# In Memory of Igor Vajda

Imre Csiszár

Rényi Institute of Mathematics, Hungarian Academy of Sciences

Information Geometry Conference
Leipzig
Aug 2-10, 2010

# Biography of Igor Vajda

- Born 1942
- Graduated 1965, Czech Technical University
- PhD 1968, Charles University, Prague
- Work: UTIA (Institute of Information Theory and Automation, Czech Academy of Sciences); member of Board of UTIA: 1990
- Visiting Professor: Catholic Universiteit Leuven, Complutense Universidad Madrid, Université de Montpellier, M. Hérnandez Universidad, Alicante
- Member of IEEE 1990, Fellow 2001
- 4 monographs, more than 100 journal publications
- Awards: Prize of the Academy of Sciences, Jacob Wolfowitz Prize, Medal of merits of Czech Technical University, several Annual prizes of UTIA

# Subject of this memorial lecture

Igor Vajda's contribution to "distances" of probability distributions (PD's) and their statistical applications

- a major direction of his research
- of main interest for this audience
- also in the speaker's research interest

Primarily: $f$-divergences in general, and their subclass called power divergences

Secondarily: other distances not in this class, as Bregman distances

# $f$-divergence $D_f(P\|Q)$

$f$ any convex function on $(0,\infty)$. The conventions

$$f(0) \triangleq \lim_{u\downarrow 0} f(u), \quad 0f\left(\frac{0}{0}\right) \triangleq 0, \quad 0f\left(\frac{u}{0}\right) \triangleq u \lim_{t\to\infty} \frac{f(t)}{t}$$

make $vf(u/v)$ convex, lower semicontinuous on $[0,\infty)^2$.
For PD's $P$, $Q$ on a set $\mathbf{X}$ (endowed with a $\sigma$-algebra)

$$D_f(P\|Q) \triangleq \begin{cases} \sum q_i f(\frac{p_i}{q_i}) & \text{discrete case} \\ \int q f(\frac{p}{q}) \mathrm{d}\mu & p \triangleq \frac{\mathrm{d}P}{\mathrm{d}\mu} \quad q \triangleq \frac{\mathrm{d}Q}{\mathrm{d}\mu} \quad \text{general case;} \end{cases}$$

does not depend on the dominating measure, one may take $\mu = P + Q$.
[Csiszár 1963, 67, Ali-Silvey 1966; books: Liese-Vajda, Teubner 1987, Vajda, Kluwer 1989.]
Definition makes sense beyond PD's but not considered here.

# Properties of $f$-divergences

Unless stated otherwise, $f(1) = 0$, and strict convexity at $1$ are assumed.

$$0 \leq D_f(P\|Q) \leq f(0) + f^*(0), \quad f^*(t) \triangleq tf\left(\frac{1}{t}\right).$$

First equality iff $P = Q$ ($D_f$ is a "distance")
Second equality, for strictly convex $f$ with $f(0) + f^*(0) < \infty$: iff $P \perp Q$.

$D_f(Q\|P) = D_{f^*}(P\|Q)$.

Data processing inequality, for partitions $\mathcal{A} = (A_1, \ldots, A_k)$ of $\mathbf{X}$:

$$D_f(P^{\mathcal{A}}\|Q^{\mathcal{A}}) \leq D_f(P\|Q), \text{ where } P^{\mathcal{A}} \triangleq (P(A_1), \ldots, P(A_k)).$$

Pardo-Vajda 1997: Characterizes $f$-divergences among distances of form $\sum \delta(p_i, q_i)$.

# Examples

- Kullback $I$-divergence (relative entropy) $D(P\|Q)$: $f(t) = t \log t$
- Power divergences $D_\alpha(P\|Q) \overset{\triangle}{=} c_\alpha(\sum p_i^\alpha q_i^{1-\alpha} - 1)$: $f(t) = c_\alpha(t^\alpha - 1)$ (Perez 1967 with $c_\alpha = \frac{1}{\alpha-1}$; Cressie-Read 1984 and Liese-Vajda 1987 with $c_\alpha = \frac{1}{\alpha(\alpha-1)}$, admitting inclusion of $D(P\|Q)$ and $D(Q\|P)$ as limits for $\alpha \to 1$ or $0$.)

  The order-$\alpha$ divergence of Rényi 1961 is a function of $D_\alpha(P\|Q)$.

  $\alpha = 2$: Pearson's $\chi^2$ $\quad \sum \frac{p_i^2}{q_i} - 1 = \sum \frac{(p_i - q_i)^2}{q_i}$

  $\alpha = 1/2$: Hellinger $\quad 2(1 - \sum \sqrt{p_i q_i}) = \sum(\sqrt{p_i} - \sqrt{q_i})^2$
- $\chi^\alpha(P, Q) \overset{\triangle}{=} \sum |p_i - q_i|^\alpha q_i^{1-\alpha}$ of Vajda 1972: $f(t) = |t - 1|^\alpha$, $\alpha \geq 1$

  $\alpha = 1$: $\chi^1(P, Q) = |P - Q| \overset{\triangle}{=} \sum |p_i - q_i|$, variation distance

# Inequalities for $f$-divergencies

Here, Vajda's first and last contributions to this subject are mentioned.

Vajda 1970: sharpened the "Pinsker inequality" $D(P\|Q) \geq \frac{1}{2}|P-Q|^2$ by adding a fourth power term.
Vajda 1972: lower and upper bounds to $D_f(P\|Q)$ in terms of $|P-Q|$, using that the minimum of $D_f(P\|Q)$ subject to $|P-Q|=V$ is attained for PD's on a two-point set. The minimum of $vf\left(\frac{u}{v}\right)+(1-v)f\left(\frac{1-u}{1-v}\right)$ subject to $2(u-v)=V$ has since been called Vajda' tight lower bound.

Harremoöes -Vajda 2010: Given convex functions $f$, $g$, the range of the map $(P, Q) \mapsto (D_f(P\|Q), D_g(P\|Q))$ is a convex set in $\mathbf{R}^2$. Each point in this set is achieved by PD's on a 4-point set, but not necessarily on a two-point set. Explicitly determined the range set for some pairs of divergences, including the power divergences of orders 2 and 3.

# Metric divergences

No $f$-divergences except constant multiples of the variation distance are metrics (Khosravifard-Fooladivanda-Gulliver 2007).
Powers of $f$-divergences may be metrics, thus symmetric and satisfy the triangle inequality, such as Hellinger distance.
Csiszár-Fischer 1962: powers of symmetrized $\alpha$-divergences, $0 < \alpha < 1$.
Kafka-Österreicher-Vincze 1991: If $f = f^*$ and $f(t)/(1-t^\beta)^{1/\beta}$ is nondecreasing for $t \in [0,1)$ then $[D_f(P\|Q)]^\beta$ is a metric.

Österreicher-Vajda 2003 and Vajda 2009: For each $\alpha \in \mathbf{R}$, the function

$$f_\alpha(t) = \frac{\text{sign}\alpha}{1-\alpha}\left[(t^{1/\alpha}+1)^\alpha - 2^{\alpha-1}(t+1)\right] \quad \alpha \neq 0,\, 1$$

$$f_1(t) = t\log t + (t+1)\log\frac{2}{t+1} \qquad f_0(t) = \frac{1}{2}|t-1|$$

is convex, meets above condition with $\beta = 1/2$ or $1/\alpha$ if $\alpha \leq 2$ or $> 2$.
The special cases $\alpha = -1,\, 0,\, 1,\, 2$ give well-known $f$-divergences.

# Chernoff distance and error probability

Given: finite family $\{P_\vartheta, \vartheta \in \Theta\}$ of PD's on $\mathbf{X}$ and prior distribution on $\Theta$; observed: $\mathbf{X}$-valued random variables $X_1, \ldots, X_n$, conditionally i.i.d. on the condition $\vartheta = i$ with distribution $P_i$.

Estimator of $\vartheta$: mapping $d : \mathbf{X}^n \to \Theta$.
Error probability $e = \Pr\{d(X_1, \ldots, X_n) \neq \vartheta\}$ is minimized by Bayes estimator: $d(X_1, \ldots, X_n)$ equals an $i \in \Theta$ with largest posterior probability $\Pr\{\vartheta = i | X_1, \ldots, X_n\}$.

Chernoff 1952: For $\Theta = \{1, 2\}$

$$\lim_{n\to\infty} \frac{1}{n} \log(e_{\mathrm{Bayes}}) = \log \inf_{0<\alpha<1} \int p_1^\alpha p_2^{1-\alpha} \mathrm{d}\mu \triangleq -D_{\mathrm{Ch}}(P_1, P_2).$$

# Further results on Bayes error

Vajda 1967: the Bayes error is exponentially small even in non-i.i.d. cases (exact exponent not specified)

Rényi 1969, Vajda 1969: Chernoff's result holds also for the conditional entropy $H(\vartheta|X_1, \ldots, X_n)$ in the role of $\log e_{\mathrm{Bayes}}$. Vajda 1969 also addressed the case $\Theta = \{1, \ldots, k\}$, replacing the Chernoff distance $D_{\mathrm{Ch}}(P_1, P_2)$ by $\min_{i \neq j} D_{\mathrm{Ch}}(P_i, P_j)$.

Vajda 1970 filled mathematical details; in the definition of $D_{\mathrm{Ch}}(P_1, P_2)$ one can take

$$\min_{0 \leq \alpha \leq 1} \int_{p_1 p_2 > 0} p_1^{\alpha} p_2^{1-\alpha} \mathrm{d}\mu,$$

restriction needed for not mutually absolutely continuous $P_1, P_2$.

# Divergence-based estimation and testing

Observing i.i.d. random variables $X_1, \ldots, X_n$ whose commmon distribution is an unknown member of a family $\{P_\vartheta, \vartheta \in \Theta\}$ of PD's on $\mathbf{X}$, one expects that the unkown distribution is close to the empirical disrtibution $\hat{P}_n$ of the sample $(X_1, \ldots, X_n)$. This suggest the estimate

$$\hat{\vartheta} \triangleq \text{argmin}_\vartheta D_f(\hat{P}_n \| P_\vartheta) \quad (f(t) = t \log t \Rightarrow MLE).$$

If it is a hypothesis to be tested that the commmon distribution of the $X_i$'s indeed belongs to the family $\{P_\vartheta, \vartheta \in \Theta\}$, a natural acceptance criterion is $\inf_\vartheta D(\hat{P}_n \| P_\vartheta) \leq c_n$, for suitable $c_n \downarrow 0$.

$f(t) = t^2 - 1$ or $t \log t$ give the classical $\chi^2$ and likelihood ratio tests. Liese-Vajda 1987 provide examples that other $f$-divergence tests may have larger power against some alternatives.

Tests of this kind are directly applicable only in the discrete case. Ways to overcome this problem will be discussed later.

# $f$-divergence tests

Provided that the $f$-estimate $\hat{\vartheta} \triangleq \mathrm{argmin}_{\vartheta} D_f(\hat{P}_n \| P_{\vartheta})$ is well defined, the test statistic $\inf_{\vartheta} D(\hat{P}_n \| P_{\vartheta})$ equals $D_f(\hat{P}_n \| P_{\hat{\vartheta}})$. Modified versions of that statistic are also used, with $\hat{\vartheta}$ minimizing $\tilde{f}$-divergence for some $\tilde{f} \neq f$. For example, $\chi^2$ test may be used with MLE $\hat{\vartheta}$.

Tests of this kind were studied by Cressie-Reed 1984, concentrating on power divergences.
Menandez-Morales-Pardo-Vajda 1995 studied asymptotic distributions of such test statistics in a general setting, also including previously not considered scenarios where $\hat{P}_n$ is not necessarily an empirical distribution (but the average of several ones). They also addressed the choice of $f$, concentrating on power divergences, extending results of Cressie-Reed.

# Convergence of quantized $f$-divergences

What sequences of partitions $\mathcal{A}_n = (A_{n1}, \ldots, A_{nk_n})$ of $\mathbf{X}$ yield

$$\lim_{n\to\infty} D_f(P^{\mathcal{A}_n} \| Q^{\mathcal{A}_n}) = D_f(P\|Q). \qquad (*)$$

Gelfand-Yaglom 1957, Perez 1957: For I-divergence ($f(t) = t\log t$), $(*)$ holds if the partitions $\mathcal{A}_n$ are refining and $\frac{\mathrm{d}P}{\mathrm{d}(P+Q)}$ is measurable with respect to the smallest $\sigma$-algebra containing each $A_{nk}$ (thus the latter $\sigma$-algebra is $(P,Q)$-sufficient).

Vajda 1972,1973: same proof works for any $f$, also for Fisher information. Given a dominated family of PD's $\{P_\vartheta, \vartheta \in \Theta\}$, where $\Theta \subset \mathbf{R}$ is an open interval, denote $\frac{\mathrm{d}P}{\mathrm{d}\mu} = p_\vartheta$, $\frac{\partial}{\partial\vartheta} p_\vartheta = \dot{p}_\vartheta$. Fisher information is defined by

$$I(\vartheta) \triangleq \int (\dot{p}_\vartheta / p_\vartheta)^2 \, p_\vartheta \mathrm{d}\mu, \quad I^{\mathcal{A}}(\vartheta) \triangleq \sum \Big( \dot{P}_\vartheta(A_i) / P_\vartheta(A_i) \Big)^2 P_\vartheta(A_i).$$

Vajda 1973 also considered $I_\alpha(\vartheta)$, replacing $2$ by $\alpha > 1$.

# Convergence of quantized $f$-divergences, continued

Partition sequences occurring in applications seldom have the refining property. A general sufficient condition for $(*)$ appears in Csiszár 1973. Vajda 2002: For $\mathbf{X} = \mathbf{R}$ and (Lebesgue) absolutely continuous $P$ and $Q$, $(*)$ holds providing for each $x \in \mathbf{R}$ the length of $A_n(x)$, the interval $A_{ni}$ containing $x$, goes to $0$ as $n \to \infty$.
An analogous result for (generalized) Fisher information also holds.
These results are extended also to $\mathbf{X} = \mathbf{R}^d$ and rectangle partitions.

Berlinet-Vajda 2006: If $\{x : q(x) > 0\}$ is an open interval then the condition $\max_j Q(A_{nj}) \to 0$ as $n \to \infty$ implies $(*)$ for each $P << Q$. This condition is also necessary for $(*)$ when $D_f(P\|Q) < \infty$. Strong results are proved also on the speed of convergence in $(*)$, for the case of $\chi^2$-divergence and partitions into nearly $Q$-equiprobable intervals.

# $f$-divergence test, continuous case

The problem of testing a (simple) hypothesis given by a continuous distribution $Q$ on $\mathbf{R}$ may be reduced to that when $Q$ is the uniform distribution on $(0,1)$. Partitioning $(0,1)$ into intervals $A_{nj} = (a_{j-1}, a_j]$, $j = 1, \ldots, k = k_n$, consider the $f$-divergence statistic

$$T_n \triangleq nD_f(Q\|\hat{P}_n) = n \sum_{j=1}^{k} Q(A_{nj}) f\left(\frac{\hat{P}_n(A_{nj})}{Q(A_{nj})}\right).$$

It is convenient to take partitions into intervals either of equal $Q$-measure (length), or of equal $\hat{P}_n$-measure $\frac{1}{n}|\{i : X_i \in A_{nj}\}|$.
In the first case the statistic $T_n$ reduces to standard ones (for usual choices of $f$) whose asymptotic behavior is, on a basic level, well understood. On the next slide, a more refined problem is considered.

# Bahadur efficiency of power divergence tests

Consider the previous hypothesis testing problem, using partitions into intervals of equal $Q$-measure (length). This leads to testing the (simple) hypothesis that the common distribution of i.i.d. random variables $X_1, \dots, X_n$ with values in $\{1, \dots, k\}$ is uniform, via the $f$-divergence test statistic

$$D_f(\hat{P}_n \| Q) = \sum_{j=1}^{k} Q(j) f\left(\frac{\hat{P}_n(j)}{Q(j)}\right) = \frac{1}{k} \sum_{j=1}^{k} f\left(\frac{k}{n} \left|\{i : X_i = j\}\right|\right).$$

Harremoës-Vajda 2008 compared such tests with different functions $f$, concentrating on power divergences with different $\alpha$ (admitting $k$ grow with $n$, subject to $k/n \to 0$). Extending a result of Quine-Robinson 1985 for $\alpha = 1$ versus $\alpha = 2$, they showed that any $\alpha \in (0, 1]$ is infinitely more Bahadur efficient than any $\alpha > 1$.

# $f$-divergences and spacings

Returning to the previous hypothesis testing problem, here take partitions into intervals of equal $\hat{P}_n$-measure.
Denote the ordered sample by $Y_1, \dots, Y_n$, set $Y_0 = 0$, suppose $m = n/k$ is an integer. Then $A_{nj} = (Y_{m(j-1)}, Y_{mj}]$, where $Y_n$ is replaced by $1$, and

$$T_n = m\left[\sum_{j=1}^{k-1} f(k(Y_{mj} - Y_{m(j-1)})) + f(k(1 - Y_{m(k-1)}))\right].$$

Statisics based on “spacings” $Y_j - Y_{j-1}$ or “ $m$-spacings” $Y_{mj} - Y_{m(j-1)}$ are frequently used to test uniformity, on an intuitive background.

Morales-Pardo-Vajda 2003, Vajda-van der Meulen 2006, Vajda 2007: the familiar test statistics using spacings are related to $T_n$ above.
Though formally not special cases of the latter, asymptotic equivalence has been demonstrated by them, and by Vajda-van der Meulen 2010.

# Barron's density estimator

The histogram estimate from an i.i.d. sample of a PD $P$ on $\mathrm{R}$, rather, of its density with respect to a given $Q$, is $f_n(x) \triangleq \hat{P}^n(A_n(x))/Q(A_n(x))$. Let the $k_n$ intervals of the underlying partition $\mathcal{A}_n$ have equal $Q$-measure. Barron 1988 and Barron-Györfi-van der Meulen 1992 proposed mixing the histogram estimate with $Q$, giving weight $\frac{k_n}{n+k_n}$ to $Q$. This yields an estimator consistent in reversed I-divergence, subject to $D(P\|Q) < \infty$.

Berlinet-Vajda-van der Meulen 1998, Györfi-Liese-Vajda-van der Meulen 1998: consistensy in reversed $\chi^2$ divergence, and in other $f$-divergences.

Vajda-van der Meulen 1998, Vajda 2001: The $\chi^2$-divergence of $P$ from the Barron estimate goes to $0$ at best rate if $k_n$ is of order $n^{1/3}$ (subject to regularity conditions). Vajda 2001 also addressed how to chose $Q$.

Beirant-Berlinet-Biau-Vajda 2002: Smooth Barron-type estimators.

# Regularized $f$-divergence

Statistical applications of $f$-divergence are facilitated by a “regularization” to get meaningful values for $P \perp Q$, retaining the original ones for $P \equiv Q$.

Liese -Vajda 2006, Broniatowski-Keziou 2006, Broniatowski-Vajda 2009:

$$\underline{D}_f(P, \tilde{P} \| Q) \triangleq \int f'_+(\frac{\mathrm{d}P}{\mathrm{d}\tilde{P}})\mathrm{d}P + \int \left[ f(\frac{\mathrm{d}P}{\mathrm{d}\tilde{P}}) - f'_+(\frac{\mathrm{d}P}{\mathrm{d}\tilde{P}})\frac{\mathrm{d}P}{\mathrm{d}\tilde{P}} \right] \mathrm{d}Q \le D_f(P\|Q)$$

for $P \equiv \tilde{P}$, and any $Q$; equality if $\tilde{P} = Q$.
Given a family $\{P_\vartheta, \vartheta \in \Theta\}$ of mutually absolutely continuous PD's, they suggested a modified $f$-divergence statistic, replacing $D_f(P_\vartheta \| \hat{P}_n)$ either by $\underline{D}_f(P_\vartheta, P_{\tilde{\vartheta}} \| \hat{P}_n)$ (for a fixed “escort parameter” $\tilde{\vartheta}$) or by the supremum of the latter subject to $\tilde{\vartheta} \in \Theta$. This supremum is equal to $D_f(P_\vartheta \| \hat{P}_n)$ in the discrete case, and is typically non-trivial also otherwise.

For $f(t) = -\log t$, both kinds of modified statistics yield the MLE.

# Other distances distances

A variant of Bregman (1967) distance, called scaled Bregman distance by Stummer-Vajda 2009: For PD's $P$, $Q$ dominated by $M$
$B_f(P,Q|M) \triangleq \int \left[ f(\frac{\mathrm{d}P}{\mathrm{d}M}) - f(\frac{\mathrm{d}Q}{\mathrm{d}M}) - f'_+(\frac{\mathrm{d}Q}{\mathrm{d}M})(\frac{\mathrm{d}P}{\mathrm{d}M} - \frac{\mathrm{d}Q}{\mathrm{d}M}) \right] \mathrm{d}M.$
They explicitly calculated $B_f$-distances for power functions $f_\alpha$, including $\alpha = 0$ and $1$, specifically for $P, Q, M$ in an exponential family.

Broniatowsky-Vajda 2009 studied more general distances of PD's, of form $\mathcal{D}_\psi(P,Q) \triangleq \int \psi(p,q)\mathrm{d}\mu$, where $\psi$ is any "decomposable" distance on $(0,\infty)$, $\psi(s,t) = \psi^0(s) + \psi^1(t) + \rho(s)t$. They give rise to (generalized) M-estimators, whose theory is well developed (Vajda substantially contributed). This class contains the previous $\underline{D}_f$, Bregman distances, and the generalized power divergences of Basu et al. 1998, studied in detail.

Györfi-Vajda-van der Meulen 1996: Consistency of minimum Kolmogorov distance estimates.

# Efficiency and robustnes

ML estimators are typically efficient but not robust. Other estimators, including $f$-divergence based ones, may be preferable for robustness. Vajda paid substantial attention to robustness problems, computed influence functions for various statistics considerd by him. His interest in $f$-divergences was partially motivated by robustness considerations.

Broniatowski-Vajda 2009: The last family of distances studied there yields estimators that in some cases outperform the best robust estimators known before.

Gyorfi-Vajda 2001: Studied "blended statistics" of Lindsay 1994, corresponding to $f$-divergences with $f(t) = (1-t)^2/[a+(1-a)t]$, robust for $0 < a < 1$ unlike the "extremal cases" $\chi^2$ and reversed $\chi^2$ ($a = 1/2$ gives the divergence of Vincze 1981 and Le Cam 1986).

# Omitted topics

Many important contributions of Vajda could not be covered.

- Development of the theory of statistical experiments (sufficiency, deficiency, etc.) via $f$-vivergences
- Generalized entropies, relation to divergences
- Statistical theory of M-estimators
- Neural nets
- Log-optimal portfolios

Igor Vajda has been a highly productive scientist, even in the last years full with ideas that he no longer had time to develop.

The scientific community will badly miss him, and his many friends even more as a truely lovable person.