L'héritage de Marco Schützenberger en statistiques et théorie de l'information

- **van Trees avant van Trees**
- **Pinsker avant Pinsker**
- **. . . et le mystère de la constante 4/3**

EDMSA'24, Caen, 15 mai 2024

Olivier Rioul

Télécom Paris,
Institut Polytechnique de Paris,
France

<olivier.rioul@telecom-paris.fr>

“No scientific discovery is named after its original discoverer” (Stigler’s Law, 1980)

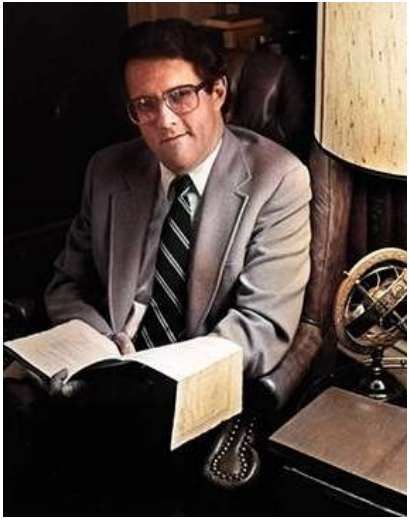
STIGLER'S LAW OF EPONYMY*

Stephen M. Stigler

Department of Statistics
University of Chicago
Chicago, Illinois 60637

No reader of Robert K. Merton's work on the reward system of science could fail to be struck by his insightful and engaging discussions of the role of eponymy in the social structure of science. The uninitiated should read (and reread) his 1957 address, "Priorities in Scientific Discovery,"[1] but for present purposes I must at least repeat his definition of eponymy, as "the practice of affixing the name of the scientist to all or part of what he has found, as with the Copernican system, Hooke's law, Planck's constant, or Halley's comet."[2] Merton went on to discuss three levels of a hierarchic order of eponymous practice: at the top there are a few men for whom an entire epoch is named, then comes a larger number of scientists designated as "father" of a particular science, and, finally, "thousands of eponymous laws, theories, theorems, hypotheses, instruments, constants, and distributions."[3] The present paper is an attempt by an Outsider to the sociology of science to shed some light on the workings of the eponymic reward system at this third level, and a report on a small statistical investigation into eponymous practices of my own field, statistics.

I have chosen as a title for this paper, and for the thesis I wish to present and discuss, "Stigler's law of eponymy." At first glance this may appear to be a flagrant violation of the "Institutional Norm of Humility,"[4] and since statisticians are even more aware of the importance of norms than are members of other disciplines, I hasten to add a humble disclaimer. If there is an idea in this paper that is not at least implicit in Merton's *The Sociology of Science*, it is either a happy accident or a likely error. Rather I have, in the Mertonian tradition of the self-confirming hypothesis, attempted to frame the self-proving theorem. For "Stigler's Law of Eponymy" in its simplest form is this: "No scientific discovery is named after its original discoverer."



“Every important inequality in statistics was discovered by Schützenberger 10 years before anyone else”

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

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

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

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 :

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

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 :

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

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

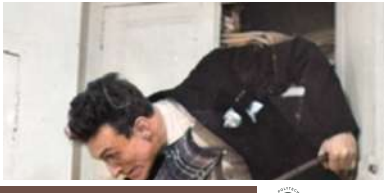
Par conséquent W est plus grand que la somme des deux premiers termes de son développement qui sont :

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

Marcel-Paul (Marco) Schützenberger

- Alsatian family (grand grand grand father was Strasbourg's mayor)

- grand grand father Paul: renowned chemist, founded the ESPCI, satirized in *Les Palmes de Mr. Schutz*



Marcel-Paul (Marco) Schützenberger

- after WWII, participates in Surrealist/Dadaist movements — appears in a short film with Boris Vian, and becomes the main character (Dr. Markus Schutz) in Boris Vian's novel *Et on tuera tous les affreux*
- member of the cabinet of Communist minister Charles Tillon, publishes articles in lattice theory and in physiology, while studying "ancient Mongolian"
- 1948 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

Marcel-Paul (Marco) Schützenberger

- applies statistical methods to the analysis of various medical problems, (e.g., discovery of trisomy 21)
- 1948, following a paper by psychologist Anne Ancelin based on his statistics, was offed a position in London; got married immediately (to be better paid) in London with Anne Ancelin (photo). Finally declined the position, the couple divorced in 1952.
- publishes papers on combinatorics in a genetics journal, on biostatitics with George Darmois, consultant to the World Health Organization
- 1952 & 1953, WHO sent him to Asia to combat infectious diseases of tropical countries. Met 2nd wife Hariati Soerosoegondo in Java.

Marcel-Paul (Marco) Schützenberger

- 1952, came to information theory from biostatistics:

APPLICATIONS BIOMÉTRIQUES DE LA THÉORIE DE L'INFORMATION

par M. P. SCHÜTZENBERGER

- published his mathematical thesis in 1953 (director: Darmois, president: Fréchet)

CONTRIBUTION

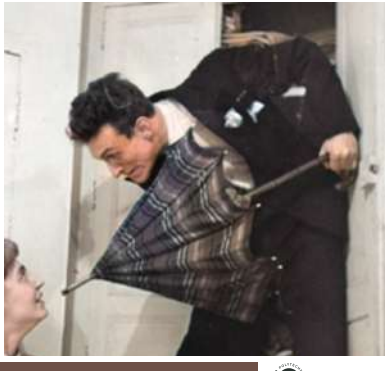
AUX

APPLICATIONS STATISTIQUES

DE LA

THÉORIE DE L'INFORMATION

par



Marcel-Paul (Marco) Schützenberger

- at the crossroads of algebra and theoretical computer science: variable-length codes, monoids, automata theory, etc.
- His seminal paper *Une théorie algébrique du codage* (1958) lays foundations of automata theory and its relationship with rational languages and semi-groups.
- mostly known for: Kleene-Schützenberger theorem (1961) in the theory of formal languages and automata, Chomsky-Schützenberger theorem (1963), a representation theorem of context-free languages
- 1970s-1980s scientific advisor for the WHO, to detect and prevent accidents due to the careless use of medicines or chemical/biological weapons

Marcel-Paul (Marco) Schützenberger

- his son Mahar, *major* of the École Polytechnique in 1976, killed in a car accident in 1980 at the age of 23

- 1988, elected to the French Academy of Sciences
- since 1991, Mahar Schützenberger Prize for Indonesians preparing their doctoral thesis in France
- *Triangle de pensées* reports the discussions with Alain Connes & André Lichnerowicz on relativity, quantum mechanics or Gödel's theorem, and the relations among mathematics, physics, philosophy...
- defended strong arguments against the Darwinian theory of evolution
- saddened by the death of his wife Nariati in 1993

Marcel-Paul (Marco) Schützenberger

Schützenberger's personality was complex and unorthodox, capable of glowing praises as well as ironic sarcasm, passionate for discussion, paradox and controversy.

Two emblematic inequalities in statistics and information theory

- Schützenberger-*van Trees inequality*, 1968~~1968~~ 1957 (Bayesian parametric estimation)

$$\mathrm{MMSE}(\hat{\theta}) \geq \left(\widetilde{J}_\theta + n \cdot \mathbb{E}_\theta(J_\theta)\right)^{-1}$$

- Schützenberger-*Pinsker inequality*, 1960~~1960~~ 1953 (statistical distance Δ vs. informational divergence D) improved by Kullback, 1967/1970~~1967/1970~~

$$D(p\|q) \geq 2\log e \cdot \Delta^2(p,q) + \frac{4}{9}\log e \cdot \Delta^4(p,q)$$

Part One: Schützenberger-van Trees inequality (Bayesian Cramér-Rao bound)

$$\mathrm{MMSE}(\hat{\theta}) \geq \left(\widetilde{J}_\theta + n \cdot \mathbb{E}_\theta(J_\theta)\right)^{-1}$$

Parametric Estimation

Observed data $\underline{x} = (x_1, x_2, \ldots, x_n)$ be a very long i.i.d. sequence $\sim p_{\theta^*} \ll \mu$.
Model $\theta \mapsto p_\theta$ is known:

$$p_\theta(\underline{x}) = p_\theta(x_1)p_\theta(x_2)\cdots p_\theta(x_n)$$

Find an asymptotically optimal estimator $\hat{\theta}(\underline{x})$ of θ^*.

- taking logarithms, asymptotically as $n \to +\infty$, by the law of large numbers,

$$\frac{1}{n}\log p_\theta(\underline{x}) = \frac{1}{n}\sum_1^n \log p_\theta(x_i) \to \mathbb{E}_{\theta^*} \log p_\theta(X) = -H(p_{\theta^*}\|p_\theta)$$

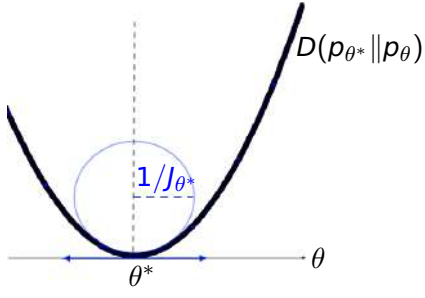
- divergence $D(p_{\theta^*}\|p_\theta) = H(p_{\theta^*}\|p_\theta) - H(p_{\theta^*}) \geq 0$ is minimum $= 0$ iff $p_\theta = p^*_\theta$. (i.e., by identifiability $\theta = \theta^*$)
- asymptotically as $n \to +\infty$,

$$\theta^* = \arg\min_\theta D(p_{\theta^*}\|p_\theta) \iff \hat{\theta}(\underline{x}) = \arg\max_\theta \frac{1}{n}\log p_\theta(\underline{x}) \text{ (maximum likelihood)}$$

Fisher's Information: Operational Definition

At the minimum $\theta = \theta^*$:

- null gradient $\frac{\partial}{\partial\theta} D(p_{\theta^*}\|p_\theta)\big|_{\theta=\theta^*} = -\mathbb{E}S_\theta(X) = 0$ where score $S_\theta(X) = \frac{\partial}{\partial\theta}\log p_\theta(X)$.
- curvature $J_{\theta^*} = \frac{\partial^2}{\partial\theta^2} D(p_{\theta^*}\|p_\theta)\big|_{\theta=\theta^*} \geq 0$ (Fisher information)

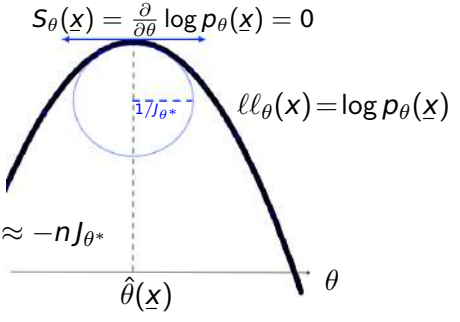


$$J_\theta = -\int \frac{\partial^2}{\partial\theta^2}\Big(\log p_\theta(x)\Big) p_\theta(x)\,\mathrm{d}\mu(x) = \underbrace{\int \Big(\frac{\partial}{\partial\theta}\log p_\theta(x)\Big)^2 p_\theta(x)\,\mathrm{d}\mu(x)}_{\mathrm{Var}(S_\theta(X))}$$

Fisher's Information: Operational Definition

Therefore, the maximum likelihood estimator $\hat{\theta}(\underline{x}) = \arg\max_{\theta} \log p_\theta(\underline{x})$ satisfies:

- asymptotically, $0 = \frac{1}{n} S_\theta(\underline{x}) \approx \mathbb{E}(S_\theta(X))$ at $\theta = \hat{\theta}$ hence $D(p_{\theta^*} \| p_{\hat{\theta}}) \to 0$ and $\boxed{\hat{\theta} \to \theta^*}$ as $n \to \infty$;
- asymptotically, $\frac{S_{\theta^*}(\underline{x}) - \overbrace{S_{\hat{\theta}}(\underline{x})}^{=0}}{\theta^* - \hat{\theta}} \to \frac{\partial^2}{\partial \theta^2} \log p_\theta(\underline{x})\big|_{\theta=\theta^*} \approx -nJ_{\theta^*}$ hence $\mathbb{E}(\hat{\theta} - \theta^*)^2 \sim \frac{nJ_{\theta^*}}{n^2 J_{\theta^*}^2} = \frac{1}{nJ_{\theta^*}}$, i.e.,



Theorem (Fisher's Estimation Theorem)

Asymptotically, ML estimation has MSE $\approx \frac{1}{nJ_{\theta^*}}$ *for observation X*

Converse theorem: Cramér-Rao bound

If $\hat{\theta}$ is unbiased, MSE $= \mathrm{Var}(\hat{\theta}) \geq \frac{1}{nJ_{\theta^*}}$

Fréchet-Darmois-Cramér-Rao bound (CRB~~CRB~~ FDB)

Theorem (CRBFDB)

In (frequentist) parametric estimation, for a regular statistical model and for any unbiased estimator,

$$\mathrm{MSE}(\hat{\theta}) \geq \frac{1}{nJ_{\theta^*}}$$

- Maurice Fréchet, 1943 "*Le contenu de ce mémoire a formé une partie de notre cours de statistique mathématique a l'Institut Henri Poincaré pendant l'hiver 1939-1940*" [président du jury de thèse de Schützenberger]
- Georges Darmois, 1945 "*Les résultats du mémoire de M. Fréchet peuvent être étendus à un nombre quelconque de paramètres*" [directeur de thèse de Schützenberger]
- C. Radhakrishna Rao, 1945
- Harald Cramér, 1946 (in his book *Mathematical Methods of Statistics*)

Extension to Bayesian Estimation: van Trees inequality

Theorem (van Trees: BCRB)

Lower Bound on the Minimum Mean-Square Error in Estimating a Random Parameter. In this section we prove the following theorem.

Theorem. Let a be a random variable and $\mathbf{r}$, the observation vector. The mean-square error of any estimate $\hat{a}(\mathbf{R})$ satisfies the inequality

$$E\{[\hat{a}(\mathbf{R}) - a]^2\} \geq \left(E\left\{\left[\frac{\partial \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A}\right]^2\right\}\right)^{-1} = \left\{-E\left[\frac{\partial^2 \ln p_{\mathbf{r},a}(\mathbf{R}, A)}{\partial A^2}\right]\right\}^{-1}. \quad (217)$$

In Bayesian parametric estimation, under some technical conditions,

Extension to Bayesian Estimation: Schützenberger-van Trees inequality

321*t*. M. P. Schützenberger: *A generalization of the Fréchet-Cramér inequality to the case of Bayes estimation.*

Let $f(x)$ be the a priori density function of x; $g(y|x)$ the conditional density function of y. For fixed x, the set of n independent y-variates is represented by s. The density function of s is $f'(s)$ and $g'(x|s)$ is the a posteriori density function of x, for given s. The a posteriori variance of the Bayes estimate is $v_s^2 = \int (x-\bar{x})^2 g'(x|s)dx$ and $v^2 = E_s v_s^2 = \int v_s^2 f'(s)ds$ is its average over s. $F = \int (\partial f(x)/\partial x)^2 (f(x))^{-1} dx$; $G = E_y G_x$ with $G_x = \int ((\partial/\partial x) g(y|x))^2 (g(y|x))^{-1} dy$; $G' = E_s G_s'$ with $G_s' = \int ((\partial/\partial x) g'(x|s))^2 (g(x|s))^{-1} dx$. The usual assumptions on f and g, which insure that F, G_x, G_s' are finite are made. Since $0 = F' = \int ((\partial/\partial x) f'(x))^2 (f'(x))^{-1} dx$, it is easily seen that $F + nG = G'$ (Third London Symposium on Information Theory, 1955, p. 18). Furthermore, it is a classical result that $v_s^2 G_s' \geq 1$. Thus $v^2 = E_s v_s^2 \geq (E_s 1/v_s^2)^{-1} \geq (E_s G_s')^{-1} = (F+nG)^{-1}$, which is the desired inequality that tends to the usual form when n goes to infinity. It reduces to an equality if and only if $v^2 = v_s^2 = (G_s')^{-1}$ for all s, that is, if and only if $g'(x|s)$ is gaussian with variance independent of s. If, furthermore, $y - x = z$ has a distribution $h(z)$ independent of x, this implies that $f(x)$ and $h(z)$ are also gaussian.

- Bulletin of the American Mathematical Society, 1957, 10 years before van Trees!
- Long version in French, 1958: "*À propos de l'inégalité de Fréchet-Cramér*"
- acknowledged by van Trees in 2007 in a footnote (with a misprint):

[12]The Bayesian Cramér–Rao bound was derived by Van Trees [Van68, Van01a]. It was derived independently by Shutzenberger [Shu57]. This latter derivation is a model of economy (1/3 of a page) but does not appear to have been noticed by either the engineering or statistical community.

Part Two: Schützenberger-Pinsker inequality

$$D(p\|q) \geq 2 \log e \cdot \Delta^2(p,q) + \frac{4}{9} \log e \cdot \Delta^4(p,q)$$

How far is one distribution to another?

Distances $\Delta(p, q)$

- Lévy-Prokhorov
- Fortet-Mourier
- Kantorovich-Rubinstein
  a.k.a. Wasserstein
- Radon
- “Hellinger” (Jeffreys)
- L^1, L^p
- (Ky Fan: between rv’s)
- ⋮
- Total variation $\Delta(p, q)$ (“statistical”)

Divergences $D(p\|q)$

- Rényi
- Bhattacharyya
- “Jensen-Shannon” (Lin)
- “Jeffreys” (sym. Kullback-Leibler)
- “Pearson” χ^2
- “Cauchy-Schwarz”
- Sundaresan
- Itakura–Saito
- ⋮
- Kullback-Leibler $D(p\|q)$ (“relative entropy”)

Definitions

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

- does *not* depend on the choice of the dominating measure $\mu \gg p, q$ (e.g., discrete/continuous cases);
- vanishes iff $p = q$;
  $\Delta(p,q) = 1 \iff p \wedge q = 0$ μ-a.e. (non-overlapping supports) $\implies D(p\|q) = +\infty$

$$\Delta = 0 \Longleftrightarrow D = 0$$

$$\Delta = 1 \Longrightarrow D = +\infty$$

- Binary case $P = (p, 1-p)$, $Q = (q, 1-q)$:

$$\boxed{\delta(p,q) = |p-q|} \qquad \boxed{d(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}}$$

Alternate definitions

Supremum taken all *partitions* into a countable number of $A_i \in \Omega$:

$$\Delta(p,q) \triangleq \frac{1}{2} \underbrace{\sup \sum_i |p(A_i) - q(A_i)|}_{\text{Jordan's total variation of } p-q} \qquad D(p\|q) \triangleq \sup \sum_i p(A_i) \log \frac{p(A_i)}{q(A_i)}$$

- enough to consider *intervals* A_i (when $\Omega = \mathbb{R}^d$): *Dobrushin's theorem* for D (1960)
- increases by subpartionning (by the triangular inequality for Δ, by the *log-sum inequality* for D). Hence $\sup = \lim$ for finer and finer partitions.
- $\Delta = \frac{1}{2} \int |p - q| \, \mathrm{d}\mu$ already for binary partition $\{p < q\}$, $\{p \geq q\}$.
  $D = \int p \log \frac{p}{q} \, \mathrm{d}\mu$ by *Gel'fand-Yaglom–Perez theorem* (1959)

Nice Properties

- Total variation distance: binary reduction property for partition A, A^c:

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

A sufficiently small value of $\Delta(p,q)$ implies that no statistical test can effectively distinguish between the two distributions p and q

- Kullback-Leibler divergence: tensorization property for product of probability measures:

$$D\Big(\bigotimes_i p_i \Big\| \bigotimes_i q_i\Big) = \sum_i D(p_i \| q_i)$$

Summary of Properties

$\Delta(p, q)$

- metric ✓
- binary reduction (bounded) ✓
- does not tensorize ✗

$D(p\|q)$

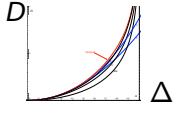
- not a metric ✗
- unbounded limit ✗
- nicely tensorizes ✓

$$\Delta = 0 \iff D = 0$$

$$\Delta = 1 \implies D = +\infty$$

Which topology is finer?

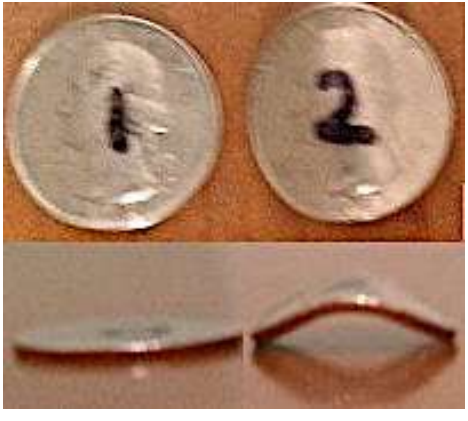
Pinsker's inequality: $\boxed{D \geq \varphi(\Delta)}$ where $\varphi(0) = 0$, φ increasing (convex)



By binary reduction, it is enough to prove it in the binary case: $\boxed{d \geq \varphi(\delta)}$

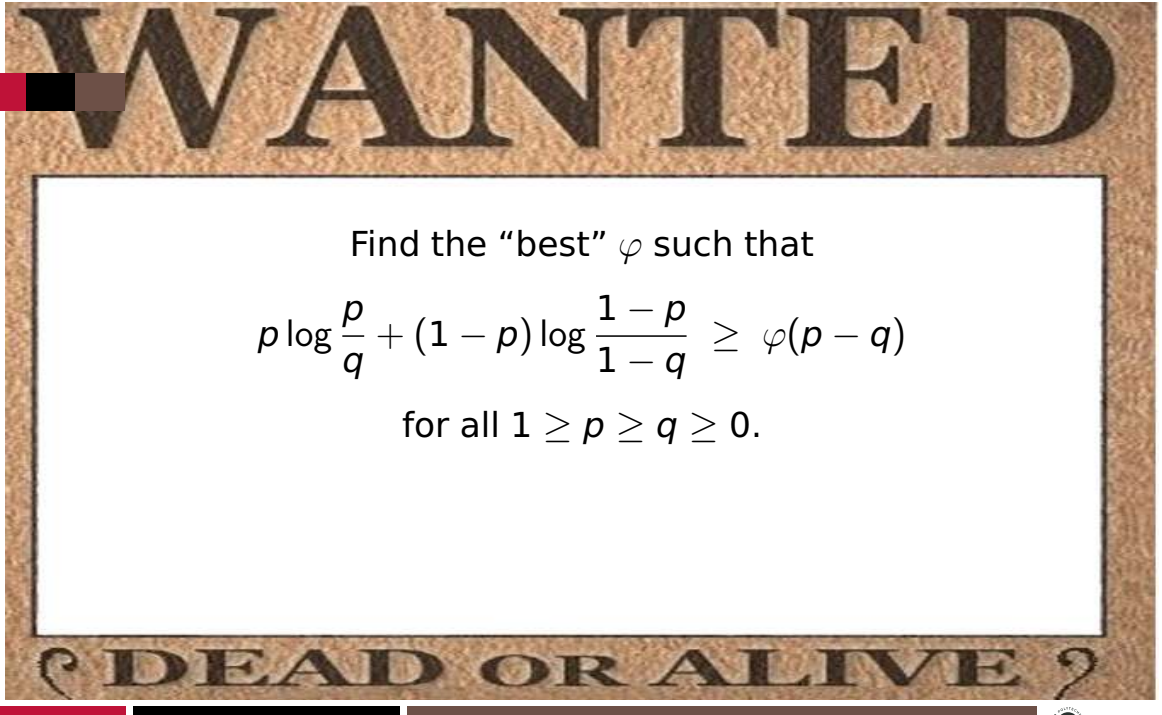
Why is PInsker's Inequality Useful?

Example: How to distinguish fair ($p = ½$) from unfair ($q \neq ½$) coin with n tosses



To be sure with probability $1 - \epsilon$, we need

$$\Delta(\otimes_{i=1}^{n} P_i, \otimes_{i=1}^{n} Q_i) \geq |(1 - \epsilon) - \epsilon| = 1 - 2\epsilon$$


WANTED
Find the “best” φ such that
$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \varphi(p-q)$$
for all $1 \geq p \geq q \geq 0$.
DEAD OR ALIVE ?

The classical “Pinsker” Inequality

$$\boxed{D \geq c \cdot \Delta^2}$$

where[1] the optimal (maximum) constant c is $\boxed{c = 2}$.

Pinsker’s book, 1960: Информация и информационная устойчивость случайных величин и процессов

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

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

- *Pinsker did not explicitly state Pinsker’s inequality* (not even in some other form $D \geq \varphi(\Delta)$)
- He proved instead two inequalities, which combined gives $\Delta \leq D + 10\sqrt{D}$

The classical “Pinsker” Inequality (cont’d)

- Volkonskii and Rozanov, 1959: $\boxed{D \geq 2\Delta - \log(1+2\Delta)}$ first explicit occurence, before the publication of Pinsker’s book!
- Sakaguchi’s book, 1964: $\boxed{D \geq c \cdot \Delta^2}$ with the suboptimal $c = 1$ first explicit occurence of the classical inequality... but remained unpublished!
- McKean, 1966: $\boxed{D \geq c \cdot \Delta^2}$ with the suboptimal $c = \frac{1}{e}$ first published occurence of the classical inequality!
- Csiszár, 1966: mentions $\boxed{D \geq c \cdot \Delta^2}$ with the optimal $c = 2$... but without proof! (he only proved $c = \frac{1}{4}$)
- Finally, Csiszár, 1967 proved $c = 2$; can be simplified as a 1-line proof:

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

The classical “Pinsker” Inequality, improved

- Kemperman, 1968: (independently?) re-derived $D \geq 2\Delta^2$ (with an ad-hoc proof).
- in a note added in proof, Csiszár mentions an earlier independent derivation of Kullback, published in the same year 1967, with an improved inequality

$$\boxed{D \geq 2\Delta^2 + \frac{4}{3}\Delta^4}$$

- but Vajda, 1970 noticed that the constant $\frac{4}{3}$ is wrong and should be replaced by the optimal constant $\frac{4}{9}$!

What happened?

- In fact, Kullback copied a earlier derivation in a 1953 French doctoral thesis by. . .

Marcel-Paul (Marco) Schützenberger's 1953 Thesis

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

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

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

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 :

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

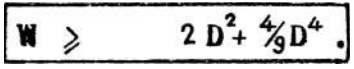
Tous les termes sont positifs car le polynome $t^{2i} - 2it + 2i - 1$ a un unique extremum pour $t = 1$ et prend en ce point la valeur 0 .

Bien plus :

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

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

Marcel-Paul (Marco) Schützenberger's 1953 Thesis



$$D \geq 2\Delta^2 + \frac{4}{9}\Delta^4$$

- optimal constants of Pinsker's inequality, first and second order
- in the binary case
- 7 years before Pinsker's book
- 14 years before Csiszár's first order term
- 17 years before Kullback's correction of the second-order term !

The $4/3$ vs. $4/9$ Mystery

$$D(p\|q) \geq 2 \log e \cdot \Delta^2(p,q) + \frac{4}{3} \log e \cdot \Delta^4(p,q)$$

The 4/3 vs. 4/9 Mystery

- 1953, the original French manuscript of Schützenberger was published with the correct constant 4/9 [Hélène S.]

$$W \geqslant 2D^2 + \tfrac{4}{9}D^4 .$$

- 1956, Kullback review of Schützenberger's thesis. "*The reader is cautioned to read out for misprints*"

For the binomial the author shows that

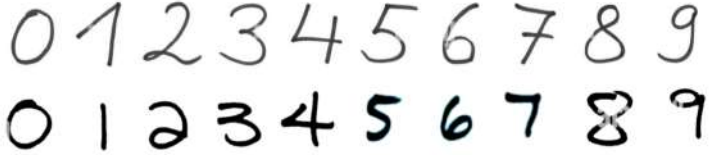
$$2(p_0-p_1)^2+(4/3)(p_0-p_1)^4 \leqq p_0 \log (p_0/p_1) + q_0 \log (q_0/q_1)$$

- 1966, Kambo and Kotz (*On exponential bounds for binomial probabilities*) copied Schützenberger's derivation verbatim, without citation and with the wrong 4/3 !

LEMMA 3. *Let $p>0$, $q>0$, $p+q=1$ and $c \geqq 0$, then*

The 4/3 vs. 4/9 Mystery

(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)

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

Schützenberger's identity

Schützenberger's derivation is correct and gives the 1st and 2nd order *optimal* constants based on his identity:

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

where $d = d(p,q)$, $x = 1-2p$ and $y = 1-2q$.

- 1969, Krafft & Schmitz used Schützenberger identity to derive 3rd-order constant $\frac{2}{9}$
- 1975, Toussaint converted it into a Pinsker inequality $D \geq 2\Delta^2 + \frac{4}{9}\Delta^4 + \frac{2}{9}\Delta^6 \ldots$ but this constant is not optimal !
- 2001, Topsøe used Schützenberger identity to derive 3rd-order *optimal* constant $\frac{32}{135}$
- 2003, Fedotov, Harremoës, Topsøe used Schützenberger's identity to derive the *4th-order* optimal constant $7072/42525$

$$\boxed{D \geq 2\Delta^2 + \frac{4}{9}\Delta^4 + \frac{32}{135}\Delta^6 + \frac{7072}{42525}\Delta^8 + \cdots}$$

More Recent Improvements on Schützenberger-Pinsker Inequality

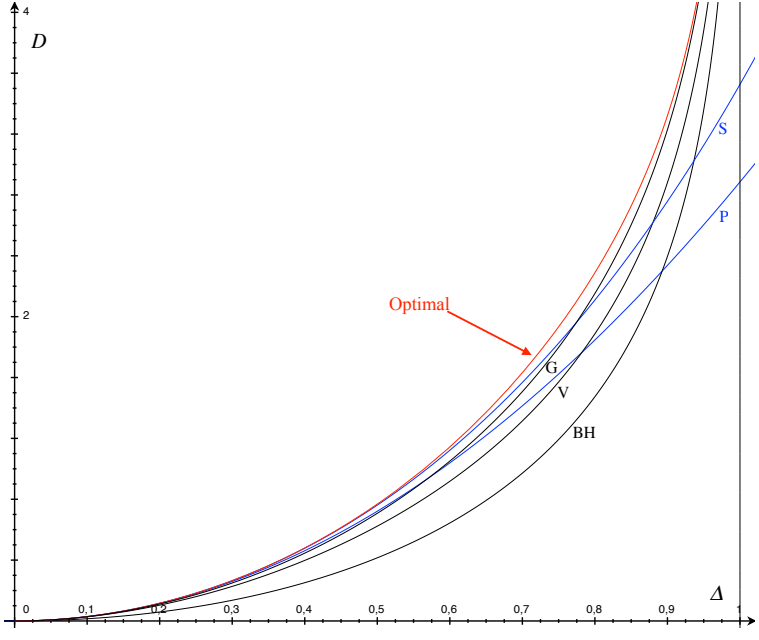
So far the inequalities become vacuous as soon as D gets large, e.g., $D \geq 2\Delta^2$ is vacuous as soon as $D > 2$: Improved bounds should reflect the fact that $\Delta = 1$ (non overlapping supports) imply $D = +\infty$

- Vajda, 1970: $D \geq \log \frac{1+\Delta}{1-\Delta} - 2 \log e \cdot \frac{\Delta}{1+\Delta}$
- Bretagnolle & Huber, 1978: weaker (but simpler) inequality $D \geq \log \frac{1}{1-\Delta^2}$ (can be explicitly reversed: $\Delta \leq \sqrt{1 - \exp(-D)}$)
- Tsybakov's classic book 2009: even weaker Bretagnolle-Huber inequality $D \geq \log \frac{1}{2(1-\Delta)}$ (or $\Delta \leq 1 - \frac{1}{2}\exp(-D)$)
- Gilardoni, 2008: best known explicit Pinsker inequality of this kind so far:

$$\boxed{D \geq \log \frac{1}{1-\Delta} - (1-\Delta)\log(1+\Delta)} \qquad = \text{BH bound} + \Delta \log(1+\Delta)$$

(simple proof in the GSI/Information Geometry paper)

Summary



Optimal Schützenberger-Pinsker Inequality

Derived by Fedotov, Harremoës, and Topsøe (2009), only in implicit form, using Legendre–Fenchel transformation as a curve parametrized by hyperbolic trigonometric functions. — Simplified expression:

Theorem

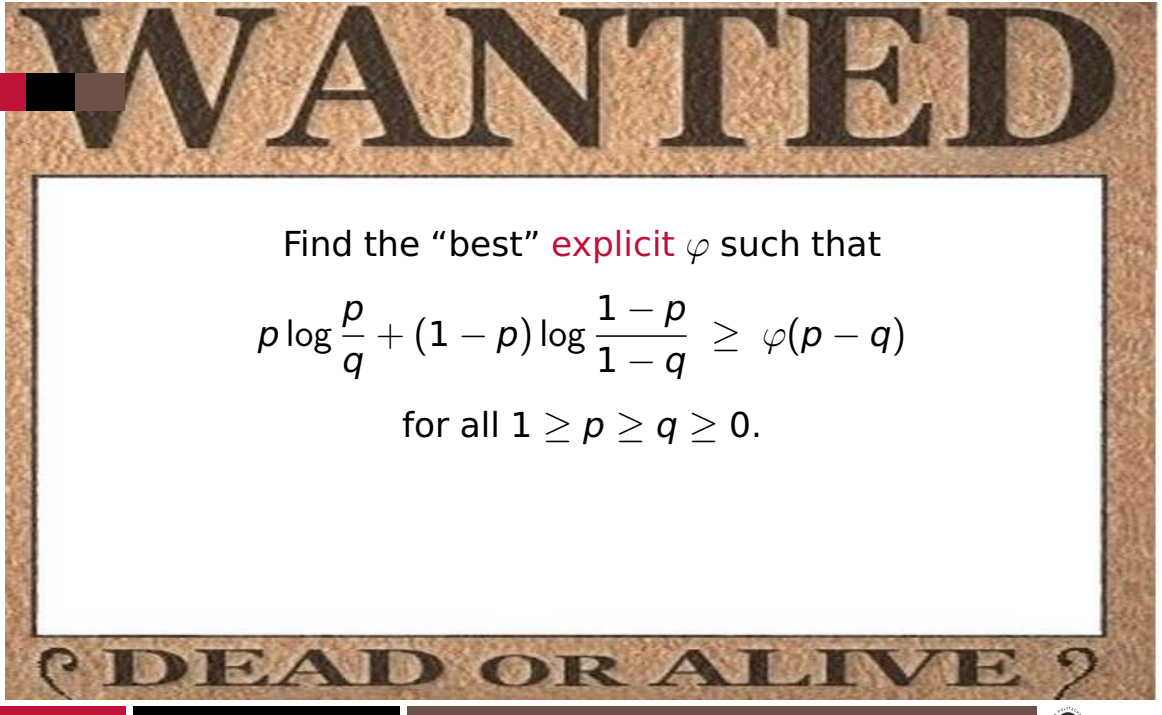
The optimal Pinsker inequality $D \geq \varphi^*(\Delta)$ is given in parametric form as

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

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.

(simple proof in the GSI/Information Geometry paper using Lagrangian) □


WANTED
Find the “best” explicit φ such that
$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \varphi(p-q)$$
for all $1 \geq p \geq q \geq 0$.
DEAD OR ALIVE ?

References

- "A historical perspective on Schützenberger-Pinsker inequalities," in Proc. 6th International Conference on Geometric Science of Information (GSI 2023), Saint Malo, France, Aug. 30-Sept. 1, 2023. Proceedings, Part I, Lecture Notes in Computer Science, Vol. 14071, pp. 291-306, Springer, 2023.
- "A historical perspective on Schützenberger-Pinsker inequalities **(Extended Version)**," pp.1–44, *Information Geometry*, to appear, 2024. *Final version available*
- "A historical perspective on the Schützenberger-van Trees Bayesian Fréchet-Darmois-Cramér-Rao bound," in preparation with Alexandre Renaux, 2025?

L'héritage de Marco Schützenberger en statistiques et théorie de l'information

- van Trees avant van Trees
- Pinsker avant Pinsker
- . . . et le mystère de la constante 4/3

Thank you!

Olivier Rioul

Télécom Paris,
Institut Polytechnique de Paris,
France

<olivier.rioul@telecom-paris.fr>