An Exact Characterization of the Generalization Error of Machine Learning Algorithms

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

▶ **To cite this version:**

Xinying Zou, Samir M. Perlaza, Iñaki Esnaola, Eitan Altman, H. Vincent Poor. An Exact Characterization of the Generalization Error of Machine Learning Algorithms. 2024. hal-04426500v1

HAL Id: hal-04426500

https://inria.hal.science/hal-04426500v1

Preprint submitted on 2 Feb 2024 (v1), last revised 14 May 2024 (v2)

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

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

Distributed under a Creative Commons Attribution 4.0 International License

An Exact Characterization of the Generalization Error of Machine Learning Algorithms

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

***Abstract*—In this paper, the notion of worst-case data-generating (WCDG) probability measure, recently introduced in [1], is leveraged to obtain an exact expression of the expected generalization gap (or generalization error) for any machine learning algorithm. This exact expression is provided in terms of information measures involving the WCDG probability measure and leads to an upper bound on the generalization error that is equal to the sum of the mutual information and the lautum information between the models and the datasets, up to a constant factor. This upper bound is achieved by a Gibbs algorithm whose parameters satisfy particular conditions. Finally, given a fixed model, it is shown that the empirical risk is a sub-Gaussian random variable when datasets are sampled from the WCDG probability measure. This observation leads to the construction of new generalization guarantees coined (ϵ, δ)-robustness.**

I. Introduction

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:

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

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

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

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

$$\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, x_i, y_i) , \quad (3)$$

where the function ℓ is defined in (2). Given a model $\boldsymbol{\theta} \in \mathcal{M}$, the expected loss induced by a measure $P \in \triangle(\mathcal{X} \times \mathcal{Y})$ is defined as follows.

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

$$\mathsf{R}_{\boldsymbol{\theta}}(P) = \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P(x, y), \quad (4)$$

where the function ℓ is defined in (2).

Given a probability measure $P_S \in \triangle(\mathcal{X} \times \mathcal{Y})$, which can be interpreted as a *prior* on the set of data points, and a model $\boldsymbol{\theta} \in \mathcal{M}$, studying the worst-case data-generating (WCDG) probability measure leads to two different optimization problems. Firstly, a *neighborhood* of candidate data-generating probability measures is constructed as the set of probability measures that satisfy a constraint expressed in terms of a *statistical distance* $\mathsf{d} : \triangle(\mathcal{X} \times \mathcal{Y}) \times \triangle(\mathcal{X} \times \mathcal{Y}) \to [0, +\infty)$. This *neighborhood* describes the set of plausible probability measures from which test datasets might be sampled from. This work focuses on the case in which the statistical distance is the relative entropy with respect to P_S. Thus, the WCDG measure is the probability measure that maximizes the expectation of the loss within this neighborhood. This view leads to the following problem, c.f. [1] and [2]:

$$\max_{P \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})} \mathsf{R}_{\boldsymbol{\theta}}(P) \quad (5a)$$

$$\text{s.t.} \qquad D(P \| P_S) \leqslant \gamma, \quad (5b)$$

where $\gamma > 0$ determines the neighborhood around P_S as the set $\{P \in \triangle(\mathcal{X} \times \mathcal{Y}) : D(P \| P_S) \leqslant \gamma\}$ and the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (4).

Secondly, the WCDG probability measure can also be interpreted as the probability measure that trades off the maximization of the expectation of the loss and the minimization of the statistical distance with respect to P_S. This alternative perspective yields the optimization problem:

$$\max_{P \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})} \mathsf{R}_{\boldsymbol{\theta}}(P) - \beta D(P \| P_S) , \quad (6a)$$

where the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (4); and $\beta > 0$ determines the trade-off between maximization of the expectation of the loss and the statistical distance to P_S.

The optimization problems in (5) and (6) exhibit a major difference in the following case. If for all $P \in \triangle_{P_S}$, the functional $\mathsf{R}_{\boldsymbol{\theta}}$ in (4) is such that $\mathsf{R}_{\boldsymbol{\theta}}(P) = c$, for some fixed $c \geqslant 0$, then all the measures in $\{P \in \triangle(\mathcal{X} \times \mathcal{Y}) : D(P\|P_S) \leqslant \gamma\}$ are solutions to the optimization problem in (5). Alternatively, in this case P_S is the unique solution to the problem in (6). Although this difference holds substantial mathematical importance, its practical relevance is limited. This observation is explained by the notion of nonseparable loss functions, which is analogous to [3, Definition 4.1]. When the function ℓ in (2) is nonseparable with respect to the reference measure P_S, it is a constant almost surely with respect to such a measure. More specifically, there exists a real $a \geq 0$, such that $P_S(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = a\}) = 1$. As a consequence, for all probability measures $P \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})$, it holds that $P(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = a\}) = 1$. If this pathological case of nonseparable loss functions is disregarded, the optimization problems in (5) and (6) exhibit considerable similarity. To recognize this similarity, consider the following. Given a model $\boldsymbol{\theta} \in \mathcal{M}$, let $\mathsf{J}_{P_S,\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$ be the function

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P_S(x,y)\right), \tag{7}$$

where the function ℓ is defined in (2). Let also the set $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ be the set

$$\mathcal{J}_{P,\boldsymbol{\theta}} \triangleq \left\{t \in (0,+\infty) : \mathsf{J}_{P,\boldsymbol{\theta}}\left(\frac{1}{t}\right) < +\infty\right\}. \tag{8}$$

Using these elements, the following theorem proves that the optimization problems in (5) and (6) exhibit the same unique solution.

Theorem 1: If the function ℓ in (2) is separable with respect to the measure P_S and $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (8), then the probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})$ that satisfies for all $(x,y) \in \operatorname{supp} P_S$,

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

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is defined in (7) and

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

is the unique solution to the optimization problems in (5) and (6).

Proof: The proof is presented in [4]. ∎

Theorem 1 presents an alternative characterization of the WCDG probability measure introduced in [1], which unveils properties that hold significant importance in the subsequent sections.

II. Contributions and Related Works

The expected generalization gap (EGG) is a central performance metric for machine learning algorithms, see for instance, [5]–[14] and [15]. Analytical formulations for the exact EGG have been established solely for the Gibbs algorithm when the reference measure is a probability measure in [5] and a σ-finite measure in [3]. For other machine learning algorithms, only upper bounds in terms of information measures have been obtained for the EGG, see for instance, [8], [16]–[26], and references therein. The first contribution of this paper is a closed-form expression for EGG for any statistical machine learning algorithm. This expression is formulated using information measures that incorporate the WCDG probability measure introduced in [1]. The derived expression establishes an upper bound on the EGG for any statistical machine learning algorithm. Such a bound is the sum of the mutual information and the lautum information between models and datasets, up to a constant for which a closed-form expression is provided. Notably, this bound is tight for a Gibbs algorithm, whose parameters are selected to satisfy a specific condition. The idea of studying the *worst-case* is also adopted in minimax distributed robust optimization, where the aim is to minimize the risk over a large ambiguity set, see for instance [27], [28] and [29]. This problem can also be cast as an instance of the maximum entropy principle [30]. Finally, this work has privileged the mathematical formulation whose solution exhibits properties that are useful for the analysis of the generalization. In particular, given a fixed model, the empirical risk is shown to be a sub-Gaussian random variable when datasets are sampled from the WCDG probability measure. This has led to the construction of new generalization guarantees coined (ϵ, δ)-robustness.

III. An Expression of the Expected Generalization Gap

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

$$G(\boldsymbol{\theta}, P_1, P_2) = \mathsf{R}_{\boldsymbol{\theta}}(P_1) - \mathsf{R}_{\boldsymbol{\theta}}(P_2), \tag{11}$$

where the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (4) and $\triangle(\mathcal{X} \times \mathcal{Y})$ is the set of all probability measures over $\mathcal{X} \times \mathcal{Y}$. In order to introduce the definition of the generalization gap, the notion of type induced by a dataset [31] is presented below.

Definition 2 (Type of a dataset): The type induced by the dataset $\boldsymbol{z}$ in (1) is a probability measure in $\triangle(\mathcal{X} \times \mathcal{Y})$, denoted by $P_{\boldsymbol{z}}$, such that for all measurable subsets $\mathcal{A}$ of $\mathcal{X} \times \mathcal{Y}$, $P_{\boldsymbol{z}}(\mathcal{A}) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{(x_t,y_t) \in \mathcal{A}\}}$.

In [1, Lemma 5.1], it is proved that the empirical risk $\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})$ in (3) can be written as $\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) = \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{\boldsymbol{z}}(x,y) = \mathsf{R}_{\boldsymbol{\theta}}(P_{\boldsymbol{z}})$, with the functional $\mathsf{R}_{\boldsymbol{\theta}}$ defined in (4). Given a model $\boldsymbol{\theta} \in \mathcal{M}$ obtained from the training dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ and assuming that the training and test data points are generated according to an independent and identically distributed probability measure $P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$, the *generalization gap* induced by the model $\boldsymbol{\theta}$ is

$$G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \mathsf{R}_{\boldsymbol{\theta}}(P_Z) - \mathsf{R}_{\boldsymbol{\theta}}(P_{\boldsymbol{z}}). \tag{12}$$

The term $\mathsf{R}_{\boldsymbol{\theta}}(P_{\boldsymbol{z}}) = \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})$ is referred to as the training empirical risk, while the term $\mathsf{R}_{\boldsymbol{\theta}}(P_Z)$ is refered as *true risk* or *population risk*.

Let $\triangle(\mathcal{M}|(\mathcal{X}\times\mathcal{Y})^n)$ be the set containing all probability measures over $\mathcal{M}$ conditioned over an element of $(\mathcal{X}\times\mathcal{Y})^n$. Let also the function $\overline{\overline{G}} : \triangle(\mathcal{M}|(\mathcal{X}\times\mathcal{Y})^n) \times \triangle((\mathcal{X}\times\mathcal{Y})^n) \to \mathbb{R}$ be such that

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) = \int\int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \quad (13)$$

where $P_{\boldsymbol{z}}$ is the type induced by the dataset $\boldsymbol{z}$; and the functional G is defined in (12). Consider the random transformation $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, such that given a training dataset $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$, the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} \in \triangle(\mathcal{M})$ is used to perform model selection. The conditional measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is referred to as a statistical learning algorithm. Under the assumption that the training and test datasets are both formed by independent and identically distributed datapoints sampled from the measure P_Z in (12), the expected generalization gap induced by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (13), where $P_{\boldsymbol{Z}} \in \triangle((\mathcal{X}\times\mathcal{Y})^n)$ is a product distribution formed by P_Z. The main result of this section is presented by the following theorem.

Theorem 2: The expected generalization gap $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (13), with $P_{\boldsymbol{Z}} \in \triangle((\mathcal{X}\times\mathcal{Y})^n)$, satisfies:

$$\begin{aligned}\frac{1}{\beta}\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) = \frac{1}{n} L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \\ + \int D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right) \mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \\ - \int D\left(P_Z \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right) \mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \quad (14)\end{aligned}$$

where the term $L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)$ is the *lautum information* [32] given by

$$L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\big(P_{\boldsymbol{\Theta}} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \quad (15)$$

and the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}$ is the WCDG probability measure in (9), with reference measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ obtained from the joint measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}P_{\boldsymbol{Z}}$ and $\beta \in \Big\{t > 0 : \forall \boldsymbol{\theta} \in \mathcal{M}, \int \exp\left(\frac{1}{t}\,\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) < +\infty\Big\}$.

Proof: The proof is presented in [4]. ∎

Theorem 2 appears to be the first exact characterization of the expected generalization gap induced by an arbitrary algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in terms of information measures. The following theorem unveils an upper bound closely linked to the celebrated Gibbs algorithm, which adds an interesting insight to the characterization of the generalization gap via the WCDG probability measure.

Theorem 3: The expected generalization gap $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (13), with $P_{\boldsymbol{Z}} \in \triangle((\mathcal{X}\times\mathcal{Y})^n)$, satisfies:

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) \leqslant \frac{\beta}{n}\left(I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) + L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)\right) (16)$$

where the term $I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)$ is the mutual information given by

$$I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| P_{\boldsymbol{\Theta}}\big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \quad (17)$$

$L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)$ is in (15); and β satisfies for all $\boldsymbol{\theta} \in \mathcal{M}$, $D\left(P_Z \| P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right) = 0$, with the measures P_Z, $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$, and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}$ defined in (14).

Proof: The proof is presented in [4]. ∎

A typical example of a statistical learning algorithm is the Gibbs algorithm, which is parametrized by a positive real λ and by a σ-measure $Q \in \triangle(\mathcal{M})$, c.f. [3] and [33]. In the following, the focus is on the case in which Q is a probability measure. Under this assumption, the probability measure representing such an algorithm, which is denoted by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}$ with $\lambda > 0$, satisfies for all $\boldsymbol{\theta} \in \operatorname{supp} Q$ and for all $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$, $\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\right)$, 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 EGG induced by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}$, under the assumption that datasets are sampled from a product distribution $P_{\boldsymbol{Z}} \in \triangle((\mathcal{X}\times\mathcal{Y})^n)$ formed by the measure P_Z, denoted by $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}, P_{\boldsymbol{Z}})$, satisfies the following property [1], [5] and [3].

Lemma 1: The expected generalization gap $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}, P_{\boldsymbol{Z}})$ satisfies

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}, P_Z) = \lambda\Big(I\Big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\Big) + L\Big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\Big)\Big), \quad (18)$$

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

$$I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\right) \triangleq \int D\Big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)} \| P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\Big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \text{ and} \quad (19)$$

$$L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\right) \triangleq \int D\Big(P_{\boldsymbol{\Theta}}^{(Q,\lambda)} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}\Big) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), \quad (20)$$

with $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$ being a measure such that for all measurable sets $\mathcal{A} \in \mathscr{B}(\mathcal{M})$,

$$P_{\boldsymbol{\Theta}}^{(Q,\lambda)}(\mathcal{A}) = \int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}(\mathcal{A}) \,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}). \quad (21)$$

Note that $I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\right)$ in (19) and $L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\right)$ in (20) also satisfy that

$$I\left(P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}; P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right) = \int D\Big(P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)} \| P_{\boldsymbol{Z}}\Big) \mathrm{d}P_{\boldsymbol{\Theta}}^{(Q,\lambda)}(\boldsymbol{\nu}); \quad (22)$$

$$L\left(P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}; P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\right) = \int D\Big(P_{\boldsymbol{Z}} \| P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}\Big) \mathrm{d}P_{\boldsymbol{\Theta}}^{(Q,\lambda)}(\boldsymbol{\nu}), \quad (23)$$

where $P_{\boldsymbol{Z}|\boldsymbol{\Theta}}^{(Q,\lambda)} \in \triangle((\mathcal{X}\times\mathcal{Y})^n | \mathcal{M})$ is the conditional probability measure that satisfies for all measurable subsets $\mathcal{B}$ of $(\mathcal{X}\times\mathcal{Y})^n$,

$$P_{\boldsymbol{Z}}(\mathcal{B}) = \int P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}(\mathcal{B}) \,\mathrm{d}P_{\boldsymbol{\Theta}}^{(Q,\beta)}(\boldsymbol{\theta}). \quad (24)$$

Moreover, for all $\boldsymbol{\theta} \in \mathcal{M}$, the measure $P^{(Q,\lambda)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n|\mathcal{M}\right)$ is a product measure formed by a measure $P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}} \in \triangle(\mathcal{X}\times\mathcal{Y}|\mathcal{M})$ that satisfies for all measurable sets $\mathcal{A}_i$ in $\mathcal{X}\times\mathcal{Y}$, with $i \in \{1,2,\ldots,n\}$,

$$P^{(Q,\lambda)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\left(\mathcal{A}_1\times\mathcal{A}_2\times\ldots\times\mathcal{A}_n\right) = \prod_{t=1}^{n} P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\left(\mathcal{A}_t\right). \quad (25)$$

Using this notation, Theorem 3 and Lemma 1 lead to the following theorem.

Theorem 4: Assume that there exits a $\lambda > 0$ that satisfies

$$\lambda \in \left\{t>0 : \forall \boldsymbol{\theta}\in\mathcal{M}, \int \exp\left(\frac{1}{nt}\ell(\boldsymbol{\theta},x,y)\right) \mathrm{d}P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y) < +\infty\right\}, \quad (26)$$

where the function ℓ is defined in (2); and the measure $P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is defined in (25). Let the measure $P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ be the WCDG probability measure of the form in (9). If λ satisfies for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$D\left(P_Z \| P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) = 0, \quad (27)$$

then, for all $P_{\boldsymbol{\Theta}|\boldsymbol{Z}} \in \triangle\left(\mathcal{M}|\left(\mathcal{X}\times\mathcal{Y}\right)^n\right)$, the expected generalization gaps $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (13) and $\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (18) satisfy

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) \leqslant \overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}). \quad (28)$$

Proof: The proof is presented in [4]. ■

Theorem 4 shows that, under the assumption that datasets are sampled from $P_{\boldsymbol{Z}}$, the EGG of any algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is upper-bounded by the expected generalization gap induced by a particular Gibbs algorithm $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$. This Gibbs algorithm induces a posterior for a given model $\boldsymbol{\theta}$, denoted by $P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (25). When such a posterior is used as a reference measure to build the WCDG probability measure for such a model $\boldsymbol{\theta}$, with parameter $n\lambda$, it leads to the probability measure $P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$. If the parameters Q and λ are chosen to satisfy (27), then for all $\boldsymbol{\theta} \in \mathcal{M}$, the WCDG probability measure is identical to the actual ground-truth data-generating probability measure P_Z. This is reminiscent of the principle of *indifference* over which the notion of equilibrium in zero-sum games with noisy observations is built [34].

The inequality in (28) reveals the central role of the Gibbs algorithm in statistical machine learning. Essentially, by studying the Gibbs algorithm $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, for which its parameters Q and λ are chosen to satisfy (27), the expected generalization gap of any algorithm facing data sampled from $P_Z = P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ can be upper bounded.

IV. Sub-Gaussianity of the Loss

The function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (7) exhibits several properties that enable the analysis of general machine learning algorithms. Let the m-th derivative of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ be denoted by $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}} : \mathbb{R}\to\mathbb{R}$, with $m \in \mathbb{N}$. Hence, for all $t \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (8), $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}(t) \triangleq \frac{\mathrm{d}^m}{\mathrm{d}s^m}\mathsf{J}_{P_S,\boldsymbol{\theta}}(s)\Big|_{s=t}$. It is proved in [4] that for a fixed $\boldsymbol{\theta} \in \mathcal{M}$, it follows that if $(X,Y) \sim P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$, with $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (9), then the random variable

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

with the function ℓ in (2), possesses a mean, variance, and third moment given by $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, and $\mathsf{J}^{(3)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, respectively. In particular, the cumulant generating function of the random variable $W_{\boldsymbol{\theta}}$, which is denoted by $\mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}} : \mathbb{R}\to\mathbb{R}$, satisfies

$$\mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y)\right), \quad (30)$$

where the function ℓ is defined in (2). It is also proved in [4] that $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) = \mathsf{R}_{\boldsymbol{\theta}}\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$, with the functional $\mathsf{R}_{\boldsymbol{\theta}}$ defined in (4). The next theorem provides an upper bound on the cumulant generating function $\mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}$ in (30).

Theorem 5: For all $t \in \{\alpha\in\mathbb{R} : \mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}(\alpha) < +\infty\}$, the function $\mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}$ in (30) satisfies the following inequality:

$$\mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}(t) \leq t\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) + \frac{1}{2}t^2\zeta^2_{P_S,\boldsymbol{\theta}}, \quad (31)$$

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

$$\zeta_{P_S,\boldsymbol{\theta}} \triangleq \sup\left\{\sqrt{\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}(\xi)} : \xi \in \left(-\infty, b-\frac{1}{\beta}\right)\right\}, \quad (32)$$

with $b \triangleq \sup\{\alpha\in\mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}(\alpha) < +\infty\}$, and the functions $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}$ and $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}$ are the first and the second derivatives of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (7).

Proof: The proof is presented in [4]. ■

The relevance of Theorem 5 lies on the fact that it implies that the random variable $W_{\boldsymbol{\theta}}$ in (29) is a sub-Gaussian random variable with sub-Gaussian parameter $\zeta_{P_S,\boldsymbol{\theta}}$ in (32). An interesting discussion on the impact of the random variable $W_{\boldsymbol{\theta}}$ being sub-Gaussian is presented in [7, Theorem 1]. This observation leads to insightful guidelines for algorithm design.

V. (ϵ,δ)-Robustness

An important observation about the dependence of the WCDG probability distribution $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (9) on the parameter γ in (10) is that $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is the measure that induces the largest expected loss within the neighborhood of P_S, defined as

$$\mathcal{N}(P_S,\gamma) \triangleq \{P\in\triangle(\mathcal{X}\times\mathcal{Y}) : D(P\|P_S) \leqslant \gamma\}. \quad (33)$$

The following lemma provides insights on how $P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}$ changes according to the expansion or the contraction of the neighborhood.

Lemma 2: The relative entropy $D\left(P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\|P_S\right)$ in (10) satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\beta}D\left(P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\|P_S\right) = -\frac{1}{\beta^3}\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) \leq 0, \quad (34)$$

where $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ is the variance of $W_{\boldsymbol{\theta}}$ in (29). Moreover, the inequality is strict if and only if the function ℓ in (2) is separable with respect to P_S.

Proof: The proof is presented in [4]. ■

From Lemma 2, it is shown that there exists a bijection between γ and β through (10). More interestingly, the variations of $D\left(P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\|P_S\right)$ and the expected loss $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ with respect to β have the following connection: $\frac{\mathrm{d}}{\mathrm{d}\beta}D\left(P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\|P_S\right) = \frac{1}{\beta}\frac{\mathrm{d}}{\mathrm{d}\beta}\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, which implies that both $D\left(P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\|P_S\right)$ and $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ derease with β, while their rates of change might differ. For instance, for larger values of β with $\beta > 1$, the variation of $D\left(P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\|P_S\right)$ is larger than $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$. The following definition describes a performance metric for a given data-generating probability measure $P \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n\right)$ and a model $\boldsymbol{\theta} \in \mathcal{M}$ that leverages the observations above and provides guidelines for the choice of the values of β (or γ).

Definition 3 ((ϵ,δ)-Robustness): Given a pair of positive reals (ϵ,δ) with $\epsilon < 1$, a model $\boldsymbol{\theta} \in \mathcal{M}$ is said to be (ϵ,δ)-robust to a probability measure $P \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n\right)$, if

$$P\left(\{\boldsymbol{z} \in (\operatorname{supp} P_S)^n : \mathsf{L}(\boldsymbol{z},\boldsymbol{\theta}) \geq \delta\}\right) \leq \epsilon. \quad (35)$$

This notion of robustness enables the study of the performance guarantees that a model $\boldsymbol{\theta}$ yields when faced with data generated by the WCDG probability measure for specific parameters β and P_S. An important issue that arises from this definition is the characterization of the largest value of γ (or smallest value of β) that achieves (ϵ,δ)-robustness, i.e. how much can the WCDG probability measure deviate from the reference P_S while the guarantee still holds. The following theorem establishes a condition for the (ϵ,δ)-robustness to hold for a given model $\boldsymbol{\theta}$.

Theorem 6: The probability measure $P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}$ in (9) satisfies that for all $\delta > \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, with $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ being the mean of $W_{\boldsymbol{\theta}}$ in (29),

$$P^{(P_S,\beta)}_{\hat{\boldsymbol{Z}}|\mathbf{\Theta}=\boldsymbol{\theta}}\Big(\{\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n : \mathsf{L}\left(\boldsymbol{\theta},\boldsymbol{z}\right) \geqslant \delta\}\Big)$$
$$\leq \exp\left(-nD\left(P^{(P_S,\beta^*)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\|P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}\right)\right), \quad (36)$$

where $\beta^* \in (0,\beta) \cap \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (8), satisfies

$$\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta^*}\right) = \delta; \quad (37)$$

the function L is defined in (3); and the measure $P^{(P_S,\beta)}_{\hat{\boldsymbol{Z}}|\mathbf{\Theta}=\boldsymbol{\theta}} \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n\right)$ is a product measure formed by $P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}$.

Proof: The proof is presented in [4]. ■

Theorem 6 describes the (ϵ,δ)-robustness of a model $\boldsymbol{\theta}$ in terms of another WCDG probability distribution $P^{(P_S,\beta^*)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}$, where $\beta^* < \beta$. Interestingly, the WCDG probability measure $P^{(P_S,\beta^*)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}$ induces an expected loss that is equal to δ and greater than the expected loss induced by $P^{(P_S,\beta)}_{\hat{Z}|\mathbf{\Theta}=\boldsymbol{\theta}}$.

The following theorem provides a (ϵ,δ)-robust guarantee to any product probability measure formed by a measure in the neighborhood $\mathcal{N}\left(P_S,\gamma\right)$ in (33).

Theorem 7: For all $\boldsymbol{\theta} \in \mathcal{M}$ and for all $P_Z \in \mathcal{N}\left(P_S,\gamma\right)$, with $\mathcal{N}\left(P_S,\gamma\right)$ in (33) and β in (10), it follows that for all $\delta > \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, with $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ being the mean of $W_{\boldsymbol{\theta}}$ in (29),

$$P_{\boldsymbol{Z}}\left(\{\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n : \mathsf{L}\left(\boldsymbol{\theta},\boldsymbol{z}\right) \geqslant \delta\}\right) \leq \frac{1}{\delta}\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), \quad (38)$$

where the function ℓ is defined in (2) and the product probability measure $P_{\boldsymbol{Z}}$ is formed by P_Z.

Proof: The proof is presented in [4]. ■

The relevance of Theorem 7 is that given a model $\boldsymbol{\theta}$, it establishes that for all $\delta > \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ and $\epsilon_\delta \triangleq \frac{1}{\delta}\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, such a model $\boldsymbol{\theta}$ is (ϵ_δ,δ)-robust to all probability measures in $\mathcal{N}\left(P_S,\gamma\right)$, with $\mathcal{N}\left(P_S,\gamma\right)$ in (33). This observation raises the question on whether performing model selection based on the minimization of $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ is an alternative to classical approaches based on empirical risk minimization (ERM) as in [35]; or statistical ERM as in [3] and [36].

VI. Conclusions and Remarks

The first explicit expression in terms of information measures of the EGG for statistical learning algorithms has been presented. This expression has led to an upper bound on the EGG consisting of the sum of the mutual and lautum information between the models and the datasets, up to a constant factor. Such a bound is tight at least for one Gibbs algorithm, whose parameters are chosen to satisfy a particular condition. This observation reveals the central role of the Gibbs algorithm in statistical machine learning. Fundamentally, an exploration into the EGG of the Gibbs algorithm facilitates the derivation of overarching insights applicable to any statistical learning algorithm. Finally, the study of the properties of the WCDG probability measure leads to new machine learning algorithm performance metrics and insightful guidelines for algorithm design.

Acknowledgment

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

REFERENCES

[1] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, "Generalization analysis of machine learning algorithms via the worst-case data-generating probability measure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2024.

[2] ——, "The Worst-Case Data-Generating Probability Measure," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9515, Aug. 2023.

[3] 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.

[4] X. Zou, S. M. Perlaza, I. Esnaola, E. Altman, and H. V. Poor, "An exact characterization of the generalization error for machine learning algorithms," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9539, Jan. 2024.

[5] 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.

[6] 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.

[7] 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.

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

[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] 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.

[12] 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.

[13] 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.

[14] ——, "Empirical risk minimization with relative entropy regularization Type-II," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9575, May. 2023.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] I. Casado, L. A. Ortega, A. R. Masegosa, and A. Pérez, "Pac-bayes-chernoff bounds for unbounded losses," arXiv preprint arXiv:2401.01148, Jan. 2024.

[24] F. Hellström, G. Durisi, B. Guedj, and M. Raginsky, "Generalization bounds: Perspectives from information theory and PAC-Bayes," arXiv preprint arXiv:2309.04381, Sep. 2023.

[25] L. P. Barnes, A. Dytso, and H. V. Poor, "Improved information-theoretic generalization bounds for distributed, federated, and iterative learning," *Entropy*, vol. 24, no. 9, 2022.

[26] G. Aminian, Y. Bu, L. Toni, M. R. D. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," *IEEE Transactions on Information Theory*, vol. 70, no. 1, pp. 632–655, 2024.

[27] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.

[28] Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization," *Optimization Online*, vol. 1, no. 2, p. 9, 2013.

[29] J. Lee and M. Raginsky, "Minimax Statistical Learning with Wasserstein distances," *Advances in Neural Information Processing Systems*, vol. 31, pp. 2687–2696, 2018.

[30] S. Mazuelas, Y. Shen, and A. Pérez, "Generalized maximum entropy for supervised classification," *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2530–2550, Jan. 2022.

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

[32] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

[33] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization: Optimality and sensitivity," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, Jul. 2022, pp. 684–689.

[34] K. Sun, S. M. Perlaza, and A. Jean-Marie, "2×2 Zero-sum games with commitments and Noisy Observations," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023.

[35] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," *Advances in neural information processing systems*, vol. 30, pp. 4151–4161, Dec. 2017.

[36] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Empirical risk minimization with f-divergence regularization in statistical learning," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9521, Oct. 2023.