Title Journal

Abstract

Abstract of the Journal

Index Terms

Supervised Learning, Empirical Risk Minimization; Relative Entropy; Regularization; Gibbs Measure; Inductive Bias; Gibbs Algorithm; Sensitivity; and Generalization.

I. Introduction

II. Folkore Theorems of the Radom-Nikodym derivative

This section introduces relevant notational conventions alongside the Radon-Nikodym theorem. In particular, some equalities presented in this paper are valid almost surely with respect to a given measure. For clarity, given a measure space $(\Omega, \mathscr{F}, P)$, the notation $\overset{a.s.}{\underset{P}{=}}$ is introduced and shall be read as "equal for all $x \in \Omega$ except on a negligible set with respect to P"; or equivalently as "equal almost surely with respect to P". Moreover, given two measures P and Q on the same measurable space, the notation $P \ll Q$ stands for "the measure P is absolutely continuous with respect to Q". Using this notation, the Radon-Nikodym derivative is introduced by the following theorem.

Theorem 1 (Radon-Nikodym theorem, [?, Theorem 2.2.1]): Let P and Q be two measures on a given measurable space $(\Omega, \mathscr{F})$, such that Q is σ-finite and $P \ll Q$. Then, there exists a nonnegative Borel measurable function $g : \Omega \to \mathbb{R}$ such that for all $\mathcal{A} \in \mathscr{F}$,

$$P(\mathcal{A}) = \int_{\mathcal{A}} g(x) \mathrm{d}Q(x). \tag{1}$$

Moreover, if another function h satisfies for all $\mathcal{A} \in \mathscr{F}$ that $P(\mathcal{A}) = \int_{\mathcal{A}} h(x) \mathrm{d}Q(x)$, then $g(x) \overset{a.s.}{\underset{Q}{=}} h(x)$.

The function g in (1) is often referred to as the Radom-Nikodym derivative of P with respect to Q; and is also written as $\frac{\mathrm{d}P}{\mathrm{d}Q}$, such that $g(x) = \frac{\mathrm{d}P}{\mathrm{d}Q}(x)$.

The Radon-Nikodym theorem is the foundational tool from which many folklore theorems in information theory originate. Some of these folklore theorems are thoroughly studied in the following sections.

III. Basic Folklore Theorems

This section focuses on basic folklore theorems, where "basic" denotes their well-established nature. One of the most common folklore theorems is often referred to as the "change of measure" theorem.

Theorem 2 (Change of Measure): Let P and Q be two measures on the measurable space $(\Omega, \mathscr{F})$ with $P \ll Q$; and Q a σ-finite measure. Let $f : \Omega \to \mathbb{R}$ be a Borel measurable function such that the integral $\int_{\Omega} f(x) \mathrm{d}P(x)$ exists. Then, for all $\mathcal{A} \in \mathscr{F}$,

$$\int_{\mathcal{A}} f(x) \mathrm{d}P(x) = \int_{\mathcal{A}} f(x) \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \mathrm{d}Q(x). \tag{2}$$

Proof: The first part of the proof is developed under the assumption that the function f is simple. That is, for all $x \in \mathcal{X}$, $f(x) = \sum_{i=1}^{m} a_i \mathbb{1}_{\mathcal{A}_i}(x)$, for finite $m \in \mathbb{N}$, disjoint sets $\mathcal{A}_1$, $\mathcal{A}_2$, …, $\mathcal{A}_m$ in $\mathscr{F}$ and reals a_1, a_2, …, a_m. For all $\mathcal{A} \in \mathscr{F}$, and for all $i \in \{1, 2, \ldots, m\}$, let $\mathcal{B}_i = \mathcal{A} \cap \mathcal{A}_i$, hence,

$$\begin{aligned}
\int_{\mathcal{A}} f(x) \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \mathrm{d}Q(x) &= \int_{\mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \sum_{i=1}^{n} a_i \mathbb{1}_{\mathcal{A}_i}(x) \mathrm{d}Q(x) && (3)\\
&= \sum_{i=1}^{n} a_i \int_{\mathcal{B}_i} \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \mathrm{d}Q(x) && (4)\\
&= \sum_{i=1}^{n} a_i P(\mathcal{B}_i), && (5)
\end{aligned}$$

where the equality in (4) follows from the linearity of the integral [**?**, Theorem 1.6.3]; and the equality in (5) follows from Theorem 1. On the other hand, for all $\mathcal{A} \in \mathscr{F}$

$$\begin{aligned}\int_{\mathcal{A}} f(x)\mathrm{d}P(x) &= \int_{\mathcal{A}} \sum_{i=1}^{n} a_i \mathbb{1}_{\mathcal{A}_i(x)} \mathrm{d}P(x) && (6)\\ &= \sum_{i=1}^{n} \int_{\mathcal{A}} a_i \mathbb{1}_{\mathcal{A}_i(x)} \mathrm{d}P(x) && (7)\\ &= \sum_{i=1}^{n} a_i \int_{\mathcal{B}_i} \mathrm{d}P(x) = \sum_{i=1}^{n} a_i P(\mathcal{B}_i), && (8)\end{aligned}$$

where the equality in (7) follows from the linearity of the integral [**?**, Theorem 1.6.3]. Hence, from Theorem 1, and equalities (5) and (8), it follows that when f is a simple function, the equality in (2) holds. This concludes the first part of the proof.

The second part of the proof proceeds by considering the following observations: (a) simple functions form a dense subset of the space of Borel measurable functions [**?**, Theorem 1.5.5(b)]; and (b) the integral is a continuous map from that space [**?**, Theorem 1.6.2]. Hence, from (a) and (b), it follows that (2) also holds for any Borel measurable function f. This completes the proof. ■

Another reputed folklore theorem, which is often referred to as the "proportional measures" theorem, establishes the explicit forms of the Radon-Nikodym derivatives between two measures, in which one is proportional to the other.

Theorem 3 (Proportional Measures): Let P and Q be two σ- finite measures on the measurable space $(\Omega, \mathscr{F})$, such that for all $\mathcal{A} \in \mathscr{F}$

$$Q(\mathcal{A}) = cP(\mathcal{A}), \tag{9}$$

with $c > 0$. Then, for all $x \in \Omega$

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x) \overset{a.s.}{\underset{Q}{=}} \frac{1}{c}, \text{ and } \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \overset{a.s.}{\underset{P}{=}} c. \tag{10}$$

Proof: First, note that (9) implies that the measures P and Q are mutually absolutely continuous. Hence, from Theorem 1, it follows that for all $\mathcal{A} \in \mathscr{F}$,

$$P(\mathcal{A}) = \int_{\mathcal{A}} \mathrm{d}P(x) = \int_{\mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\, \mathrm{d}Q(x). \tag{11}$$

On the other hand, the equality (9) also implies

$$P(\mathcal{A}) = \frac{1}{c} Q(\mathcal{A}) = \int_{\mathcal{A}} \frac{1}{c} \mathrm{d}Q(x). \tag{12}$$

Hence, it follows directly from Theorem 1 that the Radon-Nikodym derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is unique almost surely with respect to Q. Thus, $\frac{\mathrm{d}P}{\mathrm{d}Q}(x) \overset{a.s.}{\underset{Q}{=}} \frac{1}{c}$. Using similar arguments and the fact that P and Q are mutually absolutely continuous, it is verified that $\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \overset{a.s.}{\underset{P}{=}} c$. ■

In the case in which $c = 1$ in (10), measures P and Q are identical, thus, $\frac{\mathrm{d}P}{\mathrm{d}Q}(x) \overset{a.s.}{\underset{Q}{=}} \frac{\mathrm{d}Q}{\mathrm{d}P}(x) \overset{a.s.}{\underset{P}{=}} 1$.

The following folklore theorem is often referred to as the "chain rule".

Theorem 4 (Chain Rule): Let P, Q, and R be three measures on the measurable space $(\Omega, \mathscr{F})$ such that $P \ll Q$; $Q \ll R$; and Q and R are σ-finite measures. Then,

$$\frac{\mathrm{d}P}{\mathrm{d}R}(x) \overset{a.s.}{\underset{R}{=}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \frac{\mathrm{d}Q}{\mathrm{d}R}(x). \tag{13}$$

Proof: From the assumptions of the theorem, it follows that for all $\mathcal{A} \in \mathscr{F}$,

$$\begin{aligned}P(\mathcal{A}) &= \int_{\mathcal{A}} \mathrm{d}P(x) = \int_{\mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \mathrm{d}Q(x) && (14)\\ &= \int_{\mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \frac{\mathrm{d}Q}{\mathrm{d}R}(x) \mathrm{d}R(x) && (15)\\ &= \int_{\mathcal{A}} \frac{\mathrm{d}P}{\mathrm{d}R}(x) \mathrm{d}R(x), && (16)\end{aligned}$$

where the second equality in (14) follows from Theorem 1; and the equality in (15) follows from Theorem 2 . The equality in (16) holds from Theorem 1 and by noticing that $P \ll R$. Therefore, the equalities in (15) and (16) together with Theorem 1 imply (13), which completes the proof. ■

The following folklore theorem shows the connection between the Radon-Nikodym derivative and its multiplicative inverse.

Theorem 5 (Multiplicative Inverse): Let P and Q be two mutually absolutely continuous measures on the measurable space $(\Omega, \mathscr{F})$; and assume that for all $x \in \Omega$, $\frac{\mathrm{d}Q}{\mathrm{d}P}(x) > 0$. Then,

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x) \quad \overset{a.s.}{\underset{Q}{=}} \quad \left(\frac{\mathrm{d}Q}{\mathrm{d}P}(x)\right)^{-1}. \tag{17}$$

Proof: From Theorem 4, it follows that

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \overset{a.s.}{\underset{Q}{=}} \frac{\mathrm{d}Q}{\mathrm{d}Q}(x) \overset{a.s.}{\underset{Q}{=}} 1, \tag{18}$$

where the last equality follows from Theorem 3, with $c = 1$. This completes the proof. ■

The subsequent folklore theorem establishes the linearity of the Radon-Nikodym derivative.

Theorem 6 (Linearity): Let P be a σ-finite measure on $(\Omega, \mathscr{F})$ and let also $Q_1, Q_2, \ldots, Q_n$ be finite measures on $(\Omega, \mathscr{F})$ absolutely continuous with respect to P. Let $c_1, c_2, \ldots, c_n$ be positive reals; and let S be a finite measure on $(\Omega, \mathscr{F})$ such that for all $\mathcal{A} \in \mathscr{F}$, $S(\mathcal{A}) = \sum_{t=1}^{n} c_t Q_t(\mathcal{A})$. Then,

$$\frac{\mathrm{d}S}{\mathrm{d}P}(x) \quad \overset{a.s.}{\underset{P}{=}} \quad \sum_{t=1}^{n} c_t \frac{\mathrm{d}Q_t}{\mathrm{d}P}(x). \tag{19}$$

Proof: The proof starts by noticing that, from the assumptions of the theorem, it holds that $S \ll P$. Hence, for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$\int_{\mathcal{A}} \frac{\mathrm{d}S}{\mathrm{d}P}(x)\mathrm{d}P(x) = \int_{\mathcal{A}} \mathrm{d}S(x) = \sum_{t=1}^{n} c_t Q_t(\mathcal{A}) \tag{20}$$

$$= \sum_{t=1}^{n} \int_{\mathcal{A}} c_t \mathrm{d}Q_t(x) = \sum_{t=1}^{n} \int_{\mathcal{A}} c_t \frac{\mathrm{d}Q_t}{\mathrm{d}P}(x)\mathrm{d}P(x) \tag{21}$$

$$= \int_{\mathcal{A}} \sum_{t=1}^{n} c_t \frac{\mathrm{d}Q_t}{\mathrm{d}P}(x)\mathrm{d}P(x), \tag{22}$$

where the first equality in (20) and the last equality in (21) follow from Theorem 2; and the equality (22) follows from the additivity property of the integral [**?**, Corollary 1.6.4]. The proof ends by using Theorem 1, which implies the equality in (19) from (22). ■

The following folklore theorem establishes the continuity of the Radon-Nikodym derivative.

Theorem 7 (Continuity): Let P be a σ-finite measure on $(\Omega, \mathscr{F})$, and let $Q_1, Q_2, \cdots$ be an infinite sequence of σ-finite measures on $(\Omega, \mathscr{F})$, converging to a measure Q. Suppose that for all $n \in \mathbb{N}$, $Q_n \ll P$. Then, $Q \ll P$ and

$$\lim_{n\to\infty} \frac{\mathrm{d}Q_n}{\mathrm{d}P}(x) \overset{a.s.}{\underset{P}{=}} \frac{\mathrm{d}Q}{\mathrm{d}P}(x). \tag{23}$$

Proof: From the assumptions of the theorem, for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$Q(\mathcal{A}) = \lim_{n\to\infty} Q_n(\mathcal{A}) \tag{24}$$

$$= \lim_{n\to\infty} \int_{\mathcal{A}} \frac{\mathrm{d}Q_n}{\mathrm{d}P}(x)\mathrm{d}P(x) \tag{25}$$

$$= \int_{\mathcal{A}} \lim_{n\to\infty} \frac{\mathrm{d}Q_n}{\mathrm{d}P}(x)\mathrm{d}P(x), \tag{26}$$

where the equality in (25) follows from Theorem 2 ; and the equality in (26) follows from [**?**, Theorem 1.6.2]. The equality in (26) implies that $Q \ll P$. Hence, for all $\mathcal{A} \in \mathscr{F}$, it holds that

$$Q(\mathcal{A}) = \int_{\mathcal{A}} \frac{\mathrm{d}Q}{\mathrm{d}P}(x)\mathrm{d}P(x). \tag{27}$$

Therefore, the equalities in (26) and (27) jointly with Theorem 1 imply equation (23), which completes the proof. ■

The ensuing folklore theorem establishes the relation between the Radon-Nikodym derivative of a product measure with respect to its component measures.

Theorem 8 (Product of Measures): For all $i \in \{1, 2\}$, let P_i and Q_i be a finite and a σ-finite measure on $(\Omega_i, \mathscr{F}_i)$, respectively; with $P_i \ll Q_i$. Let also P_1P_2 and Q_1Q_2 be the product measures on $(\Omega_1 \times \Omega_2, \mathscr{F}_1 \times \mathscr{F}_2)$ formed by P_1 and P_2; and Q_1 and Q_2, respectively. Then,

$$\frac{\mathrm{d}P_1P_2}{\mathrm{d}Q_1Q_2}(x_1, x_2) \overset{a.s.}{=\!=}_{Q_1Q_2} \frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x_1)\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x_2). \tag{28}$$

Proof: From the assumptions of the theorem, for all $\mathcal{A} \in (\Omega_1 \times \Omega_2)$,

$$\begin{aligned} P_1P_2(\mathcal{A}) &= \int_{\mathcal{A}} \mathrm{d}P_1P_2(x_1, x_2) && (29)\\ &= \int\int_{\mathcal{A}_{x_2}} \mathrm{d}P_1(x_1)\mathrm{d}P_2(x_2) && (30)\\ &= \int\int_{\mathcal{A}_{x_2}} \frac{\mathrm{d}P_1(x_1)}{\mathrm{d}Q_1}\mathrm{d}Q_1(x_1)\mathrm{d}P_2(x_2) && (31)\\ &= \int\int_{\mathcal{A}_{x_2}} \frac{\mathrm{d}P_1(x_1)}{\mathrm{d}Q_1}\frac{\mathrm{d}P_2(x_2)}{\mathrm{d}Q_2}\mathrm{d}Q_1(x_1)\mathrm{d}Q_2(x_2) && (32)\\ &= \int_{\mathcal{A}} \frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(x_1)\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(x_2)\mathrm{d}Q_1Q_2(x_1, x_2), && (33) \end{aligned}$$

where $\mathcal{A}_{x_2}$ is the section of the set $\mathcal{A}$ determined by x_2, namely, $\mathcal{A}_{x_2} \triangleq \{x_1 \in \Omega_1 : (x_1, x_2) \in \mathcal{A}\}$; the equality in (29) arises from the definition of P_1P_2 as the product of P_1 and P_2; the equality in (31) is a direct consequence of Theorem 2; the equality in (32) follows from Theorem 1; and finally, the equality in (33) is due to the construction of Q_1Q_2 as the product measure of Q_1 and Q_2.

The proof follows by observing that from the equality in (33), it holds that $P_1P_2 \ll Q_1Q_2$. Thus, for all $\mathcal{A} \in \mathscr{F}_1 \times \mathscr{F}_2$,

$$P_1P_2(\mathcal{A}) = \int_{\mathcal{A}} \frac{\mathrm{d}P_1P_2}{\mathrm{d}Q_1Q_2}(x_1, x_2)\mathrm{d}Q_1Q_2(x_1, x_2). \tag{34}$$

The equalities in (33) and (34), together with Theorem 1, imply the equality in (28), which completes the proof. ■

IV. Advanced folklore theorems

This section requires some additional notation. In particular, denote by $\triangle(\mathcal{X}, \mathscr{F}_{\mathcal{X}})$, or simply $\triangle(\mathcal{X})$, the set of all probability measures on the measurable space $(\mathcal{X}, \mathscr{F}_{\mathcal{X}})$, where $\mathscr{F}_{\mathcal{X}}$ is a σ-algebra on $\mathcal{X}$. Using this notation, conditional probability measures can be defined as follows.

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

$$\begin{aligned} \mathcal{X} &\to [0, 1] && (35)\\ x &\mapsto P_{Y|X=x}(\mathcal{A}) && (36) \end{aligned}$$

is Borel measurable. The set of such conditional probability measures is denoted by $\triangle(\mathcal{Y}|\mathcal{X})$.

A conditional probability $P_{Y|X} \in \triangle(\mathcal{Y}|\mathcal{X})$ and a probability measure $P_X \in \triangle(\mathcal{X})$ determine two unique probability measures in $\triangle(\mathcal{X} \times \mathcal{Y})$ and $\triangle(\mathcal{Y} \times \mathcal{X})$, respectively. These probability measures are denoted by P_{XY} and P_{YX}, respectively, and for all sets $\mathcal{A} \in \mathscr{F}_{\mathcal{X}} \times \mathscr{F}_{\mathcal{Y}}$, it follows that

$$P_{XY}(\mathcal{A}) = \int P_{Y|X=x}(\mathcal{A}_x)\mathrm{d}P_X(x), \tag{37}$$

where $\mathcal{A}_x$ is the section of the set $\mathcal{A}$ determined by x, namely,

$$\mathcal{A}_x \triangleq \{y \in \mathcal{Y} : (x, y) \in \mathcal{A}\}. \tag{38}$$

Alternatively, for all sets $\mathcal{B} \in \mathscr{F}_{\mathcal{Y}} \times \mathscr{F}_{\mathcal{X}}$, it follows that

$$P_{YX}(\mathcal{B}) = \int P_{Y|X=x}(\mathcal{B}_x)\mathrm{d}P_X(x), \tag{39}$$

where $\mathcal{B}_x$ is the section of the set $\mathcal{B}$ determined by x. For all sets $\mathcal{A} \in \mathscr{F}_{\mathcal{X}} \times \mathscr{F}_{\mathcal{Y}}$, let the set $\hat{\mathcal{A}} \in \mathscr{F}_{\mathcal{Y}} \times \mathscr{F}_{\mathcal{X}}$ be such that

$$\hat{\mathcal{A}} \quad = \quad \{(y,x) \in \mathcal{Y} \times \mathcal{X} : (x,y) \in \mathcal{A}\}. \tag{40}$$

Then, from (37) and (39), it holds that

$$P_{XY}(\mathcal{A}) \quad = \quad P_{YX}\big(\hat{\mathcal{A}}\big). \tag{41}$$

Using this notation, the notion of marginal probability measures can be introduced as follows.

Definition 2 (Marginals): Given two joint probability measures $P_{XY} \in \triangle(\mathcal{X} \times \mathcal{Y})$ and $P_{YX} \in \triangle(\mathcal{Y} \times \mathcal{X})$, satisfying (41), the marginal probability measures in $\triangle(\mathcal{X})$ and $\triangle(\mathcal{Y})$, denoted by P_X and P_Y, respectively satisfy for all sets $\mathcal{A} \in \mathscr{F}_{\mathcal{X}}$ and for all sets $\mathcal{B} \in \mathscr{F}_{\mathcal{Y}}$,

$$P_X(\mathcal{A}) \quad \triangleq P_{XY}(\mathcal{A} \times \mathcal{Y}) = \quad P_{YX}(\mathcal{Y} \times \mathcal{A}); \text{ and} \tag{42}$$

$$P_Y(\mathcal{B}) \quad \triangleq P_{XY}(\mathcal{X} \times \mathcal{B}) = \quad P_{YX}(\mathcal{B} \times \mathcal{X}). \tag{43}$$

From the total probability theorem [**?**, Theorem 4.5.2], it follows that for all $\mathcal{A} \in \mathscr{F}_{\mathcal{Y}}$,

$$P_Y(\mathcal{A}) = \int \int_{\mathcal{A}} \mathrm{d}P_{Y|X=x}(y) \mathrm{d}P_X(x); \tag{44}$$

and for all $\mathcal{B} \in \mathscr{F}_{\mathcal{X}}$,

$$P_X(\mathcal{B}) = \int \int_{\mathcal{B}} \mathrm{d}P_{X|Y=y}(x) \mathrm{d}P_Y(y). \tag{45}$$

The joint probability measures P_{XY} and P_{YX} can be described via the conditional probability measure $P_{Y|X}$ and the probability measure P_X as in (37) and in (39); or via the conditional probability measure $P_{X|Y} \in \triangle(\mathcal{X}|\mathcal{Y})$ and the marginal probability measure $P_Y \in \triangle(\mathcal{Y})$. More specifically, for all sets $\mathcal{A} \in \mathscr{F}_{\mathcal{X}} \times \mathscr{F}_{\mathcal{Y}}$, it follows that

$$P_{XY}(\mathcal{A}) \quad = \quad \int P_{X|Y=y}(\mathcal{A}_y) \mathrm{d}P_Y(y), \tag{46}$$

where $\mathcal{A}_y$ is the section of the set $\mathcal{A}$ determined by y, namely,

$$\mathcal{A}_y \triangleq \{x \in \mathcal{X} : (x,y) \in \mathcal{A}\}. \tag{47}$$

Alternatively, for all sets $\mathcal{B} \in \mathscr{F}_{\mathcal{Y}} \times \mathscr{F}_{\mathcal{X}}$, it follows that

$$P_{YX}(\mathcal{B}) \quad = \quad \int P_{X|Y=y}(\mathcal{B}_y) \mathrm{d}P_Y(y), \tag{48}$$

where $\mathcal{B}_y$ is the section set of $\mathcal{B}$ determined by y.

Within this context, the following folklore theorem highlights a property of conditional measures, which is reminiscent of the unit measure axiom in probability theory.

Theorem 9 (Unit Measure): Consider the conditional probability measures $P_{Y|X} \in \triangle(\mathcal{Y}|\mathcal{X})$ and $P_{X|Y} \in \triangle(\mathcal{X}|\mathcal{Y})$; the probability measures $P_Y \in \triangle(\mathcal{Y})$ and $P_X \in \triangle(\mathcal{X})$ that satisfy (44) and (45). Assume that for all $x \in \mathcal{X}$, the probability measure $P_{Y|X=x} \ll P_Y$. Then,

$$\int \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_X(x) \quad \overset{\underset{P_Y}{a.s.}}{=} \quad 1. \tag{49}$$

Proof: For all $\mathcal{A} \in \mathscr{F}_{\mathcal{Y}}$, from (44), it holds that

$$P_Y(\mathcal{A}) \quad = \quad \int \int_{\mathcal{A}} \mathrm{d}P_{Y|X=x}(y) \mathrm{d}P_X(x) \tag{50}$$

$$\quad = \quad \int \int_{\mathcal{A}} \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_Y(y) \mathrm{d}P_X(x) \tag{51}$$

$$\quad = \quad \int_{\mathcal{A}} \int \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_X(x) \mathrm{d}P_Y(y), \tag{52}$$

where the equality in (51) follows from a change of measure (Theorem 2). Moreover, (52) is obtained using Fubini's theorem [**?**, Theorem 2.6.6]. The proof proceeds by noticing that $P_Y(\mathcal{A}) = \int_{\mathcal{A}} \mathrm{d}P_Y(y)$, and thus from Theorem 1 and the equality in (52), the statement in (49) holds. ■

The following folklore theorem is reminiscent of the Bayes rule.

Theorem 10 (Bayes-like rule): Consider the conditional probability measures $P_{Y|X}$ and $P_{X|Y}$; the probability measures P_Y and P_X that satisfy (44) and (45); and the joint probability measures P_{YX} and P_{XY} in (39) and (46) respectively. Let also $P_X P_Y \in \triangle(\mathcal{X} \times \mathcal{Y}, \mathscr{F}_{\mathcal{X}} \times \mathscr{F}_{\mathcal{Y}})$ and $P_Y P_X \in \triangle(\mathcal{Y} \times \mathcal{X}, \mathscr{F}_{\mathcal{Y}} \times \mathscr{F}_{\mathcal{X}})$ be the measures formed by the product of the marginals P_X and P_Y. Assume that:

(a) For all $x \in \mathcal{X}$, $P_{Y|X=x} \ll P_Y$; and
(b) For all $y \in \mathcal{Y}$, $P_{X|Y=y} \ll P_X$.

Then,

$$\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}(x,y) \overset{a.s.}{\underset{P_X P_Y}{=}} \frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x) \tag{53}$$

$$\overset{a.s.}{\underset{P_X P_Y}{=}} \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \tag{54}$$

$$\overset{a.s.}{\underset{P_X P_Y}{=}} \frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(y,x). \tag{55}$$

Proof: Note that assumptions (a) and (b) are sufficient for the Radon-Nikodym derivatives of P_{XY} with respect to $P_X P_Y$ and P_{YX} with respect to $P_Y P_X$ to exist. Hence, it follows that for all sets $\mathcal{A} \in \mathscr{F}_{\mathcal{X}} \times \mathscr{F}_{\mathcal{Y}}$,

$$P_{XY}(\mathcal{A}) = \int_{\mathcal{A}} \frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}(x,y) \mathrm{d}P_X P_Y(x,y), \tag{56}$$

which follows from Theorem 2. Note also that from (46), it follows that

$$P_{XY}(\mathcal{A}) = \int \int_{\mathcal{A}_y} \mathrm{d}P_{X|Y=y}(x) \mathrm{d}P_Y(y) \tag{57}$$

$$= \int \int_{\mathcal{A}_y} \frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x) \mathrm{d}P_X(x) \mathrm{d}P_Y(y) \tag{58}$$

$$= \int \int \mathbb{1}_{\mathcal{A}_y}(x) \frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x) \mathrm{d}P_X(x) \mathrm{d}P_Y(y) \tag{59}$$

$$= \int \mathbb{1}_{\mathcal{A}}(x,y) \frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x) \mathrm{d}P_X P_Y(x,y) \tag{60}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x) \mathrm{d}P_X P_Y(x,y), \tag{61}$$

where, the set $\mathcal{A}_y$ is defined in (47). Moreover, the equality in (58) follows from Assumption (b) and Theorem 1. Similarly, from (37), it follows that

$$P_{XY}(\mathcal{A}) = \int \int_{\mathcal{A}_x} \mathrm{d}P_{Y|X=x}(y) \mathrm{d}P_X(x) \tag{62}$$

$$= \int \int_{\mathcal{A}_x} \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_Y(y) \mathrm{d}P_X(x) \tag{63}$$

$$= \int \int \mathbb{1}_{\mathcal{A}_x}(y) \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_Y(y) \mathrm{d}P_X(x) \tag{64}$$

$$= \int \int \mathbb{1}_{\mathcal{A}_y}(x) \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_X(x) \mathrm{d}P_Y(y) \tag{65}$$

$$= \int \mathbb{1}_{\mathcal{A}}(x,y) \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_X P_Y(x,y) \tag{66}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_X P_Y(x,y), \tag{67}$$

where, the set $\mathcal{A}_x$ is defined in (38). Moreover, the equality in (63) follows from Assumption (a) and Theorem 1; and the equality in (65) follows by exchanging the order of integration [**?**, Theorem 2.6.6]. Finally, from (41), it follows that

$$P_{XY}(\mathcal{A}) = \int_{\hat{\mathcal{A}}} \mathrm{d}P_{YX}(y,x)$$

$$= \int_{\hat{\mathcal{A}}} \frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(y,x) \mathrm{d}P_Y P_X(y,x) \tag{68}$$

$$= \int \mathbb{1}_{\hat{\mathcal{A}}}(y,x)\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(y,x)\mathrm{d}P_Y P_X(y,x) \tag{69}$$

$$= \int\int \mathbb{1}_{\hat{\mathcal{A}}_x}(y)\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(y,x)\mathrm{d}P_Y(y)\mathrm{d}P_X(x) \tag{70}$$

$$= \int\int \mathbb{1}_{\hat{\mathcal{A}}_y}(x)\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(y,x)\mathrm{d}P_X(x)\mathrm{d}P_Y(y) \tag{71}$$

$$= \int \mathbb{1}_{\hat{\mathcal{A}}}(x,y)\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(y,x)\mathrm{d}P_X P_Y(x,y) \tag{72}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(y,x)\mathrm{d}P_X P_Y(x,y), \tag{73}$$

where the set $\hat{\mathcal{A}}$ is defined in (40). By performing a change of measure using Theorem 2 and assumption (a), the equality in (68) is obtained; and the equality in (71) follows by exchanging the order of integration [**?**, Theorem 2.6.6].

The proof is completed from Theorem 1 and by combining equations (56), (61), (67) and (73), which establish (53), (54) and (55). ■

Example 1:

From definition **??**, we want to show that there exist an $x \in \mathcal{X}$ such that $P_{Y|X=x}(A) > 0$ and $P_X(\{x \in \mathcal{X} : P_{Y|X=x}(A) > 0\}) = 0$

For all $x \in \mathbb{R}$, for $\mathcal{A} \in \mathscr{F}_Y$, for $\mathcal{B} \in \mathscr{F}_Y$ and μ the Lebesgue measure, $P_{Y|X=x}(A)$ and $P_X(B)$ are defined as follows

$$P_{Y|X=x}(A) = \mu(A \cap (-x,x)) \tag{74}$$

$$P_X(B) = \mu(B) \tag{75}$$

Given $A \in \mathcal{B}(\mathbb{R})$ such that $P_Y(A) = 0$, then

$$P_Y(A) = \int P_{Y|X=x}(A)\mathrm{d}P_X(x) \tag{76}$$

$$= \int \mu(A \cap (-x,x))\mathbb{1}_{\{x\in\mathbb{N}\}}\mathrm{d}P_X(x) \tag{77}$$

$$= \int_{\mathbb{N}} \mu(A \cap (-x,x))\mathrm{d}P_X(x) \tag{78}$$

$$= \sum_{t=0}^{\infty} \mu(A \cap (-t,t))\mu(\{t\}) \tag{79}$$

$$= 0 \tag{80}$$

The equality in (77) follows from (44); the equality (80) follows from the definition of the Lebesgue measure of a singleton. Thus, this exemple constructs $P_{Y|X}$ and P_X such that for some $\mathcal{A} \in \mathcal{B}(\mathbb{R})$ $P_Y(A) = 0$ but $P_{Y|X=x}(A) > 0$ for some $x \in \mathcal{X}$, therefore the Radom-Nikodym derivative of $P_Y|X = x$ with respect to P_Y does not exist. This proves that assumption (a) in 10 cannot be weakened and is needed for all $x \in \mathcal{X}$. A example for assumption (b) in 10 can be constructed similarly.

Theorem 11 (Inverse Bayes-like Rule): Consider the conditional probability measures $P_{Y|X}$ and $P_{X|Y}$; and the probability measures P_Y and P_X that satisfy (44) and (45); and the joint probability measures P_{YX} and P_{XY} in (39) and (46) respectively. Assume that:

(a) For all $x \in \mathcal{X}$, $P_Y \ll P_{Y|X=x}$; and
(b) For all $y \in \mathcal{Y}$, $P_X \ll P_{X|Y=y}$.

Then,

$$\frac{\mathrm{d}P_X P_Y}{\mathrm{d}P_{XY}}(x,y) \overset{a.s.}{\underset{P_{XY}}{=}} \frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x) \tag{81}$$

$$\overset{a.s.}{\underset{P_{XY}}{=}} \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y) \tag{82}$$

$$\overset{a.s.}{\underset{P_{YX}}{=}} \frac{\mathrm{d}P_Y P_X}{\mathrm{d}P_{YX}}(y,x). \tag{83}$$

Proof: The proof follows along the same lines as the proof of Theorem 10. The proof follows by observing that for all measurable sets $\mathcal{A} \in \mathcal{X} \times \mathcal{Y}$, the product measure $P_X P_Y \in \triangle(\mathcal{X} \times \mathcal{Y})$ satisfies

$$\begin{aligned} & P_X P_Y(\mathcal{A}) \\ = & \int_{\mathcal{A}} \mathrm{d}P_X P_Y(x,y) && (84) \\ = & \int \int_{\mathcal{A}_y} \mathrm{d}P_X(x)\mathrm{d}P_Y(y) && (85) \\ = & \int \int_{\mathcal{A}_y} \frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x)\mathrm{d}P_{X|Y=y}(x)\mathrm{d}P_Y(y) && (86) \\ = & \int \int \frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x)\mathbb{1}_{\{x\in\mathcal{A}_y\}}\mathrm{d}P_{X|Y=y}(x)\mathrm{d}P_Y(y) && (87) \\ = & \int \frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x)\mathbb{1}_{\{x\in\mathcal{A}_y\}}\mathrm{d}P_{XY}(x,y) && (88) \\ = & \int_{\mathcal{A}} \frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x)\mathrm{d}P_{XY}(x,y), && (89) \end{aligned}$$

where the set $\mathcal{A}_y$ is defined in (47); the equality in (**??**) follows from Assumption (b) and [**?**, Theorem 2.2.3]; and the measure P_{XY} is defined in (46).

The proof proceeds by noticing that for all measurable sets $\mathcal{A} \in \mathcal{X} \times \mathcal{Y}$, the product measure $P_X P_Y \in \triangle(\mathcal{X} \times \mathcal{Y})$ also satisfies

$$\begin{aligned} & P_X P_Y(\mathcal{A}) \\ = & \int_{\mathcal{A}} \mathrm{d}P_X P_Y(x,y) && (90) \\ = & \int \int_{\mathcal{A}_y} \mathrm{d}P_X(x)\mathrm{d}P_Y(y) && (91) \\ = & \int \int_{\mathcal{A}_x} \mathrm{d}P_Y(y)\mathrm{d}P_X(x) && (92) \\ = & \int \int_{\mathcal{A}_x} \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y)\mathrm{d}P_{Y|X=x}(y)\mathrm{d}P_X(x) && (93) \\ = & \int \int \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y)\mathbb{1}_{\{y\in\mathcal{A}_x\}}\mathrm{d}P_{Y|X=x}(y)\mathrm{d}P_X(x) && (94) \\ = & \int \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y)\mathbb{1}_{\{y\in\mathcal{A}_x\}}\mathrm{d}P_{YX}(y,x) && (95) \\ = & \int_{\hat{\mathcal{A}}} \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y)\mathrm{d}P_{YX}(y,x) && (96) \\ = & \int_{\mathcal{A}} \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y)\mathrm{d}P_{XY}(x,y), && (97) \end{aligned}$$

where the sets $\mathcal{A}_x$, $\mathcal{A}_y$, and $\hat{\mathcal{A}}$ are defined in (38), (47), and (40); and the measure P_{YX} is defined in (39). The equality in (**??**) follows by exchanging the order of the integrals [**?**, Theorem 2.6.6]; the equality in (**??**) follows from Assumption (a) and [**?**, Theorem 2.2.3].

The proof is completed by noticing that from (**??**) and (**??**), the following equalities hold:

$$\begin{aligned} P_{X,Y}(\mathcal{A}) &= \int_{\mathcal{A}} \frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x)\mathrm{d}P_{XY}(x,y), && (98) \\ &= \int_{\mathcal{A}} \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y)\mathrm{d}P_{XY}(x,y), && (99) \end{aligned}$$

which together with [**?**, Theorem 2.2.3] implies the equality in (**??**) almost surely with respect to the measure $P_{XY} \in \triangle(\mathcal{X} \times \mathcal{Y})$ in (37). This completes the proof.

■

Remark 1: An alternative proof for Theorem 9 can be written as follows, by combining Theorems 10 11 and 5.

$$\int \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y)\mathrm{d}P_X(x) \tag{100}$$

$$= \int \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y)\frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x)\mathrm{d}P_{X|Y=y}(x) \tag{101}$$

$$= \int \mathrm{d}P_{X|Y=y}(x) = 1 \tag{102}$$

Example 2:

This example shows that there exist a $z \in \mathcal{X}$ such that $P_{Y|X=z}(A) = 0$ and $P_Y(A) > 0$.

For all $x \in \mathbb{R}$, for $\mathcal{A} \in \mathscr{F}_Y$, for $\mathcal{B} \in \mathscr{F}_Y$ and μ the Lebesgue measure, $P_{Y|X=x}(A)$ and $P_X(B)$ are defined as follows

$$P_{Y|X=x}(A) = \mathbb{1}_{\{A\cap\{(x,y)\in\mathbb{R}^2 : x=y\}\neq\emptyset\}} \tag{103}$$

$$P_X(B) = \mu(B) \tag{104}$$

Given $A \in \mathcal{B}(\mathbb{R})$ such that $P_Y(A) = 0$, then

$$P_Y(A) = \int P_{Y|X=x}(A)\mathrm{d}P_X(x) \tag{105}$$

$$= \int \mathbb{1}_{\{A\cup\{(x,y)\in\mathbb{R}^2 : x=y\}\neq\emptyset\}}\mathrm{d}P_X(x) \tag{106}$$

$$= \int \mathbb{1}_{\{A\cup\{(x,x)\}\neq\emptyset\}}\mathrm{d}P_X(x) \tag{107}$$

$$= \int \mathbb{1}_{\{x\in A\}}\mathrm{d}P_X(x) \tag{108}$$

$$= \int_A \mathrm{d}P_X(x) \tag{109}$$

$$= \mu(A) \tag{110}$$

Let us define the infimum as follows,

$$\underline{x} = inf_y\{x \in \mathbb{R} : (x, y) \in A\} \tag{111}$$

Then for all $z \in \mathbb{R}$ such that $z = \underline{x} - \epsilon$ for $\epsilon > 0$,

The equality in (77) follows from (44); the equality (80) follows from the definition of the Lebesgue measure of a singleton. Thus, this exemple constructs $P_{Y|X}$ and P_X such that for some $\mathcal{A} \in \mathcal{B}(\mathbb{R})$ but $P_{Y|X=x}(A) = 0$ but $P_Y(A) > 0$ for some $x \in \mathcal{X}$, therefore the Radom-Nikodym derivative of P_Y with respect to $P_Y|X = x$ does not exist. This proves that assumption (a) in 11 cannot be weakened and is needed for all $x \in \mathcal{X}$. A example for assumption (b) in 11 can be constructed similarly.

Example 3: Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ be a probability space, with P_X absolutely continuous with respect to the Lebesgue measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. For all $x \in \mathbb{R}$ and for all $\mathcal{A} \in \mathcal{B}(\mathbb{R})$, let

$$P_{Y|X} \triangleq \{P_{Y|X=x} \in \triangle(\mathbb{R}, \mathcal{B}(\mathbb{R})) : x \in \mathbb{R}\} \tag{112}$$

such that:

$$P_{Y|X=x}(\mathcal{A}) = \mathbb{1}_{\{x\in\mathcal{A}\}} \tag{113}$$

Hence, there exists a unique measure on $\left(\mathbb{R}^2, \mathcal{B}(\mathbb{R})\right)$ denoted by P_{XY} such that:

$$P_{XY}(\mathcal{A}) = \int_{\mathbb{R}}\int_{\mathcal{A}_y} \mathrm{d}P_{Y|X=x}(y)\mathrm{d}P_X(x) \tag{114}$$

$$= \int_{\mathbb{R}} P_{Y|X=x}(\mathcal{A}_x)\mathrm{d}P_X(x) \tag{115}$$

$$= \int_{\mathbb{R}} \mathbb{1}_{\{(x,x)\in(A)\}}\mathrm{d}P_X(x) \tag{116}$$

$$= P_X(\{t \in \mathbb{R} : (t, t) \in \mathcal{A}\}). \tag{117}$$

Assuming that $\mathcal{A} = \{(x, y) \in \mathbb{R}^2 : x = y\}$, it follows that

$$P_{XY} = P_X(\{t \in \mathbb{R} : (t, t) \in \mathcal{A}\}) \tag{118}$$
$$= P_X(\mathbb{R}) \tag{119}$$
$$= 1. \tag{120}$$

Moreover, let the measure P_Y on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be such that for all $\mathcal{A} \in \mathcal{B}(\mathbb{R})$,

$$P_Y(\mathcal{A}) = P_{XY}(\mathbb{R} \times A) \tag{121}$$
$$= P_X(\{t \in \mathbb{R} : (t, t) \in \mathcal{A}\}) = P_X(\mathcal{A}), \tag{122}$$

hence P_Y is absolutely continuous with respect to μ. Hence,

$$P_X P_Y(\{(x, y) \in \mathbb{R} : x = y\}) \tag{123}$$
$$= \int_{\{(x,y)\in\mathbb{R}^2 : x=y\}} \mathrm{d}P_X P_Y(x, y) \tag{124}$$
$$= \int_{\{(x,y)\in\mathbb{R}^2 : x=y\}} \frac{\mathrm{d}P_X P_Y}{\mathrm{d}\mu\mu}(x, y)\mathrm{d}\mu\mu(x, y) \tag{125}$$
$$= \int_{\{(x,y)\in\mathbb{R}^2 : x=y\}} \frac{\mathrm{d}P_X}{\mathrm{d}\mu}(x)\frac{\mathrm{d}P_Y}{\mathrm{d}\mu}(y)\mathrm{d}\mu\mu(x, y) \tag{126}$$
$$= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_{\{\{x=y\}\}} \frac{\mathrm{d}P_X}{\mathrm{d}\mu}(x)\frac{\mathrm{d}P_Y}{\mathrm{d}\mu}(y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) \tag{127}$$
$$= \int_{\mathbb{R}}\int_{\{y\}} \frac{\mathrm{d}P_X}{\mathrm{d}\mu}(x)\frac{\mathrm{d}P_Y}{\mathrm{d}\mu}(y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) \tag{128}$$
$$= 0 \tag{129}$$

where the equality in (**??**) follows from the fact that for all $y \in \mathbb{R}, \mu(\{y\}) = 0$. Thus P_{XY} is not absolutely continuous with respect to $P_X P_Y$.

V. Information measures

A. KL

The entropy of a given random variable is defined as follows.

Definition 3 (Entropy): Let X be a discrete random variable and assume it induces the probability measure $P_X \in \triangle(\mathcal{X}, \mathcal{F}_\mathcal{X})$ and a sigma finite probability measure $Q \in \triangle(\mathcal{X}, \mathcal{F}_\mathcal{X})$, with $P_X \ll Q$, on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then, the entropy of X, denoted by $H(X)$, $H(\frac{\mathrm{d}P_X}{\mathrm{d}Q}(x))$ or $H(P_X)$, is:

$$H(X) \triangleq -\sum_{x\in\operatorname{supp} P_X} \frac{\mathrm{d}P_X}{\mathrm{d}Q}(x) \log P_X(x). \tag{130}$$

When Q is the counting measure $\dfrac{\mathrm{d}P_X}{\mathrm{d}Q}$ is the probability mass function induced by P_X, and when Q is the Lebesgue measure $\dfrac{\mathrm{d}P_X}{\mathrm{d}Q}$ is the probability density function induced by P_X.

B. Mutual information

Let $P_{XY} \in \triangle(\mathcal{X} \times \mathcal{Y})$ and $P_{YX} \in \triangle(\mathcal{Y} \times \mathcal{X})$ be two joint probability measures, satisfying (41), such that $P_{XY} \triangleq P_{X|Y}P_X$ and $P_{XY} \triangleq P_{Y|X}P_Y$. Let the marginal probability measures in $\triangle(\mathcal{X})$ and $\triangle(\mathcal{Y})$, be denoted by P_X and P_Y and satisfy for all measurable sets $\mathcal{A} \subseteq \mathcal{X}$ and for all measurable sets $\mathcal{B} \subseteq \mathcal{Y}$, (42) and (43) respectively.. Let two product measures $P_X P_Y \in \triangle(\mathcal{Y} \times \mathcal{X})$and $P_Y P_X \in \triangle(\mathcal{Y} \times \mathcal{X})$ be the product of the marginals. For all $x \in \mathcal{X}$, the probability measure $P_{Y|X=x}$ is absolutely continuous with respect to P_Y; and for all $y \in \mathcal{Y}$, the probability measure $P_{X|Y=y}$ is absolutely continuous with respect to P_X.

$$\mathrm{I}(P_{Y|X}; P_Y P_X) \triangleq \mathrm{D}(P_{YX} \| P_Y P_X) \tag{131}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_Y P_X}(x, y)\Big)\mathrm{d}P_{YX}(x, y) \tag{132}$$
$$= \int \mathrm{D}\big(P_{Y|X=x} \| P_Y\big)\mathrm{d}P_X(x), \tag{133}$$

$$\mathrm{I}(P_{X|Y};P_XP_Y) \triangleq \mathrm{D}(P_{XY}\|P_XP_Y) \tag{134}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_XP_Y}(x,y)\Big)\mathrm{d}P_{XY}(x,y) \tag{135}$$
$$= \int \mathrm{D}\big(P_{X|Y=y}\|P_X\big)\mathrm{d}P_Y(y). \tag{136}$$

$$\mathrm{D}(P_{XY}\|P_XP_Y) \tag{137}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_XP_Y}(x,y)\Big)\mathrm{d}P_{XY}(x,y) \tag{138}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\mathrm{d}P_{XY}(x,y) \tag{139}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_XP_Y}(x,y)\mathrm{d}P_XP_Y(x,y) \tag{140}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\mathrm{d}P_XP_Y(x,y) \tag{141}$$
$$= \int\int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\mathrm{d}P_X(x)\mathrm{d}P_Y(y). \tag{142}$$

Option 1:

$$\int\int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\mathrm{d}P_X(x)\mathrm{d}P_Y(y)$$
$$= \int\int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\mathrm{d}P_{X|Y=y}(x)\mathrm{d}P_Y(y) \tag{143}$$
$$= \int \mathrm{D}\big(P_{X|Y=y}\|P_X\big)\mathrm{d}P_Y(y). \tag{144}$$

Option 2:

$$\int\int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\mathrm{d}P_X(x)\mathrm{d}P_Y(y)$$
$$= \int\int \log\Big(\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y)\Big)\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y)\mathrm{d}P_X(x)\mathrm{d}P_Y(y) \tag{145}$$
$$= \int\int \log\Big(\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y)\Big)\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y)\mathrm{d}P_Y(y)\mathrm{d}P_X(x) \tag{146}$$
$$= \int\int \log\Big(\frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y)\Big)\mathrm{d}P_{Y|X=x}(y)\mathrm{d}P_X(x) \tag{147}$$
$$= \int \mathrm{D}\big(P_{Y|X=x}\|P_Y\big)\mathrm{d}P_X(x). \tag{148}$$

Option 3:

$$\int\int \log\Big(\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\Big)\frac{\mathrm{d}P_{X|Y=y}}{\mathrm{d}P_X}(x)\mathrm{d}P_X(x)\mathrm{d}P_Y(y)$$
$$= \int\int \log\Big(\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_YP_X}(y,x)\Big)\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_YP_X}(y,x)\mathrm{d}P_X(x)\mathrm{d}P_Y(y) \tag{149}$$
$$= \int\int \log\Big(\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_YP_X}(y,x)\Big)\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_YP_X}(y,x)\mathrm{d}P_Y(y)\mathrm{d}P_X(x) \tag{150}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_YP_X}(y,x)\Big)\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_YP_X}(y,x)\mathrm{d}P_YP_X(y,x) \tag{151}$$
$$= \int \log\Big(\frac{\mathrm{d}P_{YX}}{\mathrm{d}P_YP_X}(y,x)\Big)\mathrm{d}P_{YX}(y,x) \tag{152}$$
$$= \mathrm{D}(P_{YX}\|P_YP_X). \tag{153}$$

C. Lautum information

REFERENCES

[1] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA, USA: Academic Press, 2000.

[2] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing Statistical Hypotheses*, 3rd ed. New York, NY, USA: Springer, 2005.