Empirical Risk Minimization with Generalized Relative Entropy Regularization

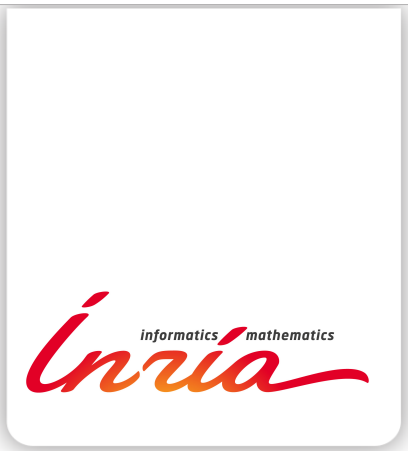
Samir M. Perlaza, Gaetan Bisson, Iñaki Esnaola, Alain Jean-Marie, Stefano Rini

Empirical Risk Minimization with Generalized Relative Entropy Regularization

Samir M. Perlaza[1,2,3], **Gaetan Bisson**[2], **Iñaki Esnaola**[3,4], **Alain Jean-Marie**[1], **and Stefano Rini**[5]

RESEARCH REPORT
N° 9454
February 7, 2022
Project-Team NEO

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



Empirical Risk Minimization with Generalized Relative Entropy Regularization

Samir M. Perlaza[(1,2,3)], Gaetan Bisson[(2)], Iñaki Esnaola[(3,4)], Alain Jean-Marie[(1)], and Stefano Rini[(5)]

Project-Team NEO

Research Report n° 9454 — February 7, 2022 — 66 pages

Abstract: The empirical risk minimization problem with relative entropy regularization (ERM-RER) is investigated considering that the reference measure is a σ-finite measure instead of a probability measure. This generalization allows for a larger degree of flexibility in the incorporation of prior knowledge over the set of models. In this setting, the interplay of the regularization parameter, the reference measure, the risk measure, and the expected empirical risk induced by the solution of the ERM-RER problem, which is proved to be unique, is characterized. For a fixed dataset, the empirical risk is shown to be a sub-Gaussian random variable, when the models follow the probability measure that is the solution to the ERM-RER problem. The sensitivity of the expected empirical risk to deviations from the solution of the ERM-RER problem is studied and upper and lower bounds on the expected empirical risk are provided. Finally, it is shown that the expectation of the sensitivity is upper bounded, up to a constant factor, by the square root of the lautum information between the models and the datasets.

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

Most of this work was done while Samir M. Perlaza was visiting the GAATI Laboratory of the University of French Polynesia between May 3 and October 31, 2021.

(1) INRIA, Centre de Recherche de Sophia Antipolis, Sophia Antipolis, France.
(2) GAATI Laboratory, University of French Polynesia, BP 6570, 98702 Faaa, French Polynesia.
(3) Department of Electrical and Computer Engineering, Princeton University, 08544 Princeton, N.J.
(4) Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, United Kingdom.
(5) Department of Electrical and Computer Engineering, National Chao Tung University. Hsinchu, Taiwan.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Minimisation du Risque Empirique avec Régularisation par l'Entropie Relative Généralisée

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

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

Contents

1 Introduction

The problem of empirical risk minimization (ERM), also known as M-estimation [42], minimum contrast estimation [4,29], sample average approximation [26], among other appellations, appears in numerous central problems in machine learning [45], information theory [32], statistics [47,48], and operations research [10, 13], etc. Depending on the field, the ERM problem might take different forms and certain mathematical objects might take different names. In the following, ERM and the main results in this work, are presented in the context of supervised machine learning using the notation and nomenclature of this application field.

The empirical risk minimization (ERM) problem with relative entropy regularization (ERM-RER) has been the workhorse for building probability measures on the set of models, without any additional assumption on the statistical description of the datasets and/or the models [2, 10, 51]. Instead of additional statistical assumptions, which is typical in Bayesian methods [36], relative entropy regularization requires a reference probability measure, which is external to the ERM problem. Often, such reference represents prior knowledge and is chosen to assign high probability to the models that induce low empirical risks. The ERM-RER problem is known to possess a unique solution, which is a Gibbs probability measure and has been studied using information theoretic notions in [39,50,53]; statistical physics [10]; PAC (Probably Approximatively Correct)-Bayesian learning theory [19, 20, 31, 41]; and proved to be of particular interest in classification problems in [22, 54].

In this report, the ERM with relative entropy regularization (ERM-RER) is generalized to incorporate a σ-finite measure with arbitrary support as the reference measure. That is, in this setting, the prior knowledge incorporated by the reference measure need not contain in its support the set of solutions of the ERM problem without regularization. The flexibility introduced by this formulation becomes particularly relevant for the case in which priors are available in the form of probability distributions that can be evaluated up to some normalizing factor, as is often the case in practical scenarios [37]. Moreover, as shown below, for some specific choices of the reference measure, the ERM-RER boils down to particular cases of special interest: the information-risk minimization problem, the ERM with differential entropy regularization; and the ERM with discrete entropy regularization. Hence, the proposed formulation yields a unified mathematical framework that comprises a large class of problems. Nonetheless, despite its general character, the ERM-RER with an arbitrary σ-finite reference measure is shown to possess a unique solution. Indeed, the solution is a Gibbs probability measure whose partition function is defined with respect to the σ-finite reference measure. More importantly, the solution does not depend on the probability distribution of the dataset and can be computed for a particular dataset realization.

Finally, using the mathematical framework described above, the sensitivity of the ERM-RER problem is studied at a given dataset. The sensitivity is defined as the absolute value of the difference between two quantities: (a) The expectation of the empirical risk at a given dataset with respect to the measure that is the solution to the ERM-RER problem; and (b) the expectation of the empirical risk at such dataset with respect to another measure. The sensitivity is upper bounded by a term that is proportional to the squared-root of the relative entropy of the alternative measure with respect to the mesure that is a solution to the ERM problem. More interestingly, the expectation of the sensitivity with respect to the probability distribution of the data sets turns out to be bounded by a term that is proportional to the squared-root of the lautum information between the models and the datasets. This bound is reminiscent to the result in [50] in which, under certain conditions, the generalization gap is upper bounded by a term that is proportional to the squared-root of the mutual information between the models and

the datasets. Together, the existing characterisations of the generalization gap and the proposed characterization of the sensitivity provide a solid theoretical framework to study the trade-off between generalization error and empirical risk.

1.1 Notation

In this work, sets are denoted by calligraphic letters. Given a set $\mathcal{A}$, the notation $\mathscr{F}(\mathcal{A})$ represents a sigma-field (σ-field) on $\mathcal{A}$. When $\mathcal{A} \subset \mathbb{R}^d$, for some $d \in \mathbb{N}$, the Borel sigma-field on $\mathcal{A}$ is denoted by $\mathscr{B}(\mathcal{A})$. The interior of $\mathcal{A}$ is denoted $\mathrm{int}\mathcal{A}$. The set of all measures that might be defined over a measurable space denoted by the tuple $(\mathcal{A}, \mathscr{F}(\mathcal{A}))$ is denoted by $\triangle(\mathcal{A}, \mathscr{F}(\mathcal{A}))$. The support of the measure P is denoted by $\operatorname{supp} P$. The *generalized relative entropy* is defined below as the extension to σ-finite measures of the relative entropy usually defined for probability measures.

Definition 1.1 (Relative Entropy)**.** *Given two σ-finite measures P and Q on the same measurable space, such that Q is absolutely continuous with respect to P, the relative entropy of Q with respect to P is*

$$D(Q\|P) = \int \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right) \mathrm{d}P(x), \tag{1}$$

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

Definition 1.2 (Mutual and Lautum Information)**.** *Consider two random variables X and Y jointly inducing the probability measure P_{XY} in $\left(\mathbb{R}^2, \mathscr{B}\left(\mathbb{R}^2\right)\right)$. Let the measure P_X and the measure P_Y, both on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, be the marginal probability measures of P_{XY}. The mutual information between X and Y is*

$$\mathtt{I}(P_{XY}) = D(P_{XY}\|P_Y P_X); \tag{2}$$

and the lautum information [33] between X and Y is

$$\mathtt{L}(P_{XY}) = D(P_Y P_X\|P_{XY}). \tag{3}$$

1.2 The Problem of Empirical Risk Minimization (ERM)

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

$$Y = f(\boldsymbol{\theta}^\star, X). \tag{4}$$

The elements of the sets $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$ are often referred to as *models*, *patterns* and *labels*, respectively. A pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is referred to as a *labeled pattern*. In practice, patterns are high dimensional structures such as digital photos, medical images, digital text, speech signals, DNA sequences, etc. Labels are often countable sets, e.g., proper names, or uncountable sets, e.g., the reals numbers. A typical example of a labeled pattern (x, y) is the case in which x is a representation of a medical image and y is the representation of an indication of whether an anomaly is observed in such image.

The model $\boldsymbol{\theta}^\star$ in (4), which is not necessarily unique and often referred to as the *ground truth model*, is unknown. Given a number of labelled patterns, the objective consists in obtaining a model $\hat{\boldsymbol{\theta}}$, such that given a new pattern $u \in \mathcal{X}$, the estimated label $f(\hat{\boldsymbol{\theta}}, u)$ is equal, or as close as possible, to the true label $f(\boldsymbol{\theta}^\star, u)$. Establishing a distance between the labels $f(\hat{\boldsymbol{\theta}}, u)$ and

$f(\boldsymbol{\theta}^\star, u)$ is made possible by the assumption that the set $\mathcal{Y}$ can be equipped with a metric to form a metric space.

The *loss* or *risk* of adopting a model $\boldsymbol{\nu} \in \mathcal{M}$ for a particular labelled pattern $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is determined by the function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty)$, which is often referred to as the *risk function*. More specifically, the model $\boldsymbol{\theta}$ induces the risk $\ell\left(f(\boldsymbol{\theta}, x), y\right)$ with respect to the labelled pattern $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The risk function is chosen such that $\ell\left(y, y\right) = 0$. Nonetheless, it might exist other models $\boldsymbol{\theta} \in \mathcal{M} \setminus \{\boldsymbol{\theta}^\star\}$ such that $\ell\left(f(\boldsymbol{\theta}, x'), y'\right) = 0$ for some specific data points (x', y').

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

$$\boldsymbol{z} = \big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \big) \in (\mathcal{X} \times \mathcal{Y})^n , \tag{5a}$$

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

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

Thus, given the data set $\boldsymbol{z}$ in (5a), the model $\boldsymbol{\theta}$ is preferred against another model $\boldsymbol{\mu} \in \mathcal{M}$, if $\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) < \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\mu}\right)$. Using this notation, the ERM problem, with respect to the data set $\boldsymbol{z}$, can be formulated as an optimization problem of the form [45]:

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) , \tag{5c}$$

whose solutions form the set denoted by

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

The ground truth model $\boldsymbol{\theta}^\star$ in (4) satisfies that $\boldsymbol{\theta}^\star \in \mathcal{T}\left(\boldsymbol{z}\right)$ and $\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}^\star\right) = 0$. That is, the model $\boldsymbol{\theta}^\star$ is one of the solutions to the ERM problem in (5), and thus, the set $\mathcal{T}\left(\boldsymbol{z}\right)$ is not empty.

Despite the apparent simplicity of the ERM problem in (5), it is a difficult problem [46]. Often, the difficulty is due in part to the representation of the patterns, which are collections of heterogeneous classes of data with large dimensions; the function f in (4), which might be computationally difficult to calculate; the choice of the risk function, which is left at the discretion of the statistician; and the large number of data points whose processing requires unprecedented computing capabilities; among many others reasons. See for instance [5] and [21] for a detailed discussion.

1.3 Stochastic Solvers

Among the most popular methods for solving the ERM problem in (5) are those based on the gradient. The stochastic gradient descent algorithm [35] and its variants fall within this class of methods. See, for instance, the literature reviews in [8, 45, 49] and references therein. Other methods are based on constructing probability measures on $(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right))$ that are used to choose a particular model from the set $\mathcal{M}$. This operation is known as *sampling* [37]. Methods based on the construction of probability measures include Bayesian methods [29,36] and PAC-Bayesian methods [18, 40]. In order to describe these methods, it is assumed that independently of the choice of representation of patterns and labels, it is possible to form the following probability space:

$$(\mathcal{X} \times \mathcal{Y}, \mathscr{F}\left(\mathcal{X} \times \mathcal{Y}\right), P_{XY}) . \tag{6}$$

Given the probability space in (6), a more formal definition of labeled pattern or data point is the following.

Definition 1.3 (Data Point)**.** *The pair (x, y) is said to be a data point if $(x, y) \in \operatorname{supp} P_{XY}$.*

Several data points form a data set, as shown hereunder.

Definition 1.4 (Data Set)**.** *Given n data points, with $n \in \mathbb{N}$, denoted by (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), a data set is represented by the tuple $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$.*

In the following, a central idea is that of *stochastic solvers*, also known as *stochastic estimators* in the realm of estimation theory [34]; and *samplers* in the realm of learning theory [10, 37].

A *stochastic solver* to the ERM problem in (5) is a probability measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ conditioned on an element of the set $(\mathcal{X} \times \mathcal{Y})^n$, i.e., a dataset.

Definition 1.5 (Stochastic Solver)**.** *A function $g : \mathscr{B}(\mathcal{M}) \times (\mathcal{X} \times \mathcal{Y})^n \to [0, 1]$ is said to be a stochastic solver to the ERM problem in (5) if it satisfies the following conditions:*
(a) For all $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$, the function $h_{\boldsymbol{u}} : \mathscr{B}(\mathcal{M}) \to [0, 1]$ that satisfies for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$ that

$$h_{\boldsymbol{u}}(\mathcal{A}) = g(\mathcal{A}, \boldsymbol{u}), \tag{7}$$

is a probability measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$; and
(b) For all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$, the function $h_{\mathcal{A}} : (\mathcal{X} \times \mathcal{Y})^n \to [0, 1]$ that satisfies for all $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$ that

$$h_{\mathcal{A}}(\boldsymbol{u}) = g(\mathcal{A}, \boldsymbol{u}), \tag{8}$$

is measurable with respect to the measurable spaces $((\mathcal{X} \times \mathcal{Y})^n, \mathscr{F}((\mathcal{X} \times \mathcal{Y})^n))$ and $([0, 1], \mathscr{B}([0, 1]))$.

In the following, stochastic solvers are denoted by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$. Hence, for all $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$, the notation $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}$ represents the probability measure from which models can be sampled from to approximate the solution to the ERM problem in (5) given the dataset $\boldsymbol{u}$. Moreover, given a set $\mathcal{A} \in \mathscr{B}(\mathcal{M})$, the function $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}(\mathcal{A}|\cdot) : (\mathcal{X} \times \mathcal{Y})^n \to [0, 1]$ is Borel measurable with respect to the product measurable space $((\mathcal{X} \times \mathcal{Y})^n, \mathscr{F}((\mathcal{X} \times \mathcal{Y})^n))$. Hence, the notations $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}(\mathcal{A}|\boldsymbol{u})$ or $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}(\mathcal{A})$ are indifferently used and represent the probability assigned by the stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the set $\mathcal{A}$ conditioned on the observation of the data set $\boldsymbol{u}$.

Using this notation, this work focuses in the case in which given a stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem in (5), a model $\hat{\boldsymbol{\theta}}$ is randomly selected from $\mathcal{M}$ according to the probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, with $\boldsymbol{z}$ being the data set $\boldsymbol{z}$ in (5a). Other choices for model selection might be, for instance, the first cumulant of the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, i.e.,

$$\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta} \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). \tag{9}$$

Another choice might be the maximum of the Radon-Nikodym derivative of the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to a given measure P on the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathcal{M}} \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}), \tag{10}$$

if the Radon-Nikodym derivative and the maximum exist. The choice in (10) is reminiscent to *maximum à posteriori* estimation and *maximum likelihood* estimation, when P is chosen to be the Lebesgue measure on $\mathcal{M}$. See for instance, [34, Chapter IV]. Other choices of $\hat{\boldsymbol{\theta}}$ obtained from $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, and the corresponding implications, are described in [18] and references therein.

In order to study the case in which a model $\hat{\boldsymbol{\theta}}$ is randomly chosen according to $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, it is assumed that the function $\bar{\ell} : \mathcal{M} \times \mathcal{X} \times \mathcal{Y} \to [0, +\infty)$, which satisfies

$$\bar{\ell}(\boldsymbol{\theta}, x, y) = \ell(f(\boldsymbol{\theta}, x), y), \tag{11}$$

where the functions f and ℓ are those in (4) and (5b), is measurable with respect to the measurable spaces $(\mathcal{M} \times \mathcal{X} \times \mathcal{Y}, \mathscr{F}(\mathcal{M} \times \mathcal{X} \times \mathcal{Y}))$ and $([0,+\infty), \mathscr{B}([0,+\infty)))$. This assumption implies that for all $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$, the empirical risk function $\mathsf{L}_{\boldsymbol{u}}$ in (5b) is measurable with respect to the measurable spaces $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $([0,+\infty), \mathscr{B}([0,+\infty)))$.

Under these assumptions, let the expected empirical risk induced by the stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ be defined as follows.

1.3.1 Expected Empirical Risk

The expected empirical risk is defined with respect to a dataset, as follows.

Definition 1.6 (Expected Empirical Risk)**.** *Given a data set $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})$, let the function $\mathsf{R}_{\boldsymbol{u}} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0,+\infty)$ be such that for all probability measures $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$,*

$$\mathsf{R}_{\boldsymbol{u}}(Q) = \int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}), \tag{12}$$

where the function $\mathsf{L}_{\boldsymbol{u}}$ is in (5b)*. The empirical risk induced by the measure Q at dataset $\boldsymbol{u}$ is $\mathsf{R}_{\boldsymbol{u}}(Q)$.*

Consider a stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem with respect to the data set $\boldsymbol{z}$ in (5), and assume that

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{T}(\boldsymbol{z})) = 1, \tag{13}$$

with the set $\mathcal{T}(\boldsymbol{z})$ in (5d). Hence, $\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = 0$, i.e., the stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ achieves zero expected empirical risk with respect to the data set $\boldsymbol{z}$ in (5). Nonetheless, given a data set $\boldsymbol{u}$ different from $\boldsymbol{z}$, the stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ might induce a nonzero expected empirical risk. That is, $\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \geqslant \mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = 0$. This implies that a desired property for any stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is that for all data sets $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$, the expected empirical risk $\mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ be small and exhibit small changes with respect to changes in the data set $\boldsymbol{u}$. In order to formalize this desired property, the expected empirical risk is compared to the expected risk.

1.3.2 Expected Risk

The expected risk is defined as follows.

Definition 1.7 (Expected Risk)**.** *Let the function $\mathsf{R} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0,+\infty)$ be such that for all probability measures $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$,*

$$\mathsf{R}(Q) = \int \left(\int \bar{\ell}(\boldsymbol{\theta}, u, v) \, \mathrm{d}P_{XY}(u,v) \right) \mathrm{d}Q(\boldsymbol{\theta}), \tag{14}$$

where the function $\bar{\ell}$ is in (11) *and the measure P_{XY} is specified in* (6)*. The expected risk induced by the measure Q is $\mathsf{R}(Q)$.*

The interest in the expected risk, i.e., the function R in (14), lies on the fact that it establishes a metric of preference among stochastic solvers. More specifically, let $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ and $Q_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ be two stochastic solvers to the ERM in (5). The stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is preferred against $Q_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, if $\mathsf{R}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) < \mathsf{R}\left(Q_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$. In particular, consider a stochastic solver $P^{\star}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem with respect to the data set $\boldsymbol{z}$ in (5) for which for all $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$,

$$P^{\star}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}(\mathcal{T}) = 1, \tag{15}$$

with $\mathcal{T}$ being the following set,

$$\mathcal{T} = \arg\min_{\boldsymbol{\theta}\in\mathcal{M}} \int \bar{\ell}\left(\boldsymbol{\theta}, u, v\right) \mathrm{d}P_{XY}\left(u, v\right), \tag{16}$$

where the function $\bar{\ell}$ is in (11). The optimal model $\boldsymbol{\theta}^\star$ in (4) satisfies $\boldsymbol{\theta}^\star \in \mathcal{T}$; and the stochastic solver $P^\star_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ satisfies that for all $\boldsymbol{u} \in (\mathcal{X}\times\mathcal{Y})^n$,

$$\mathsf{R}_{\boldsymbol{u}}\left(P^\star_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right) = \mathsf{R}\left(P^\star_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right) = 0. \tag{17}$$

In the context of the ERM problem in (5), this justifies the preference for $P^\star_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ against stochastic solvers $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ for which $\mathsf{R}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) > 0$, with $\boldsymbol{z}$ being the data set in (5a). Unfortunately, such preference metric is purely theoretical, as in practice, the function R in (14) and the stochastic solver in (15) cannot be calculated due to their dependence on the unknown probability measure P_{XY} in (6).

Despite the fact that the function R in (14) cannot be calculated, it draws particular interest due to the following equality, which holds for all measures $Q \in \triangle\left(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right)\right)$,

$$\mathsf{R}\left(Q\right) = \int \mathsf{R}_{\boldsymbol{u}}\left(Q\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right), \tag{18}$$

where the probability measure $P_{\boldsymbol{Z}}$ is a product measure in the product measurable space $\left(\left(\mathcal{X}\times\mathcal{Y}\right)^n, \mathscr{F}\left(\left(\mathcal{X}\times\mathcal{Y}\right)^n\right)\right)$. That is, for all $\mathcal{A} \in \mathscr{F}\left(\left(\mathcal{X}\times\mathcal{Y}\right)^n\right)$ of the form $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n$, such that for all $i \in \{1, 2, \ldots, n\}$, $\mathcal{A}_i \in \mathscr{F}\left(\mathcal{X}\times\mathcal{Y}\right)$, it holds that

$$P_{\boldsymbol{Z}}\left(\mathcal{A}\right) = \prod_{t=1}^{n} P_{XY}\left(\mathcal{A}_t\right). \tag{19}$$

More specifically, under the assumption that data points are independent and identically distributed according to P_{XY} in (6), the expected value with respect to $P_{\boldsymbol{Z}}$ of the expected empirical risk in (18) is equal to the expected risk in (14). Given a stochastic solver, this observation allows studying the difference between the expected empirical risk with respect to a given data set $\boldsymbol{u}$ and the expected risk, which is defined as the *generalization gap* (GG).

1.3.3 Generalization Gap

The generalization gap is defined as follows.

Definition 1.8 (Generalization Gap)**.** *Let $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ be a stochastic solver to the ERM problem in* (5)*. Let also the function $\mathsf{G} : \left(\mathcal{X}\times\mathcal{Y}\right)^n \to \mathbb{R}$ be such that for all data sets $\boldsymbol{u} \in \left(\mathcal{X}\times\mathcal{Y}\right)^n$,*

$$\mathsf{G}\left(\boldsymbol{u}\right) = \mathsf{R}_{\boldsymbol{u}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right) - \mathsf{R}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right), \tag{20}$$

where the functions $\mathsf{R}_{\boldsymbol{u}}$ and R are in (12) *and* (14)*, respectively. The generalization gap induced by the stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ at the dataset $\boldsymbol{u}$ is $\mathsf{G}\left(\boldsymbol{u}\right)$.*

The analysis of the generalization gap relies on the fact that the function G in (20) is measurable with respect to the measurable spaces $\left(\left(\mathcal{X}\times\mathcal{Y}\right)^n, \mathscr{F}\left(\left(\mathcal{X}\times\mathcal{Y}\right)^n\right)\right)$ and $\left(\mathbb{R}, \mathscr{B}\left(\mathbb{R}\right)\right)$. This is a consequence of the fact that the function $\bar{\ell}$ in (11) is assumed to be measurable with respect to the measurable spaces $\left(\mathcal{M}\times\mathcal{X}\times\mathcal{Y}, \mathscr{F}\left(\mathcal{M}\times\mathcal{X}\times\mathcal{Y}\right)\right)$ and $\left([0,+\infty), \mathscr{B}\left([0,+\infty)\right)\right)$. The measurability of the function G allows leveraging existing results in concentration inequalities for obtaining generalization guarantees.

1.4 Generalization Guarantees

Generalization guarantees are upper-bounds on the generalization gap, which can be of two types [1]. The first type consists in upper bounds on the expected value with respect to $P_{\boldsymbol{Z}}$ in (19) of the absolute value of G. See for instance, [39, 50, 52, 53]. The second type consists in upper bounds on the function G that hold with high probability with respect to the measure $P_{\boldsymbol{Z}}$. Bounds of this type are described in [11, 18–20, 30, 52, 53] and are often referred to as *probably approximately correct* (PAC) guarantees for generalization. This appellation follows from the fact that probability measures satisfying such bounds, concentrate on sets containing models that induce small generalization gaps. Therefore, models sampled according to such measures are *approximately correct*, in the sense of small generalization gaps, with high probability. The notion of PAC can be understood as a generalization of the concept first introduced in [44] and the development of bounds of this kind is a central objective in the so-called PAC-Learning [40, Chapter 3].

The following examples illustrate these two types of guarantees.

Example 1.1 (Theorem 1 in [50])**.** *Assume that there exists a real $\sigma > 0$ such that for all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $t \in \mathbb{R}$,*

$$\log\left(\int \exp\left(t\,\bar{\ell}\left(\boldsymbol{\theta}, u, v\right)\right) \mathrm{d}P_{XY}\left(u, v\right)\right) \leqslant \frac{1}{2}\sigma^2 t^2 + t\int \bar{\ell}\left(\boldsymbol{\theta}, u, v\right) \mathrm{d}P_{XY}\left(u, v\right), \tag{21}$$

with the function $\bar{\ell}$ in (11)*. Then, for all stochastic solvers $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem in* (5)*, it holds that*

$$\int \left|\mathsf{G}\left(\boldsymbol{u}\right)\right| \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right) \leqslant \sqrt{\frac{2\sigma^2}{n}\int D\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}} \| P_{\boldsymbol{\Theta}}\big) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right)}, \tag{22}$$

where the data set $\boldsymbol{z}$ is in (5a)*; the probability measure $P_{\boldsymbol{\Theta}}$ is such that for all $\mathcal{A} \in \mathscr{B}\left(\mathcal{M}\right)$,*

$$P_{\boldsymbol{\Theta}}\left(\mathcal{A}\right) = \int P_{\boldsymbol{\Theta}|\boldsymbol{Z}}\left(\mathcal{A}|\boldsymbol{u}\right) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right); \tag{23}$$

and the measure $P_{\boldsymbol{Z}}$ is in (19)*.*

Example 1.2 (Theorem 2 in [30])**.** *Given a real $\delta > 0$, a stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem in* (5)*, and a probability measure P on $(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right))$ such that $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with P, let $\mathcal{R}(\delta, P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P)$ be the set*

$$\mathcal{R}(\delta, P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P) \triangleq \left\{\boldsymbol{u} \in \left(\mathcal{X}\times\mathcal{Y}\right)^n : \mathsf{G}\left(\boldsymbol{u}\right) < \sqrt{\frac{D\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P\big) + \ln\frac{2\sqrt{n}}{\delta}}{2n}}\right\}, \tag{24a}$$

where the dataset $\boldsymbol{z}$ is in (5a)*. Then, for all $\delta > 0$, and for all stochastic solvers $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem in* (5) *for which the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with P, it holds that*

$$P_{\boldsymbol{Z}}\left(\mathcal{R}(\delta, P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P)\right) > 1 - \delta, \tag{24b}$$

where the product probability measure $P_{\boldsymbol{Z}}$ is defined in (19)*.*

In Example 1.1, the inequality in (21) implies that, for all $\boldsymbol{\theta} \in \mathcal{M}$, the random variable $\bar{\ell}\left(\boldsymbol{\theta}, X, Y\right)$, with X and Y being the random variables that induce the probability measure P_{XY} in (6), is σ sub-Gaussian. In such a case, for all stochastic solvers $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ for which

$$\int D\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}} \| \bar{P}_{\boldsymbol{\Theta}}\big) \mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{u}\right) \leqslant \gamma, \tag{25}$$

with $\gamma \in (0, +\infty)$, it is always possible to arbitrarily reduce to zero the expectation with respect to $P_{\boldsymbol{Z}}$ of the absolute value of G in (20) by increasing the number n of data points in the data set $\boldsymbol{z}$ in (5a). This observation is central as it justifies the fact that a stochastic solver that minimizes $\mathsf{R}_{\boldsymbol{z}}$ in (12) and satisfies the condition

$$\int |\mathsf{G}(\boldsymbol{u})| \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}) \leqslant \epsilon, \tag{26}$$

for some $\epsilon > 0$ arbitrarily small, can be accepted as an approximation to a stochastic solver that minimizes the expected risk R in (14). The relevance of this observation stems from the fact that the function R cannot be directly minimized as it depends on the unknown probability measure P_{XY} in (6).

Alternatively, the expectation with respect to $P_{\boldsymbol{Z}}$ of the absolute value of G in (20) can be made arbitrarily small by reducing the dependence of the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ on the data set $\boldsymbol{z}$ in (5a). See, for instance [39] and [50] in which such dependence is studied using information measures. Consider for instance a stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ that is independent of the data set $\boldsymbol{z}$. That is, for all data sets $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$,

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}} = Q, \tag{27}$$

for some probability measure Q on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. In this case, $D\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}} \| \bar{P}_{\boldsymbol{\Theta}}\big) = 0$, and thus, equality is met in (25) with $\gamma = 0$. Together with (22), this implies that for all data sets $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$, the generalization gap at data set $\boldsymbol{u}$ is zero, i.e., $\mathsf{G}(\boldsymbol{u}) = 0$.

In a nutshell, two desired properties for a stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem in (5) are:

(a) The stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ induces a small expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$; and

(b) the relative entropy of $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to a probability measure independent of the data set $\boldsymbol{z}$ is close to zero.

As shown by the previous examples, properties (a) and (b) can be simultaneously satisfied in the asymptotic regime of large data sets, under some specific conditions. This implies arbitrarily small generalization gap and small expected empirical risk. Nonetheless, in the case in which data sets contain only a finite number of labelled patterns, satisfying condition (a) and satisfying condition (b) are mutually opposing objectives. Hence, observing small generalization gaps does not imply small expected empirical risk. A class of stochastic solvers that is widely known to handle this tradeoff is that of the Gibbs stochastic solvers [16].

1.4.1 Gibbs Stochastic Solvers

A Gibbs stochastic solver to the ERM problem in (5) is often defined with respect to another probability measure on the measurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. Nonetheless, in this work, a more general definition is adopted and Gibbs stochastic solvers are defined with respect to σ-finite measures. The following definition strengthen this observation.

Definition 1.9 (Gibbs Stochastic Solver)**.** *Let P be a σ-finite measure on the mesurable space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and let λ be a positive real. A stochastic solver of the ERM problem with respect to the data set $\boldsymbol{z}$ in* (5)*, denoted by $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, is said to be a P-Gibbs stochastic solver with parameter λ, if for all $\boldsymbol{\theta} \in \mathcal{M}$, it holds that,*

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\displaystyle\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}P(\boldsymbol{\nu})}, \tag{28}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ *is defined in* (5b).

Let $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ be a Gibbs stochastic solver with respect to a given probability measure P with parameter $\lambda > 0$. Then, the main feature of $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is that the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, with $\boldsymbol{z}$ in (5a), is a Gibbs measure with respect to the measure P, c.f., [6, 7, 17] and [14, Chapter 4].

1.4.2 Empirical Risk Minimization with Relative Entropy Regularization

One of the main properties of the P-Gibbs stochastic solver $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in (28), when P is a probability measure, is that the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the unique solution to the following optimization problem [11, 18, 50],

$$\min_{Q\in\triangle(\mathcal{M},\mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}\left(Q\right) + \lambda D\left(Q\|P\right), \tag{29}$$

where the function $\mathsf{R}_{\boldsymbol{z}}$ is specified in (12). In the following, the optimization problem in (29) is referred to as ERM with relative entropy regularization (ERM-RER) problem.

From the objective function of the optimization problem in (29), it is easy to see that the P-Gibbs stochastic solver $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in (28) trades off via the parameter λ, the two competing objectives described above: (a) minimizing the expected empirical risk with respect to the data set $\boldsymbol{z}$, i.e., the function $\mathsf{R}_{\boldsymbol{z}}$; and (b) approaching, in the sense of relative entropy, the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ to another measure, i.e., P, which is independent of the data sets. This is in part, one of the reasons that justify the popularity of Gibbs stochastic solvers. Nonetheless, a prominent difficulty for adopting Gibbs stochastic solvers is the choice of the reference probability measure P, which is often restricted to probability measures.

1.4.3 Special Cases of the ERM-RER Problem

The nature of the set $\mathcal{M}$ and the choice of the reference measure P lead to special cases of the ERM-RER problem in (29). Three cases are of particular interest: (a) The set $\mathcal{M}$ satisfies $\mathcal{M} \subseteq \mathbb{R}^d$, with $d \in \mathbb{N}$; and P is the Lebesgue measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$; (b) The set $\mathcal{M}$ is countable; and the measure P is a counting measure; and (c) The set $\mathcal{M}$ and the measure P form a probability measure space $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$.

In the former, the ERM-RER in (29) satisfies the following

$$\begin{aligned}
\min_{Q\in\triangle(\mathcal{M},\mathscr{B}(\mathcal{M}))} \mathsf{R}_{\boldsymbol{z}}\left(Q\right) + \lambda D\left(Q\|P\right) =& \min_{Q\in\triangle(\mathcal{M},\mathscr{B}(\mathcal{M}))} \int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{\nu})\mathrm{d}P(\boldsymbol{\nu}) \\
&+\lambda \int \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right) \mathrm{d}P(x) && (30)\\
=&\min_{g} \int_{\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) g\left(\boldsymbol{\nu}\right) \mathrm{d}\boldsymbol{\nu} + \lambda \int_{\mathcal{M}} g(\boldsymbol{\theta}) \log\left(g(\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{\theta} && (31)\\
=&\min_{g} \int_{\mathcal{M}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right) g\left(\boldsymbol{\nu}\right) \mathrm{d}\boldsymbol{\nu} - \lambda H\left(g\right), && (32)
\end{aligned}$$

where the Radon-Nikodym derivative $\frac{\mathrm{d}Q}{\mathrm{d}P}$ in (30) is a probability density function (pdf) denoted by g, which implies that the optimization domain in (32) is the set of pdfs on $\mathcal{M}$. In (32), the notation $H\left(g\right)$ represents the differential entropy of the pdf g, c.f., Chapter 8 in [12].

In case (b), a similar analysis would show that the ERM-RER problem in (29) boils down to the ERM with discrete entropy regularization (ERM-DisER). More specifically, the Radon-Nikodym

derivative $\frac{\mathrm{d}Q}{\mathrm{d}P}$ in (30) is a probability mass function (pmf) denoted by p. This implies that the optimization domain is the set of pmfs on $\mathcal{M}$; and the entropy is that of the pmf p, c.f., Chapter 2 in [12]. Both, the ERM-DifER and ERM-DisER problems are closely related to those typically arising while using Jayne's maximum entropy principle [23–25]. In the case of classification problems, see for instance, [16, 22, 54].

Finally, case (c) is the classical formulation of ERM-RER and often referred to as the *information-risk minimization* (IRM) problem in information theory [53].

1.5 Contributions

In this report, the problem of ERM-RER with respect to the data set $\boldsymbol{z}$ in (29) is studied by assuming that the reference measure P is a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, and not necessarily a probability measure. The motivation for this generalization is that some specific choices of the reference measure lead to special cases of the ERM-RER problem that are central in learning theory, e.g., the ERM with (differential or discrete) entropy; or the IRM problem. Moreover, this generalization allows for a larger degree of flexibility in the incorporation of prior knowledge over the set of models. In a nutshell, the proposed formulation yields a unified mathematical framework that comprises a large class of problems.

Under these assumptions, the most relevant results in this work are described hereunder:

- (i) The optimal solution to the ERM-RER with respect to the dataset $\boldsymbol{z}$ in (29), when P is a σ-finite measure, is unique and forms a P-Gibbs stochastic solver $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, where $\lambda \in (0, b)$, for some $b \in \mathbb{R} \cup \{+\infty\}$.
- (ii) The expected empirical risk $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is nondecreasing with λ (Theorem 2.2). The class of loss functions for which $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is monotone an increasing with λ is introduced.
- (iii) In the limit as λ tends to zero, the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is proved to concentrate on a set, often referred to as the *limiting set*. Depending on the choice of the reference measure P, the limiting set might be empty; identical to the set $\mathcal{T}(\boldsymbol{z})$ in (5d); or exhibit an empty intersection with $\mathcal{T}(\boldsymbol{z})$. This is due to the flexibility in the choice of the reference measure in this formulation.
- (iv) The class of reference measures for which the limiting set is identical to the set of solutions to the ERM problem in (5), i.e., the set $\mathcal{T}(\boldsymbol{z})$ in (5d), is characterized.
- (v) The empirical risk, when the models are distributed according to the optimal probability measure, is shown to be a sub-Gaussian random variable.
- (vi) Necessary and sufficient conditions for the existence of a regularization parameter that achieves an arbitrarily small empirical risk with arbitrarily high probability are presented.
- (vii) The sensitivity of the expected empirical risk to deviations from the solution of the ERM-RER problem is studied. The sensitivity is then used to provide upper and lower bounds on the expected empirical risk. Moreover, it is shown that the expectation of the sensitivity is upper bounded, up to a constant factor, by the square root of the lautum information between the models and the datasets.

These findings, together with other results described in the following sections, provide a deeper understanding of the ERM-RER problem and set a theoretical base for designing algorithms

based on Gibbs stochastic solvers whose reference measures are not necessarily probability measures.

2 The Solution to the ERM Problem with Relative Entropy Regularization

This section studies the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in (29) assuming that the reference measure P is a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$; the parameter $\lambda > 0$ is fixed; and the dataset $\boldsymbol{z}$ is the one in (5a). Under these assumptions, the solution is presented in terms of the function $K_{\boldsymbol{z}} : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ and the set $\mathcal{K}_{\boldsymbol{z}} \subset (0, +\infty)$, which are both parametrized by the data set $\boldsymbol{z}$ and the measure P. The former satisfies that for all $t \in \mathbb{R}$,

$$K_{\boldsymbol{z}}(t) = \log\left(\int \exp\left(t\, \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right) \mathrm{d}P(\boldsymbol{\theta})\right), \tag{33}$$

with $\mathsf{L}_{\boldsymbol{z}}$ in (5b), whereas, the latter satisfies that

$$\mathcal{K}_{\boldsymbol{z}} \triangleq \left\{ s > 0 :\ K_{\boldsymbol{z}}\left(-\frac{1}{s}\right) < +\infty \right\}. \tag{34}$$

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

Lemma 2.1. *The set $\mathcal{K}_{\boldsymbol{z}}$ in (34) is either the empty set or a convex set that satisfies*

$$(0, b) \subset \mathcal{K}_{\boldsymbol{z}}, \tag{35}$$

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

Proof. The proof of Lemma 2.1 is presented in Appendix A. □

In the special case in which P is chosen to be a probability measure over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, it holds that for all $t > 0$, the value $K_{\boldsymbol{z}}\left(-\frac{1}{t}\right)$ is upper bounded by zero. The following corollary formalizes this observation.

Lemma 2.2. *Assume that the measure P in (33) is a probability measure. Then, the set $\mathcal{K}_{\boldsymbol{z}}$ in (34) satisfies*

$$\mathcal{K}_{\boldsymbol{z}} = (0, +\infty). \tag{36}$$

Proof. The proof of Lemma 2.2 is presented in Appendix B. □

Using this notation, the solution to the ERM-RER problem in (29) is presented by the following theorem.

Theorem 2.1. *Consider the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in (29), when P is a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34). Then, the solution to such problem is a unique measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, denoted by $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, whose Radon-Nikodym derivative with respect to P satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} P$,*

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \exp\left(-K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right), \tag{37}$$

where, the function $\mathsf{L}_{\boldsymbol{z}}$ is defined in (5b) and the function $K_{\boldsymbol{z}}$ is defined in (33).

Proof: The proof of Theorem 2.1 is presented in Appendix C. ∎

In Theorem 2.1, when the measure P is a probability measure, the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is a Gibbs measure with parameter λ with respect to the function $\mathsf{L}_{\boldsymbol{z}}$ and the measure P . Moreover, in such a case, the function $K_{\boldsymbol{z}}$ is often referred to as the *partition function*. See for instance, [13, Section 7.3.1]. In order not to disrupt with the current nomenclature, in the following, independently on whether P is a probability measure, both the function $K_{\boldsymbol{z}}$ and the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ are referred to as the partition function and the Gibbs measure with parameter λ with respect to the function $\mathsf{L}_{\boldsymbol{z}}$ and the measure P. Similarly, in Definition 1.9, the notion of stochastic solver was introduced by considering that the measure P in the ERM-RER problem in (29) was a probability measure. In the following, as long as P is a σ-finite measure, a stochastic solver $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is said to be a P-Gibbs stochastic solver if the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfies the equality in (37).

The Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ in (37) is strictly positive and bounded, as shown by the following lemma.

Lemma 2.3. *Let the set $\mathcal{T}(\boldsymbol{z})$ in* (5d) *and the σ-finite measure P in* (37) *be such that $\mathcal{T}(\boldsymbol{z}) \cap \operatorname{supp} P \neq \emptyset$. Then, for all $\boldsymbol{\theta} \in \operatorname{supp} P$, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ in* (37) *satisfies for all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{T}(\boldsymbol{z}) \times \mathcal{T}(\boldsymbol{z})$ that*

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) \leqslant \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}_1) = \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}_2). \tag{38}$$

Proof: The proof of Lemma 2.3 is presented in Appendix D. ∎

The following lemmas describe the asymptotic behaviour of the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ in (37) when $\lambda \to 0$ or $\lambda \to +\infty$.

Lemma 2.4. *Let the measure P in* (37) *be a probability measure. Then, for all $\boldsymbol{\theta} \in \operatorname{supp} P$, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ in* (37) *satisfies*

$$\lim_{\lambda \to +\infty} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = 1. \tag{39}$$

Proof: From Theorem 2.1, it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} P$,

$$\lim_{\lambda \to +\infty} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \lim_{\lambda \to +\infty} \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}P(\boldsymbol{\nu})} \tag{40}$$

$$= \frac{1}{\int \mathrm{d}P(\boldsymbol{\nu})} \tag{41}$$

$$= 1, \tag{42}$$

which completes the proof. ∎

Lemma 2.4 unveils the fact that for all $\boldsymbol{\nu} \in \operatorname{supp} P$,

$$\lim_{\lambda \to +\infty} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) = P(\boldsymbol{\nu}). \tag{43}$$

This is consistent with the fact that when λ is arbitrarily large, the optimization problem in (29) boils down to exclusively minimizing the relative entropy. Such minimum is zero and is observed when both measures $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and P are identical.

From Lemma 2.1, it follows that, when P is not a probability measure, the set $\mathcal{K}_{\boldsymbol{z}}$ might be an interval of the form $(0, b)$, with $b < \infty$. Hence, in such a case, the analysis in which λ tends to infinity is void.

In the limit when λ tends to zero, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ in (37) can be presented in terms of the set

$$\mathcal{L}_{\boldsymbol{z}}(\delta) \triangleq \left\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \mathsf{L}_{\boldsymbol{z}}\left(\hat{\boldsymbol{\theta}}\right) + \delta, \text{ with } \hat{\boldsymbol{\theta}} \in \mathcal{T}(\boldsymbol{z})\right\} \tag{44}$$

$$= \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \delta\}, \tag{45}$$

with $\delta \in [0, +\infty)$. In particular consider the positive real

$$\delta^{\star} \triangleq \inf\{\delta \in [0, +\infty) : P(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0\}, \tag{46}$$

and let $\mathcal{L}^{\star}_{\boldsymbol{z}}$ be the following set:

$$\mathcal{L}^{\star}_{\boldsymbol{z}} \triangleq \{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}^{*}) + \delta^{\star}, \text{ with } \boldsymbol{\theta}^{*} \in \mathcal{T}(\boldsymbol{z})\} \tag{47}$$

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

Lemma 2.5. *If $P(\mathcal{L}^{\star}_{\boldsymbol{z}}) > 0$, with the set $\mathcal{L}^{\star}_{\boldsymbol{z}}$ in (47) and P the σ-finite measure in (37), then for all $\boldsymbol{\theta} \in \operatorname{supp} P$, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ in (37) satisfies*

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \frac{1}{P(\mathcal{L}^{\star}_{\boldsymbol{z}})} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{L}^{\star}_{\boldsymbol{z}}\}}. \tag{49}$$

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

$$\lim_{\lambda \to 0^{+}} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \begin{cases} +\infty & \text{if } \boldsymbol{\theta} \in \mathcal{L}^{\star}_{\boldsymbol{z}} \\ 0 & \text{otherwise.} \end{cases} \tag{50}$$

Proof: The proof of Lemma 2.5 is presented in Appendix E. ∎

Lemma 2.5 describes a concentration phenomenon of the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) on the set $\mathcal{L}^{\star}_{\boldsymbol{z}}$ in (47) as λ tends to zero. This is evident from the fact that in the limit, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ is nonzero only over the elements of $\mathcal{L}^{\star}_{\boldsymbol{z}}$. This implication is formally proved in Lemma 2.12 below. Nonetheless, before a formal proof of the concentration of probability, it is interesting to study the set $\mathcal{L}^{\star}_{\boldsymbol{z}}$ in some special cases.

Consider for instance the case in which for all $\delta > 0$ it holds that $P(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0$. Then, $\delta^{\star} = 0$ and $\mathcal{L}^{\star}_{\boldsymbol{z}} = \mathcal{T}(\boldsymbol{z})$, with $\mathcal{T}(\boldsymbol{z})$ in (5d). Hence, in this particular case, the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ asymptotically concentrates on the set of solutions to the problem (5), as λ tends to zero. Interestingly, independently of whether $P(\mathcal{T}(\boldsymbol{z})) = 0$ or $P(\mathcal{T}(\boldsymbol{z})) > 0$, the concentration phenomenon takes place. Note that measures that satisfy that for all $\delta > 0$, it holds that $P(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0$, always exist. For instance, the Lebesgue measure when $\mathcal{M}$ is uncountable; or counting measures when $\mathcal{M}$ is countable. Another case of interest arises when the concentration

of probability occurs over a set that does not contain the set $\mathcal{T}(\boldsymbol{z})$. This occurs when $\delta^\star > 0$ and thus, for all $\delta < \delta^\star$, it holds that $P(\mathcal{L}_{\boldsymbol{z}}(\delta)) = 0$. In this case, as λ tends to zero, the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ concentrates on the set $\mathcal{L}^\star_{\boldsymbol{z}}$, for which $\mathcal{L}^\star_{\boldsymbol{z}} \cap \mathcal{T}(\boldsymbol{z}) = \emptyset$.

Finally, it is interesting to highlight a case in which the concentration of probability does not take place. Consider for instance, the case in which the $\mathcal{L}^\star_{\boldsymbol{z}}$ is empty. In such a case, for all $\boldsymbol{\theta} \in \operatorname{supp} P$, it holds that $\lim_{\lambda \to 0^+} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = 0$. This implies that, in the limit as λ tends to zero, the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is not a probability measure, but the trivial measure. This follows from the fact that in the context of the ERM-RER in (29), a λ arbitrarily small leads to exclusively minimizing the empirical risk, which in this case, does not have an optimum in $\operatorname{supp} P$.

These observations highlight the influence of the reference measure P on the probability measure Gibbs measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. A more in-depth analysis of this influence is presented in the following subsection via the negligible sets with respect to the measure P.

2.1 Negligible Sets and Coherent Measures

A central observation in the choice of the reference measure P in the ERM-RER in (29) is that it determines the negligible sets with respect to the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37). More specifically, for all sets $\mathcal{C} \in \mathscr{B}(\mathcal{M})$, it follows from Theorem 2.1 that if $P(\mathcal{C}) = 0$, then $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = 0$. The following lemma shows that the converse is also true.

Lemma 2.6. *The probability measures P and $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) are mutually absolutely continuous.*

Proof: The proof of Lemma 2.6 is presented in Appendix F. ∎

The relevance of Lemma 2.6 is that it proves that for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, the collection of negligible sets with respect to the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) is identical to the collection of negligible sets with respect to the measure P. That is, for all subsets $\mathcal{C} \in \mathscr{B}(\mathcal{M})$,

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) > 0 \quad \text{if and only if} \quad P(\mathcal{C}) > 0. \tag{51}$$

An immediate consequence of these observations is the following.

Lemma 2.7. *For all $(\alpha, \beta) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), assume that the measures $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfy (37) with $\lambda = \alpha$ and $\lambda = \beta$, respectively. Then, $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ are mutually absolutely continuous.*

Proof: The proof of Lemma 2.7 is presented in Appendix G. ∎

Lemma 2.7 proves that for all $(\alpha, \beta) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, the collection of negligible sets with respect to $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and the collection of negligible sets with respect to $P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ are identical. This implies that the negligible sets with respect to the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) are invariant with respect to the choice of $\lambda \in \mathcal{K}_{\boldsymbol{z}}$.

In the context of the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in (29), a desired condition for a P-Gibbs stochastic solver is that for all $\delta > 0$, the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ in (44) possesses nonzero probability measure, i.e., $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) > 0$. More explicitly, it is desired that the models that induce empirical risks smaller than δ concentrate most of the probability with respect to $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. This is motivated by the fact that if $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}_{\boldsymbol{z}}(\delta)) = 0$, the expected empirical risk

$\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (12) is bounded away from zero, c.f., Lemma 2.11. In this context, Lemma 2.6 and Lemma 2.7 rise the need of choosing the reference measure P in (37) such that $P\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0$, for all $\delta > 0$ arbitrarily small. Measures satisfying this condition are referred to as *coherent measures*.

Definition 2.1. *A measure P is said to be coherent with the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in* (29)*, if for all $\delta > 0$,*

$$P\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0, \tag{52}$$

where the set $\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)$ is defined in (44)*.*

In the case in which P is coherent, the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) satisfies for all $\delta > 0$

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}\left(\delta\right)\right) > 0, \tag{53}$$

as desired.

2.2 The Partition Function, Cumulants and Separability

This section introduces some properties of the partition function $K_{\boldsymbol{z}}$ in (33), which are central in this work. The first property is concerned with the continuity and differentiability of the function $K_{\boldsymbol{z}}$.

Lemma 2.8. *The function $K_{\boldsymbol{z}}$ in* (33) *is continuous and differentiable infinitely many times in* $(-\infty, 0)$*.*

Proof: The proof of Lemma 2.8 is presented in Appendix H. ■

More specific properties for the function $K_{\boldsymbol{z}}$ in (33) can be stated for the case in which the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable with respect to the measure P in (37).

Definition 2.2 (Separable Empirical Risk Function)**.** *The empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in* (5b) *is said to be separable with respect to the σ-finite measure P in* (37)*, 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, such that for all $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathcal{A} \times \mathcal{B}$,*

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_1\right) > c > \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}_2\right). \tag{54}$$

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

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

From this perspective, nonseparable empirical risk functions exhibit little practical interest. This follows from observing that models sampled from the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) induce the same empirical risk.

The definition of separability in Definition 2.2 is in terms of the reference measure P in (37). Nonetheless, Lemma 2.6 provides an alternative definition in terms of the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), with $\lambda \in \mathcal{K}_{\boldsymbol{z}}$.

Definition 2.3. *The empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (5b) *is said to be separable with respect to a* σ*-finite measure* P*, if and only if there exist a real* $\lambda \in \mathcal{K}_{\boldsymbol{z}}$*, with* $\mathcal{K}_{\boldsymbol{z}}$ *in* (34)*; a positive real* $c > 0$*; and two subsets* $\mathcal{A}$ *and* $\mathcal{B}$ *of* $\mathcal{M}$ *nonnegligible with respect to the probability measure* $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (37)*, such that for all* $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{A} \times \mathcal{B}$*, the inequality in* (54) *holds.*

The following lemma presents a general property of the function $K_{\boldsymbol{z}}$ in (33) for the case of separable empirical risk functions.

Lemma 2.9. *The function* $K_{\boldsymbol{z}}$ *in* (33) *is convex. If the empirical risk function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (5b) *is separable with respect to the* σ*-finite measure* P *in* (33)*, then the function* $K_{\boldsymbol{z}}$ *is strictly convex.*

Proof: The proof of Lemma 2.9 is presented in Appendix I. ■

Let the m-th derivative of the function $K_{\boldsymbol{z}}$ in (33) be denoted by $K^{(m)}_{\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$, with $m \in \mathbb{N}$. Hence, for all $s \in \mathcal{K}_{\boldsymbol{z}}$,

$$K^{(m)}_{\boldsymbol{z}}\left(-\frac{1}{s}\right) \triangleq \frac{\mathrm{d}^m}{\mathrm{d}t^m} K_{\boldsymbol{z}}(t)\Big|_{t=-\frac{1}{s}}. \tag{56}$$

The following lemma provides explicit expressions for the first, second and third derivatives of the function $K_{\boldsymbol{z}}$ in (33).

Lemma 2.10. *The first, second and third derivatives of the function* $K_{\boldsymbol{z}}$ *in* (33) *evaluated at* $-\frac{1}{\lambda}$*, with* $\lambda \in \operatorname{int}\mathcal{K}_{\boldsymbol{z}}$ *and* $\mathcal{K}_{\boldsymbol{z}}$ *in* (34)*, satisfy the following equalities,*

$$K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{57}$$

$$K^{(2)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)^2 \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \text{ and} \tag{58}$$

$$K^{(3)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)^3 \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{59}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ *is defined in* (5b) *and the measure* $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *satisfies* (37)*.*

Proof: The proof of Lemma 2.10 is presented in Appendix J. ■

From Lemma 2.10, it follows that if $\boldsymbol{\Theta}$ is the random vector that induces the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), with $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, the empirical risk $\mathsf{L}_{\boldsymbol{z}}$ in (5b) becomes the random variable

$$W \triangleq \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\Theta}), \tag{60}$$

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

2.2.1 The Mean of the Empirical Risk

The mean of the random variable W in (60) is equivalent to the expected empirical risk with respect to the data set $\boldsymbol{z}$ induced by $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, i.e., the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (12), as shown by the following corollary.

Corollary 2.1. *The probability measure* $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (37) *verifies that*

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{61}$$

where the functions $\mathsf{R}_{\boldsymbol{z}}$ *and* $K^{(1)}_{\boldsymbol{z}}$ *are defined in* (12) *and* (57)*, respectively.*

The following theorem shows an interesting property of the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$.

Theorem 2.2. *The value* $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ *in* (61) *is nondecreasing with* $\lambda \in \mathcal{K}_{\boldsymbol{z}}$. *Moreover, the function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (5b) *is separable with respect to the measure* P *if and only if the value* $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ *is strictly increasing with* $\lambda \in \mathcal{K}_{\boldsymbol{z}}$.

Proof: The proof of Theorem 2.2 is presented in Appendix K. ■

A question that arises from Theorem 2.2 is whether the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (61) can be made arbitrarily close to $\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}^{\star}\right) = 0$, with $\boldsymbol{\theta}^{\star}$ in (4), by making λ arbitrarily small. The following lemma shows that there exist cases in which the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ is bounded away from zero, even for arbitrarily small values of λ.

Lemma 2.11. *For all* $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, *with* $\mathcal{K}_{\boldsymbol{z}}$ *in* (34), *the function* $K^{(1)}_{\boldsymbol{z}}$ *in* (57) *satisfies,*

$$K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \geqslant \delta^{\star}, \tag{62}$$

where $\delta^{\star}$ *is defined in* (46). *Moreover, the function* $\mathsf{L}_{\boldsymbol{z}}$ *in* (5b) *is separable with respect to the measure* P *if and only if the inequality in* (62) *is strict.*

Proof: The proof of Lemma 2.11 is presented in Appendix L. ■

In the limit as λ tends to zero, a probability concentration phenomena takes place. This phenomena is studied in detail in Section 2.3. Nonetheless, the following lemma leads to preliminary intuitions.

Lemma 2.12. *The measure* $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (37) *and the set* $\mathcal{L}^{\star}_{\boldsymbol{z}}$ *in* (47) *satisfy,*

$$\lim_{\lambda \to 0^{+}} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}^{\star}_{\boldsymbol{z}}\right) = \begin{cases} 0 & \text{if} \quad \mathcal{L}^{\star}_{\boldsymbol{z}} = \emptyset \\ 1 & \text{if} \quad \mathcal{L}^{\star}_{\boldsymbol{z}} \neq \emptyset. \end{cases} \tag{63}$$

Proof: The proof of Lemma 2.12 is presented in Appendix M. ■

Lemma 2.12 reveals the fact that as λ tends to zero, the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ concentrates in the set $\mathcal{L}^{\star}_{\boldsymbol{z}}$. An immediate consequence of this observation is the following theorem.

Theorem 2.3. *The function* $K^{(1)}_{\boldsymbol{z}}$ *in* (57) *satisfies,*

$$\lim_{\lambda \to 0^{+}} K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \begin{cases} 0 & \text{if} \quad \mathcal{L}^{\star}_{\boldsymbol{z}} = \emptyset \\ \delta^{\star} & \text{if} \quad \mathcal{L}^{\star}_{\boldsymbol{z}} \neq \emptyset. \end{cases} \tag{64}$$

where $\delta^{\star}$ *is defined in* (46).

Proof: The proof of Theorem 2.3 is presented in Appendix N. ■

Theorem 2.3 unveils the relevance of coherent measures (Definition 2.1). More specifically, the expected empirical risk with respect to the dataset $\boldsymbol{z}$ induced by $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, i.e., the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (12), can be made arbitrarily close to zero, if the reference measure P is coherent.

The anomalous case in which $\mathcal{L}^{\star}_{\boldsymbol{z}}$ is empty leads to a zero expected empirical risk in the limit as λ tends to zero, i.e., $\lim_{\lambda \to 0^{+}} \mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = 0$. Nonetheless, in such a case, in the limit

as λ tends to zero, the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the trivial measure. That is, all sets in $\mathscr{F}(\mathcal{M})$ exhibit zero measure with respect to $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Hence, this is an anomaly instead of an optimality guarantee.

2.2.2 Second and Third Cumulants of the Emprical Risk

The monotonicity of the function $K^{(1)}_{\boldsymbol{z}}$, stated by Theorem 2.2, is not a property exhibit by the functions $K^{(2)}_{\boldsymbol{z}}$ and $K^{(3)}_{\boldsymbol{z}}$, respectively specified by (58) and (59). This section highlights this observation via the following example.

Example 2.1. *Consider the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in (29), under the assumption that P is a probability measure and the empirical risk function is of the form*

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{A} \\ 1 & \text{if} \quad \boldsymbol{\theta} \in \mathcal{M} \setminus \mathcal{A}, \end{cases} \tag{65}$$

where, the sets $\mathcal{A}$ and $\mathcal{M} \setminus \mathcal{A}$ are nonnegligible with respect to the reference probability measure P in (29). Hence, from (33), the following holds for all $\lambda > 0$,

$$K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \log\left(P(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - P(\mathcal{A}))\right). \tag{66}$$

The derivatives $K^{(1)}_{\boldsymbol{z}}$, $K^{(2)}_{\boldsymbol{z}}$, and $K^{(3)}_{\boldsymbol{z}}$ in (56) of the function $K_{\boldsymbol{z}}$ in (66) are:

$$K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{\exp\left(-\frac{1}{\lambda}\right)(1 - P(\mathcal{A}))}{P(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - P(\mathcal{A}))}; \tag{67}$$

$$K^{(2)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{P(\mathcal{A})(1 - P(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)}{\left(P(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - P(\mathcal{A}))\right)^2}; \text{ and} \tag{68}$$

$$K^{(3)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \frac{P(\mathcal{A})(1 - P(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)\left(P(\mathcal{A}) - (1 - P(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right)\right)}{\left(P(\mathcal{A}) + \exp\left(-\frac{1}{\lambda}\right)(1 - P(\mathcal{A}))\right)^3}. \tag{69}$$

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

$$P(\mathcal{A}) - (1 - P(\mathcal{A}))\exp\left(-\frac{1}{\lambda}\right) > 0. \tag{70}$$

Assume that $P(\mathcal{A}) \geqslant \frac{1}{2}$. Thus, it holds that for all $\lambda > 0$, the inequality in (70) is always satisfied. This follows from observing that for all $\lambda > 0$,

$$\exp\left(-\frac{1}{\lambda}\right) < 1 \leqslant \frac{P(\mathcal{A})}{1 - P(\mathcal{A})}. \tag{71}$$

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

$$\frac{1}{4} \geqslant K^{(2)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right) > K^{(2)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) > \ldots > 0. \tag{72}$$

Alternatively, assume that $P(\mathcal{A}) < \frac{1}{2}$. In this case, the inequality in (70) is satisfied if and only if

$$\lambda < \left(\log\left(\frac{1 - P(\mathcal{A})}{P(\mathcal{A})}\right)\right)^{-1}. \tag{73}$$

Hence, if $P(\mathcal{A}) < \frac{1}{2}$*, then for all decreasing sequences of positive reals* $\left(\log\left(\frac{1-P(\mathcal{A})}{P(\mathcal{A})}\right)\right)^{-1} > \lambda_1 > \lambda_2 > \ldots > 0$*, it holds that*

$$\frac{1}{4} > K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda_1}\right) > K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda_2}\right) > \ldots > 0. \tag{74}$$

Moreover, for all decreasing sequences of positive reals $\lambda_1 > \lambda_2 > \ldots > \left(\log\left(\frac{1-P(\mathcal{A})}{P(\mathcal{A})}\right)\right)^{-1}$*, it holds that*

$$K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda_1}\right) < K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda_2}\right) < \ldots < \frac{1}{4}. \tag{75}$$

The upperbound by $\frac{1}{4}$ *in* (72), (74) *and* (75) *follows by noticing that the value* $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$ *is maximized when* $\lambda = \left(\log\left(\frac{1-P(\mathcal{A})}{P(\mathcal{A})}\right)\right)^{-1}$ *and* $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right) = \frac{1}{4}$.

Example 2.1 provides important insights on the choice of the reference measure P. Note for instance that the optimal set of models is the set $\mathcal{A}$. That is, $\mathcal{T}(\boldsymbol{z}) = \mathcal{A}$, with $\mathcal{T}(\boldsymbol{z})$ in (5d). When the reference measure assigns a probability to the set of optimal models $\mathcal{T}(\boldsymbol{z})$ that is bigger than or equal to the probability of suboptimal models $\mathcal{M} \setminus \mathcal{T}(\boldsymbol{z})$, i.e., $P(\mathcal{T}(\boldsymbol{z})) \geqslant \frac{1}{2}$, the variance is strictly decreasing to zero when λ decreases.

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

In general, these observation suggest that reference measures P that allocate small measures to the sets containing the set $\mathcal{T}(\boldsymbol{z})$ might require reducing the value λ beyond a small threshold in order to observe small values of $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\lambda}\right)$, e.g., the variance of the random variable W in (60). These observations are central to understanding the concentration of probability that occurs when λ decreases, as discussed in the following section.

2.3 Concentration of Probability

This section describes two phenomena concerning the properties of the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37). First, this section shows that it is possible to determine a subset of $\mathcal{M}$, denoted by $\mathcal{N}(\lambda)$ and defined later in (81), over which the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ concentrates most of the probability as a consequence of reducing $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34). As λ tends to zero, the set $\mathcal{N}(\lambda)$ decreases to a set that is independent of λ, but strongly dependent on the reference measure P in (37). Second, this section shows that at the same time that the set $\mathcal{N}(\lambda)$ decreases as a consequence of reducing $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, the probability $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda))$ increases to one.

2.3.1 Preliminaries

Given a real $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), consider a partition of $\mathcal{M}$ formed by the sets $\mathcal{R}_0(\lambda)$, $\mathcal{R}_1(\lambda)$ and $\mathcal{R}_2(\lambda)$, such that

$$\mathcal{R}_0(\lambda) \triangleq \left\{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \right\}, \tag{76a}$$

$$\mathcal{R}_1(\lambda) \triangleq \left\{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \right\}, \text{ and} \tag{76b}$$

$$\mathcal{R}_2(\lambda) \triangleq \left\{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \right\}, \tag{76c}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ is in (5b).

These sets exhibit several properties that are central for proving the main results of this section.

Lemma 2.13. *The measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), satisfies*

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\lambda)) > 0, \tag{77}$$

if and only if

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\lambda)) > 0, \tag{78}$$

where the sets $\mathcal{R}_1(\alpha)$ and $\mathcal{R}_2(\alpha)$ are in (76b) and (76c), respectively.

Proof: The proof of Lemma 2.13 is presented in Appendix O. ∎

Lemma 2.13 shows that given a real $\gamma \in \mathcal{K}_{\boldsymbol{z}}$, if there exists a nonnegligible set with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ whose elements induce an empirical risk that is smaller than the expected empirical risk $K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)$, then there exists a nonnegligible set with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ whose elements induce an empirical risk that is bigger than the expected empirical risk $K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\gamma}\right)$. Moreover, the converse also holds.

A more general result can be immediately obtained by combining Lemma 2.7 and Lemma 2.13.

Lemma 2.14. *For all $\alpha \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), satisfies*

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\alpha)) > 0, \tag{79}$$

if and only if

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\alpha)) > 0, \tag{80}$$

where the sets $\mathcal{R}_1(\alpha)$ and $\mathcal{R}_2(\alpha)$ are in (76b) and (76c), respectively.

2.3.2 The Limiting Set

Consider the following set, with $\lambda \in \mathcal{K}_{\boldsymbol{z}}$,

$$\mathcal{N}(\lambda) \triangleq \left\{ \boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) \right\}, \tag{81}$$

where the function $K_{\boldsymbol{z}}^{(1)}$ is defined by (57). This section shows that, when λ decreases, the set $\mathcal{N}(\lambda)$ forms a monotonic sequence of sets that decreases to the set

$$\mathcal{N}^\star \triangleq \mathcal{L}_{\boldsymbol{z}}\left(\delta^\star\right), \tag{82}$$

where, $\delta^\star$ is defined in (46); and the set $\mathcal{L}_{\boldsymbol{z}}(\cdot)$ is defined in (44).

In the context of the ERM-RER problem in (29), when the reference measure P is a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and the data set $\boldsymbol{z}$ is the one in (5a), the set $\mathcal{N}(\lambda)$, with $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, contains all the models that induce an empirical risk that is smaller than the expected empirical risk (with respect to the dataset $\boldsymbol{z}$) induced by the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\lambda)}$ in (37). This observation unveils the existence of a relation between the set $\mathcal{N}^\star$ in (82) and the set $\mathcal{T}(\boldsymbol{z})$ in (5d), as shown by the following corollary.

Corollary 2.2. *The set $\mathcal{N}^\star$ in* (82) *satisfies*

$$\mathcal{T}(\boldsymbol{z}) \subseteq \mathcal{N}^\star, \tag{83}$$

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

$$\mathcal{T}(\boldsymbol{z}) = \mathcal{N}^\star, \tag{84}$$

if and only if the reference measure P in (37) *is coherent with the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in* (29)*.*

Corollary 2.2 shows that $\mathcal{N}^\star$ is not empty. This follows from the fact that the set $\mathcal{T}(\boldsymbol{z})$, which is not empty, is a subset of $\mathcal{N}^\star$. This observation turns out to be particularly important at the light of the following theorem.

Theorem 2.4. *For all $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in* (34) *and $\lambda_1 > \lambda_2$, the sets $\mathcal{N}(\lambda_1)$ and $\mathcal{N}(\lambda_2)$ in* (81) *satisfy*

$$\mathcal{M} \supseteq \mathcal{N}(\lambda_1) \supseteq \mathcal{N}(\lambda_2) \supseteq \mathcal{N}^\star, \tag{85}$$

with $\mathcal{M}$ and $\mathcal{N}^\star$ the sets defined in (5d) *and* (82)*. Moreover, if the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in* (5b) *is continuous on $\mathcal{M}$ and separable (Definition 2.2), then,*

$$\mathcal{M} \supset \mathcal{N}(\lambda_1) \supset \mathcal{N}(\lambda_2) \supset \mathcal{N}^\star. \tag{86}$$

Proof: The proof of Theorem 2.4 is presented in Appendix P. ∎

Theorem 2.4 shows that, when λ decreases, the set $\mathcal{N}(\lambda)$ monotonically decreases to the set $\mathcal{N}^\star$, which always contains the set of solutions $\mathcal{T}(\boldsymbol{z})$ to the ERM problem in (5). Nonetheless, for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), depending on the choice of the σ-finite measure P in (37), only a subset of $\mathcal{N}(\lambda)$ might exhibit nonzero probability with respect to the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\lambda)}$ in (37). Consider for instance a noncoherent measure P. That is, there exists a $\delta > 0$, such that for all $\gamma < \delta$, it holds that $P\left(\mathcal{L}_{\boldsymbol{z}}(\gamma)\right) = 0$. Therefore, for all $\lambda \in \left\{\alpha \in \mathcal{K}_{\boldsymbol{z}} : K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right) > \delta\right\}$, it holds that $\mathcal{L}_{\boldsymbol{z}}(\gamma) \subseteq \mathcal{N}(\lambda)$. In this case, note also that the set $\mathcal{T}(\boldsymbol{z})$ in (5d) is a subset of a zero-measure set, i.e., $\mathcal{T}(\boldsymbol{z}) \subseteq \mathcal{L}_{\boldsymbol{z}}(\gamma)$.

2.3.3 Probability of the Limiting Set

The following theorem shows that the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ over the set $\mathcal{N}(\lambda)$ does not increase when λ increases.

Theorem 2.5. *For all $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34) and $\lambda_1 > \lambda_2$, assume that the measures $P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ satisfy (37) with $\lambda = \lambda_1$ and $\lambda = \lambda_2$, respectively. Then, the set $\mathcal{N}(\lambda_2)$ in (81) satisfies*

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)) \leqslant P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)). \tag{87}$$

Moreover, the function $\mathsf{L}_{\boldsymbol{z}}$ is separable with respect to P, with P in (37), (Definition 2.2) if and only if for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, with $\lambda_1 > \lambda_2$, it holds that

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

Proof: The proof of Theorem 2.5 is presented in Appendix Q. ∎

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

Lemma 2.15. *Let the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) be separable and consider two positive reals $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34) and $\lambda_1 > \lambda_2$. Assume that*

$$P\Big(\mathcal{N}(\lambda_1) \cap \mathcal{R}_2(\lambda_2)\Big) = 0. \tag{89}$$

Then, the measures $P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and $P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which satisfy (37) with $\lambda = \lambda_1$ and $\lambda = \lambda_2$ respectively, verify

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_1)) < P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)), \tag{90}$$

where for all $t \in \{1, 2\}$, the sets $\mathcal{R}_2(\lambda_t)$ and $\mathcal{N}(\lambda_t)$ are in (76c) and (81), respectively.

Proof: The proof of Lemma 2.15 is presented in Appendix R. ∎

The following example shows the relevance of Lemma 2.15 in the case in which the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is a simple function.

Example 2.2. *Consider Example 2.1. Note that, for all $\lambda > 0$,*

$$0 < K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < 1, \tag{91}$$

which implies that given two reals λ_1 and λ_2 such that $\lambda_1 > \lambda_2 > 0$, it holds that,

$$\mathcal{N}(\lambda_1) \cap \mathcal{R}_2(\lambda_2) = \left\{\boldsymbol{\nu} \in \mathcal{M} : K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda_2}\right) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda_1}\right)\right\} \tag{92}$$

$$= \emptyset, \tag{93}$$

and moreover, $\mathcal{N}(\lambda_1) = \mathcal{N}(\lambda_2)$. Finally, from Lemma 2.15,

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

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

Theorem 2.6. *The probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), with the measure P being a σ-finite measure P on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, satisfies*

$$\lim_{\lambda\to 0^+} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda)) = \begin{cases} 0 & \text{if} \quad \mathcal{L}^\star_{\boldsymbol{z}} = \emptyset \\ 1 & \text{if} \quad \mathcal{L}^\star_{\boldsymbol{z}} \neq \emptyset, \end{cases} \tag{95}$$

where, the sets $\mathcal{N}(\cdot)$ and $\mathcal{L}^\star_{\boldsymbol{z}}$ are respectively defined in (81) and (47).

Proof: The proof of Theorem 2.6 is presented in Appendix S. ∎

2.4 Cumulant Generating Function of the empirical risk

Let λ be a real in $\mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), and consider the transport of the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) from $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ through the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b). Denote the resulting probability measure in $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ by $P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$ in $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. That is, for all $\mathcal{A} \in \mathscr{B}(\mathbb{R})$,

$$P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) = P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathsf{L}^{-1}_{\boldsymbol{z}}(\mathcal{A})\right), \tag{96}$$

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

$$\mathsf{L}^{-1}_{\boldsymbol{z}}(\mathcal{A}) \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \in \mathcal{A}\}. \tag{97}$$

Note that the random variable W in (60) induces the probability measure $P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$. The objective of this section is to prove that the random variable W is a sub-Gaussian random variable. For this purpose, note that the cumulant generating function induced by the measure $P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$, denoted by $J_{\boldsymbol{z},\lambda} : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, satisfies for all $t \in \mathbb{R}$,

$$J_{\boldsymbol{z},\lambda}(t) = \log\left(\int \exp(tz)\, \mathrm{d}P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}(z)\right) = \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{u})\right). \tag{98}$$

For all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), the following lemma provides an expression for $J_{\boldsymbol{z},\lambda}(t)$ in terms of the cumulant generating function $K_{\boldsymbol{z}}$ in (33), for all $t \in (-\infty, \frac{1}{\lambda})$.

Lemma 2.16. *Given a real $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), the cumulant generating function $J_{\boldsymbol{z},\lambda}$ in (98), verifies the following equality for all $t \in \left(-\infty, \frac{1}{\lambda}\right)$,*

$$J_{\boldsymbol{z},\lambda}(t) = K_{\boldsymbol{z}}\left(t - \frac{1}{\lambda}\right) - K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < +\infty. \tag{99}$$

Proof: The proof of Lemma 2.16 is presented in Appendix T. ∎

Denote by $J^{(m)}_{\boldsymbol{z},\lambda} : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, with $m \in \mathbb{N}$, the m-th derivative of the function $J_{\boldsymbol{z},\lambda}$ in (98). That is, for all $s \in \left(-\infty, \frac{1}{\lambda}\right)$,

$$J^{(m)}_{\boldsymbol{z},\lambda}(s) = \frac{\mathrm{d}^m}{\mathrm{d}t^m} J_{\boldsymbol{z},\lambda}(t)\Big|_{t=s}. \tag{100}$$

From Lemma 2.16, it follows that for all $m \in \mathbb{N}$, and for all $\alpha \in \left(-\infty, \frac{1}{\lambda}\right)$, the following holds,

$$J^{(m)}_{\boldsymbol{z},\lambda}(\alpha) = K^{(m)}_{\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right). \tag{101}$$

The equality in (101) establishes a relation between the cumulant generating functions $J_{\boldsymbol{z},\lambda}$ and $K_{\boldsymbol{z}}$ respectively induced by the measures $P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$ in (96) and $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), with $\lambda \in \mathcal{K}_{\boldsymbol{z}}$. The following lemma leverages these observations and presents the main result of this section.

Theorem 2.7. *The cumulant generating function $J_{\boldsymbol{z},\lambda}$ in (98) verifies the following inequality for all $\alpha \in \left(-\infty, \frac{1}{\lambda}\right)$,*

$$J_{\boldsymbol{z},\lambda}(\alpha) \leqslant \alpha K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{1}{2}\alpha^2 B^2_{\boldsymbol{z}} \tag{102}$$

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

$$B^2_{\boldsymbol{z}} = \sup_{\gamma \in \mathcal{K}_{\boldsymbol{z}}} K^{(2)}_{\boldsymbol{z}}\left(-\frac{1}{\gamma}\right) < +\infty, \tag{103}$$

with $\mathcal{K}_{\boldsymbol{z}}$ in (34); and the functions $K^{(1)}_{\boldsymbol{z}}$ and $K^{(2)}_{\boldsymbol{z}}$ are respectively defined in (57), and (58).

Proof: The function $K_{\boldsymbol{z}}$ in (33) is differentiable infinitely many times over the interior of the set $\mathcal{K}_{\boldsymbol{z}}$ (Lemma 2.8). Thus, from the Taylor-Lagrange theorem, c.f., [43, Theorem 2.5.4], it follows that for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ and for all $\alpha \in \left(-\infty, \frac{1}{\lambda}\right)$, there exists a real $\xi \in \left(-\infty, \frac{1}{\lambda}\right)$ such that

$$K_{\boldsymbol{z}}\left(\alpha - \frac{1}{\lambda}\right) = K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \alpha K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{\alpha^2}{2} K^{(2)}_{\boldsymbol{z}}(\xi)\,. \tag{104}$$

From (104) and Lemma 2.16, it holds that

$$J_{\boldsymbol{z},\lambda}(\alpha) = \alpha K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \frac{\alpha^2}{2} K^{(2)}_{\boldsymbol{z}}(\xi)\,. \tag{105}$$

Finally, the inequality in (102) follows from the maximization of the function $K^{(2)}_{\boldsymbol{z}}$ on the set $\mathcal{K}_{\boldsymbol{z}}$, which completes the proof. ∎

The main implication of Theorem 2.7 is that the random variable W in (60) is a sub-Gaussian random variable with parameter B. This result is central for studying of the generalization error of Gibbs stochastic solvers. Based on Theorem 2.7 and previous results in [50], generalization guarantees can be immediately obtained. Nonetheless, such analysis is left out of the scope of this work.

3 (δ, ϵ)-Optimality and Sensitivity

This section introduces a notion of optimality in probability, which is reminiscent to the existing PAC-generalization guarantees. In this case, the optimality is in the sense of low expected empirical risks instead of small generalization gaps.

Definition 3.1 ((δ,ϵ)-Optimality)**.** *Given a pair of positive reals (δ, ϵ), with $\epsilon < 1$, a stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ to the ERM problem with respect to the data set $\boldsymbol{z}$ in (5) is said to be (δ,ϵ)-optimal, if the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$ in (44) satisfies*

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{L}_{\boldsymbol{z}}(\delta)\right) > 1 - \epsilon. \tag{106}$$

For all $\delta > 0$, it holds that $\mathcal{T}(\boldsymbol{z}) \subset \mathcal{L}_{\boldsymbol{z}}(\delta)$, with the sets $\mathcal{T}(\boldsymbol{z})$ and $\mathcal{L}_{\boldsymbol{z}}$ in (5d) and (44), respectively. Hence, from Definition 3.1, it follows that if the stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is (δ,ϵ)-optimal, then it assigns a probability that is always bigger than $1-\epsilon$ to the set $\mathcal{L}_{\boldsymbol{z}}(\delta)$. That is,

the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ concentrates on a set that contains models that induce an empirical risk that is smaller than δ. From this perspective, particular interest is given to the smallest δ and the smallest ϵ when choosing a (δ, ϵ)-optimal stochastic solver.

3.1 (δ, ϵ)-Optimality of Gibbs Stochastic Solvers

This section shows that Gibbs stochastic solvers to the ERM-RER in (29) might be (δ, ϵ)-optimal for some values of δ and ϵ, when the reference measure is *consistent*.

Definition 3.2 (Consistent Measure)**.** *A measure P is said to be consistent with the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in* (29)*, if the set $\mathcal{L}_{\boldsymbol{z}}^{\star}$ in* (47) *is not empty.*

The relation between coherent and consistent measures is described by the following corollary.

Corollary 3.1. *Every coherent measure is consistent.*

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

Theorem 3.1. *Consider the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in* (29)*, when P is a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. If the measure P is consistent, for all $(\delta, \epsilon) \in (\delta^{\star}, +\infty) \times (0, 1)$, with $\delta^{\star}$ in* (46)*, then, there exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in* (34)*, such that the P-Gibbs stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(\lambda)}$ is (δ, ϵ)-optimal.*

Proof of Theorem 3.1: Let δ be a real in $(\delta^{\star}, +\infty)$, with $\delta^{\star}$ in (46). Let also $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ satisfy the following equality:

$$K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right) \leqslant \delta. \tag{107}$$

Note that under the assumption that P is consistent, such a λ in (107) always exists. This follows from the fact that the value $K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$ is nondecreasing with λ (Theorem 2.2) and in the limit as λ tends to zero (Theorem 2.3), the value $K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$ tends to $\delta^{\star}$. Moreover, from (44) and (81), it holds that

$$\mathcal{N}(\lambda) \subseteq \mathcal{L}_{\boldsymbol{z}}(\delta), \tag{108}$$

and thus,

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\lambda)}\left(\mathcal{L}_{\boldsymbol{z}}(\delta)\right) \geqslant P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\lambda)}\left(\mathcal{N}(\lambda)\right). \tag{109}$$

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

$$P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\gamma)}(\mathcal{N}(\gamma)) > 1 - \epsilon. \tag{110}$$

The existence of such a positive real γ follows from Theorem 2.6. Hence, from (110), it holds that,

$$1 - \epsilon < P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\gamma)}(\mathcal{N}(\gamma)) \tag{111}$$

$$\leqslant P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\gamma)}\left(\mathcal{L}_{\boldsymbol{z}}(\delta)\right), \tag{112}$$

where the inequality in (112) follows from the fact that $\mathcal{N}(\gamma) \subseteq \mathcal{N}(\lambda) \subseteq \mathcal{L}_{\boldsymbol{z}}(\delta)$. Finally, the inequality in (112) implies that the stochastic solver $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(\gamma)}$ is a (δ, ϵ)-optimal probability measure (Definition 3.1). This completes the proof. ■

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

Theorem 3.2. *Consider the ERM-RER problem with respect to the data set $\boldsymbol{z}$ in (29), when P is a σ-finite measure on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$. Then, for all $(\delta, \epsilon) \in (0, +\infty) \times (0, 1)$, there always exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), such that the P-Gibbs stochastic solver $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is (δ, ϵ)-optimal if and only if the reference measure P is coherent.*

Proof of Theorem 3.2: The proof is divided into two parts. The first part deals with the case in which the measure P is coherent. The second part deals with the converse case.

The first part is as follows. Under the assumption that the measure P is coherent it follows that $\delta^\star = 0$. Moreover, it follows from Corollary 3.1 that the measure P is also consistent. Then, from Theorem 3.1, it follows that for all $(\delta, \epsilon) \in (0, +\infty) \times (0, 1)$, there always exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), such that the P-Gibbs stochastic solver $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is (δ, ϵ)-optimal.

The second part is as follows. Under the assumption that the measure P is not coherent, there always exists a $\delta > 0$, such that for all $\gamma < \delta$, it holds that $P(\mathcal{L}_{\boldsymbol{z}}(\gamma)) = 0$. Hence, the P-Gibbs stochastic solver $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is not (γ, ϵ)-optimal, for all $\gamma < \delta$, with $\epsilon \in (0, 1)$. This completes the proof. ■

3.2 Sensitivity of the Expected Empirical Risk

The sensitivity due to a change of measure is a performance metric defined as follows.

Definition 3.3. *Given a σ-finite measure P and a real $\lambda > 0$, let*

$$\mathsf{D}_\lambda : (\mathcal{X} \times \mathcal{Y})^n \times \triangle_P(\mathcal{M}, \mathscr{B}(\mathcal{M})) \to [0, +\infty]$$

be a function such that

$$\mathsf{D}_\lambda(\boldsymbol{u}, Q) = \begin{cases} \left| \mathsf{R}_{\boldsymbol{u}}(Q) - \mathsf{R}_{\boldsymbol{u}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right) \right| & \text{if } \lambda \in \mathcal{K}_{\boldsymbol{u}} \\ +\infty & \text{otherwise,} \end{cases} \tag{113}$$

where the function $\mathsf{R}_{\boldsymbol{u}}$ is defined in (12) and the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}$ is the solution to the ERM-RER problem in (29) with respect to the data set $\boldsymbol{u}$. The sensitivity of the expected empirical risk at dataset $\boldsymbol{u}$ when the measure changes from $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}$ to Q is $\mathsf{D}_\lambda(\boldsymbol{u}, Q)$.

The sensitivity $\mathsf{D}_\lambda(\boldsymbol{z}, Q)$, with $\boldsymbol{z}$ the data set in (5a), is a positive real that indicates the variation of the expected empirical risk with respect to $\boldsymbol{z}$, i.e., the function $\mathsf{R}_{\boldsymbol{z}}$ in (61), when the measure is changed from the optimal solution to the ERM-RER problem in (29), i.e., the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, to an alternative measure Q. That is, the sensitivity $\mathsf{D}_\lambda(\boldsymbol{z}, Q)$ is a means to quantify the change of the expected empirical risk with respect to deviations from the optimal solution of the ERM-RER problem. In the aim of characterizing the sensitivity $\mathsf{D}_\lambda(\boldsymbol{z}, Q)$, consider the following lemma.

Lemma 3.1. *Given two probability measures P and Q over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$, with Q absolutely continuous with P, the following holds for all $\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n$,*

$$\int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta}) \mathrm{d}P(\boldsymbol{\theta}) \leqslant \inf_{t \in (-\infty, 0)} \left(\frac{D(Q\|P) + \log\left(\int \exp\left(t\left(\mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta}) - \mu\right)\right) \mathrm{d}P(\boldsymbol{\theta})\right)}{t} \right) \tag{114}$$

where $\mu = \int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta})$, and the function $\mathsf{L}_{\boldsymbol{u}}$ is defined in (5b).

Proof: The proof of Lemma 3.1 is presented in Appendix U. ■

Lemma 3.1 together with Theorem 2.7 lead to an upper bound on the sensitivity to a change of measure.

Theorem 3.3. *Consider the solution* $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (37) *to ERM-RER problem in* (29) *with* $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ *and* $\mathcal{K}_{\boldsymbol{z}}$ *in* (34)*. Let* Q *be a probability measure over* $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ *absolutely continuous with* P*. Then,*

$$\mathsf{D}_{\lambda}(\boldsymbol{z}, Q) \leqslant \sqrt{2B^2_{\boldsymbol{z}} D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)} \tag{115}$$

where the function D_{λ} *is defined in* (113)*; and the constant* $B_{\boldsymbol{z}}$ *is defined in* (103)*.*

Proof: The proof of Theorem 3.3 is presented in Appendix V. ■

In the context of the ERM problem in (5), Theorem 3.3 establishes an upper and a lower bound on the increase or decrease on the expected empirical risk minimization that can be obtained by deviating from the optimal solution of the ERM-RER in (29). More specifically, note that for all probability measures $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ that are absolutely continuous with P, it holds that,

$$\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) - \sqrt{2B^2_{\boldsymbol{z}} D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)} \leqslant \mathsf{R}_{\boldsymbol{z}}(Q) \leqslant \mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \sqrt{2B^2_{\boldsymbol{z}} D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}. \tag{116}$$

The probability measure Q over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ that achieves the lower bound (or the upper bound) in (116) can be explicitly calculated, as shown by the following theorem.

Theorem 3.4. *Consider the solution* $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *in* (37) *to ERM-RER problem in* (29) *with* $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ *and* $\mathcal{K}_{\boldsymbol{z}}$ *in* (34)*. Consider also the following optimization problem,*

$$\min_{Q \in \triangle_P(\mathcal{M}, \mathscr{B}(\mathcal{M}))} \quad \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}), \tag{117a}$$

$$\text{subject to:} \quad D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant c, \tag{117b}$$

where the function $\mathsf{L}_{\boldsymbol{z}}$ *is in* (5b)*; and the constant* c *is nonnegative. Then, the solution to optimization problem in* (117) *is a Gibbs probability measure* $P^{(\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ *satisfying* (37) *with* $\omega \leqslant \lambda$ *such that*

$$D\left(P^{(\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = c. \tag{118}$$

Proof: The proof of Theorem 3.4 is presented in Appendix W. ■

The relevance of Theorem 3.4 is that it shows that the measure Q that reduces the expected empirical risk beyond the expected empirical risk induced by the Gibbs measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), i.e., the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (61), subject to the constraint $D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant c$, with $c > 0$, is also a Gibbs measure $P^{(\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, with $\omega < \lambda$ chosen to satisfy (118). More specifically, the inequality in the left hand side in (116) is observed with equality when $Q = P^{(\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Similarly, it can be shown that the measure Q that increases the expected empirical risk beyond the expected empirical risk induced by the Gibbs measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), i.e., the value $\mathsf{R}_{\boldsymbol{z}}\left(P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)$ in (61), subject to the constraint $D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) \leqslant c$, with $c > 0$, is also a Gibbs measure $P^{(\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, with $\omega > \lambda$ chosen to satisfy (118), provided that such ω exists in $\mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), c.f., Lemma 2.1.

The following corollary of Theorem 3.3 studies the expectation of the sensitivity with respect to the probability measure $P_{\boldsymbol{Z}}$ in (19).

Corollary 3.2. *For all*

$$\lambda \in \bigcap_{\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n} \mathcal{K}_{\boldsymbol{u}}, \tag{119}$$

with $\mathcal{K}_{\boldsymbol{u}}$ in (34)*, given a probability measure Q over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ absolutely continuous with P, it holds that*

$$\int \mathsf{D}_\lambda(\boldsymbol{u}, Q) \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}) \leqslant \int \sqrt{2B_{\boldsymbol{u}}^2 D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}), \tag{120}$$

where $B_{\boldsymbol{u}}$ is defined in (103)*; the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}$ is the solution to the ERM-RER problem in* (29) *with respect to the data set $\boldsymbol{u}$; and the probability measure $P_{\boldsymbol{Z}}$ is defined in* (19)*.*

The set $\bigcap_{\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n} \mathcal{K}_{\boldsymbol{u}}$ in (122) can be empty for some choices of the σ-finite measure P and empirical loss function $\mathsf{L}_{\boldsymbol{z}}$ in (5b). Nonetheless, from Lemma 2.2, it follows that when P is a probability measure,

$$\bigcap_{\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n} \mathcal{K}_{\boldsymbol{u}} = (0, +\infty). \tag{121}$$

In the following, the expectation of the sensitivity with respect to the measure $P_{\boldsymbol{Z}}$ in (19) is to shown to have an upper-bounded that can be expressed in terms of the lautum information between the models and the data sets.

Theorem 3.5. *For all*

$$\lambda \in \bigcap_{\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n} \mathcal{K}_{\boldsymbol{u}}, \tag{122}$$

with $\mathcal{K}_{\boldsymbol{u}}$ in (34)*, it follows that*

$$\int \mathsf{D}_\lambda\left(\boldsymbol{u}, P^{(\lambda)}_{\boldsymbol{\Theta}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}) \leqslant \sqrt{2B^2 \int D\left(P^{(\lambda)}_{\boldsymbol{\Theta}} \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u})}, \tag{123}$$

where the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}$ is the solution to the ERM-RER problem in (29) *with respect to the data set $\boldsymbol{u}$; the probability measure $P_{\boldsymbol{Z}}$ is defined in* (19)*; the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}}$ is such that for all $\mathcal{A} \in \mathscr{B}(\mathcal{M})$,*

$$P^{(\lambda)}_{\boldsymbol{\Theta}}(\mathcal{A}) = \int P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}(\mathcal{A}) \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}); \tag{124}$$

and the constant B satisfies $B^2 = \sup_{\boldsymbol{u} \in (\mathcal{X} \times \mathcal{Y})^n} B_{\boldsymbol{u}}^2$, with $B_{\boldsymbol{u}}$ defined in (103)*.*

Proof: The proof follows from Corollary 3.5. In particular, from (120), for all probability measures Q over $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ absolutely continuous with P, it holds that

$$\int \mathsf{D}_\lambda(\boldsymbol{u}, Q) \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}) \leqslant \int \sqrt{2B_{\boldsymbol{u}}^2 D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}) \tag{125}$$

$$\leqslant \int \sqrt{2B^2 D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right)} \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}) \tag{126}$$

$$\leqslant \sqrt{2B^2 \int D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u})}. \tag{127}$$

where the inequality in (126) follows from (3.5); and the inequality in (127) follows from Jensen's inequality [3, Theorem 6.3.5].

Finally, from (124) and (127), it holds that

$$\int \mathsf{D}_\lambda\left(\boldsymbol{u}, P_{\boldsymbol{\Theta}}^{(\lambda)}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}) \leqslant \sqrt{2\,B^2\,\mathtt{L}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(\lambda)} P_{\boldsymbol{Z}}\right)}, \tag{128}$$

where the function $\mathtt{L}$ in (128) is the lautum information (Definition 1.2). This completes the proof. ∎

The right-hand side in (123) can be written in terms of the lautum information [33] between the models and the data sets, denoted by $\mathtt{L}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(\lambda)} P_{\boldsymbol{Z}}\right)$, by observing that

$$\mathtt{L}\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(\lambda)} P_{\boldsymbol{Z}}\right) = \int D\left(P_{\boldsymbol{\Theta}}^{(\lambda)} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}^{(\lambda)}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{u}). \tag{129}$$

In a nutshell, it can be concluded that as the expectation of the absolute value of the generalization gap with respect to the measure $P_{\boldsymbol{Z}}$ in (19) is upper bounded in terms of the mutual information between the models and the datasets, c.f., Example 1.1; the expectation of the sensitivity with respect to the measure $P_{\boldsymbol{Z}}$ is upper bounded by the lautum information between the models and the datasets, c.f., Theorem 3.5.

4 Discussion and Final Remarks

The ERM-RER problem in (29) has been studied under the assumption that the reference measure P is a σ-finite measure. In this context, it has been shown that the solution exists and is the unique probability measure whose Radon-Nikodym derivative with respect to P is in Theorem 2.1. Interestingly, the optimal measure is shown to belong to a extended class of Gibbs measures for which the partition function is with respect to a σ-finite measure instead of a probability measure. An important remark is that the choice of the reference measure P plays a central role in the concentration of the probability of the optimal measure; and the mean and variance of the empirical risk when the models are distributed according to the optimal measure. A class of reference measures (coherent measures) for which the highest probability is always allocated to the set of models that are solutions ERM problem (lowest empirical risk) is introduced. More interestingly, it is shown that when the reference measure does not belong to this class, the models that induce the lowest empirical risk are observed with zero probability. Finally, the empirical risk when the models are distributed according to the optimal probability measure is shown to be a sub-Gaussian random variable. This observation is leveraged to study the sensitivity of the ERM-RER problems and unveil the connections between the sensitivity and the lautum information between the models and the data sets.

Appendices

A Proof of Lemma 2.1

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

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

$$\frac{\mathrm{d}}{\mathrm{d}t} \exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) = \frac{1}{t^2} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \geqslant 0, \tag{130}$$

and thus, from (33), it follows that $K_{\boldsymbol{z}}\left(-\frac{1}{b}\right)$ is nondecreasing with b. This implies that $(0, b] \subset \mathcal{K}_{\boldsymbol{z}}$. This proves the convexity of $\mathcal{K}_{\boldsymbol{z}}$.

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

$$b^\star = \sup \mathcal{K}_{\boldsymbol{z}}. \tag{131}$$

Hence, if $b^\star = +\infty$, it follows from (34) that

$$\mathcal{K}_{\boldsymbol{z}} = (0, +\infty). \tag{132}$$

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

$$(0, b^\star) \subset \mathcal{K}_{\boldsymbol{z}}. \tag{133}$$

This completes the first part.

The second part is trivial. Under the assumption that the set $\mathcal{K}_{\boldsymbol{z}}$ in (34) is empty, there is nothing to prove.

This completes the proof.

B Proof of Lemma 2.2

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

$$\exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\right) \leqslant 1, \tag{134}$$

and thus,

$$K_{\boldsymbol{z}}\left(-\frac{1}{t}\right) = \log\left(\int \exp\left(-\frac{1}{t}\, \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right) \mathrm{d}P(\boldsymbol{u})\right) \tag{135}$$

$$\leqslant \log\left(\int \mathrm{d}P(\boldsymbol{u})\right) \tag{136}$$

$$\leqslant 0, \tag{137}$$

which implies that $(0, +\infty) \subseteq \mathcal{K}_{\boldsymbol{z}}$. Thus, from (34), it holds that $\mathcal{K}_{\boldsymbol{z}} = (0, +\infty)$, which completes the proof.

C Proof of Theorem 2.1

The objective function in the optimization problem in (29) can be written as follows:

$$\min_{Q}\left[\int \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)+\lambda\int\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)\right], \tag{138}$$

where the optimization is over all measures Q on $(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right))$ that are absolutely continuous with P and satisfy

$$\int\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)=1, \tag{139}$$

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

Let $\mathscr{M}$ be the set of nonnegative measurable functions with respect to the measurable spaces $(\operatorname{supp}P,\mathscr{B}\left(\operatorname{supp}P\right))$ and $(\mathbb{R},\mathscr{B}\left(\mathbb{R}\right))$. The Lagrangian of the optimization problem in (138) can be constructed in terms of a function in $\mathscr{M}$, instead of a measure in $\triangle\left(\operatorname{supp}P,\mathscr{B}\left(\operatorname{supp}P\right)\right)$. Let such Lagrangian be $L:\mathscr{M}\times[0,+\infty)\to\mathbb{R}$ of the form

$$\begin{aligned}L\left(\frac{\mathrm{d}Q}{\mathrm{d}P},\beta\right)=&\int\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)+\lambda\int\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)\\&+\beta\left(\int\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)-1\right),\end{aligned} \tag{140}$$

where β is a positive real that acts as a Lagrangian multiplier due to the constraint (139).

Let $g:\mathbb{R}^k\to\mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (140) at $\left(\frac{\mathrm{d}Q}{\mathrm{d}P},\beta\right)\in\mathscr{M}\times[0,+\infty)$ in the direction of g is

$$\partial L\left(\frac{\mathrm{d}Q}{\mathrm{d}P},\beta;g\right)\triangleq\frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha)\bigg|_{\alpha=0}, \tag{141}$$

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

$$\begin{aligned}r(\alpha)=&\int\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)+\alpha g\left(\boldsymbol{\nu}\right)\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)+\beta\left(\int\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)+\alpha g\left(\boldsymbol{\nu}\right)\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)-1\right)\\&+\lambda\int\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)+\alpha g\left(\boldsymbol{\nu}\right)\right)\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)+\alpha g\left(\boldsymbol{\nu}\right)\right)\mathrm{d}P\left(\boldsymbol{\nu}\right).\end{aligned} \tag{142}$$

Note that the derivative of the real function r in (142) is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha)=&\int\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)g\left(\boldsymbol{\nu}\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)+\beta\int g\left(\boldsymbol{\nu}\right)\mathrm{d}P\left(\boldsymbol{\nu}\right)\\&+\lambda\int g\left(\boldsymbol{\nu}\right)\left(1+\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)+\alpha g\left(\boldsymbol{\nu}\right)\right)\right)\mathrm{d}P\left(\boldsymbol{\nu}\right).\end{aligned} \tag{143}$$

From (141) and (143), it follows that

$$\partial L\left(\frac{\mathrm{d}Q}{\mathrm{d}P},\beta;g\right)=\int g\left(\boldsymbol{\nu}\right)\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\lambda\left(1+\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)\right)\right)+\beta\right)\mathrm{d}P\left(\boldsymbol{\nu}\right). \tag{144}$$

The relevance of the Gateaux differential in (144) stems from [27, Theorem 1, page 178], which unveils the fact that a necessary condition for the functional L in (140) to have a minimum at $\left(\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}, \beta\right) \in \mathscr{M} \times [0, +\infty)$ is that for all functions $g \in \mathscr{M}$,

$$\partial L \left(\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}; g \right) = 0. \tag{145}$$

From (145), it follows that $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ must satisfy for all functions g in $\mathscr{M}$ that

$$\int g(\boldsymbol{\nu}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \lambda \left(1 + \log \left(\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\nu}) \right) \right) + \beta \right) \mathrm{d}P(\boldsymbol{\nu}) = 0, \tag{146}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) + \lambda \left(1 + \log \left(\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\nu}) \right) \right) + \beta = 0, \tag{147}$$

and thus,

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\nu}) = \exp\left(-\frac{\beta + \lambda}{\lambda} \right) \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda} \right), \tag{148}$$

with β chosen to satisfy (139). That is,

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\nu}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda} \right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda} \right) \mathrm{d}P(\boldsymbol{\theta})} \tag{149}$$

$$= \exp\left(-K_{\boldsymbol{z}}\left(-\frac{1}{\lambda} \right) - \frac{1}{\lambda} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \right). \tag{150}$$

The proof continues by verifying that the objective function in (138) is strictly convex, and thus, the measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ that satisfies (149) is the unique minimizer. More specifically, note that the objective function in (138) is the sum of two terms. The first one, i.e., $\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{\nu}) \mathrm{d}P(\boldsymbol{\nu})$, is linear in $\frac{\mathrm{d}Q}{\mathrm{d}P}$. The second, i.e., $\int \frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{\nu}) \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(\boldsymbol{\nu}) \right) \mathrm{d}P(\boldsymbol{\nu})$, is strictly convex with $\frac{\mathrm{d}Q}{\mathrm{d}P}$. Hence, given that $\lambda > 0$, the sum of both terms is strictly convex with $\frac{\mathrm{d}Q}{\mathrm{d}P}$. This implies the uniqueness of $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and completes the proof.

D Proof of Lemma 2.3

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \geqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}), \tag{151}$$

and thus, for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), it holds that

$$\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \leqslant \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) = \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu})}{\lambda}\right), \tag{152}$$

which implies

$$\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right) \mathrm{d}P(\boldsymbol{\alpha})} \leqslant \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right) \mathrm{d}P(\boldsymbol{\alpha})} = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right) \mathrm{d}P(\boldsymbol{\alpha})}. \tag{153}$$

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

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) \leqslant \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\mu}) = \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\nu}), \tag{154}$$

which completes the proof.

E Proof of Lemma 2.5

From Theorem 2.1, it follows that for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} P$,

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}P(\boldsymbol{\nu})} \tag{155}$$

$$= \left(\exp\left(\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right) \int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right) \mathrm{d}P(\boldsymbol{\nu})\right)^{-1} \tag{156}$$

$$= \left(\int \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu})\right)^{-1}. \tag{157}$$

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

$$\mathcal{A}_0(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = 0\}, \tag{158a}$$

$$\mathcal{A}_1(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < 0\}, \text{ and} \tag{158b}$$

$$\mathcal{A}_2(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > 0\}. \tag{158c}$$

Using the sets $\mathcal{A}_0(\boldsymbol{\theta})$, $\mathcal{A}_1(\boldsymbol{\theta})$, and $\mathcal{A}_2(\boldsymbol{\theta})$ in (157), the following holds for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ and for all $\boldsymbol{\theta} \in \operatorname{supp} P$,

$$\begin{aligned}\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = &\Bigg(\int_{\mathcal{A}_0(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \\ &+ \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \\ &+ \int_{\mathcal{A}_2(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \Bigg)^{-1} \qquad (159) \\ = &\Bigg(P(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \\ &+ \int_{\mathcal{A}_2(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \Bigg)^{-1}. \qquad (160)\end{aligned}$$

Note that the sets $\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star\}$ and $\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star\}$, with $\delta^\star$ in (46), form a partition of the set $\operatorname{supp} P$. Following this observation, the rest of the proof is divided into two parts. The first part evaluates $\lim_{\lambda \to 0^+} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})$, with $\boldsymbol{\theta} \in \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star\}$. The second part considers the case in which $\boldsymbol{\theta} \in \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star\}$.

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

$$\mathcal{A}_0(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = \delta\}, \qquad (161a)$$
$$\mathcal{A}_1(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta\}, \text{ and} \qquad (161b)$$
$$\mathcal{A}_2(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta\}. \qquad (161c)$$

Consider the sets $\mathcal{A}_{2,1}(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\mu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star\}$ and $\mathcal{A}_{2,2}(\boldsymbol{\theta}) = \{\boldsymbol{\mu} \in \mathcal{M} : \delta^\star \leqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta\}$, and note that $\mathcal{A}_{2,1}(\boldsymbol{\theta})$ and $\mathcal{A}_{2,2}(\boldsymbol{\theta})$ form a partition of $\mathcal{A}_2(\boldsymbol{\theta})$, and

$$P(\mathcal{A}_{2,1}(\boldsymbol{\theta})) = 0. \qquad (162)$$

Hence, plugging the equalities in (161) and (162) in (160) yields,

$$\begin{aligned}\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = &\Bigg(P(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \\ &+ \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \Bigg)^{-1}. \qquad (163)\end{aligned}$$

The equality in (163) implies that for all $\delta > \delta^\star$ and for all $\boldsymbol{\theta} \in \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta\}$,

$$\begin{aligned}\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) =& \Bigg(P(\mathcal{A}_0(\boldsymbol{\theta})) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \\ &+ \lim_{\lambda\to 0^+} \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \Bigg)^{-1} && (164)\\ =& \Bigg(P(\mathcal{A}_0(\boldsymbol{\theta})) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_{2,2}(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \Bigg)^{-1} && (165)\\ =& \left(P(\mathcal{A}_0(\boldsymbol{\theta})) + \infty\right)^{-1} && (166)\\ =& 0, && (167)\end{aligned}$$

where the equality in (165) follows verifying that the dominated convergence theorem [3, Theorem 2.6.9] holds. That is,
(a) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that $\exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) < 1$; and
(b) For all $\boldsymbol{\nu} \in \mathcal{A}_1(\boldsymbol{\theta})$, it holds that

$$\lim_{\lambda\to 0^+} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) = 0. \tag{168}$$

This completes the first part of the proof.

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

$$\begin{aligned}\mathcal{A}_0(\boldsymbol{\theta}) &= \{\boldsymbol{\mu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) = \delta^\star\}, && (169a)\\ \mathcal{A}_1(\boldsymbol{\theta}) &= \{\boldsymbol{\mu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) > \delta^\star\}, \text{ and} && (169b)\\ \mathcal{A}_2(\boldsymbol{\theta}) &= \{\boldsymbol{\mu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\mu}) < \delta^\star\}. && (169c)\end{aligned}$$

Observing that $P(\mathcal{A}_2(\boldsymbol{\theta})) = 0$, plugging the equalities in (169) in (160) yields,

$$\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \left(P(\mathcal{A}_0(\boldsymbol{\theta})) + \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \right)^{-1}. \tag{170}$$

The equality in (170) implies that for all $\boldsymbol{\theta} \in \mathcal{L}_{\boldsymbol{z}}(\delta^\star)$,

$$\begin{aligned}\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) =& \left(P(\mathcal{A}_0(\boldsymbol{\theta})) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_1(\boldsymbol{\theta})} \exp\left(\frac{1}{\lambda}\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\right)\right) \mathrm{d}P(\boldsymbol{\nu}) \right)^{-1} \\ =& \frac{1}{P(\mathcal{A}_0(\boldsymbol{\theta}))}, && (171)\end{aligned}$$

where the equality in (171) follows from the same arguments as in (165). This completes the second part.

Finally, from (167) and (171), it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} P$,

$$\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \frac{1}{P(\mathcal{L}^\star_{\boldsymbol{z}})} \mathbb{1}_{\{\boldsymbol{\theta} \in \mathcal{L}^\star_{\boldsymbol{z}}\}}, \tag{172}$$

which completes the proof.

F Proof of Lemma 2.6

For all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), and for all $\mathcal{C} \in \mathscr{B}(\mathcal{M})$,

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = \int_{\mathcal{C}} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}), \tag{173}$$

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

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) = 0, \tag{174}$$

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

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

$$0 = P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{C}) \tag{175}$$

$$= \int_{\mathcal{C}} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}). \tag{176}$$

From Theorem 2.1, it holds that for all $\alpha \in \mathcal{K}_{\boldsymbol{z}}$ and for all $\boldsymbol{\nu} \in \operatorname{supp} P$,

$$\frac{\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\nu}) > 0, \tag{177}$$

which implies that

$$\int_{\mathcal{C}} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) = 0, \tag{178}$$

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

G Proof of Lemma 2.7

The proof that the measure $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to $P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is an immediate consequence of Lemma 2.6 and the Radon-Nikodym theorem [3, Theorem 2.2.1]. More specifically, from Lemma 2.6 it holds that for all $(\alpha, \beta) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$,
(i) the measure $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to P; and
(ii) the measure P is absolutely continuous with respect to $P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

From (i) and the Radon-Nikodym theorem [3, Theorem 2.2.1], it follows that Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ exists.

From (ii) and the Radon-Nikodym theorem [3, Theorem 2.2.1], it follows that the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}$ exists.

Hence, from [28, Theorem 3.41], the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}$ exists and for all $\boldsymbol{\theta} \in \operatorname{supp} P$,

$$\frac{\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) \left(\frac{\mathrm{d}P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) \right)^{-1}, \tag{179}$$

which implies the absolute continuity of the measure $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ with respect to $P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ ([3, Theorem 2.2.1]).

The proof that the measure $P^{(\beta)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with respect to $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is obtained by exchanging the choices of α and β and completes the proof.

H Proof of Lemma 2.8

Consider the transport of the measure P from $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ to $([0,+\infty), \mathscr{B}([0,+\infty)))$ through the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b). Denote the resulting measure in $([0,+\infty), \mathscr{B}([0,+\infty)))$ by P_V. More specifically, for all $\mathcal{A} \in \mathscr{B}([0,+\infty))$, it holds that $P_V(\mathcal{A}) = P(\{\boldsymbol{\theta} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \in \mathcal{A}\})$. Hence, the function $K_{\boldsymbol{z}}$ in (33) can be written for all $t \in \mathbb{R}$ in terms of the measure P_V as follows

$$K_{\boldsymbol{z}}(t) = \log\left(\int \exp(t\,v)\,\mathrm{d}P_V(v) \right). \tag{180}$$

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

$$\phi_V(t) = \int \exp(t\,v)\,\mathrm{d}P_V(v) = \exp(K_{\boldsymbol{z}}(-t)). \tag{181}$$

From [15, Theorem 1a (page 439)], it follows that the function ϕ_V has derivatives of all orders in $(0,+\infty)$, and thus, so does the function $K_{\boldsymbol{z}}$ in $(-\infty,0)$. This implies the continuity of $K_{\boldsymbol{z}}$ in $(-\infty,0)$, and completes the proof.

I Proof of Lemma 2.9

Let $(\gamma_1, \gamma_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$ and $\alpha \in [0,1]$ be fixed. When $\alpha = 0$ or $\alpha = 1$, the proof is trivial. Then, for all $\alpha \in (0,1)$, the following holds

$$
\begin{aligned}
&\alpha K_{\boldsymbol{z}}(\gamma_1) + (1-\alpha) K_{\boldsymbol{z}}(\gamma_2) \\
&= \alpha \log\left(\int \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right) + (1-\alpha) \log\left(\int \exp\left(\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right) && (182)\\
&= \log\left(\left(\int \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right)^{\alpha} \right) + \log\left(\left(\int \exp\left(\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right)^{(1-\alpha)} \right) && (183)\\
&= \log\left(\left(\int \exp\left(\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right)^{\alpha} \left(\int \exp\left(\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right)^{(1-\alpha)} \right) && (184)\\
&= \log\left(\left(\int \exp\left(\alpha\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)^{\frac{1}{\alpha}} \mathrm{d}P(\boldsymbol{u}) \right)^{\alpha} \left(\int \exp\left((1-\alpha)\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)^{\frac{1}{1-\alpha}} \mathrm{d}P(\boldsymbol{u}) \right)^{(1-\alpha)} \right) && (185)\\
&\geqslant \log\left(\int \exp\left(\alpha\gamma_1 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \exp\left((1-\alpha)\gamma_2 \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right) && (186)\\
&= \log\left(\int \exp\left(\left(\alpha\gamma_1 + (1-\alpha)\gamma_2\right) \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u}) \right) \mathrm{d}P(\boldsymbol{u}) \right) && (187)\\
&= K_{\boldsymbol{z}}\left(\alpha\gamma_1 + (1-\alpha)\gamma_2\right), && (188)
\end{aligned}
$$

where the inequality in (186) follows from Hölder's inequality. Note that if the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is a constant, then (186) holds with equality. Alternatively, under the assumption that there exist at least two disjoint subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{M}$ that are nonnegligible with respect to P and satisfy (54) for some positive real $c > 0$, the inequality is strict. This completes the proof.

J Proof of Lemma 2.10

For all $s \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), the equality in (56) implies the following,

$$
\begin{aligned}
K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{s}\right) &= \frac{\mathrm{d}}{\mathrm{d}t} \log\left(\int \exp\left(t \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \right)\Bigg|_{t=-\frac{1}{s}} && (189)\\
&= \frac{1}{\int \exp\left(t \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{v})\right) \mathrm{d}P(\boldsymbol{v})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u}) \exp\left(t \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) \Bigg|_{t=-\frac{1}{s}} && (190)\\
&= \frac{1}{\int \exp\left(-\frac{1}{s} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{v})\right) \mathrm{d}P(\boldsymbol{v})} \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u}) \exp\left(-\frac{1}{s} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) && (191)\\
&= \exp\left(-K_{\boldsymbol{z}}\left(-\frac{1}{s}\right)\right) \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u}) \exp\left(-\frac{1}{s} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) && (192)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u}) \exp\left(-K_{\boldsymbol{z}}\left(-\frac{1}{s}\right) - \frac{1}{s} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right) \mathrm{d}P(\boldsymbol{u}) && (193)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P^{(s)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), && (194)
\end{aligned}
$$

where the equality in (190) holds from the dominated convergence theorem [3]; the equality in (192) follows from (33); and the equality in (194) follows from (37).

For all $s \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), the equalities in (56) and (193) imply that

$$
\begin{aligned}
K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\exp\left(-K_{\boldsymbol{z}}(t)+t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)\mathrm{d}P(\boldsymbol{u})\bigg|_{t=-\frac{1}{s}} && (195)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\left(-K_{\boldsymbol{z}}^{(1)}(t)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)\exp\left(-K_{\boldsymbol{z}}(t)+t\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)\mathrm{d}P(\boldsymbol{u})\bigg|_{t=-\frac{1}{s}} && (196)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\left(-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)\exp\left(-K_{\boldsymbol{z}}\left(-\frac{1}{s}\right)-\frac{1}{s}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)\mathrm{d}P(\boldsymbol{u}) && (197)\\
&= \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\left(-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)+\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}(\boldsymbol{u}) && (198)\\
&= -K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\,\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}(\boldsymbol{u})+\int\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)^2\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}(\boldsymbol{u}) && (199)\\
&= -\left(K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)^2+\int\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})\right)^2\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}(\boldsymbol{u}) && (200)\\
&= \int\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{u})-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)^2\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}(\boldsymbol{u}), && (201)
\end{aligned}
$$

where the equality in (196) follows from the dominated convergence theorem [3]; the equality in (198) is due to a change of measure through the Radon-Nikodym derivative in (37); and the equality in (200) follows from (194).

For all $s \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), the equalities in (56) and (200) imply that

$$K_{\boldsymbol{z}}^{(3)}\left(-\frac{1}{s}\right)=\frac{\mathrm{d}}{\mathrm{d}t}\left(\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{2}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(-\frac{1}{t}\right)}\left(\boldsymbol{u}\right)-\left(K_{\boldsymbol{z}}^{(1)}\left(t\right)\right)^{2}\right)\Bigg|_{t=-\frac{1}{s}} \tag{202}$$

$$=\frac{\mathrm{d}}{\mathrm{d}t}\left(\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{2}\exp\left(-K_{\boldsymbol{z}}\left(t\right)+t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\mathrm{d}P\left(\boldsymbol{u}\right)-\left(K_{\boldsymbol{z}}^{(1)}\left(t\right)\right)^{2}\right)\Bigg|_{t=-\frac{1}{s}} \tag{203}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{2}\left(\frac{\mathrm{d}}{\mathrm{d}t}\exp\left(-K_{\boldsymbol{z}}\left(t\right)+t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\Big|_{t=-\frac{1}{s}}\right)\mathrm{d}P\left(\boldsymbol{u}\right)$$
$$-2K_{\boldsymbol{z}}^{(1)}\left(t\right)K_{\boldsymbol{z}}^{(2)}\left(t\right)\Big|_{t=-\frac{1}{s}} \tag{204}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{2}\left(\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)-K_{\boldsymbol{z}}^{(1)}\left(t\right)\right)\exp\left(-K_{\boldsymbol{z}}\left(t\right)+t\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\Big|_{t=-\frac{1}{s}}\right)\mathrm{d}P\left(\boldsymbol{u}\right)$$
$$-2K_{\boldsymbol{z}}^{(1)}\left(t\right)K_{\boldsymbol{z}}^{(2)}\left(t\right)\Big|_{t=-\frac{1}{s}} \tag{205}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{2}\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)\exp\left(-K_{\boldsymbol{z}}\left(-\frac{1}{s}\right)-\frac{1}{s}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\mathrm{d}P\left(\boldsymbol{u}\right)$$
$$-2K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right) \tag{206}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{2}\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}\left(\boldsymbol{u}\right)$$
$$-2K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right) \tag{207}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{3}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}\left(\boldsymbol{u}\right)$$
$$-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{2}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}\left(\boldsymbol{u}\right)$$
$$-2K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right) \tag{208}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{3}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}\left(\boldsymbol{u}\right)$$
$$-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\left(K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right)+\left(K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)^{2}\right)-2K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right) \tag{209}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)^{3}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}\left(\boldsymbol{u}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)^{3}-3K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{s}\right) \tag{210}$$

$$=\int\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{s}\right)\right)^{3}\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(s)}\left(\boldsymbol{u}\right), \tag{211}$$

where the equality in (203) follows from (37); the equality in (204) follows from the dominated convergence theorem [3]; the equality in (207) follows from (37); and the equality in (209) follows from (200). This completes the proof.

K Proof of Theorem 2.2

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

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \tag{212}$$

$$=\frac{\mathrm{d}}{\mathrm{d}\lambda}\int \frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\nu})}\mathrm{d}P(\boldsymbol{\theta}) \tag{213}$$

$$=\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\frac{\mathrm{d}}{\mathrm{d}\lambda}\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\nu})}\mathrm{d}P(\boldsymbol{\theta}) \tag{214}$$

$$\begin{aligned}=&\frac{1}{\lambda^2}\int\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\nu})}\mathrm{d}P(\boldsymbol{\theta})\\&-\frac{1}{\lambda^2}\int\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\alpha})}{\left(\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\nu})\right)^2}\mathrm{d}P(\boldsymbol{\theta})\end{aligned} \tag{215}$$

$$\begin{aligned}=&\frac{1}{\lambda^2}\int\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)^2\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta})\\&-\frac{1}{\lambda^2}\left(\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\,\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta})\right)^2\end{aligned} \tag{216}$$

$$=\frac{1}{\lambda^2}\int\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})-\int\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)^2\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{217}$$

$$\geqslant 0, \tag{218}$$

where the equality in (214) follows from the dominated convergence theorem [3].

The inequality in (218) implies that the the expected empirical risk $K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)$ in (33) is non-decreasing with respect to λ. Hence, given two reals λ_1 and λ_2, with $\lambda_1 > \lambda_2 > 0$, it holds that

$$K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda_1}\right)\geqslant K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda_2}\right). \tag{219}$$

The rest of the proof consists in showing that for all $\alpha\in\mathcal{K}_{\boldsymbol{z}}$, the function $K_{\boldsymbol{z}}^{(2)}$ in (56) satisfies $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right)>0$ if and only if the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable. Hence, the proof is divided into two parts. In the first part, it is shown that if for all $\alpha\in\mathcal{K}_{\boldsymbol{z}}$, $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right)>0$, then the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable. The second part of the proof, consists in showing that if the function $\mathsf{L}_{\boldsymbol{z}}$ is separable, then, for all $\alpha\in\mathcal{K}_{\boldsymbol{z}}$, $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right)>0$.

The first part is as follows. From Lemma 2.10, it holds that for all $\alpha \in \mathcal{K}_{\boldsymbol{z}}$,

$$K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right)=\int \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{220}$$

$$=\int_{\mathcal{R}_0(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{221}$$

$$+\int_{\mathcal{R}_1(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{222}$$

$$+\int_{\mathcal{R}_2(\alpha)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right), \tag{223}$$

where the sets $\mathcal{R}_0(\alpha)$, $\mathcal{R}_1(\alpha)$, and $\mathcal{R}_2(\alpha)$ are respectively defined in (76). Hence,

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

$$+\int_{\mathcal{R}_2(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \tag{224}$$

$$\leqslant\left(\inf_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right)$$

$$+\left(\sup_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right). \tag{225}$$

Under the assumption that for all $\alpha \in \mathcal{K}_{\boldsymbol{z}}$ the function $K_{\boldsymbol{z}}^{(2)}$ in (56) satisfies $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right) > 0$, it follows from (225) that

$$0<\left(\inf_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right)$$

$$+\left(\sup_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)-K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\right)^2 P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right). \tag{226}$$

Note that if

$$P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right) > 0, \tag{227}$$

then, $\inf_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \neq K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)$. Moreover, if

$$P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_2(\alpha)\right) > 0, \tag{228}$$

then, $\sup_{\boldsymbol{\theta}} \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right) \neq K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)$. Therefore, the inequality in (226) implies that at least one of the following claims is true:

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

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

Nonetheless, from Lemma 2.13, it follows that both claims **(a)** and **(b)** hold simultaneously. Hence, the sets $\mathcal{R}_1(\alpha)$ and $\mathcal{R}_2(\alpha)$ are both nonnegligible with respect to $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ and moreover, it holds that for all $(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \mathcal{R}_1(\alpha) \times \mathcal{R}_2(\alpha)$,

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}_1\right) > K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right) > \mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}_2\right). \tag{229}$$

This proves that under the assumption that for all $\alpha \in \mathcal{K}_{\boldsymbol{z}}$, $K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\alpha}\right) > 0$, the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable. This completes the first part of the proof.

The second part of the proof is simpler. Assume that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable. That is, for all $\gamma \in \mathcal{K}_{\boldsymbol{z}}$, there exist a positive real $c_\gamma > 0$; and two subsets $\mathcal{A}(\gamma)$ and $\mathcal{B}(\gamma)$ of $\mathcal{M}$ that are nonnegligible with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) and verify that for all $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{A}(\gamma) \times \mathcal{B}(\gamma)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > c_\gamma > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2). \tag{230}$$

From Lemma 2.10, it holds that for all $\gamma \in \mathcal{K}_{\boldsymbol{z}}$,

$$\begin{aligned}
K_{\boldsymbol{z}}^{(2)}\left(-\frac{1}{\gamma}\right) &= \int \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (231)\\
&= \int_{\mathcal{A}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (232)\\
&\quad + \int_{\mathcal{B}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (233)\\
&\quad + \int_{(\mathcal{A}(\gamma)\cup\mathcal{B}(\gamma))^{\mathsf{c}}} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) && (234)\\
&> 0, && (235)
\end{aligned}$$

where the inequality (235) follows from the following facts. First, if $c_\gamma < K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)$, with c_γ in (230), then for all $\boldsymbol{\nu} \in \mathcal{B}(\gamma)$, it holds that

$$\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 > \left(c - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2, \tag{236}$$

and thus,

$$\begin{aligned}
\int_{\mathcal{B}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) &> \left(c - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}(\gamma)) && (237)\\
&> 0. && (238)
\end{aligned}$$

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

$$\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 > \left(c - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2, \tag{239}$$

and thus,

$$\begin{aligned}
\int_{\mathcal{B}(\gamma)} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) &> \left(c - K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\gamma}\right)\right)^2 P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}(\gamma)) && (240)\\
&\geqslant 0. && (241)
\end{aligned}$$

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

L Proof of Lemma 2.11

Consider the following sets

$$\mathcal{A}_0 \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star\}, \tag{242}$$
$$\mathcal{A}_1 \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^\star\}, \text{ and} \tag{243}$$
$$\mathcal{A}_2 \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star\}, \tag{244}$$

with $\delta^\star$ in (46) and the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b).

From (57), for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, with $\mathcal{K}_{\boldsymbol{z}}$ in (34), it holds that,

$$K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \int_{\mathcal{A}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{A}_1} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{245}$$
$$= \int_{\mathcal{A}_0} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{246}$$
$$= \delta^\star P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}}) + \int_{\mathcal{A}_2} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{247}$$

where, the equality in (246) follows from noticing that $P(\mathcal{A}_1) = 0$ and thus, $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) = 0$ (Lemma 2.6); and the equality in (247) follows from noticing that $\mathcal{A}_0 = \mathcal{L}^\star_{\boldsymbol{z}}$, with $\mathcal{L}^\star_{\boldsymbol{z}}$ in (47).

The rest of the proof is divided into two parts. The first part considers the case in which the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable with respect to the measure P. The second part considers the converse case.

The first part is as follows. Under the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable with respect to the measure P, if $P(\mathcal{A}_0) > 0$, it follows that there exists a real $c > \delta^\star$, such that $P(\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > c\}) > 0$. By observing that $\{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > c\} \subset \mathcal{A}_2$, it follows that $P(\mathcal{A}_2) > 0$. Moreover, if $P(\mathcal{A}_0) = 0$, it holds that $P(\mathcal{A}_2) = 1$. Hence, from (247), for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, it holds that,

$$K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) > \delta^\star P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}}) + \delta^\star P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{248}$$
$$= \delta^\star, \tag{249}$$

where, the inequality in (248) follows from noticing that for all $\boldsymbol{\theta} \in \mathcal{A}_2$, it holds that $\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) > \delta^\star$.

The second part of the proof is as follows. Under the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is not separable with respect to the measure P, it holds that $P(\mathcal{A}_0) = 1$ and $P(\mathcal{A}_1) = P(\mathcal{A}_2) = 0$. Hence, from (247), it follows that

$$K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) = \delta^\star. \tag{250}$$

This completes the proof.

M Proof of Lemma 2.12

Consider the following sets

$$\mathcal{A}_0 \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) = \delta^\star\}, \tag{251}$$
$$\mathcal{A}_1 \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^\star\}, \text{ and} \tag{252}$$
$$\mathcal{A}_2 \triangleq \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) > \delta^\star\}, \tag{253}$$

with $\delta^\star$ in (46) and the function $\mathsf{L}_{\boldsymbol{z}}$ in (5b).

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

$$1 = P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) + P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{254}$$
$$= P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_2) \tag{255}$$
$$= P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_2} \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{256}$$

where, the equality in (255) follows from noticing that $P(\mathcal{A}_1) = 0$ and thus, $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_1) = 0$ (Lemma 2.6).

The above implies that

$$1 = \lim_{\lambda\to 0^+} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \lim_{\lambda\to 0^+} \int_{\mathcal{A}_2} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \tag{257}$$
$$= \lim_{\lambda\to 0^+} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0) + \int_{\mathcal{A}_2} \lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \tag{258}$$
$$= \lim_{\lambda\to 0^+} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}_0), \tag{259}$$

where, the equality in (258) follows from noticing two facts: (a) For all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, the Randon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ is positive and bounded (Lemma 2.3); and (b) For all $\boldsymbol{\theta} \in \mathcal{A}_2$, it holds that $\lim_{\lambda\to 0^+} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = 0$. Hence, the dominated convergence theorem [3, Theorem 1.6.9] holds.

Finally, by noticing that $\mathcal{A}_0 = \mathcal{L}^\star_{\boldsymbol{z}}$, it holds that

$$\lim_{\lambda\to 0^+} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}}) = 1, \tag{260}$$

which completes the proof.

N Proof of Theorem 2.3

From (247) in the proof of Lemma 2.11, it holds that

$$\lim_{\lambda\to 0^+} K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\lambda}\right)=\lim_{\lambda\to 0^+}\delta^\star P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}})+\lim_{\lambda\to 0^+}\int_{\mathcal{A}_2}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \tag{261}$$

$$=\lim_{\lambda\to 0^+}\delta^\star P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}})+\lim_{\lambda\to 0^+}\int_{\mathcal{A}_2}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}\left(\boldsymbol{\theta}\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{262}$$

$$=\lim_{\lambda\to 0^+}\delta^\star P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}})+\int_{\mathcal{A}_2}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)\lim_{\lambda\to 0^+}\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}\left(\boldsymbol{\theta}\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{263}$$

$$=\delta^\star\lim_{\lambda\to 0^+}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}}) \tag{264}$$

$$=\delta^\star, \tag{265}$$

where, the equality in (263) follows from noticing two facts: (*a*) For all $\lambda\in\mathcal{K}_{\boldsymbol{z}}$, the Randon-Nikodym derivative $\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}$ is positive and bounded (Lemma 2.3); and (*b*) For all $\boldsymbol{\theta}\in\mathcal{A}_2$, it holds that $\lim_{\lambda\to 0^+}\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}\left(\boldsymbol{\theta}\right)=0$. Hence, the dominated convergence theorem [3, Theorem 1.6.9] holds. The inequality in (264) follows from Lemma 2.12 This completes the proof.

O Proof of Lemma 2.13

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

The first part is proved by contradiction. Assume that set $\mathcal{R}_2\left(\alpha\right)$ is negligible with respect to $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Hence, from Lemma 2.10, it holds that

$$K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)=\int_{\mathcal{R}_0(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\int_{\mathcal{R}_1(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right)$$
$$+\int_{\mathcal{R}_2(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \tag{266}$$

$$=\int_{\mathcal{R}_0(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\int_{\mathcal{R}_1(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \tag{267}$$

$$=K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0(\alpha)\right)+\int_{\mathcal{R}_1(\alpha)}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right) \tag{268}$$

$$<K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0(\alpha)\right)+K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right) \tag{269}$$

$$=K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right)\left(P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_0(\alpha)\right)+P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\mathcal{R}_1(\alpha)\right)\right) \tag{270}$$

$$=K_{\boldsymbol{z}}^{(1)}\left(-\frac{1}{\alpha}\right), \tag{271}$$

which is a contradiction.

The second part of the proof follows the same arguments as in the first part. Assume that the set $\mathcal{R}_1(\alpha)$ is negligible with respect to $P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Hence, from Lemma 2.10, it holds that

$$\begin{aligned}
K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) =& \int_{\mathcal{R}_0(\alpha)} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) + \int_{\mathcal{R}_1(\alpha)} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \\
&+ \int_{\mathcal{R}_2(\alpha)} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) && (272)\\
=& \int_{\mathcal{R}_0(\alpha)} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) + \int_{\mathcal{R}_2(\alpha)} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) && (273)\\
=& K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\alpha)) + \int_{\mathcal{R}_2(\alpha)} \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) && (274)\\
>& K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\alpha)) + K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\alpha)) && (275)\\
=& K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\alpha}\right) \left(P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\alpha)) + P^{(\alpha)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\alpha))\right) && (276)\\
=& K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\alpha}\right), && (277)
\end{aligned}$$

which is also a contradiction. This completes the proof.

P Proof of Theorem 2.4

From Theorem 2.2, it follows that for all $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \frac{\mathrm{d}P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha})\, \mathrm{d}P(\boldsymbol{\alpha}) \geqslant \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \frac{\mathrm{d}P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha})\, \mathrm{d}P(\boldsymbol{\alpha}), \tag{278}$$

which implies the following inclusions:

$$\mathcal{R}_1(\lambda_2) \subseteq \mathcal{R}_1(\lambda_1), \text{ and} \tag{279a}$$
$$\mathcal{R}_2(\lambda_1) \subseteq \mathcal{R}_2(\lambda_2), \tag{279b}$$

with the sets $\mathcal{R}_1(\cdot)$ and $\mathcal{R}_2(\cdot)$ in (76). From (81), it holds that for all $i \in \{1, 2\}$,

$$\mathcal{N}(\lambda_i) = \mathcal{R}_2(\lambda_i)^{\mathsf{c}}, \tag{280}$$

where the complement is with respect to $\mathcal{M}$. Thus, the inclusion in (279b) and the equality in (280) yields,

$$\mathcal{N}(\lambda_1) \supseteq \mathcal{N}(\lambda_2). \tag{281}$$

The inclusion $\mathcal{M} \supseteq \mathcal{N}(\lambda_1)$ follows from (81). Alternatively, the inclusion $\mathcal{N}(\lambda_2) \supseteq \mathcal{N}^\star$, follows from observing that for all $\boldsymbol{\nu} \in \mathcal{N}^\star$,

$$\begin{aligned}
\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \frac{\mathrm{d}P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha})\, \mathrm{d}P(\boldsymbol{\alpha}) &\geqslant \delta^\star \int \frac{\mathrm{d}P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha})\, \mathrm{d}P(\boldsymbol{\alpha}) && (282)\\
&\geqslant \delta^\star && (283)\\
&\geqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}), && (284)
\end{aligned}$$

which implies that $\boldsymbol{\nu} \in \mathcal{N}(\lambda_2)$. This completes the proof of (85).

The proof of (86) is as follows. From the mean value theorem [3] and the assumption that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is continuous on $\mathcal{M}$, it follows that for all $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, there always exists a model $\boldsymbol{\theta} \in \mathcal{M}$, such that

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}), \tag{285}$$

which implies that $\mathcal{R}_0(\lambda)$ is not empty, and as a consequence, $\mathcal{N}(\lambda) = \mathcal{R}_0(\lambda_1) \cup \mathcal{R}_1(\lambda_1)$ is not empty. Hence, for all $\boldsymbol{\theta} \in \mathcal{R}_0(\lambda_1)$ it holds that $\boldsymbol{\theta} \notin \mathcal{N}(\lambda_2)$. This proves that the elements of $\mathcal{R}_0(\lambda_1)$ are in $\mathcal{N}(\lambda_1)$ but not in $\mathcal{N}(\lambda_2)$. This, together with (281), verifies that

$$\mathcal{N}(\lambda_1) \supset \mathcal{N}(\lambda_2). \tag{286}$$

The strict inclusion $\mathcal{M} \supset \mathcal{N}(\lambda_1)$ is proved by contradiction. Assume that there exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ such that $\mathcal{M} = \mathcal{N}(\lambda)$. Then, $\mathcal{R}_2(\lambda) = \emptyset$ and thus, $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\lambda)) = 0$, which together with Lemma 2.13, implies that $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\lambda)) = 0$ and consequently,

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\lambda)) = 1. \tag{287}$$

This contradicts the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable (Definition 2.2). Hence, $\mathcal{M} \supset \mathcal{N}(\lambda_1)$.

Finally, the strict inclusion $\mathcal{N}(\lambda_2) \supset \mathcal{N}^\star$ is proved by contradiction. Assume that there exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ such that $\mathcal{N}^\star = \mathcal{N}(\lambda)$. Hence, three cases might arise:
(a) there exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, such that $\delta^\star < K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ and the set $\left\{\boldsymbol{\nu} \in \mathcal{M} : \delta^\star < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right\} = \emptyset$;
(b) there exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, such that $\delta^\star > K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$ and the set $\left\{\boldsymbol{\nu} \in \mathcal{M} : K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant \delta^\star\right\} = \emptyset$; or
(c) there exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$, such that $\delta^\star = K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)$.

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

$$\mathcal{R}_1(\lambda) = \{\boldsymbol{\nu} \in \mathcal{M} : \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) < \delta^\star\}, \tag{288}$$

which implies that

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\lambda)) = 0. \tag{289}$$

From Lemma 2.13 and (289), it follows that,

$$P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\lambda)) = 0. \tag{290}$$

Finally, by noticing that

$$1 = P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\lambda)) + P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\lambda)) + P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\lambda)) \tag{291}$$

$$= P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\lambda)), \tag{292}$$

leads to the conclusion that the assumption that there exists a $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ such that $\mathcal{N}^\star = \mathcal{N}(\lambda)$ is a contradiction to the assumption that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable (Definition 2.2). Hence, $\mathcal{N}(\lambda_2) \supset \mathcal{T}(\boldsymbol{z})$, which completes the proof of (86).

Q Proof of Theorem 2.5

The proof of (87) is based on the analysis of the derivative of $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})$ with respect to λ, for some fixed set $\mathcal{A} \subseteq \mathcal{M}$. More specifically, given a $\gamma \in \mathcal{K}_{\boldsymbol{z}}$, it holds that

$$P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) = \int_{\mathcal{A}} \frac{\mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha})\,\mathrm{d}P(\boldsymbol{\alpha}), \tag{293}$$

and from the fundamental theorem of calculus [38], it follows that for all $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) - P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) = \int_{\lambda_2}^{\lambda_1} \frac{\mathrm{d}}{\mathrm{d}\gamma} P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\,\mathrm{d}\gamma \tag{294}$$

$$= \int_{\lambda_2}^{\lambda_1} \frac{\mathrm{d}}{\mathrm{d}\gamma} \int_{\mathcal{A}} \frac{\mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha})\,\mathrm{d}P(\boldsymbol{\alpha})\,\mathrm{d}\gamma \tag{295}$$

$$= \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{A}} \frac{\mathrm{d}}{\mathrm{d}\gamma} \frac{\mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha})\,\mathrm{d}P(\boldsymbol{\alpha})\,\mathrm{d}\gamma, \tag{296}$$

where the equality in (295) follows from (293); and the equality in (296) holds from Theorem 2.1 and the dominated convergence theorem [3].

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

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\nu})} \tag{297}$$

$$= \frac{\frac{1}{\lambda^2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\nu})} - \frac{\frac{1}{\lambda^2}\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\alpha})}{\left(\int \exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)\mathrm{d}P(\boldsymbol{\nu})\right)^2} \tag{298}$$

$$= \frac{1}{\lambda^2}\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) - \frac{1}{\lambda^2} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\nu})\,\mathrm{d}P(\boldsymbol{\nu}) \tag{299}$$

$$= \frac{1}{\lambda^2} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\theta}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right). \tag{300}$$

Plugging (300) into (296) yields,

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

$$= \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{A}} \frac{1}{\gamma^2} \frac{\mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,\mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}P(\boldsymbol{\alpha})\,\mathrm{d}\gamma \tag{301}$$

$$= \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{A}} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\,\mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right)\mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\,\mathrm{d}\gamma. \tag{302}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant 0, \tag{303}$$

and thus,

$$\int_{\mathcal{N}(\lambda_2)} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \leqslant 0. \tag{304}$$

The equalities in (302) and (304), with $\mathcal{A} = \mathcal{N}(\lambda)$, imply that

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)) - P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)) \leqslant 0. \tag{305}$$

The inequality $0 < P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2))$ in (87) is proved by contradiction. Assume that for some $\lambda \in \mathcal{K}_{\boldsymbol{z}}$ it holds that $0 = P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2))$. Then, $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_0(\lambda_2)) + P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\lambda_2)) = 0$, which implies that $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\lambda_2)) = 1$, which is a contradiction. See for instance, Lemma 2.14. This completes the proof of (87).

The proof of (88) is divided into two parts. The first part shows that if for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)) < P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)), \tag{306}$$

then the function $\mathsf{L}_{\boldsymbol{z}}$ is separable. The second part of the proof shows that if the function $\mathsf{L}_{\boldsymbol{z}}$ is separable, then, for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$, the inequality in (306) holds.

The first part is as follows. In the proof of Theorem 2.4 it is shown (see (302)) that for all pairs $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$ with $\lambda_1 > \lambda_2$,

$$\begin{aligned} & P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)) - P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)) \\ = & \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{N}(\lambda_2)} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\gamma. \end{aligned} \tag{307}$$

Assume that for a given pair $(\lambda_1, \lambda_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, with $\lambda_1 > \lambda_2$, the inequality in (306) holds. Then, from (307),

$$\begin{aligned} 0 > & \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{N}(\lambda_2)} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\gamma \\ = & \int_{\lambda_2}^{\lambda_1} \int_{\mathcal{R}_1(\lambda_2)} \frac{1}{\gamma^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\gamma \end{aligned} \tag{308}$$

where the equality in (308) follows from noticing that $\mathcal{R}_0(\lambda_2)$ and $\mathcal{R}_1(\lambda_2)$ form a partition of $\mathcal{N}(\lambda_2)$, with the sets $\mathcal{R}_0(\lambda_2)$, $\mathcal{R}_1(\lambda_2)$ and $\mathcal{N}(\lambda_2)$ defined in (76a), (76b), and (81), respectively.

The inequality in (308) implies that the set $\mathcal{R}_1(\lambda_2)$ is nonnegligible with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, for some $\gamma \in (\lambda_2, \lambda_1)$. Hence, from Lemma 2.14, it follows that both sets $\mathcal{R}_1(\lambda_2)$ and $\mathcal{R}_2(\lambda_2)$ are nonnegligible with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$.

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

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)) < P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)), \tag{309}$$

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

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2), \tag{310}$$

which together with Definition 2.3 verify that the function $\mathsf{L}_{\boldsymbol{z}}$ is separable. This ends the first part of the proof.

The second part of the proof is under the assumption that the empirical risk function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable. That is, from Definition 2.3, for all $\gamma \in \mathcal{K}_{\boldsymbol{z}}$, there exist a positive real $c_\gamma > 0$ and two subsets $\mathcal{A}(\gamma)$ and $\mathcal{B}(\gamma)$ of $\mathcal{M}$ nonnegligible with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37) that verify that for all $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) \in \mathcal{A}(\gamma) \times \mathcal{B}(\gamma)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_1) > c_\gamma > \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}_2). \tag{311}$$

In the proof of Theorem 2.4, c.f. (302), it has been proved that given a pair $(\alpha_1, \alpha_2) \in \mathcal{K}_{\boldsymbol{z}} \times \mathcal{K}_{\boldsymbol{z}}$, with $\alpha_1 > \gamma > \alpha_2$, it holds that for all subsets $\mathcal{A}$ of $\mathcal{M}$,

$$\begin{aligned}
&P^{(\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) - P^{(\alpha_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}) \\
&= \int_{\alpha_2}^{\alpha_1} \int_{\mathcal{A}} \frac{1}{\lambda^2} \frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}(\boldsymbol{\alpha}) \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right) \mathrm{d}P(\boldsymbol{\alpha})\, \mathrm{d}\lambda \qquad (312)\\
&= \int_{\alpha_2}^{\alpha_1} \int_{\mathcal{A}} \frac{1}{\lambda^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})\, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})\right) \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha})\, \mathrm{d}\lambda. \qquad (313)
\end{aligned}$$

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

$$c_\gamma < K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\gamma}\right), \tag{314}$$

with c_γ in (311). The second case considers that

$$c_\gamma \geqslant K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\gamma}\right). \tag{315}$$

In the first case, it follows from (81) that

$$\mathcal{B}(\gamma) \subset \mathcal{N}(\gamma), \tag{316}$$

which implies that

$$\begin{aligned}
P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\gamma)) &\geqslant P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{B}(\gamma)) \qquad (317)\\
&> 0, \qquad (318)
\end{aligned}$$

where, the inequality in (318) follows from the fact that $\mathcal{B}(\gamma)$ is nonnegligible with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. This implies that the set $\mathcal{N}_1(\gamma)$ is not negligible with respect $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Moreover, from (81) and (316), it follows that for all $\boldsymbol{\alpha} \in \mathcal{N}(\gamma)$ and for all $\lambda \in (\gamma, \alpha_1)$,

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) < \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - c_\gamma \tag{319}$$

$$< 0, \tag{320}$$

where the inequality in (319) follows from (314); and the inequality in (320) follows from (311). Thus,

$$\int_\gamma^{\alpha_1} \int_{\mathcal{N}(\gamma)} \frac{1}{\lambda^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\lambda < 0, \tag{321}$$

which implies, from (313), that

$$P^{(\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\gamma)) - P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\gamma)) < 0. \tag{322}$$

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

$$A(\gamma) \subseteq \mathcal{R}_2(\gamma), \tag{323}$$

which implies that

$$P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_2(\gamma)) \geqslant P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A}(\gamma)) \tag{324}$$

$$> 0, \tag{325}$$

where the inequality in (325) follows from the fact that $\mathcal{A}(\gamma)$ is nonnegligible with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. This implies that the set $\mathcal{R}_2(\gamma)$ is not negligible with respect $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. From Lemma 2.13, it follows that both $\mathcal{R}_1(\gamma)$ and $\mathcal{R}_2(\gamma)$ are nonnegligible with respect to $P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Using this result, the following holds,

$$P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\gamma)) \geqslant P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{R}_1(\gamma)) \tag{326}$$

$$> 0, \tag{327}$$

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

From (81) and Theorem 2.2, it follows that for all $\boldsymbol{\alpha} \in \mathcal{N}(\gamma)$ and for all $\lambda \in (\gamma, \alpha_1)$,

$$0 \geqslant \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\gamma)}_{\boldsymbol{\Theta}|\underline{\boldsymbol{X}}=\boldsymbol{z}=\boldsymbol{y}}(\boldsymbol{\nu}) \tag{328}$$

$$> \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\underline{\boldsymbol{X}}=\boldsymbol{z}=\boldsymbol{y}}(\boldsymbol{\nu}). \tag{329}$$

Thus,

$$\int_\gamma^{\alpha_1} \int_{\mathcal{N}(\gamma)} \frac{1}{\lambda^2} \left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\alpha}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \, \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \right) \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\alpha}) \, \mathrm{d}\lambda < 0, \tag{330}$$

which implies, from (313), that

$$P^{(\alpha_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\gamma)) - P^{(\gamma)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\gamma)) < 0. \tag{331}$$

The inequality $0 < P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2))$ in (88) has already been proved while proving (87), and thus, this completes the proof of (88).

R Proof of Lemma 2.15

The proof is based on the observation that

$$\mathcal{N}(\lambda_1)=\mathcal{N}(\lambda_2)\cup(\mathcal{N}(\lambda_1)\cap\mathcal{R}_2(\lambda_2)), \tag{332}$$

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

$$\begin{aligned} P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_1))&=P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}(\lambda_2)\cup(\mathcal{N}(\lambda_1)\cap\mathcal{R}_2(\lambda_2))\Big) &(333)\\ &=P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}(\lambda_2)\Big)+P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}(\lambda_1)\cap\mathcal{R}_2(\lambda_2)\Big) &(334)\\ &=P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{N}(\lambda_2)\Big), &(335)\end{aligned}$$

where the equality in (334) follows from Lemma 2.6 and the equality in (89).

Note that the sets $\mathcal{N}(\lambda_2)$, $\mathcal{N}(\lambda_1)\cap\mathcal{R}_2(\lambda_2)$ and $\mathcal{R}_1(\lambda_1)$ form a partition of $\mathcal{M}$ and thus, the equality in (89) implies that for all $i\in\{1,2\}$,

$$\begin{aligned} 0<P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2))&=1-P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{R}_1(\lambda_1)\Big) &(336)\\ &=P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{R}_0(\lambda_1)\Big)+P^{(\lambda_i)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\Big(\mathcal{R}_2(\lambda_1)\Big), &(337)\end{aligned}$$

where the inequality in (336) holds from Theorem 2.5. Finally, under the assumption that the empirical function $\mathsf{L}_{\boldsymbol{z}}$ in (5b) is separable, it holds from Theorem 2.5 that

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2))<P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)). \tag{338}$$

Plugging (335) into (338), with $i=1$, yields,

$$P^{(\lambda_1)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_1))<P^{(\lambda_2)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda_2)), \tag{339}$$

and this completes the proof.

S Proof of Theorem 2.6

$$\begin{aligned} \lim_{\lambda\to0^+}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda))&=\lim_{\lambda\to0^+}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}})+\lim_{\lambda\to0^+}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{N}(\lambda)) &(340)\\ &=\lim_{\lambda\to0^+}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{L}^\star_{\boldsymbol{z}}) &(341)\\ &=\begin{cases}0 & \text{if}\quad \mathcal{L}^\star_{\boldsymbol{z}}=\emptyset\\ 1 & \text{if}\quad \mathcal{L}^\star_{\boldsymbol{z}}\neq\emptyset,\end{cases} &(342)\end{aligned}$$

where, the equalities in (341) follows Lemma 2.12. This completes the proof.

T Proof of Lemma 2.16

The cumulant generating function $J_{\boldsymbol{z},\lambda}$ in (98) induced by the measure $P^{(\lambda)}_{W|\boldsymbol{Z}=\boldsymbol{z}}$ in (96) evaluated at t, with $t \leqslant \frac{1}{\lambda}$, is

$$J_{\boldsymbol{z},\lambda}(t) = \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}\left(\boldsymbol{u}\right)\mathrm{d}P\left(\boldsymbol{u}\right)\right) \tag{343}$$

$$= \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\exp\left(-K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right)\right)\mathrm{d}P\left(\boldsymbol{u}\right)\right) \tag{344}$$

$$= \log\left(\int \exp\left(\left(t-\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right) - K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\mathrm{d}P\left(\boldsymbol{u}\right)\right) \tag{345}$$

$$= \log\left(\int \exp\left(\left(t-\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right) - K_{\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right) + K_{\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right) - K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\mathrm{d}P\left(\boldsymbol{u}\right)\right) \tag{346}$$

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

$$= + \log\left(\int \exp\left(\left(t-\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{u}\right) - K_{\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right)\right)\mathrm{d}P\left(\boldsymbol{u}\right)\right) \tag{347}$$

$$= K_{\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right) - K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) + \log\left(\int \frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{\left(-\frac{1}{t-\frac{1}{\lambda}}\right)}}{\mathrm{d}P}\left(\boldsymbol{u}\right)\mathrm{d}P\left(\boldsymbol{u}\right)\right) \tag{348}$$

$$= K_{\boldsymbol{z}}\left(t-\frac{1}{\lambda}\right) - K_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right), \tag{349}$$

where the equality in (344) follows from Theorem 2.1; and the equality in (348) follows from the fact that $-\frac{1}{t-\frac{1}{\lambda}} > 0$ for all $t < \frac{1}{\lambda}$. This completes the proof.

U Proof of Lemma 3.1

From [9, Corollary 4.15, Page 100], it follows that the probability measures P and Q in $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ satisfy the following equality:

$$D\left(Q\|P\right) = \sup_{f} \int f\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) - \log\int \exp\left(f\left(\boldsymbol{\theta}\right)\right)\mathrm{d}P\left(\boldsymbol{\theta}\right), \tag{350}$$

where the supremum is over the space of all measurable functions f with respect to $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, such that $\int \exp\left(f\left(\boldsymbol{\theta}\right)\right)\mathrm{d}P\left(\boldsymbol{\theta}\right) < \infty$. Hence, for all $\boldsymbol{u} \in \left(\mathcal{X}\times\mathcal{Y}\right)^n$ and for all $t \in (-\infty, 0)$, it follows that the empirical risk function $\mathsf{L}_{\boldsymbol{u}}$ in (5b) satisfies that

$$D\left(Q\|P\) \geqslant \int t\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) - \log\int \exp\left(t\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right)\right)\mathrm{d}P\left(\boldsymbol{\theta}\right) \tag{351}$$

$$\geqslant \int t\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) - \log\int \exp\left(t\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right) + t\mu - t\mu\right)\mathrm{d}P\left(\boldsymbol{\theta}\right). \tag{352}$$

$$= \int t\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right)\mathrm{d}Q\left(\boldsymbol{\theta}\right) - \int t\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right)\mathrm{d}P\left(\boldsymbol{\theta}\right) - \log\int \exp\left(t\mathsf{L}_{\boldsymbol{u}}\left(\boldsymbol{\theta}\right) - t\mu\right)\mathrm{d}P\left(\boldsymbol{\theta}\right), \tag{353}$$

which leads to

$$\int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \leqslant \frac{D(Q\|P) + \log \int \exp\left(t\left(\mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta}) - \mu\right)\right)\mathrm{d}P(\boldsymbol{\theta})}{t}. \quad (354)$$

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

$$\int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta})\,\mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta})\,\mathrm{d}P(\boldsymbol{\theta}) \leqslant \inf_{t\in(-\infty,0)} \frac{D(Q\|P) + \log \int \exp\left(t\left(\mathsf{L}_{\boldsymbol{u}}(\boldsymbol{\theta}) - \mu\right)\right)\mathrm{d}P(\boldsymbol{\theta})}{t}, \quad (355)$$

which completes the proof.

V Proof of Theorem 3.3

From Lemma 3.1, it holds that the probability measure $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ in (37), satisfies

$$\begin{aligned}
&\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \\
&\leqslant \inf_{t\in(-\infty,0)} \left(\frac{D\left(Q\|P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \log\left(\int \exp\left(t\,\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{\boldsymbol{z}}\left(-\tfrac{1}{\lambda}\right)\right)\right)\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right)}{t} \right), \quad (356)
\end{aligned}$$

where the function $K^{(1)}_{\boldsymbol{z}}$ is defined in (57). Moreover, for all $t \in (-\infty, 0)$,

$$\begin{aligned}
&\log\left(\int \exp\left(t\,\left(\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) - K^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)\right)\right)\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right) \\
&= \log\left(\int \exp\left(t\,\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\right)\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\right) - tK^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) && (357) \\
&= J_{\boldsymbol{z},\lambda}(t) - tK^{(1)}_{\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) && (358) \\
&\leqslant \frac{1}{2}t^2 B^2_{\boldsymbol{z}}, && (359)
\end{aligned}$$

where the inequality in (358) follows from (98); the inequality in (359) follows from Theorem 2.7; and the constant $B_{\boldsymbol{z}}$ is defined in (103).

Plugging (359) into (356) yields for all $t \in (-\infty, 0)$,

$$\int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \leqslant \inf_{t\in(-\infty,0)} \frac{D\left(Q\|P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) + \frac{1}{2}t^2 B^2_{\boldsymbol{z}}}{t}: \quad (360)$$

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

$$c \triangleq \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). \quad (361)$$

Hence, for all $t \in (-\infty, 0)$,

$$c\,t - \frac{1}{2}t^2 B^2_{\boldsymbol{z}} \leqslant D\left(Q\|P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \quad (362)$$

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

$$\phi(\alpha) = \frac{1}{2}\alpha^2 B^2_{\boldsymbol{z}}, \quad (363)$$

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

$$\phi^*(x) = \max_{t \in (-\infty, 0)} xt - \phi(t). \tag{364}$$

In particular, note that

$$\phi^*(c) \leqslant D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{365}$$

Note that for all $x \in \mathbb{R}$ and for all $t \in (-\infty, 0)$, the function $\phi^\star$ in (364) satisfies

$$x\,t - \frac{1}{2}t^2 B_{\boldsymbol{z}}^2 \leqslant \phi^\star(x) = x\alpha^\star(x) - \phi\left(\alpha^\star(x)\right), \tag{366}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(x\alpha - \phi\left(\alpha\right)\right) = x - \alpha B_{\boldsymbol{z}}^2 = 0. \tag{367}$$

That is,

$$\alpha^\star(x) = \frac{x}{B_{\boldsymbol{z}}^2}. \tag{368}$$

Plugging (368) into (366) yields,

$$\phi^\star(x) = \frac{x^2}{2B_{\boldsymbol{z}}^2}. \tag{369}$$

Hence, from (365) and (366), given c in (361) for all $t \in (-\infty, 0)$,

$$c\,t - \frac{1}{2}t^2 B_{\boldsymbol{z}}^2 \leqslant \phi^\star(c) \leqslant D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right), \tag{370}$$

and thus,

$$\frac{c^2}{2B_{\boldsymbol{z}}^2} \leqslant D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right). \tag{371}$$

This implies that either,

$$c \leqslant \sqrt{2B_{\boldsymbol{z}}^2 D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)} \tag{372}$$

or

$$c \geqslant -\sqrt{2B_{\boldsymbol{z}}^2 D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}, \tag{373}$$

which leads to

$$\left| \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}Q(\boldsymbol{\theta}) - \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right| \leqslant \sqrt{2B_{\boldsymbol{z}}^2 D\left(Q \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)}, \tag{374}$$

and completes the proof.

W Proof of Theorem 3.4

The optimization problem in (117) can be re-written as follows:

$$\min_{Q} \quad \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) \frac{\mathrm{d}Q}{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu}) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}), \tag{375a}$$

$$\text{subject to:} \quad \int \frac{\mathrm{d}Q}{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu}) \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}(\boldsymbol{\nu})\right) \mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \leqslant c, \tag{375b}$$

where the optimization is over all measures Q on $(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ that are absolutely continuous with $P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu})$ and satisfy

$$\int \mathrm{d}Q(\boldsymbol{\nu}) = 1. \tag{376}$$

Let $\mathscr{M}$ be the set of nonnegative measurable functions with respect to the measurable spaces $\left(\operatorname{supp} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, \mathscr{B}\left(\operatorname{supp} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right)$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. The Lagrangian of the optimization problem in (375) can be constructed in terms of a function in $\mathscr{M}$, instead of a measure over the measurable space $\left(\operatorname{supp} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, \mathscr{B}\left(\operatorname{supp} P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)\right)$. Let such Lagrangian $L : \mathscr{M} \times [0, +\infty)^2 \to \mathbb{R}$ be of the form

$$\begin{aligned} L(g, \alpha, \beta) = & \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) g(\boldsymbol{\nu}) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) + \alpha\left(\int g(\boldsymbol{\nu}) \log(g(\boldsymbol{\nu})) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) - c\right) \\ & + \beta\left(\int g(\boldsymbol{\nu}) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) - 1\right), \end{aligned} \tag{377}$$

where g is a notation to represent the Radon-Nikodym derivative $\frac{\mathrm{d}Q}{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}$; the reals α and β are both nonnegative and act as Lagrangian multipliers due to the constraint (375b) and (139), respectively.

Let $h : \mathbb{R}^k \to \mathbb{R}$ be a function in $\mathscr{M}$. The Gateaux differential of the functional L in (377) at $(g, \alpha, \beta) \in \mathscr{M} \times [0, +\infty)^2$ in the direction of h is

$$\partial L(g, \alpha, \beta; h) \triangleq \frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma)\Big|_{\gamma=0}, \tag{378}$$

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

$$\begin{aligned} r(\gamma) = & \int \mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu}) (g(\boldsymbol{\nu}) + \gamma h(\boldsymbol{\nu})) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) + \\ & \alpha\left(\int (g(\boldsymbol{\nu}) + \gamma h(\boldsymbol{\nu})) \log(g(\boldsymbol{\nu}) + \gamma h(\boldsymbol{\nu})) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) - c\right) \\ & + \beta\left(\int (g(\boldsymbol{\nu}) + \gamma h(\boldsymbol{\nu})) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) - 1\right), \end{aligned} \tag{379}$$

Note that the derivative of the real function r in (380) is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\gamma} r(\gamma) = & \int \mathsf{L}_{\boldsymbol{z}} h(\boldsymbol{\nu}) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \\ & + \alpha \int h(\boldsymbol{\nu}) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) + \alpha \int h(\boldsymbol{\nu}) \log(g(\boldsymbol{\nu}) + \gamma h(\boldsymbol{\nu})) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \\ & + \beta \int h(\boldsymbol{\nu}) \,\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}). \end{aligned} \tag{380}$$

From (378) and (380), it follows that

$$\partial L\left(g,\alpha,\beta;h\right)=\int h\left(\boldsymbol{\nu}\right)\left(\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\alpha\left(1+\log g\left(\boldsymbol{\nu}\right)\right)+\beta\right)\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right). \tag{381}$$

From [27, Theorem 1, page 178], it holds that a necessary condition for the functional L in (377) to have a minimum at $(g,\alpha,\beta)\in\mathscr{M}\times[0,+\infty)^2$ is that for all functions $h\in C\left(\mathcal{M}\right)$

$$\partial L\left(g,\alpha,\beta;h\right)=0, \tag{382}$$

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

$$\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)+\alpha\left(1+\log g\left(\boldsymbol{\nu}\right)\right)+\beta=0. \tag{383}$$

Thus,

$$g\left(\boldsymbol{\nu}\right)=\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)}{\alpha}\right)\exp\left(-\frac{\beta+\alpha}{\alpha}\right), \tag{384}$$

where α and β are chosen to satisfy their corresponding constraints. That is,

$$\frac{\mathrm{d}Q}{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\left(\boldsymbol{\nu}\right)=\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\alpha}\right)\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}, \tag{385}$$

where α is chosen to satisfy

$$D\left(Q\|P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right)=c. \tag{386}$$

From Lemma 2.6, it follows that the probability measure Q and the σ-finite measure P satisfy,

$$\frac{\mathrm{d}Q}{\mathrm{d}P}\left(\boldsymbol{\nu}\right)=\frac{\mathrm{d}Q}{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}\left(\boldsymbol{\nu}\right)\frac{\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}P}\left(\boldsymbol{\nu}\right) \tag{387}$$

$$=\left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\alpha}\right)\mathrm{d}P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}\right)\left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\lambda}\right)\mathrm{d}P\left(\boldsymbol{\theta}\right)}\right) \tag{388}$$

$$=\left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\alpha}\right)}{\int\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\alpha}\right)\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\alpha}\right)}{\lambda}\right)\mathrm{d}P\left(\boldsymbol{\alpha}\right)}\mathrm{d}P\left(\boldsymbol{\theta}\right)}\right)\left(\frac{\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\nu})}{\lambda}\right)}{\int\exp\left(-\frac{\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\theta}\right)}{\lambda}\right)\mathrm{d}P\left(\boldsymbol{\theta}\right)}\right) \tag{389}$$

$$=\frac{\exp\left(-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right)}{\int\exp\left(-\left(\frac{1}{\alpha}+\frac{1}{\lambda}\right)\mathsf{L}_{\boldsymbol{z}}\left(\boldsymbol{\nu}\right)\right)\mathrm{d}P\left(\boldsymbol{\theta}\right)}, \tag{390}$$

which implies that Q is a P-Gibbs probability measure on $(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right))$, with parameter $\frac{\alpha\lambda}{\alpha+\lambda}$. That is, for all $\boldsymbol{\nu}\in\operatorname{supp}P$,

$$Q(\boldsymbol{v})=P^{\left(\frac{\alpha\lambda}{\alpha+\lambda}\right)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\left(\boldsymbol{\nu}\right), \tag{391}$$

where α is chosen to satisfy (386). Let the positive real ω be $\omega \triangleq \frac{\alpha\lambda}{\alpha+\lambda}$. Thus, from Theorem 2.2, it follows that $\omega \in (0, \lambda]$ and satisfies $D\left(P^{(\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\nu}) \| P^{(\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\right) = c$. The proof ends by verifying that the objective function in (377) is strictly convex, and thus, the measure $P^{(\omega)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is the unique minimizer. This completes the proof.

References

[1] I. Alabdulmohsin, "An information-theoretic route from generalization in expectation to generalization in probability," in *Proc. of the 20th International Conference on Artificial Intelligence and Statistics*, Apr. 2017, pp. 92–100.

[2] P. Alquier, J. Ridgway, and N. Chopin, "On the properties of variational approximations of Gibbs posteriors," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, 2016.

[3] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA: Harcourt/Academic Press, 1999.

[4] L. Birge and P. Massart, "Rates of convergence for minimum contrast estimators," *Probability Theory and Related Fields*, vol. 97, pp. 113–150, 1993.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 2006.

[6] L. Boltzmann, *Leçons sur la théorie des gaz (première partie).* Paris, France: Gauthier-Villars, Réédition Jacques Gabay, 1902.

[7] ——, *Leçons sur la théorie des gaz (deuxième partie).* Paris, France: Gauthier-Villars, Réédition Jacques Gabay, 1905.

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

[9] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, 1st ed. Oxford, UK: Oxford University Press, 2013.

[10] O. Catoni, *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, 1st ed. New York, NY, USA: Springer Science & Business Media, 2004, vol. 1851.

[11] ——, *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.

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

[13] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY: Springer-Verlag, 2009.

[14] A. Engel and C. Van den Broeck, *Statistical mechanics of learning.* Cambridge University Press, 2001.

[15] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. New York, NY: Jhon Wiley & Sons, 1971, vol. II.

[16] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 721–741, 1984.

[17] J. W. Gibbs, *Elementary principles in statistical mechanics*, 1st ed. New Haven, NJ: Yale University Press, 1902.

[18] B. Guedj, "A primer on PAC-Bayesian learning." in *Tutorials of the International Conference on Machine Learning (ICML)*, Jun. 2019.

[19] B. Guedj and L. Pujol, "Still no free lunches: The price to pay for tighter PAC-Bayes bounds," *Entropy*, vol. 23, no. 11, 2021.

[20] M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor, "PAC-Bayes unleashed: Generalisation bounds with unbounded losses," *Entropy*, vol. 23, no. 10, 2021.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, and Predicition*, 2nd ed. New York, NY: Springer, 2009.

[22] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," *Neural Information Processing Systems*, 1999.

[23] E. T. Jaynes, "Information theory and statistical mechanics I," *Physical Review Journals*, vol. 106, pp. 620–630, May 1957.

[24] ——, "Information theory and statistical mechanics II," *Physical Review Journals*, vol. 108, pp. 171–190, Oct. 1957.

[25] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*, 1st ed. New York, NY: Wiley, 1989.

[26] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 479–502, 2002.

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

[28] S. M. Perlaza, "INFO5147 Lecture Notes: Selected topics in information theory," École Normale Supérieure (ENS) de Lyon, August 2020.

[29] P. Massart, *Concentration inequalities and model selection*, 1st ed. New York, NY: Springer, 2007.

[30] D. A. McAllester, "Some PAC-Bayesian theorems," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, Jul. 1998, pp. 230–234.

[31] ——, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[32] M. Mezard and A. Montanari, *Information, physics, and computation*, 1st ed. New York, NY, USA: Oxford University Press, 2009.

[33] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

[34] H. V. Poor, *An introduction to signal detection and estimation*, 2nd ed. New York, NY, USA: Springer Science & Business Media, 2013.

[35] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[36] C. P. Robert, *The Bayesian choice: From decision-theoretic foundations to computational implementation*, 1st ed. New York, NY: Springer, 2007.

[37] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*, 2nd ed. New York, NY, USA: Springer, 2004.

[38] W. Rudin, *Principles of mathematical analysis*, 1st ed. New York, NY: McGraw-Hill Book Company, Inc., 1953.

[39] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, Jan. 2019.

[40] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2014.

[41] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proceedings of the tenth annual conference on Computational learning theory*, 1997, pp. 2–9.

[42] L. A. Stefanski and D. D. Boos, "The calculus of M-estimation," *The American Statistician*, vol. 56, no. 1, pp. 29–38, 2002.

[43] W. F. Trench, *Introduction to Real Analysis*, 1st ed. Hoboken, NJ: Prentice Hall/Pearson Education, 2003.

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

[45] V. Vapnik, "Principles of risk minimization for learning theory," in *Advances in neural information processing systems*, vol. 4, 1992, pp. 831–838.

[46] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[47] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, 1st ed. New York, NY, USA: Cambridge university press, 2018.

[48] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, 1st ed. New York, NY, USA: Cambridge University Press, 2019.

[49] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 102–113, May 2020.

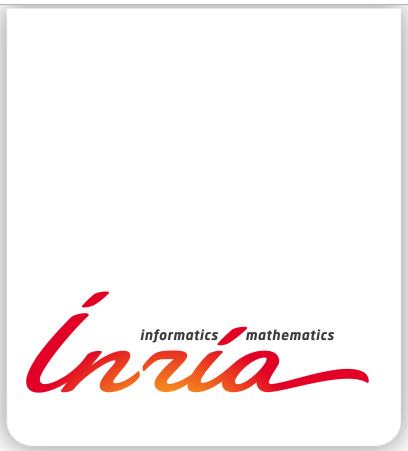
[50] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. of the Thirty-first Conference on Neural Information Processing Systems (NeurIPS)*, Dec. 2017.

[51] L. Zdeborová and F. Krzakala, "Statistical physics of inference: Thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, Aug. 2016.

[52] T. Zhang, "From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, vol. 34, no. 5, pp. 2180–2210, 2006.

[53] ——, "Information-theoretic upper and lower bounds for statistical estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, Apr. 2006.

[54] J. Zhu and E. P. Xing, "Maximum entropy discrimination Markov networks," *Journal of Machine Learning Research*, vol. 10, no. 11, 2009.



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau -
Rocquencourt
BP 105 - 78153 Le Chesnay
Cedex
inria.fr

ISSN 0249-6399