On Data-Processing and Majorization Inequalities for $f$-Divergences

Igal Sason
EE Department, Technion - Israel Institute of Technology

IZS 2020
Zurich, Switzerland
February 26-28, 2020
Introduction

$f$-Divergences

$f$-divergences form a general class of divergence measures which are commonly used in information theory, learning theory and related fields.


This Talk is Restricted to the Discrete Setting

- $f : (0, \infty) \mapsto \mathbb{R}$ is a convex function with $f(1) = 0$;
- $P, Q$ are probability mass functions defined on a (finite or countably infinite) set $\mathcal{X}$.

$f$-Divergence: Definition

The $f$-divergence from $P$ to $Q$ is given by

$$D_f(P\|Q) := \sum_{x \in \mathcal{X}} Q(x) f \left( \frac{P(x)}{Q(x)} \right)$$

with the convention that

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

$$0 f \left( \frac{0}{0} \right) := 0, \quad 0 f \left( \frac{a}{0} \right) := \lim_{t \downarrow 0} t f \left( \frac{a}{t} \right) = a \lim_{u \to \infty} \frac{f(u)}{u}, \quad a > 0.$$
$f$-divergences: Examples

- **Relative entropy**
  \[ f(t) = t \log t, \quad t > 0 \implies D_f(P \| Q) = D(P \| Q), \]
  \[ f(t) = -\log t, \quad t > 0 \implies D_f(P \| Q) = D(Q \| P). \]

- **Total variation (TV) distance**
  \[ f(t) = |t - 1|, \quad t \geq 0 \]
  \[ \implies D_f(P \| Q) = |P - Q| := \sum_{x \in \mathcal{X}} |P(x) - Q(x)|. \]

- **Chi-Squared Divergence**
  \[ f(t) = (t - 1)^2, \quad t \geq 0 \]
  \[ \implies D_f(P \| Q) = \chi^2(P \| Q) := \sum_{x \in \mathcal{X}} \frac{(P(x) - Q(x))^2}{Q(x)}. \]
\( f \)-divergences: Examples (cont.)

**\( E_\gamma \) divergence (Polyanskiy, Poor and Verdú, IEEE T-IT, 2010)**

For \( \gamma \geq 1 \),

\[
E_\gamma(P\|Q) := D_{f_\gamma}(P\|Q)
\]

(1)

with \( f_\gamma(t) = (t - \gamma)^{+} \), for \( t > 0 \), and \( x^{+} := \max\{x, 0\} \).

- \( E_1(P\|Q) = \frac{1}{2} |P - Q| \implies E_\gamma \) divergence generalizes TV distance.
- \( E_\gamma(P\|Q) = \max_{\mathcal{E} \in \mathcal{F}} (P(\mathcal{E}) - \gamma Q(\mathcal{E})) \).

**Other Important \( f \)-divergences**

- Triangular Discrimination (Vincze-Le Cam distance ’81; Topsøe 2000);
- Jensen-Shannon divergence (Lin 1991; Topsøe 2000);
- DeGroot statistical information (DeGroot ’62; Liese & Vajda ’06); see later.
- Marton’s divergence (Marton 1996; Samson 2000).
Data-Processing Inequality for $f$-Divergences

Let

- $\mathcal{X}$ and $\mathcal{Y}$ be finite or countably infinite sets;
- $P_X$ and $Q_X$ be probability mass functions that are supported on $\mathcal{X}$;
- $W_{Y|X}: \mathcal{X} \rightarrow \mathcal{Y}$ be a stochastic transformation;
- Output distributions:
  \[
P_Y := P_X W_{Y|X}, \quad Q_Y := Q_X W_{Y|X};
  \]
- $f: (0, \infty) \rightarrow \mathbb{R}$ be a convex function with $f(1) = 0$.

Then,

\[
D_f(P_Y \| Q_Y) \leq D_f(P_X \| Q_X).
\]
Contraction Coefficient for $f$-Divergences

Let

- $Q_X$ be a probability mass function defined on a set $\mathcal{X}$, and which is not a point mass;
- $W_{Y|X}: \mathcal{X} \rightarrow \mathcal{Y}$ be a stochastic transformation.

The contraction coefficient for $f$-divergences is defined as

$$
\mu_f(Q_X, W_{Y|X}) := \sup_{P_X: D_f(P_X \| Q_X) \in (0, \infty)} \frac{D_f(P_Y \| Q_Y)}{D_f(P_X \| Q_X)}.
$$

I. Sason  
IZS 2020, Zurich, Switzerland  
Strong Data Processing Inequalities (SDPI)

If $\mu_f(Q_X, W_{Y|X}) < 1$, then

$$D_f(P_Y \parallel Q_Y) \leq \mu_f(Q_X, W_{Y|X}) D_f(P_X \parallel Q_X).$$

Contraction coefficients for $f$-divergences play a key role in strong data-processing inequalities:

- Ahlswede and Gács ('76);
- Cohen et al. ('93);
- Raginsky ('16);
- Polyanskiy and Wu ('16, '17);
- Makur, Polyanskiy and Wu ('18).
Let
\[
\xi_1 := \inf_{x \in \mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [0, 1], \quad \xi_2 := \sup_{x \in \mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [1, \infty].
\]

\[c_f := c_f(\xi_1, \xi_2) \geq 0\] and \[d_f := d_f(\xi_1, \xi_2) \geq 0\] satisfy
\[
2c_f \leq \frac{f'_+(v) - f'_+(u)}{v - u} \leq 2d_f, \quad \forall u, v \in \mathcal{I}, \, u < v
\]

where \(f'_+\) is the right-side derivative of \(f\), and \(\mathcal{I} := [\xi_1, \xi_2] \cap (0, \infty)\). Then,
\[
d_f \left[ \chi^2(P_X \parallel Q_X) - \chi^2(P_Y \parallel Q_Y) \right]
\geq D_f(P_X \parallel Q_X) - D_f(P_Y \parallel Q_Y)
\geq c_f \left[ \chi^2(P_X \parallel Q_X) - \chi^2(P_Y \parallel Q_Y) \right] \geq 0.
\]
Theorem 1: SDPI (Cont.)

If $f$ is twice differentiable on $\mathcal{I}$, then the best coefficients are given by

$$ c_f = \frac{1}{2} \inf_{t \in \mathcal{I}(\xi_1, \xi_2)} f''(t), \quad d_f = \frac{1}{2} \sup_{t \in \mathcal{I}(\xi_1, \xi_2)} f''(t). $$
Theorem 1: SDPI (Cont.)

If $f$ is twice differentiable on $\mathcal{I}$, then the best coefficients are given by

$$c_f = \frac{1}{2} \inf_{t \in \mathcal{I}(\xi_1, \xi_2)} f''(t), \quad d_f = \frac{1}{2} \sup_{t \in \mathcal{I}(\xi_1, \xi_2)} f''(t).$$

This SDPI is Locally Tight

Let

$$\lim_{n \to \infty} \inf_{x \in \mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1, \quad \lim_{n \to \infty} \sup_{x \in \mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1.$$

If $f$ has a continuous second derivative at unity, then

$$\lim_{n \to \infty} \frac{D_f(P_X^{(n)} \| Q_X) - D_f(P_Y^{(n)} \| Q_Y)}{\chi^2(P_X^{(n)} \| Q_X) - \chi^2(P_Y^{(n)} \| Q_Y)} = \frac{1}{2} f''(1).$$
Advantage: Tensorization of the Chi-Squared Divergence

\[ \chi^2(P_1 \times \ldots \times P_m \parallel Q_1 \times \ldots \times Q_m) = \prod_{i=1}^{m} \left( 1 + \chi^2(P_i \parallel Q_i) \right) - 1. \]
Theorem 2: SDPI for $f$-divergences

Let $f: (0, \infty) \to \mathbb{R}$ satisfy the conditions:

- $f$ is a convex function, differentiable at 1, $f(1) = 0$, and $f(0) := \lim_{t \to 0^+} f(t) < \infty$;
- The function $g: (0, \infty) \to \mathbb{R}$, defined by $g(t) := \frac{f(t) - f(0)}{t}$ for all $t > 0$, is convex.

Let

$$
\kappa(\xi_1, \xi_2) := \sup_{t \in (\xi_1, 1) \cup (1, \xi_2)} \frac{f(t) + f'(1)(1 - t)}{(t - 1)^2}.
$$

Then,

$$
\frac{D_f(P_Y \| Q_Y)}{D_f(P_X \| Q_X)} \leq \frac{\kappa(\xi_1, \xi_2)}{f(0) + f'(1)} \cdot \frac{\chi^2(P_Y \| Q_Y)}{\chi^2(P_X \| Q_X)}.
$$
Numerical Results

The tightness of the bounds (SDPI inequalities) in Theorems 1 and 2 was exemplified numerically for transmission over a BEC and BSC.
**List Decoding**

- Decision rule outputs a list of choices.
- The extension of Fano’s inequality to list decoding, expressed in terms of $H(X|Y)$, was initiated by Ahlswede, Gacs and Körner ('66).
- Useful to prove converse results (jointly with the blowing-up lemma).

---

**Generalized Fano’s Inequality for Fixed List Size**

$$H(X|Y) \leq \log M - d\left(P_L \parallel 1 - \frac{L}{M}\right)$$

where $d(\cdot \parallel \cdot)$ denotes the binary relative entropy:

$$d(x||y) := x \log \left(\frac{x}{y}\right) + (1 - x) \log \left(\frac{1 - x}{1 - y}\right), \quad x, y \in (0, 1).$$
Theorem 3: Tightened Bound by Strong DPI (SDPI)

- Let $P_{XY}$ be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M$.
- Consider a decision rule $L : \mathcal{Y} \rightarrow \binom{\mathcal{X}}{L}$, where $\binom{\mathcal{X}}{L}$ stands for the set of subsets of $\mathcal{X}$ with cardinality $L$, and $L < M$ is fixed.
- Denote the list decoding error probability by $P_L := \mathbb{P}[X \notin \mathcal{L}(Y)]$.

If the $L$ most probable elements from $\mathcal{X}$ are selected, given $Y \in \mathcal{Y}$, then

$$H(X|Y) \leq \log M - d \left( P_L \| 1 - \frac{L}{M} \right) - \frac{\log e}{2} \cdot \frac{\mathbb{E}[P_X|Y(X|Y)] - \frac{1-P_L}{L}}{\sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_X|Y(x|y)}.$$

Proof: Use Theorem 1 (our first SDPI) with $f(t) = t \log t$, $t > 0$, $P_{X|Y=y}$, and $Q_{X|Y=y}$ be equiprobable over $\{1, \ldots, M\}$, $W_{Z|X,Y=y}$ be 1 or 0 if $X \in \mathcal{L}(y)$ or $X \notin \mathcal{L}(y)$, and average over $Y$.

Numerical experimentation exemplifies this improvement.
Generalized Fano’s Inequality for Variable List Size (1975)

- Let $P_{XY}$ be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M$;
- Consider a decision rule $\mathcal{L} : \mathcal{Y} \rightarrow 2^{\mathcal{X}}$;
- Let the (average) list decoding error probability be given by

$$P_{\mathcal{L}} := \mathbb{P}[X \notin \mathcal{L}(Y)]$$

with $|\mathcal{L}(y)| \geq 1$ for all $y \in \mathcal{Y}$.

Then,

$$H(X|Y) \leq h(P_{\mathcal{L}}) + \mathbb{E}[\log |\mathcal{L}(Y)|] + P_{\mathcal{L}} \log M.$$
Theorem: A Consequence of DPI for the $E_\gamma$-Divergence

For every $\gamma \geq 1$,

$$P_L \geq \frac{1 + \gamma}{2} - \frac{\gamma \mathbb{E}[|\mathcal{L}(Y)|]}{M} - \frac{1}{2} \mathbb{E} \left[ \sum_{x \in \mathcal{X}} \left| P_{X|Y}(x|Y) - \frac{\gamma}{M} \right| \right].$$

Conditions for the bound to hold with equality are proved in the paper.
**Simple Example**

- $X, Y$ are RVs getting values in $\mathcal{X} = \{0, 1, 2, 3, 4\}$, $\mathcal{Y} = \{0, 1\}$.
- $P_{XY}$ is their joint probability mass function, given by

\[
\begin{align*}
P_{XY}(0, 0) &= P_{XY}(1, 0) = P_{XY}(2, 0) = \frac{1}{8}, \\
P_{XY}(3, 0) &= P_{XY}(4, 0) = \frac{1}{16}, \\
P_{XY}(0, 1) &= P_{XY}(1, 1) = P_{XY}(2, 1) = \frac{1}{24}, \\
P_{XY}(3, 1) &= P_{XY}(4, 1) = \frac{3}{16}.
\end{align*}
\]

- $\mathcal{L}(0) = \{0, 1, 2\}$ and $\mathcal{L}(1) = \{3, 4\}$ are the lists in $\mathcal{X}$, given $Y \in \mathcal{Y}$.

Then,

- If $\gamma = \frac{5}{4}$, the bound holds with equality and $P_\mathcal{L} = \frac{1}{4}$.
- The generalized Fano’s inequality only gives $P_\mathcal{L} \geq 0.1206$. 
## Summary

- We focus on strong data-processing inequalities for $f$-divergences.
- We exemplify their utility for list decoding error bounds.
- Another application (see paper): Variable-to-fixed Tunstall codes.
- Majorization inequalities and an IT application presented at ITA '20.

## Journal Papers (Related Work)

More on $f$-Divergences and $f$-Informativities

- 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 & strong data processing inequalities**;
- Statistical DeGroot information & important links to $f$-divergences;
- Integral & variational representations of $f$-divergences & applications;
- Sufficiency and $\varepsilon$-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.
More on $f$-Divergences and $f$-Informativities

- **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 & strong data processing inequalities;
- Statistical DeGroot information & important links to $f$-divergences;
- Integral & variational representations of $f$-divergences & applications;
- Sufficiency and $\varepsilon$-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.

Thanks to Imre who introduced these information measures!