A note on the Hyvärinen divergence between densities of an exponential family

Frank Nielsen
Sony Computer Science Laboratories Inc

October 2016

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space where μ is a positive measure (e.g., Lebesgue or counting) with $\mathcal{X}$ denoting the sample space and $\mathcal{F}$ the σ-algebra. Hyvärinen proposed the following divergence for estimating non-normalized distributions using the method of score matching (Eq. 2 in [4]):

$$D_{\mathrm{Hyv}}[p:q] := \frac{1}{2}\int \left\| \nabla_x \log \frac{p(x)}{q(x)} \right\|^2 \; p(x)\mathrm{d}\mu(x),$$

where $p(x)$ and $q(x)$ are two densities with full support $\mathcal{X}$.

The divergence is said to be *right-sided projective*:

$$\forall \lambda > 0, \quad D_{\mathrm{Hyv}}[p:\lambda q] = D_{\mathrm{Hyv}}[p:q],$$

since $-\nabla_x \log \lambda = 0$. Thus we may consider a non-normalized distribution $\tilde{q}$ for the right-hand-side argument of the *Hyvärinen divergence*:

$$D_{\mathrm{Hyv}}[p:q] = D_{\mathrm{Hyv}}[p:\tilde{q}].$$

Let $p = p_{\theta_1}$ and $q = p_{\theta_2}$ denote two densities of an exponential family [5, 1]:

$$\left\{ p_\theta(x) = \exp\left(\sum_{i=1}^D \theta_i t_i(x) - F(\theta) + k(x) \right) \; : \; \theta \in \Theta \right\},$$

where $t(x) = (t_1(x), \ldots, t_D(x))$ is a vector of sufficient statistics which are affinely independent, θ the natural parameter space, $k(x)$ an auxiliary carrier term defining the measure $\mathrm{d}\nu = \exp(k(x))\mathrm{d}\mu$ (i.e., $\mathrm{d}\nu = \mathrm{d}\mu$ when $k(x) = 0$), and $F(\theta)$ the cumulant function normalizing the density: $F(\theta) = \log \int \sum_{i=1}^D \exp\left(\theta_i t_i(x) + k(x)\right) \mathrm{d}\mu$. The order of the d-dimensional exponential family ($d = \dim(\mathcal{X})$) is its number of parameters D. Let us rewrite the density of the exponential family as

$$p_\theta(x) = \exp\left(\langle \theta, t(x)\rangle - F(\theta) + k(x)\right),$$

where $\langle a, b\rangle = a^\top b$ is the scalar product.

Since $\nabla_x \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} = \langle \theta, \nabla_x t(x)\rangle$ (since $\langle \nabla_x \theta, t(x)\rangle = 0$) with $\Delta\theta := \theta_1 - \theta_2$, the Hyvärinen divergence becomes

$$D_{\mathrm{Hyv}}[p_{\theta_1} : p_{\theta_2}] = \frac{1}{2}\int \|\langle \theta, \nabla_x t(x)\rangle\|^2 \, p_{\theta_1}(x)\mathrm{d}\mu(x).$$

When the exponential family is natural (i.e., $t(x) = x$ and $D = d$), we have $\nabla_x t(x) = \nabla_x x = 1_d$, and we have

$$D_{\mathrm{Hyv}}[p_{\theta_1} : p_{\theta_2}] = \frac{1}{2}\left\|\langle \theta_1 - \theta_2, 1_d\rangle\right\|^2.$$

In particular, when $D = 1$, we have $D_{\mathrm{Hyv}}[p_{\theta_1} : p_{\theta_2}] = \frac{1}{2}(\theta_1 - \theta_2)^2$.

For example, the exponential family of continuous exponential distributions (with μ the Lebesgue measure)

$$\left\{p_\lambda^{\mathrm{Exp}}(x)=\lambda\exp(-\lambda x)\ :\ \lambda>0\right\}$$

is a natural exponential family with natural parameter $\theta=-\lambda$ and $k(x)=0$. We have

$$D_{\mathrm{Hyv}}[p_{\lambda_1}^{\mathrm{Exp}}:p_{\lambda_2}^{\mathrm{Exp}}]=\frac{1}{2}(\lambda_2-\lambda_1)^2.$$

Another example is the discrete Poisson NEF (with μ counting measure on $\mathcal{X}=\{0,1,\ldots\}$):

$$\left\{p_\lambda^{\mathrm{Poi}}(x)=\frac{\lambda^x\exp(-\lambda)}{x!}\ :\ \lambda>0\right\}$$

with $\theta=\log\lambda$ and $k(x)=-\log x!$. We have

$$D_{\mathrm{Hyv}}[p_{\lambda_1}^{\mathrm{Poi}}:p_{\lambda_2}^{\mathrm{Poi}}]=\frac{1}{2}\left(\log\frac{\lambda_2}{\lambda_1}\right)^2.$$

Now, consider densities of a univariate polynomial exponential family [2, 6] (PEF) with sufficient statistics $t(x)=(x,\ldots,x^D)$. The PEFs include the exponential distribution family for $t(x)=x$ and the univariate normal family for $t(x)=(x,x^2)$. Notice that the cumulant function F of a PEF is not available in closed form in general.

For univariate exponential family densities ($d=1$) of order D, we have

$$D_{\mathrm{Hyv}}[p_{\theta_1}:p_{\theta_2}]=\frac{1}{2}\int\left\|\sum_{i=1}^D\Delta\theta_i t_i'(x)\right\|^2 p_{\theta_1}(x)\mathrm{d}\mu(x).$$

For the PEFs, we have $t_i'(x)=ix^{i-1}$ for $i\in\{1,\ldots,D\}$. Thus the Hyvärinen divergence between two densities of a PEF is expressed as:

$$D_{\mathrm{Hyv}}[p_{\theta_1}^{\mathrm{PEF}}:p_{\theta_2}^{\mathrm{PEF}}]=\frac{1}{2}\sum_{i=1}^D\sum_{j=1}^D\Delta\theta_i\Delta\theta_j ijE_{p_{\theta_1}}\left[x^{i+j-2}\right]. \tag{1}$$

For the normal family [5] $\{p_{\mu,\sigma}\}$ ($D=2$), we have the natural parameter $\theta=\left(\frac{\mu}{\sigma^2},-\frac{1}{2\sigma^2}\right)$, and $E_{p_{\mu,\sigma}}[x^2]=\mu^2+\sigma^2$, $E_{p_{\mu,\sigma}}[x^3]=\mu^3+3\mu\sigma^2$, $E_{p_{\mu,\sigma}}[x^4]=\mu^4+6\mu^2\sigma^2+3\sigma^4$.

Pluggins those terms in Eq. 1 and simplifying the expression, we get:

$$D_{\mathrm{Hyv}}[p_{\mu_1,\sigma_1}^{\mathrm{Nor}}:p_{\mu_2,\sigma_2}^{\mathrm{Nor}}]=\frac{(\sigma_1^2-\sigma_2^2)^2}{2\sigma_1^2\sigma_2^4}+\frac{(\mu_1-\mu_2)^2}{2\sigma_2^4}. \tag{2}$$

Observe that it is an asymmetric divergence: $D_{\mathrm{Hyv}}[p_{\mu_1,\sigma_1}^{\mathrm{Nor}}:p_{\mu_2,\sigma_2}^{\mathrm{Nor}}]\neq D_{\mathrm{Hyv}}[p_{\mu_2,\sigma_2}^{\mathrm{Nor}}:p_{\mu_1,\sigma_1}^{\mathrm{Nor}}]$.

This formula can be verified with the following Maxima software code which calculates symbolically the definite integral:

```
p(x,m,s):=1.0/(sqrt(2*%pi)*s)*exp(-(((x-m)/s)**2)/2);
assume(s1>0);
assume(s2>0);
assume(m1);
assume(m2);
integrate((1/2)*(derivative(log(p(x,m1,s1)/p(x,m2,s2)),x,1)**2)*p(x,m1,s1),x,-inf,inf);
ratsimp(%);
```

For higher degree PEFs, the cumulant function F is not available in closed form. We can however estimate the terms $E_{ij} = E_{p_{\theta_1}}\left[x^{i+j-2}\right]$ using Monte Carlo integration with rejection sampling of the proposal distribution p_{θ_1}. Rejection sampling does not require the normalization constant $\exp(F(\theta_1))$. Therefore we approximate by Monte Carlo the terms E_{ij} by i.i.d. sampling $x_1, \ldots, x_s \sim p_{\theta_1}$ using rejection sampling, and we get

$$\hat{E}_{ij} := \frac{1}{s}\sum_{l=1}^{s} x_l^{i+j-2}.$$

Then we estimate the Hyvärinen divergence by

$$\hat{D}_{\mathrm{Hyv}}[p_{\theta_1}^{\mathrm{PEF}} : p_{\theta_2}^{\mathrm{PEF}}] := \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} \Delta\theta_i \Delta\theta_j ij \hat{E}_{ij}.$$

In [6], the double-sided projective γ-divergence [3] is used to discriminate between two PEF densities. The estimation of the Hyvärinen divergence provides an alternative method to discriminate two densities of a PEF.

References

[1] Ole Barndorff-Nielsen. *Information and exponential families: in statistical theory.* John Wiley & Sons, 2014.

[2] Loren Cobb, Peter Koppstein, and Neng Hsin Chen. Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78(381):124–130, 1983.

[3] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[4] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[5] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[6] Frank Nielsen and Richard Nock. Patch matching with polynomial exponential families and projective divergences. In *International Conference on Similarity Search and Applications*, pages 109–116. Springer, 2016.