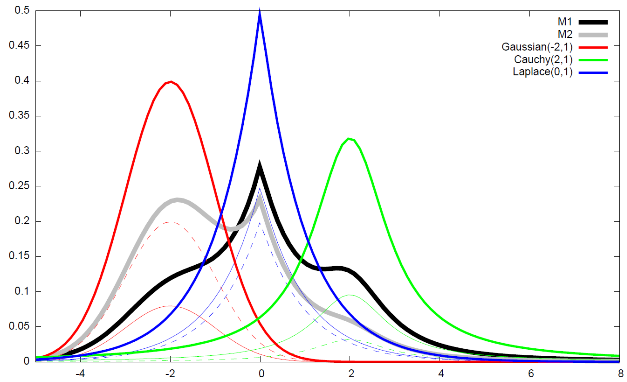
ON THE GEOMETRY OF MIXTURES OF PRESCRIBED DISTRIBUTIONS

Frank Nielsen
Sony Computer Science Laboratories, Japan
Richard Nock
Data 61, The Australian National & Sydney Universities, Australia

1/ Geometry of w-mixtures:

In probability, a *statistical mixture*:
$m(x;w) = m(x;\eta) = \sum_{i=1}^{k-1} \eta_i p_i(x) + \left(1 - \sum_{i=1}^{k-1} \eta_i\right) p_0(x)$

In information geometry, a *mixture family*:
$\mathcal{M} = \left\{ m(x;\eta) = \sum_{i=1}^{k-1} \eta_i f_i(x) + c(x), \quad \eta \in \Delta_D^\circ \right\}$
$f_i(x) = p_i(x) - p_0(x)$ for $i \in [D]$, $\quad c(x) = p_0(x)$



2/ Dually flat space from a strictly convex and smooth functional (here, statistical):

Shannon differential entropy of a mixture $m(x)$ (concave):

$$h(m) := -\int_{\mathcal{X}} m(x) \log m(x) \mathrm{d}\mu(x)$$

Shannon information as a Bregman generator (convex):

$$F^*(\eta) = \int m(x;\eta) \log m(x;\eta) \mathrm{d}\mu(x)$$

Dual Legendre convex conjugate (cross-entropy):

$$F(\theta) = -\int p_0(x) \log m(x;\eta) \mathrm{d}\mu(x)$$

Dual parameterization of η-mixtures:

$$\theta^i(\eta) = (\nabla_\eta F^*(\eta))_i = \int (p_i(x) - p_0(x)) \log m(x;\eta) \mathrm{d}\mu(x)$$

Fact: Kullback-Leibler divergence between two η-mixtures (or w-mixtures) is equivalent to a Bregman divergence defined for the Shannon negentropy generator on the η-parameters.

Corollary: The KL between w-Gaussian mixture model is a Bregman divergence for the Shannon negentropy generator.

$$\begin{aligned}
\mathrm{KL}(m_1 : m_2) &= \int m(x;\eta_1) \log \frac{m(x;\eta_1)}{m(x;\eta_2)} \mathrm{d}\mu(x) \\
&= B_{F^*}(\eta_1 : \eta_2) = B_F(\theta_2 : \theta_1) \\
&= D_{F^*,F}(\eta_1 : \theta_2) = D_{F,F^*}(\theta_2 : \eta_1)
\end{aligned}$$

where $D_{F^*,F}(\eta_1 : \theta_2) = F^*(\eta_1) + F(\theta_2) - \langle \eta_1, \theta_2 \rangle$

3/ Applications:

- Optimal KL-averaging integration:

Theorem:The KL-averaging integration of w-mixtures performed optimally without information loss.
$\hat{\eta} = \mathrm{argmin}_\eta \sum_{i=1}^m \mathrm{KL}(m(\hat{\eta}_i) : m(\eta)) \equiv \sum_{i=1}^m B_{F^*}(\hat{\eta}_i : \eta)$
$\Rightarrow \hat{\eta} = \frac{1}{m} \sum_{i=1}^m \hat{\eta}_i$ (Bregman right centroid indep. of F^*)

- Skew α-Jensen-Shannon divergence:

$\mathrm{JS}_\alpha(p : q) := (1-\alpha)\mathrm{KL}(p : m_\alpha) + \alpha \mathrm{KL}(q : m_\alpha)$, for $\alpha \in [0,1]$, and $m_\alpha = (1-\alpha)p + \alpha q$.
α-Jensen divergences
$J_{F^*,\alpha}(\eta_1 : \eta_2) := (1-\alpha)F^*(\eta_1) + \alpha F^*(\eta_2) - F^*((1-\alpha)\eta_1 + \alpha\eta_2)$, for $F^*(\eta) = -h(m(x;\eta))$.

Limit cases:

$$\lim_{\alpha \to 1^-} \frac{J_{F^*,\alpha}(\eta_1 : \eta_2)}{\alpha(1-\alpha)} = B_{F^*}(\eta_1 : \eta_2) = \mathrm{KL}(m_1 : m_2)$$

$$\lim_{\alpha \to 0^+} \frac{J_{F^*,\alpha}(\eta_1 : \eta_2)}{\alpha(1-\alpha)} = B_{F^*}(\eta_2 : \eta_1) = \mathrm{KL}(m_2 : m_1)$$

Theorem. The α-Jensen-Shannon statistical divergences between η-mixtures amount to α-Jensen divergences between their corresponding η-mixture parameters: $\mathrm{JS}_\alpha(m(x;\eta_1) : m(x;\eta_2)) = J_{F^*,\alpha}(\eta_1 : \eta_2)$.

4/ Divergence inequalities and family closure:

$m^\epsilon(p,q) = (1-\epsilon)p + \epsilon q = p + \epsilon(q-p) = m^{1-\epsilon}(q : p)$ for $\epsilon \in [0,1]$. $I_f^\epsilon(p : q) := I_f(m^\epsilon(p,q) : m^\epsilon(q,p))$.

The f-divergence $I_f(m(x;w) : m(x;w'))$ between any two w-mixtures is upper bounded by $I_f(w : w') = \sum_{i=0}^{k-1} w_i f(\frac{w_i'}{w_i})$.

$$\begin{aligned}
I_f^\epsilon(p : q) &\leq (1-\epsilon) I_f(p : q) + \epsilon I_f(q : p), \\
I_f^\epsilon(p : q) &\leq (1-\epsilon) f\left(\frac{\epsilon}{1-\epsilon}\right) + \epsilon f\left(\frac{1-\epsilon}{\epsilon}\right).
\end{aligned}$$

References:

- On w-mixtures: Finite convex combinations of prescribed component distributions, arxiv 1708.00568
- Monte Carlo Information Geometry: The dually flat case, arxiv 1803.07225