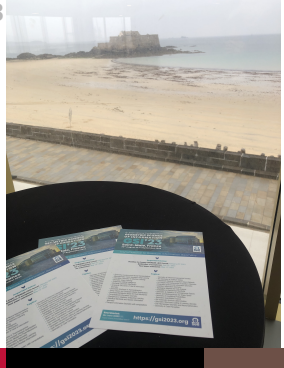
A Historical Perspective on Schützenberger-Pinsker Inequalities

GSI'23, Saint-Malo, August 30th, 2023

Olivier Rioul

Télécom Paris,
Institut Polytechnique de Paris,
France

<olivier.rioul@telecom-paris.fr>



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)
- ⋮

Divergences $D(p\|q)$

- Rényi
- Bhattacharyya
- “Jensen-Shannon” (Lin)
- “Jeffreys” (sym. Kullback-Leibler)
- “Pearson” χ^2
- “Cauchy-Schwarz”
- Sundaresan
- Itakura–Saito
- ⋮

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);

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) $\implies D(p\|q) = +\infty$

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

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

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 $A_i \in \Omega$:

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

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)

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.

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

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 ✓

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

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

Summary of Properties

$\Delta(p, q)$

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

$D(p\|q)$

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

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

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

Which topology is finer?

Summary of Properties

$\Delta(p, q)$

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

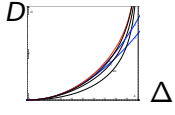
$D(p\|q)$

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

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

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

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



Summary of Properties

$\Delta(p, q)$

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

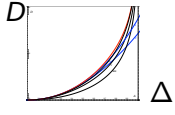
$D(p\|q)$

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

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

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

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

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

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

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

but this does not tensorize. However, divergence does:

$$D(\otimes_{i=1}^{n} P_i \| \otimes_{i=1}^{n} Q_i) = n \cdot d(p\|q) = n \cdot \log \frac{1}{\sqrt{1-4\delta^2}}$$

where $\delta = |p - q|$.

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

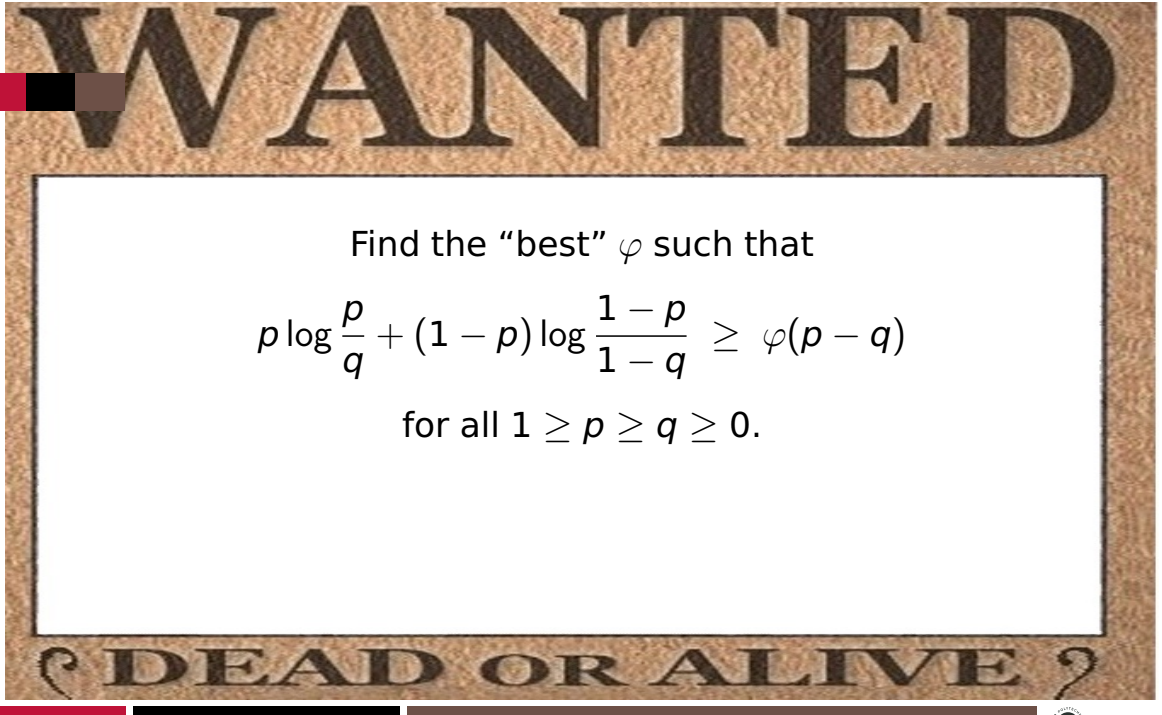
$$\Delta(\otimes_{i=1}^{n} P_i, \otimes_{i=1}^{n} Q_i) \geq |(1-\epsilon) - \epsilon| = 1 - 2\epsilon$$

but this does not tensorize. However, divergence does:

$$D(\otimes_{i=1}^{n} P_i \| \otimes_{i=1}^{n} Q_i) = n \cdot d(p\|q) = n \cdot \log \frac{1}{\sqrt{1-4\delta^2}}$$

where $\delta = |p - q|$. Then by Pinsker inequality, $\boxed{D \geq \varphi(\Delta)}$, there should be at least

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


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 $\boxed{c = 2}$.

[1] Assuming natural logarithms—otherwise divide D by $\log e$.

The classical “Pinsker” Inequality

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

where 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

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

The classical “Pinsker” Inequality

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

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

- *Pinsker did not explicitly state Pinsker’s inequality* (not even in some other form $D \geq \varphi(\Delta)$)

The classical “Pinsker” Inequality

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

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

- *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

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

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

- *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 $\Delta \leq D + 10\sqrt{D}$
- Some math (Verdú, 2014) gives $D \geq c \cdot \Delta^2$ with the suboptimal $c = \frac{1}{102}$.

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!

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!

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!

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 $c = \frac{1}{4}$)

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).

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

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

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 $\frac{4}{9}$!

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?

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

Marcel-Paul (Marco) Schützenberger

Marcel-Paul (Marco) Schützenberger

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

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

- 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*
- during WWII, appointed intern at the psychiatric hospital in Blois; active in resistance activities (worked for the Intelligence Service); and published his first mathematical paper

SÉANCE DU 31 MAI 1943. 71

THÉORIE DES GROUPES. — *Sur la théorie des structures de Dedekind.* Note de M. **Marcel-Paul Schutzenberger**, présentée par M. Gaston Julia

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*

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"

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)

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.

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

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

Marcel-Paul (Marco) Schützenberger

- 1952, came to information theory from biostatistics:
- published his mathematical thesis in 1953 (director: Darmois, president: Fréchet)

CONTRIBUTION

AUX

APPLICATIONS STATISTIQUES

DE LA

THÉORIE DE L'INFORMATION

par

M. P. SCHÜTZENBERGER

Marcel-Paul (Marco) Schützenberger

- 1952, came to information theory from biostatistics:
- published his mathematical thesis in 1953 (director: Darmois, president: Fréchet)
- invited to MIT by Claude Shannon, he spent 1956-57 there

Marcel-Paul (Marco) Schützenberger

- 1952, came to information theory from biostatistics:
- published his mathematical thesis in 1953 (director: Darmois, president: Fréchet)
- invited to MIT by Claude Shannon, he spent 1956-57 there
- at the crossroads of algebra and theoretical computer science: variable-length codes, monoids, automata theory, Kleene-Schützenberger theorem. His seminal paper *Une théorie algébrique du codage* (1958) lays foundations of automata theory and its relationship with rational languages and semi-groups.

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 :

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

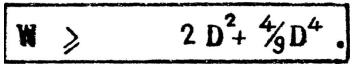
Tous les termes sont positifs car le polynome $t^{2i} - 2it + 2i - 1$ a un unique extremum pour $t = 1$ et prend en ce point la valeur 0 .

Bien plus :

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

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

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 !

From $4/3$ to $4/9$

- 1953, most likely, the original French manuscript of Schützenberger was published with the wrong constant 4/3

From $4/3$ to $4/9$

- 1953, most likely, the original French manuscript of Schützenberger was published with the wrong constant 4/3
- 1966, Kambo and Kotz (*On exponential bounds for binomial probabilities*) copied Schützenberger's derivation verbatim, without citing the original reference and without correcting the error 4/3 !

From $4/3$ to $4/9$

- 1953, most likely, the original French manuscript of Schützenberger was published with the wrong constant 4/3
- 1966, Kambo and Kotz (*On exponential bounds for binomial probabilities*) copied Schützenberger's derivation verbatim, without citing the original reference and without correcting the error 4/3 !
- 1967, Kullback mentioned Schützenberger's result to derive the improved Pinsker inequality (with the erroneous constant 4/3)

From $4/3$ to $4/9$

- 1953, most likely, the original French manuscript of Schützenberger was published with the wrong constant 4/3
- 1966, Kambo and Kotz (*On exponential bounds for binomial probabilities*) copied Schützenberger's derivation verbatim, without citing the original reference and without correcting the error 4/3 !
- 1967, Kullback mentioned Schützenberger's result to derive the improved Pinsker inequality (with the erroneous constant 4/3)
- 1969, Krafft (*A Note on On exponential bounds for binomial probabilities*) pointed the wrong constant in Kambo and Kotz's paper and corrected 4/3 to 4/9

From $4/3$ to $4/9$

- 1953, most likely, the original French manuscript of Schützenberger was published with the wrong constant 4/3
- 1966, Kambo and Kotz (*On exponential bounds for binomial probabilities*) copied Schützenberger's derivation verbatim, without citing the original reference and without correcting the error 4/3 !
- 1967, Kullback mentioned Schützenberger's result to derive the improved Pinsker inequality (with the erroneous constant 4/3)
- 1969, Krafft (*A Note on On exponential bounds for binomial probabilities*) pointed the wrong constant in Kambo and Kotz's paper and corrected 4/3 to 4/9
- 1970, Vajda pointed the wrong constant in Kullback's paper and claimed that "*Kraft corrected an inequality of Schützenberger on which Kullback's result was based*".

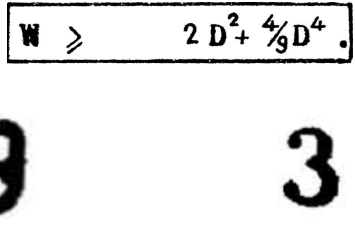
From $4/3$ to $4/9$

- 1953, most likely, the original French manuscript of Schützenberger was published with the wrong constant 4/3
- 1966, Kambo and Kotz (*On exponential bounds for binomial probabilities*) copied Schützenberger's derivation verbatim, without citing the original reference and without correcting the error 4/3 !
- 1967, Kullback mentioned Schützenberger's result to derive the improved Pinsker inequality (with the erroneous constant 4/3)
- 1969, Krafft (*A Note on On exponential bounds for binomial probabilities*) pointed the wrong constant in Kambo and Kotz's paper and corrected 4/3 to 4/9
- 1970, Vajda pointed the wrong constant in Kullback's paper and claimed that "*Kraft corrected an inequality of Schützenberger on which Kullback's result was based*".
- 1970, Correction was acknowledged by Kullback in the transactions on Information Theory

From $4/3$ to $4/9$

- 1953, most likely, the original French manuscript of Schützenberger was published with the wrong constant 4/3
- 1966, Kambo and Kotz (*On exponential bounds for binomial probabilities*) copied Schützenberger's derivation verbatim, without citing the original reference and without correcting the error 4/3 !
- 1967, Kullback mentioned Schützenberger's result to derive the improved Pinsker inequality (with the erroneous constant 4/3)
- 1969, Krafft (*A Note on On exponential bounds for binomial probabilities*) pointed the wrong constant in Kambo and Kotz's paper and corrected 4/3 to 4/9
- 1970, Vajda pointed the wrong constant in Kullback's paper and claimed that "*Kraft corrected an inequality of Schützenberger on which Kullback's result was based*".
- 1970, Correction was acknowledged by Kullback in the transactions on Information Theory
- Later, Schützenberger corrected by hand 4/3 to 4/9 in his thesis manuscript

From 4/3 to 4/9



(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.

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 $\frac{2}{9}$

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 !

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 $\frac{32}{135}$

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 $D = +\infty$

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: $D \geq \log \frac{1+\Delta}{1-\Delta} - 2 \log e \cdot \frac{\Delta}{1+\Delta}$

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: $\Delta \leq \sqrt{1 - \exp(-D)}$)

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

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:

$$\boxed{D \geq \log \frac{1}{1-\Delta} - (1-\Delta)\log(1+\Delta)}$$

(simple proof in the paper)

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 in the paper:

Theorem

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

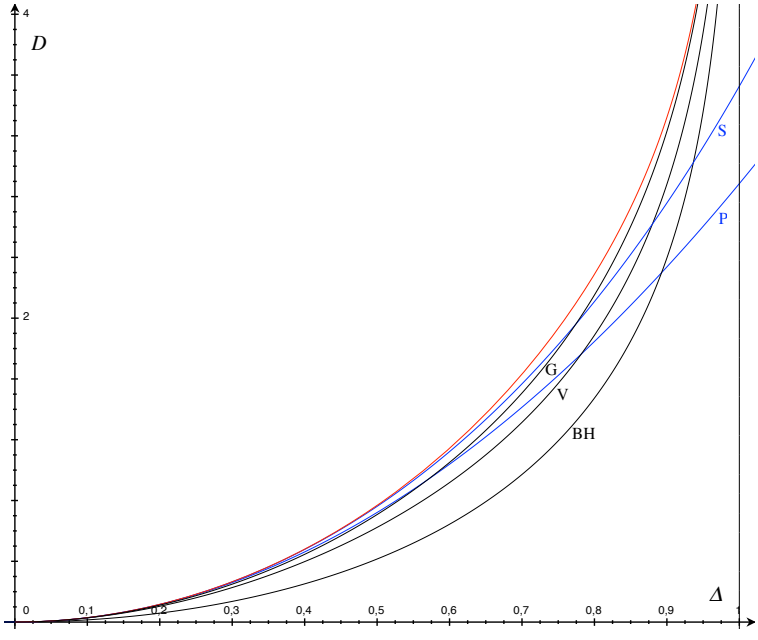
$$\begin{cases} \Delta &= \lambda(1-q)q \\ D &= \log(1-\lambda q) + \lambda q(1+\lambda(1-q))\log e \end{cases} \tag{1}$$

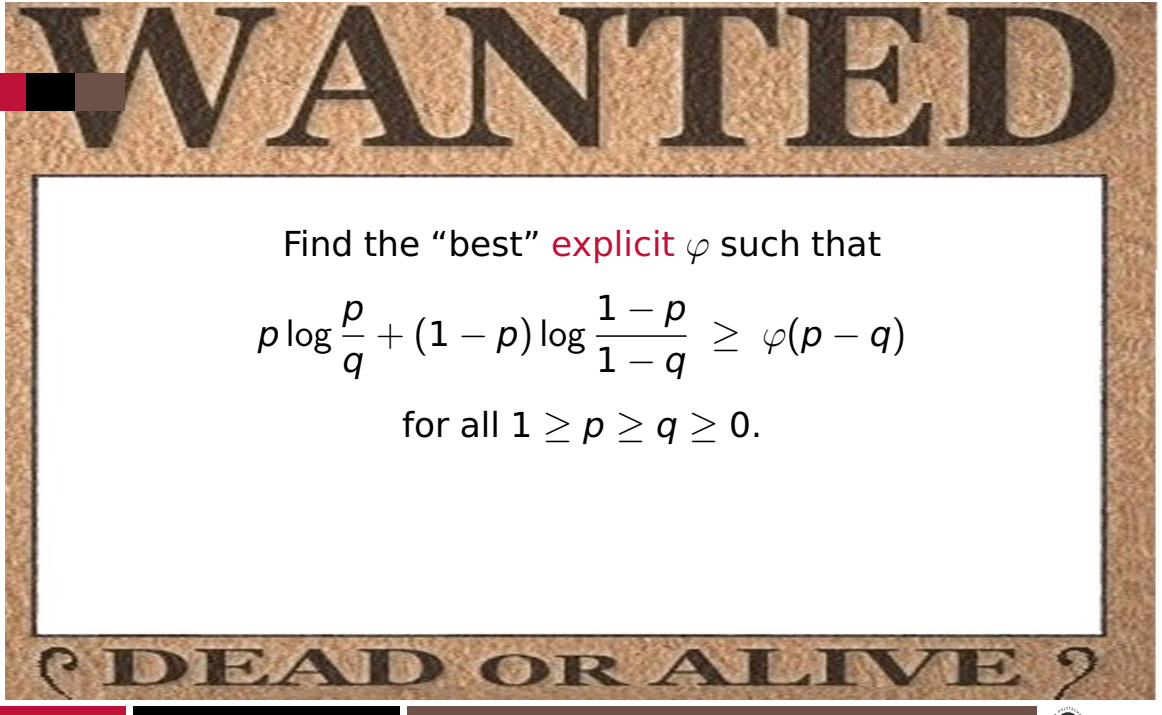
where $\lambda \geq 0$ is the parameter and $q = q(\lambda) \triangleq \frac{1}{\lambda} - \frac{1}{e^{\lambda}-1} \in [0, \frac{1}{2}]$..

Proof.

(simple proof in the paper using Lagrangian) □

Summary




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 ?

A Historical Perspective on Schützenberger-Pinsker Inequalities

Thank you!

Olivier Rioul

Télécom Paris,
Institut Polytechnique de Paris,
France

<olivier.rioul@telecom-paris.fr>