# The Worst-Case Data-Generating Probability Measure

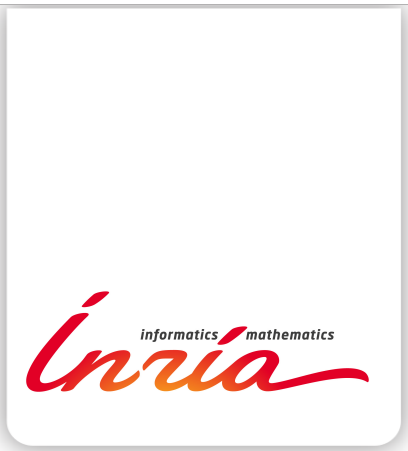
Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman


# 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 — August 2023 — 26 pages

**Abstract:** In this report, 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 solution to the maximization of the expected loss under a relative entropy constraint with respect to a reference $\sigma$-finite measure. Given a model, the central result consists of an explicit expression for the difference between the expectations of the loss with respect to any two given probability measures over the datasets. Such a difference is characterized in terms of "statistical distances" measured via KL-divergences involving the given measures; the reference measure; and the worst-case probability measure. When the given measures are the types (empirical probability measures) induced by two datasets, a closed-form expression for the difference between the corresponding empirical risks is obtained. Finally, the generalization gap induced by any arbitrary machine learning algorithm is characterized. Existing results for the Gibbs algorithm, such as the equality between the generalization gap and a sum of mutual information and lautum information, up to a constant factor, are recovered. All the above suggests a duality between the Gibbs algorithm and the worst-case measure beyond the fact that both are represented by Gibbs probability measures.

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

Xinying Zou, Samir M. Perlaza and Eitan are with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France. Iñaki Esnaola is with the ACSE Dept. at The University of Sheffield, Sheffield S1 3JD, UK. Eitan Altman is also with the Laboratoire d'Informatique d'Avignon (LIA), Université d'Avignon, France. Samir M. Perlaza and Iñaki Esnaola are also with the ECE Dept. at Princeton University, Princeton N.J. 08544, USA. Samir M. Perlaza is also with the GAATI Laboratory at the Université de la Polynésie Française, Faaa 98702, French Polynesia.

**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 $\sigma$-finie 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 [1–4] and [5]. 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[1] induced by the model and the empirical risk with respect to the training dataset. When the choice of model 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 [1]; and for the case in which the reference measure is a $\sigma$-finite measure [7]. In the case of other algorithms, the EGE is characterized by various upper-bounds leveraging different techniques. The metric of mutual information is first proposed in [8], further developed in [3] and combined with chaining methods in [9, 10] for deriving upper bounds on the EGE. Similar bounds on EGE are obtained in [4, 11–13] and references therein. Other information measures such as the Wasserstein distance [2,14,15], maximal leakage [16,17], mutual $f$-information [18], and Jensen-Shannon divergence [19] are used for providing upper bounds on EGE as well. 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 [20]; probability theory [21–23]; and information theory [1, 3, 24, 25]. 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 [5, 26]. This line of work opens the door to the study of the data-generating probability measures and their effect on the GE and EGE, as shown in the following section.

## 1.1 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 $\sigma$-finite measure is not bigger than 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

[1] Population risk, for a fixed model, refers to the expectation of the loss function, with respect to the ground-truth probability measure of the data [6]

that, if the worst-case probability measure exists, then it is a Gibbs probability measure (Theorem 3.1) 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 has been thoroughly 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 4.2). This is a transcendental result as the reference measure and the "statistical distance" threshold can be arbitrarily chosen, which leads to numerous closed-form expression for such a variation.

The second contribution leverages the observation that under the assumption that datasets are tuples of independent and identically distributed datapoints, datasets can be represented by their corresponding types, which are probability measures [27]. 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 5.1). This observation allows using Theorem 4.2 to provide an explicit expression to the difference between two empirical risks induced by the same model on two different datasets. This difference is referred to as the *sensitivity* of the empirical risk. Using the same arguments, closed-form expressions 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 of any machine learning algorithm are strongly connected with the notion of worst-case data-generating probability measure. As a byproduct, an alternative proof to the existing result (see [1] 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.

## 1.2 Notation

Given a measurable space $(\Omega, \mathscr{F})$, the notation $\triangle(\Omega)$ is used to represent the set of $\sigma$-finite measures that can be defined over $(\Omega, \mathscr{F})$. Often, when the sigma-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 $\sigma$-finite 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 $\sigma$-finite measures defined over $(\Omega, \mathscr{F})$ conditioned on an element of $\mathcal{X}$. Given two $\sigma$-finite 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

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

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:

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

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

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

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

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 $\mathsf{L} : \left(\mathcal{X} \times \mathcal{Y}\right)^n \times \mathcal{M} \to [0, +\infty]$, which satisfies

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

where the functions $f$ and $\ell$ are defined in (3) and (4).

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

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \mathsf{L}\left(\boldsymbol{z}, \boldsymbol{\theta}\right). \tag{6}$$

The underlying assumption in this work is described in terms of the following measurable spaces: $(\mathcal{X} \times \mathcal{Y}, \mathscr{F}_{\mathcal{X}\times\mathcal{Y}})$ and $(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right))$. The function $g : \mathcal{X} \times \mathcal{Y} \times \mathcal{M}$ such that $g(x, y, \boldsymbol{\theta}) = \ell\left(f(\boldsymbol{\theta}, x), y\right)$, with the functions $f$ and $\ell$ defined in (3) and (4), is measurable with respect to the product measurable space $(\mathcal{X} \times \mathcal{Y}, \mathscr{F}_{\mathcal{X}\times\mathcal{Y}}) \times (\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right))$ and the Borel measurable space $(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right))$.

# 3 An Auxiliary Optimization Problem

This section introduces an optimization problem whose solution is referred to as the worst-case probability measure. Such a probability measure, which is conditioned on a given model $\boldsymbol{\theta} \in \mathcal{M}$, is parametrized by a $\sigma$-finite measure $P_S \in \triangle(\mathcal{X} \times \mathcal{Y})$ and by a positive real $\gamma$. In a nutshell, the worst-case probability measure maximizes the expected loss while its relative entropy with respect to $P_S$ is not bigger than $\gamma$. Using this notation, the optimization problem of interest is:

$$\max_{P \in \Delta_{P_S}(\mathcal{X}\times\mathcal{Y})} \int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P(x, y) \tag{7a}$$

$$\text{s. t.} \quad D(P\|P_S) \leq \gamma \tag{7b}$$

$$\int \mathrm{d}P(x, y) = 1, \tag{7c}$$

where the functions $f$ and $\ell$ are defined in (3) and (4).

When the $\sigma$-finite measure $P_S$ in (7) is a probability measure, it can be interpreted as a prior on the probability distribution of the datasets. When, such a measure is not a probability measure, but a finite measure, it might model the case in which the probability distribution is known up to a normalization factor. When it is $\sigma$-finite measure, for instance the counting measure or the Lebesgue measure, the constraint in (7b) becomes respectively, a constraint on the discrete entropy and the differential entropy introduced by [28]. This particular case has been extensively studied in the realm of optimization theory [29].

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 $\gamma$ 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 are made clearer by studying the properties of the solution to the optimization problem in (7). 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 [30].

## 3.1 The Solution

The following theorem characterizes the solution to the optimization problem in (7) using the function $\mathsf{J}_{P_S,\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$ that satisfies

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(f(\boldsymbol{\theta}, x), y)\right) \mathrm{d}P_S(x, y)\right), \tag{8}$$

where the function $f$ and $\ell$ are in (3) and (4) respectively.

**Theorem 3.1.** *The solution to the optimization problem in* (7)*, if it exists, is denoted by* $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ *and satisfies for all* $(x, y) \in \operatorname{supp} P_S$*,*

$$\frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x, y) = \exp\left(\frac{\ell(f(\boldsymbol{\theta}, x), y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right), \tag{9}$$

*where the function* $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ *is in* (8)*, and* $\beta$ *is chosen to satisfy:*

$$D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) = \gamma. \tag{10}$$

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

Theorem 3.1 does not provide any guarantee on the existence or uniqueness of the solution to the optimization problem in (7). While $\gamma = 0$ leads to a trivial problem in (7) whose solution is unique and identical to $P_S$, a unique solution is also observed under other choices. For instance, if $P_S$ is a Gibbs probability measure, the problem in (7) always possesses a unique solution [7, Theorem 3.2]. Nonetheless, such aspects are left out of the scope of this paper. In the following, it is assumed that the model $\boldsymbol{\theta}$, the real $\gamma$, and the $\sigma$-finite measure $P_S$ in (7) are such that a solution exists. Under this assumption, it is important to highlight the following. Let the set $\mathcal{J}_{P_S,\boldsymbol{\theta}} \subset (0, +\infty)$ be:

$$\mathcal{J}_{P_S,\boldsymbol{\theta}} \triangleq \left\{ t \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) < +\infty \right\}. \tag{11}$$

Hence, a necessary condition for the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ to be a solution to the optimization problem in (7) is that $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$.

Finally, note that if a solution exists, the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (9) is a Gibbs probability measure [31]. From this perspective, the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (8) is often referred to as the log-partition function [32].

## 3.2 Mutual Absolute Continuity

The solution to the optimization problem in (7) exhibits several properties among which the mutual absolute continuity with respect to the measure $P_S$.

**Lemma 3.1.** *The probability measures* $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ *and* $P_S$ *in* (9) *are mutually absolutely continuous.*

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

An immediate consequence of the mutual absolute continuity between the measures $P_S$ and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (9) is portraited by the following lemma.

**Lemma 3.2.** *The probability measures* $P_S$ *and* $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ *in* (9) *satisfy:*

$$\begin{aligned}
&\beta \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) \\
&= \int \ell(f(\boldsymbol{\theta},x),y) \mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) - \beta D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) && (12)\\
&= \int \ell(f(\boldsymbol{\theta},x),y) \mathrm{d}P_S(x,y) + \beta D\left(P_S \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right), && (13)
\end{aligned}$$

*where the functions* $f$ *and* $\ell$ *are defined in* (3) *and* (4)*, respectively; and the function* $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ *is in* (8)*.*

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

The equality in (12) can be further simplified by noticing that $\beta$ is chosen to satisfy (10).

# 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

$$G(\boldsymbol{\theta}, P_1, P_2) = \int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_1(x, y) - \int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_2(x, y), \quad (14)$$

where the functions $f$ and $\ell$ are defined in (3) and (4), respectively. The value $G(\boldsymbol{\theta}, P_1, P_2)$ represents the variation of the expectation of the loss when the probability measure over the datasets changes from $P_2$ to $P_1$. Such a value is often referred to as the *sensitivity* of the expected loss and 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 (9) to an alternative measure.

**Theorem 4.1** (Sensitivity of the Expected Loss)**.** *For all $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ and for all $\boldsymbol{\theta} \in \mathcal{M}$,*

$$G\left(\boldsymbol{\theta}, P, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) = \beta\left(D\left(P \| P_S\right) - D\left(P \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)\right), \quad (15)$$

*where the function $G$ is in* (14)*; and the model $\boldsymbol{\theta}$ and the measures $P_S$ and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ satisfy* (9)*.*

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

The following corollary of Theorem 4.1 highlights 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 4.1.** *The probability measures $P_S$ and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (9) *satisfy*

$$G(\boldsymbol{\theta}, P_S, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}) = -\beta\left(D\left(P_S \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) + D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)\right), \quad (16)$$

*where the function $G$ is in* (14)*.*

The right-hand side of the equality in (16) is a symmetrized Kullback-Liebler divergence, also known as Jeffrey's divergence [33], between the measures $P_S$ and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$. More importantly, when $P_S$ is a probability measure, it follows from [7, Theorem 2.1] 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 bigger than or equal to the expected loss induced by the reference measure $P_S$. This is formalized by the following corollary of Theorem 4.1.

**Corollary 4.2.** *If $P_S$ in* (7) *is a probability measure, the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (9) *satisfies:*

$$\int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x, y) \geq \int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_S(x, y). \quad (17)$$

Note that the probability measure $P_S$ in Corollary 4.2 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 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 4.1, it is possible to obtain the sensitivity of the expected loss when the measure changes from a given probability distribution to any alternative probability distribution, as shown by the following theorem.

**Theorem 4.2.** *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 function $G$ in* (14) *satisfies*

$$G(\boldsymbol{\theta},P_1,P_2)=\beta\Big(D\left(P_2\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right)-D\left(P_1\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) -D\left(P_2\|P_S\right)+D\left(P_1\|P_S\right)\Big), \tag{18}$$

*where the model $\boldsymbol{\theta}$ and the measures $P_S$ and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ satisfy* (9).

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

Note that the parameters $\gamma$ and $P_S$ in (7) can be arbitrarily chosen. This is essentially because only the right-hand side of (75) depends on $P_S$ and $\beta$. Another interesting observation is that none of the terms in the right-hand side of (75) depends simultaneously on both $P_1$ and $P_2$. Interestingly, these terms dependent exclusively on the pair formed by $P_i$ and $P_S$, with $i\in\{1,2\}$. These observations highlight the significant flexibility of the expression in (75) to construct closed-form expressions for the sensitivity $G(\boldsymbol{\theta},P_1,P_2)$ in (14). The only constraint on the choice of $P_S$ is that both measures $P_1$ and $P_2$ must be absolutely continuous with respect to $P_S$. The following corollaries follow by adopting particular choices for both $P_S$.

## 4.1 Choice of Probability Measures

Two choices of $P_S$ for which the expression in the right-hand side of (75) significantly simplifies are $P_S=P_1$ and $P_S=P_2$, which leads to the following corollary of Theorem 4.2.

**Corollary 4.3.** *If $P_1$ is absolutely continuous with $P_2$, then the term $G(\boldsymbol{\theta},P_1,P_2)$ in* (14) *satisfies:*

$$G(\boldsymbol{\theta},P_1,P_2)=\beta\Big(D\left(P_2\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_2,\beta)}\right)-D\left(P_1\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_2,\beta)}\right)+D\left(P_1\|P_2\right)\Big). \tag{19}$$

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

$$G(\boldsymbol{\theta},P_1,P_2)=\beta\Big(D\left(P_2\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_1,\beta)}\right)-D\left(P_1\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_1,\beta)}\right)-D\left(P_2\|P_1\right), \tag{20}$$

*where for all $i\in\{1,2\}$, the probability measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_i,\beta)}$ satisfies* (9) *under the assumption that $P_S=P_i$.*

Interestingly, absolute continuity between $P_1$ with respect to $P_2$ or between $P_2$ with respect to $P_1$ is not necessary for obtaining an expression for the value $G(\boldsymbol{\theta}, P_1, P_2)$ in (14). Note by choosing $P_S$ as a convex combination of $P_1$ and $P_2$, always guarantees an explicit expression for $G(\boldsymbol{\theta}, P_1, P_2)$ independently of whether these measures are absolutely continuous with respect to each other.

## 4.2 Choice of the Counting Measure

This subsection adopts the assumption that the set $\mathcal{X} \times \mathcal{Y}$ is countable. Under this assumption, any probabilty measure $P \in \Delta(\mathcal{X} \times \mathcal{Y})$ is absolutely continuous with respect to the counting measure $\nu$. Moreover, the Radon-Nikodym derivative of $P$ with respect to $\nu$ is the probability mass function of $P$, denoted by $p : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. In this case,

$$D(P\|\nu) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x, y) \log(p(x, y)) \triangleq -H(P), \tag{21}$$

where the function $H$ is Shannon's discrete entropy function [28].

**Corollary 4.4.** *Let the set $\mathcal{X}\times\mathcal{Y}$ be countable and let $\nu$ be the counting measure on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X}\times\mathcal{Y}})$. The function $G$ in* (18) *satisfies*

$$G(\boldsymbol{\theta}, P_1, P_2) = \beta\Big(D\left(P_2\|P^{(\nu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_1\|P^{(\nu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) + H(p_2) - H(p_1)\Big), \quad \textit{(22)}$$

*where the probability measure $P^{(\nu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfies* (9) *under the assumption that $P_S = \nu$; and the terms $H(P_1)$ and $H(P_2)$ represent Shannon's discrete entropy of the probability measures $P_1$ and $P_2$, respectively.*

## 4.3 Choice of the Lebesgue Measure

This subsection adopts the following assumptions: $(a)$ the set $\mathcal{X} \times \mathcal{Y}$ is a Borel set and $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ is the corresponding Borel $\sigma$-field; and $(b)$ the measure $P_S$ is the Lebesgue measure $\mu$ on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X}\times\mathcal{Y}})$. Under these assumptions, any probability measure $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ posseses a probability density function denoted by $p : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. In this case,

$$D(P\|\mu) = \int p(t) \log(p(t)) \mathrm{d}\mu(t) \triangleq -h(P), \tag{23}$$

where the function $h$ is Shannon's differential entropy function.

**Corollary 4.5.** *Let $\mathcal{X} \times \mathcal{Y}$ be a Borel set and $\mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$ be the corresponding Borel $\sigma$-field. The function $G$ in* (18) *satisfies*

$$G(\boldsymbol{\theta}, P_1, P_2) = \beta\Big(D\left(P_2\|P^{(\mu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_1\|P^{(\mu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) + h(P_2) - h(P_1)\Big), \quad \textit{(24)}$$

*where $\mu$ is a Lebesgue measure; the probability measure $P^{(\mu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfies* (9) *under the assumption that $P_S = \mu$; and the terms $h(P_1)$ and $h(P_2)$ represent Shannon's differential entropy of the probability measures $P_1$ and $P_2$ respectively.*

# 5 Analysis of the Empirical-Risk

This section presents a mathematical object known as a *type* in the realm of information theory [27]. In the context of this work, a type is a probability measure induced by a dataset, as shown hereunder.

**Definition 5.1** (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 $\{(x,y)\}\in\mathscr{F}_{\mathcal{X}\times\mathcal{Y}}$,*

$$P_{\boldsymbol{z}}\left(\{(x,y)\}\right)=\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}_{\{x=x_t,y=y_t\}}(x,y). \tag{25}$$

In the following, the abuse of noting $P_{\boldsymbol{z}}\left(\{(x,y)\}\right)$ as $P_{\boldsymbol{z}}\left(x,y\right)$ is allowed for the ease of presentation.

The central observation of this section is that the empirical risk $\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})$ in (5) 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 5.1** (Empirical Risks and Types)**.** *The empirical risk $\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})$ in* (5) *satisfies*

$$\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})=\int\ell(f(\boldsymbol{\theta},x),y)\mathrm{d}P_{\boldsymbol{z}}(x,y), \tag{26}$$

*where the measure $P_{\boldsymbol{z}}$ is the type induced by the dataset $\boldsymbol{z}$ in* (2)*; and the functions $f$ and $\ell$ are defined in* (3) *and* (4)*, respectively.*

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

Equipped with the result in Lemma 5.1, 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

$$G(\boldsymbol{\theta},P_{\boldsymbol{z}_1},P_{\boldsymbol{z}_2})=\mathsf{L}(\boldsymbol{z}_1,\boldsymbol{\theta})-\mathsf{L}(\boldsymbol{z}_2,\boldsymbol{\theta}), \tag{27}$$

where the function $G$ is in (14). Assume that $P_{\boldsymbol{z}_1}$ and $P_{\boldsymbol{z}_2}$ are absolutely continuous with respect to the reference measure $P_S$ in (7). Under this assumption, the equality in (27) 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 5.1.** *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* (7)*, for all $\boldsymbol{\theta}\in\mathcal{M}$, the following holds:*

$$\begin{aligned}\mathsf{L}(\boldsymbol{z}_1,\boldsymbol{\theta})-\mathsf{L}(\boldsymbol{z}_2,\boldsymbol{\theta})=\beta\Big(&D\left(P_{\boldsymbol{z}_2}\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)-D\left(P_{\boldsymbol{z}_1}\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\\&-D\left(P_{\boldsymbol{z}_2}\|P_S\right)+D\left(P_{\boldsymbol{z}_1}\|P_S\right)\Big),\end{aligned} \tag{28}$$

*where the function $\mathsf{L}$ is in* (5)*; the model $\boldsymbol{\theta}\in\mathcal{M}$, and the measures $P_S$ and $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfy* (9)*.*

*Proof:* The proof follows from the equality in (27), which together with Theorem 4.2 completes the proof. ■

In Theorem 5.1, 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 [5].

Another interesting choice is the counting measure, whose formalization follows from Corollary 4.4 and is described by the following corollary .

**Corollary 5.1.** *Let the set $\mathcal{X} \times \mathcal{Y}$ be countable and let $\nu$ be the counting measure on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{X}\times\mathcal{Y}})$. Then, the difference $\mathsf{L}(\boldsymbol{z}_1, \boldsymbol{\theta}) - \mathsf{L}(\boldsymbol{z}_2, \boldsymbol{\theta})$ in* (28) *satisfies:*

$$\begin{aligned}\mathsf{L}(\boldsymbol{z}_1, \boldsymbol{\theta}) - \mathsf{L}(\boldsymbol{z}_2, \boldsymbol{\theta}) = \beta\Big(& H(P_{\boldsymbol{z}_2}) - H(P_{\boldsymbol{z}_1}) \\ & + D\left(P_{\boldsymbol{z}_2} \| P^{(\nu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_{\boldsymbol{z}_1} \| P^{(\nu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\Big), \end{aligned} \tag{29}$$

*where the measure $P^{(\nu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfies for all singletons $\{(x, y)\} \in \mathcal{F}_{\mathcal{X}\times\mathcal{Y}}$,*

$$P^{(\nu,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\{(x, y)\}) = \frac{\exp\left(\frac{1}{\beta}\ell(f(\boldsymbol{\theta}, x), y)\right)}{\displaystyle\sum_{(u,v)\in\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{1}{\beta}\ell(f(\boldsymbol{\theta}, u), v)\right)}; \tag{30}$$

*and the terms $H(P_{\boldsymbol{z}_1})$ and $H(P_{\boldsymbol{z}_2})$ represent Shannon's discrete entropy of the probability measures $P_{\boldsymbol{z}_1}$ and $P_{\boldsymbol{z}_2}$, respectively.*

Other choices, such as the Lebesgue measure, whose formalization follows from Corollary 4.5, lead to similar results.

From Theorem 5.1, 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 $\boldsymbol{z}$ is determined by two values: $(a)$ the difference of the "statistical distance" from the types induced by the training and test datasets to the worst-case data-generating probability measure, i.e., $D\left(P_{\boldsymbol{z}_2} \| P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_{\boldsymbol{z}_1} \| P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$; and $(b)$ the difference of the "statistical distance" from the types to the reference measure $P_S$, i.e., $D\left(P_{\boldsymbol{z}_1} \| P_S\right) - D\left(P_{\boldsymbol{z}_2} \| P_S\right)$. When each of these values is bounded by a given $\epsilon$, the difference between the test and training empirical risk, i.e., $\mathsf{L}(\boldsymbol{z}_1, \boldsymbol{\theta}) - \mathsf{L}(\boldsymbol{z}_2, \boldsymbol{\theta})$, is upperbounded by $2\beta\epsilon$, with $\beta$ being determined by the "statistical distance" threshold $\gamma$ in (7).

# 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 thanks to 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

$$G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_Z(x, y) - \int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_{\boldsymbol{z}}(x, y). \tag{31}$$

The term $\int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_{\boldsymbol{z}}(x, y) = \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})$ is an empirical risk often referred to as the training risk, training loss, or training error [6]. This is essentially the loss induced by the model with respect to the same dataset that has been used for obtaining (training) such a model. The term $\int \ell(f(\boldsymbol{\theta}, x), y) \mathrm{d}P_Z(x, y)$ is the expected loss under the assumption that the ground-truth distribution of the data is $P_Z$. Interestingly, as shown in (31), such generalization error can be written in terms of the function $G$ in (14). This observation leads to the following descriptions of the generalization gap.

**Lemma 6.1.** *The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (31) satisfies:*

$$G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \beta\Big(D\Big(P_{\boldsymbol{z}} \| P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big) - D(P_{\boldsymbol{z}} \| P_Z) - D\Big(P_Z \| P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)\Big), \tag{32}$$

*where the measure $P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is the solution to the optimization problem in (7) under the assumption that $P_S = P_Z$.*

*Proof:* The proof follows from Corollary 4.3 by noticing that the type $P_{\boldsymbol{z}}$ is absolutely continuous with respect to $P_Z$. ∎

Lemma 6.1 highlights the intuition that if the type $P_{\boldsymbol{z}}$ induced by the training dataset $\boldsymbol{z}$ is at arbitrary small "statistical distance" of the ground-truth measure $P_Z$, the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (31) is arbitrarily close to zero. This is revealed by the facts that $D(P_{\boldsymbol{z}} \| P_Z)$ would be arbitrarily small; and so would be the difference $D\left(P_{\boldsymbol{z}} \| P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_Z \| P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$.

A more general expression for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (31) is provided by the following corollary of Theorem 4.2.

**Corollary 6.1.** *The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (31) satisfies:*

$$\begin{aligned} G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \beta\Big(&D\left(P_{\boldsymbol{z}} \| P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_Z \| P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) \\ &- D(P_{\boldsymbol{z}} \| P_S) + D(P_Z \| P_S)\Big), \end{aligned} \tag{33}$$

*where the measure $P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is in (7).*

Note that several expressions for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (31) can be obtained from Corollary 6.1 by choosing the reference $P_S$ and the parameter $\gamma$ in (7), which determines the value of $\beta$.

## 6.1 Expected Generalization Gap

A conditioned probability distrubuition $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, 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 (31) 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

$$\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}}) = \int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \tag{34}$$

where the function $G$ is in (31). 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 (34). Corollary 6.1, by strategically choosing the reference measure $P_S$ and the parameter $\gamma$ in (7), leads to numerous closed-form expressions for the expected generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, when the training dataset is $\boldsymbol{z}$. Interestingly, regardless of the choice of $P_S$ and $\gamma$, the resulting expressions highlight 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 (34) 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 a product distribution formed by $P_Z$. Let $\overline{\overline{G}} : \triangle(\mathcal{M}, \mathscr{B}(\mathcal{M}) \,|\, (\mathcal{X} \times \mathcal{Y})^n \times \triangle(\mathcal{X} \times \mathcal{Y})^n) \to \mathbb{R}$ be a function such that

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

where the function $G$ is in (31). Using this notation, the doubly-expected generalization error induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z)$ in (35). In existing literature, the doubly-expected generalization gap is simply referred to as generalization gap. See for instance [3], [1], and [7]. This is due to the fact that in existing literature, the central role of the training dataset is often faded away by taking expectations. As in the case of the expected generalization gap, Corollary 6.1 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 example of a statistical learning algorithm is the Gibbs algorithm, which is parametrized by a positive real $\lambda$ and by a $\sigma$-finite 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 $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\right), \tag{36}$$

where the dataset $\boldsymbol{z}$ represents the training dataset; and the function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ satisfies$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp\left(t\, \mathsf{L}(\boldsymbol{z},\boldsymbol{\nu})\right) \mathrm{d}Q(\boldsymbol{\nu})\right)$.

The doubly-expected generalization error induced by the Gibbs algorithm with parameters $Q$ and $\lambda$, 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 6.2** (Generalization Gap of the Gibbs Algorithm)**.** *Given the conditional probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in* (36) *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*

$$\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z) = \lambda\left(I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) + L\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)\right), \tag{37}$$

*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, a mutual information and a lautum information:*

$$I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \text{ and} \tag{38}$$

$$L\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\right) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), \tag{39}$$

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

*Proof:* This proof is presented in Appendix G. ∎

Lemma 6.2 has been proved before in the case in which $Q$ is a probability measure in [1]; and in the more general case in which $Q$ is a $\sigma$-finite measure in [7]. In both [1] 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 6.2. This limitation is due to the fact that the proof of Lemma 6.2 relies on the notion of types, which is known to fail capturing the correlation between datapoints, as pointed in [27]. Nonetheless, the assumption of independent and identically distributed is widely adopted in the realm of machine learning. Despite this limitation, the relevance of Lemma 6.2 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 function $G$ in (14), and the notion of (doubly-expected) generalization gap, which is one of the main performance metrics for evaluating the generalization capabilities of machine learning algorithms.

# 7 Conclusions and Final Remarks

The worst-case data-generating probability measure in Theorem 3.1 has been shown to be a cornerstone in statistical machine learning. This is essentially 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, Theorem 10.4]. These dependences appear intriguing and suggest a 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. Nonetheless, such connection appears to be nontrivial and is suggested as a promising line of work in this area.

# Appendices

## A Proof of Theorem 3.1

The optimization problem in (7) can be written as follows:

$$\max_{P\in\Delta_{P_S}(\mathcal{X}\times\mathcal{Y})} \int \ell(f(\theta,x),y)\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y) \tag{40a}$$

$$\text{s. t.} \int \frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P_S(x,y) \leq \gamma \tag{40b}$$

$$\int \frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y) = 1. \tag{40c}$$

Let $\mathscr{S}$ be the set of nonnegative measurable functions with respect to the measurable spaces $(\mathcal{X}\times\mathcal{Y},\mathcal{F}_{\mathcal{X}\times\mathcal{Y}})$ and $(\mathbb{R},\mathscr{B}(\mathbb{R}))$. The Lagrangian of the optimization problem in (40) can be constructed in terms of a function in $\mathscr{S}$, instead of a measure in $\Delta_{P_S}(\mathcal{X}\times\mathcal{Y})$. Let such a Lagrangian be $L:\mathscr{S}\times[0,+\infty)^2\to\mathbb{R}$ such that

$$\begin{aligned} L\left(\frac{\mathrm{d}P}{\mathrm{d}P_S},\beta,\lambda\right) &= \int \ell(f(\theta,x),y)\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y) \\ &-\beta\left(\int\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P_S(x,y)-\gamma\right)+\lambda\left(\int\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y)-1\right), \end{aligned} \tag{41}$$

where $\beta$ and $\lambda$ are nonnegative reals that act as Lagrangian multipliers due to the constraints in (40b) and (40c), respectively.

Let $g:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ be a function in $\mathscr{S}$. The Gateaux differential of the function $L$ in (41) at $\left(\frac{\mathrm{d}P}{\mathrm{d}P_S},\beta,\lambda\right)\in\mathscr{S}\times[0,+\infty)^2$ in the direction of $g$ is

$$\partial L\left(\frac{\mathrm{d}P}{\mathrm{d}P_S},\beta,\lambda;g\right) \triangleq \frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha)\bigg|_{\alpha=0}, \tag{42}$$

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

$$\begin{aligned} r(\alpha) &= \int \ell(f(\theta,x),y)\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)+\alpha g(x,y)\right)\mathrm{d}P_S(x,y) \\ &-\beta\int\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)+\alpha g(x,y)\right)\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)+\alpha g(x,y)\right)\mathrm{d}P_S(x,y)-\beta\gamma \\ &+\lambda\left(\int\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)+\alpha g(x,y)\right)\mathrm{d}P_S(x,y)-1\right). \end{aligned} \tag{43}$$

Note that the derivative of the real function $r$ is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha}r(\alpha) &= \int \ell(f(\theta,x),y)g(x,y)\,\mathrm{d}P_S(x,y) \\ &+\lambda\int g(x,y)\mathrm{d}P_S(x,y)-\beta\int g(x,y)\left(1+\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)+\alpha g(x,y)\right)\right)\mathrm{d}P_S(x,y). \end{aligned} \tag{44}$$

From equation (42) and (44), it follows that

$$\begin{aligned}
&\partial L\left(\frac{\mathrm{d}P}{\mathrm{d}P_S},\beta,\lambda;g\right)\\
&= \int \ell(f(\theta,x),y)g\left(x,y\right)\mathrm{d}P_S\left(x,y\right)\\
&+\lambda\int g\left(x,y\right)\mathrm{d}P_S\left(x,y\right)-\beta\int g\left(x,y\right)\left(1+\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}\left(x,y\right)\right)\right)\mathrm{d}P_S\left(x,y\right). \quad (45)
\end{aligned}$$

A necessary condition [34][Theorem 1, Page 178] for the functional $L$ in (41) to have a minimum at $\left(\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S},\beta,\lambda\right)\in\mathscr{S}\times[0,+\infty)^2$ is that for all functions $g\in\mathscr{S}$, the following holds,

$$\partial L\left(\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S};g\right)=0. \quad (46)$$

The equality in (46) holds for all functions $g\in\mathscr{S}$ if for all $(x,y)\in\operatorname{supp}P_S$, $\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}$ satisfies:

$$\ell(f(\theta,x),y)-\beta\left(1+\log\left(\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y)\right)\right)+\lambda=0.$$

That is,

$$\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y)=\exp\left(\frac{\lambda-\beta}{\beta}\right)\exp\left(\frac{\ell(f(\theta,x),y)}{\beta}\right), \quad (47)$$

where $\beta$ and $\lambda$ are chosen to satisfy their corresponding constraints. Hence, from (47), it follows that

$$\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y)=\frac{\exp\left(\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta}\right)}{\int\exp\left(\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta}\right)\mathrm{d}P_S(x,y)}, \quad (48)$$

where $\beta$ is chosen to satisfy

$$D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right)=\gamma. \quad (49)$$

This completes the proof.

# B Proof of Lemma 3.1

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

$$P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(\mathcal{C})=\int_{\mathcal{C}}\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y), \quad (50)$$

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

$$P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{C}) = 0, \tag{51}$$

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

$$0 = P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{C}) = \int_{\mathcal{C}} \frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y)\mathrm{d}P_S(x,y). \tag{52}$$

From Theorem 3.1, it holds that for all $(x,y) \in \operatorname{supp} P_S$,

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

Note that if a solution to the optimization problem (7) 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, $\exp\left(\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta}\right) > 0$, where the strict inequality is due to the fact that for all $(x,y) \in \operatorname{supp} P_S$, the function $\ell$ in (4) is nonnegative. Hence, for all $(x,y) \in \operatorname{supp} P_S$,

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

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

# C Proof of Lemma 3.2

The equality in (12) follows from observing that:

$$D\left(P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right) = \int \log\left(\frac{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{53}$$

$$= \int \log\left(\exp\left(\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\right)\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) \tag{54}$$

$$= \int \frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta}\mathrm{d}P_S(x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), \tag{55}$$

where equality (54) follows from (9). The equality in (13) is proved as follows:

$$D\left(P_S \| P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) = \int \log\left(\frac{\mathrm{d}P_S}{\mathrm{d}P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}}(x,y)\right)\mathrm{d}P_S(x,y) \tag{56}$$

$$= \int \log\left(\exp\left(-\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta} + \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\right)\mathrm{d}P_S(x,y) \tag{57}$$

$$= -\int \frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta}\mathrm{d}P_S(x,y) + \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), \tag{58}$$

where equality in (57) follows from equation (9). This completes the proof.

# D Proof of Theorem 4.1

The proof follows from Theorem 3.1 and by noticing that the relative entropy $D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right)$ satisfies:

$$D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right) = \int \log\left(\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) \tag{59}$$

$$= \int \left(\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) \tag{60}$$

$$= \int \frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta}\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), \tag{61}$$

where the equality in (60) follows from (9). 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:

$$D\left(P\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P\|P_S\right) \tag{62}$$

$$= \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\right)\mathrm{d}P(x,y) - \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P(x,y) \tag{63}$$

$$= \int \left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\right) - \log\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\right)\right)\mathrm{d}P(x,y) \tag{64}$$

$$= \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\frac{\mathrm{d}P_S}{\mathrm{d}P}(x,y)\right)\mathrm{d}P(x,y) \tag{65}$$

$$= \int \log\left(\frac{\mathrm{d}P_S}{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}(x,y)\right)\mathrm{d}P(x,y) \tag{66}$$

$$= \int \log\left(\exp\left(-\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta} + \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\right)\mathrm{d}P(x,y) \tag{67}$$

$$= \int \left(-\frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta} + \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right)\mathrm{d}P(x,y) \tag{68}$$

$$= -\int \frac{\ell(f(\boldsymbol{\theta},x),y)}{\beta}\mathrm{d}P(x,y) + \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right). \tag{69}$$

Therefore, from (61) and (69), it follows that

$$D\left(P\|P_S\right) - D\left(P\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right) \tag{70}$$

$$= \frac{1}{\beta}\int \ell(f(\boldsymbol{\theta},x),y)\mathrm{d}P(x,y) - \frac{1}{\beta}\int \ell(f(\boldsymbol{\theta},x),y)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) \tag{71}$$

$$= \frac{1}{\beta}G\left(\boldsymbol{\theta},P,P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right), \tag{72}$$

which completes the proof.

# E Proof of Theorem 4.2

The proof follows from the following equalities:

$$
\begin{aligned}
& G(\boldsymbol{\theta}, P_1, P_2) && (73)\\
&= G(\boldsymbol{\theta}, P_1, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}) - G(\boldsymbol{\theta}, P_2, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}) && (74)\\
&= \beta\Big(D(P_1\|P_S) - D\Big(P_1\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\Big) - D\Big(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\Big)\Big)\\
&\quad -\beta\Big(D(P_2\|P_S) - D\Big(P_2\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\Big) - D\Big(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\Big) && (75)\\
&= \beta\Big(D\left(P_2\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_1\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_2\|P_S\right) + D\left(P_1\|P_S\right)\Big), && (76)
\end{aligned}
$$

where equality in (75) follows from (15). This completes the proof.

# F Proof for Lemma 5.1

The proof follows from Definition 5.1 and the following equalities:

$$
\begin{aligned}
\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) &= \frac{1}{n}\sum_{t=1}^{n} \ell\left(f(\boldsymbol{\theta}, x_t), y_t\right) && (77)\\
&= \frac{1}{n}\sum_{(x,y)\in(\mathcal{X}\times\mathcal{Y})}\sum_{t=1}^{n} \mathbb{1}_{\{x=x_t, y=y_t\}}\ell(f(\boldsymbol{\theta}, x), y) && (78)\\
&= \sum_{(x,y)\in(\mathcal{X}\times\mathcal{Y})} \ell(f(\boldsymbol{\theta}, x), y) P_{\boldsymbol{z}}(x, y) && (79)\\
&= \int \ell(f(\boldsymbol{\theta}, x), y)\mathrm{d}P_{\boldsymbol{z}}(x, y), && (80)
\end{aligned}
$$

which completes the proof.

# G Proof of Lemma 6.2

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}) && (81)\\
&= \int \left( \int \left( \int \ell(f(\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(f(\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}) && (82)\\
&= \int \left( \int \ell(f(\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(f(\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}) && (83)\\
&= \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}) && (84)\\
&= \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}) && (85)\\
&= \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}) && (86)\\
&= \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). && (87)
\end{aligned}
$$

where equality in (82) follows from (31); the equality in (84) follows with the function $\mathsf{L}$ in (5) and $P_{\boldsymbol{Z}} \in \triangle(\mathcal{X} \times \mathcal{Y})^n$ being a product measure obtained from $P_Z$; and equality in (87) follows from [7, Theorem 10.4]. This completes the proof.

## References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

[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 $\alpha$-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] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] F. Futami and T. Iwata, "Information-theoretic analysis of test data sensitivity in uncertainty," arXiv preprint arXiv:2307.12456, Jul. 2023.

[26] 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.

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

[28] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[29] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*, 1st ed. New York, NY, USA: Wiley, 1989.

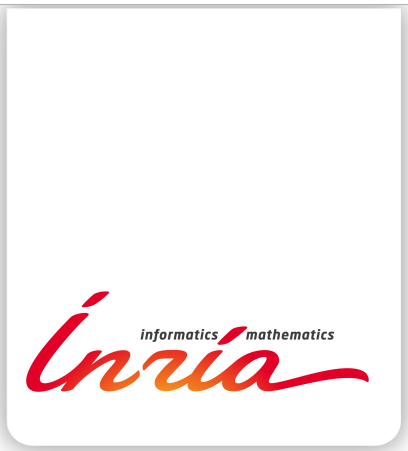
[30] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Analysis of the relative entropy asymmetry in the regularization of empirical risk minimization," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.

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

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

[33] 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.

[34] 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

ISSN 0249-6399