Proofs of Alpha Divergence Properties

Thursday, September 11, 2008
Rice University
STAT 631 / ELEC 639: **Graphical Models**

Scribe:
Ahmad Beirami

Instructor:
Dr. Volkan Cevher

Reviewers:
Beth Bower and Konstantinos Tsianos

1. Motivation

In many cases, we need to define a measure of distance between probability distributions. For example when trying to approximate a distribution with another on of a particular form, we can compute the necessary parameters by minimizing such a distance measure. A general class of distance measures is the family of α-divergences. In general, for any two probability distributions $p(x), q(x)$ the α-divergences is defined as follows: [1]

$$D_\alpha(p||q) = \frac{\int \alpha p(x) + (1-\alpha)q(x) - [p(x)]^\alpha [q(x)]^{1-\alpha} dx}{\alpha(1-\alpha)},$$

for all $\alpha \in [-\infty, +\infty]$.

In a previous lecture, there is more details about the exact usage of α-divergences. We devote this lecture to proving the following important properties:

(1) $D_\alpha(p||q)$ is convex with respect to both p and q

(2) $D_\alpha(p||q) \geq 0$. and $D_\alpha(p||q) = 0$ if and only if $p(x) = q(x)$ for all x [1]

(3) KL-divergence is a special case of an α-divergence:

$$\lim_{\alpha \to 0} D_\alpha(p||q) = \mathrm{KL}(q||p) = -\int q(x) \ln \left(\frac{p(x)}{q(x)}\right) dx$$
$$\lim_{\alpha \to 1} D_\alpha(p||q) = \mathrm{KL}(p||q) = -\int p(x) \ln \left(\frac{q(x)}{p(x)}\right) dx.$$

(4) For $\alpha \to -\infty$, the estimation q that approximates p is exclusive, i.e., $q(x) \leq p(x)$ for all x

(5) For $\alpha \to \infty$, the estimation q that approximates p is inclusive, i.e., $q(x) \geq p(x)$ for all x

(6) For $\alpha \leq 0$, the estimation q that approximates p is *zero forcing* i.e., $p(x) = 0 \Rightarrow q(x) = 0$

(7) For $\alpha > 1$, the estimation q that approximates p is *zero avoiding* i.e., $p(x) > 0 \Rightarrow q(x) > 0$

[1] Since the distributions need not be continuous, this statement from now on will refer to all the values of x for which the distributions are well-defined.

To proceed with proving properties $1-7$ first note that the definition of an α-divergence can be further simplified to

$$D_\alpha(p||q) = \frac{1 - \int [p(x)]^\alpha [q(x)]^{1-\alpha} dx}{\alpha(1-\alpha)}. \tag{1}$$

2. Proofs

Property 1. $D_\alpha(p||q)$ is convex with respect to both p and q.

Proof.

$$\begin{aligned}\frac{\partial^2 D_\alpha(p||q)}{\partial p^2} &= \frac{\int -\alpha(\alpha-1)[p(x)]^{\alpha-2}[q(x)]^{1-\alpha} dx}{\alpha(1-\alpha)} \\ &= \int [p(x)]^{\alpha-2}[q(x)]^{1-\alpha} dx \geq 0.\end{aligned} \tag{2}$$

Therefore, $D_\alpha(p||q)$ is convex with respect to p. It is straightforward to repeat this calculation for q.

Property 2. $D_\alpha(p||q) \geq 0.$ and $D_\alpha(p||q) = 0$ if and only if $p(x) = q(x)$ for all x.

Proof.

We prove this by proving the following inequality for any x:

$$f(p,q) = \frac{\alpha p(x) + (1-\alpha)q(x) - [p(x)]^\alpha [q(x)]^{1-\alpha}}{\alpha(1-\alpha)} \geq 0. \tag{3}$$

If this statement is proven, $D_\alpha(p||q) \geq 0$ will be achieved by integration of f with respect to x.

Now, let us look at f:

$$f(p,q) = \frac{\alpha p + (1-\alpha)q - [p]^\alpha [q]^{1-\alpha}}{\alpha(1-\alpha)}. \tag{4}$$

Now, we differentiate with respect to p

$$\frac{\partial f(p,q)}{\partial p} = \frac{q}{\alpha(1-\alpha)} \left(\frac{\alpha}{q} - \alpha \frac{p^{\alpha-1}}{q^\alpha} \right). \tag{5}$$

We can also find the second derivative with respect to p as given by

$$\frac{\partial^2 f(p,q)}{\partial p^2} = \frac{p^{\alpha-2}}{q^\alpha - 1}. \tag{6}$$

It is easy to observe that $f(q,q) = 0$ and $\frac{\partial f}{\partial p}(q,q) = 0$. We can also see that $\frac{\partial^2 f}{\partial p^2}(p,q) > 0$ for all $p > 0$. Therefore, $f(p,q)$ only has one minimum with respect to p and it takes its minimum where $\frac{\partial f}{\partial p}(p,q) = 0$, i.e., where $p = q$. Also, the value of the minimum is $f(q,q) = 0$.

Now, we have

$$D_\alpha(p||q) = \int f\left(p(x), q(x)\right) dx \geq 0. \tag{7}$$

The equality sign is obtained if and only if $p(x) = q(x)$ for all x.

Property 3.

$$\lim_{\alpha\to 0} D_\alpha(p||q) = KL(q||p) = -\int q(x)\ln\left(\frac{p(x)}{q(x)}\right)dx$$
$$\lim_{\alpha\to 1} D_\alpha(p||q) = KL(p||q) = -\int p(x)\ln\left(\frac{q(x)}{p(x)}\right)dx. \tag{8}$$

Proof.

When $\alpha \to 0$, it is easy to observe that both the numerator and the denominator in equation 1 will vanish. Therefore, we can use the L'Hopital rule to obtain the limit (We differentiate both of them with respect to α).

$$\lim_{\alpha\to 0} D_\alpha(p||q) = \lim_{\alpha\to 0} \frac{-\int \ln\left(\frac{p(x)}{q(x)}\right)[p(x)]^\alpha[q(x)]^{1-\alpha}dx}{1-2\alpha} = KL(q||p). \tag{9}$$

The proof of the other equality is also straightforward. Note that due to the symmetry between (p,α) and $(q,1-\alpha)$ the second equality holds.

Property 4. If we are minimizing D_α to find the best p that estimates q and $\alpha \to -\infty$, the estimation q is exclusive, i.e., $q(x) \leq p(x)$ for all x.

Proof.

We take the limit as

$$\lim_{\alpha\to -\infty} D_\alpha(p||q) = \lim_{\alpha\to -\infty} \frac{\left(\frac{q}{p}\right)^{-\alpha}}{\alpha^2}. \tag{10}$$

If p and q are continuous functions and $q(x) > p(x)$ at some point x_0, there exists a neighborhood around x_0, where $q(x) > p(x)$. Therefore, $(q/p)^{-\alpha} \to \infty$. and the result of the integral will diverge to infinity. Thus, any minimization algorithm will result in a $q(x)$, where $q(x) \leq p(x)$.

Property 5. If we are minimizing D_α to find the best p that estimates q and $\alpha \to \infty$, the estimation q is inclusive, i.e., $q(x) \geq p(x)$ for all x.

Proof.

We take the limit as

$$\lim_{\alpha\to \infty} D_\alpha(p||q) = \lim_{\alpha\to -\infty} \frac{\left(\frac{p}{q}\right)^{\alpha}}{\alpha^2}. \tag{11}$$

If p and q are continuous functions and $q(x) < p(x)$ at some point x_0, there exists a neighborhood of x_0, where $q(x) < p(x)$. Therefore, $(p/q)^\alpha \to \infty$ and the result of the integral will diverge to infinity. Thus, any minimization algorithm will result in a $q(x)$, where $q(x) \geq p(x)$.

Property 6. If we are minimizing D_α to find the best p that estimates q and $\alpha \leq 0$, then $p(x) = 0 \Rightarrow q(x) = 0$.

Proof.

We have

$$D_\alpha(p||q) = \frac{1 - \int \left(\frac{q(x)}{p(x)}\right)^{-\alpha} q(x)dx}{\alpha(1-\alpha)}. \tag{12}$$

If p and q are continuous functions and $p(x) = 0$ at some point x_0, there exists a neighborhood around x_0, where $p(x) < \epsilon$. Therefore, if $q(x) > 0$, $(q/p)((q/p)^{-\alpha})$ will be very large in that neighborhood, which makes the integral grow very large. Thus, any minimization algorithm will result in a $q(x)$, where $q(x) = 0$ if $p(x) = 0$. (Note that D_α is convex with respect to p and q).

Property 7. If we are minimizing D_α to find the best p that estimates q and $\alpha > 1$, then $p(x) > 0 \Rightarrow q(x) > 0$.

Proof.

We can achieve this property using symmetry as discussed before. We have

$$D_\alpha(p||q) = \frac{1 - \int \left(\frac{p(x)}{q(x)}\right)^{\alpha-1} p(x)dx}{\alpha(1-\alpha)}. \tag{13}$$

If p and q are continuous functions and $p(x) > 0$ at some point x_0, there exists a neighborhood around x_0, where $p(x) > k$. Therefore, if $q(x_0) = 0$, (p/q) will be very large in that neighborhood, which makes the integral grow very large. Thus, any minimization algorithm will result in a $q(x)$, where $q(x) > 0$ if $p(x) > 0$.

Keywords: *α-divergence, variational methods, KL-divergence*

References

[1] Thomas Minka. Divergence measures and message passing. Technical report, Microsoft Research Ltd, 2005.