Machine Learning SS: Kyoto U.

Information Geometry
and Its Applications to Machine Learning

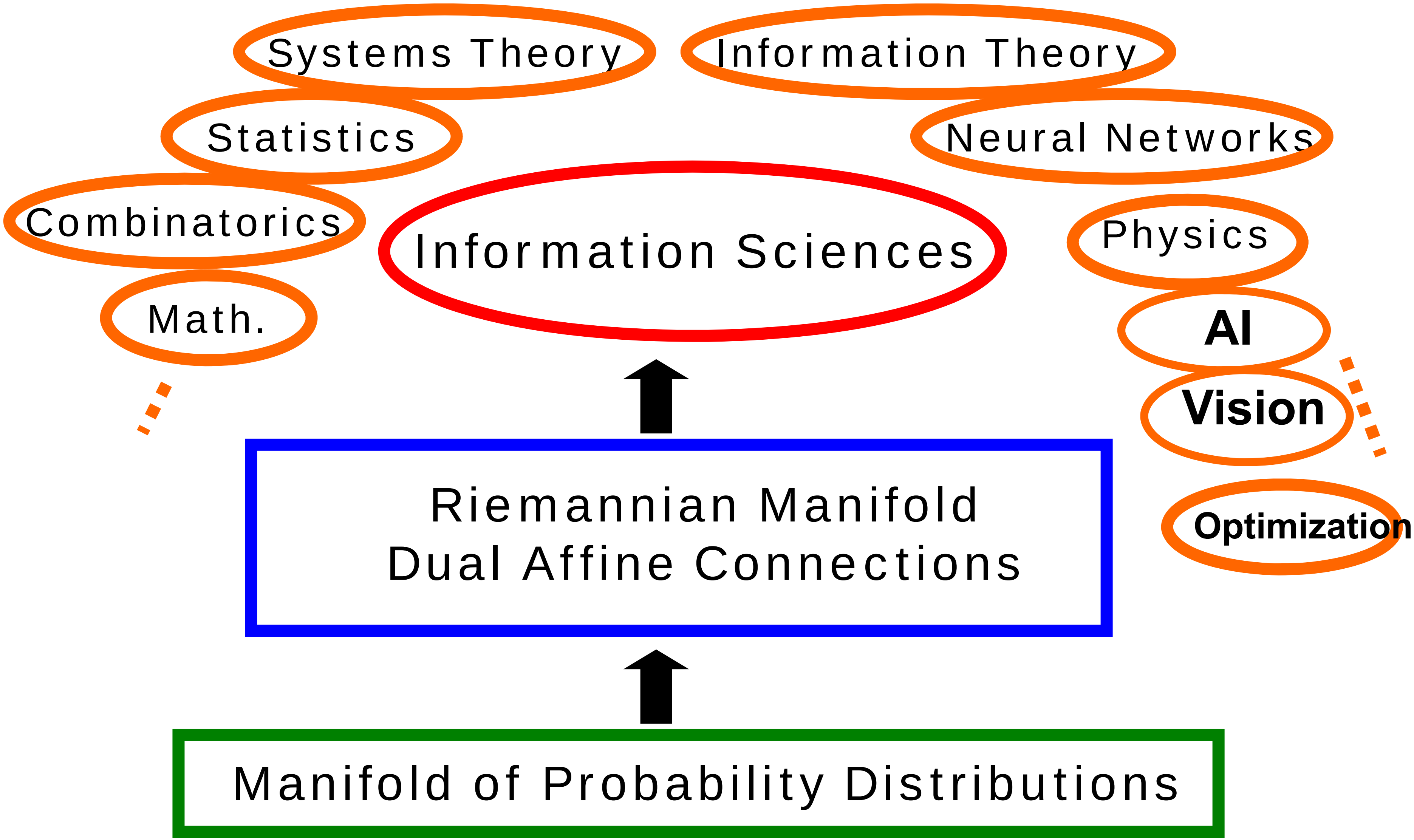
Shun-ichi Amari
RIKEN Brain Science Institute

Information Geometry

-- Manifolds of Probability Distributions

$$M = \{ p(\mathbf{x}) \}$$

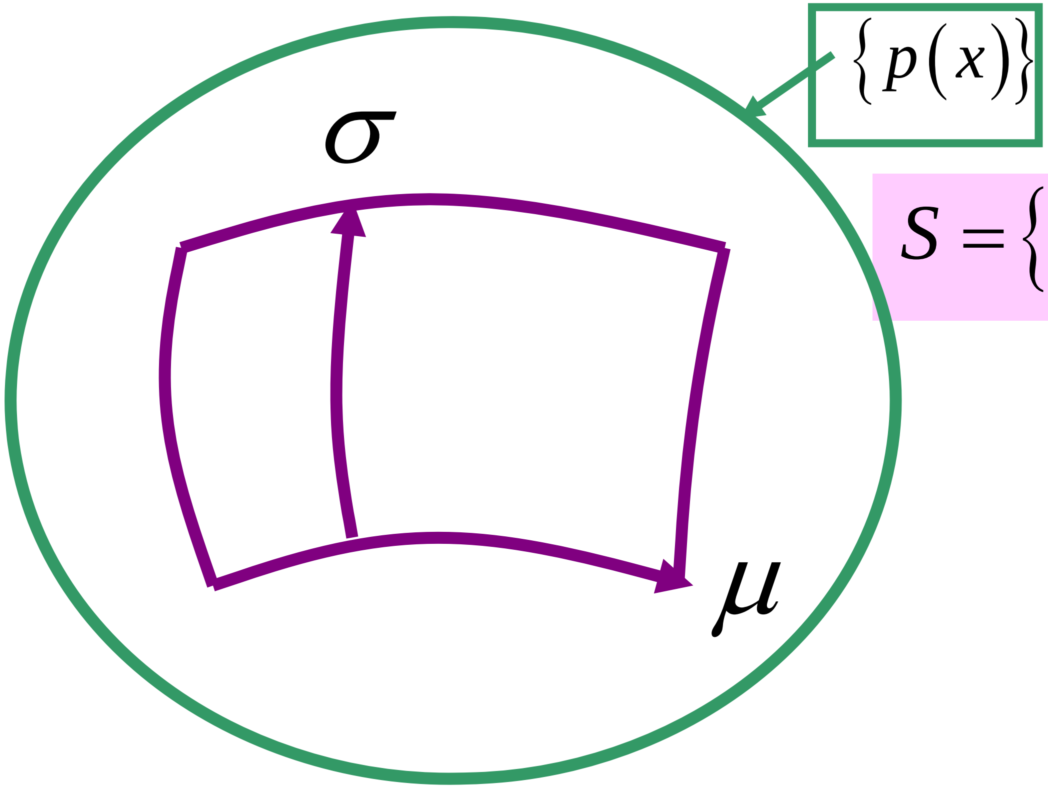
Information Geometry



Information Geometry ?

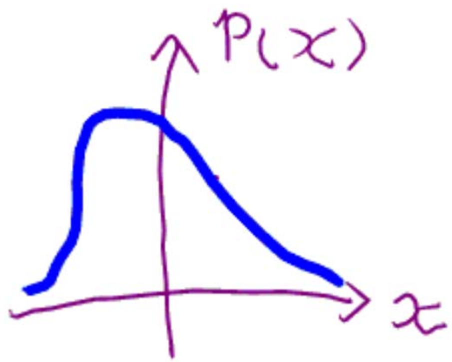
Gaussian distributions

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



$$\boldsymbol{\theta} = (\mu, \sigma)$$

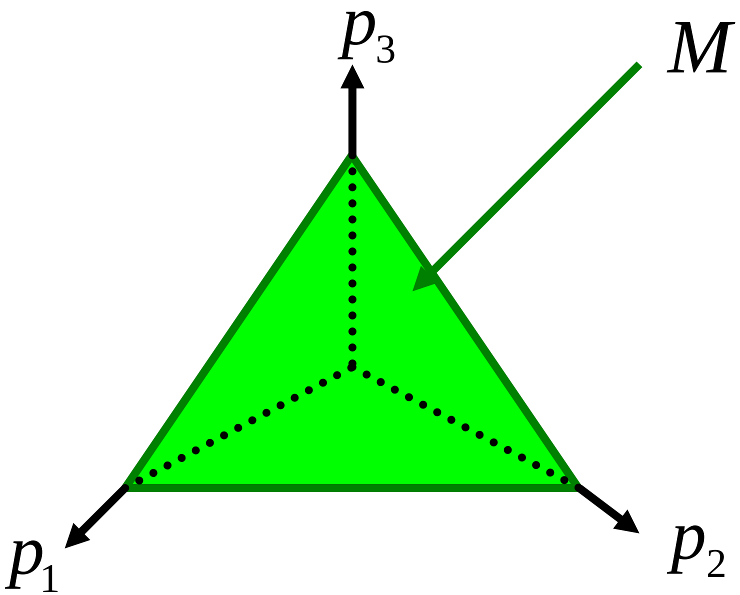
$$S = \{ p(x; \boldsymbol{\theta}) \}$$



Manifold of Probability Distributions

$$x = 1, 2, 3 \qquad S_n = \{p(x)\}$$

$$\boldsymbol{p} = (p_1, p_2, p_3) \qquad p_1 + p_2 + p_3 = 1$$



$$M = \{p(x;\theta)\}$$

Invariance

$$S = \{ p(x, \boldsymbol{\theta}) \}$$

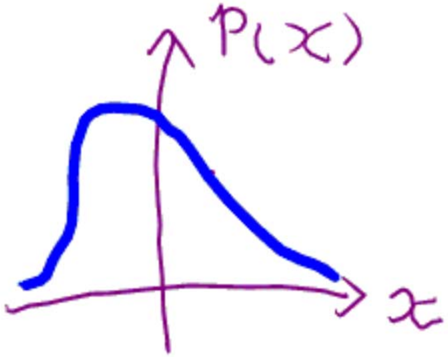
Invariant under different representation

$$y = y(x), \qquad \bar{p}(y, \boldsymbol{\theta})$$

sufficient statistics

$$\int \left| p(x, \theta_1) - p(x, \theta_2) \right|^2 dx$$

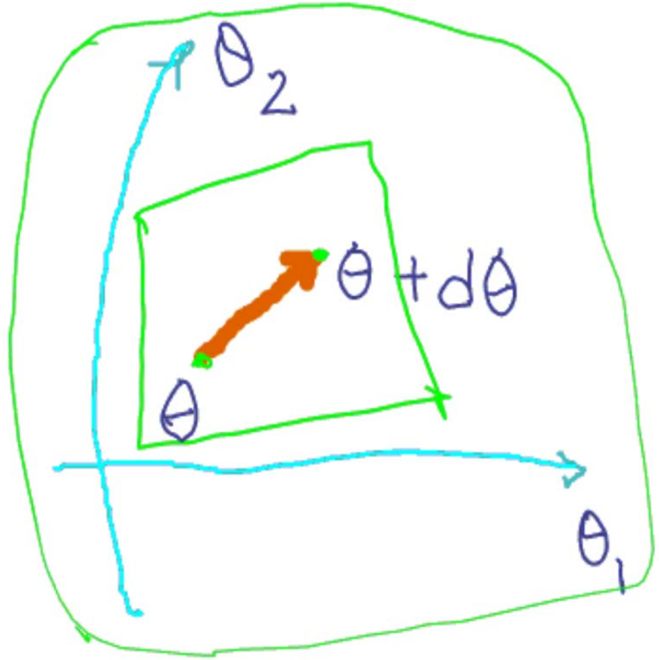
$$\neq \int \left| \bar{p}(y, \theta_1) - \bar{p}(y, \theta_2) \right|^2 dy$$



Two Geometrical Structures

Riemannian metric
affine connection --- geodesic

$$ds^2 = \sum g_{ij}(\theta) d\theta_i d\theta_j$$



Fisher information

$$g_{ij} = E\left[\frac{\partial}{\partial \theta_i}\log p \frac{\partial}{\partial \theta_j}\log p\right]$$

Orthogonality: innner product

$$< d_1\theta, d_2\theta >= d_1\theta^T G d_2\theta$$

Affine Connection

covariant derivative; parallel transport

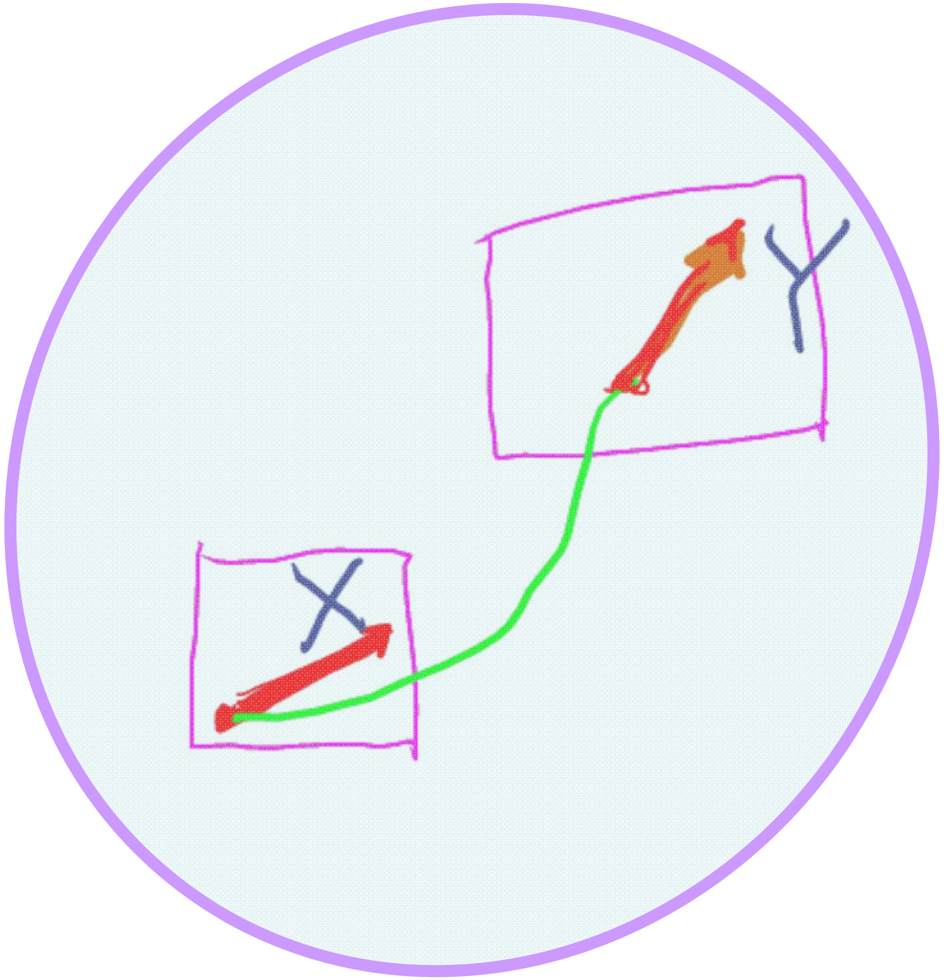
$$\nabla_X Y, \quad \Pi_c X = Y$$

geodesic $\Pi\dot{X} = \dot{X} \quad X = X(t)$

$$s = \int \sqrt{\sum g_{ij}(\theta) d\theta^i d\theta^j}$$

minimal distance

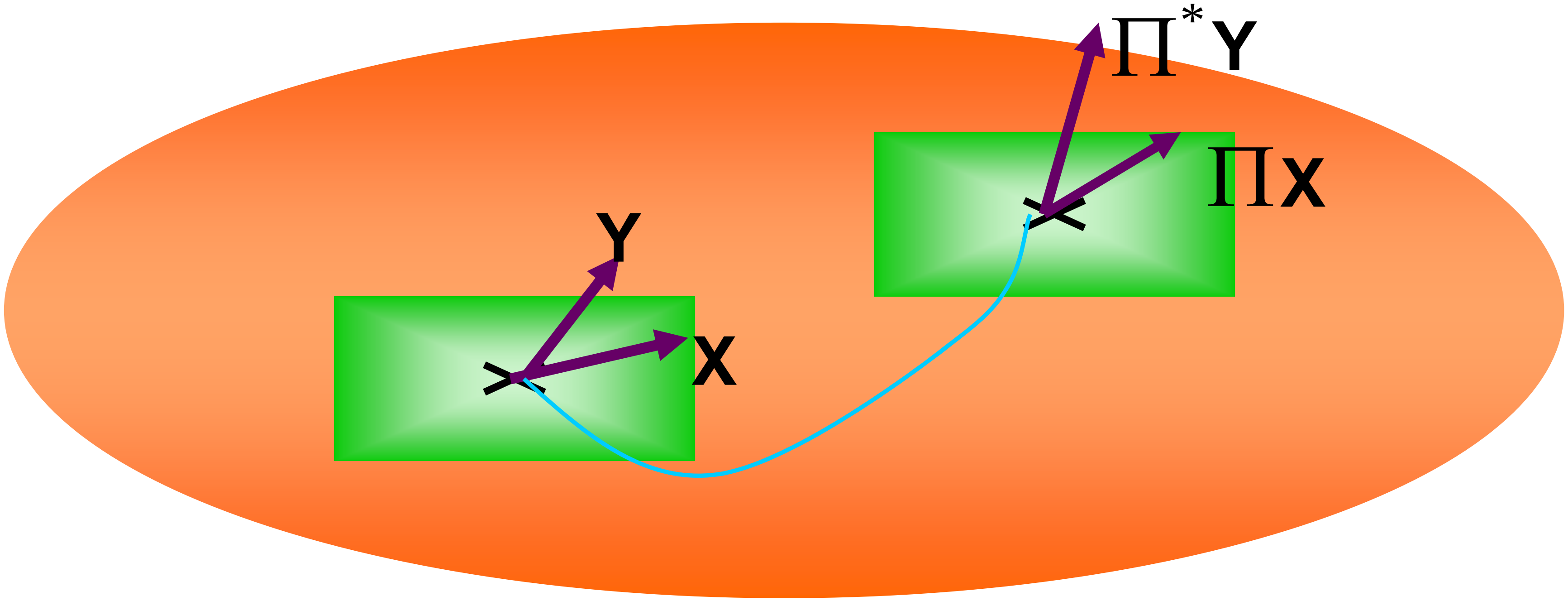
straight line



Duality: two affine connections

$$\{S, g, \nabla, \nabla^*\}$$

$$\langle X, Y \rangle = \langle \Pi X, \Pi^* Y \rangle \qquad < X, Y >= \sum g_{ij} X^i Y^j$$



Riemannian geometry: $\Pi = \Pi^*$

Dual Affine Connections

$$(\nabla, \nabla^*)$$

$$(\Pi, \Pi^*)$$

e-geodesic

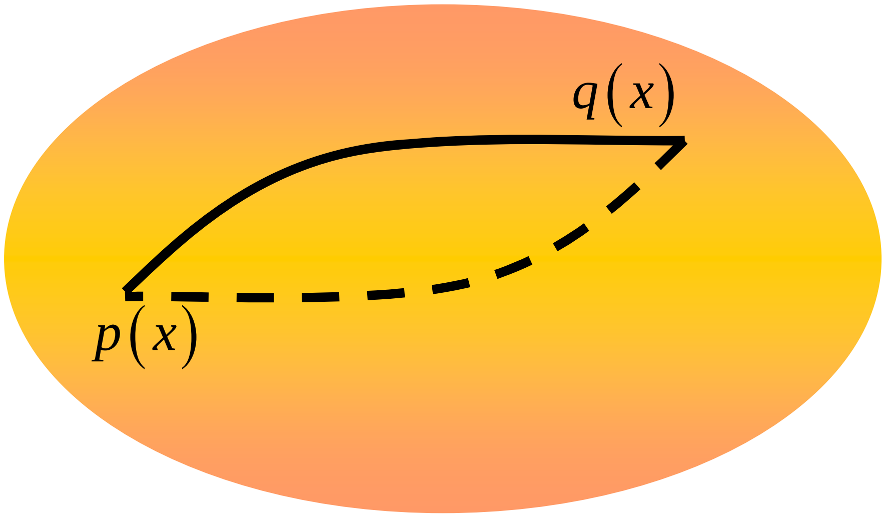
$$\log r(x,t) = t \log p(x) + (1-t)\log q(x) + c(t)$$

m-geodesic

$$r(x,t) = tp(x) + (1-t)q(x)$$

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

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



Mathematical structure of $S = \left\{ p\left(x, \xi\right) \right\}$

$$g_{ij}\left(\xi\right) = E\left[\partial_i l \partial_j l\right]$$

$$T_{ijk}\left(\xi\right) = E\left[\partial_i l \partial_j l \partial_k l\right]$$

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

$$l = \log p\left(x, \xi\right); \qquad \partial_i = \frac{\partial}{\partial \xi^i}$$

α -**connection**

$$\Gamma^{\alpha}_{ijk} = \left\{i, j; k\right\} - \alpha T_{ijk}$$

$\nabla^{\alpha} \leftrightarrow \nabla^{-\alpha}$: **dually coupled**

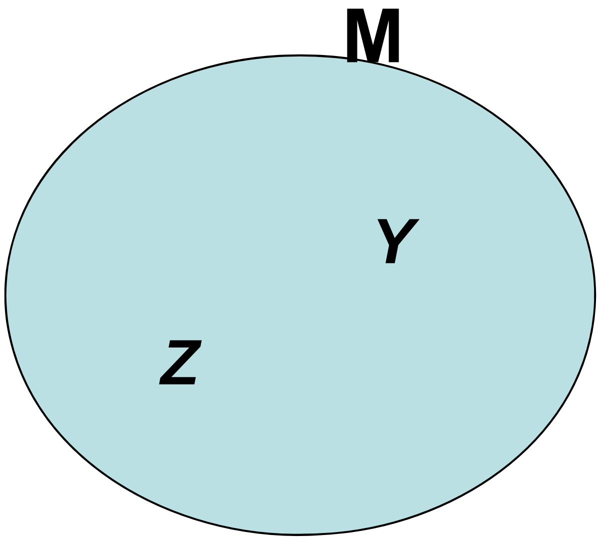
$$X\left\langle Y, Z\right\rangle = \left\langle \nabla_X Y, Z\right\rangle + \left\langle Y, \nabla^*_X Z\right\rangle$$

Divergence: $D[z:y]$

$$D[\boldsymbol{z}:\boldsymbol{y}] \geq 0$$

$$D[\boldsymbol{z}:\boldsymbol{y}] = 0, \quad \text{iff} \;\; \boldsymbol{z} = \boldsymbol{y}$$

$$D[\boldsymbol{z}:\boldsymbol{z}+d\boldsymbol{z}] = \sum g_{ij} dz_i dz_j$$

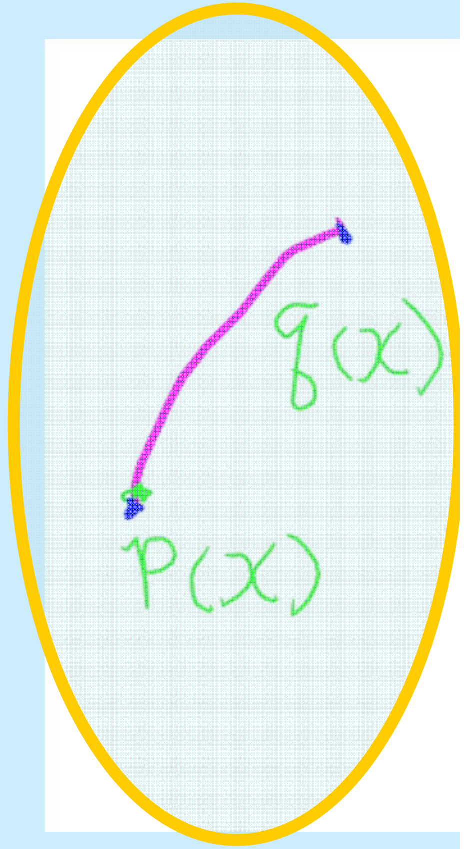


positive-definite

Kullback-Leibler Divergence

quasi-distance

$$D[p(x):q(x)] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$



$$D[p(x):q(x)] \geq 0 \quad =0 \text{ iff } p(x) = q(x)$$

$$D[p:q] \neq D[q:p]$$

f divergence of $\tilde{S}$

$$\tilde{S} = \{\tilde{\boldsymbol{p}}\}, \qquad \tilde{p}_i > 0 : (\sum \tilde{p}_i = 1 \text{ } nn \text{ holds})$$

$$D_f[\tilde{\boldsymbol{p}} : \tilde{\boldsymbol{q}}] = \sum \tilde{p}_i f\left(\frac{\tilde{q}_i}{\tilde{p}_i}\right) \geq 0 \qquad D_f[\tilde{\boldsymbol{p}} : \tilde{\boldsymbol{q}}] = 0 \Leftrightarrow \tilde{\boldsymbol{p}} = \tilde{\boldsymbol{q}}$$

not invariant under $\tilde{f}(u) = f(u) - c(u-1)$

α divergence

$$D_\alpha[\tilde{p}:\tilde{q}] = \sum \{ \frac{1-\alpha}{2}\tilde{p}_i + \frac{1+\alpha}{2}\tilde{q}_i - \tilde{p}_i^{\frac{1-\alpha}{2}} \tilde{q}_i^{\frac{1+\alpha}{2}} \}$$

KL-divergence

$$D[\tilde{p}:\tilde{q}] = \sum \{ \tilde{p}_i \log \frac{\tilde{p}_i}{\tilde{q}_i} + \tilde{p}_i - \tilde{q}_i \}$$

α

$(\alpha,\beta)-divergence$

$$D_{\alpha,\beta}[p:q]=\sum\{\frac{\alpha}{\alpha+\beta}p_i^{\alpha+\beta}+\frac{\beta}{\alpha+\beta}q_i^{\alpha+\beta}-p^{\alpha}{}_i q_i^{\beta}\}$$

$\beta=-\alpha:\quad \alpha-\text{divergence}$

$\alpha=1:\quad \beta\text{-divergence}$

Metric and Connections Induced by Divergence

Riemannian metric

(Eguchi)

$$g_{ij}\left(\boldsymbol{z}\right)=\partial_i\partial_j D\left[\boldsymbol{z}:\boldsymbol{y}\right]_{|\boldsymbol{y}=\boldsymbol{z}}: \quad D\left[\boldsymbol{z}:\boldsymbol{z}+d\boldsymbol{z}\right]=\frac{1}{2}g_{ij}\left(\boldsymbol{z}\right)dz_i dz_j$$

affine connections

$$\{\nabla,\nabla^*\}$$

$$\Gamma_{ijk}\left(\boldsymbol{z}\right)=-\partial_i\partial_j\partial_k^{'}D\left[\boldsymbol{z}:\boldsymbol{y}\right]_{|\boldsymbol{y}=\boldsymbol{z}}$$

$$\partial_i=\frac{\partial}{\partial z_i},\qquad \partial_i^{'}=\frac{\partial}{\partial y_i}$$

$$\Gamma_{ijk}^{*}\left(\boldsymbol{z}\right)=-\partial_i^{'}\partial_j^{'}\partial_k D\left[\boldsymbol{z}:\boldsymbol{y}\right]_{|\boldsymbol{y}=\boldsymbol{z}}$$

Duality: $X < Y, Z > = < \nabla_X Y, Z > + < Y, \nabla^*_X Z >$

$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma^*_{kji}$$

$$\Gamma_{ijk} = \Gamma^*_{ijk} - T_{ijk}$$

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

Dually flat manifold

exponential family; mixture family; {p(x); x discrete}

$$p(x,\theta)=\exp\left\{\sum \theta_i x_i-\psi(\theta)\right\} \ : \text{ exponential family}$$

θ-coordinates : affine coordinates, flat, geodesics

$\updownarrow$

η-coordinates: (dual) affine coordinates, flat, geodesics

not Riemannian flat

canonical divergence D(P: P') : $KL-divergence$

Dually flat manifold

potential functions $\psi(\theta), \varphi(\eta)$

$$\eta_i = \frac{\partial}{\partial \theta_i}\psi(\theta); \quad \theta_i = \frac{\partial}{\partial \eta_i}\varphi(\theta):$$ **Legendre transformation**

$$\psi(\theta) + \varphi(\eta) - \sum \theta_i \eta_i = 0$$

$$g_{ij}(\theta) = \frac{\partial^2}{\partial \theta_i \partial \theta_j}\psi(\theta) \cdots g^{ij} = \frac{\partial^2}{\partial \eta_i \partial \eta_j}\varphi(\theta)$$

$p(x,\theta) = \exp\left\{\sum \theta_i x_i - \psi(\theta)\right\}$: exponential family

ψ : cumulant generating function

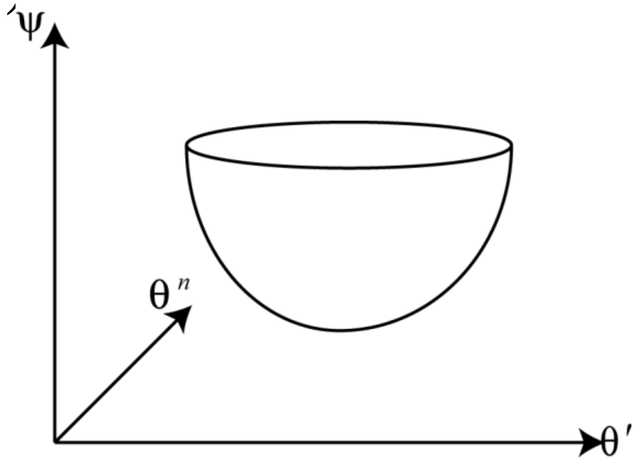
φ : negative entropy

canonical divergence D(P: P')=$\psi(\theta) + \varphi(\eta') - \sum \theta_i \eta_i'$

Manifold with Convex Function

S : **coordinates** $\boldsymbol{\theta} = \left(\theta^1, \theta^2, \cdots, \theta^n\right)$

$\psi(\boldsymbol{\theta})$: **convex function**



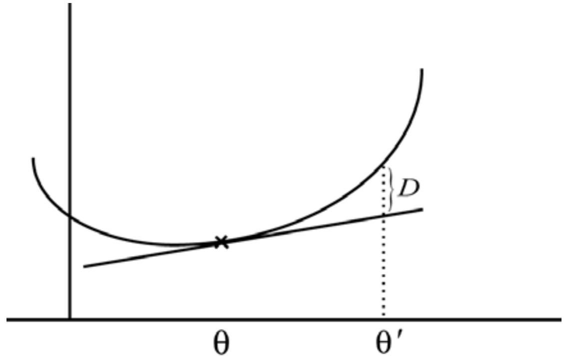
$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\sum\left(\theta^i\right)^2$$

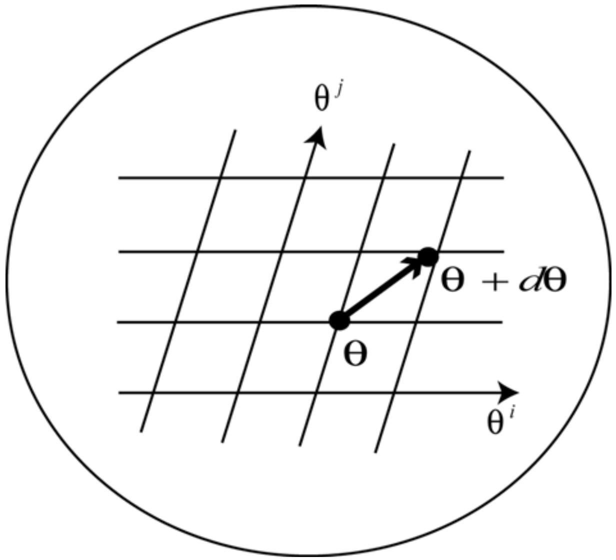
negative entropy $\varphi(p) = \int p(x) \log p(x)\, dx$

Riemannian metric and flatness (affine structure) $\{S,\ \psi(\theta),\ \theta\}$

Bregman divergence

$$D(\theta,\theta') = \psi(\theta') - (\theta' - \theta)\cdot \operatorname{grad}\ \psi(\theta)$$





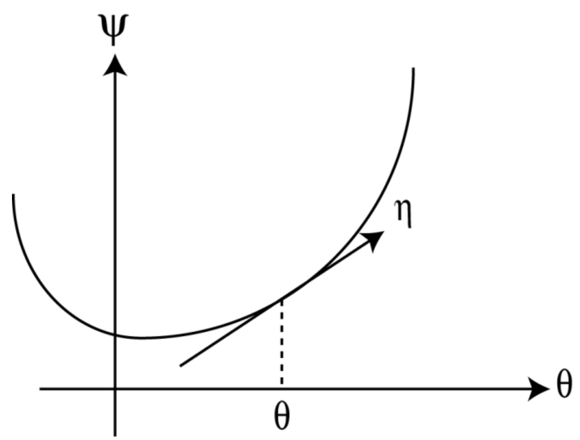
$$D(\theta,\theta+d\theta) = \frac{1}{2}\sum g_{ij}(\theta)\, d\theta^i d\theta^j$$

$$g_{ij} = \partial_i \partial_j \psi(\theta), \quad \partial_i = \frac{\partial}{\partial \theta^i}$$

Flatness (affine) θ : geodesic (not Levi-Civita)

Legendre Transformation

$$\eta_i = \partial_i \psi(\boldsymbol{\theta})$$



$\boldsymbol{\theta} \leftrightarrow \boldsymbol{\eta}$ one-to-one

$$\varphi(\boldsymbol{\eta}) + \psi(\boldsymbol{\theta}) - \theta_i \eta^i = 0$$

$$\varphi(\eta) = \max_\theta \{\theta^i \eta_i - \psi(\theta)\}$$

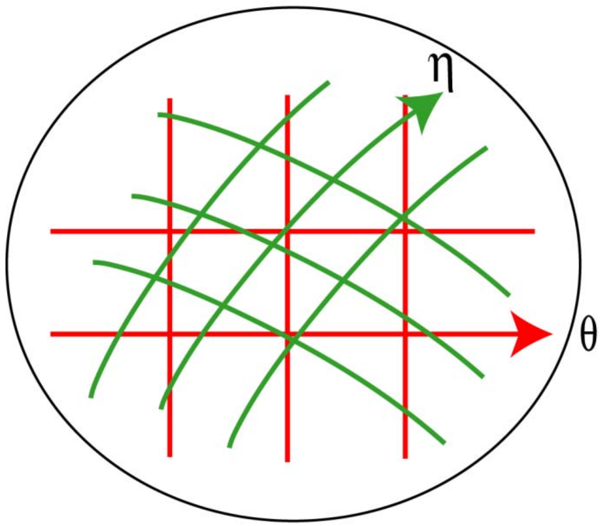
$$\theta^i = \partial^i \varphi(\boldsymbol{\eta}), \quad \partial^i = \frac{\partial}{\partial \eta_i}$$

$$D(\boldsymbol{\theta}, \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}') - \boldsymbol{\theta} \cdot \boldsymbol{\eta}'$$

Two affine coordinate systems (θ, η)

θ : geodesic (e-geodesic)

η : dual geodesic (m-geodesic)



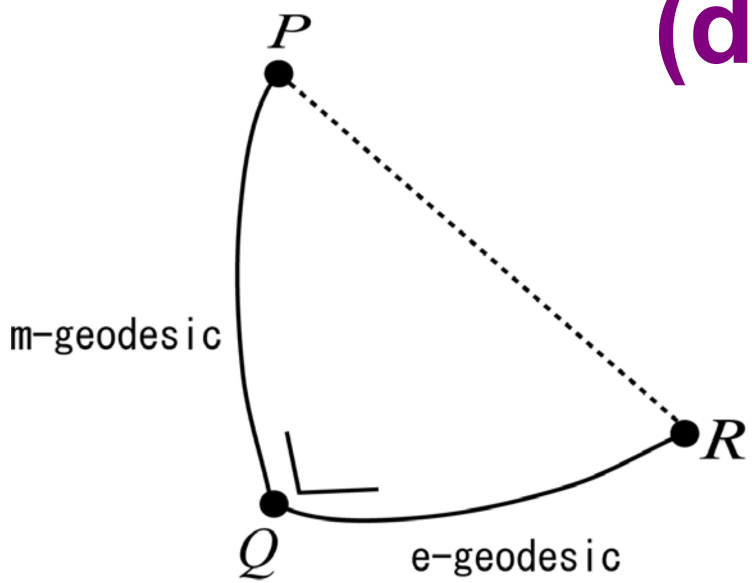
"dually orthogonal"

$$\left\langle \partial_i, \partial^j \right\rangle = \delta_i^j$$

$$\partial_i = \frac{\partial}{\partial \theta^i}, \quad \partial^i = \frac{\partial}{\partial \eta_i}$$

$$X < Y, Z > = < \nabla_X Y, Z > + < Y, \nabla^*{}_X Z >$$

Pythagorean Theorem
(dually flat manifold)

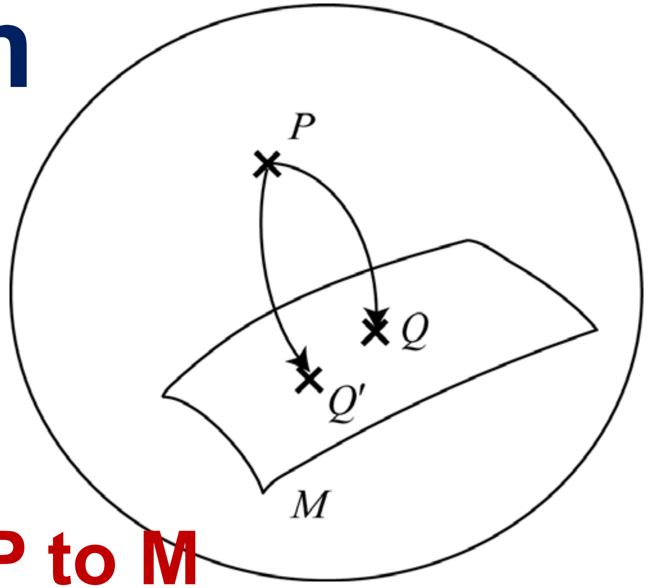


$$D[P:Q]+D[Q:R]=D[P:R]$$

Euclidean space: self-dual $\theta = \eta$

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

Projection Theorem



$$\min_{Q \in M} D[P:Q]$$

Q = m-geodesic projection of P to M

$$\min_{Q \in M} D[Q:P]$$

Q' = e-geodesic projection of P to M

Two Types of Divergence

Invariant divergence (Chentsov, Csiszar)

f-divergence: Fisher-α structure

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

Flat divergence (Bregman) – convex function

KL-divergence belongs to both classes: flat and invariant

divergence

$S = \{\mathbf{p}\}$ **: space of probability distributions**

invariance

dually flat space

invariant divergence

F-divergence
Fisher inf metric
Alpha connection

Flat divergence

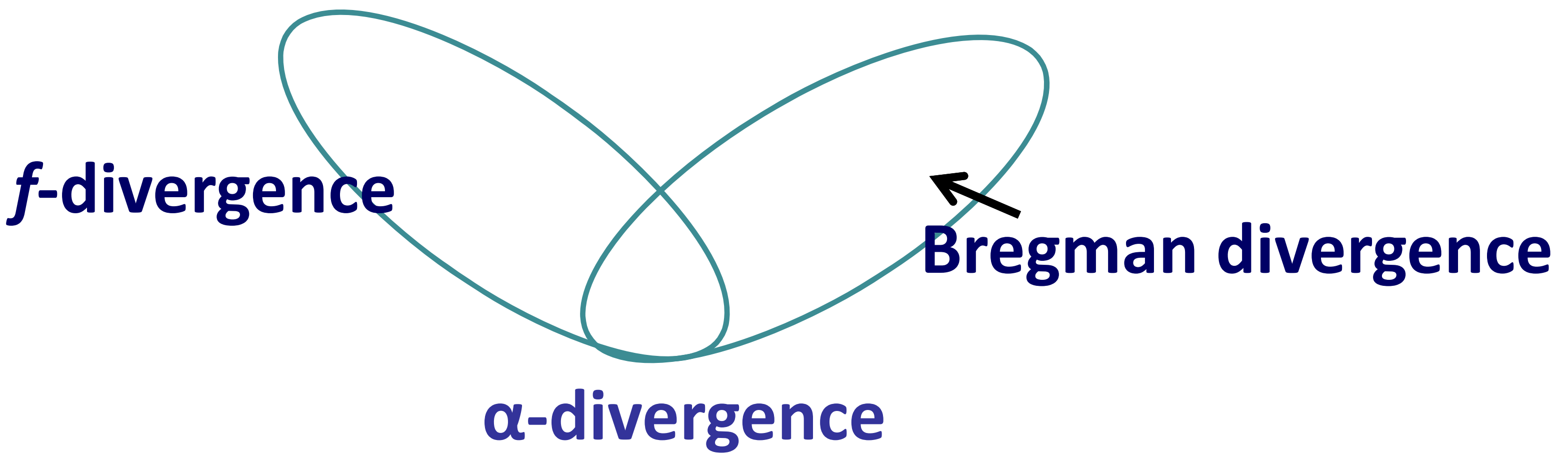
convex functions
Bregman

KL-divergence

$$\mathbf{D}[\mathbf{p} : \mathbf{q}] = \int \mathbf{p}(\mathbf{x}) \log\{\frac{\mathbf{p}(\mathbf{x})}{\mathbf{q}(\mathbf{x})}\}\mathbf{dx}$$

Space of positive measures : vectors, matrices, arrays

$$\tilde{S} = \{\tilde{\boldsymbol{p}}\}, \qquad \tilde{p}_i > 0 : (\sum \tilde{p}_i = 1 \ nn \ \text{holds})$$



Applications of Information Geometry

Statistical Inference
Machine Learning and AI
Computer Vision
Convex Programming
Signal Processing (ICA; Sparse)
Information Theory, Systems Theory
Quantum Information Geometry

Applications to Statistics

curved exponential family:

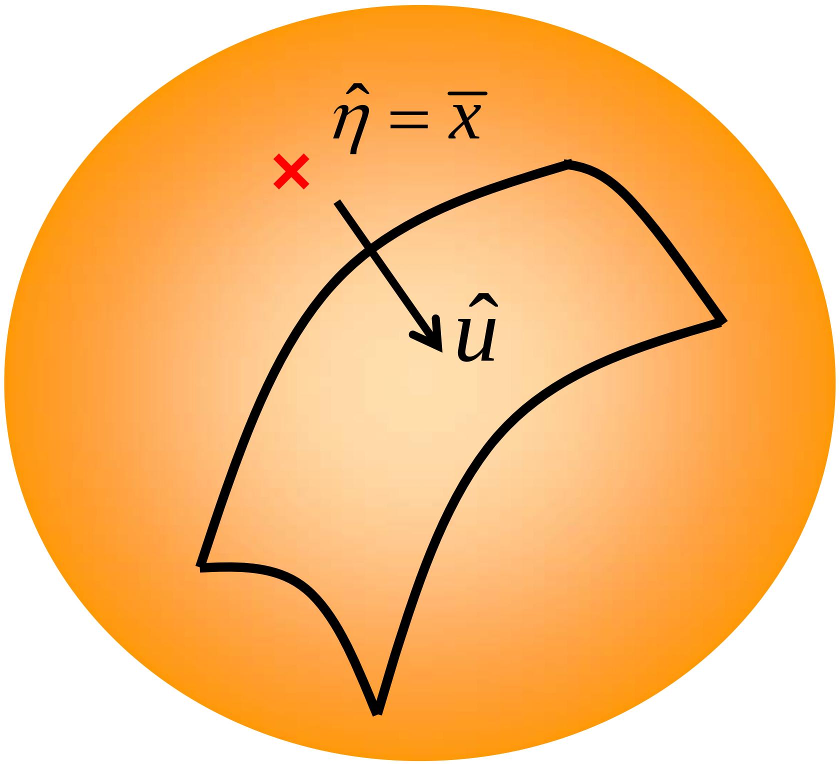
$$p(x,\theta) = \exp\{\theta \cdot x - \psi(\theta)\}$$

$$p(x,u) \; \square \; x_{1,} x_2, \ldots x_n$$

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

$\hat{u}(x_1, \ldots, x_n)$: estimator

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$



x : discrete X = {0, 1, …, n}

$$S_n = \{p(x) \mid x \in X\}: \quad \textbf{exponential family}$$

$$p(x) = \sum_{i=0}^{n} p_i \delta_i(x) = \exp[\sum_{i=1}^{n} \theta^i x_i - \psi(\theta)]$$

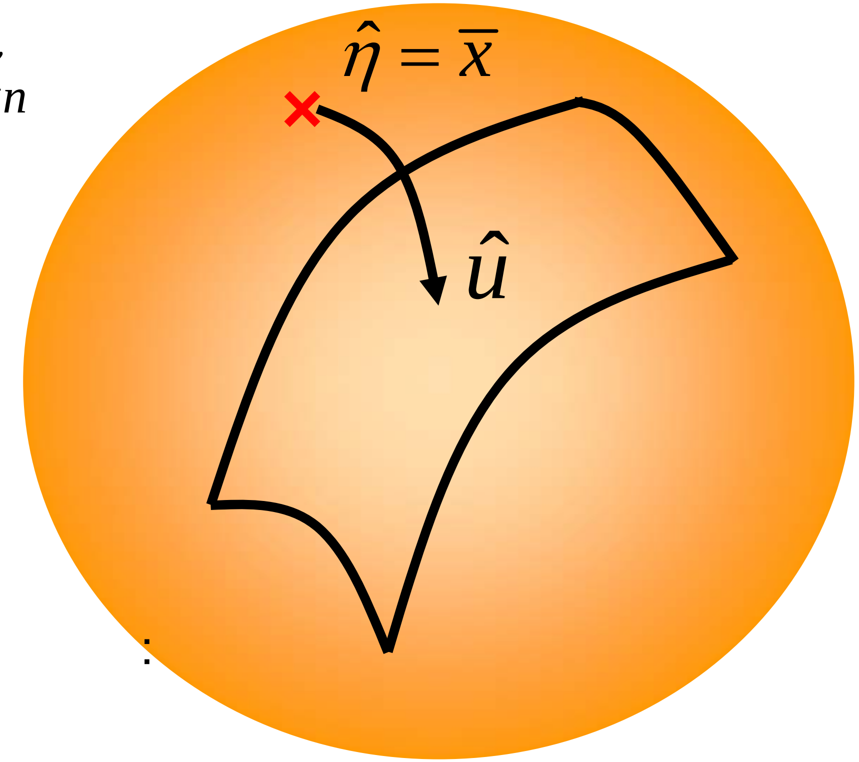
$$\theta^i = \log p_i - \log p_0; \quad x_i = \delta_i(x); \quad \psi(\theta) = -\log p_0$$

statistical model : $p(x, u)$

High-Order Asymptotics

$$p\left(x,\theta(\mathrm{u})\right) \qquad : x_1,\cdots,x_n$$

$$\hat{\mathrm{u}}=\mathrm{u}\left(x_1,\cdots,x_n\right)$$



$$e=E\left[\left(\hat{u}-u\right)\left(\hat{u}-u\right)^T\right]$$

$$e=\frac{1}{n}G_1+\frac{1}{n^2}G_2$$

$G_1 \geq G^{-1}$:**Cramér-Rao: linear theory**

$G_2 = H_M^{(e)^2} + H_A^{(m)^2} + \Gamma^{(m)^2}$ **quadratic approximation**

Information Geometry of Belief Propagation

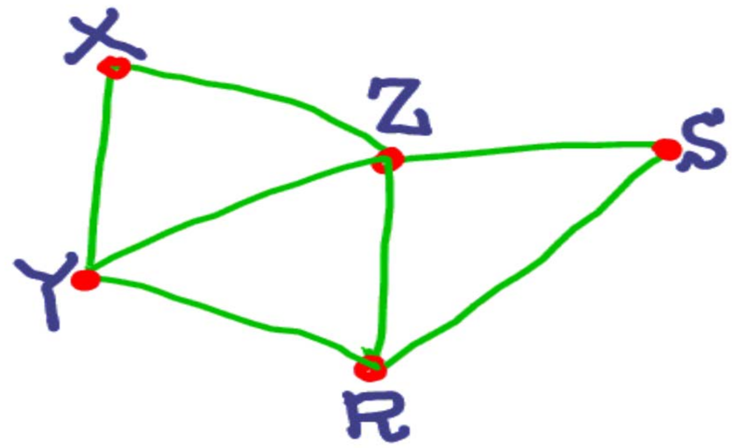
- Shun-ichi Amari (RIKEN BSI)
- Shiro Ikeda (Inst. Statist. Math.)
- Toshiyuki Tanaka (Kyoto U.)

Stochastic Reasoning

$$p(x, y, z, r, s)$$

$$p(x, y, z \mid r, s)$$

$$x, y, z, \ldots = 1, -1$$



Stochastic Reasoning

$q(x_1, x_2, x_3, \ldots | \text{observation})$

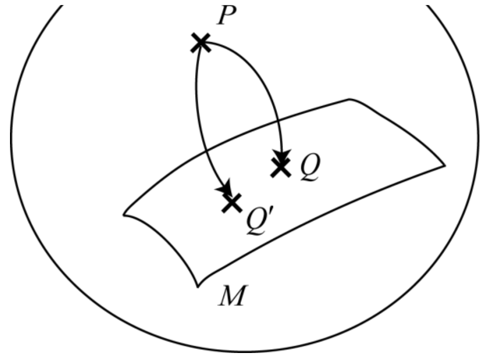
$\mathbf{X} = (x_1\ x_2\ x_3\ \ldots..)$ $\quad x = 1, -1$

$\mathbf{X} = \operatorname{argmax} q\boldsymbol{(}x_1, x_2, x_3, \ldots..\boldsymbol{)}$ $\quad$ maximum likelihood

$X_i = \operatorname{sgn} E[x_i]$ $\quad$ least bit error rate estimator

Mean Value

Marginalization: projection to independent distributions



$$\Pi_0 q(\mathbf{x}) = q_1(x_1)q_2(x_2)\ldots q_n(x_n) = q_0(\mathbf{x})$$

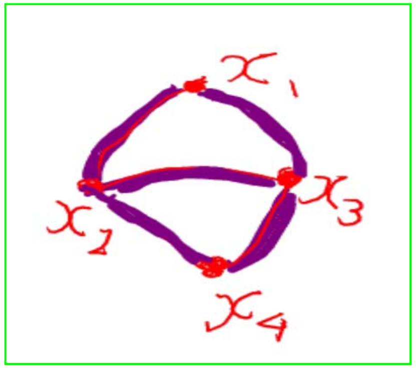
$$q_i(x_i) = \int q(x_{1,}\ldots, x_n)\, dx_1 ..d\breve{x}_i ..dx_n$$

$$\eta = \mathbf{E}_q[\mathbf{x}] = \mathbf{E}_{q_0}[\mathbf{x}]$$

$$q(\boldsymbol{x}) = \exp\left\{ \sum k_i \cdot x_i + \sum_{r=1}^{L} c_r(\boldsymbol{x}) - \psi_q \right\}$$

$$c_r(\boldsymbol{x}) = c_r x_{i_1} \cdots x_{i_s}, \qquad r = (i_1 \cdots i_s)$$

$$x_i = \{1, -1\} \qquad r = (i_1, i_2)$$



Boltzmann machine, spin glass, neural networks
Turbo Codes, LDPC Codes

Computationally Difficult

$$q(\boldsymbol{x}) \rightarrow \eta = E[\boldsymbol{x}]$$

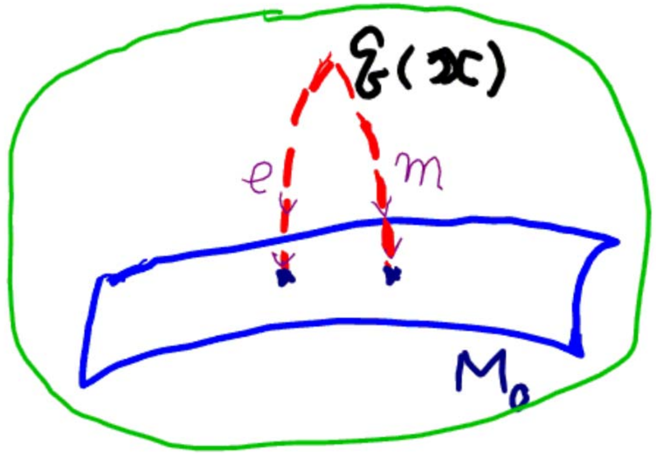
$$q(\boldsymbol{x}) = \exp\left\{\sum c_r(\boldsymbol{x}) - \psi_q\right\}$$

- **mean-field approximation**
- **belief propagation**
- **tree propagation, CCCP (convex-concave)**

Information Geometry of Mean Field Approximation

- m-projection
- e-projection

D[q:p]= $\sum_x q(x) \log \frac{q(x)}{p(x)}$



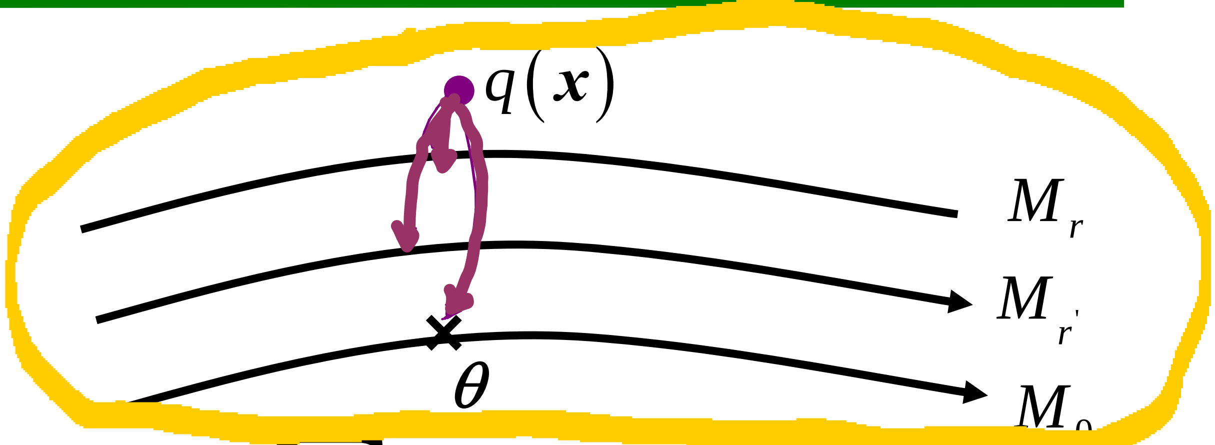
Π_0^mq = argmin D[q:p]

Π_0^eq = argmin D[p:q]

$p(x) \in M_0$

$$M_0 = \{\Pi_i p_i(x_i)\}$$

Information Geometry



$$q(x) = \exp\{\sum c_r(x) - \phi\}$$

$$M_0 = \{p_0(\boldsymbol{x}, \boldsymbol{\theta})\} = \exp\{\boldsymbol{\theta} \cdot \boldsymbol{x} - \psi_0\}$$

$$M_r = \{p_r(\boldsymbol{x}, \xi_r) = \exp\{c_r(\boldsymbol{x}) + \xi_r \cdot \boldsymbol{x} - \psi_r\}\}$$

$$r = 1, \cdots, L$$

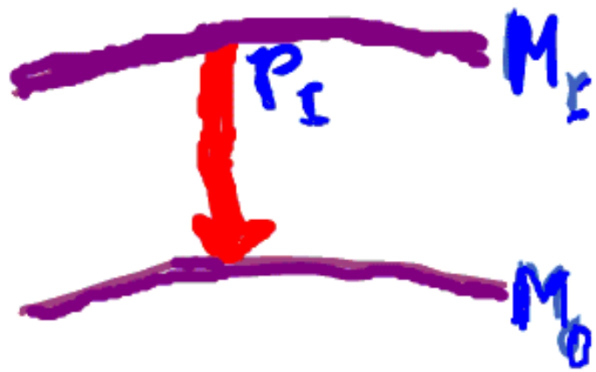
Belief Propagation

$$M_r : p_r\left(\boldsymbol{x},\xi_r\right) = \exp\left\{c_r\left(\boldsymbol{x}\right) + \xi_r \cdot \boldsymbol{x} - \psi_r\right\}$$

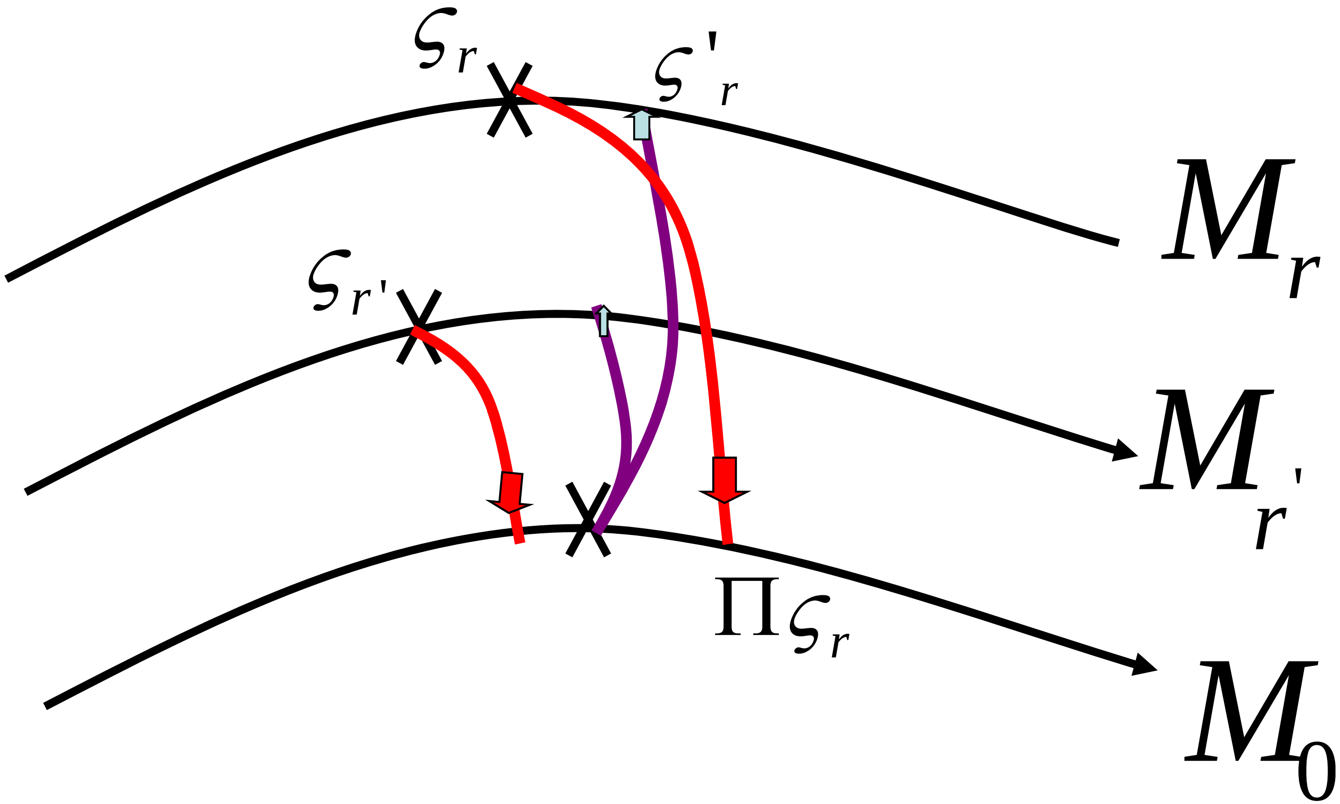
$$\Pi_0 p_r(x,\xi_r{}^t)$$

$$\boldsymbol{\theta}_r{}^{t+1} = \Pi_0 p_r(x,\xi_r{}^t) - \xi_r{}^{\mathrm{t}}$$: belief for $c_r(x)$

$$\boldsymbol{\theta}^{t+1} = \sum \boldsymbol{\theta}_r{}^{t+1}$$



Belief Prop Algorithm

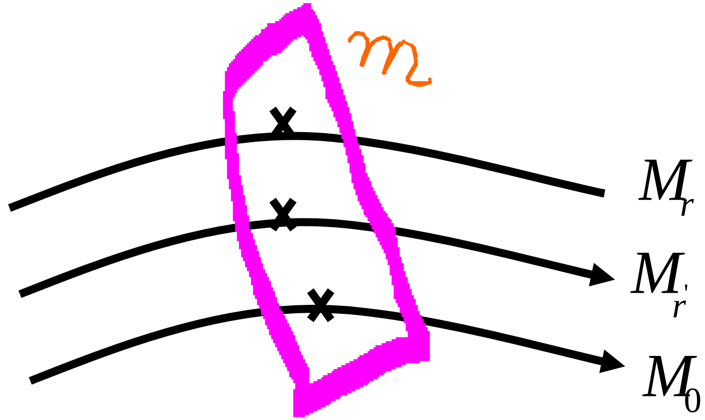


Equilibrium of BP $\left(\boldsymbol{\theta}^*, \boldsymbol{\xi}_r{}^*\right)$

1) *m*-condition

$$\boldsymbol{\theta}^* = \Pi_0 p_r\left(\boldsymbol{x}, \boldsymbol{\xi}^*{}_r\right)$$

m-flat submanifold $M\left(\theta^*\right)$

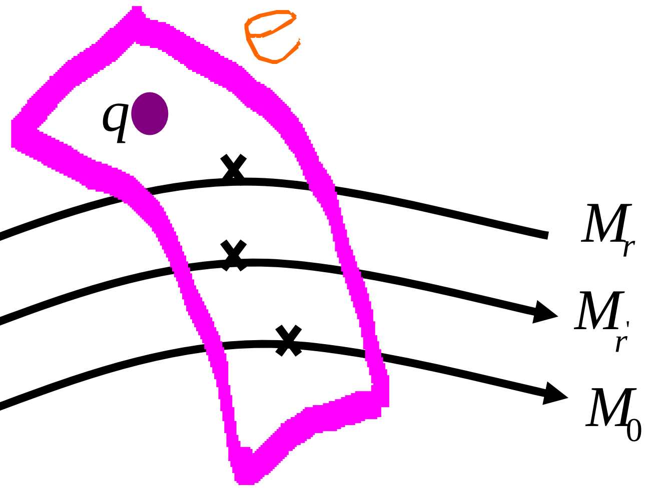


$\xi_1(\theta')$

2) *e*-condition

$$\boldsymbol{\theta}^* = \frac{1}{L-1}\sum_r \boldsymbol{\xi}^*{}_r$$

$q\left(\boldsymbol{x}\right) \in e$-flat submanifold



Free energy:

$$F\left(\theta,\zeta_1,\cdots,\zeta_L\right)=D\left[p_0:q\right]-\sum D\left[p_0:p_r\right]$$

critical point

$$\frac{\partial F}{\partial \boldsymbol{\theta}}=0 \qquad : e\text{-condition}$$

$$\frac{\partial F}{\partial \zeta_r}=0 \qquad : m\text{-condition}$$

not convex

Belief Propagation

e-condition OK

$$(\boldsymbol{\theta}; \boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_L), \qquad \boldsymbol{\theta}' = \frac{1}{L-1} \sum \boldsymbol{\xi}'_r$$

$$(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_L) \rightarrow (\boldsymbol{\xi}'_1, \boldsymbol{\xi}'_2, \ldots, \boldsymbol{\xi}'_L)$$

CCCP m-condition OK

$$\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$$

$$\boldsymbol{\xi}_1(\boldsymbol{\theta}'), \boldsymbol{\xi}_1(\boldsymbol{\theta}'), \ldots, \boldsymbol{\xi}_1(\boldsymbol{\theta}')$$

$$\boldsymbol{\theta}' = \Pi_0 p_r\left(\boldsymbol{x}, \boldsymbol{\xi}'_r\right): \quad \boldsymbol{\xi}' = \Pi_r p_0\left(\boldsymbol{x}, \boldsymbol{\theta}'\right)$$

$$\boldsymbol{\xi}_r^{t+1} = \Pi_r \boldsymbol{\theta}^{t+1} = \Pi_r p_0\left(\boldsymbol{x}, \boldsymbol{\theta}^{t+1}\right)$$

$$\boldsymbol{\theta}^{t+1} = L\boldsymbol{\theta}^t - \sum \xi_r^{t+1}$$

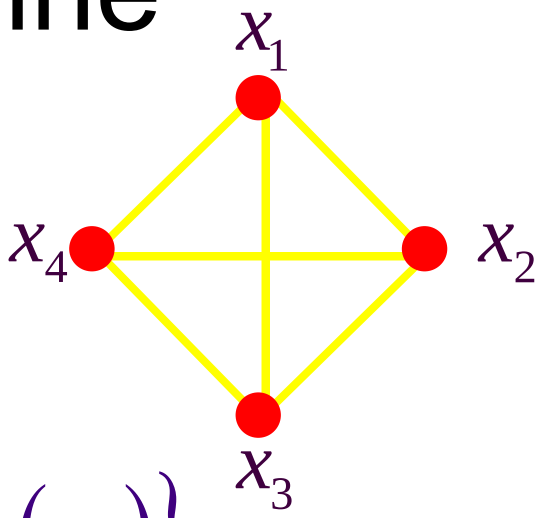
Convex-Concave Computational Procedure (CCCP) Yuille

$$F(\theta) = F_1(\theta) - F_2(\theta)$$

$$\nabla F_1(\theta^{t+1}) = \nabla F_2(\theta^t)$$

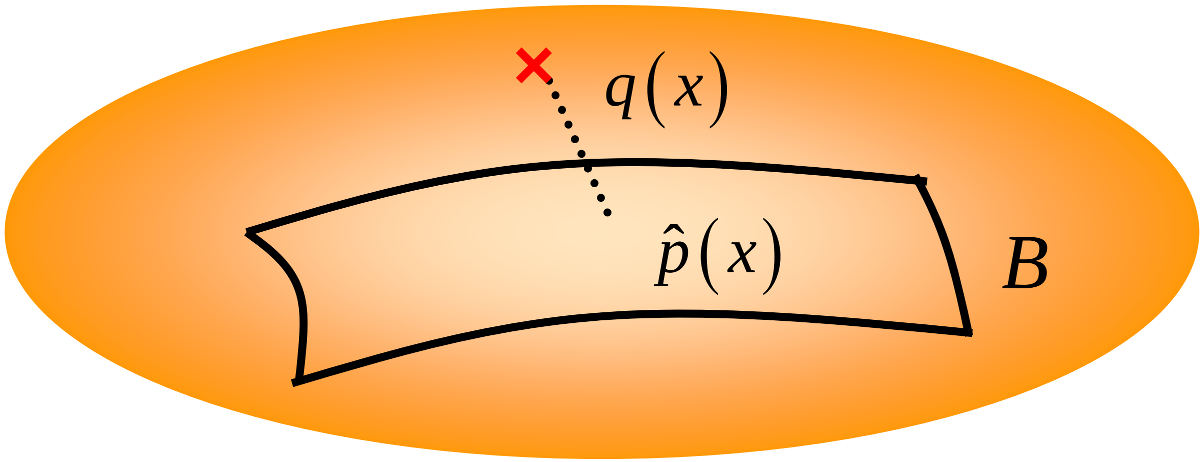
Elimination of double loops

Boltzmann Machine



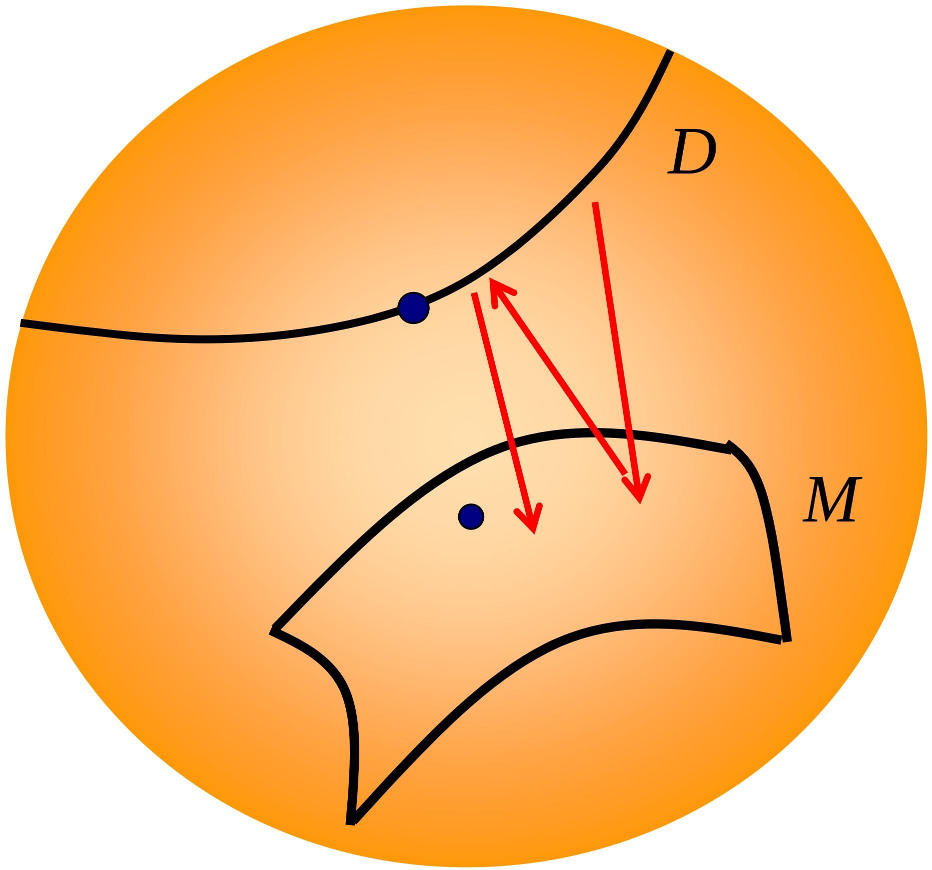
$$p(x_i = 1) = \varphi\left(\sum w_{ij} x_j - h_i\right)$$

$$p(x) = \exp\left\{\sum w_{ij} x_i x_j - \sum h_i x_i - \psi(w)\right\}$$



Boltzmann machine
---hidden units

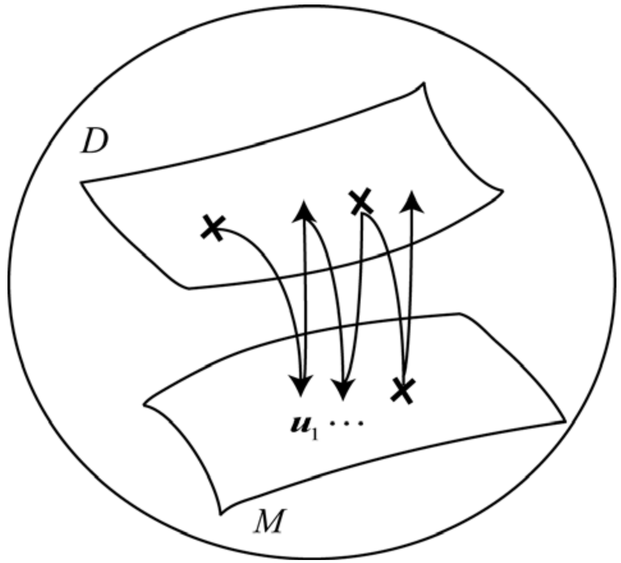
- EM algorithm
- e-projection
- m-projection



EM algorithm

hidden variables

$$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{u})$$



$$D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$$

$$M = \{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{u})\}$$

$$D_M = \{p(\boldsymbol{x}, \boldsymbol{y}) \mid p(\boldsymbol{x}) = p_D(\boldsymbol{x})\}$$

$\min KL[\hat{p}(\boldsymbol{x}, \boldsymbol{y}) : p \in M]$ m-projection to M

$\min KL[p \in D : p(\boldsymbol{x}, \boldsymbol{y}; \hat{\boldsymbol{u}})]$ e-projection to D

SVM : support vector machine

Embedding

$$z_i = \phi_i(x)$$

$$f(x) = \sum w_i \phi_i(x) = \sum \alpha_i y_i K(x_i, x)$$

Kernel

$$K(x, x') = \sum \phi_i(x) \phi_i(x')$$

Conformal change of kernel

$$K(x, x') \longrightarrow \rho(x) \rho(x') K(x, x')$$

$$\rho(x) = \exp\{-\kappa \mid f(x) \mid^2\}$$

Signal Processing

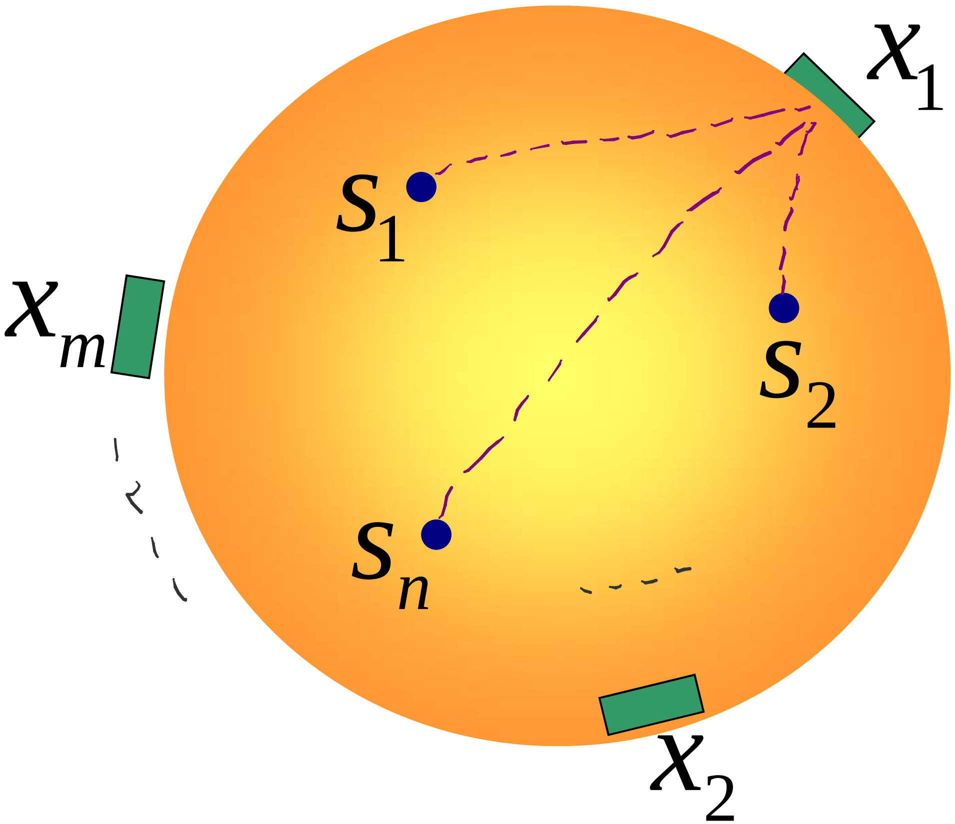
ICA : Independent Component Analysis

$$\boldsymbol{x}_t = A\boldsymbol{s}_t \qquad\qquad \boldsymbol{x}_t \to \boldsymbol{s}_t$$

sparse component analysis

positive matrix factorization

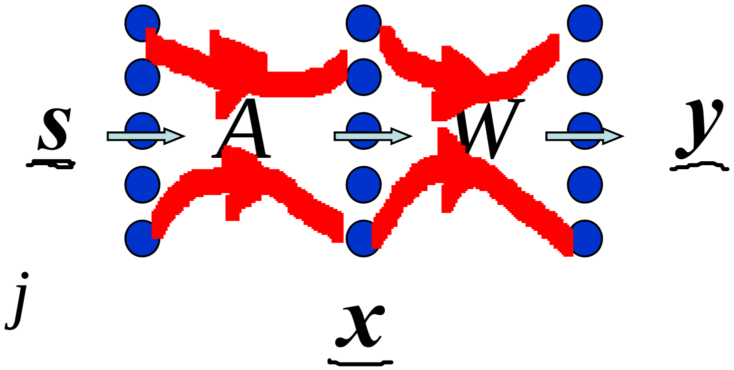
mixture and unmixture of independent signals



$$x_i = \sum_{j=1}^{n} A_{ij} s_j$$

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

Independent Component Analysis



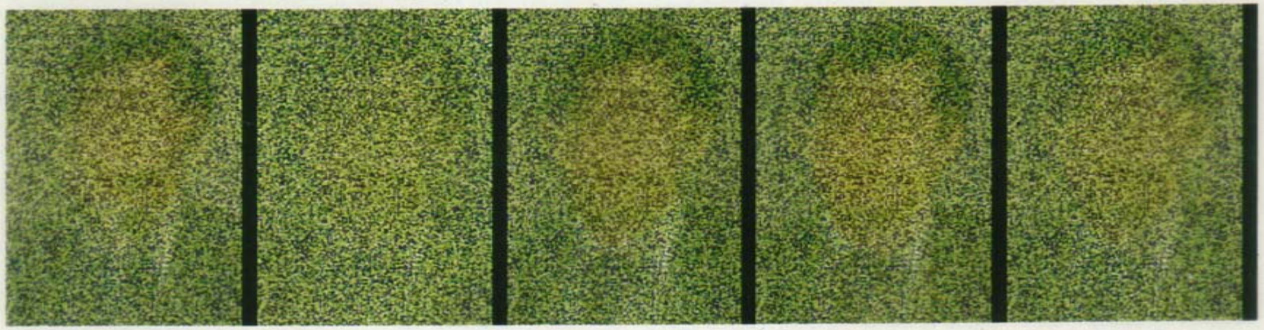
$$\boldsymbol{x} = A\boldsymbol{s} \qquad x_i = \sum A_{ij} s_j$$

$$\boldsymbol{y} = W\boldsymbol{x} \qquad W = A^{-1}$$

observations: x(1), x(2), …, x(t)
recover: s(1), s(2), …, s(t)

Example of color image separation :

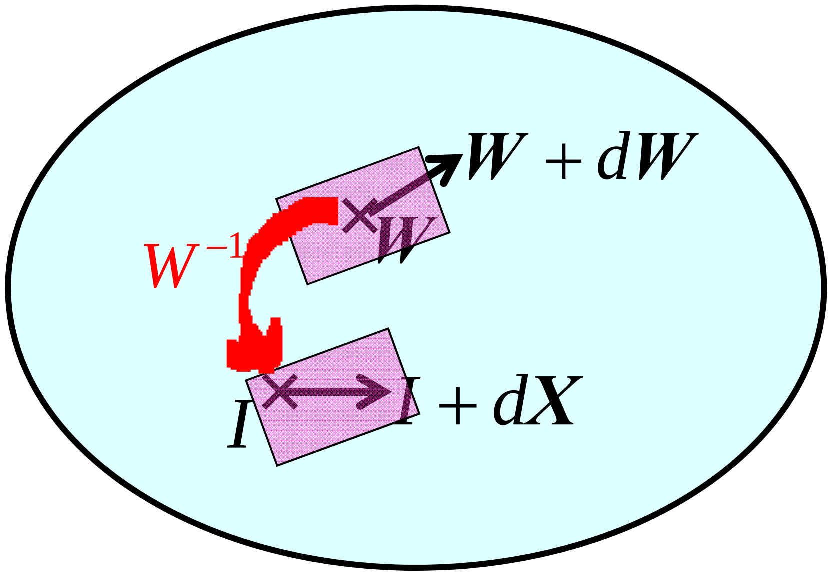
Five original images (but unknown to the neural net)



Five mixed images for separation

Final (stable states) of five separated images

Space of Matrices : Lie group



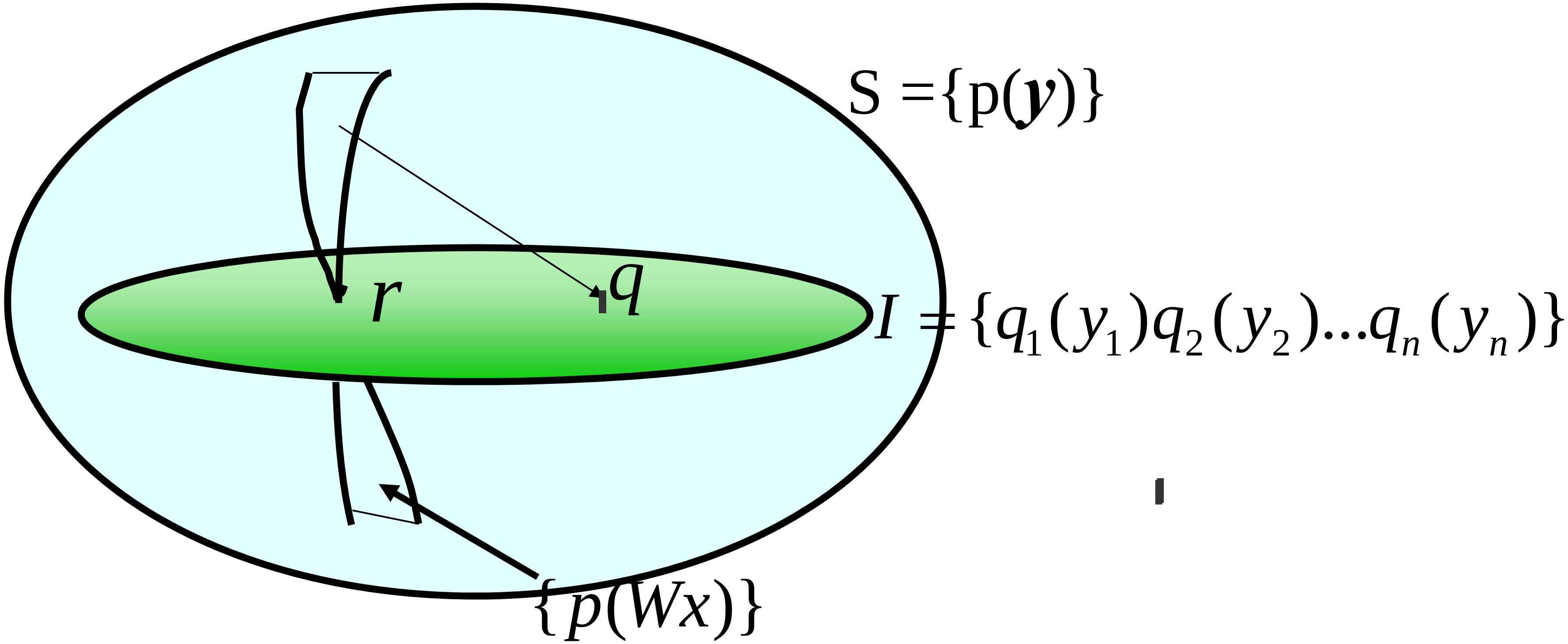
$$dX = dWW^{-1}$$

$$|dW|^2 = \mathrm{tr}\left(dXdX^T\right) = \mathrm{tr}\left(dWW^{-1}W^{-T}dW^T\right)$$

$$\tilde{\nabla} l = \frac{\partial l}{\partial W} W^T W$$

dX : **non-holonomic basis**

Information Geometry of ICA



natural gradient
estimating function
stability, efficiency

$$l(\mathbf{W}) = KL[p(\mathbf{y};\mathbf{W}):q(\mathbf{y})]$$

$$r(\mathbf{y})$$

Semiparametric Statistical Model

$$p(\mathbf{x};\mathbf{W},r)=|\mathbf{W}|\,r(\mathbf{W}\mathbf{x})$$

$\mathbf{W}=\mathbf{A}^{-1},\ \ r(s):$ unknown $\qquad r=\prod r_i$

x(1), x(2), …, x(t)

Natural Gradient

$$\Delta \boldsymbol{W} = -\eta \frac{\partial l\left(\boldsymbol{y}, \boldsymbol{W}\right)}{\partial \boldsymbol{W}} \boldsymbol{W}^T \boldsymbol{W}$$

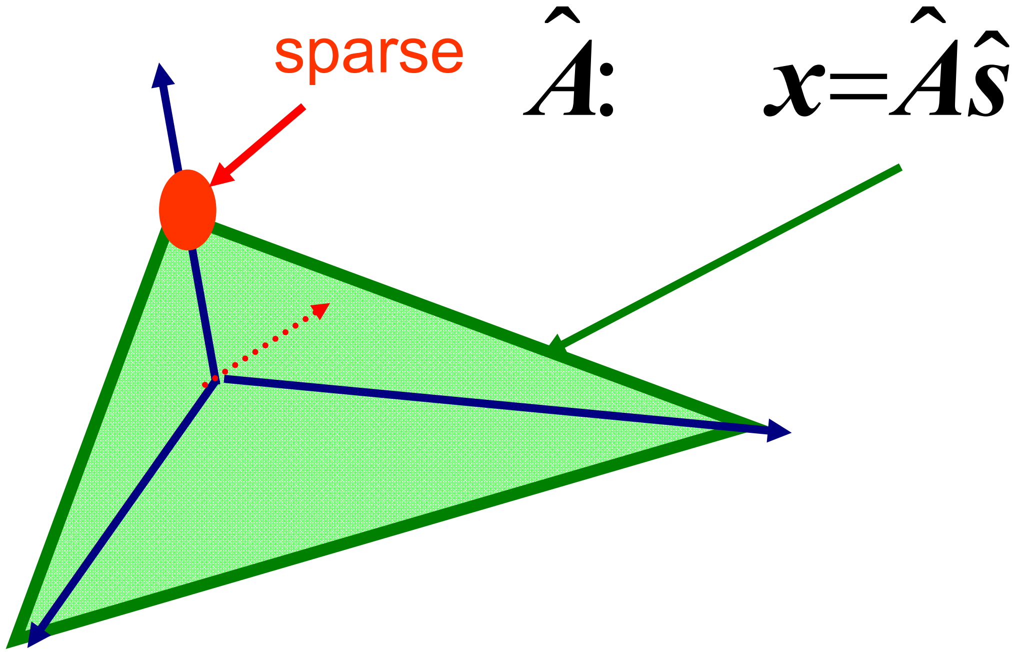
Basis Given: overcomplete case
Sparse Solution

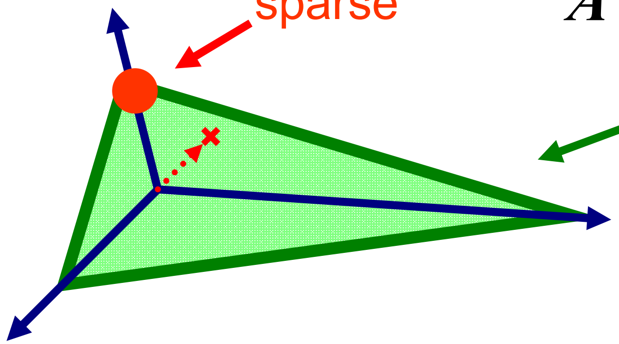
$$\boldsymbol{x} = A\boldsymbol{s} = \sum s_i \boldsymbol{a}_i$$

many solutions

many $s_i \to 0$

$$\boldsymbol{x}_t = \hat{A}\boldsymbol{s}_t$$





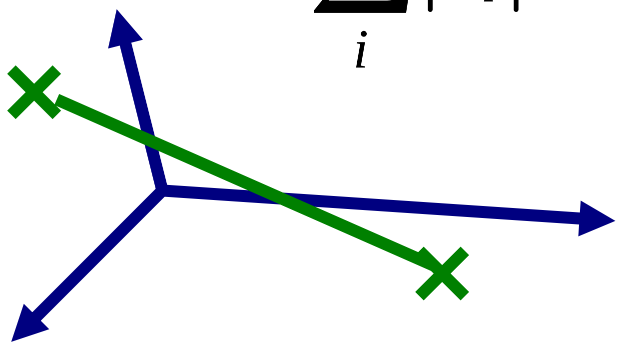
$\hat{\boldsymbol{A}}: \quad x = \hat{\boldsymbol{A}}\hat{\boldsymbol{s}}$

generalized inverse

L_2-norm: $\min \Sigma \left| \hat{\boldsymbol{s}}_i^2 \right|$

sparse solution

L_1-norm: $\min \sum_i \left| \hat{\boldsymbol{s}}_i \right|$



(a) Three binary edge images (reverse images are used in the experiment)

(b) Two edge image mixtures

(c) Reconstructed binary edge images (after reversion)

Fig. 5: Example of edge image image reconstruction: (a) the three binary edge images (reverse image copies are supplied for processing) , (b) their two mixtures, (c) the three extracted edge images (after reversion).

Overcomplete Basis and Sparse Solution

$$\boldsymbol{x} = \sum s_i \boldsymbol{a}_i = A\boldsymbol{s}$$

$$\min |\boldsymbol{s}|_1 = \sum |s_i|$$

$$\min |A\boldsymbol{s} - \boldsymbol{x}|_p + \alpha |\boldsymbol{s}|_{p'}$$

non-linear denoising

Sparse Solution

$$\min \varphi(\boldsymbol{\beta})$$

penalty $F_p(\boldsymbol{\beta}) = \sum |\boldsymbol{\beta}_i|^p$: Bayes prior

$F_0(\boldsymbol{\beta}) = \#1[\beta_i \neq 0]$: **sparsest solution**

$F_1(\boldsymbol{\beta}) = \sum |\beta_i| : L_1$ **solution**

$F_p(\boldsymbol{\beta})$: $0 \le p \le 1$

Sparse solution: overcomplete case

$F_2(\boldsymbol{\beta}) = \sum |\beta_i|^2$: **generalized inverse solution**

Optimization under Spasity Condition

$$\begin{cases} \min \varphi(\boldsymbol{\beta}) & : \text{ convex function} \\ \text{constraint} & F(\boldsymbol{\beta}) \le c \end{cases}$$

typical case:

$$\varphi(\boldsymbol{\beta}) = \frac{1}{2}\left|\boldsymbol{y} - X\boldsymbol{\beta}\right|^2 = \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T G(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$$

$$F(\boldsymbol{\beta}) = \frac{1}{p}\sum\left|\beta_i\right|^p \quad ; \qquad p = 2, \;\; p = 1, \;\; p = 1/2$$

L1-constrained optimization

P_c Problem $\min \varphi(\beta)$ under $F(\beta) \le c$ **LASSO**

solution $\beta^*(c)$: $c = 0 \to \infty$

$\beta_c^* = 0 \to \beta^*$

P_λ Problem $\min \varphi(\beta) + \lambda F(\beta)$ **LARS**

solution $\beta^*(\lambda)$ $\lambda = \infty \to 0$

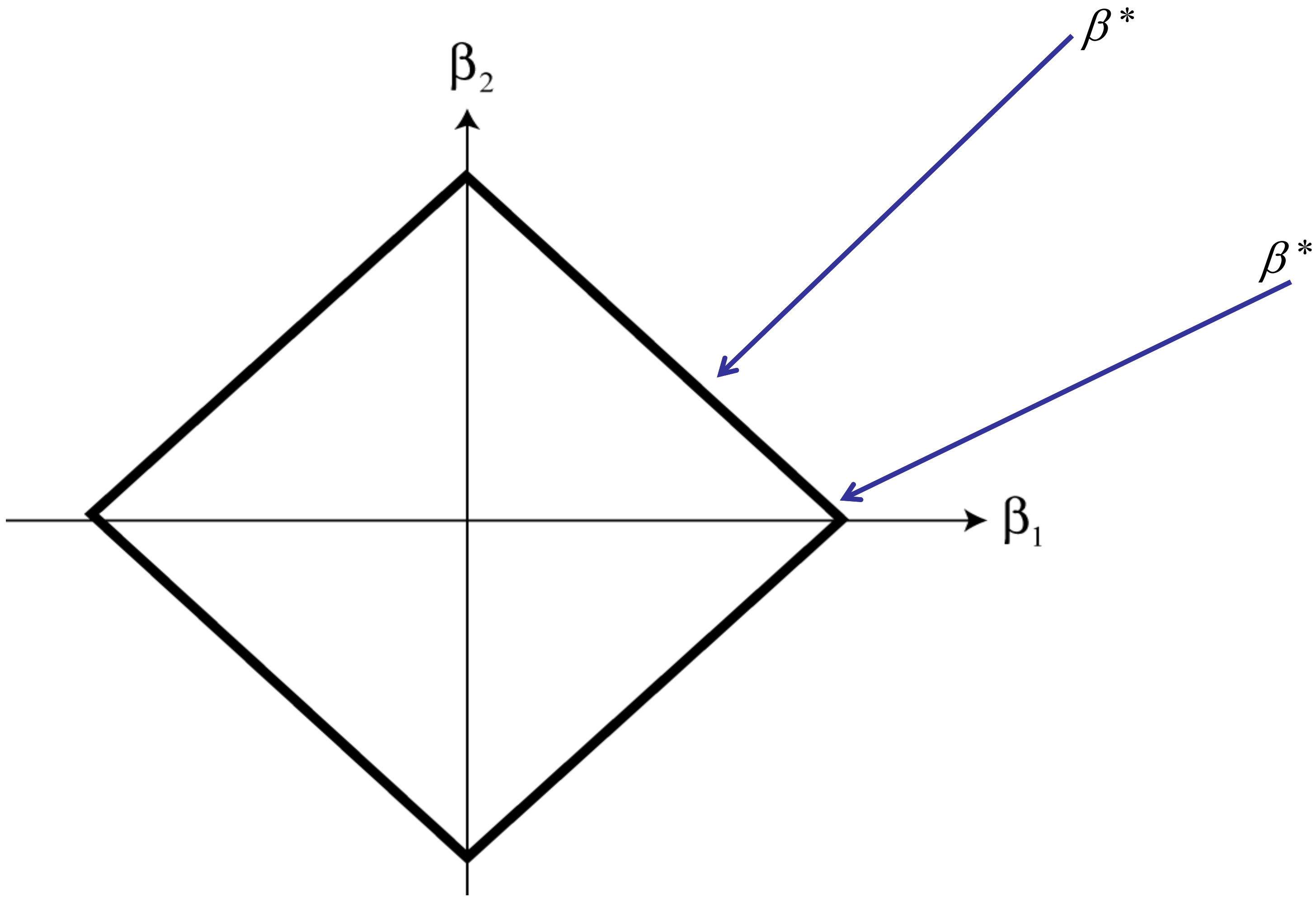
$\beta_\lambda^* = 0 \to \beta^*$

solutions β_c^* and β_λ^* : coincide, $\lambda = \lambda(c)$, $p \ge 1$

$p < 1$: $\lambda = \lambda(c)$ multiple, noncontinuous

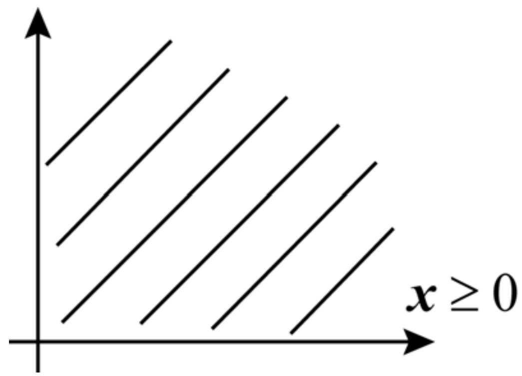
stability different

Projection from β^* to F = c (information geometry)



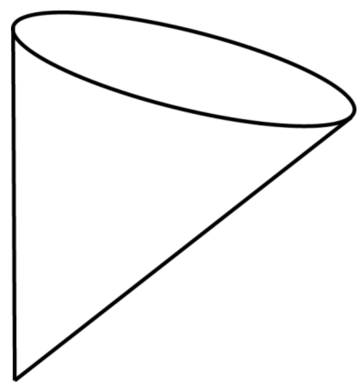
Convex Cone Programming

P : positive semi-definite matrix



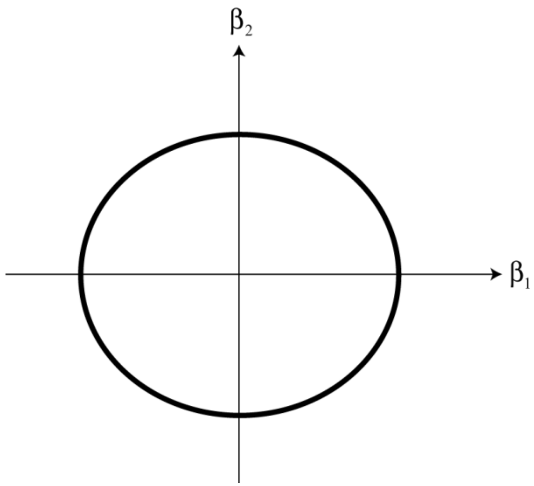
convex potential function

dual geodesic approach

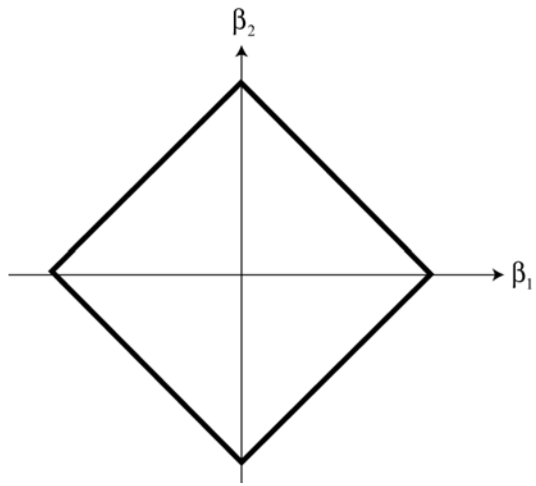


$$A\boldsymbol{x} = \boldsymbol{b}, \qquad \min \boldsymbol{c} \cdot \boldsymbol{x}$$

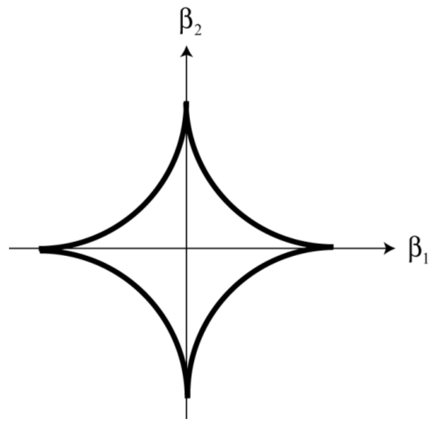
Support vector machine



a) $R_c : n = 2, \ \ p > 1$



b) $R_c : n = 2, \ p = 1$



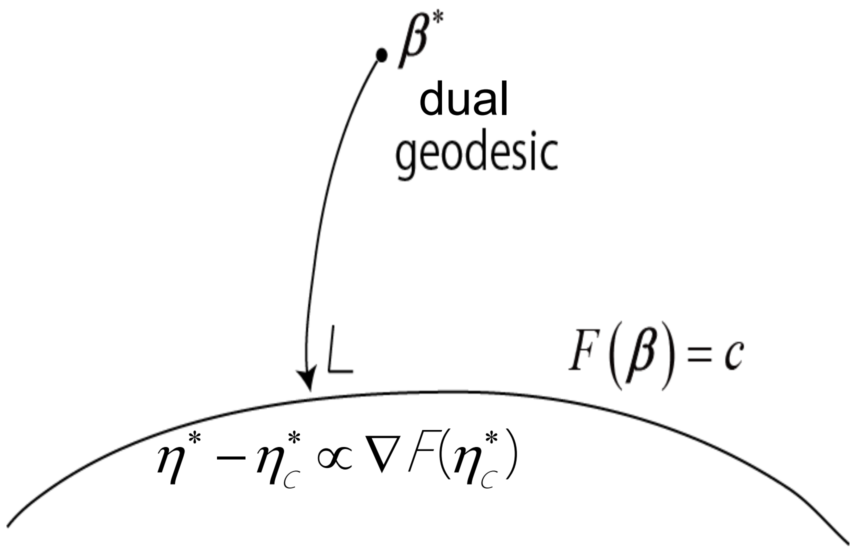
non-convex

c) $R_c : n = 2, \ p < 1$

Fig. 1

orthogonal projection, dual projection

$$\min\ \varphi(\boldsymbol{\beta}) = \mathbf{D}\left[\boldsymbol{\beta} : \boldsymbol{\beta}^{*}\right],\quad \mathbf{F}\left(\boldsymbol{\beta}\right) = \mathbf{c}:$$ **dual geodesic projection**



$$\eta_C^* \propto \nabla F(\eta_C^*)$$

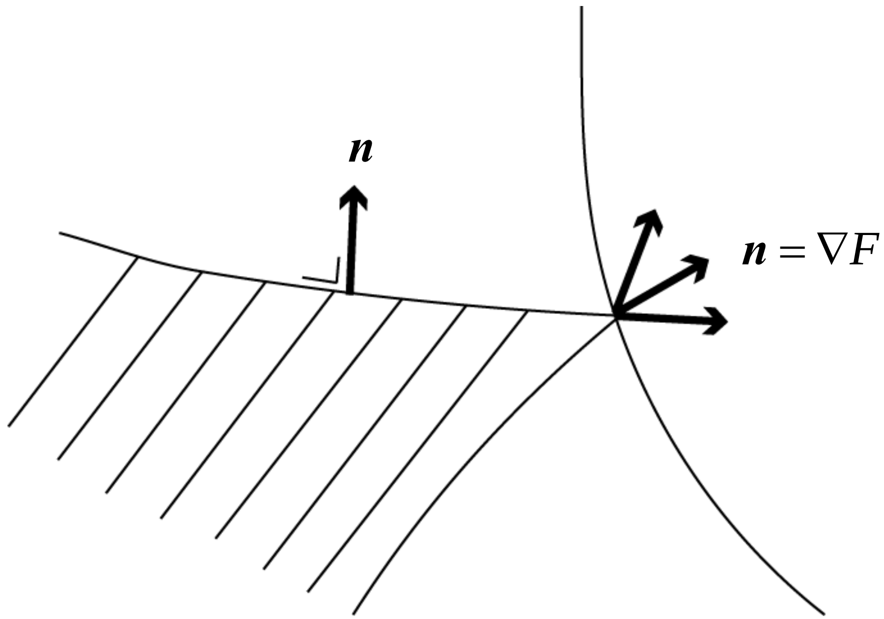


Fig. 5 subgradient

LASSO path and LARS path (stagewise solution)

$$\min \varphi(\beta) \;\; : \; F(\beta) = c$$

$$\min \varphi(\beta) + \lambda F(\beta)$$

$\beta^*(c),\ \beta^*(\lambda)$ **c $\Leftrightarrow$ λ correspondence**

Active set and gradient

$$A(\boldsymbol{\beta}) = \left\{ i \middle| \beta_i \neq 0 \right\}$$

$$\nabla F_p(\boldsymbol{\beta}) = \begin{cases} \operatorname{sgn}(\beta_i)\left|\beta_i\right|^{-(1-p)}, & i \in A \\ (-\infty, \infty), & i \notin A \\ [-1,1] & \end{cases}$$

Solution path

$$\nabla_A \varphi\left(\boldsymbol{\beta}_c^*\right) + \lambda_c \nabla_A F\left(\boldsymbol{\beta}_c^*\right) = 0, \quad \boldsymbol{\beta}_c^*$$

$$\left\{\nabla_A \nabla_A \varphi\left(\boldsymbol{\beta}_c^*\right) + \lambda_c \nabla_A \nabla_A F\left(\boldsymbol{\beta}_c^*\right)\right\} \cdot \dot{\boldsymbol{\beta}}_c = -\dot{\lambda}_c \nabla_A F\left(\boldsymbol{\beta}_c\right)$$

$$\dot{\boldsymbol{\beta}}_c = -\dot{\lambda}_c K^{-1} \nabla_A F\left(\boldsymbol{\beta}_c^*\right) \ ; \ \dot{\boldsymbol{\beta}}_c = \frac{d}{dc} \boldsymbol{\beta}_c$$

$$K = G\left(\boldsymbol{\beta}_c^*\right) + \lambda_c \nabla\nabla F\left(\boldsymbol{\beta}_c^*\right)$$

$$\left(\nabla\nabla F_1 = 0; \quad \nabla F_1 = (\operatorname{sgn} \beta_i) : \quad L_1\right)$$

Solution path in the subspace of the active set

$$\nabla_A \varphi\left(\boldsymbol{\beta}_\lambda^*\right) + \lambda \nabla_A F\left(\boldsymbol{\beta}_\lambda^*\right) = 0 \qquad \nabla_A \text{ : active direction}$$

$$\dot{\boldsymbol{\beta}}_\lambda^* = -K_A^{-1} \nabla_A F\left(\boldsymbol{\beta}_\lambda^*\right)$$

turning point $\mathbf{A} \rightarrow \mathbf{A}'$

Gradient Descent Method

min L(x+a): $g_{ij}a^i a^j = \varepsilon^2$

$$\nabla L = \{\frac{\partial}{\partial x_i} L(x)\}: \quad \text{covariant}$$

$$\tilde{\nabla} L = \{\sum g^{ji} \frac{\partial}{\partial x_i} L(x)\}: \quad \text{contravariant}$$

$$x_{t+1} = x_t - c\tilde{\nabla} L(x_t)$$

Extended LARS (p = 1) and Minkovskian grad

norm $\|\boldsymbol{a}\|_p = \sum |a_i|^p$

$\max \psi(\boldsymbol{\beta} + \varepsilon \boldsymbol{a})$ under $\|\boldsymbol{a}\|_p = 1$

$\psi(\boldsymbol{\beta} + \varepsilon \boldsymbol{a}) - \lambda \|\boldsymbol{a}\|_p$

$p = 1^+$

$$\nabla_1 \psi(\boldsymbol{\beta}) = \mathbf{1}_A \begin{cases} \operatorname{sgn} \eta_i, & |\eta_i| = \max\{|\eta_1|, \cdots, |\eta_N|\} \\ 0, & \text{otherwise} \end{cases}$$

$\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\beta})$

$$i^* = \arg\max |f_i|$$

$$\max |f_i| = |f_{i^*}| = |f_{j^*}|$$

$$\left(\tilde{\nabla} F\right)_i = \begin{cases} 1, & \text{for } i = i^* \text{ and } j^*, \\ 0 & \text{otherwise.} \end{cases}$$

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta \tilde{\nabla} F$$

LARS

$$\tilde{\nabla} F = \nabla f$$

Euclidean case

$$\tilde{\nabla} F = c\left(\operatorname{sgn} f_i\right)\left| f_i \right|^{\frac{1}{p-1}}$$

$$\tilde{\nabla} F = c\left(\operatorname{sgn} f_{i^*}\right)\begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\alpha \to 1$$

L1/2 constraint: non-convex optimization

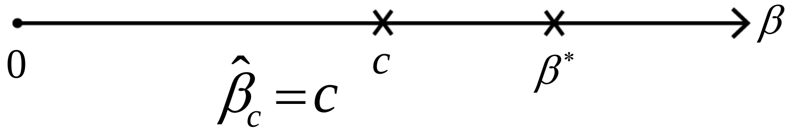
λ-trajectory and δ-trajectory

Ex. 1-dim

$$\varphi(\beta) = \frac{1}{2}(\beta - \beta *)^2$$

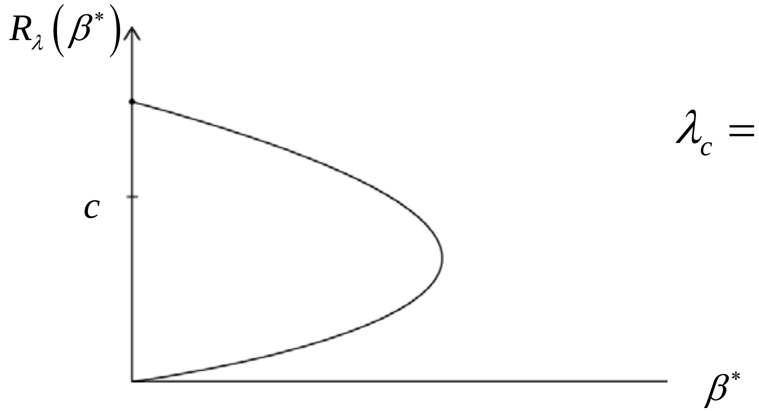
$$f_\lambda(\beta) = \phi + \lambda F = \frac{1}{2}(\beta - 2)^2 + 2\lambda\sqrt{\beta}$$

$$P_c : \min\left(\beta - \beta^*\right)^2, \quad |\beta| \le c$$



$$P_\lambda : \nabla f_\lambda = 0 \qquad \beta - \beta^* + \frac{\lambda}{\sqrt{\beta}} = 0 \qquad \hat{\beta} = R_\lambda\left(\beta^*\right)$$

: Xu Zongben's operator



$$\lambda_c = \sqrt{c}\left(\beta^* - c\right)$$

λ

ICCN-Huangshan（黄山）

Sparse Signal Analysis

Shun-ichi Ammari（甘利俊一）

RIKEN Brain Science Institute

（Collaborator:　Masahiro Yukawa, Niigata University)

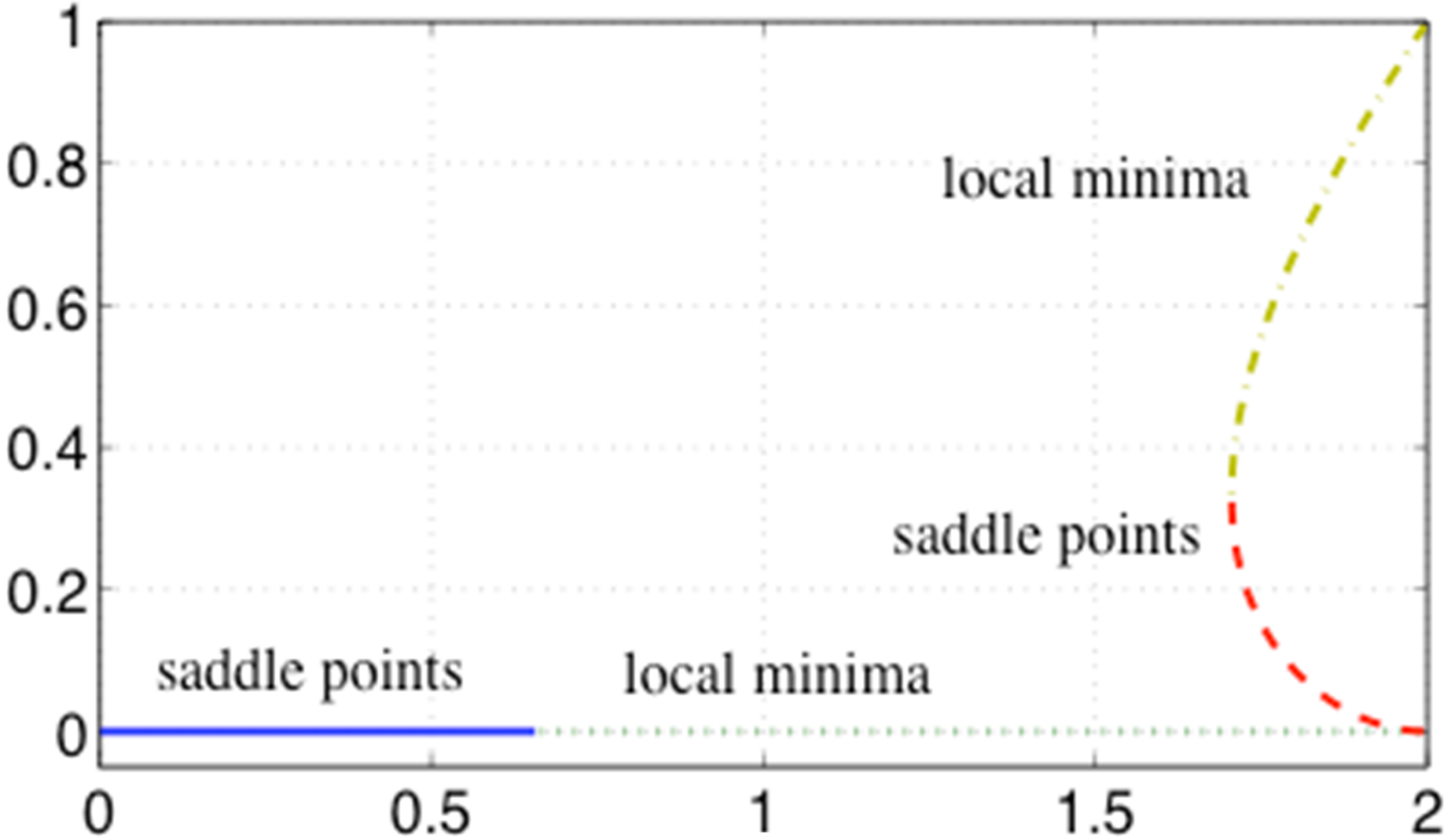
Solution Path : $\lambda \leftrightarrow c$

not continuous, not-monotone

jump

$$\beta_\lambda \Leftrightarrow \beta_c$$

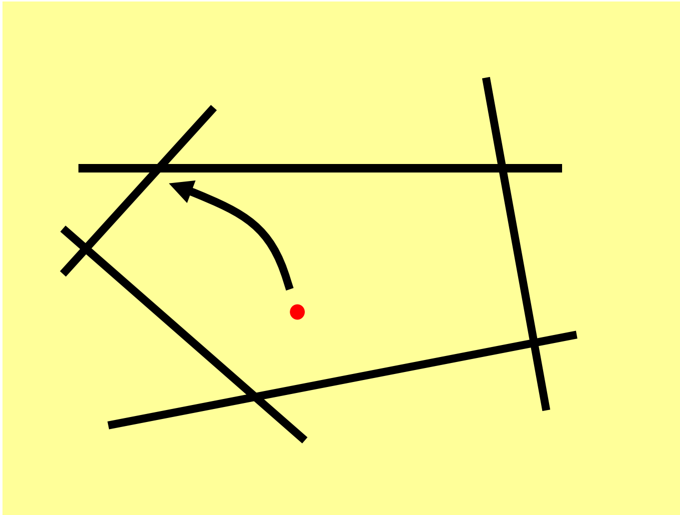
An Example of the greedy path



β_2

β_1

Linear Programming



$$\sum A_{ij} x_j \geq b_i$$

$$\max \sum c_i x_i$$

$$\psi(\boldsymbol{x}) = \sum_i \log\left(\sum A_{ij} x_j - b_i\right)$$

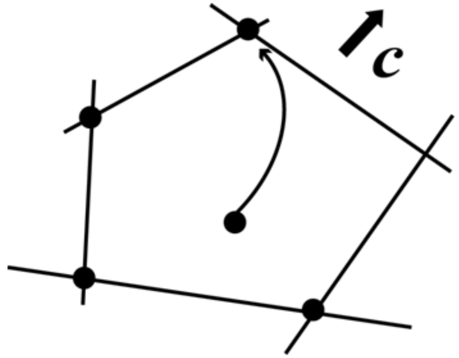
Convex Programming
— Inner Method

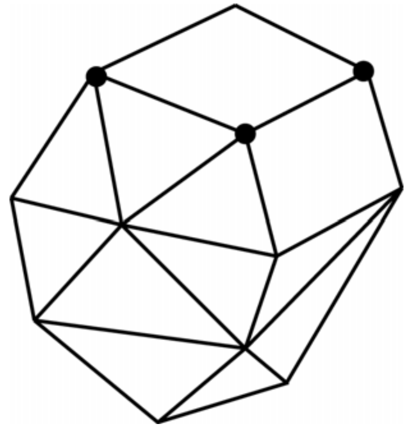
$$LP : A\boldsymbol{x} \geq \boldsymbol{b}, \qquad \boldsymbol{c} \cdot \boldsymbol{x} \geq 0$$

$$\min \qquad \boldsymbol{c} \cdot \boldsymbol{x}$$

$$\psi(\boldsymbol{x}) = \sum \log\left(\sum A_{ij} x_j - b_i\right) + \sum \log x_i$$

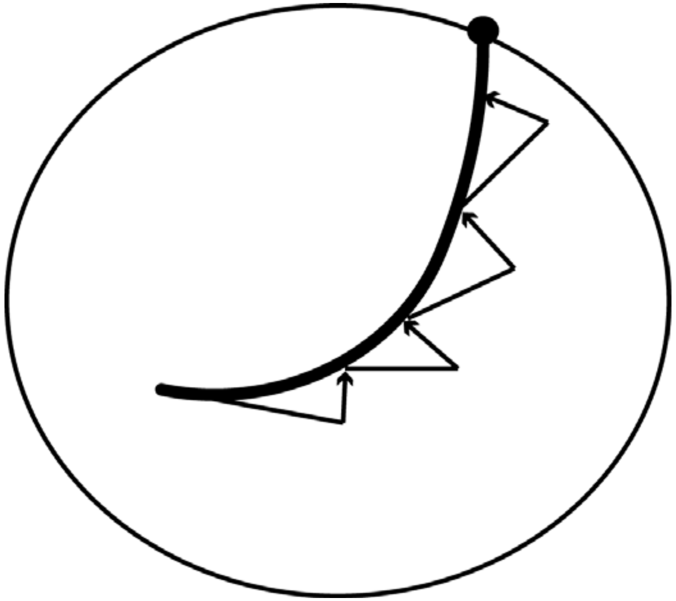
$$\boldsymbol{\eta} = \partial_i \psi(\boldsymbol{x})$$





Simplex method ; inner method

Polynomial-Time Algorithm



curvature : step-size

$$\left|\boldsymbol{H}^{(m)}\right|^2$$

$$\min : t\boldsymbol{c}\cdot\boldsymbol{x}+\psi\left(\boldsymbol{x}\right) \qquad \boldsymbol{x}=\boldsymbol{\delta}\left(t\right) \qquad \nabla^* - \text{geodesic}$$

Neural Networks

Multilayer Perceptron

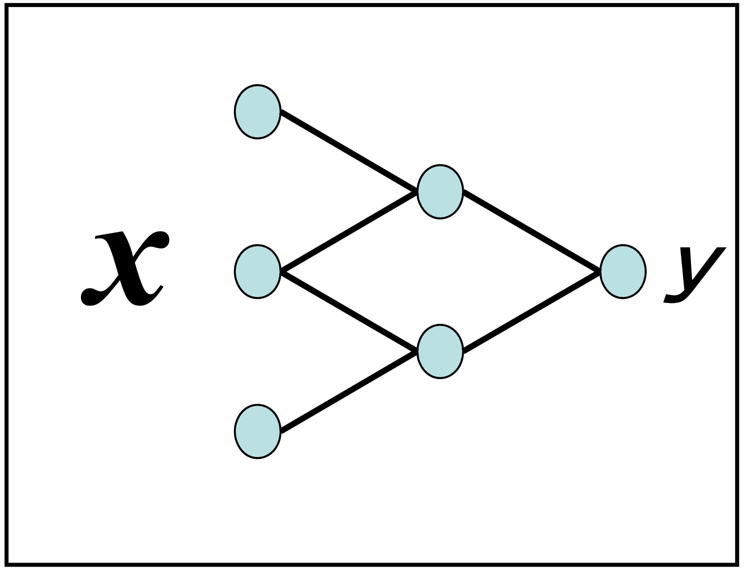
Higher-order correlations

Synchronous firing

Multilayer Perceptrons

$$y = \sum v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right) + n$$

$$x = \left(x_1, x_2, \ldots, x_n\right)$$



$$p\left(y \middle| \boldsymbol{x}; \boldsymbol{\theta}\right) = c \exp\left\{-\frac{1}{2}\left(y - f\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)^2\right\}$$

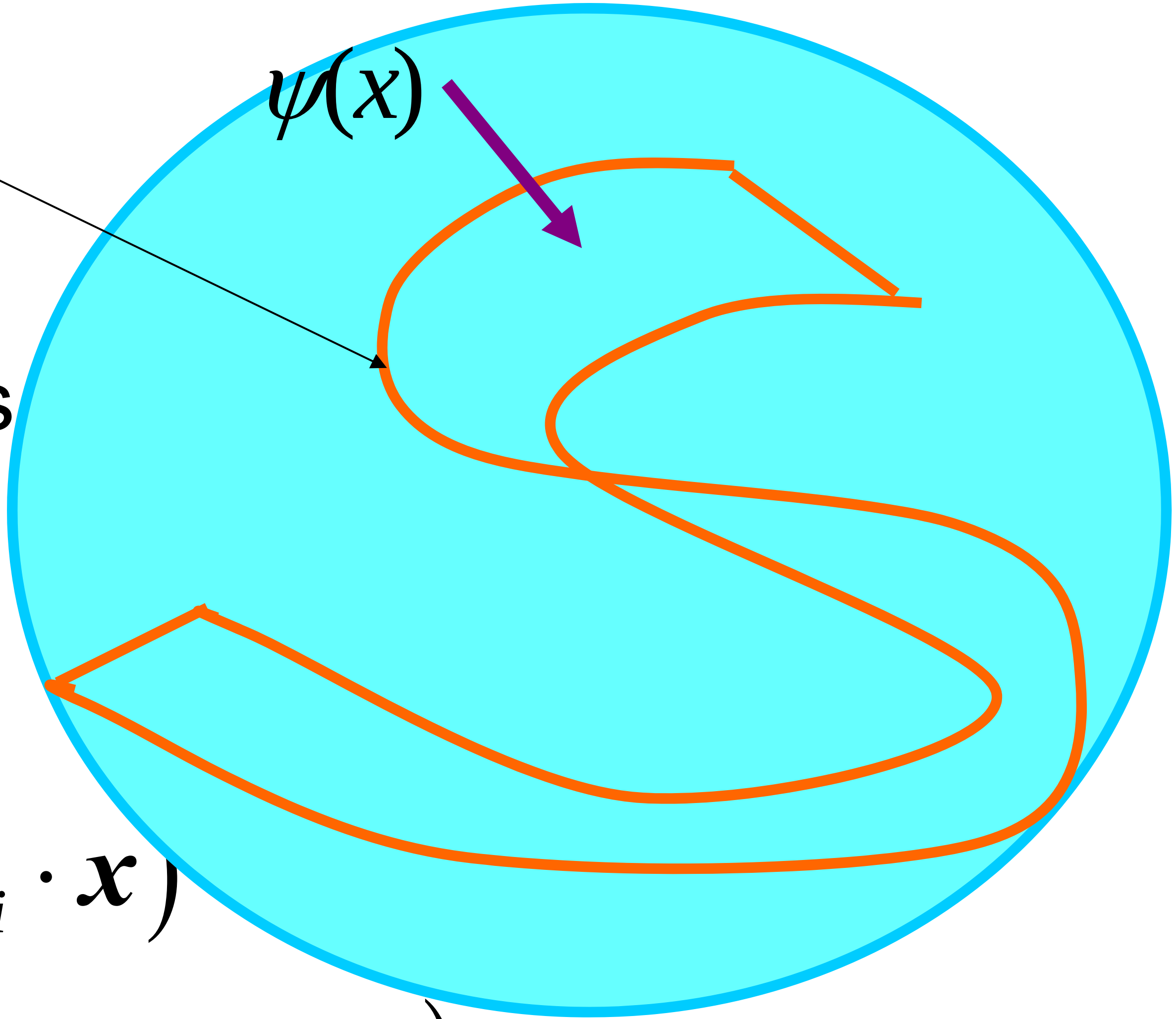
$$f\left(\boldsymbol{x}, \boldsymbol{\theta}\right) = \sum v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right)$$

$$\theta = \left(w_1, \ldots, w_m; v_1, \ldots, v_m\right)$$

Multilayer Perceptron

neuromanifold

space of functions

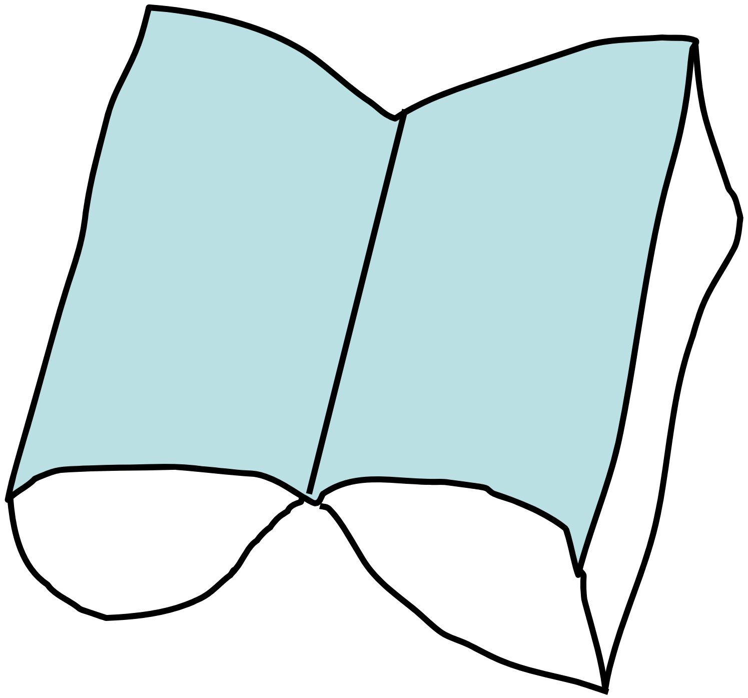


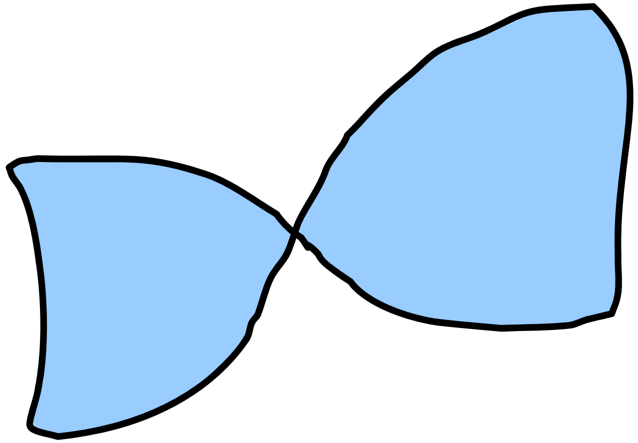
$$y = f(\boldsymbol{x}, \boldsymbol{\theta})$$

$$= \sum v_i \, \varphi \left(\boldsymbol{w}_i \cdot \boldsymbol{x} \right)$$

$$\boldsymbol{\theta} = \left(v_1 \cdots, v_m \; ; \; \boldsymbol{w}_{1,} \cdots, \boldsymbol{w}_m \right)$$

singularities

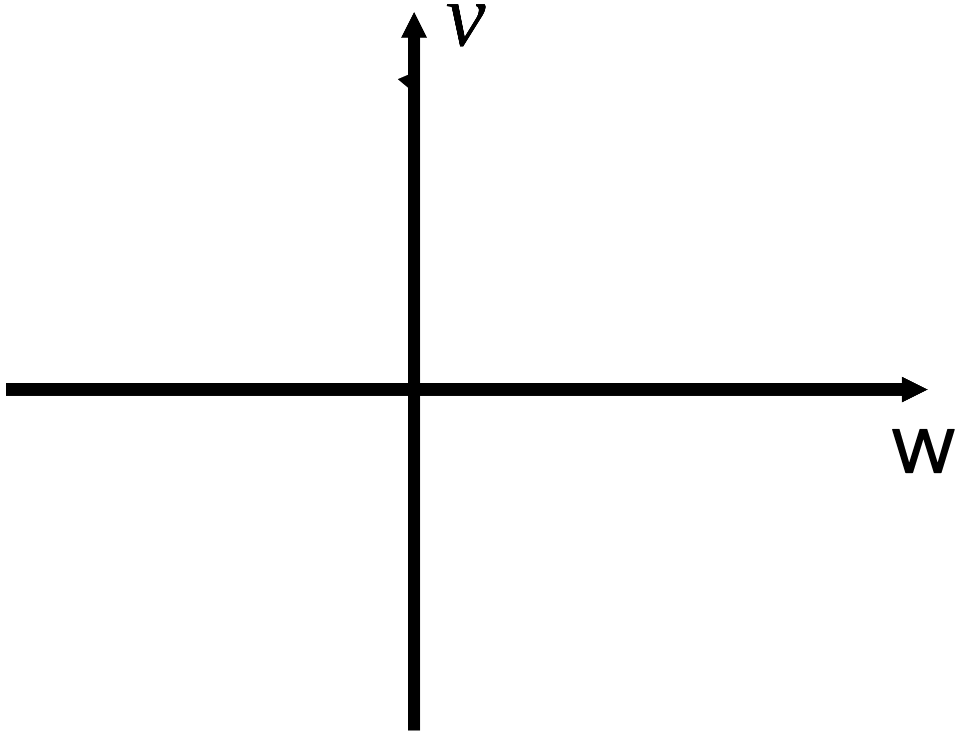


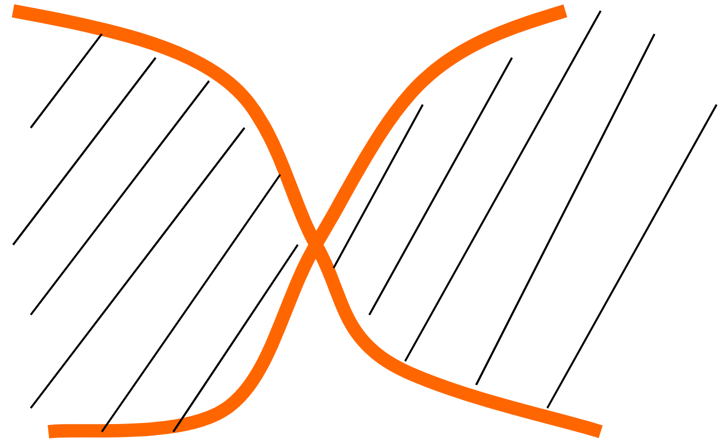


Geometry of singular model

$$y = v\varphi(\boldsymbol{w} \cdot \boldsymbol{x}) + n$$

$$v\,|\,\mathbf{w}\,| = 0$$





Backpropagation ---gradient learning

$$\text{examples}:(y_1,\boldsymbol{x}_1),\cdots(y_t,\boldsymbol{x}_t)$$

$$E=\frac{1}{2}\left|y-f(\boldsymbol{x},\boldsymbol{\theta})\right|^2=-\log p(y,\boldsymbol{x};\boldsymbol{\theta})$$

natural gradient (Riemannian)

$$\tilde{\nabla}E=G^{-1}\nabla E \text{ --steepest descent}$$

$$\Delta\boldsymbol{\theta}_t=-\eta_t\frac{\partial E}{\partial\boldsymbol{\theta}}$$

$$f(\boldsymbol{x},\boldsymbol{\theta})=\sum v_i\varphi(\boldsymbol{w}_i\cdot\boldsymbol{x})$$

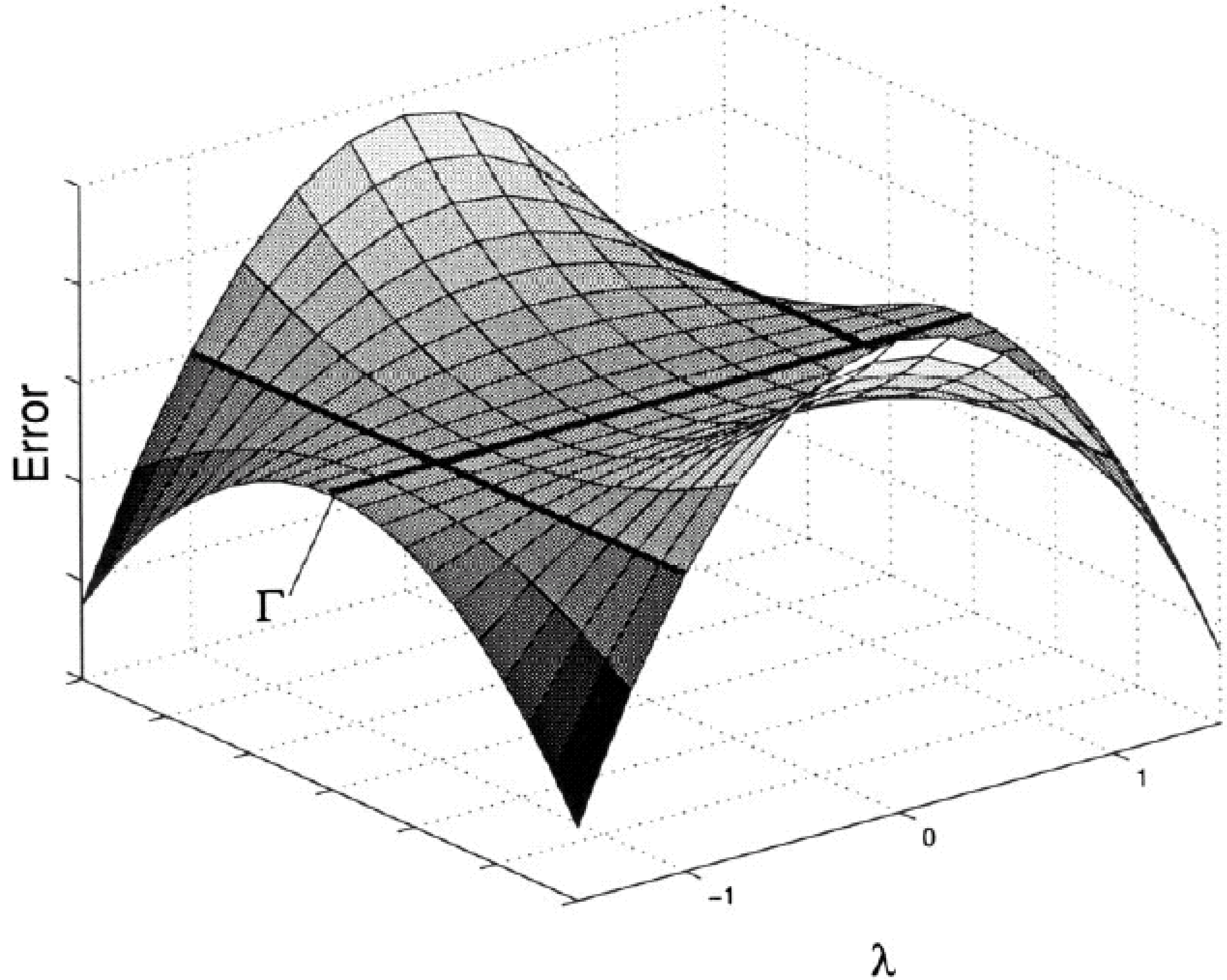
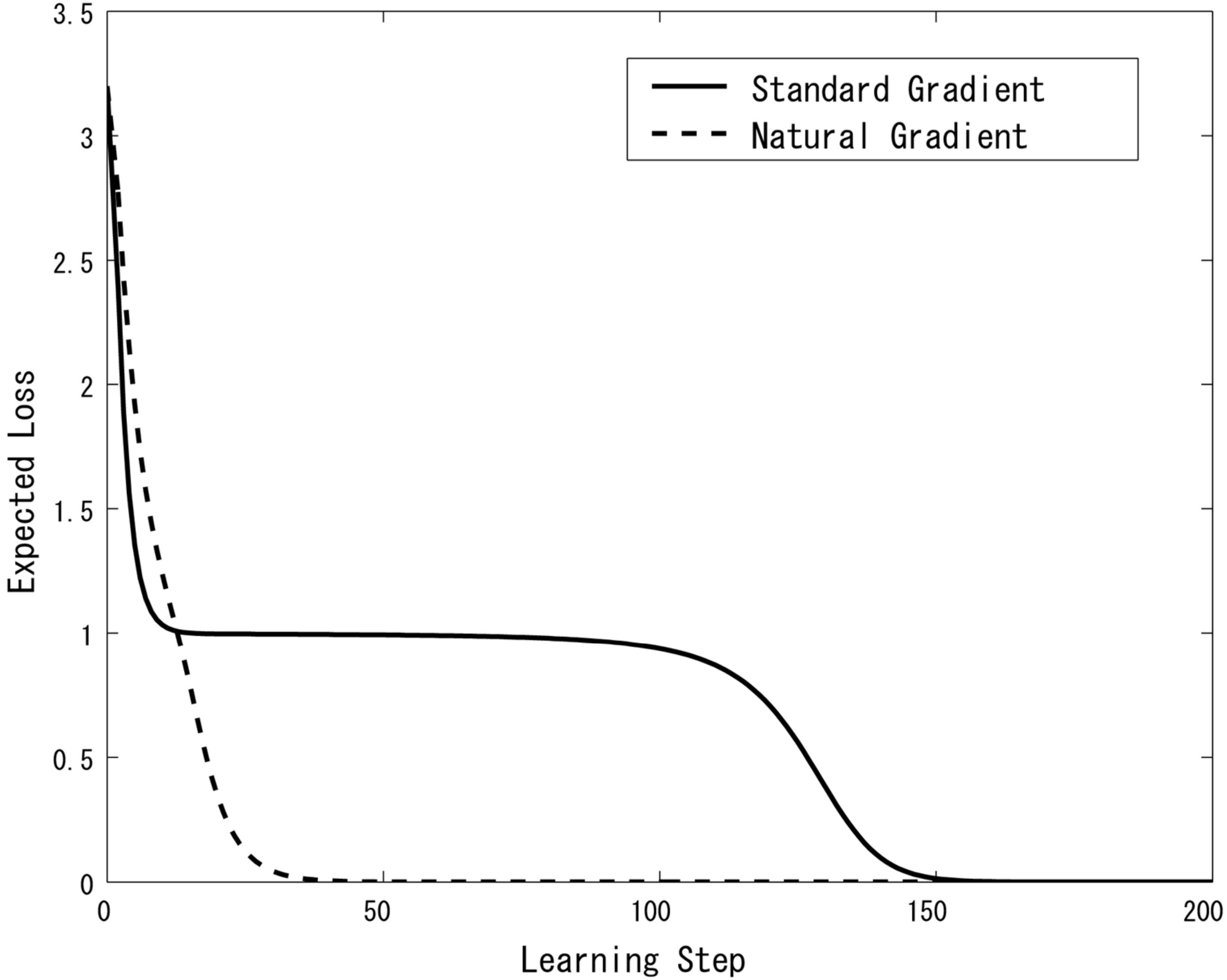


Fig. 5. Critical set with local minima and plateaus.



Standard Gradient
Natural Gradient
Expected Loss
Learning Step
3.5
3
2.5
2
1.5
1
0.5
0
0
50
100
150
200

conformal transformation

q-Fisher information

$$g_{ij}^{(q)}(p) = \frac{q}{h_q(p)} g_{ij}^{F}(p)$$

$q-divergence$

$$D_q[p(x):r(x)] = \frac{1}{(1-q)h_q(p)}(1-\int p(x)^q r(x)^{1-q} dx)$$

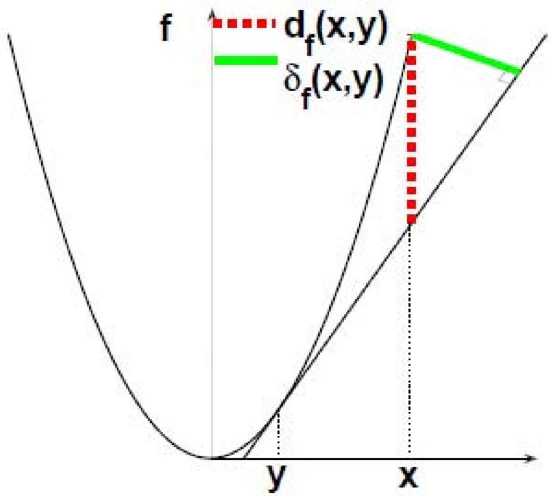
Total Bregman Divergence and its Applications to Shape Retrieval

•Baba C. Vemuri, Meizhu Liu, Shun-ichi Amari, Frank Nielsen

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010

Total Bregman Divergence

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

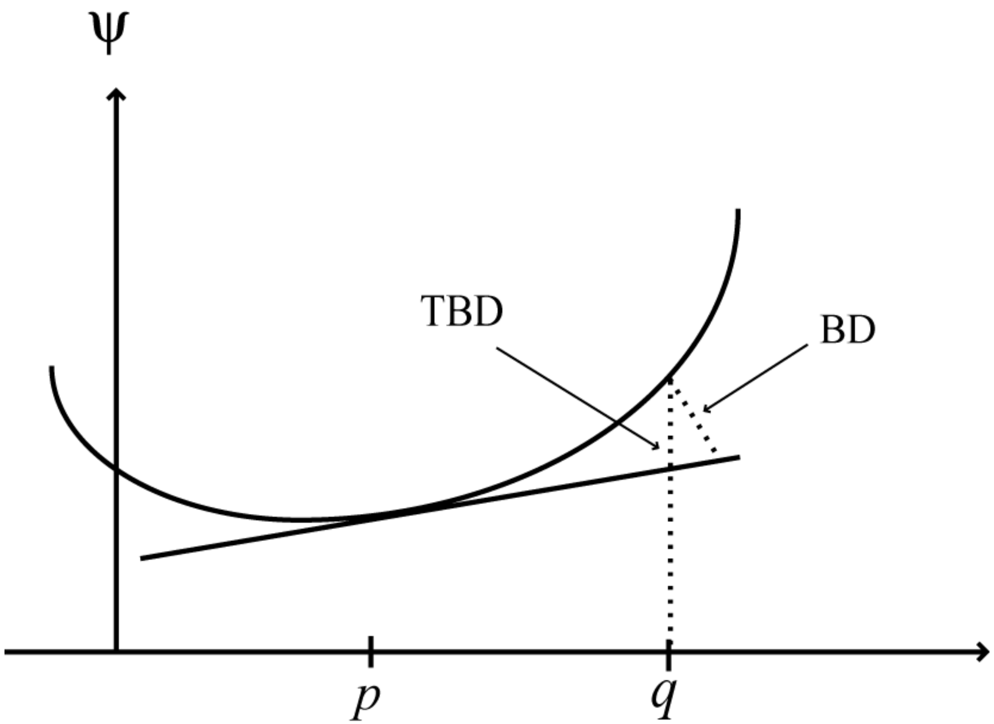


- rotational invariance
- conformal geometry

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.

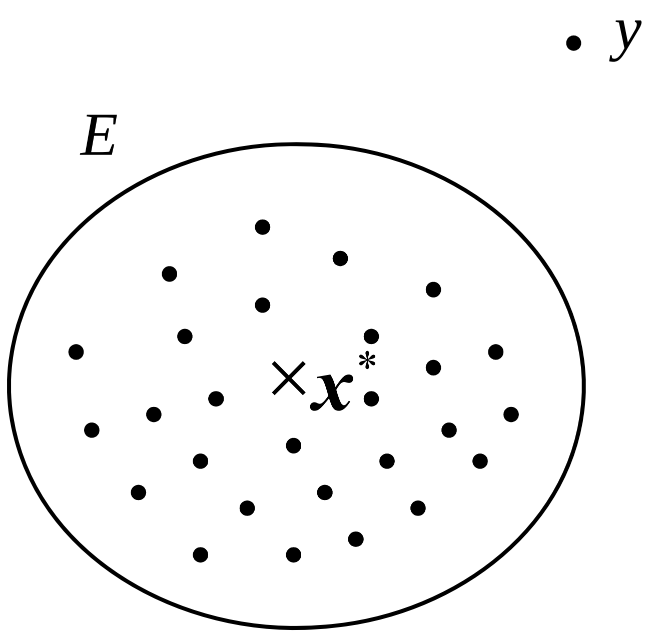
Total Bregman divergence (Vemuri)

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



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 $\boldsymbol{x}^*$

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

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

q -super-robust estimator (Eguchi)

$$\max \hat{p}(x,\xi) \to \max \frac{p(x,\xi)}{h_{q+1}}$$

bias-corrected q-estimating function

$$s_q(x,\xi) = \hat{p}(x,\xi)\left\{\partial_i \log p - c_{q+1}(\xi)\right\}$$

$$c_{q+1} = \frac{1}{q+1}\partial \log h_{q+1}(\xi)$$

$$\sum_{i=0}^{N} s_q(x_i,\xi) = 0 \qquad \Leftrightarrow \max \frac{1}{h_{q+1}} \sum \hat{p}(x_i,\xi)$$

Conformal change of divergence

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

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

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

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

t-center is robust

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

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

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

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

$$\boldsymbol{z}\left(\boldsymbol{x}^*,\boldsymbol{y}\right)=G^{-1}\frac{\nabla f\left(\boldsymbol{y}\right)-\nabla f\left(\boldsymbol{x}^*\right)}{w\left(\boldsymbol{y}\right)}$$

$$G=\frac{1}{n}\sum w\left(\boldsymbol{x}_i\right)\nabla\nabla f\left(\boldsymbol{x}_i\right)$$

Robust: $\boldsymbol{z}$ is bounded

$$\boldsymbol{z}\left(\boldsymbol{x}^*,\boldsymbol{y}\right)=G^{-1}\frac{\boldsymbol{y}}{\sqrt{1+\left\|\boldsymbol{y}\right\|^2}}$$

$$\frac{\nabla f\left(\boldsymbol{y}\right)}{w\left(\boldsymbol{y}\right)}=\frac{\nabla f\left(\boldsymbol{y}\right)}{\sqrt{1+\left\|\nabla f\left(\boldsymbol{y}\right)\right\|^2}}<\infty$$

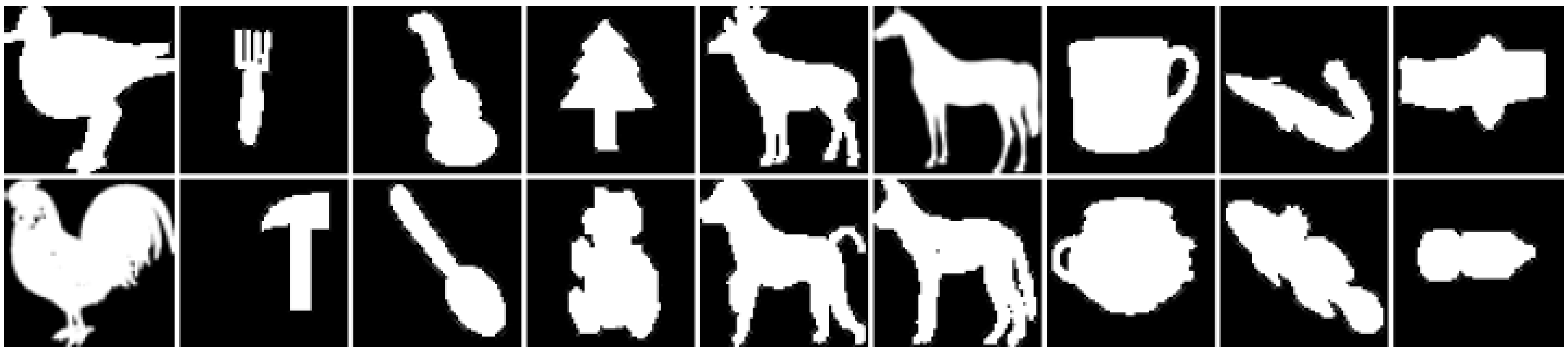
$$\boldsymbol{z}\left(\boldsymbol{x}^*,\boldsymbol{y}\right)=\boldsymbol{y}$$

Euclidean case $f=\frac{1}{2}\left|\boldsymbol{x}\right|^2$

$$w\left(\boldsymbol{y}\right)>1$$

MPEG7 database

- Great intraclass variability, and small interclass dissimilarity.



Other TBD applications

Diffusion tensor imaging (DTI) analysis [Vemuri]

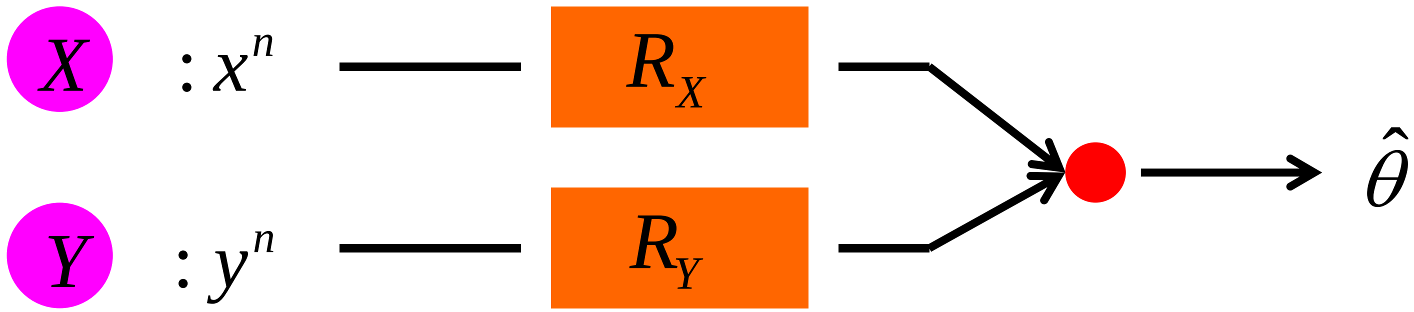
- **Interpolation**
- **Segmentation**

Baba C. Vemuri, Meizhu Liu, Shun-ichi Amari and Frank Nielsen, *Total Bregman Divergence and its Applications to DTI Analysis*, IEEE TMI, to appear

TBD application-shape retrieval

- Using MPEG7 database;
- 70 classes, with 20 shapes each class
  (Meizhu Liu)

Multiterminal Information & Statistical Inference

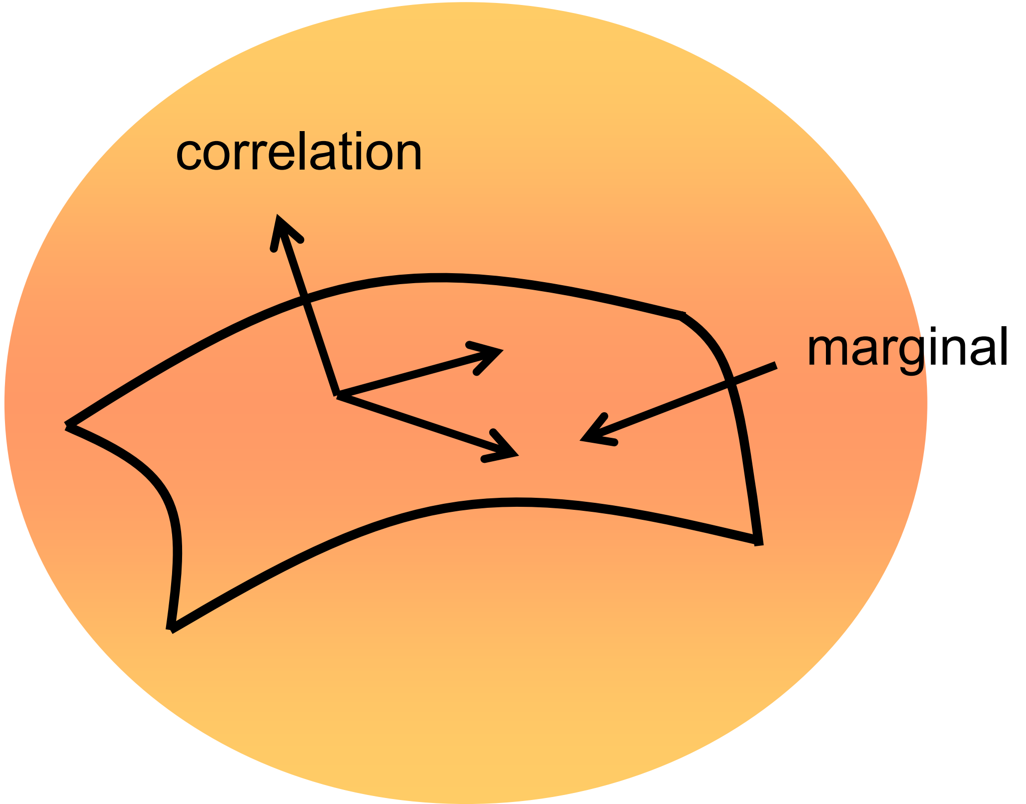


$$p\left(x, y; \theta\right)$$

$$\left|M_X\right| = 2^{R_X n} \quad \left|M_r\right| = 2^{R_Y n}$$

$p(x,y)$

$p(x,y,z)$



$$G = G_M + G_C$$

0-rate Slepian-Wolf

Linear Systems

$$u_t \longrightarrow \boxed{} \longrightarrow x_t$$

ARMA

$$x_{t+1} = \frac{1 + b_1 z^{-1} + \cdots + b_q z^{-p}}{1 + a_1 z^{-1} + \cdots + a_p z^{-p}} u_t$$

$$\theta = \left(a_1, \cdots, a_p : b_1, \cdots, b_q \right)$$

$$\boldsymbol{x}_{t+1} = f\left(\boldsymbol{\theta}, z^{-1}, u_t \right)$$

AR---e-flat

MA---m-flat

Machine Learning

Boosting : combination of weak learners

$$D = \left\{ \left(\boldsymbol{x}_1, y_1 \right), \left(\boldsymbol{x}_2, y_2 \right), \cdots, \left(\boldsymbol{x}_N, y_N \right) \right\}$$

$$y_i = \pm -1$$

$$f\left(\boldsymbol{x}, \boldsymbol{u} \right): y = h\left(\boldsymbol{x}, \boldsymbol{u} \right) = \operatorname{sgn} f\left(\boldsymbol{x}, \boldsymbol{u} \right)$$

Weak Learners

$$H(\boldsymbol{x}) = \mathrm{sgn}\left(\sum \alpha_t h_t(\boldsymbol{x})\right)$$

$$\varepsilon_t : \ \mathrm{Prob}\ \left\{h_t(\boldsymbol{x}_i) \neq y_i\right\} \qquad \Big| W_t$$

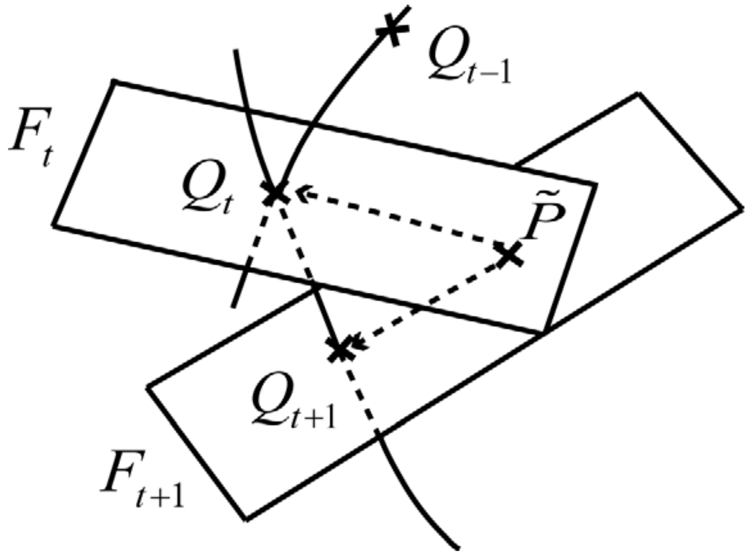
$$W_{t+1}(i) = cW_t(i)\exp\left\{-\alpha_t y_i h_t(x_i)\right\}$$

weight distribution

Boosting —generalization

$$Q_t = \left\{ Q_t\left(y|x\right) = Q_{t-1}\left(y|x\right)\exp\left\{\alpha_t y h_t\left(x\right) - \tilde{f}\right\}\right\}$$

$$F_t = \left\{ P\left(y,x\right) Eyh_t\left(x\right) = \text{const} \right\}$$



$$\alpha_t : \min D\left[\tilde{P} : Q_t\right]$$

$$D\left(\tilde{P}, Q_{t+1}\right) < D\left(\tilde{P}, Q_t\right)$$

Integration of evidences:

$$x_1, x_2, \ldots x_m$$

arithmetic mean

geometric mean

harmonic mean

α -mean

Various Means

$$\frac{a+b}{2} \quad : \quad \sqrt{ab} \quad : \quad \frac{2}{\frac{1}{a}+\frac{1}{b}}$$

arithmetic   geometric   harmonic

Any other mean?

Generalized mean: f-mean

f(u): monotone; f-representation of u

$$m_f(a,b) = f^{-1}\{\frac{f(a)+f(b)}{2}\}$$

scale free $\quad m_f(ca,cb) = cm_f(a,b)$

α-representation $\quad f_\alpha(u) = u^{\frac{1-\alpha}{2}}, \quad \alpha \neq 1$

$$\log u, \quad \alpha = 1$$

α-mean : $m_\alpha(p_1(s), p_2(s))$

$$\alpha = 1 \quad : \quad \sqrt{ab}$$

$$\alpha = -1: \quad \frac{a+b}{2}$$

$$\alpha = 0 \quad : \quad (\sqrt{a} + \sqrt{b})^2 = \frac{a+b}{4} + \frac{1}{2}\sqrt{ab}$$

$$\alpha = \infty \qquad m_\alpha = \min(a, b)$$

$$\alpha = -\infty \qquad m_\alpha = \max(a, b)$$

α — Family of Distributions

$$\{p_1(s),\cdots,p_k(s)\} \qquad p(x;\theta)=f_\alpha^{-1}\{\sum\theta_i f_\alpha(p_i(x))\}$$

mixture family :

$$p_{mix}(s)=\sum_{i=1}^{k}t_i p_i(s), \qquad \sum t_i=1$$

exponential family :

$$\log p_{\exp}(s)=\sum t_i \log p_i(s)-\psi$$