

$$= \int \frac{\partial}{\partial a}\dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.123}$$

$$= \int \frac{\mathrm{d}}{\mathrm{d}a}\dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.124}$$

$$= \int \frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a^2}\,\frac{1}{\ddot{f}\left(\dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\right)}\,\mathrm{d}Q(\boldsymbol{\theta}), \tag{C.125}$$

where (C.125) follows from Lemma A.3.2; and

$$\frac{\partial}{\partial b}F(a,b)=\frac{\partial}{\partial b}\left(\int \dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta})-1\right) \tag{C.126}$$

$$=\int \frac{\partial}{\partial b}\dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.127}$$

$$=\int \frac{\mathrm{d}}{\mathrm{d}b}\dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.128}$$

$$=\int -\frac{1}{a}\frac{1}{\ddot{f}\left(\dot{f}^{-1}\left(-\frac{b+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\right)}\mathrm{d}Q(\boldsymbol{\theta}), \tag{C.129}$$

where (C.129) follows from Lemma A.3.2. Note that, for all $(a,b)\times\mathcal{A}\times\mathcal{B}$, the partial derivative $\frac{\partial}{\partial b}F$, statisfies,

$$\frac{\partial}{\partial b}F(a,b)<0, \tag{C.130}$$

From the fact that $a>0$, and the strict convexity and twice differntiablity of f implies that for all $u\in\mathbb{R}$, the function $\ddot{f}$ satisfies, $\ddot{f}(u)>0$. Then, from *The Implicit Function Theorem* in [145, Theorem 4] the function $N_{Q,\boldsymbol{z}}$ is exists and is unique in the open interval $\mathcal{A}$ with $\mathcal{A}$ in (C.93) and for all $a\in\mathcal{A}$ satisfies that

$$N_{Q,\boldsymbol{z}}(a)=b, \tag{C.131}$$

such that

$$F(a,N_{Q,\boldsymbol{z}}(a))=0, \tag{C.132}$$

which completes the proof of continuity and uniqueness for the normalization function $N_{Q,\boldsymbol{z}}$. The third part is as follows. From [145, Theorem 4] it follows that

$$\frac{\mathrm{d}}{\mathrm{d}a}N_{Q,\boldsymbol{z}}(a)=-\left(\frac{\partial}{\partial b}F(a,N_{Q,\boldsymbol{z}}(a))\right)^{-1}\frac{\partial}{\partial a}F(a,N_{Q,\boldsymbol{z}}(a)), \tag{C.133}$$

$$=-\frac{\displaystyle\int \frac{N_{Q,\boldsymbol{z}}(a)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a^2}\frac{1}{\ddot{f}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(a)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\right)}\mathrm{d}Q(\boldsymbol{\theta})}{\displaystyle\int -\frac{1}{a}\frac{1}{\ddot{f}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(a)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{a}\right)\right)}\mathrm{d}Q(\boldsymbol{\nu})} \tag{C.134}$$

$$=\frac{\displaystyle\int \frac{N_{Q,\boldsymbol{z}}(a)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\frac{1}{\ddot{f}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(a)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\right)\right)}\mathrm{d}Q(\boldsymbol{\theta})}{\displaystyle\int \frac{1}{\ddot{f}\left(\dot{f}^{-1}\left(-\frac{N_{Q,\boldsymbol{z}}(a)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{a}\right)\right)}\mathrm{d}Q(\boldsymbol{\nu})} \tag{C.135}$$

$$=\frac{\displaystyle\int \frac{N_{Q,\boldsymbol{z}}(a)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{a}\left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)^{-1}\mathrm{d}Q(\boldsymbol{\theta})}{\displaystyle\int \left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\right)\right)^{-1}\mathrm{d}Q(\boldsymbol{\nu})} \tag{C.136}$$

$$= \frac{N_{Q,\boldsymbol{z}}(a)}{a} + \frac{1}{a} \frac{\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \left(\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \right)^{-1} \mathrm{d}Q(\boldsymbol{\theta})}{\int \left(\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \right) \right)^{-1} \mathrm{d}Q(\boldsymbol{\nu})} \tag{C.137}$$

$$= \frac{N_{Q,\boldsymbol{z}}(a)}{a} + \frac{1}{a} \frac{\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \left(\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right) \right)^{-1} \mathrm{d}Q(\boldsymbol{\theta})}{\int \left(\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \right) \right)^{-1} \mathrm{d}Q(\boldsymbol{\nu})} \tag{C.138}$$

$$= \frac{N_{Q,\boldsymbol{z}}(a)}{a} + \frac{1}{a} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\frac{1}{\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)}}{\int \frac{1}{\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \right)} \mathrm{d}Q(\boldsymbol{\nu})} \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.139}$$

The proof continues by considering a function $g_a : \mathcal{M} \to \mathbb{R}$, such that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$

$$g_a(\boldsymbol{\theta}) = \frac{\frac{1}{\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)}}{\int \frac{1}{\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \right)} \mathrm{d}Q(\boldsymbol{\nu})}. \tag{C.140}$$

Note that from the assumption that f is strictly convex and twice differentiable, the derivative $\dot{f}$ is increasing, and the second derivative $\ddot{f}$ is positive for all $\boldsymbol{\theta} \in \operatorname{supp} Q$. Also, the denominator of the fraction is the integral of the reciprocal of $\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \right)$ with respect to the measure Q. This term serves as a normalization constant ensuring that the resulting function is a proper probability density such that

$$\int g_a(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) = 1. \tag{C.141}$$

Therefore, the function g_a in (C.140) can be interpreted as the Radon-Nikodym derivative of a new probability measure $P^{(a)}$, parametrizes by the regularization factor a with respect to Q. Specifically, if we define a measure $P^{(a)}$ such that for any set $\mathcal{A} \in \mathscr{F}_{\mathcal{M}}$,

$$P^{(a)}(\mathcal{A}) = \int_{\mathcal{A}} \frac{\frac{1}{\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \right)}}{\int \frac{1}{\ddot{f} \left(\frac{\mathrm{d}P^{(Q,a)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu}) \right)} \mathrm{d}Q(\boldsymbol{\nu})} \, dQ(\boldsymbol{\theta}). \tag{C.142}$$

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

$$g_a(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(a)}}{\mathrm{d}Q}(\boldsymbol{\theta}). \tag{C.143}$$

From (C.139) and (C.143)

$$N_{Q,\boldsymbol{z}}(a) = a\frac{\mathrm{d}}{\mathrm{d}a}N_{Q,\boldsymbol{z}}(a) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\frac{\mathrm{d}P^{(a)}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.144}$$

$$= a\frac{\mathrm{d}}{\mathrm{d}a}N_{Q,\boldsymbol{z}}(a) - \mathsf{R}_{\boldsymbol{z}}\Big(P^{(a)}\Big), \tag{C.145}$$

whith $\mathsf{R}_{\boldsymbol{z}}$ defined in (2.12), which completes the proof of the ODE for the normalization function. This completes the proof. ■

C.3 Proof of Lemma 4.3.2

Proof: From the implicit function theorem the derivative of the normalization function satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}N_{Q,\boldsymbol{z}}(\lambda) = -\left(\frac{\partial}{\partial\beta}F(\lambda,\beta)\right)^{-1}\frac{\partial}{\partial\lambda}F(\lambda,\beta), \tag{C.146}$$

$$= \frac{\displaystyle\int \frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)^{-1}\mathrm{d}Q(\boldsymbol{\theta})}{\displaystyle\int\left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\right)\right)^{-1}\mathrm{d}Q(\boldsymbol{\nu})} \tag{C.147}$$

$$= \frac{\displaystyle\int -\dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)^{-1}\mathrm{d}Q(\boldsymbol{\theta})}{\displaystyle\int\left(\ddot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\right)\right)^{-1}\mathrm{d}Q(\boldsymbol{\nu})} \tag{C.148}$$

$$= \frac{\displaystyle\int -\dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\ddot{f}^*\left(\dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)\mathrm{d}Q(\boldsymbol{\theta})}{\displaystyle\int\ddot{f}^*\left(\dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\nu})\right)\right)\mathrm{d}Q(\boldsymbol{\nu})} \tag{C.149}$$

Let the function $\psi : \mathbb{R} \to \mathbb{R}$ be

$$\psi(y) = \frac{\ddot{f}^*(y)}{\dot{f}^*(y)} = \frac{\mathrm{d}}{\mathrm{d}y}\log\Big(\dot{f}^*(y)\Big). \tag{C.150}$$

Then, from (C.149) and (C.150)

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}N_{Q,\boldsymbol{z}}(\lambda) = \frac{\displaystyle\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\psi\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})}{\displaystyle\int\psi\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})}. \tag{C.151}$$

Note that under Assumptions (a), it holds that for all $v \in \mathcal{J}$, the functions $\dot{f}^{-1}$ and $\dot{f}^*$ satisfy $\dot{f}^{-1}(t) = \dot{f}^*(t)$. Hence, together with the assumption that $\dot{f}^{-1}$ is log convex, it implies that the function ψ is increasing. Therefore, the map $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mapsto \psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)$ is decreasing with respect to the empirical risk $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})$. Now, consider the integral

$$\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). \tag{C.152}$$

Note that modelts $\boldsymbol{\theta} \in \mathcal{M}$ are samples i.i.d with respect to the probability measure, therefore it holds that

$$\begin{aligned}
&\frac{1}{2}\int\int\Big(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\Big(\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)\\
&-\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\Big)\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\\
=&\frac{1}{2}\Big(\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\\
&+\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\\
&-\int\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\\
&-\int\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\Big) \qquad \text{(C.153)}\\
=&\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\\
&-\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\int\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). \qquad \text{(C.154)}
\end{aligned}$$

Observe that for all $(\boldsymbol{\theta},\boldsymbol{\nu}) \in \operatorname{supp} Q$ such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \geq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})$ then

$$\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big) \leq \psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big), \tag{C.155}$$

whichit implies that

$$\Big(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\Big(\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)-\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\Big) \leq 0. \tag{C.156}$$

Similarly, for all $(\boldsymbol{\theta},\boldsymbol{\nu}) \in \operatorname{supp} Q$ such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})$ then

$$\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big) > \psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big) \tag{C.157}$$

it holds that

$$\Big(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\Big(\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\Big)-\psi\Big(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})+\beta}{\lambda}\Big)\Big) < 0. \tag{C.158}$$

Ussing (C.154), (C.156) and (C.158)

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

$$\leq \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \int \psi\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right) P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). \tag{C.159}$$

From (C.151) and (C.159), it follows that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}N_{Q,\boldsymbol{z}}(\lambda) \leq \frac{\displaystyle\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \int \psi\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})}{\displaystyle\int \psi\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})} \tag{C.160}$$

$$= \int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{C.161}$$

$$= \int \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}). \tag{C.162}$$

Note that from the assumption in (4.10), it follows that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}N_{Q,\boldsymbol{z}}(\lambda) \leq 0. \tag{C.163}$$

whcih completes the proof. ■

C.4 Proof of Lemma 4.3.3

Proof: Given a reference measure Q, a dataset $\boldsymbol{z}$, a strictly convex and differentiable function f that induces an f-divergence and the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6), under the assumption that there exists a $\lambda \in (0,\infty)$, such that the optimization problem (4.1) has a solution, the proof is concerned with characterizing the set of all regularization factors λ for which a solution exists. This set of regularization factors is denoted by $\mathcal{A}_{Q,\boldsymbol{z}}$, where $\mathcal{A}_{Q,\boldsymbol{z}} \subseteq (0,\infty)$. The proof is divided into three parts. In the first part, the Legendre-Fenchel transform of f is connected to the Radon-Nikodym derivative of the solution to the optimization problem in (4.1) presented in Theorem 4.3.1. In the second part, the strictly increasing property of the Radon-Nikodym derivative, obtained from the connection established with the Legendre-Fenchel transform of f, is used to evaluate the real values of λ under which assumption (4.3a) holds. In the third part, the strictly increasing property is used to evaluate the real values of λ under which assumption (4.3b) holds.

The first part is as follows. The Legendre-Fenchel transform of f is defined as

$$f^*(t) \triangleq \sup_{s\in\mathcal{I}}(ts - f(s)), \tag{C.164}$$

where $f^* : \mathcal{J} \to \mathbb{R}$. From [137, Theorem 23.5] the Legendre-Fenchel transform f^* in (C.164) satisfies

$$f^*(t) = t\frac{\mathrm{d}}{\mathrm{d}t}f^*(t) - f\left(\frac{\mathrm{d}}{\mathrm{d}t}f^*(t)\right). \tag{C.165}$$

Furthermore, from [137, Corollary 23.5.1], the function $\frac{\mathrm{d}}{\mathrm{d}t}f^* : \mathcal{J} \to \mathcal{I}$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}f^*(t) = \left(\frac{\mathrm{d}f}{\mathrm{d}t}\right)^{-1}(t), \tag{C.166}$$

which is the functional inverse of the derivative of f, denoted by $\dot{f}^{-1}$ for simplicity. Note that, given the assumption that f is strictly convex and induces an f-divergence, it follows from [137, Theorem 12.2] that the function f^* in (C.164) is also strictly convex. From the strict convexity of f^*, it follows from Lemma A.3.1 that $\dot{f}^{-1}$ in (C.166) is strictly increasing. Furthermore, from [137, Corollaey 26.3.1] f^* in (C.164) is bijective with $\frac{\mathrm{d}f}{\mathrm{d}s} : \mathcal{I} \to \mathcal{J}$, which completes the first part of the proof.

The second part is as follows. Evaluating the real values of λ under which assumption in (4.3a) holds requires to show that the function $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ belongs to the set of nonnegative measurable functions. From the f-divergence in Defintion 1.5.1 and the fact that $\dot{f}^{-1}$ strictly increasing and bijective, the proof follows by showing that the limit

$$\lim_{x\to 0^+} \dot{f}(x) = t_0, \tag{C.167}$$

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

$$-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda} > t_0. \tag{C.168}$$

Note that (C.168) is suficient from the fact that the monotonicity of $\dot{f}^{-1}$ implies that for all $t > t_0$,

$$\dot{f}^{-1}(t) > 0. \tag{C.169}$$

To evaluate the real values of λ under which assumption in (4.3a) holds, three cases must be considered for the limit in (C.167).

Case 1: Assume that

$$\lim_{x\to 0^+} \dot{f}(x) = \infty. \tag{C.170}$$

Under the above assumption, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda} < \infty, \tag{C.171}$$

which implies that

$$\dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right) < 0. \tag{C.172}$$

Hence, assumption (b) of Theorem 4.3.1 is not satisfied and nothing can be stated about the solution.

Case 2: Assume that

$$\lim_{x\to 0^+} \dot{f}(x) = a, \tag{C.173}$$

where $a \in \mathbb{R}$. Under the above assumption, consider the set

$$\mathcal{D} = \{\boldsymbol{\theta} \in \operatorname{supp} Q : -\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < a\lambda + \beta\}. \tag{C.174}$$

On one hand, note that if the function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is unbounded in $\operatorname{supp} Q$, from (C.173) the set $\mathcal{D}$ in (C.174) is nonegligble and measurable, such that for all $\boldsymbol{\theta} \in \mathcal{D}$,

$$-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda} < a, \tag{C.175}$$

which implies that

$$\dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right) < 0. \tag{C.176}$$

Hence, assumption (b) of Theorem 4.3.1 is not satisfied and nothing can be stated about the solution. On the other hand, if the function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is bounded in $\operatorname{supp} Q$, such that

$$M = \sup_{\boldsymbol{\theta} \in \operatorname{supp} Q} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}). \tag{C.177}$$

Then, there exists a $\lambda_{Q,\boldsymbol{z}} \in [0, \infty)$ such that

$$-M = a\lambda_{Q,\boldsymbol{z}} + \beta. \tag{C.178}$$

From (C.178) for all $\lambda > \lambda_{Q,\boldsymbol{z}}$, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right) > 0. \tag{C.179}$$

From (C.179), consider the following conditions: If there exists a model $\bar{\boldsymbol{\theta}} \in \operatorname{supp} Q$ such that $\mathsf{L}_{\boldsymbol{z}}(\bar{\boldsymbol{\theta}}) = M$, where M is defined in (C.177), then the set of regularization factors λ for which the function $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is nonnegative is $[\lambda_{Q,\boldsymbol{z}}, \infty)$. Alternatively, if for all models $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < M$, where M is defined in (C.177), then the set of regularization factors λ for which the function $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is nonnegative is $(\lambda_{Q,\boldsymbol{z}}, \infty)$.

Case 3: Assume that

$$\lim_{x \to 0^+} \dot{f}(x) = -\infty. \tag{C.180}$$

Under the above assumption, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda} > -\infty, \tag{C.181}$$

which implies that

$$\dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right) > 0. \tag{C.182}$$

Hence, for all $\lambda \in (0, \infty)$ assumption (b) of Theorem 4.3.1 is satisfie such that the nonnegativity of function $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ is guaranteed. This completes the second part of the proof.

The third part is as follows. Evaluating the values λ under which assumption (4.3b) holds requires to show there exists a real value $\beta \in \mathbb{R}$ such that the integral of $\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}$ with respect to Q is one. From Theorem 4.3.2 the monotonicity of the normalization

function $N_{Q,\boldsymbol{z}}$ in (4.7) there is a minimum regularization factor $\lambda^{\star}_{Q,\boldsymbol{z}}$ defined in (3.17). Furthermore, from Theorem 4.3.2 the continuity of the function $N_{Q,\boldsymbol{z}}$ implies that for all $\lambda \in \left(\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right)$, there exists a unique $\beta \in \mathcal{B}_{Q,\boldsymbol{z}}$ such that assumption (b) of Theorem 4.3.1 is satisfied. From Theorem 4.3.2, it holds that

$$\lim_{\lambda \to {\lambda^{\star}_{Q,\boldsymbol{z}}}^{+}} N_{Q,\boldsymbol{z}}(\lambda) = N_{Q,\boldsymbol{z}}\big(\lambda^{\star}_{Q,\boldsymbol{z}}\big), \tag{C.183}$$

with the function $N_{Q,\boldsymbol{z}}$ defined in (4.7) and the limit from the right is well defined from the fact that the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is convex. To determine whether the infimum in (3.17) belongs to the set $\mathcal{A}_{Q,\boldsymbol{z}}$ two cases are considered.

Case 1: Assume that $\beta > N_{Q,\boldsymbol{z}}\left(\lambda^{\star}_{Q,\boldsymbol{z}}\right)$, such that

$$\int \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda^{\star}_{Q,\boldsymbol{z}}}\right) \mathrm{d}Q(\boldsymbol{\theta}) = \infty. \tag{C.184}$$

Notice that from (C.184) for all $\beta_1 \in \left[N_{Q,\boldsymbol{z}}\left(\lambda^{\star}_{Q,\boldsymbol{z}}\right), \beta\right)$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\dot{f}^{-1}\left(-\frac{\beta_1 + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda^{\star}_{Q,\boldsymbol{z}}}\right) > \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda^{\star}_{Q,\boldsymbol{z}}}\right). \tag{C.185}$$

Hence, under the above assumption, $N_{Q,\boldsymbol{z}}\left(\lambda^{\star}_{Q,\boldsymbol{z}}\right) \notin \mathcal{B}_{Q,\boldsymbol{z}}$ which implies that the set of all regularization $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7) that satisfy assumption (4.3b) is $\mathcal{A}_{Q,\boldsymbol{z}} = \left(\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right)$.

Case 2: Assume that $\beta > N_{Q,\boldsymbol{z}}\left(\lambda^{\star}_{Q,\boldsymbol{z}}\right)$, such that

$$\int \dot{f}^{-1}\left(-\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda^{\star}_{Q,\boldsymbol{z}}}\right) \mathrm{d}Q(\boldsymbol{\theta}) < \infty. \tag{C.186}$$

From the monotonicity of the solution in part one and continuity of the function $N_{Q,\boldsymbol{z}}$ in (4.7) from Theorem 4.3.2, there exists a $\beta^{\star}_{Q,\boldsymbol{z}} \in \mathcal{B}_{Q,\boldsymbol{z}}$ such that $N_{Q,\boldsymbol{z}}\left(\lambda^{\star}_{Q,\boldsymbol{z}}\right) = \beta^{\star}_{Q,\boldsymbol{z}}$, which implies that the set of all regularization factors $\mathcal{A}_{Q,\boldsymbol{z}} = \left[\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right)$.

Finally, from parts two and three of the proof the set $\mathcal{A}_{Q,\boldsymbol{z}}$ is a convex set such that the regularization factors for which the assumptions of Theorem 4.3.1 hold and are given by

$$\left(\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right) \subseteq \mathcal{A}_{Q,\boldsymbol{z}} \subseteq \left[\lambda^{\star}_{Q,\boldsymbol{z}}, \infty\right), \tag{C.187}$$

which completes the proof.

■

C.5 Proof of Lemma 4.3.4

Proof: The following proof is divided into two parts. In the first part, an auxiliary function is introduced and proven to be continuous. In the second part, a contradiction is shown under the assumption that $\dot{f}^{-1}$ is nonnegative and the continuity of the

auxiliary function. Finally, it is shown that for nonnegative $\dot{f}^{-1}$, the set of admissible regularization factors is the positive reals.

The first part is as follows. Let the function $k : \mathbb{R} \to (0, +\infty)$, be such that

$$k(b) = \int \dot{f}^{-1}\left(\frac{-b - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.188}$$

The first step is to prove that the function k in (C.188) is continuous in $\mathbb{R}$. This is proved by showing that k always exhibits a limit. Note that from Lemma A.3.1 the function $\dot{f}^{-1}$ is strictly increasing, it holds that for all $b \in \mathcal{B}_{Q,\boldsymbol{z}}$ with $\mathcal{B}_{Q,\boldsymbol{z}}$ defined in (4.7) and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, it holds that

$$\dot{f}^{-1}\left(\frac{-b - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \leq \dot{f}^{-1}\left(-\frac{b}{\lambda}\right), \tag{C.189}$$

where equality holds if and only if $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = 0$. Now, from the [137, Corollary 24.5.1] the function $\dot{f}^{-1}$ is continuous, such that for all $a \in \mathcal{B}$, it holds that

$$\lim_{b \to \beta} \dot{f}^{-1}\left(\frac{-b - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) = \dot{f}^{-1}\left(\frac{-\beta - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right). \tag{C.190}$$

Hence, from the dominated convergence theorem [140, Theorem 1.6.9], the following limit exists and satisfies

$$\begin{aligned}
\lim_{b \to \beta} k(b) &= \lim_{b \to \beta} \int \dot{f}^{-1}\left(\frac{-b - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.191)}\\
&= \int \left(\lim_{b \to \beta} \dot{f}^{-1}\left(\frac{-b - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.192)}\\
&= \int \dot{f}^{-1}\left(\frac{-\beta - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(C.193)}\\
&= k(\beta), && \text{(C.194)}
\end{aligned}$$

which proves that the function k in (C.188) is continuous.

The second part is as follows. From the assumption that $\mathcal{B}_{Q,\boldsymbol{z}}$ is nonempty, there is a $b \in \mathcal{B}_{Q,\boldsymbol{z}}$ and a $\lambda \in (0, \infty)$ such that,

$$1 = \int \dot{f}^{-1}\left(-\frac{b + \mathsf{L}_{\boldsymbol{z}}}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.195}$$

From [137, Corollary 24.5.1] and Lemma A.3.1 the function $\dot{f}^{-1}$ is continuous and strictly increasing, for all $b_1 \in \left(b^{\star}_{Q,\boldsymbol{z}}, b\right)$ and for all $b_2 \in (b, \infty)$, it holds that

$$\int \dot{f}^{-1}\left(-\frac{b_1 + \mathsf{L}_{\boldsymbol{z}}}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) > 1 > \int \dot{f}^{-1}\left(-\frac{b_2 + \mathsf{L}_{\boldsymbol{z}}}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.196}$$

Under the same argument, for all $\lambda_1 \in (0, \lambda)$ and for all $\lambda_2 \in (\lambda, \infty)$, it holds that

$$\int \dot{f}^{-1}\left(-\frac{b + \mathsf{L}_{\boldsymbol{z}}}{\lambda_1}\right) \mathrm{d}Q(\boldsymbol{\theta}) < 1 < \int \dot{f}^{-1}\left(-\frac{b + \mathsf{L}_{\boldsymbol{z}}}{\lambda_2}\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{C.197}$$

Hence, given that the function k in (C.188) is continuous, strictly decreasing, from (C.196) then, there always exists two reals b_1 and b_2 in $\mathcal{B}_{Q,\boldsymbol{z}}$ such that $k(b_1) < 1 < k(b_2)$, it follows from the intermediate-value theorem [146, Theorem 4.23] that there always

exists a unique real $b \in \mathcal{B}_{Q,\boldsymbol{z}}$ such that $k(b) = 1$. Furthermore, for all $b \in \mathcal{B}_{Q,\boldsymbol{z}}$ there always exists two reals λ_1 and λ_2 in $(0, \infty)$ such that inequality (C.197) holds, it follows from the intermediate-value theorem [146, Theorem 4.23] that there always exists a unique real $b \in \mathcal{B}_{Q,\boldsymbol{z}}$ for all $\lambda \in (0, \infty)$ such that $k(b) = 1$. Finally, from the fact that $N_{Q,\boldsymbol{z}}$ in (4.7) is continuous and strictly increasing, if $\mathcal{B}_{Q,\boldsymbol{z}} = \left(t^\star_{Q,\boldsymbol{z}}, \infty\right)$ then the set of admisible regularization factors $\mathcal{A}_{Q,\boldsymbol{z}}$ in (4.7a) is identical to $(0, \infty)$, which completes the proof. ■

C.6 Proof of Lemma 4.3.5

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) \geq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2), \tag{C.198}$$

and thus, for all $\lambda > 0$, such that $\beta \in \mathcal{B}$, with $\mathcal{B}$ defined in (4.3a), it holds that

$$-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) + \beta}{\lambda} \leq -\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) + \beta}{\lambda}. \tag{C.199}$$

From Lemma A.3.1 the function $\dot{f}^{-1}$ is strictly increasing, which implies that

$$\dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_1) + \beta}{\lambda}\right) \leq \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}_2) + \beta}{\lambda}\right). \tag{C.200}$$

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

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_1) \leq \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}_2), \tag{C.201}$$

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

C.7 Proof of Lemma 4.3.6

Proof: From Lemma 4.3.5, it follows that for all $\lambda > 0$ such that $\beta \in \mathcal{B}$, with $\mathcal{B}$ defined in (4.3a), it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$, and for all $\boldsymbol{\phi} \in \mathcal{L}^\star_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$,

$$\begin{aligned}
\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) &\leq \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\phi}) && \text{(C.202)}\\
&= \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\right) && \text{(C.203)}\\
&= \dot{f}^{-1}\left(-\frac{\delta^\star_{Q,\boldsymbol{z}} + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\right) && \text{(C.204)}\\
&< \infty, && \text{(C.205)}
\end{aligned}$$

where (C.202) follows from (4.4); (C.203) follows from the fact that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\phi}) \geq \delta^\star_{Q,\boldsymbol{z}}$; and (C.205) follows from the fact that for all $\lambda > 0$, the function $N_{Q,\boldsymbol{z}}(\lambda) < \infty$. From the definition of $\delta^\star_{Q,\boldsymbol{z}}$ in (3.16) and $\mathcal{L}^\star_{Q,\boldsymbol{z}}$ in (3.18) equality in (C.202) holds if and only if $\boldsymbol{\theta} \in \mathcal{L}^\star_{Q,\boldsymbol{z}} \cap \operatorname{supp} Q$. This completes the proof of finiteness.

For the proof of positivity, observe that from Corollary 4.3.1 it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

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

which completes the proof. ■

C.8 Proof of Lemma 4.3.7

Proof: The proof is divided into two parts. The first part shows via contradiction that for all $\lambda > 0$ such that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the solution to the optimization problems in (4.1) and in (4.2) that the expected empirical risk of $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is less than the expected empirical risk of the prior Q. The second part shows that as λ increases, the expected empirical risk increases. The first part is as follows. Under the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) is serparable (see Definition 2.4.1), there exist a real value $a > 0$ and a partition of the set $\mathcal{M}$ formed by some noneglible sets $\mathcal{A}_0$ and $\mathcal{A}_1$ with respect to Q that satisfy the following:

$$\mathcal{A}_0 \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geq a\}, \tag{C.207a}$$
$$\mathcal{A}_1 \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) < a\}. \tag{C.207b}$$

Under the assumption that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) satisfies

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) < Q(\mathcal{A}_0), \tag{C.208}$$

and it follows that

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < \mathsf{R}_{\boldsymbol{z}}(Q), \tag{C.209}$$

with $\mathsf{R}_{\boldsymbol{z}}$ defined in (2.12). From the fact that for all $P \in \triangle_Q(\mathcal{M})$, the f-divergence in (4.2) satisfies

$$\mathsf{D}_f(P\|Q) \geq 0, \tag{C.210}$$

where the equality holds if and only if for all $\mathcal{A} \in \mathscr{F}(\mathcal{M})$, the probability $P(\mathcal{A}) = Q(\mathcal{A})$. Then, under the assumption that the probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (4.4) satisfies

$$P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) > Q(\mathcal{A}_0), \tag{C.211}$$

it follows that

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) > \mathsf{R}_{\boldsymbol{z}}(Q). \tag{C.212}$$

From (C.210) and (C.212) it follows

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \lambda \mathsf{D}_f\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) \geq \mathsf{R}_{\boldsymbol{z}}(Q) + \lambda \mathsf{D}_f(Q\|Q), \tag{C.213}$$

which is a contradiction. This implies that for $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to minimize the optimization problems in (4.1) and (4.2), the expected empirical risk of $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < \mathsf{R}_{\boldsymbol{z}}(Q), \tag{C.214}$$

which completes the first part of the proof.

The second part is as follows. Given two regularization factors $(\lambda_1, \lambda_2) \in [0, \infty)^2$, such that $\lambda_1 < \lambda_2$, it holds for all $i \in \{1, 2\}$ that

$$\frac{\mathrm{d}P^{(Q,\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta_i}{\lambda_i}\right) \tag{C.215a}$$

$$= \dot{f}^{-1}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda_i)}{\lambda_i}\right) \tag{C.215b}$$

where equality (C.215b) follows from (4.7). From Theorem 4.3.2 and the fact that $\lambda_1 < \lambda_2$ it holds that $N_{Q,\boldsymbol{z}}(\lambda_1) < N_{Q,\boldsymbol{z}}(\lambda_2)$. Let a in (C.207) be

$$a = \frac{\lambda_1 N_{Q,\boldsymbol{z}}(\lambda_2) - \lambda_2 N_{Q,\boldsymbol{z}}(\lambda_2)}{\lambda_2 - \lambda_1}. \tag{C.216}$$

Note that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ such that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = a$, it holds that

$$\frac{\mathrm{d}P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}). \tag{C.217}$$

Then, for all $\boldsymbol{\theta} \in \mathcal{A}_0$, it holds that

$$\frac{\mathrm{d}P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \geq \frac{\mathrm{d}P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{C.218}$$

and for all $\boldsymbol{\theta} \in \mathcal{A}_1$, it holds that

$$\frac{\mathrm{d}P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) < \frac{\mathrm{d}P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}), \tag{C.219}$$

where inequalities (C.218) and (C.219) follow from (C.217) and the fact that that $\dot{f}^{-1}$ in (C.19) is strictly increasing. Hence, the probability measures $P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ over the set $\mathcal{A}_1$ satysfy

$$P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) = \int_{\mathcal{A}_1} \frac{\mathrm{d}P^{(Q,\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.220}$$

$$< \int_{\mathcal{A}_1} \frac{\mathrm{d}P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) \tag{C.221}$$

$$= P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1), \tag{C.222}$$

where inequality (C.221) follows from (C.68). Therefore, from (C.222) the expected empirical risk satisfies

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{C.223}$$

which implies that the expected empirical risk increases with respect to λ. This completes the proof. ■

Appendix D

Dual ERM-fDR

D.1 Proof of Theorem 5.4.1

Proof: From Theorem 4.3.1, Theorem 5.3.1 and Lemma A.3.3, the Legendre-Fenchel transform in Definition 1.5.2 satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) = -\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) - f^\star\left(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\right), \tag{D.1}$$

where (D.1) can be rearranged into

$$-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^\star\left(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\right). \tag{D.2}$$

Using (D.2), it can be shown that

$$\begin{aligned}
&\lambda\int\left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\right)\right)\left(1 - \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) \\
&= \lambda\int -\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\left(1 - \frac{\mathrm{d}P}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) && \text{(D.3)} \\
&= \lambda\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta}) - \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta}) && \text{(D.4)} \\
&= \lambda\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\,\mathrm{d}P(\boldsymbol{\theta}) - \lambda\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda)}{\lambda}\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && \text{(D.5)} \\
&= \int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) - \int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + N_{Q,\boldsymbol{z}}(\lambda) - N_{Q,\boldsymbol{z}}(\lambda) && \text{(D.6)} \\
&= \mathsf{R}_{\boldsymbol{z}}(P) - \mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), && \text{(D.7)} \\
&= \mathsf{G}\left(\boldsymbol{z}, P, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), && \text{(D.8)}
\end{aligned}$$

where (D.3) follows from (D.2), (D.4) follows from the fact that P is absolutely continuous with respect to Q; and (D.4) follows from (5.17). This completes the proof. ∎

D.2 Proof of Lemma 5.3.3

Proof: The Legendre-Fenchel transform of a strictly convex function $f : \mathcal{I} \to \mathbb{R}$ satisfies

$$f^*(t) \triangleq \sup_{s\in\mathcal{I}}(ts - f(s)). \tag{D.9}$$

From [137, Theorem 23.5] if f is strictly convex then maximizing argument of the convex conjugate f^* satisfies

$$f^*(t) = t\frac{\mathrm{d}}{\mathrm{d}t}f^*(t) - f\left(\frac{\mathrm{d}}{\mathrm{d}t}f^*(t)\right). \tag{D.10}$$

Furthermore, from [137, Corollary 23.5.1] the function $\dot{f}^{-1}$ is the derivative of the convex conjugate of f in (D.10), which implies that Differential Equations

$$f^*(t) = t\dot{f}^{-1}(t) - f\left(\dot{f}^{-1}(t)\right). \tag{D.11}$$

From Theorem 4.3.1, let $t = -\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}$ in the optimization problems in (4.1) and (4.2), then it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right) = -\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) - f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right). \tag{D.12}$$

Taking the integral of (D.12) with respect to the reference measure Q, yields

$$\begin{aligned}
&\int f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) \\
&= \int\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) - f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\right)\mathrm{d}Q(\boldsymbol{\theta}) &\text{(D.13)}\\
&= -\frac{1}{\lambda}\left(\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)+\beta\right) - \mathsf{D}_f\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right). &\text{(D.14)}
\end{aligned}$$

Arranging (D.14) results in

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \lambda\mathsf{D}_f\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|Q\right) = -\lambda\int f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right)\mathrm{d}Q(\boldsymbol{\theta}) - \beta, \tag{D.15}$$

which completes the proof.

The second part of the proof is as follows. From Lemma 4.3.1, equality (D.12) can be rewriten as

$$\begin{aligned}
&\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda} \\
&= -f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right)\frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) - \frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right). &\text{(D.16)}
\end{aligned}$$

Taking the integral of (D.16) with respect to the reference measure Q, yields

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

$$- \int \frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}). \tag{D.17}$$

Arranging (D.17) results in

$$\begin{aligned} &\mathsf{R}_{\boldsymbol{z}}(Q) + \lambda \int \frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \\ &= -\lambda \int f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right) \frac{\mathrm{d}Q}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta}) - \beta, \end{aligned} \tag{D.18}$$

which completes the proof. ■

D.3 Proof Theorem 5.4.2

Proof: From (5.17) the gap for an arbitrary dataset $\boldsymbol{z}$ and two arbitrary probability measures P_1 and P_2 satisfies

$$\mathsf{G}(\boldsymbol{z}, P_1, P_2) = \mathsf{R}_{\boldsymbol{z}}(P_1) - \mathsf{R}_{\boldsymbol{z}}(P_2) \tag{D.19}$$

From Theorem 5.4.1 the gap for an arbitrary dataset $\boldsymbol{z}$ and two arbitrary probability measures P_1 and P_2 in $\triangle_Q(\mathcal{M})$, is given satisfies

$$\begin{aligned} &\mathsf{G}(\boldsymbol{z}, P_1, P_2) \\ &\quad = \mathsf{G}\Big(\boldsymbol{z}, P_1, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) - \mathsf{G}\Big(\boldsymbol{z}, P_2, P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big) && \text{(D.20)} \\ &\quad = \lambda \int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right)\right) \left(1 - \frac{\mathrm{d}P_1}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) \\ &\qquad - \lambda \int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right)\right) \left(1 - \frac{\mathrm{d}P_2}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}) && \text{(D.21)} \\ &\quad = \lambda \int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right)\right) \left(\frac{\mathrm{d}P_2}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right. \\ &\qquad \left. - \frac{\mathrm{d}P_1}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta}). && \text{(D.22)} \end{aligned}$$

Substituting the probability measures P_1 and P_2 for the probability measures $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P_{\boldsymbol{\Theta}}$; and taking the expectation of (D.22) with respect to $P_{\boldsymbol{Z}}$ yields

$$\begin{aligned} &\overline{\overline{\mathsf{G}}}\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\big) \\ &\quad = \lambda \int \int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta}{\lambda}\right)\right) \left(\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right. \\ &\qquad \left. - \frac{\mathrm{d}P_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \end{aligned} \tag{D.23}$$

which completes the proof. ■

D.4 Proof Theorem 5.4.3

Proof: From Theorem 4.3.1 and [137, Corollary 23.5.1], the Legendre-Fenchel transform in Definition 1.5.2 satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^{\star}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right). \tag{D.24}$$

Then, from Theorem 5.4.2 and (D.24), the generalization error of the solution to the ERM-fDR problem in (4.1) satisfies

$$\begin{aligned}
&\overline{\overline{\mathsf{G}}}\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\mathbf{Z}}, P_{\boldsymbol{Z}}\right)\\
&= \lambda \int\int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^{*}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right)\right)\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right.\\
&\quad \left. - \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \qquad \text{(D.25)}\\
&= \lambda \int\int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^{*}\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right)\right)\left(1 - \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \qquad \text{(D.26)}\\
&= \lambda \int\int \dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\left(1 - \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \qquad \text{(D.27)}\\
&= \lambda\left(\int\int \dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})\right.\\
&\quad \left. - \int\int \dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}Q}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})\right) \qquad \text{(D.28)}\\
&= \lambda\left(\int\int \dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})\right.\\
&\quad \left. - \int\int \dot{f}\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})\right), \qquad \text{(D.29)}
\end{aligned}$$

where (D.27) follows from (D.24), (D.28) follows from Corollary 4.3.1. This completes the proof. ∎

D.5 Proof Remark 1

Proof: Under the assumption that the function f in (4.1) is

$$f(x) = x\log(x), \tag{D.30}$$

from the Legendre-Fenchel transform in Definition 1.5.2 it follows that

$$f^{\star}(t) = \exp(t+1). \tag{D.31}$$

Note that for the relative entropy, it also holds that

$$\frac{\mathrm{d}}{\mathrm{d}t} f^\star(t) = \exp(t+1), \tag{D.32}$$

which together with (D.31) and Theorem 4.3.1 yields

$$f^\star\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right) = \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}). \tag{D.33}$$

Then, under the assumption in (D.30), the Gibbs algorithm satisfies for all $\boldsymbol{z} \in \operatorname{supp} P_{\boldsymbol{Z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}). \tag{D.34}$$

Then, from Theorem 5.4.2 it follows that,

$$\begin{aligned}
&\overline{\overline{\mathsf{G}}}\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\mathbf{Z}}, P_{\boldsymbol{Z}}\Big) \\
&= \lambda \int\int \left(f\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + f^*\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})+\beta}{\lambda}\right)\right)\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right. \\
&\quad \left. - \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \qquad \text{(D.35)} \\
&= \lambda \int\int \left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right)\left(1\right. \\
&\quad \left. - \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \qquad \text{(D.36)} \\
&= \lambda \int\int \left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + 1\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\left(1\right. \\
&\quad \left. - \frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta})\right) \mathrm{d}Q(\boldsymbol{\theta})\, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \qquad \text{(D.37)} \\
&= \lambda \int \left(\int \left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + 1\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta})\right. \\
&\quad \left. - \int \left(\log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) + 1\right)\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}}{\mathrm{d}Q}(\boldsymbol{\theta})\, \mathrm{d}Q(\boldsymbol{\theta})\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \qquad \text{(D.38)} \\
&= \lambda \int \left[\int \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right. \\
&\quad \left. - \int \log\left(\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\right] \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \qquad \text{(D.39)}
\end{aligned}$$

where (D.35) follows from (D.30) and (D.34). Note also that (D.38) is the result of substituting $\dot{f}(x) = \log(x)+1$ into Theorem 5.4.3. This completes the proof. ■

Appendix E

Complementary

E.1 Examples

E.1.1 Example 1

Consider the ERM-fDR problem in (4.1) with reverse relative entropy and assume that: (a) $\mathcal{M} = \mathcal{X} = \mathcal{Y} = [0, \infty)$; (b) $\boldsymbol{z} = (1, 0)$; and (c) $Q \ll \mu$, with μ the Lebesgue measure, such that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu}(\boldsymbol{\theta}) = 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}). \tag{E.1}$$

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

$$f(\boldsymbol{\theta}, x) = x\boldsymbol{\theta}, \tag{E.2}$$

and the loss function ℓ in (2.5) be

$$\ell(f(\boldsymbol{\theta}, x), y) = (x\boldsymbol{\theta} - y)^2, \tag{E.3}$$

which implies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = (x\boldsymbol{\theta} - y)^2, \tag{E.4}$$

with the function $\mathsf{L}_{\boldsymbol{z}}$ defined in (2.6). Furthermore, from assumptions (a), (b), and (E.4), 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

$$\delta^\star_{Q,\boldsymbol{z}} = 0. \tag{E.5}$$

Under the current assumptions, the objective of this example is to show that $\mathcal{B}_{Q,\boldsymbol{z}} = \left[\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$. For this purpose, from Lemma 4.3.3, it is sufficient to show that the conditions in (4.3) hold. From Theorem 4.3.1, it follows that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (E.71) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}\mu}(\boldsymbol{\theta}) = \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}), \tag{E.6}$$

with β satisfying (4.3). Thus,

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

$$= \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} 4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{\theta}) \tag{E.7}$$

$$= \int_0^{\infty} \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(x\boldsymbol{\theta} - y)^2 - \delta^{\star}_{Q,\boldsymbol{z}}} \, \mathrm{d}\boldsymbol{\theta} \tag{E.8}$$

$$= \int_0^{\infty} \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}^2 - \delta^{\star}_{Q,\boldsymbol{z}}} \, \mathrm{d}\boldsymbol{\theta} \tag{E.9}$$

$$= \int_0^{\infty} \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}^2} \, \mathrm{d}\boldsymbol{\theta} \tag{E.10}$$

$$= \int_0^{\infty} 4 \exp(-2\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{E.11}$$

$$= 2, \tag{E.12}$$

where (E.7) follows from (E.1); (E.9) follows from the assumption that $(x, y) = (1, 0)$; and (E.10) follows from the fact that $\delta^{\star}_{Q,\boldsymbol{z}} = 0$. Finally, the function $N_{Q,\boldsymbol{z}}$ in (4.7) satisfies $N_{Q,\boldsymbol{z}}\left(\frac{1}{2}\right) = 0$, which implies $-\delta^{\star}_{Q,\boldsymbol{z}} \in \mathcal{B}_{Q,\boldsymbol{z}}$, thus the set $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$.

E.1.2 Example 2

Consider Example 1 in Appendix E.1.1 with $\boldsymbol{z} = ((1, 1))$. Note that (E.5) 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 4.3.3, it is sufficient to show that the conditions in (4.3) do not hold. That is,

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

$$= \int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \frac{\mathrm{d}Q}{\mathrm{d}\mu}(\boldsymbol{\theta}) \, \mathrm{d}\mu(\boldsymbol{\theta}) \tag{E.13}$$

$$= \int \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}} \, \mathrm{d}\mu(\boldsymbol{\theta}) \tag{E.14}$$

$$= \int \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(x\boldsymbol{\theta} - y)^2 - \delta^{\star}_{Q,\boldsymbol{z}}} \, \mathrm{d}\mu(\boldsymbol{\theta}) \tag{E.15}$$

$$= \int \frac{4\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \, \mathrm{d}\mu(\boldsymbol{\theta}) \tag{E.16}$$

$$= 4 \int \frac{\boldsymbol{\theta}^2 \exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta} - 1)^2} \, \mathrm{d}\mu(\boldsymbol{\theta}) \tag{E.17}$$

$$= 4 \int \left(\frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \frac{\exp(-2\boldsymbol{\theta})}{2}}{\boldsymbol{\theta} - 1} - \frac{-\frac{\exp(-2\boldsymbol{\theta})}{2} - \frac{\exp(-2\boldsymbol{\theta})}{2}}{(\boldsymbol{\theta} - 1)^2} \right) \mathrm{d}\mu(\boldsymbol{\theta}) \tag{E.18}$$

$$= 4 \left(\int \frac{\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \frac{\exp(-2\boldsymbol{\theta})}{2}}{\boldsymbol{\theta} - 1} \, \mathrm{d}\mu(\boldsymbol{\theta}) + \int \frac{\frac{\exp(-2\boldsymbol{\theta})}{2} + \frac{\exp(-2\boldsymbol{\theta})}{2}}{(\boldsymbol{\theta} - 1)^2} \, \mathrm{d}\mu(\boldsymbol{\theta}) \right) \tag{E.19}$$

$$= 4 \left(\frac{1}{2} \int \frac{2\boldsymbol{\theta} \exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta} - 1} \, \mathrm{d}\mu(\boldsymbol{\theta}) \right.$$

$$+\frac{1}{2}\int \frac{\exp(-2\boldsymbol{\theta})+\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\mu(\boldsymbol{\theta})\Big), \tag{E.20}$$

where (E.13) follows from the assumption that $Q \ll \mu$, (E.14) follows from (E.1), (E.16) follows from (E.5) and the assumption that $(x, y) = (1, 1)$. Using integration by parts on the second integral in (E.20), let the functions $\phi : \mathcal{M} \to \mathbb{R}$ and $\psi : \mathcal{M} \to \mathbb{R}$ be

$$\phi(\boldsymbol{\theta}) = \boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta}), \text{ and} \tag{E.21a}$$
$$\psi(\boldsymbol{\theta}) = -\frac{1}{\boldsymbol{\theta}-1}. \tag{E.21b}$$

The derivatives of ϕ and ψ satisfy

$$\frac{\mathrm{d}\phi}{\mathrm{d}\mu}(\boldsymbol{\theta}) = -2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) - \exp(-2\boldsymbol{\theta}), \text{ and} \tag{E.22a}$$
$$\frac{\mathrm{d}\psi}{\mathrm{d}\mu}(\boldsymbol{\theta}) = \frac{1}{(\boldsymbol{\theta}-1)^2}, \tag{E.22b}$$

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

$$\begin{aligned}
&\int_{[a,b]} \frac{\exp(-2\boldsymbol{\theta})+\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\mu(\boldsymbol{\theta}) \\
&= \int_{[a,b]} \phi(\boldsymbol{\theta})\frac{\mathrm{d}\psi}{\mathrm{d}\mu}(\boldsymbol{\theta})\mu(\boldsymbol{\theta}) && \text{(E.23)}\\
&= \Big[\phi(\boldsymbol{\theta})\psi(\boldsymbol{\theta})\Big]_a^b - \int_{[a,b]} \frac{\mathrm{d}\phi}{\mathrm{d}\mu}(\boldsymbol{\theta})\psi(\boldsymbol{\theta})\,\mathrm{d}\mu(\boldsymbol{\theta}) && \text{(E.24)}\\
&= \Big[-\frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta})+\exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1}\Big]_a^b \\
&\quad + \int_{[a,b]} \frac{-2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta})-\exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1}\,\mathrm{d}\mu(\boldsymbol{\theta}), && \text{(E.25)}
\end{aligned}$$

where (E.25) follows the equalities (E.21) and (E.22). Substituting (E.25) into (E.20) yields

$$\begin{aligned}
&\int \frac{1}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \delta^{\star}_{Q,\boldsymbol{z}}}\,\mathrm{d}Q(\boldsymbol{\theta}) \\
&= 4\int_{[0,\infty)} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\mu(\boldsymbol{\theta}) && \text{(E.26)}\\
&= 4\int_{[0,1]} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\mu(\boldsymbol{\theta}) \\
&\quad + 4\int_{(1,\infty)} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\mu(\boldsymbol{\theta}) && \text{(E.27)}\\
&\geq 4\int_{[0,1]} \frac{\boldsymbol{\theta}^2\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\mu(\boldsymbol{\theta}) && \text{(E.28)}\\
&= 2\Bigg(\int_{[0,1]} \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta})+\exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1}\,\mathrm{d}\mu(\boldsymbol{\theta}) \\
&\quad + \int_{[0,1]} \frac{\exp(-2\boldsymbol{\theta})+\exp(-2\boldsymbol{\theta})}{(\boldsymbol{\theta}-1)^2}\,\mathrm{d}\mu(\boldsymbol{\theta})\Bigg) && \text{(E.29)}
\end{aligned}$$

$$
\begin{aligned}
&= 2\Bigg(\int_{[0,1]} \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1}\, \mathrm{d}\mu(\boldsymbol{\theta}) \\
&\quad + \Big[- \frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1} \Big]_0^1 \\
&\quad - \int_{[0,1]} \frac{2\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1}\, \mathrm{d}\mu(\boldsymbol{\theta}) \Bigg) && \text{(E.30)} \\
&= 2\Big[- \frac{\boldsymbol{\theta}\exp(-2\boldsymbol{\theta}) + \exp(-2\boldsymbol{\theta})}{\boldsymbol{\theta}-1} \Big]_0^1 && \text{(E.31)} \\
&= \infty, && \text{(E.32)}
\end{aligned}
$$

where (E.26) follows from the assumption that $\mathcal{M} = [0, \infty)$, (E.28) 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 (E.29) follows from (E.20), and (E.30) follows from substituting (E.25) into (E.29). From (E.32), it follows that the function $N_{Q,\boldsymbol{z}}$ in (4.7) is undefined at zero, which implies $\delta^\star_{Q,\boldsymbol{z}} \notin \mathcal{A}_{Q,\boldsymbol{z}}$, and this, $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$.

E.1.3 Example 3

Consider the ERM-fDR problem in (4.1) with reverse relative entropy (E.69) and assume that: (a) the set $\mathcal{B}$ is a proper subset of $\mathcal{M}$, and (b) the probability measure Q satisfies

$$
\begin{aligned}
Q(\mathcal{B}) &= \epsilon, \qquad \text{and} && \text{(E.33a)} \\
Q(\mathcal{M} \setminus \mathcal{B}) &= 1 - \epsilon, && \text{(E.33b)}
\end{aligned}
$$

with $\epsilon > 0$. Let the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (2.6) be

$$
\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{B} \\ c & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{B}, \end{cases} \tag{E.34}
$$

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 $\mathcal{B}_{Q,\boldsymbol{z}} = \left(-\delta^\star_{Q,\boldsymbol{z}}, \infty\right)$ and $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$. To show this, it is necessary to characterize the function $N_{Q,\boldsymbol{z}}$ in (4.3b). Hence, from the fact that the Lagrangian multiplier β for the optimization problem in (4.1) satisfies

$$
\int \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\, \mathrm{d}Q(\boldsymbol{\nu}) = 1, \tag{E.35}
$$

which follows from Theorem 4.3.1, the empirical risk function $\mathsf{L}_{\boldsymbol{z}} : \mathcal{M} \to \mathbb{R}_0^+$ in (E.34), which is a simple function, and the probability measure Q in (E.33a), it holds that

$$
\begin{aligned}
&\int \frac{\lambda}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}\, \mathrm{d}Q(\boldsymbol{\nu}) \\
&= \lambda \Bigg(\frac{1}{\beta + c_0} Q(\mathcal{T}(\boldsymbol{z})) + \frac{1}{\beta + c_1} Q(\mathcal{M} \setminus \mathcal{T}(\boldsymbol{z})) \Bigg) && \text{(E.36)} \\
&= \lambda \Bigg(\frac{1}{\beta + c_0} Q(\mathcal{T}(\boldsymbol{z})) + \frac{1}{\beta + c_1} (1 - Q(\mathcal{T}(\boldsymbol{z}))) \Bigg) && \text{(E.37)} \\
&= \lambda \Bigg(\frac{(\beta + c_1) Q(\mathcal{T}(\boldsymbol{z})) + (\beta + c_0)(1 - Q(\mathcal{T}(\boldsymbol{z})))}{\beta^2 + \beta(c_0 + c_1) + c_0 c_1} \Bigg) && \text{(E.38)}
\end{aligned}
$$

$$= \lambda\left(\frac{(c_1 - c_0)Q(\mathcal{T}(\boldsymbol{z})) + \beta + c_0}{\beta^2 + \beta(c_0 + c_1) + c_0 c_1}\right). \tag{E.39}$$

From (E.35) and (E.39), it follows that

$$0 = \beta^2 + \beta(c_0 + c_1) + c_0 c_1 - \lambda((c_1 - c_0)Q(\mathcal{T}(\boldsymbol{z})) + \beta + c_0) \tag{E.40}$$
$$= \beta^2 + \beta(c_0 + c_1 - \lambda) + c_0 c_1 - \lambda c_0 - \lambda(c_1 - c_0)Q(\mathcal{T}(\boldsymbol{z})). \tag{E.41}$$

From (E.41) and the fact that $c_0 = 0$ in equation (E.34), it holds that

$$0 = \beta^2 + \beta(c_1 - \lambda) - \lambda c_1 Q(\mathcal{T}(\boldsymbol{z})). \tag{E.42}$$

Observe that the expression in (E.42) is a quadratic polynomial that has two roots r_1 and r_2. Hence, (E.42) in terms of r_1 and r_2 satisfies

$$0 = \beta^2 - (r_1 + r_2)\beta + r_1 r_2 \tag{E.43}$$
$$= (\beta - r_1)(\beta - r_2), \tag{E.44}$$

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

$$r_1 = -\frac{(c_1 - \lambda)}{2} - \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}, \tag{E.45}$$

and

$$r_2 = -\frac{(c_1 - \lambda)}{2} + \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}. \tag{E.46}$$

The proof continues by verifying that the roots in (E.45) and (E.46) 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

$$-\frac{c_1 - \lambda}{2} \leq \left|\frac{c_1 - \lambda}{2}\right| \tag{E.47}$$
$$= \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2} \tag{E.48}$$
$$\leq \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}. \tag{E.49}$$

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 (E.45) and (E.46) are real, which implies that r_1 and r_2 are real. From (E.49), for all $\lambda \in [0, +\infty)$ and for all $Q(\mathcal{T}(\boldsymbol{z})) \in [0, 1)$, it holds

$$r_1 < 0; \tag{E.50}$$

and following the same arguments

$$r_2 > 0. \tag{E.51}$$

Hence, the solution for the Lagrange Multiplier β that satisfies (E.35) given the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (E.34) and the probability measure Q in (E.33a) is

$$\beta = -\frac{(c_1 - \lambda)}{2} + \sqrt{\left(\frac{c_1 - \lambda}{2}\right)^2 + \lambda c_1 Q(\mathcal{T}(\boldsymbol{z}))}, \tag{E.52}$$

which implies that the function $N_{Q,\boldsymbol{z}}$ in (4.3b) under the current assumptions in (E.33) and (E.34) satisfies

$$N_{Q,\boldsymbol{z}}(\lambda) = -\frac{(c - \lambda)}{2} + \sqrt{\left(\frac{c - \lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})}. \tag{E.53}$$

The proof of equality (E.53) is presented in appendix E.1.3.

From Theorem 4.3.1, it follows that $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (E.71) satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\begin{aligned} &\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) \\ &= \frac{\lambda}{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \frac{(c-\lambda)}{2} + \sqrt{\left(\frac{c-\lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})}}. \end{aligned} \tag{E.54}$$

Under the current assumptions, from Lemma 4.3.3, it is sufficient to show that for all $c \in (0, \infty)$ and $\lambda \in (0, \infty)$ in (E.34), the function $N_{Q,\boldsymbol{z}}$ in (4.3b) is strictly greater than $-\delta^\star_{Q,\boldsymbol{z}}$. From (E.53), it holds that

$$\begin{aligned} N_{Q,\boldsymbol{z}}(\lambda) &= -\frac{(c - \lambda)}{2} + \sqrt{\left(\frac{c - \lambda}{2}\right)^2 + \lambda c Q(\mathcal{B})} && \text{(E.55)} \\ &> -\frac{(c - \lambda)}{2} + \sqrt{\left(\frac{c - \lambda}{2}\right)^2} && \text{(E.56)} \\ &= -\frac{(c - \lambda)}{2} + \left|\frac{c - \lambda}{2}\right| && \text{(E.57)} \\ &\geq 0 && \text{(E.58)} \\ &= -\delta^\star_{Q,\boldsymbol{z}}, && \text{(E.59)} \end{aligned}$$

which proves that for all $c \in (0, \infty)$ and for all $\lambda \in (0, \infty)$, it holds that $N_{Q,\boldsymbol{z}}(\lambda) > -\delta^\star_{Q,\boldsymbol{z}}$ which implies that $-\delta^\star_{Q,\boldsymbol{z}} \notin \mathcal{B}_{Q,\boldsymbol{z}}$ with the set $\mathcal{B}_{Q,\boldsymbol{z}}$ defined in (4.7) and thus $\mathcal{A}_{Q,\boldsymbol{z}} = (0, \infty)$.

E.2 ERM-fDR Table Derivation

This section presents the solutions to the optimization problem in (4.1) for specific choices of the function f.

E.2.1 Relative Entropy

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that $f(x) = x \log(x)$. The derivative of the function f satisfies

$$\dot{f}(x) = 1 + \log(x). \tag{E.60}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is the relative entropy of P with respect to Q, also known as Kullback-Leibler divergence. That is,

$$\mathsf{D}_f(P\|Q) = \mathsf{D}(P\|Q), \tag{E.61}$$

with D defined in (1.2). From (E.60) and Theorem 4.3.1, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right), \tag{E.62}$$

where β can be obtained explicitly using (4.3b), which yields

$$1 = \int \exp\left(-\frac{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}) \tag{E.63}$$

$$= \exp\left(-\frac{\beta+\lambda}{\lambda}\right) \int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{E.64}$$

which implies that

$$\beta = -\lambda + \lambda \log\left(\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta})\right). \tag{E.65}$$

Plugging (E.65) into (E.62) yields

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

This result has been independently reported by several authors in [84,92,108,113,115], and proved via a large variety of methods.

E.2.2 Reverse Relative Entropy

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that $f(x) = -\log(x)$, and note that the derivative of the function f satisfies

$$\dot{f}(x) = -\frac{1}{x}. \tag{E.67}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is the relative entropy of Q with respect to P. That is,

$$\mathsf{D}_f(P\|Q) = \mathsf{D}(Q\|P), \tag{E.68}$$

where D defined in (1.2). This result, in contrast to the previous example, justifies referring to $\mathsf{D}_f(P\|Q)$ as the *reverse relative entropy*. From (E.67) and Theorem 4.3.1, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda}{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}. \tag{E.69}$$

While a closed-form expression for β in (E.69) is unknown, the regularization factor λ and β satisfy the following

$$\lambda = \frac{1}{\int \frac{1}{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})} \, \mathrm{d}Q(\boldsymbol{\theta})}, \tag{E.70}$$

which follows from (4.3b). Note that (E.70) establishes a one-to-one relation between λ and β. This observation leads to the following

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) + \beta)^{-1}}{\displaystyle\int (\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \beta)^{-1} \, \mathrm{d}Q(\boldsymbol{\nu})}, \tag{E.71}$$

which follows from plugging (E.70) into (E.69). This result has been previously reported in [119].

E.2.3 Jeffreys Divergence

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that $f(x) = x \log(x) - \log(x)$ and note that the derivative of the function f satisfies

$$\dot{f}(x) = \log(x) + 1 - x^{-1}. \tag{E.72}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is Jeffreys divergence between P and Q, also known as symmetrized relative entropy or symmetrized Kullback-Leibler divergence, which follows from observing that

$$\mathsf{D}_f(P\|Q) = \mathsf{D}(P\|Q) + \mathsf{D}(Q\|P), \tag{E.73}$$

with D defined in (1.2). From (E.72) and Theorem 4.3.1, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(W_0\left(\exp\left(\frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right) - \frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right), \tag{E.74}$$

where the function $W_0 : [0, \infty) \to [0, \infty)$ is the Lambert function, which for a function $g : \mathbb{R} \to \mathbb{R}$ such that $g(x) = x \exp(x)$ satisfies $W_0(g(x)) = x$. The coupling induced by the Lambert function in (E.74) between the parameters β and λ prevents the characterization of β in closed-form. Hence, β must be obtained via numerical methods using

$$1 = \int \exp\left(W_0\left(\exp\left(\frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\right) - \frac{\beta + \lambda + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{E.75}$$

which follows from (4.3b).

E.2.4 Jensen-Shannon Divergence

Let the function $f : (0, +\infty) \to \mathbb{R}$ be such that $f(x) = x \log\left(\frac{2x}{x+1}\right) + \log\left(\frac{2}{x+1}\right)$. The derivative of the function f satisfies

$$\dot{f}(x) = \log(2x) - \log(x + 1). \tag{E.76}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is the Jensen-Shannon's divergence between P and Q, and similarly to the Jeffreys divergence, it is also symmetric, which follows from observing that

$$\mathsf{D}_f(P\|Q) = \mathsf{D}\left(P\|\frac{P+Q}{2}\right) + \mathsf{D}\left(Q\|\frac{P+Q}{2}\right), \tag{E.77}$$

with D defined in (1.2). From (E.76) and Theorem 4.3.1, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{1}{2\exp\left(\frac{\beta+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)-1}, \tag{E.78}$$

where $\beta = -\lambda\log(2c)$, with $c > 0$ that satisfies the fixed-point equation

$$c = \frac{1}{\displaystyle\int \frac{1}{\exp\left(\frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)-c}\mathrm{d}Q(\boldsymbol{\theta})}, \tag{E.79}$$

which follows from (4.3b).

E.2.5 Hellinger Divergence

Let the function $f : (0,+\infty) \to \mathbb{R}$ be such that $f(x) = (1-\sqrt{x})^2$. Note that the derivative of the function f satisfies

$$\dot{f}(x) = 1 - \frac{1}{\sqrt{x}}. \tag{E.80}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is Hellinger's divergence of P with respect to Q, and similarly to (E.73) and (E.77), it is also symmetric, which follows from observing that

$$\mathsf{D}_f(P\|Q) = \int\left(1-\sqrt{\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})}\right)^2 \mathrm{d}Q(\boldsymbol{\theta}). \tag{E.81}$$

From (E.80) and Theorem 4.3.1, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \left(\frac{\lambda}{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right)^2, \tag{E.82}$$

where β is chosen to satisfy

$$\int\left(\frac{\lambda}{\beta+\lambda+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}\right)^2 \mathrm{d}Q(\boldsymbol{\theta}) = 1, \tag{E.83}$$

which follows from (4.3b).

E.2.6 χ^2 Divergence

Let the function $f : (0,\infty) \to \mathbb{R}$ be such that $f(x) = x^2 - 1$, whose derivative satisfies

$$\dot{f}(x) = 2x. \tag{E.84}$$

In this case, the resulting f-divergence $\mathsf{D}_f(P\|Q)$ is the Pearson-divergence, also known as the χ^2-divergence between P and Q. From (E.84) and Theorem 4.3.1, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = -\frac{\beta + \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{2\lambda}, \tag{E.85}$$

where β satisfies (4.3b), which implies

$$1 = -\int \left(\frac{\beta}{2\lambda} + \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{2\lambda} \right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{E.86}$$

and thus,

$$\beta = -(\lambda + \mathsf{R}_{\boldsymbol{z}}(Q)). \tag{E.87}$$

Plugging (E.87) into (E.85) yields

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\lambda + \mathsf{R}_{\boldsymbol{z}}(Q) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{2\lambda}. \tag{E.88}$$

E.3 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 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 the matrix $I \in [0,1]^{28\times 28}$.

E.3.1 Features extraction of the Histogram of Oriented Gradients

The grayscale images are processed by calculating their corresponding *histogram of oriented gradients* (HOG) [147]. 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

$$\mathsf{G}_w(i,j) = \begin{cases} I(i+1,j) - I(i-1,j) & \text{if } i \in \{2,\dots,27\} \\ I(i+1,j) - I(i,j) & \text{if } i = 1 \\ I(i,j) - I(i-1,j) & \text{if } i = 28 \end{cases}, \tag{E.89}$$

and

$$\mathsf{G}_h(i,j) = \begin{cases} I(i,j+1) - I(i,j-1) & \text{if } j \in \{2,\dots,27\} \\ I(i,j+1) - I(i,j) & \text{if } j = 1 \\ I(i,j) - I(i,j-1) & \text{if } j = 28 \end{cases}, \tag{E.90}$$

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

2.) Given a pixel loacation $(i,j) \in \{1,2,...,28\}^2$, 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 ϕ :

$\{1, 2, ..., 28\}^2 \to \mathbb{R}$, such that

$$\mathsf{M}(i,j) = \sqrt{\mathsf{G}_w(i,j)^2 + \mathsf{G}_h(i,j)^2}, \text{ and} \tag{E.91}$$

$$\phi(i,j) = \arctan\left(\frac{\mathsf{G}_h(i,j)}{\mathsf{G}_w(i,j)}\right). \tag{E.92}$$

3.) The matrix I is divided into sub-matrices of size 4×4, such that the number of sub-matrices along the *width* and *height* are:

$$N_w = \frac{28}{4} = 7, \text{ and} \tag{E.93}$$

$$N_h = \frac{28}{4} = 7, \tag{E.94}$$

where N_w represents the number of sub-matrices along the width, and N_h represents the number of sub-matrices along the height. These sub-matrices, with $w \in \{1, \cdots, N_w\}$ and $h \in \{1, \cdots, N_h\}$, are referred to as *cells*, and denoted by

$$C_{w,h} = \begin{bmatrix} I(a_w, b_h) & \cdots & I(a_w + 3, b_h) \\ \vdots & \ddots & \vdots \\ I(a_w, b_h + 3) & \cdots & I(a_w + 3, b_h + 3) \end{bmatrix}, \tag{E.95}$$

where the real values a_w and b_h are

$$a_w = 4(w-1) + 1 \tag{E.96}$$

$$b_b = 4(h-1) + 1. \tag{E.97}$$

This implies that the matrix I can be represented as

$$I = \begin{bmatrix} C_{1,1} & \cdots & C_{N_w,1} \\ \vdots & \ddots & \vdots \\ C_{1,N_h} & \cdots & C_{N_w,N_h} \end{bmatrix}. \tag{E.98}$$

From (E.95), the set of all pairs (i, j) of pixel coordinates in I that lie within the cell $C_{w,h}$ is given by:

$$\mathcal{A}_{w,h} = \{a_w, a_w + 3\} \times \{b_h, b_h + 3\}, \tag{E.99}$$

with a_w in (E.96) and b_h in (E.97).

4.) For each cell $C_{w,h}$ in (E.95) the orientations $\phi(i, j)$ in (E.92) are divided into k bins, with $k \in \mathbb{N}$. That is, the n^{th} bin, with $1 \leq n \leq k$, satisfies that

$$\mathcal{B}_{w,h}^{(n)} = \left\{ \phi(i,j) \in \mathbb{R} : 180\left(\frac{n-1}{k}\right) \leq \phi(i,j) < 180\left(\frac{n}{k}\right) : (i,j) \in \mathcal{A}_{w,h} \right\} \tag{E.100}$$

Note that for the simulations, the parameter is set to $k = 9$. 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 (E.95) is given by the function $H_{w,h}(n) : \{1, 2, \ldots, k\} \to \mathbb{R}$, such that

$$H_{w,h}(n) = \sum_{(i,j) \in \mathcal{A}_{w,h}} \mathsf{M}(i,j) \mathbb{1}_{\left\{\phi(i,j) \in \mathcal{B}_{w,h}^{(n)}\right\}}, \tag{E.101}$$

with M in (E.91); and $\mathcal{B}_n$ in (E.100). Thus the histogram of gradient orientations of the cell $C_{w,h}$ is represented by the vector $\boldsymbol{H}_{w,h} \in \mathbb{R}^k$, such that

$$\boldsymbol{H}_{w,h} = [H_{w,h}(1), H_{w,h}(2), \cdots H_{w,h}(k)], \tag{E.102}$$

with the function $H_{w,h}$ in (E.101).

5.) To account for illumination and contrast variations, the histogram $\boldsymbol{H}_{w,h}$ in (E.102) is normalized. To normalize the histograms for all cells $C_{w,h}$ in (E.95), the cells are grouped into sub-matrices formed by 2×2 cells with a *cell overlap* denoted by $o \in \mathbb{N}$. For the simulations, the overlap is set to $o = 1$, such that number of sub-matrices is:

$$\begin{aligned} N_t &= (N_w - o) \times (N_h - o) && \text{(E.103a)} \\ &= (7 - 1) \times (7 - 1) && \text{(E.103b)} \\ &= 36, && \text{(E.103c)} \end{aligned}$$

with N_w in (E.93) and N_h in (E.94). These sub-matrices of the matrix I in (E.98), with $(m, s) \in \left\{1, \cdots \sqrt{N_t}\right\}^2$ are referred to as *blocks*, and denoted by

$$B_{m,s} = \begin{bmatrix} C_{m,s} & C_{m+1,s} \\ C_{m,s+1} & C_{m+1,s+1} \end{bmatrix}, \tag{E.104}$$

with $C_{m,s}$ in (E.95). From (E.98) and (E.104), 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, denoted by B, is given by the ratio of the total number of pixels in a block to the number of pixels in a cell:

$$\begin{aligned} B &= \frac{N_b}{N_c} && \text{(E.105a)} \\ &= \frac{8 \times 8}{4 \times 4} && \text{(E.105b)} \\ &= 4, && \text{(E.105c)} \end{aligned}$$

where N_b is the number of pixels in a block and $N_c$is the number of pixels in a cell. The normalized histogram of a cell $C_{w,h}$ in a block $B_{m,s}$ is denoted by the vector $\hat{\boldsymbol{H}}_{w,h}^{(m,s)} \in \mathbb{R}^k$. This normalization of is typically done using the L2-norm, such that

$$\hat{\boldsymbol{H}}_{w,h}^{(m,s)} = \frac{\boldsymbol{H}_{w,h}}{\sqrt{\sum_{(i,j)\in\{m,m+1\}\times\{s,s+1\}} \boldsymbol{H}_{i,j}^2 + \epsilon^2}}, \tag{E.106}$$

where $\boldsymbol{H}_{w,h}$ in (E.102) is the unnormalized histogram, and the $\epsilon > 0$ to avoid division by zero.

6.) For an image with 36 blocks (see (E.103)), 9 orientation bins, and a size of block 4 (see (E.105)), the size $l \in \mathbb{N}$ of the HOG feature vector $\hat{\mathbf{x}} \in \mathbb{R}^l$ is:

$$\begin{aligned} l &= N_t \times B \times k && \text{(E.107a)} \\ &= 36 \times 4 \times 9 && \text{(E.107b)} \\ &= 1296. && \text{(E.107c)} \end{aligned}$$

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

$$\begin{aligned}\hat{\mathbf{x}} = \Big[&\hat{\boldsymbol{H}}_{1,1}^{(1,1)}, \hat{\boldsymbol{H}}_{1,2}^{(1,1)}, \hat{\boldsymbol{H}}_{2,1}^{(1,1)}, \hat{\boldsymbol{H}}_{2,2}^{(1,1)}, \\ &\hat{\boldsymbol{H}}_{2,1}^{(2,1)}, \hat{\boldsymbol{H}}_{2,2}^{(2,1)}, \hat{\boldsymbol{H}}_{3,1}^{(2,1)}, \hat{\boldsymbol{H}}_{3,2}^{(2,1)}, \cdots, \\ &\hat{\boldsymbol{H}}_{6,6}^{(6,6)}, \hat{\boldsymbol{H}}_{6,7}^{(6,6)}, \hat{\boldsymbol{H}}_{7,6}^{(6,6)}, \hat{\boldsymbol{H}}_{7,7}^{(6,6)}\Big]^\top. \end{aligned} \tag{E.108}$$

E.3.2 Principal Component Analysis

The final step in the data processing is to reduce the dimensionality of the pattern $\hat{\mathbf{x}}$ in (E.108) from $\mathbb{R}^l$, with l in (E.107) 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 E.3.1. The resulting 12,183 HOG vectors $\hat{\mathbf{x}} \in \mathbb{R}^l$, with l in (E.107) and $\hat{\mathbf{x}}$ in (E.108) are reduced to $\mathbb{R}^2$ using PCA in the simulation 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:

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\mathbf{x}}_i - \boldsymbol{\mu})(\hat{\mathbf{x}}_i - \boldsymbol{\mu})^\top, \tag{E.109}$$

where $n = 12{,}183$ and $\mathbf{C} \in \mathbb{R}^{l\times l}$, with l in (E.107), and $\boldsymbol{\mu}$ is the mean of all the training patterns given by

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i. \tag{E.110}$$

2.) The next step in PCA is to perform an *eigenvalue decomposition* of the covariance matrix $\mathbf{C}$ in (E.109). The decomposition can be written as:

$$\mathbf{C} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top, \tag{E.111}$$

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

3.) Following the computation of the eigenvectors, the dimensionality is reduced from $\mathbb{R}^l$ 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 $\mathbf{W} \in \mathbb{R}^{l\times 2}$, defined as

$$\mathbf{W} = [\mathbf{w}_1\, \mathbf{w}_2]. \tag{E.112}$$

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

$$\mathbf{x} = \mathbf{W}^\top \hat{\mathbf{x}}, \tag{E.113}$$

with $\hat{\mathbf{x}}$ in (E.108), $\boldsymbol{W}$ in (E.112) and $\mathbf{x} \in \mathbb{R}^2$ is the 2-dimensional coordinates of the original pattern $\hat{\mathbf{x}}$ in the reduced-dimensional space.

E.3.3 Simulation Dataset

In this simulation, a datapoint is a tuple $(\hat{\mathbf{x}}, y) \in \mathbb{R}^l \times \{6, 7\}$, with $\hat{\mathbf{x}}$ in (E.108) and y being the label assigned by MNIST to the image I in (E.98). 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.

Bibliography

[1] R. A. Fisher, "Theory of statistical estimation," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 22, no. 5, pp. 700–725, Nov. 1925.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 2006.

[3] R. A. Meyers, *Encyclopedia of physical science and technology*, 3rd ed. Tarzana, CA, USA: Academic, 2002.

[4] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *Journal of Machine Learning Research*, vol. 12, no. 3, pp. 1069–1109, Mar 2011.

[5] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, "Entropy-SGD: Biasing gradient descent into wide valleys," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124018, Dec. 2019.

[6] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, Jun. 2007.

[7] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, Sep. 1998.

[8] A. Sinha, M. O'Kelly, R. Tedrake, and J. C. Duchi, "Neural bridge sampling for evaluating safety-critical autonomous systems," *Advances in Neural Information Processing Systems*, vol. 33, no. 1, pp. 6402–6416, Dec. 2020.

[9] E. Parliament, "EU AI act: first regulation on artificial intelligence," pp. 1–5, 2023.

[10] D. Wang, M. Ye, and J. Xu, "Differentially private empirical risk minimization revisited: Faster and more general," *Advances in Neural Information Processing Systems*, vol. 30, no. 1, Dec. 2017.

[11] 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.

[12] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularizations," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9454, Feb. 2022.

[13] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–10, Dec. 2017.

[14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, Dec. 2013.

[15] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," arXiv preprint arXiv:2106.11342, Jun. 2021.

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

[17] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, Jan. 2020.

[18] S. M. Perlaza and X. Zou, "The generalization error of machine learning algorithms," *Submitted to IEEE Transactions on Information Theory*, pp. 1–1, Nov. 2024.

[19] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Asymmetry of the relative entropy in the regularization of empirical risk minimization," *Submitted to IEEE Transactions on Information Theory*, pp. 1–1, Oct. 2024.

[20] ——, "Equivalence of empirical risk minimization to regularization on the family of f-divergences," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, Jul. 2024.

[21] F. Daunas, I. Esnaola, and S. M. Perlaza, "A dual optimization view to empirical risk minimization with f-divergence regularization," in *Submitted to the IEEE Information Theory Workshop (ITW)*, Sydney, Australia, Oct. 2025.

[22] F. Daunas, I. Esnaola, S. M. Perlaza, and G. Aminian, "Generalization error of f-divergence stabilized algorithms via duality," arXiv preprint arXiv:2502.14544, Feb. 2025.

[23] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, no. 1, pp. 299–318, Jun. 1967.

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

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

[26] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2014.

[27] S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900*, 1st ed. Cambridge, MA, USA: Harvard University Press, 1986.

[28] C. E. Shannon, "Presentation of a maze-solving machine. cybernetics: Circular, casual, and feedback mechanisms in biological and social systems,"

in *Proceedings of the Transactions Eighth Conference Cybernetics*, New York, NY, USA, Mar. 1952, pp. 169–181.

[29] A. L. Samuel, "Some studies in machine learning using the game of checkers," *Improving the Efficiency of a Problem Solver*, vol. 62, no. 1, pp. 535–554, Jun. 1959.

[30] J. N. Morgan and J. A. Sonquist, "Some results from a non-symmetrical branching process that looks for interaction effects," *Young*, vol. 8, no. 5, pp. 40–53, Dec. 1963.

[31] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, vol. 222, no. 594-604, pp. 309–368, Jan. 1922.

[32] J. Neyman and E. Pearson, "Biometrika trust," *Biometrika*, vol. 20, no. 1/2, pp. 175–240, Jul. 1928.

[33] R. Nock and P. Jappy, "Decision tree based induction of decision lists," *Intelligent Data Analysis*, vol. 3, no. 3, pp. 227–240, Jan. 1999.

[34] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4.5," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 13–19, Feb. 2014.

[35] M. Milanović and M. Stamenković, "Chaid decision tree: Methodological frame and application," *Economic Themes*, vol. 54, no. 4, pp. 563–586, 2016.

[36] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[37] R. M. Corless, D. J. Jeffrey, and D. E. Knuth, "A sequence of series for the Lambert w function," in *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, Maui, HI, USA, Jul. 1997, pp. 197–204.

[38] F. Rosenblatt, *The perceptron, a perceiving and recognizing automaton Project Para*, 1st ed. Cornell Aeronautical Laboratory, 1957.

[39] ——, "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms," *Springer*, vol. 3, no. 1, pp. 218–219, Jan. 1962.

[40] M. Minsky and S. Papert, "Perceptrons: An introduction to computational geometry," *MIT Press*, vol. 1, no. 1, 1969.

[41] V. Vapnik and A. Chervonenkis, *Theory of Pattern Recognition*, 1st ed. Moscow: Nauka, 1974.

[42] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational Learning Theory (COLT)*, Pittsburgh, PA, USA, Jul. 1992, pp. 144–152.

[43] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[44] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Stanford Univ Ca Stanford Electronics Labs, Stanford, CA, USA, Tech. Rep. AD241531, Jun. 1960.

[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[46] S. Arora, D. Bhattacharjee, M. Nasipuri, L. Malik, M. Kundu, and D. K. Basu, "Performance comparison of SVM and ANN for handwritten Devanagari character recognition," arXiv preprint arXiv:1006.5902, Jun. 2010.

[47] O. Catoni, *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, 1st ed. Beachwood, OH, USA: Institute of Mathematical Statistics Lecture Notes - Monograph Series, 2007, vol. 56.

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

[49] J. MacQueen, "Multivariate observations," in *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, vol. 1, Berkeley, CA, USA, Jun. 1967, pp. 281–297.

[50] K. Pearson, "Principal components analysis," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 6, no. 2, pp. 559–565, Jan. 2008.

[51] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, Sep. 1933.

[52] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, "Unsupervised semantic segmentation by distilling feature correspondences," arXiv preprint arXiv:2203.08414, Mar. 2022.

[53] M. N. Yein and M. F. Amasyal, "Generative diffusion models: A survey of current theoretical developments," Aug. 2024.

[54] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, Mar. 2009.

[55] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*, 1st ed. Cambridge, MA, USA: MIT press Cambridge, 1998.

[56] R. Bellman and R. Kalaba, "Dynamic programming and statistical communication theory," *Proceedings of the National Academy of Sciences*, vol. 43, no. 8, pp. 749–751, May. 1957.

[57] L. Breiman, "Bagging predictors," Aug. 1996.

[58] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[59] D. H. Wolpert, "Stacked generalization," Aug. 1992.

[60] T. M. Mitchell, *The Need for Biases in Learning Generalizations*, 1st ed. New Brunswick, NJ, USA: Department of Computer Science, Laboratory for Computer Science Research, 1980.

[61] D. F. Gordon and M. Desjardins, "Evaluation and selection of biases in machine learning," *Machine Learning*, vol. 20, no. 1, pp. 5–22, Jul. 1995.

[62] R. Kent, *Data Construction and Data Analysis for Survey Research*, 3rd ed. London, UK: Bloomsbury Publishing, 2020.

[63] J. Neyman and E. S. Pearson, "IX. On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, no. 694-706, pp. 289–337, Feb. 1933.

[64] J. Wishart and M. Bartlett, "The distribution of second order moment statistics in a normal system," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 28, no. 4, pp. 455–459, Nov. 1932.

[65] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*, 1st ed. Cambridge, MA, USA: MIT press, 1964.

[66] R. E. Kalman, "On the general theory of control systems," in *Proceedings of the First International Conference on Automatic Control*, Moscow, USSR, Jun. 1960, pp. 481–492.

[67] D. V. Lindley, *Inference*, 1st ed. Cambridge, UK: Cambridge University Press, 1965.

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

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

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

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

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

[73] 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.

[74] 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.

[75] 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.

[76] 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.

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

[78] 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.

[79] 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.

[80] C. E. Shannon, *The mathematical theory of communication, by CE Shannon (and recent contributions to the mathematical theory of communication), W. Weaver*, 1st ed. Champaign, IL, USA: University of Illinois Press, 1949.

[81] H. Cramér, *Mathematical methods of statistics*, 1st ed. Princeton, NJ, USA: Princeton university press, 1999.

[82] S. Kullback, *Information theory and statistics*, 1st ed. New York, NY, USA: Dover, 1978.

[83] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis," in *Proceedings of the Conference on Learning Theory (COLT)*, Amsterdam, Netherlands, Jun. 2017, pp. 1674–1703.

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

[85] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees.*, 1st ed. New York, NY, USA: Chapman and Hall, 1984.

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

[87] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114v10, May. 2014.

[88] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the Allerton Conference on Communication, Control and Computing*, Vancouver, Canada, Apr. 1999, pp. 368–377.

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

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

[91] 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.

[92] 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.

[93] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes: with applications to statistics*, 1st ed. New York, NY, USA: Springer, 1996.

[94] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, Jul. 2001.

[95] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, Nov. 2002.

[96] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in Neural Information Processing Systems*, vol. 4, pp. 950–957, Dec. 1991.

[97] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," arXiv preprint arXiv:1611.03530, Feb. 2017.

[98] 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.

[99] 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.

[100] 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.

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

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

[103] 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.

[104] 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.

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

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematics Doklady*, vol. 4, no. 6, pp. 1035–1038, Dec. 1963.

[111] A. E. Horel, "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, vol. 58, no. 1, pp. 54–59, Jun. 1962.

[112] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, "The worst-case data-generating probability measure," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9515, Aug. 2023.

[113] 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.

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

[115] 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.

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

[117] 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.

[118] 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.

[119] 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.

[120] 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.

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

[122] 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.

[123] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "On the generalization for transfer learning: An information-theoretic analysis," *Submitted to IEEE Transactions on Information Theory*, pp. 1–1, early access, Aug. 14, 2024.

[124] B. Rodríguez Gálvez, "An information-theoretic approach to generalization theory," PhD thesis, KTH Royal Institute of Technology, Example City, CA, Jun. 2024, available at https://example.com/thesis.pdf.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

[131] X. Zou, S. M. Perlaza, I. Esnaola, E. Altman, and H. V. Poor, "An exact characterization of the generalization error of machine learning algorithms," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9539, Jan. 2024.

[132] 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.

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

[134] A. Rényi *et al.*, "On measures of information and entropy," in *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, Berkeley, CA, USA, Jun. 1961, pp. 547–561.

[135] I. Sason and S. Verdú, "f -divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, Jun. 2016.

[136] L. Martino, V. Elvira, and G. Camps-Valls, "The recycling Gibbs sampler for efficient learning," *Digital Signal Processing*, vol. 74, no. 1, pp. 1–13, Aug. 2018.

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

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

[139] R. Gateaux, "Sur les fonctionnelles continues et les fonctionnelles analytiques," *Comptes rendus hebdomadaires des séances de l'Académie des Sciences, Paris*, vol. 157, no. 325-327, p. 65, 1913.

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

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

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

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

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

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

[146] W. Rudin, *Principles of Mathematical Analysis*, 1st ed. New York, NY, USA: McGraw-Hill Book Company, Inc., 1953.

[147] 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.