Rényi Divergence and Majorization

ISIT 2010, Austin

Tim van Erven & Peter Harremoës

Centrum Wiskunde & Informatica, Amsterdam and Copenhagen Business College

June 14th

Rényi divergences

Let P and Q be probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$, and let p and q be their densities with respect to a common σ-finite dominating measure μ. Then for $\alpha \in R \setminus \{0, 1\}$ the *Rényi divergence D_α of order α* of P from Q is defined as

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} \, \mathrm{d}\mu. \tag{1}$$

Continuity considerations lead to the following extensions:

$$\begin{aligned}
D_{-\infty}(P\|Q) &= \lim_{\alpha \downarrow -\infty} D_\alpha(P\|Q) = \log\left(\mathit{ess}\inf \frac{p}{q}\right), \\
D_0(P\|Q) &= \lim_{\alpha \downarrow 0} D_\alpha(P\|Q) = -\log Q(p > 0), \\
D_1(P\|Q) &= \lim_{\alpha \uparrow 1} D_\alpha(P\|Q) = D(P\|Q), \\
D_\infty(P\|Q) &= \lim_{\alpha \downarrow \infty} D_\alpha(P\|Q) = \log\left(\mathit{ess}\sup \frac{p}{q}\right).
\end{aligned}$$

Basic properties

Monotonicity $\alpha \to D_\alpha(P\|Q)$ is increasing.

Convexity $D_\alpha(P\|Q)$ is convex in P for $\alpha > 0$ and convex in Q for $\alpha \leq 1$. It is joint convex for $\alpha \in [0,1]$.

Additivity $D_\alpha(P_1 \times P_2 \| Q_1 \times Q_2) = D_\alpha(P_1\|Q_1) + D_\alpha(P_2\|Q_2)$

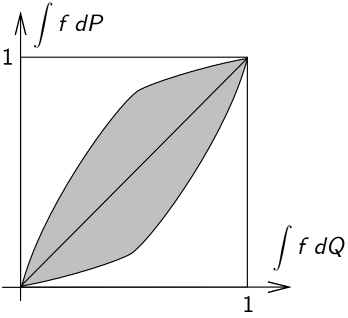
Continuity $(P, Q) \to D_\alpha(P\|Q)$ is lower semi-continuous for $\alpha > 0$ and upper semi-continuous for $\alpha < 0$.

Majorization and Markov ordering

Definition The *Lorenz diagram* of (P, Q) is the range of

$$f \mapsto \left(\int f \, \mathrm{d}P, \int f \, \mathrm{d}Q \right),$$

where f is a measurable function with values in $[0, 1]$. (Harremoës 2004)



If Q is the uniform distribution then the Lorenz diagram of (P_1, Q) is a subset of the Lorenz diagram of (P_2, Q) if and only if P_2 *majorizes* P_1.

Majorization and Markov ordering

Theorem The Lorenz diagram of (P_1, Q) is a subset of the Lorenz diagram of (P_2, Q) if and only if there exists a Markov operator that transforms P_2 into P_1 and leaves Q invariant.
Definition Let P_1, P_2 and Q be measures on the same measurable set $\mathcal{X}$. We write $P_2 \succeq_Q P_1$ if the Lorenz diagram of (P_1, Q) is a subset of the Lorenz diagram of (P_2, Q) . If the Lorenz diagrams of (P_1, Q) and (P_2, Q) are equal, then we write $P_1 \simeq_Q P_2$.
This ordering that generalizes majorization will be called the *Markov ordering* (Gorban, Gorban and Judge, 2010).
Theorem Let Q be a measure on a measurable set $\mathcal{X}$. If Q is a uniform distribution on a finite set or if Q has no atoms, then $M_+^1(\mathcal{X}) / \simeq_Q$ is a lattice, where $M_+^1(\mathcal{X})$ denotes the set of probability measures on $\mathcal{X}$.

The Lorenz curve and statistics

Definition The *Lorenz curve* of (P, Q) is the convex envelope of the Lorenz diagram, i.e. the largest convex function such that all the points in the Lorenz diagram are at or above the curve.

Proposition Let P and Q be measures on the same measurable set $\mathcal{X}$. The Lorenz curve of (P, Q) is the convex envelop of the points $(P(A_t), Q(A_t))$ where A_t are events of the form $A_t = \left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P}{\mathrm{d}Q} \leq t\right\}$.

In statistics the sets $A_t = \left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P}{\mathrm{d}Q} \leq t\right\}$ play the role of acceptance sets related to the likelihood ratio test of ratio t.

Divergence and ordering

Theorem For $\alpha > 0$ the Rényi divergence $D_\alpha(P\|Q)$ is a increasing function of P on the lattice corresponding to Q.

Using the Markov ordering we get more insight.

Theorem If $\alpha \geq 1$ and $D_\alpha(\tilde{P}\|Q) < \infty$, then the Rényi divergence $D_\alpha(P\|Q)$ is continuous in P on the set $\left\{P \mid P \preceq_Q \tilde{P}\right\}$ when the set of probability measures is equipped with the topology of total variation.

Image of Triangle

Theorem Let P_1 and P_2 denote distributions that are absolutely continuous with respect to Q. If Markov ordering is taken with respect to Q then power divergence is sub-modular and super-additive, i.e.

$$d_\alpha(P_1, Q) + d_\alpha(P_2, Q) \geq d_\alpha(P_1 \wedge P_2, Q) + d_\alpha(P_1 \vee P_2, Q)$$
$$d_\alpha(P_1, Q) + d_\alpha(P_2, Q) \leq d_\alpha(P_1 \wedge P_2, Q).$$

Since power divergence is a function of Rényi divergence one can reformulate the theorem in terms of Rényi divergence. Like Rényi divergence, the power divergence $d_\alpha(P, Q)$ tends to the information divergence $D(P\|Q)$ as $\alpha \uparrow 1$. Therefore

Corollary Let P_1 and P_2 be distributions that are absolutely continuous with respect to Q. If the Markov ordering is taken with respect to Q then information divergence is sub-modular and super-additive, i.e.

$$D(P_1\|Q) + D(P_2\|Q) \geq D(P_1 \wedge P_2\|Q) + D(P_1 \vee P_2\|Q)$$
$$D(P_1\|Q) + D(P_2\|Q) \leq D(P_1 \wedge P_2\|Q).$$

Guessing moments

Definition *A guessing function* in $\mathcal{X}$ is a function $g : \mathcal{X} \to \mathbb{R}$ such that $Q\left(\{x \mid g\left(x\right) \leq t\}\right) \leq t$ for $t \in \left[0, 1\right]$.

Definition Let P be a probability measure on $\mathcal{X}$. For each Radon-Nikodým derivative $\frac{\mathrm{d}P}{\mathrm{d}Q}$, the *ranking function* r of $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is given by

$$r\left(x\right) = Q\left(\left\{y \mid \frac{\mathrm{d}P}{\mathrm{d}Q}\left(y\right) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}\left(x\right)\right\}\right).$$

Proposition The ranking function is the guessing function that minimizes the ρ-th moment if $\rho > 0$ and maximizes the ρ-th moment if $\rho < 0$.

Asymptotics

Proposition Assume that P_1, P_2 and Q are probability measures on $\mathcal{X}$ and $P_1 \preceq_Q P_2$. Let r_1 and r_2 denote the ranking functions of P_1 and P_2. Then

$$\|r_1\|_\rho \leq \|r_2\|_\rho \text{ if } \rho > 0,$$
$$\|r_1\|_\rho \geq \|r_2\|_\rho \text{ if } \rho < 0.$$

Theorem If $\alpha = \frac{1}{1+\rho} > 0$ then for any i.i.d. sequence $X_1^n = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$ we have

$$\lim_{n\to\infty} -\frac{1}{n} \log\left(\left\|r\left(X_1^n\right)\right\|_\rho\right) = D_\alpha\left(P\|Q\right).$$

The result gives a new interpretation of Rényi divergence.

Conclusion

- Physical Rény divergence has many nice properties.
- Markov ordering and majorization plays a natural role in the theory.
- The concept of guessing moments generalize to Rényi divergence.
- Gives new interpretation of Rényi divergence.
- May help in understanding why the class of Rényi maximizers is stable under taking marginal distributions.