A historical perspective on Schützenberger-Pinsker inequalities

Olivier Rioul

A Historical Perspective on Schützenberger-Pinsker Inequalities

Olivier Rioul[0000−0002−8681−8916]

LTCI, Télécom Paris, Institut Polytechnique de Paris, France
olivier.rioul@telecom-paris.fr
https://perso.telecom-paristech.fr/rioul/

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 information theoretic 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 Schûtzenberger and Pinsker contributions as well as those of Volkonskii & Rozanov, Sakaguchi, McKean, Csiszár, Kullback, Kemperman, Vajda, Bretagnolle & Huber, Krafft & Schmitz, Toussaint, Reid & Williamson, Gilardoni, as well as the optimal derivation of Fedotov, Harremoës, & Topsøe.

Keywords: Pinsker inequality · Total variation · Kullback-Leibler divergence · Statistical Distance · Mutual Information · Data processing inequality.

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. The common viewpoint is to define a "distance" $\Delta(p, q)$ between probability distributions p and q, which should at least 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 include the Lévy-Prokhorov and the Fortet-Mourier (a.k.a. "Wasserstein" or Kantorovich-Rubinstein) distances (which metrize the weak conver-

gence or convergence in distribution), the (stronger) Kolmogorov-Smirnov distance (which metrizes the uniform convergence in distribution), the Radon distance (which metrizes the strong convergence), the Jeffreys (a.k.a. Hellinger[1]) distance, and many others[2].

In this paper, we focus on the *total variation distance*, which is one of the strongest among the preceding examples. Arguably, it is also the simplest—as a L^1-norm distance—and the most frequently used in applications, particularly those related to Bayesian inference.

In many information theoretic applications, however, other types of "distances," that do not necessarily satisfy the triangle inequality, are often preferred. Such "distances" are called *divergences* $D(p,q)$. They may not even satisfy the symmetry property: In general, $D(p,q)$ is the divergence *of* q *from* p, and not "between p and q"[3]. Examples include the Rényi α-divergence, the Bhattacharyya divergence (a variation of the Jeffreys (Hellinger) distance), Lin's "Jensen-Shannon" divergence, the triangular divergence, Pearson's χ^2 divergence, the "Cauchy-Schwarz" divergence, the (more general) Sundaresan divergence, the Itakura–Saito divergence, and many more.

In this paper, we focus on the *Kullback-Leibler divergence*[4], historically the most popular type of divergence which has become ubiquitous in information theory. 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*); and the fact that it tensorizes nicely for products of probability distributions, expressed in terms of the sum of the individual divergences[5] (which give rise to useful chain rule properties).

A Pinsker-type inequality can be thought of as a general inequality of the form

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

[1] What is generally known as the "Hellinger distance" was in fact introduced by Jeffreys in 1946. The Hellinger integral (1909) 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).

[2] Some stronger types of convergence can also be metrized, but by distances between *random variables* rather than between distributions. For example, the Ky Fan distance metrizes the convergence in probability.

[3] Evidently, such divergences can always be symmetrized by considering $\big(D(p,q) + D(q,p)\big)/2$ instead of $D(p,q)$.

[4] Two fairly general classes of divergences are Rényi's f-divergences and the Bregman divergences. Some (square root of) f-divergences also yield genuine distances, like the Jeffreys (Hellinger) distance or the square root of the Jensen-Shannon divergence. It was recently shown that the Kullback-Leibler divergence is the only divergence that is both a f-divergence and a Bregman divergence [13].

[5] 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.

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. Any such Pinsker inequality implies that the topology induced by D is finer[6] than that induced by Δ. Many Pinsker-type inequalities have been established, notably between f-divergences.

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.

2 Preliminaries

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 $\mu = (p+q)/2$ as a dominating measure since $p \ll \mu$ and $q \ll \mu$. By the the Radon-Nikodym theorem, they admit *densities* with respect to μ, which we again denote by p and q, respectively. Thus for any event[7] $A \in \mathcal{A}$, $p(A) = \int_A p \,\mathrm{d}\mu = \int_A p(x) \,\mathrm{d}\mu(x)$, and similarly for q. 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 $\int_A p \,\mathrm{d}\mu = \sum_{x \in A} p(x)$; if μ is a Lebesgue measure, then p is a continuous probability distribution with $\int_A p \,\mathrm{d}\mu = \int_A p(x) \,\mathrm{d}x$. 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 $1 - p$. This ambiguity in notation should be easily resolved from the context.

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

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

that is, half the $L^1(\mu)$-norm of the difference of densities. It is important to note that this definition 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 $\int |p' - q'| \,\mathrm{d}\mu' = \int |p - q| \,\mathrm{d}\mu$.

[6] If, in addition, a *reverse* Pinsker inequality $\Delta \geq \psi(D)$ holds, then the associated topologies are equivalent.

[7] This is an overload in notations and one should not confuse $p(\{x\})$ with $p(x)$.

That Δ is a distance (metric) is obvious from this definition. Since $\int(p-q)\,\mathrm{d}\mu = 0$, we can also write $\Delta(p,q) = \int(p-q)^+\,\mathrm{d}\mu = \int(p-q)^-\,\mathrm{d}\mu$ (positive and negative parts) or $\Delta(p,q) = \int p\vee q\ \mathrm{d}\mu - 1 = 1 - \int p\wedge q\ \mathrm{d}\mu$ in terms of the maximum and minimum. The normalization factor $1/2$ ensures that $0 \le \Delta(p,q) \le 1$, with maximum value $\Delta(p,q) = 1 - \int p\wedge q\ \mathrm{d}\mu = 1$ if and only if $p\wedge q = 0$ μ-a.e., that is, p and q have "non-overlapping" supports. Note that the total variation distance between *binary* distributions $\mathcal{B}(p)$ and $\mathcal{B}(q)$ is simply

$$\delta(p,q) = |p-q|. \tag{3}$$

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

$$\Delta(p,q) \triangleq \frac{1}{2}\sup\sum_i |p(A_i) - q(A_i)|, \tag{4}$$

where the supremum is taken all *partitions* of Ω into a countable number of (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 by Jordan in the 19th century, and justifies the name "total variation" given to Δ.

That the two definitions (2) and (4) coincide can easily be seen as follows. First, by the triangular inequality, the sum $\sum_i |p(A_i) - q(A_i)|$ in (4) can only increase by subpartitioning, hence (4) 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 $A^- = \{p \le q\}$. Then the corresponding sum already equals $\sum_i (p-q)(A_i^+) + (q-p)(A_i^-) = (p-q)(\sum_i A_i^+) + (q-p)(\sum_i A_i^-) = (p-q)(A^+) + (q-p)(A^-) = \int(p-q)^+ + (p-q)^-\,\mathrm{d}\mu = \int|p-q|\,\mathrm{d}\mu$.

As a side result, the supremum in (4) is attained for binary partitions $\{A^+, A^-\}$ of the form $\{A, A^\complement\}$, so that $\Delta(p,q) = \frac{1}{2}\sup\big(|p(A) - q(A)| + |p(A^\complement) - q(A^\complement)|\big)$, that is,

$$\Delta(p,q) = \sup_A |p(A) - q(A)| \tag{5}$$

(without the $1/2$ factor). 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. In fact, given some observation X following either p (null hypothesis H_0) or q (alternate hypothesis H_1), such a statistical test takes the form « is $X \in A$? » (then accept H_0, otherwise reject it). Then since $|p(X \in A) - q(X \in A)| \le \Delta$ is small, type-I or type-II errors have total probability $p(X \notin A) + q(X \in A) \ge 1 - \Delta$. 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.

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

$$D(p\|q) \triangleq \int p \log \frac{p}{q} \, \mathrm{d}\mu, \tag{6}$$

where since $x \log x \geq -(\log e)/e$, the negative part of the integral is finite[8]. Therefore, this integral is always meaningful and can be finite, or infinite $= +\infty$. Again note that this definition 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 $\int p' \log \frac{p'}{q'} \, \mathrm{d}\mu' = \int p \log \frac{p}{q} \, \mathrm{d}\mu$.

By Jensen's inequality applied to the convex function $x \log x$, $D(p\|q)$ is non-negative and vanishes if only if the two distributions p and q coincide. For products of distributions $p = \bigotimes_i p_i$, $q = \bigotimes_i q_i$, it is easy to establish the useful tensorization property $D(p\|q) = \sum_i D(p_i\|q_i)$. The divergence between binary distributions $\mathcal{B}(p)$ and $\mathcal{B}(q)$ is simply

$$d(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}. \tag{7}$$

The double bar notation '$\|$' (instead of a comma) is universally used but may look exotic. Kullback and Leibler did not originate this notation in their seminal paper [19]. They rather used $I(1:2)$ for alternatives p_1, p_2 with a semi colon to indicate non commutativity. Later the notation $I(P \,|\, Q)$ was used but this collides 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 [27]. This notation was soon adopted by researchers of the Hungarian school of information theory, notably Csiszár (see, e.g., [5,6,7]).

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

$$D(p\|q) \triangleq \sup \sum_i p(A_i) \log \frac{p(A_i)}{q(A_i)} \tag{8}$$

where the supremum is again taken all *partitions* of Ω into a countable number of (disjoint) $A_i \in \mathcal{A}$. By the *log-sum inequality*, the sum $\sum_i p(A_i) \log \frac{p(A_i)}{q(A_i)}$ in (8) can only increase by subpartitioning, hence (8) 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 (this is the content of Dobrushin's theorem [24, § 2]). That the two definitions (6) and (8) coincide (in particular when (8) is finite, which implies $p \ll q$) is the content of a theorem by Gel'fand & Yaglom [10] and Perez [23].

[8] The logarithm (log) is considered throughout this paper in *any* base.

Statistical Distance and Mutual Information How does some observation Y affect the probability distribution of some 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:

$$\Delta(X;Y) = \mathbb{E}_y\, \Delta(p_{X|y}, p_X) = \Delta(p_{XY}, p_X \otimes p_Y), \tag{9}$$

and using the statistical divergence, one obtains the celebrated *mutual information*[9]:

$$I(X;Y) = \mathbb{E}_y\, D(p_{X|y} \| p_X) = D(p_{XY}, \| p_X \otimes p_Y) \tag{10}$$

introduced by Fano [8], based on Shannon's works. 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 $I = I(X;Y)$:

$$I \geq \varphi(\Delta) \tag{11}$$

for any two random variables X and Y, with the same φ as in (1). In particular, in terms of sequences of random variables, $I(X_n;Y_n) \to 0$ implies $\Delta(X_n;Y_n) \to 0$, a fact first proved by Pinsker [24, §2.3].

Binary Reduction of Pinsker's Inequality A straightforward observation, that greatly simplifies the derivation of Pinsker inequalities, follows from the alternative definitions (4) and (8). We have seen that the supremum in (4) is attained for binary partitions of the form $\{A, A^{\complement}\}$. On the other hand, the supremum in (8) is obviously greater then that for such binary partitions. Therefore, any Pinsker inequality (1) is equivalent to the inequality expressed in term of binary distributions (3), (7):

$$d \geq \varphi(\delta) \tag{12}$$

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

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \varphi(|p-q|) \tag{13}$$

is equivalent to the general case, but is naturally easier to prove. This binary reduction principle was first used by Csiszár [6] but as a consequence of a more general *data processing inequality* for any transition probability kernel (whose full generality is not needed here).

[9] 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).

Comparison of Pinsker Inequalities The following is sometimes useful to compare two different Pinsker inequalities (1) 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, $\varphi_1' \geq \varphi_2'$ implies that $D \geq \varphi_1(\Delta) \geq \varphi_2(\Delta)$. This comparison principle can be stated as follows: *lower derivative φ' implies weaker Pinsker inequality.*

3 Pinsker and Other Authors in the 1960s

It is generally said that Pinsker, in his 1960 book [24], proved the classical Pinsker inequality in the form

$$D \geq c \cdot \log e \cdot \Delta^2 \tag{14}$$

with a suboptimal constant c, and that the optimal (maximal) constant $c = 2$ was later found independently by Kullback [20], Csiszár [6] and Kemperman [16], 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 vs. statistical distance with $p = p_{X,Y}$ and $q = p_X \otimes p_Y$—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[10] $\Delta \leq \int p|\log \frac{p}{q}| \, \mathrm{d}\mu \leq D + 10\sqrt{D}$ with a quite involved proof for the second inequality[11] [24, pp. 14–15]. As noticed by Verdú [34], since one can always assume $\Delta \leq D + 10\sqrt{D} \leq 1$ (otherwise the inequality is vacuous), then two Pinsker inequalities imply $\Delta^2 \leq (D + 10\sqrt{D})^2 = D(D + 20\sqrt{D}) + 100D \leq 102D$ which indeed gives (14) with the suboptimal constant $c = \frac{1}{102}$. But this was nowhere mentioned in Pinsker's book [24].

The *first explicit occurrence of a Pinsker inequality of the general form* (1) occurs even *before* the publication of Pinsker's book, by Volkonskii and Rozanov [35, Eq. (V)] in 1959. They gave a simple proof of the following inequality:

$$D \geq 2 \log e \cdot \Delta - \log(1 + 2\Delta). \tag{15}$$

It is easily checked, from the second-order Taylor expansion of $\varphi(x) = 2 \log e \cdot x - \log(1 + 2x)$, that this inequality is strictly weaker than the classical Pinsker inequality (14) with the optimal constant $c = 2$, although both are asymptotically optimal near $D = \Delta = 0$.

The *first explicit occurrence of a Pinsker inequality of the classical form* (14) appeared as an exercise in Sakaguchi's 1964 book [28, pp. 32–33]. He proved $D \geq H^2 \log e \geq \Delta^2 \log e$ where H is the Hellinger distance, which gives (14) with the suboptimal constant $c = 1$. Unfortunately, Sakaguchi's book remained unpublished.

[10] In nats (natural units), that is, when the logarithm is taken to base e.

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

The *first published occurence of a Pinsker inequality of the classical form* (14) was by McKean [22, § 9a)] in 1966, who was motivated by a problem in physics related to Boltzmann's H-theorem. He proved (14) with the suboptimal constant $c = \frac{1}{e}$ (worse than Sakaguchi's) under the (unnecessary) assumption that q is Gaussian.

The *first mention of the classical Pinsker inequality* (14) *with the optimal constant* $c = 2$ was by Csiszár [5], in a 1966 manuscript received just one month after McKean's. In his 1966 paper, however, Csiszár only proved (14) with the suboptimal constant $c = \frac{1}{4}$ [5, 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 (14) (which as we have seen was only implicit).

The *first published proof of the classical Pinsker inequality* (14) *with the optimal constant* $c = 2$ was again by Csiszár one year later [6, Thm. 4.1] using binary reduction. His proof can be written as a one-line proof as follows:

$$d(p\|q) = \underbrace{d(p\|p)}_{=0} + \int_p^q \frac{\partial d(p\|r)}{\partial r}\,\mathrm{d}r = \int_p^q \frac{r-p}{r(1-r)}\,\mathrm{d}r \geq 4\int_p^q (r-p)\,\mathrm{d}r = 2(p-q)^2, \tag{16}$$

where we used natural logarithms and the inequality $r(1-r) \leq \frac{1}{4}$ for $r \in [0,1]$. That $c = 2$ is not improvable follows from the expansion $d(p\|q) = 2(p-q)^2 + o((p-q)^2)$, which also shows that this inequality (like the Volkonskii-Rozanov inequality (15)) is asymptotically optimal near $D = \Delta = 0$.

In a note added in proof, however, Csiszár mentions an earlier independent derivation of Kullback, published in the same year 1967 in [20], 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 [33] in 1970, the constant $\frac{4}{3}$ is wrong and should be corrected as $\frac{4}{9}$ [21] (see explanation in the next section).

Finally, in an 1968 Canadian symposium presentation [15]—later published as a journal paper [16] in 1969, Kemperman, apparently unaware of the 1967 papers by Csiszár and Kullback, again derived the classical Pinsker inequality with optimal constant $c = 2$. His ad-hoc proof (repeated in the renowned textbook [32]) is based in the inequality $\frac{4+2x}{3}(x\log x - x + 1) \geq (x-1)^2$, which is much less satisfying than the one-line proof (16)

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

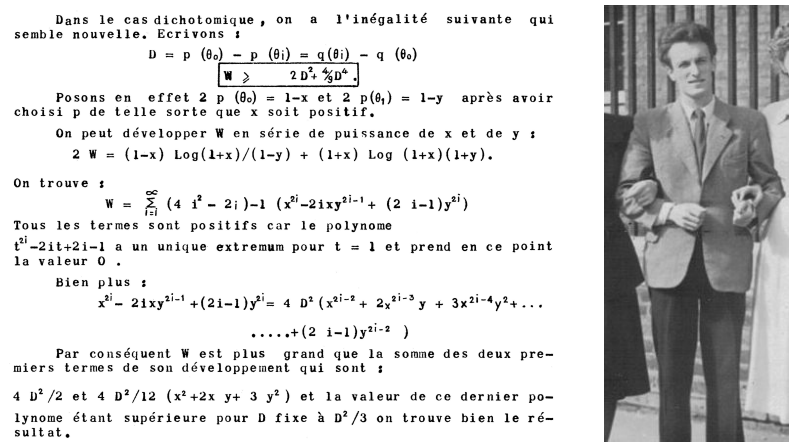
4 Schützenberger's Contribution (1953)

Seven years before the publication of Pinsker's book, the French medical doctor, geneticist, epidemiologist, and mathematician Marcel-Paul (Marco) Schützen-

berger, in his 1953 doctoral thesis [29] (see Fig 3), proved:

$$D \geq 2 \log e \cdot \Delta^2 + \frac{4}{9} \log e \cdot \Delta^4 \tag{17}$$

Not only does this contain the classical Pinsker inequality (14) with the optimal constant $c = 2$, but also the second-order improvement, with the (correct) optimal constant $\frac{4}{9}$ for the second-order term, seventeen years before Kullback! Admittedly, Schützenberger only considered the binary case, but due to the binary reduction principle, this does not entail any loss of generality.



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

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

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

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 :

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

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

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. 1: Left: Pinsker before Pinsker: In Schützenberger's notation, 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 $\frac{D^2}{3}$ instead of the correct $\frac{4D^2}{3}$. Right: Marcel-Paul (Marco) Schützenberger at his first marriage, in London, Aug. 30th, 1948.

In fact, leaving aside the use of binary reduction, Kullback's derivation [20] is just a mention of Schützenberger's inequality with the wrong constant $\frac{4}{3}$ instead of $\frac{4}{9}$. However, Vajda [33] asserts that the wrong constant comes from Schützenberger's manuscript itself, and that it was corrected in 1969 by Krafft [17]. In fact, Krafft does not refer to Schützenberger's thesis but rather to a 1966 paper by Kambo and Kotz [14] which contains a verbatim copy of Schützenberger's derivation (with the wrong constant and without citing the initial reference). While the correct constant $\frac{4}{9}$ does appear in the publicly available manuscript of Schützenberger (Fig. 3), it is apparent from the zooming in of Fig. 2 that the denominator was in fact carefully corrected by hand from a "3" to a "9". It is likely that the correction in Schützenberger's manuscript was made after 1970, when the error was discovered.

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

(b) Digits from the same manuscript.

Fig. 2: Schützenberger's correction from "3" to "9": the correction clearly follows the shape of a "3" in the original manuscript.

Nevertheless, Schützenberger's derivation is correct and gives the best constants 2 and 4/9 in (17) as an easy consequence of his identity

$$d=\sum_{k\geq 1}\frac{x^{2k}-2kxy^{2k-1}+(2k-1)y^{2k}}{2k(2k-1)}=2\delta^2\sum_{k\geq 1}\frac{x^{2k-2}+2x^{2k-3}y+\cdots+(2k-1)y^{2k-2}}{k(2k-1)} \tag{18}$$

where $x=1-2p$ and $y=1-2q$ (see Fig. 3). In 1969, Krafft and Schmitz [18] 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 [31]. But, in fact, the constant $\frac{2}{9}$ is not optimal; the optimal constant $\frac{32}{135}$ was found in 2001 by Topsøe [30]. Topsøe also derived the optimal constant for the additional term $\frac{7072}{42525}\log e\cdot\Delta^8$, whose proof is given in [9]. It is quite remarkable that all of such derivations are crucially based on the original Schützenberger's identity (18).

5 More Recent Improvements (1970s to 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 Pinsker inequality (14) with optimal constant $c=2$ become vacuous as soon as $D>2\log e$ (since $\Delta\leq 1$). Any improved Pinsker inequality of the form (1) should be such that $\varphi(1)=+\infty$ because $\Delta(p,q)=1$ (non overlapping supports) implies $D(p\|q)=+\infty$.

The first Pinsker inequality of this kind is due to Vajda in his 1970 paper [33]. He explicitly stated the problem of finding the optimal Pinsker inequality and proved

$$D\geq\log\frac{1+\Delta}{1-\Delta}-2\log e\cdot\frac{\Delta}{1+\Delta}. \tag{19}$$

where the lower bound becomes infinite as Δ approaches 1, as it should. This inequality is asymptotically optimal near $D=\Delta=0$ since $\log\frac{1+\Delta}{1-\Delta}-2\log e\cdot\frac{\Delta}{1+\Delta}=2\log e\cdot\Delta^2+o(\Delta^2)$.

In a 1978 French seminar, Bretagnolle and Huber [3,4] derived yet another Pinsker inequality similar to Vajda's (where the lower bound becomes infinite for $\Delta=1$) but with a simpler expression:

$$D\geq\log\frac{1}{1-\Delta^2}. \tag{20}$$

By the comparison principle, for natural logarithms and $0 < \Delta < 1$, $\frac{d}{d\Delta} \log \frac{1}{1-\Delta^2} = \frac{2\Delta}{1-\Delta^2} < \frac{4\Delta}{(1+\Delta)(1-\Delta^2)} = \frac{d}{d\Delta}\big(\log \frac{1+\Delta}{1-\Delta} - \frac{2\Delta}{1+\Delta}\big)$ always, since $1 + \Delta < 2$. Therefore, the Bretagnolle-Huber inequality (20) is strictly *weaker* than Vajda's inequality (19). 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 $2 \log e \cdot \Delta^2$. However, a nice property of the Bretagnolle-Huber inequality is that it can be inverted in closed form. In fact the authors expressed it as [12] $\Delta \le \sqrt{1 - \exp(-D)}$.

The Bretagnolle-Huber inequality was popularized by Tsybakov in his 2009 book on nonparametric estimation [32, Eq. (2.25)], but with a different form $\Delta \le 1 - \frac{1}{2}\exp(-D)$, or $D \ge \log \frac{1}{2(1-\Delta)}$, which is strictly *weaker* than the original, since $1 - \Delta^2 = (1 - \Delta)(1 + \Delta) < 2(1 - \Delta)$ for $0 < \Delta < 1$.

Today and to my knowledge, the best known explicit Pinsker inequality of this kind is

$$D \ge \log \frac{1}{1-\Delta} - (1 - \Delta)\log(1 + \Delta). \tag{21}$$

derived by Gilardoni in 2008 [11] (see also [12]). 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 $\delta \le p \le 1$ and $0 \le q \le 1 - \delta$. Then $d(p\|q) = (q+\delta)\log \frac{q+\delta}{q} + (1-q-\delta)\log \frac{1-q-\delta}{1-q} = \big[-q \log \frac{q+\delta}{q} - (1-q-\delta)\log \frac{1-q}{1-q-\delta}\big] + (2q + \delta)\log \frac{q+\delta}{q}$. Since $q + (1-q-\delta) = 1-\delta$ and $-\log$ is convex, the first term inside brackets is $\ge -(1 - \delta)\log(\frac{q+\delta}{1-\delta} + \frac{1-q}{1-\delta}) = (1 - \delta)\log \frac{1-\delta}{1+\delta}$. 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 $x \ge 0$ 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 $\ge (2-2\delta+\delta)\log \frac{1}{1-\delta} = (2-\delta)\log \frac{1}{1-\delta}$. Summing the two lower bounds gives the inequality. □

Note that Gilardoni's inequality adds the term $\Delta \log(1+\Delta)$ to the Bretagnolle-Huber lower bound. In fact it uniformly improves Vajda's inequality [11]. In particular, it is also asymptotically optimal near $D = \Delta = 0$, which can easily be checked directly: $\log \frac{1}{1-\Delta} - (1 - \Delta)\log(1 + \Delta) = 2 \log e \cdot \Delta^2 + o(\Delta^2)$. Also by the comparison principle, 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 $\Delta \ge 3\Delta^2$, i.e., $\Delta \le \frac{1}{3}$. Therefore, Gilardoni's inequality (21) is strictly weaker than the classical 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 better (see below).

[12] Here the exponential is relative to the base considered, e.g., $\Delta \le \sqrt{1 - e^{-D}}$ when D is expressed in nats (with natural logarithms) and $\Delta \le \sqrt{1 - 2^{-D}}$ when D is expressed in bits (with logarithms to base 2).

6 The Optimal Pinsker Inequality

The problem of finding the *optimal* Pinsker inequality (best possible lower bound in (1)) was opened by Vajda [33] in 1970. It was found in 2003 in *implicit* form, using the Legendre–Fenchel transformation, by Fedotov, Harremoës, and Topsøe in [9], as a curve parametrized by hyperbolic trigonometric functions. We give the following equivalent but simpler parametrization with the following proof that is arguably simpler as it only relies of the well-known Lagrange multiplier method.

Theorem 1 (Optimal Pinsker Inequality). *The optimal Pinsker inequality* $D \geq \varphi^*(\Delta)$ *is given in parametric form as*

$$\begin{cases} \Delta &= \lambda(1-q)q \\ D &= \log(1-\lambda q) + \lambda q(1+\lambda(1-q))\log e \end{cases} \tag{22}$$

where $\lambda \geq 0$ *is the parameter and* $q = q(\lambda) \triangleq \frac{1}{\lambda} - \frac{1}{e^\lambda - 1} \in [0, \frac{1}{2}]$..

Proof. Using binary reduction, $d(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ 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 *nats* (natural logarithms), this gives

$$\begin{cases} \frac{\partial L}{\partial p} = \log\frac{p}{q} - \log\frac{1-p}{1-q} - \lambda = 0 \\ \frac{\partial L}{\partial q} = -\frac{p}{q} + \frac{1-p}{1-q} + \lambda = 0 \end{cases} \quad \text{or} \quad \begin{cases} \lambda &= \frac{p}{q} - \frac{1-p}{1-q} \\ e^\lambda &= \frac{p}{q}\Big/\frac{1-p}{1-q} \end{cases} . \tag{23}$$

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 $d(p\|q) = \log\frac{1-p}{1-q} + p\lambda = \log(1-\lambda q) + \lambda q(1+\lambda(1-q))$. 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 $p \geq q \iff \lambda \geq 0 \iff q \in [0, 1/2]$. □

In 2009, Reid and Williamson [25,26], 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* Pinsker inequality: $D \geq \min_{|\beta|\leq 1-\Delta} \frac{1+\Delta-\beta}{2}\log\frac{1+\Delta-\beta}{1-\Delta-\beta} + \frac{1-\Delta+\beta}{2}\log\frac{1-\Delta+\beta}{1+\Delta-\beta}$. This formula,

however, is just a tautological definition of the optimal Pinsker lower bound: Indeed, by binary reduction, $d(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ 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 minimization boils down to the above expression for the lower bound.

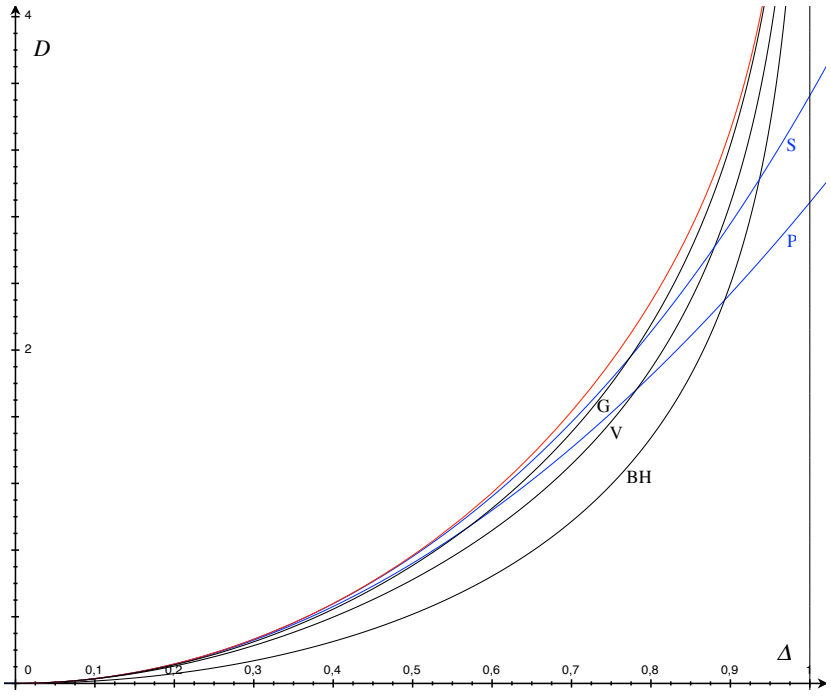


Fig. 3: Pinsker lower bounds of divergence D vs. total variation Δ. Red: Optimal (Theorem 1). Blue: Pinsker (P, Eq. 14 with $c = 2$) with optimal constant and Schützenberger (S, Eq. 17). Black: Bretagnolle-Huber (BH, Eq. 20), Vajda (V, Eq. 19) and Gilardoni (G, Eq. 21).

7 Conclusion

Fig. 3 illustrates the main Pinsker inequalities seen in this paper. As a temporary conclusion, from the implicit form using the exact parametrization of Theorem 1, it is likely that the optimal Pinsker inequality cannot be written as a *closed-form* expression with standard operations and functions. Also, the problem of finding an explicit Pinsker inequality which *uniformly* improve all the preceding ones (in

particular, the classical Pinsker inequality with optimal constant and Gilardoni's inequality) is still open.

Interestingly, 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 1;

- 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 (in nats) $D = \frac{\lambda^2}{8} + o(\lambda^2)$. Thus, one recovers that $D \sim 2\Delta^2$ near $D = \Delta = 0$. In particular, the classical Pinsker inequality (with optimal constant) and its improvements, 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 perspective, one may envision that the exact parametrization of Theorem 1 can be exploited to find new explicit bounds. Indeed, since $\lambda = \varphi^{*\prime}(\Delta)$ in the parametrization of Theorem 1, from the comparison principle, any inequality of the form $\varphi'(\Delta) \le \lambda = \varphi^{*\prime}(\Delta)$ is equivalent to a corresponding 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 Pinsker inequality (14) with optimal constant $c = 2$. Thus, the search of new 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.

References

1. Amari, S.: Information Geometry and Its Applications, Applied Mathematical Sciences, vol. 194. Springer (2016)
2. Barron, A.R.: Entropy and the central limit theorem. The Annals of Probability **14**(1), 336–342 (1986)
3. Bretagnolle, J., Huber, C.: Estimation des densités : risque minimax. In: Séminaire de probabilités (Strasbourg 1976/77). vol. 12, pp. 342–363. Springer (1978)
4. Bretagnolle, J., Huber, C.: Estimation des densités : risque minimax. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **47**, 119–137 (1979)
5. Csiszár, I.: A note on Jensen's inequality. Studia Scientiarum Mathematicarum Hungarica **1**, 185–188 (1966)
6. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungarica **2**, 299–318 (1967)
7. Csiszár, I., Körner, J.: Information Theory. Coding Theorems for Discrete Memoryless Systems. Cambridge University Press, 2nd edn. (2011) (1st edn 1981)
8. Fano, R.M.: Class notes for course 6.574: Transmission of Information. MIT, Cambridge, MA (1952)
9. Fedotov, A.A., Harremoës, P., Topsøe, F.: Refinements of Pinsker's inequality. IEEE Transactions on Information Theory **49**(6), 1491–1498 (June 2003)

10. 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)
11. Gilardoni, G.L.: An improvement on Vajda's inequality. In: In and Out of Equilibrium 2, Progress in Probability, vol. 60, pp. 299–304. Birkhäuser (Nov 2008)
12. Gilardoni, G.L.: On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. IEEE Transactions on Information Theory **56**(11), 5377–5386 (Nov 2010)
13. Jiao, J., Courtade, T.A., No, A., Venkat, K., Weissman, T.: Information measures: The curious case of the binary alphabet. IEEE Transactions on Information Theory **60**(12), 7616–7626 (2014)
14. Kambo, N.S., Kotz, S.: On exponential bounds for binomial probabilities. Ann. Inst. Statist. Math. **18**, 277–287 (1966)
15. 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 (Apr 1968)
16. Kemperman, J.H.B.: On the optimum rate of transmitting information. The Annals of Mathematical Statistics **40**(6), 2156–2177 (Dec 1969)
17. Krafft, O.: A note on exponential bounds for binomial probabilities. Ann. Institut für Mathematische Statistik **21**, 219–220 (1969)
18. Krafft, O., Schmitz, N.: A note on Hoeffding's inequality. Journal of the American Statistical Association **64**(327), 907–912 (Sep 1969)
19. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1), 79–86 (1951)
20. Kullback, S.: A lower bound for discrimination information in terms of variation. IEEE Transactions on Information Theory **13**, 126–127 (Jan 1967)
21. Kullback, S.: Correction to "A lower bound for discrimination information in terms of variation". IEEE Transactions on Information Theory **16**, 652 (Sept 1970)
22. McKean, Jr., H.P.: Speed of approach to equilibrium for Kac's caricature of a Maxwellian gas. Arch. Rational Mech. Anal. **21**, 343–367 (1966)
23. 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)
24. Pinsker, M.S.: Информация и информационная устойчивость случайных величин и процессов (Information and Information Stability of Random Variables and Processes). Izv. Akad. Nauk (1960) English translation Holden-Day, San Francisco (1964)
25. Reid, M.D., Williamson, R.C.: Generalised Pinsker inequalities. In: 22nd Annual Conference on Learning Theory (COLT 2009). Montreal, Canada (June 18–21 2009)
26. Reid, M.D., Williamson, R.C.: Information, divergence and risk for binary experiments. Journal of Machine Learning Research **12**, 731–817 (2011)
27. Rényi, A.: Az információelmélet néhány alapvető kérdése (some basic questions in information theory). Magyar Tud. Akad. Mat. Fiz. Oszt. Közl **10**, 251–282 (In Hungarian.) (1960)
28. Sakaguchi, M.: Information Theory and Decision Making. unpublished, George Washington University, Whasington D.C. (June 1964)
29. 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 (1954) Thèse de doctorat (1953)

30. Topsøe, F.: Bounds for entropy and divergence for distributions over a two-element set. Journal of Inequalities in Pure and Applied Mathematics **2**(2, Art 25), 1–13 (2001)
31. Toussaint, G.T.: Sharper lower bounds for discrimination information in terms of variation. IEEE Transactions on Information Theory **21**(1), 99–100 (Jan 1975)
32. Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer Series in Statistics, Springer (2009)
33. Vajda, I.: Note on discrimination information and variation. IEEE Transactions on Information Theory **16**, 771–773 (Nov 1970)
34. Verdú, S.: Total variation distance and the distribution of relative information. In: 2014 Information Theory and Applications Workshop (ITA). San Diego, CA, USA (Feb 9–14 2014)
35. 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)