Theoretical Sciences

Information Geometry and Its Applications

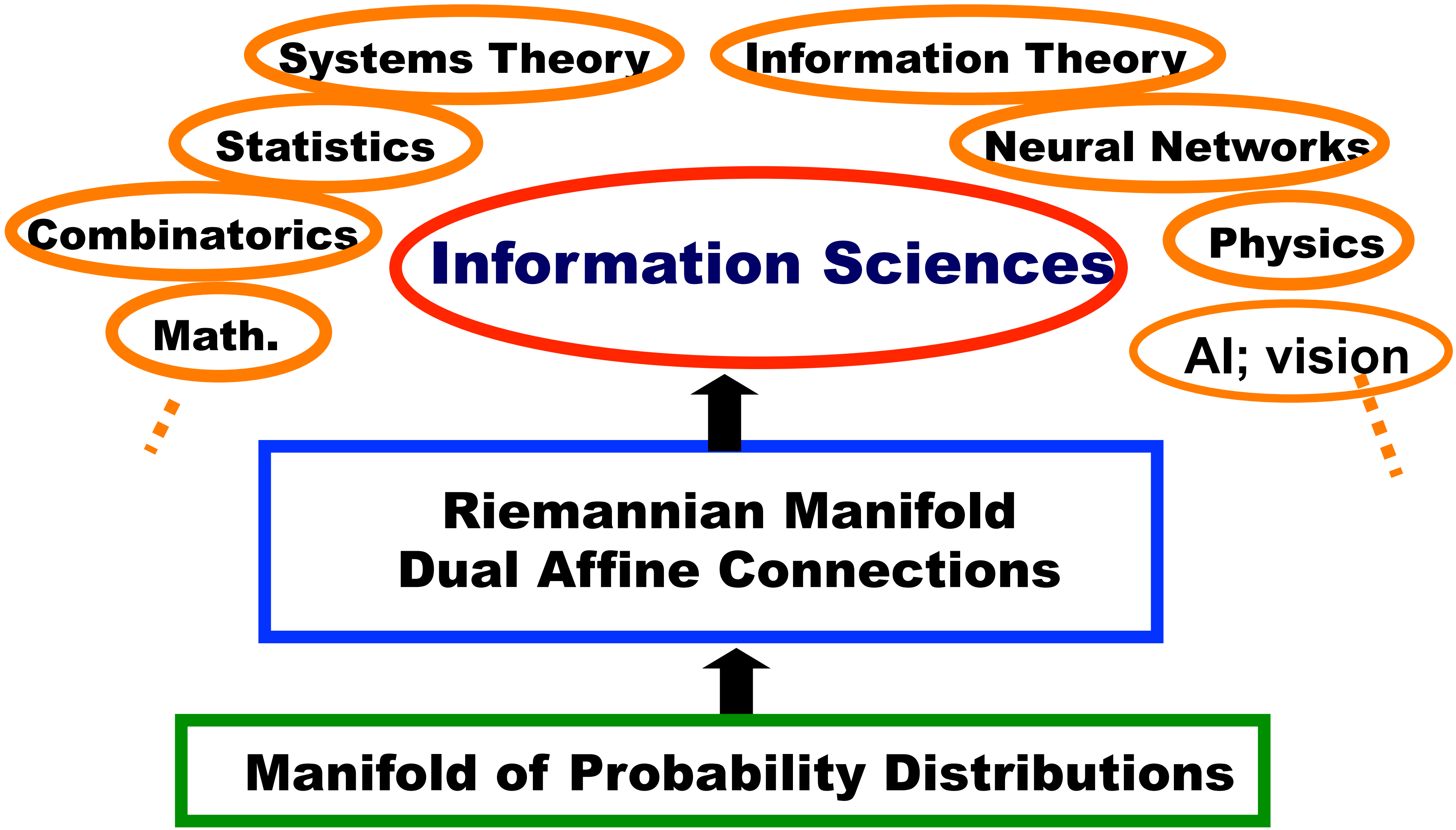
Shun-ichi Amari
RIKEN Brain Science Institute

Information Geometry:

A Unifying Framework

Statistical Inference,
Information Sciences
Signal Processing,
Machine Learning
Convex Analysis
Physics
Brain Science

Information Geometry



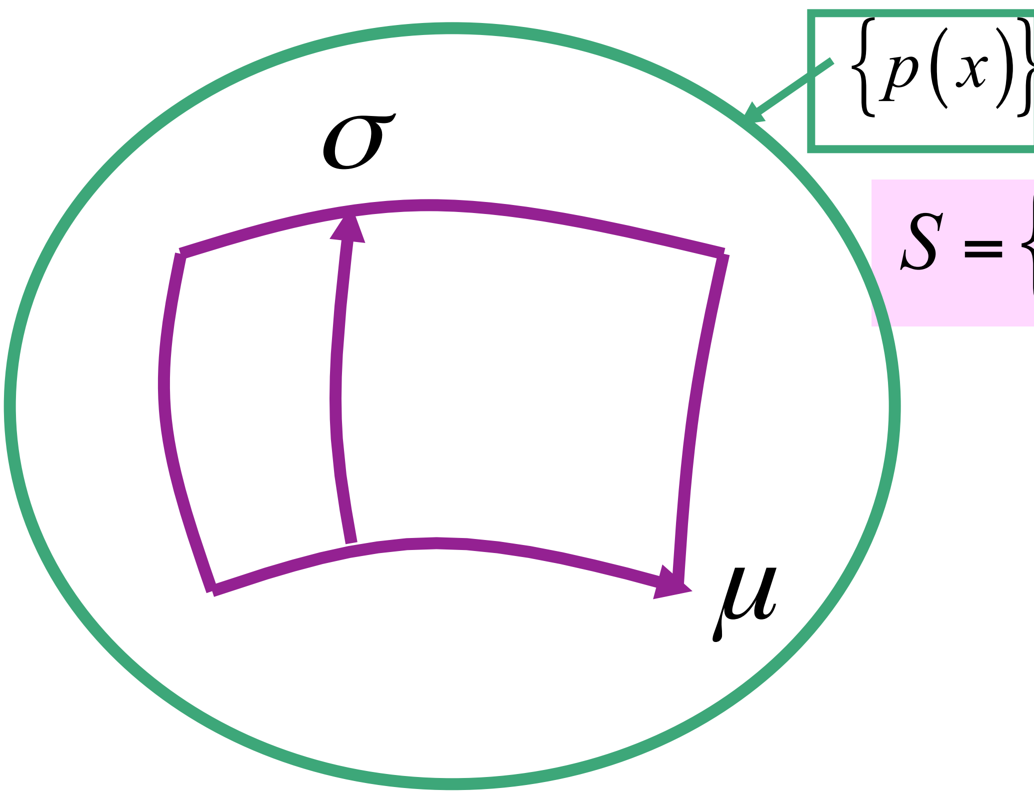
Information Geometry

-- Manifolds of Probability Distributions

Information Geometry ?

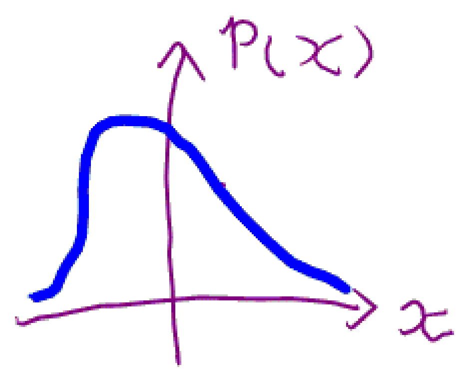
Gaussian distributions

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



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

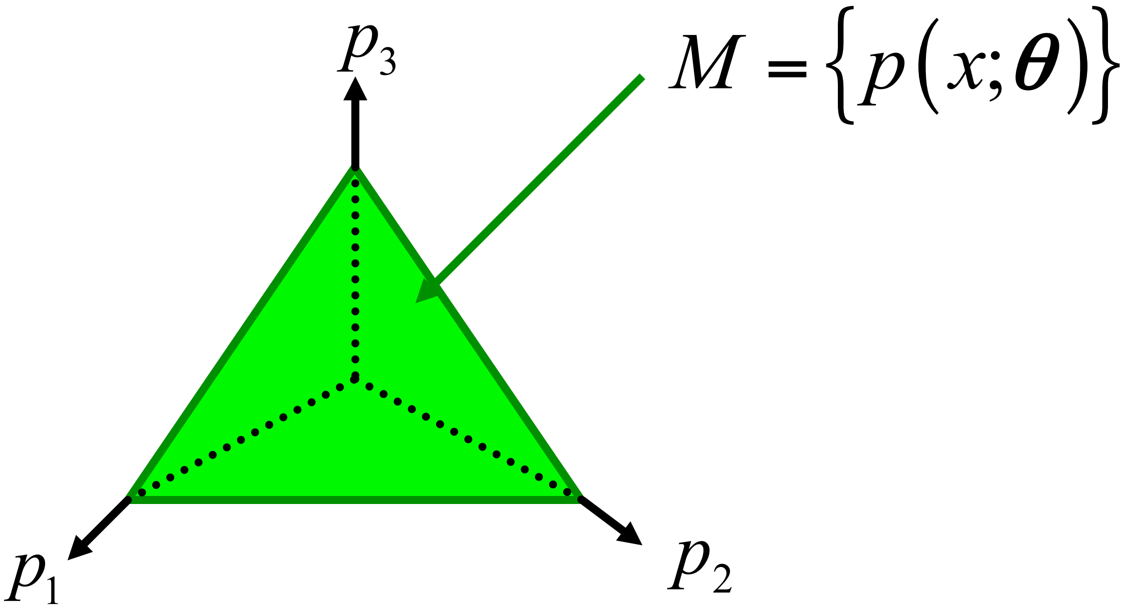
$$S=\left\{p\left(x;\boldsymbol{\theta}\right)\right\}$$



Manifold of Probability Distributions

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

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



Invariance

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

1. *Invariant under reparameterization*

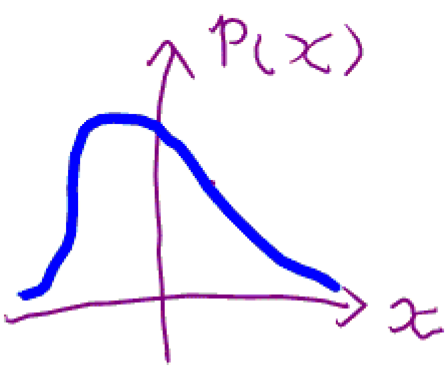
$$p(x, \theta) = \overline{p}\left(x, \boldsymbol{\xi}\right) \qquad D = \sum \theta_i^2 \neq \sum \xi_i^2$$

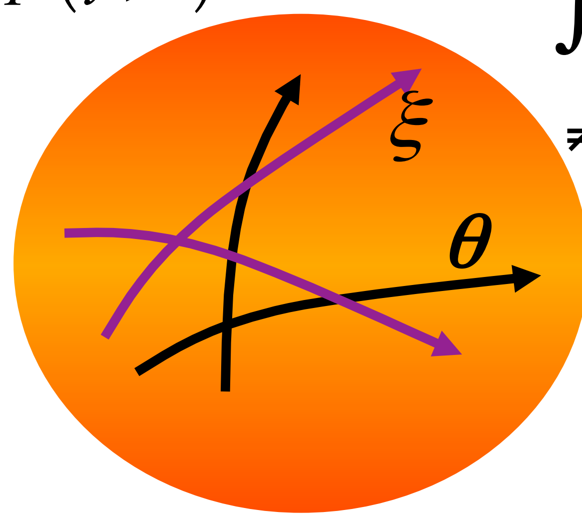
2. *Invariant under different representation*

$$y = y\left(x\right), \qquad \overline{p}\left(y, \boldsymbol{\theta}\right)$$

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

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

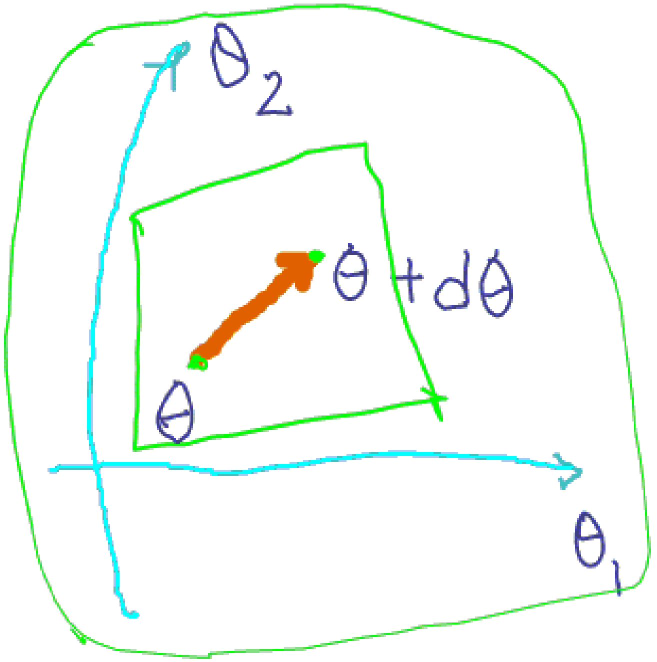




Two Structures

Riemannian metric

affine connection --- geodesic



Riemannian Structure

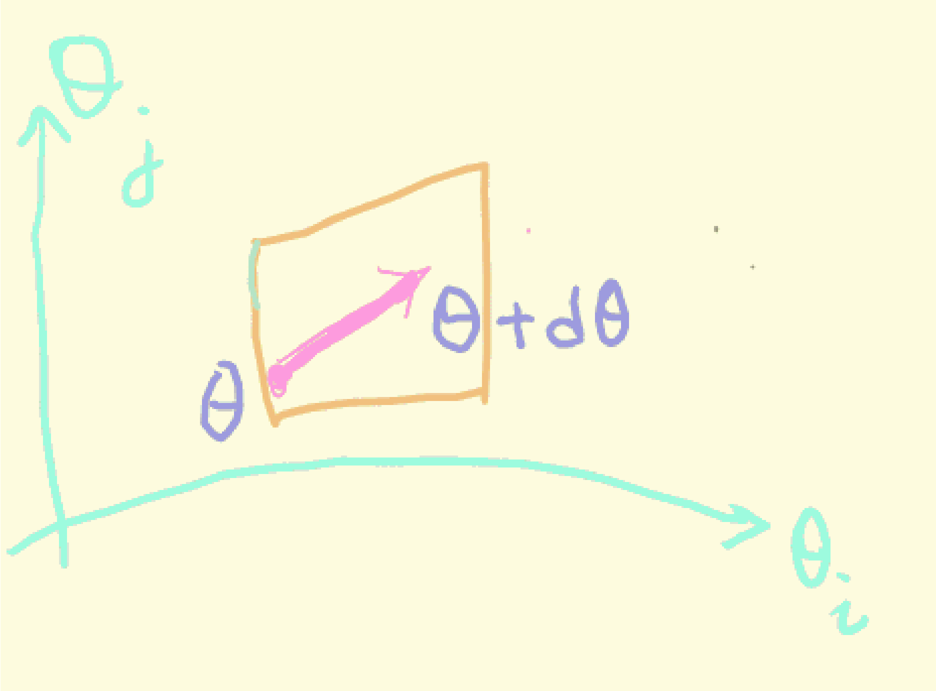
$$ds^2 = \sum g_{ij}(\theta) d\theta^i d\theta^j$$
$$= d\theta^T G(\theta) d\theta$$

Fisher information

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

$$G(\theta) = (g_{ij})$$

Euclidean $G = E$



Affine Connection

covariant derivative; parallel transport

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

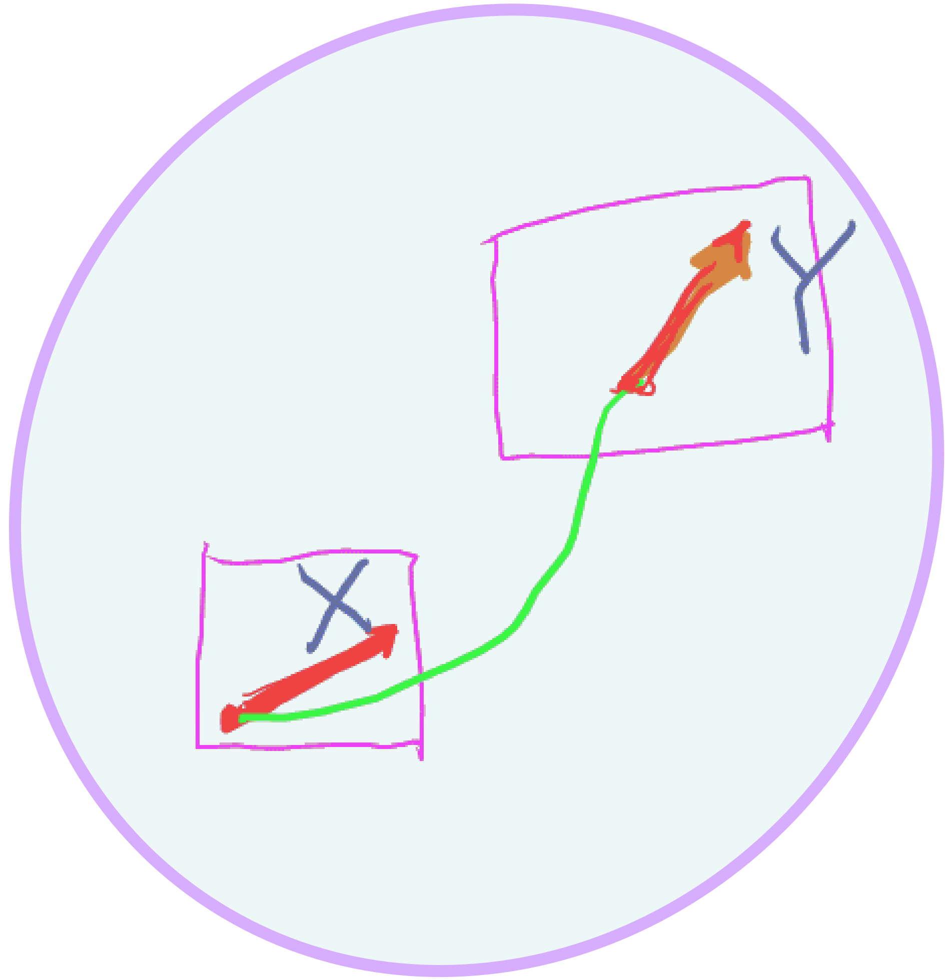
geodesic $\Pi\dot{X} = \dot{X}$ $X = X(t)$

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

minimal distance

straight line

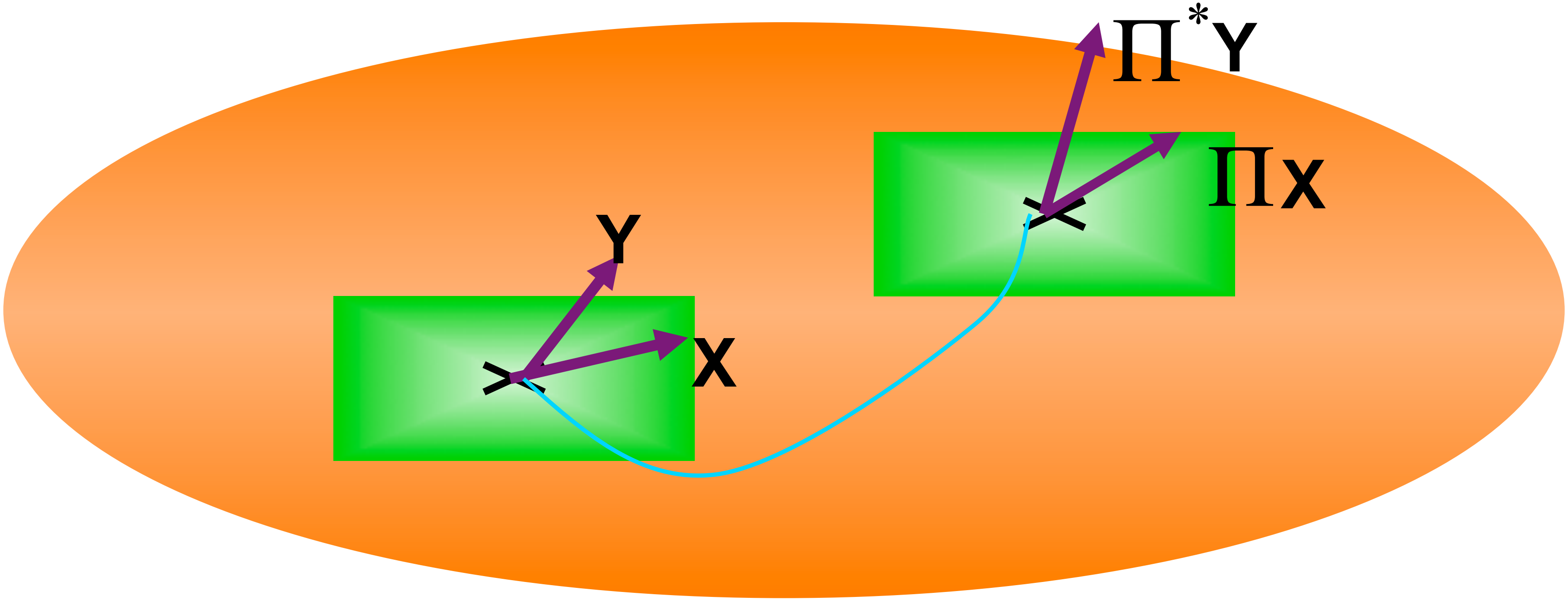
different concepts



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

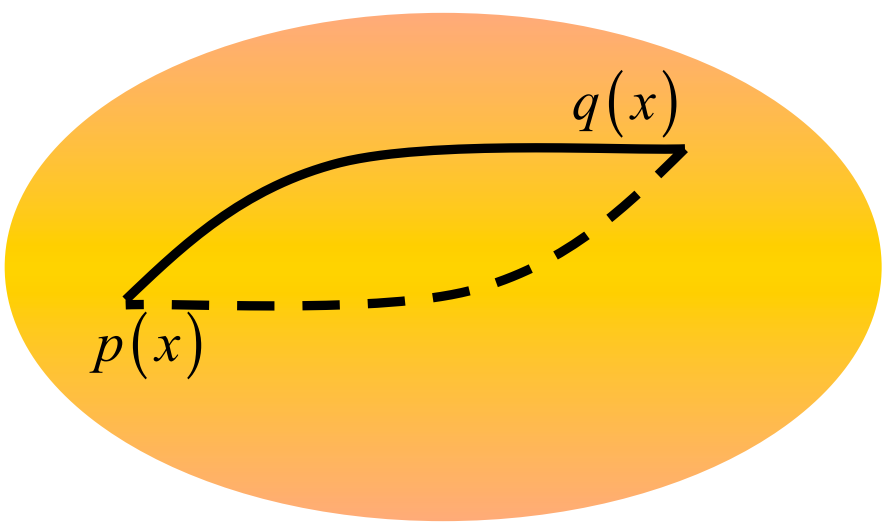
$$\log r(x,t) = t \log p(x) + (1-t) \lg q(x) + c(t)$$

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

$$(S,\ g,\ T)$$

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

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

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

α **-connection**

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

Dually flat

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

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

Dually flat manifold

θ-coordinates $\leftrightarrow$ η-coordinates

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

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

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

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

ψ : cumulant generating function

φ : negative entropy

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

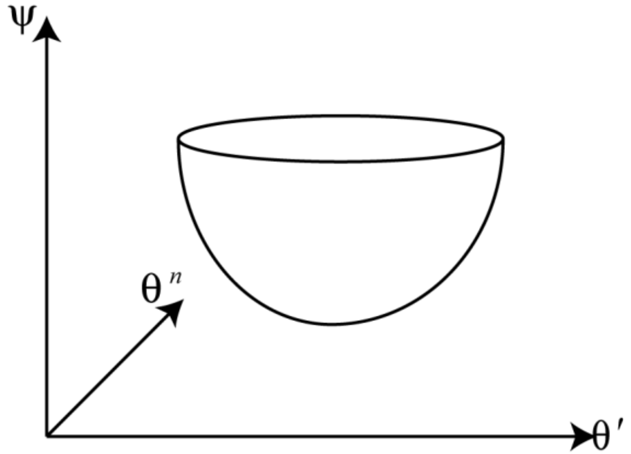
Information Geometry
-- Dually Flat Manifold

Convex Analysis
Legendre transformation
Divergence
Pythagorean theorem
I-projection

Manifold with Convex Function

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

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



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

negative entropy

$$\varphi\left(p\right) = \int p\left(x\right)\log p\left(x\right)dx$$

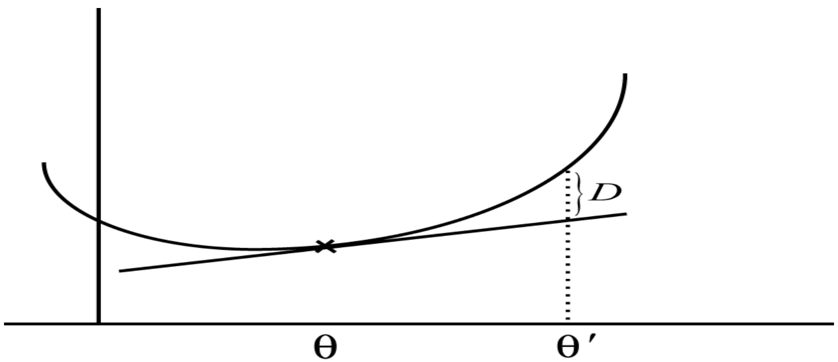
energy

mathematical programming, control systems
physics, engineering

Riemannian metric and flatness

Bregman divergence

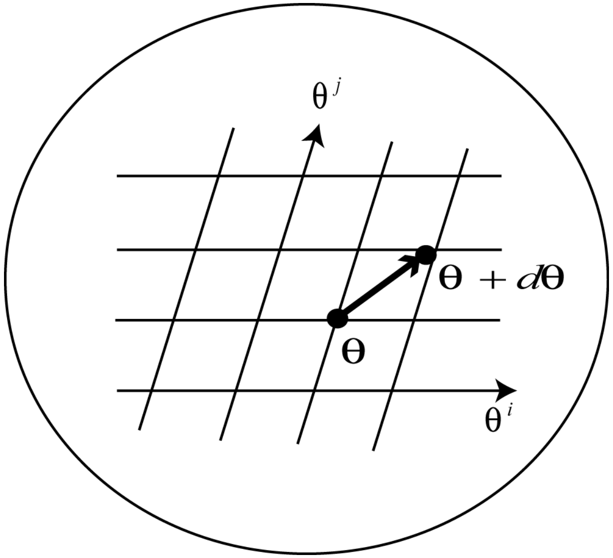
$$D(\boldsymbol{\theta}', \boldsymbol{\theta}) = \psi(\boldsymbol{\theta}') - (\boldsymbol{\theta}' - \boldsymbol{\theta}) \cdot \mathrm{grad}\, \psi(\boldsymbol{\theta})$$



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

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

$\boldsymbol{\theta}$:geodesic



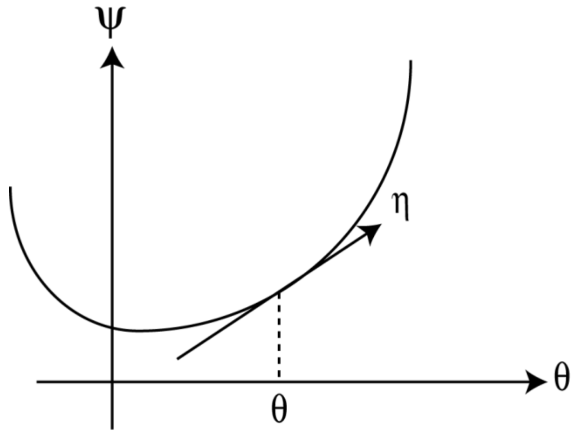
Legendre Transformation

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

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

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

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



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

Two coordinate systems of S

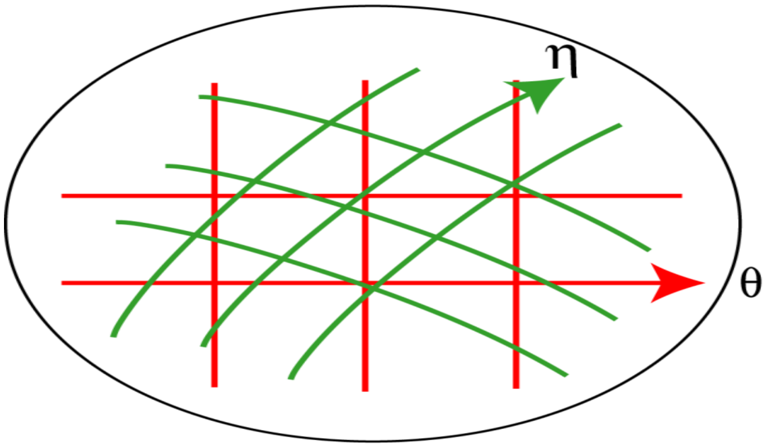
$$(\theta, \eta)$$

θ : geodesic (e-geodesic)

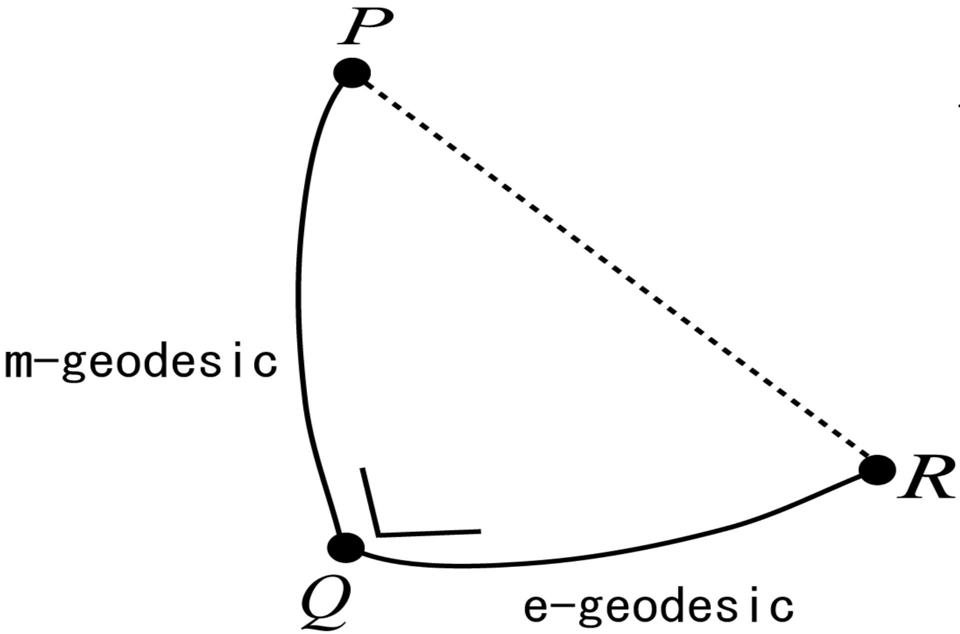
η : dual geodesic (m-geodesic)

"orthogonal"

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



Pythagorean Theorem



$$D\left[P:Q\right]+D\left[Q:R\right]=D\left[P:R\right]$$

Euclidean space: self-dual $\boldsymbol{\theta}=\boldsymbol{\eta}$

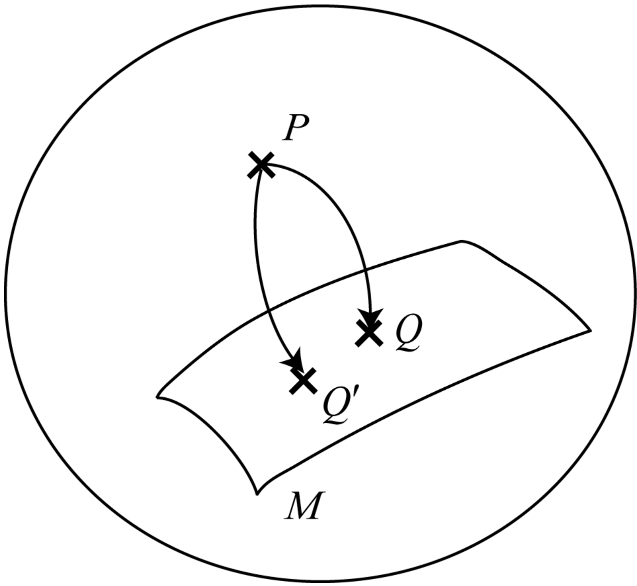
$$\psi\left(\theta\right)=\frac{1}{2}\sum\left(\theta_i\right)^2$$

Projection Theorem

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

Q = **m-geodesic projection of**

P to M



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

Q = **e-geodesic projection of** P **to** M

Information Geometry

Dually flat manifold; curved submanifold

convex potential functions $\psi(\boldsymbol{\theta}), \varphi(\boldsymbol{\eta})$

Euclidean space : self-dual

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

Probability distributions $S = \{p(x)\}$

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

$$\boldsymbol{\eta} = E[\boldsymbol{x}]$$

$\varphi(\boldsymbol{\eta})$ **: negentropy**

Dually flat manifold

θ-coordinates $\leftrightarrow$ η-coordinates

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

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

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

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

ψ : cumulant generating function

φ : negative entropy

Divergence

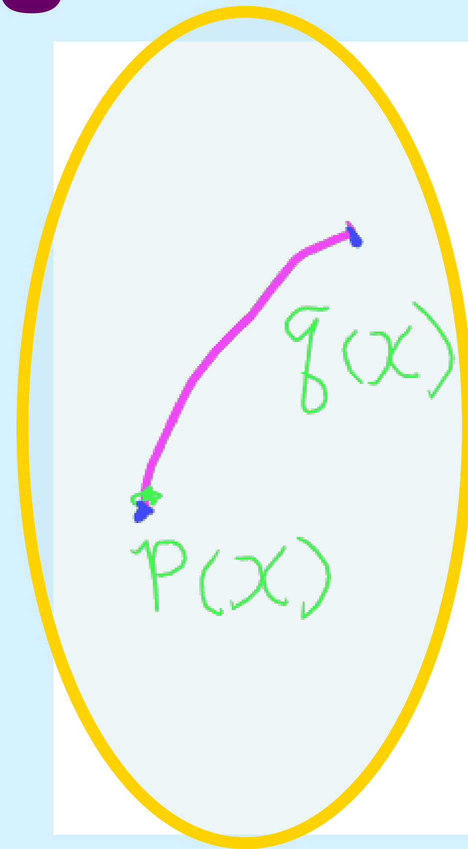
Kullback-Leibler Divergence

quasi-distance

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

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

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



Dually Flat Manifold

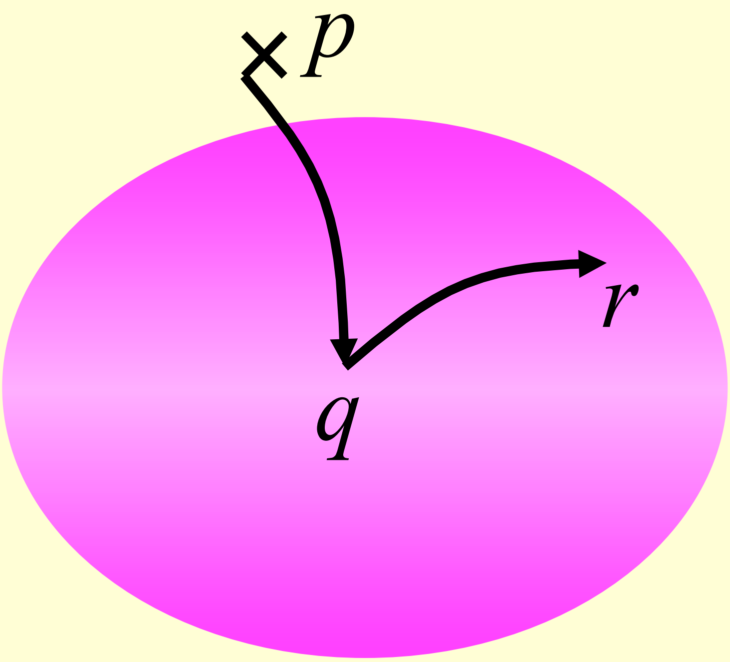
1. **Potential Functions**

   ---convex (Legendre transformation)

2. **Divergence** $D[p:q]$ Bregman divergence

3. **Pythagoras Theorem**

$$D[p:q]+D[q:r]=D[p:r]$$



4. **Projection Theorem**

5. **Dual foliation**

Applications to Statistics

curved exponential family:

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

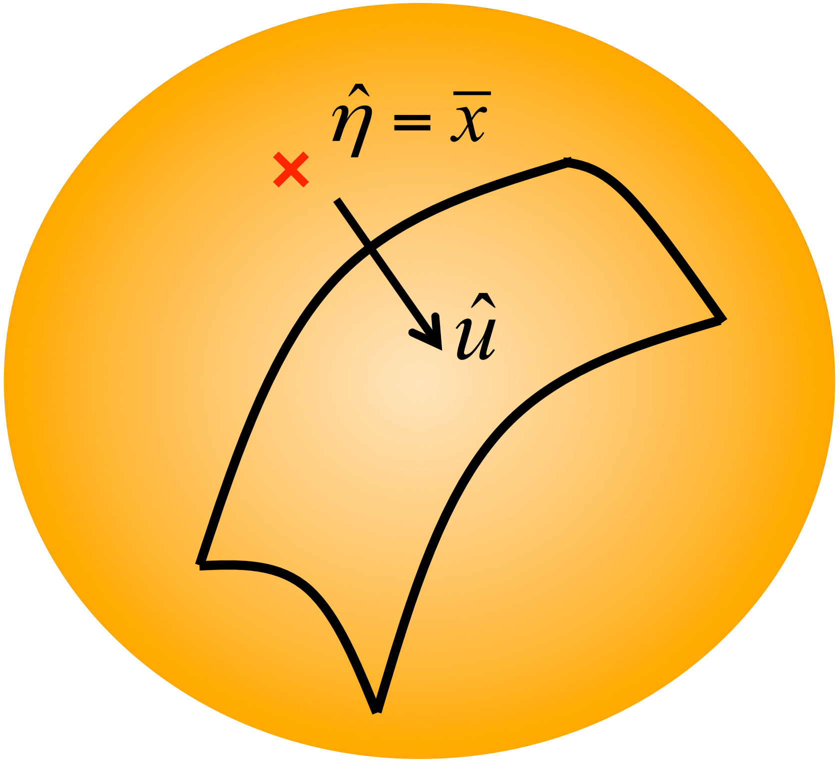
$$p(x,u): \ x_{1,} x_2, \ldots x_n$$

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

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

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

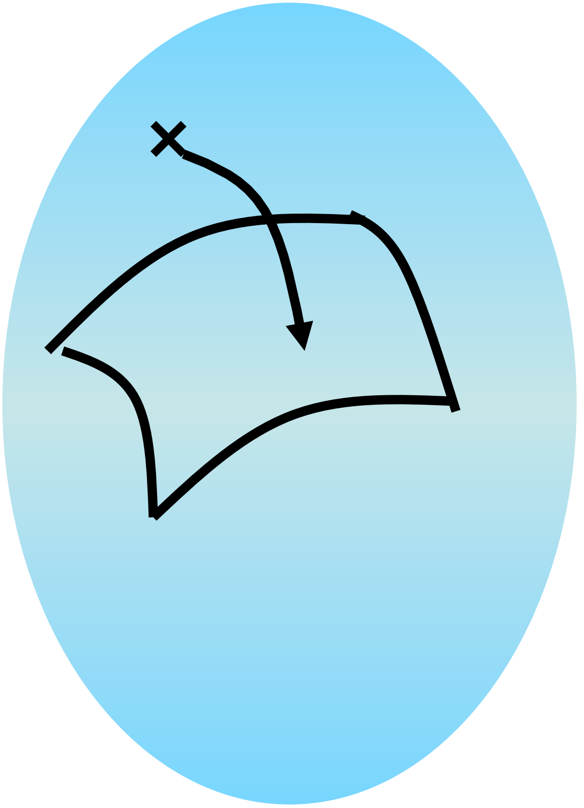
$H_0 : u = u_0$: testing



High-Order Asymptotics

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

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



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

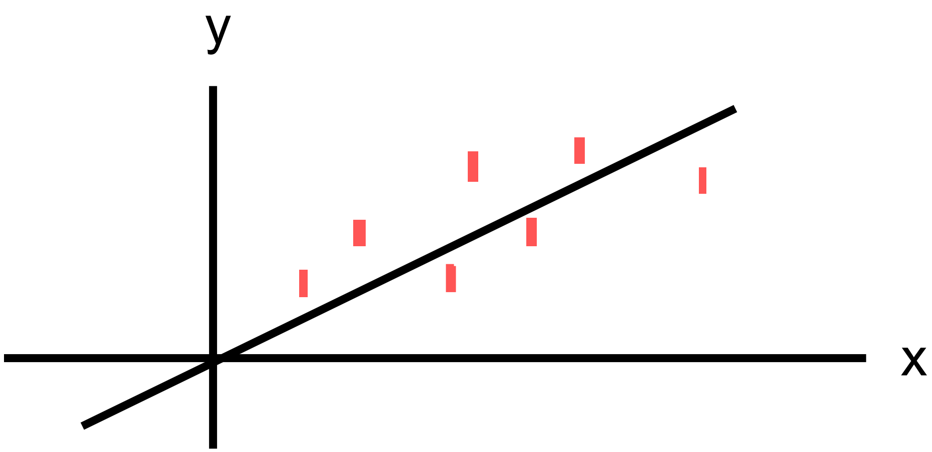
$$e = \frac{1}{n} G_1 + \frac{1}{n^2} G_2$$

$$G_1 \geq G^{-1}$$:Cramér-Rao

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

Semiparametric Statistical Model

$$M = \left\{ p\left(x, \theta, r \right) \right\}$$



linear relation

$$y = \theta x$$

$$\begin{cases} y_i = \theta \xi_i + \varepsilon_i \\ x_i = \xi_i + \varepsilon_i^{'} \end{cases}$$

$$p\left(x, y; \theta \right) = \int p\left(x, y; \xi, \theta \right) r\left(\xi \right) d\xi$$

mle, least square, total least square

Linear Regression: Semiparametrics

$$(x_1, y_1)$$

$$(x_2, y_2)$$

M

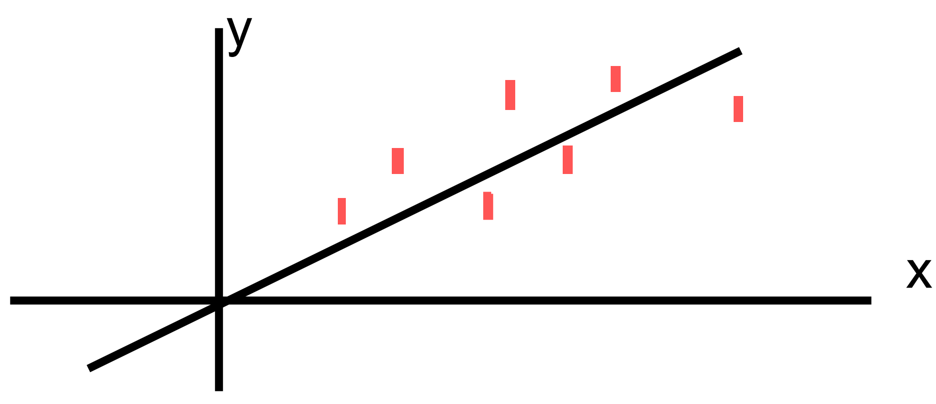
$$(x_n, y_n)$$

$$y = \theta x$$

$$x_i = \xi_i + \varepsilon_i$$

$$y_i = \theta \xi_i + \varepsilon_i'$$

$$\varepsilon_i, \varepsilon_i' : N(0, \sigma^2)$$



Statistical Model

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

$$p\left(x_i, y_i \middle| \theta, \xi_i\right) \; : \; \theta, \xi_1, \ldots, \xi_n$$

$$p\left(x, y \middle| \theta\right) = \int p\left(x, y \middle| \theta, \xi\right) Z\left(\xi\right) d\xi$$

—— **semiparametric**

Least squares?

$$L(\theta) = \sum (y_i - \theta x_i)^2 \rightarrow \min \qquad : \hat{\theta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\frac{1}{n} \sum \frac{y_i}{x_i}, \qquad \frac{\sum y_i}{\sum x_i}$$

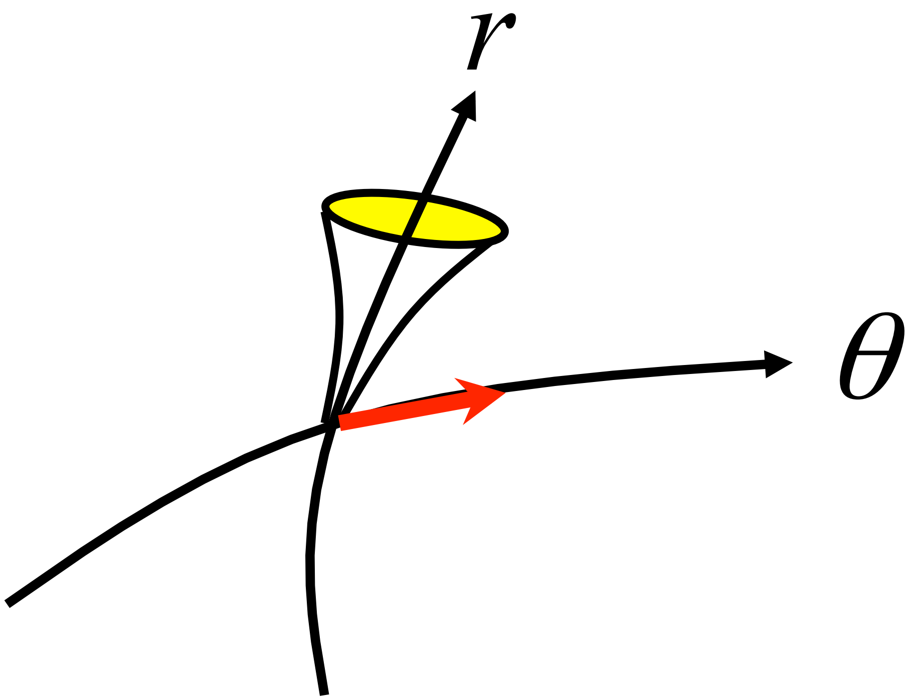
mle, TLS

$$\sum (y_i - \theta x_i)(\theta y_i + x_i) = 0$$

Neyman-Scott

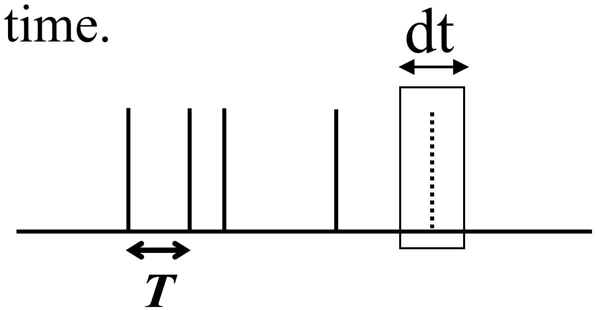
Fibre bundre

function space



Poisson process

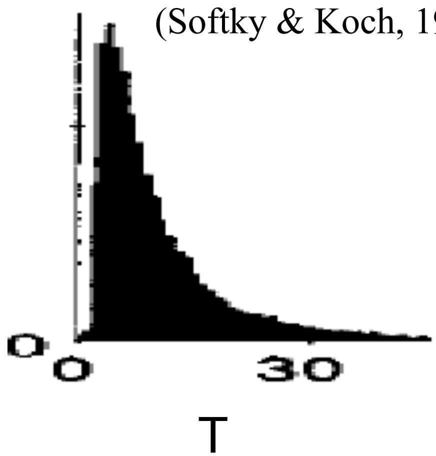
Poisson Process: Instantaneous firing rate is constant over time.



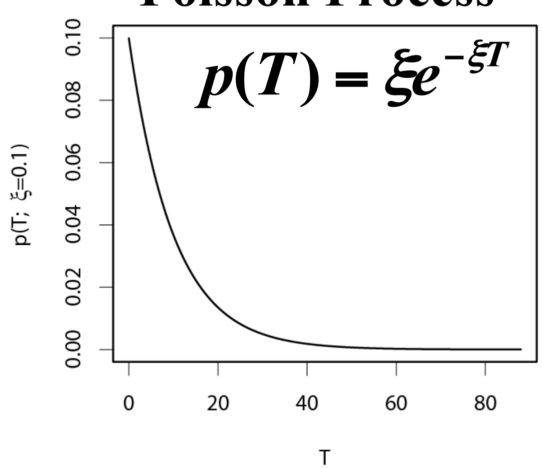
For every small time window dt, generate a spike with probability ξdt.

Cortical Neuron

(Softky & Koch, 1993)



Poisson Process



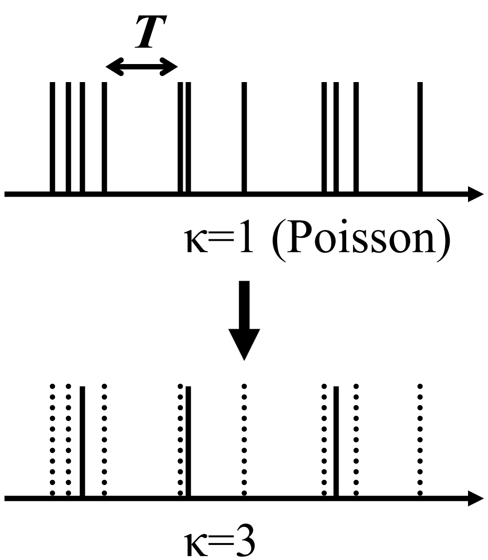
Poisson process cannot explain inter-spike interval distributions.

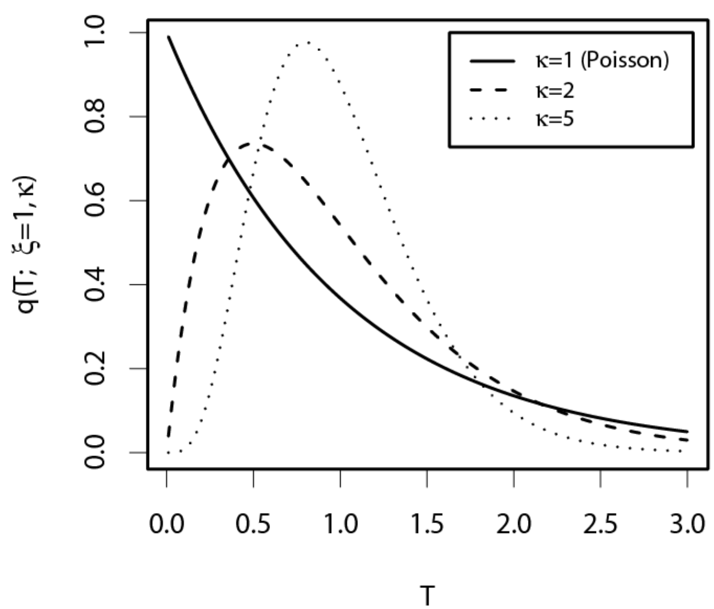
Gamma distribution

Gamma Distribution: Every κ-th spike of the Poisson process is left.

$$\mathbf{q(T;\xi,\kappa) = \frac{(\xi\kappa)^{\kappa}}{\Gamma(\kappa)} T^{\kappa-1} e^{-\xi\kappa T}.}$$

Two parameters $\begin{cases} \xi\text{: } \textbf{Firing rate} \\ \kappa\text{: } \textbf{Irregularity} \end{cases}$





Gamma distribution

$$f(T) = \frac{(r\kappa)^{\kappa}}{\Gamma(\kappa)} T^{\kappa-1} \exp\{-r\kappa T\}$$

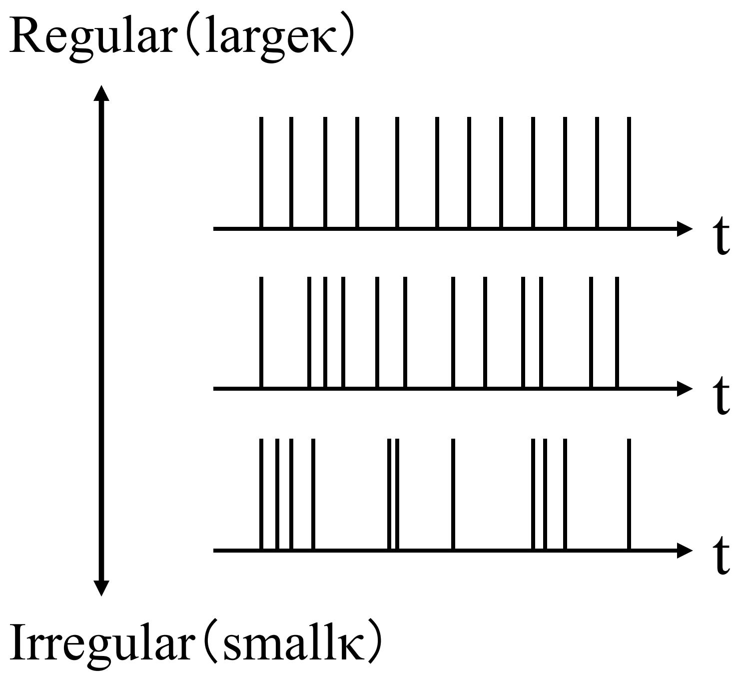
$\kappa = 1$: Poisson

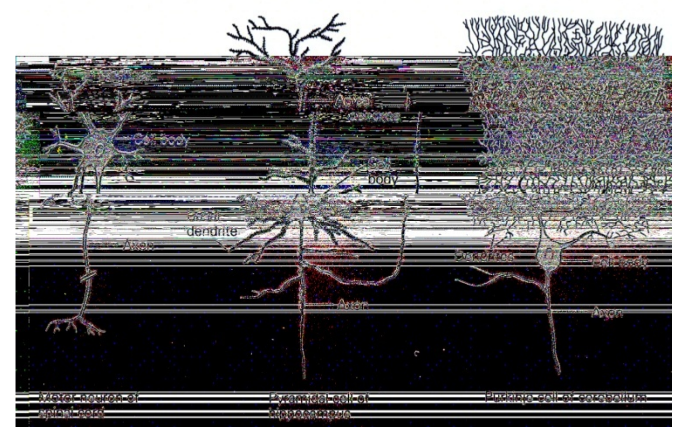
$\kappa \to \infty$: regular

Integrate-and fire

Markov model

Irregularity κ is unique to individual neurons.





Irregularity varies among neurons.
(Baker & Lemon 2000; Shinomoto et.al., 2003)

➡ **<u>We assume that κ is independent of time.</u>**

Bayesian Information Geometry

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta} \cdot \boldsymbol{x} + k(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})\right\}$$

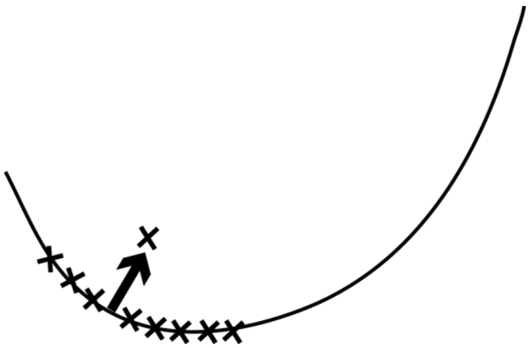
$$p(\boldsymbol{x} | \boldsymbol{\theta}), \qquad p(\boldsymbol{\theta} | \boldsymbol{x})$$

Predictive Distribution

$$D=\left\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\right\}$$

$$p\left(\boldsymbol{x}\middle|D\right)=\int p\left(\boldsymbol{x}\middle|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\middle|D\right)d\boldsymbol{\theta}$$

$$=p\left(\boldsymbol{x},\hat{\boldsymbol{\theta}}\right)+\frac{1}{N}H_{ij\kappa}\hat{v}^{\kappa}$$



$$\underset{p}{\mathrm{Min}}\,E\left[KL\left[p\left(\boldsymbol{x},\boldsymbol{\theta}_0\right):p\left(\boldsymbol{x}\middle|D\right)\right]\right]$$

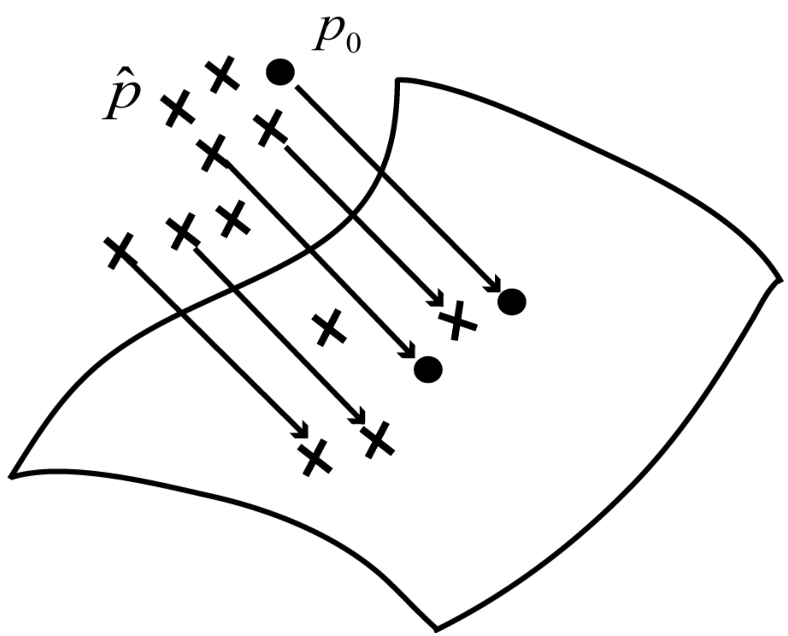
$$\mathrm{Min}\,ED_{\alpha}\left[p\left(\boldsymbol{x},\boldsymbol{\theta}_0\right):p\left(\boldsymbol{x}\middle|D\right)\right]$$

α **-predictive distribution**

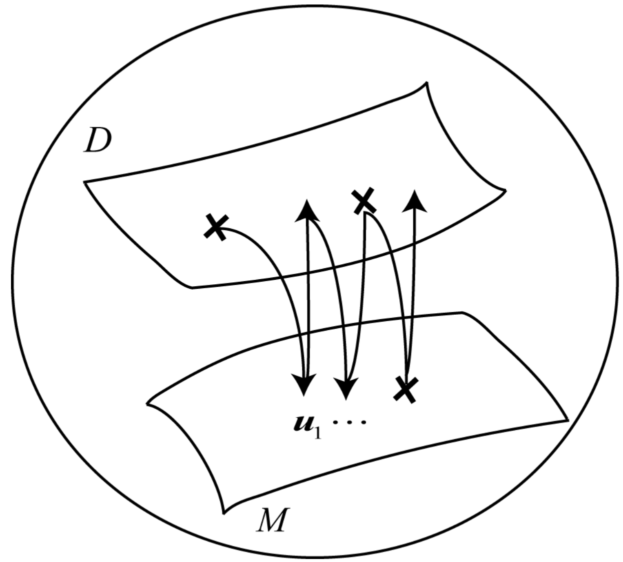
Bootstrap Method

$$D = \left\{ x_1, \underset{p_0}{\ldots}, x_N \right\} \Rightarrow \hat{p}(x)$$

resampling :



EM algorithm



hidden variables

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

$$D = \{\boldsymbol{x}_1, \mathrm{L}\ , \boldsymbol{x}_N\}$$

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

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

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

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

Tsallis q-entropy

conformal information geometry

$$H_T = E\left[\ln_q \frac{1}{p(x)}\right] = \frac{1}{1-q}\left\{\int p(x)^q \, dx - 1\right\}$$

Computer vision:

$$s(x, y) \geq 0$$

Divergence
Clustering
Center
Retrieval

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

3. Conformal change of divergence

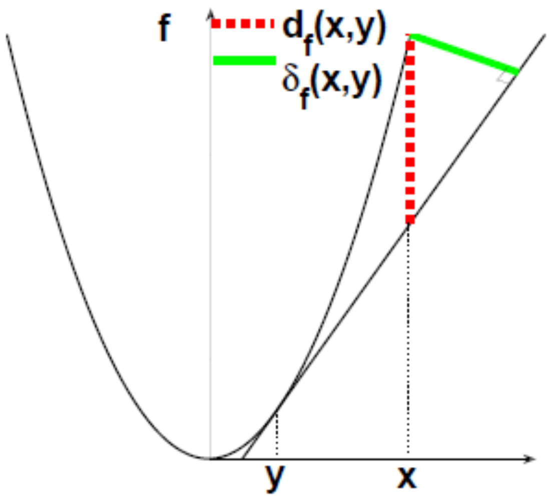
$$\tilde{D}(p:q) = \sigma(p)\sigma(q)D[p:q]$$

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

$$\tilde{\Gamma}_{ijk} = \sigma_1\sigma_2\Gamma_{ijk} + \sigma_2\partial_i\sigma_1 g_{ijk} + \sigma_2\partial_j\sigma_1 g_{ik} - \sigma_1\partial_k\sigma_2 g_{ij}$$

Conformally flat → canonical divergence

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.

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

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

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

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

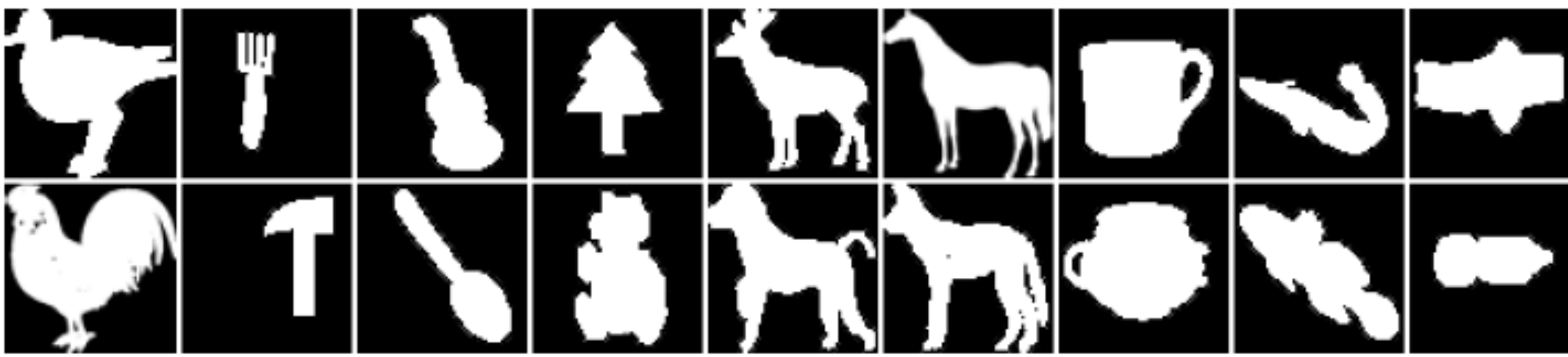
$|\boldsymbol{z}| < c \text{ as } |\boldsymbol{y}| \to \infty$: robust

TBD application-shape retrieval

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

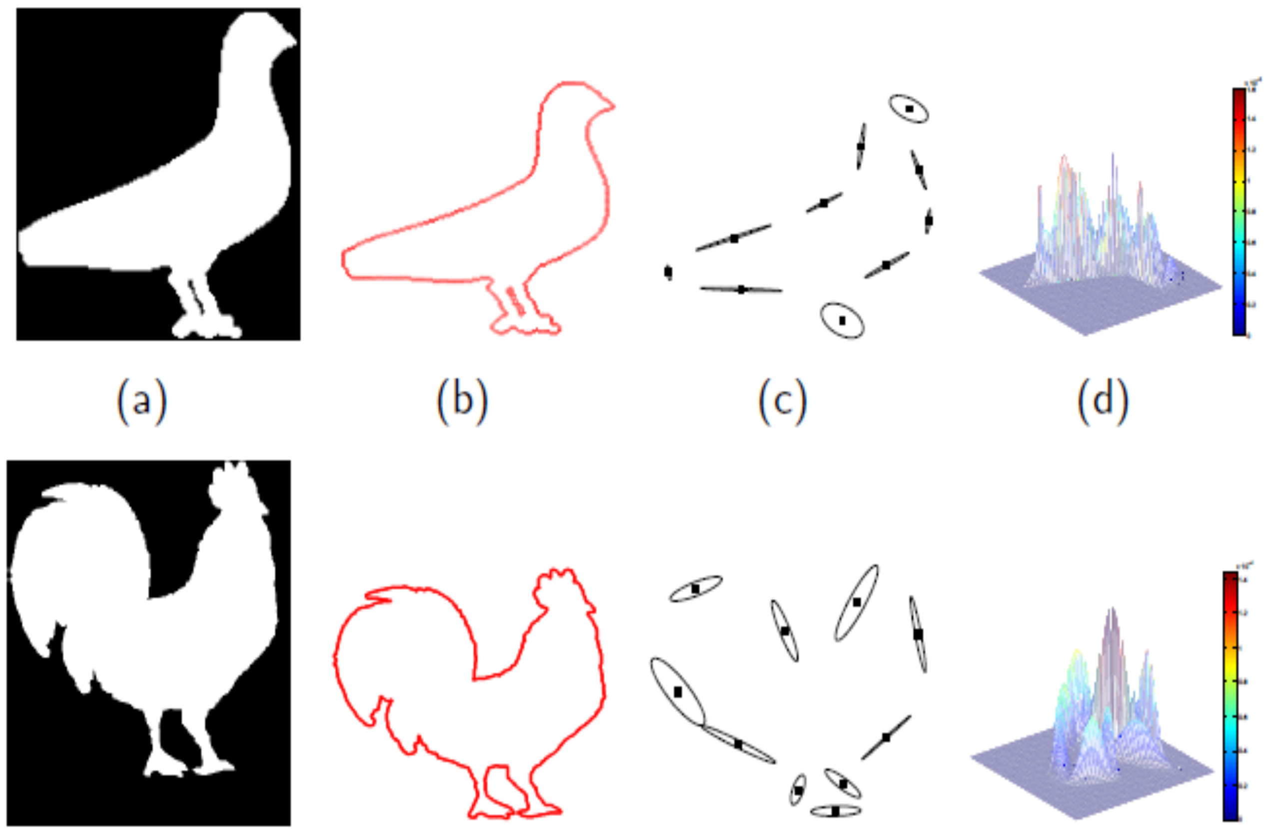
MPEG7 database

- Great intraclass variability, and small interclass dissimilarity.



Shape representation

A shape is represented using a mixture of Gaussians from the aligned boundary points.



(a) (b) (c) (d)

(e) (f) (g) (h)

Experimental results

Technique	Recognition rate (%)
Mixture of Gaussians + tSL	**89.1**
Mixture of Gaussians + χ^2	63.3
Mixture of Gaussians + SL	56.7
Shape-tree[6]	87.7
IDSC + DP + EMD[14]	86.56
Hierarchical Procrustes [15]	86.35
IDSC + DP [13]	85.4
Shape L'Âne Rouge[18]	85.25
Generative Models [21]	80.03
Curve Edit [19]	78.14
SC + TPS [3] [3]	76.51
Visual Parts [10]	76.45
CSS [16]	75.44

Table 3. Recognition rates comparison.

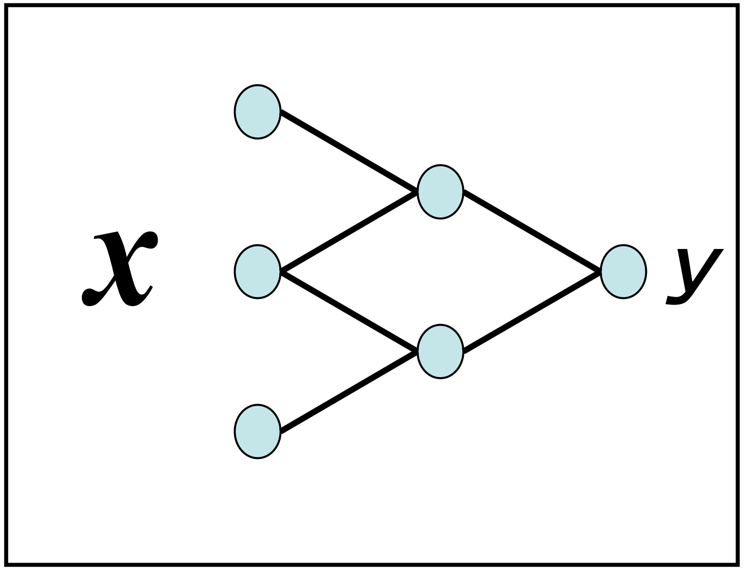
Neural Networks

Multilayer Perceptron

Higher-order correlations

Synchronous firing

Multilayer Perceptrons



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

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

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

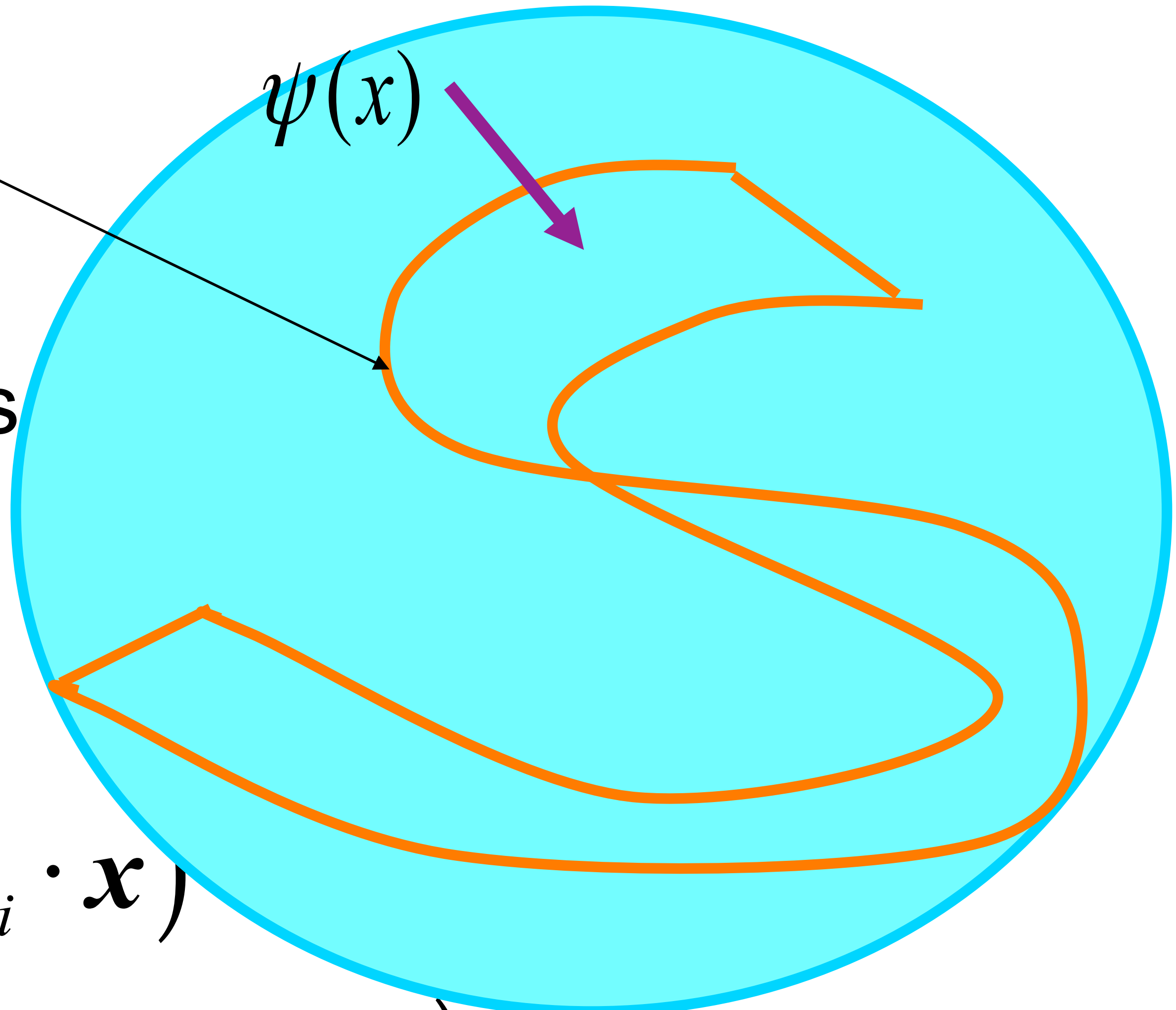
$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x})$$

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

Multilayer Perceptron

neuromanifold

space of functions

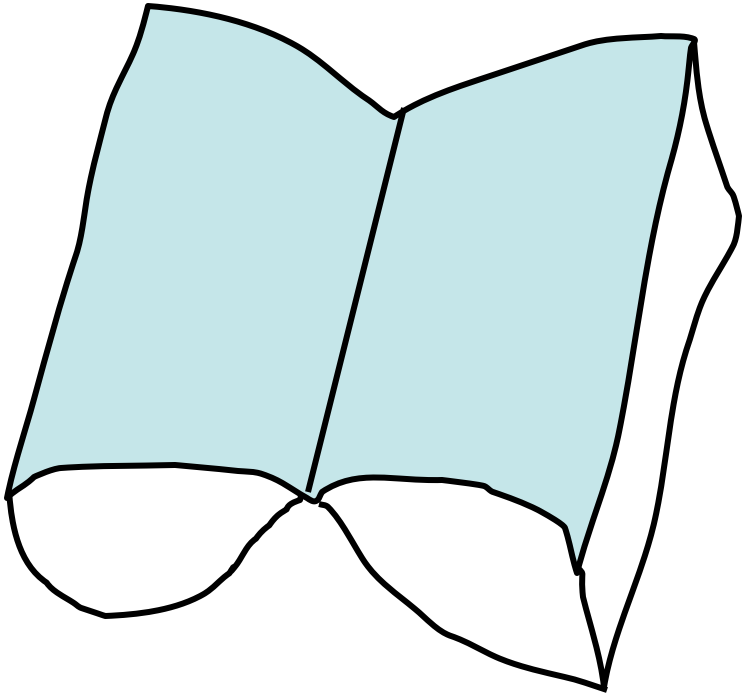


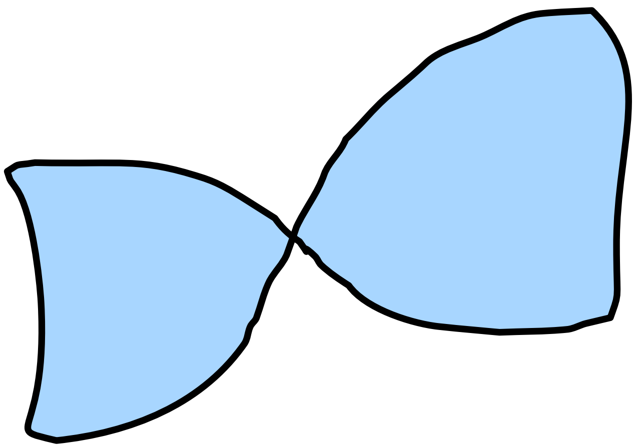
$$y = f\left(\boldsymbol{x}, \boldsymbol{\theta}\right)$$

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

$$\boldsymbol{\theta} = \left(v_1 \ldots, v_m ; \boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\right)$$

singularities

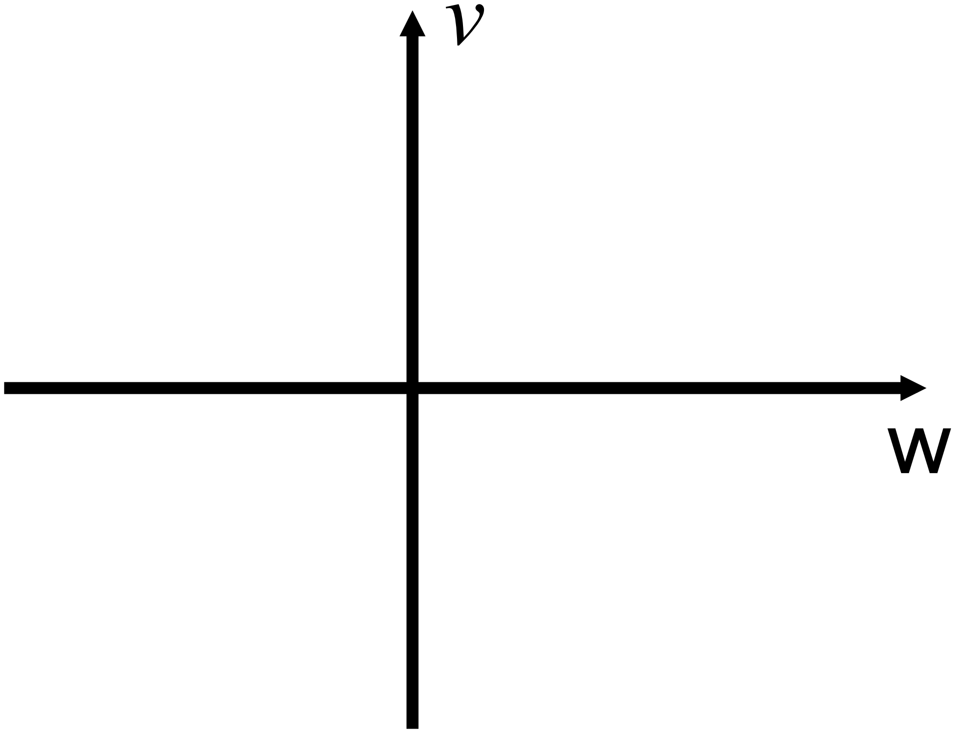


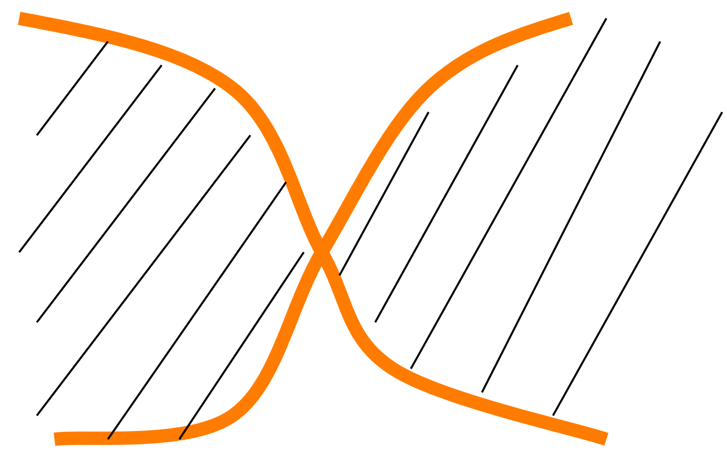


Geometry of singular model

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

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





Backpropagation ---gradient learning

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

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

natural gradient (Riemannian)

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

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

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

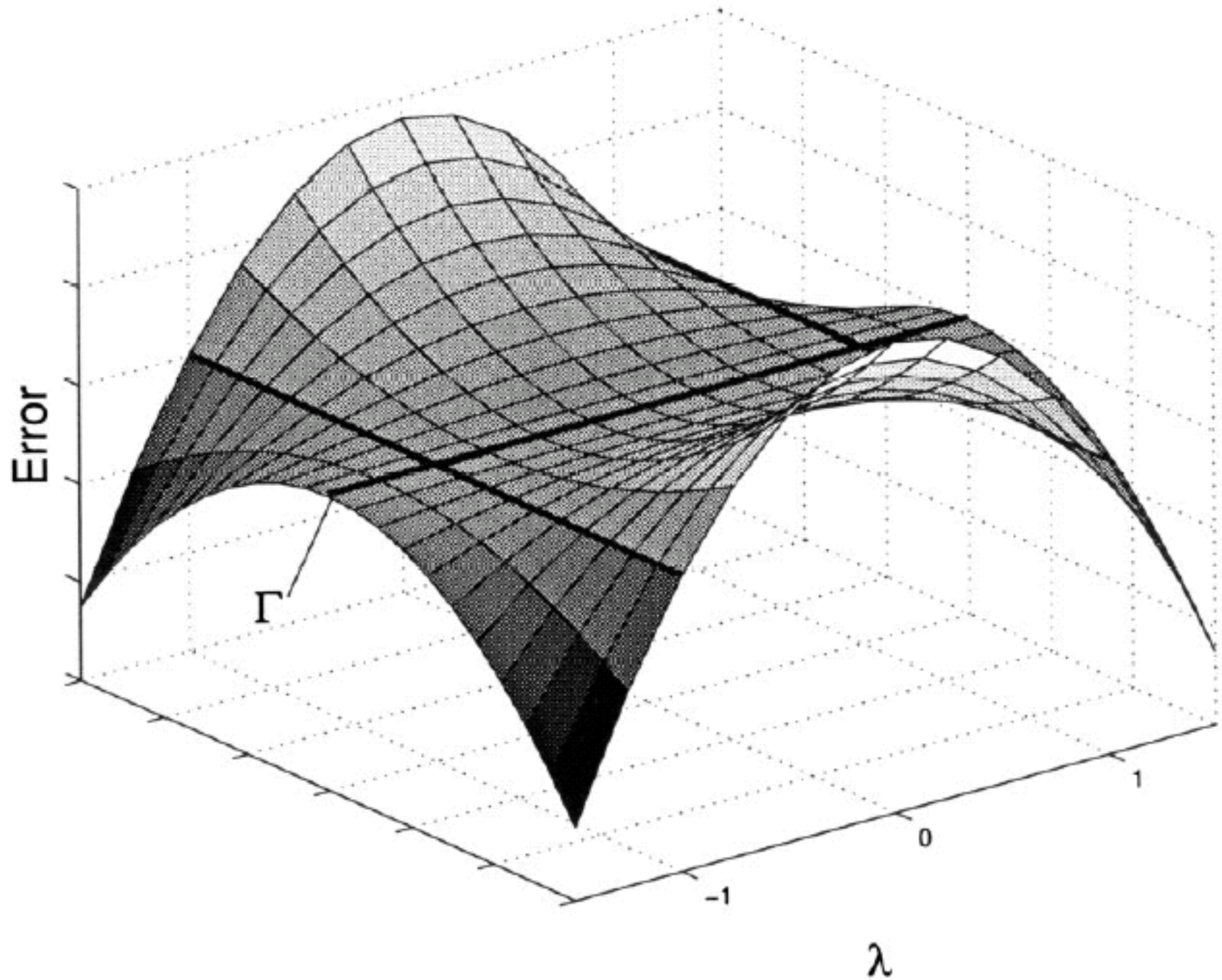
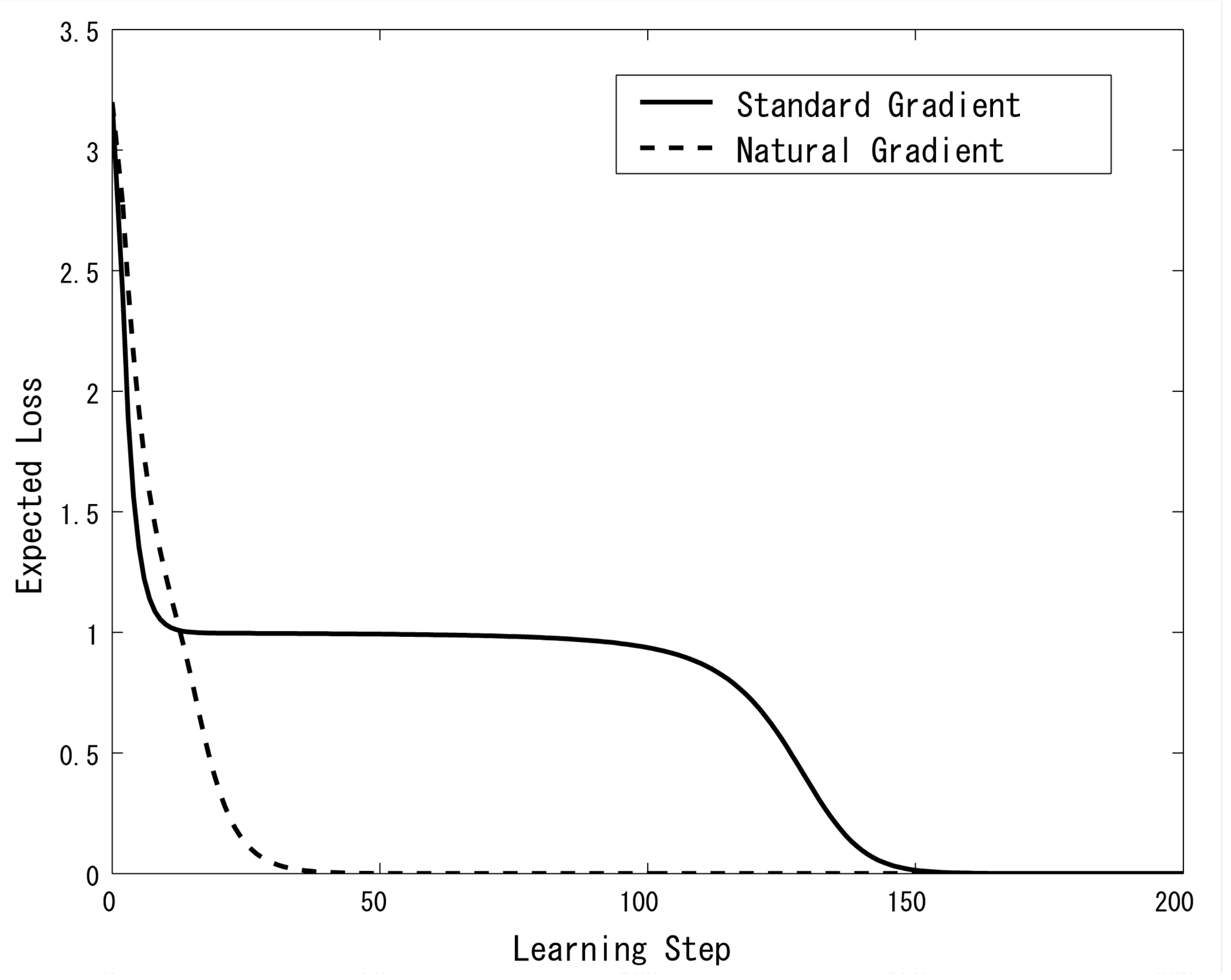


Fig. 5. Critical set with local minima and plateaus.



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

Neural Firing

$x_1 \quad x_2 \quad x_3 \quad \cdots\cdots \quad x_n$

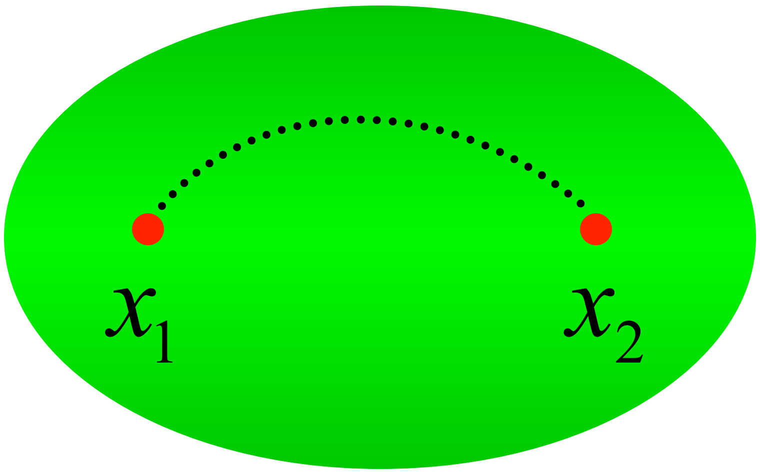
η_i : firing rate

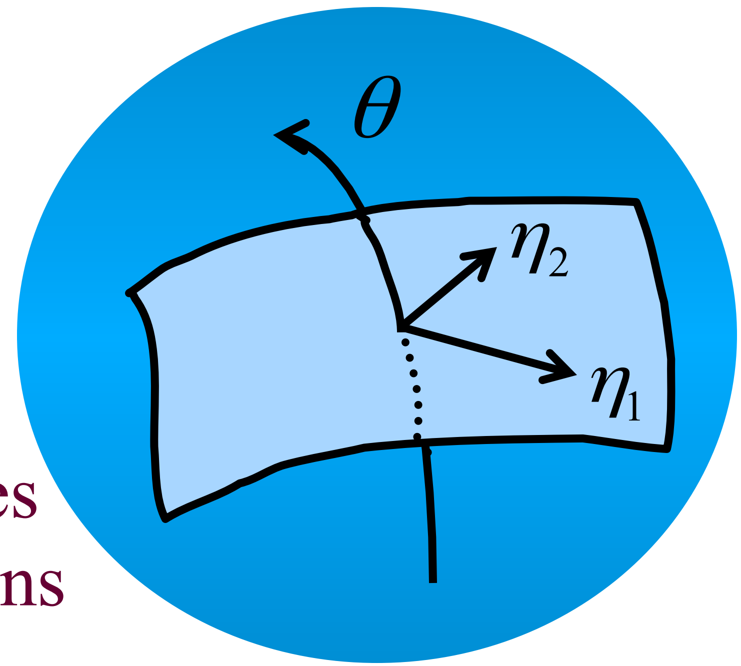
θ_{ij} : correlations

$p(x_1, x_2)$

orthogonal decomposition
higher-order correlations

Correlations of Neural Firing





$\{p(x_1, x_2)\}$

$\{p_{00}, p_{10}, p_{01}, p_{11}\}$

firing rates
correlations

$\eta_1 = p_{1\cdot}$

$\eta_2 = p_{\cdot 1}$

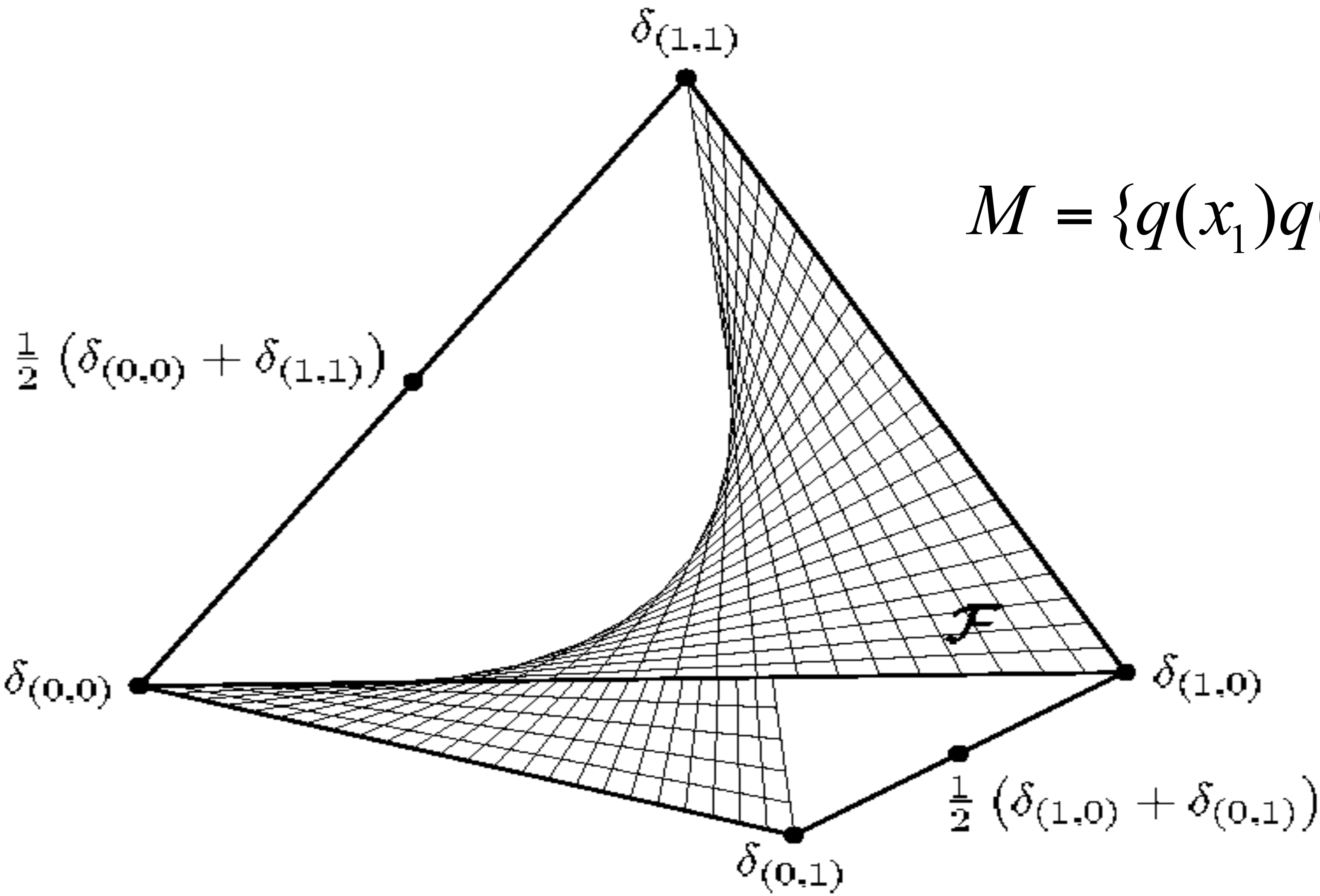
$$\theta = \log \frac{p_{11} p_{00}}{p_{10} p_{01}}$$

$\{(\eta_1, \eta_2), \theta\}$

orthogonal coordinates

Independent Distributions

$$M = \{q(x_1)q(x_2)\}$$



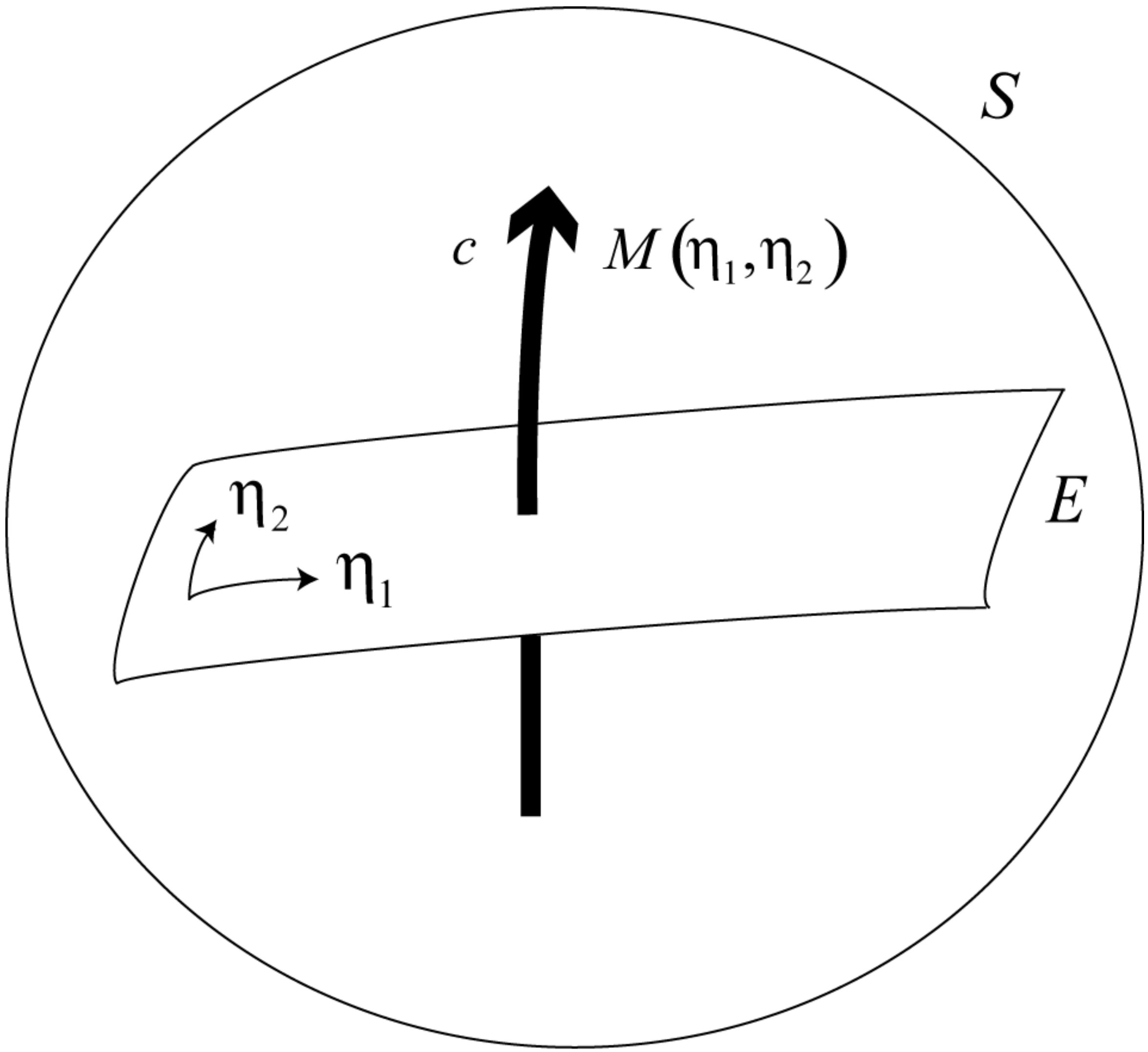
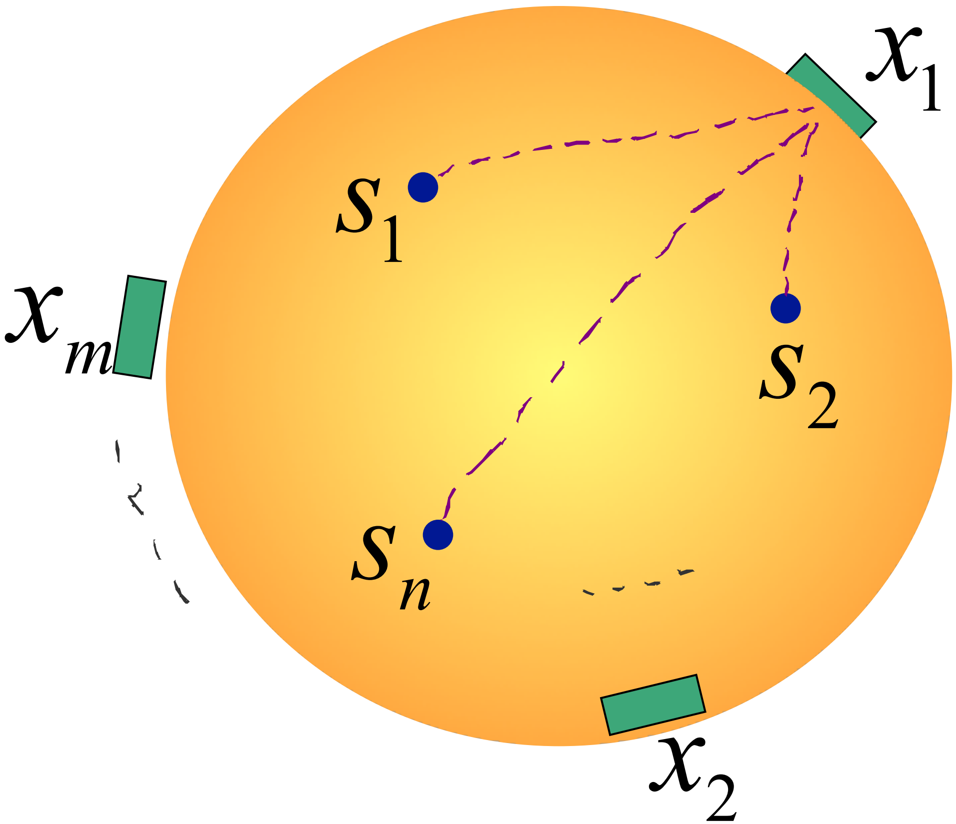


Fig. 1a

mixture and unmixture of independent signals



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

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

Signal Processing

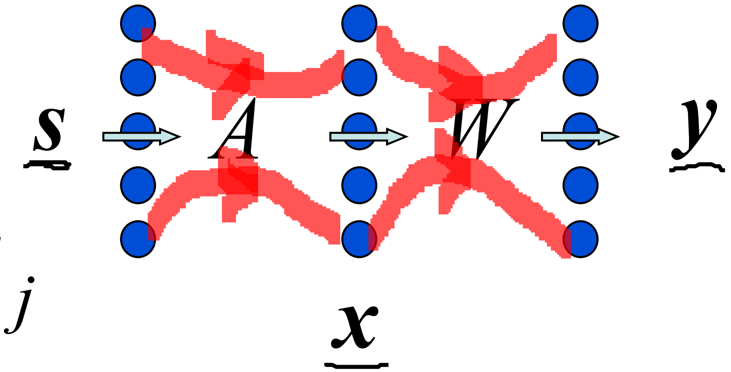
ICA : Independent Component Analysis

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

sparse component analysis

positive matrix factorization

Independent Component Analysis



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

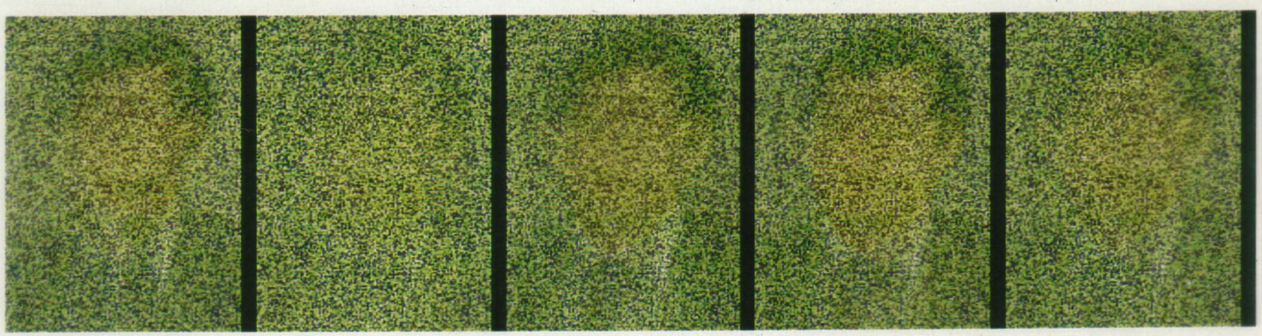
$$\boldsymbol{y} = W\boldsymbol{x} \qquad W = A^{-1}$$

observations: x(1), x(2), …, x(t)

recover: s(1), s(2), …, s(t)

Example of color image separation :

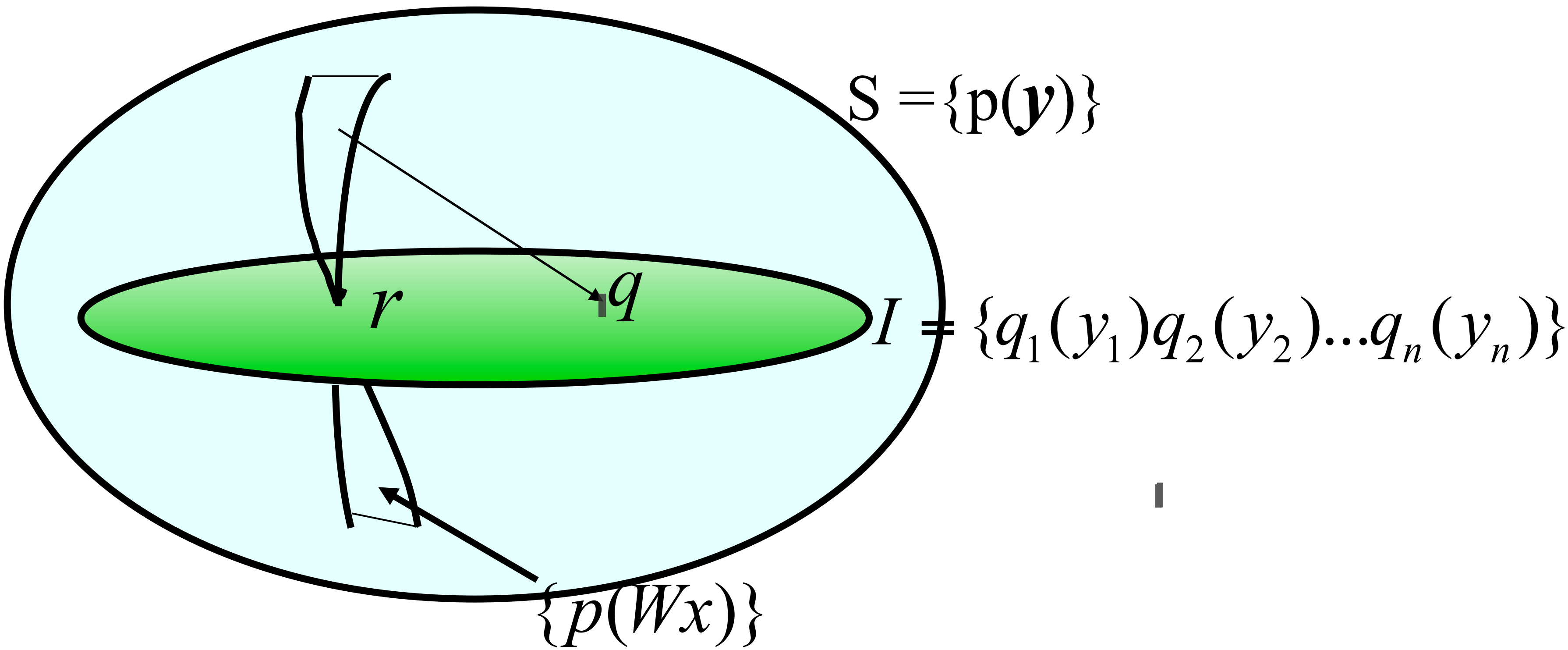
Five original images (but unknown to the neural net)



Five mixed images for separation

Final (stable states) of five separated images

Information Geometry of ICA

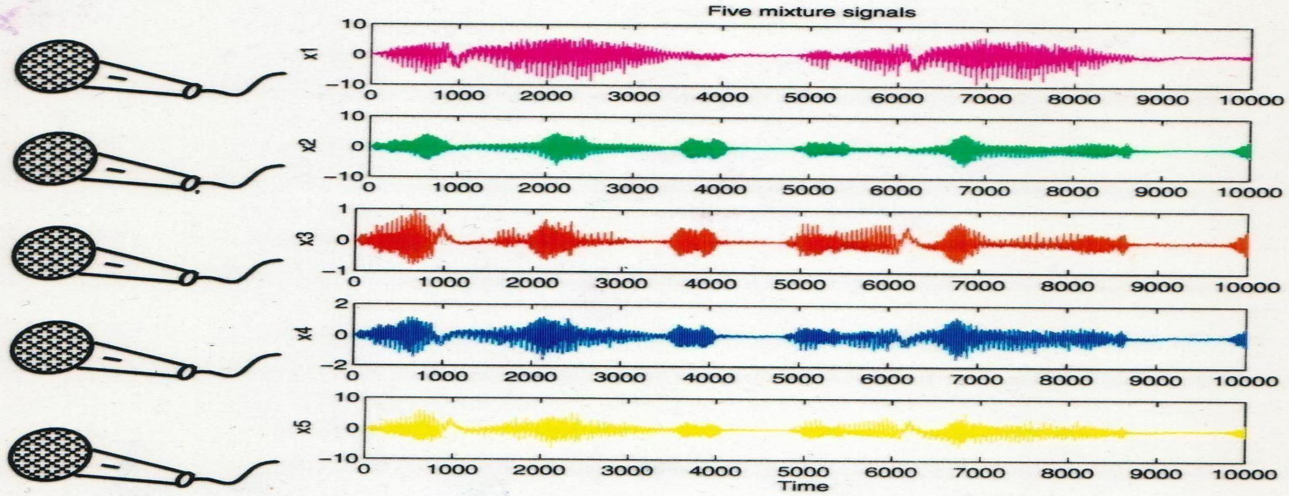


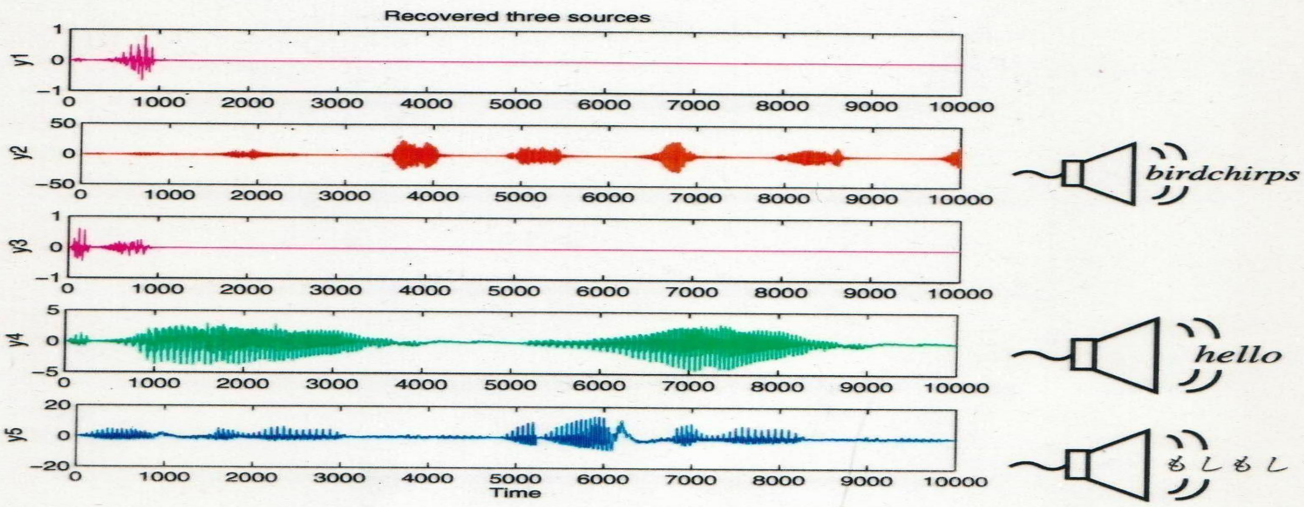
natural gradient
estimating function
stability, efficiency

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

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

Cocktail party experiment





- **5 microphones (sensors) and only 3 speakers**

Natural Gradient

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

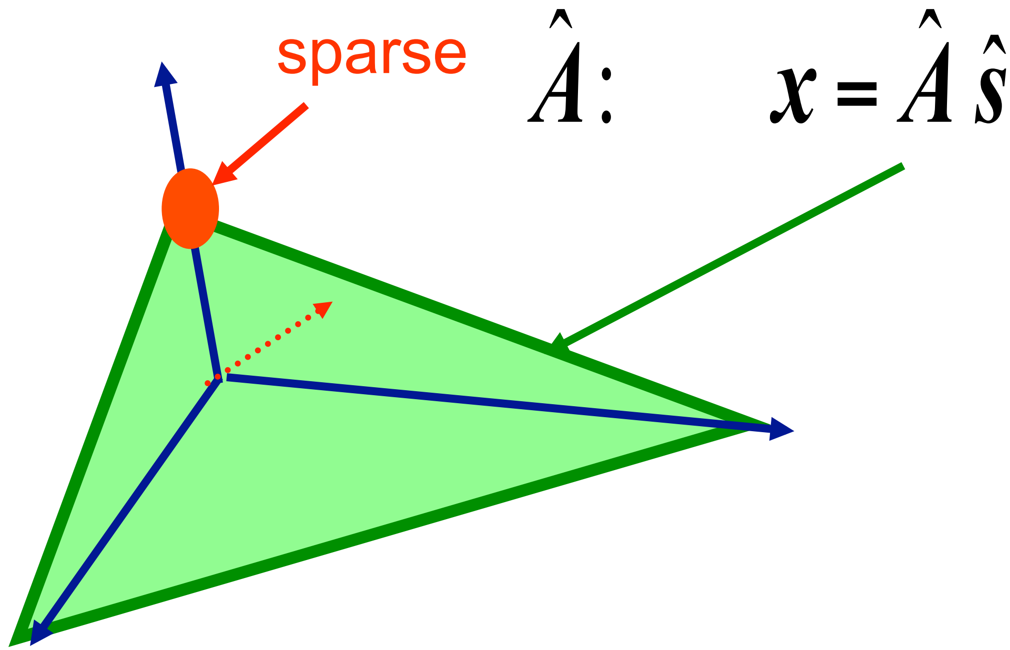
Basis Given: overcomplete case
Sparse Solution

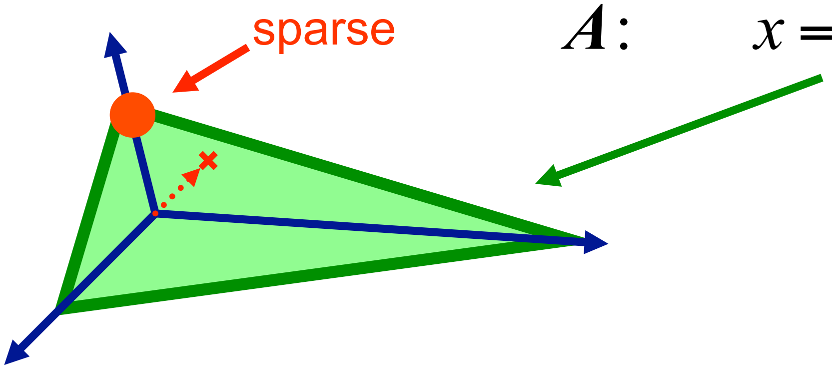
$$\boldsymbol{x} = A\boldsymbol{s} = \sum s_i \boldsymbol{a}_i$$

many solutions

many $s_i \rightarrow 0$

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





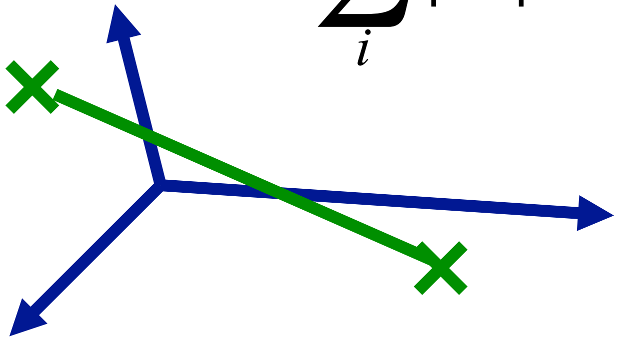
$\hat{A}: \quad x = \hat{A}\hat{s}$

generalized inverse

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

sparse solution

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



Overcomplete Basis and Sparse Solution

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

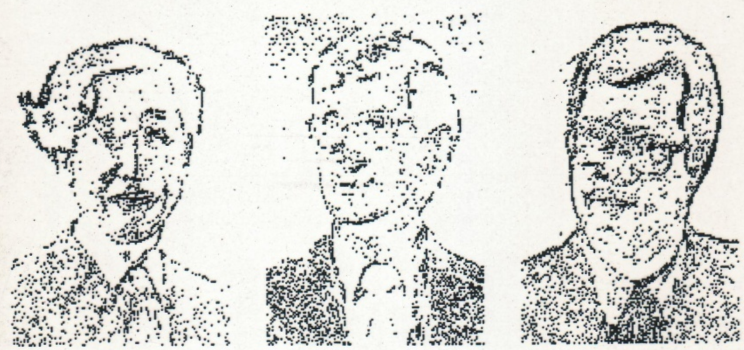
$$\min |\boldsymbol{s}|_1 = \sum |s_i|$$

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

non-linear denoising

(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).

Linear Systems

ARMA

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

$$\theta = \left(a_1, \mathrm{L} \ , a_p : b_1, \mathrm{L} \ , b_q \right)$$

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

AR---e-flat

MA---m-flat

Information Geometry of Belief Propagation

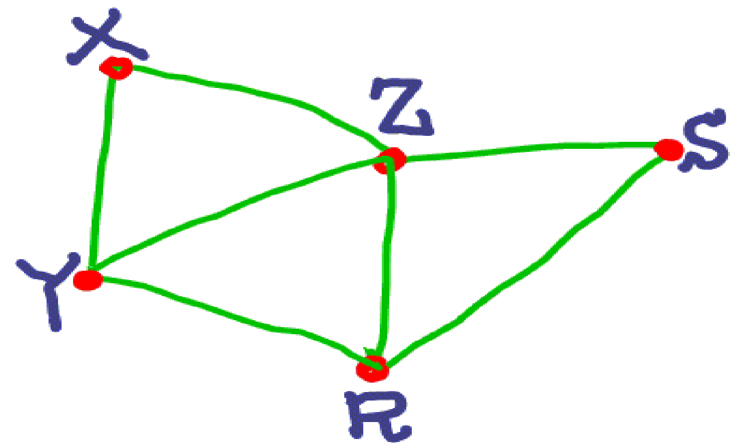
- Shun-ichi Amari (RIKEN BSI)
- Shiro Ikeda (Inst. Statist. Math.)
- Toshiyuki Tanaka (Tokyo Metropolitan U.)

Stochastic Reasoning

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

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

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



Mean Value

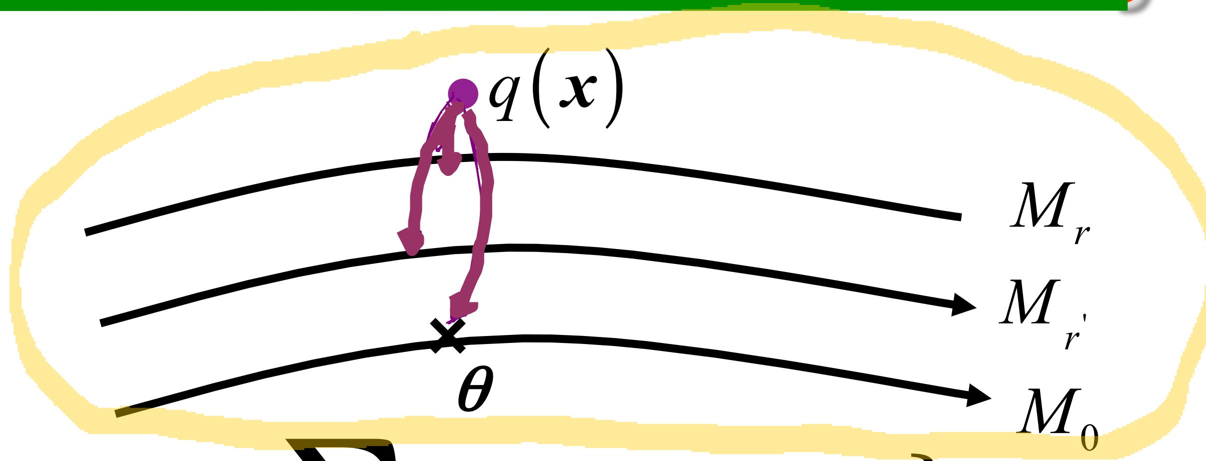
Marginlization

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

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

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

Information Geometry



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

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

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

$$r = 1, \ldots, L$$

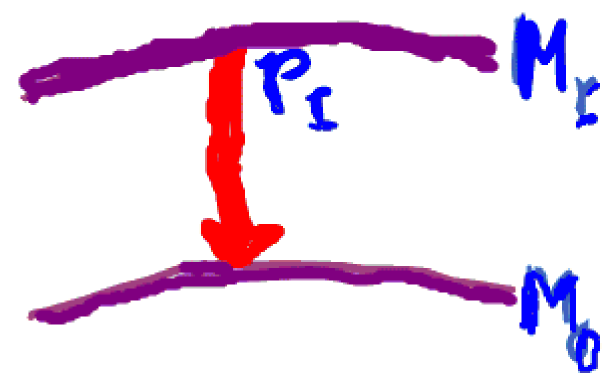
Belief Propagation

$$\Pi_0 p_r(x, \zeta_r) \qquad\qquad p(x, \varsigma_r) = \exp\{c_r(x) + \varsigma_r \cdot x - \psi_r\}$$

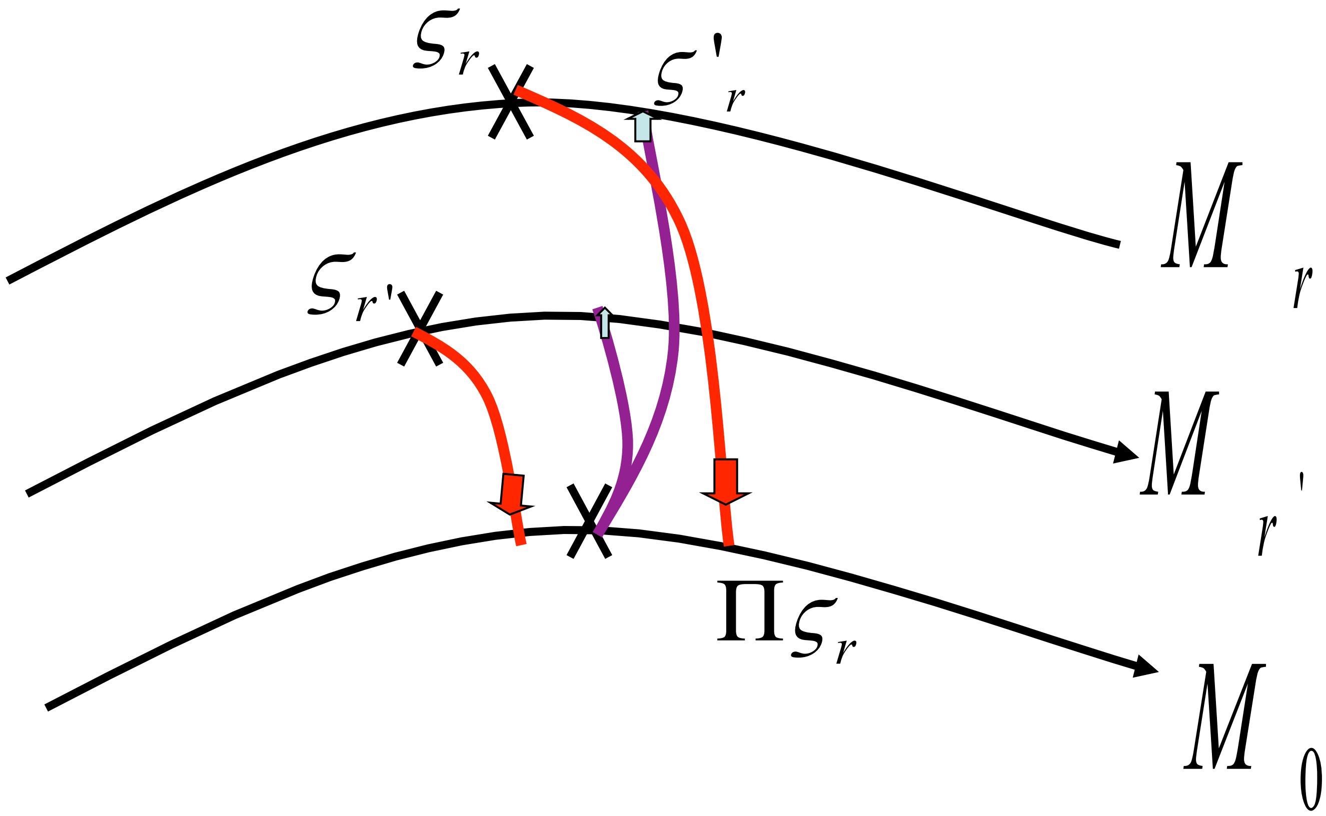
$$\xi_r^{t+1} = \Pi_0 p_r(x, \zeta_r^{t}) - \zeta_r^{\mathrm{t}} \quad : \quad \text{belief for } c_r(x)$$

$$\zeta_r^{t+1} = \sum_{r \neq r} \Pi_0 \left\{ p\left(\xi_{r'}^{t}\right) - \xi_{r'}^{t} \right\} = \sum \xi_{r'}^{t+1}$$

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



Belief Prop Algorithm



Belief Propagation

e-condition OK

$$F(\theta; \zeta_1, \zeta_2, \ldots, \zeta_L), \qquad \theta = \theta(\zeta_1, \ldots, \zeta_L)$$

$$(\zeta_1, \ldots, \zeta_L) \to (\zeta_1', \ldots, \zeta_L')$$

CCCP

m-condition OK

$$\zeta_1(\theta), \zeta_2(\theta), \zeta_L(\theta)$$

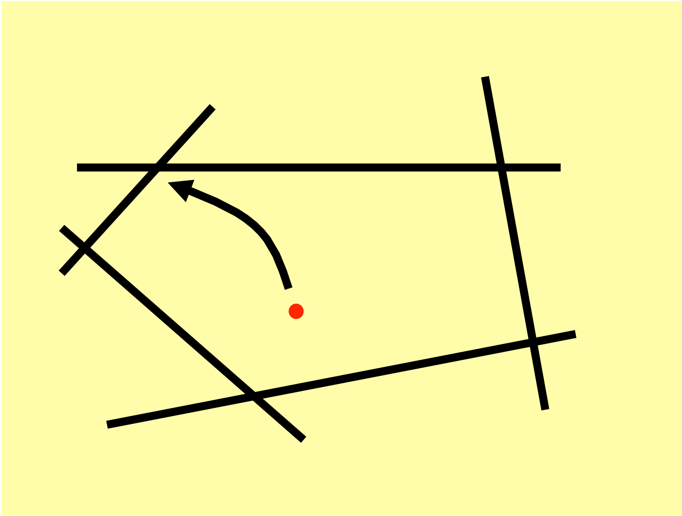
$$\theta \to \theta'$$

Linear Programming

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

$$\max \sum c_i x_i$$

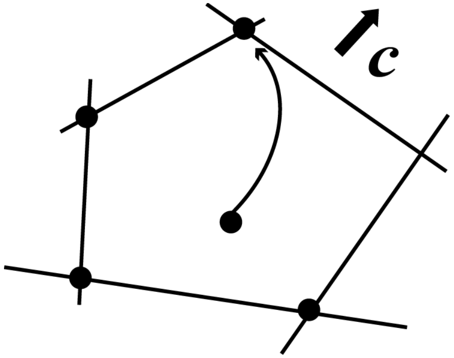
$$\psi(\boldsymbol{x}) = \sum_i \log\left(\sum A_{ij} x_j - b_i\right)$$



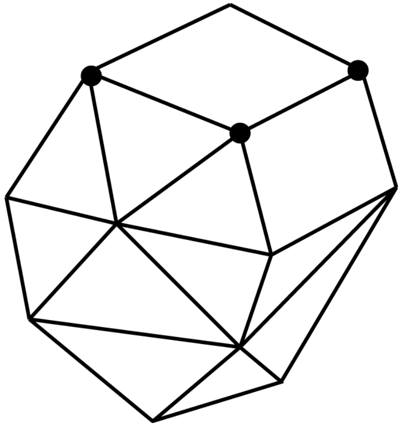
Convex Programming
— Inner Method

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

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



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

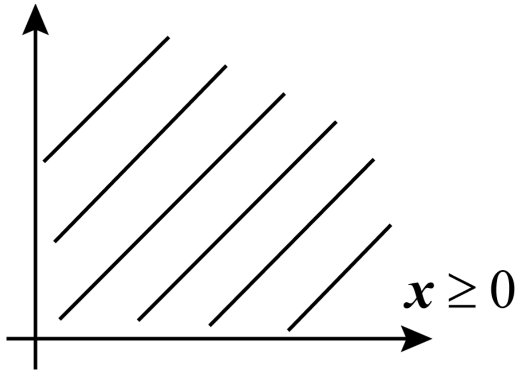


$$\boldsymbol{\eta} = \partial_i \psi(\boldsymbol{x})$$

Simplex method ; inner method

Convex Cone Programming

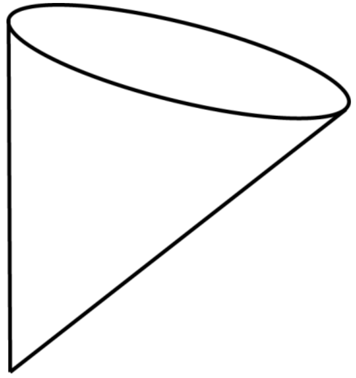
P : positive semi-definite matrix



convex potential function

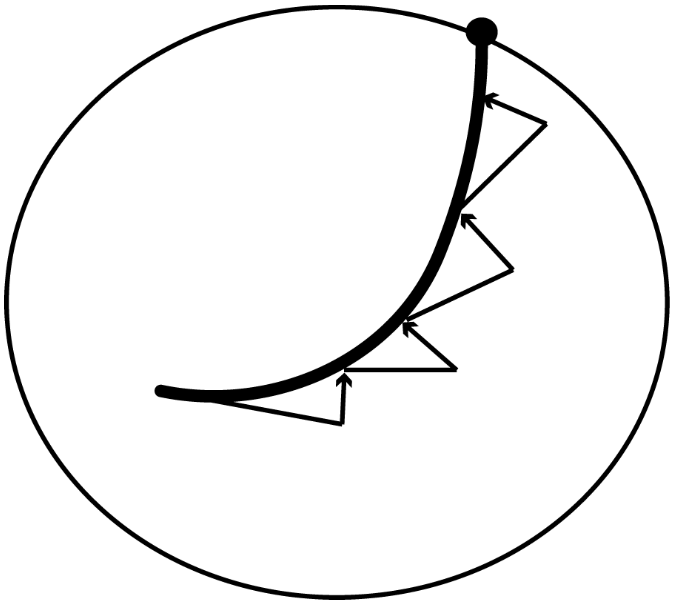
dual geodesic approach

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



Support vector machine

Polynomial-Time Algorithm



curvature : step-size

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

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

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

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

$$\varepsilon_t : \text{ Prob }\left\{h_t(\boldsymbol{x}_i) \neq y_i\right\} \quad | W_t$$

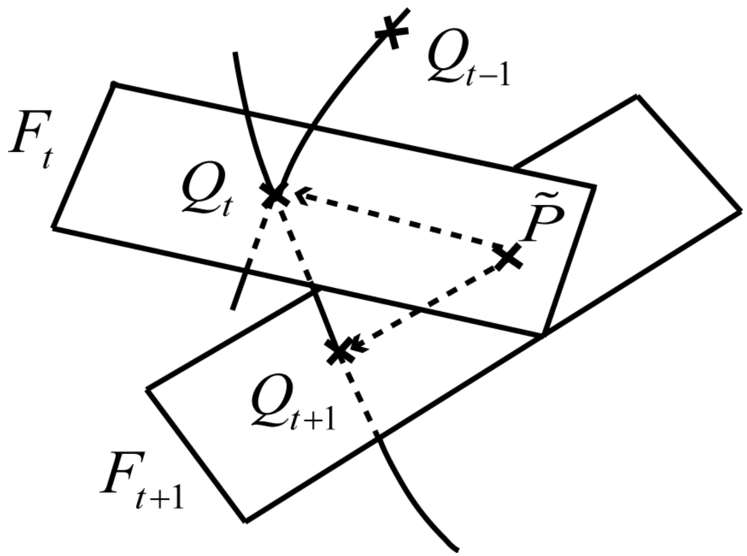
$$W_{t+1}(i) = cW_t(i)\exp\left\{-\alpha_t y_i h_t(x_i)\right\}$$

weight distribution

Boosting —generalization

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

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



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

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

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

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

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

KL-divergence, α -divergence

$$D_\alpha\left[p(x):q(x)\right]=\frac{4}{1-\alpha^2}\left\{1-\int p(x)^{\frac{1-\alpha}{2}}q(x)^{\frac{1+\alpha}{2}}dx\right\}$$

$$\alpha\to -1: D_{-1}\left[p:q\right]=KL\left[p:q\right]=\int p(x)\log\frac{p(x)}{q(x)}dx$$

$$\alpha\to +1: D_1\left[p:q\right]=KL\left[q:p\right]$$

α -representation

$$f_\alpha(p)=\frac{2}{1-\alpha}\left\{p(x)\right\}^{\frac{2}{1-\alpha}}$$

α -family : $p(x,\theta)^{\frac{1-\alpha}{2}}=c\sum\theta_i q_i(x)^{\frac{1-\alpha}{2}}$

- exponential family $(\alpha=-1)$
- mixture family $(\alpha=1)$

dually flat

Csiszar *f*-divergence, *U*-divergence

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

$$D_U\left[p(x):q(x)\right]$$

α -divergence from convex function

$$D_\alpha(P,P') = \frac{4}{1-\alpha^2}[1-\exp\{\psi(\frac{1-\alpha}{2}\theta+\frac{1+\alpha}{2}\theta') + \psi(\frac{1-\alpha}{2}\theta)+\psi(\frac{1+\alpha}{2}\theta')\}]$$

α -geodesic, duality, projection

Integration of evidences:

$$x_1, x_2, \ldots x_m$$

arithmetic mean
geometric mean
harmonic mean
α -mean