The Worst-Case Data-Generating Probability Measure

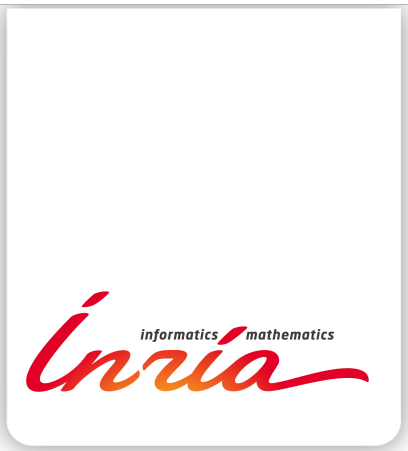
Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman

The Worst-Case Data-Generating Probability Measure

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

RESEARCH REPORT
N° 9515
August 2023
Project-Team NEO

ISSN 0249-6399 ISRN INRIA/RR--9515--FR+ENG



The Worst-Case Data-Generating Probability Measure

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

Project-Team NEO

Research Report n° 9515 — version 2 — initial version August 2023 — revised version January 2024 — 42 pages

Abstract: In this paper, the worst-case probability measure over the data is introduced as a tool for characterizing the generalization capabilities of machine learning algorithms. More specifically, the worst-case probability measure is a Gibbs probability measure and the unique solution to the maximization of the expected loss under a relative entropy constraint with respect to a reference probability measure. Fundamental generalization metrics, such as the sensitivity of the expected loss, the sensitivity of the empirical risk, and the generalization gap are shown to have closed-form expressions involving the worst-case data-generating probability measure. Existing results for the Gibbs algorithm, such as characterizing the generalization gap as a sum of mutual information and lautum information, up to a constant factor, are recovered. A novel parallel is established between the worst-case data-generating probability measure and the Gibbs algorithm. Specifically, the Gibbs probability measure is identified as a fundamental commonality of the model space and the data space for machine learning algorithms.

Key-words: Supervised Machine Learning, Worst-Case, Generalization Gap, Relative Entropy, Gibbs Algorithm, and Sensitivity.

Xinying Zou is with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France. Samir M. Perlaza is with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France; with the ECE Dept. at Princeton University, Princeton NJ 08544, USA; and also with the GAATI Laboratory at the Université de la Polynésie Française, Faaa 98702, French Polynesia. Iñaki Esnaola is with the ACSE Dept. at The University of Sheffield, Sheffield S1 3JD, UK; and also with the ECE Dept. at Princeton University, Princeton NJ 08544, USA. Eitan Altman is with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France; and also with the LIA, Université d'Avignon, France. This work has been presented at the AAAI Conference on Artificial Intelligence [1].

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

La mesure de probabilité génératrice de données dans le pire des cas

Résumé : Dans ce rapport, la mesure de probabilité du pire cas sur les données est présentée comme un outil pour caractériser les capacités de généralisation des algorithmes d'apprentissage automatique. Plus précisément, la mesure de probabilité du pire cas est une solution à la maximisation de la valeur espérée de la perte (ou risque) induite par un modèle sous une contrainte d'entropie relative par rapport à une mesure de probabilité de référence. Étant donné un modèle, le résultat central consiste en une expression explicite de la différence entre les valeurs espérées de la perte par rapport à deux mesures de probabilité quelconques sur l'ensemble de données. Cette différence est caractérisée en termes de "distances statistiques" mesurées via des divergences KL impliquant les mesures données ; la mesure de référence ; et la mesure de probabilité du pire cas. Lorsque les mesures données sont les types (mesures de probabilité empiriques) induits par deux ensembles de données, une expression sous forme fermée pour la différence entre les risques empiriques correspondants est obtenue. Enfin, l'écart de généralisation induit par un algorithme d'apprentissage quelconque est caractérisé. Les résultats existants pour l'algorithme de Gibbs, tels que l'égalité entre l'écart de généralisation et une somme d'informations mutuelles et d'informations de lautum, à un facteur constant près, sont récupérés. Tout ce qui précède suggère une dualité entre l'algorithme de Gibbs et la mesure du pire cas au-delà du fait que les deux sont représentés par des mesures de probabilité de Gibbs.

Mots-clés : Apprentissage automatique supervisé, pire cas, généralisation, entropie relative, algorithme de Gibbs, et sensibilité.

Contents

1 Introduction

The expected generalization error (GE) is a central workhorse for the analysis of generalization capabilities of machine learning algorithms, see for instance [2–5] and [6]. In a nutshell, the GE characterizes the ability of the learning algorithm to correctly find patterns in datasets that are not available during the training stage. Specifically, it is defined for a fixed training dataset and a specific model instance, as the difference between the population risk induced by the model and the empirical risk with respect to the training dataset.

When the choice of models is governed by a stochastic kernel, the expected GE (EGE) is the expectation of the GE with respect to the joint-measure of the models and the datasets. Closed-form expressions for the EGE are only known for the Gibbs algorithm in the case in which the reference measure is a probability measure [2]; and for the case in which the reference measure is a σ-finite measure [7].

1.1 Related Works

In general, the EGE of machine learning algorithms is characterized by various upper bounds leveraging different techniques. The metric of mutual information was first proposed in [8], further developed in [4] and combined with chaining methods in [9, 10] for deriving upper bounds on the EGE. Similar bounds on the EGE were obtained in [5, 11–13] and references therein. Other information measures such as the Wasserstein distance [3, 14, 15], maximal leakage [16, 17], mutual f-information [18], and Jensen-Shannon divergence [19] were used for providing upper bounds on EGE as well. In [20], the notion of *closeness* of probability measures with respect to a reference measure in terms of statistical distances was used. Therein, the authors explored the case for which the reference is the empirical measure, which is also studied in this work. Such statistical distance was formulated through f-divergences in [20], whereas in this work, the statistical distance is described in terms of relative entropy. However, the objective entailed minimizing the expected loss, while this work provides explicit expressions for the difference between empirical risks, population risk, and generalization gap. For the use of f-divergences in these optimization problems, see also [21], and references therein.

Generalization can also be studied as a local minmax problem as in [22], in which generalization bounds were given in terms of empirical risks induced by a worst-case probability measure. The set of candidate probability measures in this work was described in terms of the Wasserstein ambiguity set containing the empirical measure and the ground-truth measure almost surely. The minimax formulation was further studied by establishing a correspondence between the principle of maximum entropy and the minimax approach for decision making in [23]. To circumvent the dependence on the statistical description of the dataset, generalization analyses often rely on approaches that decouple the explicit link of the data-generating measure with the GE by using tools from combinatorics [24];

probability theory [25–27]; and information theory [2, 4, 28]. These approaches tend to distill the insight about the GE into coarse statistical descriptions of the dataset-generating measures or features of the hypothesis class that the algorithm aims to learn.

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 learning capability of the algorithm. This, in turn, results in descriptions of the EGE for which the dependence on the training dataset and the selected model is not made evident. Recent efforts for highlighting the dependence of generalization capabilities on the training dataset have led to explicit expressions for the expectation of the GE when the models are sampled using the Gibbs algorithm in [6,29]. This line of work opens the door to the study of the worst-case data-generating probability measures and their effect on the GE and EGE, as shown in the following section.

1.2 Contributions

The first contribution consists of a probability measure over the datasets coined *the worst-case data-generating* probability measure. Such a measure maximizes the expectation of the loss, while satisfying that its "*statistical distance*" to a given probability measure does not exceed a given threshold. In the following, such a "statistical distance" is measured via the KL-divergence, also known as relative entropy. Interestingly, this choice of "statistical distance" leads to the fact that, if the worst-case probability measure exists, then it is a Gibbs probability measure (Theorem 1) parametrized by the reference measure; the "statistical distance" threshold; and the loss function. The variation of the expectation of the loss when the probability measure changes from the worst-case probability measure to an alternative measure is characterized in terms of "statistical distances", also represented by relative entropies. 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 3). This is an important result as the reference measure and the "statistical distance" threshold can be arbitrarily chosen, which leads to useful 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 [30], which are also known as empirical probability measures. 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 4). This observation, in conjuction with Theorem 3 provides 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 to variations on the dataset. Using the same arguments, closed-form expressions in terms of "statistical distances" are provided for the generalization

gap induced by a given model obtained from a given training dataset.

The final contribution consists of showing that the expected generalization gap and the doubly-expected generalization gap are strongly connected with the notion of worst-case data-generating probability measure. As a byproduct, an alternative proof to the existing result (see [2] and [7]) providing a closed-form expression for the doubly-expected generalization gap of the Gibbs algorithm in terms of mutual and lautum information is presented. Despite the limitation that this alternative proof relies on the assumption of independent and identically distributed data points, its relevance is significant as it highlights an intriguing connection between the Gibbs algorithm and the worst-case data-generating probability measure.

1.3 Notation

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

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

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

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), ..., (x_n, y_n), a dataset is represented by the tuple

$$\boldsymbol{z} = \big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big) \in (\mathcal{X} \times \mathcal{Y})^n. \tag{2}$$

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

$$y = f(\boldsymbol{\theta}, x). \tag{3}$$

Let also the function

$$\hat{\ell} : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty] \tag{4}$$

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 $\hat{\ell}\left(f(\boldsymbol{\theta}, x), y\right)$. In the following, the loss function $\hat{\ell}$ is assumed to be nonnegative and for all $y \in \mathcal{Y}$, it holds that $\hat{\ell}\left(y, y\right) = 0$.

For ease of notation, let the function $\ell : \mathcal{M} \times \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ be such that

$$\ell(\boldsymbol{\theta}, x, y) = \hat{\ell}\left(f(\boldsymbol{\theta}, x), y\right). \tag{5}$$

The *empirical risk* induced by the model $\boldsymbol{\theta} \in \mathcal{M}$, with respect to the dataset $\boldsymbol{z}$ in (2), is determined by the function $\mathsf{L} : (\mathcal{X} \times \mathcal{Y})^n \times \mathcal{M} \to [0, +\infty]$, which satisfies

$$\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(\boldsymbol{\theta}, x_i, y_i\right), \tag{6}$$

where the functions f and ℓ are defined in (3) and (5).

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

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}\left(\boldsymbol{z}, \boldsymbol{\theta}\right). \tag{7}$$

The ERM problem is prone to overfitting since the set of solutions to (7) are models selected specifically for the given dataset $\boldsymbol{z}$ in (2), 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 (7). Such a regularization term can be represented by a function $R : \mathcal{M} \to \mathbb{R}$, which yields the regularized ERM problem

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}\left(\boldsymbol{z}, \boldsymbol{\theta}\right) + \lambda R\left(\boldsymbol{\theta}\right), \tag{8}$$

where λ is a nonnegative real that acts as a regularization parameter. The regularization function R in (8) 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 [31]. One common choice for R is $R\left(\boldsymbol{\theta}\right) = \|\boldsymbol{\theta}\|_p$, with $p \geq 1$. The norm is often used to account for the model complexity. Alternatively, the regularization parameter λ 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 $\boldsymbol{\theta} \in \mathcal{M}$ independently from how such a model is chosen.

3 An Auxiliary Optimization Problem

This section introduces an optimization problem whose solution is referred to as the worst-case data-generating probability measure. This probability measure,

which is conditioned on a given model $\boldsymbol{\theta} \in \mathcal{M}$, is parametrized by a probability measure $P_S \in \triangle(\mathcal{X} \times \mathcal{Y})$ and by a positive real γ. In a nutshell, the worst-case data-generating probability measure maximizes the expected loss while its relative entropy with respect to P_S is not larger than γ. Using this notation, the optimization problem of interest is:

$$\max_{P \in \Delta_{P_S}(\mathcal{X}\times\mathcal{Y})} \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P(x, y) \tag{9a}$$

$$\text{s.t.} \quad D(P\|P_S) \leq \gamma \tag{9b}$$

$$\int \mathrm{d}P(x, y) = 1, \tag{9c}$$

where the functions f and ℓ are defined in (3) and (5).

The probability measure P_S in (9) can be interpreted as a prior on the probability distribution of the datasets. From this perspective, the search of the worst-case probability measure is performed on the set of all probability measures that are at most at a "statistical distance" smaller than or equal to γ from the measure P_S. Here, such a "statistical distance" is measured in terms of the relative entropy. The benefits of the choice of relative entropy become apparent when studying the properties of the solution to the optimization problem in (9). The impact of the asymmetry of the relative entropy on this problem is left out of the scope of this work. The interested reader is referred to [32].

3.1 The Solution

The following theorem characterizes the solution to the optimization problem in (9) using the function $\mathsf{J}_{P_S,\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$, which satisfies

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P_S(x, y)\right), \tag{10}$$

with the functions f and ℓ in (3) and (5), respectively.

Theorem 1 *The solution to the optimization problem in* (9)*, if it exists, is denoted by* $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ *and satisfies for all* $(x, y) \in \operatorname{supp} P_S$*,*

$$\frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x, y) = \exp\left(\frac{\ell(\boldsymbol{\theta}, x, y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right), \tag{11}$$

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ *is defined in* (10) *and* $\beta > 0$ *satisfies*

$$D\left(P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right) = \gamma. \tag{12}$$

Proof: The proof is presented in Appendix A. ∎

Theorem 1 provides a guarantee on the uniqueness of the solution to the optimization problem in (9), whenever it exists. Nonetheless, guarantees for the

existence of a solution to (9) are not provided. In the following, it is assumed that the model $\boldsymbol{\theta}$, the real γ, and the probability measure P_S in (9) are such that a solution exists. Let the set $\mathcal{J}_{P_S,\boldsymbol{\theta}} \subset (0,+\infty)$ be:

$$\mathcal{J}_{P_S,\boldsymbol{\theta}} \triangleq \left\{ t \in (0,+\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) < +\infty \right\}. \tag{13}$$

The existence of a solution to the problem in (9) is subject to the condition $\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) < +\infty$, which involves the model $\boldsymbol{\theta}$, the loss function ℓ in (5), and the parameters β and P_S. This condition is always satisfied in the case in which the function ℓ is bounded almost surely with respect to P_S, as shown by the following example.

Example 1 *Assume that for some model $\boldsymbol{\theta} \in \mathcal{M}$, there exists a real $a \in (0,+\infty)$ such that*

$$P_S\left(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta},x,y) \leqslant a\}\right) = 1, \tag{14}$$

where the function ℓ is defined in (5)*. Note that the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ satisfies for all $t \in \mathbb{R}$,*

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) \leqslant \log\left(\int \exp\left(\frac{a}{t}\right) \mathrm{d}P(x,y)\right) \tag{15}$$

$$= \frac{a}{t} + \log\left(\int \mathrm{d}P(x,y)\right) \tag{16}$$

$$= \frac{a}{t} < +\infty, \tag{17}$$

which implies that under the assumption in (14)*, the optimization problem in* (9) *always has a solution.*

In general, if a solution to (9) exists, the measure $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (11) is a Gibbs probability measure [33]. From this perspective, the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (11) is often referred to as the log-partition function [34]. Moreover, the probability measure P_S in (9) can be interpreted as a prior on the probability distribution of the datasets.

3.2 Mutual Absolute Continuity

When the optimization problem in (9) possesses a solution, i.e., $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$ with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (13), the loss $\ell(\boldsymbol{\theta},x,y)$, with $(x,y) \in \operatorname{supp} P_S$, is finite almost surely with respect to P_S.

Lemma 1 *If the problem in* (9) *has a solution, then*

$$P_S\left(\left\{(x,y) \in \operatorname{supp} P_S : \ell(\boldsymbol{\theta},x,y) = +\infty \right\}\right) = 0, \tag{18}$$

where the function ℓ is in (5)*.*

Proof: The proof is presented in Appendix B. ■

This observation plays a key role in the proof of the main properties of the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (11). Among such properties, an important one is the mutual absolute continuity between $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ and P_S, which is formalized by the following lemma.

Lemma 2 *The probability measures $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ and P_S in* (11) *are mutually absolutely continuous.*

Proof: The proof is presented in Appendix C. ■

An immediate consequence of the mutual absolute continuity between the measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (11) is described by the following lemma.

Lemma 3 *The probability measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (11) *satisfy:*

$$\beta \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) = \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) - \beta D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right) \tag{19}$$

$$= \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) + \beta D\left(P_S\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right), \tag{20}$$

where the function ℓ is defined in (5) *and the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is defined in* (10).

Proof: This proof is presented in Appendix D. ■

4 Analysis of the Expected Loss

Let the function $G : \mathcal{M} \times \Delta(\mathcal{X}\times\mathcal{Y}) \times \Delta(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R}$ be such that

$$G(\boldsymbol{\theta},P_1,P_2) = \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_1(x,y) - \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_2(x,y), \tag{21}$$

where the function ℓ is defined in (5). The value $G(\boldsymbol{\theta},P_1,P_2)$ represents the variation of the expectation of the loss when the probability measure over the data points changes from P_2 to P_1. Such a value is often referred to as the *sensitivity* of the expected loss to variations on the probability distribution of the data points. Such a sensitivity is characterized by the following theorem for the specific case of variations from the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (11) to an alternative measure.

Theorem 2 (Sensitivity of the Expected Loss) *For all $P \in \Delta_{P_S}(\mathcal{X}\times\mathcal{Y})$ and for all $\boldsymbol{\theta}\in\mathcal{M}$,*

$$G\left(\boldsymbol{\theta},P,P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) = \beta\left(D\left(P\|P_S\right) - D\left(P\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right)\right), \tag{22}$$

where the functional G is defined in (21)*; and the model $\boldsymbol{\theta}$ and the measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ satisfy* (11).

Proof: The proof is presented in Appendix E. ■

The following corollary of Theorem 2 describes the sensitivity of the expected loss for variations from $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ to the reference measure P_S.

Corollary 1 *The probability measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (11) *satisfy:*

$$G(\boldsymbol{\theta}, P_S, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}) = -\beta \left(D\left(P_S \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) + D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)\right), \qquad (23)$$

where the functional G is in (21).

The right-hand side of the equality in (23) is a symmetrized Kullback-Liebler divergence, also known as Jeffrey's divergence [35], between the measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$.

More importantly, it holds that $D\left(P_S \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) \geqslant 0$ and $D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) \geqslant 0$, which reveals the fact that the expected loss induced by the Gibbs probability measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ is larger than or equal to the expected loss induced by the reference measure P_S. This is formalized by the following corollary of Theorem 2.

Corollary 2 *The probability measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (11) *satisfy:*

$$\int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x, y) \geq \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_S(x, y), \qquad (24)$$

where the function ℓ is defined in (5).

Note that the probability measure P_S in Corollary 2 can be arbitrarily chosen. That is, independent of the model $\boldsymbol{\theta}$. From this perspective, the measure P_S can be interpreted as a prior on the datasets, while the probability measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ can be interpreted as a posterior for the worst-case once the prior P_S is confronted with the model $\boldsymbol{\theta}$.

Equipped with the exact characterization of the sensitivity from the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ to any alternative measure P provided by Theorem 2, it is possible to obtain the sensitivity of the expected loss when the measure changes from a given probability measure to any alternative probability measure, as shown by the following theorem.

Theorem 3 *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 $\boldsymbol{\theta} \in \mathcal{M}$, the functional G in* (21) *satisfies*

$$G(\boldsymbol{\theta}, P_1, P_2) = \beta\Big(D\Big(P_2 \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\Big) - D\Big(P_1 \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\Big) - D(P_2 \| P_S) + D(P_1 \| P_S)\Big), \qquad (25)$$

where the model $\boldsymbol{\theta}$ and the measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ satisfy (11).

Proof: The proof is presented in Appendix F. ■

Note that the parameters γ and P_S in (9) can be arbitrarily chosen. This is essentially because only the right-hand side of (25) depends on P_S and β. Another interesting observation is that none of the terms in the right-hand side of (25) depends simultaneously on both P_1 and P_2. Interestingly, these terms depend 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 (25) to construct closed-form expressions for the sensitivity $G(\boldsymbol{\theta}, P_1, P_2)$ in (21). 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.

Two choices of P_S for which the expression in the right-hand side of (25) significantly simplifies are $P_S = P_1$ and $P_S = P_2$, which leads to the following corollary of Theorem 3.

Corollary 3 *If P_1 is absolutely continuous with P_2, then the value $G(\boldsymbol{\theta}, P_1, P_2)$ in (21) satisfies:*

$$G(\boldsymbol{\theta}, P_1, P_2) = \beta\Big(D\left(P_2 \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_2,\beta)}\right) - D\left(P_1 \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_2,\beta)}\right) + D\left(P_1 \| P_2\right) \Big). \tag{26}$$

Alternatively, if P_2 is absolutely continuous with P_1 then,

$$G(\boldsymbol{\theta}, P_1, P_2) = \beta\Big(D\left(P_2 \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_1,\beta)}\right) - D\left(P_1 \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_1,\beta)}\right) - D\left(P_2 \| P_1\right) \Big), \tag{27}$$

where for all $i \in \{1, 2\}$, the probability measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_i,\beta)}$ satisfies (11) under the assumption that $P_S = P_i$.

Interestingly, absolute continuity of P_1 with respect to P_2 or of 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 (21). Note that choosing P_S as a convex combination of P_1 and P_2, 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.

5 Analysis of the Empirical-Risk

This section presents a mathematical object known as a *type* in the realm of information theory [30]. In the context of this work, a type is a probability measure induced by a dataset, as shown hereunder.

Definition 1 (The Type) *The type induced by the dataset $\boldsymbol{z}$ in (2) on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathscr{F}_{\mathcal{X}\times\mathcal{Y}})$, denoted by $P_{\boldsymbol{z}}$, is such that for all singletons $\{(x, y)\} \in \mathscr{F}_{\mathcal{X}\times\mathcal{Y}}$,*

$$P_{\boldsymbol{z}}\left(\{(x, y)\}\right) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{x=x_t, y=y_t\}}(x, y). \tag{28}$$

This definition illustrates the reason why the type is often referred to as *empirical probability measure*. In the following, the abuse of noting $P_{\boldsymbol{z}}\left(\{(x, y)\}\right)$

as $P_{\boldsymbol{z}}(x, y)$ is allowed for ease of presentation. The central observation of this section is that the empirical risk $\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})$ in (6) can be written as the expectation of the loss with respect to the type $P_{\boldsymbol{z}}$. This is formalized by the following lemma.

Lemma 4 (Empirical Risks and Types) *The empirical risk* $\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})$ *in* (6) *satisfies*

$$\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) = \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{\boldsymbol{z}}(x, y), \tag{29}$$

where the measure $P_{\boldsymbol{z}}$ *is the type induced by the dataset* $\boldsymbol{z}$ *in* (2) *and the function* ℓ *is defined in* (5)*.*

Proof: The proof is presented in Appendix G. ■

Equipped with the result in Lemma 4, for a fixed model, the sensitivity of the empirical risk to changes on the datasets can be characterized using the results obtained in the previous section for the expected loss. More specifically, consider the two datasets $\boldsymbol{z}_1 \in (\mathcal{X} \times \mathcal{Y})^{n_1}$ and $\boldsymbol{z}_2 \in (\mathcal{X} \times \mathcal{Y})^{n_2}$ that induce the types $P_{\boldsymbol{z}_1}$ and $P_{\boldsymbol{z}_2}$, respectively. Hence, given a model $\boldsymbol{\theta} \in \mathcal{M}$, it follows that

$$G(\boldsymbol{\theta}, P_{\boldsymbol{z}_1}, P_{\boldsymbol{z}_2}) = \mathsf{L}(\boldsymbol{z}_1, \boldsymbol{\theta}) - \mathsf{L}(\boldsymbol{z}_2, \boldsymbol{\theta}), \tag{30}$$

where the functional G is in (21). Assume that $P_{\boldsymbol{z}_1}$ and $P_{\boldsymbol{z}_2}$ are absolutely continuous with respect to the reference measure P_S in (9). Under this assumption, the equality in (30) leads to a characterization of the sensitivity of the empirical risk induced by a given model $\boldsymbol{\theta}$ when the dataset is changed from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$.

Theorem 4 *Given two datasets* $\boldsymbol{z}_1 \in (\mathcal{X} \times \mathcal{Y})^{n_1}$ *and* $\boldsymbol{z}_2 \in (\mathcal{X} \times \mathcal{Y})^{n_2}$ *whose types* $P_{\boldsymbol{z}_1}$ *and* $P_{\boldsymbol{z}_2}$ *are absolutely continuous with respect to the measure* P_S *in* (9)*, the following holds for all* $\boldsymbol{\theta} \in \mathcal{M}$*:*

$$\begin{aligned} &\mathsf{L}(\boldsymbol{z}_1, \boldsymbol{\theta}) - \mathsf{L}(\boldsymbol{z}_2, \boldsymbol{\theta}) \\ &= \beta\Big(D\left(P_{\boldsymbol{z}_2} \| P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_{\boldsymbol{z}_1} \| P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_{\boldsymbol{z}_2} \| P_S\right) + D\left(P_{\boldsymbol{z}_1} \| P_S\right) \Big), \end{aligned} \tag{31}$$

where the function L *is in* (6)*; the model* $\boldsymbol{\theta} \in \mathcal{M}$*, and the measures* P_S *and* $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ *satisfy* (11)*.*

Proof: The proof follows from the equality in (30), which together with Theorem 3 completes the proof. ■

In Theorem 4, the reference measure P_S can be arbitrarily chosen as long as both types $P_{\boldsymbol{z}_1}$ and $P_{\boldsymbol{z}_2}$ are absolutely continuous with P_S. A choice that satisfies this constraint is the type induced by the aggregation of both datasets $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, which is denoted by $\boldsymbol{z}_0 = (\boldsymbol{z}_1, \boldsymbol{z}_2) \in (\mathcal{X} \times \mathcal{Y})^{n_0}$, with $n_0 = n_1 + n_2$. The type induced by the aggregated dataset $\boldsymbol{z}_0$, denoted by $P_{\boldsymbol{z}_0}$, is a convex combination of the types $P_{\boldsymbol{z}_1}$ and $P_{\boldsymbol{z}_2}$, that is, $P_{\boldsymbol{z}_0} = \frac{n_1}{n_0} P_{\boldsymbol{z}_1} + \frac{n_2}{n_0} P_{\boldsymbol{z}_2}$, which satisfies the absolute continuity conditions [6].

From Theorem 4, it appears that the difference between a test empirical risk $\mathsf{L}(\boldsymbol{z}_1, \boldsymbol{\theta})$ and the training empirical risk $\mathsf{L}(\boldsymbol{z}_2, \boldsymbol{\theta})$ of a given model $\boldsymbol{\theta}$ is determined by two values: (a) the difference of the "statistical distance" from the types induced by the training and test datasets to the worst-case data-generating probability measure, i.e., $D\left(P_{\boldsymbol{z}_2} \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{\boldsymbol{z}_1} \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right)$; and (b) the difference of the "statistical distance" from the types to the reference measure P_S, i.e., $D\left(P_{\boldsymbol{z}_1} \| P_S\right) - D\left(P_{\boldsymbol{z}_2} \| P_S\right)$.

6 Analysis of the Generalization Gap

The generalization gap induced by a given model $\boldsymbol{\theta} \in \mathcal{M}$, which is assumed to be obtained with a training dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, under the assumption that training and test datasets are independent and identically distributed according to the probability measure $P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$, is

$$G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_Z(x, y) - \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{\boldsymbol{z}}(x, y). \tag{32}$$

The term $\int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{\boldsymbol{z}}(x, y) = \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})$ is an empirical risk often referred to as the training risk or training loss [31]. This is essentially the loss induced by the model with respect to the dataset used for training. The term $\int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_Z(x, y)$ is the population risk, also known as true risk. That is, the expected loss under the assumption that the ground-truth probability distribution of the data points is P_Z. Interestingly, as shown in (32), such generalization error can be written in terms of the functional G in (21). This observation leads to the following description of the generalization gap.

Lemma 5 *The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (32) satisfies:*

$$G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \beta\Big(D\Big(P_{\boldsymbol{z}} \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z,\beta)}\Big) - D(P_{\boldsymbol{z}} \| P_Z) - D\Big(P_Z \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z,\beta)}\Big)\Big), \tag{33}$$

where the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z,\beta)}$ is the solution to the optimization problem in (9) under the assumption that $P_S = P_Z$.

Proof: The proof follows from Corollary 3 by noticing that the type $P_{\boldsymbol{z}}$ is absolutely continuous with respect to P_Z. ∎

Lemma 5 highlights the intuition that if the type $P_{\boldsymbol{z}}$ induced by the training dataset $\boldsymbol{z}$ is at an arbitrary small "statistical distance" of the ground-truth measure P_Z, the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (32) is arbitrarily close to zero. This is revealed by the fact that an arbitrary small value of $D\left(P_{\boldsymbol{z}} \| P_Z\right)$ implies the difference $D\left(P_{\boldsymbol{z}} \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z,\beta)}\right) - D\left(P_Z \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z,\beta)}\right)$ is also arbitrarily small.

A more general expression for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (32) is provided by the following corollary of Theorem 3.

Corollary 4 *The generalization gap* $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ *in* (32) *satisfies:*

$$G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \beta\Big(D\left(P_{\boldsymbol{z}} \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_Z \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{\boldsymbol{z}} \| P_S\right) + D\left(P_Z \| P_S\right)\Big), \tag{34}$$

where the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_Z,\beta)}$ *is in* (9).

Note that several expressions for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (32) can be obtained from Corollary 4 by choosing the reference P_S and the parameter γ in (9), which determines the value of β.

6.1 Expected Generalization Gap

A conditional probability distribuition $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, such that given a training dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in (\mathcal{M}, \mathscr{B}(\mathcal{M}))$ is used to choose models, is referred to as a statistical learning algorithm. This subsection, provides explicit expressions for the generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ and a given training dataset.

The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (32) is due to a particular model $\boldsymbol{\theta}$, which has been deterministically obtained from the training dataset $\boldsymbol{z}$. When the model is chosen by using a statistical learning algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, trained upon the dataset $\boldsymbol{z}$, the expected generalization gap is the expectation of $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ when $\boldsymbol{\theta}$ is sampled from $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$. Let $\overline{G} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M})) \times \triangle(\mathcal{X} \times \mathcal{Y}) \times \triangle(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ be such that

$$\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}}) = \int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{35}$$

where the functional G is in (32). Using this notation, the expected generalization error induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, when the training dataset is $\boldsymbol{z}$, is $\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}})$ in (35). Corollary 4, by appropriately choosing the reference measure P_S and the parameter γ in (9), leads to numerous closed-form expressions for the expected generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ for the training dataset $\boldsymbol{z}$. Interestingly, regardless of the choice of P_S and γ, the resulting expressions describe the impact of the training dataset $\boldsymbol{z}$ on the expected generalization gap.

6.2 Doubly-Expected Generalization Gap

The expected generalization gap $\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}})$ in (35) depends on the training dataset $\boldsymbol{z}$. The doubly-expected generalization gap is obtained by taking the expectation of $\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}})$ when $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ is sampled from $P_{\boldsymbol{Z}}$, which is assumed to be the product distribution formed by P_Z. Let $\overline{\overline{G}} : \triangle\Big(\mathcal{M}|\left(\mathcal{X} \times \mathcal{Y}\right)^n\Big) \times \triangle\Big(\left(\mathcal{X} \times \mathcal{Y}\right)^n\Big) \to \mathbb{R}$ be a functional such that

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z) = \int\int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{36}$$

where the functional G is in (32). Using this notation, the doubly-expected generalization error induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z)$ in (36). In existing literature, the doubly-expected generalization gap is simply referred to as generalization error. See for instance [4], [2], and [7]. Note that in these previous works, the dependence on a particular training dataset is not explicit due to results being presented for the case in which the expectation is taken with respect to all sources of randomness in the corresponding expression. As in the case of the expected generalization gap, Corollary 4 leads to numerous closed-form expressions for the doubly-expected generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$.

6.3 The Gibbs Algorithm

A typical statistical learning algorithm is the Gibbs algorithm, which is parametrized by a positive real λ and by a σ-measure $Q \in \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}))$ [7]. The probability measure representing such an algorithm, which is denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ and for all $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$,

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

where the dataset $\boldsymbol{z}$ represents the training dataset; the function L is defined in (6); and the function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ satisfies

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp\left(t\, \mathsf{L}(\boldsymbol{z}, \boldsymbol{\nu})\right) \mathrm{d}Q(\boldsymbol{\nu})\right). \tag{38}$$

The doubly-expected generalization error induced by the Gibbs algorithm with parameters Q and λ, under the assumption that datasets are sampled from a product distribution formed by the measure P_Z, denoted $\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z)$ satisfies the following property.

Lemma 6 (Generalization Gap of the Gibbs Algorithm) *Given the conditional probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in (37) and a probability measure $P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$, the generalization gap $\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z)$ satisfies*

$$\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z) = \lambda\left(I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) + L\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)\right), \tag{39}$$

where $P_{\boldsymbol{Z}} \in \triangle(\mathcal{X} \times \mathcal{Y})^n$ is a product measure obtained from P_Z; and the terms $I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)$ and $L\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)$ are, respectively, the mutual information and the lautum information given by

$$I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \text{ and} \tag{40}$$

$$L\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), \tag{41}$$

with the probability measure $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ *being such that for all sets* $\mathcal{A} \in \mathscr{B}(\mathcal{M})$, $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}(\mathcal{A}) = \int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}(\mathcal{A}) \, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu})$.

Proof: This proof is presented in Appendix H. ■

Lemma 6 has been proved before for the case in which Q is a probability measure in [2]; and in the more general case in which Q is a σ-finite measure in [7]. In both [2] and [7], the result is shown without the assumption that the measure $P_{\boldsymbol{Z}}$ is a product measure, which is an assumption in Lemma 6. This limitation is due to the fact that the proof of Lemma 6 relies on the notion of types, which is known to fail capturing the correlation between datapoints, as pointed in [30]. Nonetheless, the independent and identically distributed assumption is widely adopted in the realm of machine learning. Despite this limitation, the relevance of Lemma 6 stems from the fact that a connection has been made between the notion of sensitivity to deviations from the worst-case data-generating measure, which is captured by the functional G in (21), and the notion of (doubly-expected) generalization gap, which is a central performance metric for evaluating the generalization capabilities of machine learning algorithms.

7 Conclusions and Final Remarks

The worst-case data-generating probability measure in Theorem 1 has been shown to be a cornerstone in statistical machine learning. This is due to 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. The dependence of these performance metrics on the worst-case data-generating probability measure is shown to exist via the sensitivity of the expectation of the loss function to changes from the worst-case data-generating probability measure to any alternative probability measure. This observation is reminiscent of the dependence of the expected generalization gap and the doubly-expected generalization gap on a Gibbs probability measure on the measurable space of the models as shown in [7, 29, 36]. These dependences suggest an intriguing relation between the probability measure (on the models) describing the Gibbs algorithm and the worst-case probability measure (on the datasets) introduced in this work, which is also a Gibbs probability measure. The connection appears to be nontrivial and is suggested as a promising line of work in this area.

Appendices

Miscellaneous Results

In this section, preliminary results for proving the main results in this work are introduced.

The first result introduces the Leibniz integral rule for measurable functions that are Lipschitz continuous.

Theorem 5 *Given a probability measure space $(\Omega, \mathscr{F}, \mu)$ and an open subset $\mathcal{A}$ of $\mathbb{R}$, let the function $f : \mathcal{A} \times \Omega \to \mathbb{R}$ be measurable with respect to $(\mathcal{A} \times \Omega, \mathscr{F}_{\mathcal{A}\times\Omega})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. If for all $v \in \Omega$, the function $f(\cdot, v) : \mathcal{A} \to \mathbb{R}$ is Lipschitz continuous and for some $u \in \mathcal{A}$, $\int f(u, v)\, \mathrm{d}\mu(v) < +\infty$, then*

$$\frac{\mathrm{d}}{\mathrm{d}t} \int f(t, v)\, \mathrm{d}\mu(v)\Big|_{t=u} = \int \frac{\mathrm{d}}{\mathrm{d}t} f(t, v)\Big|_{t=u} \mathrm{d}\mu(v). \tag{42}$$

Proof: Note that

$$\frac{\mathrm{d}}{\mathrm{d}t} \int f(t, v)\, \mathrm{d}\mu(v)\Big|_{t=u} = \lim_{\delta\to 0} \frac{\int f(u, v)\, \mathrm{d}\mu(v) - \int f(u-\delta, v)\, \mathrm{d}\mu(v)}{\delta} \tag{43}$$

$$= \lim_{\delta\to 0} \int \frac{f(u, v) - f(u-\delta, v)}{\delta} \mathrm{d}\mu(v), \tag{44}$$

where the equality in (44) follows from [37, Theorem 1.6.3]. The assumption that for all $v \in \Omega$, the function $f(\cdot, v)$ is Lipschitz continuous implies that for all $u \in \mathcal{A}$ and some $\delta \in \mathbb{R}$,

$$|f(u, v) - f(u-\delta, v)| \leq L\,|\delta|, \tag{45}$$

with $L < +\infty$. And thus, dividing the RHS and LHS of (45) by $|\delta|$ yields

$$\left|\frac{f(u, v) - f(u-\delta, v)}{\delta}\right| \leq L, \tag{46}$$

which implies that

$$\int \left|\frac{f(u, v) - f(u-\delta, v)}{\delta}\right| \mathrm{d}\mu(v) \leq L < +\infty. \tag{47}$$

This allows using the dominated convergence theorem [37, Theorem 1.6.9] as follows. From (44), the following holds:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int f(t, v)\, \mathrm{d}\mu(v)\Big|_{t=u} = \lim_{\delta\to 0} \int \frac{f(u, v) - f(u-\delta, v)}{\delta} \mathrm{d}\mu(v) \tag{48}$$

$$= \int \lim_{\delta\to 0} \frac{f(u, v) - f(u-\delta, v)}{\delta} \mathrm{d}\mu(v) \tag{49}$$

$$= \int \frac{\mathrm{d}}{\mathrm{d}t} f(t, v)\Big|_{t=u} \mathrm{d}\mu(v), \tag{50}$$

where the equality in (49) follows from the dominated convergence theorem [37, Theorem 1.6.9]. This completes the proof. ■

This second result is the definition of separable functions, which is already introduced in [7, Definition 4.1].

Definition 2 *[7, Definition 4.1] Given a model $\boldsymbol{\theta} \in \mathcal{M}$, the function ℓ in (5), 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 $\mathcal{X} \times \mathcal{Y}$ that are nonnegligible with respect to P, and for all $(x_1, y_1) \in \mathcal{A}$ and for all $(x_2, y_2) \in \mathcal{B}$, it holds that*

$$\ell(\boldsymbol{\theta}, x_1, y_1) < c < \ell(\boldsymbol{\theta}, x_2, y_2) < +\infty. \tag{51}$$

When the function ℓ in (5) is nonseparable with respect to a probability measure P, it is a constant almost surely with respect to such a measure. More specifically, there exists a real $a \geq 0$, such that

$$P(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = a\}) = 1. \tag{52}$$

The next lemma introduces some properties of the log-partition function $\mathsf{J}_{P,\boldsymbol{\theta}}$ in (10).

Lemma 7 *The function $\mathsf{J}_{P,\boldsymbol{\theta}}$ in (10) is convex, nondecreasing, and differentiable infinitely many times in the interior of $\{t \in \mathbb{R} : \mathsf{J}_{P,\boldsymbol{\theta}}(t) < +\infty\}$. If the loss function ℓ in (5) is a separable function with respect to P, then the function $\mathsf{J}_{P,\boldsymbol{\theta}}$ is strictly convex.*

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 $\mathsf{J}_{P,\boldsymbol{\theta}}(\gamma_1) < +\infty$ and $\mathsf{J}_{P,\boldsymbol{\theta}}(\gamma_2) < +\infty$. Then, for all $\alpha \in (0, 1)$, the following holds

$$\begin{aligned}
&\alpha \mathsf{J}_{P,\boldsymbol{\theta}}(\gamma_1) + (1-\alpha)\mathsf{J}_{P,\boldsymbol{\theta}}(\gamma_2) \\
&= \alpha \log\left(\int \exp(\gamma_1 t) \mathrm{d}P(t)\right) + (1-\alpha)\log\left(\int \exp(\gamma_2 t)\mathrm{d}P(t)\right) && (53)\\
&= \log\left(\left(\int \exp(\gamma_1 t)\mathrm{d}P(t)\right)^{\alpha}\right) + \log\left(\left(\int \exp(\gamma_2 t)\mathrm{d}P(t)\right)^{1-\alpha}\right) && (54)\\
&= \log\left(\left(\int \exp(\gamma_1 t)\mathrm{d}P(t)\right)^{\alpha}\left(\int \exp(\gamma_2 t)\mathrm{d}P(t)\right)^{1-\alpha}\right) && (55)\\
&= \log\left(\left(\int \exp(\gamma_1 \alpha t)^p \mathrm{d}P(t)\right)^{\frac{1}{p}}\left(\int \exp(\gamma_2 (1-\alpha) t)^q \mathrm{d}P(t)\right)^{\frac{1}{q}}\right) && (56)\\
&\geq \log\left(\int \exp(\gamma_1 \alpha t)\exp(\gamma_2(1-\alpha)t)\mathrm{d}P(t)\right) && (57)\\
&= \log\left(\int \exp(\gamma_1 \alpha t + \gamma_2(1-\alpha)t)\mathrm{d}P(t)\right) && (58)\\
&= \mathsf{J}_{P_S,\boldsymbol{\theta}}(\gamma_1\alpha + \gamma_2(1-\alpha)), && (59)
\end{aligned}$$

where the equality in (56) follows with $\alpha \triangleq \frac{1}{p}$ and $1-\alpha \triangleq \frac{1}{q}$; and the inequality in (57) follows from Hölder's inequality [37, Theorem 2.4.5]. This proves the convexity of the function $\mathsf{J}_{P,\boldsymbol{\theta}}$.

Note that, also from Hölder's inequality [37, Theorem 2.4.5], the equality in (57) holds if and only if there exist two constants β_1 and β_2, not simultaneously equal to zero, such that the set

$$\mathcal{A} = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : \beta_1 \exp\left(\gamma_1 \ell(\boldsymbol{\theta}, x, y)\right) = \beta_2 \exp\left(\gamma_2 \ell(\boldsymbol{\theta}, x, y)\right)\} \tag{60}$$

$$= \left\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \exp\left((\gamma_1 - \gamma_2)\ell(\boldsymbol{\theta}, x, y)\right) = \frac{\beta_2}{\beta_1}\right\} \tag{61}$$

$$= \left\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = \frac{\log\left(\frac{\beta_2}{\beta_1}\right)}{(\gamma_1 - \gamma_2)}\right\}, \tag{62}$$

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

The proof of monotonicity follows from noticing that for all$(t_1, t_2) \in \left\{t \in \mathbb{R} : \mathsf{J}_{P,\boldsymbol{\theta}}(t) < +\infty\right\}^2$, such that $t_1 < t_2$, it follows that for all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $(x,y) \in \operatorname{supp} P$, the inequality $\exp\left(t_1 \ell(\boldsymbol{\theta}, x, y)\right) \leq \exp\left(t_2 \ell(\boldsymbol{\theta}, x, y)\right)$ holds. This implies that $\mathsf{J}_{P,\boldsymbol{\theta}}(t_1) \leq \mathsf{J}_{P,\boldsymbol{\theta}}(t_2) < +\infty$, which proves that the function is nondecreasing. Moreover, the equality holds if and only if $P\Big(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = 0\}\Big) = 1$.

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

$$\lim_{t \to \alpha} \mathsf{J}_{P,\boldsymbol{\theta}}(t) = \lim_{t \to \alpha} \log\left(\int \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P(x,y)\right) \tag{63}$$

$$= \log\left(\lim_{t \to \alpha} \int \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P(x,y)\right) \tag{64}$$

$$= \log\left(\int \lim_{t \to \alpha} \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P(x,y)\right) \tag{65}$$

$$= \log\left(\int \exp\left(\alpha\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P(x,y)\right) \tag{66}$$

$$= \mathsf{J}_{P,\boldsymbol{\theta}}(\alpha), \tag{67}$$

where equalities in (64) and (66) follow from the fact that both the logarithmic and exponential functions are continuous; and equality in (65) follows from the dominated convergence theorem [37, Theorem 1.6.9]. This proves that the function $\mathsf{J}_{P,\boldsymbol{\theta}}$ is continuous in $\{t \in \mathbb{R} : \mathsf{J}_{P,\boldsymbol{\theta}}(t) < +\infty\}$.

The proof of differentiability follows by considering the transport of the measure P from $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X} \times \mathcal{Y}})$ to $([0, +\infty], \mathscr{B}([0, +\infty]))$ through the function $g_{\boldsymbol{\theta}} : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ such that $g_{\boldsymbol{\theta}}(x,y) = \ell(\boldsymbol{\theta}, x, y)$, with ℓ in (5). Denote the resulting

probability measure in $([0,+\infty],\mathscr{B}([0,+\infty]))$ by $T_{P,\boldsymbol{\theta}}$. That is, for all $\mathcal{A} \in \mathscr{B}([0,+\infty])$,

$$T_{P,\boldsymbol{\theta}}\left(\mathcal{A}\right) = P\left(g_{\boldsymbol{\theta}}^{-1}\left(\mathcal{A}\right)\right), \tag{68}$$

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

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

Hence, the following holds

$$\mathsf{J}_{P,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P(x,y)\right) \tag{69}$$

$$= \log\left(\int \exp\left(tv\right)\mathrm{d}T_{P,\boldsymbol{\theta}}(v)\right), \tag{70}$$

where equality in (70) follows from [37, Theorem 1.6.12].

Denote by ϕ the Laplace transform of the measure P. That is for all $t \in \{x \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(x\right) < +\infty\}$,

$$\phi(t) = \int \exp\left(tv\right)\mathrm{d}T_{P,\boldsymbol{\theta}}(v). \tag{71}$$

Hence, $\phi(t) = \exp\left(\mathsf{J}_{P,\boldsymbol{\theta}}(t)\right)$. From [38, Theorem 1$a$ (page 439)], it follows that the function ϕ has derivatives of all orders in $\{t \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(t\right) < +\infty\}$, and thus, so does the function $\mathsf{J}_{P,\boldsymbol{\theta}}$ in the interior of $\{t \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(t\right) < +\infty\}$. This completes the proof. ■

Let the m-th derivative of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (10) be denoted by $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$, with $m \in \mathbb{N}$. Hence, for all $t \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (13),

$$\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}(t) \triangleq \frac{\mathrm{d}^m}{\mathrm{d}s^m}\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(s\right)\Big|_{s=t}. \tag{72}$$

The following lemma provides explicit expressions for the first and second derivatives of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (10).

Lemma 8 *For all t in the interior of $\{s \in (0,+\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right) < +\infty\}$, the first and second derivatives of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in* (10)*, denoted respectively by $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}$ and $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}$, satisfy that*

$$\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y), \text{ and} \tag{73}$$

$$\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = \int \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y), \tag{74}$$

where the measure $P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is in (11) *and the function ℓ is defined in* (5)*. Moreover, $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$ is strictly positive if and only if the function ℓ in* (5) *is separable with respect to the measure P_S.*

Proof: Note that for all t in the interior of $\{s \in (0,+\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right) < +\infty\}$,

$$\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = \frac{\mathrm{d}}{\mathrm{d}s}\log\left(\int \exp\left(s\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)\right)\Bigg|_{s=\frac{1}{t}} \tag{75}$$

$$= \frac{\frac{\mathrm{d}}{\mathrm{d}s}\int\left(\exp\left(s\ell(\boldsymbol{\theta},x,y)\right)\right)\mathrm{d}P_S(x,y)}{\int \exp\left(s\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)}\Bigg|_{s=\frac{1}{t}} \tag{76}$$

$$= \frac{\int \frac{\mathrm{d}}{\mathrm{d}s}\left(\exp\left(s\ell(\boldsymbol{\theta},x,y)\right)\right)\mathrm{d}P_S(x,y)}{\int \exp\left(s\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)}\Bigg|_{s=\frac{1}{t}} \tag{77}$$

$$= \frac{\int \ell(\boldsymbol{\theta},x,y)\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)}{\int \exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)} \tag{78}$$

$$= \exp\left(-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\int \ell(\boldsymbol{\theta},x,y)\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) \tag{79}$$

$$= \int \ell(\boldsymbol{\theta},x,y)\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y) \tag{80}$$

$$= \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y), \tag{81}$$

where the equality in (77) follows from the dominated convergence theorem [37, Theorem 1.6.9]; and the equality in (81) follows from (11). This completes the proof of (73).

The proof of (74) is as folllows,

$$\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = \frac{\mathrm{d}}{\mathrm{d}s}\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}(s)\bigg|_{s=\frac{1}{t}} \tag{82}$$

$$= \frac{\mathrm{d}}{\mathrm{d}s}\Big(\int \ell(\boldsymbol{\theta},x,y)\exp\left(s\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}(s)\right)\mathrm{d}P_S(x,y)\Big)\bigg|_{s=\frac{1}{t}} \tag{83}$$

$$= \int \frac{\mathrm{d}}{\mathrm{d}s}\Big(\ell(\boldsymbol{\theta},x,y)\exp\left(s\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}(s)\right)\Big)\mathrm{d}P_S(x,y)\bigg|_{s=\frac{1}{t}} \tag{84}$$

$$= \int \ell(\boldsymbol{\theta},x,y)(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right))$$
$$\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y) \tag{85}$$

$$= \int\left(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2\exp\left(t\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y)$$
$$+\int \ell(\boldsymbol{\theta},x,y)\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y)$$
$$-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)^2\int\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y) \tag{86}$$

$$= \int\left(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y)$$
$$+\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)^2-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)^2\int\mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{87}$$

$$= \int\left(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y) \tag{88}$$

$$= \int\left(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2\mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y), \tag{89}$$

where the equality in (84) follows from the dominated convergence theorem [37, Theorem 1.6.9]; and the equalities in (87) and (89) follow from (11). This completes the proof of (74).

The rest of the proof of Lemma 8 is divided into two parts. First, it is shown that if the function ℓ in (5) is nonseparable with respect to the measure P_S (Definition 2), then for all $t \in \{s \in (0,+\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right) < +\infty\}$, $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = 0$. The second part of the proof shows that if the function ℓ is separable, then, for all $t \in \{s \in (0,+\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right) < +\infty\}$, $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) > 0$.

The first part is as follows. Assume that the function ℓ in (5) is nonseparable with respect to the measure P_S, i.e., there exists a real $a \geq 0$, such that

$$P_S\left(\{(x,y) \in \mathcal{X}\times\mathcal{Y} : \ell(\boldsymbol{\theta},x,y) = a\}\right) = 1. \tag{90}$$

Note that from (11) and (73), it holds that

$$\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{91}$$

$$= \int \ell(\boldsymbol{\theta},x,y)\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y) \tag{92}$$

$$= \int a\exp\left(\frac{a}{t}-\frac{a}{t}\right)\mathrm{d}P_S(x,y) \tag{93}$$

$$= a. \tag{94}$$

Note also that from (11)and (74), it holds that

$$\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = \int \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{95}$$

$$= \int \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)\mathrm{d}P_S(x,y) \tag{96}$$

$$= \int (a-a)\exp\left(\frac{a}{t}-\frac{a}{t}\right)\mathrm{d}P_S(x,y) \tag{97}$$

$$= 0, \tag{98}$$

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

The second part of the proof is as follows. Assume that the function ℓ 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 nonnegaligible with respect to P_S and verify that for all $(x_1,y_1)\in\mathcal{A}$ and for all $(x_2,y_2)\in\mathcal{B}$,

$$\ell\left(\boldsymbol{\theta},x_1,y_1\right) < a < \ell\left(\boldsymbol{\theta},x_2,y_2\right). \tag{99}$$

Note that the sets $\mathcal{A}$ and $\mathcal{B}$ are disjoint. Hence, from (74), it holds that

$$\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) = \int \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{100}$$

$$= \int_{\mathcal{A}} \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y)$$

$$+ \int_{\mathcal{B}} \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{101}$$

$$+ \int_{\mathcal{X}\times\mathcal{Y}\setminus(\mathcal{A}\cup\mathcal{B})} \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{102}$$

$$> 0, \tag{103}$$

where the inequality in (103) is based on the folllowing facts. First, if $a < \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$, with a in (99), then for all $(x,y)\in\mathcal{A}$, it holds that $\ell(\boldsymbol{\theta},x,y) < a < \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$, and thus,

$$\left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 > \left(a - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2. \tag{104}$$

This implies that

$$\int_{\mathcal{A}} \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{105}$$

$$> \int_{\mathcal{A}} \left(a - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y)) \tag{106}$$

$$> \left(a - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}) \tag{107}$$

$$> 0, \tag{108}$$

where the inequality in (108) follows from the fact that the probability measures $P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ and P_S are mutually absolutely continuous (Lemma 2). Second, if $a \geq \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$, with a in (99), then for all $(x,y) \in \mathcal{B}$, it holds that $\ell(\boldsymbol{\theta},x,y) > a \geq \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$, and thus,

$$\left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 > \left(a - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2, \tag{109}$$

which implies

$$\int_{\mathcal{B}} \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{110}$$

$$> \int_{\mathcal{B}} \left(a - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 \mathrm{d}P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{111}$$

$$> \left(a - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)\right)^2 P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{B}) \tag{112}$$

$$> 0, \tag{113}$$

where the inequality in (113) follows from the fact that the probability measures $P^{(P_S,t)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ and P_S are mutually absolutely continuous (Lemma 2). This completes the proof.

■

A Proof of Theorem 1

This section is divided into two parts. First, some preliminary results that are needed for the proof are introduced. The second part presents the proof.

A.1 Preliminaries

Lemma 9 *Let $\mathscr{M}$ be the set of measurable functions $\mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ with respect to the measurable spaces $(\mathcal{X}\times\mathcal{Y}, \mathcal{F}_{\mathcal{X}\times\mathcal{Y}})$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Let $\mathscr{S}$ be the subset of*

$\mathscr{M}$ including all nonnegative functions that are absolutely integrable with respect to a probability measure P_S. That is, for all $h \in \mathscr{S}$, it holds that

$$\int |h(x,y)| \, \mathrm{d}P_S(x,y) < \infty. \tag{114}$$

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

$$\hat{r}(\alpha) = \int \left(g(x,y) + \alpha h(x,y)\right) \log\left(g(x,y) + \alpha h(x,y)\right) \mathrm{d}P_S(x,y), \tag{115}$$

for some functions g and h in $\mathscr{S}$ and $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily small. Then, the function $\hat{r}$ in (115) *is differentiable at zero.*

Proof: The objective is to prove that the function $\hat{r}$ in (115) is differentiable at zero, which boils down to proving that the limit

$$\lim_{\delta \to 0} \frac{1}{\delta} \left(\hat{r}(\alpha + \delta) - \hat{r}(\alpha)\right) \tag{116}$$

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

$$q(x) = x \log x. \tag{117}$$

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

$$\hat{r}(\alpha) = \int q\Big(h(x,y) + \alpha g\left(x,y\right)\Big) \mathrm{d}P_S(x,y). \tag{118}$$

The proof of the existence of such a limit in (116) relies on the fact that the function q is strictly convex and differentiable, which implies that q is also Lipschitz continuous. Hence, it follows that

$$\left|q\left(h(x,y) + (\alpha + \delta)g(x,y)\right) - q\left(h(x,y) + \alpha g(x,y)\right)\right| \leq c \left|g(x,y)\right| \left|\delta\right|, \tag{119}$$

for some constant c positive and finite, which implies that

$$\frac{\left|q\left(h(x,y) + (\alpha + \delta)g(x,y)\right) - q\left(h(x,y) + \alpha g(x,y)\right)\right|}{|\delta|} \leq c \left|g(x,y)\right|, \tag{120}$$

and thus, given that $g \in \mathscr{S}$, it follows that

$$\int \frac{\left|q\left(h(x,y) + (\alpha + \delta)g(x,y)\right) - q\left(h(x,y) + \alpha g(x,y)\right)\right|}{|\delta|} \mathrm{d}P_S(x,y) < +\infty. \tag{121}$$

This allows using the dominated convergence theorem as follows. From the fact that the function q is differentiable, let $q^{(1)} : (0, +\infty) \to \mathbb{R}$ be the first derivative of q. The limit in (116) satisfies for $\alpha \in (-\epsilon, \epsilon)$, with $\epsilon > 0$ arbitrarily

small,

$$\lim_{\delta\to 0}\frac{1}{\delta}\left(\hat{r}(\alpha+\delta)-\hat{r}(\alpha)\right) \tag{122}$$

$$=\lim_{\delta\to 0}\frac{1}{\delta}\Big(\int q\left(h(x,y)+(\alpha+\delta)g(x,y)\right)\mathrm{d}P_S(x,y)$$

$$-\int q\left(h(x,y)+\alpha g(x,y)\right)\mathrm{d}P_S(x,y)\Big) \tag{123}$$

$$=\lim_{\delta\to 0}\int\frac{1}{\delta}\Big(q(h(x,y)+(\alpha+\delta)g(x,y))-q(h(x,y)+\alpha g(x,y))\Big)\mathrm{d}P_S(x,y) \tag{124}$$

$$=\int\lim_{\delta\to 0}\frac{1}{\delta}\Big(q(h(x,y)+(\alpha+\delta)g(x,y))-q(h(x,y)+\alpha g(x,y))\Big)\mathrm{d}P_S(x,y) \tag{125}$$

$$=\int q^{(1)}\left(h(x,y)+\alpha g(x,y)\right)\mathrm{d}P_S(x,y) \tag{126}$$

$$<+\infty, \tag{127}$$

where the equalities in (125) and (127) follow from the dominated convergence theorem [37, Theorem 1.6.9]. From (127), it follows that the function $\hat{r}$ in (137) is differentiable at zero. This completes the proof.

■

A.2 The Proof

The optimization problem in (9) can be re-written in terms of the Radon-Nikodym derivative of the optimization measure P with respect to the measure P_S, denoted by $\frac{\mathrm{d}P}{\mathrm{d}P_S}:\mathcal{M}\to[0,\infty)$, which yields:

$$\max_{P\in\Delta_{P_S}(\mathcal{X}\times\mathcal{Y})}\int\ell(\boldsymbol{\theta},x,y)\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y) \tag{128a}$$

$$\text{s. t.}\quad\int\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\log\Big(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\Big)\mathrm{d}P_S(x,y)\leq\gamma \tag{128b}$$

$$\int\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y)=1. \tag{128c}$$

The remainder of the proof focuses on the problem in which the optimization is over the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}P_S}$ 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 $\frac{\mathrm{d}P}{\mathrm{d}P_S}$ is unique, up to sets of zero measure with respect to P_S.

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

$$\int|h(x,y)|\,\mathrm{d}P_S(x,y)<\infty. \tag{129}$$

Note that the set $\mathscr{M}$ forms a real vector space and the set $\mathscr{S}$ is a convex subset of $\mathscr{M}$. Note also that the constraints (128b) and (128c) are satisfied by the probability measure P_S, which also satisfies $P_S \in \Delta_{P_S(\mathcal{X}\times\mathcal{Y})}$. Hence, the constraints do not induce an empty feasible set. Finally, note that the optimization problem in (128) can be written as a minimization of the form:

$$\min_{g\in\mathscr{S}} \quad -\int g(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) \tag{130a}$$

$$\text{s. t.} \quad \frac{1}{\gamma}\int g(x,y)\log\left(g(x,y)\right)\mathrm{d}P_S(x,y) \leq 1, \text{ and} \tag{130b}$$

$$\int g(x,y)\mathrm{d}P_S(x,y) = 1, \tag{130c}$$

where the expression $-\int g(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y)$ is linear with g; the expression $\frac{1}{\gamma}\int g(x,y)\log\left(g(x,y)\right)\mathrm{d}P_S(x,y)$ is convex with g; and $\int g(x,y)\mathrm{d}P_S(x,y)$ is linear with g.

The proof continues by assuming that the problem in (130) possesses a solution, which is denoted by $g^\star \in \mathscr{S}$. Let $\mu_0 \in [0,+\infty)$ be

$$\mu_0 \triangleq \min_{g\in\mathscr{S}} -\int g(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) \tag{131a}$$

$$\text{s. t.} \frac{1}{\gamma}\int g(x,y)\log\left(g(x,y)\right)\mathrm{d}P_S(x,y) \leq 1 \tag{131b}$$

$$\int g(x,y)\mathrm{d}P_S(x,y) = 1 \tag{131c}$$

$$= \quad -\int g^\star(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y). \tag{131d}$$

From [39, Theorem 1, Section 8.3, page 217], it holds that there exist two tuples (a_1,b_1) and (a_2,b_2) in $\mathbb{R}^2$ such that

$$\mu_0 = \min_{g\in\mathscr{S}}\Big\{ -\int g(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) + \frac{a_1}{\gamma}\int g(x,y)\log\left(g(x,y)\right)\mathrm{d}P_S(x,y) + b_1 + a_2\int g(x,y)\mathrm{d}P_S(x,y) + b_2\Big\}, \tag{132a}$$

and moreover,

$$0 = \frac{a_1}{\gamma}\int g^\star(x,y)\log\left(g^\star(x,y)\right)\mathrm{d}P_S(x,y) + b_1, \text{ and} \tag{132b}$$

$$0 = a_2\int g^\star(x,y)\mathrm{d}P_S(x,y) + b_2. \tag{132c}$$

Hence, the proof continues by solving the ancillary optimization problem in (132a), which is without constraints. This is essentially because the tuples (a_1,b_1) and (a_2,b_2) are such that the equalities (132b) and (132c) are already satisfied.

Let the functional $L : \mathscr{S} \to \mathbb{R}$ be such that

$$L(g) = -\int g(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) + \frac{a_1}{\gamma}\int g(x,y)\log\left(g(x,y)\right)\mathrm{d}P_S(x,y) + b_1 \\ +a_2\int g(x,y)\mathrm{d}P_S(x,y) + b_2. \tag{133}$$

Let $h : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ be a function in $\mathscr{S}$. The Gateaux differential of the functional L in (133) at $g \in \mathscr{S}$ in the direction of h is

$$\partial L(g;h) \triangleq \frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha)\Big|_{\alpha=0}, \tag{134}$$

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

$$r(\alpha) = -\int \left(g(x,y)+\alpha h(x,y)\right)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) \\ +\frac{a_1}{\gamma}\int \left(g(x,y)+\alpha h(x,y)\right)\log\left(g(x,y)+\alpha h(x,y)\right)\mathrm{d}P_S(x,y) \\ +b_1 + a_2\int \left(g(x,y)+\alpha h(x,y)\right)\mathrm{d}P_S(x,y) + b_2, \tag{135}$$

which can be rewritten as follows,

$$r(\alpha) = \alpha\int h(x,y)\left(a_2 - \ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) \\ +\frac{a_1}{\gamma}\int \left(g(x,y)+\alpha h(x,y)\right)\log\left(g(x,y)+\alpha h(x,y)\right)\mathrm{d}P_S(x,y) \\ +\int g(x,y)\left(a_2-\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) + b_1 + b_2. \tag{136}$$

Note that the first term in (136) is linear with α; the second term can be written using the function $\hat{r} : \mathbb{R} \to \mathbb{R}$ in (115) such that

$$\hat{r}(\alpha) = \int \left(g(x,y)+\alpha h(x,y)\right)\log\left(g(x,y)+\alpha h(x,y)\right)\mathrm{d}P_S(x,y); \tag{137}$$

and the third term is independent of α.

Hence, based on the fact the function $\hat{r}$ in (137) is differentiable at zero (Lemma 9), so is the function r in (136), which implies that the Gateaux differential of $\partial L(g;h)$ in (134) exists.

The derivative of the real function r in (136) is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\Big(& \alpha \int h(x,y)\left(a_2 - \ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) \\ & + \frac{a_1}{\gamma}\int \left(g(x,y)+\alpha h(x,y)\right)\log\left(g(x,y)+\alpha h(x,y)\right)\mathrm{d}P_S(x,y) \\ & + \int g(x,y)\left(a_2 - \ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) + b_1 + b_2\Big) \qquad (138) \\ = & \int h(x,y)(a_2 - \ell(\boldsymbol{\theta},x,y))\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) \\ & + \frac{a_1}{\gamma}\int \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(g(x,y)+\alpha h(x,y)\right)\log\left(g(x,y)+\alpha h(x,y)\right)\mathrm{d}P_S(x,y) \qquad (139) \\ = & \int h(x,y)(a_2 - \ell(\boldsymbol{\theta},x,y))\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) \\ & + \frac{a_1}{\gamma}\int h(x,y)\left(1+\log\left(g(x,y)+\alpha h(x,y)\right)\right)\mathrm{d}P_S(x,y), \qquad (140)\end{aligned}$$

where the equality in (139) follows from Theorem 5.

From equations (134) and (140), it follows that

$$\begin{aligned}\partial L\left(g;h\right) = & \int h(x,y)(a_2 - \ell(\boldsymbol{\theta},x,y))\mathrm{d}P_S(x,y) \\ & + \frac{a_1}{\gamma}\int h(x,y)\left(1+\log g(x,y)\right)\mathrm{d}P_S(x,y) \qquad (141) \\ = & \int h(x,y)\Big(a_2 - \ell(\boldsymbol{\theta},x,y) + \frac{a_1}{\gamma}\Big(1+\log\left(g^\star(x,y)\right)\Big)\Big)\mathrm{d}P_S(x,y). \qquad (142)\end{aligned}$$

A necessary condition [39, Theorem 1, Page 178] for the functional L in (133) to have a minimum at $g^\star$ is that for all functions $h \in \mathscr{S}$,

$$\partial L\left(g^\star;h\right) = 0. \qquad (143)$$

The equality in (143) holds for all functions $h \in \mathscr{S}$ if for all $(x,y) \in \operatorname{supp} P_S$, the function $g^\star$ satisfies:

$$a_2 - \ell(\boldsymbol{\theta},x,y) + \frac{a_1}{\gamma}\left(1+\log\left(g^\star(x,y)\right)\right) = 0. \qquad (144)$$

Assuming that

$$a_1 \neq 0, \qquad (145)$$

the equality in (144) implies

$$g^\star(x,y) = \exp\Big(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)\Big)\exp\Big(\frac{-a_2\gamma}{a_1} - 1\Big), \qquad (146)$$

where the values a_1 and a_2 satisfy (132b), and (132c), and (145).

The remainder of the proof focuses on determining the values of a_1, a_2, b_1, and b_2, which must also be such that g^* in (146) satisfies the constraints (130b) and (130c). For instance, from (130c) and (132c), it follows that a_2 must be such that

$$\exp\left(\frac{-a_2\gamma}{a_1}-1\right)=\frac{1}{\int\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)}, \tag{147}$$

that is,

$$a_2=\frac{a_1}{\gamma}\left(\log\left(\int\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)\right)-1\right) \tag{148}$$

$$=\frac{a_1}{\gamma}\left(\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)-1\right), \tag{149}$$

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is in (10). Plugging (149) into (146) yields

$$g^\star(x,y)=\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right), \tag{150}$$

which implies, from (132c), that

$$b_2=-a_2. \tag{151}$$

Moreover, from (130c) and (150), it holds that

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)<+\infty. \tag{152}$$

The function $g^\star$ in (150) represents the Radon-Nikodym derivative (with respect to P_S) of a solution $P^\star\in\Delta_{P_S}(\mathcal{X}\times\mathcal{Y})$ to the problem in (9a). Hence, the equality in (132b) can be written as follows:

$$\frac{a_1}{\gamma}D(P^\star\|P_S)+b_1{=}0, \tag{153}$$

which implies

$$b_1=-\frac{a_1}{\gamma}D(P^\star\|P_S). \tag{154}$$

Concerning a_1, note that if $a_1<0$, given two labelled patterns (x_1,y_1) and (x_2,y_2) in $\mathcal{Z}$, such that $\ell(\boldsymbol{\theta},x_1,y_1)<\ell(\boldsymbol{\theta},x_2,y_2)$, it holds that

$$g^\star(x_1,y_1){\geqslant}g^\star(x_2,y_2). \tag{155}$$

Thus, the focus in the remainder of the proof is the case in which

$$a_1>0. \tag{156}$$

Note that the measure $P^\star$ in (153) depends on a_1 through its Radon-Nikodym derivative with respect to P_S, i.e., $g^\star$ in (150). The proof is finalized by providing a condition to uniquely identify a_1. To this aim, note that

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}a_1}\int g^\star(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y)\\
&=\frac{\mathrm{d}}{\mathrm{d}a_1}\int \exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) &(157)\\
&=\int\frac{\mathrm{d}}{\mathrm{d}a_1}\left(\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) &(158)\\
&=\frac{-\gamma}{a_1^2}\int\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\\
&\quad\left(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) &(159)\\
&=\frac{-\gamma}{a_1^2}\left(\int\ell(\boldsymbol{\theta},x,y)^2\mathrm{d}P^\star(x,y)-\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\int\ell(\boldsymbol{\theta},x,y)\mathrm{d}P^\star(x,y)\right) &(160)\\
&=\frac{-\gamma}{a_1^2}\left(\int\ell(\boldsymbol{\theta},x,y)^2\mathrm{d}P^\star(x,y)-\left(\int\ell(\boldsymbol{\theta},x,y)\mathrm{d}P^\star(x,y)\right)^2\right) &(161)\\
&=\frac{-\gamma}{a_1^2}\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right) &(162)\\
&\leq 0, &(163)
\end{aligned}
$$

where the functions $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}$ and $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}$ are defined in (73) and (74), respectively; the equality in (158) follows the dominated convergence theorem [37, Theorem 1.6.9]; the equality in (161) follows from (73); the equality in (162) follows from (74); and the inequality in (163) follows from the fact that $\gamma>0$ and $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\geq 0$. Note that strict inequality in (163) holds if and only if the function ℓ defined in (5) is separable with respect to the measure P_S (Lemma 8).

Note also that

$$
\begin{aligned}
&D\left(P^\star\|P_S\right)\\
&=\int g^\star(x,y)\log\left(g^\star(x,y)\right)\mathrm{d}P_S(x,y) &(164)\\
&=\int\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\mathrm{d}P_S(x,y) &(165)\\
&=\int\frac{\gamma}{a_1}\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right), &(166)
\end{aligned}
$$

and thus,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}a_1} D\left(P^\star \| P_S\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}a_1}\left(\int \frac{\gamma}{a_1} \exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right) && (167)\\
&= \int \frac{\mathrm{d}}{\mathrm{d}a_1}\left(\frac{\gamma}{a_1}\exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) \\
&\quad - \left(\frac{-\gamma}{a_1^2}\right)\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right) && (168)\\
&= \left(\frac{-\gamma}{a_1^2}\right)\int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P^\star(x,y) + \left(\frac{\gamma}{a_1^2}\right)\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right) \\
&\quad + \left(\frac{-\gamma^2}{a_1^3}\right)\int \exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right) \\
&\quad \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) && (169)\\
&= \left(\frac{-\gamma^2}{a_1^3}\right)\int \exp\left(\frac{\gamma}{a_1}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right) \\
&\quad \left(\ell(\boldsymbol{\theta},x,y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right)\right)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y) && (170)\\
&= \left(\frac{-\gamma^2}{a_1^3}\right)\left(\int \ell(\boldsymbol{\theta},x,y)^2\mathrm{d}P^\star(x,y) - \left(\int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P^\star(x,y)\right)^2\right) && (171)\\
&= \left(\frac{-\gamma^2}{a_1^3}\right)\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{\gamma}{a_1}\right) && (172)\\
&\leq 0, && (173)
\end{aligned}$$

where the functions $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}$ and $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}$ are defined in (73) and (74), respectively; the equality in (168) follows from the dominated convergence theorem [37, Theorem 1.6.9]; the equalities in (171) follows from (73); the equality in (172) follows from (74); and the inequality follows from the fact that both γ and a_1 are positive. Note that strict inequality in (173) holds if and only if the function ℓ defined in (5) is separable with respect to the measure P_S (Lemma 8).

Hence, if the function ℓ defined in (5) is separable with respect to the measure P_S (Lemma 8), then the term $\int g^\star(x,y)\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_S(x,y)$ in (131d) and $\int g^\star(x,y)\log\left(g^\star(x,y)\right)\mathrm{d}P_S(x,y)$ in (164) are both simultaneously strictly decreasing with a_1. This implies that $a_1 > 0$ shall be chosen such that

$$D\left(P^\star \| P_S\right) = \gamma, \tag{174}$$

and justify the uniqueness of the solution.

For the case in which the loss function ℓ in (5) is nonseparable (Definition 2), the objective function in (9a) is a constant and thus the problem is ill-posed.

In a nutshell, choosing $\beta = \frac{a_1}{\gamma}$ and denoting the solution $P^\star$ as $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$, it holds that $g^\star$ in (150) can be written as $\frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}$, and thus, for all $(x,y) \in \operatorname{supp} P_S$,

$$\frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y)=\exp\left(\frac{1}{\beta}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right), \tag{175}$$

where β is such that $D\left(P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right) = \gamma$. This completes the proof.

B Proof of Lemma 1

Proof: Taking into account that β in (9) is chosen from $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (13), it follows that

$$\log\left(\int \exp\left(\frac{1}{\beta}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)\right) < +\infty, \tag{176}$$

which implies that

$$\int \exp\left(\frac{1}{\beta}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y) < +\infty. \tag{177}$$

From [37, Theorem 1.6.6], the inequality in (177) implies that

$$P_S\left(\left\{(x,y) \in \operatorname{supp} P_S : \exp\left(\frac{1}{\beta}\ell(\boldsymbol{\theta},x,y)\right) = +\infty \right\}\right) = 0, \tag{178}$$

which is equivalent to

$$P_S\left(\left\{(x,y) \in \operatorname{supp} P_S : \ell(\boldsymbol{\theta},x,y) = +\infty \right\}\right) = 0 \tag{179}$$

from the fact that $\beta \in (0,+\infty)$. This completes the proof. ∎

C Proof of Lemma 2

For all $\mathcal{C} \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$,

$$P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{C}) = \int_{\mathcal{C}} \frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y), \tag{180}$$

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

$$P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{C}) = 0, \tag{181}$$

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

$$0= P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{C}) = \int_{\mathcal{C}} \frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y). \tag{182}$$

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

$$\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y) = \exp\left(\frac{\ell(\boldsymbol{\theta},x,y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right).$$

Note that if a solution to the optimization problem (9) exists, then $\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) < +\infty$. Thus, $\exp\left(-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right) > 0$. Moreover, $\exp\left(\frac{\ell(\boldsymbol{\theta},x,y)}{\beta}\right) > 0$, where the strict inequality is due to the fact that for all $(x, y) \in \operatorname{supp} P_S$, the function ℓ in (5) is nonnegative. Hence, for all $(x, y) \in \operatorname{supp} P_S$,

$$\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y) = \exp\left(\frac{\ell(\boldsymbol{\theta},x,y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right) > 0,$$

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

D Proof of Lemma 3

The equality in (19) follows from observing that:

$$\begin{aligned}
D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) &= \int \log\left(\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y)\right) \mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) && (183)\\
&= \int \log\left(\exp\left(\frac{\ell(\boldsymbol{\theta},x,y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\right) \mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) && (184)\\
&= \int \frac{\ell(\boldsymbol{\theta},x,y)}{\beta} \mathrm{d}P_S(x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), && (185)
\end{aligned}$$

where equality (184) follows from (11). The equality in (20) is proved as follows:

$$\begin{aligned}
D\left(P_S \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) &= \int \log\left(\frac{\mathrm{d}P_S}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\right) \mathrm{d}P_S(x,y) && (186)\\
&= \int \log\left(\exp\left(-\frac{\ell(\boldsymbol{\theta},x,y)}{\beta} + \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\right) \mathrm{d}P_S(x,y) && (187)\\
&= -\int \frac{\ell(\boldsymbol{\theta},x,y)}{\beta} \mathrm{d}P_S(x,y) + \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), && (188)
\end{aligned}$$

where equality in (187) follows from equation (11). This completes the proof.

E Proof of Theorem 2

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

$$D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right)=\int\log\left(\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) \tag{189}$$

$$=\int\left(\frac{\ell(\boldsymbol{\theta},x,y)}{\beta}-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) \tag{190}$$

$$=\int\frac{\ell(\boldsymbol{\theta},x,y)}{\beta}\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), \tag{191}$$

where the equality in (190) follows from (11). The proof continues by noticing that the relative entropies $D\left(P\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right)$ and $D\left(P\|P_S\right)$ satisfy:

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

$$=\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\right)\mathrm{d}P(x,y)-\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P(x,y) \tag{192}$$

$$=\int\left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\right)-\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\right)\right)\mathrm{d}P(x,y \tag{193}$$

$$=\int\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\frac{\mathrm{d}P_S}{\mathrm{d}P}(x,y)\right)\mathrm{d}P(x,y) \tag{194}$$

$$=\int\log\left(\frac{\mathrm{d}P_S}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\right)\mathrm{d}P(x,y) \tag{195}$$

$$=\int\log\left(\exp\left(-\frac{\ell(\boldsymbol{\theta},x,y)}{\beta}+\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\right)\mathrm{d}P(x,y) \tag{196}$$

$$=\int\left(-\frac{\ell(\boldsymbol{\theta},x,y)}{\beta}+\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\mathrm{d}P(x,y) \tag{197}$$

$$=-\int\frac{\ell(\boldsymbol{\theta},x,y)}{\beta}\mathrm{d}P(x,y)+\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right). \tag{198}$$

Therefore, from (191) and (198), it follows that

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

$$=\frac{1}{\beta}\int\ell(\boldsymbol{\theta},x,y)\mathrm{d}P(x,y)-\frac{1}{\beta}\int\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) \tag{199}$$

$$=\frac{1}{\beta}G\left(\boldsymbol{\theta},P,P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right), \tag{200}$$

which completes the proof.

F Proof of Theorem 3

The proof follows from the following equalities:

$$
\begin{aligned}
& G(\boldsymbol{\theta}, P_1, P_2) \\
&= G(\boldsymbol{\theta}, P_1, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}) - G(\boldsymbol{\theta}, P_2, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}) && (201)\\
&= \beta\Big(D\left(P_1\|P_S\right) - D\left(P_1\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right)\Big) \\
&\quad -\beta\Big(D\left(P_2\|P_S\right) - D\left(P_2\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right) && (202)\\
&= \beta\Big(D\left(P_2\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_1\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_2\|P_S\right) + D\left(P_1\|P_S\right)\Big), && (203)
\end{aligned}
$$

where equality in (202) follows from (22). This completes the proof.

G Proof for Lemma 4

The proof follows from Definition 1 and the following equalities:

$$
\begin{aligned}
\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) &= \frac{1}{n}\sum_{t=1}^{n} \ell\left(f(\boldsymbol{\theta}, x_t), y_t\right) && (204)\\
&= \frac{1}{n}\sum_{(x,y)\in(\mathcal{X}\times\mathcal{Y})}\sum_{t=1}^{n} \mathbb{1}_{\{x=x_t, y=y_t\}}\ell(\boldsymbol{\theta}, x, y) && (205)\\
&= \sum_{(x,y)\in(\mathcal{X}\times\mathcal{Y})} \ell(\boldsymbol{\theta}, x, y) P_{\boldsymbol{z}}(x, y) && (206)\\
&= \int \ell(\boldsymbol{\theta}, x, y)\mathrm{d}P_{\boldsymbol{z}}(x, y), && (207)
\end{aligned}
$$

which completes the proof.

H Proof of Lemma 6

The proof follows from the following equalites:

$$
\begin{aligned}
&\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z) \\
&= \int \int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\, \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) && (208)\\
&= \int \left(\int \left(\int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_Z(x,y) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \\
&- \int \left(\int \left(\int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{\boldsymbol{z}}(x,y) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) && (209)\\
&= \int \left(\int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_Z(x,y) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \\
&- \int \left(\int \left(\int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{\boldsymbol{z}}(x,y) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) && (210)\\
&= \int \left(\int \frac{1}{n} \sum_{t=1}^{n} \ell(f(\boldsymbol{\theta}, x_t), y_t) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \\
&- \int \left(\int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) && (211)\\
&= \int \left(\int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) \right) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \\
&- \int \left(\int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) && (212)\\
&= \int \left(\int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) - \int \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}) && (213)\\
&= \lambda \Big(L\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) + I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \Big). && (214)
\end{aligned}
$$

where equality in (209) follows from (32); the equality in (211) follows with the function L in (6) and $P_{\boldsymbol{Z}} \in \triangle(\mathcal{X} \times \mathcal{Y})^n$ being a product measure obtained from P_Z; and equality in (214) follows from [7, Theorem 10.4]. This completes the proof.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Y. Chu and M. Raginsky, "A unified framework for information-theoretic generalization bounds," arXiv preprint arXiv:2305.11042, May 2023.

[6] 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, pp. 328–333.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] A. R. Esposito, M. Gastpar, and I. Issa, "Robust generalization via α-mutual information," arXiv preprint arXiv:2001.06399, Jan. 2020.

[18] 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.

[19] 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.

[20] J. C. Duchi, P. W. Glynn, and H. Namkoong, "Statistics of robust optimization: A generalized empirical likelihood approach," *Mathematics of Operations Research*, vol. 46, no. 3, pp. 946–969, Aug. 2021.

[21] 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.

[22] J. Lee and M. Raginsky, "Minimax statistical learning with wasserstein distances," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[23] 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.

[24] 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.

[25] D. A. McAllester, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, Apr. 2003.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] I. Csiszár, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[31] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2014.

[32] 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, pp. 340–345.

[33] H.-O. Georgii, *Gibbs measures and phase transitions*, 2nd ed. New York, NY, USA: De Gruyter, 2011.

[34] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.

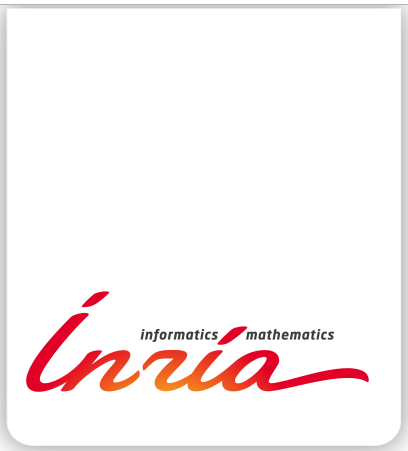
[35] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, Sep. 1946.

[36] 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.

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

[38] W. Feller, *An Introduction to Probability Theory and Its Applications II*, 2nd ed. New York, NY, USA: Jhon Wiley & Sons, 1971.

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



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