Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 9(12): 35038–35090.
DOI: 10.3934/math.20241668
Received: 25 July 2024
Revised: 04 November 2024
Accepted: 09 December 2024
Published: 16 December 2024

Research article

Relative information spectra with applications to statistical inference

Sergio Verdú

Independent researcher, Princeton, NJ 08540, USA

* **Correspondence:** Email: verdu@informationtheory.org.

Abstract: For any pair of probability measures defined on a common space, their relative information spectra—specifically, the distribution functions of the loglikelihood ratio under either probability measure—fully encapsulate all that is relevant for distinguishing them. This paper explores the properties of the relative information spectra and their connections to various measures of discrepancy including total variation distance, relative entropy, Rényi divergence, and general f-divergences. A simple definition of sufficient statistics, termed I-sufficiency, is introduced and shown to coincide with longstanding notions under the assumptions that the data model is dominated and the observation space is standard. Additionally, a new measure of discrepancy between probability measures, the NP-divergence, is proposed and shown to determine the area of the error probability pairs achieved by the Neyman-Pearson binary hypothesis tests. For independent identically distributed data models, that area is shown to approach 1 at a rate governed by the Bhattacharyya distance.

Keywords: information theory; statistical inference; sufficient statistics; hypothesis testing; Neyman-Pearson tests; information spectrum method; relative entropy; Kullback-Leibler divergence; f-divergence; total variation distance; Bhattacharyya distance
Mathematics Subject Classification: 62B05, 62B10, 62C10, 62F03, 94A15, 94A17

1. Introduction

Shortly after the advent of information theory [1], Kullback and Leibler [2] introduced *relative entropy* (or Kullback-Leibler divergence) as a means of generalizing to arbitrary alphabets Shannon's foundational information measures—entropy, differential entropy, and mutual information. They recognized that relative entropy could play a pivotal role, not in Shannon's data compression and transmission problems, but in statistical inference, in particular, in the theory of sufficient statistics, which had recently been put on a sound mathematical footing by Halmos and Savage [3]. The application of information theory to statistical inference, initiated in [2], continued with Fano's inequality [4]—a lower bound on error probability in Bayesian M-ary hypothesis testing based on conditional entropy. Lindley [5] suggested using mutual information to explore sufficiency. Chernoff [6] found an asymptotic operational role for relative entropy in another fundamental pillar of statistical inference, the theory of hypothesis testing pioneered by Neyman and Pearson in [7]. Soon after, Sanov [8] showed that relative entropy plays a pivotal role in the theory of large deviations pioneered two decades earlier by Cramér in [9]. For the purpose of statistical modeling, Jaynes [10,11] and Kullback [12] advocated the maximization of entropy and the minimization of relative entropy with a fixed nominal reference measure, respectively.

Other information theoretic measures would prove useful in statistical inference. *Rényi divergence* [13] and *Chernoff information* [14] emerged as key tools in the asymptotic analysis of non-Bayesian and Bayesian hypothesis testing, respectively. Csiszár [15] showed that the role of relative entropy in sufficient statistics, found by Kullback and Leibler, could be extended to *f-divergences*, a much wider collection of discrepancy measures that obey the data processing principle (no processing can increase them). Among the many f-divergences that have found widespread applications in statistical inference are total variation distance, χ^2-divergence [16], Hellinger distance [17], Hellinger divergence [14], Vincze-Le Cam divergence [18,19], and de Groot statistical information [20].

Moving forward to the last decade of the XXth century, [21] started a new direction in information theory: The *information spectrum* approach, whose original goal was to generalize the flagship asymptotic results in information theory without assumptions of discreteness, memorylessness, ergodicity, or even stationarity. Working with very little structure has the benefit of bringing out the essential aspects that allow Shannon's results [1] to transcend their original habitat. A price to be paid for the generality of those results is that entropy, relative entropy, and mutual information are no longer sufficient to express the asymptotic fundamental limits. Those information measures are expectations of random variables whose distributions, dubbed *information spectra* in [21], emerge as the crucial ingredients in the solution. Han's monograph [22] provides a comprehensive overview of the application of the information spectrum method to the asymptotic fundamental limits in various domains, including lossless and lossy data compression, data transmission, hypothesis testing, channel resolvability, and random number generation. Started in [23], another trend in information theory seeks to determine *non-asymptotic* fundamental limits, e.g., what is the transmission rate compatible with a blocklength of 1000 and a probability of decoding error of 10^{-2}? Approximate answers to this type of questions can be obtained through upper and lower bounds that depend on the information spectra.

Entropy is a special case of mutual information, which in turn is a special case of relative entropy. The relative entropy, $D(P_X \,\|\, P_Y)$, of probability measures P_X and P_Y defined on the same space is the expectation of the random variable $\imath_{X\|Y}(X)$, where $\imath_{X\|Y}(x)$ stands for the *relative information* defined as the logarithm of the Radon-Nikodym derivative $\frac{\mathrm{d}P_X}{\mathrm{d}P_Y}(x)$, or more generally, $\log \frac{\mathrm{d}P_X}{\mathrm{d}\mu}(x) - \log \frac{\mathrm{d}P_Y}{\mathrm{d}\mu}(x)$, where μ dominates both probability measures. The key objects of interest in this paper are the *relative information spectra*, namely, the cumulative distribution functions of $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$. In addition to a number of properties satisfied by the relative information spectra, we show new results in both sufficient statistics and binary hypothesis testing through the application of those properties.

To enhance readability and ease of reference, the rest of this work is organized in one hundred items grouped into eight sections, plus an appendix.

Section 2 contains most of the terminology and notation used throughout the paper, as well as several auxiliary results used in the sequel. As no restrictions (including absolute continuity) are placed on the pairs of probability measures under purview, the notions of *relative support* and *coefficient of* *absolute discontinuity* prove to be of central importance in the subsequent development.

Section 3 deals with the fundamental properties of the *relative information*, including the change of measure formulas without requiring absolute continuity. It also explores properties of Rényi divergence and *information density*—a special case of relative information whose expectation is the mutual information, and which proves useful in Section 7.

Section 4 focuses on the interplay of the distributions of the random variables $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$. The key notion of *equivalent pairs* of probability measures, proposed recently in [24] in the special case of absolutely continuous probability measures, is given here in full generality, along with several necessary and sufficient conditions involving Rényi divergence and f-divergences.

Section 5 shows various ways to express and bound *total variation distance* as a function of the relative information spectra, a problem initially undertaken by Le Cam [19].

A new measure of discrepancy between probability measures, dubbed the NP-*divergence*, is introduced in Section 6. Although it satisfies the data processing principle, NP-divergence is not an f-divergence. Its main operational role, which justifies its name, is revealed in Section 8.

Section 7 presents a new notion of sufficient statistics, *I-sufficiency*, based on equivalent pairs. To put this notion in perspective, Section 7 also includes a discussion of the leading existing definitions of sufficient statistics, such as classical (Fisher) sufficiency, Blackwell sufficiency, and Bayes sufficiency. I-sufficiency is a natural bridge between those notions and criteria based on the equality of the data processing inequality for f-divergence. All those notions turn out to be equivalent under the assumption that the data model is dominated and defined on a standard space.

Section 8 gives a self-contained solution to the non-asymptotic fundamental tradeoff region consisting of the set of achievable conditional error probabilities in non-Bayesian binary hypothesis testing, at a level of detail apparently not available elsewhere. A scalar proxy is often sought to quantify how well probability measures can be distinguished. In non-Bayesian hypothesis testing, the area of the tradeoff region is arguably the most natural scalar measure. Section 8 demonstrates that this area equals one-half of the NP-*divergence*. This establishes an interesting relationship between the hypothesis testing problems

$$\begin{array}{ll} \mathsf{H}_0\colon\ y \sim P_0, & \qquad \mathsf{H}_L\colon\ (y_1, y_2) \sim P_0 \otimes P_1, \\ \mathsf{H}_1\colon\ y \sim P_1, & \qquad \mathsf{H}_R\colon\ (y_1, y_2) \sim P_1 \otimes P_0. \end{array}$$

The area of the fundamental tradeoff region for $\{\mathsf{H}_0, \mathsf{H}_1\}$ is shown to equal $1 - 2\epsilon_{\min}(\mathsf{H}_L, \mathsf{H}_R)$, where $\epsilon_{\min}(\mathsf{H}_L, \mathsf{H}_R)$ is the minimum (Bayesian) error probability when $\{\mathsf{H}_L, \mathsf{H}_R\}$ are equally likely. A new asymptotic operational role is found for the *Bhattacharyya distance* in the setting of independent identically distributed data.

Section 9 gives a recap of the main new results found in the paper.

2. Preliminaries

This section introduces basic terminology and notation, along with supporting results used in the remainder of the paper.

1. $\mathscr{P}_{\mathcal{A}}$ denotes the set of probability measures defined on the measurable space $(\mathcal{A}, \mathscr{F})$.
2. For $P \in \mathscr{P}_{\mathcal{A}}$, $X \sim P$ means that $\mathbb{P}[X \in E] = P(E)$ for all $E \in \mathscr{F}$.

3. A *random transformation*

$$P_{Y|X}\colon (\mathcal{A},\mathscr{F}) \to (\mathcal{B},\mathscr{G}) \tag{2.1}$$

is a collection $\{P_{Y|X=a} \in \mathscr{P}_{\mathcal{B}},\ a \in \mathcal{A}\}$ of probability measures defined on the measurable space $(\mathcal{B},\mathscr{G})$, such that for every $B \in \mathscr{G}$, $f_B\colon \mathcal{A} \to [0,1]$ given by $f_B(a) = P_{Y|X=a}(B)$ is a Borel $\mathscr{F}$-measurable function. In the literature, random transformations are also referred to as *Markov kernels*. The sets $\mathcal{A}$ and $\mathcal{B}$ are known as the input and output alphabets of the random transformation, respectively. Note that a joint probability measure P_{XY} need not be defined notwithstanding the notation in (2.1). If in addition to the random transformation (2.1), a probability measure $P_X \in \mathscr{P}_{\mathcal{A}}$ is defined, then the *input/output* joint probability measure P_{XY} on $(\mathcal{A}\times\mathcal{B},\mathscr{F}\otimes\mathscr{G})$ is given by

$$P_{XY}(A\times B) = \int_A P_{Y|X=a}(B)\,\mathrm{d}P_X(a), \quad A\times B \in \mathscr{F}\otimes\mathscr{G}. \tag{2.2}$$

The marginal output probability measure, P_Y, also known as the *response* of $P_{Y|X}$ to P_X, is denoted by

$$P_X \to P_{Y|X} \to P_Y. \tag{2.3}$$

4. If $(P,Q) \in \mathscr{P}^2_{\mathcal{A}}$, $P \ll Q$ means that Q *dominates* P, or alternatively, P is *absolutely continuous* with respect to Q, i.e., $P(A) = 0$ for any $A \in \mathscr{F}$ such that $Q(A) = 0$. More generally, a collection $\mathscr{P} \subset \mathscr{P}_{\mathcal{A}}$ is said to be dominated by Q if $P \ll Q$ for all $P \in \mathscr{P}$. If Q dominates $\mathscr{P}$ and $Q(E) = 0$ whenever $E \in \mathscr{F}$ is such that $P(E) = 0$ for all $P \in \mathscr{P}$, Q is said to be *equivalent* to $\mathscr{P}$. The same terminology applies to general measures on $(\mathcal{A},\mathscr{F})$.

5. If $(P,Q) \in \mathscr{P}^2_{\mathcal{A}}$ and $P \ll Q \ll P$, then we write $P \ll\gg Q$ and P and Q are said to be *mutually absolutely continuous* or *equivalent*.

6. **Lemma 1.** *[3, Lemma 7] Assume that there exists a σ-finite measure on $(\mathcal{A},\mathscr{F})$ that dominates the collection $\mathscr{P} \subset \mathscr{P}_{\mathcal{A}}$. Then, there exists a probability measure in $\mathscr{P}_{\mathcal{A}}$ that is equivalent to the collection $\mathscr{P}$. In fact, there exists a finite or countably infinite collection $\{P_i \in \mathscr{P},\ i \in \mathcal{I}\}$, such that $\sum_{i\in\mathcal{I}} \pi_i P_i$ is equivalent to $\mathscr{P}$ for every probability mass function π on $\mathcal{I}$ with $\pi_i > 0$ for all $i \in \mathcal{I}$.*

   In light of Lemma 1, we frequently refer to a collection of probability measures as being *dominated*, without specifying the dominating measure, which is understood to be either a probability measure, or not more generally, a σ-finite measure. Non-σ-finite measures are of no interest in this paper. Any finite or countably infinite collection of probability measures is dominated. Examples of undominated collections of probability measures on $(\mathbb{R},\mathscr{B})$ are:

   - $\{\delta_\theta,\ \theta \in [0,1]\}$, with δ_θ the point mass on $(\mathbb{R},\mathscr{B})$ that assigns probability one to $\{\theta\}$.
   - $\{P\colon P(B) = P(\{\omega\colon -\omega \in B\}),\ \text{for all } B \in \mathscr{B}\}$.

   Despite a contrary claim in [25], undominated collections are more often the exception than the rule in most applications commonly encountered in statistical inference and information theory.

7. If $(P,Q) \in \mathscr{P}^2_{\mathcal{A}}$, P and Q are said to be *mutually singular* or *orthogonal*, $P \perp Q$, if there exists an event $F \in \mathscr{F}$ with $P(F) = 0$ and $Q(F) = 1$.

8. If $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$, the *coefficient of absolute discontinuity* of P relative to Q is defined as

$$\Pi(P \,\|\, Q) = \min_{\substack{A \in \mathscr{F}: \\ Q(A) = 1}} P(A). \tag{2.4}$$

Note that

$$P \ll Q \quad \Longleftrightarrow \quad \Pi(P \,\|\, Q) = 1, \tag{2.5}$$
$$P \perp Q \quad \Longleftrightarrow \quad \Pi(P \,\|\, Q) = 0 \quad \Longleftrightarrow \quad \Pi(Q \,\|\, P) = 0. \tag{2.6}$$

9. If $P \in \mathscr{P}_{\mathcal{A}}$ and $Q \in \mathscr{P}_{\mathcal{B}}$ (the set of probability measures defined on the measurable space $(\mathcal{B}, \mathscr{G})$), then $P \otimes Q$ denotes the product measure on the measurable space $(\mathcal{A} \times \mathcal{B}, \mathscr{F} \otimes \mathscr{G})$. The coefficient of absolute discontinuity for product probability measures is

$$\Pi(P_1 \otimes Q_1 \,\|\, P_0 \otimes Q_0) = \min_{\substack{A \in \mathscr{F} \otimes \mathscr{G}: \\ [P_0 \otimes Q_0](A) = 1}} [P_1 \otimes Q_1](A) \tag{2.7}$$
$$\leq \min_{\substack{F \in \mathscr{F}: \\ P_0(F) = 1}} P_1(F) \min_{\substack{G \in \mathscr{G}: \\ Q_0(G) = 1}} Q_1(G) \tag{2.8}$$
$$= \Pi(P_1 \,\|\, Q_1) \cdot \Pi(P_0 \,\|\, Q_0). \tag{2.9}$$

In fact, equality holds in (2.8) since we can lower bound the left side replacing $[P_1 \otimes Q_1](A)$ by $[P_1 \otimes Q_1](F \times G)$ for any $F \times G \subset A$.

10. If $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$ and $P \ll Q$, then the *Radon-Nikodym derivative* (or *density*) of P with respect to Q is the Borel $\mathscr{F}$-measurable nonnegative function

$$\frac{\mathrm{d}P}{\mathrm{d}Q} : \mathcal{A} \to [0, \infty) \tag{2.10}$$

such that any nonnegative Borel $\mathscr{F}$-measurable function $f : \mathcal{A} \to [0, \infty)$ satisfies the *change of measure* formula

$$\mathbb{E}[f(X)] = \mathbb{E}\left[\frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\, f(Y)\right], \quad X \sim P,\; Y \sim Q. \tag{2.11}$$

11. If $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$, define —up to an event of zero $P + Q$—the *support of P relative to Q* as

$$\mathcal{S}_{P\|Q} = \left\{ a \in \mathcal{A} : \frac{\mathrm{d}P}{\mathrm{d}\mu}(a) > 0 \right\} \in \mathscr{F}, \tag{2.12}$$

where μ is any measure that dominates $\{P, Q\}$, such as $P + Q$.

Lemma 2. *For any $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$,*

$$P(\mathcal{S}_{P\|Q}) = 1, \tag{2.13}$$
$$Q(\mathcal{S}_{P\|Q}) = Q(\mathcal{S}_{P\|Q} \cap \mathcal{S}_{Q\|P}) = \min_{\substack{A \in \mathscr{F}: \\ P(A) = 1}} Q(A) = \Pi(Q \,\|\, P). \tag{2.14}$$

Proof. To verify (2.13), simply note that $P(\mathcal{S}^c_{P\|Q}) = \int 1\{a \notin \mathcal{S}_{P\|Q}\} \frac{\mathrm{d}P}{\mathrm{d}\mu}(a)\, \mathrm{d}\mu(a) = 0$ for any $\mu \gg \{P, Q\}$. To justify (2.14), we need to show that if $A \in \mathscr{F}$ with $Q(A) < Q(\mathcal{S}_{P\|Q})$, then $P(A) < 1$. If $G \in \mathscr{F}$ is such that $G \subset \mathcal{S}_{P\|Q}$ and $Q(G) > 0$, then

$$P(G) = \int_G \frac{\mathrm{d}P}{\mathrm{d}\mu}(a)\, \mathrm{d}\mu(a) > 0, \tag{2.15}$$

because $\mu(G) > 0$ and $\frac{\mathrm{d}P}{\mathrm{d}\mu}(a) > 0$ if $a \in G$. Since $Q(\mathcal{S}_{P\|Q} \cap A^c) \geq Q(\mathcal{S}_{P\|Q}) - Q(A) > 0$, we obtain $0 < P(\mathcal{S}_{P\|Q} \cap A^c) = P(A^c)$. □

12. Apart from their essential contribution in our framework, the concepts in Items 8 and 11 merit broader popularity in probability theory. For example, they lead to an elementary constructive proof (cf. the standard proof in [26, p. 135]) of the *Lebesgue decomposition theorem*: If $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$, there exist probability measures $(P_1, P_{\mathbb{0}}) \in \mathscr{P}^2_{\mathcal{A}}$ such that $P_1 \ll Q$, $P_{\mathbb{0}} \perp Q$, and P is the mixture

$$P = \lambda\, P_1 + (1 - \lambda)\, P_{\mathbb{0}}, \tag{2.16}$$

for some $\lambda \in [0, 1]$. First, observe that the constructions for the cases $P \ll Q$ and $P \perp Q$ are trivial: If $P \ll Q$, then $(P_1, \lambda) = (P, 1)$; if $P \perp Q$, then $(P_{\mathbb{0}}, \lambda) = (P, 0)$. In the nontrivial case $P \not\ll Q$ and $P \not\perp Q$, we have

$$0 < \Pi(P \,\|\, Q) = P(\mathcal{S}_{P\|Q}) < 1, \tag{2.17}$$

and the law of total probability yields, for any $A \in \mathscr{F}$,

$$P(A) = P\left(A \mid \mathcal{S}_{Q\|P}\right)\, P\left(\mathcal{S}_{Q\|P}\right) + P\left(A \mid \mathcal{S}^c_{Q\|P}\right)\, P\left(\mathcal{S}^c_{Q\|P}\right). \tag{2.18}$$

So in (2.16), we have $\lambda = \Pi(P \,\|\, Q)$, $P_1 = P(\cdot \mid \mathcal{S}_{Q\|P}) \ll Q$, and $P_{\mathbb{0}} = P(\cdot \mid \mathcal{S}^c_{Q\|P}) \perp Q$.

13. The *moment generating function* and *cumulant generating function* of a $[-\infty, +\infty)$-valued random variable U are defined, respectively, by

$$\mathrm{M}_U(t) = \mathbb{E}\left[\mathrm{e}^{t\,U}\right], \tag{2.19}$$

$$\Lambda_U(t) = \log_{\mathrm{e}} \mathrm{M}_U(t). \tag{2.20}$$

Note that $\lim_{t\downarrow 0} \Lambda_U(t)$ is either infinite or equal to $\log \mathbb{P}[U > -\infty]$. If there exists $t_0 > 0$ such that $\Lambda_U(t) = \Lambda_V(t) < \infty$ for $t \in (0, t_0)$, then U and V have identical distributions (e.g., [27, p. 337]). Since $\Lambda_U(-t) = \Lambda_{-U}(t)$, U and V have identical distributions if $\Lambda_U(t) = \Lambda_V(t) < \infty$ for $t \in (-t_0, 0)$.

3. Relative information

14. If $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$ and $P \ll Q$, then the *relative information* of P with respect to Q is the Borel $\mathscr{F}$-measurable function

$$\imath_{P\|Q}(a) = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(a) \in [-\infty, \infty). \tag{3.1}$$

More generally, without requiring $P \ll Q$, if ρ is a probability (or σ-finite) measure which dominates $\{P, Q\}$, and the respective densities are denoted by $p = \frac{\mathrm{d}P}{\mathrm{d}\rho}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\rho}$, then the (generalized) relative information is defined as

$$\imath_{P\|Q}(a) = \log \frac{p(a)}{q(a)} = \begin{cases} +\infty, & a \in \mathcal{S}_{P\|Q} \cap \mathcal{S}^c_{Q\|P}; \\ \imath_{P\|\rho}(a) - \imath_{Q\|\rho}(a) \in \mathbb{R}, & a \in \mathcal{S}_{P\|Q} \cap \mathcal{S}_{Q\|P}; \\ -\infty, & a \in \mathcal{S}^c_{P\|Q} \cap \mathcal{S}_{Q\|P}. \end{cases} \tag{3.2}$$

If $a \notin \mathcal{S}_{P\|Q} \cup \mathcal{S}_{Q\|P}$, it is immaterial how to define $\imath_{P\|Q}(a)$. Therefore, any identity involving relative informations (or densities) is to be understood almost surely with respect to any measure dominating both probability measures. It follows from (3.2) that relative information satisfies the skew-symmetry property

$$\imath_{P\|Q}(a) = -\imath_{Q\|P}(a), \quad a \in \mathcal{A}. \tag{3.3}$$

In the discrete case, i.e., $\mathcal{A}$ is finite or countably infinite, if $P(a) + Q(a) > 0$, then

$$\imath_{P\|Q}(a) = \imath_Q(a) - \imath_P(a), \tag{3.4}$$

where the (absolute) information is $\imath_P(a) = \log \frac{1}{P(a)}$. The base of the logarithms in (3.1) and (3.2) determines the units of the relative information. Unless specifically indicated, it can be chosen by the reader. If the chosen base is $\mathrm{b} > 1$, then $\exp(t) = \mathrm{b}^t$. If $b = \mathrm{e}$ [resp., 2], the unit is called nat [resp., bit]. By convention, $\exp(-\infty) = 0$ and $\log 0 = -\infty$. The generalized relative information in bits is equal to $\imath_{P\|Q}(a) = \upsilon(\imath_{P\|R}(a))$ where $R = \frac{1}{2}P + \frac{1}{2}Q$, $\imath_{P\|R}(a)$ is also in bits, and $\upsilon\colon [-\infty, 1] \to [-\infty, +\infty]$ is $\upsilon(t) = t - \log_2(2 - 2^t)$.

15. If $a \in \mathcal{A}$, $(P, Q, R) \in \mathscr{P}^3_{\mathcal{A}}$ and R dominates $\{P, Q\}$, then (3.2) implies the *chain rule*

$$\imath_{P\|R}(a) - \imath_{Q\|R}(a) = \imath_{P\|Q}(a). \tag{3.5}$$

16. Often (recall Item 2) we denote $X \sim P_X$ and $Y \sim P_Y$, in which case we abbreviate $\imath_{P_X\|P_Y}$ as $\imath_{X\|Y}$. The same convention applies to the coefficient of absolute discontinuity in Item 8 and the relative support in Item 11, as well as to relative entropy and other information measures (except total variation distance) considered in the remainder of the paper. This notational convention was popularized by [28] in the context of the entropy function.

17. It follows from (2.13), (2.14), and (3.2) that

$$\mathbb{P}[\imath_{X\|Y}(X) = +\infty] = P_X(\mathcal{S}_{X\|Y} \cap \mathcal{S}^c_{Y\|X}) = 1 - \Pi(X\|Y), \tag{3.6}$$

$$\mathbb{P}[\imath_{X\|Y}(X) = -\infty] = P_X(\mathcal{S}_{Y\|X} \cap \mathcal{S}^c_{X\|Y}) = 0, \tag{3.7}$$

$$\mathbb{P}[\imath_{X\|Y}(Y) = +\infty] = P_Y(\mathcal{S}_{X\|Y} \cap \mathcal{S}^c_{Y\|X}) = 0, \tag{3.8}$$

$$\mathbb{P}[\imath_{X\|Y}(Y) = -\infty] = P_Y(\mathcal{S}_{Y\|X} \cap \mathcal{S}^c_{X\|Y}) = 1 - \Pi(Y\|X). \tag{3.9}$$

18. *Change of measure.* Without the assumption of absolute continuity, the basic change of measure formula (2.11) needs to be modified as follows.

Lemma 3. *For any nonnegative Borel measurable $f\colon \mathcal{A} \to [0,\infty)$,*

$$\begin{aligned}\mathbb{E}\left[f(Y)\,\exp(\imath_{X\|Y}(Y))\right] &= \mathbb{E}\left[f(X)\,1\{\imath_{X\|Y}(X)\in\mathbb{R}\}\right] &(3.10)\\ &= \mathbb{E}\left[f(X)\,1\{\imath_{X\|Y}(X)<\infty\}\right] &(3.11)\\ &= \mathbb{E}\left[f(X)\right], \quad \textit{if } P_X \ll P_Y; &(3.12)\\ \mathbb{E}\left[f(X)\,\exp(-\imath_{X\|Y}(X))\right] &= \mathbb{E}\left[f(Y)\,1\{\imath_{X\|Y}(Y)\in\mathbb{R}\}\right] &(3.13)\\ &= \mathbb{E}\left[f(Y)\,1\{\imath_{X\|Y}(Y)>-\infty\}\right] &(3.14)\\ &= \mathbb{E}\left[f(Y)\right], \quad \textit{if } P_Y \ll P_X, &(3.15)\end{aligned}$$

regardless of whether the expectations are finite or $+\infty$. More generally, if $f\colon \mathcal{A} \to \mathbb{R}$, the random variable on the left side of (3.10) *[resp.,* (3.13)*] is integrable if and only if so is the random variable on the right side, in which case* (3.10) *[resp.,* (3.13)*] holds.*

Proof. Identities (3.11) and (3.14) follow from (3.7) and (3.8), respectively. Suppose that ρ dominates $\{P_X, P_Y\}$ and the respective densities are denoted by p_X and p_Y. The random variable in the expectation on the left side of (3.10) is equal to zero if $Y \notin \mathcal{S}_{X\|Y} \cap \mathcal{S}_{Y\|X}$. Therefore, the left side of (3.10) equals

$$\begin{aligned}&\mathbb{E}\left[f(Y)\,\exp(\imath_{X\|Y}(Y))\,1\{Y\in\mathcal{S}_{X\|Y}\cap\mathcal{S}_{Y\|X}\}\right]\\ &= \int_{\mathcal{S}_{X\|Y}\cap\mathcal{S}_{Y\|X}} f(\omega)\,\frac{p_X(\omega)}{p_Y(\omega)}\,p_Y(\omega)\,\mathrm{d}\rho(\omega) = \mathbb{E}\left[f(X)\,1\{\imath_{X\|Y}(X)\in\mathbb{R}\}\right], \qquad (3.16)\end{aligned}$$

where we have used (3.2). If $P_X \ll P_Y$, then $\mathbb{P}[\imath_{X\|Y}(X)\in\mathbb{R}] = 1$, yielding (3.12). Alternatively, we can prove it directly from the change of measure formula (2.11). To show the claimed result for $f\colon \mathcal{A}\to\mathbb{R}$, we decompose $f = [f]^+ - [-f]^+$, and the desired result follows from the definition of expectation $\mathbb{E}[f(V)] = \mathbb{E}[f^+(V)] - \mathbb{E}[f^-(V)]$ for any V once we apply (3.10) to both $[f]^+$ and $[-f]^+$. Swapping $X \leftrightarrow Y$ and recalling (3.3) results in (3.13)–(3.15). □

Corollary 1. *For any $\beta > 0$, and nonnegative measurable function $g\colon \mathcal{A} \to [0,\infty)$,*

$$\begin{aligned}\beta\,\mathbb{E}\left[g(Y)1\{\imath_{X\|Y}(Y)\geq\log\beta\}\right] &\leq \mathbb{E}\left[g(X)1\{\log\beta\leq\imath_{X\|Y}(X)<\infty\}\right], &(3.17)\\ \mathbb{E}\left[g(X)1\{\imath_{X\|Y}(X)\leq\log\beta\}\right] &\leq \beta\,\mathbb{E}\left[g(Y)1\{-\infty<\imath_{X\|Y}(Y)\leq\log\beta\}\right]. &(3.18)\end{aligned}$$

Proof.

- (3.17) $\Longleftarrow$ (3.10) with $f(a) = g(a)\,1\{\log\beta \leq \imath_{X\|Y}(a)\}$.
- (3.18) $\Longleftarrow$ (3.10) with $f(a) = g(a)\,1\{-\infty < \imath_{X\|Y}(a) \leq \log\beta\}$.

□

19. *Invariance to labeling.*

Lemma 4. *Fix measurable spaces $(\mathcal{A},\mathscr{F})$ and $(\mathcal{B},\mathscr{G})$, and let the $(\mathscr{F},\mathscr{G})$-measurable function $f\colon \mathcal{A}\to\mathcal{B}$ be injective. Then,*

$$\imath_{f(X)\|f(Y)}(f(a)) = \imath_{X\|Y}(a), \quad a\in\mathcal{A}. \qquad (3.19)$$

Conversely, if the $(\mathscr{F},\mathscr{G})$-measurable function $g\colon \mathcal{A}\to\mathcal{B}$ is such that $\imath_{X\|Y}(a)$ depends on $a\in\mathcal{A}$ only through $g(a)$, then

$$\imath_{X\|Y}(a) = \imath_{g(X)\|g(Y)}(g(a)). \qquad (3.20)$$

Proof. Suppose that P_Z dominates $\{P_X, P_Y\}$. Then,

$$P_X(A) = \mathbb{P}[f(X) \in f(A)] \tag{3.21}$$

$$= \mathbb{E}[1\{f(Z) \in f(A)\}\,\varrho(Z)] \tag{3.22}$$

$$= \mathbb{E}[1\{Z \in A\}\,\varrho(Z)], \tag{3.23}$$

where

- (3.21) and (3.23) $\Longleftarrow 1\{a \in A\} = 1\{f(a) \in f(A)\}$ for any $A \in \mathscr{F} \Longleftarrow f$ is injective.
- (3.22) $\Longleftarrow$ Lemma 3 with $\varrho(a) = \exp(\imath_{f(X)\|f(Z)}(f(a)))$.

Therefore, $\varrho(a) = \frac{\mathrm{d}P_X}{\mathrm{d}P_Z}(a)$. In particular, $\mathcal{S}_{X\|Z} = \{a \in \mathscr{A}\colon f(a) \in \mathcal{S}_{f(X)\|f(Z)}\}$. If $P_X \ll P_Y$, we are done since we can just let $P_Z = P_Y$. Otherwise, we can follow the same reasoning with $X \leftarrow Y$, and (3.19) follows from (3.2). To show the converse part, assume for now that $P_X \ll P_Y$. Suppose that $\frac{\mathrm{d}P_X}{\mathrm{d}P_Y}(a) = \psi(g(a))$. Then, again invoking Lemma 3, we get

$$P_{g(X)}(B) = \mathbb{E}[1\{g(X) \in B\}] = \mathbb{E}[1\{g(Y) \in B\}\,\psi(g(Y))], \;\; B \in \mathscr{G}, \tag{3.24}$$

and, consequently, $\frac{\mathrm{d}P_{g(X)}}{\mathrm{d}P_{g(Y)}}(t) = \psi(t)$, which implies (3.20). Without assuming $P_X \ll P_Y$, note that $\imath_{X\|Y}(a) = \infty$ implies $a \in \mathcal{S}^c_{Y\|X} \subset g^{-1}\left(\mathcal{S}^c_{g(Y)\|g(X)}\right)$ and (3.24) continues to hold if $B \subset \mathcal{S}_{g(Y)\|g(X)}$. □

20. *Relative information of relative informations.*

Lemma 5. *Let $(P_X, P_Y) \in \mathscr{P}^2_{\mathscr{A}}$. Define the extended-valued random variables $W = \imath_{X\|Y}(X)$ and $Z = \imath_{X\|Y}(Y)$. Then, using the same units for all three relative informations,*

$$\imath_{W\|Z}(x) = x, \quad x \in [-\infty, \infty]. \tag{3.25}$$

Proof. Since the relative information need not be an injective function, we cannot invoke Lemma 4. The probability of a Borel set $A \in \mathscr{B}$, $A \subset \mathbb{R}$ can be expressed as

$$P_W(A) = \mathbb{P}\left[\imath_{X\|Y}(X) \in A\right] \tag{3.26}$$

$$= \mathbb{E}\left[1\{\imath_{X\|Y}(Y) \in A\}\exp(\imath_{X\|Y}(Y))\right] \tag{3.27}$$

$$= \mathbb{E}\left[1\{Z \in A\}\exp(Z)\right], \tag{3.28}$$

where (3.27) follows from (3.10). Hence, we are free to choose $\frac{\mathrm{d}P_W}{\mathrm{d}P_Z}(a) = \exp(a)$ for all $a \in \mathbb{R}$. If $\Pi(X\|Y) = 1 = \Pi(Y\|X)$, i.e., $P_X \ll\gg P_Y$, then $\mathcal{S}_{W\|Z} \cup \mathcal{S}_{Z\|W} = \mathbb{R}$ and it is immaterial how to define $\imath_{W\|Z}(\infty)$ and $\imath_{W\|Z}(-\infty)$. If $\Pi(X\|Y) < 1 = \Pi(Y\|X)$, then Item 17 implies

$$\mathbb{P}[W = \infty] > 0 = \mathbb{P}[Y = \infty], \tag{3.29}$$

$$\mathbb{P}[W = -\infty] = 0 = \mathbb{P}[Y = -\infty], \tag{3.30}$$

so $\imath_{W\|Z}(\infty) = \infty$ and it is immaterial how to define $\imath_{W\|Z}(-\infty)$. The same reasoning shows that if $\Pi(X\|Y) = 1 > \Pi(Y\|X)$, then $\imath_{W\|Z}(-\infty) = -\infty$ and it is immaterial how to define $\imath_{W\|Z}(\infty)$. If $\Pi(X\|Y) < 1$ and $\Pi(Y\|X) < 1$, then Item 17 gives

$$\mathbb{P}[W = \infty] > 0 = \mathbb{P}[Y = \infty], \tag{3.31}$$

$$\mathbb{P}[W = -\infty] = 0 < \mathbb{P}[Y = -\infty], \tag{3.32}$$

which implies that $\imath_{W\|Z}(\infty) = \infty$ and $\imath_{W\|Z}(-\infty) = -\infty$. □

21. If $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$, $D(P \,\|\, Q)$ stands for the *relative entropy* (or Kullback-Leibler divergence [2]), which, with the convention in Item 16, satisfies

$$\mathbb{E}[\imath_{X\|Y}(X)] = D(X \,\|\, Y), \tag{3.33}$$

$$\mathbb{E}[\imath_{X\|Y}(Y)] = -D(Y \,\|\, X). \tag{3.34}$$

The binary relative entropy function is the continuous extension to the domain $[0, 1]^2$ of the function $d(p \,\|\, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$. The data processing lemma for relative entropy implies

$$D(P \,\|\, Q) \geq \max_{A \in \mathscr{F}} d\left(P(A) \,\|\, Q(A)\right) \geq \max_{\substack{A \in \mathscr{F}: \\ P(A) = 1}} d\left(1 \,\|\, Q(A)\right) = \log \frac{1}{\Pi(Q \,\|\, P)}. \tag{3.35}$$

22. If $P_X \ll P_Y$, the change of measure formula (2.11) implies

$$\mathbb{E}\left[\exp(\imath_{X\|Y}(Y))\right] = 1. \tag{3.36}$$

Without assuming $P_X \ll P_Y$, we have

$$\mathbb{E}\left[\exp(\imath_{X\|Y}(Y))\right] = \Pi(X\|Y), \tag{3.37}$$

$$\mathbb{E}\left[\exp(-\imath_{X\|Y}(X))\right] = \Pi(Y\|X). \tag{3.38}$$

To verify (3.37) we let $f(a) = 1$ in (3.11) and recall (3.6). Swapping $X \leftrightarrow Y$ in (3.37) yields (3.38) in light of (3.3). The χ^2*-divergence* introduced by Pearson in [16] is

$$\chi^2(X \,\|\, Y) = \mathsf{Var}\left[\exp(\imath_{X\|Y}(Y))\right] \tag{3.39}$$

$$= \mathbb{E}\left[\exp\left(\imath_{X\|Y}(X)\right)\right] - 1, \tag{3.40}$$

where (3.40) holds if $P_X \ll P_Y$; otherwise, $\chi^2(X \,\|\, Y) = \infty$.

23. For $(P_X, P_Y) \in \mathscr{P}^2_{\mathcal{A}}$ and $\alpha \in (0, 1) \cup (1, \infty)$, the α-order *Rényi divergence* [13] is

$$D_\alpha(X \,\|\, Y) = \frac{1}{\alpha - 1} \log \mathbb{E}\left[\exp\left((\alpha - 1)\, \imath_{X\|Y}(X)\right)\right] \tag{3.41}$$

$$= \frac{1}{\alpha - 1} \log \mathbb{E}\left[\exp\left(\alpha\, \imath_{X\|Y}(Y)\right)\right], \tag{3.42}$$

where (3.42) holds for $\alpha \in (0, 1)$. Equivalently, if $Z \sim R$, with R a probability measure that dominates $\{P, Q\}$, then

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log \mathbb{E}\left[\exp\left(\alpha\, \imath_{P\|R}(Z) + (1 - \alpha)\, \imath_{Q\|R}(Z)\right)\right], \tag{3.43}$$

which, using the generalized relative information (3.2), also holds without requiring that $R \gg \{P, Q\}$. In addition, define

$$D_1(X \,\|\, Y) = D(X \,\|\, Y), \tag{3.44}$$

$$D_\infty(X\,\|\,Y) = \inf\{v \in \mathbb{R} \colon \mathbb{P}[\imath_{X\|Y}(X) \le v] = 1\} = \lim_{\alpha\to\infty} D_\alpha(X\,\|\,Y). \tag{3.45}$$

Along with (3.10) specialized to $f(a) = \exp\left((\alpha-1)\imath_{X\|Y}(a)\right)$, the skew-symmetry of relative information (3.3) results in the skew-symmetry of Rényi divergence

$$(1-\alpha)D_\alpha(X\,\|\,Y) = \alpha\, D_{1-\alpha}(Y\,\|\,X), \quad \alpha \in (0,1). \tag{3.46}$$

The coefficients of absolute discontinuity can be obtained from the Rényi divergence by means of

$$\lim_{\alpha\downarrow 0} D_\alpha(X\,\|\,Y) = \log\frac{1}{\Pi(Y\,\|\,X)}, \tag{3.47}$$

$$\lim_{\alpha\downarrow 0} \alpha\, D_{1-\alpha}(X\,\|\,Y) = \log\frac{1}{\Pi(X\,\|\,Y)}, \tag{3.48}$$

where (3.48) follows from (3.46) and (3.47). To show (3.47), note that

$$\exp\left(\alpha\, z - z\right) \le (1-\alpha)\exp(-z) + \alpha, \quad (\alpha, z) \in [0,1]\times(-\infty,+\infty], \tag{3.49}$$

so the dominated convergence theorem implies that

$$\lim_{\alpha\downarrow 0} \mathbb{E}\left[\exp\left((\alpha-1)\imath_{X\|Y}(X)\right)\right] = \mathbb{E}\left[\exp\left(-\imath_{X\|Y}(X)\right)\right], \tag{3.50}$$

which is equivalent to (3.47) in view of (3.38) and (3.41).

24. Section 8 shows a new operational role for the *Bhattacharyya distance* [29],

$$B(P\,\|\,Q) = \tfrac{1}{2}D_{\frac{1}{2}}(P\,\|\,Q) = \log\frac{1}{\int\sqrt{p\,q}\,\mathrm{d}\mu}, \tag{3.51}$$

where $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$ and $\{P, Q\} \ll \mu$.

25. A couple of properties of Rényi divergence used in the sequel are (e.g., [30]):
    (a) If $\alpha \ge 1$, then

$$P \not\ll Q \Longrightarrow D_\alpha(P\,\|\,Q) = \infty. \tag{3.52}$$

    (b) The following circular implications hold:

$$\begin{aligned} &\exists\alpha_0 \in (0,1) \text{ s.t. } D_{\alpha_0}(P\,\|\,Q) = \infty \\ &\qquad\Downarrow \qquad\qquad\qquad (3.53)\\ &\qquad P \perp Q \\ &\qquad\Downarrow \qquad\qquad\qquad (3.54)\\ &D_\alpha(P\,\|\,Q) = \infty,\ \alpha \in (0,\infty]. \end{aligned}$$

26. Given $P_X \in \mathscr{P}_{\mathcal{A}}$ and a random transformation $P_{Y|X}\colon (\mathcal{A},\mathscr{F}) \to (\mathcal{B},\mathscr{G})$, the following special case of relative information is known as the *information density*,

$$\imath_{X;Y}(x;y) = \imath_{P_{XY}\|P_X\otimes P_Y}(x,y) = \imath_{P_{Y|X=x}\|P_Y}(y), \tag{3.55}$$

where $P_X \to P_{Y|X} \to P_Y$. We use the same notation in non-Bayesian settings in which P_X need not be defined and $P_Y \in \mathscr{P}_{\mathcal{B}}$ on the rightmost term in (3.55) is an arbitrary unconditional probability measure. For future use, we observe that information density satisfies the chain rule

$$\imath_{XZ;Y}(a, c; b) = \imath_{X;Y}(a; b) + \imath_{Y;Z|X}(b; c \,|\, a). \tag{3.56}$$

Note that *mutual information* is $I(X; Y) = \mathbb{E}[\imath_{X;Y}(X; Y)]$, with $(X, Y) \sim P_X P_{Y|X}$.

27. Following [31], whenever $P_{XY} \ll P_X \otimes P_Y$, the dependence between X and Y is said to be *regular*. The following result gives sufficient conditions for regularity.

Lemma 6. *Fix $P_X \in \mathscr{P}_{\mathcal{A}}$ and a random transformation $P_{Y|X} \colon (\mathcal{A}, \mathscr{F}) \to (\mathcal{B}, \mathscr{G})$. With $P_X \to P_{Y|X} \to P_Y$, the following hold.*

$$P_{XY} \ll P_X \otimes P_Y$$

$$\Uparrow \tag{3.57}$$

$$\exists A_0 \in \mathscr{F} \colon\ P_X(A_0) = 1 \textit{ and } \{P_{Y|X=x},\ x \in A_0\} \textit{ is dominated by } P_Y$$

$$\Uparrow \tag{3.58}$$

$$\{P_{Y|X=x},\ \ x \in \mathcal{A}\}\ \ \textit{is dominated.}$$

Proof. To show (3.57) by contraposition, let $D \in \mathscr{F} \otimes \mathscr{G}$ be such that

$$(P_X \otimes P_Y)(D) = 0 < P_{XY}(D), \tag{3.59}$$

and denote

$$f(x, y) = 1\{(x, y) \in D\}, \tag{3.60}$$

$$D_x = \{y \in \mathcal{B} \colon\ (x, y) \in D\} \in \mathscr{G}. \tag{3.61}$$

The function $P_{Y|X=x}(D_x) = \mathbb{E}[f(X, Y)|X = x]$ is Borel $\mathscr{F}$-measurable with mean

$$\int P_{Y|X=x}(D_x)\, \mathrm{d}P_X(x) = P_{XY}(D) > 0. \tag{3.62}$$

Therefore, there exists $A_1 \in \mathscr{F}$ with $P_X(A_1) > 0$ and $P_{Y|X=x}(D_x) > 0$ for all $x \in A_1$. Likewise, the expectation of the Borel $\mathscr{F}$-measurable $P_Y(D_x) = \mathbb{E}[f(x, Y)]$, with $Y \sim P_Y$ is simply

$$\int P_Y(D_x)\, \mathrm{d}P_X(x) = (P_X \otimes P_Y)(D) = 0. \tag{3.63}$$

Consequently, there exists $A_2 \in \mathscr{F}$ with $P_X(A_2) = 1$ and $P_Y(D_x) = 0$ for all $x \in A_2$. We conclude that $P_Y(D_x) = 0 < P_{Y|X=x}(D_x)$ and, thus, $P_{Y|X=x} \not\ll P_Y$, for all $x \in A_1 \cap A_2$ and $P_X(A_1 \cap A_2) > 0$.

To show (3.58), we assume without loss of generality that $\{P_{Y|X=x}\, x \in \mathcal{A}\} \cup \{P_Y\}$ is dominated by $Q_Y \in \mathscr{P}_{\mathcal{A}}$. Denote by $p_{Y|X=x}$ and p_Y the corresponding densities of $P_{Y|X=x}$ and P_Y with respect to Q_Y, and define the Borel $(\mathscr{F} \otimes \mathscr{G})$-measurable function $f(x, y) = 1\{p_Y(y) = 0\}\, p_{Y|X=x}(y)$. If $(X, Y, \widehat{Y}) \sim P_{XY} \otimes Q_Y$, then

$$0 = \mathbb{P}\left[p_Y(Y) = 0\right] = \mathbb{E}[f(X, \widehat{Y})] \tag{3.64}$$

$$= \mathbb{E}\left[f_1(X)\right], \tag{3.65}$$

where $f_1(x) = \mathbb{E}[f(x, \widehat{Y})] = \int 1\{p_Y(y) = 0\}\, \mathrm{d}P_{Y|X=x}(y)$ and (3.65) follows from Fubini's theorem. The proof of (3.58) is complete since $f_1(x) = 0$ if and only if $P_{Y|X=x} \ll P_Y$. □

4. Relative information spectra

In Items 21 and 22, we saw that the expectations of the random variables $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$, as well as their exponentials, are well-known quantities in probability theory. We now consider the cumulative distribution functions of those $[-\infty, +\infty]$-valued random variables, which, unlike relative entropy, χ^2-divergence, or total variation distance, capture everything that serves to distinguish the probability measures P_X and P_Y.

28. **Definition 1.** *The* relative information spectra *of probability measures* $(P_X, P_Y) \in \mathscr{P}_{\mathcal{A}}^2$ *are the cumulative distribution functions of the relative information evaluated at X and Y, respectively,*

$$\mathbb{F}_{X\|Y}(\alpha) = \mathbb{P}\left[\imath_{X\|Y}(X) \le \alpha\right], \quad \alpha \in \mathbb{R}, \tag{4.1}$$

$$\overline{\mathbb{F}}_{X\|Y}(\alpha) = \mathbb{P}\left[\imath_{X\|Y}(Y) \le \alpha\right], \quad \alpha \in \mathbb{R}. \tag{4.2}$$

The arguments of $\mathbb{F}_{X\|Y}$ and $\overline{\mathbb{F}}_{X\|Y}$ have units inherited by the units of the relative information. In [22], the relative information spectra are referred to as *divergence spectra*.

29. If X and Y are discrete random variables, so are $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$. If X and Y are absolutely continuous random variables with probability density functions f_X and f_Y, then

$$\mathbb{F}_{X\|Y}(\alpha) = \mathbb{P}\left[f_X(X) \le \exp(\alpha)\, f_Y(X)\right], \tag{4.3}$$

$$\overline{\mathbb{F}}_{X\|Y}(\alpha) = \mathbb{P}\left[f_X(Y) \le \exp(\alpha)\, f_Y(Y)\right], \tag{4.4}$$

which need not be continuous.

30. Due to (3.6)–(3.9),

$$\lim_{\alpha\to\infty} \mathbb{F}_{X\|Y}(\alpha) = \Pi(X\,\|\,Y), \tag{4.5}$$

$$\lim_{\alpha\to-\infty} \mathbb{F}_{X\|Y}(\alpha) = 0, \tag{4.6}$$

$$\lim_{\alpha\to\infty} \overline{\mathbb{F}}_{X\|Y}(\alpha) = 1, \tag{4.7}$$

$$\lim_{\alpha\to-\infty} \overline{\mathbb{F}}_{X\|Y}(\alpha) = 1 - \Pi(Y\,\|\,X). \tag{4.8}$$

Although (4.5) is less than 1 if $P_X \not\ll P_Y$ and (4.8) is positive if $P_Y \not\ll P_X$, the relative information spectra are monotonically increasing and right-continuous.

31. As a result of the skew-symmetry (3.3) of the relative information,

$$\mathbb{F}_{Y\|X}(-\alpha) = 1 - \lim_{x\uparrow\alpha} \overline{\mathbb{F}}_{X\|Y}(x) = \mathbb{P}\left[\imath_{X\|Y}(Y) \ge \alpha\right], \quad \alpha \in \mathbb{R}, \tag{4.9}$$

$$\overline{\mathbb{F}}_{Y\|X}(-\alpha) = 1 - \lim_{x\uparrow\alpha} \mathbb{F}_{X\|Y}(x) = \mathbb{P}\left[\imath_{X\|Y}(X) \ge \alpha\right], \quad \alpha \in \mathbb{R}. \tag{4.10}$$

32. If $P_X = P_Y$, then $\mathbb{F}_{X\|Y}(\alpha) = \overline{\mathbb{F}}_{X\|Y}(\alpha) = 1\{\alpha \ge 0\}$.
If $P_X \perp P_Y$, then $\mathbb{F}_{X\|Y}(\alpha) = 0$ and $\overline{\mathbb{F}}_{X\|Y}(\alpha) = 1$ for all $\alpha \in \mathbb{R}$.

Lemma 7. *Let* $(P_X, P_Y) \in \mathscr{P}_{\mathcal{A}}^2$. *Then,*

(a) $P_X \neq P_Y \Longleftrightarrow \mathbb{F}_{X\|Y}(0) < 1$.

(b) If, in addition, $P_X \ll P_Y$, then $P_X \neq P_Y \iff \overline{\mathbb{F}}_{X\|Y}(0) < 1$.

Proof.

(a) $\mathbb{F}_{X\|Y}(0) = 1 \implies 0 = \mathbb{E}[[\imath_{X\|Y}(X)]^+] \geq D(X\,\|\,Y) \implies P_X = P_Y$.

(b) For any $\alpha > 0$,

$$\begin{aligned}1 - \overline{\mathbb{F}}_{X\|Y}(0) &= \mathbb{P}\left[\imath_{Y\|X}(Y) < 0\right] && (4.11)\\ &= \mathbb{E}\left[1\{\imath_{Y\|X}(X) < 0\}\,\exp(\imath_{Y\|X}(X))\right] && (4.12)\\ &\geq \mathbb{E}\left[1\{-\alpha < \imath_{Y\|X}(X) < 0\}\,\exp(\imath_{Y\|X}(X))\right] && (4.13)\\ &\geq \exp(-\alpha)\,\mathbb{P}\left[-\alpha < \imath_{Y\|X}(X) < 0\right] && (4.14)\\ &= \exp(-\alpha)\,\mathbb{P}\left[0 < \imath_{X\|Y}(X) < \alpha\right], && (4.15)\end{aligned}$$

where

- (4.11) $\Longleftarrow$ (4.10) with $X \leftrightarrow Y$.
- (4.12) $\Longleftarrow$ (3.12) with $f(a) = 1\{\imath_{Y\|X}(a) < 0\}\,\exp(-\imath_{X\|Y}(a))$.
- (4.15) $\Longleftarrow$ (3.3).

On account of (a), (2.5), and (4.5), $\lim_{\gamma\to\infty} \mathbb{F}_{X\|Y}(\gamma) = 1 > \mathbb{F}_{X\|Y}(0)$. Therefore, there must exist $\alpha > 0$ such that $\mathbb{P}[0 < \imath_{X\|Y}(X) < \alpha] > 0$.

An example of $P_X \neq P_Y$ with $\overline{\mathbb{F}}_{X\|Y}(0) = 1$ is $P_X = [\frac{1}{2}\ \ \frac{1}{2}]$, $P_Y = [1\ \ 0]$. □

33. *Example:* If $X \sim \mathcal{N}\big(\mu, \sigma^2\big)$, $Y \sim \mathcal{N}\big(0, \sigma^2\big)$, and $\mathrm{Q}(t) = \int_t^\infty \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\mathrm{d}x$, then, with α in nats,

$$\mathbb{F}_{X\|Y}(\alpha) = \mathrm{Q}\left(\frac{\mu}{2\sigma} - \frac{\alpha\,\sigma}{\mu}\right) \text{ and } \overline{\mathbb{F}}_{X\|Y}(\alpha) = \mathrm{Q}\left(-\frac{\mu}{2\sigma} - \frac{\alpha\,\sigma}{\mu}\right). \tag{4.16}$$

34. *Example:* Let $\xi = \frac{1}{2}\log_{\mathrm{e}} \frac{\sigma_X^2}{\sigma_Y^2}$, with $\sigma_X^2 > \sigma_Y^2 > 0$, $X \sim \mathcal{N}\big(0, \sigma_X^2\big)$, and $Y \sim \mathcal{N}\big(0, \sigma_Y^2\big)$. If $\alpha > -\xi$, then

$$\mathbb{F}_{X\|Y}(\alpha) = 1 - 2\,\mathrm{Q}\left(\sqrt{\frac{2\,\sigma_Y^2(\alpha+\xi)}{\sigma_X^2 - \sigma_Y^2}}\right), \tag{4.17}$$

$$\overline{\mathbb{F}}_{X\|Y}(\alpha) = 1 - 2\,\mathrm{Q}\left(\sqrt{\frac{2\,\sigma_X^2(\alpha+\xi)}{\sigma_X^2 - \sigma_Y^2}}\right), \tag{4.18}$$

while $\mathbb{F}_{X\|Y}(\alpha) = \overline{\mathbb{F}}_{X\|Y}(\alpha) = 0$ if $\alpha \leq -\xi$.

35. *Example:* [24] Suppose that V is standard Cauchy, $\lambda_0\lambda_1 \neq 0$, $X_1 = \lambda_1 V + \mu_1$, and $X_0 = \lambda_0 V + \mu_0$. Then, $\overline{\mathbb{F}}_{X_1\|X_0}(\alpha) = 1 - \mathbb{F}_{X_1\|X_0}(-\alpha)$ and

$$\mathbb{F}_{X_1\|X_0}(\log\beta) = \begin{cases} 1, & \zeta + \sqrt{\zeta^2 - 1} \leq \beta; \\ \frac{1}{2} + \frac{1}{\pi}\arcsin\frac{\beta-\zeta}{\sqrt{\zeta^2-1}}, & \zeta - \sqrt{\zeta^2-1} < \beta < \zeta + \sqrt{\zeta^2-1}; \\ 0, & 0 < \beta \leq \zeta - \sqrt{\zeta^2-1}, \end{cases} \tag{4.19}$$

$$\text{with } \zeta = \frac{\lambda_1^2 + \lambda_0^2 + (\mu_1-\mu_0)^2}{2|\lambda_0\lambda_1|} \geq 1. \tag{4.20}$$

36. *Example:* Suppose that U is uniform on $[0, 1]$. Define the probability measures on $(\mathbb{R}, \mathscr{B})$,

$$P_1 = \tfrac{1}{2}\delta_2 + \tfrac{1}{2}P_{2U}, \quad \text{and} \quad P_{\mathtt{0}} = \tfrac{1}{3}\delta_1 + \tfrac{1}{3}\delta_2 + \tfrac{1}{3}P_{3U+1}. \tag{4.21}$$

Then, $\Pi(P_1 \,\|\, P_{\mathtt{0}}) = \frac{3}{4}$, $\Pi(P_{\mathtt{0}} \,\|\, P_1) = \frac{4}{9}$, and

$$\mathbb{F}_{P_1\|P_{\mathtt{0}}}(\alpha) = \begin{cases} \frac{3}{4}, & \alpha \geq 2\log\frac{3}{2}; \\ \frac{1}{2}, & \log\frac{3}{2} \leq \alpha < 2\log\frac{3}{2}; \\ 0, & \alpha < \log\frac{3}{2}, \end{cases} \quad \overline{\mathbb{F}}_{P_1\|P_{\mathtt{0}}}(\alpha) = \begin{cases} 1, & \alpha \geq 2\log\frac{3}{2}; \\ \frac{8}{9}, & \log\frac{3}{2} \leq \alpha < 2\log\frac{3}{2}; \\ \frac{5}{9}, & \alpha < \log\frac{3}{2}. \end{cases} \tag{4.22}$$

37. A key aspect of the relative information spectra is that $\mathbb{F}_{P_1\|P_{\mathtt{0}}}$ and $\overline{\mathbb{F}}_{P_1\|P_{\mathtt{0}}}$ determine each other through the following result.

Theorem 1. *Fix arbitrary* $(P_X, P_Y) \in \mathscr{P}^2_{\mathcal{A}}$.

(a) For all $\beta > 0$,

$$\overline{\mathbb{F}}_{X\|Y}(\log\beta) = 1 - \int_0^{\frac{1}{\beta}} \left(\mathbb{F}_{X\|Y}\left(\log\frac{1}{t}\right) - \mathbb{F}_{X\|Y}(\log\beta)\right) \mathrm{d}t. \tag{4.23}$$

(b) For all $\beta > 0$,

$$\mathbb{F}_{X\|Y}(\log\beta) = \int_0^{\beta} \left(\overline{\mathbb{F}}_{X\|Y}(\log\beta) - \overline{\mathbb{F}}_{X\|Y}(\log t)\right) \mathrm{d}t. \tag{4.24}$$

(c) For all $\beta > 0$,

$$\mathbb{P}\left[\imath_{X\|Y}(X) = \log\beta\right] = \beta\,\mathbb{P}\left[\imath_{X\|Y}(Y) = \log\beta\right]. \tag{4.25}$$

(d) $\imath_{X\|Y}(X) \in (-\infty, +\infty]$ *is discrete if and only if* $\imath_{X\|Y}(Y) \in [-\infty, +\infty)$ *is discrete.* $\imath_{X\|Y}(X)$ *is absolutely continuous, except for a possible mass at* $+\infty$, *if and only if* $\imath_{X\|Y}(Y)$ *is absolutely continuous, except for a possible mass at* $-\infty$. *Then, the density functions* $f_{X\|Y}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{F}_{X\|Y}(x)$ *and* $\bar{f}_{X\|Y}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\overline{\mathbb{F}}_{X\|Y}(x)$ *satisfy*

$$f_{X\|Y}(\log t) = t\,\bar{f}_{X\|Y}(\log t), \quad t > 0. \tag{4.26}$$

(e)

$$\begin{aligned} \Pi(Y\|X) &= \mathbb{P}[\imath_{X\|Y}(Y) > -\infty] && (4.27) \\ &= \mathbb{E}\left[\exp(-\imath_{X\|Y}(X))\right] && (4.28) \\ &= \int_0^{\infty} \mathbb{F}_{X\|Y}\left(\log\frac{1}{\beta}\right) \mathrm{d}\beta && (4.29) \\ &= \int_{-\infty}^{\infty} \exp(-t)\,\mathrm{d}\mathbb{F}_{X\|Y}(t). && (4.30) \end{aligned}$$

$$\Pi(X\|Y) = \lim_{\alpha\to\infty} \mathbb{F}_{X\|Y}(\alpha) \tag{4.31}$$

$$= \mathbb{E}\left[\exp(\imath_{X\|Y}(Y))\right] \tag{4.32}$$

$$= \int_0^\infty \left(1 - \bar{\mathbb{F}}_{X\|Y}(\log\beta)\right) \mathrm{d}\beta \tag{4.33}$$

$$= \int_{-\infty}^\infty \exp(t)\, \mathrm{d}\bar{\mathbb{F}}_{X\|Y}(t). \tag{4.34}$$

(f) If $g \colon \mathbb{R} \to [0, \infty)$*, then*

$$\int_{-\infty}^\infty g(t)\ \exp(t)\, \mathrm{d}\bar{\mathbb{F}}_{X\|Y}(t) = \int_{-\infty}^\infty g(t)\, \mathrm{d}\mathbb{F}_{X\|Y}(t), \tag{4.35}$$

$$\int_{-\infty}^\infty g(t)\ \exp(-t)\, \mathrm{d}\mathbb{F}_{X\|Y}(t) = \int_{-\infty}^\infty g(t)\, \mathrm{d}\bar{\mathbb{F}}_{X\|Y}(t). \tag{4.36}$$

(g) The cumulant generating functions of $\imath_{X\|Y}(X)$ *and* $\imath_{X\|Y}(Y)$ *(nats) satisfy*

$$\Lambda_{\imath_{X\|Y}(X)}(t) = \Lambda_{\imath_{X\|Y}(Y)}(t+1), \quad \begin{cases} t \in \mathbb{R}, & P_X \ll P_Y \ll P_X; \\ t \in (-1, \infty), & P_X \ll P_Y \not\ll P_X; \\ t \in (-\infty, 0), & P_X \not\ll P_Y \ll P_X; \\ t \in (-1, 0), & P_X \not\ll P_Y \not\ll P_X, \end{cases} \tag{4.37}$$

and $\Lambda_{\imath_{X\|Y}(X)}(-1) = \log_{\mathrm{e}} \Pi(Y\|X)$ *and* $\Lambda_{\imath_{X\|Y}(Y)}(1) = \log_{\mathrm{e}} \Pi(X\|Y)$.

Proof.

(a) Fix $\alpha \in \mathbb{R}$ and let P_Z dominate $\{P_X, P_Y\}$. Then,

$$\mathbb{E}\left[\exp\left(-\imath_{X\|Y}(X)\right)\ 1\left\{\alpha < \imath_{X\|Y}(X)\right\}\right]$$

$$= \mathbb{E}\left[\exp\left(-\imath_{X\|Y}(X)\right)\ 1\left\{\alpha < \imath_{X\|Y}(X) < \infty\right\}\right] \tag{4.38}$$

$$= \mathbb{E}\left[\exp\left(\frac{\imath_{Y\|Z}(X)}{\imath_{X\|Z}(X)}\right)\ 1\left\{\alpha < \imath_{X\|Y}(X) < \infty\right\}\right] \tag{4.39}$$

$$= \mathbb{E}\left[\exp\left(\imath_{Y\|Z}(Z)\right)\ 1\left\{\alpha < \imath_{X\|Y}(Z) < \infty\right\}\right] \tag{4.40}$$

$$= \mathbb{P}\left[\alpha < \imath_{X\|Y}(Y) < \infty\right] \tag{4.41}$$

$$= 1 - \bar{\mathbb{F}}_{X\|Y}(\alpha), \tag{4.42}$$

where

- (4.38) $\Longleftarrow$ $\exp(-\infty) = 0$.
- (4.39) $\Longleftarrow$ (3.2) and the random variable in the expectation in the left side can be positive only if $X \in \mathcal{S}_{X\|Y} \cap \mathcal{S}_{Y\|X}$.
- (4.40) and (4.41) $\Longleftarrow$ change of measure (2.11).
- (4.42) $\Longleftarrow$ (3.8).

Then, we have

$$\mathbb{F}_{X\|Y}(\alpha) + \exp(\alpha)\left(1 - \bar{\mathbb{F}}_{X\|Y}(\alpha)\right) = \mathbb{E}\left[\exp\left(-[\imath_{X\|Y}(X) - \alpha]^+\right)\right] \tag{4.43}$$

$$= \int_0^1 \mathbb{P}\left[\exp\left(-[\imath_{X\|Y}(X) - \alpha]^+\right) \geq \tau\right] \mathrm{d}\tau \tag{4.44}$$

$$= \int_0^1 \mathbb{F}_{X\|Y}\left(\alpha + \log\frac{1}{\tau}\right) \mathrm{d}\tau, \tag{4.45}$$

where

- (4.43) $\Longleftarrow$ (4.38)–(4.42).
- (4.44) $\Longleftarrow$ $\mathbb{E}[T] = \int_0^1 \mathbb{P}[T \geq \tau]\, \mathrm{d}\tau$ if $T \in [0,1]$.

Rearranging the outer terms in (4.43)–(4.45) with $\alpha = \log\beta$ and changing the integration variable to $t = \frac{\tau}{\beta}$ yields (4.23).

(b) Denote $V = \exp(\imath_{X\|Y}(Y))$. By change of measure on $\mathcal{S}_{X\|Y} \cap \mathcal{S}_{Y\|X}$, we obtain

$$\mathbb{F}_{X\|Y}(\log\beta) = \mathbb{E}\left[1\left\{\exp(\imath_{X\|Y}(X) \leq \beta\right\}\, 1\left\{X \in \mathcal{S}_{X\|Y} \cap \mathcal{S}_{Y\|X}\right\}\right] \tag{4.46}$$

$$= \mathbb{E}\left[V\, 1\left\{0 < V \leq \beta\right\}\right] \tag{4.47}$$

$$= \int_0^\infty \mathbb{P}\left[1\left\{0 < V \leq \beta\right\} V > t\right]\, \mathrm{d}t \tag{4.48}$$

$$= \int_0^\beta \mathbb{P}\left[t < V \leq \beta\right]\, \mathrm{d}t \tag{4.49}$$

$$= \int_0^\beta \left(\overline{\mathbb{F}}_{X\|Y}(\log\beta) - \overline{\mathbb{F}}_{X\|Y}(\log t)\right)\, \mathrm{d}t. \tag{4.50}$$

(c) $\Longleftarrow$ (3.10) with $f(a) = 1\{\imath_{X\|Y}(a) = \log\beta\}$.

(d) Lemma 5 implies that, when restricted to $\mathbb{R}$, the probability measures of W and Z are mutually absolutely continuous; therefore, one is discrete [resp., absolutely continuous] if and only if the other one is discrete [resp., absolutely continuous]. Differentiating (4.23) with respect to t yields (4.26).

(e)
- (4.27) and (4.31) are (3.9) and (3.6), respectively.
- (4.28) and (4.32) are (3.38) and (3.37), respectively.
- (4.29) $\Longleftarrow$ (4.8) and (4.23).
- (4.33) $\Longleftarrow$ (4.5) and (4.24).
- (4.30) and (4.34) are (4.28) and (4.32), respectively, since those expectations are unchanged when restricted to $\imath_{X\|Y}(X) \in \mathbb{R}$ and $\imath_{X\|Y}(Y) \in \mathbb{R}$.

(f) $\Longleftarrow$ Lemma 3 with $f(a) = g\left(\imath_{X\|Y}(a)\right)$.
Note that the left sides of (3.10) and (3.13) are unchanged if the random variables inside the expectations are multiplied by $1\{\imath_{X\|Y}(Y) \in \mathbb{R}\}$ and $1\{\imath_{X\|Y}(X) \in \mathbb{R}\}$, respectively.

(g) The formulas for $\Lambda_{\imath_{X\|Y}(X)}(-1)$ and $\Lambda_{\imath_{X\|Y}(Y)}(1)$ follow from (4.28) and (4.32), respectively. In addition to $\Lambda_{\imath_{X\|Y}(X)}(0) = 0 = \Lambda_{\imath_{X\|Y}(Y)}(0)$, the following expressions ($\infty \cdot 0 = 0$) for the moment generating functions at $t \notin \{0, -1\}$ yield (4.37):

$$\mathrm{M}_{\imath_{X\|Y}(X)}(t) = \int_{\mathcal{S}_{X\|Y} \cap \mathcal{S}_{Y\|X}} \left(\frac{p_X(\omega)}{p_Y(\omega)}\right)^t \mathrm{d}P_X(\omega) + \infty \cdot 1\{P_X \not\ll P_Y\} \cdot 1\{t > 0\}, \tag{4.51}$$

$$\mathrm{M}_{\imath_{X\|Y}(Y)}(t+1) = \int_{\mathcal{S}_{X\|Y} \cap \mathcal{S}_{Y\|X}} \left(\frac{p_X(\omega)}{p_Y(\omega)}\right)^t \mathrm{d}P_X(\omega) + \infty \cdot 1\{P_Y \not\ll P_X\} \cdot 1\{t < -1\}. \tag{4.52}$$

□

38. In this item and the next, we upper bound $\mathbb{F}_{X\|Y}$ in terms of $\overline{\mathbb{F}}_{X\|Y}$.

    Lemma 8. *For $\beta > 0$,*

$$\mathbb{F}_{X\|Y}(\log\beta) \leq \min\{1,\beta\}, \tag{4.53}$$

$$\mathbb{F}_{X\|Y}(\log\beta) \leq \inf_{B\in\mathscr{F}} \{P_X(B^c) + \beta\, P_Y(B)\}\,, \tag{4.54}$$

$$2\,\mathbb{F}_{X\|Y}(\log\beta) \leq \beta\,\overline{\mathbb{F}}_{X\|Y}(\log\beta) + 1, \tag{4.55}$$

$$\mathbb{F}_{X\|Y}(\log\beta) \leq \beta\,\overline{\mathbb{F}}_{X\|Y}(\log\beta) + \beta\,\Pi(Y\|X) - \beta, \tag{4.56}$$

$$\mathbb{F}_{X\|Y}(\log\beta) \leq \beta\,\overline{\mathbb{F}}_{X\|Y}(\log\beta) + \Pi(X\|Y) - \beta, \tag{4.57}$$

$$\mathbb{F}_{X\|Y}(\log\beta) < \overline{\mathbb{F}}_{X\|Y}(\log\beta) + \mathrm{e}^{-\frac{1}{\beta}} - 1 + \Pi(Y\,\|\,X). \tag{4.58}$$

    Proof. Let $L_\beta = \{a \in \mathscr{A}\colon\ \imath_{X\|Y}(a) \leq \log\beta\}$.

    - (4.53) $\Longleftarrow$ (3.11) with $f(a) = 1\{a \in L_\beta\}$.
    - (4.54) is Lemma 4.1.2 in [22]. $P_X(L_\beta) \leq P_X(B^c) + P_X(L_\beta \cap B) \leq P_X(B^c) + \beta\, P_Y(B)$, where the second inequality follows from (3.18) with $g(a) = 1\{a \in B\}$.
    - (4.55) $\Longleftarrow$ (4.54) with the suboptimal choice $B = L_\beta$.
    - (4.56) $\Longleftarrow$ (4.24) and for $t > 0$, $\overline{\mathbb{F}}_{X\|Y}(\log t) \geq \lim_{\tau\to 0} \overline{\mathbb{F}}_{X\|Y}(\log\tau) = 1 - \Pi(Y\|X)$.
    - (4.53) $\Longleftarrow$ $\max\{\overline{\mathbb{F}}_{X\|Y}(\log\beta), \Pi(Y\|X)\} \leq 1$.
    - (4.57) $\Longleftarrow$ (3.6), (4.25), and (3.17) with $g(a) \leftarrow 1$.
    - (4.58) $\Longleftarrow$ $\left(1 - \frac{1}{\beta}\right)\mathbb{F}_{X\|Y}(\log\beta) \leq 1 - \frac{1}{\beta} < \mathrm{e}^{-\frac{1}{\beta}}$ and upper bound $\frac{1}{\beta}\mathbb{F}_{X\|Y}(\log\beta)$ by means of (4.56).

    □

39. The following bound is instrumental in hypothesis testing (Section 8).

    Lemma 9. *Let $(P_X, P_Y) \in \mathscr{P}^2_{\mathscr{A}}$. For any $\beta > 0$, and measurable function $g\colon \mathscr{A} \to [0,1]$,*

$$\mathbb{F}_{X\|Y}(\log\beta) - \mathbb{E}\left[g(X)\right] \leq \beta\,\overline{\mathbb{F}}_{X\|Y}(\log\beta) - \beta\,\mathbb{E}\left[g(Y)\right]. \tag{4.59}$$

    Proof. For all $a \in \mathscr{A}, \beta > 0$, and measurable function $g\colon \mathscr{A} \to [0,1]$,

$$(\beta - \exp(\imath_{X\|Y}(a)))\,(1\{\imath_{X\|Y}(a) \leq \log\beta\} - g(a)) \geq 0, \tag{4.60}$$

    because when the first factor is positive [resp., negative], then the second factor is $1 - g(a)$ [resp., $-g(a)$]. Averaging (4.60) with respect to $a \leftarrow Y$, we obtain (4.59) invoking (3.10) twice with $f(a) \leftarrow 1\{\imath_{X\|Y}(a) \leq \log\beta\}$ and $f(a) \leftarrow g(a)$, respectively, where in the second case the nonnegativity of g yields $\mathbb{E}[g(X)1\{\imath_{X\|Y}(X) \in \mathbb{R}\}] \leq \mathbb{E}[g(X)]$. □

40. **Definition 2.** $(P_1, P_{\mathbb{0}}) \in \mathscr{P}^2_{\mathscr{A}}$ *and* $(Q_1, Q_{\mathbb{0}}) \in \mathscr{P}^2_{\mathscr{B}}$ *are said to be* equivalent pairs, *denoted as* $(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}})$, *if*

$$\mathbb{F}_{P_1\|P_{\mathbb{0}}}(\alpha) = \mathbb{F}_{Q_1\|Q_{\mathbb{0}}}(\alpha), \quad \alpha \in \mathbb{R}, \tag{4.61}$$

    i.e., $\frac{\mathrm{d}P_1}{\mathrm{d}P_{\mathbb{0}}}(X_1)$ *and* $\frac{\mathrm{d}Q_1}{\mathrm{d}Q_{\mathbb{0}}}(Y_1)$ *are identically distributed when* $X_1 \sim P_1$ *and* $Y_1 \sim Q_1$.

A word of caution is that a different notion of equivalence for pairs of real-valued random variables (not pairs of probability measures) was proposed by Halmos and Savage in [3]: Suppose (X_1, Y_1, X_2, Y_2) are real-valued random variables such that $\mathbb{P}[X_i = Y_i = 0] = 0$, $i = 1, 2$; then (X_1, Y_1) and (X_2, Y_2) are equivalent in the sense of [3], if there is a fifth random variable such that $\mathbb{P}[F = 0] = 0$ and $\mathbb{P}[(X_1, Y_1) = (F \cdot X_2, F \cdot Y_2)] = 1$.

41. Definition 2 and Theorem 1 result in

$$\begin{aligned} &\{\mathbb{F}_{P_1\|P_{\mathbb{0}}}(\alpha) = \mathbb{F}_{Q_1\|Q_{\mathbb{0}}}(\alpha),\ \alpha \in \mathbb{R}\} \\ &\qquad\Updownarrow \qquad\qquad (4.62) \\ &(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}}) \\ &\qquad\Updownarrow \qquad\qquad (4.63) \\ &\left\{\overline{\mathbb{F}}_{P_1\|P_{\mathbb{0}}}(\alpha) = \overline{\mathbb{F}}_{Q_1\|Q_{\mathbb{0}}}(\alpha),\ \alpha \in \mathbb{R}\right\}. \end{aligned}$$

The remainder of the section is devoted to finding necessary and sufficient conditions for the equivalence of pairs. The relevance of such conditions will be apparent in Section 7.

42. In view of (4.9) and the fact that the relative information spectra are right-continuous, (4.62)–(4.63) imply

$$(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}}) \quad \Longleftrightarrow \quad (P_{\mathbb{0}}, P_1) \equiv (Q_{\mathbb{0}}, Q_1). \tag{4.64}$$

However, $(P_1, P_{\mathbb{0}}) \equiv (P_{\mathbb{0}}, P_1)$ is more the exception than the rule. In addition to $\mathbb{R}^n$-valued random variables that differ by a constant, one of the most notable cases satisfying this property is the Cauchy case described in Item 35.

43. **Theorem 2.** *For $(P_1, P_{\mathbb{0}}) \in \mathscr{P}^2_{\mathcal{A}}$ and $(Q_1, Q_{\mathbb{0}}) \in \mathscr{P}^2_{\mathcal{B}}$, the following circular implications hold:*

$$\begin{aligned} &\exists \alpha_0 > 0 \text{ s.t. } \{D_\alpha(P_1 \,\|\, P_{\mathbb{0}}) = D_\alpha(Q_1 \,\|\, Q_{\mathbb{0}}),\ \alpha \in (0, \alpha_0)\} \\ &\qquad\Downarrow \qquad\qquad (4.65) \\ &(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}}) \\ &\qquad\Downarrow \qquad\qquad (4.66) \\ &D_\alpha(P_1 \,\|\, P_{\mathbb{0}}) = D_\alpha(Q_1 \,\|\, Q_{\mathbb{0}}),\ \alpha \in (0, \infty]. \end{aligned}$$

Proof. If $D_\alpha(P_1 \,\|\, P_{\mathbb{0}}) = D_\alpha(Q_1 \,\|\, Q_{\mathbb{0}}) = \infty$ for some $\alpha < 1$, then (4.65) follows from Item 25-(b) since $P_1 \perp P_{\mathbb{0}}$ and $Q_1 \perp Q_{\mathbb{0}}$ implies $(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}})$. If $D_\alpha(P_1 \,\|\, P_{\mathbb{0}}) = D_\alpha(Q_1 \,\|\, Q_{\mathbb{0}}) < \infty$ for $0 < \alpha < \alpha_1 < 1$, recall from (3.42) that $(\alpha - 1) D_\alpha(X\|Y) = \Lambda_{\imath_{X\|Y}(Y)}(\alpha)$ (assuming nats for convenience). Then, Item 13 implies that the values of $D_\alpha(X\|Y)$ in a neighborhood of the origin determine the function $\overline{\mathbb{F}}_{X\|Y}$; therefore, (4.63) yields (4.65). On account of (3.41), we have

$$\exp\left((\alpha - 1)\, D_\alpha(P \,\|\, Q)\right) = \begin{cases} \displaystyle\int_0^\infty \mathbb{F}_{P\|Q}\left(\frac{\log \beta}{\alpha - 1}\right) \mathrm{d}\beta, & \alpha \in (0, 1); \\ \displaystyle\int_0^\infty \left(1 - \mathbb{F}_{P\|Q}\left(\frac{\log \beta}{\alpha - 1}\right)\right) \mathrm{d}\beta, & \alpha > 1, \end{cases} \tag{4.67}$$

which shows (4.66) for $\alpha \in (0,1) \cup (1,\infty)$. For $\alpha = 1$, we recall the definition of relative entropy, or, equivalently,

$$D(P \,\|\, Q) = \int_{-\infty}^{\infty} \left(1\{x > 0\} - \mathbb{F}_{P\|Q}(x)\right) \mathrm{d}x. \tag{4.68}$$

For $\alpha = \infty$, note that according to (3.45), $D_\infty(P \,\|\, Q) = \inf\{v \in \mathbb{R}\colon \mathbb{F}_{P\|Q}(v) = 1\}$. □

44. The following concentration bound for the relative information spectrum holds as a function of the Rényi divergence of order $\alpha > 1$: If $\delta > 0$, then

$$\mathbb{F}_{P\|Q}\left(D_\alpha(P \,\|\, Q) + \delta\right) \geq \Pi(P \,\|\, Q) - \exp\left((1-\alpha)\delta\right). \tag{4.69}$$

To verify (4.69), let $L_\beta = \{a \in \mathcal{A}\colon \imath_{P\|Q}(a) \leq \log\beta\}$, and $X \sim P$. Then,

$$\begin{aligned}
&\Pi(P \,\|\, Q) - \mathbb{F}_{P\|Q}(\log\beta) \\
&\quad= \mathbb{P}[\log\beta < \imath_{P\|Q}(X) < \infty] && (4.70)\\
&\quad= \int 1\{\log\beta < \imath_{P\|Q}(a) < \infty\} \exp\left((1-\alpha)\imath_{P\|Q}(a) + (\alpha-1)\imath_{P\|Q}(a)\right) \mathrm{d}P(a) && (4.71)\\
&\quad\leq \exp\left((1-\alpha)\left(\log\beta - D_\alpha(P \,\|\, Q)\right)\right), && (4.72)
\end{aligned}$$

on account of (3.43). Letting $\log\beta = \delta + D_\alpha(P \,\|\, Q)$ yields (4.69).

45. Let $\mathtt{F}$ be the collection of convex functions $f\colon (0,\infty) \to \mathbb{R}$. For $f \in \mathtt{F}$, the *f-divergence*, introduced and shown to satisfy the data processing principle in [15, 32, 33], can be expressed in terms of the relative information spectrum as

$$D_f(P \,\|\, Q) = \int_{-\infty}^{\infty} f\left(\exp(t)\right) \mathrm{d}\overline{\mathbb{F}}_{P\|Q}(t) + \left(1 - \Pi(Q \,\|\, P)\right) f(0) + \left(1 - \Pi(P \,\|\, Q)\right) f^\dagger(0), \tag{4.73}$$

where $f(0) = \lim_{t\downarrow 0} f(t)$ and $f^\dagger(0) = \lim_{t\downarrow 0} t\, f\left(\frac{1}{t}\right)$. Other integral representations of f-divergence as a function of the relative information spectrum, the deGroot statistical information (Item 48), and the E_γ-divergence (Item 49), can be found in [34], [35], and [36], respectively.

46. Lemmas 10 and 11 are used in Section 6 to show that the NP-divergence is not an f-divergence.

Lemma 10. *[37, (9.4)] Suppose that the convex functions $f\colon (0,\infty) \to \mathbb{R}$ and $g\colon (0,\infty) \to \mathbb{R}$ are such that $D_f(P \,\|\, Q) = D_g(P \,\|\, Q)$ for all $(P,Q) \in \mathscr{P}^2_{\mathcal{A}}$, where $|\mathcal{A}| = 2$. Then, $f(t) - g(t) = \alpha\, t - \alpha$ for some $\alpha \in \mathbb{R}$.*

Csiszár showed in [38, Theorem 1] that a discrepancy measure that satisfies the data processing inequality and the property in Lemma 11 must be an f-divergence.

Lemma 11. *Whenever $(P_1, P_{\mathbb{0}}, Q_1, Q_{\mathbb{0}}) \in \mathscr{P}^4_{\mathcal{A}}$ are such that there exists an event $\mathcal{A}_{\mathbb{0}} \in \mathscr{F}$ such that $Q_{\mathbb{0}}(\mathcal{A}_{\mathbb{0}}) = 1 = P_{\mathbb{0}}(\mathcal{A}_{\mathbb{0}})$ and $Q_1(\mathcal{A}_{\mathbb{0}}) = 0 = P_1(\mathcal{A}_{\mathbb{0}})$,*

$$D_f(\lambda\, P_1 + (1-\lambda) P_{\mathbb{0}} \,\|\, \lambda\, Q_1 + (1-\lambda) Q_{\mathbb{0}}) = \lambda\, D_f(P_1 \,\|\, Q_1) + (1-\lambda) D_f(P_{\mathbb{0}} \,\|\, Q_{\mathbb{0}}), \tag{4.74}$$

for all $\lambda \in [0,1]$ and $f \in \mathtt{F}$.

Proof. Denoting the corresponding densities with respect to a common dominating σ-finite measure μ by $p_{\mathbb{0}}$, p_1, $q_{\mathbb{0}}$, and q_1, we have $p_{\mathbb{0}}(x) = q_{\mathbb{0}}(x) = 0$ if $x \notin \mathcal{A}_{\mathbb{0}}$ and $p_1(x) = q_1(x) = 0$ if $x \in \mathcal{A}_{\mathbb{0}}$. Furthermore, we can express the densities of the mixtures by $p_\lambda = \lambda\, p_1 + (1-\lambda)p_{\mathbb{0}}$, and $q_\lambda = \lambda\, q_1 + (1-\lambda)q_{\mathbb{0}}$, respectively. Then, with the usual conventions $0 \cdot f\left(\frac{p}{0}\right) = p\, f^\dagger(0)$ if $p \geq 0$, and $f(0)\cdot 0 = f^\dagger(0)\cdot 0 = 0$,

$$
\begin{aligned}
&D_f(\lambda\, P_1 + (1-\lambda)P_{\mathbb{0}} \,\|\, \lambda\, Q_1 + (1-\lambda)Q_{\mathbb{0}}) \\
&\quad = \int_{\mathcal{A}_{\mathbb{0}}} q_\lambda\, f\left(\frac{p_\lambda}{q_\lambda}\right) \mathrm{d}\mu + \int_{\mathcal{A}_{\mathbb{0}}^{\mathrm{c}}} q_\lambda\, f\left(\frac{p_\lambda}{q_\lambda}\right) \mathrm{d}\mu && (4.75)\\
&\quad = (1-\lambda) \int_{\mathcal{A}_{\mathbb{0}}} q_{\mathbb{0}}\, f\left(\frac{p_{\mathbb{0}}}{q_{\mathbb{0}}}\right) \mathrm{d}\mu + \lambda \int_{\mathcal{A}_{\mathbb{0}}^{\mathrm{c}}} q_1\, f\left(\frac{p_1}{q_1}\right) \mathrm{d}\mu && (4.76)\\
&\quad = (1-\lambda) D_f(P_{\mathbb{0}} \,\|\, Q_{\mathbb{0}}) + \lambda\, D_f(P_1 \,\|\, Q_1). && (4.77)
\end{aligned}
$$

□

47. The convex functions

$$
f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1} \tag{4.78}
$$

result in an important special case of f-divergence known as the *Hellinger divergence* of order $\alpha \in (0,1) \cup (1,\infty)$,

$$
\begin{aligned}
\mathscr{H}_\alpha(P \,\|\, Q) = D_{f_\alpha}(P \,\|\, Q) &= \frac{1}{\alpha - 1} \left(\mathbb{E}\left[\exp\left(\alpha\, \imath_{P\|R}(Z) + (1-\alpha)\imath_{Q\|R}(Z)\right)\right] - 1\right) && (4.79)\\
&= \frac{1}{1-\alpha}\left(1 - \int p^\alpha q^{1-\alpha} \mathrm{d}\rho\right), && (4.80)
\end{aligned}
$$

which use the same notation as in (3.2) and (3.43). Furthermore, we let $\mathscr{H}_1(P \,\|\, Q) = D(P \,\|\, Q)$. The *squared Hellinger distance* is

$$
\mathscr{H}^2(P \,\|\, Q) = \tfrac{1}{2}\mathscr{H}_{\frac{1}{2}}(P \,\|\, Q) = 1 - \exp\left(-B(P \,\|\, Q)\right) = D_f(P \,\|\, Q) \leq 1, \tag{4.81}
$$

with $B(P \,\|\, Q)$ defined in Item 24, and $f(t) = 1 - \sqrt{t}$ or $f(t) = \frac{1}{2}(1 - \sqrt{t})^2$.

Theorem 3. *For $(P_1, P_{\mathbb{0}}) \in \mathscr{P}_{\mathcal{A}}^2$ and $(Q_1, Q_{\mathbb{0}}) \in \mathscr{P}_{\mathcal{B}}^2$, the following circular implications hold:*

$$
\begin{aligned}
&\exists \alpha_0 > 0 \text{ s.t. } \{\mathscr{H}_\alpha(P_1 \,\|\, P_{\mathbb{0}}) = \mathscr{H}_\alpha(Q_1 \,\|\, Q_{\mathbb{0}}),\ \alpha \in (0, \alpha_0)\} \\
&\Downarrow && (4.82)\\
&(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}}) \\
&\Downarrow && (4.83)\\
&D_f(P_1 \,\|\, P_{\mathbb{0}}) = D_f(Q_1 \,\|\, Q_{\mathbb{0}}), \quad \textit{for all } f \in \mathtt{F}.
\end{aligned}
$$

Proof.

- (4.82) $\Longleftarrow$ (4.65) because although Rényi divergence is not an f-divergence, it can be put in a one-to-one correspondence with $\mathscr{H}_\alpha(P \,\|\, Q)$ by means of

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log\left(1 + (\alpha - 1)\mathscr{H}_\alpha(P \,\|\, Q)\right), \tag{4.84}$$

in light of (3.43) and (4.79).
- (4.83) $\Longleftarrow$ (4.73).

□

48. For $p \in (0,1)$, the *deGroot statistical information* [20] is defined as the ϕ_p-divergence

$$\mathcal{I}_p(P \,\|\, Q) = D_{\phi_p}(P \,\|\, Q), \tag{4.85}$$

with the convex function $\phi_p \colon (0,\infty) \to (-1, \frac{1}{2})$,

$$\phi_p(t) = \min\{p, 1-p\} - \min\{p\,t, 1-p\} = \begin{cases} \min\{p, 1-p\} - p\,t, & 0 < t \leq \frac{1}{p} - 1; \\ -[1 - 2\,p]^+, & t > \frac{1}{p} - 1. \end{cases} \tag{4.86}$$

Theorem 4. *For* $(P_1, P_{\mathbb{0}}) \in \mathscr{P}^2_{\mathcal{A}}$ *and* $(Q_1, Q_{\mathbb{0}}) \in \mathscr{P}^2_{\mathcal{B}}$,

$$(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}}) \quad \Longleftrightarrow \quad \{\mathcal{I}_p(P_1 \,\|\, P_{\mathbb{0}}) = \mathcal{I}_p(Q_1 \,\|\, Q_{\mathbb{0}}),\ p \in (0,1)\}. \tag{4.87}$$

Proof. $\Longrightarrow$ follows from (4.83). To show $\Longleftarrow$, we use the fact that as long as f is convex and twice differentiable, the f-divergence can be expressed as [34, 39–41]

$$D_f(P \,\|\, Q) = \int_0^1 \mathcal{I}_p(P \,\|\, Q) \cdot \frac{1}{p^3} \cdot \ddot{f}\left(\frac{1-p}{p}\right) \mathrm{d}p. \tag{4.88}$$

Therefore, $\{\mathcal{I}_p(P_1 \,\|\, P_{\mathbb{0}}) = \mathcal{I}_p(Q_1 \,\|\, Q_{\mathbb{0}}),\ p \in (0,1)\} \Longrightarrow D_f(P_1 \,\|\, P_{\mathbb{0}}) = D_f(Q_1 \,\|\, Q_{\mathbb{0}})$. Since (4.78) is convex and twice differentiable, $\Longleftarrow$ in (4.87) follows from (4.82). Alternatively, we can invoke the representation of the relative information spectrum in [34, Theorem 4]:

$$\mathbb{F}_{P\|Q}\left(\log \tfrac{1-p}{p}\right) = \begin{cases} -\mathcal{I}_p(P \,\|\, Q) - (1-p)\,\dot{\mathcal{I}}_p(P \,\|\, Q) + 1, & p \in (0, \frac{1}{2}); \\ -\mathcal{I}_p(P \,\|\, Q) - (1-p)\,\dot{\mathcal{I}}_p(P \,\|\, Q), & p \in (\frac{1}{2}, 1), \end{cases} \tag{4.89}$$

and $\mathbb{F}_{P\|Q}(0) = \lim_{\alpha \downarrow 0} \mathbb{F}_{P\|Q}(\alpha)$. □

49. For $\gamma \geq 1$, denote $g_\gamma(t) = [t - \gamma]^+$, and define the E_γ *divergence* as the g_γ-divergence

$$E_\gamma(P \,\|\, Q) = D_{g_\gamma}(P \,\|\, Q). \tag{4.90}$$

Theorem 5. *For* $(P_1, P_{\mathbb{0}}) \in \mathscr{P}^2_{\mathcal{A}}$ *and* $(Q_1, Q_{\mathbb{0}}) \in \mathscr{P}^2_{\mathcal{B}}$,

$$\begin{gathered} (P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}}) \\ \Updownarrow \\ \{E_\gamma(P_1 \,\|\, P_{\mathbb{0}}) = E_\gamma(Q_1 \,\|\, Q_{\mathbb{0}}) \ \textit{and}\ E_\gamma(P_{\mathbb{0}} \,\|\, P_1) = E_\gamma(Q_{\mathbb{0}} \,\|\, Q_1), \quad \gamma \geq 1\}. \end{gathered} \tag{4.91}$$

Proof.

$\Downarrow$ We can invoke (4.64) and either (4.83) or the representation in [34, (112)],

$$E_\gamma(P \,\|\, Q) = \gamma \int_\gamma^\infty \frac{1 - \mathbb{F}_{P\|Q}(\log \beta)}{\beta^2} \mathrm{d}\beta. \tag{4.92}$$

$\Uparrow$ We can rely on Theorem 4 and

$$\mathcal{I}_p(P \,\|\, Q) = \begin{cases} p\, E_{\frac{1-p}{p}}(P \,\|\, Q), & p \in (0, \frac{1}{2}]; \\ (1-p) E_{\frac{p}{1-p}}(Q \,\|\, P), & p \in [\frac{1}{2}, 1). \end{cases} \tag{4.93}$$

Alternatively, we can capitalize on Theorem 3 of [34], namely,

$$\mathbb{F}_{P\|Q}(\log \gamma) = \begin{cases} 1 - E_\gamma(P \,\|\, Q) + \gamma\, \dot{E}_\gamma(P \,\|\, Q), & \gamma > 1; \\ -\lim_{\beta \downarrow 1} \dot{E}_\beta(Q \,\|\, P), & \gamma = 1; \\ \dot{E}_\beta(Q \,\|\, P)|_{\beta \leftarrow \frac{1}{\gamma}}, & 0 < \gamma < 1. \end{cases} \tag{4.94}$$

□

50. The fact (stated in Item 45) that no random transformation can increase the f-divergence between a pair of input probability measures suggests the possibility that the input relative information may *stochastically dominate* the output relative information. In other words, is it true that

$$\mathbb{F}_{P_1\|P_\emptyset}(x) \leq \mathbb{F}_{Q_1\|Q_\emptyset}(x), \qquad x \in \mathbb{R}, \tag{4.95}$$

for all $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$ and $(P_1, P_\emptyset) \in \mathscr{P}^2_{\mathcal{A}}$, with $P_\emptyset \to P_{Y|X} \to Q_\emptyset$ and $P_1 \to P_{Y|X} \to Q_1$? There are indeed cases in which (4.95) not only holds but holds with strict inequality on an interval of the real line. For example, if Y is independent of the input, then $\mathbb{F}_{Q_1\|Q_\emptyset}(x) = 1\{x \geq 0\}$ while $\mathbb{F}_{P_1\|P_\emptyset}(x) = 1\{x \geq 1 \text{ bit}\}$ if $P_1 = [0 \;\; 1], P_\emptyset = [\frac{1}{2} \;\; \frac{1}{2}]$. However, as long as $P_\emptyset \ll P_1$, it is impossible for (4.95) to hold and be strict in any interval because that would mean

$$\begin{aligned} 1 &= \Pi(P_\emptyset \,\|P_1) && (4.96) \\ &= \int_0^\infty \mathbb{F}_{P_1\|P_\emptyset}\left(\log \frac{1}{t}\right) \mathrm{d}t && (4.97) \\ &< \int_0^\infty \mathbb{F}_{Q_1\|Q_\emptyset}\left(\log \frac{1}{t}\right) \mathrm{d}t && (4.98) \\ &= \Pi(Q_\emptyset \,\|Q_1), && (4.99) \end{aligned}$$

where (4.97) and (4.99) follow from (4.33). Therefore, we reach the contradiction that a coefficient of absolute discontinuity is strictly greater than 1.

5. Total variation distance

In this section we turn our attention to the interplay between the relative information spectra and *total variation distance*

$$|P - Q| = 2 \max_{A \in \mathscr{F}} |P(A) - Q(A)|. \tag{5.1}$$

51. **Theorem 6.** *The total variation distance between $(P_X, P_Y) \in \mathscr{P}^2_{\mathscr{A}}$ can be expressed in terms of the relative information spectra through*

$$
\begin{aligned}
\tfrac{1}{2}\,|P_X - P_Y| &= \overline{\mathbb{F}}_{X\|Y}(0) - \mathbb{F}_{X\|Y}(0) && (5.2)\\
&= \int_0^1 \overline{\mathbb{F}}_{X\|Y}(\log\beta)\,\mathrm{d}\beta && (5.3)\\
&= 1 - \int_0^1 \mathbb{F}_{X\|Y}\left(\log\frac{1}{\beta}\right)\mathrm{d}\beta && (5.4)\\
&= 1 - \Pi(X\,\|\,Y) + \int_1^\infty \left(1 - \overline{\mathbb{F}}_{X\|Y}(\log\beta)\right)\mathrm{d}\beta && (5.5)\\
&= 1 - \Pi(Y\,\|\,X) + \int_{-\infty}^0 (1 - \exp(t))\;\mathrm{d}\overline{\mathbb{F}}_{X\|Y}(t) && (5.6)\\
&= 1 - \Pi(X\,\|\,Y) - \int_0^\infty (1 - \exp(t))\;\mathrm{d}\overline{\mathbb{F}}_{X\|Y}(t) && (5.7)\\
&= 1 - \tfrac{1}{2}\Pi(X\,\|\,Y) - \tfrac{1}{2}\Pi(Y\,\|\,X) + \tfrac{1}{2}\int_{-\infty}^\infty \left|1 - \exp(t)\right|\;\mathrm{d}\overline{\mathbb{F}}_{X\|Y}(t) && (5.8)\\
&= 1 - \tfrac{1}{2}\Pi(X\,\|\,Y) - \tfrac{1}{2}\Pi(Y\,\|\,X) + \tfrac{1}{2}\int_{-\infty}^\infty \left|1 - \exp(-t)\right|\;\mathrm{d}\mathbb{F}_{X\|Y}(t) && (5.9)\\
&= 1 - \Pi(X\,\|\,Y) + \int_0^\infty (1 - \exp(-t))\;\mathrm{d}\mathbb{F}_{X\|Y}(t) && (5.10)\\
&= 1 - \Pi(Y\,\|\,X) - \int_{-\infty}^0 (1 - \exp(-t))\;\mathrm{d}\mathbb{F}_{X\|Y}(t) && (5.11)\\
&= 1 - \Pi(Y\,\|\,X) + \int_1^\infty \mathbb{F}_{X\|Y}\left(\log\frac{1}{\beta}\right)\mathrm{d}\beta && (5.12)\\
&= \mathbb{E}\left[\left|\tanh\left(\tfrac{1}{2}\imath_{X\|Y}(W)\right)\right|\right], \quad W \sim \tfrac{1}{2}P_X + \tfrac{1}{2}P_Y, && (5.13)
\end{aligned}
$$

where the relative information in (5.13) *is in nats and* tanh(±∞) = ±1.

Proof.

- (5.2) Let $\mathscr{A}_+ = \{a \in \mathscr{A} \colon \imath_{X\|Y}(a) > 0\}$. Then,
  a) $P_X(\mathscr{A}_+) - P_Y(\mathscr{A}_+) = \overline{\mathbb{F}}_{X\|Y}(0) - \mathbb{F}_{X\|Y}(0)$;
  b) the absolute value in (5.1) is superfluous $\Longleftarrow$ $P_X(A) - P_Y(A) = P_Y(A^c) - P_X(A^c)$;
  c) $\mathscr{A}_+$ achieves the maximum in (5.1) because for any $E \in \mathscr{F}$,

$$
\begin{aligned}
P_X(\mathscr{A}_+) - P_X(E) &= P_X(\mathscr{A}_+ - E) - P_X(E - \mathscr{A}_+) && (5.14)\\
&\geq P_Y(\mathscr{A}_+ - E) - P_Y(E - \mathscr{A}_+) && (5.15)\\
&= P_Y(\mathscr{A}_+) - P_Y(E). && (5.16)
\end{aligned}
$$

- (5.3) $\Longleftarrow$ (4.24) with $\beta = 1$.
- (5.4) $\Longleftarrow$ (4.23) with $\beta = 1$.
- (5.5) $\Longleftarrow$ its right side is the right side of (5.3) $\Longleftarrow$ (4.33).

- (5.6) Let μ dominate $\{P_X, P_Y\}$ and let $p_X = \frac{\mathrm{d}P_X}{\mathrm{d}\mu}$ and $p_Y = \frac{\mathrm{d}P_Y}{\mathrm{d}\mu}$. Then,

$$\begin{aligned}\tfrac{1}{2}|P_X - P_Y| &= \int [p_Y - p_X]^+ \,\mathrm{d}\mu \qquad (5.17)\\ &= \int_{\mathcal{S}_{Y\|X}\cap\mathcal{S}^c_{X\|Y}} [p_Y - p_X]^+ \,\mathrm{d}\mu + \int_{\mathcal{S}_{Y\|X}\cap\mathcal{S}_{X\|Y}} [p_Y - p_X]^+ \,\mathrm{d}\mu \qquad (5.18)\\ &= P_Y(\mathcal{S}_{Y\|X}\cap\mathcal{S}^c_{X\|Y}) + \mathbb{E}\left[1\{\imath_{X\|Y}(Y)\in\mathbb{R}\}\,[1-\exp(\imath_{X\|Y}(Y))]^+\right] \qquad (5.19)\\ &= 1 - \Pi(Y\|X) + \int_{-\infty}^{0} (1-\exp(t))\,\mathrm{d}\bar{\mathbb{F}}_{X\|Y}(t), \qquad (5.20)\end{aligned}$$

where we have used (3.9).

- (5.7) Swapping $X \leftrightarrow Y$ in (5.17),

$$\begin{aligned}\tfrac{1}{2}|P_X - P_Y| &= \int [p_X - p_Y]^+ \,\mathrm{d}\mu \qquad (5.21)\\ &= \int_{\mathcal{S}_{X\|Y}\cap\mathcal{S}^c_{Y\|X}} [p_X - p_Y]^+ \,\mathrm{d}\mu + \int_{\mathcal{S}_{X\|Y}\cap\mathcal{S}_{Y\|X}} [p_X - p_Y]^+ \,\mathrm{d}\mu \qquad (5.22)\\ &= P_X(\mathcal{S}_{X\|Y}\cap\mathcal{S}^c_{Y\|X}) + \mathbb{E}\left[1\{\imath_{X\|Y}(Y)\in\mathbb{R}\}\,[\exp(\imath_{X\|Y}(Y))-1]^+\right] \qquad (5.23)\\ &= 1 - \Pi(X\|Y) + \int_{0}^{\infty} (\exp(t)-1)\,\mathrm{d}\bar{\mathbb{F}}_{X\|Y}(t), \qquad (5.24)\end{aligned}$$

where we have used (3.6).

- (5.8) $\Longleftarrow$ its right side is the arithmetic mean of the right sides of (5.6) and (5.7).
- (5.9) $\Longleftarrow X \leftrightarrow Y$ in (5.8) and (3.3).
- (5.10) $\Longleftarrow X \leftrightarrow Y$ in (5.6) and (3.3).
- (5.11) $\Longleftarrow X \leftrightarrow Y$ in (5.7) and (3.3).
- (5.12) $\Longleftarrow$ its right side is the right side of (5.4) $\Longleftarrow$ (4.29).
- (5.13) $\Longleftarrow$ choose $\mu = \frac{1}{2}P_X + \frac{1}{2}P_Y$ in $|P_X - P_Y| = \int |p_X - p_Y|\,\mathrm{d}\mu$ and note that

$$\left|\tanh\left(\tfrac{1}{2}\log_{\mathrm{e}}\frac{p_X}{p_Y}\right)\right| = \frac{|p_X - p_Y|}{p_X + p_Y} \quad \text{if } (p_X, p_Y) \in [0,\infty)^2 - \{(0,0)\}. \qquad (5.25)$$

□

52. Under the assumption $P_X \ll P_Y$, several of the representations in Theorem 6 can be found in [36, Theorem 12] and earlier in [42]. In addition, [36, Theorem 15] gives upper bounds on total variation distance as a function of the relative information spectrum if $P_X \ll P_Y$. Since those results are based on (5.3)–(5.4), which continue to hold in general, they too hold without restrictions on absolute continuity. In particular, the monotonicity of the relative information spectra and (5.3)–(5.4) result in

$$\tfrac{1}{2}\,|P_X - P_Y| \le \inf_{\delta>0}\left\{(1-\exp(-\delta))\;\bar{\mathbb{F}}_{X\|Y}(0) + \exp(-\delta)\,\bar{\mathbb{F}}_{X\|Y}(-\delta)\right\}, \qquad (5.26)$$

$$\tfrac{1}{2}\,|P_X - P_Y| \le 1 - \sup_{\delta>0}\left\{(1-\exp(-\delta))\;\mathbb{F}_{X\|Y}(0) + \exp(-\delta)\,\mathbb{F}_{X\|Y}(\delta)\right\}, \qquad (5.27)$$

which coincide with Le Cam's upper bounds in [19, p. 51], except that he weakens (5.27) by forbidding $\delta > \log 2$. As noted in [36], further strengthening of (5.26) [resp., (5.27)] is possible if $\overline{\mathbb{F}}_{X\|Y}(-\Delta)] = 0$ [resp., $\mathbb{F}_{X\|Y}(\Delta)] = 0$] for some $\Delta > 0$.

53. We can also lower bound total variation distance using Theorem 6 and the monotonicity of the relative information spectra. The following result supersedes Le Cam's lower bound in [19, p. 50], as well as the lower bounds in [36, Lemmas 17 and 18] claimed under $P_X \ll P_Y$.

Theorem 7. *For arbitrary* $(P_X, P_Y) \in \mathscr{P}^2_{\mathcal{A}}$ *and* $\delta > 0$,

$$\tfrac{1}{2}\,|P_X - P_Y| \geq \exp(-\delta)\,(1 - \Pi(X\,\|\,Y)) + (1 - \exp(-\delta))\,\mathbb{P}[\imath_{X\|Y}(X) \geq \delta], \tag{5.28}$$

$$\tfrac{1}{2}\,|P_X - P_Y| \geq 1 - \Pi(Y\,\|\,X) + (\exp(\delta) - 1)\;\mathbb{F}_{X\|Y}(-\delta), \tag{5.29}$$

$$\tfrac{1}{2}\,|P_X - P_Y| \geq 1 - \Pi(X\,\|\,Y) + (\exp(\delta) - 1)\;\mathbb{P}[\imath_{X\|Y}(Y) \geq \delta], \tag{5.30}$$

$$\tfrac{1}{2}\,|P_X - P_Y| \geq \exp(-\delta)\,(1 - \Pi(Y\,\|\,X)) + (1 - \exp(-\delta))\;\overline{\mathbb{F}}_{X\|Y}(-\delta). \tag{5.31}$$

Proof.

- (5.28) $\Longleftarrow$ (5.4) and $\mathbb{F}_{X\|Y}(t) \leq \Pi(X\|Y)\, 1\{t \geq \delta\} + (1 - \mathbb{P}[\imath_{X\|Y}(X) \geq \delta])\, 1\{t < \delta\}$.
- (5.29) $\Longleftarrow$ (5.12) and $\mathbb{F}_{X\|Y}(t) \geq \mathbb{F}_{X\|Y}(-\delta)\, 1\{t \geq -\delta\}$.
- (5.30) $\Longleftarrow$ $X \leftrightarrow Y$ in (5.29) and (4.9).
- (5.31) $\Longleftarrow$ $X \leftrightarrow Y$ in (5.28) and (4.10).

□

54. Even if $P_{X_1} \in \mathscr{P}_{\mathcal{A}}$ and $P_{X_2} \in \mathscr{P}_{\mathcal{A}}$ are close in total variation distance, their relative informations with respect to a third probability measure may behave quite differently.
Example. [19, p. 50]. Let $\mathcal{A} = [0, \infty)$, and suppose that P_{X_1}, P_{X_2}, and P_Y are uniform on $[0, n^2]$, $[1, n^2]$, and $[0, 1]$, respectively. Then, for all $\alpha \in \mathbb{R}$,

$$\overline{\mathbb{F}}_{X_1\|Y}(\alpha) = 1\{\alpha \geq -\log n^2\}, \tag{5.32}$$

$$\overline{\mathbb{F}}_{X_2\|Y}(\alpha) = 1, \tag{5.33}$$

$$|P_{X_1} - P_{X_2}| = \frac{2}{n^2}, \tag{5.34}$$

$$\mathbb{P}[\imath_{X_1\|Y}(Y) - \imath_{X_2\|Y}(Y) = \infty] = 1. \tag{5.35}$$

6. NP-divergence

55. The NP-*divergence* between $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$ is defined as

$$S(P\,\|\,Q) = |P \otimes Q - Q \otimes P|. \tag{6.1}$$

The terminology is motivated by an important operational role for $S(P\,\|\,Q)$ shown in Section 8 in the context of non-Bayesian hypothesis testing. For now, we point out that NP-divergence satisfies a simple Bayesian hypothesis testing operational role. Recall that the minimum average probability of error is equal to $\frac{1}{2} - \frac{1}{4}|P - Q|$ for equally likely P and Q. Now, suppose that we obtain a pair of observations one drawn from P and the other from Q, but we do not know the order of the pair and have no reason to favor one ordering over the other. Therefore, we have the equally likely hypotheses

$$\mathsf{H}_L\colon\ (y_1, y_2) \sim P_{\emptyset} \otimes P_1$$
$$\mathsf{H}_R\colon\ (y_1, y_2) \sim P_1 \otimes P_{\emptyset}$$

and the minimum probability of erroneous ordering is $\frac{1}{2} - \frac{1}{4} S(P \,\|\, Q)$.

56. *Blind wine tasting.* Offered a glass of 1982 Château Pétrus and a glass of 1990 Château Margaux, we are asked to identify which one is which. Suppose that for a given set of environmental conditions (temperature, lighting, etc.), P and Q stand for the probability measures of the respective wines on the space of visual, olfactory, and gustatory sensations. The probability of error is equal to $\frac{1}{2} - \frac{1}{4} S(P \,\|\, Q)$ since, a priori, the contents of the glasses are equally likely. Wanting to show off, a confident wine connoisseur makes a decision on the basis of tasting only one of the glasses. Then, the probability of error is $\frac{1}{2} - \frac{1}{4}|P - Q|$. Indeed, as shown below, $|P - Q| \leq S(P \,\|\, Q)$. If we do not condition on a given set of environmental conditions, the tasting sensations of both wines are dependent mainly because of their dependence on temperature. In that case, $S(P \,\|\, Q)$ is generalized to $|P_{XY} - P_{YX}|$, which can be applied whenever X and Y are defined on the same space. The potential utility of such measure of asymmetry of joint probability measures is yet to be explored.
57. *Example.* If $P = [\, p \;\; 1 - p \,]$ and $Q = [\, q \;\; 1 - q \,]$, then $S\,(P \,\|\, Q) = |P - Q| = 2\,|p - q|$.
58. *Example.* If $P = \left[\, \frac{1}{2} \; \frac{1}{2} \; 0 \,\right]$ and $Q = \left[\, 0 \; \frac{1}{2} \; \frac{1}{2} \,\right]$, then $|P - Q| = 1$ while $S(P \,\|\, Q) = \frac{3}{2}$.
59. *Example.* $S\,\left(\mathcal{N}\big(\mu_1, \sigma^2\big) \,\|\, \mathcal{N}\big(\mu_0, \sigma^2\big)\right) = 2 - 4\,\mathrm{Q}\left(\frac{|\mu_1 - \mu_0|}{\sqrt{2}\sigma}\right) = \left|\mathcal{N}\big(\mu_1, \frac{\sigma^2}{2}\big) - \mathcal{N}\big(\mu_0, \frac{\sigma^2}{2}\big)\right|$.
60. The NP-divergence satisfies the following properties.

Theorem 8. *Let $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$. Then,*

(a) $S(P \,\|\, Q) = S(Q \,\|\, P)$.

(b)

$$0 \leq S(P \,\|\, Q) \leq 2, \tag{6.2}$$

with equality on the left if and only if $P = Q$, and equality on the right if and only if $P \perp Q$.

(c) $S(P \,\|\, Q)$ does not satisfy the triangle inequality.

(d)

$$S(P_{X_1} \otimes \cdots \otimes P_{X_m} \,\|\, Q_{X_1} \otimes \cdots \otimes Q_{X_m}) \leq \sum_{i=1}^{m} S(P_i \,\|\, Q_i). \tag{6.3}$$

(e) If $P \neq Q$, then

$$\tfrac{1}{2} S(P^{\otimes n} \,\|\, Q^{\otimes n}) = 1 - \exp\left(-2n\, B(P \,\|\, Q) + o(n)\right), \tag{6.4}$$

where $B(P\|Q)$ is the Bhattacharyya distance *in Item 24.*

(f) If $|\mathcal{A}| = 2$, then $S(P \,\|\, Q) = |P - Q|$. In general,

$$|P - Q| \leq S(P \,\|\, Q) \leq 2\,|P - Q| - \tfrac{1}{2}|P - Q|^2. \tag{6.5}$$

(g)

$$\tfrac{1}{2} S(P \,\|\, Q) \geq 1 - \Pi(P \,\|\, Q) \cdot \Pi(Q \,\|\, P). \tag{6.6}$$

(h) Data processing inequality. *If $P_{X_1} \to P_{Y|X} \to P_{Y_1}$ and $P_{X_0} \to P_{Y|X} \to P_{Y_0}$, for some random transformation $P_{Y|X} \colon \mathcal{A} \to \mathcal{B}$, then*

$$S(Y_1 \,\|\, Y_0) \le S(X_1 \,\|\, X_0). \tag{6.7}$$

(i) No convex $f \colon (0, \infty) \to \mathbb{R}$ exists so that $D_f(P \,\|\, Q) = S(P \,\|\, Q)$ for all $(P, Q) \in \mathscr{P}^2_{\mathcal{A}}$.

(j) With $(X, Y) \sim P_X \otimes P_Y$,

$$\tfrac{1}{2}|P_X \otimes P_Y - P_Y \otimes P_X| = \mathbb{P}[\imath_{X\|Y}(X) > \imath_{X\|Y}(Y)] - \mathbb{P}[\imath_{X\|Y}(X) < \imath_{X\|Y}(Y)]. \tag{6.8}$$

(k)

$$(P_1, P_0) \equiv (Q_1, Q_0) \implies S(P_1 \,\|\, P_0) = S(Q_1 \,\|\, Q_0). \tag{6.9}$$

Proof.

(a) $\Longleftarrow |P - Q| = |Q - P|$.

(b) The inequalities follow because $S(P\|Q)$ is a total variation distance. Moreover,

$$S(P \,\|\, Q) = 0 \iff P \otimes Q = Q \otimes P \iff P = Q, \tag{6.10}$$
$$S(P \,\|\, Q) = 2 \iff P \otimes Q \perp Q \otimes P \iff P \perp Q. \tag{6.11}$$

(c) $P = \left[\, \tfrac{1}{2} \; \tfrac{1}{2} \; 0 \,\right]$, $Q = \left[\, \tfrac{1}{3} \; \tfrac{1}{3} \; \tfrac{1}{3} \,\right]$, $R = \left[\, 0 \; \tfrac{1}{2} \; \tfrac{1}{2} \,\right]$. Then, $S(P\|Q) + S(Q\|R) < S(P\|R)$, since

$$S(P\|Q) = S(Q\|R) = \tfrac{2}{3}, \quad \text{and} \quad S(P\|R) = \tfrac{3}{2}. \tag{6.12}$$

(d) Total variation distance satisfies the tensorization bound

$$|P_{X_1} \otimes \cdots \otimes P_{X_m} - Q_{X_1} \otimes \cdots \otimes Q_{X_m}| \le \sum_{i=1}^{m} |P_{X_i} - Q_{X_i}|. \tag{6.13}$$

Letting $(P_{X_i}, Q_{X_i}) \leftarrow (P_{X_i} \otimes Q_{X_i}, Q_{X_i} \otimes P_{X_i})$ yields (6.3).

(e) The proof consists of three building blocks:

i. As shown by Chernoff [14], if $P \neq Q$,

$$\tfrac{1}{2}|P^{\otimes n} - Q^{\otimes n}| = 1 - \exp\left(-n\, C(P \,\|\, Q) + o(n)\right), \tag{6.14}$$

where the *Chernoff information* is defined as

$$C(P \,\|\, Q) = \sup_{\alpha \in (0,1)} (1 - \alpha) D_\alpha(P \,\|\, Q). \tag{6.15}$$

ii. By relabeling of indices,

$$S(P^{\otimes n} \,\|\, Q^{\otimes n}) = |P^{\otimes n} \otimes Q^{\otimes n} - Q^{\otimes n} \otimes P^{\otimes n}| = |(P \otimes Q)^{\otimes n} - (Q \otimes P)^{\otimes n}|. \tag{6.16}$$

iii.

$$
\begin{aligned}
C(P\otimes Q\,\|\,Q\otimes P) &= \sup_{\alpha\in(0,1)} (1-\alpha)D_\alpha(P\otimes Q\,\|\,Q\otimes P) && (6.17)\\
&= \sup_{\alpha\in(0,1)} \{(1-\alpha)D_\alpha(P\,\|\,Q) + (1-\alpha)D_\alpha(Q\,\|\,P)\} && (6.18)\\
&\geq \tfrac{1}{2}D_{\frac{1}{2}}(P\,\|\,Q) + \tfrac{1}{2}D_{\frac{1}{2}}(Q\,\|\,P) && (6.19)\\
&= 2\,B(P\,\|\,Q), && (6.20)
\end{aligned}
$$

according to (3.51). To show that equality holds in (6.19), note that the function of α within {} is concave (e.g., [30]) with derivative

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}(1-\alpha)\left(D_\alpha(P\,\|\,Q)+D_\alpha(Q\,\|\,P)\right) = \frac{\int\left(\frac{p}{q}\right)^\alpha q\,\log\frac{q}{p}\,\mathrm{d}\mu}{\int\left(\frac{p}{q}\right)^\alpha q\,\mathrm{d}\mu} + \frac{\int\left(\frac{q}{p}\right)^\alpha p\,\log\frac{p}{q}\,\mathrm{d}\mu}{\int\left(\frac{q}{p}\right)^\alpha p\,\mathrm{d}\mu}, \tag{6.21}$$

which equals 0 at $\alpha = \frac{1}{2}$.

(f) For any joint probability measures P_{XY} and Q_{XY} on the same product space,

$$|P_{XY} - Q_{XY}| \geq \max\{|P_X - Q_X|, |P_Y - Q_Y|\}. \tag{6.22}$$

Letting $P_{XY} = P\otimes Q$ and $Q_{XY} = Q\otimes P$ yields $S(P\,\|\,Q) \geq |P-Q|$. As we saw in Item 57, equality holds for binary $\mathcal{A}$. The right inequality in (6.5) is a special case of

$$|P_0\otimes P_1 - Q_0\otimes Q_1| = |P_0 - Q_0| + |P_1 - Q_1| - \tfrac{1}{2}|P_0 - Q_0|\cdot|P_1 - Q_1|, \tag{6.23}$$

proved in [43] in the discrete case by means of the Strassen-Dobrushin coupling representation of total variation [44,45], which requires that (P_0, Q_0) be probability measures on a measurable space $(\mathcal{A}_0, \mathscr{F}_0)$ such that $\{(a,a)\colon a\in\mathcal{A}_0\}\in\mathscr{F}_0\otimes\mathscr{F}_0$, and analogously for (P_1, Q_1). This is satisfied by any Polish $\mathcal{A}_0$ endowed with its Borel field.

(g) Use the fact that $S(P\,\|\,Q) = D_f(P\otimes Q\,\|\,Q\otimes P)$ with $f(t) = |1-t|$, and dropping the first term on the right side of (4.73), we obtain

$$
\begin{aligned}
S(P\,\|\,Q) &\geq 2 - \Pi(P\otimes Q\,\|\,Q\otimes P) - \Pi(Q\otimes P\,\|\,P\otimes Q) && (6.24)\\
&= 2 - 2\,\Pi(P\,\|\,Q)\cdot\Pi(Q\,\|\,P), && (6.25)
\end{aligned}
$$

where (6.25) follows from Item 9.

(h) Given $P_{Y|X}\colon(\mathcal{A},\mathscr{F})\to(\mathcal{B},\mathscr{G})$, construct the random transformation $P_{Y_1Y_2|X_1X_2}\colon(\mathcal{A}^2,\mathscr{F}^2)\to(\mathcal{B}^2,\mathscr{G}^2)$ defined by

$$P_{Y_1Y_2|X_1X_2}(G_1\times G_2\,|\,a_1,a_2) = P_{Y|X}(G_1\,|\,a_1)P_{Y|X}(G_2\,|\,a_2), \quad (G_1,G_2,a_1,a_2)\in\mathscr{F}^2\times\mathcal{A}^2.$$

Note that

$$P_{X_0}\otimes P_{X_1}\to P_{Y_1Y_2|X_1X_2}\to P_{Y_0}\otimes P_{Y_1}$$

and

$$P_{X_1}\otimes P_{X_0}\to P_{Y_1Y_2|X_1X_2}\to P_{Y_1}\otimes P_{Y_0}.$$

Applying the data processing inequality for total variation distance to $P_{Y_1Y_2|X_1X_2}$ with inputs $P_{X_0}\otimes P_{X_1}$ and $P_{X_1}\otimes P_{X_0}$, we obtain

$$S(X_1\,\|\,X_0) = |P_{X_0}\otimes P_{X_1} - P_{X_1}\otimes P_{X_0}| \geq |P_{Y_0}\otimes P_{Y_1} - P_{Y_1}\otimes P_{Y_0}| = S(Y_1\,\|\,Y_0). \tag{6.26}$$

(i) Let's proceed by contradiction and assume that there exists a convex $f\colon (0,\infty)$ such that $S(P\,\|\,Q) = D_f(P\,\|\,Q)$ for all $(P,Q) \in \mathscr{P}^2_{\mathcal{A}}$. Since $S(P\,\|\,Q) = |P - Q|$ in the special case $|\mathcal{A}| = 2$, Lemma 10 implies that there exists $\alpha \in \mathbb{R}$ such that $f(t) = |1 - t| + \alpha\, t - \alpha$, i.e., $S(P\,\|\,Q) = |P - Q|$ which contradicts the examples in Items 58 and 59. An alternative route is to verify that NP-divergence fails to satisfy Lemma 11 by considering the special case $\lambda = \frac{1}{2}$, $P_1 = Q_1 = [\,\frac{1}{4}\ \frac{3}{4}\ 0\ 0\,]$, $P_{\mathbb{0}} = [\,0\ 0\ \frac{1}{4}\ \frac{3}{4}\,]$, $Q_{\mathbb{0}} = [\,0\ 0\ \frac{3}{4}\ \frac{1}{4}\,]$.

(j) With the notation used in the proof of Theorem 6,

$$\mathbb{P}[\imath_{X\|Y}(X) > \imath_{X\|Y}(Y)] = \iint p_X(a)\, p_Y(b)\, 1\{p_X(a)\, p_Y(b) > p_X(b)\, p_Y(a)\}\, \mathrm{d}\mu\, \mathrm{d}\mu, \tag{6.27}$$

$$\begin{aligned}\mathbb{P}[\imath_{X\|Y}(X) < \imath_{X\|Y}(Y)] &= \iint p_X(a)\, p_Y(b)\, 1\{p_X(a)\, p_Y(b) < p_X(b)\, p_Y(a)\}\, \mathrm{d}\mu\, \mathrm{d}\mu \\ &= \iint p_X(b)\, p_Y(a)\, 1\{p_X(b)\, p_Y(a) < p_X(a)\, p_Y(b)\}\, \mathrm{d}\mu\, \mathrm{d}\mu. \end{aligned} \tag{6.28}$$

Then, (6.8) follows from (5.17) and (6.1).

(k) The terms in the right side of (6.8) are determined by the relative information spectra:

$$\mathbb{P}[\imath_{X\|Y}(X) > \imath_{X\|Y}(Y)] = 1 - \mathbb{E}\left[\mathbb{F}_{X\|Y}(\imath_{X\|Y}(Y))\right], \tag{6.29}$$

$$\mathbb{P}[\imath_{X\|Y}(X) < \imath_{X\|Y}(Y)] = 1 - \mathbb{E}\left[\overline{\mathbb{F}}_{X\|Y}(\imath_{X\|Y}(X))\right]. \tag{6.30}$$

□

Any pair such that $(P,Q) \not\equiv (Q,P)$ provides a counterexample to $\Longleftarrow$ in (6.9).

61. While not an f-divergence, Theorem 8–(f) implies that NP-divergence is a g-divergence [46]. Several properties for the measure of dependence $\inf_{Q_Y \in \mathscr{P}_{\mathcal{B}}} S(P_{XY}\,\|\,P_X \otimes Q_Y)$ can be obtained by specializing [46, Theorem 8].

62. It may be useful to generalize the NP-divergence by replacing the total variation distance by any other f-divergence, i.e., define

$$S_f(P\,\|\,Q) = D_f(P \otimes Q\,\|\,Q \otimes P), \tag{6.31}$$

which satisfies $S_f(P\,\|\,Q) = S_f(Q\,\|\,P)$ even if D_f is not symmetric.

7. Sufficient statistics

Since its inception by Ronald Fisher in [47], the concept of sufficient statistics has played a fundamental role in mathematical statistics. This section offers a brief review of the various notions of sufficient statistics proposed in the literature, as well as their interrelationships emphasizing the connections with information theory. Moreover, we propose a new notion of sufficient statistics building upon the notion of equivalent pairs.

63. The basic setup in this section has the following ingredients:
   - measurable spaces $(\mathcal{Y}, \mathscr{F})$ and $(\mathcal{Z}, \mathscr{G})$;
   - a parameter set Θ;

- a data model (collection of distributions on $(\mathcal{Y},\mathscr{F})$): $\mathscr{P} = \{P_{Y|V=\theta} \in \mathscr{P}_{\mathcal{Y}}, \theta \in \Theta\}$;
- a random transformation: $P_{Z|Y} \colon (\mathcal{Y},\mathscr{F}) \to (\mathcal{Z},\mathscr{G})$.

An inference on the unknown parameter $\theta \in \Theta$ is made on the basis of the output of the random transformation $P_{Z|Y}$ when its input is distributed according to $P_{Y|V=\theta}$. Recall from Item 3 that no joint distribution P_{VY} is assumed to exist. In fact, the setting is non-Bayesian: No distribution is assumed on the set of parameters Θ, i.e., P_V need not be defined. The question to be formalized is: Under what conditions does the random transformation $P_{Z|Y}$ preserve all the information in Y that is relevant for inferring the parameter? Before proceeding, note that most of the statistical literature restricts attention to deterministic transformations, i.e., $Z = f(Y)$ for a $(\mathscr{F},\mathscr{G})$-measurable $f \colon \mathcal{Y} \to \mathcal{Z}$. Allowing random transformations (as in [15, 48, 49]) is practically useful, since sometimes the data is observed through an inherently random mechanism which, nevertheless, does not spoil the relevant information. For example, if $\mathcal{Y} = \mathscr{B}^n$, and under each $\theta \in \Theta$, $P_{Y|V=\theta} = P_\theta \otimes \cdots \otimes P_\theta$ with $P_\theta \in \mathscr{P}_{\mathscr{B}}$, then a random interleaver $P_{Z|Y} \colon \mathscr{B}^n \to \mathscr{B}^n$ preserves all the information in the observed n-tuple relevant to the inference of $\theta \in \Theta$ because $\{P_{Y|V=\theta} \in \mathscr{P}_{\mathcal{Y}},\ \theta \in \Theta\} = \{P_{Z|V=\theta} \in \mathscr{P}_{\mathcal{Y}},\ \theta \in \Theta\}$. At any rate, allowing random transformations is really a matter of mathematical convenience/elegance; in fact it does not widen the scope since the randomness can be incorporated into the data model: Letting $P_{Y|V=\theta} \leftarrow P_{Z|Y}P_{Y|V=\theta}$ and $f(z, y) = z$ subsumes the notion of random transformations as sufficient statistics into the classical deterministic transformations.

64. Fisher's notion [47] states that Z is a *sufficient statistic of* Y for the collection $\mathscr{P} = \{P_{Y|V=\theta}, \theta \in \Theta\}$ if

$$P_{Y|Z,V=\theta} \text{ does not depend on } \theta,$$

where, given that the unknown parameter is $V = \theta$, the joint probability measure of Y and Z is $P_{YZ|V=\theta} = P_{Y|V=\theta}P_{Z|Y}$. Following [50], when distinguishing from other notions of sufficient statistics, the sufficiency in the sense of this item is referred to as *classical sufficiency*.

65. Although $P_{YZ|V=\theta}$ is always well defined, we have to face the unfortunate fact that the conditional probability measure $P_{Y|Z,V=\theta}$ need not exist when $\mathcal{Y}$ is uncountable. The existence of such a conditional probability measure requires that for every $B \in \mathscr{F}$, there exist a $\mathscr{G}$-measurable $\phi_B \colon \mathcal{Z} \to [0, 1]$ such that for all $(\theta, B_0) \in \Theta \times \mathscr{G}$,

$$P_{YZ|V=\theta}(B \times B_0) = \mathbb{E}[\phi_B(Z)1\{Z \in B_0\}|V = \theta], \tag{7.1}$$

and $\phi_{\cdot}(z) \in \mathscr{P}_{\mathcal{Y}}$ for $z \in \mathcal{Z}_\theta$, with $P_{Z|V=\theta}(\mathcal{Z}_\theta) = 1$ for all $\theta \in \Theta$. To guarantee that this is the case, it is customary to abide by the restriction that $(\mathcal{Y},\mathscr{F})$ is a *standard* measurable space (i.e., it is isomorphic to $(E, \mathscr{B}_E)$ for some Borel subset of the real line $E \in \mathscr{B}$). Without such a restriction, Dieudonné [51] showed a counterexample where the required conditional probability does not exist, in which case the notion of sufficient statistics in the classical sense is vacuous. Whenever classical sufficiency is considered, it is typically assumed that the observation space is standard, even if this is not explicitly stated. As we see in Item 66, we need to place another restriction on the data model to make the notion of classical sufficiency well-behaved.

66. Bahadur [49] introduced the slightly more succinct notion of a *sufficient σ-field* $\overline{\mathscr{F}} \subset \mathscr{F}$, meaning that for every $B \in \mathscr{F}$ there exists a $\overline{\mathscr{F}}$-measurable $\varphi_B \colon \mathcal{Y} \to [0, 1]$, such that for all $(\theta, B_0) \in$

$\Theta \times \overline{\mathscr{F}}$,

$$P_{Y|V=\theta}(B \cap B_0) = \mathbb{E}[\varphi_B(Y) 1\{Y \in B_0\}|V = \theta]. \tag{7.2}$$

Then, a measurable function $f\colon \mathscr{Y} \to \mathscr{Z}$ is sufficient if and only if the σ-field it induces is sufficient. Curiously, the sufficiency of $\overline{\mathscr{F}}$ does not guarantee the sufficiency of every σ-field $\hat{\mathscr{F}}$ such that $\overline{\mathscr{F}} \subset \hat{\mathscr{F}} \subset \mathscr{F}$ [25]. Indeed, there may exist $f(y) = h(g(y))$ that is sufficient even though g is not sufficient. Fortunately, if the data model $\mathscr{P}$ is *dominated*, not only is that anomalous behavior impossible [49], but the notions of sufficient random transformation and sufficient σ-field are equivalent [52] (see also [37, (6.38)]).

67. Due to Halmos and Savage [3, Corollary 1], formalizing earlier ideas of Fisher [47, p. 331] and Neyman [53, Theorem II] in restricted settings, the following result is known as the *factorization theorem*.

Theorem 9. *Suppose that $(\mathscr{Y}, \mathscr{F})$ is standard, $\mathscr{P}$ is dominated, and $P_{Z|Y}$ is a deterministic transformation, i.e., $P_{Z|Y=y} = \delta_{f(y)}$, for a $(\mathscr{F}, \mathscr{G})$-measurable $f\colon \mathscr{Y} \to \mathscr{Z}$. Then, Z is a classically sufficient statistic of Y for $\mathscr{P}$ if and only if there exist Borel-measurable $w\colon \mathscr{Y} \to [-\infty, \infty)$ and $v\colon \Theta \times \mathscr{Z} \to [-\infty, \infty)$ such that, for all $\theta \in \Theta$, $v(\theta, \cdot)$ is Borel-measurable and*

$$\imath_{V;Y}(\theta; Y) = v(\theta, f(Y)) + w(Y), \ \text{a.s. } Y \sim P_{Y|V=\theta}, \tag{7.3}$$

where the information density is non-Bayesian (Item 26) with an arbitrary reference measure dominating $\mathscr{P}$.

Since the second term on the right side of (7.3) reflects the choice of the dominating measure, alternatively, we can express the condition for classical sufficiency in Theorem 9 as the existence of a dominating measure μ such that $\frac{\mathrm{d}P_{Y|V=\theta}}{\mathrm{d}\mu}$ is measurable with respect to the σ-field generated by f, for all $\theta \in \Theta$.

68. An important application of the factorization theorem is the following corollary to Corollary 3 in [3].

Theorem 10. *If $|\Theta| = 2$ and $(\mathscr{Y}, \mathscr{F})$ is standard, then $\imath_{P_1\|P_{\mathbb{0}}}(Y)$ is a sufficient statistic of Y for $\mathscr{P} = \{P_{Y|V=\theta},\ \theta \in \Theta\} = \{P_{\mathbb{0}}, P_1\}$.*

Proof. Define the function $g\colon (0,1) \times \Theta \times [-\infty, +\infty] \to (0, \infty]$,

$$g(p, \theta, z) = \begin{cases} 1 - p + p\exp(z), & \theta = \mathbb{0},\ z \in \mathbb{R}; \\ p + (1-p)\exp(-z), & \theta = 1,\ z \in \mathbb{R}; \\ 1 - p, & (\theta, z) = (\mathbb{0}, -\infty); \\ p, & (\theta, z) = (1, \infty); \\ \infty, & (\theta, z) = (1, -\infty) \text{ or } (\theta, z) = (\mathbb{0}, \infty). \end{cases} \tag{7.4}$$

We can easily verify that with $\bar{P} = \frac{1}{2}P_{\mathbb{0}} + \frac{1}{2}P_1$,

$$\imath_{P_{\mathbb{0}}\|\bar{P}}(Y) = -\log g\left(\tfrac{1}{2}, \mathbb{0}, \imath_{P_1\|P_{\mathbb{0}}}(Y)\right), \ \text{a.s. } Y \sim P_{\mathbb{0}}, \tag{7.5}$$

$$\imath_{P_1\|\bar{P}}(Y) = -\log g\left(\tfrac{1}{2}, 1, \imath_{P_1\|P_{\mathbb{0}}}(Y)\right), \ \text{a.s. } Y \sim P_1. \tag{7.6}$$

Consequently, letting $v(\theta, z) = -\log g(\frac{1}{2}, \theta, z)$ and $w(y) = \imath_{\bar{P}\|R}(y)$, (7.3) holds if the non-Bayesian information density on the left side is defined with reference measure R that dominates both P_0 and P_1. □

69. **Theorem 11.** *Let $\mathscr{P} = \{P_{Y|V=\theta},\ \theta \in \Theta\} = \{P_0, P_1\}$ be such that $D(P_1 \,\|\, P_0) < \infty$. Fix $P_{Z|Y}$ and denote $P_0 \to P_{Z|Y} \to Q_0$ and $P_1 \to P_{Z|Y} \to Q_1$. A necessary and sufficient condition for Z to be a classically sufficient statistic of Y for $\mathscr{P}$ is*

$$D(P_1 \,\|\, P_0) = D(Q_1 \,\|\, Q_0). \tag{7.7}$$

Introducing relative entropy in [2], Kullback and Leibler identified Theorem 11 as its most important property. However, they gave the result without the condition $D(P_1 \,\|\, P_0) < \infty$, in which case it need not hold.

70. The following generalization of Theorem 11 is due to Csiszár [15, 32].

Theorem 12. *Let $\mathscr{P} = \{P_{Y|V=\theta},\ \theta \in \Theta\} = \{P_0, P_1\}$ be such that $D_f(P_1 \,\|\, P_0) < \infty$, where $f\colon (0,\infty) \to \mathbb{R}$ is strictly convex. Fix $P_{Z|Y}$ and denote $P_0 \to P_{Z|Y} \to Q_0$ and $P_1 \to P_{Z|Y} \to Q_1$. A necessary and sufficient condition for Z to be a classically sufficient statistic of Y for $\mathscr{P}$ is*

$$D_f(P_1 \,\|\, P_0) = D_f(Q_1 \,\|\, Q_0). \tag{7.8}$$

71. Since Theorems 11 and 12 exclude pairs such that $D_f(P_1 \,\|\, P_0) = \infty$, it is interesting to see if there are any f-divergences such that $f\colon (0,\infty) \to \mathbb{R}$ is strictly convex and $D_f(P_1 \,\|\, P_0)$ is bounded for any pair $(P_1, P_0) \in \mathscr{P}^2_{\mathcal{Y}}$. The answer is affirmative: In view of (4.80), any order-α Hellinger divergence with $\alpha \in (0,1)$ (including the squared Hellinger distance (4.81)) is finite regardless of the pair of probability measures. Even though unbounded, the Bhattacharyya distance also qualifies since it is in one-to-one correspondence with the squared Hellinger distance. Also fitting the bill is the f-divergence with $f(t) = \frac{(t-1)^2}{t+1}$ known as the Vincze-LeCam divergence [18, 19],

$$\Delta(P \,\|\, Q) = D_f(P \,\|\, Q) \leq |P - Q| \leq 2. \tag{7.9}$$

Although outside of the scope of Theorem 12, could the simpler total variation distance serve the same purpose? The answer is negative as we verify with a simple counterexample in Item 84.

72. The binary case in Items 68–70 is particularly important: Z is said to be a *pairwise sufficient* statistic of Y for $\mathscr{P} = \{P_{Y|V=\theta},\ \theta \in \Theta\}$ if it is a sufficient statistic for $\{P_{Y|V=\theta}, P_{Y|V=\vartheta}\}$, for all $\theta \neq \vartheta \in \Theta$. Every sufficient statistic is pairwise sufficient. The converse holds if $\mathscr{P}$ is dominated [3, 37, 52, 54]. Therefore, as long as the data model is dominated, we need not wander beyond binary models to deal with classically sufficient statistics.

73. Introduced by Kolmogorov [55], Z is said to be a *Bayes sufficient statistic* of Y for $\mathscr{P}$ if for all $P_V \in \mathscr{P}_\Theta$, V and Y are conditionally independent given Z. While in Items 64–66 we did not impose the condition that the collection $\mathscr{P} = \{P_{Y|V=\theta} \in \mathscr{P}_{\mathcal{Y}}, \theta \in \Theta\}$ be a random transformation (Item 3), in this case we are indeed imposing the corresponding measurability requirement for which a σ-field $\mathscr{H}$ of the subsets of Θ is also specified. Therefore, in this setting we have the Markov chain

$$P_V \to P_{Y|V} \to P_{Z|Y} \to P_Z. \tag{7.10}$$

Classical sufficiency (Item 64) implies Bayes sufficiency, because once a probability measure is defined on V, $P_{VY|Z} = P_{V|Z}P_{Y|Z,V} = P_{V|Z}P_{Y|Z}$ if the classical criterion (7.1) is satisfied; therefore, V and Y are conditionally independent given Z. Conversely, if the collection $\mathscr{P}$ is dominated, then [3, 50] shows that pairwise Bayes sufficiency implies pairwise sufficiency, which in turn implies classical sufficiency as we saw in Item 72.

74. **Theorem 13.** *Suppose that $\mathscr{P} = \{P_{Y|V=\theta} \in \mathscr{P}_{\mathcal{Y}}, \theta \in \Theta\}$ is dominated. Then, Z is a Bayes sufficient statistic of Y for $\mathscr{P}$ if for all $P_V \in \mathscr{P}_\Theta$,*

$$\imath_{V;Y}(V;Y) = \imath_{V;Z}(V;Z), \text{ a.s. } (V,Y,Z) \sim P_V P_{Y|V} P_{Z|Y}. \tag{7.11}$$

Proof. Fix $P_V \in \mathscr{P}_\Theta$. We need to show that (7.11) is equivalent to the conditional independence of V and Y given Z. Particularizing the chain rule in (3.56),

$$\imath_{V;Y}(a;b) = \imath_{V;YZ}(a;b,c) = \imath_{V;Z}(a;c) + \imath_{V;Y|Z}(a;b|c). \tag{7.12}$$

According to Lemma 6, $P_{VY} \ll P_V \otimes P_Y$; therefore, $\mathbb{P}[\imath_{V;Y}(V;Y) \in \mathbb{R}] = 1$. We conclude that (7.11) is equivalent to $\imath_{V;Y|Z}(V;Y|Z) = 0$ a.s. □

75. In the binary case, Theorem 13 simplifies as follows.

Theorem 14. *Let $\mathscr{P} = \{P_{\mathbb{0}}, P_1\}$. Denote $P_{\mathbb{0}} \to P_{Z|Y} \to Q_{\mathbb{0}}$ and $P_1 \to P_{Z|Y} \to Q_1$ for a fixed $P_{Z|Y}$. Then, Z is a Bayes sufficient statistic of Y for $\mathscr{P}$ if and only if*

$$\imath_{P_1\|P_{\mathbb{0}}}(Y) = \imath_{Q_1\|Q_{\mathbb{0}}}(Z), \quad \text{a.s. for both } (Y,Z) \sim P_1 P_{Z|Y} \text{ and } (Y,Z) \sim P_{\mathbb{0}} P_{Z|Y}. \tag{7.13}$$

Proof. For $P_V = [0 \;\; 1]$ or $[1 \;\; 0]$, both sides of (7.11) are 0. Fix $P_V(1) = p \in (0,1)$, and denote by p_1 and $p_{\mathbb{0}}$ the densities of P_1 and $P_{\mathbb{0}}$, respectively, with respect to the dominating measure $p\,P_1 + (1-p)P_{\mathbb{0}}$. Analogously, denote by q_1 and $q_{\mathbb{0}}$ the densities of Q_1 and $Q_{\mathbb{0}}$, respectively, with respect to the dominating measure $p\,Q_1 + (1-p)Q_{\mathbb{0}}$. It readily follows that, with the notation in (7.4),

$$p_1(y) = \frac{1}{g\,(p, 1, \imath_{P_1\|P_{\mathbb{0}}}(y))} \quad \text{and} \quad p_{\mathbb{0}}(y) = \frac{1}{g\,(p, \mathbb{0}, \imath_{P_1\|P_{\mathbb{0}}}(y))}, \quad y \in \mathcal{Y}, \tag{7.14}$$

$$q_1(z) = \frac{1}{g\,(p, 1, \imath_{Q_1\|Q_{\mathbb{0}}}(z))} \quad \text{and} \quad q_{\mathbb{0}}(z) = \frac{1}{g\,(p, \mathbb{0}, \imath_{Q_1\|Q_{\mathbb{0}}}(z))}, \quad z \in \mathcal{Z}. \tag{7.15}$$

The condition in (7.11) is equivalent to

$$p_1(Y_1) = q_1(Z_1), \quad \text{a.s.}\, (Y_1, Z_1) \sim P_1 P_{Z|Y}, \tag{7.16}$$

$$p_{\mathbb{0}}(Y_{\mathbb{0}}) = q_{\mathbb{0}}(Z_{\mathbb{0}}), \quad \text{a.s.}\, (Y_{\mathbb{0}}, Z_{\mathbb{0}}) \sim P_{\mathbb{0}} P_{Z|Y}, \tag{7.17}$$

which in turn is equivalent to (7.13) in view of (7.14)–(7.15) and the strict monotonicity of the function $g\,(p, \theta, \cdot)$ for all $(p, \theta) \in (0,1) \times \{\mathbb{0}, 1\}$. □

76. **Theorem 15.** *Suppose that $\mathscr{P} = \{P_{Y|V=\theta} \in \mathscr{P}_{\mathcal{Y}}, \theta \in \Theta\}$ is dominated. Then, Z is a Bayes sufficient statistic of Y for $\mathscr{P}$ if and only if $I(V;Y) = I(V;Z)$ for all those P_V supported on two elements of Θ.*

    Proof. Because of the domination assumption, Bayes sufficiency is equivalent to pairwise Bayes sufficiency (Item 73). Therefore, we can restrict attention to those P_V supported on two elements of Θ. Note that $I(V;Y) \leq 1$ bit with those input distributions. Since $I(V;Y)$ is finite and $I(V;Z|Y) = 0$, the chain rule of mutual information

    $$I(V;Y,Z) = I(V;Z) + I(V;Y|Z) = I(V;Y) + I(V;Z|Y) \tag{7.18}$$

    implies that $I(V;Y) = I(V;Z)$ is equivalent to $I(V;Y|Z) = 0$, which, in turn, is equivalent to conditional independence of V and Y given Z. □

    Without imposing the domination assumption, a related claim can be found in [56, p. 36]. However, note that if $I(V;Y) = I(V;Z) = \infty$, (7.18) does not guarantee $I(V;Y|Z) = 0$. Apparently unaware of the notion of Bayes sufficiency, Lindley [5] had proposed $I(V;Y) = I(V;Z)$ for all P_V as a criterion for sufficiency, which he noticed to be implied by classical sufficiency.

77. Following [37,52,57,58], $P_{Z|Y}$ is called *Blackwell sufficient* for $\mathscr{P} = \{P_{Y|V=\theta}, \theta \in \Theta\}$ (sometimes also called *exhaustive* [59]) if there exists $P_{Y|Z}\colon (\mathcal{Z}, \mathscr{G}) \to (\mathcal{Y}, \mathscr{F})$ (dependent on $P_{Z|Y}$ and $\mathscr{P}$) such that for all $\theta \in \Theta$,

    $$P_{Y|V=\theta} \to P_{Z|Y} \to P_{Y|Z} \to P_{Y|V=\theta}. \tag{7.19}$$

    Therefore, $P_{Y|Z}$ acts as an "inverse random transformation" as long as the input to $P_{Z|Y}$ is drawn from $\mathscr{P}$. As shown in [60,61], (see also [52] and [37, (6.51)]), for dominated collections defined on standard spaces, classical sufficiency is the same as Blackwell sufficiency.

78. Let $\mathscr{P}_{\mathcal{Z}}$ stand for the collection of probability measures defined on $(\mathcal{Z}, \mathscr{G})$. In the terminology introduced by Blackwell [48,57], $\{P_{Y|V=\theta} \in \mathscr{P}_{\mathcal{Y}}, \theta \in \Theta\}$ is *at least as informative as* $\{P_{Z|V=\theta} \in \mathscr{P}_{\mathcal{Z}}, \theta \in \Theta\}$ if there exists a random transformation $P_{Z|Y}\colon (\mathcal{Y}, \mathscr{F}) \to (\mathcal{Z}, \mathscr{G})$ such that

    $$P_{Y|V=\theta} \to P_{Z|Y} \to P_{Z|V=\theta}, \quad \theta \in \Theta. \tag{7.20}$$

    So, $P_{Z|Y}$ is Blackwell sufficient if and only if $\{P_{Z|V=\theta}, \theta \in \Theta\}$ and $\{P_{Y|V=\theta}, \theta \in \Theta\}$ are *equally informative*.

79. Taking stock of the various notions of sufficient statistics reviewed so far in this section, the notion of Bayes sufficiency is, in principle, easier to apply than the classical notion in Item 64 and does not require the topological assumption of a standard space. On the other hand, the factorization theorem (Theorem 9) typically provides a convenient method for verifying the sufficiency of deterministic transformations. Although the Blackwell criterion (Item 77) is intuitively appealing, identifying the required inverse random transformation (or showing that none exists) is not always straightforward. Building on Definition 2, next we introduce a new notion of sufficient statistic that is both easy to verify and equivalent to the foregoing notions for dominated models in standard spaces.

Definition 3. *Fix $\{P_0, P_1\}$ and $P_{Z|Y}$, and denote $P_0 \to P_{Z|Y} \to Q_0$ and $P_1 \to P_{Z|Y} \to Q_1$. Then, Z is an I-sufficient statistic of Y for $\{P_0, P_1\}$ if*

$$(P_0, P_1) \equiv (Q_0, Q_1). \tag{7.21}$$

More generally, Z is an I-sufficient statistic of Y for $\{P_{Y|V=\theta},\ \theta \in \Theta\}$ if it is I-sufficient for every pair (θ, ϑ), $\theta \neq \vartheta \in \Theta$.

80. *Example.* For any $(P_0, P_1) \in \mathscr{P}^2_{\mathcal{Y}}$, $\imath_{P_1\|P_0}(Y)$ is an I-sufficient statistic of Y for $\{P_0, P_1\}$ because in view of Lemma 5,

$$\imath_{Q_1\|Q_0}(\imath_{P_1\|P_0}(Y_1)) = \imath_{P_1\|P_0}(Y_1) \text{ a.s. } Y_1 \sim P_1. \tag{7.22}$$

81. *Example.* Suppose that $\Theta = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}^n$ and $Y = (Y_1, \ldots, Y_n) = (\theta + X_1, \ldots, \theta + X_n)$, with $(X_1, \ldots, X_n)$ independent geometrically distributed with known parameter $q \in (0, 1)$. To verify that

$$Z = \min_{i=1,\ldots,n} Y_i \tag{7.23}$$

is an I-sufficient statistic of $(Y_1, \ldots, Y_n)$ for this undominated data model, first note that $P_{Y|V=\theta+\ell} \perp P_{Y|V=\theta}$ and $P_{Z|V=\theta+\ell} \perp P_{Z|V=\theta}$, unless ℓ is an integer. With $\ell \in \{1, 2, \ldots\}$, we obtain

$$\imath_{P_{Z|V=\theta+\ell}\|P_{Z|V=\theta}}(t) = \begin{cases} \ell\, n \log \dfrac{1}{1-q}, & t \in \{\theta + \ell, \theta + \ell + 1, \ldots\}; \\ -\infty, & t \in \{\theta, \ldots, \theta + \ell - 1\}; \\ \text{arbitrary}, & \text{otherwise}. \end{cases} \tag{7.24}$$

Moreover, we can easily check that $\imath_{P_{Y|V=\theta+\ell}\|P_{Y|V=\theta}}(y_1, \ldots, y_n) = \imath_{P_{Z|V=\theta+\ell}\|P_{Z|V=\theta}}(\min_{i=1,\ldots,n} y_i)$.

82. **Theorem 16.** *Assume that the data model $\mathscr{P} = \{P_{Y|V=\theta},\ \theta \in \Theta\}$ is dominated and fix $P_{Z|Y}$.*

   (a) If Z is a Bayes sufficient statistic of Y for $\mathscr{P}$, then Z is an I-sufficient statistic of Y for $\mathscr{P}$.

   (b) Assume that the observation space $(\mathcal{Y}, \mathscr{F})$ is standard. If Z is an I-sufficient statistic of Y for $\mathscr{P}$, then Z is a classically sufficient statistic of Y for $\mathscr{P}$.

   Proof. In view of Items 72 and 73, the domination assumption allows us to restrict attention to the $|\Theta| = 2$ case.

   (a) If Z is a Bayes sufficient statistic of Y for $\mathscr{P}$, then Theorem 14 implies that the random variables $\imath_{P_1\|P_0}(Y_1)$ and $\imath_{Q_1\|Q_0}(Z_1)$ with $Y_1 \sim P_1$ and $Z_1 \sim Q_1$ must have identical cumulative distribution functions. Therefore, (P_1, P_0) and (Q_1, Q_0) are equivalent pairs.

   (b) If Z is an I-sufficient statistic of Y for $\mathscr{P}$, then Theorem 3 implies that $D_f(P_1 \,\|\, P_0) = D_f(Q_1 \,\|\, Q_0)$ for all convex $f \colon (0, \infty) \to \mathbb{R}$. In particular, this encompasses the functions allowed in Theorem 12, and, consequently, Z is a classically sufficient statistic of Y for $\mathscr{P}$. Recall from Item 71 that the set of functions allowed in Theorem 12 is nonempty regardless of $(P_1, P_0) \in \mathscr{P}^2_{\mathcal{Y}}$.

□

83. The notions in Items 40 and 78 are related as follows.

Theorem 17. *Suppose that $(\mathscr{Y},\mathscr{F})$ and $(\mathscr{Z},\mathscr{G})$ are standard spaces. For any $(P_1,P_{\mathbb{0}})\in\mathscr{P}^2_{\mathscr{Y}}$ and $(Q_1,Q_{\mathbb{0}})\in\mathscr{P}^2_{\mathscr{Z}}$,*

$$\begin{gathered}(P_1,P_{\mathbb{0}})\equiv(Q_1,Q_{\mathbb{0}})\\ \Updownarrow\\ \{P_1,P_{\mathbb{0}}\}\ \textit{and}\ \{Q_1,Q_{\mathbb{0}}\}\ \textit{are equally informative models.}\end{gathered}\tag{7.25}$$

Proof.

$\Uparrow$ $\{P_1,P_{\mathbb{0}}\}$ and $\{Q_1,Q_{\mathbb{0}}\}$ are equally informative $\Rightarrow$ there exists $P_{Z|Y}\colon(\mathscr{Y},\mathscr{F})\to(\mathscr{Z},\mathscr{G})$ which is Blackwell sufficient for $\{P_1,P_{\mathbb{0}}\}$, and $P_1\to P_{Z|Y}\to Q_1$, $P_{\mathbb{0}}\to P_{Z|Y}\to Q_{\mathbb{0}}$. Since the model is dominated and lives in a standard space, Item 77 and Theorem 16 imply that $P_{Z|Y}$ is I-sufficient; therefore, $(P_1,P_{\mathbb{0}})\equiv(Q_1,Q_{\mathbb{0}})$.

$\Downarrow$ As we saw in Theorem 10, the deterministic transformation $P_{X|Y}$ that outputs $X=\imath_{P_1\|P_{\mathbb{0}}}(Y)$ is a classically sufficient statistic for $\{P_1,P_{\mathbb{0}}\}$, and, analogously, the deterministic transformation $P_{\bar{X}|Z}$ that outputs $\bar{X}=\imath_{Q_1\|Q_{\mathbb{0}}}(Z)$ is a classically sufficient statistic for $\{Q_1,Q_{\mathbb{0}}\}$. Because the spaces are standard and the models are dominated, those statistics are Blackwell sufficient; therefore, there exist $P_{Y|X}$ and $P_{Z|\bar{X}}$ such that

$$P_1\to P_{X|Y}\to P_{Y|X}\to P_1\tag{7.26}$$
$$P_{\mathbb{0}}\to P_{X|Y}\to P_{Y|X}\to P_{\mathbb{0}}\tag{7.27}$$
$$Q_1\to P_{\bar{X}|Z}\to P_{Z|\bar{X}}\to Q_1\tag{7.28}$$
$$Q_{\mathbb{0}}\to P_{\bar{X}|Z}\to P_{Z|\bar{X}}\to Q_{\mathbb{0}}.\tag{7.29}$$

Now by definition of $(P_1,P_{\mathbb{0}})\equiv(Q_1,Q_{\mathbb{0}})$, the response of $P_{X|Y}$ to P_1 is the same as the response of $P_{\bar{X}|Z}$ to Q_1, and the response of $P_{X|Y}$ to $P_{\mathbb{0}}$ is the same as the response of $P_{\bar{X}|Z}$ to $Q_{\mathbb{0}}$. Therefore,

$$Q_1\to P_{\bar{X}|Z}\to P_{Y|X}\to P_1\tag{7.30}$$
$$Q_{\mathbb{0}}\to P_{\bar{X}|Z}\to P_{Y|X}\to P_{\mathbb{0}}\tag{7.31}$$

which implies that $\{Q_{\mathbb{0}},Q_1\}$ is at least as informative as $\{P_{\mathbb{0}},P_1\}$. Reversing the roles $(P_1,P_{\mathbb{0}})$ and $(Q_1,Q_{\mathbb{0}})$, we conclude that $\{P_{\mathbb{0}},P_1\}$ is at least as informative as $\{Q_{\mathbb{0}},Q_1\}$.

□

84. *Example.* Let $\mathscr{B}=\{-,0,+\}$, $C=\{-,+\}$, and the random transformation $P_{Z|Y=0}(+)=\frac{1}{2}$, $P_{Z|Y=+}(+)=P_{Z|Y=-}(-)=1$. Furthermore, consider $\{P_{Y|V=\theta},\theta\in\Theta\}=\{P_1,P_{\mathbb{0}}\}$ with

$$P_1=\begin{bmatrix}\frac{1}{3}&\frac{2}{3}&0\end{bmatrix}\to P_{Z|Y}\to Q_1=\begin{bmatrix}\frac{2}{3}&\frac{1}{3}\end{bmatrix},\tag{7.32}$$
$$P_{\mathbb{0}}=\begin{bmatrix}0&\frac{2}{3}&\frac{1}{3}\end{bmatrix}\to P_{Z|Y}\to Q_{\mathbb{0}}=\begin{bmatrix}\frac{1}{3}&\frac{2}{3}\end{bmatrix}.\tag{7.33}$$

Then, $|P_1-P_{\mathbb{0}}|=|Q_1-Q_{\mathbb{0}}|=\frac{2}{3}$. Although $P_{Z|Y}$ preserves total variation distance, Z is not a sufficient statistic of Y because $(P_1,P_{\mathbb{0}})\not\equiv(Q_1,Q_{\mathbb{0}})$.

85. Summarizing the various results in this section as well as the necessary and sufficient conditions for equivalent pairs in Section 4, we have the following result.

Theorem 18. *Suppose that the data model* $\mathscr{P} = \{P_{Y|V=\theta} \in \mathscr{P}_{\mathcal{Y}}, \theta \in \Theta\}$ *is dominated and* $(\mathcal{Y}, \mathscr{F})$ *is a standard space. Fix any random transformation:* $P_{Z|Y} \colon (\mathcal{Y}, \mathscr{F}) \to (\mathcal{Z}, \mathscr{G})$*, and denote* $P_{\mathbb{0}} \to P_{Z|Y} \to Q_{\mathbb{0}}$ *and* $P_1 \to P_{Z|Y} \to Q_1$ *for* $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$*. The following are equivalent.*

(a) Z is a classically sufficient statistic of Y for $\mathscr{P}$*.*

(b) Z is a Bayes sufficient statistic of Y for $\mathscr{P}$*.*

(c) Z is a Blackwell sufficient statistic of Y for $\mathscr{P}$*.*

(d) Z is an I-sufficient statistic of Y for $\mathscr{P}$*.*

(e) For all $P_V \in \mathscr{P}_\Theta$*,*

$$\imath_{V;Y}(V;Y) = \imath_{V;Z}(V;Z), \text{ a.s. } (V,Y,Z) \sim P_V P_{Y|V} P_{Z|Y}. \tag{7.34}$$

(f) $I(V;Y) = I(V;Z)$ *for all those* $P_V \in \mathscr{P}_\Theta$ *supported on two elements of* Θ*.*

(g) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$ *and* $\alpha \in (0, \infty]$*,*

$$D_\alpha(P_1 \,\|\, P_{\mathbb{0}}) = D_\alpha(Q_1 \,\|\, Q_{\mathbb{0}}). \tag{7.35}$$

(h) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$ *and convex functions* $f \colon (0, \infty) \to \mathbb{R}$*,*

$$D_f(P_1 \,\|\, P_{\mathbb{0}}) = D_f(Q_1 \,\|\, Q_{\mathbb{0}}). \tag{7.36}$$

(i) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$*,*

$$\mathscr{H}^2(P_1 \,\|\, P_{\mathbb{0}}) = \mathscr{H}^2(Q_1 \,\|\, Q_{\mathbb{0}}). \tag{7.37}$$

(j) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$*,*

$$B(P_1 \,\|\, P_{\mathbb{0}}) = B(Q_1 \,\|\, Q_{\mathbb{0}}). \tag{7.38}$$

(k) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$*,*

$$\Delta(P_1 \,\|\, P_{\mathbb{0}}) = \Delta(Q_1 \,\|\, Q_{\mathbb{0}}). \tag{7.39}$$

(l) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$ *and* $p \in (0,1)$*,*

$$\mathcal{I}_p(P_1 \,\|\, P_{\mathbb{0}}) = \mathcal{I}_p(Q_1 \,\|\, Q_{\mathbb{0}}). \tag{7.40}$$

(m) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$ *and* $\gamma \geq 1$*,*

$$E_\gamma(P_1 \,\|\, P_{\mathbb{0}}) = E_\gamma(Q_1 \,\|\, Q_{\mathbb{0}}) \text{ and } E_\gamma(P_{\mathbb{0}} \,\|\, P_1) = E_\gamma(Q_{\mathbb{0}} \,\|\, Q_1). \tag{7.41}$$

(n) For all $(P_{\mathbb{0}}, P_1) \in \mathscr{P}^2$*,*

$$S(P_1 \,\|\, P_{\mathbb{0}}) = S(Q_1 \,\|\, Q_{\mathbb{0}}). \tag{7.42}$$

See Item 100 for the justification of Theorem 18-(n).

86. When compelled to choose among various non-sufficient statistics, it is helpful to assign a figure of merit to every random transformation $P_{Z|Y}$ providing an indication of how close it is to being sufficient. Motivated by Theorem 18-(i), a couple of possibilities are

$$\inf_{P_0 \neq P_1 \in \mathscr{P}} \frac{\mathscr{H}^2(Q_0 \,\|\, Q_1)}{\mathscr{H}^2(P_0 \,\|\, P_1)} \leq 1, \quad \text{and} \quad \inf_{P_0 \neq P_1 \in \mathscr{P}} \frac{\Delta(Q_0 \,\|\, Q_1)}{\Delta(P_0 \,\|\, P_1)} \leq 1, \tag{7.43}$$

where the inequalities follow from the fact that the both the squared Hellinger distance and the Vincze-Le Cam divergence are f-divergences and equality occurs if and only if $P_{Z|Y}$ is a sufficient statistic. Alternatively, Theorem 18-(d) suggests using

$$\sup_{P_0 \neq P_1 \in \mathscr{P}} K(\mathbb{F}_{P_1 \| P_0}, \mathbb{F}_{Q_1 \| Q_0}), \tag{7.44}$$

where $K(\mathbb{F}, \mathbb{G}) = \sup_{x \in \mathbb{R}} |\mathbb{F}(x) - \mathbb{G}(x)|$ is the Kolmogorov-Smirnov distance between the cumulative distribution functions $\mathbb{F}$ and $\mathbb{G}$. Then, $P_{Z|Y}$ is a sufficient statistic if and only if (7.44) is zero. Naturally, in (7.44) we can substitute $K(\mathbb{F}_{P_1 \| P_0}, \mathbb{F}_{Q_1 \| Q_0})$ by $|P_{\imath_{P_1 \| P_0}} - P_{\imath_{Q_1 \| Q_0}}|$ or any other measure of distance between probability measures.

8. Hypothesis testing

The information spectra of the absolute information $\imath_X(X)$ and of the information density $\imath_{X;Y}(X;Y)$ prove to be instrumental in determining the fundamental limits of lossless and lossy compression, respectively, as well as data transmission, in the latter case. Unfortunately, explicit solutions are not feasible and we must be contented with bounds, which become tight under stationary/ergodic assumptions in the limit of long data blocks. In contrast, the relative information spectra determine exactly the non-asymptotic fundamental tradeoff in *hypothesis testing*. This section gives a full detailed solution of that tradeoff in non-Bayesian hypothesis testing including an operational role for the NP-divergence. No restrictions are placed on the pair of probability measures that govern the observation under the respective hypotheses:

$$\begin{aligned} &\mathsf{H}_0\colon\ y \sim P_0, \\ &\mathsf{H}_1\colon\ y \sim P_1. \end{aligned}$$

Since we place no restrictions on P_0 and P_1, this "single-shot" setting encompasses the popular special case in which the observations are n independent drawings from a given distribution, $P_0 = \mathsf{P}_0^{\otimes n}$ and $P_1 = \mathsf{P}_1^{\otimes n}$.

87. Let $(P_1, P_0) \in \mathscr{P}_{\mathcal{Y}}^2$. A (randomized) *hypothesis test* is a deterministic measurable function $\phi\colon \mathcal{Y} \to [0,1]$, such that $\phi(y)$ is the probability of guessing P_1 if $y \in \mathcal{Y}$ is observed. A test ϕ is said to be deterministic if its range is $\{0,1\}$, i.e., $\phi(y) = 1\{y \in A\}$ for some measurable subset $A \subset \mathcal{Y}$. The performance of test ϕ is determined by the *conditional probabilities of error*,

$$\pi_{0|1} = \mathbb{P}[\text{test decides } \mathsf{H}_0 \,|\, \mathsf{H}_1] = 1 - \mathbb{E}[\phi(Y_1)], \quad Y_1 \sim P_1, \tag{8.1}$$

$$\pi_{1|0} = \mathbb{P}[\text{test decides } \mathsf{H}_1 \,|\, \mathsf{H}_0] = \mathbb{E}[\phi(Y_0)], \qquad Y_0 \sim P_0. \tag{8.2}$$

88. The hypothesis testing *fundamental tradeoff region* consists of the set of achievable error probability pairs,

$$C(P_1, P_{\mathbb{0}}) = \bigcup_{\phi\colon \mathcal{Y} \to [0,1]} \{(\mathbb{E}[\phi(Y_{\mathbb{0}})], 1 - \mathbb{E}[\phi(Y_1)])\}, \qquad Y_{\mathbb{0}} \sim P_{\mathbb{0}},\ Y_1 \sim P_1. \tag{8.3}$$

In other words, $(\pi_{1|\mathbb{0}}, \pi_{\mathbb{0}|1}) \in C(P_1, P_{\mathbb{0}})$ if there is a hypothesis test for $(P_1, P_{\mathbb{0}})$ achieving conditional error probabilities $(\pi_{1|\mathbb{0}}, \pi_{\mathbb{0}|1})$. Elementary properties of the fundamental tradeoff region include:

Theorem 19.

(a) $C(P_1, P_{\mathbb{0}})$ *is a convex set.*
(b) $C(P_1, P_{\mathbb{0}})$ *is a closed set.*
(c) $(a, b) \in C(P_1, P_{\mathbb{0}}) \Longleftrightarrow (1 - a, 1 - b) \in C(P_1, P_{\mathbb{0}})$.
(d) $(a, b) \in C(P_1, P_{\mathbb{0}}) \Longleftrightarrow (b, a) \in C(P_{\mathbb{0}}, P_1)$.

Proof.

(a) If ϕ_0 and ϕ_1 attain $(a_0, b_0) \in C(P_1, P_{\mathbb{0}})$ and $(a_1, b_1) \in C(P_1, P_{\mathbb{0}})$, respectively, and $\alpha \in (0, 1)$, then the test $(1 - \alpha)\phi_0 + \alpha\,\phi_1$ attains $(1 - \alpha)(a_0, b_0) + \alpha\,(a_1, b_1)$.
(b) The mapping $\phi \mapsto (\mathbb{E}[\phi(Y_{\mathbb{0}})], 1 - \mathbb{E}[\phi(Y_1)])$ is linear.
(c) The test $1 - \phi$ achieves $(1 - \pi_{1|\mathbb{0}}, 1 - \pi_{\mathbb{0}|1})$ if ϕ achieves $(\pi_{1|\mathbb{0}}, \pi_{\mathbb{0}|1})$.
(d) Interchanging $\phi \leftrightarrow 1 - \phi$ and $P_{\mathbb{0}} \leftrightarrow P_1$ in (8.3).

□

89. In view of Theorem 19-(c), the set of points in $C(P_1, P_{\mathbb{0}})$ above the $(0, 1)$—$(1, 0)$ diagonal is redundant. The set of Pareto optimal error probability pairs is the lower boundary of $C(P_1, P_{\mathbb{0}})$ below the diagonal, which we refer to as the *fundamental tradeoff function* $\{\alpha_\nu \in [0, 1], \nu \in [0, 1]\}$ defined by

$$\alpha_\nu(P_1, P_{\mathbb{0}}) = \min\{y \in [0, 1]\colon (\nu, y) \in C(P_1, P_{\mathbb{0}})\} \tag{8.4}$$

$$= \min_{\phi\colon \pi_{1|\mathbb{0}} \leq \nu} \pi_{\mathbb{0}|1} = 1 - \max_{\phi\colon \mathbb{E}[\phi(Y_{\mathbb{0}})] \leq \nu} \mathbb{E}[\phi(Y_1)]. \tag{8.5}$$

As a consequence of Theorem 19-(a), $\alpha_\nu(P_1, P_{\mathbb{0}})$ is convex on $[0, 1]$. Although the fundamental tradeoff region $C(P_1, P_{\mathbb{0}})$ and the fundamental tradeoff function $\alpha_\nu(P_1, P_{\mathbb{0}})$ determine each other, it is advantageous to work with both simultaneously, as we see below.

90. The diagonal connecting $(0, 1)$—$(1, 0)$ belongs to the fundamental tradeoff region

$$\{(p, 1 - p) \in [0, 1]^2\colon p \in [0, 1]\} \subset C(P_1, P_{\mathbb{0}}), \tag{8.6}$$

since $(p, 1 - p)$ is attained by the blind test $\phi(y) = p$, $y \in \mathcal{Y}$. If $P_1 = P_{\mathbb{0}}$, then blind tests are optimal and equality holds in (8.6). Note that in this case the area of the fundamental tradeoff region satisfies $|C(P_1, P_1)| = 0$, and the fundamental tradeoff function is $\alpha_\nu(P_1, P_1) = 1 - \nu$, $\nu \in [0, 1]$.

91. At the other extreme, if $P_1 \perp P_0$, then $C(P_1, P_0) = [0,1]^2$, $\alpha_\nu(P_1, P_0) = 0$, $\nu \in [0,1]$, and the area of the fundamental tradeoff region satisfies $|C(P_1, P_0)| = 1$. To see this, recall (Item 7) that there exists an event $F \in \mathscr{F}$ such that $P_1(F) = 1$ and $P_0(F) = 0$. The deterministic test $\phi(y) = 1\{y \in F\}$ achieves the point $(0,0)$, while the test $\phi(y) = 1\{y \notin F\}$ achieves the point $(1,1)$. All other points in the square are achievable because of (8.6) and Theorem 19-(a).

92. Inspired by radar, the function $\nu \mapsto 1 - \alpha_\nu(P_1, P_0)$ is frequently (e.g., [62]) referred to as the *receiver operating characteristic*, or ROC. In the radar application, P_0 is the distribution of the observations under the absence of target return. In fact, this terminology is applied not just to the best possible curve but to the tradeoff between $\pi_{1|1}$ and $\pi_{1|0}$ achieved by any particular family of tests. The so-called *area under the (*ROC*) curve*, commonly abbreviated as AUC,

$$\int_0^1 (1 - \alpha_\nu(P_1, P_0))\, \mathrm{d}\nu = \tfrac{1}{2} + \tfrac{1}{2}\, |C(P_1, P_0)| \in [\tfrac{1}{2}, 1], \tag{8.7}$$

is frequently used as a scalar proxy to evaluate the degree to which P_0 and P_1 can be distinguished. It ranges from $\frac{1}{2}$ if $P_0 = P_1$ to 1 if $P_0 \perp P_1$.

93. *Data processing theorem for the fundamental tradeoff region.* If $P_Y \to P_{Z|Y} \to P_Z$ and $Q_Y \to P_{Z|Y} \to Q_Z$, then

$$\alpha_\nu(P_Y, Q_Y) \leq \alpha_\nu(P_Z, Q_Z), \quad \nu \in [0,1], \tag{8.8}$$

$$C(P_Z, Q_Z) \subset C(P_Y, Q_Y), \tag{8.9}$$

since we always have the option of incorporating $P_{Z|Y}$ as a front end of the hypothesis test. Equality holds in (8.8)–(8.9) if Z is a Blackwell sufficient statistic of Y for (P_Y, Q_Y), because from Z (along, with possibly additional randomness) we can synthesize data whose conditional distributions are P_Y and Q_Y. Feeding that data to ϕ results in the same $(\pi_{1|0}, \pi_{0|1})$ as feeding the original Y to ϕ.

94. The minimal error probabilities compatible with *zero* error probability of the other kind are denoted by $\underline{\pi}_{0|1}$ and $\underline{\pi}_{1|0}$, i.e., they are defined by

- $(\underline{\pi}_{1|0}, 0) \in C(P_1, P_0)$ but $(\pi_{1|0}, 0) \notin C(P_1, P_0)$ if $\pi_{1|0} < \underline{\pi}_{1|0}$;
- $(0, \underline{\pi}_{0|1}) \in C(P_1, P_0)$ but $(0, \pi_{0|1}) \notin C(P_1, P_0)$ if $\pi_{0|1} < \underline{\pi}_{0|1}$.

By definition,

$$\alpha_0(P_1, P_0) = \underline{\pi}_{0|1}, \tag{8.10}$$

$$\alpha_\nu(P_1, P_0) = 0 \iff \nu \in [\underline{\pi}_{1|0}, 1]. \tag{8.11}$$

Theorem 20. *For any* $(P_1, P_0) \in \mathscr{P}_{\mathcal{Y}}^2$,

(a) $\underline{\pi}_{1|0} = \Pi(P_0 \,\|\, P_1)$*, achieved by the test* $\phi(y) = 1\{y \in \mathcal{S}_{P_1\|P_0}\}$*;*

(b) $\underline{\pi}_{0|1} = \Pi(P_1 \,\|\, P_0)$*, achieved by the test* $\phi(y) = 1\{y \notin \mathcal{S}_{P_0\|P_1}\}$*.*

Proof. For any test ϕ achieving error probabilities $(\pi_{1|0}, \pi_{0|1})$,

$$\pi_{0|1} = 0 \quad \iff \quad P_1(\phi^{-1}(1)) = 1, \tag{8.12}$$

and

$$\begin{aligned}\pi_{1|\mathbb{0}} &\geq P_{\mathbb{0}}(\phi^{-1}(1)) && (8.13)\\ &\geq P_{\mathbb{0}}(\mathcal{S}_{P_1\|P_{\mathbb{0}}}), && (8.14)\end{aligned}$$

where (8.14) holds if $\pi_{\mathbb{0}|1} = 0$ because of (2.4) and (8.12). Furthermore, as we saw in (2.14), if $\phi(y) = 1\{y \in \mathcal{S}_{P_1\|P_{\mathbb{0}}}\}$, then the right side of (8.12) is satisfied and (8.13)–(8.14) become identities. Recalling (2.14) completes the proof of (a). The proof of (b) is identical. □

95. Introduced in [7], for $(\gamma, \lambda) \in \mathbb{R} \times [0, 1]$, a *Neyman-Pearson test* between P_1 and $P_{\mathbb{0}}$ is

$$\phi_{\gamma,\lambda}(y) = \begin{cases} 1, & \imath_{P_1\|P_{\mathbb{0}}}(y) > \gamma; \\ \lambda, & \imath_{P_1\|P_{\mathbb{0}}}(y) = \gamma; \\ 0, & \imath_{P_1\|P_{\mathbb{0}}}(y) < \gamma. \end{cases} \tag{8.15}$$

The tests $\phi_{\gamma,0}$ and $\phi_{\gamma,1}$ are known as *deterministic Neyman-Pearson* tests. The *limiting Neyman-Pearson* tests are the deterministic tests

$$\lim_{\gamma\to\infty} \phi_{\gamma,\lambda}(y) = 1\{y \notin \mathcal{S}_{P_{\mathbb{0}}\|P_1}\}, \tag{8.16}$$

$$\lim_{\gamma\to-\infty} \phi_{\gamma,\lambda}(y) = 1\{y \in \mathcal{S}_{P_1\|P_{\mathbb{0}}}\}. \tag{8.17}$$

96. With $Y_{\mathbb{0}} \sim P_{\mathbb{0}}$ and $Y_1 \sim P_1$, the Neyman-Pearson test (8.15) achieves the conditional error probabilities

$$\pi_{\mathbb{0}|1}(\gamma, \lambda) = 1 - \mathbb{E}[\phi_{\gamma,\lambda}(Y_1)] = \mathbb{F}_{P_1\|P_{\mathbb{0}}}(\gamma) - \lambda\,\mathbb{P}\left[\imath_{P_1\|P_{\mathbb{0}}}(Y_1) = \gamma\right], \tag{8.18}$$

$$\pi_{1|\mathbb{0}}(\gamma, \lambda) = \mathbb{E}[\phi_{\gamma,\lambda}(Y_{\mathbb{0}})] = 1 - \overline{\mathbb{F}}_{P_1\|P_{\mathbb{0}}}(\gamma) + \lambda\,\mathbb{P}\left[\imath_{P_1\|P_{\mathbb{0}}}(Y_{\mathbb{0}}) = \gamma\right]. \tag{8.19}$$

The randomization serves to obtain convex combinations of the performances obtained by deterministic Neyman-Pearson tests,

$$\pi_{\mathbb{0}|1}(\gamma, \lambda) = \lambda\,\pi_{\mathbb{0}|1}(\gamma, 1) + (1-\lambda)\,\pi_{\mathbb{0}|1}(\gamma, 0), \tag{8.20}$$

$$\pi_{1|\mathbb{0}}(\gamma, \lambda) = \lambda\,\pi_{1|\mathbb{0}}(\gamma, 1) + (1-\lambda)\,\pi_{1|\mathbb{0}}(\gamma, 0), \tag{8.21}$$

where

$$\pi_{\mathbb{0}|1}(\gamma, 0) = \mathbb{F}_{P_1\|P_{\mathbb{0}}}(\gamma), \tag{8.22}$$

$$\pi_{1|\mathbb{0}}(\gamma, 0) = 1 - \overline{\mathbb{F}}_{P_1\|P_{\mathbb{0}}}(\gamma), \tag{8.23}$$

$$\pi_{\mathbb{0}|1}(\gamma, 1) = \lim_{\alpha\uparrow\gamma} \pi_{\mathbb{0}|1}(\alpha, 0) = \lim_{\alpha\uparrow\gamma} \mathbb{F}_{P_1\|P_{\mathbb{0}}}(\alpha), \tag{8.24}$$

$$\pi_{1|\mathbb{0}}(\gamma, 1) = \lim_{\alpha\uparrow\gamma} \pi_{1|\mathbb{0}}(\alpha, 0) = 1 - \lim_{\alpha\uparrow\gamma} \overline{\mathbb{F}}_{P_1\|P_{\mathbb{0}}}(\alpha). \tag{8.25}$$

97. The following venerable result states that the non-limiting Neyman-Pearson tests are Pareto-optimal.

Lemma 12. Neyman-Pearson *[7]. Let* $Y_\mathtt{0} \sim P_\mathtt{0}$ *and* $Y_1 \sim P_1$*. For any* $\lambda \in [0,1]$*,* $\gamma \in \mathbb{R}$*, and measurable function* $\phi \colon \mathcal{Y} \to [0,1]$*,*

$$\mathbb{E}[\phi(Y_1)] > \mathbb{E}[\phi_{\gamma,\lambda}(Y_1)] \quad \Longrightarrow \quad \mathbb{E}[\phi(Y_\mathtt{0})] > \mathbb{E}[\phi_{\gamma,\lambda}(Y_\mathtt{0})]. \tag{8.26}$$

Proof. Invoking (8.18)–(8.19), (4.25), and Lemma 9 with $g(a) = 1 - \phi(a)$, we obtain

$$\mathbb{E}[\phi(Y_\mathtt{0})] - \mathbb{E}[\phi_{\gamma,\lambda}(Y_\mathtt{0})] \geq \exp(-\gamma)\left(\mathbb{E}[\phi(Y_1)] - \mathbb{E}[\phi_{\gamma,\lambda}(Y_1)]\right). \tag{8.27}$$

□

98. In addition to giving the fundamental tradeoff in terms of the relative information spectra, the following result finds an operational role for the NP-divergence.

Theorem 21. *Let* $(P_1, P_\mathtt{0}) \in \mathscr{P}^2_\mathcal{Y}$ *such that* $P_\mathtt{0} \not\perp P_1$*.*

(a) The limiting Neyman-Pearson tests $\phi(y) = 1\{y \in \mathcal{S}_{P_1 \| P_\mathtt{0}}\}$ *and* $\phi(y) = 1\{y \notin \mathcal{S}_{P_\mathtt{0} \| P_1}\}$ *achieve the Pareto-optimal points* $(\underline{\pi}_{1|\mathtt{0}}, 0) \in C(P_1, P_\mathtt{0})$ *and* $(0, \underline{\pi}_{\mathtt{0}|1}) \in C(P_1, P_\mathtt{0})$*, respectively.*

(b) The (limiting and nonlimiting) Neyman-Pearson tests achieve $\mathcal{P}$*, the set of Pareto-optimal points of* $C(P_1, P_\mathtt{0})$*.* $\mathcal{P}$ *is the convex closure of*

$$C_0 = \bigcup_{\gamma \in \mathbb{R}} \left\{ \left(1 - \overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(\gamma), \mathbb{F}_{P_1 \| P_\mathtt{0}}(\gamma)\right) \right\}. \tag{8.28}$$

Therefore, the intersection of $C(P_1, P_\mathtt{0})$ *and the triangle below the diagonal* $(0,1)$—$(1,0)$ *is the convex hull of* $\mathcal{P} \cup (0,1) \cup (1,0)$*.*

(c) For $\nu \in (0, \underline{\pi}_{1|\mathtt{0}})$*, the Neyman-Pearson test that achieves* $\alpha_\nu(P_1, P_\mathtt{0})$ *is* $\phi_{\gamma^\star, \lambda^\star}$ *given by:*

- *if* $\overline{\mathbb{F}}^{-1}_{P_1 \| P_\mathtt{0}}(1 - \nu) \neq \emptyset$*, then* $\lambda^\star = 0$ *and* $\gamma^\star$ *is any solution to*

$$1 - \nu = \overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(\gamma^\star), \tag{8.29}$$

in which case,

$$\alpha_\nu(P_1, P_\mathtt{0}) = \mathbb{F}_{P_1 \| P_\mathtt{0}}(\gamma^\star), \tag{8.30}$$

- *if* $\overline{\mathbb{F}}^{-1}_{P_1 \| P_\mathtt{0}}(1 - \nu) = \emptyset$*, then* $\gamma^\star$ *is the unique scalar such that*

$$\lim_{x \uparrow \gamma^\star} \overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(x) < 1 - \nu < \overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(\gamma^\star), \tag{8.31}$$

and

$$\lambda^\star = \frac{\overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(\gamma^\star) - 1 + \nu}{\overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(\gamma^\star) - \lim_{x \uparrow \gamma^\star} \overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(x)} \in (0,1), \tag{8.32}$$

in which case,

$$\begin{aligned} \alpha_\nu(P_1, P_\mathtt{0}) &= \lambda^\star \lim_{x \uparrow \gamma^\star} \mathbb{F}_{P_1 \| P_\mathtt{0}}(x) + (1 - \lambda^\star)\, \mathbb{F}_{P_1 \| P_\mathtt{0}}(\gamma^\star) & (8.33) \\ &= \mathbb{F}_{P_1 \| P_\mathtt{0}}(\gamma^\star) + \exp(\gamma^\star)\left(1 - \nu - \overline{\mathbb{F}}_{P_1 \| P_\mathtt{0}}(\gamma^\star)\right). & (8.34) \end{aligned}$$

For $\nu \in (0, \underline{\pi}_{1|\mathbb{0}})$, the fundamental tradeoff function satisfies

$$\alpha_\nu(P_1, P_{\mathbb{0}}) = \max_{\gamma \in \mathbb{R}} \left\{ \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma) - \exp(\gamma) \left(\nu - 1 + \overline{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(\gamma) \right) \right\}, \tag{8.35}$$

where the maximum is achieved by $\gamma^\star$.

(d)

$$(P_1, P_{\mathbb{0}}) \equiv (Q_1, Q_{\mathbb{0}}) \iff C(P_1, P_{\mathbb{0}}) = C(Q_1, Q_{\mathbb{0}}). \tag{8.36}$$

(e)

$$|C(P_1, P_{\mathbb{0}})| = \tfrac{1}{2}\, S(P_1 \,\|\, P_{\mathbb{0}}). \tag{8.37}$$

(f) If $P \neq Q$, then

$$|C(P^{\otimes n}, Q^{\otimes n})| = 1 - \exp\left(-2n\, B(P \,\|\, Q) + o(n)\right). \tag{8.38}$$

Proof.

(a) $\Longleftarrow$ Theorem 20, (8.16)–(8.17), and the fact that $(\underline{\pi}_{1|\mathbb{0}}, 0)$ and $(0, \underline{\pi}_{\mathbb{0}|1})$ are Pareto optimal by definition.

(b) From (8.22)–(8.23) observe that the elements in C_0 are the conditional error probability pairs achieved by the deterministic Neyman-Pearson tests $\phi_{\gamma,0}$. The error probability pairs achieved by deterministic Neyman-Pearson tests $\phi_{\gamma,1}$ belong to the closure of C_0 in view of (8.24)–(8.25). The error probability pairs achieved by the randomized tests $\phi_{\gamma,\lambda}$, with $\lambda \in (0, 1)$ are the convex combinations of the pairs achieved by deterministic tests, as we saw in (8.20)–(8.21). Moreover, Theorem 20, (4.5), and (4.8) indicate that the closure of C_0 includes

$$(\underline{\pi}_{1|\mathbb{0}}, 0) = \lim_{\gamma \to -\infty} \left(1 - \overline{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(\gamma), \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma) \right), \tag{8.39}$$

$$(0, \underline{\pi}_{\mathbb{0}|1}) = \lim_{\gamma \to \infty} \left(1 - \overline{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(\gamma), \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma) \right). \tag{8.40}$$

We conclude that the set of points achieved by the (limiting and nonlimiting) Neyman-Pearson tests is equal to the convex closure of C_0. According to Lemma 12, the nonlimiting Neyman-Pearson tests achieve Pareto-optimal points of $C(P_1, P_{\mathbb{0}})$. But since the convex closure of C_0 connects the Pareto-optimal points $(\underline{\pi}_{1|\mathbb{0}}, 0)$ and $(0, \underline{\pi}_{\mathbb{0}|1})$ without any gaps there can be no Pareto-optimal points of $C(P_1, P_{\mathbb{0}})$ other than the convex closure of C_0.

(c) (8.29)–(8.30) follow from (8.22)–(8.23). Observe that whenever $\overline{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(\gamma)$ is equal to $1 - \nu$ on an interval, then $\mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma)$ is also constant on that interval in view of Theorem 1–(b). Therefore, there is no ambiguity in (8.30). Plugging (8.32) into (8.20)–(8.21), we verify that $\phi_{\gamma^\star, \lambda^\star}$ achieves $\pi_{1|\mathbb{0}}(\gamma^\star, \lambda^\star) = \nu$ and $\pi_{\mathbb{0}|1}(\gamma^\star, \lambda^\star)$ given by (8.33)–(8.34). With the aid of (4.24) we can express the function within {} in (8.35) as

$$f_\nu(\gamma) = \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma) + \exp(\gamma) \left(1 - \nu - \overline{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(\gamma) \right) \tag{8.41}$$

$$= (1 - \nu) \exp(\gamma) - \frac{1}{\log \mathrm{e}} \int_{-\infty}^{\gamma} \exp(t)\, \overline{\mathbb{F}}_{X \| Y}(t)\, \mathrm{d}t. \tag{8.42}$$

Its right- and left-derivatives at $\gamma \in \mathbb{R}$ are

$$\dot{f}_\nu^+(\gamma) = \lim_{\epsilon \downarrow 0} \frac{f_\nu(\gamma + \epsilon) - f_\nu(\gamma)}{\epsilon} = \left(1 - \nu - \overline{\mathbb{F}}_{X\|Y}(\gamma)\right) \frac{\exp(\gamma)}{\log \mathrm{e}}, \tag{8.43}$$

$$\dot{f}_\nu^-(\gamma) = \lim_{\epsilon \downarrow 0} \frac{f_\nu(\gamma) - f_\nu(\gamma - \epsilon)}{\epsilon} = (1 - \nu - \lim_{x \uparrow \gamma} \overline{\mathbb{F}}_{X\|Y}(x)) \frac{\exp(\gamma)}{\log \mathrm{e}}, \tag{8.44}$$

respectively. Consequently,

i. $\dot{f}_\nu^+(\gamma) > 0$ and $\dot{f}_\nu^-(\gamma) > 0$ at those $\gamma \in \mathbb{R}$ such that $\overline{\mathbb{F}}_{X\|Y}(\gamma) < 1 - \nu$;
ii. $\dot{f}_\nu^+(\gamma) < 0$ and $\dot{f}_\nu^-(\gamma) < 0$ at those $\gamma \in \mathbb{R}$ such that $\lim_{x\uparrow\gamma} \overline{\mathbb{F}}_{X\|Y}(x) > 1 - \nu$;
iii. If $\overline{\mathbb{F}}^{-1}_{P_1\|P_\emptyset}(1 - \nu) \neq \emptyset$, then $\dot{f}_\nu^+(\gamma^\star) = \dot{f}_\nu^-(\gamma^\star) = 0$ at any solution of (8.29).
iv. If $\overline{\mathbb{F}}^{-1}_{P_1\|P_\emptyset}(1-\nu) = \emptyset$, then at the unique $\gamma^\star$ that satisfies (8.31), $\dot{f}_\nu^-(\gamma^\star) > 0$ and $\dot{f}_\nu^+(\gamma^\star) < 0$.

Therefore, we have shown that the non-concave function to be maximized satisfies $f_\nu(\gamma) < f_\nu(\gamma^\star)$ for any γ that does not satisfy either (8.29) or (8.31). The fact that $f_\nu(\gamma^\star) = \alpha_\nu(P_1, P_\emptyset)$ follows from (8.29)–(8.30) if $\overline{\mathbb{F}}^{-1}_{P_1\|P_\emptyset}(1 - \nu) \neq \emptyset$, and (8.34) otherwise.

(d) Recalling Item 41,

$$(P_1, P_\emptyset) \equiv (Q_1, Q_\emptyset) \Longleftrightarrow \left\{\mathbb{F}_{P_1\|P_\emptyset} = \mathbb{F}_{Q_1\|Q_\emptyset} \text{ and } \overline{\mathbb{F}}_{P_1\|P_\emptyset} = \overline{\mathbb{F}}_{Q_1\|Q_\emptyset}\right\} \tag{8.45}$$

$$\Longrightarrow C(P_1, P_\emptyset) = C(Q_1, Q_\emptyset), \tag{8.46}$$

where (8.46) follows from (c). To prove the reverse implication, we must show that the function $\alpha_\nu(P_1, P_\emptyset)$ determines $\mathbb{F}_{P_1\|P_\emptyset}$ and $\overline{\mathbb{F}}_{P_1\|P_\emptyset}$. The explicit dependence is given in Theorem 22 in the Appendix.

(e) Recalling the symmetry property in Theorem 19-(c),

$$|C(P_1, P_\emptyset)| = 1 - 2 \int_0^{\underline{\pi}_{1|\emptyset}} \alpha_\nu(P_1, P_\emptyset)\, \mathrm{d}\nu. \tag{8.47}$$

Because of the convexity of α_ν (Item 89), its derivative is a non-decreasing function which may have at most a countable number of discontinuities on the interval $[0, \underline{\pi}_{1|\emptyset}]$. We partition the integral in (8.47) as the finite or countably infinite sum of subintegrals of differentiable sections, distinguishing between the sections in which α_ν is a straight line (corresponding to jumps in the relative information spectra) and those in which it is not. Recall that the non-straight-line sections are due to portions of the relative information spectra that are strictly monotonically increasing. Flat portions in the spectra only affect the kinks—points of discontinuous derivative—in α_ν, which do not contribute to its integral. Therefore, we have

$$\int_0^{\underline{\pi}_{1|\emptyset}} \alpha_\nu\, \mathrm{d}\nu = \sum_{\gamma\in\Gamma} \int_{\nu^-(\gamma)}^{\nu^+(\gamma)} \alpha_\nu\, \mathrm{d}\nu + \sum_{i\in\mathcal{I}} \int_{\nu_i}^{\nu_{i+1}} \alpha_\nu\, \mathrm{d}\nu, \tag{8.48}$$

where Γ is the finite, or countably infinite, set of abscissas at which the jumps in the relative information spectra occur, α_ν is a straight line on the intervals $[\nu^-(\gamma), \nu^+(\gamma)]$, and α_ν is differentiable but not a straight line on the intervals $[\nu_i, \nu_{i+1}]$.

i. We saw in (d) that each $\gamma \in \Gamma$ contributes a straight line in the fundamental tradeoff function of slope $-\exp(\gamma)$ between the abscissas $\nu^{-}(\gamma) = 1 - \bar{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(\gamma)$ and $\nu^{+}(\gamma) = 1 - \lim_{x \uparrow \gamma} \bar{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(x)$. In view of (8.33), observe that

$$\alpha_{\nu^{-}(\gamma)} = \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma), \tag{8.49}$$

$$\alpha_{\nu^{-}(\gamma)} - \alpha_{\nu^{+}(\gamma)} = \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_1) = \gamma]. \tag{8.50}$$

Therefore, the trapezoidal area is

$$\int_{\nu^{-}(\gamma)}^{\nu^{+}(\gamma)} \alpha_\nu \, \mathrm{d}\nu = \tfrac{1}{2} \left(\alpha_{\nu^{-}(\gamma)} + \alpha_{\nu^{+}(\gamma)} \right) \left(\nu^{+}(\gamma) - \nu^{-}(\gamma) \right) \tag{8.51}$$

$$= \left(\mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma) - \tfrac{1}{2} \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_1) = \gamma] \right) \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}}) = \gamma]. \tag{8.52}$$

Then, the sum of the subintegrals (8.52) due to the straight-line segments equals

$$\sum_{\gamma \in \Gamma} \int_{\nu^{-}(\gamma)}^{\nu^{+}(\gamma)} \alpha_\nu \, \mathrm{d}\nu = \sum_{\gamma \in \Gamma} \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma) \, \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}}) = \gamma] - \tfrac{1}{2} \, \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_1) = \imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}})], \tag{8.53}$$

where $Y_{\mathbb{0}}$ and Y_1 are independent.

ii. For a section between ν_0 and ν_1 on which α_ν is differentiable and not a straight line, the parametric solution in (8.18)–(8.19) reduces to

$$\alpha_\nu = \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(\gamma), \tag{8.54}$$

$$1 - \nu = \bar{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(\gamma), \tag{8.55}$$

whose definite integral can be written as the Lebesgue-Stieltjes integral

$$\int_{\nu_0}^{\nu_1} \alpha_\nu \, \mathrm{d}\nu = \int_{\bar{\mathbb{F}}^{-1}_{P_1 \| P_{\mathbb{0}}}(1-\nu_1)}^{\bar{\mathbb{F}}^{-1}_{P_1 \| P_{\mathbb{0}}}(1-\nu_0)} \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(t) \, \mathrm{d}\bar{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(t). \tag{8.56}$$

Summing (8.53) and all subintegrals of the non-straight-line portions in (8.56) yields

$$\int_0^{\underline{\pi}_{1|\mathbb{0}}} \alpha_\nu(P_1, P_{\mathbb{0}}) \, \mathrm{d}\nu = \int_{-\infty}^{\infty} \mathbb{F}_{P_1 \| P_{\mathbb{0}}}(t) \, \mathrm{d}\bar{\mathbb{F}}_{P_1 \| P_{\mathbb{0}}}(t) - \tfrac{1}{2} \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_1) = \imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}})] \tag{8.57}$$

$$= \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_1) \leq \imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}})] - \tfrac{1}{2} \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_1) = \imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}})]. \tag{8.58}$$

Plugging (8.58) into (8.47) yields

$$|C(P_1, P_{\mathbb{0}})| = \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}}) \leq \imath_{P_1 \| P_{\mathbb{0}}}(Y_1)] - \mathbb{P}[\imath_{P_1 \| P_{\mathbb{0}}}(Y_1) \leq \imath_{P_1 \| P_{\mathbb{0}}}(Y_{\mathbb{0}})] \tag{8.59}$$

$$= \tfrac{1}{2} |P_1 \otimes P_{\mathbb{0}} - P_{\mathbb{0}} \otimes P_1|, \tag{8.60}$$

in light of Theorem 8-(i).

(f) $\Longleftarrow$ (6.4) and (8.38).

□

99. A folk theorem (e.g., [63,64]) is that the *area under the curve* (Item 92) is the "probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance". The ambiguity in whether "higher" means $\geq$ or $>$ is inconsequential if the relative information spectra are continuous. Otherwise, we must split the difference as (8.7) together with Theorem 21-(f) yields, with $(Y_0, Y_1) \sim P_0 \otimes P_1$,

$$\int_0^1 (1 - \alpha_\nu(P_1, P_0))\, \mathrm{d}\nu = \mathbb{P}[\imath_{P_1\|P_0}(Y_1) > \imath_{P_1\|P_0}(Y_0)] + \tfrac{1}{2}\mathbb{P}[\imath_{P_1\|P_0}(Y_1) = \imath_{P_1\|P_0}(Y_0)]. \tag{8.61}$$

100. A corollary to a result by Pfanzagl [61] is that $C(P_1, P_0) = C(Q_1, Q_0)$ (in the notation of Theorems 11 and 12) is a sufficient condition for Z to be a sufficient statistic of Y for $\{P_0, P_1\}$. In fact, Theorems 16 and 21-(e) imply that the preservation of the fundamental tradeoff region in hypothesis testing is an equivalent criterion for pairwise sufficiency. Therefore, Theorem 18-(n) will follow from

$$C(P_1, P_0) = C(Q_1, Q_0)$$

$$\Updownarrow \tag{8.62}$$

$$|C(P_1, P_0)| = |C(Q_1, Q_0)|$$

$$\Updownarrow \tag{8.63}$$

$$S(P_1, P_0) = S(Q_1, Q_0).$$

To justify (8.62), recall from Item 93 that $C(Q_1, Q_0) \subset C(P_1, P_0)$. Therefore, $|C(P_1, P_0)| > |C(Q_1, Q_0)|$ unless $C(P_1, P_0) = C(Q_1, Q_0)$. Theorem 21-(f) implies (8.63).

9. Conclusions

One of the defining features of information theory is the study of random variables such as $\imath_X(X) = \log \frac{1}{P_X(X)}$, $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$, where the probability mass function of X is evaluated at X and the log density function of P_X with respect to P_Y is evaluated at X or Y, respectively. The averages of those random variables, entropy and relative entropy, are the pillars that sustain the asymptotics of the fundamental limits in data compression, hypothesis testing, and data transmission in stationary ergodic models. Beyond averages, the study of the distributions of those random variables, also known as information spectra and relative information spectra, is the key to non-asymptotic fundamental limits.

This paper has studied the relative information spectra for arbitrary pairs of probability measures defined on the same measurable space. To that end, the formalization of the concepts of relative support and coefficient of absolute discontinuity has proven valuable. Particular emphasis has been placed on the interplay of the distributions of $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$, which determine each other, as well as their relationships with measures of discrepancy such as total variation distance, relative entropy, Rényi divergence and f-divergences. Equivalent pairs of probability measures (possibly belonging to different measurable spaces) are those with identically-distributed relative informations.

The exposition of the applications to statistical inference has emphasized their connections to the literature. Based on equivalent pairs, we have introduced the conceptually simple notion of I-sufficiency, which can be checked easily even without the usual assumptions of deterministic statistics and dominated collections on standard spaces. When those assumptions are satisfied, the necessary and sufficient condition given by the Halmos-Savage factorization necessary and sufficient condition (Theorem 9) remains the gold standard for verifying the sufficiency of deterministic transformations.

The non-asymptotic (Neyman-Pearson) fundamental tradeoff region of conditional error probabilities in binary hypothesis testing is a major application of the relative information spectra. We have given a detailed description of the region without any assumptions of absolute continuity. The *area* of the Neyman-Pearson tradeoff region is a normalized measure of the discrepancy between the probability measures, equal to zero [resp., one] for identical [resp., orthogonal] probability measures, which is popular in applications in a slightly modified form referred to as the *area under the curve* (AUC). We have shown that the area of the Neyman-Pearson tradeoff region is equal to (one-half) the NP-*divergence*, $|P_0 \otimes P_1 - P_1 \otimes P_0|$, a new discrepancy measure between probability measures P_0 and P_1. Along with Chernoff information, it appears to be one of the most interesting divergences among those that satisfy the data processing inequality but are not f-divergences. We have shown that the preservation of the NP-*divergence* is a necessary and sufficient condition for the statistic to be sufficient. An immediate operational role is inherited from total variation distance, as the NP-divergence governs the error probability of the Bayesian test that identifies the order of a pair of observations, one drawn from P_0 and the other from P_1. A new asymptotic operational role for the *Bhattachrayya distance* has been shown for independent identically distributed observations: The rate of approach to 1 of the area of the fundamental non-Bayesian tradeoff region decays exponentially in twice the number of observations times the Bhattachrayya distance. In contrast, as shown in [14] in the Bayesian setting, the exponential decay of the minimum error probability is governed by the *Chernoff information* regardless of the values of the nonzero a priori probabilities.

A. Appendix 1: Relative information spectra from fundamental tradeoff function

On account of the convexity of α_ν, its derivative on $(0, \underline{\pi}_{1|0})$ is negative monotonic non-decreasing with a finite, or countably infinite, number of discontinuities. Those discontinuities determine the locations of the jumps of $\mathbb{F}_{P_1\|P_0}$, which are the same as those of $\overline{\mathbb{F}}_{P_1\|P_0}$. For $\nu \in (0, \underline{\pi}_{1|0})$, denote the left/right derivatives by

$$\dot{\alpha}_\nu^- = \lim_{\epsilon \downarrow 0} \frac{\alpha_\nu - \alpha_{\nu-\epsilon}}{\epsilon} \leq \lim_{\epsilon \downarrow 0} \frac{\alpha_{\nu+\epsilon} - \alpha_\nu}{\epsilon} = \dot{\alpha}_\nu^+ < 0. \tag{A.1}$$

Naturally, we drop the superscript whenever $\dot{\alpha}_\nu^- = \dot{\alpha}_\nu^+$. The following result gives $\mathbb{F}_{P_1\|P_0}$ and $\overline{\mathbb{F}}_{P_1\|P_0}$ as a function of $\{\alpha_\nu, \nu \in [0, \underline{\pi}_{1|0})\}$, with $\underline{\pi}_{1|0} = \max\{\nu \in [0,1] \colon \alpha_\nu = 0\}$, as per (8.11). The fact that the relative information spectrum and the fundamental tradeoff region determine each other validates the opening sentence in the abstract.

Theorem 22.

1. $\lim_{t\to\infty} \mathbb{F}_{P_1\|P_0}(t) = \underline{\pi}_{0|1} = \alpha_0$.
2. $\lim_{t\to-\infty} \overline{\mathbb{F}}_{P_1\|P_0}(t) = 1 - \underline{\pi}_{1|0}$.
3. *Fix* $\gamma \in \mathbb{R}$. *To determine* $\mathbb{F}_{P_1\|P_0}(\gamma)$ *and* $\overline{\mathbb{F}}_{P_1\|P_0}(\gamma)$, *there are two possibilities.*

   (a) There is a unique $\bar{\nu}_\gamma \in (0, \underline{\pi}_{1|0})$ *such that*

$$\dot{\alpha}_{\bar{\nu}_\gamma}^- \leq -\exp(\gamma) \leq \dot{\alpha}_{\bar{\nu}_\gamma}^+. \tag{A.2}$$

Then, $\mathbb{F}_{P_1\|P_0}(\gamma) = \alpha_{\bar{\nu}_\gamma}$ *and* $\overline{\mathbb{F}}_{P_1\|P_0}(\gamma) = 1 - \bar{\nu}_\gamma$*.*

(b) Let $(\nu_\gamma^-, \nu_\gamma^+) \subset [0, \underline{\pi}_{1|0}]$ *be the largest open interval such that*

$$\dot{\alpha}_\nu = -\exp(\gamma),\ \textit{for}\ \nu \in (\nu_\gamma^-, \nu_\gamma^+). \tag{A.3}$$

Then, $\mathbb{F}_{P_1\|P_0}(\gamma) = \alpha_{\nu_\gamma^-}$ *and* $\overline{\mathbb{F}}_{P_1\|P_0}(\gamma) = 1 - \nu_\gamma^-$*. Furthermore,* $\mathbb{F}_{P_1\|P_0}$ *experiences a jump at* γ *of height* $\alpha_{\nu_\gamma^-} - \alpha_{\nu_\gamma^+}$*, while the jump at* $\overline{\mathbb{F}}_{P_1\|P_0}(\gamma)$ *has height* $\nu_\gamma^+ - \nu_\gamma^-$*.*

Proof.

1) $\Longleftarrow$ (8.10) and (8.39).

2) $\Longleftarrow$ (8.40).

3) As we saw in Theorem 21-(d), $\mathbb{F}_{P_1\|P_0}$ and $\overline{\mathbb{F}}_{P_1\|P_0}$ experience a jump at γ if and only if the function α_ν has a straight line such that case 3b) holds.

3a) Since $\mathbb{F}_{P_1\|P_0}$ and $\overline{\mathbb{F}}_{P_1\|P_0}$ are continuous at γ, Theorem 21-(d) gives

$$\alpha_\nu = \mathbb{F}_{P_1\|P_0}(\gamma), \tag{A.4}$$

$$1 - \nu = \overline{\mathbb{F}}_{P_1\|P_0}(\gamma). \tag{A.5}$$

At those $\nu \in (0, \underline{\pi}_{1|0})$ such that $\dot{\alpha}_\nu^- = \dot{\alpha}_\nu^+$, we can differentiate (A.4) and (A.5) with respect to ν and γ, respectively, to conclude, with the aid of (4.26), that

$$\dot{\alpha}_\nu = -\exp(\gamma). \tag{A.6}$$

If $\dot{\alpha}_\nu^- < \dot{\alpha}_\nu^+$, the discontinuity in the derivative is caused by the fact that there is an interval of values of γ on which both $\mathbb{F}_{P_1\|P_0}(\gamma)$ and $\overline{\mathbb{F}}_{P_1\|P_0}(\gamma)$ are constant; therefore, according to (A.4)–(A.5), those values of γ result in a single Pareto-optimal point $(\bar{\nu}_\gamma, \alpha_{\bar{\nu}_\gamma})$. The interval of values of γ is indeed (A.2) since any slope strictly lower than $\dot{\alpha}_{\bar{\nu}_\gamma}^-$, or strictly higher than $\dot{\alpha}_{\bar{\nu}_\gamma}^+$, corresponds to Pareto-optimal points other than $(\bar{\nu}_\gamma, \alpha_{\bar{\nu}_\gamma})$.

3b) Since $\mathbb{F}_{P_1\|P_0}$ and $\overline{\mathbb{F}}_{P_1\|P_0}$ experience a jump at γ, α_ν has a straight line with slope

$$-\frac{\mathbb{P}\left[\imath_{P_1\|P_0}(Y_1) = \gamma\right]}{\mathbb{P}\left[\imath_{P_1\|P_0}(Y_0) = \gamma\right]} = -\exp(\gamma), \tag{A.7}$$

according to (8.33)–(8.32) and (4.25). The range of abscissas ν of that straight line is given by (8.31), thereby indicating that

$$\nu_\gamma^- = 1 - \overline{\mathbb{F}}_{P_1\|P_0}(\gamma), \tag{A.8}$$

$$\nu_\gamma^+ = 1 - \lim_{x \uparrow \gamma} \overline{\mathbb{F}}_{P_1\|P_0}(x). \tag{A.9}$$

Again, according to (8.33)–(8.32), the corresponding ordinates of those points are

$$\alpha_{\nu_\gamma^-} = \mathbb{F}_{P_1\|P_0}(\gamma), \tag{A.10}$$

$$\alpha_{\nu_\gamma^+} = \lim_{x \uparrow \gamma} \mathbb{F}_{P_1\|P_0}(x). \tag{A.11}$$

□

Use of Generative-AI tools declaration

The author declares he has not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

I am indebted to guest-editor Prof. Igal Sason for graciously extending the invitation for this submission and his meticulous reading of the manuscript. Reference [43] was kindly brought to my attention by an anonymous reviewer.

Conflict of interest

The author declares no conflict of interest.

References

1. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, **27** (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

2. S. Kullback, R. A. Leibler, On information and sufficiency, *Ann. Math. Stat.*, **22** (1951), 79–86. https://doi.org/10.1214/aoms/1177729694

3. P. R. Halmos, L. J. Savage, Application of the Radon-Nikodym theorem to the theory of sufficient statistics, *Ann. Math. Stat.*, **20** (1949), 225–241. https://doi.org/10.1214/aoms/1177730032

4. R. M. Fano, *Class notes for course 6.574: Statistical theory of information*, Massachusetts Institute of Technology, Cambridge, Mass., 1953.

5. D. V. Lindley, On a measure of the information provided by an experiment, *Ann. Math. Stat.*, **27** (1956), 986–1005. https://doi.org/10.1214/aoms/1177728069

6. H. Chernoff, Large-sample theory: Parametric case, *Ann. Math. Stat.*, **27** (1956), 1–22. Available from: `https://www.jstor.org/stable/2236974.`

7. J. Neyman, E. S. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Philos. T. Roy. Soc. London Ser. A*, **231** (1933), 289–337. https://doi.org/10.1098/rsta.1933.0009

8. I. N. Sanov, On the probability of large deviations of random variables, *Mat. Sb.*, **42** (1957), 11–44. https://doi.org/10.2307/3197345

9. H. Cramér, Sur un nouveau théorème-limite de la théorie des probabilités, *Actual. Sci. Ind.*, **736** (1938), 5–23.

10. E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev. Ser. II*, **106** (1957), 620–630. https://doi.org/10.1103/PhysRev.106.620

11. E. T. Jaynes, Information theory and statistical mechanics II, *Phys. Rev. Ser. II*, **108** (1957), 171–190. https://doi.org/10.1103/PhysRev.108.171

12. S. Kullback, *Information theory and statistics*, Dover: New York, 1968.

13. A. Rényi, *On measures of information and entropy*, In: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press: Berkeley, California, 1961, 547–561.

14. H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Stat.*, **23** (1952), 493–507. https://doi.org/10.1214/aoms/1177729330

15. I. Csiszár, Information-type measures of difference of probability distributions and indirect observations, *Stud. Sci. Math. Hung.*, **2** (1967), 299–318. https://doi.org/10.1016/S0010-8545(00)80126-5

16. K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *London Edinb. Dublin Philos. Mag. J. Sci.*, **50** (1900), 157–175. https://doi.org/10.1080/14786440009463897

17. H. Jeffreys, An invariant form for the prior probability in estimation problems, *Proc. Roy. Soc. London Ser. A Math. Phys. Sci.*, **186** (1946), 453–461. https://doi.org/10.1098/rspa.1946.0056

18. I. Vincze, *On the concept and measure of information contained in an observation*, In: Contributions to Probability: A Collection of Papers Dedicated to Eugene Lukacs, Academic Press: New York, 1981, 207–214. https://doi.org/10.1016/0091-3057(81)90179-9

19. L. Le Cam, *Asymptotic methods in statistical decision theory*, Springer: New York, 1986.

20. M. H. DeGroot, Uncertainty, information, and sequential experiments, *Ann. Math. Stat.*, **33** (1962), 404–419. https://doi.org/10.1214/aoms/1177704567

21. T. S. Han, S. Verdú, Approximation theory of output statistics, *IEEE T. Inform. Theory*, **39** (1993), 752–772. https://doi.org/10.1109/18.256486

22. T. S. Han, *Information spectrum methods in information theory*, Springer: Heidelberg, Germany, 2003.

23. Y. Polyanskiy, H. V. Poor, S. Verdú, Channel coding rate in the finite blocklength regime, *IEEE T. Inform. Theory*, **56** (2010), 2307–2359. https://doi.org/10.1109/TIT.2010.2043769

24. S. Verdú, The Cauchy distribution in information theory, *Entropy*, **25** (2023), 1–48. https://doi.org/10.3390/e25010048

25. D. Burkholder, Sufficiency in the undominated case, *Ann. Math. Stat.*, **32** (1961), 1191–1200. https://doi.org/10.1214/aoms/1177704859

26. P. R. Halmos, *Measure theory*, Springer: New York, 1974.

27. P. Billingsley, *Probability and measure*, 4 Eds., Wiley-Interscience: New York, 2012.

28. I. Csiszár, J. Körner, *Information theory: Coding theorems for discrete memoryless systems*, Academic: New York, 1981.

29. J. Bhattacharyya, On some analogues of the amount of information and their use in statistical estimation, *Sankhyā Indian J. Stat.*, **8** (1946), 1–14.

30. T. van Erven, P. Harremoës, Rényi divergence and Kullback-Leibler divergence, *IEEE T. Inform. Theory*, **60** (2014), 3797–3820. https://doi.org/10.1109/TIT.2014.2320500

31. A. Rényi, New version of the probabilistic generalization of the large sieve, *Acta Math. Hung.*, **10** (1959), 217–226. https://doi.org/10.1007/BF02063300

32. I. Csiszár, Eine Informationstheorische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Publ. Math. Inst. Hung. Acad. Sci.*, **8** (1963), 85–108. https://real.mtak.hu/201426/

33. S. M. Ali, S. D. Silvey, A general class of coefficients of divergence of one distribution from another, *J. Roy. Stat. Soc. Ser. B*, **28** (1966), 131–142. https://doi.org/10.2307/4441277

34. I. Sason, On f-divergences: Integral representations, local behavior, and inequalities, *Entropy*, **20** (2018), 1–32. https://doi.org/10.3390/e20010032

35. F. Liese, I. Vajda, *f-divergences: Sufficiency, deficiency and testing of hypotheses*, In: Advances in Inequalities from Probability Theory and Statistics, Nova Science: New York, 2008, 113–158.

36. I. Sason, S. Verdú, f-divergence inequalities, *IEEE T. Inform. Theory*, **62** (2016), 5973–6006. https://doi.org/10.1109/TIT.2016.2603151

37. S. Vajda, *Theory of statistical inference and information*, Kluwer: Dordrecht, The Netherlands, 1989.

38. I. Csiszár, *Information measures: A critical survey*, In: Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1974, 73–86. https://doi.org/10.1111/j.1559-3584.1974.tb03703.x

39. F. Oesterreicher, I. Vajda, Statistical information and discrimination, *IEEE T. Inform. Theory*, **39** (1993), 1036–1039. https://doi.org/10.1109/18.256536

40. F. Liese, I. Vajda, On divergences and informations in statistics and information theory, *IEEE T. Inform. Theory*, **52** (2006), 4394–4412. https://doi.org/10.1109/TIT.2006.881731

41. F. Liese, ϕ-divergences, sufficiency, Bayes sufficiency, and deficiency, *Kybernetika*, **48** (2012), 690–713. Available from: `https://www.kybernetika.cz/content/2012/4/690.`

42. S. Verdú, *Total variation distance and the distribution of relative information*, In: Proceedings of the 2014 Workshop on Information Theory and Applications, University of California: La Jolla, California, 2014.

43. A. Kontorovich, Obtaining measure concentration from Markov contraction, *Markov Process. Relat.*, **18** (2012), 613–638.

44. V. Strassen, The existence of probability measures with given marginals, *Ann. Math. Stat.*, **36** (1965), 423–439. https://doi.org/10.1214/aoms/1177700153

45. R. L. Dobrushin, Prescribing a system of random variables by conditional distributions, *Theor. Probab. Appl.*, **15** (1970), 458–486. https://doi.org/10.1137/1115049

46. Y. Polyanskiy, S. Verdú, *Arimoto channel coding converse and Rényi divergence*, In: Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing, University of Illinois: Monticello, Illinois, 2010, 1327–1333.

47. R. A. Fisher, On the mathematical foundations of theoretical statistics, *Proc. Roy. Soc. London Ser. A Math. Phys. Sci.*, **222** (1922), 309–368. https://doi.org/10.1098/rsta.1922.0009

48. D. Blackwell, Equivalent comparisons of experiments, *Ann. Math. Stat.*, **24** (1953), 265–272. https://doi.org/10.1214/aoms/1177729032

49. R. R. Bahadur, Sufficiency and statistical decision functions, *Ann. Math. Stat.*, **25** (1954), 423–462. https://doi.org/10.1214/aoms/1177728715

50. D. Blackwell, R. V. Ramamoorthi, A Bayes but not classically sufficient statistic, *Ann. Stat.*, **10** (1982), 1025–1026. https://doi.org/10.1016/0305-750X(82)90014-6

51. J. Dieudonné, Sur le théoréme de Lebesgue-Nikodym, *Ann. Math.*, **42** (1941), 547–555. https://doi.org/10.1016/S0002-9378(16)40717-9

52. H. Heyer, *Theory of statistical experiments*, Springer: New York, 1982.

53. J. Neyman, Su un teorema concernente le cosiddette statistiche sufficienti, *Istituto Italiano degli Attuari*, **6** (1935), 320–334.

54. T. P. Speed, A note on pairwise sufficiency and completions, *Sankhyā Indian J. Stat. Ser. A*, **38** (1976), 194–196.

55. A. N. Kolmogorov, Definition of center of dispersion and measure of accuracy from a finite number of observations, *Izv. Akad. Nauk SSSR Ser. Mat.*, **6** (1942), 4–32.

56. T. M. Cover, J. A. Thomas, *Elements of information theory*, 2 Eds., Wiley: New York, 2006.

57. D. Blackwell, *Comparison of experiments*, In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press: Berkeley, California, **18** (1951), 93–102. https://doi.org/10.2307/1438094

58. R. D. Reiss, *Approximate distributions of order statistics: With applications to nonparametric statistics*, Springer: New York, 2012.

59. H. Strasser, *Mathematical theory of statistics: Statistical experiments and asymptotic decision theory*, Walter de Gruyter: Berlin, 1985.

60. R. R. Bahadur, A characterization of sufficiency, *Ann. Math. Stat.*, **26** (1955), 286–293. https://doi.org/10.1214/aoms/1177728545

61. J. Pfanzagl, A characterization of sufficiency by power functions, *Metrika*, **21** (1974), 197–199. https://doi.org/10.1080/0156655740210307

62. H. L. van Trees, *Detection, estimation and modulation theory. 1. Detection, estimation and linear modulation theory*, John Wiley, 1968.

63. D. J. Hand, R. J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, *Mach. Learn.*, **45** (2001), 171–186. https://doi.org/10.1023/A:1010920819831

64. T. Fawcett, An introduction to ROC analysis, *Pattern Recogn. Lett.*, **27** (2006), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)