Limits from l'Hôpital rule: Shannon entropy as limit cases of Rényi and Tsallis entropies

Frank Nielsen
École Polytechnique/Sony Computer Science Laboratorie, Inc
http://www.informationgeometry.org

August 16, 2010

Let us prove that both Rényi extensive R_α and Tsallis non-extensive T_α entropies of order α tend to Shannon entropy when $\alpha \to 1$ using l'Hôpital rule. Following the same principle, we then show that both Rényi and Tsallis information divergences tend to the celebrated Kullback-Leibler divergence when $\alpha \to 1$.

L'Hôpital rule dates back to the 17th century, and states that the limit of the indeterminate ratio of functions equals to the limit of the ratio of their derivatives provided that (i) the limits of both the numerator and denominator coincide (either to $-\infty$, 0, or ∞), and that (ii) the limit of the ratio of the derivatives also exists.

That is, if $\lim_{x\to\alpha} f(x) = \lim_{x\to\alpha} g(x) = 0$ and $\lim_{x\to\alpha} f'(x)/g'(x) = l$ exists, then

$$\lim_{x\to\alpha} \frac{f(x)}{g(x)} = \lim_{x\to\alpha} \frac{f'(x)}{g'(x)} = l.$$

(Although the limits of f and g can alternatively be chosen as $\pm\infty$ instead of 0, we shall use below the theorem to solve 0/0 indeterminate forms.)

1 Rényi entropy

Consider Rényi entropy [2] $R_\alpha(P) = \frac{1}{1-\alpha} \log \sum_{i=1}^n p_i^\alpha$ (for $\alpha > 0$ and $\alpha \neq 1$), and let us prove that $R_\alpha(P)$ tends to Shannon entropy $S(P) = -\sum_{i=1}^n p_i \log p_i$

when $\alpha \to 1$. Set $f(\alpha) = \log \sum_{i=1}^n p_i^\alpha$ (for any fixed distribution P) and $g(\alpha) = 1 - \alpha$. Then $\frac{\mathrm{d}g(\alpha)}{\mathrm{d}\alpha} = -1$ and

$$\frac{\mathrm{d}f(\alpha)}{\mathrm{d}\alpha} = \frac{\sum_{i=1}^n \frac{\mathrm{d}}{\mathrm{d}\alpha}(p_i^\alpha)}{\sum_{i=1}^n p_i^\alpha}$$

after applying the derivative chain rule. Since $\frac{\mathrm{d}}{\mathrm{d}\alpha}(p_i^\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} e^{\alpha \log p_i} = (\log p_i) e^{\alpha \log p_i} = p_i^\alpha \log p_i$, we get

$$\frac{f'(\alpha)}{g'(\alpha)} = -\sum_{i=1}^n p_i^\alpha \log p_i, \text{ and } \lim_{\alpha \to 1} \frac{f'(\alpha)}{g'(\alpha)} = -\sum_{i=1}^n p_i \log p_i.$$

Since $\lim_{\alpha\to 1} f(\alpha) = \lim_{\alpha \to 1} g(\alpha) = 0$ and $\lim_{\alpha\to 1} \frac{f'(\alpha)}{g'(\alpha)} = -\sum_{i=1}^n p_i \log p_i$, we deduce from l'Hôpital rule that $\lim_{\alpha \to 1} R_\alpha(P) = H(P)$. That is, Rényi entropy tends to Shannon entropy as $\alpha \to 1$.

2 Tsallis entropy

Consider now the Tsallis entropy [3] T_α of order α (for $\alpha \in \mathbb{R}$)): $T_\alpha(P) = \frac{1 - \sum_{i=1} p_i^\alpha}{\alpha - 1}$. Let us prove using l'Hôpital rule that $\lim_{\alpha\to 1} T_\alpha(P) = S(p) = -\sum_i p \log p$, Shannon entropy. For a fixed distribution P, write $T_\alpha(P) = \frac{f(\alpha)}{g(\alpha)}$ with $g(\alpha) = \alpha - 1$ and $g'(\alpha) = 1$. We have $f(\alpha) = 1 - \sum_{i=1} p_i^\alpha$ and $f'(\alpha) = -\sum_i p_i^\alpha \log p_i$ since $(p_i^\alpha)' = (e^{\alpha \log p_i})' = p_i^\alpha \log p_i$. Since both limits of $f(\alpha)$ and $g(\alpha)$ tend to 0 as $\alpha \to 1$, we can apply l'Hôpital rule and get $\lim_{\alpha \to 1} T_\alpha(P) = \lim_{\alpha\to 1} \frac{f'(\alpha)}{g'(\alpha)} = -\sum_i p_i \log p_i = H(P)$, namely Shannon entropy.

3 Rényi and Tsallis information divergences

Based on Rényi and Tsallis entropies, we define the corresponding information divergences [4, 1]:

$$\begin{aligned} R_\alpha(P : Q) &= \frac{1}{\alpha - 1} \log \sum_i p_i^\alpha q_i^{1-\alpha}, \alpha > 0 \text{ and } \alpha \neq 0 \\ T_\alpha(P : Q) &= \frac{1}{\alpha - 1} (\sum_i p_i^\alpha q_i^{1-\alpha} - 1), \alpha \in \mathbb{R}^*. \end{aligned}$$

Let us again apply l'Hôpital rule to show that those parametric information divergences tend to the Kullback-Leibler divergence when $\alpha \to 1$. Consider $f(\alpha) = \sum_{i=1} p_i^\alpha q_i^{1-\alpha}$ (for fixed distributions P and Q), using the derivative chain rule we get $f'(\alpha) = \sum_i p_i^\alpha q_i^{1-\alpha} \log p_i - p_i^\alpha q_i^{1-\alpha} \log q_i = \sum_i p_i^\alpha q_i^{1-\alpha} \log \frac{p_i}{q_i}$, using the fact that $(q_i^{1-\alpha})' = (e^{(1-\alpha)\log q_i})' = -\log q_i (e^{(1-\alpha)\log q_i}) = -q_i^{1-\alpha} \log q_i$.

Let $g(\alpha) = \alpha - 1$, and $r(\alpha) = \log f(\alpha)$. We have $g'(\alpha) = 1$ and $r'(\alpha) = \frac{f'(\alpha)}{f(\alpha)} = \frac{\sum_i p_i^\alpha q_i^{1-\alpha} \log \frac{p_i}{q_i}}{\sum_{i=1} p_i^\alpha q_i^{1-\alpha}}$. We are ready to apply l'Hôpital rule since $\lim_{\alpha\to 1} r(x) = \lim_{\alpha \to 1} g(x) = 0$ and $\lim_{\alpha\to 1} r'(x) = \sum_i p_i \log \frac{p_i}{q_i}$ and $\lim_{\alpha \to 1} g'(x) = 1$ to conclude that Rényi information divergence tends to

$$\lim_{\alpha\to 1} R_\alpha(P:Q) = \sum_i p_i \log \frac{p_i}{q_i} = \mathrm{KL}(P:Q),$$

i.e., the Kullback-Leibler divergence, also known as the relative entropy with respect to Shannon entropy.

Similarly, for Tsallis information divergence, we set $t(\alpha) = f(\alpha) - 1 = \sum_i p_i^\alpha q_i^{1-\alpha} - 1$ and get $t'(\alpha) = f'(\alpha) = \sum_i p_i^\alpha q_i^{1-\alpha} \log \frac{p_i}{q_i}$. We apply l'Hôpital rule and conclude that Tsallis information tends to the Kullback-Leibler divergence:

$$\lim_{\alpha\to 1} T_\alpha(P:Q) = \sum_i p_i \log \frac{p_i}{q_i} = \mathrm{KL}(P:Q).$$

Note that those Rényi and Tsallis information divergences are mutually related to each other by *monotonic* transformations:

$$R_\alpha(P:Q) = \frac{1}{\alpha-1} \log(1 + (\alpha-1) T(P:Q)), \quad (1)$$

$$T_\alpha(P:Q) = \frac{e^{(\alpha-1)R_\alpha(P:Q)} - 1}{\alpha - 1} \quad (2)$$

In particular for $\alpha \to 1$, using a first-order approximation $\log(1+x) \simeq_{x\to 0} x$, we find that

$$\lim_{\alpha\to 1} R_\alpha(P:Q) \simeq_{\alpha\to 1} \frac{1}{\alpha-1}(\alpha-1)T(P:Q) \simeq_{\alpha\to 1} T_\alpha(P:Q)$$

(Similarly, we have $\exp(x) \simeq x \to 1 + x$, so that using E q. **??** we deduce that $T_\alpha(P:Q) \simeq_{\alpha\to 1} R_\alpha(P:Q)$.)

Although both extensive Rényi and non-extensive Tsallis information divergences meet in the limit case $\alpha = 1$, those divergences behave very differently from the classical Kullback-Leibler divergence. Indeed, the Kullback-Leibler divergence can be decomposed as $\mathrm{KL}(P : Q) = H(P, Q) - H(P, P)$, that is the cross-entropy $H(P : Q) = -\sum_i p_i \log q_i$ minus the entropy $H(P) = H(P, P) = -\sum_i p_i \log p_i$. Such a sum-decomposition does not apply to Rényi/Tsallis information divergences. (Unfortunately, Rényi/Tsallis information divergences are also sometimes misleadingly called cross-entropies in the literature. Indeed, this is confusing since in that case the entropy cannot be considered as a self cross-entropy, that would otherwise yields an inconsistent zero!)

4 Proof of l'Hôpital rule

Consider the case of the indeterminate ratio 0/0. The proof of l'Hôpital rule relies on *Cauchy mean value theorem* that states for a C^1-function f on $[a, b]$ there exists $c \in (a, b)$ such that

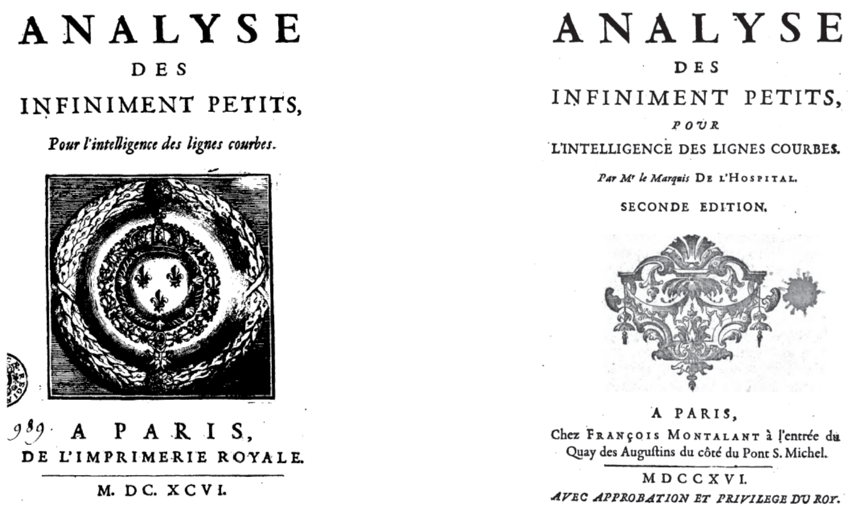
$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

First, let us define by continuity $f(\alpha) = g(\alpha) = 0$, and let $\lim_{x\rightarrow\alpha} \frac{f'(\alpha)}{g'(\alpha)} = l$. There exists $x \in (\alpha - \delta, \alpha + \delta)\backslash\{\alpha\}$ such that both $f'(x)$ and $g'(x)$ exist and $g'(x) \neq 0$. Suppose without loss of generality that $x \in (\alpha, \alpha + \delta)$, then the mean value theorem implies that $g(x) \neq 0$ (otherwise by contradiction, we would have $y \in (\alpha, x)$ with $g'(y) = 0$). Thus according to the mean value theorem there exists a point $\epsilon_x \in (\alpha, x)$ such that

$$\frac{f(x)}{g(x)} = \frac{f'(\epsilon_x)}{g'(\epsilon_x)}$$

When $x \rightarrow \alpha$, we have $\epsilon_x \rightarrow \alpha$ and since the ratio of the derivatives exists, we deduce that

$$\lim_{x\rightarrow\alpha} \frac{f(x)}{g(x)} = \lim_{x\rightarrow\alpha} \frac{f'(\epsilon_x)}{g'(\epsilon_x)} = \lim_{x\rightarrow\alpha} \frac{f'(x)}{g'(x)} = l.$$


ANALYSE
DES
INFINIMENT PETITS,
Pour l'intelligence des lignes courbes.
A PARIS,
DE L'IMPRIMERIE ROYALE.
M. DC. XCVI.

ANALYSE
DES
INFINIMENT PETITS,
POUR
L'INTELLIGENCE DES LIGNES COURBES.
Par Mr le Marquis DE L'HOSPITAL.
SECONDE EDITION.
A PARIS,
Chez FRANÇOIS MONTALANT à l'entrée du Quay des Augustins du côté du Pont S. Michel.
MDCCXVI.
AVEC APPROBATION ET PRIVILEGE DU ROY.

Figure 1: L'Hospital textbook on infinitesimal calculus: Cover pages of the original *anonymous* edition of M.DC.XCVI=1696 (left) and of the second edition of MDCCXVI=1716 (right).

5 Historical notes

Although the theorem is nowadays cited as l'Hôpital rule, it is commonly believed to be partly his work since French nobleman Guillaume de l'Hôpital (1661-1704, spelled l'Hospital in old French) hired Swiss mathematician Johann Bernoulli (1667–1748) to lecture him on mathematics. L'Hôpital wrote and published in 1696 the very first textbook on infinitesimal calculus (a precursor of modern differential calculus omitting voluntarily integration) titled: *Analysis of the infinitely small to understand curves* (see Figure 1). The name of Bernoulli is acknowledged in the preface as the mathematician that pursued on the inspiring preliminary work of Leibniz (1646–1716).

References

[1] Andrzej Cichocki, Anh Huy Phan, and Rafal Zdunek. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation; electronic version.* Wiley, 2009.

[2] Alfred Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pages 547–561, 1961.

[3] Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Physics*, 52, 1988.

[4] Tim van Erven and Peter Harremoës. Rényi divergence and its properties. *CoRR*, abs/1001.4448, 2010.