統計数理研 2013-9-26

情報幾何入門

理研 脳科学総合研究センター

甘利俊一

Manifold of Probability Distributions

Fisher metric : 1929 H. Hotelling
1945 C. R. Rao

Invariance connections : 1972 N. N. Chentsov

Curvature : 1978 B. Efron (A. P. Dawid)

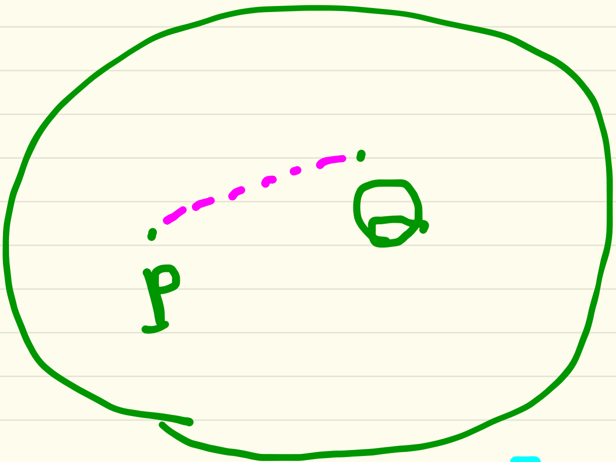
Duality : 1982 S. Amari
H. Nagaoka & S. Amari

.... many others G. Pistone J. L. Koszul

applications

Manifold and Divergence

$\boxed{\text{Dually Flat Structure}}$

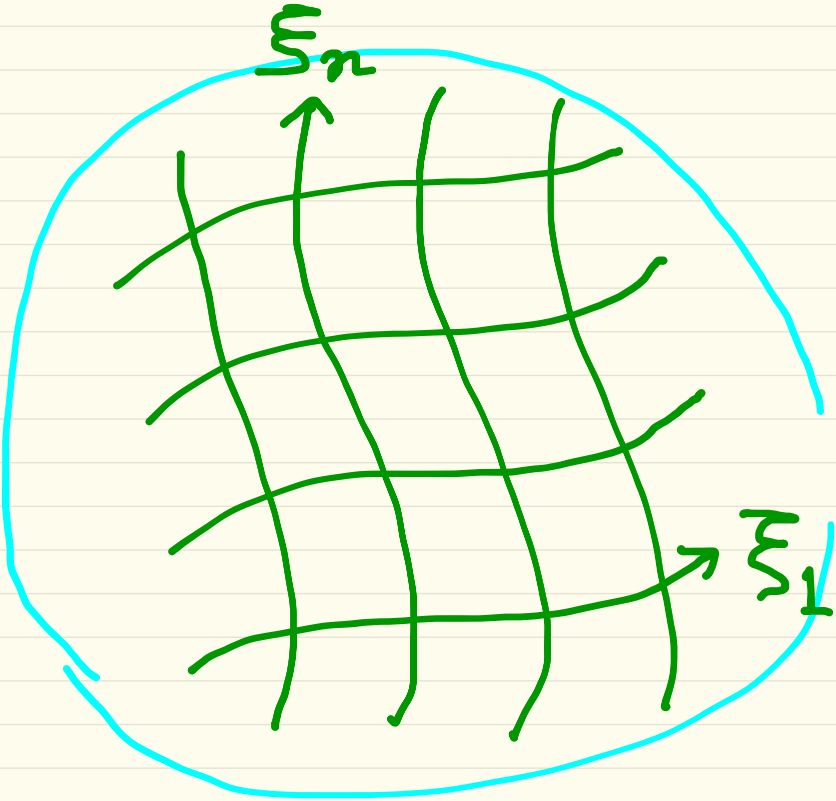


$D[P:Q] \geq 0$

distance

$D[P:Q] \neq D[Q:P]$

Manifold



coordinate system

$$\xi = (\xi_1, \cdots, \xi_n)$$

Euclidean space

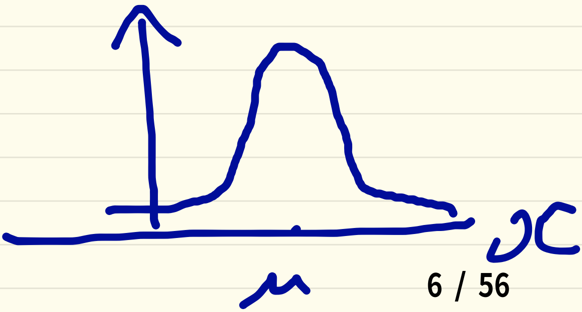
Sphere

Probability Distributions

Gaussian distribution

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

$$\xi = (\mu, \sigma)$$



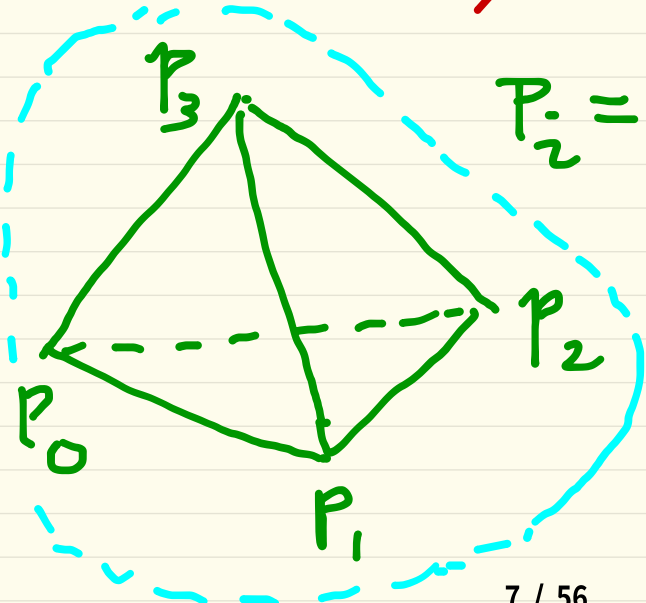
Probability Simplex S_n

$$S_n = \{ p(x) \}$$

$$x = \{0, 1, 2, \cdots, n\}$$

$$\mathbb{P} = (p_0, \cdots, p_n)$$

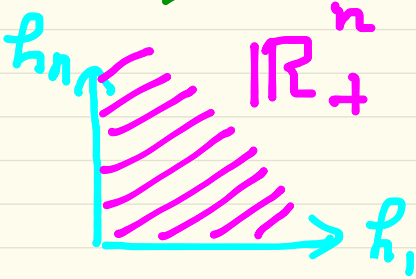
$$p_i = \text{Prob}\{x = i\}$$



S_3

histgram

$\mathbb{h} = (h_1, \cdots, h_n), \quad h_i > 0$



$\mathbb{R}_+^n$

vision

$S(x, y)$

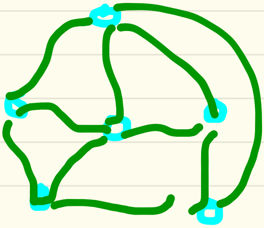
$S_{ij} = S(i, j)$

$\mathbb{R}_+^{n^2}$

Positiv-definite matrix

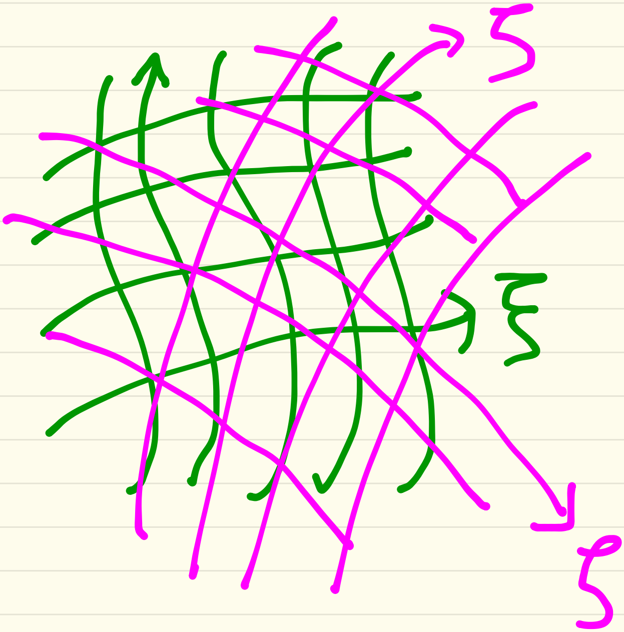
$\{P\}$ $\quad P_{ij}, \ (i \le j)$

Neural networks



$\{w_{ij}\}$

Coordinate transformation



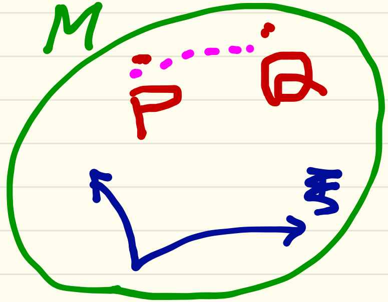
$$\xi = f(\zeta)$$

$\left\{ \begin{array}{l} \text{invariant} \\ \text{convenient} \end{array} \right.$

$$\xi_i = f_i(\zeta_1, \cdots, \zeta_n)$$

Manifold with Divergence

$D[P:Q] \qquad D[\xi_P : \xi_Q]$



1) $D[P:Q] \geq 0$; 0 iff $P=Q$

2) $D[P:P+dP] = \frac{1}{2} g_{ij}\, d\xi^i d\xi^j$

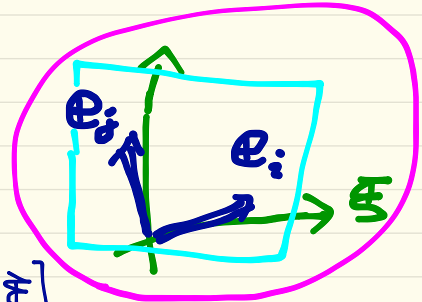
Positive-definite

$D[P:Q] \neq D[Q:P]$

in general

Riemannian metric

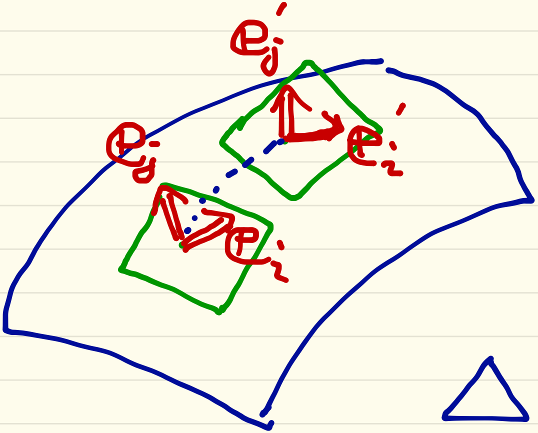
$$e_i = \frac{\partial}{\partial \xi_i} \qquad \langle e_i, e_j \rangle = g_{ij}$$



$$ds^2 = g_{ij}\, d\xi_i\, d\xi_j = \frac{1}{2} D[\xi : \xi + d\xi]$$

T_{ijk} : cubic, symmetric

アファイン接続

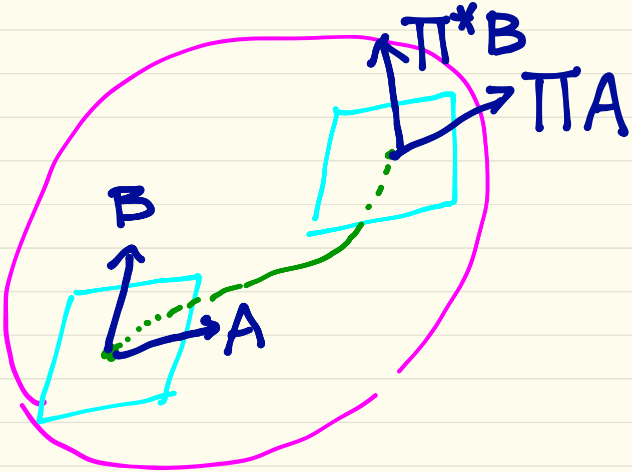


$$e_i' \sim e_i$$

$$\Delta e_i = \Pi e_i' - e_i$$

$$\nabla_{e_j} e_i = \Gamma_{ji}{}^{k} e_k$$

Dual connections (Γ, Γ^*) (∇, ∇^*)



Parallel transport

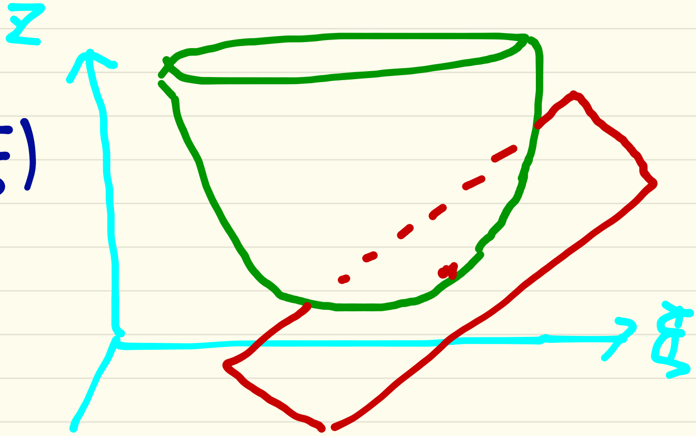
ΠA, $\Pi^* B$

$$\langle A, B \rangle = \langle \Pi A, \Pi^* B \rangle$$

$(\langle A, B \rangle = \langle \Pi A, \Pi B \rangle)$: Levi-Civita.

Convex Function

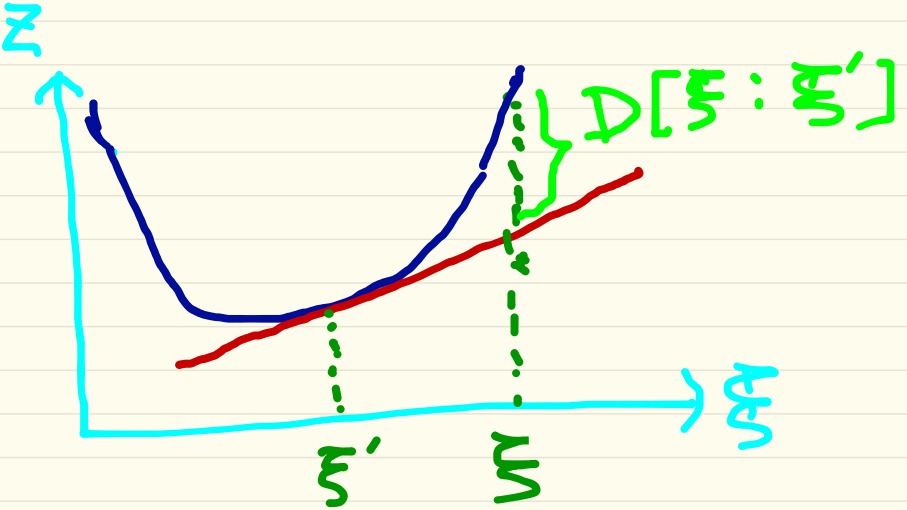
$z = \psi(\xi)$



$$\psi(\alpha \xi_1 + (1-\alpha)\xi_2) < \alpha \psi(\xi_1) + (1-\alpha)\psi(\xi_2)$$

$$a_{ij} = \frac{\partial^2 \psi(\xi)}{\partial \xi_i \, \partial \xi_j} > 0$$

Bregman Divergence



$$z = \psi(\xi') + \nabla\psi(\xi')\cdot(\xi - \xi')$$

$$D[\xi : \xi'] = \psi(\xi) - \psi(\xi') - \nabla\psi(\xi')\cdot(\xi - \xi')$$

$$g_{ij} = \frac{\partial^2 \psi(\xi)}{\partial \xi_i \partial \xi_j} > 0$$

Examples

$$\psi(\xi) = \frac{1}{2}\sum \xi_i^2$$ $\Rightarrow$ Euclid

$$\psi(\xi) = -\sum \log \xi_i$$

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

KL-divergence

$$D[p(x) : q(x)]$$

$$= \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

指数型分布族

曲指数型分布族

$$p(x, \theta) = \exp\{\theta \cdot x - \psi(\theta)\}$$

$\psi(\theta)$: 凸関数

$$g_{ij} = \frac{\partial^2}{\partial\theta^i \partial\theta^j} \psi(\theta)$$

$$T_{ijk} = \frac{\partial^3}{\partial\theta^i \partial\theta^j \partial\theta^k} \psi(\theta)$$

関数空間

flat-divergence

$\Rightarrow$ exponential family

Banerjee et al.

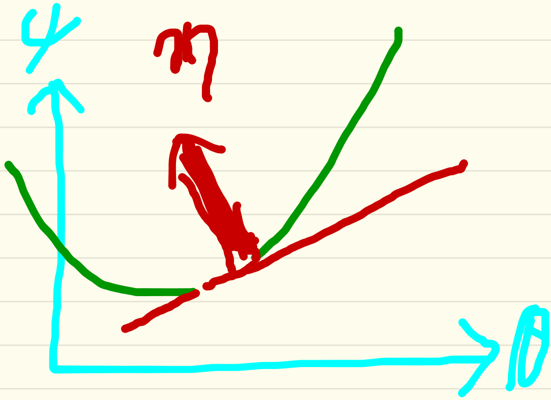
$$p(x, \theta) = \exp\{\theta \cdot x - \psi(\theta)\}$$

$$\eta = \nabla\psi(\theta) = E[x]$$

$$D[\eta : \eta'] = \psi(\theta) + \varphi(\eta') - \theta \cdot \eta \qquad \theta(\eta)$$

$$p(x, \theta) = \exp\{-D[x : \eta]\} b(x)$$

Legendre Transformation



$$\eta = \nabla\psi(\theta)$$

$$\theta = \nabla\varphi(\eta)$$

$$\underline{\theta \Leftrightarrow \eta}$$

$$\psi(\theta) + \varphi(\eta) - \theta\cdot\eta = 0$$

$$D[\theta : \theta'] = \psi(\theta) + \varphi(\eta') - \theta\cdot\eta'$$

$$\varphi(\eta) = \max_{\theta}\{\theta\cdot\eta - \psi(\theta)\}$$

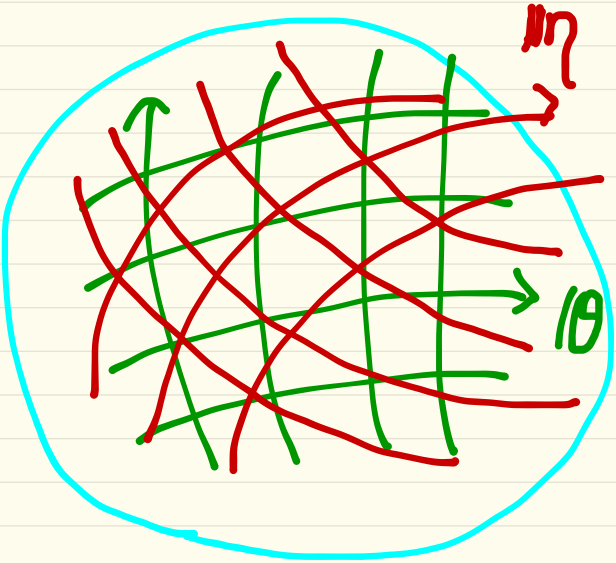
Affine Coordinates

flatt (θ, η)

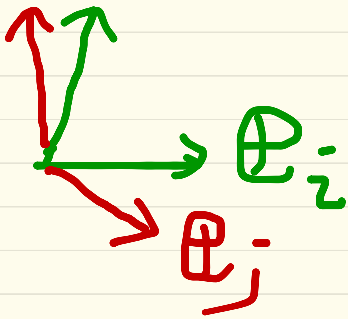
θ : flat

η : dual flat

biorthogonal



$$\langle e_i , e^{*j} \rangle = \delta_{ij}$$



Riemmanian metric

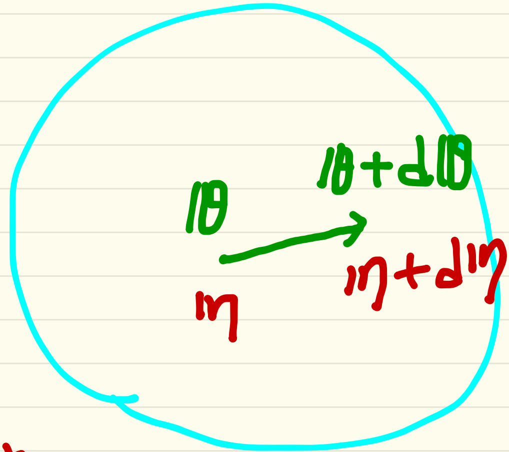
$$d\theta = \sum d\theta_i e_i$$

$$d\eta = \sum d\eta_i e_i^*$$

$$ds^2 = \langle d\theta, d\theta \rangle$$

$$= \sum g_{ij}\, d\theta_i\, d\theta_j$$

$$g_{ij} = \langle e_i, e_j \rangle$$



$$ds^2 = \sum g^*_{ij}\, d\eta_i\, d\eta_j$$

$$g^*_{ij} = \langle e_i^*, e_j^* \rangle$$

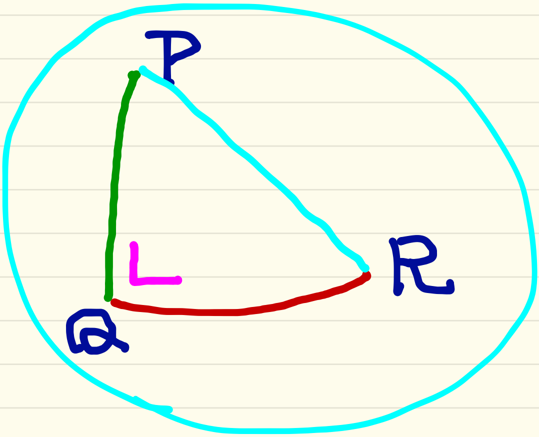
$$G = G^{*-1}$$

Pythagorean Theorem

dual flat

— geodesic

— dual geodesic



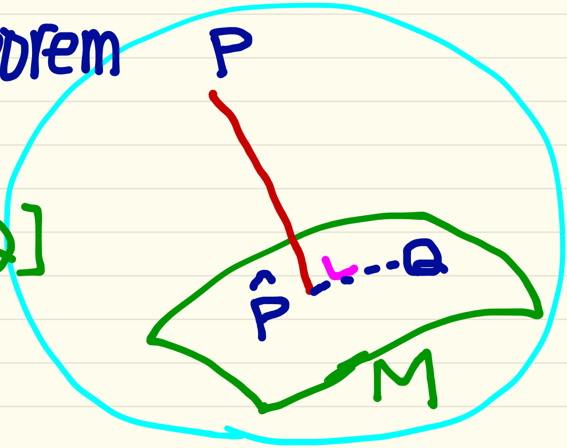
$$D[P:Q] + D[Q:R] = D[P:R]$$

Euclidean space

$$\psi(\theta) = \frac{1}{2} \sum \theta_i^2 \quad : \quad \eta_i = \theta_i$$

Proof

Projection Theorem



$$\hat{P} = \underset{Q \in M}{\arg\min}\, D[P:Q]$$

$$D^*[P:Q] = D[Q:P]$$

$\varphi(\eta)$

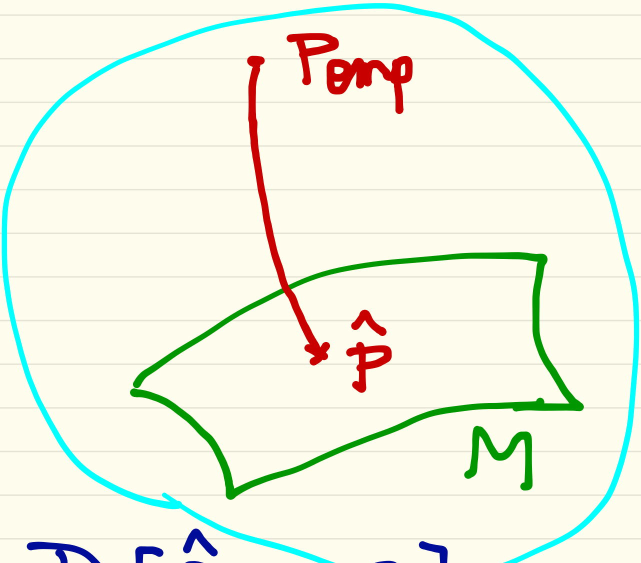
$$\hat{P}^* = \underset{Q \in M}{\arg\min}\, D[Q:P]$$

Applications :

Statistical Inference

$M = \{p(x, \xi)\}$

$x_1, \dots, x_N$



$$\hat{\xi} = \underset{Q = p(x, \xi)}{\operatorname{argmin}} \, D[\hat{P}_{emp} : Q]$$

D : KL-divergence : MLE (maximum likelihood estimator)

Invariance :

Probability distributions

$D[p(x):q(x)]$ Information monotone

$y = k(x)$

$p(x) \to \bar{p}(y)\,dy = p(x)\,dx$

$D[p(x):q(x)] \geq \bar{D}[\bar{p}(y):\bar{q}(y)]$

y : sufficient statistics

$\iff D = \bar{D}$

$p(x,\xi) = \bar{p}(y,\xi)\,r(x)$

f-divergence

$f(1) = f'(1) = 0, \quad f''(1) = 1$

$f(u)$: convex function

$$D_f[p(x) : q(x)] = \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx$$

invariant divergence

$$D_f[q(x) : p(x)] = D_{f^*}[p(x) : q(x)]$$

$$f^*(u) = u f\left(\frac{1}{u}\right)$$

$f = -\log u \qquad$ KL-divergence

S_n : discrete

$\mathbb{P} = (p_0, p_1, \cdots, p_n)$. $\quad p_i = \mathrm{Prob}\{x = i\}$

$$\circ\circ\circ \cdots\cdots \circ \qquad x$$

$$\underbrace{\circ\circ\circ}_{B_1} \quad \underbrace{\quad}_{B_2} \qquad \underbrace{\circ}_{B_m} \qquad y$$

coarse graining

$\mathbb{P} \longrightarrow \overline{\mathbb{P}}$

$p(x) \longrightarrow \overline{p}(y)$

$$\overline{p}_\alpha = \mathrm{Prob}\{y = \alpha\}$$
$$= \mathrm{Prob}\{x \in B_\alpha\}$$
$$= \sum_{i \in B_\alpha} p_i$$

Information monotone

$$D[p:q] \geq \bar{D}[\bar{p}:\bar{q}]$$

$y = k(x)$ sufficient

$$\mathrm{Prob}_p\{x \mid y \in B_\alpha\} = \mathrm{Prob}_q\{x \mid y \in B_\alpha\}$$

$$D[p:q] = \bar{D}[\bar{p}:\bar{q}]$$

decomposable divergence

$$D[p:q] = \sum k(p_i, q_i)$$

Theorem decomp. invariant d

$$: D_f[p:q] = \sum p_i f\left(\frac{q_i}{p_i}\right)$$

Flat & Invariant Divergence

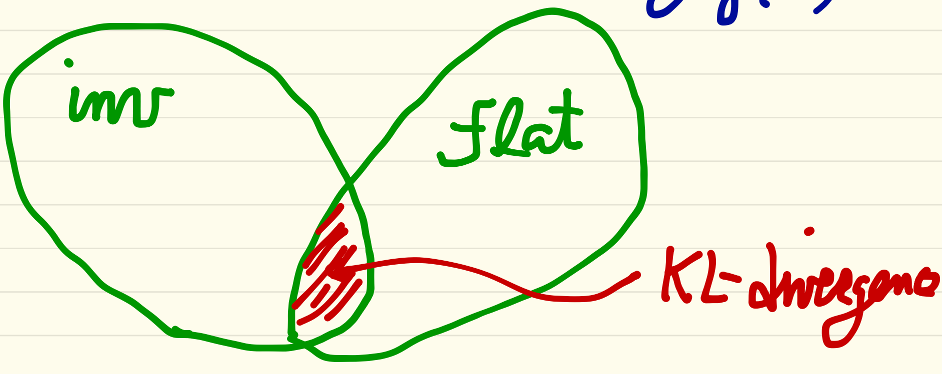
↑ Bregman

↑ f-

$$\begin{cases} \theta = \log \frac{p_i}{p_0} \\ \eta = p_i \end{cases}$$

KL-divergence

$$D_{KL}[p : q] = \sum p(x) \log \frac{p(x)}{q(x)}$$



α-divergence : invariant

$$f_\alpha(u) = \frac{-4}{1-\alpha^2} u^{\frac{1+\alpha}{2}}$$

$$D_\alpha[p : q] = \sum_i \left\{ \frac{1-\alpha}{2} p_i + \frac{1+\alpha}{2} q_i - p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}} \right\}$$

$$= 1 - \sum p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}} \quad : \text{p.d.}$$

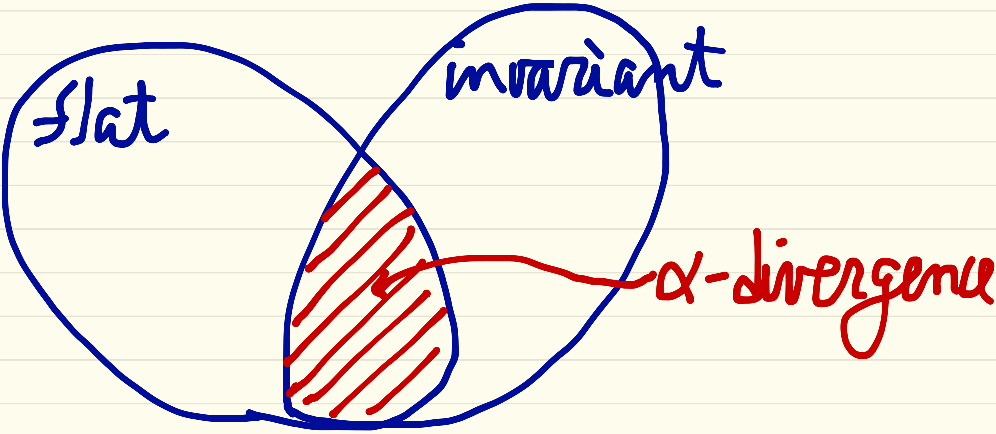
$\alpha = -1$ KL-divergence

$\alpha = 1$ dual KL

$\alpha = 0$ Hellinger

$$\frac{1}{2} \sum_i \left(\sqrt{p_i} - \sqrt{q_i} \right)^2$$

Flat & Invariant Div.
in $\mathbb{R}^n_+$: positive measure



$$\theta_i = p_i^{\frac{1+\alpha}{2}}, \quad \eta_i = p_i^{\frac{1-\alpha}{2}}$$

$$D_\alpha[p : q] = D_{-\alpha}[q : p]$$

$$\psi(\theta) = \sum \theta_i^{\frac{2}{1+\alpha}}, \quad \varphi(\eta) = \sum \eta_i^{\frac{2}{1-\alpha}}$$

$\alpha \longleftrightarrow -\alpha$ duality

α-structure

α-mean

α-family of prob. distributions

α-projection

α-optimality

Tsallis q-entropy

$$H(p) = \frac{1}{1-q}\left(\sum p_i^q - 1\right)$$

$$\alpha = 2q - 1$$

$$p_i^{\frac{1+\alpha}{2}}$$

α-mean

$x, y > 0$

$$m_f(x,y) = f^{-1}\left(\frac{f(x)+f(y)}{2}\right)$$

scale-free

$$m_f(cx, cy) = c\, m_f(x,y)$$

$$f_\alpha(u) = \begin{cases} u^{\frac{1-\alpha}{2}} \\ \log u, & \alpha = 1 \end{cases}$$

$\alpha = 1$: geometric mean $\sqrt{xy}$

$\alpha = -1$: arithmetic mean $\frac{x+y}{2}$

$\alpha = 0$: $\frac{1}{2}\left(\frac{1}{2}(x+y) + \sqrt{xy}\right)$

$\alpha = 3$: harmonic mean $\frac{2}{\frac{1}{x}+\frac{1}{y}}$

$\alpha = \infty$: $\min\{x, y\}$

fuzzy

$\alpha = -\infty$ $\max\{x, y\}$

$$m_\alpha(x,y) \geq m_{\alpha'}(x,y), \quad \alpha \leq \alpha'$$

pessimistic mean : optimistic mean

α-family of Prob. distr.

$\delta_1(x), \cdots, \delta_m(x)$

$\Rightarrow$

$$P_\alpha(x; w) = c f_\alpha^{-1}\left\{\sum_{i=1}^{m} w_i f_\alpha(\delta_i(x))\right\}$$

$\alpha = -1$ $\quad P_\alpha(x) = \sum w_i \delta_i(x)$

mixture family

$\alpha = 1$ $\quad P_\alpha(x) = \exp\left\{\sum w_i \log \delta_i(x) - \psi\right\}$

exp. family

α-integration of $q_1(x), \cdots, q_m(x)$

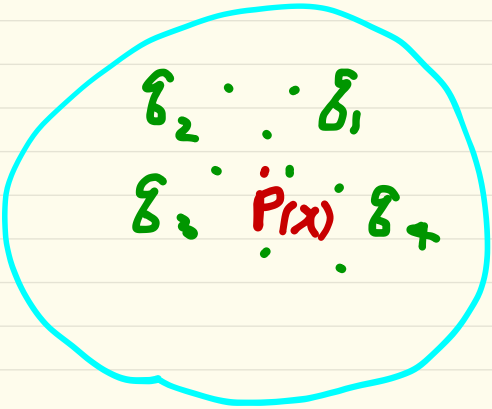
$$p(x) = f_\alpha^{-1}\left\{ \sum w_i f_\alpha\{p_i(x)\} \right\}$$

$$R[p(x)] = \sum w_i D[q_i(x) : p(x)]$$

min $R_\alpha[p(x)]$

: α-integration

mixture of experts



flat divergence (non-invariant)

(α,β)-divergence : $\mathbb{R}_+^n$

$$D_{\alpha\beta}[P:Q] = \frac{1}{\alpha\beta(\alpha+\beta)} \sum \left\{ \alpha p_i^{\alpha+\beta} + \beta q_i^{\alpha+\beta} - (\alpha+\beta) p_i^{\alpha} q_i^{\beta} \right\}$$

$$\theta_i = \frac{1}{\alpha} p_i^{\alpha}, \quad \eta_i = \frac{1}{\beta} p_i^{\beta}$$

$$\psi(\theta) = c \sum \theta_i^{\frac{\alpha+\beta}{\alpha}}, \quad \varphi(\eta) = c \sum \eta_i^{\frac{\alpha+\beta}{\beta}}$$

(U, V)-divergence

$u(s), v(s)$: monotone incr.

$$\theta_i = U(p_i), \quad \eta_i = V(p_i)$$

$$\psi(\theta) = \sum \int^{p_i} v(p) u'(p) \, dp$$

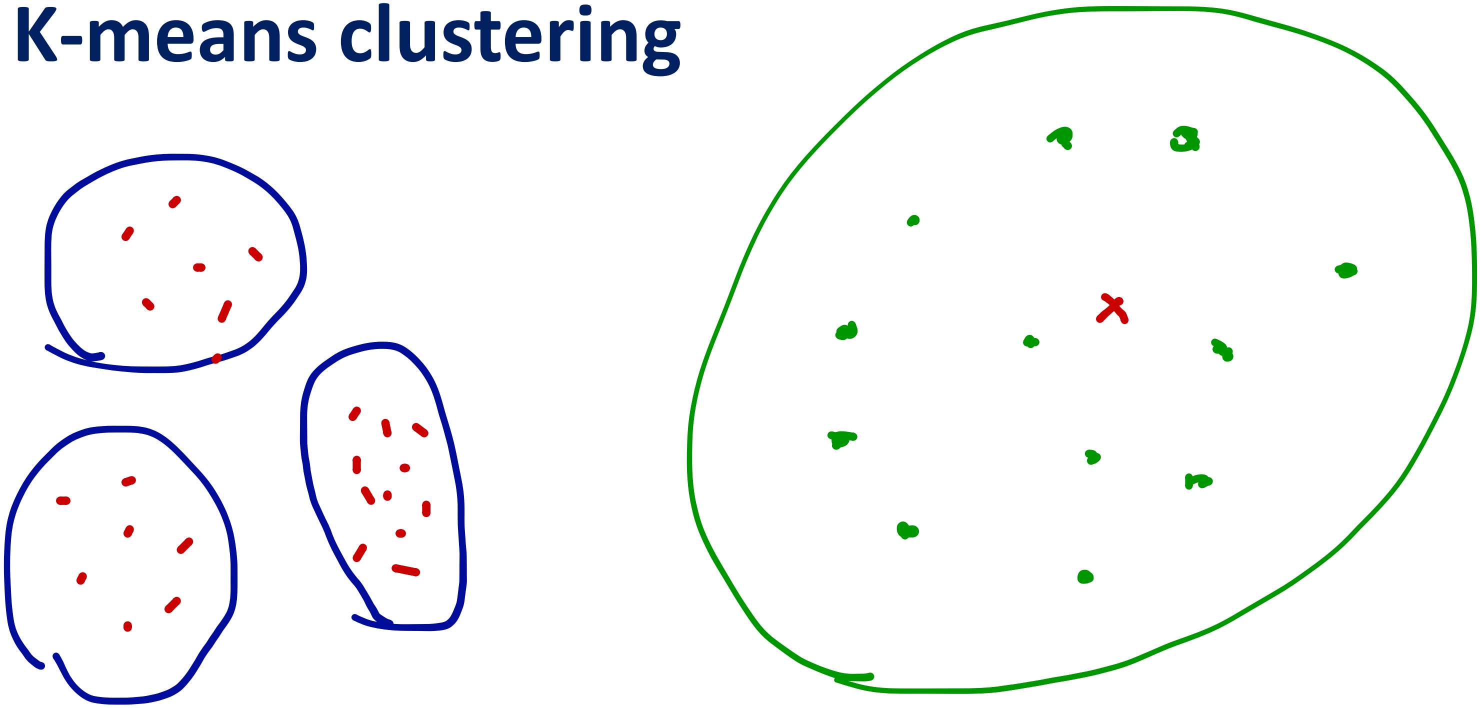
$$\varphi(\eta) = \sum \int^{p_i} u(p) v'(p) \, dp$$

$$D[p : q] = \sum_i \Big[\int^{p_i} v(p) u'(p) \, dp + \int^{q_i} u v' \, dp - u(p_i) v(q_i) \Big]$$

Center of a cluster

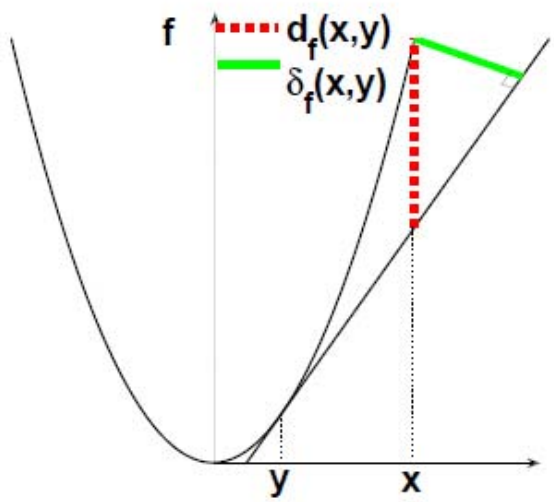
$$\boldsymbol{x}^* = \arg\min \sum_i D[\boldsymbol{x}, \boldsymbol{x}_i]$$

K-means clustering



Total Bregman Divergence

$$TD[\boldsymbol{x}:\boldsymbol{y}] = \frac{D[\boldsymbol{x}:\boldsymbol{y}]}{\sqrt{1+\|\nabla\psi\|^2}}$$



- **rotational invariance**
- **conformal geometry**

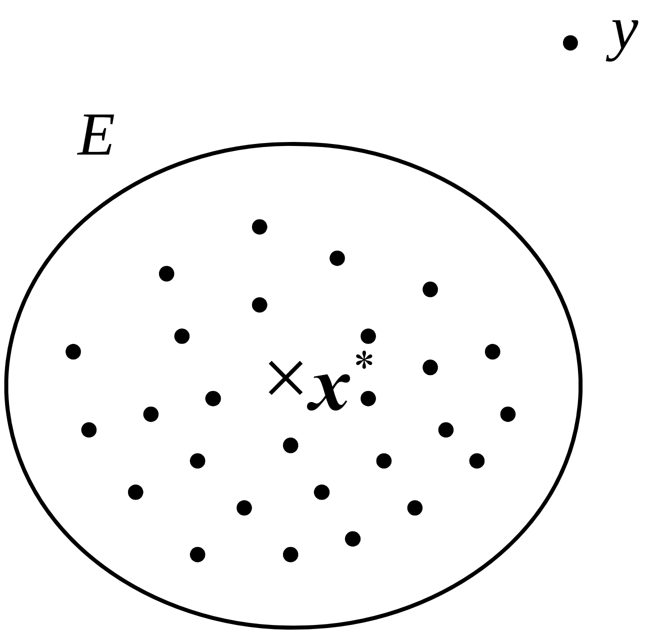
Figure: $d_f(x,y)$ (dotted red line) is BD, $\delta_f(x,y)$ (bold green line) is TBD, and the two arrows indicate the coordinate system.Note that $d_f(x,y)$ changes with rotation unlike $\delta_f(x,y)$ which is invariant to rotation.

Clustering : *t*-center

$$E = \{x_1, \cdots, x_m\}$$

T-center of E

$$\boldsymbol{x}^* = \arg\min \sum_i TD[\boldsymbol{x}, \boldsymbol{x}_i]$$



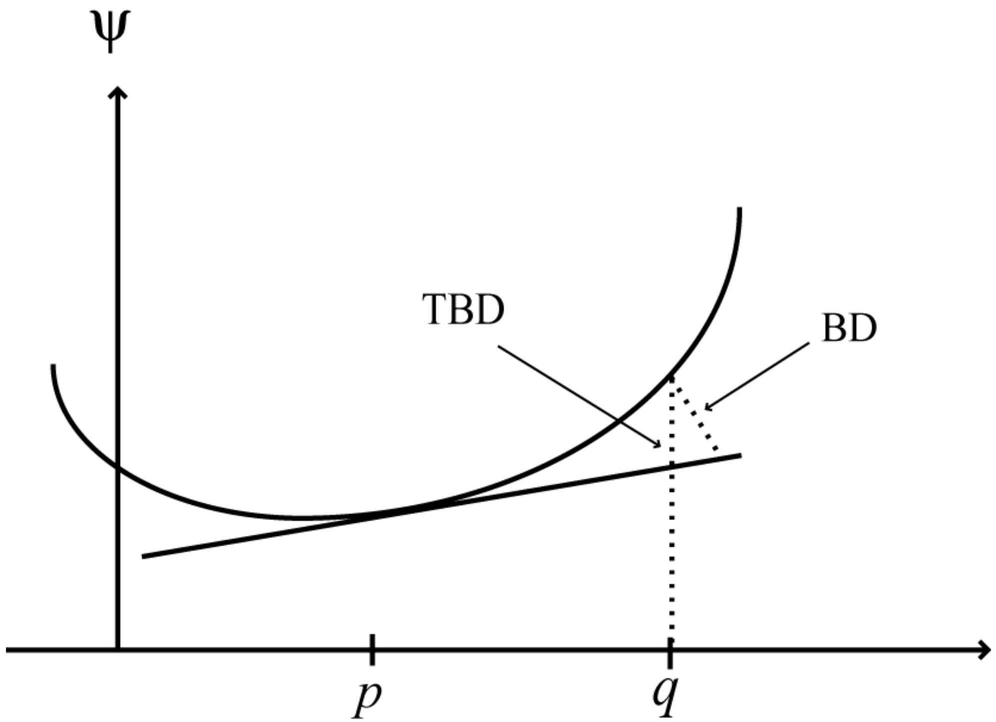
***t*-center $\boldsymbol{x}^*$**

$$\nabla \psi\left(\boldsymbol{x}^*\right) = \frac{\sum w_i \nabla \psi\left(\boldsymbol{x}_i\right)}{\sum w_i}$$

$$w_i = \frac{1}{\sqrt{1 + \left\| \nabla \psi\left(\boldsymbol{x}_i\right) \right\|^2}}$$

Total Bregman divergence (Vemuri)

$$\mathrm{TBD}(p:q) = \frac{\varphi(p) - \varphi(q) - \nabla\varphi(q)\cdot(p-q)}{\sqrt{1 + |\nabla\varphi(q)|^2}}$$



Conformal change of divergence

$$\tilde{D}\left(p:q\right)=\sigma\left(p\right)D\left[p:q\right]$$

$$\tilde{g}_{ij}=\sigma\left(p\right)g_{ij}$$

$$\tilde{T}_{ijk}=\sigma(T_{ijk}+s_k g_{ij}+s_j g_{ik}+s_i g_{jk})$$

$$s_i=\partial_i \log \sigma$$

t-center is robust

$$E^{*} = \left\{ \boldsymbol{x}_1, \cdots, \boldsymbol{x}_n ; \boldsymbol{y} \right\}$$

$$\tilde{\boldsymbol{x}}^{*} = \boldsymbol{x}^{*} + \varepsilon \boldsymbol{z}\left(\boldsymbol{x}^{*} ; \boldsymbol{y} \right), \quad \varepsilon = \frac{1}{n}$$

influence function $\boldsymbol{z}\left(\boldsymbol{x}^{*} ; \boldsymbol{y} \right)$

$$\left| \boldsymbol{z} \right| < c \quad \text{as} \quad \left| \boldsymbol{y} \right| \to \infty \quad : \text{robust}$$

Positive-Definite Matrices

(α, β)-divergence in $\mathcal{P} = \{P > 0\}$

$$D_{\alpha,\beta}[P : Q] = \mathrm{tr}\left\{ \frac{\alpha}{\alpha+\beta} P^{\alpha+\beta} + \frac{\beta}{\alpha+\beta} Q^{\alpha+\beta} - P^{\alpha} Q^{\beta} \right\}$$

$D_{u,v}$

flat divergence in S_n

: escort probability & conformal geometry

$$h_\alpha(\mathbb{P}) = \sum p_i^{\frac{1+\alpha}{2}}$$

(α-exp. fam.

$$\tilde{D}_\alpha[\mathbb{P} : \mathbb{Q}] = \frac{2}{1-\alpha} \frac{1}{h_\alpha(\mathbb{Q})} \left[1 - \sum p_i^{\frac{1+\alpha}{2}} q_i^{\frac{1-\alpha}{2}}\right]$$

$$\theta_i = \frac{2}{1-\alpha} \left[p_i^{\frac{1-\alpha}{2}} - p_0^{\frac{1-\alpha}{2}} \right]$$

$$\eta_i = \frac{1}{h_\alpha(\mathbb{P})} p_i^{\frac{1+\alpha}{2}} \approx \tilde{\mathbb{P}}$$

$$\tilde{g}_{ij}(\theta) = \frac{1}{h_\alpha(\theta)} g_{ij}(\theta)$$

conformal : unique (u-escort)

Divergence and Geometry

$D[\xi : \xi']$ $\quad \nabla_\xi = \frac{\partial}{\partial \xi_i}, \ \nabla'_{\xi'} = \frac{\partial}{\partial \xi'_i}$

Riemannian Metric

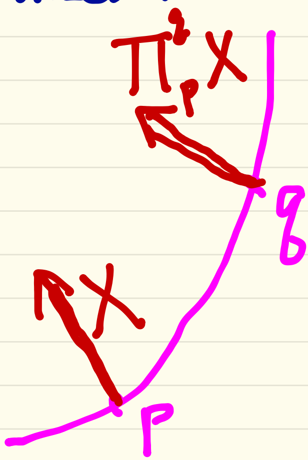
$g_{ij} = -\nabla_{\xi_i} \nabla'_{\xi'_j} D[\xi : \xi']_{\xi = \xi'}$: positive-definite

cubic tenor

$T_{ijk} = \nabla_{\xi_i} \nabla_{\xi_j} \nabla'_{\xi'_k} D - \nabla'_{\xi'_i} \nabla'_{\xi'_j} \nabla_{\xi_k} D$

$D \to \{M, g, \dot{T}\}$: T symmetric

$\{M, g\}$



covariant derivatives

$\nabla \Longleftrightarrow \nabla^*$

$\Pi_p^q X, \ \Pi_p^{*q} X$: parallel

Duall Affine Connections

$$\Gamma_{ijk} = \{i,j;k\} - \frac{1}{2} T_{ijk}$$

$$\Gamma^*_{ijk} = \{i,j;k\} + \frac{1}{2} T_{ijk}$$

$$\Gamma^\alpha_{ijk} = \{i,j;k\} - \frac{\alpha}{2} T_{ijk}$$

$\pm\alpha$ duality

$$D_Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla^*_Z Y\rangle$$

$$\langle X, Y\rangle_p = \langle \Pi_p^q X, \Pi_p^{*q} Y\rangle_q$$

Two geodesics

$\xi(t)$

$$\nabla_{\dot{\xi}} \dot{\xi}(t) = 0$$

$$\ddot{\xi}_i + \sum \Gamma_{kji} \dot{\xi}_k \dot{\xi}_j = 0$$

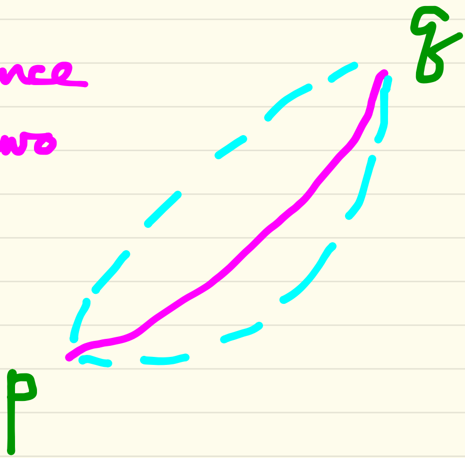
$$\nabla^*_{\dot{\xi}} \dot{\xi}(t) = 0$$

Euclidean : minimal distance
straight

Riemannian : Levi-Civita connection

dual geometry :

minimal distance
splits into two



Flat Manifold & Canonical Divergence

dually flat : $R=0 \Leftrightarrow R^*=0$

$\exists$ affine coordinates : θ, η

$\exists$ convex functions : $\psi(\theta)$, $\varphi(\eta)$

canonical divergence

$D[\theta : \theta'] = \psi(\theta) + \varphi(\eta') - \theta \cdot \eta'$

manifold of prob. distributions

invariance & flat $\Rightarrow$ KL-divergence flat-divergence

$\Rightarrow$ exponential family

Banerjee et al.

$$p(x, \theta) = \exp\{\theta \cdot x - \psi(\theta)\}$$

$$\eta = \nabla\psi(\theta) = E[x]$$

$$D[\eta : \eta'] = \psi(\theta) + \varphi(\eta') - \theta \cdot \eta \qquad \theta(\eta)$$

$$p(x, \theta) = \exp\{-D[x : \eta]\}\, b(x)$$

Le Theorem

$\{M, g, T\}$:

realization in probability model

$M = \{p(x, \xi)\}$ embedding curvature

invariance $\Rightarrow$

uniqueness of

g_{ij}, T_{ijk}

Hessian manifold (Shima)

$$g_{ij}(\xi) = \nabla_{\xi_i} \nabla_{\xi_j} \psi(\xi)$$

$$\{M, g\} \Rightarrow \{M, g, T\}$$

dually flat ?

given $g_{ij} \Rightarrow \exists\, T_{ijk}$: flatten