Generalization Analysis of Machine Learning Algorithms via the Worst-Case Data-Generating Probability Measure

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

Generalization Analysis of Machine Learning Algorithms via the Worst-Case Data-Generating Probability Measure

Xinying Zou[1], **Samir M. Perlaza**[1, 2, 3] , **Iñaki Esnaola**[2, 4], **and Eitan Altman**[1, 5]
[1] INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France
[2]Department of Electrical and Computer Engineering, Princeton University, Princeton NJ 08544, USA
[3]GAATI Laboratory, Université de la Polynésie Française, Faaa 98702, French Polynesia
[4]Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, UK
[5]Laboratoire d'Informatique d'Avignon, Université d'Avignon, France
xinying.zou@inria.fr, samir.perlaza@inria.fr, esnaola@sheffield.ac.uk, eitan.altman@inria.fr

1 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.

2 Introduction

The expected generalization error (GE) is a central workhorse for the analysis of generalization capabilities of machine learning algorithms, see for instance (Aminian et al. 2021, 2022; Chu and Raginsky 2023; Xu and Raginsky 2017) and (Perlaza et al. 2023). 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 (Aminian et al. 2021); and for the case in which the reference measure is a σ-finite measure (Perlaza et al. 2022a).

2.1 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 (Russo and Zou 2016), further developed in (Xu and Raginsky 2017) and combined with chaining methods in (Asadi, Abbe, and Verdú 2018; Asadi and Abbe 2020) for deriving upper bounds on the EGE. Similar bounds on the EGE were obtained in (Bu, Zou, and Veeravalli 2020; Chu and Raginsky 2023; Hafez-Kolahi et al. 2020; Hellström and Durisi 2020) and references therein. Other information measures such as the Wasserstein distance (Aminian et al. 2022; Lopez and Jog 2018; Wang et al. 2019), maximal leakage (Esposito, Gastpar, and Issa 2020; Issa, Esposito, and Gastpar 2019), mutual f-information (Masiha, Gohari, and Yassaee 2023), and Jensen-Shannon divergence (Aminian, Toni, and Rodrigues 2021) were used for providing upper bounds on EGE as well. In (Duchi, Glynn, and Namkoong 2021), 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 (Duchi, Glynn, and Namkoong 2021), 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 (Daunas et al. 2023b), and references therein.

Generalization can also be studied as a local minmax problem as in (Lee and Raginsky 2018), 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 (Mazuelas, Shen, and Pérez 2022). To circumvent the dependence on the statistical description of the dataset, gen-

eralization analyses often rely on approaches that decouple the explicit link of the data-generating measure with the GE by using tools from combinatorics (Cherkassky et al. 1999); probability theory (Cullina, Bhagoji, and Mittal 2018; Haddouche et al. 2021; McAllester 2003); and information theory (Aminian et al. 2021; Russo and Zou 2019; Xu and Raginsky 2017). 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 (Perlaza et al. 2022c, 2023). 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.

2.2 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 4.1) 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 5.2). 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 (Csiszár 1998), 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 6.1). This observation, in conjuction with Theorem 5.2 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 (Aminian et al. 2021) and (Perlaza et al. 2022a)) 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.

2.3 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

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

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

3 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

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

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

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

Let also the function $\hat{\ell} : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty]$ be such that given a data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the loss induced by a model $\boldsymbol{\theta} \in \mathcal{M}$ is $\hat{\ell}(f(\boldsymbol{\theta}, x), y)$. In the following, the loss function $\hat{\ell}$ is assumed to be nonnegative and for all $y \in \mathcal{Y}$, it holds that $\hat{\ell}(y, y) = 0$.

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

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

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

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

where the functions f and ℓ 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 ERM problem is prone to overfitting since the set of solutions to (6) 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 (6). Such a regularization term can be represented by a function $R : \mathcal{M} \to \mathbb{R}$, which yields the regularized ERM problem

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

where λ is a nonnegative real that acts as a regularization parameter. The regularization function R in (7) 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 (Shalev-Shwartz and Ben-David 2014). 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.

4 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 $P_S \in \triangle\left(\mathcal{X} \times \mathcal{Y}\right)$ 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:

$$\max_{P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})} \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P(x, y) \tag{8a}$$

$$\text{s.t.} \qquad D\left(P \| P_S\right) \leq \gamma \tag{8b}$$

$$\int \mathrm{d}P(x, y) = 1, \tag{8c}$$

where the functions f and ℓ are defined in (3) and (4).

The probability measure P_S in (8) 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 (8). 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 (Daunas et al. 2023a).

4.1 The Solution

The following theorem characterizes the solution to the optimization problem in (8) using the function $\mathsf{J}_{P_S,\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$, which satisfies

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P_S(x, y)\right), \tag{9}$$

with the functions f and ℓ in (3) and (4), respectively.

Theorem 4.1. *The solution to the optimization problem in* (8)*, if it exists, is denoted by* $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ *and satisfies for all* $(x, y) \in \operatorname{supp} P_S$*,*

$$\frac{\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}}{\mathrm{d}P_S}(x, y) = \exp\left(\frac{\ell(\boldsymbol{\theta}, x, y)}{\beta} - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right), \tag{10}$$

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ *is defined in* (9) *and* $\beta > 0$ *satisfies*

$$D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) = \gamma. \tag{11}$$

Proof: The proof is presented in (Zou et al. 2023, Appendix A). ∎

Theorem 4.1 provides a guarantee on the uniqueness of the solution to the optimization problem in (8), whenever it exists. Nonetheless, guarantees for the existence of a solution to (8) are not provided. In the following, it is assumed that the model $\boldsymbol{\theta}$, the real γ, and the probability measure P_S in (8) are such that a solution exists. Let the set $\mathcal{J}_{P_S,\boldsymbol{\theta}} \subset (0, +\infty)$ be:

$$\mathcal{J}_{P_S,\boldsymbol{\theta}} \triangleq \left\{t \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) < +\infty\right\}. \tag{12}$$

The existence of a solution to the problem in (8) 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 (4), 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 4.1. *Assume that for some model* $\boldsymbol{\theta} \in \mathcal{M}$*, there exists a real* $a \in (0, +\infty)$ *such that*

$$P_S\left(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) \leqslant a\}\right) = 1, \tag{13}$$

where the function ℓ is defined in (4)*. Note that the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ satisfies for all $t \in \mathbb{R}$,*

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) \leqslant \log\left(\int \exp\left(\frac{a}{t}\right) \mathrm{d}P(x,y)\right) \quad (14)$$

$$= \frac{a}{t} + \log\left(\int \mathrm{d}P(x,y)\right) \quad (15)$$

$$= \frac{a}{t} < +\infty, \quad (16)$$

which implies that under the assumption in (13)*, the optimization problem in* (8) *always has a solution.*

In general, if a solution to (8) exists, the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (10) is a Gibbs probability measure (Georgii 2011). From this perspective, the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (10) is often referred to as the log-partition function (Dembo and Zeitouni 2009). Moreover, the probability measure P_S in (8) can be interpreted as a prior on the probability distribution of the datasets.

4.2 Mutual Absolute Continuity

When the optimization problem in (8) possesses a solution, i.e., $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$ with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (12), 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 4.1. *If the problem in* (8) *has a solution, then*

$$P_S\left(\left\{(x,y) \in \operatorname{supp} P_S : \ell(\boldsymbol{\theta}, x, y) = +\infty\right\}\right) = 0, \quad (17)$$

where the function ℓ is in (4)*.*

Proof: The proof is presented in (Zou et al. 2023, 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 (10). 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 4.2. *The probability measures $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ and P_S in* (10) *are mutually absolutely continuous.*

Proof: The proof is presented in (Zou et al. 2023, 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 (10) is described by the following lemma.

Lemma 4.3. *The probability measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (10) *satisfy:*

$$\beta \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$$

$$= \int \ell(\boldsymbol{\theta},x,y) \mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y) - \beta D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) \quad (18)$$

$$= \int \ell(\boldsymbol{\theta},x,y) \mathrm{d}P_S(x,y) + \beta D\left(P_S \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right), \quad (19)$$

where the functions f and ℓ are defined in (3) *and* (4)*, respectively; and the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is in* (9)*.*

Proof: The proof is presented in (Zou et al. 2023, Appendix D). ■

5 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

$$G(\boldsymbol{\theta}, P_1, P_2) \quad (20)$$
$$= \int \ell(\boldsymbol{\theta},x,y) \mathrm{d}P_1(x,y) - \int \ell(\boldsymbol{\theta},x,y) \mathrm{d}P_2(x,y),$$

where the functions f and ℓ 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 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 (10) to an alternative measure.

Theorem 5.1 (Sensitivity of the Expected Loss)**.** *For all $P \in \Delta_{P_S}(\mathcal{X} \times \mathcal{Y})$ and for all $\boldsymbol{\theta} \in \mathcal{M}$,*

$$G\left(\boldsymbol{\theta}, P, P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) \quad (21)$$
$$= \beta\left(D(P\|P_S) - D\left(P\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right) - D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\|P_S\right)\right),$$

where the functional G is defined in (20)*; and the model $\boldsymbol{\theta}$ and the measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ satisfy* (10)*.*

Proof: The proof is presented in (Zou et al. 2023, Appendix E). ■

The following corollary of Theorem 5.1 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 5.1. *The probability measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (10) *satisfy:*

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

where the functional G is in (20)*.*

The right-hand side of the equality in (22) is a symmetrized Kullback-Liebler divergence, also known as Jeffrey's divergence (Jeffreys 1946), 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 5.1.

Corollary 5.2. *The probability measures P_S and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in* (10) *satisfy:*

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

where the function ℓ is defined in (4)*.*

Note that the probability measure P_S in Corollary 5.2 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 $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ 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 $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ to any alternative measure P provided by Theorem 5.1, 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 5.2. *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 functional G in* (20) *satisfies*

$$G(\boldsymbol{\theta},P_1,P_2)=\beta\Big(D\left(P_2\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)-D\left(P_1\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) -D\left(P_2\|P_S\right)+D\left(P_1\|P_S\right)\Big), \tag{23}$$

where the model $\boldsymbol{\theta}$ and the measures P_S and $P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfy (10)*.*

Proof: The proof is presented in (Zou et al. 2023, Appendix F). ∎

Note that the parameters γ and P_S in (8) can be arbitrarily chosen. This is essentially because only the right-hand side of (23) depends on P_S and β. Another interesting observation is that none of the terms in the right-hand side of (23) 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 (23) to construct closed-form expressions for the sensitivity $G(\boldsymbol{\theta},P_1,P_2)$ in (20). 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 (23) significantly simplifies are $P_S=P_1$ and $P_S=P_2$, which leads to the following corollary of Theorem 5.2.

Corollary 5.3. *If P_1 is absolutely continuous with P_2, then the value $G(\boldsymbol{\theta},P_1,P_2)$ in* (20) *satisfies:*

$$G(\boldsymbol{\theta},P_1,P_2) =\beta\Big(D\Big(P_2\|P^{(P_2,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)-D\Big(P_1\|P^{(P_2,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)+D(P_1\|P_2)\Big). \tag{24}$$

Alternatively, if P_2 is absolutely continuous with P_1 then,

$$G(\boldsymbol{\theta},P_1,P_2) =\beta\Big(D\Big(P_2\|P^{(P_1,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)-D\Big(P_1\|P^{(P_1,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)-D(P_2\|P_1), \tag{25}$$

where for all $i\in\{1,2\}$, the probability measure $P^{(P_i,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfies (10) *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 (20). 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.

6 Analysis of the Empirical-Risk

This section presents a mathematical object known as a *type* in the realm of information theory (Csiszár 1998). In the context of this work, a type is a probability measure induced by a dataset, as shown hereunder.

Definition 6.1 (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 $\{(x,y)\}\in\mathscr{F}_{\mathcal{X}\times\mathcal{Y}}$,*

$$P_{\boldsymbol{z}}\left(\{(x,y)\}\right)=\frac{1}{n}\sum_{t=1}^{n}\mathbb{1}_{\{x=x_t,y=y_t\}}(x,y). \tag{26}$$

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 $P_{\boldsymbol{z}}\left(x,y\right)$ is allowed for 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 6.1 (Empirical Risks and Types)**.** *The empirical risk $\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})$ in* (5) *satisfies*

$$\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})=\int\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_{\boldsymbol{z}}(x,y), \tag{27}$$

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

Proof: The proof is presented in (Zou et al. 2023, Appendix G). ∎

6.1 Sensitivity of the Empirical Risk

Equipped with the result in Lemma 6.1, 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

$$G(\boldsymbol{\theta},P_{\boldsymbol{z}_1},P_{\boldsymbol{z}_2})=\mathsf{L}(\boldsymbol{z}_1,\boldsymbol{\theta})-\mathsf{L}(\boldsymbol{z}_2,\boldsymbol{\theta}), \tag{28}$$

where the functional G is in (20). Assume that $P_{\boldsymbol{z}_1}$ and $P_{\boldsymbol{z}_2}$ are absolutely continuous with respect to the reference measure P_S in (8). Under this assumption, the equality in (28) 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 6.1. *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* (8)*, the*

following holds for all $\boldsymbol{\theta} \in \mathcal{M}$:

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

where the function 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* (10)*.*

Proof: The proof follows from the equality in (28), which together with Theorem 5.2 completes the proof. ■

In Theorem 6.1, 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 (Perlaza et al. 2023).

From Theorem 6.1, 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 $\boldsymbol{\theta}$ 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., $D\left(P_{\boldsymbol{z}_2}\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_{\boldsymbol{z}_1}\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$; and (b) the difference of the "statistical distance" from the types to the reference measure P_S, i.e., $D\left(P_{\boldsymbol{z}_1}\|P_S\right) - D\left(P_{\boldsymbol{z}_2}\|P_S\right)$.

7 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

$$\begin{aligned}&G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})\\ &= \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_Z(x,y) - \int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_{\boldsymbol{z}}(x,y). \qquad (30)\end{aligned}$$

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 (Shalev-Shwartz and Ben-David 2014). 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 (30), such generalization error can be written in terms of the functional G in (20). This observation leads to the following description of the generalization gap.

Lemma 7.1. *The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in* (30) *satisfies:*

$$\begin{aligned}&G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) =\\ &\beta\Big(D\Big(P_{\boldsymbol{z}}\|P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big) - D(P_{\boldsymbol{z}}\|P_Z) - D\Big(P_Z\|P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)\Big), \qquad (31)\end{aligned}$$

where the measure $P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is the solution to the optimization problem in (8) *under the assumption that $P_S = P_Z$.*

Proof: The proof follows from Corollary 5.3 by noticing that the type $P_{\boldsymbol{z}}$ is absolutely continuous with respect to P_Z. ■

Lemma 7.1 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 (30) 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 $D\left(P_{\boldsymbol{z}}\|P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_Z\|P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$ is also arbitrarily small.

A more general expression for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (30) is provided by the following corollary of Theorem 5.2.

Corollary 7.1. *The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in* (30) *satisfies:*

$$\begin{aligned}G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \beta\Big(&D\left(P_{\boldsymbol{z}}\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_Z\|P^{(P_S,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\\ &-D\left(P_{\boldsymbol{z}}\|P_S\right) + D\left(P_Z\|P_S\right)\Big), \qquad (32)\end{aligned}$$

where the measure $P^{(P_Z,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is in (8)*.*

Note that several expressions for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (30) can be obtained from Corollary 7.1 by choosing the reference P_S and the parameter γ in (8), which determines the value of β.

7.1 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 (30) 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

$$\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}}) = \int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}), \quad (33)$$

where the functional G is in (30). 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 (33). Corollary 7.1, by appropriately choosing the reference measure P_S and the parameter γ in (8), 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.

7.2 Doubly-Expected Generalization Gap

The expected generalization gap $\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}})$ in (33) 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

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

where the functional G is in (30). 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 (34). In existing literature, the doubly-expected generalization gap is simply referred to as generalization error. See for instance (Xu and Raginsky 2017), (Aminian et al. 2021), and (Perlaza et al. 2022a). 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 7.1 leads to numerous closed-form expressions for the doubly-expected generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$.

7.3 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 $Q \in \triangle\left(\mathcal{M}, \mathscr{B}\left(\mathcal{M}\right)\right)$ (Perlaza et al. 2022a). 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 $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\right), \quad (35)$$

where the dataset $\boldsymbol{z}$ represents the training dataset; the function L is defined in (5); and the function $K_{Q,\boldsymbol{z}} : \mathbb{R} \to \mathbb{R}$ satisfies $K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp\left(t\,\mathsf{L}\left(\boldsymbol{z},\boldsymbol{\nu}\right)\right) \mathrm{d}Q(\boldsymbol{\nu})\right)$.

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 7.2 (Generalization Gap of the Gibbs Algorithm)**.** *Given the conditional probability measure $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in (35) and a probability measure $P_Z \in \triangle\left(\mathcal{X}\times\mathcal{Y}\right)$, the generalization gap $\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z)$ satisfies*

$$\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_Z) = \lambda\Big(I\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\Big) + L\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\Big)\Big), \quad (36)$$

where $P_{\boldsymbol{Z}} \in \triangle\left(\mathcal{X}\times\mathcal{Y}\right)^n$ is a product measure obtained from P_Z; and $I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)$ 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

$$I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\Big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \text{ and } (37)$$

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

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

Proof: The proof is presented in (Zou et al. 2023, Appendix H). ■

Lemma 7.2 has been proved before for the case in which Q is a probability measure in (Aminian et al. 2021); and in the more general case in which Q is a σ-finite measure in (Perlaza et al. 2022a). In both (Aminian et al. 2021) and (Perlaza et al. 2022a), the result is shown without the assumption that the measure $P_{\boldsymbol{Z}}$ is a product measure, which is an assumption in Lemma 7.2. This limitation is due to the fact that the proof of Lemma 7.2 relies on the notion of types, which is known to fail capturing the correlation between datapoints, as pointed in (Csiszár 1998). Nonetheless, the independent and identically distributed assumption is widely adopted in the realm of machine learning. Despite this limitation, the relevance of Lemma 7.2 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 functional G in (20), and the notion of (doubly-expected) generalization gap, which is a central performance metric for evaluating the generalization capabilities of machine learning algorithms.

8 Conclusions and Final Remarks

The worst-case data-generating probability measure in Theorem 4.1 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 (Perlaza et al. 2022a,b,c). 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.

9 Acknowledgments

This work is funded in part by the ANR Project PARFAIT under grant ANR-21-CE25-0013 and the INRIA Exploratory Action IDEM.

References

Aminian, G.; Bu, Y.; Toni, L.; Rodrigues, M.; and Wornell, G. 2021. An Exact Characterization of the Generalization Error for the Gibbs Algorithm. *Advances in Neural Information Processing Systems*, 34: 8106–8118.

Aminian, G.; Bu, Y.; Wornell, G. W.; and Rodrigues, M. R. 2022. Tighter expected generalization error bounds via convexity of information measures. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2481–2486. Aalto, Finland.

Aminian, G.; Toni, L.; and Rodrigues, M. R. 2021. Jensen-Shannon information based characterization of the generalization error of learning algorithms. In *Proceedings of the IEEE Information Theory Workshop (ITW)*, 1–5. Kanazawa, Japan.

Asadi, A.; Abbe, E.; and Verdú, S. 2018. Chaining mutual information and tightening generalization bounds. *Advances in Neural Information Processing Systems*, 31: 7245–7254.

Asadi, A. R.; and Abbe, E. 2020. Chaining Meets Chain Rule: Multilevel Entropic Regularization and Training of Neural Networks. *The Journal of Machine Learning Research*, 21(1): 5453–5484.

Bu, Y.; Zou, S.; and Veeravalli, V. V. 2020. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130.

Cherkassky, V.; Shao, X.; Mulier, F. M.; and Vapnik, V. N. 1999. Model complexity control for regression using VC generalization bounds. *IEEE Transactions on Neural Networks*, 10(5): 1075–1089.

Chu, Y.; and Raginsky, M. 2023. A unified framework for information-theoretic generalization bounds. arXiv preprint arXiv:2305.11042.

Csiszár, I. 1998. The method of types. *IEEE Transactions on Information Theory*, 44(6): 2505–2523.

Cullina, D.; Bhagoji, A. N.; and Mittal, P. 2018. PAC-learning in the presence of adversaries. *Advances in Neural Information Processing Systems*, 31(1): 1–12.

Daunas, F.; Esnaola, I.; Perlaza, S. M.; and Poor, H. V. 2023a. Analysis of the Relative Entropy Asymmetry in the Regularization of Empirical Risk Minimization. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 340–345. Taipei, Taiwan.

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

Dembo, A.; and Zeitouni, O. 2009. *Large Deviations Techniques and Applications*. New York, NY, USA: Springer-Verlag, 2nd edition.

Duchi, J. C.; Glynn, P. W.; and Namkoong, H. 2021. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969.

Esposito, A. R.; Gastpar, M.; and Issa, I. 2020. Robust Generalization via α-Mutual Information. arXiv preprint arXiv:2001.06399.

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

Haddouche, M.; Guedj, B.; Rivasplata, O.; and Shawe-Taylor, J. 2021. PAC-Bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10): 1–20.

Hafez-Kolahi, H.; Golgooni, Z.; Kasaei, S.; and Soleymani, M. 2020. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 16457–16467.

Hellström, F.; and Durisi, G. 2020. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1(3): 824–839.

Issa, I.; Esposito, A. R.; and Gastpar, M. 2019. Strengthened information-theoretic bounds on the generalization error. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 582–586. Paris, France.

Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007): 453–461.

Lee, J.; and Raginsky, M. 2018. Minimax statistical learning with Wasserstein distances. *Advances in Neural Information Processing Systems*, 31: 2687–2696.

Lopez, A. T.; and Jog, V. 2018. Generalization error bounds using Wasserstein distances. In *Proceedings of the IEEE Information Theory Workshop (ITW)*, 1–5. Guangzhou, China.

Masiha, S.; Gohari, A.; and Yassaee, M. H. 2023. f-divergences and their applications in lossy compression and bounding generalization error. *IEEE Transactions on Information Theory*, 69(12): 7245–7254.

Mazuelas, S.; Shen, Y.; and Pérez, A. 2022. Generalized maximum entropy for supervised classification. *IEEE Transactions on Information Theory*, 68(4): 2530–2550.

McAllester, D. A. 2003. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1): 5–21.

Perlaza, S. M.; Bisson, G.; Esnaola, I.; Jean-Marie, A.; and Rini, S. 2022a. Empirical Risk Minimization with Relative Entropy Regularization. Technical Report RR-9454, INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France.

Perlaza, S. M.; Bisson, G.; Esnaola, I.; Jean-Marie, A.; and Rini, S. 2022b. Empirical Risk Minimization with Relative Entropy Regularization: Optimality and Sensitivity. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 684–689. Espoo, Finland.

Perlaza, S. M.; Esnaola, I.; Bisson, G.; and Poor, H. V. 2022c. Sensitivity of the Gibbs Algorithm to Data Aggregation in Supervised Machine Learning. Technical Report RR-9474, INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France.

Perlaza, S. M.; Esnaola, I.; Bisson, G.; and Poor, H. V. 2023. On the Validation of Gibbs Algorithms: Training Datasets, Test Datasets and their Aggregation. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 328–333. Taipei, Taiwan.

Russo, D.; and Zou, J. 2016. Controlling Bias in Adaptive Data Analysis Using Information Theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, 1232–1240. Cadiz, Spain.

Russo, D.; and Zou, J. 2019. How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1): 302–323.

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

Wang, H.; Diaz, M.; Santos Filho, J. C. S.; and Calmon, F. P. 2019. An information-theoretic view of generalization via Wasserstein distance. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 577–581. Paris, France.

Xu, A.; and Raginsky, M. 2017. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30: 1–10.

Zou, X.; Perlaza, S. M.; Esnaola, I.; and Altman, E. 2023. The Worst-Case Data-Generating Probability Measure. Technical Report RR-9515, INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France.