An Exact Characterization of the Generalization Error of Machine Learning Algorithms

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

An Exact Characterization of the Generalization Error of Machine Learning Algorithms

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

***Abstract*—In this paper, exact expressions for the generalization error of general machine learning algorithms are presented in terms of information measures. These expressions can be broadly divided into two classes. The first one involves a worst-case data generating (WCDG) probability measure, while the second one involves a Gibbs algorithm. These expressions are formed by the sum of three terms. The first term is either a mutual or lautum information induced by the algorithm between the models and the datasets. The second and third terms are different for the two classes. In the first class, the second term compares via Kullback-Leibler (KL) divergence the posterior probability distribution induced by the algorithm on the datasets with a WCDG probability measure; and the third term compares via KL divergence the marginal of such posterior with the same WCDG probability measure. In the second class, the second term compares via KL divergence the distributions induced on the models by the algorithm and a Gibbs algorithm when both have been training with the same dataset; the third term compares via KL divergence the marginal of the distribution on the models induced by the algorithm with the same Gibbs algorithm. For Gibbs algorithms, the second and third terms jointly become either mutual or lautum information, recovering existing results.**

I. Introduction

An exact expression for the generalization error is known only for the Gibbs algorithm [1]–[4]. For machine learning algorithms, other than the Gibbs algorithm, the generalization error is often characterized via upper bounds. Complete literature reviews of these bounds are presented in [5] and [6]. Nonetheless, it has been pointed out that such upper bounds are often loose, see for instance [7] and [8].

This paper introduces exact expressions for the generalization error that apply to any algorithm. Due to space constraints, the proofs of the main results are presented in [9].

Notation

Sets are denoted by caligraphic uppercase letters, except sets of probability measures. Given a set $\mathcal{M}$ and a sigma-field $\mathscr{F}$ on $\mathcal{M}$, the set of all probability measures that can be defined on the measurable space $(\mathcal{M}, \mathscr{F})$ is denoted by $\triangle(\mathcal{M}, \mathscr{F})$ or simply $\triangle(\mathcal{M})$, when the sigma-field is fixed in the analysis. If $\mathcal{M} \subseteq \mathbb{R}$, then the Borel sigma-field on $\mathcal{M}$ is denoted by $\mathscr{B}(\mathcal{M})$. The subset of measures in $\triangle(\mathcal{M})$ that are absolutely continuous with a probability measure Q is denoted by $\triangle_Q(\mathcal{M})$. Given a measure $P \in \triangle_Q(\mathcal{M})$, the Radon-Nikodym derivative of P with respect to Q is denoted by $\frac{\mathrm{d}P}{\mathrm{d}Q}$. Given a set $\mathcal{N}$, the set of all probability measures defined on $(\mathcal{M}, \mathscr{F})$ conditioned on an element of $\mathcal{N}$ is denoted by $\triangle(\mathcal{M}|\mathcal{N})$. More specifically, given a measure $P_{M|N} \in \triangle(\mathcal{M}|\mathcal{N})$ and $n \in \mathcal{N}$, the measure $P_{M|N=n}$ is in $\triangle(\mathcal{M})$. Given a probability measure $P_N \in \triangle(\mathcal{N})$, the mutual information [10] and lautum information [11] induced by the measures $P_{M|N}$ and P_N are denoted respectively by

$$I\left(P_{M|N}; P_N\right) \triangleq \int D\left(P_{M|N=n} \| P_M\right) \mathrm{d}P_N(n), \text{ and} \quad (1)$$

$$L\left(P_{M|N}; P_N\right) \triangleq \int D\left(P_M \| P_{M|N=n}\right) \mathrm{d}P_N(n), \quad (2)$$

with $P_M \in \triangle(\mathcal{M})$ the marginal induced by $P_{M|N}$ and P_N; and $D(\cdot\|\cdot)$ the Kullback-Leibler divergence.

II. Problem Formulation

Let $\mathcal{M}$, $\mathcal{X}$ and $\mathcal{Y}$, with $\mathcal{M} \subseteq \mathbb{R}^d$, 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*. Let the probability measure

$$P_Z \in \triangle(\mathcal{X} \times \mathcal{Y}) \quad (3)$$

be the one from which data points are independently sampled. While the probability measure P_Z is unknown in a practical setting, a finite set of datapoints is made available via a *training dataset*. A training dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$ is a tuple of n datapoints of the form:

$$\boldsymbol{z} = \left((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \right) \in (\mathcal{X} \times \mathcal{Y})^n. \quad (4)$$

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 $f(\boldsymbol{\theta}, x)$. The function $f(\boldsymbol{\theta}, \cdot) : \mathcal{X} \to \mathcal{Y}$ might for instance represent a neural network whose weights are vectorized into $\boldsymbol{\theta}$. Let the function $\hat{\ell} : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty]$, be the risk or loss function. Such a function is often of the form $\hat{\ell}(\mu, \nu) = |\mu - \nu|^p$, with $p \geqslant 1$; or $\hat{\ell}(\mu, \nu) = \mathbb{1}_{\{\mu = \nu\}}$. In general, given a labelled

pattern $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the risk induced by a model $\boldsymbol{\theta} \in \mathcal{M}$ is $\hat{\ell}\left(f(\boldsymbol{\theta}, x), y\right)$. In the following, the risk function $\hat{\ell}$ is assumed to be a general nonnegative function and for all $y \in \mathcal{Y}$, $\hat{\ell}\left(y, y\right) = 0$. For the ease of notation, consider the function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{M} \to [0, +\infty]$ such that

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

A. The Problem of Supervised Machine Learning

The problem of supervised machine learning is essentially devising an algorithm that takes as input a training dataset, as $\boldsymbol{z}$ in (4), and outputs a model $\boldsymbol{\theta}$, with certain probability. This notion is formally defined as follows.

Definition 1 (Algorithm): A conditional probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}} \in \triangle\left(\mathcal{M}|\left(\mathcal{X} \times \mathcal{Y}\right)^n\right)$ is said to represent a supervised machine learning algorithm if it satisfies the following conditions:

- (a) For all $\boldsymbol{z} \in \left(\mathcal{X} \times \mathcal{Y}\right)^n$, the set-function $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}} : \mathscr{B}\left(\mathcal{M}\right) \to [0, 1]$ is a probability measure in $\triangle\left(\mathcal{M}\right)$.
- (b) For all $\mathcal{A} \in \mathscr{B}\left(\mathcal{M}\right)$, the function $h_{\mathcal{A}} : \left(\mathcal{X} \times \mathcal{Y}\right)^n \to [0, 1]$ that satisfies $h_{\mathcal{A}}(\boldsymbol{z}) = P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})$ is measurable with respect to the measurable spaces $\left(\left(\mathcal{X} \times \mathcal{Y}\right)^n, \mathscr{Z}_n\right)$ and $\left([0, 1], \mathscr{B}\left([0, 1]\right)\right)$, where $\mathscr{Z}_n$ is a σ-algebra on $\left(\mathcal{X} \times \mathcal{Y}\right)^n$.

Let $P_{\boldsymbol{\Theta}|\boldsymbol{Z}} \in \triangle\left(\mathcal{M}|\left(\mathcal{X} \times \mathcal{Y}\right)^n\right)$ be an algorithm. Hence, the instance of such an algorithm trained upon the dataset $\boldsymbol{z}$ in (4) is denoted by $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, which is simply a probability measure in $\triangle\left(\mathcal{M}\right)$.

B. Empirical Risk and True Risk

The empirical risk is the average risk induced by a model among all datapoints of a given dataset. Let the function $\mathsf{L} : \left(\mathcal{X} \times \mathcal{Y}\right)^n \times \mathcal{M} \to [0, +\infty]$ be such that

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

where $\boldsymbol{z}$ is a dataset of the form in (4) and the function ℓ is defined in (5). Using this notation, the empirical risk is defined as follows.

Definition 2 (Empirical Risk): The *empirical risk* induced by a model $\boldsymbol{\theta} \in \mathcal{M}$ with respect to the dataset $\boldsymbol{z}$ in (4) is $\mathsf{L}\left(\boldsymbol{z}, \boldsymbol{\theta}\right)$ in (6).

In the following, the expectation of ℓ in (5) with respect to a probability measure P in $\triangle\left(\mathcal{X} \times \mathcal{Y}\right)$, for a fixed model $\boldsymbol{\theta}$ is denoted using the functional $\mathsf{R}_{\boldsymbol{\theta}} : \triangle\left(\mathcal{X} \times \mathcal{Y}\right) \to [0, +\infty]$, which satisfies

$$\mathsf{R}_{\boldsymbol{\theta}}\left(P\right) = \int \ell(x, y, \boldsymbol{\theta}) \mathrm{d}P(x, y). \tag{7}$$

Using this notation, the true risk is defined as follows.

Definition 3 (True Risk): The *true risk* induced by a model $\boldsymbol{\theta} \in \mathcal{M}$, under the assumption that datasets are formed by i.i.d. datapoints sampled from P_Z in (3), is $\mathsf{R}_{\boldsymbol{\theta}}\left(P_Z\right)$, where the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined (7).

Using the notion of *type induced by a dataset*, the empirical risk $\mathsf{L}\left(\boldsymbol{z}, \boldsymbol{\theta}\right)$ in (6) can be written in terms of the function $\mathsf{R}_{\boldsymbol{\theta}}$ in (7). A type is defined as follows.

Definition 4 (Type of a dataset): The type induced by the dataset $\boldsymbol{z}$ in (4) is a probability measure in $\triangle\left(\mathcal{X} \times \mathcal{Y}\right)$, denoted by $P_{\boldsymbol{z}}$, such that for all measurable subsets $\mathcal{A}$ of $\mathcal{X} \times \mathcal{Y}$,

$$P_{\boldsymbol{z}}\left(\mathcal{A}\right) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{(x_t, y_t) \in \mathcal{A}\}}. \tag{8}$$

The empirical risk $\mathsf{L}\left(\boldsymbol{z}, \boldsymbol{\theta}\right)$ in (6) satisfies

$$\mathsf{L}_{\boldsymbol{z}}(\boldsymbol{\theta}) = \mathsf{R}_{\boldsymbol{\theta}}\left(P_{\boldsymbol{z}}\right), \tag{9}$$

where the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (7); and the probability measure $P_{\boldsymbol{z}}$ is the type in (8) induced by the dataset $\boldsymbol{z}$. See for instance, [4, Lemma 6.1].

C. Generalization Error

Generalization metrics, including the generalization error, are often defined in terms of differences of certain quantities. In the following, those differences are referred to as *gaps* and are quantified via the following functional: $\mathsf{G} : \mathcal{M} \times \triangle\left(\mathcal{X} \times \mathcal{Y}\right) \times \triangle\left(\mathcal{X} \times \mathcal{Y}\right) \to \mathbb{R}$, such that

$$\mathsf{G}\left(\boldsymbol{\theta}, P_1, P_2\right) = \mathsf{R}_{\boldsymbol{\theta}}\left(P_1\right) - \mathsf{R}_{\boldsymbol{\theta}}\left(P_2\right). \tag{10}$$

Let the functional $\overline{\overline{G}} : \triangle\left(\mathcal{M}|\left(\mathcal{X} \times \mathcal{Y}\right)^n\right) \times \triangle\left(\mathcal{X} \times \mathcal{Y}\right)^n \to \mathbb{R}$ be such that

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) = \int\int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) \mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}) \mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{11}$$

where $P_{\boldsymbol{z}}$ is the type induced by the dataset $\boldsymbol{z}$; the functional G is defined in (10); the probability measure $P_{\boldsymbol{Z}} \in \triangle\left(\left(\mathcal{X} \times \mathcal{Y}\right)^n\right)$ is a product distribution formed by P_Z in (3); and $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is an algorithm (Definition 1).

The term $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (11) is the difference between the true risk $\mathsf{R}_{\boldsymbol{\theta}}\left(P_Z\right)$ and the empirical risk $\mathsf{R}_{\boldsymbol{\theta}}\left(P_{\boldsymbol{z}}\right)$ induced by a model $\boldsymbol{\theta}$ with respect to a dataset $\boldsymbol{z}$. Hence, $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ is the expectation of such a difference with respect to a joint probability measure in $\triangle\left(\mathcal{M} \times \left(\mathcal{X} \times \mathcal{Y}\right)^n\right)$ formed by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ and the product measure $P_{\boldsymbol{Z}}$. Using this notation, the *generalization error* of a machine learning algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}} \in \triangle\left(\mathcal{M}|\left(\mathcal{X} \times \mathcal{Y}\right)^n\right)$ is defined as follows.

Definition 5 (Generalization Error): The generalization error induced by an algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}} \in \triangle\left(\mathcal{M}|\left(\mathcal{X} \times \mathcal{Y}\right)^n\right)$, under the assumption that datasets are formed by i.i.d. datapoints sampled from P_Z in (3), is $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (11), with $P_{\boldsymbol{Z}}$ a product distribution formed by P_Z.

A more useful expression for the generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (11) can be obtained in terms of the conditional measure $P_{\boldsymbol{Z}|\boldsymbol{\Theta}} \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n|\mathcal{M}\right)$ and the measure $P_{\boldsymbol{\Theta}} \in \triangle(\mathcal{M})$, which are particular for the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ and the product measure $P_{\boldsymbol{Z}}$, as described hereunder. The measure $P_{\boldsymbol{\Theta}}$ satisfies for all measurable subsets $\mathcal{A}$ of $\mathcal{M}$,

$$P_{\boldsymbol{\Theta}}(\mathcal{A}) = \int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{A})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z})\,; \tag{12}$$

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

$$P_{\boldsymbol{Z}}(\mathcal{B}) = \int P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{B})\,\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,. \tag{13}$$

Note that under the assumption that datasets are formed by i.i.d. datapoints sampled from P_Z in (3), i.e., $P_{\boldsymbol{Z}}$ in (11) is a product measure formed by P_Z, it follows that for all $\boldsymbol{\theta} \in \mathcal{M}$, the measure $P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is a product measure formed by a probability measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}} \in \triangle(\mathcal{X}\times\mathcal{Y})$, which for all measurable sets of the form $\mathcal{A}_1\times\mathcal{A}_2\times\cdots\times\mathcal{A}_n$ in $(\mathcal{X}\times\mathcal{Y})^n$, satisfies

$$P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}_1\times\mathcal{A}_2\times\ldots\times\mathcal{A}_n) = \prod_{t=1}^{n} P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}_t)\,, \tag{14}$$

and for all $t\in\{1,2,\ldots,n\}$,

$$P_Z(\mathcal{A}_t) = \int P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}_t)\,\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,, \tag{15}$$

where the probability measure $P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is defined in (12). Using this notation, the generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (11) can be re-written as follows:

$$\begin{aligned}
&\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})\\
&= \int\int G(\boldsymbol{\theta},P_Z,P_{\boldsymbol{z}})\mathrm{d}P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\boldsymbol{z})\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), && (16)\\
&= \int\int\left(\int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_Z(x,y)\right)\mathrm{d}P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\boldsymbol{z})\,\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\\
&\quad -\int\int\left(\int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_{\boldsymbol{z}}(x,y)\right)\mathrm{d}P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\boldsymbol{z})\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) && (17)\\
&= \int\int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_Z(x,y)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}),\\
&\quad -\int\int\left(\frac{1}{n}\sum_{t=1}^{n}\ell(\boldsymbol{\theta},x_t,y_t)\right)\mathrm{d}P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\boldsymbol{z})\,\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) && (18)\\
&= \int\int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_Z(x,y)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\\
&\quad -\int\int \ell(\boldsymbol{\theta},x,y)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\boldsymbol{z})\,\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) && (19)\\
&= \int \mathsf{G}\left(\boldsymbol{\theta},P_Z,P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\,, && (20)
\end{aligned}$$

where the equality in (16) follows from (12) and (13); the equality in (17) follows from (10); the equality in (18) follows from (8) and (9); the equality in (19) follows from the fact that the probability measure $P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is a product measure as in (14); and the equality in (20) follows from (10).

The equality in (20) turns out to be instrumental in the remaining of this work.

III. The Gibbs Algorithm

A typical example of a statistical learning algorithm is the Gibbs algorithm, which is parametrized by a positive real λ and by a probability measure $Q \in \triangle(\mathcal{M},\mathscr{B}(\mathcal{M}))$ [3]. The probability measure representing such an algorithm is denoted by $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}} \in \triangle\left(\mathcal{M}|(\mathcal{X}\times\mathcal{Y})^n\right)$. When the Gibbs algorithm $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is trained with the training dataset $\boldsymbol{z}$ in (4), it satisfies for all $\boldsymbol{\theta}\in\operatorname{supp}Q$,

$$\frac{\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}}{\mathrm{d}Q}(\boldsymbol{\theta}) = \exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right) - \frac{1}{\lambda}\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\right), \tag{21}$$

where the function $K_{Q,\boldsymbol{z}}:\mathbb{R}\to\mathbb{R}$ satisfies

$$K_{Q,\boldsymbol{z}}(t) = \log\left(\int \exp\left(t\,\mathsf{L}(\boldsymbol{z},\boldsymbol{\theta})\right)\mathrm{d}Q(\boldsymbol{\theta})\right), \tag{22}$$

and the function L is defined in (6).

The generalization error induced by the Gibbs algorithm with parameters Q and λ, under the assumption that datasets are sampled from a product distribution $P_{\boldsymbol{Z}} \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n\right)$ formed by the measure P_Z is

$$\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}), \tag{23}$$

where the fuctional $\overline{\overline{G}}$ is defined in (11). Such a generalization error satisfies the following property.

Lemma 1 (Generalization Gap of the Gibbs Algorithm): The generalization error $\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (23) satisfies

$$\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) = \lambda\Big(I\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\Big) + L\Big(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\Big)\Big). \tag{24}$$

Proof: This result has been proved before in the case in which Q is a probability measure in [1]; and in the more general case in which Q is a σ-finite measure in [3]. A proof for the particular case in which Q is a probability measure and datasets are formed by independent and identically distributed datapoints is presented in [4] ∎

IV. The Worst-Case Data-Generating Probability Measure and Generalization Error

The WCDG probability measure was introduced in [4] and further studied in [12]. The WCDG probability measure, which is parametrized by a real $\beta>0$ and a probability measure P_S in $\triangle(\mathcal{X}\times\mathcal{Y})$, is specific for a given model $\boldsymbol{\theta}$ and thus, denoted by $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$. Note that $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is a measure in $\triangle_{P_S}(\mathcal{X}\times\mathcal{Y})$. An explicit expression for $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ requires defining the following function. Let $\mathsf{J}_{P_S,\boldsymbol{\theta}}:\mathbb{R}\to\mathbb{R}$ be the function

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_S(x,y)\right), \tag{25}$$

where the function ℓ is defined in (5). The probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfies for all $(x,y) \in \operatorname{supp} P_S$,

$$\frac{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y) = \exp\left(\frac{1}{\beta}\ell(\boldsymbol{\theta},x,y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right), \quad (26)$$

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is defined in (25). The existence of $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is ensured as long as $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, where

$$\mathcal{J}_{P_S,\boldsymbol{\theta}} \triangleq \left\{ t \in (0,+\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) < +\infty \right\}. \quad (27)$$

The relevance of the WCDG probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ stems from the fact that a closed-form expression for the term $\mathsf{G}(\boldsymbol{\theta},P_1,P_2)$ in (10) can be obtained in terms of $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$. The following theorem shows such an expression.

Lemma 2 (Theorem 8 in [12]): The gap $G(\boldsymbol{\theta},P_1,P_2)$ in (10) satisfies

$$G(\boldsymbol{\theta},P_1,P_2) = \beta\Big(D\left(P_2\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_1\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_2\|P_S\right) + D\left(P_1\|P_S\right)\Big), \quad (28)$$

where the parameter β, the model $\boldsymbol{\theta}$, and the measures P_S and $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfy (26); and additionally, $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with the set $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ defined in (27).

In Lemma 2, the choice of β and P_S is flexible up to some constraints. Note that P_S can be any measure in $\triangle(\mathcal{X}\times\mathcal{Y})$, as long as both P_1 and P_2 are absolutely continuous with P_S; and β is in the set $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (27). This flexibility is instrumental to obtain the main results. More specifically, the remaining of this work consists in studying the following expression, which follows immediately from the equality in (20) and Lemma 2,

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}) = \beta\int\left(D\Big(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big) - D\Big(P_Z\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big) - D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_S\right) + D(P_Z\|P_S)\right)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \quad (29)$$

where for all $\boldsymbol{\theta} \in \operatorname{supp} P_{\boldsymbol{\Theta}}$, the probability measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is absolutely continuous with P_S; and for all $\boldsymbol{\theta} \in \operatorname{supp} P_{\boldsymbol{\Theta}}$, $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with the set $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ defined in (27).

V. Main Results

The main results are closed-form expressions for the generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (11), where $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is an arbitrary algorithm. Some expressions involve a WCDG probability measure, while others involve a Gibbs algorithm.

A. Expressions Involving a WCDG Measure

The following theorem shows one closed-form expression for the generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (11) involving a mutual information.

Theorem 1: The generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (11) satisfies:

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}) = -\frac{\beta}{n}I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right) + \beta\int D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) - \beta\int D\left(P_Z\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \quad (30)$$

where the measure $P_{\boldsymbol{\Theta}}$ is defined in (12); the conditional measure $P_{Z|\boldsymbol{\Theta}}$ is defined in (15); the measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is the WCDG probability measure in (26), with P_S being such that for all $\boldsymbol{\theta} \in \operatorname{supp} P_{\boldsymbol{\Theta}}$, the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is absolutely continuous with P_S; and β being such that for all $\boldsymbol{\theta} \in \operatorname{supp} P_{\boldsymbol{\Theta}}$, $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with the set $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ defined in (27).

From Theorem 1, it follows that the generalization error can be strictly negative for a special class of algorithms. The following lemma formalizes this observation.

Lemma 3: Let $P_{\boldsymbol{\Theta}|\boldsymbol{Z}} \in \triangle\left(\mathcal{M}|\left(\mathcal{X}\times\mathcal{Y}\right)^n\right)$ be an algorithm such that for all $\boldsymbol{\theta} \in \mathcal{M}$, the corresponding probability measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (15) is identical to the WCDG probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (26), for some reference P_S and some real $\beta > 0$. Then, the expected generalization gap $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (11) satisfies:

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}) = -\frac{\beta}{n}\left(I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right) + L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)\right) \leqslant 0. \quad (31)$$

Note that if the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in Theorem 1 is such that models are chosen independently of the training dataset, i.e., for all $(\boldsymbol{u},\boldsymbol{v}) \in (\operatorname{supp} P_Z)^n \times (\operatorname{supp} P_Z)^n$, it holds that $D\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{u}}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{v}}\right) = 0$, then, $I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right) = 0$ and moreover, for all $\boldsymbol{\theta} \in \operatorname{supp} P_{\boldsymbol{\Theta}}$, it holds that

$$D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) = D\left(P_Z\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right), \quad (32)$$

which implies that $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}) = 0$. That is, algorithms that neglect the training datasets induce zero generalization error, which is a well-known result.

The following lemma introduces another expression for the generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (11) involving a lautum information.

Lemma 4 (Theorem 10 in [12]): The generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (11) satisfies:

$$\begin{aligned}\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})=&\frac{\beta}{n}L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)\\&+\beta\int D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)\mathrm{d}P_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\right)\\&-\beta\int D\left(P_{Z}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)\mathrm{d}P_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\right)\end{aligned}\quad(33)$$

where the measure $P_{\boldsymbol{\Theta}}$ is defined in (12); the conditional measure $P_{Z|\boldsymbol{\Theta}}$ is defined in (15); the measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}$ is the WCDG probability measure in (26), with reference measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ and β such that for all $\boldsymbol{\theta}\in\operatorname{supp}P_{\boldsymbol{\Theta}}$, $\beta\in\mathcal{J}_{P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}$, with the set $\mathcal{J}_{P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}$ defined as in (27).

In [12, Theorem 11], it has been suggested that there might exist a β that satisfies for all $\boldsymbol{\theta}\in\operatorname{supp}P_{\boldsymbol{\Theta}}$,

$$D\left(P_{Z}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)=0,\quad(34)$$

which leads to the equality $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) = \frac{\beta}{n}\left(I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)+L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)\right)$, from the fact that in such a case for all $\boldsymbol{\theta}\in\operatorname{supp}P_{\boldsymbol{\Theta}}$, $D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right) = D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_{Z}\right)$, and thus,

$$\int D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)\mathrm{d}P_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\right)=\frac{1}{n}I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}\right).\quad(35)$$

This observation together with Lemma 2 might lead to conjecturing that the Gibbs algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}\in\triangle\left(\mathcal{M}|\left(\mathcal{X}\times\mathcal{Y}\right)^{n}\right)$ in (21), with $\beta=n\lambda$, satisfies the equality in (34). The following lemma shows that it is true under a very restrictive condition. Before presenting such a lemma, note that the measures $P_{\boldsymbol{\Theta}}$ in (12); $P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (13); and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (14) that are specific to the Gibbs algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}$ in (21) are denoted $P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$; $P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}$; and $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}$, respectively.

Lemma 5: Let $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}\in\triangle\left(\mathcal{M}|\left(\mathcal{X}\times\mathcal{Y}\right)^{n}\right)$ be the Gibbs algorithm in (21). Then, for all $\boldsymbol{\theta}\in\operatorname{supp}P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$,

$$D\left(P_{Z}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)},n\lambda\right)}\right)=0,\quad(36)$$

if and only if

$$P_{\boldsymbol{Z}}\left(\left\{\boldsymbol{z}\in\left(\mathcal{X}\times\mathcal{Y}\right)^{n}:K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)=c\right\}\right)=1,\quad(37)$$

for some real $c\leqslant 0$, where the function $K_{Q,\boldsymbol{z}}$ is defined in (22).

The equality in (37) implies that the function $K_{Q,\boldsymbol{z}}$ is invariant with respect to $\boldsymbol{z}$, almost surely with $P_{\boldsymbol{Z}}$. Note that this is always true if for all $\boldsymbol{\theta}\in\mathcal{M}$, the function $\ell\left(\cdot,\cdot,\boldsymbol{\theta}\right)$ in (5) is nonseparable with respect to P_Z [12, Definition 2]. That is, for all $\boldsymbol{\theta}\in\mathcal{M}$,

$$P_{Z}\left(\left\{(x,y)\in\mathcal{X}\times\mathcal{Y}:\ell\left(x,y,\boldsymbol{\theta}\right)=a(\boldsymbol{\theta})\right\}\right)=1,\quad(38)$$

for some function $a:\mathcal{M}\to[0,+\infty)$. Nonetheless, this case is uninteresting in the context of machine learning, as the loss function ℓ in (5) is invariant with respect to the datapoints almost surely with P_Z in (3). Alternatively, if the function ℓ is separable, it is not clear whether the equality in (37) can be met.

The following lemma presents a property of the Gibbs algorithm that leads to the equality $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}, P_{\boldsymbol{Z}}) = \lambda\left(I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\right)+L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\right)\right)$, which is the claim of Lemma 1, without the need of verifying (36).

Lemma 6: Let $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}\in\triangle\left(\mathcal{M}|\left(\mathcal{X}\times\mathcal{Y}\right)^{n}\right)$ be the Gibbs algorithm in (21). Let $\beta\triangleq n\lambda$ be such that for all $\boldsymbol{\theta}\in\operatorname{supp}P_{\boldsymbol{\Theta}}^{(Q,\lambda)}$, $\beta\in\mathcal{J}_{P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)},\boldsymbol{\theta}}$, with the set $\mathcal{J}_{P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)},\boldsymbol{\theta}}$ defined as in (27). Then,

$$\begin{aligned}\frac{1}{n}I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}; P_{\boldsymbol{Z}}\right)=&\int D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)},\beta\right)}\right)\mathrm{d}P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\left(\boldsymbol{\theta}\right)\\&-\int D\left(P_{Z}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(Q,\lambda)},\beta\right)}\right)\mathrm{d}P_{\boldsymbol{\Theta}}^{(Q,\lambda)}(\boldsymbol{\theta}).\end{aligned}\quad(39)$$

Note that Lemma 6, together with Theorem 4, form an alternative proof for Lemma 1.

B. Expressions Involving a Gibbs Algorithm

The following lemma introduces another expression for the generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (11) involving a mutual information and a Gibbs algorithm.

Theorem 2: The generalization error $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})$ in (11) satisfies:

$$\begin{aligned}&\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}})=\lambda I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)\\&+\lambda\int\left(D\Big(P_{\boldsymbol{\Theta}}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\Big)-D\Big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\Big)\right)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}),\end{aligned}\quad(40)$$

where the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}\in\triangle\left(\mathcal{M}|\left(\mathcal{X}\times\mathcal{Y}\right)^{n}\right)$ is defined in (21), with $\lambda>0$ and the measure Q being such that for all $\boldsymbol{z}\in\operatorname{supp}P_{\boldsymbol{Z}}$, the measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$ is absolutely continuous with Q.

Theorem 2 allows the calculation of the generalization error of an arbitrary algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ by comparing it with a Gibbs algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)}\in\triangle\left(\mathcal{M}|\left(\mathcal{X}\times\mathcal{Y}\right)^{n}\right)$, whose reference measure Q is chosen such that all the instances of the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, obtained with all possible training datasets $\boldsymbol{z}$, are absolutely continuous with Q.

ACKNOWLEDGMENT

This work is funded in part by the ANR Project PARFAIT under grant ANR-21-CE25-0013; in part by the H2020 RISE Project TESTBED2 under EU Grant 872172; in part by the U.S. National Science Foundation under Grant ECCS-2335876; in part by the C3.ai Digital Transformation Institute; and in part by Princeton Language + Intelligence.

REFERENCES

[1] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8106–8118, Dec. 2021.

[2] G. Aminian, Y. Bu, 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.

[3] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, 2024.

[4] 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.

[5] 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.

[6] B. Rodríguez Gálvez, "An information-theoretic approach to generalization theory," Ph.D. dissertation, KTH Royal Institute of Technology, Apr. 2024.

[7] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them," in *Proceedings of the International Conference on Learning Representations*, Addis Ababa, Ethiopia, Apr. 2020.

[8] M. Gastpar, I. Nachum, J. Shafer, and T. Weinberger, "Fantastic generalization measures are nowhere to be found," in *International Conference on Learning Representations (ICLR)*, Vienna, Austria, May 2024.

[9] 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.

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

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

[12] X. Zou, S. M. Perlaza, I. Esnaola, E. Altman, and H. V. Poor, "The worst-case data-generating probability measure in statistical learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 175–189, 2024.