Expectation parameter of exponential family

$$p(x) = \exp(-\psi(\theta) + F(x)\theta), \quad \eta = \mathbb{E}_p[F(x)]$$

We use the normalizing condition

$$\int p(x)\,dx = \exp(-\psi(\theta)) \int \exp(F(x)\theta)\,dx = 1$$

$$\therefore \quad e^{\psi(\theta)} = \int \exp(F(x)\theta)\,dx$$

$$\therefore \quad \psi(\theta) = \log\left[\int \exp(F(x)\theta)\,dx\right]$$

Then,

$$\eta := \nabla_\theta \psi(\theta) = \frac{\partial}{\partial\theta} \log\left[\int \exp(F(x)\theta)\,dx\right]$$

$$= \frac{\int F(x)\exp(F(x)\theta)\,dx}{\int \exp(F(x)\theta)\,dx}$$

$$= \frac{\int F(x)\exp(-\psi(\theta) + F(x)\theta)\,dx}{\int \exp(-\psi(\theta) + F(x)\theta)\,dx}$$

$$\therefore y = \frac{1}{\int p(\mathbf{x})d\mathbf{x}} \int \mathbf{F}(\mathbf{x})\, p(\mathbf{x})d\mathbf{x}$$

$$= \int \mathbf{F}(\mathbf{x})\, p(\mathbf{x})d\mathbf{x} = \mathbb{E}_p\left[\mathbf{F}(\mathbf{x})\right]$$

Legendre dual of $\psi(\theta)$

$$\varphi(y) = \max_{\theta} \left(\theta \cdot y - \psi(\theta) \right)$$

$$= \int p(x) \log p(x)\, dx$$

<u>proof</u>

$$p(x) = \exp\left(-\psi(\theta) + F(x)\,\theta\right)$$

$$\log p(x) = -\psi(\theta) + F(x)\,\theta$$

$$p(x) \log p(x) = -\psi(\theta)\, p(x) + p(x) F(x)\,\theta$$

$$\int p(x) \log p(x)\, dx = -\psi(\theta) + \theta \int dx\, p(x) F(x)$$

$$= -\psi(\theta) + \theta \cdot E_p\left[F(x)\right]$$

$$= -\psi(\theta) + \theta \cdot y$$

The KL divergence is a Bregman divergence generated by negative entropy $\psi(\eta) = \int p(x) \log p(x) \, dx$.

<u>proof</u>

$$p_T(x) = \exp\left(-\psi(\theta_T) + F(x)\theta_T\right)$$

$$P_p(x) = \exp\left(-\psi(\theta_p) + F(x)\theta_p\right)$$

$$\therefore \log P_T(x) - \log P_p(x) = -\psi(\theta_T) + \psi(\theta_p)$$
$$F(x)(\theta_T - \theta_p)$$

Take expectation w.r.t. P_T,

$$\int (P_T \log P_T - P_T \log P_p) \, dx$$

$$= -\psi(\theta_T) \int P_T(x) dx$$
$$+ \psi(\theta_p) \int P_\theta(x) dx$$
$$+ \int F(x) \underset{\substack{\wedge \\ P_T(x)}}{\,} dx \cdot (\theta_T - \theta_p)$$

$$\therefore D_{KL}(T, p) = -\psi(\theta_T) + \psi(\theta_p) + \eta_T \cdot (\theta_T - \theta_p)$$

We put $\psi(\theta) = \theta \cdot \eta - \varphi(\eta)$, then,

$$D_{KL}(\tau, p) = -(\theta_\tau \cdot \eta_\tau - \varphi(\eta_\tau)) + (\theta_p \cdot \eta_p - \varphi(\eta_p)) + \eta_\tau(\theta_\tau - \theta_p)$$

$$= \varphi(\eta_\tau) - \varphi(\eta_p) - \theta_\tau \cdot \eta_\tau + \theta_p \cdot \eta_p + \eta_\tau \theta_\tau - \eta_\tau \cdot \theta_p$$

$$= \varphi(\eta_\tau) - \varphi(\eta_p) - \theta_p(\eta_\tau - \eta_p)$$

$$= \varphi(\eta_\tau) - \varphi(\eta_p) - \nabla_\theta \psi(\theta_p) \cdot (\eta_\tau - \eta_p)$$

$\boxed{/}$

Exponential family includes Gaussian dist.

Let's consider just 1D Gaussians. We define

$$F(x) = (x, x^2), \quad \theta = (\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$$

and

$$\psi(\theta) := -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-\theta_2) + \frac{1}{2}\log\pi$$

$$= \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\frac{1}{2\sigma^2} + \frac{1}{2}\log\pi$$

$$= \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\pi\cdot 2\sigma^2 = \frac{\mu^2}{2\sigma^2} + \log\sqrt{2\pi}\,\sigma$$

then...

$$p(x) = \exp(-\psi(\theta) + F(x)\cdot\theta)$$

$$= \exp\left(-\frac{\mu^2}{2\sigma^2} - \log\sqrt{2\pi}\sigma + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x^2+\mu^2+2\mu x)\right)$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Duality of the Bregman divergence.

For dual functions ψ and φ, it holds that

$$D_\psi(T,P) = D_\varphi(P,T)$$

proof

$$D_\psi(T,P) = \psi(\theta^T) - \psi(\theta^P) - \nabla_\theta \psi(\theta^P)\cdot(\theta^T - \theta^P)$$

$$= (\theta^T\cdot\eta^T - \varphi(\eta^T)) - (\theta^P\cdot\eta^P - \varphi(\eta^P)) - \eta^P\cdot(\theta^T - \theta^P)$$

$$= \theta^T\cdot\eta^T - \varphi(\eta^T) - \theta^P\cdot\eta^P + \varphi(\eta^P) - \eta^P\cdot\theta^T + \eta^P\cdot\theta^P$$

$$= \varphi(\eta^P) - \varphi(\eta^T) + \theta^T(\eta^T - \eta^P)$$

$$= \varphi(\eta^P) - \varphi(\eta^T) - \nabla_\eta \varphi(\eta^T)(\eta^P - \eta^T)$$

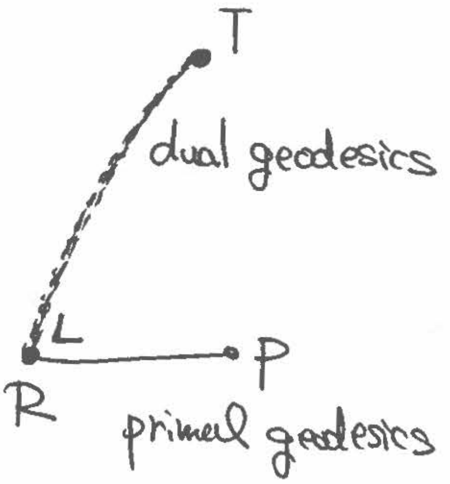
$$= D_\varphi(P,T)$$

where we used $\psi(\theta) = \theta\cdot\eta - \varphi(\eta)$.

$\blacksquare$

Pythagorean theoream

$$D_\varphi(T,P) = D_\varphi(T,R) + D_\varphi(R,P)$$



proof.

$$D_\varphi(T,R) + D_\varphi(R,P) - D_\varphi(T,P)$$

$$= \varphi(\eta^T) - \varphi(\eta^R) - \theta^R \cdot (\eta^T - \eta^R)$$
$$+ \varphi(\eta^R) - \varphi(\eta^P) - \theta^P \cdot (\eta^R - \eta^P)$$
$$- \varphi(\eta^T) + \varphi(\eta^P) + \theta^P (\eta^T - \eta^P)$$

$$= \varphi(\eta^T) - \varphi(\eta^P) - \theta^R \cdot (\eta^T - \eta^R) - \theta^P \cdot (\eta^R - \eta^P)$$
$$- \varphi(\eta^T) + \varphi(\eta^P) + \theta^P (\eta^T - \eta^P)$$

$$= \theta^P \eta^T - \theta^P \eta^P - \theta^R \eta^T + \theta^R \eta^R - \theta^P \eta^R + \theta^P \eta^P$$

$$= (\theta^P - \theta^R)(\eta^T - \eta^R) = 0$$

Because $\langle \dot{\eta}^{RT}(0), \dot{\theta}^{RP}(0) \rangle = 0$

↑ orthogonality at point R.