



**HAL**  
open science

# An Exact Characterization of the Generalization Error of Machine Learning Algorithms

Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, H. Vincent Poor

► **To cite this version:**

Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, H. Vincent Poor. An Exact Characterization of the Generalization Error of Machine Learning Algorithms. RR-9539, INRIA. 2024, pp.65. hal-04442566v1

**HAL Id: hal-04442566**

**<https://inria.hal.science/hal-04442566v1>**

Submitted on 7 Feb 2024 (v1), last revised 16 May 2024 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



# An Exact Characterization of the Generalization Error of Machine Learning Algorithms

Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, and  
H. Vincent Poor

**RESEARCH  
REPORT**

**N° 9539**

February 2024

Project-Team NEO

ISRN INRIA/RR--9539--FR+ENG

ISSN 0249-6399





# An Exact Characterization of the Generalization Error of Machine Learning Algorithms

Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman,  
and H. Vincent Poor

Project-Team NEO

Research Report n° 9539 — February 2024 — 66 pages

**Abstract:** The worst-case data-generating (WCDG) probability measure is introduced as a tool for characterizing the generalization capabilities of machine learning algorithms. Such a WCDG probability measure is shown to be the unique solution to two different optimization problems: (a) The maximization of the expected loss over the set of probability measures on the datasets whose relative entropy with respect to a *reference measure* is not larger than a given threshold; and (b) The maximization of the expected loss with regularization by relative entropy with respect to the reference measure. Such a reference measure can be interpreted as a prior on the datasets. Closed-form expressions are presented for the sensitivity of the expected loss for a fixed model. These tools result in the characterization of a novel expression for the generalization error of arbitrary machine learning algorithms. This exact expression is provided in terms of the WCDG probability measure and leads to an upper bound that is equal to the sum of the mutual information and the lautum information between the models and the datasets, up to a constant factor. This upper bound is shown to be achieved by a Gibbs algorithm.

**Key-words:** Statistical Machine Learning, Generalization Error, and Gibbs Algorithm

Xinying Zou, Samir M. Perlaza, and Eitan Altman are with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France. Iñaki Esnaola is with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, UK. H. Vincent Poor, Samir M. Perlaza, and Iñaki Esnaola are with the Department of Electrical and Computer Engineering, Princeton University, Princeton NJ 08544, USA. Samir M. Perlaza is also with the GAATI Laboratory at the Université de la Polynésie Française, Faaa 98702, French Polynesia. Eitan Altman is also with the Laboratoire d'Informatique d'Avignon (LIA), Université d'Avignon, France.

**RESEARCH CENTRE  
SOPHIA ANTIPOLIS – MÉDITERRANÉE**

2004 route des Lucioles - BP 93  
06902 Sophia Antipolis Cedex

# Une caractérisation exacte de l'erreur de généralisation pour des algorithmes d'apprentissage automatique

**Résumé :** La mesure de probabilité de génération de données dans le pire des cas (GDPC) est présentée comme un outil permettant de caractériser les capacités de généralisation des algorithmes d'apprentissage automatique. Une telle mesure de probabilité de GDPC s'avère être la solution unique à deux problèmes d'optimisation différents : (a) La maximisation de la perte (ou risque) espérée sur l'ensemble des mesures de probabilité des données dont l'entropie relative par rapport à une *mesure de référence* n'est pas supérieure à un seuil donné ; et (b) La maximisation de la perte espérée avec une régularisation par entropie relative par rapport à la mesure de référence. Une telle mesure de référence peut être interprétée comme un a priori sur les données. Les cumulants de la perte espérée lorsque les données sont obtenues en échantillonnant la mesure de GDPC s'avèrent majorés et minorés en termes des cumulants de la mesure de référence. Des expressions sous forme fermée sont présentées pour la sensibilité de la perte espérée pour un modèle fixe. Ces outils aboutissent à la caractérisation d'une nouvelle expression de l'erreur de généralisation pour les algorithmes d'apprentissage automatique. Cette expression exacte est en fonction de la mesure de probabilité de GDPC et conduit à une borne supérieure égale à la somme de l'information mutuelle et l'information lautum entre les modèles et les données, à un facteur près. Cette limite supérieure est obtenue par un algorithme de Gibbs. Cette découverte révèle qu'une exploration de l'erreur de généralisation de l'algorithme de Gibbs facilite la dérivation de conclusions globales applicables à tout algorithme d'apprentissage automatique.

**Mots-clés :** Apprentissage Statistique, Error de généralisation, et algorithme de Gibbs.

## Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Existing Results . . . . .	6
1.2	Contributions . . . . .	7
1.3	Notation . . . . .	8
<b>2</b>	<b>Problem Formulation</b>	<b>8</b>
<b>3</b>	<b>The Worst-Case Data-Generating Probability Measure</b>	<b>9</b>
3.1	General Discussion . . . . .	9
3.2	Relative Entropy Case . . . . .	11
3.3	Choice of the Reference Measure . . . . .	14
<b>4</b>	<b>Cumulants of the Expected Loss</b>	<b>15</b>
<b>5</b>	<b><math>(\epsilon, \delta)</math>-Robustness and Algorithm Design</b>	<b>19</b>
<b>6</b>	<b>Sensitivity of the Expected Loss</b>	<b>21</b>
6.1	Sensitivity of the Empirical-Risk . . . . .	23
6.2	Analysis of the Generalization Gap . . . . .	24
<b>7</b>	<b>The Worst-Case Data-Generating Probability Measure and Generalization</b>	<b>26</b>
7.1	An Exact Expression for the Doubly-Expected Generalization Gap	27
7.2	The Gibbs Algorithm . . . . .	29
<b>8</b>	<b>Conclusions and Final Remarks</b>	<b>32</b>
	<b>Appendices</b>	<b>33</b>
<b>A</b>	<b>Proof of Lemma 1</b>	<b>33</b>
<b>B</b>	<b>Proof of Lemma 2</b>	<b>35</b>
<b>C</b>	<b>Proof of Theorem 1</b>	<b>35</b>
<b>D</b>	<b>Proof of Lemma 3</b>	<b>39</b>
<b>E</b>	<b>Proof of Lemma 4</b>	<b>40</b>
<b>F</b>	<b>Proof of Lemma 5</b>	<b>41</b>
<b>G</b>	<b>Proof of Theorem 2</b>	<b>45</b>
<b>H</b>	<b>Proof of Theorem 3</b>	<b>46</b>
<b>I</b>	<b>Proof of Lemma 6</b>	<b>48</b>

<b>J Proof of Theorem 4</b>	<b>48</b>
<b>K Proof of Theorem 5</b>	<b>52</b>
<b>L Proof of Theorem 6</b>	<b>53</b>
<b>M Proof of Theorem 7</b>	<b>54</b>
<b>N Proof of Theorem 8</b>	<b>58</b>
<b>O Proof for Lemma 7</b>	<b>58</b>
<b>P Proof of Theorem 10</b>	<b>59</b>
<b>Q Proof of Theorem 11</b>	<b>60</b>

## 1 Introduction

The problem of supervised machine learning is often formulated as an empirical risk minimization (ERM) problem in which the optimization domain is a set of models [1]. This formulation is based on the observation that the empirical risk is the expectation of the loss function with respect to the empirical probability measure induced by the training dataset. See for instance, Lemma 7. This empirical measure is known in the realm of information theory as a type [2] (see Definition 3). If the ground-truth data-generating (GTDG) probability measure were available, optimal models would be the minimizers of the expectation of the loss with respect to the GTDG probability measure, also known as the *true risk* or *population risk* [1]. From this perspective, the type induced by the training dataset is a replacement for the GTDG probability measure. That is, models are chosen as the minimizers of the empirical risk instead of minimizers of the true risk. Interestingly, large datasets might lead to types that are sufficiently *close* to the GTDG probability measure with high probability [3]. Typically, such a closeness is often measured in terms of a *statistical distance*, e.g., the relative entropy of the type with respect to the GTDG probability measure.

The driving idea in this work is that the GTDG probability measure is likely to be within a *neighbourhood* of the type induced by the training dataset. Side information might also lead to a *prior* on the data, and thus, the GTDG probability measure is also likely to be within a *neighbourhood* of such a prior. From this perspective, one can adopt a *reference measure*, which can be the type, a prior, or a mixture of both, and form a unique *neighbourhood* around such a *reference measure*. This *neighbourhood* includes all probability measures on the datasets that are at a *statistical distance* smaller than or equal to a given threshold. A robust choice of models is choosing them as the minimizers of the expectation of the loss with respect to the worst-case data-generating (WCDG) probability measure within such a neighbourhood. In this case, the WCDG probability measure is assumed to be the measure that maximizes the expectation of the loss for a fixed model. This problem formulation is a distributionally robust optimization (DRO) problem [4, 5] in which the *statistical distance* might be the relative entropy as in [6], a Wasserstein distance as in [7, 8], an  $f$ -divergence as in [9, 10], among others. This problem can also be formulated via the maximum entropy principle [11, 12] as in [13, 14]. Every choice of *statistical distance* leads to a WCDG probability measure with particular properties.

When the *statistical distance* is the relative entropy, the resulting WCDG probability measure exhibits numerous properties that are central in the analysis of key generalization metrics, namely, the generalization gap (GG); the expected generalization gap (EGG); and the doubly-expected generalization gap (D-EGG). In the literature, the D-EGG is also referred to as the *generalization error* (GE). The GG is calculated for a specific model and training dataset as the difference between the true risk and the empirical risk induced by the model on such a training dataset. While the GG does not provide a generalization guarantee for a given model, it allows studying the impact of the training



dataset on the generalization capabilities. This observation is central to the results presented in this work. When models are chosen by sampling a probability measure conditioned on the training dataset, the EGG is the expectation of the GG with respect to such a probability measure. This probability measure is often referred to as the statistical learning algorithm. In this setting, the dependence of the EGG on the training dataset is twofold. Firstly, via the algorithm, which depends on the training dataset; and secondly, via the GG, which involves the calculation of the empirical risk on the training dataset. The D-EGG is the expectation of the EGG with respect to the GTDG probability measure.

## 1.1 Existing Results

The GG, EGG, and D-EGG (or GE) are central performance metrics for the analysis of the generalization capabilities of machine learning algorithms, see for instance [15–19] and [20]. In particular, the D-EGG characterizes the ability of the learning algorithm to correctly find patterns in datasets that are not available during the training stage. Closed-form expressions for the EGG are only known for the Gibbs algorithm in the case in which the reference measure is a probability measure [15]; and for the case in which the reference measure is a  $\sigma$ -finite measure [21]. In the case of other algorithms, the D-EGG is characterized by various upper-bounds leveraging different techniques. The metric of mutual information is first proposed in [22], further developed in [17] and combined with chaining methods in [23, 24]. Similar bounds are obtained in [18, 25–28] and references therein. Other information measures such as the Wasserstein distance [16, 29, 30], maximal leakage [31, 32], mutual  $f$ -information [33], and Jensen-Shannon divergence [34] are also explored in the literature. Unfortunately, none of the existing bounds has been shown to be uniformly tight in relevant practical cases [35, 36].

To circumvent the dependence of the D-EGG on the statistical description of the training dataset and the specific learning algorithm in generalization analyses, tools from combinatorics [37]; probability theory [38–40]; and information theory [15, 17, 41] have been used with partial success. The main drawback of these analytical approaches is that they provide guarantees that entail worst-case dataset generation analysis but do not identify the data-generating measures that curtail the generalization capability of the algorithm [36]. This, in turn, results in descriptions of the D-EGG for which the dependence on the training dataset and the selected algorithm is not made evident. Recent efforts for highlighting the dependence of generalization capabilities on the training dataset have led to explicit expressions for the GG and EGG when the models are sampled using the Gibbs algorithm in [20, 42]. This line of work paved the way to the study in this paper of the WCDG probability measures and their effect on the D-EGG.

## 1.2 Contributions

The first contribution consists of the derivation of a probability measure over the datasets coined WCDG probability measure. The WCDG probability measure can be defined as the measure that maximizes the expectation of the loss over a set of measures that are at a “statistical distance” with respect to a *reference measure* that is not larger than a given threshold. Alternatively, the WCDG probability measure can be defined as the measure that maximizes the expectation of the loss subject to a regularization by such a “statistical distance” with respect to the *reference measure*. When the “statistical distance” is the relative entropy, both definitions are shown to be identical, under specific conditions. Despite the limitations of relative entropy concerning its asymmetry [10, 43] and the need of absolute continuity with respect to the reference measure [9], the resulting WCDG probability measure is a Gibbs probability measure (Theorem 1) parametrized by the reference measure; the regularization parameter; and the loss function. This Gibbs probability measure is shown to exhibit relevant properties in statistical machine learning. In particular, for a fixed model, the loss resulting from the assumption that datapoints are sampled from the WCDG probability measure is a sub-Gaussian random variable. Moreover, the variation of the expectation of the loss when the probability measure changes from the WCDG probability measure to an alternative measure has an explicit expression involving relative entropy terms. Using this result, the variation of the expectation of the loss when the measure changes from an arbitrary measure to any alternative measure is presented (Theorem 8). This is a significant result as the reference measure and the regularization parameter can be arbitrarily chosen, which leads to numerous closed-form expressions for such a variation.

The second contribution leverages the observation that under the assumption that datasets are tuples of independent and identically distributed datapoints, datasets can be represented by their corresponding types, which are empirical probability measures [2]. Interestingly, the empirical risk induced by a model with respect to a given dataset is proved to be equal to the expectation of the loss with respect to the corresponding type (Lemma 7). This observation allows using Theorem 8 to provide an explicit expression to the difference between two empirical risks induced by the same model on two different datasets. This difference is referred to as the *sensitivity* of the empirical risk. Using the same arguments, closed-form expressions involving relative entropy terms are provided for the GG induced by a fixed model.

The final contribution consists in using the WCDG probability measure to obtain an exact expression of the D-EGG for any machine learning algorithm. This exact expression is provided in terms of information measures involving the WCDG probability measure and leads to an upper bound on the D-EGG that is equal to the sum of the mutual information and the lautum information between the models and the datasets, up to a constant factor. This upper bound is shown to be achieved by a Gibbs algorithm whose parameters satisfy

particular conditions. This reveals the central place of the Gibbs algorithm in statistical machine learning. Indeed, the Gibbs algorithm facilitates the derivation of overarching insights into the D-EGG applicable to any machine learning algorithm.

### 1.3 Notation

Given a measurable space  $(\Omega, \mathcal{F})$ , the notation  $\Delta(\Omega)$  is used to represent the set of  $\sigma$ -finite measures that can be defined over  $(\Omega, \mathcal{F})$ . Often, when the sigma-algebra  $\mathcal{F}$  is fixed, it is hidden to ease notation. Given a measure  $Q \in \Delta(\Omega)$ , the subset  $\Delta_Q(\Omega)$  of  $\Delta(\Omega)$  contains all  $\sigma$ -finite measures that are absolutely continuous with respect to the measure  $Q$ . Given a second measurable space  $(\mathcal{X}, \mathcal{G})$ , the notation  $\Delta(\Omega|\mathcal{X})$  is used to represent the set of  $\sigma$ -finite measures defined over  $(\Omega, \mathcal{F})$  conditioned on an element of  $\mathcal{X}$ .

## 2 Problem Formulation

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

$$\mathbf{z} = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n. \quad (1)$$

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

$$y = f(\boldsymbol{\theta}, x). \quad (2)$$

Let also the function

$$\hat{\ell} : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, +\infty] \quad (3)$$

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

For the ease of notation, let the function  $\ell : \mathcal{M} \times \mathcal{X} \times \mathcal{Y} \rightarrow [0, +\infty]$  be such that for all  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ ,

$$\ell(\boldsymbol{\theta}, x, y) = \hat{\ell}(f(\boldsymbol{\theta}, x), y). \quad (4)$$

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

$$\mathsf{L}(\mathbf{z}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i, y_i), \quad (5)$$

where the function  $\ell$  is defined in (4).

Using this notation, the model selection problem is formulated as an empirical risk minimization ERM problem, which consists of the following optimization problem:

$$\min_{\theta \in \mathcal{M}} L(\mathbf{z}, \theta). \quad (6)$$

The ERM problem is prone to overfitting since the set of solutions to (6) are models selected specifically for the given data set  $\mathbf{z}$  in (1), which limits the generalization capability of the resulting optimal model. One way to compensate for overfitting and adding more stability to the learning algorithm is by adding a regularization term to the optimization problem in (6). Such a regularization term can be represented by a function  $R : \mathcal{M} \rightarrow \mathbb{R}$ , which yields the regularized ERM problem

$$\min_{\theta \in \mathcal{M}} L(\mathbf{z}, \theta) + \lambda R(\theta), \quad (7)$$

where  $\lambda$  is a nonnegative real that acts as a regularization parameter. The regularization function constraints the choice of the model, which can be interpreted as requiring a finite space for the models or limiting the “complexity” of the model [1]. One common choice for  $R$  is  $R(\theta) = \|\theta\|_p$ , with  $p \geq 1$ . The norm is often used to account for the model complexity. Alternatively, the regularization parameter  $\lambda$  determines the weight that regularization carries in the model selection.

The main interest in this work is to study the generalization capability for a given model  $\theta \in \mathcal{M}$  independently from how such a model is chosen.

### 3 The Worst-Case Data-Generating Probability Measure

#### 3.1 General Discussion

Given a probability measure  $P_S \in \Delta(\mathcal{X} \times \mathcal{Y})$ , which can be interpreted as a *prior* on the set of data points, and a model  $\theta \in \mathcal{M}$ , looking for a WCDG probability measure might lead to two different optimization problems. Both can be defined in terms of the expected loss induced by a measure  $P \in \Delta(\mathcal{X} \times \mathcal{Y})$ , for a given a model  $\theta \in \mathcal{M}$ .

**Definition 1 (Expected Loss)** *Let  $P$  be a probability measure in  $\Delta(\mathcal{X} \times \mathcal{Y})$ . The expected loss with respect to a fixed model  $\theta \in \mathcal{M}$  induced by the measure  $P$  is*

$$R_\theta(P) = \int \ell(\theta, x, y) dP(x, y), \quad (8)$$

where the function  $\ell$  is defined in (4).

Firstly, a *neighborhood* around  $P_S$  is established through a *statistical distance*  $\mathbf{d} : \Delta(\mathcal{X} \times \mathcal{Y}) \times \Delta(\mathcal{X} \times \mathcal{Y}) \rightarrow [0, +\infty)$ . In view of this, the WCDG measure is a probability measure that maximizes the expectation of the loss within such a neighborhood. This view leads to the following problem:

$$\max_{P \in \Delta(\mathcal{X} \times \mathcal{Y})} \mathbf{R}_\theta(P) \quad (9a)$$

$$\text{s.t.} \quad \mathbf{d}(P, P_S) \leq \gamma, \quad (9b)$$

where  $\gamma > 0$  determines the neighborhood around  $P_S$  as the set  $\{P \in \Delta(\mathcal{X} \times \mathcal{Y}) : \mathbf{d}(P, P_S) \leq \gamma\}$  and the functional  $\mathbf{R}_\theta$  is defined in (8). Secondly, a WCDG measure is also interpreted as a probability measure that trades off the maximization of the expectation of the loss and the minimization of the statistical distance with respect to  $P_S$ . This point of view leads to an optimization problem of the form:

$$\max_{P \in \Delta(\mathcal{X} \times \mathcal{Y})} \mathbf{R}_\theta(P) - \beta \mathbf{d}(P, P_S), \quad (10)$$

where  $\beta > 0$  determines the trade-off between maximization of the expectation of the loss and the statistical distance to  $P_S$ ; and the functional  $\mathbf{R}_\theta$  is defined in (8). In this case, the convexity of  $\mathbf{d}(P, P_S)$  with respect to  $P$  is a valuable property for solving the optimization problem.

Depending on the choice of the statistical distance  $\mathbf{d}$  in (9) and (10), numerous WCDG probability measures might be obtained. In either case, such WCDG probability measures are subject to the limitations concerning the asymmetry of  $\mathbf{d}$ ; limitations concerning the absolute continuity requirements with respect to  $P_S$ ; or limitations derived from the need for  $\mathcal{X} \times \mathcal{Y}$  to be equipped with a metric so as to form a Polish metric space. These limitations can lead the problems in (9) and (10) to be ill posed, difficult to solve, to exhibit no solution, or to exhibit solutions with poor properties for the analysis of generalization in statistical machine learning. In particular, the statistical distance  $\mathbf{d}$  can be chosen as an  $f$ -divergence [44]. See for instance the examples in [9]. In this case,

$$\mathbf{d}(P, P_S) = \int f\left(\frac{dP}{dP_S}(x, y)\right) dP(x, y), \quad (11)$$

where the function  $f : (0, +\infty) \rightarrow \mathbb{R}$  is assumed to be concave and  $f(1) = 0$ ; and the function  $\frac{dP}{dP_S} : \mathcal{X} \times \mathcal{Y} \rightarrow [0, +\infty)$  is the Radon-Nikodym derivative of  $P$  with respect to  $P_S$ . For some choices of the function  $f$ , the corresponding  $f$ -divergence can be symmetric, e.g., Jeffrey's divergence [45]; or be asymmetric, e.g., Kullback-Leibler divergence [46]. An interesting discussion on the impact of asymmetry in optimization problems similar to the one in (10) is presented in [10] and [9]. In any case, if  $\mathbf{d}$  is chosen as an  $f$ -divergence, the optimization domain shall be restricted to the set of measures that are absolutely continuous with  $P_S$ . This is essentially because the existence of a Radon-Nikodym derivative with respect to  $P_S$  is required.

An alternative choice for the statistical distance  $d$  in (9) and (10) is

$$d(P, P_S) = \sup \left\{ \int f(x, y) dP(x, y) - \int f(x, y) dP_S(x, y) : f \in \mathcal{D} \right\}, \quad (12)$$

where  $\mathcal{D}$  is a particular set of functions  $\mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ . In such a formulation for  $d$  in (12), while  $P$  is not required to be absolutely continuous with respect to  $P_S$ , other requirements must be ensured. For instance, in the case in which the set  $\mathcal{D}$  is such that if  $f \in \mathcal{D}$ , then for all  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  and for all  $(u, v) \in \mathcal{X} \times \mathcal{Y}$ ,

$$|f(x, y) - f(u, v)| \leq d((x, y), (u, v)), \quad (13)$$

with  $d$  being a metric on the set  $\mathcal{X} \times \mathcal{Y}$ , the resulting statistical distance is the Kantorovich-Rubinstein distance, also known as the Wasserstein distance of order one [47, Remark 6.5]. Note that this choice of  $\mathcal{D}$  imposes the need of the set  $\mathcal{X} \times \mathcal{Y}$  to be equipped with the metric  $d$  in (13). Moreover,  $(\mathcal{X} \times \mathcal{Y}, d)$  must be a Polish metric space.

From the above, it becomes clear that every choice for the statistical distance  $d$  comes with particular requirements. Thus, the search for an optimal choice for  $d$  reduces to evaluating the properties of the resulting WCDG probability measures and their impact in the analysis of generalization of machine learning algorithms. In the following, it is argued that the choice of  $d$  as the relative entropy, despite the limitations concerning its asymmetry and the restriction of the optimization domains to measures that are absolutely continuous with  $P_S$ , leads to a variety of properties that are central in the analysis of generalization.

### 3.2 Relative Entropy Case

This section focuses on the special cases in which the statistical distance  $d$  in (9) and (10) is chosen as the relative entropy. The relative entropy is the  $f$ -divergence resulting from the choice of  $f(x) = x \log(x)$  in (11). See for instance [48]. Given two probability measures  $P$  and  $Q$  on the same measurable space such that  $P$  is absolutely continuous with respect to  $Q$ , the relative entropy of  $P$  with respect to  $Q$  is

$$D(P||Q) = \int \frac{dP}{dQ}(x) \log \left( \frac{dP}{dQ}(x) \right) dQ(x), \quad (14)$$

where the function  $\frac{dP}{dQ}$  is the Radon-Nikodym derivative of  $P$  with respect to  $Q$ . Using this notation, the optimization problems of interest are

$$\max_{P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})} \mathbf{R}_\theta(P) \quad (15a)$$

$$\text{s.t.} \quad D(P||P_S) \leq \gamma, \quad (15b)$$

and

$$\max_{P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})} \mathbf{R}_\theta(P) - \beta D(P||P_S). \quad (16)$$

The optimization problems in (15) and (16) exhibit a major difference in the following case. If for all  $P \in \Delta_{P_S}$ , the functional  $R_\theta$  in (8) is such that  $R_\theta(P) = c$ , for some fixed  $c \geq 0$ , then all the measures in  $\{P \in \Delta(\mathcal{X} \times \mathcal{Y}) : D(P||P_S) \leq \gamma\}$  are solutions to the optimization problem in (15). Alternatively, the problem in (16) exhibits a unique solution, which is  $P_S$ . Hence, while the former exhibits infinitely many solutions, the latter exhibits only one. Although this difference holds substantial mathematical importance, its practical relevance is limited, as it only emerges in settings for which the expectation of the loss  $R_\theta(P)$  is the same independently of the measure  $P$ . In order to avoid the above case, the notion of separable loss functions, which is analogous to [21, Definition 4.1], is introduced as follows.

**Definition 2** *Given a model  $\theta \in \mathcal{M}$ , the function  $\ell$  in (4), is said to be separable with respect to the probability measure  $P \in \Delta(\mathcal{X} \times \mathcal{Y})$ , if there exist a positive real  $c > 0$  and two subsets  $\mathcal{A}$  and  $\mathcal{B}$  of  $\mathcal{M}$  that are nonnegligible with respect to  $P$ , and for all  $((x_1, y_1), (x_2, y_2)) \in \mathcal{A} \times \mathcal{B}$ ,*

$$\ell(\theta, x_1, y_1) < c < \ell(\theta, x_2, y_2) < +\infty. \quad (17)$$

When the function  $\ell$  in (4) is nonseparable with respect to the reference measure  $P_S$ , it is a constant almost surely with respect to such a measure. More specifically, there exists a real  $a \geq 0$ , such that

$$P_S(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \ell(\theta, x, y) = a\}) = 1, \quad (18)$$

and as a consequence, for all probability measures  $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ , it holds that  $P(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \ell(\theta, x, y) = a\}) = 1$ . With this pathological case of nonseparable loss functions out of the way, the optimization problems in (15) and (16) exhibit considerable similarity. This similarity is formalized in Theorem 1 below. Given a model  $\theta \in \mathcal{M}$ , let  $J_{P_S, \theta} : \mathbb{R} \rightarrow \mathbb{R}$  be the function

$$J_{P_S, \theta}(t) = \log \left( \int \exp(t\ell(\theta, x, y)) dP_S(x, y) \right), \quad (19)$$

where the function  $\ell$  is defined in (4). The following lemma describes some properties of the function  $J_{P_S, \theta}$  in (19).

**Lemma 1** *The function  $J_{P_S, \theta}$  in (19) is convex, nondecreasing, and differentiable infinitely many times in the interior of  $\{t \in \mathbb{R} : J_{P_S, \theta}(t) < +\infty\}$ . If the function  $\ell$  in (4) is separable with respect to  $P_S$ , then the function  $J_{P_S, \theta}$  is strictly convex.*

*Proof:* The proof is presented in Appendix A. ■

Let also the set  $\mathcal{J}_{P_S, \theta}$  be the set

$$\mathcal{J}_{P_S, \theta} \triangleq \left\{ t \in (0, +\infty) : 0 \leq J_{P_S, \theta} \left( \frac{1}{t} \right) < +\infty \right\}, \quad (20)$$

which exhibits the following property.

**Lemma 2** *The set  $\mathcal{J}_{P_S, \theta}$  in (20) is either the empty set or an interval satisfying  $(b, +\infty) \subseteq \mathcal{J}_{P_S, \theta}$ , for some  $b \in [0, +\infty)$ .*

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

Using this notation, the following theorem formalizes the fact that the optimization problems in (15) and (16) exhibit the same unique solution.

**Theorem 1** *If the function  $\ell$  in (4) is separable with respect to the measure  $P_S$  and  $\beta \in \mathcal{J}_{P_S, \theta}$ , with  $\mathcal{J}_{P_S, \theta}$  in (20), then the probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$  that satisfies for all  $(x, y) \in \text{supp } P_S$ ,*

$$\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) = \exp\left(\frac{1}{\beta}\ell(\theta, x, y) - \mathcal{J}_{P_S, \theta}\left(\frac{1}{\beta}\right)\right), \quad (21)$$

where the function  $\mathcal{J}_{P_S, \theta}$  is defined in (19) and

$$D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right) = \gamma, \quad (22)$$

is the unique solution to the optimization problems in (15) and (16).

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

The probability measure  $P_S$  in Theorem 1 can be arbitrarily chosen, that is, independent of the model  $\theta$ . From this perspective, when the measure  $P_S$  is interpreted as a prior on the datasets, the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  can be interpreted as the worst probability measure for model  $\theta$  in the *neighborhood* of the prior  $P_S$ . The reference measure  $P_S$  can also be chosen to be dependent on the models. This case is studied in Section 7.2.

The parameter  $\gamma$  of the optimization problem in (15) does not explicitly appear in the expression of the solution  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21), the probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  does depend on  $\gamma$ . This dependence is shown via the equality in (22). Later, in Lemma 6, it is shown that  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right)$  in (22) is monotone with  $\beta$ . More importantly, if the function  $\ell$  in (4) is separable, the relative entropy  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right)$  is strictly decreasing with  $\beta$ . This implies that there exists a one-to-one mapping between  $\beta$  and  $\gamma$ , and thus, the analysis can be indistinctly carried out either for  $\beta$  or  $\gamma$ .

The probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) is a Gibbs probability measure [49].

In the remainder of this work, the probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  is referred to as the *WCDG probability measure* and the function  $\mathcal{J}_{P_S, \theta}$  in (19) is referred to as the *log-partition function* [50]. The WCDG probability measure exhibits a number of interesting properties. The following lemma introduces two of these properties, which are shown to be central in the remainder of this work.



**Lemma 3** *The probability measures  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  and  $P_S$  in (21) are mutually absolutely continuous. Moreover, they satisfy:*

$$\beta \mathbb{J}_{P_S, \theta} \left( \frac{1}{\beta} \right) = \mathbf{R}_\theta \left( P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right) - \beta D \left( P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S \right) \quad (23)$$

$$= \mathbf{R}_\theta(P_S) + \beta D \left( P_S \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right), \quad (24)$$

where the functional  $\mathbf{R}_\theta$  and the function  $\mathbb{J}_{P_S, \theta}$  are in (8) and (19), respectively.

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

Lemma 3 implicitly characterizes the difference of the expected losses induced by the WCDG probability measure and the reference measure. That is,

$$\mathbf{R}_\theta \left( P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right) - \mathbf{R}_\theta(P_S) = \beta \left( D \left( P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S \right) + D \left( P_S \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right) \right), \quad (25)$$

where  $D \left( P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S \right) + D \left( P_S \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right)$  is Jeffrey's divergence between  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  and  $P_S$ .

In the remainder of this work, the GTDG probability measure of the datapoints is denoted by  $P_Z \in \Delta(\mathcal{X} \times \mathcal{Y})$ . The probability measure  $P_Z$  is the measure induced by a random variable  $Z$  on the measurable space of the datapoints. Under the assumption that datasets are formed by  $n$  independent and identically distributed datapoints, the GTDG probability measure of such datasets is denoted by  $P_{\mathbf{Z}} \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$ , which is a product measure formed by  $P_Z$ . Alternatively, the probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) is induced by a random variable  $\hat{Z}$  on the measurable space of the datapoints. Under the assumption that datasets are formed by  $n$  independent datapoints sampled from the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$ , the probability measure of such datasets is denoted by  $P_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}$ , which is the product measure formed by  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$ . In a nutshell, datapoints and datasets sampled from the GTDG probability measure are represented by the random variables  $Z$  and  $\mathbf{Z}$ , respectively. Similarly, datapoints and datasets sampled from a WCDG probability measure are represented by the random variables  $\hat{Z}$  and  $\hat{\mathbf{Z}}$ , respectively.

### 3.3 Choice of the Reference Measure

Probably, the most intuitive choice of  $P_S$  is to be based on the training datasets. In order to introduce this intuition, the notion of a type induced by a dataset [2] is presented in the following.

**Definition 3 (Type of a dataset)** *The type induced by the dataset  $\mathbf{z}$  in (1) is a probability measure in  $\Delta(\mathcal{X} \times \mathcal{Y})$ , denoted by  $P_{\mathbf{z}}$ , such that for all measurable subsets  $\mathcal{A}$  of  $\mathcal{X} \times \mathcal{Y}$ ,*

$$P_{\mathbf{z}}(\mathcal{A}) = \frac{1}{n} \sum_{t=1}^n \mathbb{1}_{\{(x_t, y_t) \in \mathcal{A}\}}. \quad (26)$$

For large values of  $n$ , it might be expected that the type  $P_{\mathbf{z}}$  induced by the training dataset  $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$  be within a neighborhood of the GTDG probability distribution  $P_Z$ . More specifically, under the assumption that the sets  $\mathcal{X}$  and  $\mathcal{Y}$  are discrete, such a neighborhood might be of the form

$$\mathcal{N}(P_Z, \gamma) \triangleq \{P \in \Delta(\mathcal{X} \times \mathcal{Y}) : D(P \| P_Z) \leq \gamma\}, \quad (27)$$

for some  $\gamma > 0$ . Nonetheless, the GTDG probability measure  $P_Z$  being unknown, it appears reasonable to choose the reference measure  $P_S$  identical to the type  $P_{\mathbf{z}}$ . This leads to an optimization problem of the form in (15), whose optimization domain is

$$\mathcal{N}(P_{\mathbf{z}}, \gamma) \triangleq \{P \in \Delta_{P_{\mathbf{z}}}(\mathcal{X} \times \mathcal{Y}) : D(P \| P_{\mathbf{z}}) \leq \gamma\}. \quad (28)$$

This choice presumes that the training dataset is sufficiently large to consider that the GTDG probability measure is absolutely continuous with respect to the type  $P_{\mathbf{z}}$  such that the GTDG probability measure  $P_Z$  is within  $\mathcal{N}(P_{\mathbf{z}}, \gamma)$ . Another alternative for the choice of  $P_S$  in the optimization problems in (15) and (16) is for instance a convex combination between the type  $P_{\mathbf{z}}$  and an arbitrary probability measure known to be absolutely continuous with  $P_Z$ .

In any case, the choice of  $P_S$  imposes a condition on the optimization domain of the problems in (15) and (16) that translates into a strong inductive bias that dominates the evidence provided by the training data. More specifically, the optimization domain becomes a subset of the set of measures that are absolutely continuous with respect to  $P_S$ .

## 4 Cumulants of the Expected Loss

The log-partition function  $J_{P_S, \theta}$  in (19) can be interpreted also as the cumulant generating function of the random variable

$$V_{\theta} \triangleq \ell(\theta, X, Y), \quad (29)$$

with the function  $\ell$  in (4) and  $(X, Y) \sim P_S$ , with  $P_S$  in (21), for some given model  $\theta$  that remains fixed. In particular, if the function  $J_{P_S, \theta}$  is differentiable around zero, its derivative of order  $n$  evaluated at zero reveals the  $n$ -th cumulant of the random variable  $V_{\theta}$  in (29). More interestingly, the function  $J_{P_S, \theta}$  in (19) is intimately related to the cumulant generating function of the random variable

$$W_{\theta} \triangleq \ell(\theta, X, Y), \quad (30)$$

where  $(X, Y) \sim P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$ , with  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21), for some given model  $\theta$  that remains fixed. In particular, the cumulant generating function of the random variable  $W_{\theta}$ , which is denoted by  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta} : \mathbb{R} \rightarrow \mathbb{R}$ , satisfies

$$J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(t) = \log \left( \int \exp(t\ell(\theta, x, y)) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) \right), \quad (31)$$

where the function  $\ell$  is defined in (4). The following lemma shows the exact relation between these two cumulant generating functions.

**Lemma 4** *The function  $J_{P_S, \theta}$  in (21) and the function  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}$  in (31) satisfy*

$$J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(t) = J_{P_S, \theta}\left(t + \frac{1}{\beta}\right) - J_{P_S, \theta}\left(\frac{1}{\beta}\right). \quad (32)$$

*Proof:* The proof is presented in Appendix E. ■

Let the  $m$ -th derivative of the function  $J_{P_S, \theta}$  in (19) be denoted by  $J_{P_S, \theta}^{(m)}$ , with  $m \in \mathbb{N}$ . Hence, for all  $t \in \mathbb{R}$  such that  $J_{P_S, \theta}(t) < +\infty$ ,

$$J_{P_S, \theta}^{(m)}(t) \triangleq \frac{d^m}{ds^m} J_{P_S, \theta}(s) \Big|_{s=t}. \quad (33)$$

Moreover, let the  $m$ -th derivative of the function  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}$  in (19) be denoted by  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}^{(m)}$ , with  $m \in \mathbb{N}$ . Hence, for all  $t \in \mathbb{R}$  such that  $J_{P_S, \theta}\left(t + \frac{1}{\beta}\right) < +\infty$ ,

$$J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}^{(m)}(t) \triangleq J_{P_S, \theta}^{(m)}\left(t + \frac{1}{\beta}\right). \quad (34)$$

The following lemma capitalizes on the observations above and provides explicit expressions for the first, second and third derivatives of the function  $J_{P_S, \theta}$  in (19) and its connections with the cumulants of the random variables  $V_\theta$  and  $W_\theta$  in (29) and (30), respectively.

**Lemma 5** *For all  $t \in \mathcal{J}_{P_S, \theta}$ , with  $\mathcal{J}_{P_S, \theta}$  in (20), the first, second and third derivatives of the function  $J_{P_S, \theta}$  in (19), denoted respectively by  $J_{P_S, \theta}^{(1)}$ ,  $J_{P_S, \theta}^{(2)}$  and  $J_{P_S, \theta}^{(3)}$ , satisfy that*

$$J_{P_S, \theta}^{(1)}\left(\frac{1}{t}\right) = \int \ell(\theta, x, y) dP_{\hat{Z}|\Theta=\theta}^{(P_S, t)}(x, y), \quad (35)$$

$$J_{P_S, \theta}^{(2)}\left(\frac{1}{t}\right) = \int \left(\ell(\theta, x, y) - J_{P_S, \theta}^{(1)}\left(\frac{1}{t}\right)\right)^2 dP_{\hat{Z}|\Theta=\theta}^{(P_S, t)}(x, y), \quad (36)$$

and

$$J_{P_S, \theta}^{(3)}\left(\frac{1}{t}\right) = \int \left(\ell(\theta, x, y) - J_{P_S, \theta}^{(1)}\left(\frac{1}{t}\right)\right)^3 dP_{\hat{Z}|\Theta=\theta}^{(P_S, t)}(x, y), \quad (37)$$

where the function  $\ell$  is in (4) and the measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, t)}$  is in (21). The value  $J_{P_S, \theta}^{(2)}\left(\frac{1}{t}\right)$  is strictly positive if and only if the function  $\ell$  is separable with respect to  $P_S$ . Moreover, if there exists real  $\delta > 0$  such that the function  $J_{P_S, \theta}$  is

differentiable within  $(-\delta, \delta)$ , then

$$J_{P_S, \theta}^{(1)}(0) = \int \ell(\theta, x, y) dP_S(x, y), \text{ and} \quad (38)$$

$$J_{P_S, \theta}^{(2)}(0) = \int \left( \ell(\theta, x, y) - J_{P_S, \theta}^{(1)}(0) \right)^2 dP_S(x, y), \text{ and} \quad (39)$$

$$J_{P_S, \theta}^{(3)}(0) = \int \left( \ell(\theta, x, y) - J_{P_S, \theta}^{(1)}(0) \right)^3 dP_S(x, y). \quad (40)$$

*Proof:* The proof is presented in Appendix F. ■

From Lemma 5, it follows that the random variable  $W_\theta$  in (30) possesses a mean, variance, and third cumulant that are equivalent to  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  in (35),  $J_{P_S, \theta}^{(2)}\left(\frac{1}{\beta}\right)$  in (36), and  $J_{P_S, \theta}^{(3)}\left(\frac{1}{\beta}\right)$  in (37), respectively. Alternatively, under the assumptions of the lemma, the random variable  $V_\theta$  in (29) possesses a mean, variance, and third cumulant that are equivalent to  $J_{P_S, \theta}^{(1)}(0)$  in (38),  $J_{P_S, \theta}^{(2)}(0)$  in (39), and  $J_{P_S, \theta}^{(3)}(0)$  in (40), respectively. The following corollary of Lemma 5 highlights the monotonicity of  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  with respect to  $\beta$ .

**Corollary 1** *The expected loss  $J_{P_S, \theta}^{(1)}\left(\frac{1}{t}\right)$  in (35) is nonincreasing with respect to  $t$  in the interior of  $\{s \in \mathbb{R} : J_{P_S, \theta}\left(\frac{1}{s}\right) < +\infty\}$ . Moreover, if the loss function  $\ell$  in (4) is separable with respect to  $P_S$ ,  $J_{P_S, \theta}^{(1)}\left(\frac{1}{t}\right)$  is strictly decreasing with  $t$  in  $\{s \in \mathbb{R} : J_{P_S, \theta}\left(\frac{1}{s}\right) < +\infty\}$ .*

More generally, under the assumptions of the lemma, the random variables  $V_\theta$  in (29) and  $W_\theta$  in (30) possess finite cumulants of all orders. That is, the values  $J_{P_S, \theta}^{(m)}(0)$  and  $J_{P_S, \theta}^{(m)}\left(\frac{1}{\beta}\right)$ , which are the  $m$ -th cumulants of the random variables  $V_\theta$  and  $W_\theta$ , respectively, for all integers  $m > 0$ , are both finite. This observation together with the mean value theorem [51, Theorem 5.10] lead to the following characterization of the cumulants  $J_{P_S, \theta}^{(m)}\left(\frac{1}{\beta}\right)$  and  $J_{P_S, \theta}^{(m)}(0)$ .

**Theorem 2** *Assume that there exists real  $\delta > 0$  such that the function  $J_{P_S, \theta}$  in (19) is differentiable within  $(-\delta, \delta)$ . Then, for all  $m \in \mathbb{N}$ , the function  $J_{P_S, \theta}^{(m)}$  in (33) satisfies for all  $t \in \mathcal{J}_{P_S, \theta}$ , with  $\mathcal{J}_{P_S, \theta}$  in (20),*

$$J_{P_S, \theta}^{(m)}(0) + \frac{1}{t}c_{m,2} \leq J_{P_S, \theta}^{(m)}\left(\frac{1}{t}\right) \leq J_{P_S, \theta}^{(m)}(0) + \frac{1}{t}c_{m,1}, \quad (41)$$

and

$$c_{m,1} = \max_{s \in (\beta, +\infty)} J_{P_S, \theta}^{(m+1)}\left(\frac{1}{s}\right), \text{ and} \quad (42)$$

$$c_{m,2} = \min_{s \in (\beta, +\infty)} J_{P_S, \theta}^{(m+1)}\left(\frac{1}{s}\right), \quad (43)$$

are both finite.

*Proof:* The proof is presented in Appendix G. ■

In Theorem 2, the case  $m = 1$  is of little interest as  $J_{P_S, \theta}^{(1)}(0) = R_\theta(P_S)$  and  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) = R_\theta\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right)$ , and thus, the difference  $R_\theta\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) - R_\theta(P_S)$  is well characterized by Jeffrey's divergence in (25). Nonetheless, while the inequality  $J_{P_S, \theta}^{(1)}(0) < J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  is always verified, Theorem 2 shows that such inequality is not necessarily observed on the higher order cumulants. In the case of the variance ( $m = 2$ ), it might be observed that  $J_{P_S, \theta}^{(2)}(0) > J_{P_S, \theta}^{(2)}\left(\frac{1}{\beta}\right)$  if  $c_{2,1} < 0$ . That is, the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) leads to a larger expected loss than the one obtained with the reference measure  $P_S$  and at the same time it induces a lower variance than the variance associated with the reference measure  $P_S$ . Interestingly, the case in which the WCDG probability measure induces both a larger expected loss and higher variance is also possible when  $c_{2,2} > 0$ . A similar observation holds for cumulants of order  $m > 2$ .

The following theorem shows that the cumulant generating function  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}$  of the random variable  $W_\theta$  in (30) exhibits an important upper bound.

**Theorem 3** *For all  $t \in \{\alpha \in \mathbb{R} : J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(\alpha) < +\infty\}$ , the function  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}$  in (31) satisfies the following inequality:*

$$J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(t) \leq t J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) + \frac{1}{2} t^2 \zeta_{P_S, \theta}^2, \quad (44)$$

where  $\zeta_{P_S, \theta}$  is finite, and satisfies

$$\zeta_{P_S, \theta} \triangleq \sup \left\{ \sqrt{J_{P_S, \theta}^{(2)}(\xi)} : \xi \in \left(-\infty, b - \frac{1}{\beta}\right) \right\}, \quad (45)$$

with

$$b \triangleq \sup \{\alpha \in \mathbb{R} : J_{P_S, \theta}(\alpha) < +\infty\}, \quad (46)$$

and the functions  $J_{P_S, \theta}^{(1)}$  and  $J_{P_S, \theta}^{(2)}$  are defined in (33).

*Proof:* The proof is presented in Appendix H. ■

The relevance of Theorem 3 lies on the fact that it implies that the random variable  $W_\theta$  in (30) is a sub-Gaussian random variable with sub-Gaussian parameter  $\zeta_{P_S, \theta}$  in (45). An interesting discussion on the impact of the random variable  $W_\theta$  being sub-Gaussian is presented in [17, Theorem 1].

Finally, leveraging the properties of sub-Gaussian random variables, e.g., [52, Equation 2.9], the following holds for all  $\alpha > 0$ :

$$P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \left( \left\{ (x, y) \in \mathcal{X} \times \mathcal{Y} : \left| \ell(\theta, x, y) - J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) \right| \geq \alpha \right\} \right) \leq 2 \exp \left( -\frac{\alpha^2}{2\zeta_{P_S, \theta}^2} \right), \quad (47)$$

where  $\zeta_{P_S, \theta}$  is defined in (45) and  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  is the expected loss in (35). This type of concentration inequality allows the study of the robustness of the model  $\theta$  when it faces data sampled from WCDG probability distributions, which leads to interesting guidelines for algorithm design.

## 5 $(\epsilon, \delta)$ -Robustness and Algorithm Design

An important observation on the dependence of the WCDG probability distribution  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) on the parameter  $\gamma$  in (22) is that for all  $P \in \mathcal{N}(P_S, \gamma)$ , with  $\mathcal{N}(P_S, \gamma)$  in (27),

$$R_\theta(P) \leq R_\theta\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right). \quad (48)$$

Essentially, the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  is the measure that induces the largest expected loss within the neighborhood  $\mathcal{N}(P_S, \gamma)$  of  $P_S$ . The following lemma unveils the fact that such a neighborhood expands as  $\gamma$  increases or equivalently, as  $\beta$  decreases.

**Lemma 6** *The relative entropy  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right)$  in (22) satisfies*

$$\frac{d}{d\beta} D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right) = -\frac{1}{\beta^3} J_{P_S, \theta}^{(2)}\left(\frac{1}{\beta}\right) \leq 0, \quad (49)$$

where the function  $J_{P_S, \theta}^{(2)}$  is defined in (36). Moreover, the inequality is strict if and only if the function  $\ell$  in (4) is separable with respect to  $P_S$ .

*Proof:* The proof is presented in Appendix I. ■

The key observation from Lemma 6 is that if the function  $\ell$  is separable with respect to  $P_S$ , there exists a bijection between  $\gamma$  and  $\beta$  induced by the equality in (22). Another interesting observation is that the variations of  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right)$  in (22) and the expected loss  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  in (35), with respect to  $\beta$  (or  $\gamma$ ), obey a revealing relation. Note that

$$\frac{d}{d\beta} D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right) = \frac{1}{\beta} \frac{d}{d\beta} J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right), \quad (50)$$

which implies that both  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right)$  and  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  are decreasing with  $\beta$ . Nonetheless, their variations to changes in  $\beta$  significantly differ. For values of  $\beta$  around one, given that  $1 \in \mathcal{J}_{P_S, \theta}$ , with  $\mathcal{J}_{P_S, \theta}$  in (20), the variations of  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right)$  and  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  to changes in  $\beta$  are comparable. On the contrary, for values away from one, the variations of these quantities might be radically different. For instance, for large values of  $\beta \in \mathcal{J}_{P_S, \theta}$ , the variation of  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  is larger than the variation of  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right)$ . This is consistent

with the fact that greater values of  $\beta$  imply smaller values of  $\gamma$ . Hence, slightly increasing the size of a small neighborhood leads to including more candidates to be the WCDG probability measure, which appear to be more impactful in the maximization of the expected loss. Alternatively, for smaller values of  $\beta \in \mathcal{J}_{P_S, \theta}$ , which imply larger values of  $\gamma$ , the variation of  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right)$  is larger than the variation of  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  with respect to  $\beta$ . Hence, slightly varying the size of a neighborhood that is already big does not lead to important variations in the expected loss induced by WCDG probability measure.

The following definition describes a performance metric for a given data-generating probability measure  $P \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$  and a model  $\theta \in \mathcal{M}$  that leverages the observations above and provides guidelines for the choice of the values of  $\beta$  (or  $\gamma$ ).

**Definition 4** ( $(\epsilon, \delta)$ -Robustness) *Given a pair of positive reals  $(\delta, \epsilon)$  with  $\epsilon < 1$ , a model  $\theta \in \mathcal{M}$  is said to be  $(\delta, \epsilon)$ -robust to a probability measure  $P \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$ , if*

$$P(\{z \in (\text{supp } P_S)^n : \mathsf{L}(z, \theta) \geq \delta\}) \leq \epsilon. \quad (51)$$

This notion of robustness enables the study of the performance guarantees that a model  $\theta$  yields when faced with data generated by the WCDG probability measure for specific parameters  $\beta$  and  $P_S$ . An important issue that arises from this definition is the characterization of the largest value of  $\gamma$  (or smallest value of  $\beta$ ) that achieves  $(\epsilon, \delta)$ -robustness, i.e. how much can the WCDG probability measure deviate from the reference  $P_S$  while the guarantee still holds. The following theorem establishes a link between  $\beta$ , the reference  $P_S$ , the expected loss, and the  $(\epsilon, \delta)$ -robustness that a given model  $\theta$  achieves.

**Theorem 4** *The probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) satisfies that for all  $\delta > J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$ , with  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  in (35),*

$$P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(\{z \in (\mathcal{X} \times \mathcal{Y})^n : \mathsf{L}(\theta, z) \geq \delta\}) \leq \exp\left(-nD\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right)\right), \quad (52)$$

where  $\beta^* \in (0, \beta) \cap \mathcal{J}_{P_S, \theta}$ , with  $\mathcal{J}_{P_S, \theta}$  in (20), satisfies

$$J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta^*}\right) = \delta; \quad (53)$$

the function  $\mathsf{L}$  is defined in (5); and the measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$  is a product measure formed by  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$ .

*Proof:* This proof is presented in Appendix J. ■

Theorem 4 describes the  $(\epsilon, \delta)$ -robustness of a model  $\theta$  to the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$ . Note such a probability measure describes

test datasets that are formed by  $n$  datapoints independently and identically distributed with  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21). Notably, such a description is in terms of another WCDG probability distribution  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}$ , where  $\beta^* < \beta$ . Interestingly, the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}$  induces an expected loss that is equal to  $\delta$  and is greater than the expected loss induced by the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$ . In a nutshell, for all  $t \in (0, \beta) \cap \mathcal{I}_{P_S, \theta}$ , let  $\delta_t = nJ_{P_S, \theta}^{(1)}\left(\frac{1}{t}\right)$  and  $\epsilon_t = \exp\left(-nD\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, t)}\|P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right)\right)$ . Then, the model  $\theta$  is  $(\epsilon_t, \delta_t)$ -robust to the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$ .

The following theorem provides a  $(\epsilon, \delta)$ -robust guarantee to any product probability measure formed by a measure in the neighborhood  $\mathcal{N}(P_S, \gamma)$  in (27).

**Theorem 5** *For all models  $\theta \in \mathcal{M}$  and for all  $P_Z \in \mathcal{N}(P_S, \gamma)$ , with  $\mathcal{N}(P_S, \gamma)$  in (27) and  $\beta$  in (22), it follows that for all  $\delta > J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$ , with  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  in (35),*

$$P_Z(\{z \in (\mathcal{X} \times \mathcal{Y})^n : \mathbb{L}(\theta, z) \geq \delta\}) \leq \frac{1}{\delta} J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right), \quad (54)$$

where the function  $\ell$  is defined in (4) and the product probability measure  $P_Z$  is formed by  $P_Z$ .

*Proof:* The proof is presented in Appendix K. ■

The relevance of Theorem 5 is that given a model  $\theta$ , it establishes that for all  $\delta > J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  and  $\epsilon_\delta \triangleq \frac{1}{\delta} J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$ , such a model  $\theta$  is  $(\epsilon_\delta, \delta)$ -robust to all probability measures in  $\mathcal{N}(P_S, \gamma)$ , with  $\mathcal{N}(P_S, \gamma)$  in (27). This observation raises the question of whether performing model selection based on the minimization of  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  is an alternative for classical approaches based on empirical risk minimization (ERM) as in [53]; or statistical ERM as in [21] and [9].

## 6 Sensitivity of the Expected Loss

Given model  $\theta \in \mathcal{M}$ , the variation of the expected loss when the probability measure from which data points are sampled from varies to another measure is referred to as the *sensitivity* of the expected loss. Such a sensitivity can be quantified using the functional  $G : \mathcal{M} \times \Delta(\mathcal{X} \times \mathcal{Y}) \times \Delta(\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R}$ ,

$$G(\theta, P_1, P_2) = \mathbf{R}_\theta(P_1) - \mathbf{R}_\theta(P_2), \quad (55)$$

where the functional  $\mathbf{R}_\theta$  is defined in (8). The value  $G(\theta, P_1, P_2)$  represents the sensitivity of the expected loss  $\mathbf{R}_\theta$  when the data-generating probability measure changes from  $P_2$  to  $P_1$ . This section characterizes the *sensitivity*  $G(\theta, P_1, P_2)$  for arbitrary measures  $P_1$  and  $P_2$ . To achieve this goal, the first result is the characterization of the sensitivity of the expected loss to variations from the



measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) to an alternative measure, which is presented in the following theorem.

**Theorem 6 (Sensitivity of the Expected Loss)** *The probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) satisfies, for all  $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ ,*

$$G(\theta, P, P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}) = \beta \left( D(P \| P_S) - D\left(P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) - D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right) \right), \quad (56)$$

where the functional  $G$  is defined in (55).

*Proof:* The proof is presented in Appendix L. ■

The following theorem introduces an upper bound on the absolute value of the sensitivity  $G(\theta, P, P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}) \leq 0$  in (56), which requires the calculation of only one of the relative entropies in Theorem 6.

**Theorem 7** *The probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21) satisfies for all  $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ ,*

$$\left| G(\theta, P, P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}) \right| \leq \sqrt{2\hat{\zeta}_{P_S, \theta}^2 D\left(P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right)}, \quad (57)$$

where  $\hat{\zeta}_{P_S, \theta}$  satisfies

$$\hat{\zeta}_{P_S, \theta}^2 \triangleq \sup \left\{ \sqrt{J_{P_S, \theta}^{(2)}(\xi)} : \xi \in \left(0, b - \frac{1}{\beta}\right) \right\}, \quad (58)$$

with

$$b \triangleq \sup \{ \alpha \in (0, +\infty) : J_{P_S, \theta}(\alpha) < +\infty \}, \quad (59)$$

where the function  $J_{P_S, \theta}$  is defined in (19); the function  $J_{P_S, \theta}^{(2)}$  is defined in (36); and the functional  $G$  is defined in (55).

*Proof:* The proof is presented in Appendix M. ■

Equipped with the exact characterization of the sensitivity from the measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  to any alternative measure  $P$  provided by Theorem 6, it is possible to characterize the variation from and to arbitrary measures, as shown by the following theorem.

**Theorem 8** *For all  $P_1 \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$  and  $P_2 \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ , and for all  $\theta \in \mathcal{M}$ ,*

$$G(\theta, P_1, P_2) = \beta \left( D\left(P_2 \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) - D\left(P_1 \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) - D(P_2 \| P_S) + D(P_1 \| P_S) \right), \quad (60)$$

where the functional  $G$  is defined in (55); and the parameter  $\beta$ , the model  $\theta$ , and the measures  $P_S$  and  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  satisfy (21).

*Proof:* The proof is presented in Appendix N. ■

Note that the parameters  $\beta$  and  $P_S$  in (60) can be arbitrarily chosen. This is essentially because only the right-hand side of (60) depends on  $P_S$  and  $\beta$ . Another interesting observation is that none of the terms in the right-hand side of (60) depends simultaneously on both  $P_1$  and  $P_2$ . Interestingly, these terms dependent exclusively on the pair formed by  $P_i$  and  $P_S$ , with  $i \in \{1, 2\}$ . These observations highlight the significant flexibility of the expression in (60) to construct closed-form expressions for the sensitivity  $G(\boldsymbol{\theta}, P_1, P_2)$  in (55). The only constraint on the choice of  $P_S$  is that both measures  $P_1$  and  $P_2$  must be absolutely continuous with respect to  $P_S$ . The following corollary follows by adopting particular choices for  $P_S$ . Two choices of  $P_S$  for which the expression in the right-hand side of (60) significantly simplifies are  $P_S = P_1$  and  $P_S = P_2$ , which leads to the following corollary of Theorem 8.

**Corollary 2** *If  $P_1$  is absolutely continuous with  $P_2$ , then the value  $G(\boldsymbol{\theta}, P_1, P_2)$  in (55) satisfies:*

$$G(\boldsymbol{\theta}, P_1, P_2) = \beta \left( D \left( P_2 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_2, \beta)} \right) - D \left( P_1 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_2, \beta)} \right) + D(P_1 \| P_2) \right). \quad (61)$$

*Alternatively, if  $P_2$  is absolutely continuous with  $P_1$  then,*

$$G(\boldsymbol{\theta}, P_1, P_2) = \beta \left( D \left( P_2 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_1, \beta)} \right) - D \left( P_1 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_1, \beta)} \right) - D(P_2 \| P_1) \right), \quad (62)$$

*where for all  $i \in \{1, 2\}$ , the probability measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_i, \beta)}$  satisfies (21) under the assumption that  $P_S = P_i$ .*

Interestingly, absolute continuity of  $P_1$  with respect to  $P_2$  or  $P_2$  with respect to  $P_1$  is not necessary for obtaining an expression for the value  $G(\boldsymbol{\theta}, P_1, P_2)$  in (55). Note by choosing  $P_S$  as a convex combination of  $P_1$  and  $P_2$ , always guarantees an explicit expression for  $G(\boldsymbol{\theta}, P_1, P_2)$  independently of whether these measures are absolutely continuous with respect to each other.

## 6.1 Sensitivity of the Empirical-Risk

The following lemma shows that the empirical risk  $L(\mathbf{z}, \boldsymbol{\theta})$  in (5) can be written as the expectation of the loss with respect to the type  $P_{\mathbf{z}}$  in (26). This is formalized by the following lemma.

**Lemma 7 (Empirical Risks Via Type)** *The empirical risk  $L(\mathbf{z}, \boldsymbol{\theta})$  in (5) satisfies*

$$L(\mathbf{z}, \boldsymbol{\theta}) = \int \ell(\boldsymbol{\theta}, x, y) dP_{\mathbf{z}}(x, y) = R_{\boldsymbol{\theta}}(P_{\mathbf{z}}), \quad (63)$$

*where the measure  $P_{\mathbf{z}}$  is the type induced by the dataset  $\mathbf{z}$ ; and the function  $\ell$  and the functional  $R_{\boldsymbol{\theta}}$  are defined in (4) and (8), respectively.*

*Proof:* The proof is presented in Appendix O. ■

Equipped with the result in Lemma 7, for a fixed model, the sensitivity of the empirical risk to changes on the datasets can be characterized using the function  $G$  in (55). Consider two datasets  $\mathbf{z}_1 \in (\mathcal{X} \times \mathcal{Y})^{n_1}$  and  $\mathbf{z}_2 \in (\mathcal{X} \times \mathcal{Y})^{n_2}$  that induce the types  $P_{\mathbf{z}_1}$  and  $P_{\mathbf{z}_2}$ , respectively. The sensitivity of the empirical risk when the dataset changes from  $\mathbf{z}_2$  to  $\mathbf{z}_1$ , for a fixed model  $\boldsymbol{\theta} \in \mathcal{M}$ , is

$$G(\boldsymbol{\theta}, P_{\mathbf{z}_1}, P_{\mathbf{z}_2}) = \mathsf{L}(\mathbf{z}_1, \boldsymbol{\theta}) - \mathsf{L}(\mathbf{z}_2, \boldsymbol{\theta}). \quad (64)$$

where the functional  $G$  is defined in (55). The following theorem characterizes the sensitivity of the empirical risk.

**Theorem 9** *Given two datasets  $\mathbf{z}_1 \in (\mathcal{X} \times \mathcal{Y})^{n_1}$  and  $\mathbf{z}_2 \in (\mathcal{X} \times \mathcal{Y})^{n_2}$ , whose types  $P_{\mathbf{z}_1}$  and  $P_{\mathbf{z}_2}$  are absolutely continuous with respect to the measure  $P_S$  in (21), it holds that for all  $\boldsymbol{\theta} \in \mathcal{M}$ ,*

$$G(\boldsymbol{\theta}, P_{\mathbf{z}_1}, P_{\mathbf{z}_2}) = \beta \left( D \left( P_{\mathbf{z}_2} \| P_{\hat{\mathcal{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \right) - D \left( P_{\mathbf{z}_1} \| P_{\hat{\mathcal{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \right) - D(P_{\mathbf{z}_2} \| P_S) + D(P_{\mathbf{z}_1} \| P_S) \right), \quad (65)$$

where the functional  $G$  is in (55); the model  $\boldsymbol{\theta}$ , and the measures  $P_S$  and  $P_{\hat{\mathcal{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}$  satisfy (21).

*Proof:* The proof follows from the equality in (64), which together with Theorem 8 completes the proof. ■

In Theorem 9, the reference measure  $P_S$  can be arbitrarily chosen as long as both types  $P_{\mathbf{z}_1}$  and  $P_{\mathbf{z}_2}$  are absolutely continuous with  $P_S$ . A choice that satisfies this constraint is the type induced by the aggregation of both datasets  $\mathbf{z}_1$  and  $\mathbf{z}_2$ , which is denoted by  $\mathbf{z}_0 = (\mathbf{z}_1, \mathbf{z}_2) \in (\mathcal{X} \times \mathcal{Y})^{n_0}$ , with  $n_0 = n_1 + n_2$ . The type induced by the aggregated dataset  $\mathbf{z}_0$ , denoted by  $P_{\mathbf{z}_0}$ , is a convex combination of the types  $P_{\mathbf{z}_1}$  and  $P_{\mathbf{z}_2}$ , that is,  $P_{\mathbf{z}_0} = \frac{n_1}{n_0} P_{\mathbf{z}_1} + \frac{n_2}{n_0} P_{\mathbf{z}_2}$  [20], which satisfies the absolute continuity condition in Theorem 9.

From Theorem 9, it appears that the difference between a test empirical risk  $\mathsf{L}(\mathbf{z}_1, \boldsymbol{\theta})$  and the training empirical risk  $\mathsf{L}(\mathbf{z}_2, \boldsymbol{\theta})$  of a given model  $\mathbf{z}$  is determined by two values: (a) the difference of the “statistical distance” from the types induced by the training and test datasets to the WCDG probability measure, i.e.,  $D \left( P_{\mathbf{z}_2} \| P_{\hat{\mathcal{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \right) - D \left( P_{\mathbf{z}_1} \| P_{\hat{\mathcal{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \right)$ ; and (b) the difference of the “statistical distance” from the types to the reference measure  $P_S$ , i.e.,  $D(P_{\mathbf{z}_1} \| P_S) - D(P_{\mathbf{z}_2} \| P_S)$ .

## 6.2 Analysis of the Generalization Gap

Given a model  $\boldsymbol{\theta} \in \mathcal{M}$  obtained from the training dataset  $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$ , the *generalization gap* it induces, under the assumption that training and test data-points are independent and identically distributed according to the probability

measure  $P_Z \in \Delta(\mathcal{X} \times \mathcal{Y})$ , is

$$G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}}) = \mathbf{R}_{\boldsymbol{\theta}}(P_Z) - \mathbf{R}_{\boldsymbol{\theta}}(P_{\mathbf{z}}). \quad (66)$$

where the functional  $G$  is defined in (55) and the function  $\mathbf{R}_{\boldsymbol{\theta}}$  is defined in (8). The term  $\mathbf{R}_{\boldsymbol{\theta}}(P_{\mathbf{z}}) = \mathbf{L}(\mathbf{z}, \boldsymbol{\theta})$  is the training empirical risk, see for instance [1, Equation (3.4)]. Alternatively, the term  $\mathbf{R}_{\boldsymbol{\theta}}(P_Z)$  is the expected loss under the assumption that the data-generating probability measure is the GTDG probability measure  $P_Z$ . This term is often referred to as the *true risk* or *population risk*. In view of this, the value  $G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}})$  in (66) describes the variation of the expected empirical risk when the distribution of the datasets changes from the type  $P_{\mathbf{z}}$  to the GTDG probability measure  $P_Z$ , which is characterized by the following lemma.

**Lemma 8** *The generalization gap  $G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}})$  in (66) satisfies:*

$$G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}}) = \beta \left( D \left( P_{\mathbf{z}} \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z, \beta)} \right) - D(P_{\mathbf{z}} \| P_Z) - D \left( P_Z \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z, \beta)} \right) \right), \quad (67)$$

where the measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z, \beta)}$  is defined in (21) under the assumption that  $P_S = P_Z$ .

*Proof:* The proof follows from Corollary 2 by noticing that the type  $P_{\mathbf{z}}$  is absolutely continuous with respect to  $P_Z$ . ■

Lemma 8 highlights the intuition that if the type  $P_{\mathbf{z}}$  induced by the training dataset  $\mathbf{z}$  is at arbitrary small “statistical distance” of the GTDG probability measure  $P_Z$ , the generalization gap  $G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}})$  in (66) is arbitrarily close to zero. This is revealed by the facts that  $D(P_{\mathbf{z}} \| P_Z)$  would be arbitrarily small; and so would be the difference  $D \left( P_{\mathbf{z}} \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z, \beta)} \right) - D \left( P_Z \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z, \beta)} \right)$ .

A more general expression for the generalization gap  $G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}})$  in (66) is provided by the following corollary of Theorem 8.

**Corollary 3** *The generalization gap  $G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}})$  in (66) satisfies:*

$$G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}}) = \beta \left( D \left( P_{\mathbf{z}} \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \right) - D \left( P_Z \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \right) - D(P_{\mathbf{z}} \| P_S) + D(P_Z \| P_S) \right), \quad (68)$$

where the measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}$  is in (21).

Note that several expressions for the generalization gap  $G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}})$  in (66) can be obtained from Corollary 3 by choosing the reference  $P_S$  and the parameter  $\beta$  in (68).

## 7 The Worst-Case Data-Generating Probability Measure and Generalization

Consider a conditioned probability measure  $P_{\Theta|Z}$ , such that given a training dataset  $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$ , the measure  $P_{\Theta|Z=\mathbf{z}} \in \Delta(\mathcal{M})$  is used to perform model selection, e.g., by sampling such a measure. Hence, the conditional measure  $P_{\Theta|Z}$  is referred to as a statistical learning algorithm. This section provides explicit expressions for the generalization gap induced by the algorithm  $P_{\Theta|Z=\mathbf{z}}$ , for some training dataset  $\mathbf{z}$ ; and also explicit expressions for the generalization gap induced by the algorithm  $P_{\Theta|Z}$  when training datapoints are assumed to be sampled from a particular probability measure in  $\Delta(\mathcal{X} \times \mathcal{Y})$ . These generalization metrics are referred to as *expected generalization gap* and *doubly-expected generalization gap*.

In order to formally defined the *expected generalization gap*, let  $\bar{G} : \Delta(\mathcal{M}) \times \Delta(\mathcal{X} \times \mathcal{Y}) \times \Delta(\mathcal{X} \times \mathcal{Y}) \rightarrow \mathbb{R}$  be a functional such that

$$\bar{G}(Q, P_1, P_2) = \int G(\boldsymbol{\theta}, P_1, P_2) dQ(\boldsymbol{\theta}), \quad (69)$$

where the functional  $G$  is defined in (66). Using this notation, the expected generalization gap induced by the algorithm  $P_{\Theta|Z}$ , when the training dataset is  $\mathbf{z}$ , is

$$\bar{G}(P_{\Theta|Z=\mathbf{z}}, P_Z, P_{\mathbf{z}}) = \int G(\boldsymbol{\theta}, P_Z, P_{\mathbf{z}}) dP_{\Theta|Z=\mathbf{z}}(\boldsymbol{\theta}). \quad (70)$$

From Corollary 3, by strategically choosing the reference measure  $P_S$  and the parameter  $\beta$  in (68), numerous closed-form expressions for the expected generalization gap induced by the algorithm  $P_{\Theta|Z}$ , when the training dataset is  $\mathbf{z}$ , are obtained. Interestingly, regardless of the choice of  $P_S$  and  $\beta$ , the resulting expressions highlight the impact of the training dataset  $\mathbf{z}$  on the expected generalization gap  $\bar{G}(P_{\Theta|Z=\mathbf{z}}, P_Z, P_{\mathbf{z}})$  in (70).

The expected generalization gap  $\bar{G}(P_{\Theta|Z=\mathbf{z}}, P_Z, P_{\mathbf{z}})$  depends on the training dataset  $\mathbf{z}$ . The doubly-expected generalization gap is obtained by taking the expectation of  $\bar{G}(P_{\Theta|Z=\mathbf{z}}, P_Z, P_{\mathbf{z}})$  when  $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$  is sampled from  $P_Z \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$ , which is assumed to be a product distribution formed by  $P_Z$  in (70).

In order to formally defined the *doubly-generalization gap*, let  $\bar{\bar{G}} : \Delta(\mathcal{M} | (\mathcal{X} \times \mathcal{Y})^n) \times \Delta(\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathbb{R}$  be a function such that

$$\bar{\bar{G}}(P_{\Theta|Z}, P) = \int \int G(\boldsymbol{\theta}, P, P_{\mathbf{z}}) dP_{\Theta|Z=\mathbf{z}}(\boldsymbol{\theta}) dP(\mathbf{z}), \quad (71)$$

where  $P_{\mathbf{z}}$  is the type induced by a dataset  $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n$  and the functional  $G$  is defined in (66).

Using this notation, the doubly-expected generalization gap induced by the algorithm  $P_{\Theta|Z}$ , when the training and test datasets are both formed by independent and identically distributed datapoints sampled from the measure  $P_Z$ , is

$$\overline{G}(P_{\Theta|Z}, P_Z) = \int \int G(\theta, P_Z, P_z) dP_{\Theta|Z=z}(\theta) dP_Z(z), \quad (72)$$

where the generalization gap  $G(\theta, P_Z, P_z)$  is defined in (66); and  $P_Z$  is a product measure formed by  $P_Z$ . In existing literature, the doubly-expected generalization gap is simply referred to as generalization gap or generalization error, and thus, the distinction with generalization gaps that are dependent on specific models and training datasets is often neglected. See for instance [15], [17] and [21]. In this work, such a distinction is important.

## 7.1 An Exact Expression for the Doubly-Expected Generalization Gap

As in the case of the expected generalization gap, Corollary 3 leads to numerous closed-form expressions for the doubly-expected generalization gap induced by an algorithm  $P_{\Theta|Z}$ , when the training and test datasets are both formed by independent and identically distributed datapoints sampled from the measure  $P_Z$ . One of such expressions is of particular interest and is thoroughly studied in this subsection. Nonetheless, before presenting this expression, the following notation is introduced:

$$I(P_{\Theta|Z}; P_Z) \triangleq \int D(P_{\Theta|Z=\nu} \| P_{\Theta}) dP_Z(\nu) \quad (73)$$

$$= \int D(P_{Z|\Theta=\theta} \| P_Z) dP_{\Theta}(\theta); \text{ and} \quad (74)$$

$$L(P_{\Theta|Z}; P_Z) \triangleq \int D(P_{\Theta} \| P_{\Theta|Z=\nu}) dP_Z(\nu) \quad (75)$$

$$= \int D(P_Z \| P_{Z|\Theta=\theta}) dP_{\Theta}(\theta), \quad (76)$$

where  $P_Z$  is the product measure formed by  $P_Z$ ; and  $P_{\Theta}$  is a probability measure such that for all measurable subsets  $\mathcal{A}$  of  $\mathcal{M}$ ,

$$P_{\Theta}(\mathcal{A}) = \int P_{\Theta|Z=\nu}(\mathcal{A}) dP_Z(\nu); \quad (77)$$

and the conditional probability measure  $P_{Z|\Theta}$  satisfies for all measurable subsets  $\mathcal{B}$  of  $(\mathcal{X} \times \mathcal{Y})^n$ ,

$$P_Z(\mathcal{B}) = \int P_{Z|\Theta=\theta}(\mathcal{B}) dP_{\Theta}(\theta). \quad (78)$$

Note that for all  $\theta \in \mathcal{M}$ , the measure  $P_{Z|\Theta=\theta}$  is a product measure formed by a probability measure  $P_{Z|\Theta=\theta} \in \Delta(\mathcal{X} \times \mathcal{Y})$ . More specifically, for all measurable

sets  $\mathcal{A}_i$  in  $\mathcal{X} \times \mathcal{Y}$ , with  $i \in \{1, 2, \dots, n\}$ , it follows that

$$P_{\mathbf{Z}|\Theta=\theta}(\mathcal{A}_1 \times \mathcal{A}_2 \times \dots \times \mathcal{A}_n) = \prod_{t=1}^n P_{Z|\Theta=\theta}(\mathcal{A}_t). \quad (79)$$

The quantity  $I(P_{\Theta|\mathbf{Z}}; P_{\mathbf{Z}})$  in (73) is the *mutual information* [54] between the random variables  $\Theta$  and  $\mathbf{Z}$ , which represent the models and (training and test) datasets, when they are jointly sampled from the probability measure  $P_{\Theta\mathbf{Z}} \in \Delta(\mathcal{M} \times (\mathcal{X} \times \mathcal{Y})^n)$ , whose marginals are  $P_{\Theta}$  and  $P_{\mathbf{Z}}$  in (77) and (78), respectively. Alternatively, the quantity  $L(P_{\Theta|\mathbf{Z}}; P_{\mathbf{Z}})$  in (75) is the *loutum information* [55] between such random variables  $\Theta$  and  $\mathbf{Z}$ . The equality in (79) justifies the following equalities

$$\frac{1}{n} I(P_{\Theta|\mathbf{Z}}; P_{\mathbf{Z}}) = \int D(P_{Z|\Theta=\theta} \| P_{\mathbf{Z}}) dP_{\Theta}(\theta) \quad \text{and} \quad (80)$$

$$\frac{1}{n} L(P_{\Theta|\mathbf{Z}}; P_{\mathbf{Z}}) = \int D(P_{\mathbf{Z}} \| P_{Z|\Theta=\theta}) dP_{\Theta}(\theta), \quad (81)$$

which are well known properties of mutual information [56] and loutum information [55].

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

**Theorem 10** *The doubly-expected generalization gap  $\overline{\overline{G}}(P_{\Theta|\mathbf{Z}}, P_{\mathbf{Z}})$  in (72), with  $P_{\mathbf{Z}} \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$ , satisfies*

$$\begin{aligned} \frac{1}{\beta} \overline{\overline{G}}(P_{\Theta|\mathbf{Z}}, P_{\mathbf{Z}}) &= \frac{1}{n} L(P_{\Theta|\mathbf{Z}}; P_{\mathbf{Z}}) + \int D\left(P_{Z|\Theta=\theta} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}\right) dP_{\Theta}(\theta) \\ &\quad - \int D\left(P_{\mathbf{Z}} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}\right) dP_{\Theta}(\theta) \end{aligned} \quad (82)$$

where the term  $L(P_{\Theta|\mathbf{Z}}; P_{\mathbf{Z}})$  is in (75); the measure  $P_{Z|\Theta}$  is defined in (79); and the measure  $P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}$  is the WCDG probability measure in (21), with reference measure  $P_{Z|\Theta=\theta}$  and

$$\beta \in \left\{ t > 0 : \forall \theta \in \mathcal{M}, \int \exp\left(\frac{1}{t} \ell(\theta, x, y)\right) dP_{Z|\Theta=\theta}(x, y) < +\infty \right\}. \quad (83)$$

*Proof:* The proof is presented in Appendix P. ■

Theorem 10 appears to be the first exact characterization of the generalization gap induced by an arbitrary algorithm  $P_{\Theta|\mathbf{Z}}$  in terms of information measures. The following corollary of Theorem 10 unveils an upper bound, which is shown to be intimately linked to the celebrated Gibbs algorithm, and adds interest to the characterization of the generalization gap via the WCDG probability measure.

**Theorem 11** *The doubly-expected generalization gap  $\overline{\overline{G}}(P_{\Theta|Z}, P_Z)$ , in (72), with  $P_Z \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$ , satisfies*

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) \leq \frac{\beta}{n} (I(P_{\Theta|Z}; P_Z) + L(P_{\Theta|Z}; P_Z)), \quad (84)$$

where the terms  $I(P_{\Theta|Z}; P_Z)$  and  $L(P_{\Theta|Z}; P_Z)$  are in (73) and (75); and  $\beta$  satisfies for all  $\theta \in \mathcal{M}$ ,  $D\left(P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_Z|\Theta=\theta, \beta)}\right) = 0$ , with the measures  $P_Z$  and  $P_{\hat{Z}|\Theta=\theta}^{(P_Z|\Theta=\theta, \beta)}$  defined in (82).

*Proof:* The proof is presented in Appendix Q. ■

Note that the exact calculation of the expression in Theorem 10, and even the upper bound in Theorem 11, requires solving a problem of the form in (15) or (16), with reference measure  $P_{Z|\Theta=\theta}$ , for all  $\theta \in \mathcal{M}$  and some regularization parameter  $\beta > 0$  that satisfies (83). In a nutshell, one WCDG probability measure must be calculated for each element of the set of models  $\mathcal{M}$ . Calculating the upper bound in Theorem 11 is not easier. Obtaining the  $\beta$  in (84) requires solving the equality  $D\left(P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_Z|\Theta=\theta, \beta)}\right) = 0$  at least once for all models in  $\mathcal{M}$  to verify (83). Despite this disproportionate complexity, thanks to this characterization in terms of the WCDG probability measure, it is easy to show that there exists a Gibbs probability measure that achieves such an upper bound in Theorem 11.

## 7.2 The Gibbs Algorithm

A typical example of a statistical learning algorithm is the Gibbs algorithm, which is parametrized by a positive real  $\lambda$  and by a  $\sigma$ -measure  $Q \in \Delta(\mathcal{M}, \mathcal{B}(\mathcal{M}))$  [21]. In the following, the focus is on the case in which  $Q$  is a probability measure. Under this assumption, the probability measure representing such an algorithm, which is denoted by  $P_{\Theta|Z}^{(Q, \lambda)}$  with  $\lambda > 0$ , satisfies for all  $\theta \in \text{supp } Q$  and for all  $z \in (\mathcal{X} \times \mathcal{Y})^n$ ,

$$\frac{dP_{\Theta|Z=z}^{(Q, \lambda)}}{dQ}(\theta) = \exp\left(-K_{Q,z}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}L(z, \theta)\right), \quad (85)$$

where the dataset  $z$  represents the training dataset; and the function  $K_{Q,z} : \mathbb{R} \rightarrow \mathbb{R}$  satisfies

$$K_{Q,z}(t) = \log\left(\int \exp(tL(z, \theta)) dQ(\theta)\right). \quad (86)$$

The doubly-expected generalization gap induced by the Gibbs algorithm with parameters  $Q$  and  $\lambda$ , under the assumption that datasets are sampled from a product distribution  $P_Z \in \Delta((\mathcal{X} \times \mathcal{Y})^n)$  formed by the measure  $P_Z$  is

$$\overline{\overline{G}}(P_{\Theta|Z}^{(Q, \lambda)}, P_Z), \quad (87)$$



where the functional  $\overline{G}$  is defined in (72) and the measure  $P_{\mathbf{Z}} \in \Delta(\mathcal{X} \times \mathcal{Y})^n$  is a product measure obtained from  $P_{\mathbf{Z}}$ . Such a doubly-expected generalization gap satisfies the following property.

**Lemma 9 (Generalization Gap of the Gibbs Algorithm)** *The generalization gap  $\overline{G}(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}, P_{\mathbf{Z}})$  in (87) satisfies*

$$\overline{G}(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}, P_{\mathbf{Z}}) = \lambda \left( I(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}}) + L(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}}) \right), \quad (88)$$

where the terms  $I(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}})$  and  $L(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}})$  are, respectively, a mutual information and a lautum information:

$$I(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}}) \triangleq \int D(P_{\Theta|\mathbf{Z}=\nu}^{(Q,\lambda)} \| P_{\Theta}^{(Q,\lambda)}) dP_{\mathbf{Z}}(\nu); \text{ and} \quad (89)$$

$$L(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}}) \triangleq \int D(P_{\Theta}^{(Q,\lambda)} \| P_{\Theta|\mathbf{Z}=\nu}^{(Q,\lambda)}) dP_{\mathbf{Z}}(\nu), \quad (90)$$

with  $P_{\Theta}^{(Q,\lambda)}$  being a measure such that for all measurable subsets  $\mathcal{A}$  of  $\mathcal{M}$ ,

$$P_{\Theta}^{(Q,\lambda)}(\mathcal{A}) = \int P_{\Theta|\mathbf{Z}=\mathbf{z}}^{(Q,\lambda)}(\mathcal{A}) dP_{\mathbf{Z}}(\mathbf{z}). \quad (91)$$

*Proof:* This result has been proved before in the case in which  $Q$  is a probability measure in [15]; and in the more general case in which  $Q$  is a  $\sigma$ -finite measure in [21]. A proof for the particular case in which  $Q$  is a probability measure and datasets are formed by independent and identically distributed datapoints is presented in [57]  $\blacksquare$

Note that  $I(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}})$  in (89) and  $L(P_{\Theta|\mathbf{Z}}^{(Q,\lambda)}; P_{\mathbf{Z}})$  in (90) also satisfy that

$$I(P_{\mathbf{Z}|\Theta=\theta}^{(Q,\lambda)}; P_{\Theta}^{(Q,\lambda)}) = \int D(P_{\mathbf{Z}|\Theta=\theta}^{(Q,\lambda)} \| P_{\mathbf{Z}}) dP_{\Theta}^{(Q,\lambda)}(\theta); \quad (92)$$

and

$$L(P_{\mathbf{Z}|\Theta=\theta}^{(Q,\lambda)}; P_{\Theta}^{(Q,\lambda)}) = \int D(P_{\mathbf{Z}} \| P_{\mathbf{Z}|\Theta=\theta}^{(Q,\lambda)}) dP_{\Theta}^{(Q,\lambda)}(\theta), \quad (93)$$

where  $P_{\mathbf{Z}|\Theta}^{(Q,\beta)} \in \Delta((\mathcal{X} \times \mathcal{Y})^n | \mathcal{M})$  is the conditional probability measure that satisfies

$$P_{\mathbf{Z}}(\mathcal{B}) = \int P_{\mathbf{Z}|\Theta=\theta}^{(Q,\beta)}(\mathcal{B}) dP_{\Theta}^{(Q,\beta)}(\theta). \quad (94)$$

Moreover, for all  $\theta \in \mathcal{M}$ , the measure  $P_{\mathbf{Z}|\Theta=\theta}^{(Q,\beta)} \in \Delta((\mathcal{X} \times \mathcal{Y})^n | \mathcal{M})$  is a product measure formed by a measure  $P_{\mathcal{Z}_i|\Theta=\theta}^{(Q,\beta)} \in \Delta(\mathcal{X} \times \mathcal{Y} | \mathcal{M})$  that satisfies for all measurable sets  $\mathcal{A}_i$  in  $\mathcal{X} \times \mathcal{Y}$ , with  $i \in \{1, 2, \dots, n\}$ ,

$$P_{\mathbf{Z}|\Theta=\theta}^{(Q,\beta)}(\mathcal{A}_1 \times \mathcal{A}_2 \times \dots \times \mathcal{A}_n) = \prod_{t=1}^n P_{\mathcal{Z}_t|\Theta=\theta}^{(Q,\beta)}(\mathcal{A}_t). \quad (95)$$

Using this notation, Theorem 11 and Lemma 9 lead to the following theorem.

**Theorem 12** *Assume that there exists a  $\lambda > 0$  that satisfies*

$$\lambda \in \left\{ t > 0 : \forall \boldsymbol{\theta} \in \mathcal{M}, \int \exp\left(\frac{1}{nt} \ell(\boldsymbol{\theta}, x, y)\right) dP_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}(x, y) < +\infty \right\}, \quad (96)$$

where the function  $\ell$  is defined in (4); and the measure  $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}$  is defined in (95). Let the measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}, n\lambda)}$  be the WCDG probability measure of the form in (21). If  $\lambda$  satisfies for all  $\boldsymbol{\theta} \in \mathcal{M}$ ,

$$D\left(P_Z \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}, n\lambda)}\right) = 0, \quad (97)$$

then, for all  $P_{\boldsymbol{\Theta}|Z} \in \Delta(\mathcal{M} | (\mathcal{X} \times \mathcal{Y})^n)$ , the doubly-expected generalization gaps  $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|Z}, P_Z)$  in (72) and  $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|Z}^{(Q,\lambda)}, P_Z)$  in (87) satisfy

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|Z}, P_Z) \leq \overline{\overline{G}}(P_{\boldsymbol{\Theta}|Z}^{(Q,\lambda)}, P_Z). \quad (98)$$

*Proof:* The proof follows by choosing  $\lambda = \frac{\beta}{n}$  in Theorem 11 and verifying that under such a choice the resulting expressions in (84) and (88) are identical. ■

Theorem 12 shows that, under the assumption that datasets are sampled from  $P_Z$ , the doubly-expected generalization gap of any algorithm  $P_{\boldsymbol{\Theta}|Z}$  is upper-bounded by the doubly-expected generalization gap induced by a particular Gibbs algorithm  $P_{\boldsymbol{\Theta}|Z}^{(Q,\lambda)}$ . Such a particular Gibbs algorithm induces a posterior for a given model  $\boldsymbol{\theta}$ , denoted  $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}$  in (95). When such a posterior is used as a reference measure to build the WCDG probability measure for such a model  $\boldsymbol{\theta}$ , with parameter  $n\lambda$ , it leads to the probability measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}, n\lambda)}$ . Surprisingly, from (97), it follows that for all  $\boldsymbol{\theta} \in \mathcal{M}$ , such a WCDG probability measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}, n\lambda)}$  is identical to the actual GTDG probability measure  $P_Z$ . Hence, one can conclude that by choosing the reference measure  $P_S$  in (21) to be dependent on the models, e.g.,  $P_S = P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}$ , it is possible to observe that the resulting WCDG probability measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}, n\lambda)}$  becomes invariant with respect to  $\boldsymbol{\theta}$ . This is reminiscent of the principle of *indifference* over which is built the notion of equilibrium in zero-sum games with noisy observations. Nonetheless, this game-theoretic analysis is beyond the scope of this paper. The interested reader is referred to [58] and references therein.

The inequality in (98) reveals the central role of the Gibbs algorithm in statistical machine learning. Essentially, by studying the Gibbs algorithm  $P_{\boldsymbol{\Theta}|Z}^{(Q,\lambda)}$ , for

which its parameters  $Q$  and  $\lambda$  are chosen to satisfy (97), the doubly-expected generalization gap of any algorithm facing data sampled from  $P_Z = P_{\hat{Z}|\Theta=\theta}^{(P_Z^{(Q,\lambda)}, n\lambda)}$  can be upper bounded.

## 8 Conclusions and Final Remarks

The WCDG probability measure in Theorem 1 has been shown to be a cornerstone of statistical machine learning. This is backed by the fact that fundamental performance metrics, such as the sensitivity of the expected loss, the sensitivity of the empirical risk, the expected generalization gap, and the doubly-expected generalization gap are shown to have closed-form expressions involving such a measure. Interestingly, the WCDG probability measure is a Gibbs probability measure that, for a fixed model, induces an empirical risk that is a sub-Gaussian random variable. All the cumulants of the WCDG are finite and explicit expressions for upper and lower bounds on all cumulants are derived in terms of the cumulants of the reference measure. This analysis has led to the notion of  $(\epsilon, \delta)$ -robustness, which allows the study of the generalization capabilities of any model when it faces data generated from the WCDG probability measure. Finally, thanks to the WCDG, the first explicit expression in terms of information measures of the doubly-expected generalization gap (or generalization error) for any statistical learning algorithm has been presented. This expression has been distilled to obtain an upper-bound on the doubly-expected generalization gap consisting of the sum of the mutual information and the lautum information between the models and the datasets. The bound is shown to be tight for a Gibbs algorithm. This observation reveals the central role of the Gibbs algorithm in the characterization of the doubly-expected generalization gap.

# Appendices

## A Proof of Lemma 1

*Proof:* The proof of convexity is as follows. Let  $(\gamma_1, \gamma_2) \in \mathbb{R}^2$ , with  $\gamma_1 \neq \gamma_2$  being fixed. Assume that  $J_{P_S, \theta}(\gamma_1) < +\infty$  and  $J_{P_S, \theta}(\gamma_2) < +\infty$ . Then, for all  $\alpha \in (0, 1)$ , the following holds

$$\begin{aligned} & \alpha J_{P_S, \theta}(\gamma_1) + (1 - \alpha) J_{P_S, \theta}(\gamma_2) \\ &= \alpha \log \left( \int \exp(\gamma_1 t) dP_S(t) \right) + (1 - \alpha) \log \left( \int \exp(\gamma_2 t) dP_S(t) \right) \end{aligned} \quad (99)$$

$$= \log \left( \left( \int \exp(\gamma_1 t) dP_S(t) \right)^\alpha \right) + \log \left( \left( \int \exp(\gamma_2 t) dP_S(t) \right)^{1-\alpha} \right) \quad (100)$$

$$= \log \left( \left( \int \exp(\gamma_1 t) dP_S(t) \right)^\alpha \left( \int \exp(\gamma_2 t) dP_S(t) \right)^{1-\alpha} \right) \quad (101)$$

$$= \log \left( \left( \int \exp(\gamma_1 \alpha t)^p dP_S(t) \right)^{\frac{1}{p}} \left( \int \exp(\gamma_2 (1 - \alpha)t)^q dP_S(t) \right)^{\frac{1}{q}} \right) \quad (102)$$

$$\geq \log \left( \int \exp(\gamma_1 \alpha t) \exp(\gamma_2 (1 - \alpha)t) dP_S(t) \right) \quad (103)$$

$$= \log \left( \int \exp(\gamma_1 \alpha t + \gamma_2 (1 - \alpha)t) dP_S(t) \right) \quad (104)$$

$$= J_{P_S, \theta}(\gamma_1 \alpha + \gamma_2 (1 - \alpha)), \quad (105)$$

where the equality in (102) follows with  $\alpha \triangleq \frac{1}{p}$  and  $1 - \alpha \triangleq \frac{1}{q}$ ; and the inequality in (103) follows from Hölder's inequality [59, Theorem 2.4.5]. This proves the convexity of the function  $J_{P_S, \theta}$ .

Note that, also from Hölder's inequality [59, Theorem 2.4.5], the equality in (103) holds if and only if there exist two constants  $\beta_1$  and  $\beta_2$ , not simultaneously equal to zero, such that the set

$$\mathcal{A} = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : \beta_1 \exp(\gamma_1 \ell(\boldsymbol{\theta}, x, y)) = \beta_2 \exp(\gamma_2 \ell(\boldsymbol{\theta}, x, y))\} \quad (106)$$

$$= \left\{ (x, y) \in \mathcal{X} \times \mathcal{Y} : \exp((\gamma_1 - \gamma_2) \ell(\boldsymbol{\theta}, x, y)) = \frac{\beta_2}{\beta_1} \right\} \quad (107)$$

$$= \left\{ (x, y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = \frac{\log\left(\frac{\beta_2}{\beta_1}\right)}{(\gamma_1 - \gamma_2)} \right\}, \quad (108)$$

satisfies  $P_S(\mathcal{A}) = 1$ . Note that if such a set  $\mathcal{A}$  exists, the loss function  $\ell$  is nonseparable (Definition 2). This proves the strict convexity for separable loss functions.

The proof of monotonicity follows from noticing that for all  $(t_1, t_2) \in \{t \in \mathbb{R} : J_{P_S, \theta}(t) < +\infty\}^2$ , such that  $t_1 < t_2$ , it follows that for all  $\boldsymbol{\theta} \in \mathcal{M}$  and for all

$(x, y) \in \text{supp } P_S$ , the inequality  $\exp(t_1 \ell(\boldsymbol{\theta}, x, y)) \leq \exp(t_2 \ell(\boldsymbol{\theta}, x, y))$  holds. This implies that  $J_{P_S, \boldsymbol{\theta}}(t_1) \leq J_{P_S, \boldsymbol{\theta}}(t_2) < +\infty$ , which proves that the function is nondecreasing. Moreover, the equality holds if and only if  $P_S(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = 0\}) = 1$ .

The continuity of the function  $J_{P_S, \boldsymbol{\theta}}$  follows from observing that for all  $\alpha \in \{t \in \mathbb{R} : J_{P_S, \boldsymbol{\theta}}(t) < +\infty\}$ , it holds that

$$\lim_{t \rightarrow \alpha} J_{P_S, \boldsymbol{\theta}}(t) = \lim_{t \rightarrow \alpha} \log \left( \int \exp(t \ell(\boldsymbol{\theta}, x, y)) dP_S(x, y) \right) \quad (109)$$

$$= \log \left( \lim_{t \rightarrow \alpha} \int \exp(t \ell(\boldsymbol{\theta}, x, y)) dP_S(x, y) \right) \quad (110)$$

$$= \log \left( \int \lim_{t \rightarrow \alpha} \exp(t \ell(\boldsymbol{\theta}, x, y)) dP_S(x, y) \right) \quad (111)$$

$$= \log \left( \int \exp(\alpha \ell(\boldsymbol{\theta}, x, y)) dP_S(x, y) \right) \quad (112)$$

$$= J_{P_S, \boldsymbol{\theta}}(\alpha), \quad (113)$$

where equalities in (110) and (112) follow from the fact that both the logarithmic and exponential functions are continuous; and equality in (111) follows from the dominated convergence theorem [59, Theorem 1.6.9]. This proves that the function  $J_{P_S, \boldsymbol{\theta}}$  is continuous in the interior of  $\{t \in \mathbb{R} : J_{P_S, \boldsymbol{\theta}}(t) < +\infty\}$ .

The proof of differentiability follows by considering the transport of the measure  $P_S$  from  $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X} \times \mathcal{Y}})$  to  $([0, +\infty], \mathcal{B}([0, +\infty]))$  through the function  $g_{\boldsymbol{\theta}} : \mathcal{X} \times \mathcal{Y} \rightarrow [0, +\infty]$  such that  $g_{\boldsymbol{\theta}}(x, y) = \ell(\boldsymbol{\theta}, x, y)$ , with  $\ell$  in (4). Denote the resulting probability measure in  $([0, +\infty], \mathcal{B}([0, +\infty]))$  by  $T_{P_S, \boldsymbol{\theta}}$ . That is, for all  $\mathcal{A} \in \mathcal{B}([0, +\infty])$ ,

$$T_{P_S, \boldsymbol{\theta}}(\mathcal{A}) = P(g_{\boldsymbol{\theta}}^{-1}(\mathcal{A})), \quad (114)$$

where the term  $g_{\boldsymbol{\theta}}^{-1}(\mathcal{A})$  represents the set

$$g_{\boldsymbol{\theta}}^{-1}(\mathcal{A}) \triangleq \{(x, y) \in \mathcal{X} \times \mathcal{Y} : g_{\boldsymbol{\theta}}(x, y) \in \mathcal{A}\}.$$

Hence, the following holds

$$J_{P_S, \boldsymbol{\theta}}(t) = \log \left( \int \exp(t \ell(\boldsymbol{\theta}, x, y)) dP_S(x, y) \right) \quad (115)$$

$$= \log \left( \int \exp(tv) dT_{P_S, \boldsymbol{\theta}}(v) \right), \quad (116)$$

where equality in (116) follows from [59, Theorem 1.6.12].

Denote by  $\phi$  the Laplace transform of the measure  $P_S$ . That is for all  $t \in \{x \in \mathbb{R} : J_{P_S, \boldsymbol{\theta}}(x) < +\infty\}$ ,

$$\phi(t) = \int \exp(tv) dT_{P_S, \boldsymbol{\theta}}(v). \quad (117)$$

Hence,  $\phi(t) = \exp(\mathcal{J}_{P_S, \theta}(t))$ . From [60, Theorem 1a (page 439)], it follows that the function  $\phi$  has derivatives of all orders in  $\{t \in \mathbb{R} : \mathcal{J}_{P_S, \theta}(t) < +\infty\}$ , and thus, so does the function  $\mathcal{J}_{P_S, \theta}$  in the interior of  $\{t \in \mathbb{R} : \mathcal{J}_{P_S, \theta}(t) < +\infty\}$ . This completes the proof. ■

## B Proof of Lemma 2

*Proof:* Under the assumption that the set  $\mathcal{J}_{P_S, \theta}$  is not empty, there always exists a positive  $b \in \mathcal{J}_{P_S, \theta}$ , such that  $\mathcal{J}_{P_S, \theta}(\frac{1}{b}) < +\infty$ . From Lemma 1, it follows that the function  $\mathcal{J}_{P_S, \theta}$  is nondecreasing, which implies that  $[b, +\infty) \in \mathcal{J}_{P_S, \theta}$ . Moreover, let  $b^* \in [0, +\infty)$  be

$$b^* \triangleq \inf \mathcal{J}_{P_S, \theta}. \quad (118)$$

Hence, it follows that  $(b^*, +\infty) \in \mathcal{J}_{P_S, \theta}$ , which completes the proof. ■

## C Proof of Theorem 1

*Proof:* The optimization problem in (16) can be written as follows:

$$\begin{aligned} \max_{P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})} \quad & \int \ell(\theta, x, y) \frac{dP}{dP_S}(x, y) dP_S(x, y) \\ & - \beta \int \frac{dP}{dP_S}(x, y) \log \left( \frac{dP}{dP_S}(x, y) \right) dP_S(x, y) \end{aligned} \quad (119a)$$

$$\text{s.t.} \quad \int \frac{dP}{dP_S}(x, y) dP_S(x, y) = 1. \quad (119b)$$

The remainder of the proof focuses on the problem in which the optimization is over the Radon-Nikodym derivative  $\frac{dP}{dP_S}$  instead of the measure  $P$ . This is due to the fact that for all  $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ , the Radon-Nikodym derivative  $\frac{dP}{dP_S}$  is unique up to sets of zero measure with respect to  $P_S$ .

Let  $\mathcal{M}$  be the set of measurable functions  $\mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  with respect to the measurable spaces  $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X} \times \mathcal{Y}})$  and  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . Let  $\mathcal{S}$  be the subset of  $\mathcal{M}$  including all nonnegative functions that are absolutely integrable with respect to  $P_S$ . That is, for all  $h \in \mathcal{S}$ , it holds that

$$\int |h(x, y)| dP_S(x, y) < \infty. \quad (120)$$

Note that the set  $\mathcal{M}$  forms a real vector space and the set  $\mathcal{S}$  is a convex subset of  $\mathcal{M}$ . Note also that the constraint in (119b) is satisfied by the probability measure  $P_S$ , which also satisfies  $P_S \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ . Hence, the constraint does not induce an empty feasible set. Finally, note that the optimization problem

in (119) can be written as a minimization of the form:

$$\begin{aligned} \min_{g \in \mathcal{S}} \quad & - \int g(x, y) \ell(\boldsymbol{\theta}, x, y) dP_S(x, y) \\ & + \beta \int g(x, y) \log(g(x, y)) dP_S(x, y), \end{aligned} \quad (121a)$$

$$\text{s. t.} \quad \int g(x, y) dP_S(x, y) = 1, \quad (121b)$$

where the expression  $-\int g(x, y) \ell(\boldsymbol{\theta}, x, y) dP_S(x, y)$  is linear with  $g$ ; the expression  $\beta \int g(x, y) \log(g(x, y)) dP_S(x, y)$  is convex with  $g$ ; and  $\int g(x, y) dP_S(x, y)$  is linear with  $g$ .

The proof continues by assuming that the problem in (121) possesses a solution, which is denoted by  $g^* \in \mathcal{S}$ . Let  $\mu_0 \in [0, +\infty)$  be

$$\begin{aligned} \mu_0 \triangleq \min_{g \in \mathcal{S}} \quad & - \int g(x, y) \ell(\boldsymbol{\theta}, x, y) dP_S(x, y) + \beta \int g(x, y) \log(g(x, y)) dP_S(x, y) \\ \text{s. t.} \quad & \int g(x, y) dP_S(x, y) = 1 \end{aligned} \quad (122a)$$

$$= - \int g^*(x, y) \ell(\boldsymbol{\theta}, x, y) dP_S(x, y) + \beta \int g^*(x, y) \log(g^*(x, y)) dP_S(x, y). \quad (122b)$$

Let the Lagrangian of the optimization problem in (121) be the functional  $L : \mathcal{S} \times \mathbb{R} \rightarrow \mathbb{R}$  such that

$$\begin{aligned} L(g, \lambda) = & - \int g(x, y) \ell(\boldsymbol{\theta}, x, y) dP_S(x, y) + \beta \int g(x, y) \log(g(x, y)) dP_S(x, y) \\ & + \lambda \left( \int g(x, y) dP_S(x, y) - 1 \right), \end{aligned} \quad (123)$$

where  $\lambda$  is a real that acts as Lagrangian multiplier for the constraint in (119b). Let  $h : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  be a function in  $\mathcal{S}$ . The Gateaux differential of the function  $L$  in (123) at  $(h, \lambda) \in \mathcal{S} \times \mathbb{R}$  in the direction of  $g$ , if it exists, is

$$\partial L(g, \lambda; h) \triangleq \left. \frac{d}{d\alpha} r(\alpha) \right|_{\alpha=0}, \quad (124)$$

where the function  $r : \mathbb{R} \rightarrow \mathbb{R}$  is such that for all  $\alpha \in (-\epsilon, \epsilon)$ , with  $\epsilon > 0$  arbitrarily small,

$$\begin{aligned} r(\alpha) = & - \int (g(x, y) + \alpha h(x, y)) \ell(\boldsymbol{\theta}, x, y) dP_S(x, y) \\ & + \beta \int (g(x, y) + \alpha h(x, y)) \log(g(x, y) + \alpha h(x, y)) dP_S(x, y) \\ & + \lambda \int (g(x, y) + \alpha h(x, y)) dP_S(x, y) - \lambda, \end{aligned} \quad (125)$$

which can be rewritten as follows,

$$r(\alpha) = \int (\lambda - \ell(\boldsymbol{\theta}, x, y)) (g(x, y) + \alpha h(x, y)) dP_S(x, y) \quad (126)$$

$$+ \beta \int (g(x, y) + \alpha h(x, y)) \log (g(x, y) + \alpha h(x, y)) dP_S(x, y) - \lambda. \quad (127)$$

Note that the first term in (125) is linear with  $\alpha$ ; the second term can be written using the function  $\hat{r} : \mathbb{R} \rightarrow \mathbb{R}$  such that

$$\hat{r}(\alpha) = \int (g(x, y) + \alpha h(x, y)) \log (g(x, y) + \alpha h(x, y)) dP_S(x, y); \quad (128)$$

and the third term is independent of  $\alpha$ .

Hence, based on the fact the function  $\hat{r}$  in (128) is differentiable at zero [61, Lemma 9], so is the function  $r$  in (125), which implies that the Gateaux differential of  $\partial L(g, \lambda; h)$  in (124) exists.

The derivative of the real function  $r$  is

$$\begin{aligned} \frac{d}{d\alpha} r(\alpha) &= \frac{d}{d\alpha} \left( \int (\lambda - \ell(\boldsymbol{\theta}, x, y)) (g(x, y) + \alpha h(x, y)) dP_S(x, y) \right. \\ &\quad \left. + \beta \int (g(x, y) + \alpha h(x, y)) \log (g(x, y) + \alpha h(x, y)) dP_S(x, y) - \lambda \right) \quad (129) \end{aligned}$$

$$\begin{aligned} &= \int (\lambda - \ell(\boldsymbol{\theta}, x, y)) h(x, y) dP_S(x, y) \\ &\quad + \beta \int \frac{d}{d\alpha} \left( (g(x, y) + \alpha h(x, y)) \log (g(x, y) + \alpha h(x, y)) \right) dP_S(x, y) \quad (130) \end{aligned}$$

$$\begin{aligned} &= \int (\lambda - \ell(\boldsymbol{\theta}, x, y)) h(x, y) dP_S(x, y) \\ &\quad + \beta \int h(x, y) (1 + \log (g(x, y) + \alpha h(x, y))) dP_S(x, y), \quad (131) \end{aligned}$$

where equality in (130) follows from [61, Theorem 5].

From equation (124) and (131), it follows that

$$\begin{aligned} \partial L(g, \lambda; h) &= \int (\lambda - \ell(\boldsymbol{\theta}, x, y)) h(x, y) dP_S(x, y) \\ &\quad + \beta \int h(x, y) (1 + \log (g(x, y))) dP_S(x, y), \quad (132) \end{aligned}$$

$$= \int (\lambda - \ell(\boldsymbol{\theta}, x, y) + \beta (1 + \log (g(x, y)))) h(x, y) dP_S(x, y). \quad (133)$$

A necessary condition [62, Theorem 1, Page 178] for the functional  $L$  in (123) to have an extremum at  $\left( \frac{dP_{\mathcal{Z}}^{(P_S, \beta)}}{dP_S} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}}, \lambda \right) \in \mathcal{S} \times \mathbb{R}$  is that for all functions  $h \in \mathcal{S}$ ,



the following holds,

$$\begin{aligned} 0 &= \partial L \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}, \lambda; h \right) \\ &= \int h(x, y) \left( \lambda - \ell(\theta, x, y) + \beta \left( 1 + \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) \right) \right) \right) dP_S(x, y). \end{aligned} \quad (134)$$

The equality in (134) holds for all functions  $h \in \mathcal{S}$  if and only if for all  $(x, y) \in \text{supp } P_S$ , the function  $\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}$  satisfies:

$$\lambda - \ell(\theta, x, y) + \beta \left( 1 + \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) \right) \right) = 0. \quad (135)$$

That is,

$$\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) = \exp \left( \frac{-\lambda - \beta}{\beta} \right) \exp \left( \frac{\ell(\theta, x, y)}{\beta} \right), \quad (136)$$

where  $\lambda$  is chosen to satisfy (119b). That is,

$$\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) = \frac{\exp \left( \frac{\ell(\theta, x, y)}{\beta} \right)}{\int \exp \left( \frac{\ell(\theta, x, y)}{\beta} \right) dP_S(x, y)} \quad (137)$$

$$= \exp \left( \frac{\ell(\theta, x, y)}{\beta} - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right). \quad (138)$$

The proof continues by verifying that the measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  that satisfies (137) is the unique solution to the optimization problem in (16). Such verification is done by showing that the objective function in (16) is strictly concave with the optimization variable. Let  $P_1$  and  $P_2$  be two different probability measures in  $\Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$  and let  $\alpha$  be in  $(0, 1)$ . Hence,

$$\begin{aligned} &\int \ell(\theta, x, y) d(\alpha P_1 + (1 - \alpha)P_2)(x, y) - \beta D(\alpha P_1 + (1 - \alpha)P_2 \| P_S) \\ &= \alpha \int \ell(\theta, x, y) dP_1(x, y) + (1 - \alpha) \int \ell(\theta, x, y) dP_2(x, y) \\ &\quad - \beta D(\alpha P_1 + (1 - \alpha)P_2 \| P_S) \end{aligned} \quad (139)$$

$$\begin{aligned} &< \alpha \left( \int \ell(\theta, x, y) dP_1(x, y) - \beta D(P_1 \| P_S) \right) \\ &\quad + (1 - \alpha) \left( \int \ell(\theta, x, y) dP_2(x, y) - \beta D(P_2 \| P_S) \right), \end{aligned} \quad (140)$$

where equality in (139) follows from properties of the Lebesgue integral, while the inequality in (140) follows from [21, Theorem 2.2]. Hence, the uniqueness

of the solution is due to the strict concavity of the objective function. This completes the proof of uniqueness.

Based on [57, Theorem 3.1], for the case in which the function  $\ell$  is separable (Definition 18), it is proved that if  $\beta \in \mathcal{J}_{P_S, \theta}$ , is chosen to satisfy

$$D \left( P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S \right) = \gamma, \quad (141)$$

with  $\gamma$  in (15b), then the solutions to the optimization problems in (15) and (16), denoted by  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$ , are identical and satisfy

$$\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) = \exp \left( \frac{\ell(\theta, x, y)}{\beta} - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right), \quad (142)$$

where the functions  $\ell$  and  $J_{P_S, \theta}$  are defined in (4) and (19), respectively. This completes the proof.  $\blacksquare$

## D Proof of Lemma 3

First, the proof of absolute continuity is presented. For all  $\mathcal{C} \in \mathcal{F}_{\mathcal{X} \times \mathcal{Y}}$ ,

$$P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(\mathcal{C}) = \int_{\mathcal{C}} \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) dP_S(x, y), \quad (143)$$

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

$$P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(\mathcal{C}) = 0, \quad (144)$$

which implies the absolute continuity of  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  with respect to  $P_S$ . Alternatively, given a set  $\mathcal{C} \in \mathcal{F}_{\mathcal{X} \times \mathcal{Y}}$ , assume that  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(\mathcal{C}) = 0$ . Hence, it follows that

$$0 = P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(\mathcal{C}) = \int_{\mathcal{C}} \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) dP_S(x, y). \quad (145)$$

From Theorem 1, it holds that for all  $(x, y) \in \text{supp } P_S$ ,

$$\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) = \exp \left( \frac{\ell(\theta, x, y)}{\beta} - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right). \quad (146)$$

Note that if a solution to the optimization problem (16) exists, then  $J_{P_S, \theta} \left( \frac{1}{\beta} \right) < +\infty$ , which implies that  $\exp \left( -J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right) > 0$ . Moreover,  $\exp \left( \frac{\ell(\theta, x, y)}{\beta} \right) > 0$ , where the strict inequality is due to the fact that for all  $(x, y) \in \text{supp } P_S$ , the function  $\ell$  in (4) is nonnegative. Hence, for all  $(x, y) \in \text{supp } P_S$ ,

$$\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) = \exp \left( \frac{\ell(\theta, x, y)}{\beta} - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right) > 0, \quad (147)$$

which implies that  $P_S(\mathcal{C}) = 0$  and the absolute continuity of  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  with respect to  $P_S$ . This completes the first part of the proof.

The second part of the proof continues with the following equalities:

$$D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\|P_S\right) = \int \log\left(\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y)\right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) \quad (148)$$

$$= \int \log\left(\exp\left(\frac{\ell(\theta, x, y)}{\beta} - J_{P_S, \theta}\left(\frac{1}{\beta}\right)\right)\right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) \quad (149)$$

$$= \int \frac{\ell(\theta, x, y)}{\beta} dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) - J_{P_S, \theta}\left(\frac{1}{\beta}\right) \quad (150)$$

$$= \frac{1}{\beta} \mathbf{R}_\theta\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) - J_{P_S, \theta}\left(\frac{1}{\beta}\right), \quad (151)$$

where equality in (149) follows from (21) and the functional  $\mathbf{R}_\theta$  is defined in (8). This proves the equality in (23)

The equality in (24) is proved as follows:

$$D\left(P_S\|P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) = \int \log\left(\frac{dP_S}{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(x, y)\right) dP_S(x, y) \quad (152)$$

$$= \int \log\left(\exp\left(-\frac{\ell(\theta, x, y)}{\beta} + J_{P_S, \theta}\left(\frac{1}{\beta}\right)\right)\right) dP_S(x, y) \quad (153)$$

$$= - \int \frac{\ell(\theta, x, y)}{\beta} dP_S(x, y) + J_{P_S, \theta}\left(\frac{1}{\beta}\right) \quad (154)$$

$$= -\frac{1}{\beta} \mathbf{R}_\theta(P_S) + J_{P_S, \theta}\left(\frac{1}{\beta}\right), \quad (155)$$

where equality in (153) follows from equation (21). This completes the proof.

## E Proof of Lemma 4

The proof of equality in (32) is based on the fact that for a given  $t \in \{\nu \in \mathbb{R} : J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(\nu) < +\infty\}$ , the function  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}$  satisfies

$$J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(t) = \log\left(\int \exp(t\ell(\theta, x, y)) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y)\right) \quad (156)$$

$$= \log\left(\int \exp(t\ell(\theta, x, y)) \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) dP_S(x, y)\right) \quad (157)$$

$$= \log\left(\int \exp\left(\left(t + \frac{1}{\beta}\right)\ell(\theta, x, y)\right) dP_S(x, y) - J_{P_S, \theta}\left(\frac{1}{\beta}\right)\right) \quad (158)$$

$$= J_{P_S, \theta}\left(t + \frac{1}{\beta}\right) - J_{P_S, \theta}\left(\frac{1}{\beta}\right), \quad (159)$$

where the equality in (158) follows from (21). Moreover, from Lemma 1, it follows that the function  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}$  is continuous and nondecreasing. Hence, for all  $s \in (-\infty, t)$ , it holds that  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(s) \leq J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(t) < +\infty$ .

Let  $t^* \in \mathbb{R} \cup \{+\infty\}$  be defined by

$$t^* \triangleq \sup\{\nu \in \mathbb{R} : J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(\nu) < \infty\}. \quad (160)$$

If  $t^* = +\infty$ , then for all  $s \in \mathbb{R}$ ,  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(s) < +\infty$ , and the proof for (32) is completed. Alternatively, if  $t^* < +\infty$ , it follows that for all  $s > t^*$ ,  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(s) = +\infty$ , which implies that  $J_{P_S, \theta}(s + \frac{1}{\beta}) = +\infty$ , as the function  $J_{P_S, \theta}$  is also continuous (Lemma 1) and  $J_{P_S, \theta}(\frac{1}{\beta}) < +\infty$  (due to the choice of  $\beta$ ). Hence, in this case, the equality in (32) is of the form  $+\infty = +\infty$ . This completes the proof of (32).

## F Proof of Lemma 5

*Proof:* Note that for all  $t \in \text{int}\mathcal{J}_{P_S, \theta}$ , it holds that

$$J_{P_S, \theta}^{(1)}\left(\frac{1}{t}\right) = \frac{d}{ds} \log \left( \int \exp(s\ell(\theta, x, y)) dP_S(x, y) \right) \Big|_{s=\frac{1}{t}} \quad (161)$$

$$= \frac{\frac{d}{ds} \int (\exp(s\ell(\theta, x, y))) dP_S(x, y)}{\int \exp(s\ell(\theta, x, y)) dP_S(x, y)} \Big|_{s=\frac{1}{t}} \quad (162)$$

$$= \frac{\int \frac{d}{ds} (\exp(s\ell(\theta, x, y))) dP_S(x, y)}{\int \exp(s\ell(\theta, x, y)) dP_S(x, y)} \Big|_{s=\frac{1}{t}} \quad (163)$$

$$= \frac{\int \ell(\theta, x, y) \exp\left(\frac{1}{t}\ell(\theta, x, y)\right) dP_S(x, y)}{\int \exp\left(\frac{1}{t}\ell(\theta, x, y)\right) dP_S(x, y)} \quad (164)$$

$$= \exp\left(-J_{P_S, \theta}\left(\frac{1}{t}\right)\right) \int \ell(\theta, x, y) \exp\left(\frac{1}{t}\ell(\theta, x, y)\right) dP_S(x, y) \quad (165)$$

$$= \int \ell(\theta, x, y) \exp\left(\frac{1}{t}\ell(\theta, x, y) - J_{P_S, \theta}\left(\frac{1}{t}\right)\right) dP_S(x, y) \quad (166)$$

$$= \int \ell(\theta, x, y) dP_{\hat{Z}|\Theta=\theta}^{(P_S, t)}(x, y), \quad (167)$$

where the equality in (163) follows from the dominated convergence theorem [59, Theorem 1.6.9]; and the equality in (167) follows from (21). This completes the proof of (35).

The proof of (36) is as follows,

$$J_{P_S, \boldsymbol{\theta}}^{(2)} \left( \frac{1}{t} \right) = \frac{d}{ds} J_{P_S, \boldsymbol{\theta}}^{(1)}(s) \Big|_{s=\frac{1}{t}} \quad (168)$$

$$= \frac{d}{ds} \left( \int \ell(\boldsymbol{\theta}, x, y) \exp(s\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}(s)) dP_S(x, y) \right) \Big|_{s=\frac{1}{t}} \quad (169)$$

$$= \int \frac{d}{ds} \left( \ell(\boldsymbol{\theta}, x, y) \exp(s\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}(s)) \right) dP_S(x, y) \Big|_{s=\frac{1}{t}} \quad (170)$$

$$= \int \ell(\boldsymbol{\theta}, x, y) (\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right)) \exp \left( \frac{1}{t} \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) dP_S(x, y) \quad (171)$$

$$= \int \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 \exp \left( t\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) dP_S(x, y) \\ + \int \ell(\boldsymbol{\theta}, x, y) J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \exp \left( \frac{1}{t} \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) dP_S(x, y) \\ - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right)^2 \int \exp \left( \frac{1}{t} \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) dP_S(x, y) \quad (172)$$

$$= \int \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 \exp \left( \frac{1}{t} \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) dP_S(x, y) \\ + J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right)^2 - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right)^2 \int dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (173)$$

$$= \int \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 \exp \left( \frac{1}{t} \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) dP_S(x, y) \quad (174)$$

$$= \int \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y), \quad (175)$$

where the equality in (170) follows from the dominated convergence theorem [59, Theorem 1.6.9]; and the equalities in (173) and (175) follow from (21). This completes the proof of (36).

The second part of the proof of Lemma 1 is divided into two parts. First, it is shown that if the function  $\ell$  in (4) is nonseparable with respect to the measure  $P_S$  (Definition 2), then for all  $t \in \{s \in (0, +\infty) : J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{s} \right) < +\infty\}$ ,  $J_{P_S, \boldsymbol{\theta}}^{(2)} \left( \frac{1}{t} \right) = 0$ . Secondly, it shows that if the function  $\ell$  is separable, then, for all  $t \in \{s \in (0, +\infty) : J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{s} \right) < +\infty\}$ ,  $J_{P_S, \boldsymbol{\theta}}^{(2)} \left( \frac{1}{t} \right) > 0$ .

The first part is as follows. Assume that the function  $\ell$  in (4) is nonseparable with respect to the measure  $P_S$ , i.e., there exists a real  $a \geq 0$ , such that

$$P_S(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = a\}) = 1. \quad (176)$$

Note that from (21) and (36), it holds that

$$J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right) = \int \ell(\boldsymbol{\theta}, x, y) dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (177)$$

$$= \int \ell(\boldsymbol{\theta}, x, y) \exp\left(\frac{1}{t}\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}\left(\frac{1}{t}\right)\right) dP_S(x, y) \quad (178)$$

$$= \int a \exp\left(\frac{a}{t} - \frac{a}{t}\right) dP_S(x, y) \quad (179)$$

$$= a. \quad (180)$$

Note also that from (21) and (36), it holds that

$$J_{P_S, \boldsymbol{\theta}}^{(2)}\left(\frac{1}{t}\right) = \int \left(\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (181)$$

$$= \int \left(\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2 \exp\left(\frac{1}{t}\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}\left(\frac{1}{t}\right)\right) dP_S(x, y) \quad (182)$$

$$= \int (a - a) \exp\left(\frac{a}{t} - \frac{a}{t}\right) dP_S(x, y) \quad (183)$$

$$= 0, \quad (184)$$

which proves that if the function  $\ell$  is nonseparable, then  $J_{P_S, \boldsymbol{\theta}}^{(2)}\left(\frac{1}{t}\right) = 0$ . This completes the first part of the proof.

The second part is as follows. Assume that the function  $\ell$  is separable with respect to the measure  $P_S$ . That is, there exists a positive  $a$ ; and two subsets  $\mathcal{A}$  and  $\mathcal{B}$  of  $\mathcal{X} \times \mathcal{Y}$  that are nonnegligible with respect to  $P_S$  and verify that for all  $(x_1, y_1) \in \mathcal{A}$  and for all  $(x_2, y_2) \in \mathcal{B}$ ,

$$\ell(\boldsymbol{\theta}, x_1, y_1) < a < \ell(\boldsymbol{\theta}, x_2, y_2). \quad (185)$$

Note that the sets  $\mathcal{A}$  and  $\mathcal{B}$  are disjoint. Hence, from (36), it holds that

$$J_{P_S, \boldsymbol{\theta}}^{(2)}\left(\frac{1}{t}\right) = \int \left(\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (186)$$

$$= \int_{\mathcal{A}} \left(\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y)$$

$$+ \int_{\mathcal{B}} \left(\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y)$$

$$+ \int_{\mathcal{X} \times \mathcal{Y} \setminus (\mathcal{A} \cup \mathcal{B})} \left(\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (187)$$

$$> 0, \quad (188)$$

where the inequality in (188) is based on the following facts. First, if  $a < J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)$ , with  $a$  in (185), then for all  $(x, y) \in \mathcal{A}$ , it holds that  $\ell(\boldsymbol{\theta}, x, y) < a < J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)$ , and thus,

$$\left(\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2 > \left(a - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2. \quad (189)$$

This implies that

$$\int_{\mathcal{A}} \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (190)$$

$$> \int_{\mathcal{A}} \left( a - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (191)$$

$$> \left( a - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(\mathcal{A}) \quad (192)$$

$$> 0, \quad (193)$$

where the inequality in (193) follows from the fact that the probability measures  $P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}$  and  $P_S$  are mutually absolutely continuous (Lemma 3). Second, if  $a \geq J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right)$ , with  $a$  in (185), then for all  $(x, y) \in \mathcal{B}$ , it holds that  $\ell(\boldsymbol{\theta}, x, y) > a \geq J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right)$ , and thus,

$$\left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 > \left( a - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2, \quad (194)$$

which implies

$$\int_{\mathcal{B}} \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (195)$$

$$> \int_{\mathcal{B}} \left( a - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \quad (196)$$

$$> \left( a - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^2 P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(\mathcal{B}) \quad (197)$$

$$> 0, \quad (198)$$

where the inequality in (198) follows from the fact that the probability measures  $P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}$  and  $P_S$  are mutually absolutely continuous (Lemma 3). This completes the second part of the proof.

Finally, the proof of (37) is as follows,

$$\begin{aligned} & J_{P_S, \boldsymbol{\theta}}^{(3)} \left( \frac{1}{t} \right) \\ &= \frac{d}{ds} J_{P_S, \boldsymbol{\theta}}^{(2)}(s) \Big|_{s=\frac{1}{t}} \end{aligned} \quad (199)$$

$$= \int \frac{d}{ds} \left( \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)}(s) \right)^2 \exp(s\ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}(s)) \right) \Big|_{s=\frac{1}{t}} dP_S(x, y) \quad (200)$$

$$\begin{aligned} &= \int 2J_{P_S, \boldsymbol{\theta}}^{(2)} \left( \frac{1}{t} \right) \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right) \exp \left( \frac{1}{t} \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) \\ &\quad + \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^3 \exp \left( \frac{1}{t} \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{t} \right) \right) dP_S(x, y) \end{aligned} \quad (201)$$

$$\begin{aligned} &= 2J_{P_S, \boldsymbol{\theta}}^{(2)} \left( \frac{1}{t} \right) \int \ell(\boldsymbol{\theta}, x, y) dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) - 2J_{P_S, \boldsymbol{\theta}}^{(2)} \left( \frac{1}{t} \right) J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \\ &\quad + \int \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^3 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y) \end{aligned} \quad (202)$$

$$= \int \left( \ell(\boldsymbol{\theta}, x, y) - J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{t} \right) \right)^3 dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, t)}(x, y), \quad (203)$$

where the equality in (200) follows from the dominated convergence theorem [59, Theorem 1.6.9]; and the equalities in (202) and (203) follow from (21), (35) and (36). This completes the proof of (37).

The proofs for the equalities in (38), (39) and (40) follow along the same arguments as above. ■

## G Proof of Theorem 2

*Proof:* A direct result from Lemma 1 and the mean value theorem [51, Theorem 5.10] is

$$J_{P_S, \boldsymbol{\theta}}^{(m)} \left( \frac{1}{t} \right) - J_{P_S, \boldsymbol{\theta}}^{(m)}(0) = \frac{1}{t} J_{P_S, \boldsymbol{\theta}}^{(m+1)} \left( \frac{1}{s} \right) < +\infty, \quad (204)$$

for some  $s \in (t, +\infty)$ . Under the assumption that there exists real  $\delta > 0$  such that the function  $J_{P_S, \boldsymbol{\theta}}$  in (19) is differentiable within  $(-\delta, \delta)$ , it follows that the function  $J_{P_S, \boldsymbol{\theta}}^{(m+1)}$  is continuous in  $(-\delta, +\infty)$ , where  $\delta > 0$ . Hence, for all  $t > 0$ , the function  $J_{P_S, \boldsymbol{\theta}}^{(m+1)}$  achieves a maximum and a minimum within the interval  $[0, \frac{1}{t}]$ . Such extremas allow providing lower and upper bounds on  $J_{P_S, \boldsymbol{\theta}}^{(m)} \left( \frac{1}{t} \right)$  along with  $J_{P_S, \boldsymbol{\theta}}^{(m)}(0)$ . This completes the proof. ■



## H Proof of Theorem 3

*Proof:* This section is divided into two parts. First, a preliminary result that is needed for the proof of Theorem 3.

**Theorem 13** *For all  $t \in \mathbb{R}$ , the function  $J_{P_{\frac{Z}{\Theta}=\theta}^{(P_S, \beta), \theta}}$  in (31) verifies the following equality*

$$J_{P_{\frac{Z}{\Theta}=\theta}^{(P_S, \beta), \theta}}(t) = tJ_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) + \frac{1}{2}t^2J_{P_S, \theta}^{(2)}(\zeta_t), \quad (205)$$

with

$$\zeta_t \in \left( \min\left\{\frac{1}{\beta}, t + \frac{1}{\beta}\right\}, \max\left\{\frac{1}{\beta}, t + \frac{1}{\beta}\right\} \right), \quad (206)$$

where the functions  $J_{P_S, \theta}^{(1)}$  and  $J_{P_S, \theta}^{(2)}$  are defined in (35) and (36), respectively.

*Proof:* From Lemma 1, it follows that the function  $J_{P_S, \theta}$  is continuous, monotone, increasing and differentiable infinitely many times. Then, a Taylor expansion of  $J_{P_S, \theta}$  at the point  $\frac{1}{\beta}$  yields for all  $t \in \{\nu \in \mathbb{R} : J_{P_S, \theta}(\nu) < +\infty\}$ ,

$$J_{P_S, \theta}(t) = J_{P_S, \theta}\left(\frac{1}{\beta}\right) + \sum_{s=1}^{+\infty} \frac{J_{P_S, \theta}^{(s)}\left(\frac{1}{\beta}\right)}{s!} \left(t - \frac{1}{\beta}\right)^s. \quad (207)$$

Choosing  $t = \alpha + \frac{1}{\beta}$  in (207) with  $\alpha \in \{\nu \in \mathbb{R} : J_{P_S, \theta}\left(\alpha + \frac{1}{\beta}\right) < +\infty\}$ , it holds from the Taylor-Lagrange theorem [63, Theorem 2.5.4] that

$$J_{P_S, \theta}\left(\alpha + \frac{1}{\beta}\right) = J_{P_S, \theta}\left(\frac{1}{\beta}\right) + \alpha J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) + \frac{1}{2}\alpha^2 J_{P_S, \theta}^{(2)}(\xi), \quad (208)$$

where  $\xi \in \left( \min\left\{\frac{1}{\beta}, \alpha + \frac{1}{\beta}\right\}, \max\left\{\frac{1}{\beta}, \alpha + \frac{1}{\beta}\right\} \right)$ .

The equality in (208) leads to the following inequality:

$$J_{P_S, \theta}\left(\alpha + \frac{1}{\beta}\right) - J_{P_S, \theta}\left(\frac{1}{\beta}\right) = \alpha J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) + \frac{1}{2}\alpha^2 J_{P_S, \theta}^{(2)}(\zeta_t), \quad (209)$$

where  $\zeta_t$  is chosen as in (206).

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

$$s^* \triangleq \sup\{\alpha \in \mathbb{R} : J_{P_S, \theta}\left(\alpha + \frac{1}{\beta}\right) < \infty\} \quad (210)$$

If  $s^* = +\infty$ , then for all  $\alpha \in \mathbb{R}$ ,  $J_{P_S, \theta}\left(\alpha + \frac{1}{\beta}\right) - J_{P_S, \theta}\left(\frac{1}{\beta}\right) < +\infty$ , thus, the proof is completed by noticing that from Lemma 4, it holds that  $J_{P_{\frac{Z}{\Theta}=\theta}^{(P_S, \beta), \theta}}(\alpha) = J_{P_S, \theta}\left(\alpha + \frac{1}{\beta}\right) - J_{P_S, \theta}\left(\frac{1}{\beta}\right)$ . Hence, choosing  $\zeta_t$  as in (206) leads to the desired equality in (205).

Alternatively, if  $s^* < +\infty$ , it follows that for all  $\alpha > s^*$ ,  $J_{P_S, \theta} \left( \alpha + \frac{1}{\beta} \right) = +\infty$ . From Lemma 4, it follows that  $J_{P_{\mathcal{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(\alpha) = +\infty$ , which implies that  $+\infty = \alpha J_{P_S, \theta}^{(1)} \left( \frac{1}{\beta} \right) + \frac{1}{2} \alpha^2 J_{P_S, \theta}^{(2)}(\zeta_t)$ , and thus  $J_{P_S, \theta}^{(2)}(\zeta_t) = +\infty$ . Hence, in this case, the inequality in (44) is of the form  $+\infty = +\infty$ . This completes the proof. ■

The proof of Theorem 3 continues from here. The proof of the inequality in (44) is trivial from Theorem 13 and the choice of  $\zeta_{P_S, \theta}$  in (45). Hence, the remainder of the proof focuses on proving that  $\zeta_{P_S, \theta} < +\infty$ . From Lemma 1 and Lemma 4, it holds that

$$\{\alpha \in \mathbb{R} : J_{P_{\mathcal{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(\alpha) < +\infty\} = \left\{ \alpha \in \mathbb{R} : J_{P_S, \theta} \left( \alpha + \frac{1}{\beta} \right) < +\infty \right\}, \quad (211)$$

which implies that the set  $\{\alpha \in \mathbb{R} : J_{P_{\mathcal{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(\alpha) < +\infty\}$  is an interval of the form  $(-\infty, b - \frac{1}{\beta})$ .

For all  $\alpha \in (-\infty, b - \frac{1}{\beta})$ , the function  $J_{P_S, \theta}^{(2)}$  is continuous (Lemma 1). Hence, for all  $t \in \left( \min \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\}, \max \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\} \right) \subseteq (-\infty, b)$ , the value  $J_{P_S, \theta}^{(2)}(t)$  is finite. Moreover, the values  $J_{P_S, \theta}^{(2)} \left( \min \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\} \right)$  and  $J_{P_S, \theta}^{(2)} \left( \max \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\} \right)$  are both finite. This implies that the function  $J_{P_S, \theta}^{(2)}$  achieves a minimum and maximum within the closed interval  $\left[ \min \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\}, \max \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\} \right]$ . Thus, the corresponding term  $J_{P_S, \theta}^{(2)}(\xi_\alpha)$ , with  $\xi_\alpha \in \left( \min \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\}, \max \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\} \right)$ , is finite.

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

$$\lim_{\alpha \rightarrow -\infty} \left( \min \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\}, \max \left\{ \frac{1}{\beta}, \alpha + \frac{1}{\beta} \right\} \right) = \left( -\infty, \frac{1}{\beta} \right) \quad (212)$$

The function  $J_{P_S, \theta}^{(2)}$  is continuous in  $(-\infty, \frac{1}{\beta})$ , and thus, for all  $t \in (-\infty, \frac{1}{\beta})$ ,  $J_{P_S, \theta}^{(2)}(t) < +\infty$ . Moreover, given the fact that  $\beta \in \mathcal{J}_{P_S, \theta}$ , with  $\mathcal{J}_{P_S, \theta}$  in (20), it holds that

$$\lim_{t \rightarrow \frac{1}{\beta}} J_{P_S, \theta}^{(2)}(t) = J_{P_S, \theta}^{(2)} \left( \frac{1}{\beta} \right) < +\infty. \quad (213)$$

Alternatively,

$$\lim_{t \rightarrow -\infty} J_{P_S, \theta}^{(2)}(t) < +\infty. \quad (214)$$

Hence, from (212), (213) and (214), it follows that

$$\lim_{\alpha \rightarrow -\infty} J_{P_S, \theta}^{(2)}(\xi_\alpha) < +\infty. \quad (215)$$

On the otherhand, when  $\alpha \rightarrow b - \frac{1}{\beta}$ , it follows that  $\xi_\alpha \in (\frac{1}{\beta}, b)$ . Based on the choice of  $b$  in (46), it follows that

$$\lim_{t \rightarrow b} J_{P_S, \theta}^{(2)}(t) < +\infty. \quad (216)$$

Thus,

$$\lim_{\alpha \rightarrow b - \frac{1}{\beta}} J_{P_S, \theta}^{(2)}(\xi_\alpha) < +\infty. \quad (217)$$

From all the above, it holds that for all  $\alpha \in \{t \in \mathbb{R} : J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(t) < +\infty\}$ , the value  $J_{P_S, \theta}^{(2)}(\xi_\alpha)$  is finite, and this completes the proof. ■

## I Proof of Lemma 6

*Proof:* The proof continues with the following equalities,

$$\frac{d}{d\beta} D(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S) = \frac{d}{d\beta} \left( \frac{1}{\beta} R_\theta(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}) - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right) \quad (218)$$

$$= \frac{d}{d\beta} \left( \frac{1}{\beta} J_{P_S, \theta}^{(1)} \left( \frac{1}{\beta} \right) - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right) \quad (219)$$

$$= -\frac{1}{\beta^2} J_{P_S, \theta}^{(1)} \left( \frac{1}{\beta} \right) - \frac{1}{\beta^3} J_{P_S, \theta}^{(2)} \left( \frac{1}{\beta} \right) + \frac{1}{\beta^2} J_{P_S, \theta}^{(1)} \left( \frac{1}{\beta} \right) \quad (220)$$

$$= -\frac{1}{\beta^3} J_{P_S, \theta}^{(2)} \left( \frac{1}{\beta} \right), \quad (221)$$

where the equality in (218) follows from (23) and the equality in (219) follows from (35). This completes the proof. ■

## J Proof of Theorem 4

*Proof:* Let  $\mathcal{A}$  be a set such that

$$\mathcal{A} \triangleq \{z \in (\mathcal{X} \times \mathcal{Y})^n : \exp(tL(\theta, z)) \geq \exp(t\delta)\}, \quad (222)$$

with  $t \in \mathcal{J}_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(t)$  and  $\delta > J_{P_S, \theta}^{(1)}(\frac{1}{\beta})$ , where the set  $\mathcal{J}_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(t)$  is defined in (20). Hence, it follows that

$$\exp\left(J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(t)\right) = \int \exp(tL(\theta, z)) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(z) \quad (223)$$

$$\geq \int_{\mathcal{A}} \exp(tL(\theta, z)) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(z) \quad (224)$$

$$\geq \exp(t\delta) \int_{\mathcal{A}} dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(z) \quad (225)$$

$$= \exp(t\delta) P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(\mathcal{A}), \quad (226)$$

where equality in (223) follows from (19) with  $P_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}$  as a product measure constructed by  $P_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}$  in (21).

Note that

$$\mathcal{A} \triangleq \{\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n : \exp(t\mathbf{L}(\boldsymbol{\theta}, \mathbf{z})) \geq \exp(t\delta)\}, \quad (227)$$

is equivalent to

$$\mathcal{A} \triangleq \{\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n : \mathbf{L}(\boldsymbol{\theta}, \mathbf{z}) \geq \delta\}, \quad (228)$$

from the fact that  $t > 0$ .

Hence, it follows from (226) and (228) that

$$\begin{aligned} & P_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}(\{\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n : \mathbf{L}(\boldsymbol{\theta}, \mathbf{z}) \geq \delta\}) \\ & \leq \frac{\exp\left(\mathbf{J}_{P_{\hat{\mathbf{Z}}|\Theta=\theta, \boldsymbol{\theta}}^{(P_S, \beta)}}(t)\right)}{\exp(t\delta)} \end{aligned} \quad (229)$$

$$= \frac{\int \exp(t\mathbf{L}(\boldsymbol{\theta}, \mathbf{z})) dP_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}(\mathbf{z})}{\exp(t\delta)} \quad (230)$$

$$= \frac{\int \exp\left(\frac{t}{n} \sum_{i=1}^n \ell(\boldsymbol{\theta}, x_i, y_i)\right) dP_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}(\mathbf{z})}{\exp(t\delta)} \quad (231)$$

$$= \frac{\left(\int \exp\left(\frac{t}{n} \ell(\boldsymbol{\theta}, x, y)\right) dP_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}(x, y)\right)^n}{\exp(t\delta)} \quad (232)$$

$$= \frac{\left(\exp\left(\mathbf{J}_{P_{\hat{\mathbf{Z}}|\Theta=\theta, \boldsymbol{\theta}}^{(P_S, \beta)}}\left(\frac{t}{n}\right)\right)\right)^n}{\exp(t\delta)} \quad (233)$$

$$= \exp\left(n\mathbf{J}_{P_S, \boldsymbol{\theta}}\left(\frac{t}{n} + \frac{1}{\beta}\right) - n\mathbf{J}_{P_S, \boldsymbol{\theta}}\left(\frac{1}{\beta}\right) - t\delta\right), \quad (234)$$

where equality in (234) follows from (32).

Furthermore,

$$\begin{aligned} & P_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}(\{\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n : \mathbf{L}(\boldsymbol{\theta}, \mathbf{z}) \geq \delta\}) \\ & \leq \exp\left(n\mathbf{J}_{P_S, \boldsymbol{\theta}}\left(\frac{t^*}{n} + \frac{1}{\beta}\right) - n\mathbf{J}_{P_S, \boldsymbol{\theta}}\left(\frac{1}{\beta}\right) - t^*\delta\right) \end{aligned} \quad (235)$$

$$= \exp\left(n\left(\mathbf{J}_{P_S, \boldsymbol{\theta}}\left(\frac{t^*}{n} + \frac{1}{\beta}\right) - \mathbf{J}_{P_S, \boldsymbol{\theta}}\left(\frac{1}{\beta}\right) - \frac{t^*}{n}\delta\right)\right), \quad (236)$$

where  $t^* \in \mathcal{J}_{P_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}, \boldsymbol{\theta}}$  is

$$t^* \in \arg \min_{t \in \mathcal{J}_{P_{\hat{\mathbf{Z}}|\Theta=\theta}^{(P_S, \beta)}, \boldsymbol{\theta}}} \mathbf{J}_{P_S, \boldsymbol{\theta}}\left(\frac{t}{n} + \frac{1}{\beta}\right) - \frac{t}{n}\delta. \quad (237)$$

Since the function  $J_{P_S, \boldsymbol{\theta}} \left( \frac{t}{n} + \frac{1}{\beta} \right) - \frac{t}{n} \delta$  is strictly convex, there always exists a  $t$  that satisfies the following:

$$0 = \frac{d}{dt} \left( J_{P_S, \boldsymbol{\theta}} \left( \frac{t}{n} + \frac{1}{\beta} \right) - \frac{t}{n} \delta \right) \quad (238)$$

$$= \frac{1}{n} J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{t}{n} + \frac{1}{\beta} \right) - \frac{1}{n} \delta. \quad (239)$$

To ensure the existence of  $t^*$ , since the function  $J_{P_S, \boldsymbol{\theta}}^{(1)}$  is nondecreasing, based on the equality in (36),  $\delta$  must satisfy

$$\delta \geq J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{t^*}{n} + \frac{1}{\beta} \right). \quad (240)$$

Let  $\beta^* > 0$  be

$$\frac{1}{\beta^*} \triangleq \frac{t^*}{n} + \frac{1}{\beta}. \quad (241)$$

Hence, from (240), it holds that  $\beta^*$  must satisfy

$$\delta = J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{\beta^*} \right), \quad (242)$$

and the inequality in (235) becomes

$$\begin{aligned} & P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \left( \{ \mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^n : \mathbf{L}(\boldsymbol{\theta}, \mathbf{z}) \geq \delta \} \right) \\ & \leq \exp \left( n \left( J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{\beta^*} \right) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{\beta} \right) - \left( \frac{1}{\beta^*} - \frac{1}{\beta} \right) J_{P_S, \boldsymbol{\theta}}^{(1)} \left( \frac{1}{\beta^*} \right) \right) \right), \end{aligned} \quad (243)$$

Note that

$$\begin{aligned} & J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{\beta^*} \right) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{\beta} \right) - \left( \frac{1}{\beta^*} - \frac{1}{\beta} \right) \int \ell(\boldsymbol{\theta}, x, y) dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)}(x, y) \\ & = \frac{1}{\beta^*} \int \ell(\boldsymbol{\theta}, x, y) dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)}(x, y) - D \left( P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)} \| P_S \right) - J_{P_S, \boldsymbol{\theta}} \left( \frac{1}{\beta} \right) \\ & \quad - \left( \frac{1}{\beta^*} - \frac{1}{\beta} \right) \int \ell(\boldsymbol{\theta}, x, y) dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)}(x, y) \end{aligned} \quad (244)$$

$$\begin{aligned} & = \frac{1}{\beta} \left( \int \ell(\boldsymbol{\theta}, x, y) dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)}(x, y) - \int \ell(\boldsymbol{\theta}, x, y) dP_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}(x, y) \right) \\ & \quad + D \left( P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \| P_S \right) - D \left( P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)} \| P_S \right), \end{aligned} \quad (245)$$

$$= \frac{1}{\beta} \left( \mathbf{R}_{\boldsymbol{\theta}} \left( P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)} \right) - \mathbf{R}_{\boldsymbol{\theta}} \left( P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \right) \right) \quad (246)$$

$$+ D \left( P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \| P_S \right) - D \left( P_{\hat{\mathbf{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta^*)} \| P_S \right), \quad (247)$$

where equalities in (244) and (247) both follow from (23); and the function  $\mathbf{R}_{\boldsymbol{\theta}}$  is in (8).

Note that the relative entropy  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right)$  satisfies:

$$D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right) = \int \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}{dP_S}(x, y) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) \quad (248)$$

$$= \int \left( \frac{\ell(\theta, x, y)}{\beta} - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) \quad (249)$$

$$= \int \frac{\ell(\theta, x, y)}{\beta} dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) - J_{P_S, \theta} \left( \frac{1}{\beta} \right) \quad (250)$$

$$= \frac{1}{\beta} \mathbf{R}_\theta(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}) - J_{P_S, \theta} \left( \frac{1}{\beta} \right), \quad (251)$$

where the equality in (249) follows from (21).

The proof continues by noticing that the relative entropies  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right)$  and  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_S\right)$  satisfy:

$$\begin{aligned} & D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) - D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_S\right) \\ &= \int \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}}{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(x, y) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \\ & \quad - \int \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}}{dP_S}(x, y) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \end{aligned} \quad (252)$$

$$= \int \left( \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}}{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(x, y) \right) - \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}}{dP_S}(x, y) \right) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \quad (253)$$

$$= \int \log \left( \frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}}{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(x, y) \frac{dP_S}{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}}(x, y) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \quad (254)$$

$$= \int \log \left( \frac{dP_S}{dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}}(x, y) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \quad (255)$$

$$= \int \log \left( \exp \left( -\frac{\ell(\theta, x, y)}{\beta} + J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \quad (256)$$

$$= \int \left( -\frac{\ell(\theta, x, y)}{\beta} + J_{P_S, \theta} \left( \frac{1}{\beta} \right) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \quad (257)$$

$$= - \int \frac{\ell(\theta, x, y)}{\beta} dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) + J_{P_S, \theta} \left( \frac{1}{\beta} \right) \quad (258)$$

$$= -\frac{1}{\beta} \mathbf{R}_\theta(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}) + J_{P_S, \theta} \left( \frac{1}{\beta} \right). \quad (259)$$

Therefore, from (251) and (259), it follows that

$$\begin{aligned} & D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_S\right) - D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) - D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \| P_S\right) \\ &= \frac{1}{\beta} \mathbf{R}_\theta(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}) - \frac{1}{\beta} \mathbf{R}_\theta(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}). \end{aligned} \quad (260)$$

Plugging (260) into (247) yields

$$\begin{aligned} & \mathbf{J}_{P_S, \theta}\left(\frac{1}{\beta^*}\right) - \mathbf{J}_{P_S, \theta}\left(\frac{1}{\beta}\right) - \left(\frac{1}{\beta^*} - \frac{1}{\beta}\right) \int \ell(\theta, x, y) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)}(x, y) \\ &= -D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right). \end{aligned} \quad (261)$$

Plugging (261) into (243) yields

$$P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\left(\{z \in (\mathcal{X} \times \mathcal{Y})^n : \mathbf{L}(\theta, z) \geq \delta\}\right) \leq \exp\left(-nD\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta^*)} \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right)\right). \quad (262)$$

This completes the proof.  $\blacksquare$

## K Proof of Theorem 5

Consider the random variable  $\ell(\theta, X, Y)$ , with  $(X, Y) \sim P_Z$ , with  $P_Z \in \mathcal{N}(P_S, \gamma)$ . From Chebyshev's inequality [59, Section 2.4.9], it follows that

$$P_Z(\{z \in (\mathcal{X} \times \mathcal{Y})^n : \mathbf{L}(\theta, z) \geq \delta\}) \leq \frac{1}{\delta} \mathbf{R}_\theta(P_Z) \quad (263)$$

$$\leq \frac{1}{\delta} \mathbf{J}_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right), \quad (264)$$

where the functional  $\mathbf{R}_\theta$  is defined in (8) and the inequality in (264) follows from the fact that for all  $P_Z \in \mathcal{N}(P_S, \gamma)$ ,

$$\mathbf{R}_\theta(P_Z) \leq \mathbf{R}_\theta\left(P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) = \mathbf{J}_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right), \quad (265)$$

with the measure  $P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}$  in (21). This completes the proof.

## L Proof of Theorem 6

The proof follows from Theorem 1 and by noticing that the relative entropy  $D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}\|P_S\right)$  satisfies:

$$D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}\|P_S\right) = \int \log\left(\frac{dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}}{dP_S}(x,y)\right) dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}(x,y) \quad (266)$$

$$= \int \left(\frac{\ell(\theta, x, y)}{\beta} - J_{P_S, \theta}\left(\frac{1}{\beta}\right)\right) dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}(x,y) \quad (267)$$

$$= \int \frac{\ell(\theta, x, y)}{\beta} dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}(x,y) - J_{P_S, \theta}\left(\frac{1}{\beta}\right) \quad (268)$$

$$= \frac{1}{\beta} R_{\theta}(P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}) - J_{P_S, \theta}\left(\frac{1}{\beta}\right), \quad (269)$$

where the equality in (267) follows from (21) and the function  $R_{\theta}$  is in (8). The proof continues by noticing that the relative entropies  $D\left(P\|P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}\right)$  and  $D(P\|P_S)$  satisfy:

$$\begin{aligned} & D\left(P\|P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}\right) - D(P\|P_S) \\ &= \int \log\left(\frac{dP}{dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}}(x,y)\right) dP(x,y) - \int \log\left(\frac{dP}{dP_S}(x,y)\right) dP(x,y) \quad (270) \end{aligned}$$

$$= \int \left(\log\left(\frac{dP}{dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}}(x,y)\right) - \log\left(\frac{dP}{dP_S}(x,y)\right)\right) dP(x,y) \quad (271)$$

$$= \int \log\left(\frac{dP}{dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}}(x,y) \frac{dP_S}{dP}(x,y)\right) dP(x,y) \quad (272)$$

$$= \int \log\left(\frac{dP_S}{dP_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}}(x,y)\right) dP(x,y) \quad (273)$$

$$= \int \log\left(\exp\left(-\frac{\ell(\theta, x, y)}{\beta} + J_{P_S, \theta}\left(\frac{1}{\beta}\right)\right)\right) dP(x,y) \quad (274)$$

$$= \int \left(-\frac{\ell(\theta, x, y)}{\beta} + J_{P_S, \theta}\left(\frac{1}{\beta}\right)\right) dP(x,y) \quad (275)$$

$$= -\int \frac{\ell(\theta, x, y)}{\beta} dP(x,y) + J_{P_S, \theta}\left(\frac{1}{\beta}\right) \quad (276)$$

$$= -\frac{1}{\beta} R_{\theta}(P) + J_{P_S, \theta}\left(\frac{1}{\beta}\right). \quad (277)$$



Therefore, from (269) and (277), it follows that

$$\begin{aligned} & D(P\|P_S) - D\left(P\|P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}\right) - D\left(P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}\|P_S\right) \\ &= \frac{1}{\beta}R_{\theta}(P) - \frac{1}{\beta}R_{\theta}(P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}) \end{aligned} \quad (278)$$

$$= \frac{1}{\beta}G\left(\theta, P, P_{\hat{Z}|\Theta=\theta}^{(P_S,\beta)}\right), \quad (279)$$

which completes the proof.

## M Proof of Theorem 7

The proof of Theorem 7 requires the following lemma.

**Lemma 10** *Given two probability measures  $P_1$  and  $P_2$  over the measurable space  $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X} \times \mathcal{Y}})$ , with  $P_2$  absolutely continuous with respect to  $P_1$ , the following holds for all  $\theta \in \mathcal{M}$ ,*

$$\begin{aligned} G(\theta, P_2, P_1) \leq \inf_{t>0} & \left( \frac{D(P_2\|P_1)}{t} + \frac{1}{t} \log \left( \int \exp \left( t \left( \ell(\theta, x, y) \right. \right. \right. \right. \\ & \left. \left. \left. - \int \ell(\theta, x, y) dP_1(x, y) \right) \right) dP_1(x, y) \right), \end{aligned} \quad (280)$$

where the function  $\ell$  and the functional  $G(\theta, P_2, P_1)$  are defined in (4) and in (55), respectively.

*Proof:* From [64, Corollary 4.15, Page 100], it follows that the probability measures  $P_1$  and  $P_2$  in  $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X} \times \mathcal{Y}})$  satisfy the following equality:

$$D(P_2\|P_1) = \sup_f \int f(x, y) dP_2(x, y) - \log \int \exp(f(x, y)) dP_1(x, y), \quad (281)$$

where the supremum is over the space of all measurable functions  $f$  with respect to  $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X} \times \mathcal{Y}})$  and  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , such that  $\int \exp(f(x, y)) dP_1(x, y) < \infty$ . Hence, for all  $z \in (\mathcal{X} \times \mathcal{Y})^n$  and for all  $\theta \in \mathcal{M}$ ; and for all  $t \in (0, +\infty)$ , it

follows that the function  $\ell$  in (4) satisfies that

$$\begin{aligned} & D(P_2 \| P_1) \\ & \geq \int t\ell(\boldsymbol{\theta}, x, y) dP_2(x, y) - \log \int \exp(t\ell(\boldsymbol{\theta}, x, y)) dP_1(x, y) \end{aligned} \quad (282)$$

$$\begin{aligned} & = \int t\ell(\boldsymbol{\theta}, x, y) dP_2(x, y) - \log \int \exp\left(t\ell(\boldsymbol{\theta}, x, y) + t \int \ell(\boldsymbol{\theta}, u, v) dP_1(u, v) \right. \\ & \quad \left. - t \int \ell(\boldsymbol{\theta}, u, v) dP_1(u, v)\right) dP_1(x, y) \end{aligned} \quad (283)$$

$$\begin{aligned} & = \int t\ell(\boldsymbol{\theta}, x, y) dP_2(x, y) - t \int \ell(\boldsymbol{\theta}, u, v) dP_1(u, v) - \log \int \exp\left(t\ell(\boldsymbol{\theta}, x, y) \right. \\ & \quad \left. - t \int \ell(\boldsymbol{\theta}, u, v) dP_1(u, v)\right) dP_1(x, y) \end{aligned} \quad (284)$$

$$= tG(\boldsymbol{\theta}, P_2, P_1) - \log \int \exp\left(t\ell(\boldsymbol{\theta}, x, y) - t \int \ell(\boldsymbol{\theta}, u, v) dP_1(u, v)\right) dP_1(x, y), \quad (285)$$

which leads to

$$\begin{aligned} & G(\boldsymbol{\theta}, P_2, P_1) \\ & \leq \frac{D(P_2 \| P_1)}{t} + \frac{1}{t} \log \left( \int \exp\left(t \left( \ell(\boldsymbol{\theta}, x, y) - \int \ell(\boldsymbol{\theta}, u, v) dP_1(u, v) \right)\right) dP_1(x, y) \right). \end{aligned} \quad (286)$$

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

$$\begin{aligned} & G(\boldsymbol{\theta}, P_2, P_1) \\ & \leq \inf_{t>0} \left( \frac{D(P_2 \| P_1)}{t} + \frac{1}{t} \log \left( \int \exp\left(t \left( \ell(\boldsymbol{\theta}, x, y) \right. \right. \right. \right. \\ & \quad \left. \left. \left. - \int \ell(\boldsymbol{\theta}, x, y) dP_1(x, y) \right)\right) dP_1(x, y) \right) \right), \end{aligned} \quad (287)$$

which completes the proof of the lemma.  $\blacksquare$

The proof of Theorem M continues from here. From Lemma 10, it holds that the probability measure  $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}$  in (21), satisfies for all  $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ ,

$$\begin{aligned} & G\left(\boldsymbol{\theta}, P, P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}\right) \\ & \leq \inf_{t>0} \left( \frac{D\left(P \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}\right)}{t} + \frac{1}{t} \log \left( \int \exp\left(t \left( \ell(\boldsymbol{\theta}, x, y) \right. \right. \right. \right. \\ & \quad \left. \left. \left. - J_{P_S, \boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)\right)\right) dP_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}(x, y) \right) \right), \end{aligned} \quad (288)$$

where  $J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right)$  is in (35). Moreover, for all  $t \in (0, +\infty)$ ,

$$\begin{aligned} & \log \left( \int \exp \left( t \left( \ell(\theta, x, y) - J_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) \right) \right) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) \right) \\ &= \log \left( \int \exp(t\ell(\theta, x, y)) dP_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}(x, y) \right) - tJ_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) \end{aligned} \quad (289)$$

$$= J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}(t) - tJ_{P_S, \theta}^{(1)}\left(\frac{1}{\beta}\right) \quad (290)$$

$$\leq \frac{1}{2}t^2 \hat{\zeta}_{P_S, \theta}^2, \quad (291)$$

where the function  $J_{P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}, \theta}$  in (290) is defined in (19); the inequality in (291) follows from Theorem 3; and

$$\hat{\zeta}_{P_S, \theta}^2 \triangleq \sup \left\{ \sqrt{J_{P_S, \theta}^{(2)}(\xi)} : \xi \in \left(0, b - \frac{1}{\beta}\right) \right\}, \quad (292)$$

with

$$b \triangleq \sup \{ \alpha \in (0, +\infty) : J_{P_S, \theta}(\alpha) < +\infty \}. \quad (293)$$

Plugging (291) into (288) yields

$$G\left(\theta, P, P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right) \leq \inf_{t>0} \left( \frac{D\left(P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right)}{t} + \frac{1}{2}t^2 \hat{\zeta}_{P_S, \theta}^2 \right). \quad (294)$$

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

$$c \triangleq G\left(\theta, P, P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right). \quad (295)$$

Hence, from (294), it follows that for all  $t \in (0, +\infty)$ ,

$$ct - \frac{1}{2}t^2 \hat{\zeta}_{P_S, \theta}^2 \leq D\left(P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)}\right).$$

The rest of the proof consists in finding an explicit expression for the absolute value of  $c$  in (295). To this aim, consider the function  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  such that

$$\phi(\alpha) = \frac{1}{2}\alpha^2 \hat{\zeta}_{P_S, \theta}^2. \quad (296)$$

Let the Legendre-Fenchel transform of  $\phi$  be the function  $\phi^* : \mathbb{R} \rightarrow \mathbb{R}$ , and thus for all  $x \in \mathbb{R}$ ,

$$\phi^*(x) = \max_{t \in (0, +\infty)} xt - \phi(t). \quad (297)$$

In particular, note that

$$\phi^*(c) \leq D \left( P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right). \quad (298)$$

Note that for all  $x \in \mathbb{R}$  and for all  $t \in (0, +\infty)$ , the function  $\phi^*$  in (297) satisfies

$$xt - \frac{1}{2}t^2\hat{\zeta}_{P_S, \theta}^2 \leq \phi^*(x) = x\alpha^*(x) - \phi(\alpha^*(x)), \quad (299)$$

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

$$\frac{d}{d\alpha} (x\alpha - \phi(\alpha)) = x - \alpha\zeta^2 = 0. \quad (300)$$

That is

$$\alpha^*(x) = \frac{x}{\hat{\zeta}_{P_S, \theta}^2}. \quad (301)$$

Plugging (301) into (299) gives

$$\phi^*(x) = \frac{x^2}{2\hat{\zeta}_{P_S, \theta}^2}. \quad (302)$$

Hence, from (298) and (299), given  $c$  in (295) for all  $t \in (0, +\infty)$ ,

$$ct - \frac{1}{2}t^2\hat{\zeta}_{P_S, \theta}^2 \leq \phi^*(c) \leq D \left( P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right), \quad (303)$$

and thus,

$$\frac{c^2}{2\hat{\zeta}_{P_S, \theta}^2} \leq D \left( P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right). \quad (304)$$

This implies that

$$c \leq \sqrt{2\hat{\zeta}_{P_S, \theta}^2 D \left( P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right)}, \text{ and} \quad (305)$$

$$c \geq -\sqrt{2\hat{\zeta}_{P_S, \theta}^2 D \left( P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right)}, \quad (306)$$

which leads to

$$\left| G \left( \theta, P, P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right) \right| \leq \sqrt{2\hat{\zeta}_{P_S, \theta}^2 D \left( P \| P_{\hat{Z}|\Theta=\theta}^{(P_S, \beta)} \right)}. \quad (307)$$

This completes the proof.

## N Proof of Theorem 8

The proof follows from the following equalities:

$$G(\boldsymbol{\theta}, P_1, P_2) = G(\boldsymbol{\theta}, P_1, P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}) - G(\boldsymbol{\theta}, P_2, P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}) \quad (308)$$

$$\begin{aligned} &= \beta \left( D(P_1 \| P_S) - D\left(P_1 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}\right) - D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \| P_S\right) \right) \\ &\quad - \beta \left( D(P_2 \| P_S) - D\left(P_2 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}\right) - D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)} \| P_S\right) \right) \end{aligned} \quad (309)$$

$$= \beta \left( D\left(P_2 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}\right) - D\left(P_1 \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S, \beta)}\right) \right) \quad (310)$$

$$- D(P_2 \| P_S) + D(P_1 \| P_S), \quad (311)$$

where equality in (309) follows from (56). This completes the proof.

## O Proof for Lemma 7

The proof follows from Definition 3 and the following equalities:

$$\mathsf{L}(\mathbf{z}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^n \ell(\boldsymbol{\theta}, x_t, y_t) \quad (312)$$

$$= \frac{1}{n} \sum_{(x, y) \in (\mathcal{X} \times \mathcal{Y})} \ell(\boldsymbol{\theta}, x, y) \sum_{t=1}^n \mathbb{1}_{\{x=x_t, y=y_t\}} \quad (313)$$

$$= \sum_{(x, y) \in (\mathcal{X} \times \mathcal{Y})} \ell(\boldsymbol{\theta}, x, y) P_{\mathbf{z}}(x, y) \quad (314)$$

$$= \int \ell(\boldsymbol{\theta}, x, y) dP_{\mathbf{z}}(x, y) \quad (315)$$

$$= \mathsf{R}_{\boldsymbol{\theta}}(P_{\mathbf{z}}), \quad (316)$$

where the functional  $\mathsf{R}_{\boldsymbol{\theta}}$  is defined in (8). This completes the proof.

## P Proof of Theorem 10

*Proof:* The doubly-expected generalization gap  $\overline{\overline{G}}(P_{\Theta|Z}, P_Z)$  in (72) satisfies:

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) = \int \int G(\boldsymbol{\theta}, P_Z, P_Z) dP_{\Theta|Z=z}(\boldsymbol{\theta}) dP_Z(z), \quad (317)$$

$$= \int \int G(\boldsymbol{\theta}, P_Z, P_Z) dP_{Z|\Theta=\boldsymbol{\theta}}(z) dP_{\Theta}(\boldsymbol{\theta}), \quad (318)$$

$$= \int \int \left( \int \ell(\boldsymbol{\theta}, x, y) dP_Z(x, y) \right) dP_{Z|\Theta=\boldsymbol{\theta}}(z) dP_{\Theta}(\boldsymbol{\theta}) \\ - \int \int \left( \int \ell(\boldsymbol{\theta}, x, y) dP_Z(x, y) \right) dP_{Z|\Theta=\boldsymbol{\theta}}(z) dP_{\Theta}(\boldsymbol{\theta}) \quad (319)$$

$$= \int \int \ell(\boldsymbol{\theta}, x, y) dP_Z(x, y) dP_{\Theta}(\boldsymbol{\theta}), \\ - \int \int \left( \frac{1}{n} \sum_{t=1}^n \ell(\boldsymbol{\theta}, x_t, y_t) \right) dP_{Z|\Theta=\boldsymbol{\theta}}(z) dP_{\Theta}(\boldsymbol{\theta}) \quad (320)$$

$$= \int \int \ell(\boldsymbol{\theta}, x, y) dP_Z(x, y) dP_{\Theta}(\boldsymbol{\theta}) \\ - \int \int \ell(\boldsymbol{\theta}, x, y) dP_{Z|\Theta=\boldsymbol{\theta}}(z) dP_{\Theta}(\boldsymbol{\theta}) \quad (321)$$

$$= \int G(\boldsymbol{\theta}, P_Z, P_{Z|\Theta=\boldsymbol{\theta}}) dP_{\Theta}(\boldsymbol{\theta}), \quad (322)$$

where the equality in (318) follows from (77) and (78); the equality in (319) follows from (66); the equality in (320) follows from (77) and (26); the equality in (321) follows from the fact that the probability measure  $P_{Z|\Theta=\boldsymbol{\theta}}$  is a product measure as in (79); and the equality in (322) follows from (55).

Finally, from Corollary 2 (Equation (61)), it follows that

$$\frac{1}{\beta} \overline{\overline{G}}(P_{\Theta|Z}, P_Z) = \int D\left(P_{Z|\Theta=\boldsymbol{\theta}} \| P_{\hat{Z}|\Theta=\boldsymbol{\theta}}^{(P_{Z|\Theta=\boldsymbol{\theta}}, \beta)}\right) dP_{\Theta}(\boldsymbol{\theta}) \\ - \int D\left(P_Z \| P_{\hat{Z}|\Theta=\boldsymbol{\theta}}^{(P_{Z|\Theta=\boldsymbol{\theta}}, \beta)}\right) dP_{\Theta}(\boldsymbol{\theta}) \\ + \int D(P_Z \| P_{Z|\Theta=\boldsymbol{\theta}}) dP_{\Theta}(\boldsymbol{\theta}), \quad (323)$$

where  $\int D(P_Z \| P_{Z|\Theta=\boldsymbol{\theta}}) dP_{\Theta}(\boldsymbol{\theta}) = \frac{1}{n} L(P_{\Theta|Z}; P_Z)$ , which follows from (81). This completes the proof.  $\blacksquare$

## Q Proof of Theorem 11

*Proof:* The term  $\int \left( D \left( P_{Z|\Theta=\theta} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) - D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) \right) dP_{\Theta}(\theta)$  in (82) satisfies

$$\begin{aligned} & \int \left( D \left( P_{Z|\Theta=\theta} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) - D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) \right) dP_{\Theta}(\theta) \\ & \leq \int D \left( P_{Z|\Theta=\theta} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) dP_{\Theta}(\theta), \end{aligned} \quad (324)$$

with equality if for all  $\theta \in \mathcal{M}$ , it holds that  $D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) = 0$ . That is, if for all  $\theta \in \mathcal{M}$ , the probability measures  $P_Z$  and  $P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}$  are identical. Under such a condition, it holds from (80) that  $\int D \left( P_{Z|\Theta=\theta} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) dP_{\Theta}(\theta) = \frac{1}{n} I(P_{\Theta|Z}; P_Z)$ , where the term  $I(P_{\Theta|Z}; P_Z)$  is the mutual information in (73). Hence, if for all  $\theta \in \mathcal{M}$ , it holds that  $D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) = 0$ , then, it holds that:

$$\overline{\overline{G}}(P_{\Theta|Z}, P_Z) \leq \frac{\beta}{n} (I(P_{\Theta|Z}; P_Z) + L(P_{\Theta|Z}; P_Z)). \quad (325)$$

The rest of the proof consists in showing that if (325) holds with equality, then, for all  $\theta \in \mathcal{M}$ , it holds that  $D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) = 0$ . This second part of the proof is as follows:

$$\begin{aligned} & \overline{\overline{G}}(P_{\Theta|Z}, P_Z) \\ & = \frac{\beta}{n} (I(P_{\Theta|Z}; P_Z) + L(P_{\Theta|Z}; P_Z)) \end{aligned} \quad (326)$$

$$= \beta (I(P_{\Theta|Z}; P_Z) + L(P_{\Theta|Z}; P_Z)) \quad (327)$$

$$\begin{aligned} & = \beta \left( \int D \left( P_{Z|\Theta=\theta} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) dP_{\Theta}(\theta) - \int D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) dP_{\Theta}(\theta) \right. \\ & \quad \left. + \int D(P_Z \| P_{Z|\Theta=\theta}) dP_{\Theta}(\theta) \right), \end{aligned} \quad (328)$$

where the equality in (326) is the initial assumption; the equality in (327) follows from (80) and (81); and the equality in (328) follows from Corollary 2. Finally, from (327) and (328), it follows that

$$\begin{aligned} & \int D \left( P_{Z|\Theta=\theta} \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) dP_{\Theta}(\theta) - \int D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) dP_{\Theta}(\theta) \\ & = \int D(P_{Z|\Theta=\theta} \| P_Z) dP_{\Theta}(\theta), \end{aligned} \quad (329)$$

which leads to

$$\int \int \log \left( \frac{dP_Z}{dP_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}}(x, y)} \right) dP_{Z|\Theta=\theta}(x, y) dP_{\Theta}(\theta) \quad (330)$$

$$= \int \int \log \left( \frac{dP_Z}{dP_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}}(x, y)} \right) dP_Z(x, y) dP_{\Theta}(\theta). \quad (331)$$

The equality (331) implies that for all  $\theta \in \text{supp } P_{\Theta}$  and for all  $(x, y) \in \text{supp } P_Z$ , either

$$\frac{P_{Z|\Theta=\theta}}{P_Z}(x, y)=1; \text{ or} \quad (332)$$

$$\frac{dP_Z}{dP_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}}(x, y)=1. \quad (333)$$

The equality in (332) implies that the algorithm  $P_{\Theta|Z}$  leads to the case in which models and datasets are independent. That is, for all  $\theta \in \mathcal{M}$ , it follows that the measures  $P_{Z|\Theta=\theta}$  and  $P_Z$  are identical. Alternatively, the equality in (333) implies that for all  $\theta \in \text{supp } P_{\Theta}$ , the WCDG probability measure  $P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)}$  is identical to the ground-truth data-generating probability measure  $P_Z$ . That is, for all  $\theta \in \mathcal{M}$ ,  $D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) = 0$ . The former case is disregarded given that statistical machine learning algorithms are probability measures on  $\Delta(\mathcal{M})$  dependent on the training datasets. Hence, it follows that if (325) holds with equality, then for all  $\theta \in \mathcal{M}$ ,  $D \left( P_Z \| P_{\hat{Z}|\Theta=\theta}^{(P_{Z|\Theta=\theta}, \beta)} \right) = 0$ . This completes the proof. ■



## References

- [1] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2014.
- [2] I. Csiszár, “The method of types,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.
- [3] I. N. Sanov, “On the probability of large deviations of random variables,” *Selected Translations in Mathematical Statistics and Probability*, vol. 1, pp. 213–244, 1961.
- [4] H. Rahimian and S. Mehrotra, “Frameworks and results in distributionally robust optimization,” *Open Journal of Mathematical Optimization*, vol. 3, pp. 1–85, 2022.
- [5] E. Delage and Y. Ye, “Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems,” *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.
- [6] Z. Hu and L. J. Hong, “Kullback-Leibler divergence constrained distributionally robust optimization,” *Optimization Online*, vol. 1, no. 2, p. 9, 2013.
- [7] J. Lee and M. Raginsky, “Minimax Statistical Learning with Wasserstein distances,” *Advances in Neural Information Processing Systems*, vol. 31, pp. 2687–2696, 2018.
- [8] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, “Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning,” in *Operations research & management science in the age of analytics*. Informs, 2019, pp. 130–166.
- [9] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, “Empirical risk minimization with  $f$ -divergence regularization in statistical learning,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9521, Oct. 2023.
- [10] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, “Analysis of the relative entropy asymmetry in the regularization of empirical risk minimization,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.
- [11] E. T. Jaynes, “Information theory and statistical mechanics I,” *Physical Review Journals*, vol. 106, no. 4, pp. 620–630, May 1957.
- [12] —, “Information theory and statistical mechanics II,” *Physical Review Journals*, vol. 108, no. 2, pp. 171–190, Oct. 1957.
- [13] S. Mazuelas, Y. Shen, and A. Pérez, “Generalized maximum entropy for supervised classification,” *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2530–2550, Jan. 2022.

- 
- [14] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*, 1st ed. New York, NY, USA: Wiley, 1989.
- [15] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, “An exact characterization of the generalization error for the Gibbs algorithm,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 8106–8118, Dec. 2021.
- [16] G. Aminian, Y. Bu, G. W. Wornell, and M. R. Rodrigues, “Tighter expected generalization error bounds via convexity of information measures,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Aalto, Finland, Jun. 2022, pp. 2481–2486.
- [17] A. Xu and M. Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 1–10, Dec. 2017.
- [18] Y. Chu and M. Raginsky, “A unified framework for information-theoretic generalization bounds,” arXiv preprint arXiv:2305.11042, May 2023.
- [19] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, “Empirical risk minimization with relative entropy regularization: Optimality and sensitivity,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, Jul. 2022, pp. 684–689.
- [20] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, “On the validation of Gibbs algorithms: Training datasets, test datasets and their aggregation,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.
- [21] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, “Empirical risk minimization with relative entropy regularization,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9454, Feb. 2022.
- [22] D. Russo and J. Zou, “Controlling bias in adaptive data analysis using information theory,” in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, vol. 51, Cadiz, Spain, May 2016, pp. 1232–1240.
- [23] A. Asadi, E. Abbe, and S. Verdú, “Chaining mutual information and tightening generalization bounds,” *Advances in Neural Information Processing Systems*, vol. 31, pp. 7245–7254, Dec. 2018.
- [24] A. R. Asadi and E. Abbe, “Chaining meets chain rule: Multilevel entropic regularization and training of neural networks.” *J. Mach. Learn. Res.*, vol. 21, pp. 139–1, Jun. 2020.
- [25] L. P. Barnes, A. Dytso, and H. V. Poor, “Improved information-theoretic generalization bounds for distributed, federated, and iterative learning,” *Entropy*, vol. 24, no. 9, 2022.

- [26] Y. Bu, S. Zou, and V. V. Veeravalli, “Tightening mutual information-based bounds on generalization error,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, Jan. 2020.
- [27] F. Hellström and G. Durisi, “Generalization bounds via information density and conditional information density,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 824–839, Nov. 2020.
- [28] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, “Conditioning and processing: Techniques to improve information-theoretic generalization bounds,” *Advances in Neural Information Processing Systems*, pp. 16 457–16 467, Dec. 2020.
- [29] A. T. Lopez and V. Jog, “Generalization error bounds using wasserstein distances,” in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Guangzhou, China, Nov. 2018, pp. 1–5.
- [30] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, “An information-theoretic view of generalization via Wasserstein distance,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 577–581.
- [31] I. Issa, A. R. Esposito, and M. Gastpar, “Strengthened information-theoretic bounds on the generalization error,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Paris, France, Jul. 2019, pp. 582–586.
- [32] A. R. Esposito, M. Gastpar, and I. Issa, “Robust generalization via  $\alpha$ -mutual information,” arXiv preprint arXiv:2001.06399, Jan. 2020.
- [33] S. Masiha, A. Gohari, and M. H. Yassaee, “f-divergences and their applications in lossy compression and bounding generalization error,” *IEEE Transactions on Information Theory*, pp. 7245–7254, Apr. 2023.
- [34] G. Aminian, L. Toni, and M. R. Rodrigues, “Jensen-Shannon information based characterization of the generalization error of learning algorithms,” in *Proceedings of the IEEE Information Theory Workshop (ITW)*, Kanazawa, Japan, Oct. 2021, pp. 1–5.
- [35] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, “Fantastic generalization measures and where to find them,” in *Proceedings of the International Conference on Learning Representations*, Addis Ababa, Ethiopia, Apr. 2020.
- [36] M. Gastpar, I. Nachum, J. Shafer, and T. Weinberger, “Fantastic generalization measures are nowhere to be found,” *arXiv preprint arXiv:2309.13658*, 2023.
- [37] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik, “Model complexity control for regression using VC generalization bounds,” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1075–1089, Sep. 1999.

- [38] D. A. McAllester, “PAC-Bayesian stochastic model selection,” *Machine Learning*, vol. 51, no. 1, pp. 5–21, Apr. 2003.
- [39] D. Cullina, A. N. Bhagoji, and P. Mittal, “PAC-learning in the presence of adversaries,” *Advances in Neural Information Processing Systems*, vol. 31, no. 1, pp. 1–12, Dec. 2018.
- [40] M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor, “PAC-Bayes unleashed: Generalisation bounds with unbounded losses,” *Entropy*, vol. 23, no. 10, pp. 1–20, Oct. 2021.
- [41] D. Russo and J. Zou, “How much does your data exploration overfit? Controlling bias via information usage,” *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, Jan. 2019.
- [42] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, “Sensitivity of the Gibbs algorithm to data aggregation in supervised machine learning,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9474, Jun. 2022.
- [43] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, “Empirical risk minimization with relative entropy regularization type-II,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9508, May. 2023.
- [44] I. Csiszár, “Information-type measures of difference of probability distributions and indirect observation,” *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, no. 1, pp. 299–318, Jun. 1967.
- [45] H. Jeffreys, “An invariant form for the prior probability in estimation problems,” *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, Sep. 1946.
- [46] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [47] C. Villani, *Optimal transport: Old and new*, 1st ed. Berlin, Germany: Springer, 2009.
- [48] I. Sason and S. Verdú, “ $f$ -divergence inequalities,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, Jun. 2016.
- [49] H.-O. Georgii, *Gibbs measures and phase transitions*, 2nd ed. New York, NY, USA: De Gruyter, 2011.
- [50] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.
- [51] W. Rudin, *Principles of mathematical analysis*, 1st ed. New York, NY, USA: McGraw-Hill Book Company, Inc., 1953.
- [52] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, 1st ed. New York, NY, USA: Cambridge University Press, 2019.

- 
- [53] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, “The marginal value of adaptive gradient methods in machine learning,” *Advances in neural information processing systems*, vol. 30, pp. 4151–4161, Dec. 2017.
- [54] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, Jul. 1948.
- [55] D. P. Palomar and S. Verdú, “Lautum information,” *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.
- [56] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
- [57] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, “Generalization analysis of machine learning algorithms via the worst-case data-generating probability measure,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2024.
- [58] K. Sun, S. M. Perlaza, and A. Jean-Marie, “ $2 \times 2$  Zero-sum games with commitments and Noisy Observations,” in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.
- [59] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA, USA: Harcourt Academic Press, 2000.
- [60] W. Feller, *An Introduction to Probability Theory and Its Applications II*, 2nd ed. New York, NY, USA: Jhon Wiley & Sons, 1971.
- [61] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, “The Worst-Case Data-Generating Probability Measure,” INRIA, Centre Inria d’Université Côte d’Azur, Sophia Antipolis, France, Tech. Rep. RR-9515, Aug. 2023.
- [62] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: Wiley, 1997.
- [63] W. F. Trench, *Introduction to Real Analysis*, 1st ed. Hoboken, NJ, USA: Prentice Hall Pearson Education, 2003.
- [64] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Non-asymptotic Theory of Independence*, 1st ed. Oxford, UK: Oxford University Press, 2013.



**RESEARCH CENTRE  
SOPHIA ANTIPOLIS – MÉDITERRANÉE**

2004 route des Lucioles - BP 93  
06902 Sophia Antipolis Cedex

Publisher  
Inria  
Domaine de Voluceau -  
Rocquencourt  
BP 105 - 78153 Le Chesnay  
Cedex  
[inria.fr](http://inria.fr)