# Proofs for Folklore Theorems on the Radon-Nikodym Derivative

Yaiza Bermudez*, Gaetan Bisson†, Iñaki Esnaola‡§, and Samir M. Perlaza*†§
*INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France.
†Laboratoire GAATI, Université de la Polynésie française, Fa'a'ā, French Polynesia.
‡School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, United Kingdom.
§ECE Dept. Princeton University, Princeton, 08544 NJ, USA.

*Abstract*—**Rigorous statements and formal proofs are presented for both foundational and advanced folklore theorems on the Radon-Nikodym derivative. The cases of product and marginal measures are carefully considered; and the hypothesis under which the statements hold are rigorously enumerated.**

## I. Introduction

In mathematics, folklore theorems refer to results that are widely accepted and frequently utilized by experts but are often not formally proven or explicitly documented. In game theory for example, the original Folk theorem earned its name because, although it was widely recognized among game theorists during the 1950s, it remains unpublished and without attribution to particular authors. See for instance [1] and [2]. Folklore theorems populate all areas in mathematics. In information theory, a large set of folklore theorems involve the Radon-Nikodym derivative (RND), first introduced by Radon [3]; and later generalized by Nikodym [4]. The existence of such folklore theorems in this area arises in part from the fact that all Shannon's information measures can be defined in terms of the RND. Interestingly, most properties of the RND are often presented as comments in most textbooks on measure theory and probability theory, c.f, [5]–[11]

Claude Shannon did not use the RND in his foundational papers [12], [13] to define entropy and mutual information. Instead, Shannon opted for restricting his publications to the case in which measures either possess a probability mass function or a probability density function, which are both instances of RNDs. This choice significantly influenced the presentation of most subsequent results in information theory and established the style in which classical textbooks were written [14]–[30]. Nonetheless, the RND has been increasingly adopted in modern textbooks [31] and in the definition of new information measures, e.g., lautum information [32], to privilege a unified presentation. That is, independently of the measure used as a reference, e.g., the counting measure, the Lesbegue measure, etc. Some recent results whose presentation relies on the RND are for instance [33]–[42]. In the light of the central role of the RND in information theory, this paper presents rigorous statements and formal proofs for the most common folklore theorems on the RND. These include the change of measure theorem, the proportional-measures theorem, the chain rule, the multiplicative inverse, linearity and continuity theorems, as well as the product-measure theorem. Less common theorems such as the unit measure and two Bayes-like rules of the RND are also formally proved. This is the first time such material appears in peer-reviewed literature. Finally, it is important to highlight that particular attention has been put on rigorously stating the conditions under which these theorems hold, hopefully providing a valuable reference for researchers and students in information theory.

## II. Preliminaries

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, [5, Theorem 2.2.1]):* Let $P$ and $Q$ be two measures on a given measurable space $(\Omega, \mathscr{F})$, such that $Q$ is $\sigma$-finite and $P \ll Q$. Then, there exists a nonnegative Borel measurable function $g: \Omega \rightarrow \mathbb{R}$ 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 $g(x) = \frac{\mathrm{d}P}{\mathrm{d}Q}(x)$.

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.

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 $\sigma$-finite measure. Let $f : \Omega \rightarrow \mathbb{R}$ 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 [5, 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 [5, 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 [5, Theorem 1.5.5(b)]; and $(b)$ the integral is a continuous map from that space [5, 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 $\sigma$-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 $\sigma$-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,

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

*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 $\sigma$-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,

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

*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 [5, 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 $\sigma$-finite measure on $(\Omega, \mathscr{F})$, and let $Q_1, Q_2, \cdots$ be an infinite sequence of $\sigma$-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 [5, 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 $\sigma$-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,

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

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

$$P_1P_2\left(\mathcal{A}\right) = \int_{\mathcal{A}} \mathrm{d}P_1P_2\left(x_1, x_2\right) \tag{29}$$

$$= \int\int_{\mathcal{A}_{x_2}} \mathrm{d}P_1\left(x_1\right)\mathrm{d}P_2\left(x_2\right) \tag{30}$$

$$= \int\int_{\mathcal{A}_{x_2}} \frac{\mathrm{d}P_1\left(x_1\right)}{\mathrm{d}Q_1}\mathrm{d}Q_1\left(x_1\right)\mathrm{d}P_2\left(x_2\right) \tag{31}$$

$$= \int\int_{\mathcal{A}_{x_2}} \frac{\mathrm{d}P_1(x_1)}{\mathrm{d}Q_1}\frac{\mathrm{d}P_2(x_2)}{\mathrm{d}Q_2}\mathrm{d}Q_1(x_1)\mathrm{d}Q_2(x_2) \tag{32}$$

$$= \int_{\mathcal{A}} \frac{\mathrm{d}P_1}{\mathrm{d}Q_1}\left(x_1\right)\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}\left(x_2\right)\mathrm{d}Q_1Q_2\left(x_1, x_2\right), \tag{33}$$

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 $\mathcal{A} \in \mathscr{F}_1 \times \mathscr{F}_2$,

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

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 $\sigma$-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

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

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

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

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 $\mathcal{B} \in \mathscr{F}_{\mathcal{Y}}$,

$$\begin{aligned} P_X(\mathcal{A}) &\triangleq P_{XY}(\mathcal{A} \times \mathcal{Y}) = P_{YX}(\mathcal{Y} \times \mathcal{A}); \text{ and} && (42)\\ P_Y(\mathcal{B}) &\triangleq P_{XY}(\mathcal{X} \times \mathcal{B}) = P_{YX}(\mathcal{B} \times \mathcal{X}). && (43) \end{aligned}$$

From the total probability theorem [5, 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

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

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

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

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,

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

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

$$\begin{aligned} P_Y(\mathcal{A}) &= \int \int_{\mathcal{A}} \mathrm{d}P_{Y|X=x}(y) \mathrm{d}P_X(x) && (50)\\ &= \int \int_{\mathcal{A}} \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_Y(y) \mathrm{d}P_X(x) && (51)\\ &= \int_{\mathcal{A}} \int \frac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_Y}(y) \mathrm{d}P_X(x) \mathrm{d}P_Y(y), && (52) \end{aligned}$$

where the equality in (51) follows from a change of measure (Theorem 2). Moreover, (52) is obtained using Fubini's theorem [5, 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 [5, 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 [5, 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). ■

*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,

$$\frac{\mathrm{d}P_X P_Y}{\mathrm{d}P_{XY}}(x,y) \overset{a.s.}{\underset{P_{XY}}{=}} \frac{\mathrm{d}P_X}{\mathrm{d}P_{X|Y=y}}(x)$$
$$\overset{a.s.}{\underset{P_{XY}}{=}} \frac{\mathrm{d}P_Y}{\mathrm{d}P_{Y|X=x}}(y) \overset{a.s.}{\underset{P_{YX}}{=}} \frac{\mathrm{d}P_Y P_X}{\mathrm{d}P_{YX}}(y,x). \tag{74}$$

*Proof:* The proof follows along the same lines as the proof of Theorem 10. ■

## V. Conclusions and Final Remarks

This paper provides proofs for several well-known folklore theorems, as well as some other lesser-known results in information theory involving the Radon-Nikodym derivative. Notably, these theorems serve as fundamental tools for establishing various properties of Shannon's information measures in a unified framework, that is, regardless of whether the underlying measures admit a probability mass function or a probability density function. Nonetheless, the proof of these properties are left out of the scope of this paper, due to space constraints. Finally, it is important to highlight that the presented results hold in full generality, as the reference measure is assumed to be any $\sigma$-finite measure.

# References

[1] D. Fudenberg and E. Maskin, “The Folk theorem in repeated games with discounting or with incomplete information,” *Econometrica*, vol. 54, no. 3, pp. 533–554, 1986.

[2] D. Fudenberg, *Game Theory*, 1st ed. Cambridge, MA, USA: MIT Press, 1991.

[3] J. Radon, *Theorie und Anwendungen der absolut additiven Mengenfunktionen*, 1st ed. Vienna, Austria: Hölder, 1913.

[4] O. Nikodym, “Sur une généralisation des intégrales de mj radon,” *Fundamenta Mathematicae*, vol. 15, no. 1, pp. 131–179, 1930.

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

[6] P. R. Halmos, *Measure Theory*, 1st ed. Princeton, NJ, USA: Van Nostrand, 1950.

[7] P. Billingsley, *Probability and Measure*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2012.

[8] H. L. Royden and P. M. Fitzpatrick, *Real Analysis*, 4th ed. Beijing, PRC: China Machine Press, 2010.

[9] S. Axler, *Measure, Integration & Real Analysis*, 1st ed. New York, NY, USA: Springer, 2020.

[10] R. Durrett, *Probability: Theory and Examples*, 5th ed. Cambridge, UK: Cambridge University Press, 2019.

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

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

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

[14] S. Kullback, *Information Theory and Statistics*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1959.

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

[16] R. B. Ash, *Information Theory*, 1st ed. Mineola, NY, USA: Dover Publications, 1990.

[17] M. Bloch and J. Barros, *Physical-Layer Security: From Information Theory to Security Engineering*, 1st ed. Cambridge, UK: Cambridge University Press, 2011.

[18] E. Çınlar, *Probability and Stochastics*, 1st ed. New York, NY, USA: Springer, 2011.

[19] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 1st ed. Cambridge, UK: Cambridge University Press, 2011.

[20] A. El Gamal and Y.-H. Kim, *Network Information Theory*, 1st ed. Cambridge, UK: Cambridge University Press, 2011.

[21] R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*, 1st ed. Cambridge, MA, USA: MIT Press, 1961.

[22] A. Feinstein, *Foundations of Information Theory*, 1st ed. New York, NY, USA: McGraw-Hill, 1958.

[23] R. G. Gallager, *Information Theory and Reliable Communication*, 1st ed. New York, NY, USA: Wiley, 1968.

[24] T. S. Han, *Information-Spectrum Methods in Information Theory*, 1st ed. Berlin, Germany: Springer, 2003.

[25] J. Kapur, *Maximum-Entropy Models in Science and Engineering*, 1st ed. New York, NY, USA: John Wiley & Sons, 1989.

[26] R. J. McEliece, *The Theory of Information and Coding*, 1st ed. Cambridge, UK: Cambridge University Press, 2002.

[27] M. Mezard and A. Montanari, *Information, Physics, and Computation*, 1st ed. Oxford, UK: Oxford University Press, 2009.

[28] M. Pinsker, *Information and Information Stability of Random Variables and Processes*, 1st ed. San Francisco, CA, USA: Holden-Day, 1964.

[29] J. Wolfowitz, *Coding Theorems of Information Theory*, 1st ed. Berlin, Germany: Springer, 1964.

[30] R. W. Yeung, *Information Theory and Network Coding*, 1st ed. New York, NY, USA: Springer, 2008.

[31] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*, 1st ed. Cambridge, UK: Cambridge University Press, 2024.

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

[33] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, “Asymmetry of the relative entropy in the regularization of empirical risk minimization,” *arXiv preprint arXiv:2410.02833*, 2024.

[34] R. Agrawal and T. Horel, “Optimal bounds between $f$-divergences and integral probability metrics,” *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 5662–5720, Jan. 2021.

[35] 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.

[36] 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*, vol. 1, no. 1, pp. 1–24, Apr. 2023.

[37] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, “An information-theoretic approach to generalization theory,” Ph.D. dissertation, KTH Royal Institute of Technology, 2024.

[38] R. C. Yavas, V. Kostina, and M. Effros, “Variable-length sparse feedback codes for point-to-point, multiple access, and random access channels,” *IEEE Transactions on Information Theory*, 2023.

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

[40] S. Verdú, “Relative information spectra with applications to statistical inference,” *AIMS Mathematics*, vol. 9, no. 12, pp. 35 038–35 090, 2024.

[41] 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.

[42] 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.