On Csiszár's f-Divergences and Informativities with Applications

Igal Sason

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa, Israel

Conference for Celebrating the 80th Birthday of Imre Csiszár

Alfréd Rényi Institute of Mathematics
Hungarian Academy of Sciences
Budapest, Hungary, June 4–5, 2018

f-Divergences

- Probability theory, information theory, learning theory, statistical signal processing and many other disciplines, greatly benefit from divergence measures.

f-Divergences

- Probability theory, information theory, learning theory, statistical signal processing and many other disciplines, greatly benefit from divergence measures.
- f-divergences (Csiszár, 1963) form a large class of divergence measures, indexed by convex functions f, which include as special cases:
  - I-divergences (relative entropies);
  - χ^2-divergence;
  - squared Hellinger distance;
  - total variation distance;
  - DeGroot statistical information;
  - etc.

f-Divergences

- Probability theory, information theory, learning theory, statistical signal processing and many other disciplines, greatly benefit from divergence measures.
- f-divergences (Csiszár, 1963) form a large class of divergence measures, indexed by convex functions f, which include as special cases:
  - I-divergences (relative entropies);
  - χ^2-divergence;
  - squared Hellinger distance;
  - total variation distance;
  - DeGroot statistical information;
  - etc.
- f-divergences satisfy the data processing inequality.

f-Informativities

f-Informativities (Csiszár, 1972) form a generalization of the mutual information:

- KL divergence $\implies$ Shannon's Mutual Information;
- In general, f-divergence $\implies$ f-informativity.

The Origins

- I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, Jan. 1963.
- I. Csiszár, "A note on Jensen's inequality,' *Studia Scientiarum Mathematicarum Hungarica*, vol. 1, pp. 185–188, 1966.
- I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, Jan. 1967.
- I. Csiszár, "On topological properties of f-divergences," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 329–339, Jan. 1967.
- I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematicarum Hungarica*, vol. 2, pp. 191–213, Mar. 1972.
- S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistics Society*, series B, vol. 28, no. 1, pp. 131–142, Jan. 1966.

Scope of this talk

Properties, and applications of f-divergences and f-informativities.

Notation

- $\mathcal{C}$ denotes the set of convex functions $f\colon (0,\infty) \mapsto \mathbb{R}$ with $f(1) = 0$;
- P and Q are probability measures;
- $P, Q \ll \mu$ (e.g., $\mu = \frac{1}{2}(P + Q)$), and $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $q := \frac{\mathrm{d}Q}{\mathrm{d}\mu}$.

Notation

- $\mathcal{C}$ denotes the set of convex functions $f\colon (0,\infty) \mapsto \mathbb{R}$ with $f(1) = 0$;
- P and Q are probability measures;
- $P, Q \ll \mu$ (e.g., $\mu = \frac{1}{2}(P+Q)$), and $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $q := \frac{\mathrm{d}Q}{\mathrm{d}\mu}$.

f-Divergence: Definition

The f-divergence from P to Q is given, independently of μ, by

$$D_f(P\|Q) := \int q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu \tag{1}$$

with the convention that

$$f(0) := \lim_{t \downarrow 0} f(t), \tag{2}$$

$$0f\left(\frac{0}{0}\right) := 0, \qquad 0f\left(\frac{a}{0}\right) := \lim_{t \downarrow 0} tf\left(\frac{a}{t}\right) = a \lim_{u \to \infty} \frac{f(u)}{u},\ a > 0. \tag{3}$$

f-divergences: Examples

- Relative entropy

$$f(t) = t \log t, \quad t > 0 \implies D_f(P\|Q) = D(P\|Q), \tag{4}$$

$$f(t) = -\log t, \; t > 0 \implies D_f(P\|Q) = D(Q\|P). \tag{5}$$

f-divergences: Examples

- Relative entropy

$$f(t) = t \log t, \quad t > 0 \implies D_f(P\|Q) = D(P\|Q), \tag{4}$$

$$f(t) = -\log t, \; t > 0 \implies D_f(P\|Q) = D(Q\|P). \tag{5}$$

- Total variation (TV) distance

$$f(t) = |t-1|, \quad t \geq 0 \tag{6}$$

$$\Rightarrow D_f(P\|Q) = |P-Q| := \int \left| \frac{\mathrm{d}P}{\mathrm{d}\mu} - \frac{\mathrm{d}Q}{\mathrm{d}\mu} \right| \mathrm{d}\mu, \quad P, Q \ll \mu. \tag{7}$$

f-divergences: Examples

- Relative entropy

$$f(t) = t \log t, \quad t > 0 \implies D_f(P\|Q) = D(P\|Q), \tag{4}$$

$$f(t) = -\log t, \; t > 0 \implies D_f(P\|Q) = D(Q\|P). \tag{5}$$

- Total variation (TV) distance

$$f(t) = |t-1|, \quad t \geq 0 \tag{6}$$

$$\Rightarrow D_f(P\|Q) = |P-Q| := \int \left| \frac{\mathrm{d}P}{\mathrm{d}\mu} - \frac{\mathrm{d}Q}{\mathrm{d}\mu} \right| \mathrm{d}\mu, \quad P, Q \ll \mu. \tag{7}$$

- Power divergence of order $\alpha \in (0,1) \cup (1,\infty)$:

$$f_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}, \quad t \geq 0 \tag{8}$$

$$\Rightarrow \mathcal{I}_\alpha(P\|Q) := D_{f_\alpha}(P\|Q) := \tfrac{1}{\alpha(\alpha-1)} \left(\int \left(\tfrac{\mathrm{d}P}{\mathrm{d}\mu}\right)^\alpha \left(\tfrac{\mathrm{d}Q}{\mathrm{d}\mu}\right)^{1-\alpha} \mathrm{d}\mu - 1 \right).$$

f-divergences: Examples (cont.)

- χ^2-divergence:

$$\chi^2(P\|Q) := \int \frac{(p-q)^2}{q}\,\mathrm{d}\mu = \tfrac{1}{2}\,\mathcal{I}_2(P\|Q). \qquad (9)$$

f-divergences: Examples (cont.)

- χ^2-divergence:

$$\chi^2(P\|Q) := \int \frac{(p-q)^2}{q}\, \mathrm{d}\mu = \tfrac{1}{2}\,\mathcal{I}_2(P\|Q). \tag{9}$$

- Relative entropies: continuous extension at $\alpha = 0$ and $\alpha = 1$ yield

$$\mathcal{I}_1(P\|Q) = \tfrac{1}{\log e}\, D(P\|Q), \quad \mathcal{I}_0(P\|Q) = \tfrac{1}{\log e}\, D(Q\|P). \tag{10}$$

f-divergences: Examples (cont.)

- χ^2-divergence:

$$\chi^2(P\|Q) := \int \frac{(p-q)^2}{q}\,\mathrm{d}\mu = \tfrac{1}{2}\,\mathcal{I}_2(P\|Q). \tag{9}$$

- Relative entropies: continuous extension at $\alpha = 0$ and $\alpha = 1$ yield

$$\mathcal{I}_1(P\|Q) = \tfrac{1}{\log e}\,D(P\|Q), \quad \mathcal{I}_0(P\|Q) = \tfrac{1}{\log e}\,D(Q\|P). \tag{10}$$

- Squared Hellinger distance:

$$\mathscr{H}^2(P\|Q) := \tfrac{1}{2}\int (\sqrt{p}-\sqrt{q})^2\,\mathrm{d}\mu = 1 - \int \sqrt{pq}\,\mathrm{d}\mu = \tfrac{1}{4}\,\mathcal{I}_{\frac{1}{2}}(P\|Q). \tag{11}$$

f-divergences: Examples (cont.)

- χ^2-divergence:

$$\chi^2(P\|Q) := \int \frac{(p-q)^2}{q} \, \mathrm{d}\mu = \tfrac{1}{2} \mathcal{I}_2(P\|Q). \tag{9}$$

- Relative entropies: continuous extension at $\alpha = 0$ and $\alpha = 1$ yield

$$\mathcal{I}_1(P\|Q) = \tfrac{1}{\log e} \, D(P\|Q), \quad \mathcal{I}_0(P\|Q) = \tfrac{1}{\log e} \, D(Q\|P). \tag{10}$$

- Squared Hellinger distance:

$$\mathscr{H}^2(P\|Q) := \tfrac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\mu = 1 - \int \sqrt{pq} \, \mathrm{d}\mu = \tfrac{1}{4} \mathcal{I}_{\frac{1}{2}}(P\|Q). \tag{11}$$

- Rényi divergence of order $\alpha \in (0, 1) \cup (1, \infty)$:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \, \log\big(1 + \alpha(\alpha - 1) \, \mathcal{I}_\alpha(P\|Q)\big). \tag{12}$$

Measures of Dependence (Rényi 1959, Csiszár 1967)

- Rényi formulated postulates for dependence measures between two random variables, and studied properties of such measures.

Measures of Dependence (Rényi 1959, Csiszár 1967)

- Rényi formulated postulates for dependence measures between two random variables, and studied properties of such measures.
- Csiszár suggested using f-divergences as dependence measures: $D_f(P_{XY} \| P_X \times P_Y)$ fulfills the postulates by Rényi if $f \in \mathcal{C}$ is strictly convex at 1, and $\lim_{t \to \infty} \frac{f(t)}{t} = +\infty$.

Measures of Dependence (Rényi 1959, Csiszár 1967)

- Rényi formulated postulates for dependence measures between two random variables, and studied properties of such measures.
- Csiszár suggested using f-divergences as dependence measures: $D_f(P_{XY} \| P_X \times P_Y)$ fulfills the postulates by Rényi if $f \in \mathcal{C}$ is strictly convex at 1, and $\lim_{t \to \infty} \frac{f(t)}{t} = +\infty$.
  - Mutual information:

$$f(t) = t \log t, \; (t > 0) \implies D_f(P_{XY} \| P_X \times P_Y) = I(X;Y). \tag{13}$$

  - Mean square contingency: $f(t) = (t-1)^2, \; (t \geq 0)$

$$\implies D_f(P_{XY} \| P_X \times P_Y) = \chi^2(P_{XY} \| P_X \times P_Y) := \phi^2(X,Y). \tag{14}$$

Reflexivity: If $f \in \mathcal{C}$, then $D_f(P\|Q) \geq 0$.
If f is also strictly convex at 1, then $D_f(P\|Q) = 0 \iff P = Q$.

Convexity: $D_f(P\|Q)$ is convex in (P, Q).

Uniqueness: f and g-divergences are identical if and only if there exists a constant $c \in \mathbb{R}$ such that

$$f(t) - g(t) = c\,(t-1), \quad t > 0.$$

Symmetry: let f^* be the $*$-conjugate function of $f \in \mathcal{C}$, given by

$$f^*(t) = t\, f\left(\tfrac{1}{t}\right) \tag{15}$$

for all $t > 0$. Then, $f^* \in \mathcal{C}$, and

$$D_f(P\|Q) = D_{f^*}(Q\|P). \tag{16}$$

Distance Metrics

- No f-divergence, except for positive constant multiples of the total variation distance, is a distance metric (Gulliver et al., "Confliction of the convexity and metric properties in f-divergences," 2007).

Distance Metrics

- No f-divergence, except for positive constant multiples of the total variation distance, is a distance metric (Gulliver et al., "Confliction of the convexity and metric properties in f-divergences," 2007).
- Csiszár and Fischer considered powers of symmetrized α divergences for $\alpha \in (0, 1)$ which are distance metrics:

$$f_\alpha(t) = 1 + t - (t^\alpha + t^{1-\alpha}), \quad t > 0.$$

Distance Metrics

- No f-divergence, except for positive constant multiples of the total variation distance, is a distance metric (Gulliver et al., "Confliction of the convexity and metric properties in f-divergences," 2007).
- Csiszár and Fischer considered powers of symmetrized α divergences for $\alpha \in (0, 1)$ which are distance metrics:

$$f_\alpha(t) = 1 + t - (t^\alpha + t^{1-\alpha}), \quad t > 0.$$

- Kafka et al. (1991): If $f = f^*$ and $f(t)(1 - t^\beta)^{-\frac{1}{\beta}}$ is monotonically non-decreasing on $t \in [0, 1)$, then $D_f^\beta(P\|Q)$ is a distance metric.
- Ostreicher-Vajda (2003) and Vajda (2009) studied explicit f-divergences satisfying the above conditions by Kafka et al.
- Square-roots of f-divergences which are bounded distance metrics:
  - $d_1(P, Q) = \sqrt{\mathscr{H}^2(P\|Q)}$;
  - $d_2(P, Q) = \sqrt{D\left(P\|\frac{1}{2}(P+Q)\right) + D\left(Q\|\frac{1}{2}(P+Q)\right)}$.

Data Processing Inequality (Csiszár, 1967)

Let

- $f \in \mathcal{C}$;
- $(\mathcal{X}, \mathscr{X})$ and $(\mathcal{Y}, \mathscr{Y})$ be measurable spaces;
- P and Q be probability measures on $\mathcal{X}$;
- for all $x \in \mathcal{X}$, $K(\cdot|x)$ is a probability measure that is $\mathscr{Y}$-measurable;
- KP and KQ are prob. measures on $\mathcal{Y}$ such that, for every $\mathcal{B} \in \mathscr{Y}$,

$$KP(\mathcal{B}) := \int_{\mathcal{X}} K(\mathcal{B}|x)\, \mathrm{d}P(x), \quad KQ(\mathcal{B}) := \int_{\mathcal{X}} K(\mathcal{B}|x)\, \mathrm{d}Q(x).$$

Then,

$$D_f(KP \| KQ) \le D_f(P \| Q). \tag{17}$$

Range of Values Theorem (Vajda, 1972)

- The range of an f-divergence is given by

$$0 \leq D_f(P\|Q) \leq f(0) + f^*(0) \tag{18}$$

where

$$f^*(0) := \lim_{t \downarrow 0} f^*(t) = \lim_{u \to \infty} \frac{f(u)}{u}, \tag{19}$$

and

  - $D_f(P\|Q) = 0$ if $P = Q$;
  - $D_f(P\|Q) = f(0) + f^*(0)$ if $P \perp Q$ (i.e., $\text{supp}(P) \cap \text{supp}(Q) = \emptyset$);
  - every value in this range is attainable by a suitable pair of (P, Q).

Strengthened Version (Feldman and Österreicher, 1989)

$$\sup_{P \neq Q} \frac{D_f(P\|Q)}{|P-Q|} = \tfrac{1}{2}\big(f(0) + f^*(0)\big). \tag{20}$$

Sup. is arbitrarily approached by (P, Q) defined on a ternary alphabet.

Strengthened Version (Feldman and Österreicher, 1989)

$$\sup_{P \neq Q} \frac{D_f(P\|Q)}{|P-Q|} = \tfrac{1}{2}\big(f(0) + f^*(0)\big). \tag{20}$$

Sup. is arbitrarily approached by (P, Q) defined on a ternary alphabet.

Implication

$$D_f(P\|Q) \leq \tfrac{1}{2}\big(f(0) + f^*(0)\big)\,|P-Q| \tag{21}$$

if $f(0), f^*(0) < \infty$.

Local Behavior of f-divergences (Csiszár, 1967)

If $f \in \mathcal{C}$ is strictly convex at 1, then $\exists\ \psi_f : [0,\infty) \to [0,\infty)$ such that

- $\lim_{x \downarrow 0} \psi_f(x) = 0$;
- $|P - Q| \leq \psi_f\big(D_f(P\|Q)\big)$.

Local Behavior of f-divergences (Csiszár, 1967)

If $f \in \mathcal{C}$ is strictly convex at 1, then $\exists\, \psi_f : [0, \infty) \to [0, \infty)$ such that

- $\lim_{x \downarrow 0} \psi_f(x) = 0$;
- $|P - Q| \le \psi_f\big(D_f(P \| Q)\big)$.

Corollary

If $f \in \mathcal{C}$ is strictly convex at 1, then

$$\lim_{n \to \infty} D_f(P_n \| Q_n) = 0 \;\Rightarrow\; \lim_{n \to \infty} |P_n - Q_n| = 0. \tag{22}$$

Special case:

Convergence to 0 in relative entropy $\Longrightarrow$ Convergence to 0 in TV distance.

Local Behavior of f-divergences (Pardo and Vajda, 2003)

Let

- $\{P_n\}$ be a sequence of probability measures on $(\mathcal{A}, \mathscr{F})$;
- the sequence $\{P_n\}$ converge to a prob. measure Q in the sense that

$$\lim_{n\to\infty} \operatorname{ess\,sup} \frac{\mathrm{d}P_n}{\mathrm{d}Q}(Y) = 1, \quad Y \sim Q \tag{23}$$

  where $P_n \ll Q$ for all sufficiently large n.
- f be convex on $(0, \infty)$, and f'' be continuous at 1.

Then,

$$\lim_{n\to\infty} \frac{D_f(P_n \| Q)}{\chi^2(P_n \| Q)} = \tfrac{1}{2} f''(1), \quad \lim_{n\to\infty} \chi^2(P_n \| Q) = 0. \tag{24}$$

M. Pardo and I. Vajda, "On asymptotic properties of information-theoretic divergences," *IEEE T-IT*, vol. 49, pp. 1860–1868, July 2003.

Local Behavior: Example

$$P_n \to Q \quad \Longrightarrow \quad \lim_{n\to\infty} \frac{D(P_n\|Q)}{\chi^2(P_n\|Q)} = \tfrac{1}{2}\log e. \tag{25}$$

Proof

Let

$$f(t) = t\log t.$$

Then,

$$D_f(P_n\|Q) = D(P_n\|Q), \quad f''(1) = \log e.$$

Local Behavior of Relative Entropy (Csiszár, 1975)

In one of his famous papers, Csiszár proved that

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \, D(\lambda P + (1-\lambda) Q \,\|\, Q) = 0. \tag{26}$$

Reference

I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.

Local Behavior of f-divergences (I.S., 2018)

As a continuation to Csiszár's result (1975), we strengthened it as follows: Let

- P and Q be prob. measures on $(\mathcal{A}, \mathscr{F})$, and suppose that

$$\operatorname{ess\,sup} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y) < \infty, \quad Y \sim Q; \tag{27}$$

- $f \in \mathcal{C}$, and its second derivative is continuous at 1.

Then,

$$\lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_f(\lambda P + (1-\lambda)Q \,\|\, Q) = \lim_{\lambda \downarrow 0} \frac{1}{\lambda^2} D_f(Q \,\|\, \lambda P + (1-\lambda)Q) \tag{28}$$

$$= \tfrac{1}{2} f''(1)\, \chi^2(P \| Q). \tag{29}$$

I. Sason, "On f-Divergences: integral representations, local behavior, and inequalities," *Entropy*, May 2018.

$D(P\|Q)$ and $\chi^2(P\|Q)$ (I.S.)

Let P and Q be probability measures. Then,

$$\frac{1}{\log e}\, D(P\|Q) = \int_0^1 \chi^2(P \,\|\, (1-\lambda)P + \lambda Q)\, \frac{\mathrm{d}\lambda}{\lambda}, \tag{30}$$

$$\tfrac{1}{2}\, \chi^2(Q\|P) = \int_0^1 \chi^2((1-\lambda)P + \lambda Q \,\|\, P)\, \frac{\mathrm{d}\lambda}{\lambda}. \tag{31}$$

Csiszár-Kemperman-Kullback-Pinsker inequality (∼1967)

$$D(P\|Q) \geq \frac{1}{2}\,|P - Q|^2\,\log e.$$

A Simple Proof (I.S.)

- By invoking the Cauchy-Schwartz, it readily follows that

$$\chi^2(P\|Q) \geq |P-Q|^2. \tag{32}$$

- Using (32), we get

$$\begin{aligned}
\tfrac{1}{\log e}\, D(P\|Q) &= \int_0^1 \chi^2(P \,\|\, (1-\lambda)P + \lambda Q)\, \frac{\mathrm{d}\lambda}{\lambda} \\
&\geq \int_0^1 \underbrace{\bigl|P - \bigl((1-\lambda)P + \lambda Q\bigr)\bigr|^2}_{=\lambda^2\,|P-Q|^2} \frac{\mathrm{d}\lambda}{\lambda} \\
&= |P-Q|^2 \int_0^1 \lambda\, \mathrm{d}\lambda \\
&= \tfrac{1}{2}\, |P-Q|^2.
\end{aligned}$$

Other Simple Proofs

Apart of the Csiszár-Kemperman-Kullback-Pinsker inequality, the identity

$$\frac{1}{\log e}\, D(P\|Q) = \int_0^1 \chi^2(P\,\|\,(1-\lambda)P + \lambda Q)\, \frac{\mathrm{d}\lambda}{\lambda}$$

enables us to prove several (new and old) f-divergence inequalities:

$$D(P\|Q) \leq \tfrac{1}{3}\,\chi^2(P\|Q) + \tfrac{1}{6}\,\chi^2(Q\|P), \tag{33}$$

$$D(P\|Q) \leq \tfrac{1}{2}\,\chi^2(P\|Q) + \tfrac{1}{4}\,|P-Q|. \tag{34}$$

Pinsker-Type Inequalities for f-divergences

Theorem (Csiszár, 1966)

Let $f \in \mathcal{C}$ be twice differentiable, and let $r_0 \in (0, 1)$ and $a > 0$ satisfy

$$f''(u) \geq a > 0, \quad \forall\, u \in (1 - r_0, 1 + r_0). \tag{35}$$

Let $\delta \leq r_0^2$. Then,

$$D_f(P\|Q) \leq \delta \quad \Longrightarrow \quad |P - Q| \leq c\sqrt{\delta}, \quad (c := \tfrac{2}{a} + 1). \tag{36}$$

$f(t) = t \ln t$ for $t > 0$, and $r_0 = \frac{1}{2}$, $a = \frac{2}{3}$ $\Rightarrow |P - Q| \leq 4\sqrt{D(P\|Q)}$ nats.

1-to-1 correspondence $\mathcal{I}_\alpha \leftrightarrow D_\alpha \implies$ extendable to the Rényi divergence.

I. Csiszár, "A note on Jensen's inequality,' *Studia Scient. Math. Hungarica*, 1966.

Csiszár-Kemperman-Kullback-Pinsker Inequality

$$\inf_{P \neq Q} \frac{D(P\|Q)}{|P-Q|^2} = \frac{1}{2} \log e \quad \Longrightarrow \quad D(P\|Q) \geq \frac{1}{2} |P-Q|^2 \log e. \tag{37}$$

Csiszár-Kemperman-Kullback-Pinsker Inequality

$$\inf_{P \neq Q} \frac{D(P\|Q)}{|P-Q|^2} = \tfrac{1}{2} \log e \quad \Longrightarrow \quad D(P\|Q) \geq \tfrac{1}{2}\, |P-Q|^2 \, \log e. \qquad (37)$$

Question

Is there a reverse Pinsker inequality which provides an upper bound on the relative entropy as a function of the TV distance ?

Csiszár-Kemperman-Kullback-Pinsker Inequality

$$\inf_{P \neq Q} \frac{D(P\|Q)}{|P-Q|^2} = \tfrac{1}{2} \log e \quad \Longrightarrow \quad D(P\|Q) \geq \tfrac{1}{2} \, |P-Q|^2 \, \log e. \qquad (37)$$

Question

Is there a reverse Pinsker inequality which provides an upper bound on the relative entropy as a function of the TV distance ?

No, for every $\varepsilon > 0$, $\exists \; (P, Q)$ s.t. $|P - Q| \leq \varepsilon, \; D(P\|Q) = \infty$. ☹

Csiszár-Kemperman-Kullback-Pinsker Inequality

$$\inf_{P \neq Q} \frac{D(P\|Q)}{|P-Q|^2} = \tfrac{1}{2}\log e \quad \Longrightarrow \quad D(P\|Q) \geq \tfrac{1}{2}\,|P-Q|^2\,\log e. \qquad (37)$$

Question

Is there a reverse Pinsker inequality which provides an upper bound on the relative entropy as a function of the TV distance ?

No, for every $\varepsilon > 0$, $\exists\ (P, Q)$ s.t. $|P - Q| \leq \varepsilon,\ D(P\|Q) = \infty$. ☹

However, we can obtain a reverse Pinsker inequality when the relative information is bounded. ☺

Reverse Pinsker Inequality: Finite Alphabet (Csiszár & Talata, 2006)

If $\mathcal{A}$ is a finite set, P and Q are probability measures defined on $\mathcal{A}$, and $Q_{\min} := \min_{x \in \mathcal{A}} Q(x) > 0$, then

$$D(P\|Q) \leq \frac{\log e}{Q_{\min}} \cdot |P - Q|^2. \tag{38}$$

Recent Applications of (38)

- I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE T-IT*, Mar. 2006.
- V. Kostina and S. Verdú, "Channels with cost constraints: strong converse and dispersion," *IEEE T-IT*, May 2015.
- K. Marton, *Distance-divergence inequalities*: rate of decrease of divergence (from stationary distribution) for Gibbs samplers, ISIT 2013 Shannon lecture, July 2013.
- M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE T-IT*, Nov. 2013.

β_1 and β_2

Given a pair of probability measures (P, Q) on the same measurable space, denote $\beta_1, \beta_2 \in [0, 1]$ by

$$\beta_1 = \exp\big(-\operatorname{ess\,sup}\ \imath_{P\|Q}(Y)\big), \tag{39}$$

$$\beta_2 = \operatorname{ess\,inf}\ \exp\big(\imath_{P\|Q}(Y)\big) \tag{40}$$

with $Y \sim Q$.

A Reverse Pinsker Inequality (I.S. and S. Verdú, 2016)

If $\beta_1 \in (0,1)$ and $\beta_2 \in [0,1)$, then,

$$D(P\|Q) \leq \tfrac{1}{2} \left(\varphi(\beta_1^{-1}) - \varphi(\beta_2)\right) |P - Q| \qquad (41)$$

where $\varphi\colon [0,\infty) \mapsto [0,\infty)$ is given by

$$\varphi(t) = \begin{cases} 0 & t = 0 \\ \dfrac{t \log t}{t-1} & t \in (0,1) \cup (1,\infty) \\ \log e & t = 1. \end{cases} \qquad (42)$$

A Reverse Pinsker Inequality (I.S. and S. Verdú, 2016)

If $\beta_1 \in (0,1)$ and $\beta_2 \in [0,1)$, then,

$$D(P\|Q) \le \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2)\right)|P - Q| \tag{41}$$

where $\varphi\colon [0,\infty) \mapsto [0,\infty)$ is given by

$$\varphi(t) = \begin{cases} 0 & t = 0 \\ \dfrac{t \log t}{t-1} & t \in (0,1) \cup (1,\infty) \\ \log e & t = 1. \end{cases} \tag{42}$$

Generalized to Rényi divergences of order $\alpha \in (0,\infty)$.

I.S. and S. Verdú, "f-divergence inequalities," *IEEE T-IT*, Nov. 2016.

A Reverse Pinsker Inequality (Binette, 2018)

For fixed $\delta \in [0, 2]$, $\beta_1, \beta_2 \in [0, 1]$, let $\mathcal{D}(\delta, \beta_1, \beta_2)$ denote the set of all probability measures P and Q with

$$|P - Q| = \delta, \tag{43}$$

$$\beta_1 = \exp\bigl(-\operatorname{ess\,sup}\, \imath_{P\|Q}(Y)\bigr), \tag{44}$$

$$\beta_2 = \operatorname{ess\,inf}\, \exp\bigl(\imath_{P\|Q}(Y)\bigr) \tag{45}$$

where $Y \sim Q$. Then,

$$\max_{(P,Q)\in\mathcal{D}(\delta,\beta_1,\beta_2)} D_f(P\|Q) = \tfrac{1}{2}\,\delta\, \left(\frac{f(\beta_1^{-1})}{\beta_1^{-1} - 1} + \frac{f(\beta_2)}{1 - \beta_2}\right) \tag{46}$$

and the maximum is attained by P and Q defined on a set of size 3. Specialized to the result for relative entropy with a similar proof's concept.

O. Binette, "Note on reverse Pinsker inequalities," May 15, 2018. [Online]. Available at https://arxiv.org/abs/1805.05135.

f-Informativity (Csiszár 1972)

I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematicarum Hungarica*, vol. 2, pp. 191–213, Mar. 1972.

f-informativity measures generalize Shannon's mutual information, and Gallager's function E_0 in the random coding error exponent.

f-Informativity (Cont.)

Let

- $f \in \mathcal{C}$;
- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability measures defined on $\mathcal{X}$;
- w be a probability measure defined on Θ.

The f-informativity, $I_f(w, \mathcal{P})$, is defined as

$$I_f(w, \mathcal{P}) := \inf_Q \int_\Theta D_f(P_\theta \| Q) \, \mathrm{d}w(\theta) \tag{47}$$

where the infimum is taken over all probability measures Q on $\mathcal{X}$.

Special Case of f-informativity: Mutual Information

Let $f(t) = t \log t$ for $t > 0$, then $D_f(\cdot\|\cdot) = D(\cdot\|\cdot)$, and

$$\begin{aligned} I_f(w, \mathcal{P}) &:= \inf_Q \int_\Theta D(P_\theta \| Q) \, \mathrm{d}w(\theta) \\ &= \int_\Theta D(P_\theta \| Q^*) \, \mathrm{d}w(\theta) \end{aligned} \tag{48}$$

with

$$Q^*(x) := \int_\Theta P_\theta(x) \, \mathrm{d}w(\theta), \quad \forall x \in \mathcal{X}. \tag{49}$$

This follows from the identity by Topsøe (Stud. Sci. Math. Hung., 1967):

$$\int_\Theta D(P_\theta \| Q) \, \mathrm{d}w(\theta) = \int_\Theta D(P_\theta \| Q^*) \, \mathrm{d}w(\theta) + D(Q^* \| Q). \tag{50}$$

Hence, the f-informativity is specialized to the mutual information:

$$I_f(w, \mathcal{P}) = I(X; \theta). \tag{51}$$

Properties

In the Bayesian case, f-informativities share several useful properties of the mutual information, such as the data processing inequality.

Absolute f-Informativity (f-Radius)

$$\rho_f(\mathcal{P}) = \inf_Q \sup_{\theta \in \Theta} D_f(P_\theta \| Q). \tag{52}$$

Hence,

$$0 \leq I_f(w, \mathcal{P}) \leq \rho_f(\mathcal{P}), \tag{53}$$

so, the non-negative f-informativity is upper bounded by the f-radius.

Absolute f-Informativity (f-Radius)

$$\rho_f(\mathcal{P}) = \inf_Q \sup_{\theta \in \Theta} D_f(P_\theta \| Q). \tag{52}$$

Hence,

$$0 \leq I_f(w, \mathcal{P}) \leq \rho_f(\mathcal{P}), \tag{53}$$

so, the non-negative f-informativity is upper bounded by the f-radius.

- For observation channels without prior probabilities, f-informativities have the geometric interpretation of a radius.
- In view of the redundancy-capacity theorem, the f-radius is a generalization of the channel capacity (let $f(t) = t \log t$ for $t > 0$).

Parameter Estimation: Basic Model

- The estimand is an unknown parameter θ;
- Θ is the parameter space for θ;
- $\mathcal{X}$ is the sample space for the observed data X;
- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is the model for the data X conditioned on θ;
- $\mathcal{A}$ is the action space for the estimation $\hat{\theta}$, based on the data X;
- $L\colon \Theta \times \mathcal{A} \mapsto [0, \infty)$ is a loss function for estimating θ by $\hat{\theta}$.

Estimator

Let $T\colon \mathcal{X} \mapsto \mathcal{A}$ be an arbitrary mapping where $\hat{\theta} = T(x)$ for $x \in \mathcal{X}$.

Estimator

Let $T\colon \mathcal{X} \mapsto \mathcal{A}$ be an arbitrary mapping where $\hat{\theta} = T(x)$ for $x \in \mathcal{X}$.

Risks

- Minimax risk: expected loss for the best estimator & worst prior

$$R_{\text{minimax}}(L;\Theta) := \inf_{T\colon \mathcal{X}\mapsto\mathcal{A}} \sup_{\theta\in\Theta} \mathbb{E}\big[L\big(\theta, T(X)\big)\big] \tag{54}$$

  where the expectation is taken over $X \sim P_\theta$.

- Bayes risk: expected loss for the best estimator with a prior w on Θ

$$R_{\text{Bayes}}(w, L;\Theta) := \inf_{T\colon \mathcal{X}\mapsto\mathcal{A}} \int_\Theta \mathbb{E}\big[L\big(\theta, T(X)\big)\big]\, \mathrm{d}w(\theta). \tag{55}$$

Estimator

Let $T\colon \mathcal{X} \mapsto \mathcal{A}$ be an arbitrary mapping where $\hat{\theta} = T(x)$ for $x \in \mathcal{X}$.

Risks

- Minimax risk: expected loss for the best estimator & worst prior

$$R_{\text{minimax}}(L;\Theta) := \inf_{T\colon \mathcal{X}\mapsto\mathcal{A}} \sup_{\theta\in\Theta} \mathbb{E}\big[L\big(\theta, T(X)\big)\big] \tag{54}$$

where the expectation is taken over $X \sim P_\theta$.

- Bayes risk: expected loss for the best estimator with a prior w on Θ

$$R_{\text{Bayes}}(w, L;\Theta) := \inf_{T\colon \mathcal{X}\mapsto\mathcal{A}} \int_\Theta \mathbb{E}\big[L\big(\theta, T(X)\big)\big]\, \mathrm{d}w(\theta). \tag{55}$$

$$\Longrightarrow\ R_{\text{minimax}}(L;\Theta) \geq R_{\text{Bayes}}(w, L;\Theta) \qquad \forall \text{ prior } w.$$

Bayes Risk

- If the prior distribution w is known, then the Bayes estimator attains the Bayes risk; ☺
- In general, however, the Bayes estimator is computationally hard to evaluate $\Longrightarrow$ Bayes risk has, often, no closed-form expression. ☹

Bayes Risk Lower Bound

A lower bound on the Bayes risk

- characterizes the fundamental limit of any estimator given the prior knowledge;
- serves as a lower bound on the minimax risk (for the worst prior).

Bayes Risk Lower Bounds (Cont.)

Approach

- Derivation of Bayes risk lower bounds relies heavily on the data processing inequality for f-divergences.
- First derived for $0 - 1$ loss functions, and then extended to an arbitrary non-negative loss function.

Reference

X. Chen, A. Guntuboyina, and Y. Zhang, "On Bayes risk lower bounds," *Journal of Machine Learning Research*, vol. 17, pp. 1–58, Dec. 2016.

Notation

Let $f \in \mathcal{C}$, and define the function $\phi_f \colon [0,1]^2 \mapsto \mathbb{R}$ as follows:

$$\phi_f(a,b) := \begin{cases} bf\left(\frac{a}{b}\right) + (1-b)f\left(\frac{1-a}{1-b}\right), & (a,b) \in [0,1] \times (0,1) \\ af^*(0) + f(1-a), & (a,b) \in [0,1] \times \{0\} \\ f(a) + (1-a)f^*(0), & (a,b) \in [0,1] \times \{1\}. \end{cases}$$

Bayes Risk Lower Bounds for Arbitrary Non-Negative Loss Functions

For arbitrary

- $f \in \mathcal{C}$;
- upper bound $\overline{I}_f$ on the f-informativity $I_f(w, \mathcal{P})$,

let $u_f \colon [0, \infty) \mapsto \left[\frac{1}{2}, 1\right]$ be the monotonically non-decreasing function:

$$u_f(x) := \inf \left\{ \tfrac{1}{2} \leq b \leq 1 : \phi_f\left(\tfrac{1}{2}, b\right) > x \right\}, \quad x \geq 0 \tag{56}$$

and if $\phi_f\left(\frac{1}{2}, b\right) \leq x$ for every $b \in \left[\frac{1}{2}, 1\right]$, then $u_f(x) := 1$. Then,

$$R_{\text{Bayes}}(w, L; \Theta) \geq \tfrac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\big(B_t(a, L)\big) < 1 - u_f\left(\overline{I}_f\right) \right\} \tag{57}$$

where, for $a \in \mathcal{A}$ and $t > 0$,

$$B_t(a, L) := \Big\{ \theta \in \Theta : L(\theta, a) < t \Big\}. \tag{58}$$

Bayes Risk Lower Bounds (Cont., Chen *et al.*, 2016)

Specialization to specific f-divergences yields the following lower bounds:

- Relative entropy ($f(t) = t \log t$ for $t > 0$):

$$R_{\mathsf{Bayes}}(w, L; \Theta) \geq \tfrac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\big(B_t(a, L)\big) < \tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - \exp\big(-2\bar{I}_{\mathsf{KL}}\big)} \right\} \quad (59)$$

- χ^2 divergence ($f(t) = t^2 - 1$ for $t > 0$):

$$R_{\mathsf{Bayes}}(w, L; \Theta) \geq \tfrac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\big(B_t(a, L)\big) < \tfrac{1}{2} - \tfrac{1}{2}\sqrt{\frac{\bar{I}_{\chi^2}}{1 + \bar{I}_{\chi^2}}} \right\}. \quad (60)$$

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;
- Contraction coefficients, and strong data processing inequalities;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;
- Contraction coefficients, and strong data processing inequalities;
- Statistical DeGroot information & important links to f-divergences;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;
- Contraction coefficients, and strong data processing inequalities;
- Statistical DeGroot information & important links to f-divergences;
- Integral & variational representations of f-divergences & applications;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;
- Contraction coefficients, and strong data processing inequalities;
- Statistical DeGroot information & important links to f-divergences;
- Integral & variational representations of f-divergences & applications;
- Sufficiency and ε-sufficiency of observation channels & implications;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;
- Contraction coefficients, and strong data processing inequalities;
- Statistical DeGroot information & important links to f-divergences;
- Integral & variational representations of f-divergences & applications;
- Sufficiency and ε-sufficiency of observation channels & implications;
- Zakai & Ziv's extension of rate-distortion theory with f-divergences;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;
- Contraction coefficients, and strong data processing inequalities;
- Statistical DeGroot information & important links to f-divergences;
- Integral & variational representations of f-divergences & applications;
- Sufficiency and ε-sufficiency of observation channels & implications;
- Zakai & Ziv's extension of rate-distortion theory with f-divergences;
- Asymptotic methods in statistical decision theory with f-divergences;

f-Divergences and f-Informativities: Theory and Applications

- I-divergence (relative entropy), and generalization to f-divergences;
- Mutual information, and generalization by means of f-informativities;
- Risk lower bounds in estimation and learning problems;
- Exact locus of the joint range of f-divergences & tensorization;
- Contraction coefficients, and strong data processing inequalities;
- Statistical DeGroot information & important links to f-divergences;
- Integral & variational representations of f-divergences & applications;
- Sufficiency and ε-sufficiency of observation channels & implications;
- Zakai & Ziv's extension of rate-distortion theory with f-divergences;
- Asymptotic methods in statistical decision theory with f-divergences;
- Robustness of f-divergence based estimators.

f-Divergences and f-Informativities: Theory and Applications

- **I**-divergence (relative entropy), and generalization to f-divergences;
- **M**utual information, and generalization by means of f-informativities;
- **R**isk lower bounds in estimation and learning problems;
- **E**xact locus of the joint range of f-divergences & tensorization;

- **C**ontraction coefficients, and strong data processing inequalities;
- **S**tatistical DeGroot information & important links to f-divergences;
- **I**ntegral & variational representations of f-divergences & applications;
- **S**ufficiency and ε-sufficiency of observation channels & implications;
- **Z**akai & Ziv's extension of rate-distortion theory with f-divergences;
- **A**symptotic methods in statistical decision theory with f-divergences;
- **R**obustness of f-divergence based estimators.

Best wishes ! **Legjobbakat kvánom !**