Some contributions to the theory of distances

Frank Nielsen
Sony Computer Science Laboratories Inc., Tokyo, Japan
E-mail:`Frank.Nielsen@acm.org`

November 6, 2020

1 Calculating statistical distances, relative entropies, cross-entropies and entropies

- **Cumulant-free closed-form formulas for some common (dis)similarities between densities of an exponential family** (`https://arxiv.org/abs/2003.02469`)

Since the Jensen and Bregman convex generators $F(\theta)$ are defined modulo an *affine term* $\langle a,\theta\rangle+\mathrm{b}$ (i.e., $J_F(\theta_1:\theta_2)=J_G(\theta_1:\theta_2)$ and $B_F(\theta_1:\theta_2)=B_G(\theta_1:\theta_2)$ with $G(\theta)=F(\theta)+\langle a,\theta\rangle+\mathrm{b}$), we can choose the equivalent generator $G(\theta):=-\log p_\theta(x)=F(\theta)-\langle t(x),\theta\rangle-k(x)$, and express the Kullback-Leibler divergence, the skewed Bhattacharrya divergences, the α-divergences and many other statistical distances between densities of a natural exponential family $\{p_\theta(x)=1_{\mathcal{X}}(x)\ \exp\left(\langle t(x),\theta\rangle-F(\theta)+k(x)\right)\}$ without *explicitly* using the log-normalizer $F(\theta)=\log\left(\int_{x\in\mathcal{X}}\exp\left(\langle t(x),\theta\rangle+k(x)\right)\mathrm{d}\mu(x)\right)$ of the exponential family.

For example, the Bhattacharyya coefficient is expressed as:

$$
\begin{aligned}
\rho[p_{\theta_1},p_{\theta_2}] &:= \int_{x\in\mathcal{X}}\sqrt{p_{\theta_1}(x)\ p_{\theta_2}(x)}\mathrm{d}\mu(x),\\
&= \exp(-J_F(\theta_1:\theta_2))=\exp(-J_{-\log p_\theta(\omega)}(\theta_1:\theta_2)),\quad \forall\ \omega\in\mathcal{X},\\
&= \frac{p_{\bar\theta}(\omega)}{\sqrt{p_{\theta_1}(\omega)p_{\theta_2}(\omega)}},\quad \forall\ \omega\in\mathcal{X},
\end{aligned}
$$

where $\bar\theta:=\frac{\theta_1+\theta_2}{2}$. For generic exponential families parameterized by $\lambda(\theta)$ (i.e., not in natural form), we need to explicit the mid-parameter $\bar\lambda:=\lambda(\bar\theta)$ from the *partial* factorization of the exponential family (the λ-mean corresponding to the θ-mean).

For the Kullback-Leibler divergence, using the fact that $D_{\mathrm{KL}}[p_{\theta_1}:p_{\theta_2}]=B_F[\theta_2:\theta_1]=B_G[\theta_2:\theta_1]$ (better written as $D^*_{\mathrm{KL}}[p_{\theta_2}:p_{\theta_1}]=B_F[\theta_2:\theta_1]$ where D^*_{KL} is the reverse divergence) with the equivalent generator $G(\theta)=-\log p_\theta(x)$, we get

$$
D_{\mathrm{KL}}[p_{\theta_1}:p_{\theta_2}]=\log\left(\frac{p_{\theta_1}(\omega)}{p_{\theta_2}(\omega)}\right)+(\theta_2-\theta_1)^\top(t(\omega)-\nabla F(\theta_1)),\quad \forall\ \omega\in\mathcal{X}.
$$

Choosing ω such that $t(\omega)=\nabla F(\theta_1)=E_{p_{\theta_1}}[t(x)]=:\eta_1$, we express the KLD as a log density ratio: $D_{\mathrm{KL}}[p_{\theta_1}:p_{\theta_2}]=\log\left(\frac{p_{\theta_1}(\omega)}{p_{\theta_2}(\omega)}\right)$. In general we may need several ω_i's so that $\frac{1}{s}\sum_i t(\omega_i)=\nabla F(\theta_1)=\eta_1$.

For example, we can write the KLD between two multivariate normal distributions as

$$
D_{\mathrm{KL}}[p_{\mu_1,\Sigma_1}:p_{\mu_2,\Sigma_2}] = \frac{1}{2d}\sum_{i=1}^d\left(\log\left(\frac{p_{\mu_1,\Sigma_1}\left(\mu_1-\sqrt{d\lambda_i}e_i\right)}{p_{\mu_2,\Sigma_2}\left(\mu_1-\sqrt{d\lambda_i}e_i\right)}\right)+\log\left(\frac{p_{\mu_1,\Sigma_1}\left(\mu_1+\sqrt{d\lambda_i}e_i\right)}{p_{\mu_2,\Sigma_2}\left(\mu_1+\sqrt{d\lambda_i}e_i\right)}\right)\right),
$$

where $[\sqrt{d\Sigma_1}]_{\cdot,i}=\sqrt{\lambda_i}e_i$ denotes the vector extracted from the i-th column of the square root matrix of $d\Sigma_1$.