α-centroids and α-barycenters of probability measures: Average divergence minimizers with respect to α-divergences

Frank Nielsen

June 2008, revised December 2009

Abstract

We summarize the paper: Shun-ichi Amari, *Integration of Stochastic Models by Minimizing α-Divergence*, Neural Computation (NECO), (19)10:2780-2796, October 2007.
It is proved that weighted α-means are barycenters defined with respect to α-divergences. Applications to α-risk optimal experts and α-Bayesian predictive distributions are described. We revisited and extended these results (to β-divergences) using the framework of representational Bregman divergences (ISVD'09).

Key words: α-mean, α-mixture (ie., centroid of α-representation of distribution), α-integration (ie., barycenter), α-risk and α-experts (product or mixture of experts), α-predictive distribution for Bayesian predictive distribution estimation.

1 Barycenters of α-divergences are weighted α-means.

For a strictly monotonous and differentiable function f (bijective mapping chosen modulo affine terms $ax+b$), a *weighted f-mean* M_f of a *population* $p_1, ..., p_n$ is defined by $M_f(p_1, ..., p_n; w) = f^{-1}(\sum_i w_i f(p_i))$ with $w_i \geq 0, \forall i$ and $\sum_i w_i = 1$. Generalized means extend the familiar arithmetic ($f(x) = x$), geometric ($f(x) = \log x$) and harmonic ($f(x) = \frac{1}{x}$) Pythagorean means, among many others. *Linear scale free* means further satisfy the essential property $M_f(cp_1, ..., cp_n; w) = cM_f(p_1, ..., p_n; w)$, for any nonnegative constant $c > 0$. Hardy et al. proved in 1952 that linear scale free f-means are obtained *for and only for* generators $f_\alpha(x) = x^{\frac{1-\alpha}{2}}$ with $\alpha \neq 1$, and $f_1(x) = \log x$. A barycenter is defined as the unique solution of a weighted minimum average optimization problem (the minimum is called α-information; for Bregman divergences it is the Bregman information that is a Burbea-Rao divergence. That is, a generalized Jensen-Shannon divergence). Linear scale free means are α-means that satisfy the composition rule.

The α-means m_α (shortcut for f_α-means with uniform weight w) of *nonnegative* numbers are given by $c_\alpha(\sum_i p_i^{\frac{1-\alpha}{2}})^{\frac{2}{1-\alpha}}$, where c_α is a proper constant that yields $m_\alpha(a, ..., a) = a$. The α-mean is inversely monotone wrt. to α: $m_\alpha < m_\beta$ for $\beta < \alpha$.

For a probability density function $p(x)$, define its *α-representation* $f_\alpha(p(x))$, and the *α-family mixture* as $p_\alpha(x) = cf_\alpha^{-1}(\frac{1}{n}\sum f_\alpha(p_i(x)))$, where c is the normalization coefficient. A weighted α-mixture is called the α-integration (ie., a barycenter of pdfs). For $\alpha = -1$, we obtain the traditional *linear mixture model*. For $\alpha = 1$, we get the *exponential family* $m_1(x) = \exp(\sum_i w_i \log p_i(x) - F(w))$,

where $c = \exp -F(w)$ is the *cumulant generating function* (also called log normalizer). The α-*divergence* is a parametric family f_α of distortion measures that can be derived from Csiszár f-divergence[1] as follows:

$$D_\alpha(p||q) = \begin{cases} q - p + p\log\frac{p}{q} & \alpha = -1, \\ p - q + q\log\frac{q}{p} & \alpha = 1, \\ \frac{2}{1+\alpha}p + \frac{2}{1-\alpha}q - \frac{4}{1-\alpha^2}p^{\frac{1-\alpha}{2}}q^{\frac{1+\alpha}{2}} & \alpha \neq \pm 1. \end{cases}$$

D_α divergences are asymmetric except for $\alpha = 0$, and $D_\alpha(p||q) = D_{-\alpha}(q||p)$. In particular, $D_{-1}(p||q) = \mathrm{KL}(p||q)$ and $D_1(p||q) = \mathrm{KL}(q||p)$, and $D_0(p||q) = 2\int(\sqrt{p(x)} - \sqrt{q(x)})\mathrm{d}x$ bears the name of *squared Hellinger*[2] distance. Let the *information radius* be defined as $R_\alpha(q) = \sum_i w_i D_\alpha(p_i||q)$ (right-side barycenter) and attained for $q^* = \arg\min R_\alpha(q)$. The α-*integration* $cf_\alpha^{-1}(\sum w_i f_\alpha(p_i))$ is optimal wrt. to divergence D_α.

2 Applications: α-Experts and α-Bayesian predictive distribution

Applications of (1) finite α-integration of experts, and (2) continuous α-integration for α-Bayesian predictive distributions are then discussed:

Integration of experts. Let w_i denote the *reliability* of expert E_i that given some input signal s produces a pdf (or a relaxed positive measure) $e_i(x|s)$. The α-*risk* is defined in terms of α divergence by $R_\alpha(q|s) = \sum w_i D_\alpha(p_i(x|s)||q(x))$. The α-*expert machine* produces a signal $q(s|x) = f_\alpha^{-1}(\sum w_i(x) f_\alpha(p_i(s|x)))$ that is optimal wrt. to α-risk. The α-expert machine generalized former mixture of experts ($\alpha = 1$) and product of experts ($\alpha = -1$). A key issue is to determine the "'best"' α values. Amari addresses this issue on determining the weight $w_i(x)$ when a teacher is provided as an oracle.

Bayesian predictive distribution. In the Bayesian framework, given $D = \{x_1, ..., x_n\}$ independent observations and $\pi(\theta)$ a prior distribution, the *Bayesian predictive distribution* $p(x|D) = \int p(x|\theta)\frac{\pi(\theta)\prod_{i=1}^n p(x_i|\theta)}{\int p(D,\theta)d\theta}$ (a continuously infinite generalized means) is shown optimal wrt. (minimizing the expectation of) the risk $R(q(x|D)) = \int \pi(\theta)p(D|\theta)\mathrm{KL}(p(x|\theta)||q(x|D)]\mathrm{d}\theta\mathrm{d}D$ with $dD = \mathrm{d}x_1...\mathrm{d}x_n$. That is, the Bayesian predictive distribution minimizes the mean Kullback-Leibler divergence from *true distribution* $p(x|\theta)$ to *test distribution* $p(x|D)$. This result is extended to D_α divergence: The α-predictive distribution $p_\alpha(x|D) = cf_\alpha^{-1}(\int f_\alpha(p(x|\theta))p(\theta|D))\mathrm{d}\theta$ is optimal wrt. to the α-risk $R_\alpha(q(x|D)) = \int \pi(\theta)p(D|\theta)D_\alpha(p(x|\theta)||q(x|D))\mathrm{d}\theta\mathrm{d}D$.

In human brain, α-integration (ie., barycenter computations) seems to take part. See for example, Weber-Fechner and Stevens laws that describe *population coding* in medial temporal (MT) and medial superior temporal (MST) cortex areas.

[1] Csiszár f-divergence are defined by $C_f(p||q) = \int p(x) f(\frac{q(x)}{p(x)})\mathrm{d}x$, with f convex. The dual Csiszár divergence $C_{f^*}(p||q) = C_f(q||p)$ is obtained for generator $f^*(x) = xf(\frac{1}{x})$. It follows that symmetrized Csiszár divergences are Csiszár divergences for generator $s_f(x) = \frac{f(x)+xf(\frac{1}{x})}{2}$. Csiszár divergences satisfy the information monotonicity property.

[2] Hellinger distance can be viewed as the L_2-norm of space $\sqrt{p(x)}$.

Reference

F. Nielsen and R. Nock, The dual Voronoi diagrams with respect to representational Bregman divergences, International Symposium on Voronoi Diagrams (ISVD), June 2009. F. Nielsen and R. Nock, Sided and Symmetrized Bregman Centroids, IEEE transactions on information theory (2009), vol. 55, no. 6, pp. 2882-2904