Information Geometry
https://doi.org/10.1007/s41884-024-00138-z

RESEARCH

A historical perspective on Schützenberger-Pinsker inequalities (extended version)

Olivier Rioul[1]

Received: 20 February 2024 / Revised: 11 May 2024 / Accepted: 18 May 2024

Abstract

This paper presents a tutorial overview of so-called Pinsker inequalities which establish a precise relationship between information and statistics, and whose use have become ubiquitous in many applications. According to Stigler's law of eponymy, no scientific discovery is named after its original discoverer. Pinsker's inequality is no exception: Years before the publication of Pinsker's book in 1960, the French medical doctor, geneticist, epidemiologist, and mathematician Marcel-Paul (Marco) Schützenberger, in his 1953 doctoral thesis, not only proved what is now called Pinsker's inequality (with the optimal constant that Pinsker himself did not establish) but also the optimal second-order improvement, more than a decade before Kullback's derivation of the same inequality. We review Schützenberger and Pinsker contributions as well as those of Volkonskii and Rozanov, Sakaguchi, McKean, Csiszár, Kullback, Kemperman, Vajda, Bretagnolle and Huber, Krafft and Schmitz, Toussaint, Reid and Williamson, Gilardoni, as well as the optimal derivation of Fedotov, Harremoës, and Topsøe. We also present some historical elements on the life and work of Schützenberger, and discuss an interesting problem of an erroneous constant in the Schützenberger-Pinsker inequality.

Keywords Pinsker inequality · Total variation · Kullback–Leibler divergence · Statistical distance · Mutual information · Data processing inequality

Communicated by Frank Nielsen.

This paper is an extended version of the work presented at the 6th international conference on Geometric Science of Information (GSI 2023), Saint Malo, France, Aug. 30-Sept. 1, 2023.

✉ Olivier Rioul
olivier.rioul@telecom-paris.fr

[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France

Published online: 30 May 2024

1 Introduction

How far is one probability distribution from another? This question finds many different answers in information geometry, statistics, coding and information theory, cryptography, game theory, learning theory, and even biology or social sciences. Distances and divergences are often used to quantify how close two distributions may be, and it is particularly interesting to establish tight bounds between distances and divergences.

1.1 Preliminaries and notations

We assume that all considered probability distributions over a given measurable space $(\Omega, \mathcal{A})$ admit a σ-finite *dominating measure* μ, with respect to which they are absolutely continuous. This can always be assumed when considering finitely many distributions. For example, p and q admit $\mu = \frac{p+q}{2}$ as a dominating measure since $p \ll \mu$ and $q \ll \mu$. By the Radon-Nikodym theorem, they admit *densities* (Radon-Nikodym derivatives) with respect to μ, which we again denote by p and q, respectively. This ambiguity in notation should be easily resolved from the context.

For any event $A \in \mathcal{A}$, $p(A) = \int_A p \,\mathrm{d}\mu = \int_A p(x)\,\mathrm{d}\mu(x)$, and similarly for q. This is an overload in notations and one should not confuse $p(\{x\})$ with $p(x)$. Two distributions p, q are equal if $p(A) = q(A)$ for all $A \in \mathcal{A}$, that is, $p = q$ μ-a.e. in terms of densities.

If μ is a counting measure, then p is a discrete probability distribution with $\int_A p \,\mathrm{d}\mu = \sum_{x\in A} p(x)$, where the density $p(x)$ is a p.m.f. (probability mass function); if μ is a Lebesgue measure, then p is a continuous probability distribution with $\int_A p \,\mathrm{d}\mu = \int_A p(x)\,\mathrm{d}x$, where the density $p(x)$ is a p.d.f. (probability density function). For short we simply write $\int p = \int p \,\mathrm{d}\mu$. When p, q are one-dimensional distributions defined over $\mathbb{R}$, the corresponding c.d.f.'s (cumulative distribution functions) are denoted by uppercase letters P, Q.

We also consider the important case where p and q are binary (Bernoulli) distributions with parameters again denoted p and q, respectively. Thus for $p \sim \mathcal{B}(p)$ we have $p(x) = p$ or $1 - p$. Again this ambiguity in notation should be easily resolved from the context.

The logarithm (log) is considered throughout this paper in *any* base. Similarly, the exponential ($\exp = \log^{-1}$) is relative to the base considered, e.g., natural exponential $\exp x = e^x$ and natural logarithm $\log x = \ln x$ in base e.

We use the following notations for sets $\{p < q\} = \{x \;;\; p(x) < q(x)\}$, etc., minimum $p \wedge q$, maximum $p \vee q$, positive and negative parts $a^+ = a \vee 0$, $a^- = (-a) \vee 0 = -(a \wedge 0)$, with decompositions $p + q = p \wedge q + p \vee q$, $a = a^+ - a^-$, and $|a| = a^+ + a^-$.

1.2 Distances between distributions

In order to quantify how close two distributions are, a common viewpoint is to define a "distance" $\Delta(p, q)$ between probability distributions p and q, which should at least

Table 1 Some distances and the corresponding types of convergence they metrize

Distance $\Delta(p, q)$	Convergence
Lévy [2]-Prokhorov [3] $\inf\{\epsilon > 0 \mid \forall A,\, p(A) \leq q(A^\epsilon) + \epsilon\}$	
where A^ϵ = Borel set of points at distance $\leq \epsilon$ from A	Weak (in distribution)
Fortet-Mourier [4] $\sup_{\|f\| \leq 1} \left\| \int f(p - q) \right\|$ (bounded Lipschitz f)	
a.k.a. "Wasserstein" [5] or Kantorovich [6]-Rubinstein [7]	Weak (in distribution)
Kolmogorov [8]-Smirnov [9] $\|P - Q\|_\infty$	Uniform (in distribution)
Radon	Strong
Total variation $\frac{1}{2} \int \|p - q\| = \frac{1}{2} \|p - q\|_1$	L^1
Hellinger[a] (Jeffreys [10]) $\sqrt{\frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2} = \|\sqrt{p} - \sqrt{q}\|_2 / \sqrt{2}$	L^1
"Jensen-Shannon" [11] $\sqrt{\int p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q}}$	L^1
Vincze [12]-Le Cam [13] $\sqrt{\frac{1}{2} \int \frac{(p-q)^2}{p+q}} = \| \frac{p-q}{p+q} \|_2$ norm w.r.t. $\mu = \frac{p+q}{2}$	L^1
Ky Fan [14]	In probability

[a] What is generally known as the "Hellinger distance" was in fact introduced by Jeffreys in 1946 [10]. The Hellinger integral (1909) [15] is just a general method of integration that can be used to define the Jeffreys distance. The Jeffreys ("Hellinger") distance should not be confused with the "Jeffreys divergence", which was studied by Kullback as a symmetrized Kullback–Leibler divergence (see below)

satisfy the basic property that it is *nonnegative* and *vanishes only when the two probability distributions coincide*: $p = q$ in the given statistical manifold [1]. Strictly speaking, distances $\Delta(p, q)$ should also satisfy the two usual requirements of *symmetry* $\Delta(p, q) = \Delta(q, p)$ and *triangle inequality* $\Delta(p, q) + \Delta(q, r) \geq \Delta(p, r)$. In this case the probability distribution space becomes a *metric space*.

Examples of (metric) distances, along with the corresponding convergences that they metrize are given in Table 1. Some other types of convergence can also be metrized, but by distances between *random variables* rather than between distributions. For example, the Ky Fan distance metrizes convergence in probability, which is stronger than convergence in distribution.

Some distances between distributions are "strongly" equivalent, in the sense that they satisfy inequalities in both directions such as $\frac{2}{3}\Delta_{LP}^2(p, q) \leq \Delta_{FM}(p, q) \leq 2\Delta_{LP}(p, q)$ for Lévy-Prokhorov and Fortet-Mourier distances, which both metrize weak convergence (convergence in distribution). Also, $\frac{1}{2}\Delta_H^2(p, q) \leq \Delta_{TV}(p, q) \leq \Delta_H(p, q)$ for Hellinger and total variation distances, $\Delta_H(p, q) \leq \Delta_{VC}(p, q) \leq \sqrt{2}\Delta_H(p, q)$ for Hellinger and Vincze-Le Cam distances [13, § 4.2], $\Delta_{VC}^2(p, q) \leq \Delta_{TV}(p, q) \leq \Delta_{VC}(p, q)$ for Vincze-Le Cam and total variation distances and $\Delta_{VC}(p, q)\sqrt{\log e} \leq \Delta_{JS}(p, q) \leq \Delta_{VL}(p, q)\sqrt{2 \log 2}$ for Vincze-Le Cam and Jensen-Shannon distances [16]. Thus total variation, Hellinger, Jensen-Shannon and Vincze-Le Cam distances are all strongly equivalent and all define the same topology.

It is well known that L^1 convergence of densities implies convergence in distribution (see, e.g., [17]). For example in the one-dimensional case of real random variables, taking $A = (-\infty, x]$ in Corollary 2 below gives $\sup_x |P(x) - Q(x)| \leq \Delta_{TV}(p, q)$ where P, Q denote c.d.f.s corresponding to p, q, respectively. Thus $\Delta_{TV}(p, q) \to 0$

implies $\|P - Q\|_\infty \to 0$ (Kolmogorov-Smirnov distance). This uniform convergence of c.d.f.s in turn implies pointwise convergence hence convergence in distribution.

In this paper, we focus on the *total variation distance*, implying the strongest type of convergence among the preceding examples. Arguably, it is also the simplest—as an L^1-norm distance—and the most frequently used in applications, particularly those related to Bayesian inference and statistical tests (see § 2 below).

1.3 Divergences from one distribution to another

In many information theoretic applications, other types of "distances", that do not necessarily satisfy the triangle inequality, are often preferred. Such "distances" are called *divergences* $D(p\|q)$. A formal definition [1, Def. 1.1] of a divergence $D(p\|q)$, defined for any p, q in a differentiable statistical manifold, is that it should not only be nonnegative and vanish only when $p = q$, but also that locally, $D(p\|p + dp)$ is a positive definite quadratic form for infinitesimal displacements dp from p. In other words, at $q = p$, the gradient $\partial_q D(p\|q)$ vanishes and the Hessian $\partial^2_{q,q} D(p\|q) > 0$ is positive definite.

Divergences, however, may not satisfy the symmetry property: In general, $D(p\|q)$ is the divergence *of* q *from* p, and not "between p and q". Evidently, such divergences can always be symmetrized by considering either $D(p\|q) + D(q\|p)$ or $D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2}) = D(p\|\mu) + D(q\|\mu)$ instead of $D(p\|q)$; but even so, divergences do not satisfy the triangle inequality in general.

Divergences were first introduced in statistics and information theory, in relation to the notion of entropy, and have found numerous applications. From their definition, they also provide the statistical manifold with a dually flat structure equipped with a Riemannian metric [1], making them fundamentally useful in information geometry.

Examples of divergences are given in Table 2. Two fairly general classes of divergences are the f-divergences and the Bregman divergences. The only f-divergence that is also a distance is the total variation distance [28]. However, some square roots of (symmetrized) f-divergences also yield genuine distances. For example, the Jeffreys ("Hellinger") distance is the square root of the Bhattacharyya divergence: $\Delta_H(p, q) = \sqrt{1 - \int \sqrt{pq}}$, the "Jensen-Shannon" distance is the square root of a symmetrized Kullback–Leibler divergence: $\Delta_{JS}(p, q) = \sqrt{D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2})}$, and the Vincze-Le Cam distance is the square root of a symmetrized χ^2 divergence: $\Delta_{VC}(p, q) = \sqrt{\frac{1}{2}(\chi^2(p\|\frac{p+q}{2}) + \chi^2(q\|\frac{p+q}{2}))}$. See [29] for a review of many examples.

In this paper, we focus on the *Kullback–Leibler divergence*, historically the most popular type of divergence which has become ubiquitous in information theory. It is the only divergence that is both a f-divergence and a Bregman divergence [30, Appendix D]. Two of the reasons of its popularity are its relation to Shannon's entropy—the Kullback–Leibler divergence is also known as the *relative entropy* (see Sect. 3); and the fact that it *tensorizes* nicely for products of probability distributions, expressed in terms of the sum of the individual divergences (see Prop. 6 below).

Table 2 Most divergences studied in the literature fall into three general classes: α-divergences ($\alpha > 0$), f-divergences (f convex), Bregman divergences (ϕ convex)

Rényi α-divergences [18] $D_\alpha(p\|q) = \frac{1}{\alpha-1}\log\int p^\alpha q^{1-\alpha}$
Kullback–Leibler [19] $D_{\mathrm{KL}}(p\|q) = \lim_{\alpha\to 1} D_\alpha(p\|q)$
Sundaresan [20] $D'_\alpha(p\|q) = D_{\frac{1}{\alpha}}(\frac{p^\alpha}{\int p^\alpha}\|\frac{q^\alpha}{\int q^\alpha})$
"Cauchy-Schwarz" [21, Eq. (31)] $D'_2(p\|q) = \log\frac{\int p^2 \int q^2}{(\int pq)^2}$
Csiszár f-divergences [22] $D_f(p\|q) = \int q f(\frac{p}{q})$
Kullback–Leibler [19] $D_{\mathrm{KL}}(p\|q) = \int p\log\frac{p}{q}$
Total variation $\frac{1}{2}\int \|p-q\|$
Pearson [23] $\chi^2(p\|q) = \int \frac{(p-q)^2}{q}$
Bhattacharyya [24] $1 - \int\sqrt{pq}$
Bregman divergences [25] $D_\phi(p\|q) = \phi(p) - \phi(q) - \nabla\phi(q)\cdot(p-q)$
Squared Euclidean $\int(p-q)^2$, squared Mahalanobis distance [26]
Itakura-Saito [27] $\int \frac{p}{q} - \log\frac{p}{q} - 1$
Kullback–Leibler [19] $D_{\mathrm{KL}}(p\|q)$
Kullback–Leibler divergence [19] is a member of all three

1.4 Pinsker inequalities

Definition 1 *(Pinsker Inequality)* A Pinsker-type inequality is any general inequality of the form

$$D \geq \varphi(\Delta) \tag{1}$$

relating divergence $D = D(p\|q)$ to distance $\Delta = \Delta(p, q)$ and holding for any probability distributions p and q. Here $\varphi(x)$ should assume positive values for $x > 0$ with $\varphi(0) = 0$ in accordance with the property that both $D(p\|q)$ and $\Delta(p, q)$ vanish only when $p = q$.

Typically φ is also increasing, differentiable, and often convex.

The existence of any such Pinsker inequality implies the following statement: Convergence in the sense of divergence D implies convergence in the sense of distance Δ. Intuitively, this means that the "topology" induced by D is *finer* than that induced by Δ. In fact, because $D(p\|q)$ is generally not symmetric in (p, q), it induces two separate topologies based on neighborhoods with respect to either the first or second argument (see, e.g., [31])—but not on both arguments as noticed in [32]. Convergence in either topology implies metric convergence in Δ:

$$D(p_n\|p) \to 0 \text{ or } D(p\|p_n) \to 0 \text{ implies } \Delta(p_n, p) = \Delta(p, p_n) \to 0 \tag{2}$$

as $n \to \infty$. The first type of convergence $D(p_n \| p) \to 0$ is often used to provide strong limit theorems, e.g., the entropic central limit theorem of Barron [17] for the Kullback-Leibler divergence, which is stronger than the usual central limit theorem in distribution.

If, in addition, a *reverse* Pinsker inequality $\Delta \geq \psi(D)$ holds, then the associated topologies are equivalent. Many Pinsker-type inequalities (direct or reverse) have been established, notably between f-divergences. See e.g., [33] and [34, § 7.5,7.6] for some examples.

In this paper, we present historical considerations of the classical *Pinsker inequality* where D is the Kullback–Leibler divergence and Δ is the total variation distance. This inequality is by far the most renowned inequality of its kind, and finds many applications, e.g., in statistics, information theory, and computer science. Many considerations in this paper, however, equally apply to other types of distances and divergences.

1.5 Outline

The remainder of this paper is organized as follows. Section 2 and 3 present basic properties of total variation distance and Kullback–Leibler divergence, respectively. Section 4 gives a motivating example illustrating the usefulness of Pinsker's inequality. Section 5 discusses the related notions of statistical distance and mutual information, which were originally considered by Pinsker. Some useful ingredients for proving Pinsker inequalities are presented in Sect. 6. The contributions from Pinsker and other authors in the 1960 s are reviewed in Sect. 7. Section 8 is devoted to Schützenberger's key contribution as well as some elements of his life and work. Other recent improvements of Pinsker's inequality are reviewed in Sects. 8 and 9. Finally, Sect. 10 discusses the *optimal* Schützenberger-Pinsker inequality and concludes.

This tutorial article is both historical and educational. For completeness, in all its sections, proofs are provided to illustrate the ideas. Most of these proofs and some of the statements are simplified versions of the original ones, particularly Prop. 26 and Theorem 27.

2 Total variation distance: basic properties

In this section, we review some basic definitions and properties of the total variation distance.

The *total variation distance* $\Delta(p, q)$ can be defined in two different ways. The simplest is the following

Definition 2 *(Total Variation Distance—First Definition)*

$$\Delta(p, q) \triangleq \frac{1}{2} \int |p - q| \, \mathrm{d}\mu, \tag{3}$$

that is, half the $L^1(\mu)$-norm of the difference of densities. In particular, the total variation distance between *binary* distributions $\mathcal{B}(p)$ and $\mathcal{B}(q)$ is simply

$$\delta(p,q) = \frac{|p-q| + |(1-p)-(1-q)|}{2} = |p-q|. \tag{4}$$

It is important to note that the above definition of $\Delta(p,q)$ is coherent in the sense that it does *not* depend on the choice of the dominating measure μ. Indeed, if $\mu \ll \mu'$, with density $\frac{d\mu}{d\mu'} = f$, then the densities w.r.t. μ' become $p' = pf$ and $q' = qf$ so that $\int |p'-q'| \, d\mu' = \int |p-q| \, d\mu$.

That Δ is a *distance* (metric) is obvious from this definition (apart from the $1/2$ factor, it is the L^1-norm distance).

Remark 1 (Random variable notation) Some authors (e.g., [35, Chap. 8]) define the "statistical distance" of two random variables $X \sim p_X$ and $Y \sim p_Y$ as

$$\Delta(X,Y) \triangleq \Delta(p_X, q_X).$$

Here $\Delta(X;Y)$ depends only on the (marginal) distributions of X and Y, not on their joint distribution. This is not to be confused with Definition 6 below.

Strictly speaking, Δ is *not* a distance between random variables X and Y since the fact that they share the same distribution does not necessarily imply $X = Y$ (or even $X = Y$ almost everywhere). However it becomes a distance on random variables if we agree to identify two *equivalent* random variables, that is, variables having the same distribution. In other words, Δ is then a distance on the quotient space of random variables modulo this equivalence relation. One always has symmetry $\Delta(X;Y) = \Delta(Y;X)$ and the triangle inequality $\Delta(X;Z) \le \Delta(X;Y) + \Delta(Y;Z)$.

Since $\int |p-q| \, d\mu = \int (p-q)^+ \, d\mu + \int (p-q)^- \, d\mu$ and $\int (p-q)^+ \, d\mu - \int (p-q)^- \, d\mu = \int (p-q) \, d\mu = 0$, taking half sum we may also write

$$\Delta(p,q) = \int (p-q)^+ \, d\mu = \int (p-q)^- \, d\mu \tag{5}$$

in term of positive and negative parts. Also, since $(p-q)^+ = p - p \wedge q = p \vee q - q$ and $(p-q)^- = q - p \wedge q = p \vee q - p$ we have

$$\Delta(p,q) = \int p \vee q \; d\mu - 1 = 1 - \int p \wedge q \; d\mu \tag{6}$$

in terms of the maximum and minimum.

Thus, the normalization factor $1/2$ in the definition ensures that $0 \le \Delta(p,q) \le 1$, with

- minimum value $\Delta(p,q) = 0$ if and only if $p = q$;
- maximum value $\Delta(p,q) = 1 - \int p \wedge q \; d\mu = 1$ if and only if $p \wedge q = 0$ μ-a.e., also noted $p \perp q$, that is, p and q have "non-overlapping" supports.

The alternate definition of the total variation distance is to proceed from the discrete case to the general case as follows.

Definition 3 *(Total variation distance—second definition)*

$$\Delta(p,q) \triangleq \frac{1}{2} \sup \sum_i |p(A_i) - q(A_i)|, \tag{7}$$

where the supremum is taken over all *partitions* of Ω into a countable number of (pairwise disjoint) $A_i \in \mathcal{A}$.

When $\Omega \subset \mathbb{R}$, this supremum can simply be taken over partitions of *intervals* A_i, and (apart from the factor $1/2$) this exactly corresponds to the usual notion of *total variation* of the corresponding cumulative distribution f of the signed measure $p-q$. This is a well-known measure of the one-dimensional arclength of the curve $y = f(x)$, introduced as the *oscillation totale* (total oscillation) by the French mathematician Camille Jordan [36] in the 19th century, and justifies the name "total variation" given to Δ.

Proposition 1 *The above two definitions of total variation coincide.*

Proof First, by the triangle inequality, the sum $\sum_i |p(A_i) - q(A_i)|$ in (7) can only increase by subpartitioning (refining the partition), hence (7) can be seen as a limit for finer and finer partitions. Second, consider the subpartition $A_i^+ = A_i \cap A^+$, $A_i^- = A_i \cap A^-$, where, say, $A^+ = \{p > q\}$ and $A^- = \{p \le q\}$, or more generally, any two complementary sets satisfying

$$\begin{cases} \{p > q\} \subseteq A_+ \subseteq \{p \ge q\} \\ \{p < q\} \subseteq A_- \subseteq \{p \le q\}. \end{cases}$$

Then the corresponding sum in (7) already equals

$$\begin{aligned} \sum_i (p-q)(A_i^+) + (q-p)(A_i^-) &= (p-q)\left(\sum_i A_i^+\right) + (q-p)\left(\sum_i A_i^-\right) \\ &= (p-q)(A^+) + (q-p)(A^-) \\ &= \int (p-q)^+ + (p-q)^- \,\mathrm{d}\mu \\ &= \int |p-q| \,\mathrm{d}\mu \end{aligned} \tag{8}$$

which is (3). □

A key property in the sequel is that the maximum in (7) is attained for *binary* partitions.

Corollary 2 *(Total variation distance—third definition)*

$$\Delta(p,q) = \sup_A |p(A) - q(A)| \tag{9}$$

(without the $1/2$ *factor).*

Proof From the above proof, the supremum in (7) is attained for binary partitions $\{A^+, A^-\}$ of the form $\{A, A^{\complement}\}$, so that $\Delta(p,q) = \frac{1}{2}\sup\big(|p(A) - q(A)| + |p(A^{\complement}) - q(A^{\complement})|\big)$, which is (9). □

An important consequence is the following result on *binary hypothesis testing* with two hypotheses on the distributions of some data X:

$$\begin{cases} \mathbf{H_0}: & X \sim p \\ \mathbf{H_1}: & X \sim q \end{cases} \tag{10}$$

with any possible deterministic test T such that the null hypothesis $\mathbf{H_0}$ is rejected if $X \in T$, accepted otherwise. The two types of error are

$$\begin{cases} \textbf{Type I}: \text{(false positive)} & p(X \in T) \\ \textbf{Type II}: \text{(false negative)} & q(X \notin T). \end{cases} \tag{11}$$

Proposition 3 *For any test* T*, the sum of type-1 and type-2 errors is lower bounded by*

$$p(X \in T) + q(X \notin T) \geq 1 - \Delta(p,q). \tag{12}$$

Proof Obvious from the inequality $\Delta(p,q) = \sup_A |p(A) - q(A)| \geq q(T) - p(T)$. □

Remark 2 (Statistical equivalence) This important property ensures that *a sufficiently small value of* $\Delta(p,q)$ *implies that no statistical test can effectively distinguish between the two distributions* p and q, since type-I or type-II errors have total probability $1 - \Delta$ arbitrarily close to one. Thus in this sense the two hypotheses p and q are Δ-undistinguishable.

For the case of independent observations we are faced with the evaluation of the total variation distance for products of distributions. In this situation, Pinsker's inequality is particularly useful since it relates it to the Kullback–Leibler divergence which nicely tensorizes, thus allowing a simple evaluation (see § 4 below).

3 Kullback–Leibler divergence: basic properties

The *Kullback–Leibler divergence* [19], also known as statistical divergence, or simply divergence, can similarly be defined in two different ways. One can first define:

Definition 4 *(Kullback–Leibler divergence—first definition)*

$$D(p\|q) \triangleq \int p \log \frac{p}{q} \, \mathrm{d}\mu. \tag{13}$$

Since $x \log x \geq -\frac{\log e}{e}$, the negative part of the integral is finite. Therefore, this integral is always meaningful and can be finite, or infinite $= +\infty$.

In particular, the divergence between binary distributions $\mathcal{B}(p)$ and $\mathcal{B}(q)$ is simply

$$d(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}. \tag{14}$$

Again note that the above definition of $D(p\|q)$ does *not* depend on the choice of the dominating measure μ. Indeed, if $\mu \ll \mu'$, with density $\frac{d\mu}{\mathrm{d}\mu'} = f$, then the densities w.r.t. μ' become $p' = pf$ and $q' = qf$ so that $\int p' \log \frac{p'}{q'} \, \mathrm{d}\mu' = \int p \log \frac{p}{q} \, \mathrm{d}\mu$.

Remark 3 (Relative entropy) The Kullback–Leibler divergence is also known in information theory as the *relative entropy* because it can be written as

$$D(p\|q) = \int p \log \frac{1}{q} \, \mathrm{d}\mu - \int p \log \frac{1}{p} \, \mathrm{d}\mu \tag{15}$$

where the first term is the cross entropy of q relative to p and the second term $H(p) = \int p \log \frac{1}{p} \, \mathrm{d}\mu$ is the Shannon entropy of p. Note that $H(p)$, contrary to $D(p\|q)$, does depend on the choice of the dominating measure. This explains in particular why differential entropy (when μ is the Lebesgue measure) is very different in nature from the discrete entropy (when μ is a counting measure). In particular, the differential entropy can be negative, and not invariant under invertible changes of variables [34, 37, 38].

Remark 4 (Absolute continuity of p w.r.t. q) The (Kullback–Leibler) divergence is traditionally defined by the above expression when $p \, \mathrm{d}\mu$ is absolutely continuous w.r.t. $q \, \mathrm{d}\mu$ ($p \ll q$), otherwise it is defined as $+\infty$ (see, e.g., [34, Defn. 2.1]). However, whenever $p \not\ll q$, then the function $\log(p/q)$ equals $+\infty$ on a set of positive p-measure, hence the above definition already gives $D(p\|q) = +\infty$. It is still possible, however, that $D(p\|q) = +\infty$ even when $p \ll q$. This is the case, for example, when q is Gaussian and p is a pdf with infinite variance such as the Cauchy distribution.

Remark 5 (Double bar notation) The (Kullback–Leibler) divergence is not symmetric in (p, q), which seems to be the reason for which the double bar notation '$\|$' (instead of a comma) is used. The origin of this exotic notation is not well-known. Kullback and Leibler themselves did not originate this notation in their seminal paper [19]. They rather used $I(1:2)$ for alternatives p_1, p_2 with a colon ":" to indicate non commutativity. Later the notation $I(P \mid Q)$ was used but this collided with the notation '$|$' for conditional distributions. The first occurence of the double bar notation I could find was by Rényi in the form $I(P\|Q)$ in the same paper that introduced Rényi entropies and divergences [18]. This notation was soon adopted by researchers of the Hungarian school of information theory, notably Csiszár (see, e.g., [32, 39, 40]).

Divergence is not symmetric, nor does it satisfy the triangle inequality. However, $D(p\|q)$ is nonnegative and vanishes if only if the two distributions p and q coincide:

Proposition 4 (Nonnegativity of divergence)

$$D(p\|q) \geq 0 \text{ with equality } D(p\|q) = 0 \iff p = q. \tag{16}$$

Proof By Jensen's inequality applied to the convex function $f(x) = x \log x$, $D(p\|q) = \int q f(\frac{p}{q})\, \mathrm{d}\mu \geq f(\int q \cdot \frac{p}{q}\, \mathrm{d}\mu) = f(1) = 0$ with equality iff $\frac{p}{q}$ is constant, i.e., p and q coincide.

An alternate proof is to use the well-known inequality $\log \frac{q}{p} \leq (\frac{q}{p} - 1) \log e$, hence $p \log \frac{p}{q} \geq (p - q) \log e$, which gives $D(p\|q) \geq \int (p - q) \log e \ \mathrm{d}\mu = 0$ with the same equality condition. □

Remark 6 (Random variable notation) Some authors (e.g., [38]) define the divergence of two random variables $X \sim p_X$ and $Y \sim p_Y$ as

$$D(X\|Y) \triangleq D(p_X\|q_X).$$

Here $D(X\|Y)$ depends only on the (marginal) distributions of X and Y, not on their joint distribution. This is not to be confused with Definition 7 below.

Strictly speaking, D is *not* a divergence between random variables X, Y in the sense that $D(X\|Y) \geq 0$ yet $D(X\|Y) = 0$ does not imply $X = Y$, but only that they are equivalent: $X \equiv Y$ in the sense that they share the same distribution.

The range of values taken by total variation distance and divergence can be summarized as follows.

- minimum value $D(p\|q) = 0 \iff \Delta(p, q) = 0$ if and only if $p = q$;
- maximum value $\Delta(p, q) = 1 \implies D(p\|q) = +\infty$ since $p \perp q \implies p \not\ll q$.

The alternate definition of divergence is again to proceed from the discrete case to the general case as follows.

Definition 5 *(Kullback–Leibler divergence—second definition)*

$$D(p\|q) \triangleq \sup \sum_i p(A_i) \log \frac{p(A_i)}{q(A_i)} \tag{17}$$

where the supremum is again taken over all *partitions* of Ω into a countable number of (pairwise disjoint) $A_i \in \mathcal{A}$.

Similarly as for total variation, the sum $\sum_i p(A_i) \log \frac{p(A_i)}{q(A_i)} \geq 0$ in (17) can only increase by subpartitioning (refining the partition) by the well-known *log-sum inequality* (see e.g., [41]); hence (17) can be seen as a limit for finer and finer partitions. Also, when $\Omega \subset \mathbb{R}$ or R^d, this supremum can simply be taken over partitions of *intervals* A_i, by Dobrushin's theorem [42, § 2].

Proposition 5 *The above two definitions of Kullback–Leibler divergence coincide.*

Proof A 1959 theorem by Gel'fand & Yaglom [43] and Perez [44] proves that the two definitions (13) and (17) coincide—in particular when (17) is finite, which implies $p \ll q$. For a modern proof see [34, § 4.2]. □

In contrast to total variation, Kullback–Leibler divergence "tensorizes" nicely for products of probability distributions:

Proposition 6 (Tensorization property) *For products of n distributions* $p = \bigotimes_{i=1}^n p_i$, $q = \bigotimes_{i=1}^n q_i$, *one has*

$$D(p\|q) = D\Big(\bigotimes_{i=1}^n p_i \,\Big\|\, \bigotimes_{i=1}^n q_i\Big) = \sum_i D(p_i\|q_i) \tag{18}$$

Proof Obvious from the definition. □

Incidentally, this tensorization property implies that the corresponding divergence is unbounded, while, by contrast, most of the above examples of distances (like the total variation distance) are bounded and can always be normalized to assume values between 0 and 1.

4 A motivating example

To understand why Pinsker's inequality $D \geq \varphi(\Delta)$ can be useful, consider the problem of distinguishing a fair coin (Bernoulli distribution $\mathcal{B}(p)$ with $p = 1/2$) from a unfair coin (Bernoulli distribution $\mathcal{B}(q)$ with $q \neq 1/2$) using only the result of a certain number n of i.i.d. tosses—we are not allowed to examine the coin which could be slightly bent to make it unfair.

By Proposition 3, to be sure to identify whether the coin is fake or not with probability ϵ requires errors of both types less than ϵ, hence $2\epsilon \geq 1 - \Delta$, that is,

$$\Delta\Big(\bigotimes_{i=1}^n p_i, \bigotimes_{i=1}^n q_i\Big) \geq 1 - 2\epsilon \tag{19}$$

for independent tosses $i = 1$ to n. The problem is that total variation $\Delta\Big(\bigotimes_{i=1}^n p_i, \bigotimes_{i=1}^n q_i\Big)$ does not nicely tensorize (is not scalable). However, Kullback-Leibler divergence is scalable by (18):

$$D\Big(\bigotimes_{i=1}^n p_i \,\Big\|\, \bigotimes_{i=1}^n q_i\Big) = n \cdot d(p\|q) \tag{20}$$

where

$$d(p\|q) = d(\frac{1}{2}\|q) = \log \frac{1}{2\sqrt{q(1-q)}} = \log \frac{1}{\sqrt{1-4\delta^2}} \tag{21}$$

using the notation $\delta = \delta(p, q) = |p - q|$ for binary total variation.

By Pinsker's inequality $D \geq \varphi(\Delta)$, this gives the following estimate:

$$n \geq \frac{\varphi(1-2\epsilon)}{\log \frac{1}{\sqrt{1-4\delta^2}}} \text{ tosses} \tag{22}$$

are required to distinguish a fair coin from an unfair one. This estimate is easily computable given φ for some particular Pinsker inequality, and as expected is all the more large as $\delta = |p - q|$ is small.

As this example shows, Pinsker's inequality is particularly useful to carry tensorization properties for independent distributions to statistical (total variation) distance. In summary, it is useful to make nonscalable quantities scalable.

5 Statistical distance and mutual information

Pinsker, in his seminal work [42], did not actually consider probability distributions in general but rather random variables: How does some observation Y affect the probability distribution of some given random variable X? This can be measured as the distance or divergence of X from X given Y, averaged over the observation Y.

Using the total variation distance, one obtains the notion of *statistical distance* between the two random variables X and Y, which is often used in computer science (see, e.g., [45] and [35, Chap. 8]).

Definition 6 *(Statistical distance, a.k.a. total variation information)*

$$\Delta(X; Y) \triangleq \mathbb{E}_y \, \Delta(p_{X|y}, p_X) = \Delta(p_{XY}, p_X \otimes p_Y). \tag{23}$$

Using the Kullback–Leibler divergence, one obtains the celebrated *mutual information* introduced by Fano [46], based on Shannon's seminal work [47]:

Definition 7 *(Mutual information)*

$$I(X; Y) \triangleq \mathbb{E}_y \, D(p_{X|y} \| p_X) = D(p_{XY} \| p_X \otimes p_Y). \tag{24}$$

Here the semicolon ";" is often used to separate the variables. The comma "," rather denotes joint variables and has higher precedence than ";" as in $I(X; Y, Z)$ which denotes the mutual information between X and (Y, Z). Basic properties of $I(X; Y)$ can be found in [34, 37, 38].

Remark 7 (Markov Kernel) In this information theoretic setting, a rigorous definition of $p_{X|Y}$ is based on the notion of *Markov kernel*, which can be obtained by disintegration of a joint distribution $p_{X,Y}$ under some technical conditions (see.g., [34, § 2.4]). Such considerations are not needed if one assumes that joint p_{XY} and product $p_X \otimes p_Y$ distributions are dominated by a product measure $d\mu(x)d\nu(y)$ with corresponding Radon-Nikodym derivatives $p(x, y) = p_{XY}(x, y)$ and $q(x, y) = p_X \otimes p_Y(x, y) =$

$p_X(x)p_Y(y)$. In this case one can simply define $p_{X|y}(x) = \frac{p_{XY}(x,y)}{p_Y(y)}$ for p_Y-a.e. y so that the above relations hold.

From these definitions, it follows that any Pinsker inequality (1) can also be interpreted as an inequality relating statistical distance $\Delta = \Delta(X;Y)$ to mutual information $I = I(X;Y)$:

Proposition 7 (Informational Pinsker inequality)

$$I(X;Y) \geq \varphi(\Delta(X;Y)) \tag{25}$$

for any two random variables X and Y, with the same φ as in (1)*.*

Proof Obvious from (1) by setting $p = p_{XY}$ and $q = p_X \otimes p_Y$. □

6 Some Ingredients for Proving Pinsker Inequalities

6.1 Binary reduction of Pinsker's inequality

A straightforward observation, that greatly simplifies the derivation of Pinsker inequalities, follows from the alternative definitions (7) and (17).

Theorem 8 (Binary reduction) *Any Pinsker inequality* (1) *is equivalent to the inequality expressed in term of binary distributions* (4)*,* (14)*:*

$$d \geq \varphi(\delta) \tag{26}$$

relating binary divergence $d = d(p\|q)$ to binary distance $\delta = |p - q|$ and holding for any parameters $p, q \in [0,1]$. Thus, the binary case, which writes

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \varphi(|p-q|) \tag{27}$$

is equivalent to the general case, but is naturally easier to prove.

Proof By Corollary 2, the supremum in (7) is attained for binary partitions of the form $\{A, A^{\complement}\}$. On the other hand, the supremum in (17) is obviously greater than that supremum for such binary partitions. Therefore, if Pinsker's inequality holds in the binary case, then

$$D(p\|q) \geq d(p(A)\|q(A)) \geq \varphi(\delta(p(A), q(A))) = \varphi(\Delta(p,q)) \tag{28}$$

which is Pinsker's inequality in the general case. □

Remark 8 (Symmetries) In any Pinsker inequality, since $\Delta(p, q)$ is symmetric in (p, q), both $D(p\|q)$ and $D(q\|p)$ admit the same lower bound $\varphi(\Delta(p, q))$. Furthermore, in the binary case, one has $d(p\|q) = d(1 - p\|1 - q)$. Therefore, it is enough to prove only (27) under the condition $p > q$, that is

$$p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \geq \varphi(p - q) \tag{29}$$

for any $1 \geq p > q \geq 0$.

Remark 9 (Data processing inequality) The above binary reduction principle was first used by Csiszár [32] but as a consequence of a more general *data processing inequality* for any Markov kernel $p_{Y|X}$, which states that if the Markov kernel sends p_X to p_Y and q_X to q_Y, then

$$D(p_Y\|q_Y) \leq D(p_X\|q_X). \tag{30}$$

In fact this holds for any f-divergence, including total variation distance. This was first proved by Csiszár [22, 32]. Taking the "deterministic channel" $Y = 1_{X\in A}$ one recovers the binary reduction $d(p_X(A)\|q_X(A)) \leq D(p_X\|q_X)$.

The full generality of the data processing inequality, however, is not needed in Theorem 8.

6.2 Comparison of two Pinsker inequalities

The following is sometimes useful to compare two different Pinsker inequalities (1) (provided that φ is differentiable and always satisfies the condition $\varphi(0) = 0$):

Theorem 9 (Pinsker comparison) *A Pinsker inequality $D \geq \varphi_1(\Delta)$ is (uniformly) stronger then another Pinsker inequality $D \geq \varphi_2(\Delta)$, that is, $\varphi_1(\Delta) \geq \varphi_2(\Delta)$ for all $0 \leq \Delta \leq 1$, if the derivatives satisfy the inequality $\varphi_1' \geq \varphi_2'$.*

This comparison principle can be stated as follows: *lower derivative φ' implies weaker Pinsker inequality*.

Proof Consider two Pinsker inequalities of the form $D \geq \varphi_1(\Delta)$ and $D \geq \varphi_2(\Delta)$ where both φ_1 and φ_2 are nonnegative differentiable functions such that $\varphi_1(0) = \varphi_2(0) = 0$. By comparison of derivatives, $\varphi_1' \geq \varphi_2'$ and $\varphi_1(0) = \varphi_2(0) = 0$ imply $D \geq \varphi_1(\Delta) \geq \varphi_2(\Delta)$. □

6.3 Note on the choice of the base of the logarithm

The inequalities shown in this paper are more easily expressed with natural logarithms ($\ln = \log_e$ to base e) because it simplifies the mathematical derivations. Accordingly, many papers do assume, sometimes implicitly, natural logarithms. It has become, however, common practice in information theory that the logarithm (log) is considered in *any* base, so that one can express inequalities with logarithms to base 2 or 10 instead

of natural logarithms. In this general case, any logarithmic measure such as D has to be divided by $\log e$. Thus, if the inequality $D \geq \varphi(\Delta)$ holds for natural logarithms, then the general case writes $D \geq \varphi(\Delta) \cdot \log e$. As a result, some constants in the inequalities have to be multiplied by $\log e$. We follow this approach here. The "natural case" can always be recovered by removing the constants $\log e$.

7 Pinsker and other authors in the 1960 s

7.1 Pinsker's original derivation

It is generally said that Pinsker, in his 1960 book [42], proved the classical Pinsker inequality in the form

$$D \geq c \cdot \Delta^2 \tag{31}$$

with a suboptimal constant c, and that the optimal (maximal) constant $c = 2 \log e$ was later found independently by Kullback [48], Csiszár [32] and Kemperman [49], hence the alternative name Kullback-Csiszár-Kemperman inequality.

In fact, Pinsker did not explicitly state Pinsker's inequality in this form, not even in the general form (1) for some other function φ. First of all, he only investigated mutual information $I(X;Y)$ vs. statistical distance $\Delta(X;Y)$ with $p = p_{X,Y}$ and $q = p_X \otimes p_Y$ (see § 5)—yet his results do easily carry over to the general case of arbitrary distributions p and q. More important, he actually showed two separate inequalities:

Proposition 10 (Pinsker's 1960 contribution)

$$\Delta \cdot \log e \leq \int p \left| \log \frac{p}{q} \right| \mathrm{d}\mu \leq D + 10\sqrt{D \cdot \log e}. \tag{32}$$

Proof By (5) and the same inequality $(p-q) \log e \leq p \log \frac{p}{q}$ as in the proof of Prop. 4, one has

$$\Delta \cdot \log e = \int (p-q)^+ \log e \, \mathrm{d}\mu \leq \int p \left(\log \frac{p}{q} \right)^+ \mathrm{d}\mu \leq \int p \left| \log \frac{p}{q} \right| \mathrm{d}\mu.$$

Pinsker's proof of the second inequality amounts to upper bounding the negative part $\int p \left(\log \frac{p}{q} \right)^- \mathrm{d}\mu \leq 5\sqrt{D \cdot \log e}$, and is much more involved [42, pp. 14–15], see Fig. 1. □

Remark 10 (Barron's 1986 derivation [17]) Decades later in 1986, Barron [17, Cor. p. 339] proved this second inequality in (32) (with the better constant $\sqrt{2}$ instead of 10) as an easy consequence of Pinsker's inequality itself with the optimal constant

При малых значениях $I(\xi, \eta)$ оказывается полезным следующее неравенство

$$\mathcal{J}(\xi, \eta) \leqslant I(\xi, \eta) + \Gamma \sqrt{I(\xi, \eta)}, \qquad (2.3.3)$$

where Γ is a positive constant which does not depend upon (ξ, η).

To verify (2.3.3) it suffices to show that the integral of the negative part of $i_{\xi\eta}(x, y)$ with respect to $P_{\xi\eta}$ is not less than $-\Gamma_1\sqrt{I(\xi, \eta)}$, where $\Gamma_1 > 0$ is a fixed number.

Let Q_ϵ denote the set of points (x, y) for which $i_{\xi\eta}(x, y) < -\epsilon$ and consider the expression

$$P_{\xi\eta}(Q_\epsilon)\log\frac{P_{\xi\eta}(Q_\epsilon)}{P_{\xi\times\eta}(Q_\epsilon)} + (1 - P_{\xi\eta}(Q_\epsilon))\log\frac{1 - P_{\xi\eta}(Q_\epsilon)}{1 - P_{\xi\times\eta}(Q_\epsilon)}. \qquad (2.3.4)$$

This sum (corresponding to the partition of $X \times Y$ into Q_ϵ and $X \times Y - Q_\epsilon$) has the same form as the sums appearing in the definition (2.1.1) of the information $I(\xi, \eta)$. Hence (2.3.4) is not greater than $I(\xi, \eta)$. Further, since $i_{\xi\eta}(x, y) = \log a_{\xi\eta}(x, y) < -\epsilon$ for $(x, y) \in Q_\epsilon$,

$$\frac{P_{\xi\eta}(Q_\epsilon)}{P_{\xi\times\eta}(Q_\epsilon)} = \frac{\int_{Q_\epsilon} a_{\xi\eta}(x, y)P_{\xi\times\eta}(dx\,dy)}{P_{\xi\times\eta}(Q_\epsilon)} = \frac{\int_{Q_\epsilon} e^{i_{\xi\eta}(x,y)}P_{\xi\times\eta}(dx\,dy)}{P_{\xi\times\eta}(Q_\epsilon)}$$

$$< \frac{\int_{Q_\epsilon} e^{-\epsilon}P_{\xi\times\eta}(dx\,dy)}{P_{\xi\times\eta}(Q_\epsilon)} = e^{-\epsilon}. \qquad (2.3.5)$$

Combining (2.3.5) with the remark that (2.3.4) does not exceed $I(\xi, \eta)$, we conclude that $P_{\xi\eta}(Q_\epsilon)$ does not exceed the supremum over all values of u which satisfy the system of inequalities

$$\begin{cases} u\log\frac{u}{r} + (1 - u)\log\frac{1 - u}{1 - r} \leqslant I(\xi, \eta); \\ \frac{u}{r} \leqslant e^{-\epsilon}; \qquad 0 \leqslant u \leqslant r \leqslant 1. \end{cases} \qquad (2.3.6)$$

Differentiating the left side of the first of the above inequalities with respect to u, we see that it is a decreasing function of u for $u \leqslant r$. Thus, without decreasing the supremum over u, we may replace the inequality $u/r \leqslant e^{-\epsilon}$ by the equality $u = re^{-\epsilon}$. Substituting $r = ue^{\epsilon}$ into the first of the inequalities (2.3.6), we obtain instead of (2.3.6) the inequality

$$-u\epsilon + (1 - u)\log\frac{1 - u}{1 - ue^{\epsilon}} \leqslant I(\xi, \eta); \qquad 0 \leqslant u \leqslant 1 \qquad (2.3.7)$$

We estimate the term $\log\frac{1 - u}{1 - ue^{\epsilon}}$ by using the inequality $\log u \geqslant 1 - 1/u$. The result is

$$\log\frac{1 - u}{1 - ue^{\epsilon}} \geqslant 1 - \frac{1 - ue^{\epsilon}}{1 - u} = \frac{u(e^{\epsilon} - 1)}{1 - u}.$$

Substituting $\frac{u(e^{\epsilon} - 1)}{1 - u}$ for $\log\frac{1 - u}{1 - ue^{\epsilon}}$ in (2.3.7), we obtain the inequality

$$u(-\epsilon + e^{\epsilon} - 1) \leqslant I(\xi, \eta). \qquad (2.3.8)$$

It follows, from what we have done, that every value u which satisfies the system (2.3.6) also satisfies (2.3.8), and therefore $P_{\xi\eta}(Q_\epsilon)$ does not exceed the largest value of u satisfying (2.3.8). Thus:

$$P_{\xi\eta}(Q_\epsilon) \leqslant u_{\max} \leqslant \frac{I(\xi, \eta)}{e^{\epsilon} - 1 - \epsilon} < \frac{I(\xi, \eta)}{\epsilon^2/2} \qquad (2.3.9)$$

We can now estimate the integral with respect to $P_{\xi\eta}$ of the negative part of $i_{\xi\eta}(x, y)$, which may be written in the form

$$\int_{i_{\xi\eta}(x,y)<0} i_{\xi\eta}(x, y)P_{\xi\eta}(dx\,dy) = \int_{-\sqrt{I(\xi,\eta)}<i_{\xi\eta}(x,y)<0} i_{\xi\eta}(x, y)P_{\xi\eta}(dx\,dy)$$
$$+ \int_{i_{\xi\eta}(x,y)<-\sqrt{I(\xi,\eta)}} i_{\xi\eta}(x, y)P_{\xi\eta}(dx\,dy). \qquad (2.3.10)$$

Obviously

$$\int_{-\sqrt{I(\xi,\eta)}\leqslant i_{\xi\eta}(x,y)<0} i_{\xi\eta}(x, y)P_{\xi\eta}(dx\,dy) \geqslant \int_{X\times Y} -\sqrt{I(\xi, \eta)}P_{\xi\eta}(dx\,dy)$$
$$= -\sqrt{I(\xi, \eta)} \qquad (2.3.11)$$

and

$$\int_{i_{\xi\eta}(x,y)<-\sqrt{I(\xi,\eta)}} i_{\xi\eta}(x, y)P_{\xi\eta}(dx\,dy) = \int_{-\infty}^{-\sqrt{I(\xi,\eta)}} \epsilon\,d_\epsilon P\{i(\xi, \eta) \leqslant \epsilon\}$$
$$= \int_{-\infty}^{-\sqrt{I(\xi,\eta)}} \epsilon\,d_\epsilon P_{\xi\eta}(Q_{-\epsilon}) \geqslant \int_{-\infty}^{-\sqrt{I(\xi,\eta)}} \epsilon\,d_\epsilon\frac{2I(\xi, \eta)}{\epsilon^2} \qquad (2.3.12)$$
$$= -\int_{-\infty}^{-\sqrt{I(\xi,\eta)}} \frac{4I(\xi, \eta)}{\epsilon^3}d\epsilon = -4\sqrt{I(\xi, \eta)}.$$

Comparing (2.3.10) with (2.3.11) and (2.3.12), we obtain

$$\int_{i_{\xi\eta}(x,y)<0} i_{\xi\eta}(x, y)P_{\xi\eta}(dx\,dy) \geqslant -\Gamma_1\sqrt{I(\xi, \eta)}, \qquad (2.3.13)$$

where Γ_1 is a fixed number.

Fig. 1 Pinsker's second inequality in (32): Original statement (in Russian) in [42] where $\Gamma \leq 10$ for natural logarithms, and the proof, pages 14–15 of [42], translated into English

$c = 2\log e$. Indeed, one has $\int p\left|\log\frac{p}{q}\right| = D + 2\int p\left(\log\frac{p}{q}\right)^-$, where

$$\int p\left(\log\frac{p}{q}\right)^- = \int_{p<q} p\log\frac{q}{p} \leq \int_{p<q}(q - p)\log e$$
$$= \int (q - p)^+\log e = \Delta\log e \leq \sqrt{\frac{D\log e}{2}},$$

hence $\int p\left|\log\frac{p}{q}\right| \leq D + \sqrt{2D\log e}$.

Corollary 11 (Verdú's 2014 observation [50]) *Pinsker's original inequalities* (32) *imply the following "Pinsker's inequality"* (31) *with suboptimal constant* $c = \frac{\log e}{102}$*:*

$$D \geq \frac{\log e}{102}\Delta^2. \qquad (33)$$

However, this was nowhere mentioned in Pinsker's book [42].

Proof Since $0 \leq \Delta \leq 1$ always, one can always assume

$$\Delta\log e \leq D + 10\sqrt{D\log e} \leq \log e,$$

otherwise the inequality is vacuous. Then by (32),

$$\begin{aligned}\Delta^2(\log e)^2 &\le (D+10\sqrt{D\log e})^2\\ &= (D+20\sqrt{D\log e})D+100D\log e\\ &\le 2D\log e+100D\log e=102D\log e,\end{aligned}$$

which gives (33). □

Remark 11 (Continuity properties) By Pinsker's inequality (31), if divergence is arbitrarily small, then so is total variation. Thus for distributions, the strongest type of convergence is in divergence, followed by total variation, followed by convergence in distribution.

What motivated Pinsker in his 1960 book [42, § 2.3] was to prove a "continuity property" with respect to mutual information in terms of sequences of random variables:

$$I(X_n;Y_n)\to 0 \implies \Delta(X_n;Y_n)\to 0 \tag{34}$$

This is obvious from the Pinsker inequality $I(X_n;Y_n)\ge c\cdot\Delta^2(X_n;Y_n)$ which is (31) with $p=p_{X,Y}$ and $q=p_X\otimes p_Y$.

7.2 First occurrences of Pinsker inequalities in the 1960 s

7.2.1 Volkonskii and aozanov (1959)

The *first explicit occurrence of a Pinsker inequality of the general form* $D\ge\varphi(\Delta)$ (1) occurs even *before* the publication of Pinsker's book, by Volkonskii and Rozanov [51, Eq. (V)] in 1959.

Proposition 12 (Volkonskii and Rozanov 1959 contribution)

$$D\ge 2\log e\cdot\Delta-\log(1+2\Delta). \tag{35}$$

They referred to M. S. Pinsker (probably an earlier manuscript version of his book [42]) for his continuity property (34) and gave a simple proof of their inequality as follows.

Proof Since $f(x)=x\log e-\log(1+x)\ge f(|x|)\ge 0$ is convex, one has

$$\begin{aligned}D(p\|q)&=\int p\left(\left(\frac{q}{p}-1\right)\log e-\log\frac{q}{p}\right)=\int pf\left(\frac{q}{p}-1\right)\\ &\ge\int pf\left(\left|\frac{q}{p}-1\right|\right)\\ &\ge f\left(\int p\left|\frac{q}{p}-1\right|\right)=f(2\Delta).\end{aligned}$$

□

Since for $x > 0$, one has $0 < f(x) = x \log e - \log(1+x) < \frac{x^2}{2} \log e$ in the above proof, the lower bound is nonnegative but strictly weaker than the classical Pinsker inequality (31) with optimal constant $c = 2 \log e$. As we shall see, however, both are asymptotically optimal near $D = \Delta = 0$ because $f(x) = \frac{x^2}{2} \log e + o(x^2)$.

7.2.2 Sakaguchi (1964)

The *first explicit occurrence of a Pinsker inequality of the classical form* (31) appeared as an exercise in Minoru Sakaguchi's 1964 book [52, pp. 32–33]. Unfortunately, Sakaguchi's book remained unpublished.

Proposition 13 (Sakaguchi's 1964 Contribution)

$$D \geq \log e \cdot \Delta^2 \tag{36}$$

which is (31) *with the suboptimal constant* $c = \log e < 2 \log e$.

Proof Sakaguchi actually proved $D \geq 2\Delta_H^2 \log e \geq \Delta^2 \log e$ where Δ_H is the Hellinger distance: Considering the distribution $q' = \frac{\sqrt{pq}}{\int \sqrt{pq}}$, one has

$$\begin{aligned} D(p\|q) &= 2 \int p \log \frac{p}{\sqrt{pq}} = 2D(p\|q') - 2 \log \int \sqrt{pq} \\ &\geq -2 \log \int \sqrt{pq} \\ &\geq 2(1 - \int \sqrt{pq}) \log e = 2\Delta_H^2 \log e \end{aligned}$$

since $\log x \leq (x-1) \log e$. Now, since $\Delta_H^2 \geq 0$, $\int \sqrt{pq} \leq 1$, so

$$\begin{aligned} 2\Delta_H^2 &\geq \left(1 + \int \sqrt{pq}\right)\left(1 - \int \sqrt{pq}\right) \\ &= \frac{1}{4} \int (\sqrt{p} + \sqrt{q})^2 \int (\sqrt{p} - \sqrt{q})^2 \\ &\geq \frac{1}{4} \Big(\int (\sqrt{p} + \sqrt{q})(\sqrt{p} - \sqrt{q}) \Big)^2 = \Delta^2 \end{aligned}$$

by Cauchy-Schwarz inequality. □

7.2.3 McKean (1966)

The *first published occurence of a Pinsker inequality of the classical form* (31) was by McKean [53, § 9a)] in 1966, who was motivated by a problem in physics related to Boltzmann's H-theorem.

Proposition 14 (McKean's 1966 contribution)

$$D \geq \frac{\log e}{e} \cdot \Delta^2 \tag{37}$$

which is (31) *with the suboptimal constant* $c = \frac{\log e}{e}$ *(worse than Sakaguchi's).*

McKean was aware that his constant c is "not the best possible constant". His proof is originally under the (unnecessary) assumption that q is Gaussian, and can be rephrased as follows.

Proof The convex function $x \log x - (x-1)\log e$ is lower bounded by $\frac{(x-1)^2}{2e} \log e$ when $x \geq e$ and by $\frac{x-1}{2e} \log e$ for $0 < x < e$. Splitting the integral it follows that

$$\begin{aligned} D &= \int q\Big(\frac{p}{q} \log \frac{p}{q} - \Big(\frac{p}{q} - 1\Big)\log e\Big) \\ &\geq \frac{1}{2e}\Big(\int_{\frac{p}{q}<e} q\Big|\frac{p}{q} - 1\Big|^2 + \int_{\frac{p}{q}\geq e} q\Big|\frac{p}{q} - 1\Big|\Big)\log e \\ &\geq \frac{1}{2e}\Big[\Big(\int_{\frac{p}{q}<e} q\Big|\frac{p}{q} - 1\Big|\Big)^2 + \Big(\int_{\frac{p}{q}\geq e} q\Big|\frac{p}{q} - 1\Big|\Big)^2\Big]\log e \end{aligned}$$

by the convexity of x^2 and the fact that $\int_{\frac{p}{q}\geq e} q\big|\frac{p}{q} - 1\big| = \int_{\frac{p}{q}\geq e} p - q \leq \int p = 1$. Thus since $a^2 + b^2 \geq \frac{(a+b)^2}{2}$,

$$\begin{aligned} D &\geq \frac{1}{2e}\Big[\Big(\int_{\frac{p}{q}<e} |p-q|\Big)^2 + \Big(\int_{\frac{p}{q}\geq e} |p-q|\Big)^2\Big]\log e \\ &\geq \frac{\log e}{4e}\Big(\int |p-q|\Big)^2 = \frac{\log e}{e}\Delta^2. \end{aligned}$$

□

7.2.4 Csiszár (1966)

The *first mention of the classical Pinsker inequality* (31) *with the optimal constant* $c = 2\log e$ was by Csiszár [39], in a manuscript received just one month after McKean's. In this 1966 paper, however, Csiszár only proved (31) with the suboptimal constant $c = \frac{\log e}{4}$ [39, Eq. (13)], which is worse than McKean's. But he also acknowledged the preceding result of Sakaguchi (with the better constant $c = 1$) and stated (without proof) that the best constant is $c = 2$. He also mentioned the possible generalization to f-divergences. On this occasion he credited Pinsker for having found an inequality of the type (31) (which as we have seen was only implicit).

Proposition 15 (Csiszár's 1966 contribution)

$$D \geq \frac{\log e}{4} \cdot \Delta^2 \tag{38}$$

which is (31) *with the suboptimal constant* $c = \frac{\log e}{4}$ *(worse than McKean's).*

Csiszár's proof, which can be adapted to other types of f-divergences, is as follows.

Proof One can always assume $D \leq \frac{\log e}{4}$ since $\Delta \leq 1$. The convex function $x \log x - (x-1)\log e$ has null derivative at $x = 1$ and has second derivative $\frac{\log e}{x} \geq \frac{2\log e}{3}$ when $|x-1| \leq \sqrt{D/\log e} \leq \frac{1}{2}$. It follows that $x \log x - (x-1)\log e \geq \frac{\log e}{3}(x-1)^2$ there. Therefore,

$$\begin{aligned}
2\Delta &= \int q\left|\frac{p}{q} - 1\right| \\
&\leq \sqrt{D/\log e} + \int_{|\frac{p}{q}-1|>\sqrt{D/\log e}} q\left|\frac{p}{q} - 1\right| \\
&\leq \sqrt{D/\log e} + \frac{1}{\sqrt{D/\log e}} \int_{|\frac{p}{q}-1|>\sqrt{D/\log e}} q\left(\frac{p}{q} - 1\right)^2 \\
&\leq \sqrt{D/\log e} + \frac{3}{\sqrt{D/\log e}} \int p \log \frac{p}{q} - (p-q)\log e \\
&= \sqrt{D/\log e} + \frac{3D/\log e}{\sqrt{D/\log e}} = 4\sqrt{D/\log e}.
\end{aligned}$$

□

Following his derivation, Csiszár declares: "the best constant can be calculated in a straightforward way […] I intend to return to this question in another paper".

7.2.5 Csiszár again (1967)

The *first published proof of the classical Pinsker inequality* (31) *with the optimal constant* $c = 2\log e$ was again by Csiszár one year later [32, Thm. 4.1] using binary reduction—which he obtained as a particular case of the data processing inequality, see Remark 9.

Proposition 16 (Csiszár's 1967 Contribution)

$$D \geq 2\log e \cdot \Delta^2$$

where the constant $2\log e$ *is optimal.*

His proof can be written as an essentially one-line proof as follows.

Proof Using binary reduction (Theorem 8), it suffices to prove that $d = d(p\|q) \geq 2\log e \cdot \delta^2$ where $\delta = |p-q|$. Now by the "fundamental theorem of calculus",

$$\begin{aligned}
d(p\|q) &= \underbrace{d(p\|p)}_{=0} + \int_p^q \frac{\partial d(p\|r)}{\partial r}\,\mathrm{d}r = \int_p^q \frac{r-p}{r(1-r)}\,\mathrm{d}r \log e \\
&\geq 4\int_p^q (r-p)\,\mathrm{d}r \log e = 2(p-q)^2 \log e
\end{aligned}$$

where we used the inequality $r(1-r) \leq \frac{1}{4}$ for $r \in [0,1]$. That $c = 2\log e$ is not improvable follows from the expansion $d(p\|q) = 2(p-q)^2 \log e + o((p-q)^2)$. □

As a side result, this inequality (like the Volkonskii-Rozanov inequality (35)) is asymptotically optimal near $D = \Delta = 0$.

7.2.6 Kullback (1967)

In a note added in proof, however, Csiszár mentions an earlier independent derivation of Kullback, published in the same year 1967 in [48], with an improved inequality of the form $D \geq 2\log e \cdot \Delta^2 + \frac{4}{3}\log e \cdot \Delta^4$. In his correspondance, Kullback acknowledged the preceding result of Volkonskii and Rozanov. Unfortunately, as Vajda noticed [54] in 1970, the constant $\frac{4}{3}$ is wrong and should be corrected as $\frac{4}{9}$ [55] (see explanation below).

Proposition 17 (Kullback's 1967–1970 Contribution)

$$D \geq 2\log e \cdot \Delta^2 + \frac{4}{9}\log e \cdot \Delta^4 \tag{39}$$

where the initial constant for the second term was $4/3$ *and has been corrected to* $4/9$ *in 1970.*

Kullback's derivation uses again binary reduction (obtained as in the proof of Theorem 8) and then invokes an inequality of Schützenberger. This is explained in greater detail in the following sections. As it turns out, both constants $2\log e$ and $\frac{4}{9}\log e$ are optimal.

7.2.7 Kemperman (1968)

Finally, in an 1968 Canadian symposium presentation [56]—later published as a journal paper [49] in 1969, Kemperman, apparently unaware of the 1967 papers by Csiszár and Kullback, again derived the classical Pinsker inequality with optimal constant $c = 2\log e$.

Proposition 18 (Kemperman's 1968 Contribution)

$$D \geq 2\log e \cdot \Delta^2 \tag{40}$$

where the constant $2\log e$ *is optimal.*

Kemperman's ad-hoc proof (repeated in the renowned textbook [57]) is based on the inequality $\frac{4+2x}{3}(x\log x - (x-1)\log e) \geq (x-1)^2 \log e$ for any $x \geq 0$, which is much less satisfying than the one-line proof of Prop. 16:

Proof By Cauchy-Schwarz,

$$\begin{aligned}\sqrt{2D} &= \sqrt{\int q\frac{4+2(p/q)}{3}\cdot\int q\big(\frac{p}{q}\log\frac{p}{q}-\big(\frac{p}{q}-1\big)\log e\big)}\\ &\geq \int q\sqrt{\frac{4+2(p/q)}{3}}\sqrt{\frac{p}{q}\log\frac{p}{q}-\big(\frac{p}{q}-1\big)\log e}\\ &\geq \int\big|\frac{p}{q}-1\big|q\sqrt{\log e}=2\Delta\sqrt{\log e}.\end{aligned}$$

□

To acknowledge all the above contributions, it is perhaps permissible to rename Pinsker's inequality as the *Pinsker-Volkonskii-Rozanov-Sakaguchi-McKean-Csiszár-Kullback-Kemperman inequality.* However, this would unfairly obliterate the pioneer contribution of Schützenberger, as we now show.

8 Schützenberger's contribution (1953)

8.1 Schützenberger's life and work

First of all, it is worth saying a few words about the life and work of Marcel-Paul (Marco) Schützenberger. He is such an extraordinary personality, one of the most original and prolific researchers in many diverse areas such as genetics, statistics, information theory, variable length codes, combinatorics, automata, formal languages, etc. At the same time he also had an extraordinary personal life.

Marco Schützenberger is of Alsatian origin. His great-great-grandfather Georges Schützenberger was mayor of Strasbourg (the author's birthplace), and the name "Schützenberger" is still well known today as a beer brand. His great grandfather Paul Schützenberger moved from Alsace to Paris just before the 1870 Franco-German war, and was a renowned chemist. He founded the ESPCI (École supérieure de physique et de chimie industrielles de la Ville de Paris), where many important physicists worked, and was the subject of a satire in *Les Palmes de M. Schutz*, a famous play and movie in which he [Monsieur Schutz] insistently relies on the discoveries of Pierre and Marie Curie to obtain the palmes académiques (a national order awarded to eminent academics). Marco's father Pierre Schützenberger was a psychiatric physician, expert witness in the famous Papin sisters' case [58].

During World War II, the young Marco was appointed intern at a psychiatric hospital. At the same time he was active in resistance activities—he apparently worked for the Intelligence Service—and published his first mathematical paper on group theory in 1943. After World War II, Schützenberger participates in surrealist and Dadaist movements—he appears in a short film with Boris Vian, and becomes the main character ("Dr. Markus Schutz") in Vian's famous novel *Et on tuera tous les affreux* ("To Hell with the Ugly"). He is also a member of the cabinet of Communist minister Charles Tillon, publishes articles in lattice theory and in physiology, while studying

Fig. 2 A domestic scene staged by Marcel-Paul (Marco) Schützenberger and his wife, world-renowned psychotherapist Anne Ancelin Schützenberger, entitled "Psycho-drama of a 1948 marriage" during the Saint Germain des Prés period in Paris

"ancient Mongolian". In 1948 he defends his doctoral thesis entitled *Contribution à l'étude du sexe à la naissance* (Contribution to the study of sex at birth), awarded by the French Academy of Medicine. He applies statistical methods to the analysis of various medical problems, and later contributes to the discovery of trisomy 21.

Also in 1948, following a paper by French psychologist Anne Ancelin based on his statistics, he was offed a position in London. In order to be better paid, he got married immediately in London with Anne Ancelin (see photo in Fig. 2)…but eventually declined the position. The couple had one daughter, Hélène, and soon divorced in 1952. Schützenberger publishes papers on combinatorics in a genetics journal, on biostatistics with statistician George Darmois, and is consultant epidemiologist for the World Health Organization (WHO). In 1952 and 1953, the WHO sent him to Asia to combat infectious diseases of tropical countries. He met his second wife Hariati Soerosoegondo in Java, Indonesia.

In 1952 he came to information theory from biostatistics and defended his mathematical thesis in 1953 (advisor: Georges Darmois, president: Maurice Fréchet) entitled *Contribution aux applications statistiques de la théorie de l'information* (Contribution to statistical applications of information theory) [59]. He also has made several

other discoveries in statistics in the 1950 s, notably on what is now called the Bayesian Cramér-Rao bound [60, 61].

In the present paper, we focus on a particular result obtained by Schützenberger in his 1953 thesis, in the context of the interplay between statistics and information theory. In fact, Schützenberger was apparently the first to discover the profound connections between information measures such as Shannon's entropy, "Wald's information" (now better known as Kullback–Leibler divergence) and Fisher's information in estimation theory. Presumably because of these results, while affiliated to the CNRS (the French national center of scientific research), he was invited by Claude Shannon to spend the year 1956–1957 at the MIT Research Laboratory of Electronics in Cambridge, where his son Mahar (contraction of Marco and Hariati) was born.

According to Dominique Foata [62], his first maths student, and Dominique Perrin [63], another of his students, both renowned researchers in discrete mathematics and computer science, there is no evidence of Schützenberger interacting directly with Shannon. Perhaps they didn't get along very well at MIT, or already had very different preoccupations in 1956? As a matter of fact, Schützenberger was rather silent about this period of his professional life, and was apparently not inclined to ever mention his thesis and his early work in mathematical statistics and information theory. According to Foata [62], he wasn't even convinced he had written a good thesis.

During his stay at MIT and later on, Schützenberger was much more interested in variable length codes with R. S. Markus and context-free languages with N. Chomsky. He had already worked on lattices, statistics, block-designs and other combinatorial problems, and progressively on semigroups, automata, and codes. He is known for his Schützenberger groups in semigroup theory. In the algebraic theory of variable-length codes, the Kleene-Schützenberger theorem [64] is a fundamental theorem in the theory of formal languages and automata. He is also a pioneer of formal language theory with N. Chomsky. The famous Chomsky-Schützenberger theorem [65] is a representation theorem of context-free languages. With S. Eilenberg, he developed the theory of pseudo-varieties of semigroups. He is known for his works on the combinatorics of words. He created the combinatorics of the "plactic monoid", and its applications in the study of the symmetric group. Overall Schützenberger was renowned to all as a specialist in discrete mathematics and computer science, at the interface between algebra, probability and combinatorics. His early work on information theory remains little known.

In the 1970 s and 1980 s, Marco Schützenberger also served as scientific advisor for the WHO, to detect and prevent accidents due to the careless use of medicines or chemical and biological weapons. His son Mahar, *major* (ranked first in the admission exam) of the école Polytechnique in 1976, was killed in a car accident in 1980 at the age of 23, which deeply affected Marco Schützenberger. After serving as a corresponding member, Marco Schützenberger was elected to the French Academy of Sciences in 1988. Since 1991, the Mahar Schützenberger Prize has been awarded to Indonesian researchers preparing their doctoral thesis in France.

According to his students and friends, Schützenberger's personality was complex and unorthodox. He always held strong—sometimes contradictory—opinions on very diverse subjects such as the Darwinian theory of evolution, which he considered incompatible with any serious statistical analysis. Capable of glowing praises as well as ironic

sarcasm, he was passionate for discussion, paradox and controversy. The book *Triangle de pensées* (Triangle of thought) reports the discussions between Alain Connes, André Lichnerowicz, and Marco Schützenberger on general relativity, quantum mechanics or Gödel's theorem, and more generally the relations among mathematics, physics, philosophy, and other sciences. Schützenberger was saddened by the death of his wife Hariati in 1993. Inveterate smoker [66], he died a few years after her, in 1996.

8.2 Schützenberger's expansion

Seven years before the publication of Pinsker's book, 14 years before Csiszár derivation of the first order term, and 17 years before Kullback's correction of the second-order term, Marco Schützenberger, in his 1953 doctoral thesis [59], proved the following

Theorem 19 (Schützenberger's Inequality [59, p. 58])

$$D \geq 2\log e \cdot \Delta^2 + \frac{4}{9}\log e \cdot \Delta^4 \tag{41}$$

with optimal constants $2\log e$ *and* $\frac{4}{9}\log e$.

Optimality here means that the constants 2, $\frac{4}{9}$ are (in turn) not improvable: If $D \geq c_1\Delta^2 + c_2\Delta^4$ with $c1, c2 \geq 0$ then not only $c_1 \leq 2\log e$ but also $c_1 = 2\log e$ implies $c_2 \leq \frac{4}{9}\log e$, i.e., the constants' pair $(2, \frac{4}{9})$ is maximal in lexicographic order.

Schützenberger's derivation is based on the following remarkable identity.

Lemma 20 (Schützenberger's expansion) *For any* $x, y \in [-1, 1]$,

$$\begin{aligned} d(x,y) &\triangleq \frac{1-x}{2}\log\frac{1-x}{1-y} + \frac{1+x}{2}\log\frac{1+x}{1+y} \\ &= \sum_{k\geq 1} \frac{x^{2k} - 2kxy^{2k-1} + (2k-1)y^{2k}}{2k(2k-1)}\log e \\ &= (x-y)^2 \sum_{k\geq 1} \frac{x^{2k-2} + 2x^{2k-3}y + 3x^{2k-4}y^2 + \cdots + (2k-1)y^{2k-2}}{2k(2k-1)}\log e \end{aligned} \tag{42}$$

where all terms are nonnegative.

Proof Write $d(x, y) = e(x, y) + e(-x, -y)$ where $e(x, y)$ expands as

$$e(x,y) = \tfrac{1-x}{2}\log\tfrac{1-x}{1-y} = \tfrac{1-x}{2}\sum_{n\geq 1}\tfrac{y^n-x^n}{n}\log e = \tfrac{1}{2}\sum_{n\geq 1}\tfrac{(y^n-x^n)+(x^{n+1}-xy^n)}{n}\log e.$$

Adding $e(-x, -y)$ amounts to restricting the sum to even values $n = 2k$ for $(y^n - x^n)$ and to odd values $n = 2k - 1$ for $(x^{n+1} - xy^n)$ and multiplying by 2. This gives

$$d(x,y) = \sum_{k\geq 1}\tfrac{y^{2k}-x^{2k}}{2k}\log e + \tfrac{x^{2k}-xy^{2k-1}}{2k-1}\log e = \sum_{k\geq 1}\tfrac{x^{2k}-2kxy^{2k-1}+(2k-1)y^{2k}}{2k(2k-1)}\log e$$

as announced.

Now the polynomial $x^{2k} - 2kxy^{2k-1} + (2k-1)y^{2k}$ vanishes for $x = y$ and has derivative in x for fixed y equal to $2kx^{2k-1} - 2ky^{2k-1}$, which vanishes only when $x = y$. Therefore, this polynomial is always nonnegative and divisible by $(x-y)^2$. More precisely, the following calculation gives the announced factorization:

$$\begin{aligned} x^{2k} - 2kxy^{2k-1} + (2k-1)y^{2k} &= x^{2k} - y^{2k} - 2k(x-y)y^{2k-1} \\ &= (x-y)\big(x^{2k-1} + x^{2k-2}y + \cdots + xy^{2k-2} + y^{2k-1} - 2ky^{2k-1}\big) \\ &= (x-y)\big(x^{2k-1} + 2x^{2k-2}y + 3x^{2k-3}y^2 + \cdots + (2k-1)xy^{2k-2} \\ &\qquad - x^{2k-2}y - 2x^{2k-3}y^2 - \cdots - (2k-2)xy^{2k-2} - (2k-1)y^{2k-1}\big) \\ &= (x-y)^2\big(x^{2k-2} + 2x^{2k-3}y + 3x^{2k-4}y^2 + \cdots + (2k-1)y^{2k-2}\big) \end{aligned} \tag{43}$$

where both factors are nonnegative. □

Proof of Schützenberger's Theorem 19 Consider the binary case with $d = d(p\|q)$ and define $x = 1-2p \in [-1, 1]$ and $y = 1-2q \in [-1, 1]$. Then $\delta = (p-q) = \frac{y-x}{2}$ and Schützenberger's expansion $d = 4\delta^2 \sum_{k\geq 1} \frac{x^{2k-2}+2x^{2k-3}y+\cdots+(2k-1)y^{2k-2}}{2k(2k-1)} \log e$ gives

$$d = 2\log e \cdot \delta^2 + \frac{\delta^2}{3} \log e \cdot (x^2 + 2xy + 3y^2) + \cdots \tag{44}$$

In particular $d \geq 2\log e \cdot \delta^2$ where the constant $2\log e$ is optimal because $d \sim 2\log e\,\delta^2$ when $x, y \to 0$.

In the second-order term, one has $x^2 + 2xy + 3y^2 = \frac{(x-y)^2}{3} + 2\frac{(x+2y)^2}{3} \geq \frac{4\delta^2}{3}$. This gives $d \geq 2\log e \cdot \delta^2 + \frac{\delta^2}{3} \log e \cdot \frac{4\delta^2}{3} = 2\log e \cdot \delta^2 + \frac{4}{9} \log e \cdot \delta^4$. Again the constant $\frac{4}{9} \log e$ is optimal because $d \sim 2\log e \cdot \delta^2 + \frac{4}{9} \log e \cdot \delta^4$ as $x = -2y \to 0$. □

Schützenberger's original derivation (in French) is reproduced in Fig 3. The framed formula "*qui semble nouvelle*" (that seems to be new) is Schützenberger's inequality (41) where W ("*information de Wald*") is the Kullback–Leibler divergence and D is the total variation distance. It is the only framed equation in the whole thesis, which should indicate that Schützenberger was aware of its importance.

Admittedly, Schützenberger only considered the binary case, yet due to the binary reduction principle (Theorem 8), we know that this does not entail any loss of generality. There is also a typo at the end of the derivation that says that minimizing $x^2 + 2xy + 3y^2$ for fixed $2\delta = y - x$ gives $\frac{\delta^2}{3}$ instead of the correct $\frac{4\delta^2}{3}$.

The fact remains that quite remarkably, not only does Schützenberger's thesis contain the classical Pinsker inequality (31) with the optimal constant $c = 2\log e$, but also the second-order improvement, with the (correct) optimal constant $\frac{4}{9} \log e$ for the second-order term, 17 years before Kullback!

Dans le cas dichotomique, on a l'inégalité suivante qui semble nouvelle. Ecrivons :

$$D = p(\theta_0) - p(\theta_1) = q(\theta_1) - q(\theta_0)$$

$$\boxed{W \geqslant 2D^2 + \tfrac{4}{3}D^4.}$$

Posons en effet $2\,p(\theta_0) = 1-x$ et $2\,p(\theta_1) = 1-y$ après avoir choisi p de telle sorte que x soit positif.

On peut développer W en série de puissance de x et de y :

$$2W = (1-x)\,\mathrm{Log}(1+x)/(1-y) + (1+x)\,\mathrm{Log}\,(1+x)(1+y).$$

On trouve :

$$W = \sum_{i=1}^{\infty} (4i^2 - 2i)^{-1}\,(x^{2i} - 2ixy^{2i-1} + (2i-1)y^{2i})$$

Tous les termes sont positifs car le polynome $t^{2i} - 2it + 2i - 1$ a un unique extremum pour $t = 1$ et prend en ce point la valeur 0.

Bien plus :

$$x^{2i} - 2ixy^{2i-1} + (2i-1)y^{2i} = 4D^2(x^{2i-2} + 2x^{2i-3}y + 3x^{2i-4}y^2 + \ldots$$

$$\ldots + (2i-1)y^{2i-2})$$

Par conséquent W est plus grand que la somme des deux premiers termes de son développement qui sont :

$4D^2/2$ et $4D^2/12\,(x^2 + 2xy + 3y^2)$ et la valeur de ce dernier polynome étant supérieure pour D fixe à $D^2/3$ on trouve bien le résultat.

Fig. 3 Pinsker before Pinsker: In Schützenberger's notation [59, p. 58], W is for Wald's information, which is Kullback–Leibler divergence, and $D = p - q$. There is a typo at the end: Minimizing $x^2 + 2xy + 3y^2$ for fixed $2D = y - x$ is said to give $D^2/3$ instead of the correct $4D^2/3$

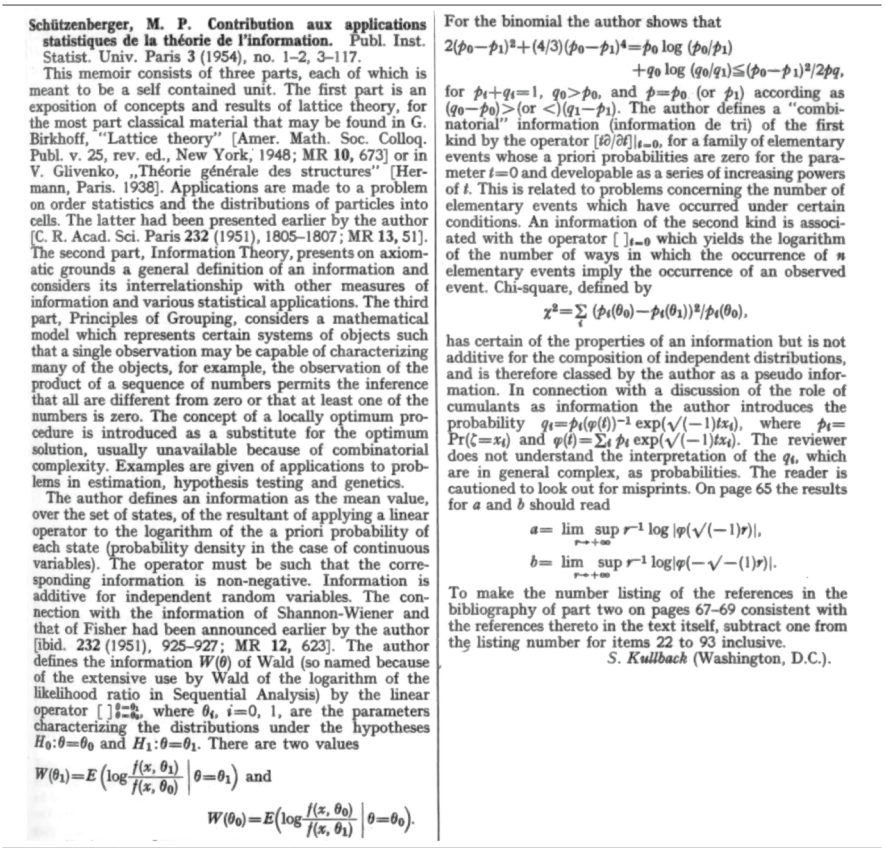
8.3 The 4/3 vs. 4/9 Mystery

In fact, leaving aside the use of binary reduction, Kullback's 1967 derivation [48] is just a mention of Schützenberger's inequality with the wrong constant $\frac{4}{3}$ instead of $\frac{4}{9}$. One may wonder why such a wrong constant appeared instead of the correct one found in Schützenberger's thesis published in 1954 (Fig 3).

Somewhat mysteriously, just before the apprearence of Kullback's 1967 paper, Kambo and Kotz presented a verbatim copy of Schützenberger's derivation in their 1966 paper [67], in a different context (the derivation of a variant of a Chernoff bound), without citing the original reference of Schützenberger, and also with the wrong constant $\frac{4}{3}$! A bit later in 1969, the constant $\frac{4}{3}$ was corrected to $\frac{4}{9}$ by Krafft [68], referring only to the Kambo-Kotz 1966 paper.

One year later, in the context of Pinsker inequalities, Vajda [54] pointed the wrong constant in Kullback's paper and claimed that "*Kraft corrected an inequality of Schützenberger on which Kullback's result was based*". Almost immediately, the correction was acknowledged by Kullback in the IEEE transactions on information theory [55].

From these facts, it appeared plausible to us that, even though the correct constant $\frac{4}{9}$ does appear in the publicly available 1954 published thesis of Schützenberger, a previous version of the manuscript, available to Kullback and other researchers, had the erroneous $\frac{4}{3}$ which was later corrected by Marco Schützenberger from a "3" to



Schützenberger, M. P. Contribution aux applications statistiques de la théorie de l'information. Publ. Inst. Statist. Univ. Paris **3** (1954), no. 1–2, 3–117.

This memoir consists of three parts, each of which is meant to be a self contained unit. The first part is an exposition of concepts and results of lattice theory, for the most part classical material that may be found in G. Birkhoff, "Lattice theory" [Amer. Math. Soc. Colloq. Publ. v. 25, rev. ed., New York, 1948; MR **10,** 673] or in V. Glivenko, „Théorie générale des structures" [Hermann, Paris. 1938]. Applications are made to a problem on order statistics and the distributions of particles into cells. The latter had been presented earlier by the author [C. R. Acad. Sci. Paris **232** (1951), 1805–1807; MR **13,** 51]. The second part, Information Theory, presents on axiomatic grounds a general definition of an information and considers its interrelationship with other measures of information and various statistical applications. The third part, Principles of Grouping, considers a mathematical model which represents certain systems of objects such that a single observation may be capable of characterizing many of the objects, for example, the observation of the product of a sequence of numbers permits the inference that all are different from zero or that at least one of the numbers is zero. The concept of a locally optimum procedure is introduced as a substitute for the optimum solution, usually unavailable because of combinatorial complexity. Examples are given of applications to problems in estimation, hypothesis testing and genetics.

The author defines an information as the mean value, over the set of states, of the resultant of applying a linear operator to the logarithm of the a priori probability of each state (probability density in the case of continuous variables). The operator must be such that the corresponding information is non-negative. Information is additive for independent random variables. The connection with the information of Shannon-Wiener and that of Fisher had been announced earlier by the author [ibid. **232** (1951), 925–927; MR **12,** 623]. The author defines the information $W(\theta)$ of Wald (so named because of the extensive use by Wald of the logarithm of the likelihood ratio in Sequential Analysis) by the linear operator $[\,]_{\theta=\theta_0}^{\theta=\theta_1}$, where θ_i, $i=0$, 1, are the parameters characterizing the distributions under the hypotheses $H_0:\theta=\theta_0$ and $H_1:\theta=\theta_1$. There are two values

$$W(\theta_1)=E\left(\log\frac{f(x,\theta_1)}{f(x,\theta_0)}\,\middle|\,\theta=\theta_1\right) \text{ and}$$

$$W(\theta_0)=E\left(\log\frac{f(x,\theta_0)}{f(x,\theta_1)}\,\middle|\,\theta=\theta_0\right).$$

For the binomial the author shows that

$$2(p_0-p_1)^2+(4/3)(p_0-p_1)^4=p_0\log(p_0/p_1)+q_0\log(q_0/q_1)\leqq(p_0-p_1)^2/2pq,$$

for $p_i+q_i=1$, $q_0>p_0$, and $p=p_0$ (or p_1) according as $(q_0-p_0)>$ (or $<$)(q_1-p_1). The author defines a "combinatorial" information (information de tri) of the first kind by the operator $[\partial^i/\partial t^i]_{t=0}$, for a family of elementary events whose a priori probabilities are zero for the parameter $t=0$ and developable as a series of increasing powers of t. This is related to problems concerning the number of elementary events which have occurred under certain conditions. An information of the second kind is associated with the operator $[\,]_{t=0}$ which yields the logarithm of the number of ways in which the occurrence of n elementary events imply the occurrence of an observed event. Chi-square, defined by

$$\chi^2=\sum_i(p_i(\theta_0)-p_i(\theta_1))^2/p_i(\theta_0),$$

has certain of the properties of an information but is not additive for the composition of independent distributions, and is therefore classed by the author as a pseudo information. In connection with a discussion of the role of cumulants as information the author introduces the probability $q_t=p_t(\varphi(t))^{-1}\exp(\sqrt{(-1)}tx_t)$, where $p_t=\Pr(\zeta=x_t)$ and $\varphi(t)=\sum_t p_t\exp(\sqrt{(-1)}tx_t)$. The reviewer does not understand the interpretation of the q_t, which are in general complex, as probabilities. The reader is cautioned to look out for misprints. On page 65 the results for a and b should read

$$a=\lim_{r\to+\infty}\sup r^{-1}\log|\varphi(\sqrt{(-1)}r)|,$$

$$b=\lim_{r\to+\infty}\sup r^{-1}\log|\varphi(-\sqrt{-(1)}r)|.$$

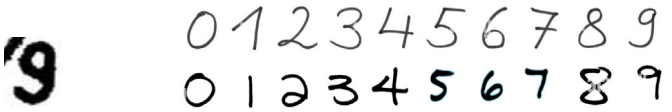
To make the number listing of the references in the bibliography of part two on pages 67–69 consistent with the references thereto in the text itself, subtract one from the listing number for items 22 to 93 inclusive.

S. Kullback (Washington, D.C.).

Fig. 4 Review of Schützenberger's thesis by Kullback in 1956 [70]. The erroneous $\frac{4}{3}$ appears in the first formula at the top of the second column. The first equal sign $=$ in this formula must in fact be a "lesser or equal" sign $\leq$

a "9". As it turns out, however, this is not the case. Thanks to his daughter Hélène Schützenberger [69] who showed us the original 1953 manuscript, it clearly appeared that the $\frac{4}{9}$ constant was correct from the start and has never been in error.

Dominique Foata, his first mathematical student [62], pointed out to us that the review of Schützenberger's thesis was in fact made in the Mathematical Reviews journal by Kullback himself in 1956 [70]. At the end of his review, Kullback declares: "*The reader is cautioned to read out for misprints*". This, however, did not prevent him from making several misprints in his review, notably on the main Schützenberger inequality as shown in Fig. 4, where the constant $\frac{4}{3}$ is incorrect. Therefore, as it turns out, Kullback was wrong about this constant as early as 1956, and reproduced the error in his 1967 paper. It is plausible that the unconventional typography of the tiny 9 in the fraction 4/9 on page 58 of Schützenberger's thesis (Fig. 3) is more reminiscent of a 3 than a 9, at least for Americans (see Fig. 5). This could also be an explanation as to why Kambo and Kotz (independently?) made the same mistake.



(a) Denominator in the fraction 4/9, zoomed in.

(b) Digits are not written exactly the same way in France (top) and in the USA (bottom)

Fig. 5 A tiny "9" that can be read as a "3" in the boxed equation in the original manuscript (Fig. 3): It is likely that in the USA, it rather follows the shape of a "3"

8.4 Subsequent works based on Schützenberger's expansion

In 1969, Krafft and Schmitz [71] extended Schützenberger's derivation by one additional term in $\frac{2}{9}\log e\cdot\Delta^6$, which was converted into a Pinsker inequality in 1975 by Toussaint [72]. But, in fact, the constant $\frac{2}{9}\log e$ is *not* optimal; the optimal constant $\frac{32}{135}\log e$ was found in 2001 by Topsøe [73]. Topsøe also derived the optimal constant for the additional term $\frac{7072}{42525}\log e\cdot\Delta^8$. It is quite remarkable that all of such derivations are crucially based on the original Schützenberger's expansion (42), which the authors of these works wrongly refer to as the Kambo-Kotz expansion (see the discussion in the above Subsection).

Proposition 21 (Topsøe's 2001–2003 contribution)

$$D \geq 2\log e\cdot\Delta^2+\frac{4}{9}\log e\cdot\Delta^4+\frac{32}{135}\log e\cdot\Delta^6+\frac{7072}{42525}\log e\cdot\Delta^8 \tag{45}$$

where the constants $2\log e$, $\frac{4}{9}\log e$, $\frac{32}{135}\log e$ *and* $\frac{7072}{42525}\log e$ *are optimal.*

Again optimality means that these constants are, in turn, not improvable: If $D\geq c_1\Delta^2+c_2\Delta^4+c_3\Delta^6+c_4\Delta^8$ with $c_i\geq 0$ then $(c_1,c_2,c_3,c_4)=(2\log e,\frac{4}{9}\log e,\frac{32}{135}\log e,\frac{7072}{42525}\log e)$ is maximal in lexicographic order.

Proof We extend the proof of Schützenberger's Theorem 19 for the third term $\frac{32}{135}\log e\cdot\Delta^6$ with a somewhat simplified presentation compared to [73], to illustrate the use of Schützenberger's expansion. The proof for the fourth term $\frac{7072}{42525}\log e\cdot\Delta^8$ is more complicated: Its sketch can be found in [74].

Following the proof of Theorem 19, taking the first four terms $k=1,2,3,4$ in Schützenberger's expansion (recall that all terms in the expansion are nonnegative) gives

$$\begin{aligned} d &\geq 2\delta^2+\frac{4}{9}\delta^4+\frac{2\delta^2}{9}(x+2y)^2+\frac{x^6-6xy^5+5y^6}{30}+\frac{x^8-8xy^7+7y^8}{56}\\ &\geq 2\delta^2+\frac{4}{9}\delta^4+\frac{1}{30}(x^6-6xy^5+5y^6)+|\delta|\sqrt{\frac{1}{63}(x+2y)^2(x^8-8xy^7+7y^8)} \end{aligned}$$

where we have used that $a+b\geq\pm2\sqrt{ab}$. In the particular case $x=-2y$, i.e., $\delta=\frac{y-x}{2}=\frac{3y}{2}$, the last term vanishes and we already have $\frac{x^6-6xy^5+5y^6}{30}=\frac{64+12+5}{30}y^6=$

$\frac{81}{30}\left(\frac{2}{3}\right)^6\delta^6 = \frac{32}{135}\delta^6$. This will prove that the constant $\frac{32}{135}$ is optimal, if we prove that in the general case,

$$\pm\delta\sqrt{\frac{1}{63}(x+2y)^2(x^8-8xy^7+7y^8)} \geq \frac{32}{135}\delta^6 - \frac{1}{30}(x^6-6xy^5+5y^6).$$

Taking the square and using Schützenberger's factorization by $(x-y)^2$ this amounts to proving the inequality $90^2\frac{\delta^2}{7}(x-y)^2(x+2y)^2(x^6+2x^5y+3x^4y^2+4x^3y^3+5x^2y^4+6xy^5+7y^6) \geq \left((x-y)^6-9(x-y)^2(x^4+2x^3y+3x^2y^2+4xy^3+5y^4)\right)^2$. Simplifying by the factors $(x-y)^2$ and $(x+2y)^2$ gives $45^2(x^6+2x^5y+3x^4y^2+4x^3y^3+5x^2y^4+6xy^5+7y^6) \geq 7(8x^3+6x^2y+9xy^2+22y^3)^2$, which is equivalent to $1577x^6+3378x^5y+4815x^4y^2+4880x^3y^3+7710^2y^4+9378xy^5+10787y^6 \geq 0$.

Since $at^2+bt \geq -\frac{b^2}{4a}$ we obtain successively $1577x^6+3378x^5y+4815x^4y^2 \geq (4815-\frac{3378^2}{4\cdot 1577})x^4y^2 = 3006.0456\ldots x^4y^2$, then $3006.0456\ldots x^4y^2+4880x^3y^3+7710x^2y^4 \geq 5729.45789\ldots x^2y^4$, and finally $5729.45789\ldots x^2y^4+9378xy^5+10787y^6 \geq 6949.5128\ldots\ldots y^6 \geq 0$, which ends the proof. □

In view of Schützenberger's essential contribution, which predates all other works on Pinsker's inequality, it is perhaps legitimate to rename the Pinsker inequality definitively as the *Schützenberger-Pinsker inequality*.

9 Other recent improvements (1970–2000s)

So far, all derived Schützenberger-Pinsker inequalities are only useful when D and Δ are small, and become uninteresting as D or Δ increases. For example, the classical inequality (31) with optimal constant $c = 2\log e$ become vacuous as soon as $D > 2\log e$ (since $\Delta \leq 1$). Any improved Schützenberger-Pinsker inequality of the form $D \geq \varphi(\Delta)$ (1) should be such that $\varphi(1) = +\infty$ because $\Delta(p,q) = 1$ (non overlapping supports) implies $D(p\|q) = +\infty$.

9.1 Vajda (1970)

The first Schützenberger-Pinsker inequality of this kind is due to Vajda in his 1970 paper [54]. He explicitly stated the problem of finding the optimal Schützenberger-Pinsker inequality (see Sect. 10) and proved the following

Proposition 22 (Vajda's 1970 contribution)

$$D \geq \log\frac{1+\Delta}{1-\Delta} - 2\log e\cdot\frac{\Delta}{1+\Delta}. \tag{46}$$

Notice that the lower bound becomes infinite as Δ approaches 1, as it should. This inequality is also asymptotically optimal near $D = \Delta = 0$ since $\log\frac{1+\Delta}{1-\Delta} - 2\log e\cdot\frac{\Delta}{1+\Delta} = 2\log e\cdot\Delta^2 + o(\Delta^2)$. We provide a simple proof along the lines of Vajda's proof [54].

Proof By binary reduction (Theorem 8), it suffices to prove that

$$d = p\log\frac{p}{p-\delta} + (1-p)\log\frac{1-p}{1-p+\delta} \geq \log\frac{1+\delta}{1-\delta} - \frac{2\delta}{1+\delta}\log e$$

where $\delta = p - q > 0$. For fixed $\delta \in [0,1]$, $d = d(p)$ is convex in $p = \delta + q \in [v, 1]$ because

$$d'(p) = \left(\log\frac{p}{p-\delta} - \frac{\delta}{p-\delta}\right) - \left(\log\frac{1-p}{1-p+\delta} + \frac{\delta}{1-p+\delta}\right)$$

and

$$d''(p) = \frac{\delta^2}{p(p-\delta)^2} + \frac{\delta^2}{(1-p)(1-p+\delta)^2} \geq 0.$$

It follows that

$$d(p) \geq d\left(\frac{1+\delta}{2}\right) + \left(p - \frac{1+\delta}{2}\right)d'\left(\frac{1+\delta}{2}\right)$$

where

$$d'\left(\frac{1+\delta}{2}\right) = 2\log\frac{1+\delta}{1-\delta} - \frac{2\delta}{1-\delta} - \frac{2\delta}{1+\delta} = 2\left(\log\frac{1+\delta}{1-\delta} - \frac{2\delta}{1-\delta^2}\right)$$

is nonpositive—indeed, as shown by Toussaint [72],

$$\log\tfrac{1+\delta}{1-\delta} = \log(1+\delta) - \log(1-\delta) = \int_0^\delta \frac{1}{1+t} + \frac{1}{1-t}\,\mathrm{d}t = \int_0^\delta \frac{2}{1-t^2}\,\mathrm{d}t$$
$$\leq \frac{2}{1-\delta^2}\int_0^\delta \mathrm{d}t = \frac{2\delta}{1-\delta^2}.$$

Now since $p \leq 1$,

$$d(p) \geq d\left(\frac{1+\delta}{2}\right) + \left(1 - \frac{1+\delta}{2}\right)d'\left(\frac{1+\delta}{2}\right)$$
$$= \delta\log\frac{1+\delta}{1-\delta} + 2\frac{1-\delta}{2}\left(\log\frac{1+\delta}{1-\delta} - \frac{2\delta}{1-\delta^2}\right) = \log\frac{1+\delta}{1-\delta} - \frac{2\delta}{1+\delta}.$$

□

9.2 Bretagnolle and Huber (1978)

In a 1978 French seminar, Bretagnolle and Huber [75] derived yet another Schützenberger-Pinsker inequality similar to Vajda's (where the lower bound becomes infinite for $\Delta = 1$) but with a simpler expression:

Proposition 23 (Bretagnolle–Huber 1978 contribution)

$$D \geq \log \frac{1}{1-\Delta^2}. \tag{47}$$

A nice property of this inequality is that it can be inverted in closed form. In fact the authors expressed it as

$$\Delta \leq \sqrt{1-\exp(-D)}. \tag{48}$$

Proof Write

$$\begin{aligned}\exp(-D) &= \exp \int p \log \frac{q}{p} = \left(\exp \int p \log \sqrt{\frac{q}{p}}\right)^2 \\ &\leq \left(\int p \exp \log \sqrt{\frac{q}{p}}\right)^2 = \left(\int \sqrt{pq}\right)^2\end{aligned}$$

by Jensen's inequality applied to the exponential. Therefore, by the Cauchy-Schwarz inequality,

$$\exp(-D) \leq \left(\int \sqrt{p \wedge q}\sqrt{p \vee q}\right)^2 \leq \int p \wedge q \int p \vee q = (1-\Delta)(1+\Delta) = 1-\Delta^2.$$

□

Remark 12 (Comparison to Vajda's inequality) Even though their work is more recent, Bretagnolle and Huber were probably unaware of Vadja's inequality (46), which is uniformly stronger. In fact, by the comparison principle (Theorem 9), for $0 < \Delta < 1$, $\frac{d}{d\Delta} \log \frac{1}{1-\Delta^2} = \frac{2\Delta}{1-\Delta^2} < \frac{4\Delta}{(1+\Delta)(1-\Delta^2)} = \frac{d}{d\Delta}\left(\log \frac{1+\Delta}{1-\Delta} - \frac{2\Delta}{1+\Delta}\right)$ always, since $1+\Delta < 2$. Therefore, the Bretagnolle-Huber inequality (47) is strictly *weaker* than Vajda's inequality (46).

Moreover, it is not asymptotically optimal near $D = \Delta = 0$ since $\log \frac{1}{1-\Delta^2} \sim \log e \cdot \Delta^2$ is worse than the asymptotically optimal $2 \log e \cdot \Delta^2$.

Remark 13 (Tsybakov's 2009 version) The Bretagnolle-Huber inequality was popularized in their subsequent 1979 paper [76], taken up by Tsybakov in his 2009 book on nonparametric estimation [57, Eq. (2.25)], under the form

$$D \geq \log \frac{1}{2(1-\Delta)} \tag{49}$$

or

$$\Delta \leq 1 - \frac{1}{2}\exp(-D). \tag{50}$$

This, however, is strictly *weaker* than the original, since $1-\Delta^2 = (1-\Delta)(1+\Delta) < 2(1-\Delta)$ for $0 < \Delta < 1$. Moreover, it becomes vacuous as soon as $\Delta < \frac{1}{2}$ since then

the lower bound in (49) is negative. Therefore, such an inequality is not a genuine Pinsker-type inequality in the sense of Definition 1, and cannot be used for small values of Δ and D, e.g., to prove continuity properties (see Remark 11).

Recently, an improvement of (49) was proposed in [77, § 8.3]:

Proposition 24 (Gerchinovitz–Menard–Stoltz 2020 contribution)

$$D \geq \log \frac{\gamma}{1-\Delta} \tag{51}$$

or

$$\Delta \leq 1 - \gamma \exp(-D). \tag{52}$$

where $\gamma = e^{-1/e} \approx 0.6922 > \frac{1}{2}$.

Proof Using binary reduction (Theorem 8), we may assume that $\delta = p - q > 0$. Since $x \log x \geq -\frac{\log e}{e} = \log \gamma$ (see Definition 4) and $1 - q < 1$, one has

$$d \geq p \log \frac{p}{q} + p \log \frac{1}{1-q} + \log \gamma \geq p \log \frac{p}{q} + \log \gamma = -p \log \frac{q}{p\gamma} - (1-p) \log \frac{1}{\gamma},$$

hence by Jensen's inequality for the exponential,

$$\exp(-d) \leq p \frac{q}{p\gamma} + \frac{1-p}{\gamma} = \frac{1-\delta}{\gamma},$$

which gives the announced inequality. □

Remark 14 (Comparison to Vajda's inequality) Such an inequality suffers from the same disadvantage as in Remark 13 because it cannot be used for small values of Δ or D. Furthermore, just like the original Bretagnolle-Huber inequality (Remark 12), it is uniformly weaker than Vajda's inequality as can be easily checked.

Even more recently, Canonne [78] has shown that an inequality similar to (but uniformly worse than) the Bretagnolle-Huber inequality can be derived form the classical Schützenberger-Pinsker inequality:

Proposition 25 (Canonne's 2023 contribution)

$$D \geq \log \frac{1}{1-\beta\Delta^2} \tag{53}$$

or

$$\Delta \leq \sqrt{\frac{1-\exp(-D)}{\beta}}; \tag{54}$$

where $\beta = 1 - e^{-2} \approx 0.86466 < 1$.

Proof Since $\exp((\log e)x) = e^x$, the classical Schützenberger-Pinsker inequality $D \geq 2\log e \cdot \Delta^2$ rewrites

$$1 - e^{-2\Delta^2} \leq 1 - \exp(-D) \tag{55}$$

Now for $x \in [0,1]$, $f(x) = 1 - e^{-2x}$ is concave so that it lies above its chord $f(x) \geq f(0) + (f(1) - f(0))(x - 0)$, that is, $1 - e^{-2x} \geq (1 - e^{-2})x = \beta x$. Substituting $x = \Delta^2$ yields the announced inequality. □

Remark 15 (Comparison to other inequalities) Even though it is uniformly weaker than the Bretagnolle-Huber inequality, Canonne's inequality can still be used for small values of Δ and D, contrary to the Tsybakov inequality (49) and its improved version (51). However, because it is derived from (and uniformly weaker than) the classical Schützenberger-Pinsker inequality, it becomes vacuous as soon as $D > 2\log e$.

Remark 16 (Testing fair vs. unfair coin) Going back to the testing example from Sect. 4, Canonne [78] has studied the estimate (22) for some of the inequalities $D \geq \varphi(\Delta)$ of this subsection. Since the ϵ and δ parameters should be small, only those Pinsker-type inequalities which do not become vacuous for large D can be used. The original Bretagnolle-Huber inequality (47) yields a particularly simple estimate, which can be written as

$$(1 - 4\delta^2)^{n/2} \leq 4(\epsilon - \epsilon^2). \tag{56}$$

The Tsybakov version (49) yields a slightly weaker (and simpler) inequality, where $(\epsilon - \epsilon^2)$ is replaced by ϵ.

9.3 Gilardoni (2008)

Today and to the knowledge of the author, the best known explicit Schützenberger-Pinsker inequality of this kind was derived by Gilardoni in 2008 [79] (see also [80]):

Proposition 26 (Gilardoni's 2008 contribution)

$$D \geq \log \frac{1}{1-\Delta} - (1-\Delta)\log(1+\Delta). \tag{57}$$

Gilardoni's proof is based on considerations on symmetrized f-divergences. A simple proof is as follows:

Proof One can always assume that $\delta = p - q > 0$, where $\delta \leq p \leq 1$ and $0 \leq q \leq 1 - \delta$. Then

$$\begin{aligned} d(p\|q) &= (q+\delta)\log\frac{q+\delta}{q} + (1-q-\delta)\log\frac{1-q-\delta}{1-q} \\ &= \left[-q\log\frac{q+\delta}{q} - (1-q-\delta)\log\frac{1-q}{1-q-\delta}\right] + (2q+\delta)\log\frac{q+\delta}{q}. \end{aligned}$$

Since $q + (1-q-\delta) = 1-\delta$ and $-\log$ is convex, the first term inside brackets is $\geq -(1-\delta)\log(\frac{q+\delta}{1-\delta} + \frac{1-q}{1-\delta}) = (1-\delta)\log\frac{1-\delta}{1+\delta}$. The second term writes $\delta\frac{(2+x)\log(1+x)}{x}$ where $x = \frac{\delta}{q}$. Now $(2+x)\log(1+x)$ is convex for $x \geq 0$ and vanishes for $x = 0$, hence the slope $\frac{(2+x)\log(1+x)}{x}$ is minimal for minimal x, that is, for maximal $q = 1-\delta$. Therefore, the second term is $\geq (2-2\delta+\delta)\log\frac{1}{1-\delta} = (2-\delta)\log\frac{1}{1-\delta}$. Summing the two lower bounds gives the inequality. □

Remark 17 (Comparison to other inequalities) Note that Gilardoni's inequality adds the term $\Delta\log(1+\Delta)$ to the Bretagnolle-Huber lower bound, hence uniformly improves it. In fact, it also uniformly improves the stronger Vajda's inequality since by the comparison principle (Theorem 9) for $\Delta > 0$, $\frac{d}{d\Delta}\big(\log\frac{1}{1-\Delta} - (1-\Delta)\log(1+\Delta)\big) = \Delta\frac{3-\Delta}{1-\Delta^2} + \log(1+\Delta) > \Delta\frac{3-\Delta}{1-\Delta^2} + \Delta - \frac{\Delta^2}{2} = \frac{4\Delta}{(1+\Delta)(1-\Delta^2)} - \Delta\frac{1-\Delta}{(1+\Delta)^2} + \Delta(1-\frac{\Delta}{2}) > \frac{4\Delta}{(1+\Delta)(1-\Delta^2)} = \frac{d}{d\Delta}\big(\log\frac{1+\Delta}{1-\Delta} - \frac{2\Delta}{1+\Delta}\big)$. In particular, it is also asymptotically optimal near $D = \Delta = 0$, which can easily be checked directly: $\log\frac{1}{1-\Delta} - (1-\Delta)\log(1+\Delta) = 2\log e\cdot\Delta^2 + o(\Delta^2)$.

However, Gilardoni's inequality is still weaker than the classical Schützenberger-Pinsker inequality for small Δ or D. In fact, by the comparison principle (Theorem 9), for natural logarithms and $\Delta > 0$, $\frac{d}{d\Delta}\big(\log\frac{1}{1-\Delta} - (1-\Delta)\log(1+\Delta)\big) = \Delta\frac{3-\Delta}{1-\Delta^2} + \log(1+\Delta) < 3\Delta+\Delta = 4\Delta = \frac{d}{d\Delta}(2\Delta^2)$ as soon as $\Delta \geq 3\Delta^2$, i.e., $\Delta \leq \frac{1}{3}$. Therefore, Gilardoni's inequality (57) is strictly weaker than the classical Schützenberger-Pinsker inequality *at least* for $0 < \Delta < 1/3$ (in fact for $0 < \Delta < 0.569\ldots$). For Δ close to 1, however, Gilardoni's inequality is obviously better.

10 The optimal Schützenberger-Pinsker inequality

The problem of finding the *optimal* Schützenberger-Pinsker inequality, that is, the best possible lower bound in (1):

$$\varphi^*(\Delta) = \inf_{\Delta(p,q)=\Delta} D(p\|q) \tag{58}$$

was opened by Vajda [54] in 1970.

10.1 Fedotov, Harremoës, and Topsøe (2003)

Vajda's problem was solved in 2003 in *implicit* form, using the Legendre-Fenchel transformation, by Fedotov, Harremoës, and Topsøe in [74], as a curve parametrized by hyperbolic trigonometric functions. We give the following equivalent but simpler parametrization.

Theorem 27 (Optimal Schützenberger-Pinsker inequality) *The optimal Schützenberger-Pinsker inequality $D \geq \varphi^*(\Delta)$ is given in parametric form as*

$$\begin{cases} \Delta &= \lambda(1-q)q \\ D &= \log(1-\lambda q) + \lambda q(1+\lambda(1-q))\log e \end{cases} \tag{59}$$

where $\lambda \geq 0$ is the parameter and

$$q = q(\lambda) \triangleq \frac{1}{\lambda} - \frac{1}{e^\lambda - 1} \in [0, \tfrac{1}{2}]. \tag{60}$$

The following proof that is arguably simpler as the one in [74] since it only relies of the well-known Lagrange multiplier method.

Proof Using binary reduction (Theorem 8), $d(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is to be minimized under the linear constraint $p - q = \delta \in [-1, 1]$. It is well known that divergence $d(p\|q)$ is strictly convex in (p, q). Given that the objective function is convex and the constraint is linear, the solution can be given by the Lagrange multiplier method. The Lagrangian is $L(p, q) = d(p\|q) - \lambda(p-q)$ and the solution is obtained as global minimum of L, which by convexity is obtained by setting the gradient w.r.t. p and q to zero. Assuming natural logarithms for simplification, this gives

$$\begin{cases} \frac{\partial L}{\partial p} = \log\frac{p}{q} - \log\frac{1-p}{1-q} - \lambda = 0 \\ \frac{\partial L}{\partial q} = -\frac{p}{q} + \frac{1-p}{1-q} + \lambda = 0 \end{cases} \quad \text{or} \quad \begin{cases} e^\lambda = \frac{p}{q}\Big/\frac{1-p}{1-q} \\ \lambda = \frac{p}{q} - \frac{1-p}{1-q} \end{cases}. \tag{61}$$

Therefore, $\frac{p}{q} = \lambda + \frac{1-p}{1-q} = e^\lambda\frac{1-p}{1-q}$, and we have $\frac{1-p}{1-q} = \frac{\lambda}{e^\lambda-1}$ and $\frac{p}{q} = \frac{\lambda e^\lambda}{e^\lambda-1}$. Solving for q, then for p, one obtains $1 = 1 - p + p = (1-q)\frac{\lambda}{e^\lambda-1} + q\frac{\lambda e^\lambda}{e^\lambda-1}$, which gives $q = q(\lambda) = \frac{1}{\lambda} - \frac{1}{e^\lambda-1}$ as announced above and $p = q\lambda(1 + \frac{1}{e^\lambda-1}) = q\lambda(1 + \frac{1}{\lambda} - q) = q(1+\lambda(1-q))$. Therefore, we obtain the desired parametrization $\delta = p - q = \lambda(1-q)q$ and $d(p\|q) = \log\frac{1-p}{1-q} + p\lambda = \log(1-\lambda q) + \lambda q(1+\lambda(1-q))$ with natural logarithm (which is multiplied by $\log e$ for logarithm to any base).

Finally, observe that the transformation $(p, q) \mapsto (1-p, 1-q)$ leaves $d = d(p\|q)$ unchanged but changes $\delta \mapsto -\delta$. In the parametrization, this changes $\lambda \mapsto -\lambda$ and $q(\lambda) \mapsto q(-\lambda) = 1 - q(\lambda)$. Accordingly, this change of parametrization changes $(\delta, d) \mapsto (-\delta, d)$ as can be easily checked. Therefore, the resulting optimal φ^* is even. Restricting to $\delta = |p-q| = p - q \geq 0$ amounts to $p \geq q \iff \lambda \geq 0 \iff q \in [0, 1/2]$. □

Remark 18 In 2009, Reid and Williamson [81, 82], using a particularly lengthy proof mixing learning theory, 0-1 Bayesian risks, and integral representations of f-divergences, claimed the following "explicit form" of the *optimal* Schützenberger-Pinsker inequality:

$$D \geq \min_{|\beta| \leq 1-\Delta} \frac{1+\Delta-\beta}{2}\log\frac{1+\Delta-\beta}{1-\Delta-\beta} + \frac{1-\Delta+\beta}{2}\log\frac{1-\Delta+\beta}{1+\Delta-\beta}. \tag{62}$$

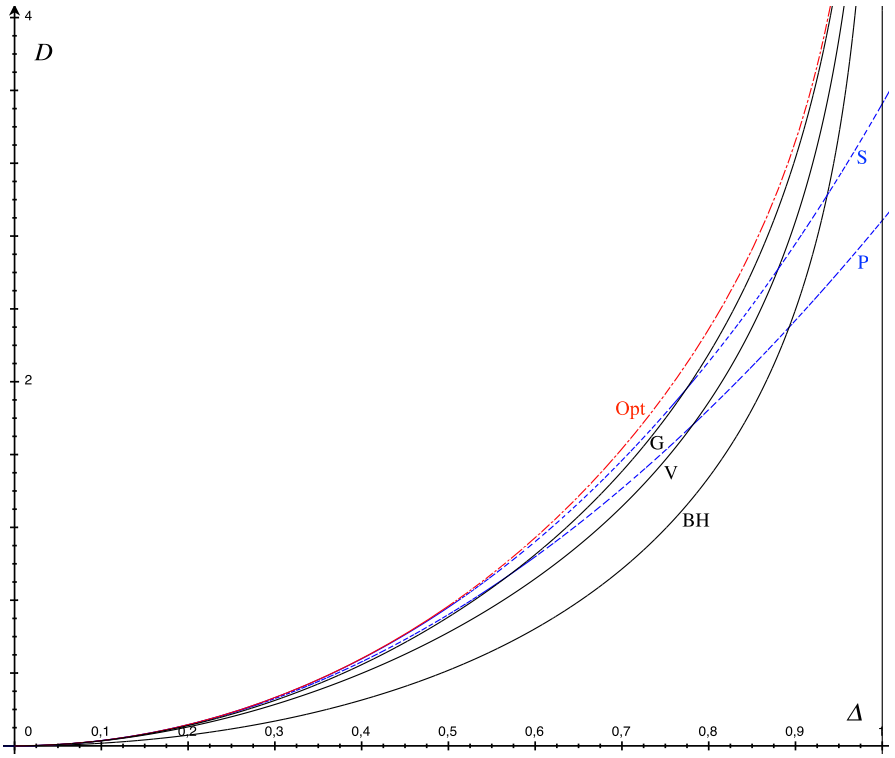


Fig. 6 Schützenberger-Pinsker lower bounds of divergence D (with logarithm to base 2) vs. total variation Δ. Red, dashdotted: Optimal (Opt, Theorem 27). Blue, dashed: Pinsker (P, Eq. 31 with $c = 2\log e$) with optimal constant and Schützenberger (S, Eq. 41). Black, solid: Bretagnolle-Huber (BH, Eq. 47), Vajda (V, Eq. 46) and Gilardoni (G, Eq. 57)

This formula, however, is just a trivial reformulation of the optimal lower bound: Indeed, by binary reduction (Theorem 8), $d(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is to be minimized under the constraint $\delta = p-q$, hence $\delta \le p \le 1$ and $q \le 1-\delta$. Letting $\beta = 1-p-q$, this amounts to minimizing over β in the interval $[\delta-1, 1-\delta]$ for fixed $\delta = p-q$. Since $p = \frac{1+\delta-\beta}{2}$ and $q = p-v = \frac{1-\delta-\beta}{2}$, this boils down to the above expression (62) for the lower bound, where the minimization problem over β is not solved in [81, 82].

10.2 Concluding remarks

Figure 6 illustrates the main Schützenberger-Pinsker inequalities seen in this paper. From the implicit form using the exact parametrization of Theorem 27, we conjecture that the optimal Schützenberger-Pinsker inequality cannot be written as a *closed-form* expression with standard operations and functions. Also, the problem of finding an explicit and reasonably simple Schützenberger-Pinsker inequality which *uniformly*

improves all the preceding ones (in particular, the classical inequality with optimal constant and Gilardoni's inequality) is still open.

Referring to Fig. 6, asymptotic optimality near the two extremes ($V = D = 0$ as $\lambda \to 0$ or $V = 1$, $D = +\infty$ as $\lambda \to \infty$) can easily be obtained from the parametrization of Theorem 27:

- As $\lambda \to 0$, by Taylor expansion one obtains $q = \frac{1}{2} - \frac{\lambda}{12} + o(\lambda)$, $\Delta = \frac{\lambda}{4} + o(\lambda)$, and $D = \frac{\lambda^2}{8} \log e + o(\lambda^2)$. Thus, one recovers that $D \sim 2\Delta^2 \log e$ near $D = \Delta = 0$. In particular, the classical inequality (with optimal constant) and its improvements with higher-order terms, as well as Vajda's and Gilardoni's inequality, are asymptotically optimal near $D = \Delta = 0$.
- As $\lambda \to +\infty$, $q = \frac{1}{\lambda} + o(\frac{1}{\lambda})$, $\exp d = \frac{\lambda}{e^\lambda - 1} e^{\lambda + o(1)} \sim \lambda \sim \frac{1}{1-\Delta}$. Thus it follows that $\exp D \sim \frac{1}{1-\Delta}$ near $\Delta = 1$ and $D = +\infty$. Vajda's and the Bretagnolle-Huber inequalities are such that $\exp D \sim \frac{c}{1-\Delta}$ there, with suboptimal constants $c = \frac{2}{e} = 0.7357\ldots < 1$ and $c = \frac{1}{2} < 1$, respectively. Only Gilardoni's inequality is optimal in this region with $c = 1$.

As a mathematical perspective, one may envision that the exact parametrization of Theorem 27 can be exploited to find new explicit bounds. Indeed, since $\lambda = \varphi^{*\prime}(\Delta)$ in the parametrization of Theorem 27, from the comparison principle (Theorem 9), any inequality of the form $\varphi'(\Delta) \le \lambda = \varphi^{*\prime}(\Delta)$ is equivalent to a corresponding Schützenberger-Pinsker inequality (1) associated to φ. For example, since $4\Delta = 4\lambda(1-q)q \le \lambda$ always in the parametrization, one recovers the classical inequality (31) with optimal constant $c = 2 \log e$. Thus, the search of new Schützenberger-Pinsker inequality amounts to solving the inequality in $\lambda > 0$: $\varphi'\big(\lambda\big(1 - q(\lambda)\big)q(\lambda)\big) \le \lambda$ for φ.

On the historical side, the influence of Schützenberger on the derivation of the classical "Pinsker's inequality" still has some mysteries, particularly considering the erroneous constant $\frac{4}{3}$ and the fact that this error was made apparently independently by different authors. It is important to note, however, that Marco Schützenberger did derive the correct constants in 1953, the first-order constant 7 years before Pinsker and the second-order constant 17 years before it was re-established by Kullback and Vajda.

As another historical perspective [83], it turns out that Schützenberger had derived in 1957 another important inequality in statistics, the Bayesian version of the Fréchet-Darmois-Cramér-Rao inequality [60, 61], more commonly known as the van Trees inequality, which van Trees discovered independently in 1968. Thus, Schützenberger did prove the famous "van Trees inequality" 11 years before van Trees, and this inequality should be called as the Schützenberger-van Trees inequality—just as the famous Pinsker inequality should be called the Schützenberger-Pinsker inequality.

Acknowledgements The author is indebted to his beloved friend Hélène Fache, teacher at the Lycée Hoche in Versailles, who recalled that she had taught François Bojarski in class, who happens to be Marco Schützenberger's grandson. Many thanks also to his mother Hélène Schützenberger, who showed her father's original thesis manuscript, told many anecdotes about him and put the author in touch with several people who knew him well. Jean Berstel's website http://www-igm.univ-mlv.fr/~berstel/Mps/ is a gold mine of information on Schützenberger. Fruitful discussions with his students Dominique Foata and Dominique Perrin are also gratefully acknowledged.

Author Contributions O.R. wrote the manuscript.

Data Availability Not applicable. No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The author states that there is no Conflict of interest.

References

1. Amari, S.: Information geometry and its applications. Applied mathematical sciences, vol. 194. Springer, Berlin (2016)
2. Lévy, P.: Théorie de l'addition des variables aléatoires. Gauthier-Villars, Paris (1937)
3. Prokhorov, Y.V.: Convergence of random processes and limit theorems in probability theory. Theory Probab. Appl. **1**(2), 157–214 (1956)
4. Fortet, R., Mourier, E.: Convergence de la répartition empirique vers la répartition théorique. Annales scientifiques de l'École Normale Supérieure **70**(3), 267–285 (1953)
5. Vaserštein, L.N.: Processes over denumerable products of spaces, describing large systems of automata. Probl. Peredachi Inf. **5**(3), 64–72 (1969)
6. Kantorovich, L.V.: Mathematical methods of organizing and planning production (in russian). Manage. Sci. **6**(4), 366–422 (1939)
7. Kantorovich, L.V., Rubinstein, G.S.: On a functional space and certain extremal problems (in russian). Dokl. Akad. Nauk SSSR **115**, 1058–1061 (1957)
8. Kolmogorov, A.N.: Sulla determinazione empirica di una legge di distribuzione. G. Ist. Ital. Attuari **4**, 83–91 (1933)
9. Smirnov, N.: Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat. **19**(2), 279–281 (1948)
10. Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proc. R. Soc. Lond. A **186**(1007), 453–461 (1946)
11. Endres, D.M., Schindelin, J.E.: A new metric for probability distributions. IEEE Trans. Inf. Theory **49**(7), 1858–1860 (2003)
12. Vincze, I.: On the concept and measure of information contained in an observation. In: Gani, J., Rohatgi, V.F. (eds.) Contributions to probability, pp. 207–214. Academic Press, New York (1981)
13. Le Cam, L.: Asymptotic methods in statistical decision theory. Springer Series in Statistics. Springer, Berlin (1986)
14. Fan, K.: Entfernung zweier zufälligen größen and die konvergenz nach wahrscheinlichkeit. Math. Z. **49**, 681–683 (1944)
15. Hellinger, E.: Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. Jfür die reine und angewandte Mathematik **136**, 210–271 (1909)
16. Topsøe, F.: Some inequalities for information divergence and related measures of discrimination. IEEE Trans. Inf. Theory **46**(4), 1602–1609 (2000)
17. Barron, A.R.: Entropy and the central limit theorem. Ann. Probab. **14**(1), 336–342 (1986)
18. Rényi, A.: Az információelmélet néhány alapvető kérdése (Some basic questions in information theory). Magyar Tud. Akad. Mat. Fiz. Oszt. Közl **10**, 251–282 (1960)
19. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
20. Sundaresan, R.: Guessing under source uncertainty. IEEE Trans. Inf. Theory **53**(1), 269–287 (2007)
21. Principe, J.C., Xu, D., Fisher, J.W., III.: Information-theoretic learning. Unsupervised Adapt Filter **1**, 265–319 (2000). (**Chap. 7**)
22. Csiszár, I.: Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. Publ. Math. Inst. Hungar. Acad. Sci. A **8**, 85–108 (1963)
23. Pearson, K.: 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. Philos Mag Ser 5 **50**(302), 157–175 (1900)

24. Bhattacharyya, A.K.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. **35**, 99–109 (1943)
25. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. Math. Phys. **7**(3), 200–217 (1967)
26. Mahalanobis, P.C.: On the generalised distance in statistics. Proc Natl Inst Sci India **2**(1), 49–55 (1936)
27. Itakura, F., Saito, S.: Analysis Synthesis Telephony Based on the Maximum Likelihood Method. In: Proc. 6th International Congress on Acoustics, Los Alamitos, CA, pp. 17–20 (1968)
28. Khosravifard, M., Fooladivanda, D., Gulliver, T.A.: Confliction of the convexity and metric properties in f-divergences. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E90–A**(9), 1848–1853 (2007)
29. Vajda, I.: On metric divergences of probability measures. Kybernetika **45**(6), 885–900 (2009)
30. Jiao, J., Courtade, T.A., No, A., Venkat, K., Weissman, T.: Information measures: The curious case of the binary alphabet. IEEE Trans. Inf. Theory **60**(12), 7616–7626 (2014)
31. Belavkin, R.V.: Asymmetric topologies on statistical manifolds. In: Nielsen, F., Barbaresco, F. (eds.) Geometric science of information. Lecture notes in computer science, vol. 9389, pp. 203–210. Springer, Berlin (2015)
32. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Stud. Sci. Math. Hung. **2**, 299–318 (1967)
33. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. Int. Stat. Rev. **70**(3), 419–435 (2002)
34. Polyanskiy, Y., Wu, Y.: Information theory: from coding to learning, 1st edn. Cambridge University Press, Cambridge (2023)
35. Shoup, V.: A computational introduction to number theory and algebra, 2nd edn. Cambridge University Press, Cambridge (2009)
36. Jordan, C.: Sur la série de Fourier. C. R. Hebd. Seances Acad. Sci. **92**, 228–230 (1881)
37. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, Hoboken. (1st Ed. 1990, 2nd Ed. 2006)
38. Rioul, O.: This is IT A primer on Shannon's entropy and information. In: Duplantier, B., Rivasseau, V. (eds.) Information Theory Poincaré Seminar 2018. Progress in Mathematical Physics, pp. 49–86. Springer, Cham (2021)
39. Csiszár, I.: A note on Jensen's inequality. Stud. Sci. Math. Hung. **1**, 185–188 (1966)
40. Csiszár, I., Körner, J.: Information Theory. Coding Theorems for Discrete Memoryless Systems, 2nd edn. Cambridge University Press, Cambridge (2011) (1st edn., 1981)
41. Csiszár, I., Shields, P.C.: Information theory and statistics: a tutorial. Foundations and trends in communications and information theory. Now Publishers Inc., Hanover (2004)
42. Pinsker, M.S.: Information and Information Stability of Random Variables and Processes. Izv. Akad. Nauk, Moscow (1960) English translation Holden-Day, San Francisco (1964)
43. Gel'fand, S.I., Yaglom, A.M.: О вычислении количества информации о случайной функции, содержащейся в другой такой функции (calculation of the amount of information about a random function contained in another such function). Usp. Mat. Nauk. **12**(1), 3–52 (1959)
44. Perez, A.: Information theory with an abstract alphabet. Generalized forms of McMillan's limit theorem for the case of discrete and continuous times. Theory of Probability & Its Applications **4**(1), 99–102 (1959)
45. Prest, T., Goudarzi, D., Martinelli, A., Passelègue, A.: Unifying Leakage Models on a Rényi Day. In: Advances in Cryptology - CRYPTO 2019 - 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2019, Proceedings, Part I, pp. 683–712 (2019)
46. Fano, R.M.: Class Notes for Course 6.574: Transmission of Information. MIT, Cambridge. (1952)
47. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3 & 4), 379–423623656 (1948)
48. Kullback, S.: A lower bound for discrimination information in terms of variation. IEEE Trans. Inf. Theory **13**, 126–127 (1967)
49. Kemperman, J.H.B.: On the optimum rate of transmitting information. Ann. Math. Stat. **40**(6), 2156–2177 (1969)
50. Verdú, S.: Total Variation Distance and the Distribution of Relative Information. In: 2014 Information Theory and Applications Workshop (ITA), San Diego, CA, USA (2014)

51. Volkonskii, V.A., Rozanov, Y.A.: Some limit theorems for random functions. I (English translation from Russian). Theory of Probability and its Applications **IV**(2), 178–197 (1959)
52. Sakaguchi, M.: Information Theory and Decision Making. unpublished, George Washington University, Whasington D.C. (1964)
53. McKean, H.P., Jr.: Speed of approach to equilibrium for Kac's caricature of a Maxwellian gas. Arch. Rational Mech. Anal. **21**, 343–367 (1966)
54. Vajda, I.: Note on discrimination information and variation. IEEE Trans. Inf. Theory **16**, 771–773 (1970)
55. Kullback, S.: Correction to "A lower bound for discrimination information in terms of variation". IEEE Trans. Inf. Theory **16**, 652 (1970)
56. Kemperman, J.H.B.: On the optimum rate of transmitting information. In: Proceedings of the International Symposium on Probability and Information Theory, pp. 126–169. Springer, Hamilton, Ontario, Canada (1968)
57. Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer Series in Statistics. Springer, Berlin (2009)
58. Edwards, R., Reader, K.: The Papin Sisters. Oxford Studies in Modern European Culture. Oxford University Press, USA (1984)
59. Schützenberger, M.-P.: Contribution aux Applications Statistiques de la Théorie de L'information vol. 3, No 1–2. Institut de statistique de l'Université de Paris, Paris (1954) Thèse de doctorat (1953)
60. Schützenberger, M.-P.: A generalization of the Fréchet-Cramér inequality to the case of Bayes estimation. Bulletin of the American Mathematical Society **63** (1957)
61. Schützenberger, M.-P.: À propos de l'inégalité de Fréchet-Cramér. Publications de l'Institut de statistique de l'Université de Paris **7**(3–4), 3–6 (1958)
62. Foata, D.: Private communication (2023)
63. Perrin, D.: Private communication (2023)
64. Schützenberger, M.-P.: On the definition of a family of automata. Inf. Control **4**(2–3), 245–270 (1961)
65. Chomsky, N., Schützenberger, M.-P.: The algebraic theory of context-free languages. In: Braffort, P., Hirschberg, D. (eds.) Computer programming and formal languages, pp. 118–161. North Holland, Amsterdam (1963)
66. Bojarski, F.: Private communication (2023)
67. Kambo, N.S., Kotz, S.: On exponential bounds for binomial probabilities. Ann. Inst. Statist. Math. **18**, 277–287 (1966)
68. Krafft, O.: A note on exponential bounds for binomial probabilities. Ann. Institut für Mathematische Statistik **21**, 219–220 (1969)
69. Schützenberger, H.: Private communication (2023)
70. Kullback, S.: Review of "contribution aux applications statistiques de la théorie de l'information" by M. P. Schützenberger. Available: https://mathscinet.ams.org/mathscinet/article?mr=77816. Mathematical Reviews **17**(10), 1099 (1956). Accessed 11 May 2024
71. Krafft, O., Schmitz, N.: A note on Hoeffding's inequality. J. Am. Stat. Assoc. **64**(327), 907–912 (1969)
72. Toussaint, G.T.: Sharper lower bounds for discrimination information in terms of variation. IEEE Trans. Inf. Theory **21**(1), 99–100 (1975)
73. Topsøe, F.: Bounds for entropy and divergence for distributions over a two-element set. J Inequal Pure Appl Math **2**(2, Art 25), 1–13 (2001)
74. Fedotov, A.A., Harremoës, P., Topsøe, F.: Refinements of Pinsker's inequality. IEEE Trans. Inf. Theory **49**(6), 1491–1498 (2003)
75. Bretagnolle, J., Huber, C.: Estimation des densités : risque minimax. In: Séminaire de Probabilités (Strasbourg 1976/77), vol. 12, pp. 342–363. Springer, Berlin (1978)
76. Bretagnolle, J., Huber, C.: Estimation des densités: risque minimax. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **47**, 119–137 (1979)
77. Gerchinovitz, S., Ménard, P., Stoltz, G.: Fano's inequality for random variables. Stat. Sci. **35**(2), 178–201 (2020)
78. Canonne, C.L.: A short note on an inequality between KL and TV. arxiv:2202.07198v2 (2023). Accessed 11 May 2024
79. Gilardoni, G.L.: An improvement on Vajda's inequality. In: cIn and out of equilibrium 2. Progress in probability, vol. 60, pp. 299–304. Birkhäuser, Basel (2008)
80. Gilardoni, G.L.: On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. IEEE Trans. Inf. Theory **56**(11), 5377–5386 (2010)

81. Reid, M.D., Williamson, R.C.: Generalised Pinsker Inequalities. In: 22nd Annual Conference on Learning Theory (COLT 2009), Montreal, Canada (2009)
82. Reid, M.D., Williamson, R.C.: Information, divergence and risk for binary experiments. J. Mach. Learn. Res. **12**, 731–817 (2011)
83. Renaux, A., Rioul, O.: In preparation (2024)