The Worst-Case Data-Generating Probability Measure in Statistical Learning

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

The Worst-Case Data-Generating Probability Measure in Statistical Learning

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

***Abstract*—The worst-case data-generating (WCDG) probability measure is introduced as a tool for characterizing the generalization capabilities of machine learning algorithms. Such a WCDG probability measure is shown to be the unique solution to two different optimization problems: (a) The maximization of the expected loss over the set of probability measures on the datasets whose relative entropy with respect to a *reference measure* is not larger than a given threshold; and (b) The maximization of the expected loss with regularization by relative entropy with respect to the reference measure. Such a reference measure can be interpreted as a prior on the datasets. The WCDG cumulants are finite and bounded in terms of the cumulants of the reference measure. To analyze the concentration of the expected empirical induced by the WCDG probability measure, the notion of (ϵ, δ)-robustness of models is introduced. Closed-form expressions are presented for the sensitivity of the expected loss for a fixed model. These tools result in the characterization of a novel expression for the generalization error of arbitrary machine learning algorithms. This exact expression is provided in terms of the WCDG probability measure and leads to an upper bound 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. This finding reveals that an exploration into the generalization error of the Gibbs algorithm facilitates the derivation of overarching insights applicable to any machine learning algorithm.**

***Index Terms*—Supervised Machine Learning, Worst-Case, Generalization Gap, Relative Entropy, Gibbs Algorithm, and Sensitivity.**

I. Introduction

The problem of supervised machine learning is often formulated as an empirical risk minimization (ERM) problem in which the optimization domain is a set of models [3]. This formulation is based on the observation that the empirical risk is the expectation of the loss function with respect to the empirical probability measure induced by the training dataset. See for instance, Lemma 7. This empirical measure is known in the realm of information theory as a type [4] (see Definition 3). If the ground-truth data-generating (GTDG) probability measure were available, optimal models would be the minimizers of the expectation of the loss with respect to the GTDG probability measure, also known as the *true risk* or *population risk* [3]. From this perspective, the type induced by the training dataset is a replacement for the GTDG probability measure. That is, models are chosen as the minimizers of the empirical risk instead of minimizers of the true risk. Interestingly, large datasets might lead to types that are sufficiently *close* to the GTDG probability measure with high probability [5]. Typically, such a closeness is often measured in terms of a *statistical distance*, e.g., the relative entropy of the type with respect to the GTDG probability measure.

The driving idea in this work is that the GTDG probability measure is likely to be within a *neighbourhood* of the type induced by the training dataset. Side information might also lead to a *prior* on the data, and thus, the GTDG probability measure is also likely to be within a *neighbourhood* of such a prior. From this perspective, one can adopt a *reference measure*, which can be the type, a prior, or a mixture of both, and form a unique *neighbourhood* around such a *reference measure*. This *neighbourhood* includes all probability measures on the datasets that are at a *statistical distance* smaller than or equal to a given threshold. A robust choice of models is choosing them as the minimizers of the expectation of the loss with respect to the worst-case data-generating (WCDG) probability measure within such a neighbourhood. In this case, the WCDG probability measure is assumed to be the measure that maximizes the expectation of the loss for a fixed model. This problem formulation is a distributionally robust optimization (DRO) problem [6], [7] in which the *statistical distance* might be the relative entropy as in [8], a Wasserstein distance as in [9], [10], an f-divergence as in [11], [12], among others. This problem can also be formulated via the maximum entropy principle [13], [14] as in [15], [16]. Every choice of *statistical distance* leads to a WCDG probability measure with particular properties.

When the *statistical distance* is the relative entropy, the resulting WCDG probability measure exhibits numerous properties that are central in the analysis of key generalization metrics, namely, the generalization gap (GG); the expected generalization gap (EGG); and the doubly-expected generalization

Xinying Zou is with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France.

Samir M. Perlaza is with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France; the ECE Dept. at Princeton University, Princeton N.J. 08544, USA; also with the GAATI Laboratory at the Université de la Polynésie Française, Faaa 98702, French Polynesia.

Iñaki Esnaola is also with the ACSE Dept. at The University of Sheffield, Sheffield S1 3JD, UK; and also with the ECE Dept. at Princeton University, Princeton N.J. 08544, USA.

Eitan Altman is with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France; and also with the Laboratoire d'Informatique d'Avignon (LIA), Université d'Avignon, France.

H. Vincent Poor is with the ECE Dept. at Princeton University, Princeton N.J. 08544, USA.

This work is funded in part by the ANR Project PARFAIT under grant ANR-21-CE25-0013; in part by a grant from the C3.ai Digital Transformation Institute; and in part by the INRIA Exploratory Action IDEM. Parts of this work were presented in the 38th Annual AAAI Conference on Artificial Intelligence [1]; the INRIA technical report [2]; and were submitted to the International Symposium on Information Theory in 2024.

gap (D-EGG). In the literature, the D-EGG is also referred to as the *generalization error* (GE). The GG is calculated for a specific model and training dataset as the difference between the true risk and the empirical risk induced by the model on such a training dataset. While the GG does not provide a generalization guarantee for a given model, it allows studying the impact of the training dataset on the generalization capabilities. This observation is central to the results presented in this work. When models are chosen by sampling a probability measure conditioned on the training dataset, the EGG is the expectation of the GG with respect to such a probability measure. This probability measure is often referred to as the statistical learning algorithm. In this setting, the dependence of the EGG on the training dataset is twofold. Firstly, via the algorithm, which depends on the training dataset; and secondly, via the GG, which involves the calculation of the empirical risk on the training dataset. The D-EGG is the expectation of the EGG with respect to the GTDG probability measure.

A. Existing Results

The GG, EGG, and D-EGG (or GE) are central performance metrics for the analysis of the generalization capabilities of machine learning algorithms, see for instance [17]–[21] and [22]. In particular, the D-EGG characterizes the ability of the learning algorithm to correctly find patterns in datasets that are not available during the training stage. Closed-form expressions for the EGG are only known for the Gibbs algorithm in the case in which the reference measure is a probability measure [17]; and for the case in which the reference measure is a σ-finite measure [23]. In the case of other algorithms, the D-EGG is characterized by various upper-bounds leveraging different techniques. The metric of mutual information is first proposed in [24], further developed in [19] and combined with chaining methods in [25], [26]. Similar bounds are obtained in [20], [27]–[30] and references therein. Other information measures such as the Wasserstein distance [18], [31], [32], maximal leakage [33], [34], mutual f-information [35], and Jensen-Shannon divergence [36] are also explored in the literature. Unfortunately, none of the existing bounds has been shown to be uniformly tight in relevant practical cases [37], [38].

To circumvent the dependence of the D-EGG on the statistical description of the training dataset and the specific learning algorithm in generalization analyses, tools from combinatorics [39]; probability theory [40]–[42]; and information theory [17], [19], [43] have been used with partial success. 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 generalization capability of the algorithm [38]. This, in turn, results in descriptions of the D-EGG for which the dependence on the training dataset and the selected algorithm is not made evident. Recent efforts for highlighting the dependence of generalization capabilities on the training dataset have led to explicit expressions for the GG and EGG when the models are sampled using the Gibbs algorithm in [22], [44]. This line of work paved the way to the study in this paper of the WCDG probability measures and their effect on the D-EGG.

B. Contributions

The first contribution consists of the derivation of a probability measure over the datasets coined WCDG probability measure. The WCDG probability measure can be defined as the measure that maximizes the expectation of the loss over a set of measures that are at a "statistical distance" with respect to a *reference measure* that is not larger than a given threshold. Alternatively, the WCDG probability measure can be defined as the measure that maximizes the expectation of the loss subject to a regularization by such a "statistical distance" with respect to the *reference measure*. When the "statistical distance" is the relative entropy, both definitions are shown to be identical, under specific conditions. Despite the limitations of relative entropy concerning its asymmetry [12], [45] and the need of absolute continuity with respect to the reference measure [11], the resulting WCDG probability measure is a Gibbs probability measure (Theorem 1) parametrized by the reference measure; the regularization parameter; and the loss function. This Gibbs probability measure is shown to exhibit relevant properties in statistical machine learning. In particular, for a fixed model, the loss resulting from the assumption that datapoints are sampled from the WCDG probability measure is a sub-Gaussian random variable. Moreover, the variation of the expectation of the loss when the probability measure changes from the WCDG probability measure to an alternative measure has an explicit expression involving relative entropy terms. 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 8). This is a significant result as the reference measure and the regularization parameter can be arbitrarily chosen, which leads to numerous 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, which are empirical probability measures [4]. 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 7). This observation allows using Theorem 8 to provide an explicit expression to the difference between two empirical risks induced by the same model on two different datasets. This difference is referred to as the *sensitivity* of the empirical risk. Using the same arguments, closed-form expressions involving relative entropy terms are provided for the GG induced by a fixed model.

The final contribution consists in using the WCDG probability measure to obtain an exact expression of the D-EGG 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 D-EGG 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 shown to be achieved by a Gibbs algorithm whose parameters satisfy particular conditions. This reveals the central place of the Gibbs algorithm in statistical machine learning. Indeed, the Gibbs algorithm facilitates the derivation of overarching insights into the D-EGG applicable to any machine learning algorithm.

C. Notation

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

II. 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:

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

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

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

Let also the function

$$\hat{\ell} : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty] \quad (3)$$

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 $y \in \mathcal{Y}$, $\hat{\ell}(y, y) = 0$.

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

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

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

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

where the function ℓ is defined in (4).

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

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

The ERM problem is prone to overfitting since the set of solutions to (6) are models selected specifically for the given data set $\boldsymbol{z}$ in (1), 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

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

where λ is a nonnegative real that acts as a regularization parameter. The regularization function 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 [3]. One common choice for R is $R(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_p$, 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.

III. The Worst-Case Data-Generating Probability Measure

A. General Discussion

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 $\boldsymbol{\theta} \in \mathcal{M}$, looking for a WCDG probability measure might lead to two different optimization problems. Both can be defined in terms of the expected loss induced by a measure $P \in \Delta(\mathcal{X} \times \mathcal{Y})$, for a given a model $\boldsymbol{\theta} \in \mathcal{M}$.

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

$$\mathsf{R}_{\boldsymbol{\theta}}(P) = \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P(x, y), \quad (8)$$

where the function ℓ is defined in (4).

Firstly, a *neighborhood* around P_S is established through a *statistical distance* $\mathsf{d} : \triangle(\mathcal{X} \times \mathcal{Y}) \times \triangle(\mathcal{X} \times \mathcal{Y}) \to [0, +\infty)$. In view of this, the WCDG measure is a probability measure that maximizes the expectation of the loss within such a neighborhood. This view leads to the following problem:

$$\max_{P \in \Delta(\mathcal{X} \times \mathcal{Y})} \mathsf{R}_{\boldsymbol{\theta}}(P) \quad (9a)$$

$$\text{s.t.} \quad \mathsf{d}(P, P_S) \leqslant \gamma, \quad (9b)$$

where $\gamma > 0$ determines the neighborhood around P_S as the set $\{P \in \triangle(\mathcal{X} \times \mathcal{Y}) : \mathsf{d}(P, P_S) \leqslant \gamma\}$ and the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (8). Secondly, a WCDG measure is also interpreted as a 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 point of view leads to an optimization problem of the form:

$$\max_{P\in\Delta(\mathcal{X}\times\mathcal{Y})} \mathsf{R}_{\boldsymbol{\theta}}\left(P\right) - \beta \mathsf{d}\left(P, P_S\right), \tag{10}$$

where $\beta > 0$ determines the trade-off between maximization of the expectation of the loss and the statistical distance to P_S; and the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (8). In this case, the convexity of $\mathsf{d}\left(P, P_S\right)$ with respect to P is a valuable property for solving the optimization problem.

Depending on the choice of the statistical distance d in (9) and (10), numerous WCDG probability measures might be obtained. In either case, such WCDG probability measures are subject to the limitations concerning the asymmetry of d; limitations concerning the absolute continuity requirements with respect to P_S; or limitations derived from the need for $\mathcal{X}\times\mathcal{Y}$ to be equipped with a metric so as to form a Polish metric space. These limitations can lead the problems in (9) and (10) to be ill posed, difficult to solve, to exhibit no solution, or to exhibit solutions with poor properties for the analysis of generalization in statistical machine learning. In particular, the statistical distance d can be chosen as an f-divergence [46]. See for instance the examples in [11]. In this case,

$$\mathsf{d}\left(P, P_S\right) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}P_S}(x,y)\right)\mathrm{d}P(x,y), \tag{11}$$

where the function $f:(0,+\infty)\to\mathbb{R}$ is assumed to be concave and $f(1)=0$; and the function $\frac{\mathrm{d}P}{\mathrm{d}P_S}:\mathcal{X}\times\mathcal{Y}\to[0,+\infty)$ is the Radon-Nikodym derivative of P with respect to P_S. For some choices of the function f, the corresponding f-divergence can be symmetric, e.g., Jeffrey's divergence [47]; or be asymmetric, e.g., Kullback-Leibler divergence [48]. An interesting discussion on the impact of asymmetry in optimization problems similar to the one in (10) is presented in [12] and [11]. In any case, if d is chosen as an f-divergence, the optimization domain shall be restricted to the set of measures that are absolutely continuous with P_S. This is essentially because the existence of a Radon-Nikodym derivative with respect to P_S is required.

An alternative choice for the statistical distance d in (9) and (10) is

$$\begin{aligned}&\mathsf{d}\left(P, P_S\right) \\ &= \sup\left\{\int f(x,y)\mathrm{d}P(x,y) - \int f(x,y)\mathrm{d}P_S(x,y) : f\in\mathcal{D}\right\},\end{aligned} \tag{12}$$

where $\mathcal{D}$ is a particular set of functions $\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$. In such a formulation for d in (12), while P is not required to be absolutely continuous with respect to P_S, other requirements must be ensured. For instance, in the case in which the set $\mathcal{D}$ is such that if $f\in\mathcal{D}$, then for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and for all $(u,v)\in\mathcal{X}\times\mathcal{Y}$,

$$|f(x,y) - f(u,v)| \leqslant d\left((x,y),(u,v)\right), \tag{13}$$

with d being a metric on the set $\mathcal{X}\times\mathcal{Y}$, the resulting statistical distance is the Kantorovich-Rubinstein distance, also known as the Wasserstein distance of order one [49, Remark 6.5]. Note that this choice of $\mathcal{D}$ imposes the need of the set $\mathcal{X}\times\mathcal{Y}$ to be equipped with the metric d in (13). Moreover, $(\mathcal{X}\times\mathcal{Y}, d)$ must be a Polish metric space.

From the above, it becomes clear that every choice for the statistical distance d comes with particular requirements. Thus, the search for an optimal choice for d reduces to evaluating the properties of the resulting WCDG probability measures and their impact in the analysis of generalization of machine learning algorithms. In the following, it is argued that the choice of d as the relative entropy, despite the limitations concerning its asymmetry and the restriction of the optimization domains to measures that are absolutely continuous with P_S, leads to a variety of properties that are central in the analysis of generalization.

B. Relative Entropy Case

This section focuses on the special cases in which the statistical distance d in (9) and (10) is chosen as the relative entropy. The relative entropy is the f-divergence resulting from the choice of $f(x) = x\log(x)$ in (11). See for instance [50]. 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

$$D\left(P\|Q\right) = \int \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right)\mathrm{d}Q(x), \tag{14}$$

where the function $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative of P with respect to Q. Using this notation, the optimization problems of interest are

$$\max_{P\in\triangle_{P_S}(\mathcal{X}\times\mathcal{Y})} \mathsf{R}_{\boldsymbol{\theta}}\left(P\right) \tag{15a}$$

$$\text{s.t.} \qquad D\left(P\|P_S\right) \leqslant \gamma, \tag{15b}$$

and

$$\max_{P\in\Delta_{P_S}(\mathcal{X}\times\mathcal{Y})} \mathsf{R}_{\boldsymbol{\theta}}\left(P\right) - \beta D\left(P\|P_S\right). \tag{16}$$

The optimization problems in (15) and (16) exhibit a major difference in the following case. If for all $P\in\triangle_{P_S}$, the functional $\mathsf{R}_{\boldsymbol{\theta}}$ in (8) is such that $\mathsf{R}_{\boldsymbol{\theta}}\left(P\right) = c$, for some fixed $c\geqslant 0$, then all the measures in $\{P\in\triangle\left(\mathcal{X}\times\mathcal{Y}\right) : D\left(P\|P_S\right)\leqslant\gamma\}$ are solutions to the optimization problem in (15). Alternatively, the problem in (16) exhibits a unique solution, which is P_S. Hence, while the former exhibits infinitely many solutions, the latter exhibits only one. Although this difference holds substantial mathematical importance, its practical relevance is limited, as it only emerges in settings for which the expectation of the loss $\mathsf{R}_{\boldsymbol{\theta}}(P)$ is the same independently of the measure P. In order to avoid the above case, the notion of separable loss functions, which is analogous to [23, Definition 4.1], is introduced as follows.

Definition 2: Given a model $\boldsymbol{\theta}\in\mathcal{M}$, the function ℓ in (4), is said to be separable with respect to the probability measure $P\in\Delta\left(\mathcal{X}\times\mathcal{Y}\right)$, if there exist a positive real $c>0$ and two subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{M}$ that are nonnegligible with respect to P, and for all $((x_1,y_1),(x_2,y_2))\in\mathcal{A}\times\mathcal{B}$,

$$\ell(\boldsymbol{\theta},x_1,y_1) < c < \ell(\boldsymbol{\theta},x_2,y_2) < +\infty. \tag{17}$$

When the function ℓ in (4) 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

$$P_S\left(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = a\}\right) = 1, \tag{18}$$

and as a consequence, for all probability measures $P \in \triangle_{P_S}(\mathcal{X} \times \mathcal{Y})$, it holds that $P\left(\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \ell(\boldsymbol{\theta}, x, y) = a\}\right) = 1$. With this pathological case of nonseparable loss functions out of the way, the optimization problems in (15) and (16) exhibit considerable similarity. This similarity is formalized in Theorem 1 below. Given a model $\boldsymbol{\theta} \in \mathcal{M}$, let $\mathsf{J}_{P_S,\boldsymbol{\theta}} : \mathbb{R} \to \mathbb{R}$ be the function

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}(t) = \log\left(\int \exp\left(t\ell(\boldsymbol{\theta}, x, y)\right) \mathrm{d}P_S(x,y)\right), \tag{19}$$

where the function ℓ is defined in (4). The following lemma describes some properties of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19).

Lemma 1: The function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19) is convex, nondecreasing, and differentiable infinitely many times in the interior of $\{t \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}(t) < +\infty\}$. If the function ℓ in (4) is separable with respect to P_S, then the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is strictly convex.

Proof: The proof is presented in [51, Appendix A]. ■

Let also the set $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ be the set

$$\mathcal{J}_{P_S,\boldsymbol{\theta}} \triangleq \left\{t \in (0,+\infty) : 0 \leq \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) < +\infty\right\}, \tag{20}$$

which exhibits the following property.

Lemma 2: The set $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (20) is either the empty set or an interval satisfying $(b,+\infty) \subseteq \mathcal{J}_{P_S,\boldsymbol{\theta}}$, for some $b \in [0,+\infty)$.

Proof: The proof is presented in [51, Appendix B]. ■

Using this notation, the following theorem formalizes the fact that the optimization problems in (15) and (16) exhibit the same unique solution.

Theorem 1: If the function ℓ in (4) 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 (20), 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 $(x,y) \in \operatorname{supp} P_S$,

$$\frac{\mathrm{d}P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}}{\mathrm{d}P_S}(x,y) = \exp\left(\frac{1}{\beta}\ell(\boldsymbol{\theta}, x, y) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right), \tag{21}$$

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is defined in (19) and

$$D\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right) = \gamma, \tag{22}$$

is the unique solution to the optimization problems in (15) and (16).

Proof: The proof is presented in [51, Appendix C]. ■

The probability measure P_S in Theorem 1 can be arbitrarily chosen, that is, independent of the model $\boldsymbol{\theta}$. From this perspective, when the measure P_S is interpreted as a prior on the datasets, the WCDG probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ can be interpreted as the worst probability measure for model $\boldsymbol{\theta}$ in the *neighborhood* of the prior P_S. The reference measure P_S can also be chosen to be dependent on the models. This case is studied in Section VII-B.

The parameter γ of the optimization problem in (15) does not explicitly appear in the expression of the solution $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (21), the probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ does depend on γ. This dependence is shown via the equality in (22). Later, in Lemma 6, it is shown that $D\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right)$ in (22) is monotone with β. More importantly, if the function ℓ in (4) is separable, the relative entropy $D\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right)$ is strictly decreasing with β. This implies that there exists a one-to-one mapping between β and γ, and thus, the analysis can be indistinctly carried out either for β or γ.

The probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (21) is a Gibbs probability measure [52]. In the remainder of this work, the probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is referred to as the *WCDG probability measure* and the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19) is referred to as the log-partition function [53]. The WCDG probability measure exhibits a number of interesting properties. The following lemma introduces two of these properties, which are shown to be central in the remainder of this work.

Lemma 3: The probability measures $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ and P_S in (21) are mutually absolutely continuous. Moreover, they satisfy:

$$\beta \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) = \mathsf{R}_{\boldsymbol{\theta}}\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - \beta D\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right) \tag{23}$$

$$= \mathsf{R}_{\boldsymbol{\theta}}(P_S) + \beta D\left(P_S \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right), \tag{24}$$

where the functional $\mathsf{R}_{\boldsymbol{\theta}}$ and the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ are in (8) and (19), respectively.

Proof: The proof is presented in [51, Appendix D]. ■

Lemma 3 implicitly characterizes the difference of the expected losses induced by the WCDG probability measure and the reference measure. That is,

$$\mathsf{R}_{\boldsymbol{\theta}}\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - \mathsf{R}_{\boldsymbol{\theta}}\left(P_S\right) = \beta\left(D\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right) + D\left(P_S \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\right), \tag{25}$$

where $D\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_S\right) + D\left(P_S \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$ is Jeffrey's divergence between $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ and P_S.

In the remainder of this work, the GTDG probability measure of the datapoints is denoted by $P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$. The probability measure P_Z is the measure induced by a random variable Z on the measurable space of the datapoints. Under the assumption that datasets are formed by n independent and identically distributed datapoints, the GTDG probability measure of such datasets is denoted by $P_{\boldsymbol{Z}} \in \triangle\left((\mathcal{X} \times \mathcal{Y})^n\right)$, which is a product measure formed by P_Z. Alternatively, the probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (21) is induced by a random variable $\hat{Z}$ on the measurable space of the datapoints. Under the assumption that datasets are formed by n independent

datapoints sampled from the WCDG probability measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$, the probability measure of such datasets is denoted by $P_{\hat{\boldsymbol{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$, which is the product measure formed by $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$. In a nutshell, datapoints and datasets sampled from the GTDG probability measure are represented by the random variables Z and $\boldsymbol{Z}$, respectively. Similarly, datapoints and datasets sampled from a WCDG probability measure are represented by the random variables $\hat{Z}$ and $\hat{\boldsymbol{Z}}$, respectively.

C. Choice of the Reference Measure

Probably, the most intuitive choice of P_S is to be based on the training datasets. In order to introduce this intuition, the notion of a type induced by a dataset [4] is presented in the following.

Definition 3 (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 $\mathcal{X}\times\mathcal{Y}$,

$$P_{\boldsymbol{z}}(\mathcal{A}) = \frac{1}{n}\sum_{t=1}^{n}\mathbb{1}_{\{(x_t,y_t)\in\mathcal{A}\}}. \tag{26}$$

For large values of n, it might be expected that the type $P_{\boldsymbol{z}}$ induced by the training dataset $\boldsymbol{z}\in(\mathcal{X}\times\mathcal{Y})^n$ be within a neighborhood of the GTDG probability distribution P_Z. More specifically, under the assumption that the sets $\mathcal{X}$ and $\mathcal{Y}$ are discrete, such a neighborhood might be of the form

$$\mathcal{N}(P_Z,\gamma)\triangleq\{P\in\triangle(\mathcal{X}\times\mathcal{Y}) : D(P\|P_Z)\leqslant\gamma\}, \tag{27}$$

for some $\gamma>0$. Nonetheless, the GTDG probability measure P_Z being unknown, it appears reasonable to choose the reference measure P_S identical to the type $P_{\boldsymbol{z}}$. This leads to an optimization problem of the form in (15), whose optimization domain is

$$\mathcal{N}(P_{\boldsymbol{z}},\gamma)\triangleq\{P\in\triangle_{P_{\boldsymbol{z}}}(\mathcal{X}\times\mathcal{Y}) : D(P\|P_{\boldsymbol{z}})\leqslant\gamma\}. \tag{28}$$

This choice presumes that the training dataset is sufficiently large to consider that the GTDG probability measure is absolutely continuous with respect to the type $P_{\boldsymbol{z}}$ such that the GTDG probability measure P_Z is within $\mathcal{N}(P_{\boldsymbol{z}},\gamma)$. Another alternative for the choice of P_S in the optimization problems in (15) and (16) is for instance a convex combination between the type $P_{\boldsymbol{z}}$ and an arbitrary probability measure known to be absolutely continuous with P_Z.

In any case, the choice of P_S imposes a condition on the optimization domain of the problems in (15) and (16) that translates into a strong inductive bias that dominates the evidence provided by the training data. More specifically, the optimization domain becomes a subset of the set of measures that are absolutely continuous with respect to P_S.

IV. Cumulants of the Expected Loss

The log-partition function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19) can be interpreted also as the cumulant generating function of the random variable

$$V_{\boldsymbol{\theta}} \triangleq \ell(\boldsymbol{\theta},X,Y), \tag{29}$$

with the function ℓ in (4) and $(X,Y)\sim P_S$, with P_S in (21), for some given model $\boldsymbol{\theta}$ that remains fixed. In particular, if the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is differentiable around zero, its derivative of order n evaluated at zero reveals the n-th cumulant of the random variable $V_{\boldsymbol{\theta}}$ in (29). More interestingly, the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19) is intimately related to the cumulant generating function of the random variable

$$W_{\boldsymbol{\theta}} \triangleq \ell(\boldsymbol{\theta},X,Y), \tag{30}$$

where $(X,Y)\sim P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$, with $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (21), for some given model $\boldsymbol{\theta}$ that remains fixed. In particular, the cumulant generating function of the random variable $W_{\boldsymbol{\theta}}$, which is denoted by $\mathsf{J}_{P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)},\boldsymbol{\theta}}:\mathbb{R}\to\mathbb{R}$, satisfies

$$\mathsf{J}_{P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)},\boldsymbol{\theta}}(t) = \log\left(\int\exp\left(t\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}(x,y)\right), \tag{31}$$

where the function ℓ is defined in (4). The following lemma shows the exact relation between these two cumulant generating functions.

Lemma 4: The function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (21) and the function $\mathsf{J}_{P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)},\boldsymbol{\theta}}$ in (31) satisfy

$$\mathsf{J}_{P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)},\boldsymbol{\theta}}(t) = \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(t+\frac{1}{\beta}\right) - \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right). \tag{32}$$

Proof: The proof is presented in [51, Appendix E]. ■

Let the m-th derivative of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19) be denoted by $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(m)}$, with $m\in\mathbb{N}$. Hence, for all $t\in\mathbb{R}$ such that $\mathsf{J}_{P_S,\boldsymbol{\theta}}(t)<+\infty$,

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(m)}(t)\triangleq\frac{\mathrm{d}^m}{\mathrm{d}s^m}\mathsf{J}_{P_S,\boldsymbol{\theta}}(s)\Big|_{s=t}. \tag{33}$$

Moreover, let the m-th derivative of the function $\mathsf{J}_{P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)},\boldsymbol{\theta}}$ in (19) be denoted by $\mathsf{J}_{P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)},\boldsymbol{\theta}}^{(m)}$, with $m\in\mathbb{N}$. Hence, for all $t\in\mathbb{R}$ such that $\mathsf{J}_{P_S,\boldsymbol{\theta}}\left(t+\frac{1}{\beta}\right)<+\infty$,

$$\mathsf{J}_{P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)},\boldsymbol{\theta}}^{(m)}(t)\triangleq\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(m)}\left(t+\frac{1}{\beta}\right). \tag{34}$$

The following lemma capitalizes on the observations above and provides explicit expressions for the first, second and third derivatives of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19) and its connections with the cumulants of the random variables $V_{\boldsymbol{\theta}}$ and $W_{\boldsymbol{\theta}}$ in (29) and (30), respectively.

Lemma 5: For all $t\in\mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (20), the first, second and third derivatives of the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19), denoted respectively by $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}$, $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(2)}$ and $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(3)}$, satisfy that

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right) = \int\ell(\boldsymbol{\theta},x,y)\mathrm{d}P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,t)}(x,y), \tag{35}$$

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(2)}\left(\frac{1}{t}\right) = \int\left(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^2\mathrm{d}P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,t)}(x,y), \tag{36}$$

and

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(3)}\left(\frac{1}{t}\right) = \int\left(\ell(\boldsymbol{\theta},x,y)-\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)\right)^3\mathrm{d}P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,t)}(x,y), \tag{37}$$

where the function ℓ is in (4) and the measure $P^{(P_S,t)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is in (21). The value $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$ is strictly positive if and only if the function ℓ is separable with respect to P_S. Moreover, if there exists real $\delta > 0$ such that the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is differentiable within $(-\delta, \delta)$, then

$$\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}(0) = \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_S(x, y), \text{and} \tag{38}$$

$$\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}(0) = \int \left(\ell(\boldsymbol{\theta}, x, y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}(0)\right)^2 \mathrm{d}P_S(x, y), \text{and} \tag{39}$$

$$\mathsf{J}^{(3)}_{P_S,\boldsymbol{\theta}}(0) = \int \left(\ell(\boldsymbol{\theta}, x, y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}(0)\right)^3 \mathrm{d}P_S(x, y). \tag{40}$$

Proof: The proof is presented in [51, Appendix F]. ■

From Lemma 5, it follows that the random variable $W_{\boldsymbol{\theta}}$ in (30) possesses a mean, variance, and third cumulant that are equivalent to $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ in (35), $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ in (36), and $\mathsf{J}^{(3)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ in (37), respectively. Alternatively, under the assumptions of the lemma, the random variable $V_{\boldsymbol{\theta}}$ in (29) possesses a mean, variance, and third cumulant that are equivalent to $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}(0)$ in (38), $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}(0)$ in (39), and $\mathsf{J}^{(3)}_{P_S,\boldsymbol{\theta}}(0)$ in (40), respectively. The following corollary of Lemma 5 highlights the monotonicity of $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ with respect to β.

Corollary 1: The expected loss $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$ in (35) is nonincreasing with respect to t in the interior of $\left\{s \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right) < +\infty\right\}$. Moreover, if the loss function ℓ in (4) is separable with respect to P_S, $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right)$ is strictly decreasing with t in $\left\{s \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right) < +\infty\right\}$.

More generally, under the assumptions of the lemma, the random variables $V_{\boldsymbol{\theta}}$ in (29) and $W_{\boldsymbol{\theta}}$ in (30) possess finite cumulants of all orders. That is, the values $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}(0)$ and $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, which are the m-th cumulants of the random variables $V_{\boldsymbol{\theta}}$ and $W_{\boldsymbol{\theta}}$, respectively, for all integers $m > 0$, are both finite. This observation together with the mean value theorem [54, Theorem 5.10] lead to the following characterization of the cumulants $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ and $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}(0)$.

Theorem 2: Assume that there exists real $\delta > 0$ such that the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ in (19) is differentiable within $(-\delta, \delta)$. Then, for all $m \in \mathbb{N}$, the function $\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}$ in (33) satisfies for all $t \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (20),

$$\mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}(0) + \frac{1}{t}c_{m,2} \leq \mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{t}\right) \leq \mathsf{J}^{(m)}_{P_S,\boldsymbol{\theta}}(0) + \frac{1}{t}c_{m,1}, \tag{41}$$

and

$$c_{m,1} = \max_{s \in (\beta, +\infty)} \mathsf{J}^{(m+1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right), \text{ and} \tag{42}$$

$$c_{m,2} = \min_{s \in (\beta, +\infty)} \mathsf{J}^{(m+1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{s}\right), \tag{43}$$

are both finite.

Proof: The proof is presented in [51, Appendix G]. ■

In Theorem 2, the case $m = 1$ is of little interest as $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}(0) = \mathsf{R}_{\boldsymbol{\theta}}(P_S)$ and $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) = \mathsf{R}_{\boldsymbol{\theta}}\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$, and thus, the difference $\mathsf{R}_{\boldsymbol{\theta}}\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - \mathsf{R}_{\boldsymbol{\theta}}(P_S)$ is well characterized by Jeffrey's divergence in (25). Nonetheless, while the inequality $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}(0) < \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ is always verified, Theorem 2 shows that such inequality is not necessarily observed on the higher order cumulants. In the case of the variance ($m = 2$), it might be observed that $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}(0) > \mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ if $c_{2,1} < 0$. That is, the WCDG probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (21) leads to a larger expected loss than the one obtained with the reference measure P_S and at the same time it induces a lower variance than the variance associated with the reference measure P_S. Interestingly, the case in which the WCDG probability measure induces both a larger expected loss and higher variance is also possible when $c_{2,2} > 0$. A similar observation holds for cumulants of order $m > 2$.

The following theorem shows that the cumulant generating function $\mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}$ of the random variable $W_{\boldsymbol{\theta}}$ in (30) exhibits an important upper bound.

Theorem 3: 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 (31) satisfies the following inequality:

$$\mathsf{J}_{P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}},\boldsymbol{\theta}}(t) \leq t\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right) + \frac{1}{2}t^2\zeta^2_{P_S,\boldsymbol{\theta}}, \tag{44}$$

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

$$\zeta_{P_S,\boldsymbol{\theta}} \triangleq \sup\left\{\sqrt{\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}(\xi)} : \xi \in \left(-\infty, b - \frac{1}{\beta}\right)\right\}, \tag{45}$$

with

$$b \triangleq \sup\{\alpha \in \mathbb{R} : \mathsf{J}_{P_S,\boldsymbol{\theta}}(\alpha) < +\infty\}, \tag{46}$$

and the functions $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}$ and $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}$ are defined in (33).

Proof: The proof is presented in [51, Appendix H]. ■

The relevance of Theorem 3 lies on the fact that it implies that the random variable $W_{\boldsymbol{\theta}}$ in (30) is a sub-Gaussian random variable with sub-Gaussian parameter $\zeta_{P_S,\boldsymbol{\theta}}$ in (45). An interesting discussion on the impact of the random variable $W_{\boldsymbol{\theta}}$ being sub-Gaussian is presented in [19, Theorem 1].

Finally, leveraging the properties of sub-Gaussian random variables, e.g., [55, Equation 2.9], the following holds for all $\alpha > 0$:

$$P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\left(\left\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \left|\ell(\boldsymbol{\theta}, x, y) - \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)\right| \geq \alpha\right\}\right)$$
$$\leqslant 2\exp\left(-\frac{\alpha^2}{2\zeta^2_{P_S,\boldsymbol{\theta}}}\right), \tag{47}$$

where $\zeta_{P_S,\boldsymbol{\theta}}$ is defined in (45) and $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ is the expected loss in (35). This type of concentration inequality allows the study of the robustness of the model $\boldsymbol{\theta}$ when it faces data sampled from WCDG probability distributions, which leads to interesting guidelines for algorithm design.

V. (ϵ, δ)-Robustness and Algorithm Design

An important observation on the dependence of the WCDG probability distribution $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (21) on the parameter γ in (22) is that for all $P \in \mathcal{N}(P_S, \gamma)$, with $\mathcal{N}(P_S, \gamma)$ in (27),

$$\mathsf{R}_{\boldsymbol{\theta}}(P) \leqslant \mathsf{R}_{\boldsymbol{\theta}}\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right). \tag{48}$$

Essentially, the WCDG probability measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ is the measure that induces the largest expected loss within the neighborhood $\mathcal{N}(P_S, \gamma)$ of P_S. The following lemma unveils the fact that such a neighborhood expands as γ increases or equivalently, as β decreases.

Lemma 6: The relative entropy $D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)$ in (22) satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\beta} D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) = -\frac{1}{\beta^3} \mathsf{J}_{P_S,\boldsymbol{\theta}}^{(2)}\left(\frac{1}{\beta}\right) \leqslant 0, \tag{49}$$

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(2)}$ is defined in (36). Moreover, the inequality is strict if and only if the function ℓ in (4) is separable with respect to P_S.

Proof: The proof is presented in [51, Appendix I]. ■

The key observation from Lemma 6 is that if the function ℓ is separable with respect to P_S, there exists a bijection between γ and β induced by the equality in (22). Another interesting observation is that the variations of $D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)$ in (22) and the expected loss $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)$ in (35), with respect to β (or γ), obey a revealing relation. Note that

$$\frac{\mathrm{d}}{\mathrm{d}\beta} D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right) = \frac{1}{\beta}\frac{\mathrm{d}}{\mathrm{d}\beta} \mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right), \tag{50}$$

which implies that both $D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)$ and $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)$ are decreasing with β. Nonetheless, their variations to changes in β significantly differ. For values of β around one, given that $1 \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (20), the variations of $D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)$ and $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)$ to changes in β are comparable. On the contrary, for values away from one, the variations of these quantities might be radically different. For instance, for large values of $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, the variation of $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)$ is larger than the variation of $D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)$. This is consistent with the fact that greater values of β imply smaller values of γ. Hence, slightly increasing the size of a small neighborhood leads to including more candidates to be the WCDG probability measure, which appear to be more impactful in the maximization of the expected loss. Alternatively, for smaller values of $\beta \in \mathcal{J}_{P_S,\boldsymbol{\theta}}$, which imply larger values of γ, the variation of $D\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \| P_S\right)$ is larger than the variation of $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)$ with respect to β. Hence, slightly varying the size of a neighborhood that is already big does not lead to important variations in the expected loss induced by WCDG probability measure.

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

$$P\left(\{\boldsymbol{z} \in (\operatorname{supp} P_S)^n : \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) \geq \delta\}\right) \leq \epsilon. \tag{51}$$

This notion of robustness enables the study of the performance guarantees that a model $\boldsymbol{\theta}$ yields when faced with data generated by theWCDG 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 link between β, the reference P_S, the expected loss, and the (ϵ, δ)-robustness that a given model $\boldsymbol{\theta}$ achieves.

Theorem 4: The probability measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (21) satisfies that for all $\delta > \mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)$, with $\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta}\right)$ in (35),

$$\begin{aligned} &P_{\hat{\boldsymbol{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\left(\{\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n : \mathsf{L}(\boldsymbol{\theta}, \boldsymbol{z}) \geqslant \delta\}\right) \\ &\leq \exp\left(-nD\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta^*)} \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right)\right), \end{aligned} \tag{52}$$

where $\beta^* \in (0, \beta) \cap \mathcal{J}_{P_S,\boldsymbol{\theta}}$, with $\mathcal{J}_{P_S,\boldsymbol{\theta}}$ in (20), satisfies

$$\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{\beta^*}\right) = \delta; \tag{53}$$

the function L is defined in (5); and the measure $P_{\hat{\boldsymbol{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \in \triangle\left((\mathcal{X} \times \mathcal{Y})^n\right)$ is a product measure formed by $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$.

Proof: This proof is presented in [51, Appendix J]. ■

Theorem 4 describes the (ϵ, δ)-robustness of a model $\boldsymbol{\theta}$ to the WCDG probability measure $P_{\hat{\boldsymbol{Z}}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)} \in \triangle\left((\mathcal{X} \times \mathcal{Y})^n\right)$. Note such a probability measure describes test datasets that are formed by n datapoints independently and identically distributed with $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ in (21). Notably, such a description is in terms of another WCDG probability distribution $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta^*)}$, where $\beta^* < \beta$. Interestingly, the WCDG probability measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta^*)}$ induces an expected loss that is equal to δ and is greater than the expected loss induced by the WCDG probability measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$. In a nutshell, for all $t \in (0, \beta) \cap \mathcal{J}_{P_S,\boldsymbol{\theta}}$, let $\delta_t = n\mathsf{J}_{P_S,\boldsymbol{\theta}}^{(1)}\left(\frac{1}{t}\right)$ and $\epsilon_t = \exp\left(-nD\left(P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,t)} \| P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}\right)\right)$. Then, the model $\boldsymbol{\theta}$ is (ϵ_t, δ_t)-robust to the WCDG probability measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$.

The following theorem provides a (ϵ, δ)-robust guarantee to any product probability measure formed by a measure in the neighborhood $\mathcal{N}(P_S, \gamma)$ in (27).

Theorem 5: For all models $\boldsymbol{\theta} \in \mathcal{M}$ and for all $P_Z \in \mathcal{N}(P_S, \gamma)$, with $\mathcal{N}(P_S, \gamma)$ in (27) and β in (22), it follows that for all $\delta > \mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$, with $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ in (35),

$$P_{\boldsymbol{Z}}\left(\{\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n : \mathsf{L}(\boldsymbol{\theta}, \boldsymbol{z}) \geqslant \delta\}\right) \leq \frac{1}{\delta}\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right), \quad (54)$$

where the function ℓ is defined in (4) and the product probability measure $P_{\boldsymbol{Z}}$ is formed by P_Z.

Proof: The proof is presented in [51, Appendix K]. ■

The relevance of Theorem 5 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 $(\epsilon_\delta, \delta)$-robust to all probability measures in $\mathcal{N}(P_S, \gamma)$, with $\mathcal{N}(P_S, \gamma)$ in (27). This observation raises the question of whether performing model selection based on the minimization of $\mathsf{J}^{(1)}_{P_S,\boldsymbol{\theta}}\left(\frac{1}{\beta}\right)$ is an alternative for classical approaches based on empirical risk minimization (ERM) as in [56]; or statistical ERM as in [23] and [11].

VI. Sensitivity of the Expected Loss

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

$$G(\boldsymbol{\theta}, P_1, P_2) = \mathsf{R}_{\boldsymbol{\theta}}(P_1) - \mathsf{R}_{\boldsymbol{\theta}}(P_2), \quad (55)$$

where the functional $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (8). The value $G(\boldsymbol{\theta}, P_1, P_2)$ represents the sensitivity of the expected loss $\mathsf{R}_{\boldsymbol{\theta}}$ when the data-generating probability measure changes from P_2 to P_1. This section characterizes the *sensitivity* $G(\boldsymbol{\theta}, P_1, P_2)$ for arbitrary measures P_1 and P_2. To achieve this goal, the first result is the characterization of the sensitivity of the expected loss to variations from the measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (21) to an alternative measure, which is presented in the following theorem.

Theorem 6 (Sensitivity of the Expected Loss): The probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (21) satisfies, for all $P \in \Delta_{P_S}(\mathcal{X}\times\mathcal{Y})$,

$$\begin{aligned} &G\left(\boldsymbol{\theta}, P, P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) \\ &= \beta\Big(D(P\|P_S) - D\left(P\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_S\right)\Big), \end{aligned} \quad (56)$$

where the functional G is defined in (55).

Proof: The proof is presented in [51, Appendix L]. ■

The following theorem introduces an upper bound on the absolute value of the sensitivity $G\left(\boldsymbol{\theta}, P, P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) \leqslant 0$ in (56), which requires the calculation of only one of the relative entropies in Theorem 6.

Theorem 7: The probability measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (21) satisfies for all $P \in \Delta_{P_S}(\mathcal{X}\times\mathcal{Y})$,

$$\left|G\left(\boldsymbol{\theta}, P, P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\right| \leq \sqrt{2\hat{\zeta}^2_{P_S,\boldsymbol{\theta}} D\left(P\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)}, \quad (57)$$

where $\hat{\zeta}_{P_S,\boldsymbol{\theta}}$ satisfies

$$\hat{\zeta}^2_{P_S,\boldsymbol{\theta}} \triangleq \sup\left\{\sqrt{\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}(\xi)} : \xi \in \left(0, b - \frac{1}{\beta}\right)\right\}, \quad (58)$$

with

$$b \triangleq \sup\{\alpha \in (0, +\infty) : \mathsf{J}_{P_S,\boldsymbol{\theta}}(\alpha) < +\infty\}, \quad (59)$$

where the function $\mathsf{J}_{P_S,\boldsymbol{\theta}}$ is defined in (19); the function $\mathsf{J}^{(2)}_{P_S,\boldsymbol{\theta}}$ is defined in (36); and the functional G is defined in (55).

Proof: The proof is presented in [51, Appendix M]. ■

Equipped with the exact characterization of the sensitivity from the measure $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ to any alternative measure P provided by Theorem 6, it is possible to characterize the variation from and to arbitrary measures, as shown by the following theorem.

Theorem 8: 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 $\boldsymbol{\theta} \in \mathcal{M}$,

$$\begin{aligned} G(\boldsymbol{\theta}, P_1, P_2) = \beta\Big(&D\left(P_2\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_1\|P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) \\ &- D(P_2\|P_S) + D(P_1\|P_S)\Big), \end{aligned} \quad (60)$$

where the functional G is defined in (55); and the parameter β, the model $\boldsymbol{\theta}$, and the measures P_S and $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfy (21).

Proof: The proof is presented in [51, Appendix N]. ■

Note that the parameters β and P_S in (60) can be arbitrarily chosen. This is essentially because only the right-hand side of (60) depends on P_S and β. Another interesting observation is that none of the terms in the right-hand side of (60) depends simultaneously on both P_1 and P_2. Interestingly, these terms dependent exclusively on the pair formed by P_i and P_S, with $i \in \{1, 2\}$. These observations highlight the significant flexibility of the expression in (60) to construct closed-form expressions for the sensitivity $G(\boldsymbol{\theta}, P_1, P_2)$ in (55). The only constraint on the choice of P_S is that both measures P_1 and P_2 must be absolutely continuous with respect to P_S. The following corollary follows by adopting particular choices for P_S. Two choices of P_S for which the expression in the right-hand side of (60) significantly simplifies are $P_S = P_1$ and $P_S = P_2$, which leads to the following corollary of Theorem 8.

Corollary 2: If P_1 is absolutely continuous with P_2, then the value $G(\boldsymbol{\theta}, P_1, P_2)$ in (55) satisfies:

$$\begin{aligned} &G(\boldsymbol{\theta}, P_1, P_2) \\ &= \beta\Big(D\left(P_2\|P^{(P_2,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_1\|P^{(P_2,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) + D(P_1\|P_2)\Big). \end{aligned} \quad (61)$$

Alternatively, if P_2 is absolutely continuous with P_1 then,

$$\begin{aligned} &G(\boldsymbol{\theta}, P_1, P_2) \\ &= \beta\Big(D\left(P_2\|P^{(P_1,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_1\|P^{(P_1,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D(P_2\|P_1)\Big), \end{aligned} \quad (62)$$

where for all $i \in \{1, 2\}$, the probability measure $P^{(P_i,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfies (21) under the assumption that $P_S = P_i$.

Interestingly, absolute continuity of P_1 with respect to P_2 or 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 (55). Note by choosing P_S as a convex combination of P_1 and P_2, always guarantees an explicit expression for $G(\boldsymbol{\theta}, P_1, P_2)$ independently of whether these measures are absolutely continuous with respect to each other.

A. Sensitivity of the Empirical-Risk

The following lemma shows 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 $P_{\boldsymbol{z}}$ in (26). This is formalized by the following lemma.

Lemma 7 (Empirical Risks Via Type): The empirical risk $\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})$ in (5) satisfies

$$\mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta}) = \int \ell(\boldsymbol{\theta}, x, y) \mathrm{d}P_{\boldsymbol{z}}(x, y) = \mathsf{R}_{\boldsymbol{\theta}}(P_{\boldsymbol{z}}), \tag{63}$$

where the measure $P_{\boldsymbol{z}}$ is the type induced by the dataset $\boldsymbol{z}$; and the function ℓ and the functional $\mathsf{R}_{\boldsymbol{\theta}}$ are defined in (4) and (8), respectively.

Proof: The proof is presented in [51, Appendix O]. ■

Equipped with the result in Lemma 7, for a fixed model, the sensitivity of the empirical risk to changes on the datasets can be characterized using the function G in (55). Consider 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. The sensitivity of the empirical risk when the dataset changes from $\boldsymbol{z}_2$ to $\boldsymbol{z}_1$, for a fixed model $\boldsymbol{\theta} \in \mathcal{M}$, is

$$G(\boldsymbol{\theta}, P_{\boldsymbol{z}_1}, P_{\boldsymbol{z}_2}) = \mathsf{L}(\boldsymbol{z}_1, \boldsymbol{\theta}) - \mathsf{L}(\boldsymbol{z}_2, \boldsymbol{\theta}). \tag{64}$$

where the functional G is defined in (55). The following theorem characterizes the sensitivity of the empirical risk.

Theorem 9: 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 (21), it holds that for all $\boldsymbol{\theta} \in \mathcal{M}$,

$$\begin{aligned} G(\boldsymbol{\theta}, P_{\boldsymbol{z}_1}, P_{\boldsymbol{z}_2}) = \beta\Big(& D\left(P_{\boldsymbol{z}_2} \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) \\ & - D\left(P_{\boldsymbol{z}_1} \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_{\boldsymbol{z}_2} \| P_S\right) + D\left(P_{\boldsymbol{z}_1} \| P_S\right)\Big), \end{aligned} \tag{65}$$

where the functional G is in (55); the model $\boldsymbol{\theta}$, and the measures P_S and $P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ satisfy (21).

Proof: The proof follows from the equality in (64), which together with Theorem 8 completes the proof. ■

In Theorem 9, 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, $P_{\boldsymbol{z}_0} = \frac{n_1}{n_0} P_{\boldsymbol{z}_1} + \frac{n_2}{n_0} P_{\boldsymbol{z}_2}$ [22], which satisfies the absolute continuity condition in Theorem 9.

From Theorem 9, 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 WCDG probability measure, i.e., $D\left(P_{\boldsymbol{z}_2} \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_{\boldsymbol{z}_1} \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$; and (b) the difference of the "statistical distance" from the types to the reference measure P_S, i.e., $D\left(P_{\boldsymbol{z}_1} \| P_S\right) - D\left(P_{\boldsymbol{z}_2} \| P_S\right)$.

B. Analysis of the Generalization Gap

Given a model $\boldsymbol{\theta} \in \mathcal{M}$ obtained from the training dataset $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^n$, the *generalization gap* it induces, under the assumption that training and test datapoints are independent and identically distributed according to the probability measure $P_Z \in \triangle(\mathcal{X} \times \mathcal{Y})$, is

$$G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \mathsf{R}_{\boldsymbol{\theta}}(P_Z) - \mathsf{R}_{\boldsymbol{\theta}}(P_{\boldsymbol{z}}). \tag{66}$$

where the functional G is defined in (55) and the function $\mathsf{R}_{\boldsymbol{\theta}}$ is defined in (8). The term $\mathsf{R}_{\boldsymbol{\theta}}(P_{\boldsymbol{z}}) = \mathsf{L}(\boldsymbol{z}, \boldsymbol{\theta})$ is the training empirical risk, see for instance [3, Equation (3.4)]. Alternatively, the term $\mathsf{R}_{\boldsymbol{\theta}}(P_Z)$ is the expected loss under the assumption that the data-generating probability measure is the GTDG probability measure P_Z. This term is often referred to as the *true risk* or *population risk*. In view of this, the value $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (66) describes the variation of the expected empirical risk when the distribution of the datasets changes from the type $P_{\boldsymbol{z}}$ to the GTDG probability measure P_Z, which is characterized by the following lemma.

Lemma 8: The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (66) satisfies:

$$\begin{aligned} & G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \\ & \beta\Big(D\Big(P_{\boldsymbol{z}} \| P^{(P_Z,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big) - D(P_{\boldsymbol{z}} \| P_Z) - D\Big(P_Z \| P^{(P_Z,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)\Big), \end{aligned} \tag{67}$$

where the measure $P^{(P_Z,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is defined in (21) under the assumption that $P_S = P_Z$.

Proof: The proof follows from Corollary 2 by noticing that the type $P_{\boldsymbol{z}}$ is absolutely continuous with respect to P_Z. ■

Lemma 8 highlights the intuition that if the type $P_{\boldsymbol{z}}$ induced by the training dataset $\boldsymbol{z}$ is at arbitrary small "statistical distance" of the GTDG probability measure P_Z, the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (66) is arbitrarily close to zero. This is revealed by the facts that $D\left(P_{\boldsymbol{z}} \| P_Z\right)$ would be arbitrarily small; and so would be the difference $D\left(P_{\boldsymbol{z}} \| P^{(P_Z,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_Z \| P^{(P_Z,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)$.

A more general expression for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (66) is provided by the following corollary of Theorem 8.

Corollary 3: The generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (66) satisfies:

$$\begin{aligned} G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}}) = \beta\Big(& D\left(P_{\boldsymbol{z}} \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) - D\left(P_Z \| P^{(P_S,\beta)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right) \\ & - D\left(P_{\boldsymbol{z}} \| P_S\right) + D\left(P_Z \| P_S\right)\Big), \end{aligned} \tag{68}$$

where the measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{(P_S,\beta)}$ is in (21).

Note that several expressions for the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ in (66) can be obtained from Corollary 3 by choosing the reference P_S and the parameter β in (68).

VII. The Worst-Case Data-Generating Probability Measure and Generalization

Consider a conditioned probability measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, 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, e.g., by sampling such a measure. Hence, the conditional measure $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is referred to as a statistical learning algorithm. This section provides explicit expressions for the generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}$, for some training dataset $\boldsymbol{z}$; and also explicit expressions for the generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ when training datapoints are assumed to be sampled from a particular probability measure in $\triangle(\mathcal{X}\times\mathcal{Y})$. These generalization metrics are referred to as *expected generalization gap* and *doubly-expected generalization gap*.

In order to formally defined the *expected generalization gap*, let $\overline{G} : \triangle(\mathcal{M}) \times \triangle(\mathcal{X}\times\mathcal{Y}) \times \triangle(\mathcal{X}\times\mathcal{Y}) \to \mathbb{R}$ be a functional such that

$$\overline{G}(Q, P_1, P_2) = \int G(\boldsymbol{\theta}, P_1, P_2)\mathrm{d}Q(\boldsymbol{\theta}), \tag{69}$$

where the functional G is defined in (66). Using this notation, the expected generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, when the training dataset is $\boldsymbol{z}$, is

$$\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}}) = \int G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta}). \tag{70}$$

From Corollary 3, by strategically choosing the reference measure P_S and the parameter β in (68), numerous closed-form expressions for the expected generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, when the training dataset is $\boldsymbol{z}$, are obtained. Interestingly, regardless of the choice of P_S and β, the resulting expressions highlight the impact of the training dataset $\boldsymbol{z}$ on the expected generalization gap $\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}})$ in (70).

The expected generalization gap $\overline{G}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}, P_Z, P_{\boldsymbol{z}})$ 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 $P_{\boldsymbol{Z}} \in \triangle((\mathcal{X}\times\mathcal{Y})^n)$, which is assumed to be a product distribution formed by P_Z in (70).

In order to formally defined the *doubly-generalization gap*, let $\overline{\overline{G}} : \triangle(\mathcal{M}|(\mathcal{X}\times\mathcal{Y})^n) \times \triangle(\mathcal{X}\times\mathcal{Y})^n \to \mathbb{R}$ be a function such that

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P) = \int\int G(\boldsymbol{\theta},P,P_{\boldsymbol{z}})\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P(\boldsymbol{z}), \tag{71}$$

where $P_{\boldsymbol{z}}$ is the type induced by a dataset $\boldsymbol{z} \in (\mathcal{X}\times\mathcal{Y})^n$ and the functional G is defined in (66).

Using this notation, the doubly-expected generalization gap induced by the algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, when the training and test datasets are both formed by independent and identically distributed datapoints sampled from the measure P_Z, is

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}, P_{\boldsymbol{Z}}) = \int\int G(\boldsymbol{\theta},P_Z,P_{\boldsymbol{z}})\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}(\boldsymbol{\theta})\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{72}$$

where the generalization gap $G(\boldsymbol{\theta}, P_Z, P_{\boldsymbol{z}})$ is defined in (66); and $P_{\boldsymbol{Z}}$ is a product measure formed by P_Z. In existing literature, the doubly-expected generalization gap is simply referred to as generalization gap or generalization error, and thus, the distinction with generalization gaps that are dependent on specific models and training datasets is often neglected. See for instance [17], [19] and [23]. In this work, such a distinction is important.

A. An Exact Expression for the Doubly-Expected Generalization Gap

As in the case of the expected generalization gap, Corollary 3 leads to numerous closed-form expressions for the doubly-expected generalization gap induced by an algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$, when the training and test datasets are both formed by independent and identically distributed datapoints sampled from the measure P_Z. One of such expressions is of particular interest and is thoroughly studied in this subsection. Nonetheless, before presenting this expression, the following notation is introduced:

$$I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}} \| P_{\boldsymbol{\Theta}}\big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \tag{73}$$

$$= \int D\big(P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}} \| P_{\boldsymbol{Z}}\big)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}); \text{ and} \tag{74}$$

$$L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right) \triangleq \int D\big(P_{\boldsymbol{\Theta}} \| P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}\big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}) \tag{75}$$

$$= \int D\big(P_{\boldsymbol{Z}} \| P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\big)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \tag{76}$$

where $P_{\boldsymbol{Z}}$ is the product measure formed by P_Z; and $P_{\boldsymbol{\Theta}}$ is a probability measure such that for all measurable subsets $\mathcal{A}$ of $\mathcal{M}$,

$$P_{\boldsymbol{\Theta}}(\mathcal{A}) = \int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}(\mathcal{A})\,\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \tag{77}$$

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

$$P_{\boldsymbol{Z}}(\mathcal{B}) = \int P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{B})\,\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \tag{78}$$

Note 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 $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}} \in \triangle(\mathcal{X}\times\mathcal{Y})$. More especifically, for all measurable sets $\mathcal{A}_i$ in $\mathcal{X}\times\mathcal{Y}$, with $i \in \{1, 2, \ldots, n\}$, it follows that

$$P_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_n) = \prod_{t=1}^{n} P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}_t). \tag{79}$$

The quantity $I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}; P_{\boldsymbol{Z}}\right)$ in (73) is the *mutual information* [57] between the random variables $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$, which represent the models and (training and test) datasets, when they are jointly sampled from the probability measure $P_{\boldsymbol{\Theta}\boldsymbol{Z}} \in \triangle(\mathcal{M}\times(\mathcal{X}\times\mathcal{Y})^n)$, whose marginals are $P_{\boldsymbol{\Theta}}$ and $P_{\boldsymbol{Z}}$ in (77) and (78), respectively. Alternatively, the quantity

$L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)$ in (75) is the *lautum information* [58] between such random variables $\boldsymbol{\Theta}$ and $\boldsymbol{Z}$. The equality in (79) justifies the following equalities

$$\frac{1}{n}I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)=\int D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_Z\right)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \text{ and } \quad (80)$$

$$\frac{1}{n}L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)=\int D\left(P_Z\|P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)\mathrm{d}P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \quad (81)$$

which are well known properties of mutual information [59] and lautum information [58].

Using this notation, the main result of this section is presented by the following theorem.

Theorem 10: The doubly-expected generalization gap $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (72), with $P_{\boldsymbol{Z}} \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n\right)$, satisfies

$$\begin{aligned}\frac{1}{\beta}\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})=&\frac{1}{n}L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)\\&+\int D\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)\mathrm{d}P_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\right)\\&-\int D\left(P_Z\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)\mathrm{d}P_{\boldsymbol{\Theta}}\left(\boldsymbol{\theta}\right) \quad (82)\end{aligned}$$

where the term $L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)$ is in (75); the measure $P_{Z|\boldsymbol{\Theta}}$ is defined in (79); and the measure $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}$ is the WCDG probability measure in (21), with reference measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ and

$$\beta\in\left\{t>0:\forall\boldsymbol{\theta}\in\mathcal{M},\int\exp\left(\frac{1}{t}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y)<+\infty\right\}. \quad (83)$$

Proof: The proof is presented in [51, Appendix P]. ■

Theorem 10 appears to be the first exact characterization of the generalization gap induced by an arbitrary algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ in terms of information measures. The following corollary of Theorem 10 unveils an upper bound, which is shown to be intimately linked to the celebrated Gibbs algorithm, and adds interest to the characterization of the generalization gap via the WCDG probability measure.

Theorem 11: The doubly-expected generalization gap $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$, in (72), with $P_{\boldsymbol{Z}} \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n\right)$, satisfies

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})\leqslant\frac{\beta}{n}\left(I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)+L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)\right), \quad (84)$$

where the terms $I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)$ and $L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)$ are in (73) and (75); and β satisfies for all $\boldsymbol{\theta}\in\mathcal{M}$, $D\left(P_Z\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)=0$, with the measures P_Z and $P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}$ defined in (82).

Proof: The proof is presented in [51, Appendix Q]. ■

Note that the exact calculation of the expression in Theorem 10, and even the upper bound in Theorem 11, requires solving a problem of the form in (15) or (16), with reference measure $P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$, for all $\boldsymbol{\theta}\in\mathcal{M}$ and some regularization parameter $\beta>0$ that satisfies (83). In a nutshell, one WCDG probability measure must be calculated for each element of the set of models $\mathcal{M}$. Calculating the upper bound in Theorem 11 is not easier. Obtaining the β in (84) requires solving the equality $D\left(P_Z\|P_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}^{\left(P_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},\beta\right)}\right)=0$ at least once for all models in $\mathcal{M}$ to verify (83). Despite this disproportionate complexity, thanks to this characterization in terms of the WCDG probability measure, it is easy to show that there exists a Gibbs probability measure that achieves such an upper bound in Theorem 11.

B. 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 σ-measure $Q\in\triangle\left(\mathcal{M},\mathscr{B}\left(\mathcal{M}\right)\right)$ [23]. 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 $\boldsymbol{z}\in(\mathcal{X}\times\mathcal{Y})^n$,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}}{\mathrm{d}Q}(\boldsymbol{\theta})=\exp\left(-K_{Q,\boldsymbol{z}}\left(-\frac{1}{\lambda}\right)-\frac{1}{\lambda}\mathsf{L}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\right), \quad (85)$$

where the dataset $\boldsymbol{z}$ represents the training dataset; and the function $K_{Q,\boldsymbol{z}}:\mathbb{R}\to\mathbb{R}$ satisfies

$$K_{Q,\boldsymbol{z}}\left(t\right)=\log\left(\int\exp\left(t\,\mathsf{L}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\right)\mathrm{d}Q(\boldsymbol{\theta})\right). \quad (86)$$

The doubly-expected generalization gap 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

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)},P_{\boldsymbol{Z}}), \quad (87)$$

where the fuctional $\overline{\overline{G}}$ is defined in (72) and the measure $P_{\boldsymbol{Z}}\in\triangle\left(\mathcal{X}\times\mathcal{Y}\right)^n$ is a product measure obtained from P_Z. Such a doubly-expected generalization gap satisfies the following property.

Lemma 9 (Generalization Gap of the Gibbs Algorithm): The generalization gap $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)},P_{\boldsymbol{Z}})$ in (87) satisfies

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)},P_Z)=\lambda\Big(I\Big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)};P_{\boldsymbol{Z}}\Big)+L\Big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)};P_{\boldsymbol{Z}}\Big)\Big), \quad (88)$$

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:

$$I\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)};P_{\boldsymbol{Z}}\right)\triangleq\int D\Big(P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}\|P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\Big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}); \text{ and } \quad (89)$$

$$L\left(P_{\boldsymbol{\Theta}|\boldsymbol{Z}}^{(Q,\lambda)};P_{\boldsymbol{Z}}\right)\triangleq\int D\Big(P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\|P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{\nu}}^{(Q,\lambda)}\Big)\mathrm{d}P_{\boldsymbol{Z}}(\boldsymbol{\nu}), \quad (90)$$

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

$$P_{\boldsymbol{\Theta}}^{(Q,\lambda)}\left(\mathcal{A}\right)=\int P_{\boldsymbol{\Theta}|\boldsymbol{Z}=\boldsymbol{z}}^{(Q,\lambda)}\left(\mathcal{A}\right)\mathrm{d}P_{\boldsymbol{Z}}\left(\boldsymbol{z}\right). \quad (91)$$

Proof: This result has been proved before in the case in which Q is a probability measure in [17]; and in the more general case in which Q is a σ-finite measure in [23]. 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 [1] ■

Note that $I\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)$ in (89) and $L\left(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}};P_{\boldsymbol{Z}}\right)$ in (90) also satisfy that

$$I\left(P^{(Q,\lambda)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}};P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right)=\int D\Big(P^{(Q,\lambda)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\|P_{\boldsymbol{Z}}\Big)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\nu}); \quad (92)$$

and

$$L\left(P^{(Q,\lambda)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}};P^{(Q,\lambda)}_{\boldsymbol{\Theta}}\right)=\int D\Big(P_{\boldsymbol{Z}}\|P^{(Q,\lambda)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\Big)\mathrm{d}P^{(Q,\lambda)}_{\boldsymbol{\Theta}}(\boldsymbol{\nu}), \quad (93)$$

where $P^{(Q,\beta)}_{\boldsymbol{Z}|\boldsymbol{\Theta}} \in \triangle\left((\mathcal{X}\times\mathcal{Y})^n|\mathcal{M}\right)$ is the conditional probability measure that satisfies

$$P_{\boldsymbol{Z}}(\mathcal{B}) = \int P^{(Q,\beta)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{B})\,\mathrm{d}P^{(Q,\beta)}_{\boldsymbol{\Theta}}(\boldsymbol{\theta}). \quad (94)$$

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

$$P^{(Q,\beta)}_{\boldsymbol{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}_1\times\mathcal{A}_2\times\ldots\times\mathcal{A}_n)=\prod_{t=1}^{n}P^{(Q,\beta)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathcal{A}_t). \quad (95)$$

Using this notation, Theorem 11 and Lemma 9 lead to the following theorem.

Theorem 12: Assume that there exits a $\lambda>0$ that satisfies

$$\lambda\in\left\{t>0:\forall\boldsymbol{\theta}\in\mathcal{M},\int\exp\left(\frac{1}{nt}\ell(\boldsymbol{\theta},x,y)\right)\mathrm{d}P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}(x,y)<+\infty\right\}, \quad (96)$$

where the function ℓ is defined in (4); and the measure $P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is defined in (95). 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 (21). If λ satisfies for all $\boldsymbol{\theta}\in\mathcal{M}$,

$$D\left(P_Z\|P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}\right)=0, \quad (97)$$

then, for all $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}\in\triangle\left(\mathcal{M}|\,(\mathcal{X}\times\mathcal{Y})^n\right)$, the doubly-expected generalization gaps $\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (72) and $\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})$ in (87) satisfy

$$\overline{\overline{G}}(P_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}})\leqslant\overline{\overline{G}}(P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}},P_{\boldsymbol{Z}}). \quad (98)$$

Proof: The proof follows by choosing $\lambda=\frac{\beta}{n}$ in Theorem 11 and verifying that under such a choice the resulting expressions in (84) and (88) are identical. ■

Theorem 12 shows that, under the assumption that datasets are sampled from $P_{\boldsymbol{Z}}$, the doubly-expected generalization gap of any algorithm $P_{\boldsymbol{\Theta}|\boldsymbol{Z}}$ is upper-bounded by the doubly-expected generalization gap induced by a particular Gibbs algorithm $P^{(Q,\lambda)}_{\boldsymbol{\Theta}|\boldsymbol{Z}}$. Such a particular Gibbs algorithm induces a posterior for a given model $\boldsymbol{\theta}$, denoted $P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ in (95). 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 $P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$. Surprisingly, from (97), it follows that for all $\boldsymbol{\theta}\in\mathcal{M}$, such a WCDG probability measure $P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ is identical to the actual GTDG probability measure P_Z. Hence, one can conclude that by choosing the reference measure P_S in (21) to be dependent on the models, e.g., $P_S=P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}}$, it is possible to observe that the resulting WCDG probability measure $P^{\left(P^{(Q,\lambda)}_{Z|\boldsymbol{\Theta}=\boldsymbol{\theta}},n\lambda\right)}_{\hat{Z}|\boldsymbol{\Theta}=\boldsymbol{\theta}}$ becomes invariant with respect to $\boldsymbol{\theta}$. This is reminiscent of the principle of *indifference* over which is built the notion of equilibrium in zero-sum games with noisy observations. Nonetheless, this game-theoretic analysis is beyond the scope of this paper. The interested reader is referred to [60] and references therein.

The inequality in (98) 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 (97), the doubly-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.

VIII. Conclusions and Final Remarks

The WCDG probability measure in Theorem 1 has been shown to be a cornerstone of statistical machine learning. This is backed by 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. Interestingly, the WCDG probability measure is a Gibbs probability measure that, for a fixed model, induces an empirical risk that is a sub-Gaussian random variable. All the cumulants of the WCDG are finite and explicit expressions for upper and lower bounds on all cumulants are derived in terms of the cumulants of the reference measure. This analysis has led to the notion of (ϵ,δ)-robustness, which allows the study of the generalization capabilities of any model when it faces data generated from the WCDG probability measure. Finally, thanks to the WCDG, the first explicit expression in terms of information measures of the doubly-expected generalization gap (or generalization error) for any statistical learning algorithm has been presented. This expression has been distilled to obtain an upper-bound on the doubly-expected genealization gap consisting of the sum of the mutual information and the lautum information between the models and the datasets. The bound is shown to be tight for a Gibbs algorithm. This observation reveals the central role of the Gibbs algorithm in the characterization of the doubly-expected generalization gap.

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

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

[5] I. N. Sanov, "On the probability of large deviations of random variables," *Selected Translations in Mathematical Statistics and Probability*, vol. 1, pp. 213–244, 1961.

[6] H. Rahimian and S. Mehrotra, "Frameworks and results in distributionally robust optimization," *Open Journal of Mathematical Optimization*, vol. 3, pp. 1–85, 2022.

[7] 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.

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

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

[10] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning," in *Operations research & management science in the age of analytics*. Informs, 2019, pp. 130–166.

[11] 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.

[12] 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.

[13] E. T. Jaynes, "Information theory and statistical mechanics I," *Physical Review Journals*, vol. 106, no. 4, pp. 620–630, May 1957.

[14] ——, "Information theory and statistical mechanics II," *Physical Review Journals*, vol. 108, no. 2, pp. 171–190, Oct. 1957.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, vol. 51, Cadiz, Spain, May 2016, pp. 1232–1240.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] S. Masiha, A. Gohari, and M. H. Yassaee, "f-divergences and their applications in lossy compression and bounding generalization error," *IEEE Transactions on Information Theory*, pp. 7245–7254, Apr. 2023.

[36] 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.

[37] 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.

[38] M. Gastpar, I. Nachum, J. Shafer, and T. Weinberger, "Fantastic generalization measures are nowhere to be found," *arXiv preprint arXiv:2309.13658*, 2023.

[39] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik, "Model complexity control for regression using VC generalization bounds," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1075–1089, Sep. 1999.

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

[41] D. Cullina, A. N. Bhagoji, and P. Mittal, "PAC-learning in the presence of adversaries," *Advances in Neural Information Processing Systems*, vol. 31, no. 1, pp. 1–12, Dec. 2018.

[42] M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor, "PAC-Bayes unleashed: Generalisation bounds with unbounded losses," *Entropy*, vol. 23, no. 10, pp. 1–20, Oct. 2021.

[43] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, Jan. 2019.

[44] 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.

[45] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Empirical risk minimization with relative entropy regularization type-II," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9508, May. 2023.

[46] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, no. 1, pp. 299–318, Jun. 1967.

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

[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[49] C. Villani, *Optimal transport: Old and new*, 1st ed. Berlin, Germany: Springer, 2009.

[50] I. Sason and S. Verdú, "f -divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, Jun. 2016.

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

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

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

[54] W. Rudin, *Principles of mathematical analysis*, 1st ed. New York, NY, USA: McGraw-Hill Book Company, Inc., 1953.

[55] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, 1st ed. New York, NY, USA: Cambridge University Press, 2019.

[56] 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.

[57] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, Jul. 1948.

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

[59] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.

[60] 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.