f-Divergence Inequalities

Igal Sason (Technion) Sergio Verdú (Princeton)

2016 ICSEE International Conference on
the Science of Electrical Engineering

Eilat, Israel
November 16–18, 2016

Motivation for the Talk

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

Motivation for the Talk

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

- Enable proving convergence of prob. measures with various metrics

Motivation for the Talk

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

- Enable proving convergence of prob. measures with various metrics
- Inequalities related to strong data processing and maximal correlation

Motivation for the Talk

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

- Enable proving convergence of prob. measures with various metrics
- Inequalities related to strong data processing and maximal correlation
- Learning and statistics

Motivation for the Talk

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

- Enable proving convergence of prob. measures with various metrics
- Inequalities related to strong data processing and maximal correlation
- Learning and statistics
- Analysis of rates of convergence

Motivation for the Talk

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

- Enable proving convergence of prob. measures with various metrics
- Inequalities related to strong data processing and maximal correlation
- Learning and statistics
- Analysis of rates of convergence
- Transportation-cost inequalities and concentration of measures

Motivation for the Talk **and Location**

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

- **E**nable proving convergence of prob. measures with various metrics
- **I**nequalities related to strong data processing and maximal correlation
- **L**earning and statistics
- **A**nalysis of rates of convergence
- **T**ransportation-cost inequalities and concentration of measures

Motivation for the Talk **and Location**

Many metrics fall under the paradigm of an f-divergence.

$\Rightarrow$ bounds among f-divergences are useful in many instances:

- **I**nequalities related to strong data processing and maximal correlation
- **C**oncentration of measures and transportation-cost inequalities
- **S**tatistics and learning
- **E**nable proving convergence of prob. measures with various metrics
- **E**valuation of rates of convergence

Outline of this work

- Developing systematic approaches to derive f-divergence inequalities, dealing with probability measures on arbitrary alphabets.

Outline of this work

- Developing systematic approaches to derive f-divergence inequalities, dealing with probability measures on arbitrary alphabets.
- Functional domination is one such approach; we find the best constants upper/lower bounding a ratio of f-divergences.

Outline of this work

- Developing systematic approaches to derive f-divergence inequalities, dealing with probability measures on arbitrary alphabets.
- Functional domination is one such approach; we find the best constants upper/lower bounding a ratio of f-divergences.
- Another 2 approaches rely on moment inequalities and log-convexity, and on a derivation of a strengthened Jensen's inequality.

Outline of this work

- Developing systematic approaches to derive f-divergence inequalities, dealing with probability measures on arbitrary alphabets.
- Functional domination is one such approach; we find the best constants upper/lower bounding a ratio of f-divergences.
- Another 2 approaches rely on moment inequalities and log-convexity, and on a derivation of a strengthened Jensen's inequality.
- Special attention is devoted to the total variation distance and its relation to the relative information and relative entropy, including "reverse Pinsker inequalities".

Outline of this work

- Developing systematic approaches to derive f-divergence inequalities, dealing with probability measures on arbitrary alphabets.
- Functional domination is one such approach; we find the best constants upper/lower bounding a ratio of f-divergences.
- Another 2 approaches rely on moment inequalities and log-convexity, and on a derivation of a strengthened Jensen's inequality.
- Special attention is devoted to the total variation distance and its relation to the relative information and relative entropy, including "reverse Pinsker inequalities".
- Derivation of an inequality linking the relative entropy and relative information spectrum.

Outline of this work

- Developing systematic approaches to derive f-divergence inequalities, dealing with probability measures on arbitrary alphabets.
- Functional domination is one such approach; we find the best constants upper/lower bounding a ratio of f-divergences.
- Another 2 approaches rely on moment inequalities and log-convexity, and on a derivation of a strengthened Jensen's inequality.
- Special attention is devoted to the total variation distance and its relation to the relative information and relative entropy, including "reverse Pinsker inequalities".
- Derivation of an inequality linking the relative entropy and relative information spectrum.
- Derivation of integral expressions of the Rényi divergence in terms of the relative information spectrum, leading to bounds on the Rényi divergence in terms of the variational distance or relative entropy.

Journal Paper

I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. on Information Theory*, vol. 62, no. 11, pp. 5973–6006, **November 2016**.

f-Divergence

Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function, and let $P \ll Q$. The *f-divergence* from P to Q is given by

$$D_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q. \tag{1}$$

If $P, Q \ll \mu$, $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$, then

$$D_f(P\|Q) = \int q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu. \tag{2}$$

Basic property

If $f\colon (0,\infty) \to \mathbb{R}$ is convex and $f(1) = 0$, $P \ll Q$, then

$$D_f(P\|Q) \geq 0. \tag{3}$$

If, furthermore, f is strictly convex at $t = 1$, then equality in (3) holds if and only if $P = Q$.

Examples of f-divergences

- Relative entropy

$$D(P\|Q) = D_r(P\|Q) \tag{4}$$

where

$$r(t) = t\log t + (1-t)\log e, \quad t > 0.$$

Examples of f-divergences

- Relative entropy

$$D(P\|Q) = D_r(P\|Q) \tag{4}$$

where

$$r(t) = t\log t + (1-t)\log e, \quad t > 0.$$

- χ^2-divergence: $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$,

$$\chi^2(P\|Q) = D_f(P\|Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q. \tag{5}$$

Examples of f-divergences

- Relative entropy

$$D(P\|Q) = D_r(P\|Q) \tag{4}$$

where

$$r(t) = t\log t + (1-t)\log e, \quad t > 0.$$

- χ^2-divergence: $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$,

$$\chi^2(P\|Q) = D_f(P\|Q) = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q. \tag{5}$$

- Total variation (TV) distance: Setting $f(t) = |t-1|$ results in

$$|P - Q| = D_f(P\|Q) \tag{6}$$

$$= 2 \sup_{\mathcal{F}\in\mathscr{F}} \big(P(\mathcal{F}) - Q(\mathcal{F})\big). \tag{7}$$

Examples of f-divergences (cont.)

- Marton's divergence:

$$d_2^2(P,Q) = \min \mathbb{E}\left[\mathbb{P}^2[X \neq Y \,|\, Y]\right] \tag{8}$$
$$= D_s(P\|Q) \tag{9}$$

where the minimum is over all probability measures P_{XY} with respective marginals $P_X = P$ and $P_Y = Q$, and

$$s(t) = (t-1)^2\, 1\{t < 1\}. \tag{10}$$

1) Marton's divergence satisfies the triangle inequality (Marton '96);
2) $d_2(P,Q) = 0$ implies $P = Q$;
3) however, due to its asymmetry, it is not a distance measure.

Theorem 1: Functional Domination

Let $P \ll Q$, and assume

- f and g are convex on $(0, \infty)$ with $f(1) = g(1) = 0$;
- $g(t) > 0$ for all $t \in (0, 1) \cup (1, \infty)$.

Denote the function $\kappa\colon (0, 1) \cup (1, \infty) \to \mathbb{R}$

$$\kappa(t) = \frac{f(t)}{g(t)}, \quad t \in (0, 1) \cup (1, \infty) \tag{11}$$

and

$$\bar{\kappa} = \sup_{t \in (0,1) \cup (1,\infty)} \kappa(t). \tag{12}$$

Theorem 1 (cont.)

Then,

a)

$$D_f(P\|Q) \leq \bar{\kappa}\, D_g(P\|Q). \tag{13}$$

b) If, in addition, $f'(1) = g'(1) = 0$, then

$$\sup_{P \neq Q} \frac{D_f(P\|Q)}{D_g(P\|Q)} = \bar{\kappa}. \tag{14}$$

Theorem 2 - Samson's inequality (2000)

If $P \ll Q$, then

$$d_2^2(P,Q) + d_2^2(Q,P) \leq \frac{2}{\log e}\, D(P\|Q) \tag{15}$$

Theorem 1 $\Rightarrow$ Theorem 2 (an alternative proof of Samson's inequality).

Theorem 2 - Samson's inequality (2000)

If $P \ll Q$, then

$$d_2^2(P,Q) + d_2^2(Q,P) \leq \frac{2}{\log e}\, D(P\|Q) \tag{15}$$

Theorem 1 $\Rightarrow$ Theorem 2 (an alternative proof of Samson's inequality).

concentration of measure

Samson's inequality strengthens the Pinsker-type inequality in Marton '96:

$$d_2^2(P,Q) \leq \frac{2}{\log e}\, \min\big\{D(P\|Q),\, D(Q\|P)\big\}, \tag{16}$$

useful for proving Marton's conditional transportation inequality.

Definition of 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

$$\beta_1 = \exp\big(-D_\infty(P\|Q)\big), \tag{17}$$

$$\beta_2 = \exp\big(-D_\infty(Q\|P)\big) \tag{18}$$

with the convention that if $D_\infty(P\|Q) = \infty$, then $\beta_1 = 0$, and if $D_\infty(Q\|P) = \infty$, then $\beta_2 = 0$.

Definition of 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

$$\beta_1 = \exp\big(-D_\infty(P\|Q)\big), \tag{17}$$

$$\beta_2 = \exp\big(-D_\infty(Q\|P)\big) \tag{18}$$

with the convention that if $D_\infty(P\|Q) = \infty$, then $\beta_1 = 0$, and if $D_\infty(Q\|P) = \infty$, then $\beta_2 = 0$.

- if $\beta_1 > 0$, then $P \ll Q$, while $\beta_2 > 0$ implies $Q \ll P$.
- if $P \lll\ggg Q$, then with $Y \sim Q$,

$$\beta_1 = \operatorname{ess\,inf} \frac{\mathrm{d}Q}{\mathrm{d}P}(Y) = \left(\operatorname{ess\,sup} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\right)^{-1}, \tag{19}$$

$$\beta_2 = \operatorname{ess\,inf} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y) = \left(\operatorname{ess\,sup} \frac{\mathrm{d}Q}{\mathrm{d}P}(Y)\right)^{-1}. \tag{20}$$

Since $\beta_1 = 1 \Leftrightarrow \beta_2 = 1 \Leftrightarrow P = Q$, it is advisable to avoid trivialities by excluding that case.

Theorem 3: Bounded Relative Information

Let f and g satisfy the assumptions in Theorem 1, and assume that $(\beta_1, \beta_2) \in [0, 1)^2$. Then,

$$D_f(P\|Q) \leq \kappa^\star \, D_g(P\|Q) \tag{21}$$

where

$$\kappa^\star = \sup_{\beta \in (\beta_2, 1) \cup (1, \beta_1^{-1})} \kappa(\beta) \tag{22}$$

and $\kappa(\cdot)$ is defined in Theorem 1.

Application — Theorem 4: Reverse Samson's inequality

Let $(\beta_1, \beta_2) \in (0,1)^2$. Then,

$$\inf \frac{d_2^2(P,Q) + d_2^2(Q,P)}{D(P\|Q)} = \min\left\{\kappa(\beta_1^{-1}),\, \kappa(\beta_2)\right\} \tag{23}$$

where the infimum is over all $P \ll Q$ with given (β_1, β_2), and where $\kappa\colon (0,1) \cup (1,\infty) \to \left(0, \frac{2}{\log e}\right)$ is defined as

$$\kappa(t) = \frac{(t-1)^2}{r(t)\, \max\{1,t\}}, \quad t \in (0,1) \cup (1,\infty) \tag{24}$$

$$\lim_{t\to 1} \kappa(t) = \frac{2}{\log e} = \bar{\kappa}, \tag{25}$$

$\kappa(\cdot)$ is monotonically increasing on $(0,1)$, and decreasing on $(1,\infty)$.

Application – Theorem 5: Reverse Pinsker's inequality

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

$$D(P\|Q) \leq \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2)\right)\,|P-Q| \tag{26}$$

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

$$\varphi(t) = \begin{cases} 0 & t = 0 \\ \frac{t\log t}{t-1} & t \in (0,1)\cup(1,\infty) \\ \log e & t = 1. \end{cases} \tag{27}$$

Application – Theorem 5: Reverse Pinsker's inequality

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

$$D(P\|Q) \leq \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2)\right)\,|P-Q| \tag{26}$$

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

$$\varphi(t) = \begin{cases} 0 & t=0 \\ \frac{t\log t}{t-1} & t \in (0,1)\cup(1,\infty) \\ \log e & t=1. \end{cases} \tag{27}$$

More on Theorem 5 and Pinsker's inequality

1) Improves an earlier bound by Verdú (ITA '14);
2) Generalized in our work to Rényi divergence of order $\alpha \in (0,\infty)$.
3) Pinsker's inequality is extended to E_γ divergence, generalizing the TV distance $\Rightarrow$ linking relative entropy & relative information spectrum.

Theorem 6: Local Behavior of f-divergences

Suppose that $\{P_n\}$, a sequence of probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$, converges to Q in the sense that, for $Y \sim Q$,

$$\lim_{n\to\infty} \operatorname{ess\,sup} \frac{\mathrm{d}P_n}{\mathrm{d}Q}(Y) = 1 \tag{28}$$

where $P_n \ll Q$ for all sufficiently large n. If f and g are convex on $(0,\infty)$ and they are positive except at $t = 1$ (where they are 0), then

$$\lim_{n\to\infty} D_f(P_n\|Q) = \lim_{n\to\infty} D_g(P_n\|Q) = 0, \tag{29}$$

$$\min\{\kappa(1^-), \kappa(1^+)\} \le \lim_{n\to\infty} \frac{D_f(P_n\|Q)}{D_g(P_n\|Q)} \le \max\{\kappa(1^-), \kappa(1^+)\} \tag{30}$$

where we have indicated the left and right limits of the function $\kappa(\cdot)$ at 1 by $\kappa(1^-)$ and $\kappa(1^+)$, respectively.

Applications: Local Behavior of f-divergences

Corollary

Let $\{P_n \ll Q\}$ converge to Q in the sense of (28). Then,

$$\lim_{n\to\infty} D(P_n\|Q) = 0, \tag{31}$$

$$\lim_{n\to\infty} D(Q\|P_n) = 0, \tag{32}$$

$$\lim_{n\to\infty} \frac{D(P_n\|Q)}{D(Q\|P_n)} = 1, \tag{33}$$

$$\lim_{n\to\infty} \frac{D(P_n\|Q)}{\chi^2(P_n\|Q)} = \frac{1}{2}\log e. \tag{34}$$

Note that (34) is known in the finite alphabet case.

Concluding Remarks

- **Thank you all for coming to this early talk today !**
- I focused here on f-divergence inequalities via functional domination. This is only one approach in this work, nevertheless ...
- Those interested are very welcome to read the (just) published paper.

Journal Paper

I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. on Information Theory*, vol. 62, no. 11, pp. 5973–6006, **November 2016**.