Variations on the Expectation Due to Changes in the Probability Measure

Samir M. Perlaza*†‡ and Gaetan Bisson‡
*INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France.
†ECE Dept. Princeton University, Princeton, 08544 NJ, USA.
‡Laboratoire GAATI, University of French Polynesia.

***Abstract*—Closed-form expressions are presented for the variation of the expectation of a given function due to changes in the probability measure used for the expectation. They unveil interesting connections with Gibbs probability measures, the mutual information, and the lautum information.**

I. Introduction

Let m be a positive integer and denote by $\triangle(\mathbb{R}^m)$ the set of all probability measures on the measurable space $(\mathbb{R}^m, \mathscr{B}(\mathbb{R}^m))$, with $\mathscr{B}(\mathbb{R}^m)$ being the Borel σ-algebra on $\mathbb{R}^m$. Given a Borel measurable function $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, consider the functional $\mathsf{G}_h : \mathbb{R}^n \times \triangle(\mathbb{R}^m) \times \triangle(\mathbb{R}^m) \to \mathbb{R}$ such that

$$\mathsf{G}_h(x, P_1, P_2) = \int h(x,y) \mathrm{d}P_1(y) - \int h(x,y) \mathrm{d}P_2(y), \quad (1)$$

which quantifies the variation of the expectation of the measurable function h due to changing the probability measure from P_2 to P_1. Such a functional is defined when both integrals exist and are finite.

In order to define the expectation of $\mathsf{G}_h(x, P_1, P_2)$ with respect to x, the structure formalized below is required.

Definition 1: A family $P_{Y|X} \triangleq (P_{Y|X=x})_{x \in \mathbb{R}^n}$ of elements of $\triangle(\mathbb{R}^m)$ indexed by $\mathbb{R}^n$ is said to be a conditional probability measure if, for all sets $\mathcal{A} \in \mathscr{B}(\mathbb{R}^m)$, the map

$$\begin{aligned} \mathbb{R}^n &\to [0,1] \\ x &\mapsto P_{Y|X=x}(\mathcal{A}) \end{aligned}$$

is Borel measurable. The set of all such conditional probability measures is denoted by $\triangle(\mathbb{R}^m|\mathbb{R}^n)$.

In this setting, consider the functional $\bar{\mathsf{G}}_h : \triangle(\mathbb{R}^m|\mathbb{R}^n) \times \triangle(\mathbb{R}^m|\mathbb{R}^n) \times \triangle(\mathbb{R}^n) \to \mathbb{R}$ such that

$$\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right) = \int \mathsf{G}_h\left(x, P^{(1)}_{Y|X=x}, P^{(2)}_{Y|X=x}\right) \mathrm{d}P_X(x). \quad (2)$$

This quantity can be interpreted as the variation of the integral (expectation) of the function h when the probability measure changes from the joint probability measure $P^{(1)}_{Y|X} P_X$ to another joint probability measure $P^{(2)}_{Y|X} P_X$, both in $\triangle(\mathbb{R}^m \times \mathbb{R}^n)$. This follows from (2) by observing that

$$\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right) = \int h(x,y) \mathrm{d}P^{(1)}_{Y|X} P_X(y,x) - \int h(x,y) \mathrm{d}P^{(2)}_{Y|X} P_X(y,x). \quad (3)$$

Special attention is given to the quantity $\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right)$, for some $P_{Y|X} \in \triangle(\mathbb{R}^m|\mathbb{R}^n)$, with P_Y being the marginal of the joint probability measure $P_{Y|X} \cdot P_X$. That is, for all sets $\mathcal{A} \in \mathscr{B}(\mathbb{R}^m)$,

$$P_Y(\mathcal{A}) = \int P_{Y|X=x}(\mathcal{A}) \mathrm{d}P_X(x). \quad (4)$$

Its relevance stems from the fact that it captures the variation of the expectation of the function h when the probability measure changes from the joint probability measure $P_{Y|X} P_X$ to the product of its marginals $P_Y P_X$. That is,

$$\begin{aligned} &\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right) \\ &= \int \left(\int h(x,y) \mathrm{d}P_Y(y) - \int h(x,y) \mathrm{d}P_{Y|X=x}(y) \right) \mathrm{d}P_X(x) \\ &= \int h(x,y) \mathrm{d}P_Y P_X(y,x) - \int h(x,y) \mathrm{d}P_{Y|X} P_X(y,x). \quad (5) \end{aligned}$$

A. Contributions

This work makes two key contributions: First, a closed-form expression for the variation $\mathsf{G}_h(x, P_1, P_2)$ in (1) is provided for a fixed $x \in \mathbb{R}^n$ and two arbitrary measures P_1 and P_2, expressed in terms of information measures. Second, a closed-form expression for the expected variation $\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)$ in (2) is presented also in terms of information measures, for arbitrary conditional probability measures $P^{(1)}_{Y|X}$ and $P^{(2)}_{Y|X}$, along with an arbitrary probability measure P_X.

As a byproduct, specific closed-form expressions are provided for the variation $\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right)$ in (5) in terms of both mutual information [1], [2], and lautum information [3]. The specific case in which $P_{Y|X}$ is a Gibbs conditional probability measure is highlighted as $\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right)$ is equal (up to a constant factor) to the sum of mutual and lautum information of the joint probability measure $P_{Y|X} P_X$.

This work is supported in part by the European Commission through the H2020-MSCA-RISE-2019 project 872172; the French National Agency for Research (ANR) through the Project ANR-21-CE25-0013 and the project ANR-22-PEFT-0010 of the France 2030 program PEPR Réseaux du Futur; and in part by the Agence de l'innovation de défense (AID) through the project UK-FR 2024352.

B. Relevance and Applications

The relevance of the study of the variation of the integral (expectation) of h (for some fixed $x \in \mathbb{R}^n$) due to a measure change from P_2 to P_1, i.e., the value $\mathsf{G}_h(x, P_1, P_2)$ in (1), is evidenced by its central role in the definition of *integral probability metrics* (IPMs) [4], [5]. Using the notation in (1), an IPM results from the optimization problem

$$\sup_{h \in \mathcal{H}} \left| \mathsf{G}_h(x, P_1, P_2) \right|, \tag{6}$$

for some fixed $x \in \mathbb{R}^n$ and a particular class of functions $\mathcal{H}$. Note for instance that the maximum mean discrepancy is an IPM [6], as well as the Wasserstein distance of order one [7]–[10].

Other areas of mathematics in which the variation $\mathsf{G}_h(x, P_1, P_2)$ in (1) plays a key role is distributionally robust optimization (DRO) [11], [12] and optimization with relative entropy regularization [13], [14]. In these areas, the variation $\mathsf{G}_h(x, P_1, P_2)$ is a central tool. See for instance, [15], [16].

Variations of the form $\mathsf{G}_h(x, P_1, P_2)$ in (1) have also been studied in [17] and [18] in the particular case of statistical machine learning for the analysis of generalization error. The central observation is that the generalization error of machine learning algorithms can be written in the form $\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right)$ in (5). This observation is the main building block of the *method of gaps* introduced in [18], which leads to a number of closed-form expressions for the generalization error involving mutual information, lautum information, among other information measures.

The results of the present paper unify and generalize many special cases that were obtained in some of the articles discussed above.

II. Preliminaries

The main results presented in this work involve Gibbs conditional probability measures. Such measures are parametrized by a Borel measurable function $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$; a σ-finite measure Q on $\mathbb{R}^m$; and a vector $x \in \mathbb{R}^n$. Note that the variable x will remain inactive until Section IV. Although it is introduced now for consistency, it could be removed altogether from all results presented in this section and Section III.

Denote by $\mathsf{K}_{h,Q,x} : \mathbb{R} \to \mathbb{R}$ the function that satisfies

$$\mathsf{K}_{h,Q,x}(t) = \log\left(\int \exp\left(t\, h(x,y)\right) \mathrm{d}Q(y) \right). \tag{7}$$

Under the assumption that Q is a probability measure, the function $\mathsf{K}_{h,Q,x}$ in (7) is the cumulant generating function of the random variable $h(x, Y)$, for some fixed $x \in \mathbb{R}^n$ and $Y \sim Q$. Using this notation, the definition of the Gibbs conditional probability measure is presented hereunder.

Definition 2 (Gibbs Conditional Probability Measure): Given a Borel measurable function $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$; a σ-finite measure Q on $\mathbb{R}^m$; and a $\lambda \in \mathbb{R}$, the probability measure $P^{(h,Q,\lambda)}_{Y|X} \in \triangle(\mathbb{R}^m | \mathbb{R}^n)$ is said to be an (h, Q, λ)-Gibbs conditional probability measure if

$$\forall x \in \mathbb{R}^n,\ \mathsf{K}_{h,Q,x}(-\lambda) < +\infty; \tag{8}$$

and for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$\frac{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}{\mathrm{d}Q}(y) = \exp\left(-\lambda h(x,y) - \mathsf{K}_{h,Q,x}(-\lambda)\right), \tag{9}$$

where the function $\mathsf{K}_{h,Q,x}$ is defined in (7).

Note that, while $P^{(h,Q,\lambda)}_{Y|X}$ is an (h, Q, λ)-Gibbs conditional probability measure, the measure $P^{(h,Q,\lambda)}_{Y|X=x}$, obtained by conditioning it upon a given vector $x \in \mathbb{R}^n$, is referred to as an (h, Q, λ)-Gibbs probability measure.

Condition 8 is easily met under certain conditions. For instance, if h is a nonnegative function and Q is a finite measure, then it holds for all $\lambda \in (-\infty, 0)$. Let $\triangle_Q(\mathbb{R}^m) \triangleq \{P \in \triangle(\mathbb{R}^m) : P \ll Q\}$, with $P \ll Q$ standing for "P absolutely continuous with respect to Q". The relevance of (h, Q, λ)-Gibbs probability measures relies on the fact that under some conditions, they are the unique solutions to problems of the form,

$$\min_{P \in \triangle_Q(\mathbb{R}^m)} \int h(x,y) \mathrm{d}P(y) + \frac{1}{\lambda} D\left(P \| Q\right), \text{ and} \tag{10}$$

$$\max_{P \in \triangle_Q(\mathbb{R}^m)} \int h(x,y) \mathrm{d}P(y) + \frac{1}{\lambda} D\left(P \| Q\right), \tag{11}$$

where $\lambda \in \mathbb{R} \setminus \{0\}$, $x \in \mathbb{R}$, and $D(P\|Q)$ denotes the relative entropy (or KL divergence) of P with respect to Q.

Lemma 1: Assume that the optimization problem in (10) (respectively, in (11)) admits solutions. Then, if $\lambda > 0$ (respectively, if $\lambda < 0$), the probability measure $P^{(h,Q,\lambda)}_{Y|X=x}$ in (9) is the unique solution.

Proof: The uniqueness of the solutions to the optimization problems in (10) and (11) arises from the nature of their objective functions: the objective function in (10) is strictly convex with respect to the measure P when $\lambda > 0$, while the function in (11) is strictly concave when $\lambda < 0$. See for instance, [13, Theorem 2]. The proofs that these unique solutions correspond to (h, Q, λ)-Gibbs probability measures follow the same approach as the proofs of [13, Theorem 3] and [17, Theorem 1]. ■

The following lemma highlights a key property of (h, Q, λ)-Gibbs conditional probability measures.

Lemma 2: Given an (h, Q, λ)-Gibbs probability measure, denoted by $P^{(h,Q,\lambda)}_{Y|X=x}$, with $x \in \mathbb{R}^n$,

$$-\frac{1}{\lambda}\mathsf{K}_{h,Q,x}(-\lambda)$$
$$=\int h(x,y)\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}(y) + \frac{1}{\lambda}D\left(P^{(h,Q,\lambda)}_{Y|X=x}\|Q\right) \tag{12}$$
$$=\int h(x,y)\mathrm{d}Q(y) - \frac{1}{\lambda}D\left(Q\|P^{(h,Q,\lambda)}_{Y|X=x}\right); \tag{13}$$

moreover, if $\lambda > 0$, this further equals

$$=\min_{P\in\triangle_Q(\mathbb{R}^m)} \int h(x,y)\mathrm{d}P(y) + \frac{1}{\lambda}D(P\|Q); \tag{14}$$

alternatively, if $\lambda < 0$,

$$=\max_{P\in\triangle_Q(\mathbb{R}^m)} \int h(x,y)\mathrm{d}P(y) + \frac{1}{\lambda}D(P\|Q), \tag{15}$$

where the function $\mathsf{K}_{h,Q,x}$ is defined in (7).

Proof: The proof of (12) follows from taking the logarithm of both sides of (9) and integrating with respect to $P^{(h,Q,\lambda)}_{Y|X=x}$. As for the proof of (13), it follows by noticing that for all $(x, y) \in \mathbb{R}^n \times \operatorname{supp} Q$, the Radon-Nikodym derivative $\frac{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}{\mathrm{d}Q}(y)$ in (9) is strictly positive. Thus, $\frac{\mathrm{d}Q}{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}(y) = \left(\frac{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}{\mathrm{d}Q}(y)\right)^{-1}$. Hence, taking the negative logarithm of both sides of (9) and integrating with respect to Q leads to (13). Finally, the equalities in (14) and (15) follow from Lemma 1 and (12). ∎

The following lemma introduces the main building block of this work, which is a characterization of the deviation $\mathsf{G}_h\left(x, P, P^{(h,Q,\lambda)}_{Y|X=x}\right)$.

Lemma 3: Consider an (h, Q, λ)-Gibbs probability measure, denoted by $P^{(h,Q,\lambda)}_{Y|X=x} \in \triangle(\mathbb{R}^m)$, with $\lambda \neq 0$ and $x \in \mathbb{R}$. For all $P \in \triangle_Q(\mathbb{R}^m)$,

$$\mathsf{G}_h\left(x, P, P^{(h,Q,\lambda)}_{Y|X=x}\right)$$
$$=\frac{1}{\lambda}\left(D\left(P\|P^{(h,Q,\lambda)}_{Y|X=x}\right) + D\left(P^{(h,Q,\lambda)}_{Y|X=x}\|Q\right) - D(P\|Q)\right). \tag{16}$$

Proof: The proof follows by noticing that for all $P \in \triangle_Q(\mathbb{R}^m)$,

$$D\left(P\|P^{(h,Q,\lambda)}_{Y|X=x}\right)$$
$$=\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}(y)\right)\mathrm{d}P(y) \tag{17}$$
$$=\int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}(y)\frac{\mathrm{d}P}{\mathrm{d}Q}(y)\right)\mathrm{d}P(y) \tag{18}$$
$$=\int \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}(y)\right)\mathrm{d}P(y) + D(P\|Q) \tag{19}$$
$$=\lambda\int h(x,y)\mathrm{d}P(y) + \mathsf{K}_{h,Q,x}(-\lambda) + D(P\|Q) \tag{20}$$
$$=\lambda\mathsf{G}_h\left(x, P, P^{(h,Q,\lambda)}_{Y|X=x}\right) - D\left(P^{(h,Q,\lambda)}_{Y|X=x}\|Q\right) + D(P\|Q), \tag{21}$$

where (20) follows from (9); and (21) follows from (12). ∎

It is interesting to highlight that $\mathsf{G}_h\left(x, P, P^{(h,Q,\lambda)}_{Y|X=x}\right)$ in (16) characterizes the variation of the function $h(x,\cdot) : \mathbb{R}^m \to \mathbb{R}$ from the solutions to the optimization problems in (10) and (11), if they exist, to an alternative measure P.

III. Characterization of $\mathsf{G}_h(x, P_1, P_2)$ in (1)

The main result of this section is the following theorem.

Theorem 4: For all probability measures P_1 and P_2, both absolutely continuous with respect to a given σ-finite measure Q on $\mathbb{R}^m$, the variation $\mathsf{G}_h(x, P_1, P_2)$ in (1) satisfies,

$$\mathsf{G}_h(x, P_1, P_2) = \frac{1}{\lambda}\Big(D\left(P_1\|P^{(h,Q,\lambda)}_{Y|X=x}\right) - D\left(P_2\|P^{(h,Q,\lambda)}_{Y|X=x}\right)$$
$$+D(P_2\|Q) - D(P_1\|Q)\Big), \tag{22}$$

where the probability measure $P^{(h,Q,\lambda)}_{Y|X=x}$, with $\lambda \neq 0$, is an (h, Q, λ)-Gibbs probability measure.

Proof: The proof follows from Lemma 3 and by observing that $\mathsf{G}_h(x, P_1, P_2) = \mathsf{G}_h\left(x, P_1, P^{(h,Q,\lambda)}_{Y|X=x}\right) - \mathsf{G}_h\left(x, P_2, P^{(h,Q,\lambda)}_{Y|X=x}\right)$. ∎

Theorem 4 might be particularly simplified in the case in which the reference measure Q is a probability measure. Consider for instance the case in which P_1 is absolutely continuous with respect to P_2 (or P_2 is absolutely continuous with respect to P_1). In such a case, the reference measure might be chosen as P_2 (or P_1), as shown hereunder.

Corollary 5: Consider the variation $\mathsf{G}_h(x, P_1, P_2)$ in (1). If the probability measure P_1 is absolutely continuous with respect to P_2, then,

$$\mathsf{G}_h(x, P_1, P_2) = \frac{1}{\lambda}\Big(D\left(P_1\|P^{(h,P_2,\lambda)}_{Y|X=x}\right) - D\left(P_2\|P^{(h,P_2,\lambda)}_{Y|X=x}\right)$$
$$-D(P_1\|P_2)\Big). \tag{23}$$

Alternatively, if the probability measure P_2 is absolutely continuous with respect to P_1, then,

$$\mathsf{G}_h(x, P_1, P_2) = \frac{1}{\lambda}\Big(D\left(P_1\|P^{(h,P_1,\lambda)}_{Y|X=x}\right) - D\left(P_2\|P^{(h,P_1,\lambda)}_{Y|X=x}\right)$$
$$+D(P_2\|P_1)\Big), \tag{24}$$

where the probability measures $P^{(h,P_1,\lambda)}_{Y|X=x}$ and $P^{(h,P_2,\lambda)}_{Y|X=x}$ are respectively (h, P_1, λ)- and (h, P_2, λ)-Gibbs probability measures, with $\lambda \neq 0$.

In the case in which neither P_1 is absolutely continuous with respect to P_2; nor P_2 is absolutely continuous with respect to P_1, the reference measure Q in Theorem 4 can always be chosen as a convex combination of P_1 and P_2. That is, for all Borel sets $\mathcal{A} \in \mathscr{B}(\mathbb{R}^m)$, $Q(\mathcal{A}) = \alpha P_1(\mathcal{A}) + (1-\lambda) P_2(\mathcal{A})$, with $\alpha \in (0,1)$.

Theorem 4 can be specialized to the specific cases in which Q is the Lebesgue or the counting measure.

a) If Q is the Lebesgue measure: the probability measures P_1 and P_2 in (22) admit probability density functions f_1 and f_2, respectively. Moreover, the terms $-D(P_1\|Q)$ and $-D(P_2\|Q)$ are Shannon's differential entropies [1] induced by P_1 and P_2, denoted by $\mathsf{h}(P_1)$ and $\mathsf{h}(P_2)$, respectively. That is, for all $i \in \{1,2\}$,

$$\mathsf{h}(P_i) \triangleq -\int f_i(x) \log f_i(x) \mathrm{d}x. \tag{25}$$

The probability measure $P^{(h,Q,\lambda)}_{Y|X=x}$, with $\lambda \neq 0$, $x \in \mathbb{R}^n$, and Q the Lebesgue measure, possesses a probability density function, denoted by $f^{(h,Q,\lambda)}_{Y|X=x} : \mathbb{R}^m \to (0,+\infty)$, which satisfies

$$f^{(h,Q,\lambda)}_{Y|X=x}(y) = \frac{\exp(-\lambda h(x,y))}{\int \exp(-\lambda h(x,y)) \mathrm{d}y}. \tag{26}$$

b) If Q is the counting measure: the probability measures P_1 and P_2 in (22) admit probability mass functions $p_1 : \mathcal{Y} \to [0,1]$ and $p_2 : \mathcal{Y} \to [0,1]$, with $\mathcal{Y}$ a countable subset of $\mathbb{R}^m$. Moreover, $-D(P_1\|Q)$ and $-D(P_2\|Q)$ are respectively Shannon's discrete entropies [1] induced by P_1 and P_2, denoted by $\mathsf{H}(P_1)$ and $\mathsf{H}(P_2)$, respectively. That is, for all $i \in \{1,2\}$,

$$\mathsf{H}(P_i) \triangleq -\sum_{y \in \mathcal{Y}} p_i(y) \log p_i(y). \tag{27}$$

The probability measure $P^{(h,Q,\lambda)}_{Y|X=x}$, with $\lambda \neq 0$ and Q the counting measure, possesses a conditional probability mass function, denoted by $p^{(h,Q,\lambda)}_{Y|X=x} : \mathcal{Y} \to (0,+\infty)$, which satisfies

$$p^{(h,Q,\lambda)}_{Y|X=x}(y) = \frac{\exp(-\lambda h(x,y))}{\sum_{y \in \mathcal{Y}} \exp(-\lambda h(x,y))}. \tag{28}$$

IV. Characterizations of $\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)$ in (2)

The main result of this section is a characterization of $\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)$ in (2).

Theorem 6: Consider the variation $\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)$ in (2) and assume that for all $x \in \mathbb{R}^n$, the probability measures $P^{(1)}_{Y|X=x}$ and $P^{(2)}_{Y|X=x}$ are both absolutely continuous with respect to a σ-measure Q. Then,

$$\begin{aligned}&\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)\\ &= \frac{1}{\lambda}\int \Big(D\left(P^{(1)}_{Y|X=x}\|P^{(h,Q,\lambda)}_{Y|X=x}\right) - D\left(P^{(2)}_{Y|X=x}\|P^{(h,Q,\lambda)}_{Y|X=x}\right)\\ &\quad + D\left(P^{(2)}_{Y|X=x}\|Q\right) - D\left(P^{(1)}_{Y|X=x}\|Q\right)\Big) \mathrm{d}P_X(x),\end{aligned} \tag{29}$$

where the probability measure $P^{(h,Q,\lambda)}_{Y|X}$, with $\lambda \neq 0$, is an (h,Q,λ)-Gibbs conditional probability measure.

Proof: The proof follows from (2) and Theorem 4. ■

Note that, from (2), it follows that the general expression for the expected variation $\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)$ might be simplified according to Corollary 5. For instance, if for all $x \in \mathbb{R}^m$, the probability measure $P^{(1)}_{Y|X=x}$ is absolutely continuous with respect to $P^{(2)}_{Y|X=x}$, the measure $P^{(2)}_{Y|X=x}$ can be chosen to be the reference measure in the calculation of $\mathsf{G}_h\left(x, P^{(1)}_{Y|X=x}, P^{(2)}_{Y|X=x}\right)$ in (2). This observation leads to the following corollary of Theorem 6.

Corollary 7: Consider the variation $\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)$ in (2) and assume that for all $x \in \mathbb{R}^n$, the probability measures $P^{(1)}_{Y|X=x}$ is absolutely continuous with respect to $P^{(2)}_{Y|X=x}$. Then,

$$\begin{aligned}&\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)\\ &= \frac{1}{\lambda}\int \Bigg(D\left(P^{(1)}_{Y|X=x}\|P^{\left(h,P^{(2)}_{Y|X=x},\lambda\right)}_{Y|X=x}\right)\\ &\quad - D\left(P^{(2)}_{Y|X=x}\|P^{\left(h,P^{(2)}_{Y|X=x},\lambda\right)}_{Y|X=x}\right)\\ &\quad - D\left(P^{(1)}_{Y|X=x}\|P^{(2)}_{Y|X=x}\right)\Bigg) \mathrm{d}P_X(x).\end{aligned} \tag{30}$$

Alternatively, if for all $x \in \mathbb{R}^n$, the probability measure $P^{(2)}_{Y|X=x}$ is absolutely continuous with respect to $P^{(1)}_{Y|X=x}$, then,

$$\begin{aligned}&\bar{\mathsf{G}}_h\left(P^{(1)}_{Y|X}, P^{(2)}_{Y|X}, P_X\right)\\ &= \frac{1}{\lambda}\int \Bigg(D\left(P^{(1)}_{Y|X=x}\|P^{\left(h,P^{(1)}_{Y|X=x},\lambda\right)}_{Y|X=x}\right)\\ &\quad - D\left(P^{(2)}_{Y|X=x}\|P^{\left(h,P^{(1)}_{Y|X=x},\lambda\right)}_{Y|X=x}\right)\\ &\quad + D\left(P^{(2)}_{Y|X=x}\|P^{(1)}_{Y|X=x}\right)\Bigg) \mathrm{d}P_X(x),\end{aligned} \tag{31}$$

where the measures $P^{\left(h,P^{(1)}_{Y|X=x},\lambda\right)}_{Y|X=x}$ and $P^{\left(h,P^{(2)}_{Y|X=x},\lambda\right)}_{Y|X=x}$ are $(h, P^{(1)}_{Y|X=x}, \lambda)$- and $(h, P^{(2)}_{Y|X=x}, \lambda)$-Gibbs probability measures, respectively.

The Gibbs probability measures $P^{\left(h,P^{(1)}_{Y|X=x},\lambda\right)}_{Y|X=x}$ and $P^{\left(h,P^{(2)}_{Y|X=x},\lambda\right)}_{Y|X=x}$ in Corollary 7 are particularly interesting as their reference measures depend on x. Gibbs measures of this form appear, for instance, in [13, Corollary 10].

Two special cases are particularly noteworthy.

a) When the reference measure Q is the Lebesgue measure: observe that the terms $-\int D\left(P^{(1)}_{Y|X=x}\|Q\right) \mathrm{d}P_X(x)$ and $-\int D\left(P^{(2)}_{Y|X=x}\|Q\right) \mathrm{d}P_X(x)$ in (29) both become Shannon's differential conditional entropy, denoted by

$\mathsf{h}\left(P^{(1)}_{Y|X}|P_X\right)$ and $\mathsf{h}\left(P^{(2)}_{Y|X}|P_X\right)$, respectively. That is, for all $i \in \{1,2\}$,

$$\mathsf{h}\left(P^{(i)}_{Y|X}|P_X\right) \triangleq \int \mathsf{h}\left(P^{(i)}_{Y|X=x}\right) \mathrm{d}P_X(x), \qquad (32)$$

where h is the entropy functional in (25).

b) When the reference measure Q is the counting measure: the terms $-\int D\left(P^{(1)}_{Y|X=x}\|Q\right) \mathrm{d}P_X(x)$ and $-\int D\left(P^{(2)}_{Y|X=x}\|Q\right) \mathrm{d}P_X(x)$ in (29) both become Shannon's discrete conditional entropies, denoted by $\mathsf{H}\left(P^{(1)}_{Y|X}|P_X\right)$ and $\mathsf{H}\left(P^{(2)}_{Y|X}|P_X\right)$, respectively. That is, for all $i \in \{1,2\}$,

$$\mathsf{H}\left(P^{(i)}_{Y|X}|P_X\right) \triangleq \int \mathsf{H}\left(P^{(i)}_{Y|X=x}\right) \mathrm{d}P_X(x), \qquad (33)$$

where H is the entropy functional in (27).

V. Characterizations of $\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right)$ in (5)

The main result of this section is a characterization of $\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right)$ in (5), which describes the variation of the expectation of the function h when the probability measure changes from the joint probability measure $P_{Y|X}P_X$ to the product of its marginals $P_Y \cdot P_X$. This result is presented hereunder and involves the mutual information $I\left(P_{Y|X}; P_X\right)$ and lautum information $L\left(P_{Y|X}; P_X\right)$, defined as follows:

$$I\left(P_{Y|X}; P_X\right) \triangleq \int D\left(P_{Y|X=x}\|P_Y\right) \mathrm{d}P_X(x); \text{ and} \qquad (34)$$

$$L\left(P_{Y|X}; P_X\right) \triangleq \int D\left(P_Y\|P_{Y|X=x}\right) \mathrm{d}P_X(x). \qquad (35)$$

Theorem 8: Consider the expected variation $\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right)$ in (5) and assume that, for all $x \in \mathbb{R}^n$:

(a) The probability measures P_Y and $P_{Y|X=x}$ are both absolutely continuous with respect to a given σ-finite measure Q; and

(b) The probability measures P_Y and $P_{Y|X=x}$ are mutually absolutely continuous.

Then, it follows that

$$\begin{aligned}
&\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right) \\
=&\frac{1}{\lambda}\Bigg(I\left(P_{Y|X}; P_X\right) + L\left(P_{Y|X}; P_X\right) \\
&+\int\int \log\left(\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}(y)\right) \mathrm{d}P_Y(y)\mathrm{d}P_X(x) \\
&-\int\int \log\left(\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}(y)\right) \mathrm{d}P_{Y|X=x}(y)\mathrm{d}P_X(x)\Bigg), \qquad (36)
\end{aligned}$$

where the probability measure $P^{(h,Q,\lambda)}_{Y|X}$, with $\lambda \neq 0$, is an (h,Q,λ)-Gibbs conditional probability measure.

Proof: The proof follows from Theorem 6, which holds under assumption (a) and leads to

$$\begin{aligned}
&\bar{\mathsf{G}}_h\left(P_Y, P_{Y|X}, P_X\right) \\
=&\frac{1}{\lambda}\int\Bigg(D\left(P_Y\|P^{(h,Q,\lambda)}_{Y|X=x}\right) - D\left(P_{Y|X=x}\|P^{(h,Q,\lambda)}_{Y|X=x}\right) \\
&+D\left(P_{Y|X=x}\|Q\right) - D\left(P_Y\|Q\right)\Bigg)\mathrm{d}P_X(x). \qquad (37)
\end{aligned}$$

The proof continues by noticing that

$$\int D\left(P_{Y|X=x}\|Q\right) \mathrm{d}P_X(x) = I\left(P_{Y|X}; P_X\right) + D\left(P_Y\|Q\right), \qquad (38)$$

and

$$\begin{aligned}
&\int D\left(P_Y\|P^{(h,Q,\lambda)}_{Y|X=x}\right) \mathrm{d}P_X(x) = L\left(P_{Y|X}; P_X\right) \\
&+\int\int \log\left(\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}(y)\right) \mathrm{d}P_Y(y)\mathrm{d}P_X(x). \qquad (39)
\end{aligned}$$

Finally, using (38) and (39) in (37) yields (36), which completes the proof. ■

An interesting observation from Theorem 8 is that the last two terms in the right-hand side of (36) are both zero in the case in which $P_{Y|X}$ is an (h,Q,λ)-Gibbs conditional probability measure. This is observation is highlighted by the following corollary.

Corollary 9: Consider an (h,Q,λ)-Gibbs conditional probability measure, denoted by $P^{(h,Q,\lambda)}_{Y|X} \in \triangle\left(\mathbb{R}^m|\mathbb{R}^n\right)$, with $\lambda \neq 0$; and a probability measure $P_X \in \triangle\left(\mathbb{R}^n\right)$. Let the measure $P^{(h,Q,\lambda)}_{Y} \in \triangle\left(\mathbb{R}^m\right)$ be such that for all sets $\mathcal{A} \in \mathscr{B}\left(\mathbb{R}^m\right)$,

$$P^{(h,Q,\lambda)}_{Y}\left(\mathcal{A}\right) = \int P^{(h,Q,\lambda)}_{Y|X=x}\left(\mathcal{A}\right) \mathrm{d}P_X(x). \qquad (40)$$

Then,

$$\begin{aligned}
&\bar{\mathsf{G}}_h\left(P^{(h,Q,\lambda)}_{Y}, P^{(h,Q,\lambda)}_{Y|X}, P_X\right) \\
=&\frac{1}{\lambda}\left(I\left(P^{(h,Q,\lambda)}_{Y|X}; P_X\right) + L\left(P^{(h,Q,\lambda)}_{Y|X}; P_X\right)\right). \qquad (41)
\end{aligned}$$

VI. Conclusion

Closed-form expressions for the variation of the integral of a given measurable function due to changes in the probability measure has been presented. In these expressions, the Gibbs probability measure plays a central role and brings significant flexibility as some of its parameters can be chosen up to mild conditions. In the case of joint probability measures, the focus has been on two particular measure changes, which unveil connections with both mutual and lautum information. First, one of the marginal probability measures remains the same after the change; and second, the joint probability measure changes to the product of its marginals.

References

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

[2] ——, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 623–656, Oct. 1948.

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

[4] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in applied probability*, vol. 29, no. 2, pp. 429–443, 1997.

[5] V. M. Zolotarev, "Probability metrics," *Teoriya Veroyatnostei i ee Primeneniya*, vol. 28, no. 2, pp. 264–287, 1983.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.

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

[8] W. Liu, G. Yu, L. Wang, and R. Liao, "An information-theoretic framework for out-of-distribution generalization with applications to stochastic gradient Langevin dynamics," *arXiv preprint arXiv:2403.19895*, 2024.

[9] ——, "An information-theoretic framework for out-of-distribution generalization," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, July 2024, pp. 2670–2675.

[10] R. Agrawal and T. Horel, "Optimal bounds between f-divergences and integral probability metrics," *Journal of Machine Learning Research*, vol. 22, no. 128, pp. 1–59, 2021.

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

[12] C. Xu, J. Lee, X. Cheng, and Y. Xie, "Flow-based distributionally robust optimization," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 62 – 77, Apr. 2024.

[13] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 5122 – 5161, Jul. 2024.

[14] 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.

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

[16] 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.

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

[18] S. M. Perlaza and X. Zou, "The generalization error of machine learning algorithms," *arXiv preprint arXiv:2411.12030*, 2024.