IGAIA 4　Bohemia

Information Geometry

–– Historical Episodes and Future with Recent Developments

Shun-ichi Amari
RIKEN Brain Science Institute

Prehistory --- Riemannian Geometry

H. Hotteling 1929

Riemannian metric and Fisher information

location-scale model : constant curvature

P. Ch. Maharanobis 1936 Euclidean distance (multivariate-Gaussian)

C. R. Rao 1945 Cramer-Rao Theorem; Riemannian

H. Jeffreys 1946 Bayesian theory and Jeffreys invariant prior

Dual Geometry, Invariance

N. Chentsov 1972 invariance, $\{g, T\}$, α-connection

B. Efron 1975 (A. P. Dawid) statistical curvature; higher-order asymptotics

O. Barndorff-Nielsen 1976 exponential family; Legendre transform

S. Amari 1982 duality; curvature and statistics (M. Kumon)

H. Nagaoka and S. Amari 1982 duality, Pythagorean theorem

Amari's personal history

1958: statistics seminar (master course at U Tokyo)

S. Kullback, "Information and Statistics"

Riemannian metric (suggested by S. Moriguti)

Gausssian $N(\mu,\sigma^2)$: geodesic, constant curvature (Poincare-half plane)

beautiful structure → essential meaning?

mathematical engineering

graph and topology of networks: homology

non-Riemannian geometry of materials manifold: dislocations

information systems, learning and neural networks

Statistical curvature and higher-order inference

B. Efron, 1975

Fisher's idea; exponential connection and mixture connection

A. P. Dawid, 1975 e- and m-connections

S. Amari : α-geometry $Error^2 = \frac{1}{n} G^{-1} + \frac{1}{n^2}(H_e{}^2 + H_m{}^2 + \Gamma_m{}^2) + \frac{1}{n^3} K$

(Rao, Kano K? Fisher's dream)

Amari and M.Kumon higher-order power of statistical test

Amari paper : Ann. Statist. 1982:

Reviewers (S. Lauritzen and A.P. Dawid)

Chentsov work (handwritten manuscript)

H. Nagaoka and S. Amari 1982 (Technical Report)

Ann. Probability Theory 7 reviewers

Zeitshrift fur Wahrsceinlichkeitstheorie und VerwandteGebiete

geometry has nothing to do with statistics

IEEE Trans. Inf. Theory Shannon Theory, now well-known

London Workshop: 1984 (D. Cox)

Cox visited Japan in 1983

patron of information geometry

Rao, Efron, Dawid, Barndorff-Nielsen, Lauritzen
Kass, Eguchi many others

Dodson, Critchley, Marriot, Komaki, Zhang, Ay, Pistone, Giblisco, Nielsen, …

Information Geometry --- lucky naming

Applications area:

statistics, time-series and systems,
machine learning, signal processing, optimization
brain theory, consciousness
physics, economics, mathematics
(Banach manifold, affine differential geometry and beyond)
quantum information, Tsallis entropy

International Conferences

IGAIA series; GSIS series, …

Many monographs

new journal   (Jun Zhang); where to publish

mailing list and society

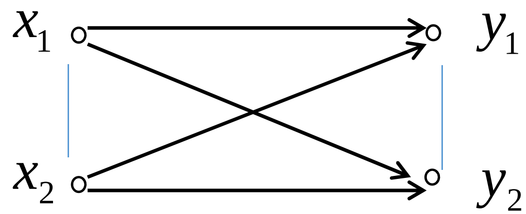
still small community; united and cooperative, blessed by all

My recent works

1. Systems complexity and consciousness (IIT)
2. Geometry of score matching (Hyvarinen score)
3. Natural gradient descent and topology of deep learning)
4. Canonical divergence
5. Multi-terminal statistical inference
6. Information geometry and Wasserstein distance

Information Integration and Complexity of Systems

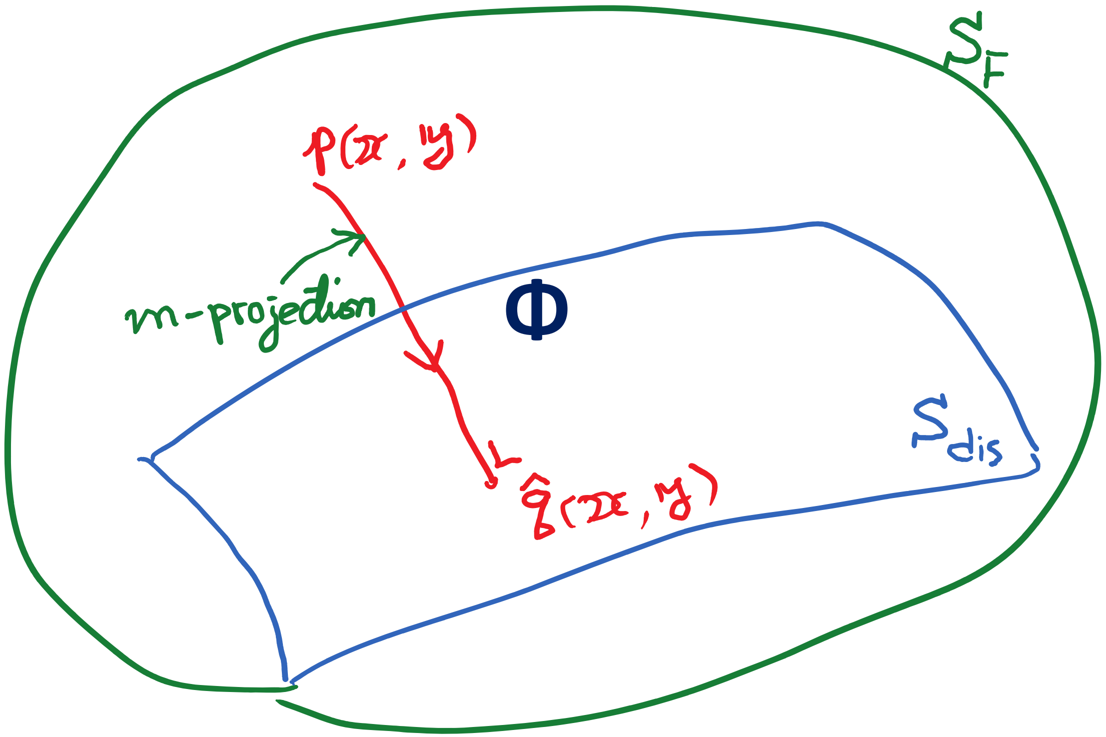
-- Stochastic approach



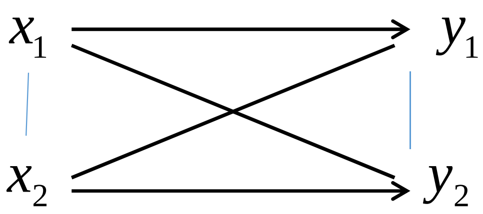
$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})\, p(\mathbf{y} \mid \mathbf{x})$$

x: state of the brain **y: next state of the brain**

Integrated Information Theory G. Tononi



Necessary condition; sufficient?



full model: $S_F = \{ p(\mathbf{x}, \mathbf{y}) \}$ **Disconnected model:**

$$S_{dis} = \{ q(\mathbf{x}, \mathbf{y}) \} \quad q(\mathbf{y} \mid \mathbf{x}) = \Pi q(y_i \mid x_i)$$

measure of interaction : N. Ay

information integration : Tononi

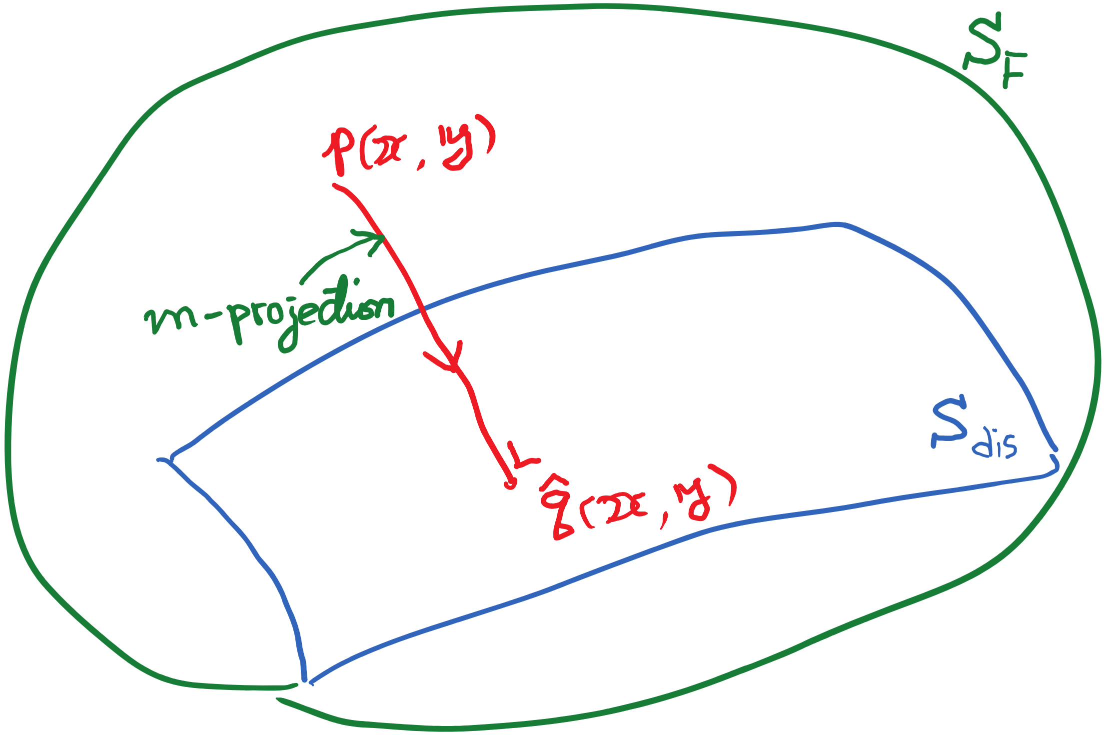
Barrett and Seth

Many other Φ

Measure of information integration, or system complexity Φ

Information Geometry N. Ay

$$\Phi = D_{KL}[p : \hat{q}]$$



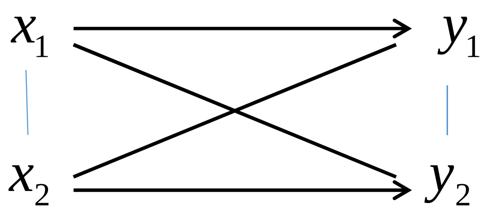
Definition of Φ : Postulates

1) $$\Phi = \min_{q} D[p(x,y) : q(x,y)], \quad q \in S_{dis}$$

2) $$D = D_{KL}[p:q] = \int p(x,y) \log \frac{p(x,y)}{q(x,y)}$$

3) **Disconnected model: S_{dis} Markov conditions**

Markov Condition



(1→2) branch deleted: Markov condition: $x_1 \to x_2 \to y_2$

$$p(x_1, y_2 \mid x_2) = p(x_1 \mid x_2)\, p(y_2 \mid x_2)$$

$$X_1 - X_2 - Y_2$$

$$X_2 - X_1 - Y_1$$

$$S_{dis}: \quad \text{all } x_i \to y_j \quad (i \neq j) \quad \text{deleted}$$

Why KL-divergence?

1) $\mathbf{D[p:q] \geq 0, \;\; = 0, \text{ when and only when } p = q}$

2) $D[p:q]$ **invariant under transformations of** $\mathbf{X}$

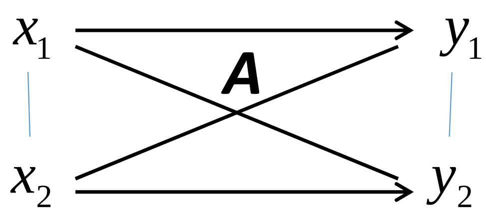
3) $\mathbf{D[p:q] = \sum_i d[p(x_i), q(x_i)]}$

4) $\mathrm{D}[p:q]$ **induces flat structure dually**

Geometric degree of information integration

$$\Phi_{geo} = \min_{q} D_{KL}[\,p(x,y) : q(x,y)], \quad q \in S_{dis}$$

Gaussian case

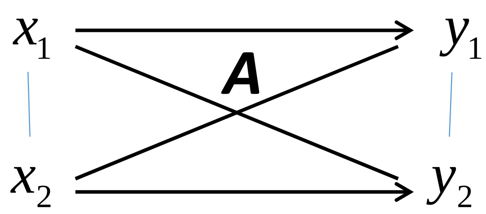


$$\mathbf{y} = A\mathbf{x} + \mathbf{e} \qquad \Sigma = \mathrm{E}[\mathbf{e}\mathbf{e}^T]$$

$$\mathbf{y} = A'\mathbf{x} + \mathbf{e}' \qquad \Sigma' = \mathrm{E}[\mathbf{e}'\mathbf{e}'^T] \qquad A' : \text{diagonal}$$

$$\Phi_{geo} = \log \frac{|\Sigma'|}{|\Sigma|}$$

Gaussian case



$$\mathbf{y} = A\mathbf{x} + \mathbf{e} \qquad \Sigma = \mathrm{E}[\mathbf{e}\mathbf{e}^T]$$

$$\mathbf{y} = A'\mathbf{x} + \mathbf{e}' \qquad \Sigma' = \mathrm{E}[\mathbf{e}'\mathbf{e}'^T] \qquad A' : \text{diagonal}$$

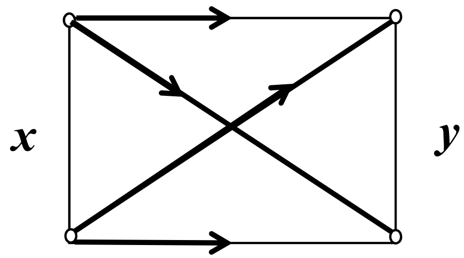
$$\Phi_{geo} = \log \frac{|\Sigma'|}{|\Sigma|}$$

Many other definitions of Φ

Full model

Disconnected models

Full model S_F **: graphical model**



$$p(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{\sum \theta_i^X x_i + \sum \theta_i^Y y_i + \theta_{12}^X x_1 x_2 + \theta_{12}^Y y_1 y_2 + \sum \theta_{ij}^{XY} x_i y_j \right.$$
$$\left. + \text{ higher-order terms } - \psi(\boldsymbol{\theta})\right\}$$

exponential family

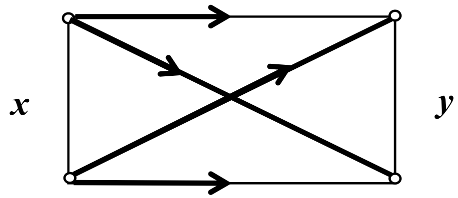
θ **-coordinates** $\quad \theta$

η **-coordinates** $\quad \eta_i^X = E[x_i], \cdots, \eta_{ij}^{XY} = E[x_i y_j], \cdots$

There are many disconnected models!!

Split Model S_H : Ay, Barrett & Seth

$$q(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{x}) \prod q(y_i | x_i)$$

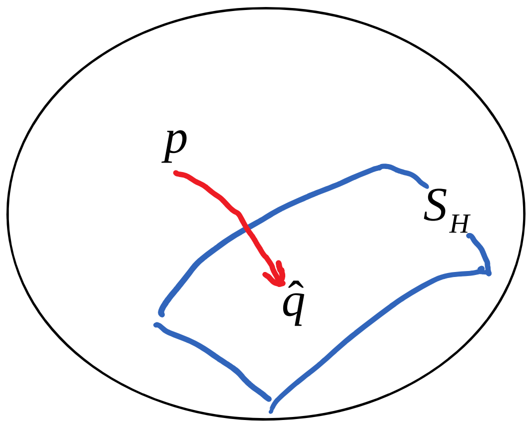


$$\theta_{12}^{XY} = \theta_{21}^{XY} = \theta_{12}^{Y} = 0$$

$Y_1 - X_1 - X_2 - Y_2$

$$\Phi_H = D_{KL}[p : S_H] = \min_{q \in S_H} D_{KL}[p : q]$$

$$\hat{q} = \prod_{M_S} p \quad : \quad \hat{q}(\boldsymbol{y}|\boldsymbol{x}) = \prod p(y_i | x_i)$$



$$\Phi_H = \sum H[Y_i | X_i] - H[\boldsymbol{Y} | \boldsymbol{X}]$$

$$S_H: \quad \theta_{12}^{XY} = \theta_{21}^{XY} = \theta_{12}^{Y} = 0; \qquad \hat{\boldsymbol{\eta}}_q = \boldsymbol{\eta}_q$$

Mixed Coordinates :

$$\xi = \left(\eta_i^X, \eta_{12}^X, \eta_i^Y, \eta_{11}^{XY}, \eta_{22}^{XY}; \theta_{12}^{XY}, \theta_{21}^{XY}, \theta_{12}^{Y}\right)$$

$$\hat{q} = \prod p = q\left(\boldsymbol{x}, \boldsymbol{y}; \hat{\xi}\right)$$

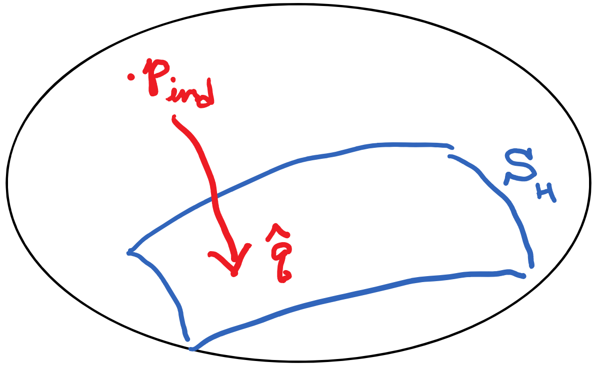
Markovian Condition

$$Y_1 - X_1 - X_2 - Y_2 \quad \Rightarrow$$

$$X_1 - X_2 - Y_2; \; Y_1 - X_1 - X_2$$

problem $0 \le \Phi[p] \le I[X:Y]$

$$p_{ind} = p_X(\boldsymbol{x})\, p_Y(\boldsymbol{y}) \Rightarrow I = 0, \quad \Phi > 0$$



Split Model S_{Gr}

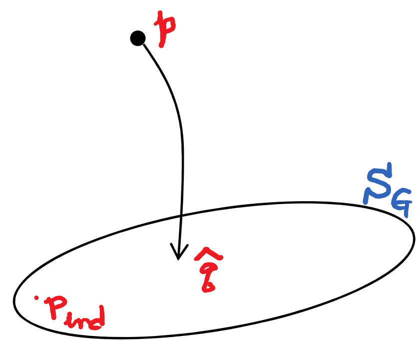
graphical model

$$q(\boldsymbol{x},\boldsymbol{y}) = q_X(\boldsymbol{x})\tilde{q}_Y(\boldsymbol{y})\prod q(y_i|x_i)$$

$$\theta_{12}^{XY} = \theta_{21}^{XY} = 0$$

$$q(x_1, y_2|x_2, y_1) = q(x_1|x_2, y_1)q(y_2|x_2, y_1)$$

$$0 \le \Phi \le I(X:Y)$$



$$\hat{q}_X(\boldsymbol{x}) = p_Y(\boldsymbol{x}), \qquad \hat{q}_Y(\boldsymbol{y}) = p_Y(\boldsymbol{y})$$

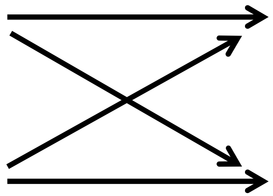
$$\hat{q}(y_i|x_i) = p(y_i|x_i)$$

Problem: Gaussian channel

$$p(\boldsymbol{x},\boldsymbol{y}):\ \ \boldsymbol{y}=A\boldsymbol{x}+\boldsymbol{\varepsilon}\ \Rightarrow\ \hat{q}:\ \ \boldsymbol{y}=\hat{A}\boldsymbol{x}+\boldsymbol{\varepsilon}'\in S_G$$

$$p(\boldsymbol{x},\boldsymbol{y})=\exp\left[-\frac{1}{2}\left\{\boldsymbol{x}'\sum\nolimits_X^{-1}\boldsymbol{x}+(\boldsymbol{y}-A\boldsymbol{x})'\sum\nolimits_\varepsilon^{-1}(\boldsymbol{y}-A\boldsymbol{x})-\psi\right\}\right]$$

$\hat{A}$ **is not diagonal** $\Leftarrow\ \theta_{12}^{XY}=\theta_{21}^{XY}=0$



A

:

Mismatched Decoding Model S_M

$$\mathbf{y} \rightarrow \hat{\mathbf{x}}$$

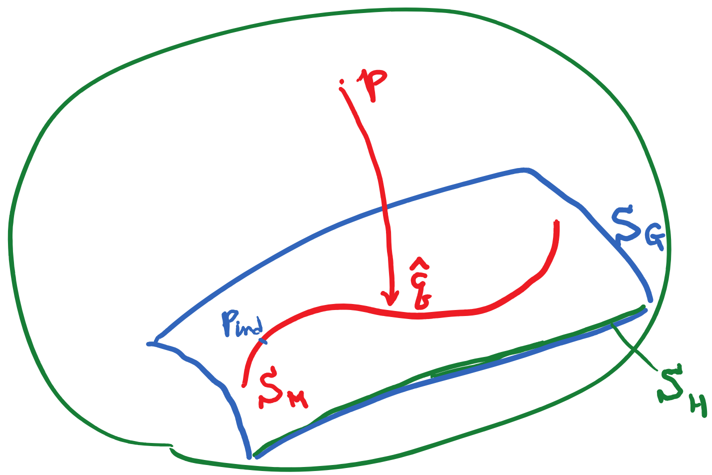
Best mismatched decoding from y to x

$$x_1 \longrightarrow y_1$$

$$x_2 \longrightarrow y_2$$

$$S_M = \{ q_\beta(\mathbf{x},\mathbf{y}) = p(\mathbf{x})\, p_\beta(\mathbf{y}) \Pi p(\mathrm{y}_i \mid \mathrm{x}_i)^\beta \}$$

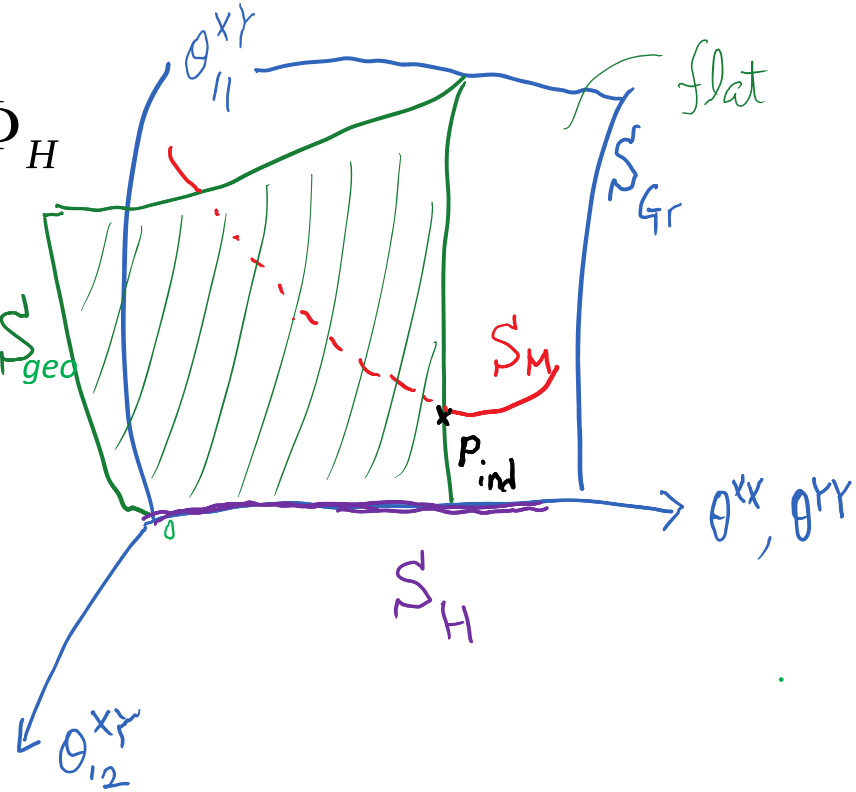
$$\Phi^* = D_{KL}[\, p : S_M \,]$$



$S_H \subset S_{Gr}, S_{Geo}: \quad S_{Gr}, S_H$ dually flat

$S_M \subset S_{Gr}: \quad S_M, S_{Geo}$ not flat

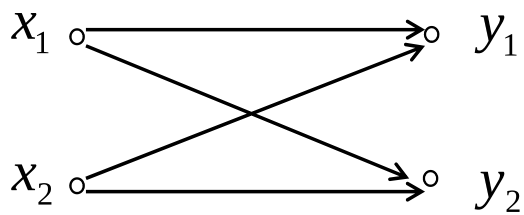
$\Phi_{Gr} \le \Phi_H, \Phi_M; \quad \Phi_{Geo} \le \Phi_H$



Transfer Entropy; Granger causality

$$TE[x_i \to y_j] = \min D_{KL}[p:q] \quad q \in (i \to j) \text{ disconnected}$$

$$= H[Y_j \mid X - X_i] - H[Y_j \mid X]$$

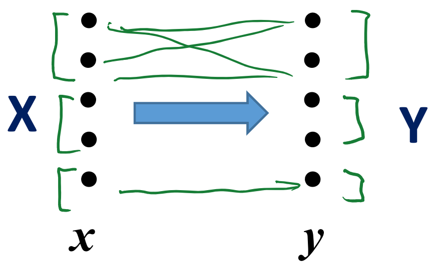


Non-additive

$$TE[x_i \to y_j; x_k \to y_m] \neq TE[x_i \to y_j] + TE[x_k \to y_m]$$

Hierarchy: transfer entropy

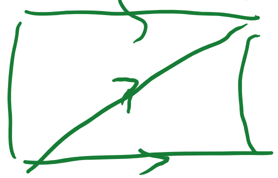
Partition of X



cutting branches

split models

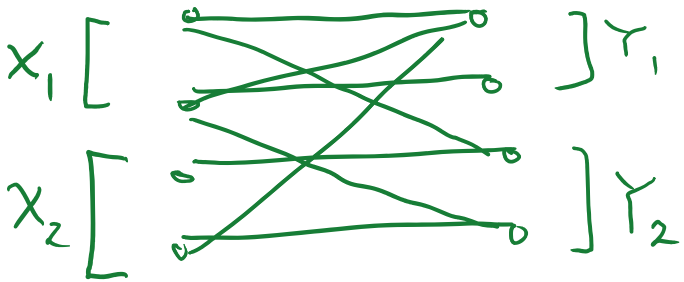
transfer entropy



$$\bigcup X_i = X, \qquad X_i \cap X_j = \phi$$

$$\bigcup Y_i = Y, \qquad Y_i \cap Y_j = \phi$$

Partition



subadditivity

Information Geometry of Hyvärinen Game Score

Following Grinwald, Dawid, Parry, Lauritzen, Hyvärinen

$$L(p,q) = E_{p(x)}\left[l(x,a)\right] : a = q(x)$$

$$S(x,q) = l(x,q) \qquad l(x,q) = \log q(x)$$

S-entropy $\quad H_S[p:q] = E_p\left[S(x,q)\right]$

S-divergence $\quad D_S[p:q] = H_S[p:q] - H_S[p:p]$

Hyvärinen score

$$S(x,q) = \ddot{l}(x,\xi) + \frac{1}{2}\left\{\dot{l}(x,\xi)\right\}^2$$

$$D_S[p:q] = \frac{1}{2}E_p\left[\frac{d}{dx}\left\{\log p(x) - \log q(x,\theta)\right\}^2\right]$$

$$D[p:cq] = D[p:q]$$

$$\boldsymbol{s}(x,\xi) = \partial_\xi \ddot{l}(x,\xi) + \dot{l}(x,\xi)\partial_\xi \dot{l}(x,\xi)$$

parametric case $M = \{q(\boldsymbol{x}, \boldsymbol{\xi})\} : A$

$$\min_{\xi} D_S\left[p : q(\boldsymbol{x}, \boldsymbol{\xi}) \right]$$

$$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\xi}) = \partial_{\xi} S(\boldsymbol{x}, \boldsymbol{\xi})$$: estimating function

$$\sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{x}_i, \boldsymbol{\xi}) = 0$$: estimating equation

Information geometry of $D_S[p : q]$

Asymptotic Analysis of estimator

$$E\left[\Delta\boldsymbol{\xi}\Delta\boldsymbol{\xi}^{T}\right]=\frac{1}{N}K^{-1}VK^{-T}\geq G^{-1}$$

$$K=E\left[\partial_{\xi}\boldsymbol{s}\left(\boldsymbol{x},\boldsymbol{\xi}\right)\right]$$

$$V=E\left[\boldsymbol{s}\left(\boldsymbol{x},\boldsymbol{\xi}\right)\boldsymbol{s}\left(\boldsymbol{x},\boldsymbol{\xi}\right)^{T}\right]\to G$$

$$E\left[\Delta\boldsymbol{\xi}\Delta\boldsymbol{\xi}^{T}\right]=G^{-1}AG^{-1}$$

$$A=E\left[\boldsymbol{a}\left(\boldsymbol{x},\boldsymbol{\xi}\right)\boldsymbol{a}\left(\boldsymbol{x},\boldsymbol{\xi}\right)^{T}\right]$$

$$\boldsymbol{s}\left(\boldsymbol{x},\boldsymbol{\xi}\right)=c\left\{\partial_{\xi}\log\left(\boldsymbol{x},\boldsymbol{\xi}\right)+\boldsymbol{a}\left(\boldsymbol{x},\boldsymbol{\xi}\right)\right\}$$

$\boldsymbol{a}\perp\partial_{\xi}\log q$: efficient

Hyvärinen score

$$S(x,q) = \ddot{l}(x,\xi) + \frac{1}{2}\left\{\dot{l}(x,\xi)\right\}^2$$

$$D_S[p:q] = \frac{1}{2}E_p\left[\frac{d}{dx}\left\{\log p(x) - \log q(x,\theta)\right\}^2\right]$$

$$D[p:cq] = D[p:q]$$

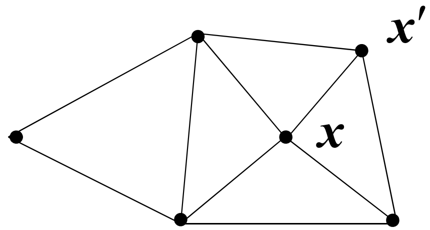
$$\boldsymbol{s}(x,\xi) = \partial_\xi \ddot{l}(x,\xi) + \dot{l}(x,\xi)\partial_\xi \dot{l}(x,\xi)$$

Hyvärinen estimator

Fisher efficient $\Longleftrightarrow$ q **multivariate Gaussian**

Discrete case : $\boldsymbol{x} \in$ **graph nodes**

$$\Delta f\left(\boldsymbol{x}\right) = \frac{1}{N_x} \sum_{\boldsymbol{x}' \in N} \left\{ f\left(\boldsymbol{x}\right) - f\left(\boldsymbol{x}'\right) \right\}$$



$$N_x = \mathrm{const}$$

$$S(\boldsymbol{x},q)=\left\{\frac{\Delta q(\boldsymbol{x})}{q(\boldsymbol{x})}\right\}^2-2\Delta\left\{\frac{\Delta q(\boldsymbol{x})}{q(\boldsymbol{x})}\right\}$$

$$D_S[\boldsymbol{\xi}:\boldsymbol{\xi}']=E_{q(\boldsymbol{x},\boldsymbol{\xi})}\left\{\frac{\Delta q(\boldsymbol{x},\boldsymbol{\xi})}{q(\boldsymbol{x},\boldsymbol{\xi})}-\frac{\Delta q(\boldsymbol{x},\boldsymbol{\xi}')}{q(\boldsymbol{x},\boldsymbol{\xi}')}\right\}^2$$

$$\sum_{\boldsymbol{x}}\Delta f(\boldsymbol{x})h(\boldsymbol{x})=\sum_{\boldsymbol{x}}f(\boldsymbol{x})\Delta' h(\boldsymbol{x})$$

$$f,h\to 0$$

$$x\to\pm\infty$$

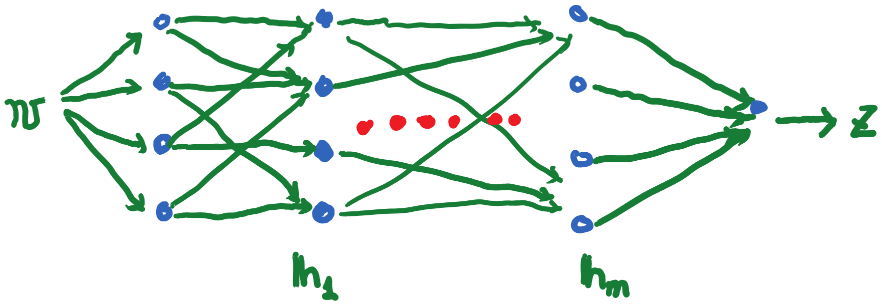
$$\int f'(x)h(x)\,dx=-\int f(x)h'(x)\,dx$$

Deep Learning

Self-Organization + Supervised Learning

RBM: Restricted Boltzmann Machine

Auto-Encoder, Recurrent Net



tricks !!

ideas !

Dropout

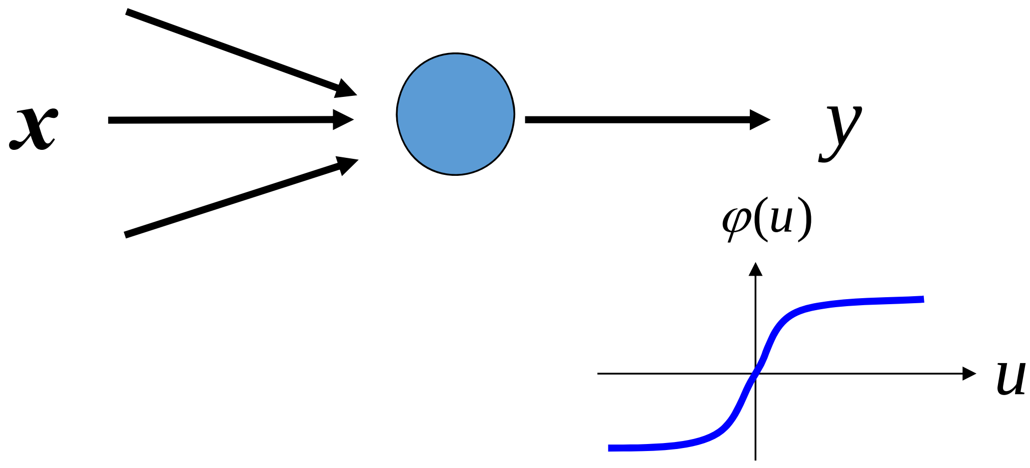
Contrastive divergence

bi-directional

convolution

Mathematical Neurons

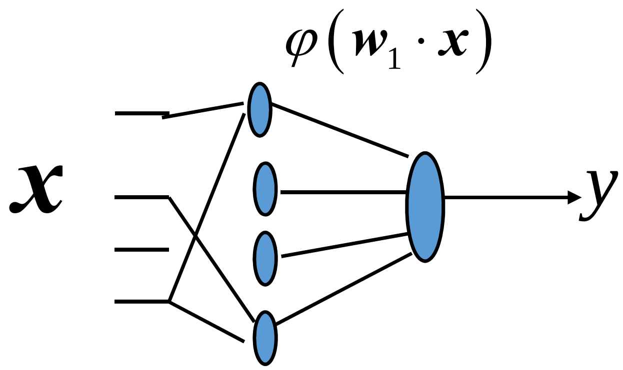
$$y=\varphi\left(\sum w_i x_i - h\right)=\varphi\left(\boldsymbol{w}\cdot\boldsymbol{x}\right)$$



Multilayer Perceptrons

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

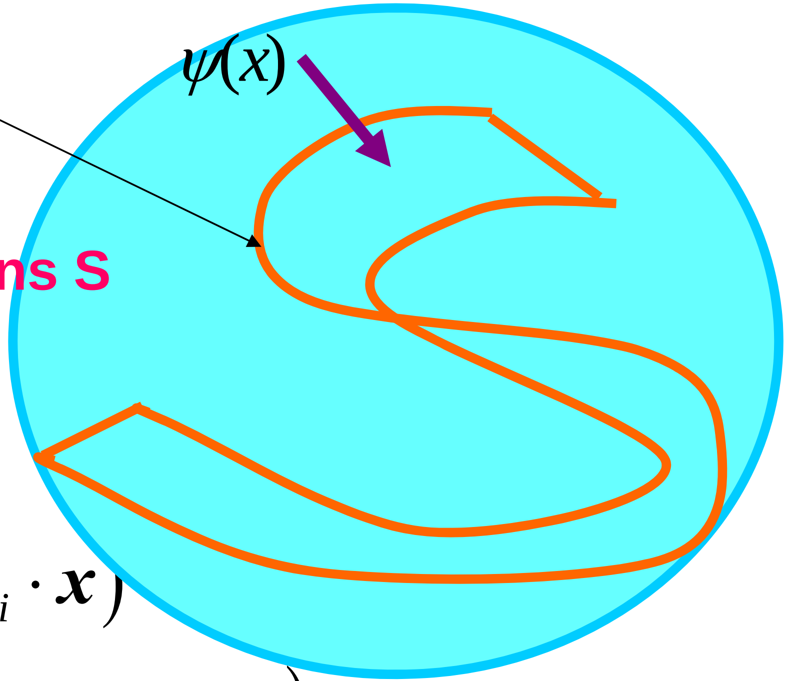
$$x = (x_1, x_2, ..., x_n)$$



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

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

Multilayer Perceptron



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

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

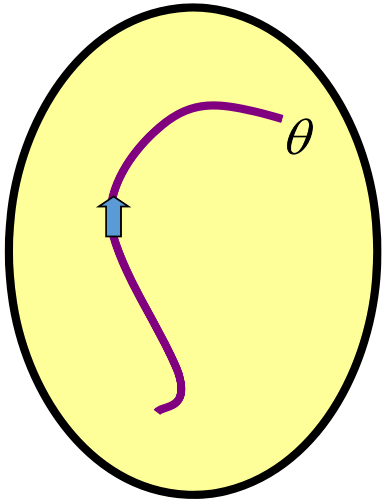
$$\boldsymbol{\theta} = \left(v_1 \cdots, v_m\ ;\ \boldsymbol{w}_{1,} \cdots, \boldsymbol{w}_m\right)$$

Backpropagation --- stochastic gradient learning

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

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

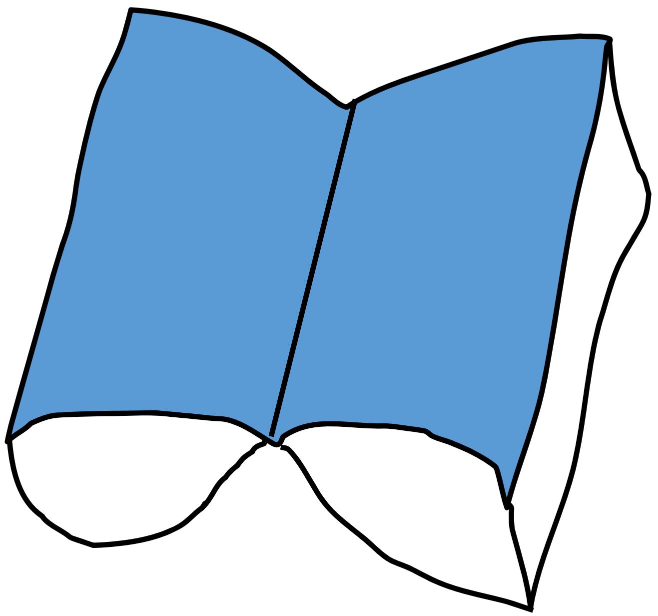
$$= -\log p(y, \boldsymbol{x}; \boldsymbol{\theta})$$

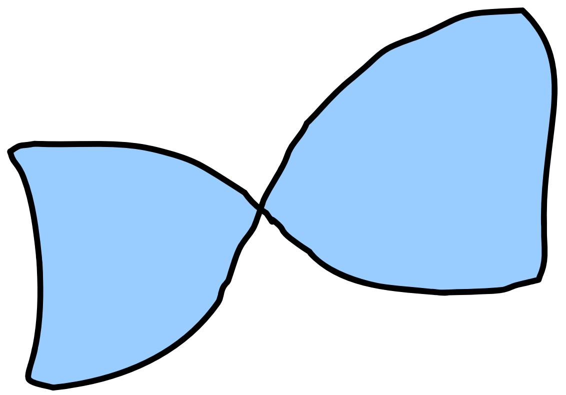


$$\Delta \boldsymbol{\theta}_t = -\eta_t \frac{\partial l(y_t, x_t; \theta_t)}{\partial \boldsymbol{\theta}}$$

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

singularities

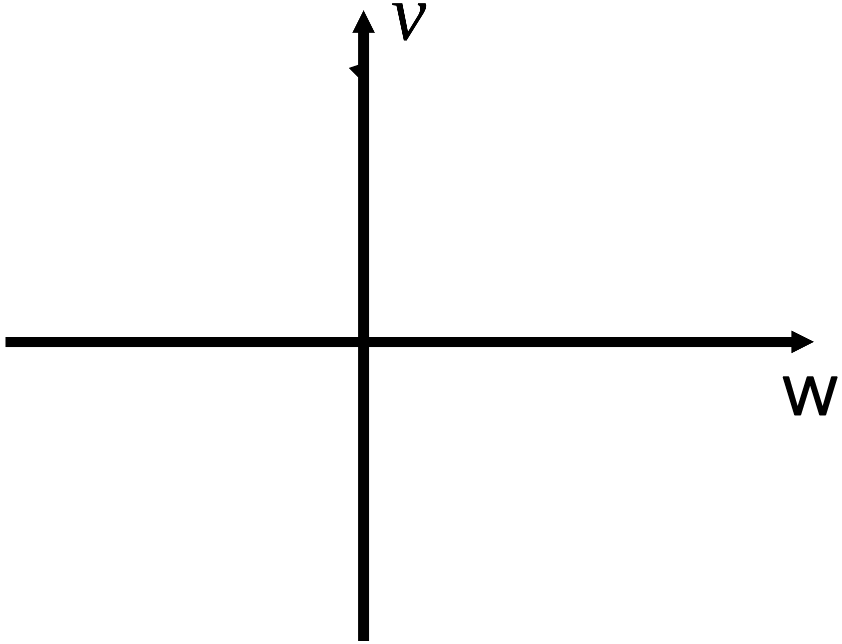


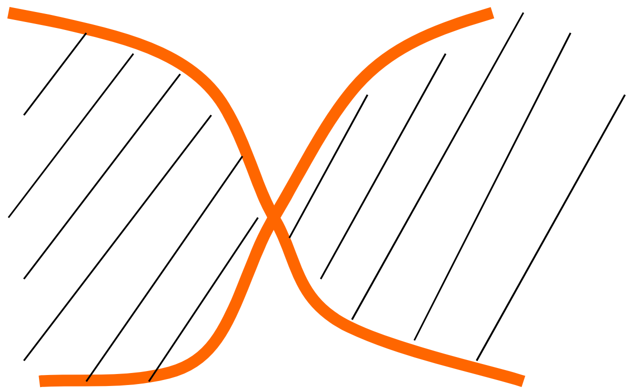


Geometry of singular model

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

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





model: 2 hidden neurons

$$f\left(\boldsymbol{x},\boldsymbol{\theta}\right)=w_1\varphi\left(\boldsymbol{J}_1\cdot\boldsymbol{x}\right)+w_2\varphi\left(\boldsymbol{J}_2\cdot\boldsymbol{x}\right)$$

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

$$\varphi\left(u\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u}e^{-\frac{t^2}{2}}dt$$

loss function: $l\left(\boldsymbol{x}, y; \boldsymbol{\theta}\right) = \frac{1}{2}\left\{y - f\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right\}^2$

y : teacher signal : $\boldsymbol{\theta}_0$ **stochastic descent learning**

$$\dot{\boldsymbol{\theta}} = -\eta \left\langle \frac{\partial l\left(\boldsymbol{x}_t, y_t, \boldsymbol{\theta}_t\right)}{\partial \boldsymbol{\theta}} \right\rangle$$ **backprop : vanilla gradient**

Natural Gradient Stochastic Descent

$$\dot{\boldsymbol{\theta}} = -\eta G^{-1}\left(\boldsymbol{\theta}_t\right)\left\langle \nabla_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, y_t, \boldsymbol{\theta}_t\right)\right\rangle$$

$$\nabla_{\boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}$$

$$G\left(\boldsymbol{\theta}\right) = \left\langle \nabla_{\boldsymbol{\theta}} l \nabla_{\boldsymbol{\theta}} l \right\rangle$$: **Fisher Information Matrix**

invarint; steepest descent

Natural gradient is superior

Steepest descent; invariant   Yan Ollivier

Fisher-efficient

Natural gradient is non-vanishing even in multiple layers

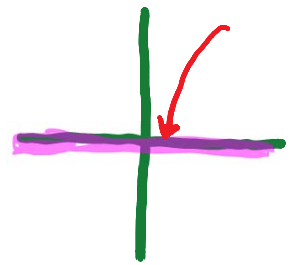
Good at singular regions   (avoid plateaus: Milnor attractor)

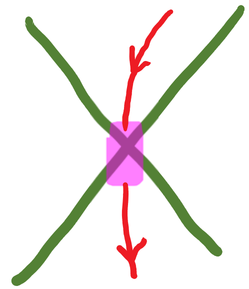
Adaptive Natural Gradient

$$G_{t+1}^{-1} = (1+\varepsilon) G_t^{-1} - \varepsilon G_t^{-1} \nabla l(x_t) \nabla l(x_t)^T G_t^{-1}$$

$G^{-1} \to \infty$, $\nabla l \to 0$ at singularities

$\underline{G^{-1} \nabla l}$





Singular Region in Parameter Space

$$R\left(w,\boldsymbol{J}\right)=\left\{\boldsymbol{\theta}\middle|\boldsymbol{J}_1=\boldsymbol{J}_2=\boldsymbol{J},w_1+w_2=w\right\}$$

$$\cup\left\{\boldsymbol{\theta}\middle|w_1=0,w_2=w,\boldsymbol{J}_2=\boldsymbol{J}\right\}$$

$$\cup\left\{\boldsymbol{\theta}\middle|w_1=w,w_2=0,\boldsymbol{J}_1=\boldsymbol{J}\right\}$$

$$f\left(\boldsymbol{x},\boldsymbol{\theta}\right)=w_1\varphi\left(\boldsymbol{J}_1\cdot\boldsymbol{x}\right)+w_2\varphi\left(\boldsymbol{J}_2\cdot\boldsymbol{x}\right)$$

Coordinate transformation

$$\boldsymbol{v} = \frac{w_1 \boldsymbol{J}_1 + w_2 \boldsymbol{J}_2}{w_1 + w_2},$$

$$w = w_1 + w_2,$$

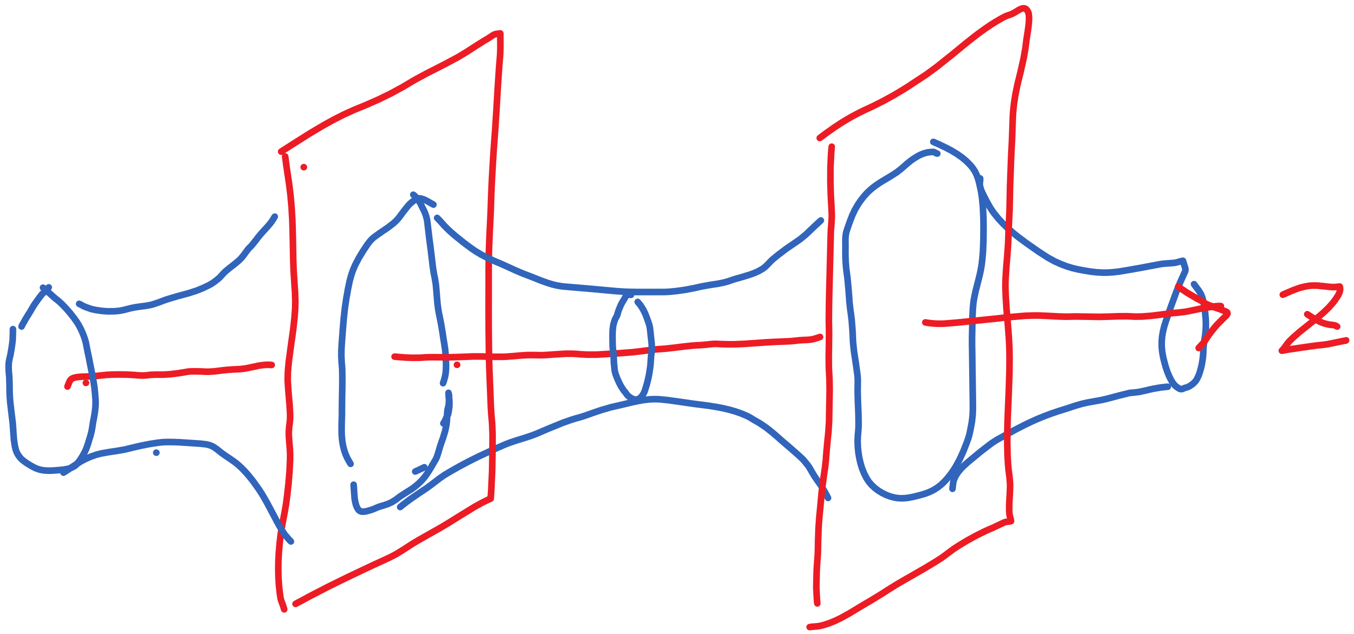
$$\boldsymbol{u} = \boldsymbol{J}_2 - \boldsymbol{J}_1,$$

$$z = \frac{w_2 - w_1}{w_1 + w_2}$$

$$\boldsymbol{\xi} = \left(\boldsymbol{v}, w, \boldsymbol{u}, z\right)$$

Singular Region

$$R(w, \boldsymbol{J}) = \{\boldsymbol{u} = 0\} \bigcup \{z = \pm 1\}$$



Milnor attractor

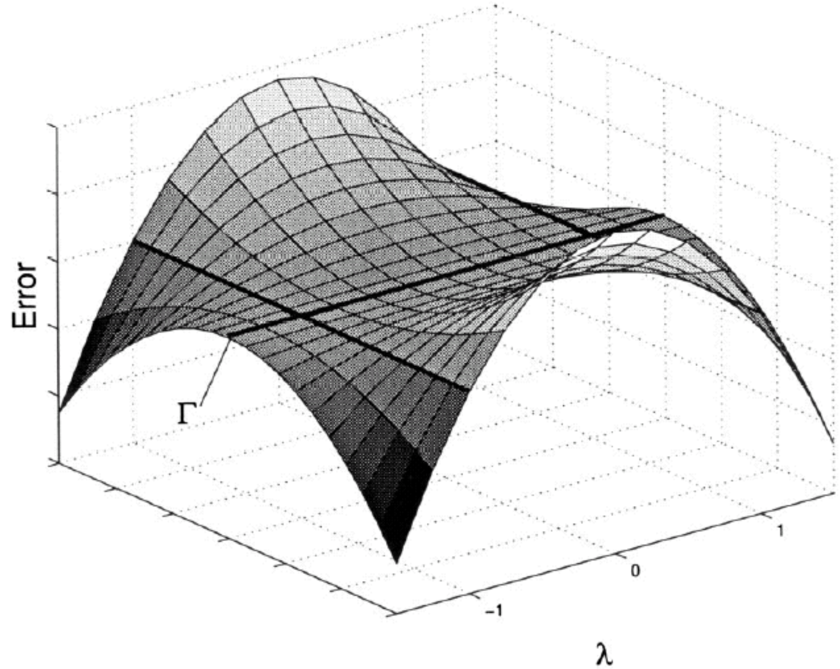


Fig. 5. Critical set with local minima and plateaus.

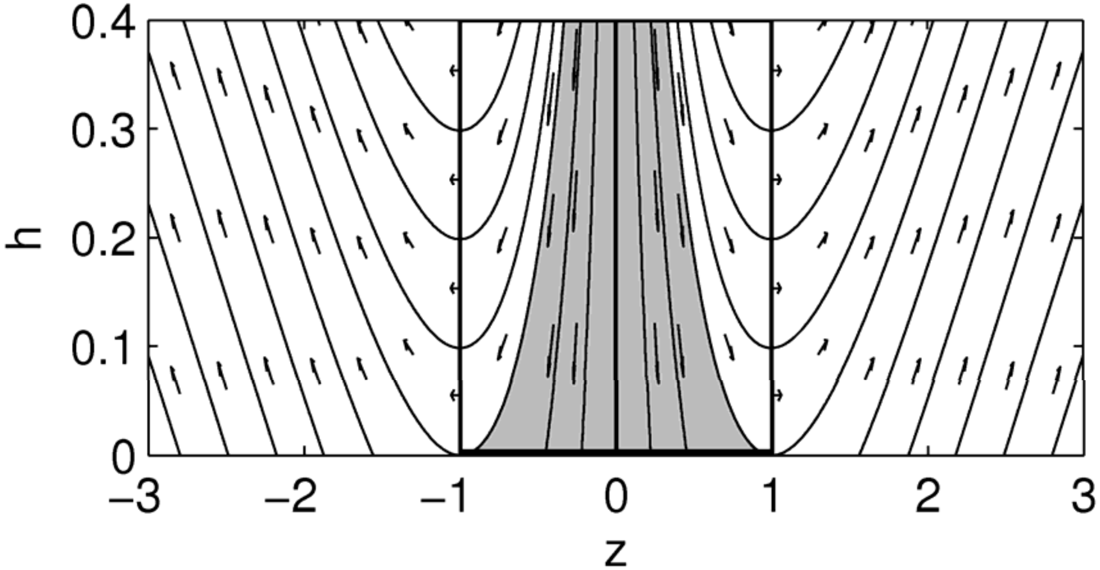
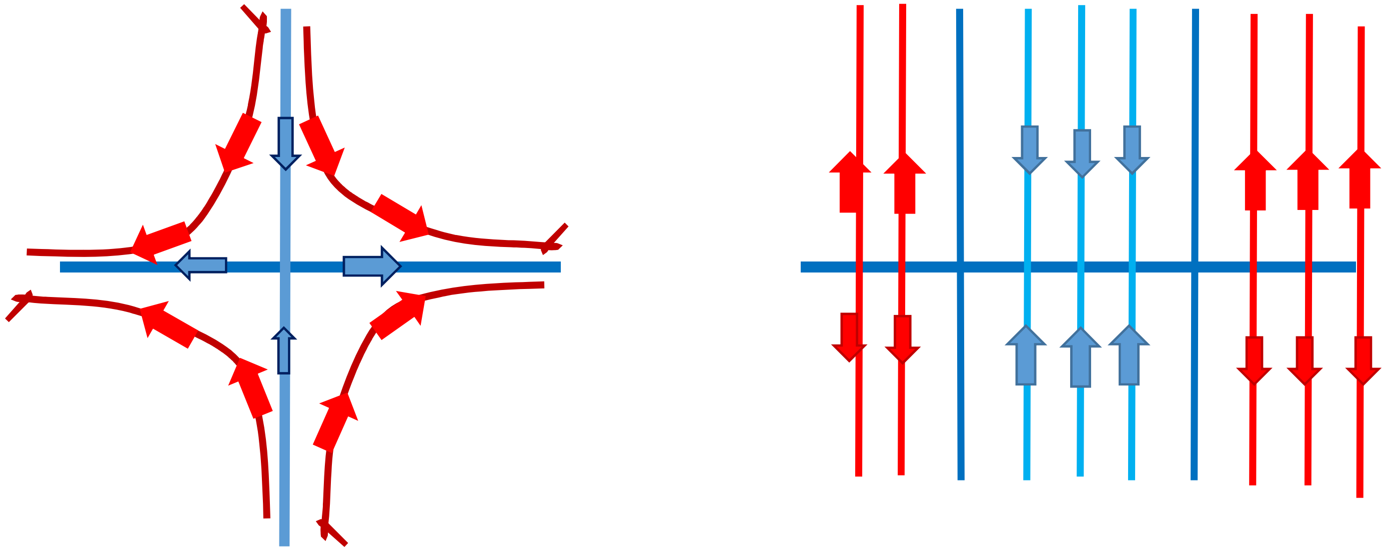


Fig. 2: trajectories

Saddle and plateau



Topology of singular R

blow-down coordinates ： $\boldsymbol{\alpha}=(\tau,\sigma,\boldsymbol{e})$

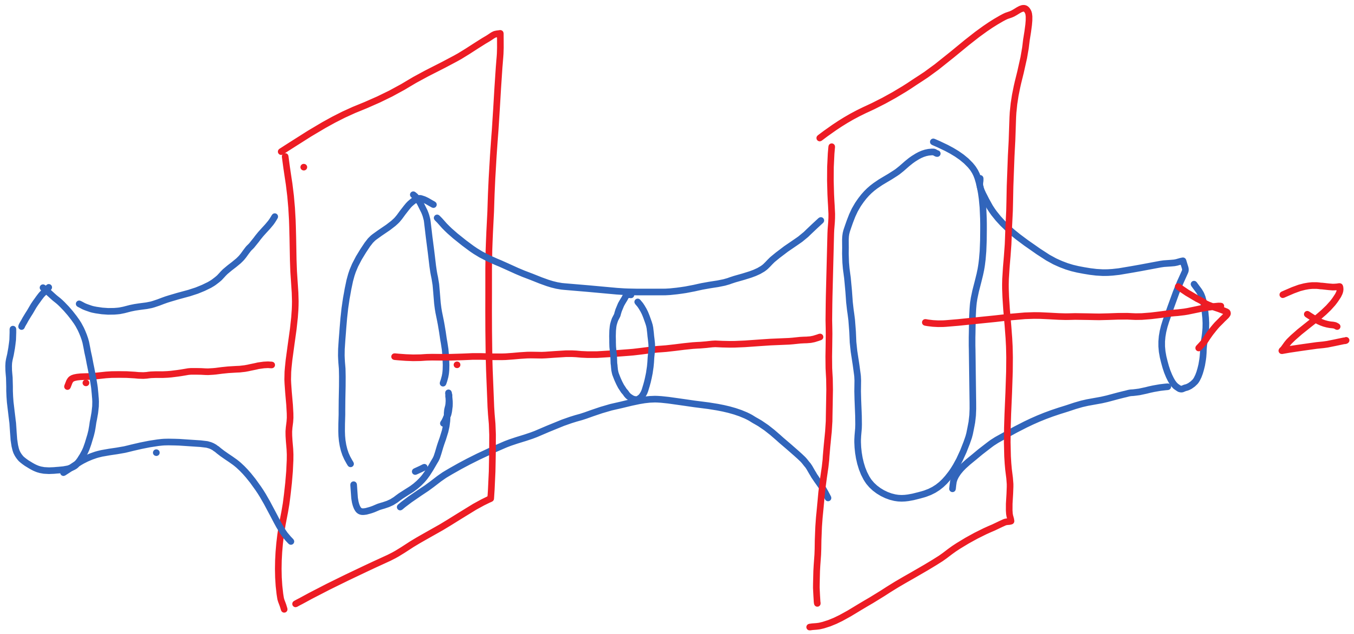
$$\tau = c_1\left(1-z^2\right)u^2, \qquad u = |\boldsymbol{u}|$$

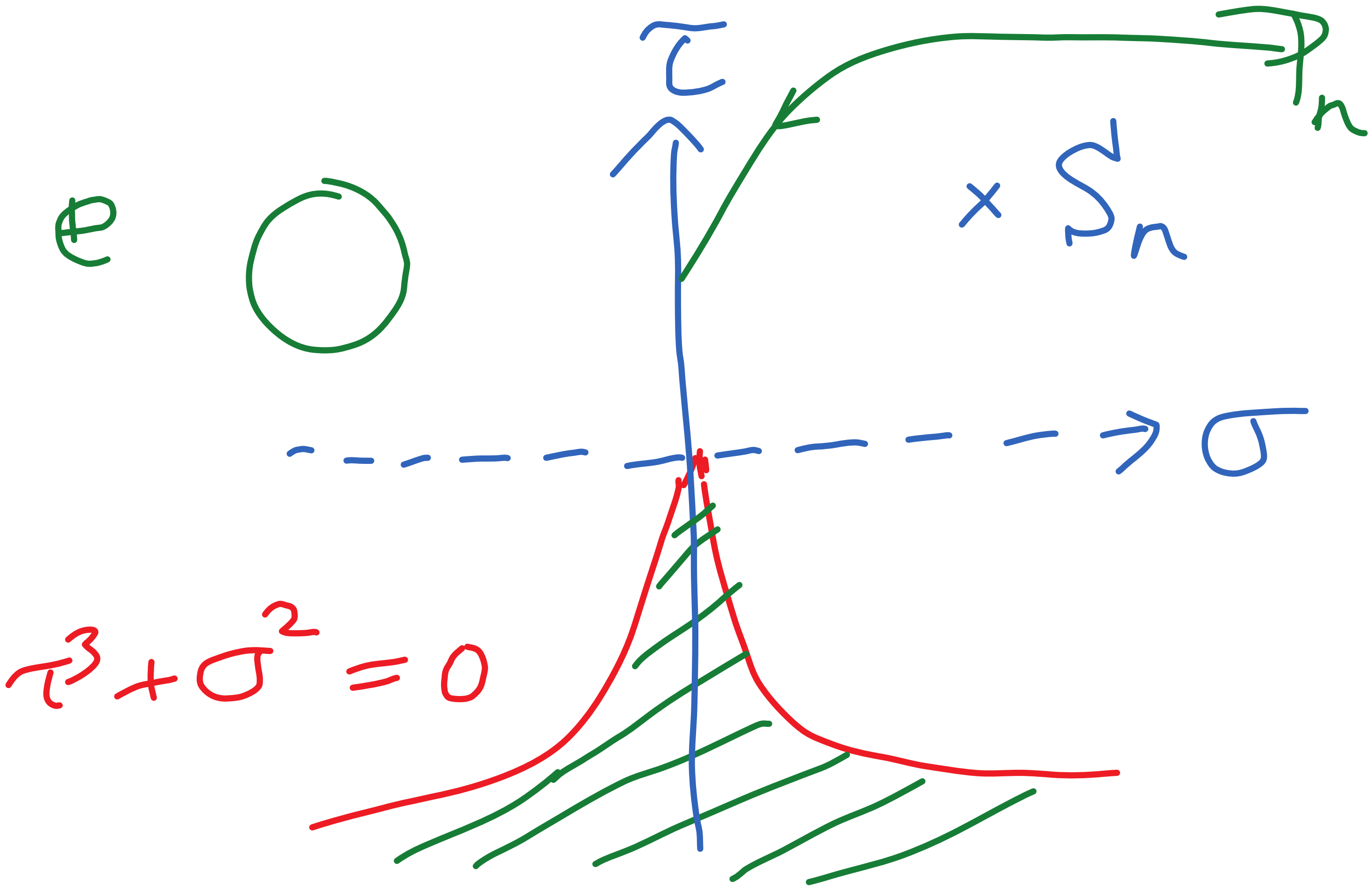
$$\sigma = c_2 z\left(1-z^2\right)u^3,$$

$$\boldsymbol{e} = \frac{\boldsymbol{u}}{|\boldsymbol{u}|} \in S_n, \qquad |\boldsymbol{e}| = 1$$

Singular Region

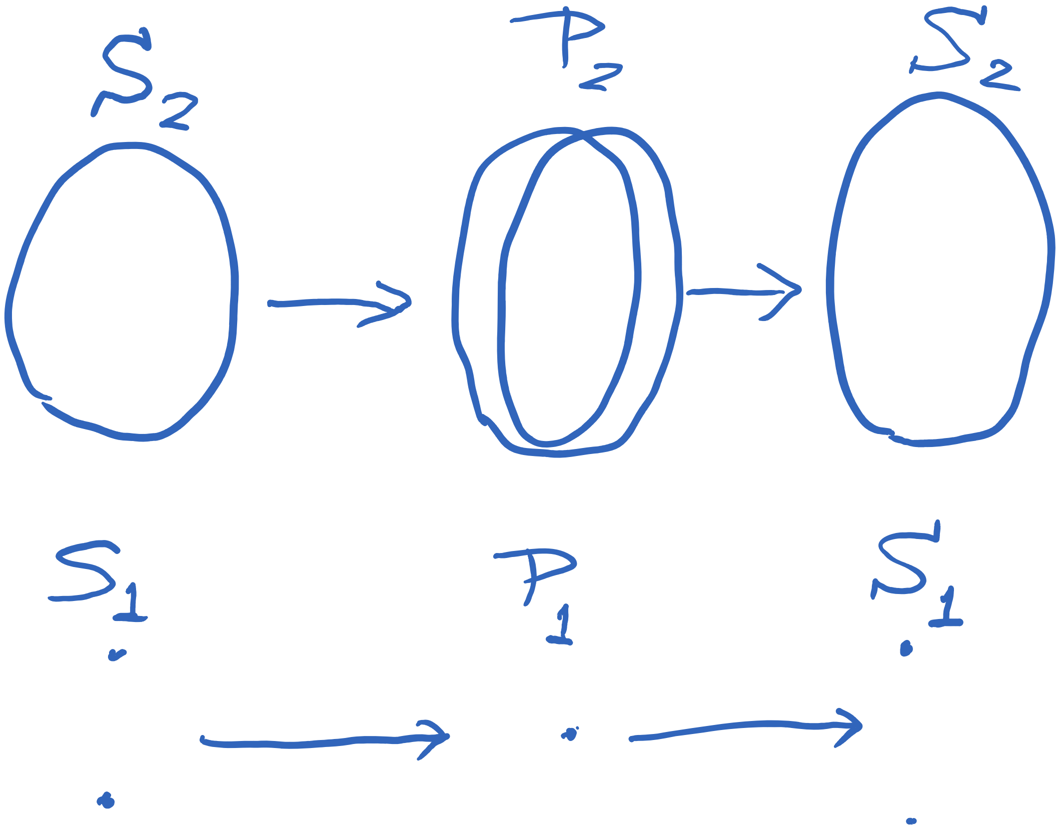
$$R(w, \boldsymbol{J}) = \{\boldsymbol{u} = 0\} \bigcup \{z = \pm 1\}$$





τ
σ
× S_n
P_n
τ^3 + σ^2 = 0

Sphere *Sn* and Projective space *Pn*



natural gradient learning near singularity

$$\frac{d}{dt}\begin{pmatrix}\tau\\ \sigma\end{pmatrix} = -\eta\begin{pmatrix}\tau\\ \sigma\end{pmatrix} \qquad : \quad \text{true model} \in R$$

$$\frac{d}{dt}\begin{pmatrix}\tau\\ \sigma\end{pmatrix} = O(1) \qquad : \quad \text{true model} \notin R$$

Milnor attractor

Canonical Divergence in Manifold of Dual Affine Connections

Nihat Ay and *S. Amari*

Divergence and metric

$$D\left[p:q\right]\geq 0$$

$$D\left[\boldsymbol{\xi}:\boldsymbol{\xi}+d\boldsymbol{\xi}\right]=\frac{1}{2}g_{ij}\left(\boldsymbol{\xi}\right)d\xi^{i}d\xi^{j}+O\left(\left|d\boldsymbol{\xi}\right|^{3}\right)$$

G : **Riemannian metric, positive-definite**

Divergence and dual affine connections

$$\Gamma_{ijk} \sim \nabla \qquad\qquad \Gamma^{*}_{ijk} \sim \nabla^{*}$$

$$\Gamma_{ijk} = -\partial_i \partial_j \partial'_k D\left[\xi : \xi'\right]_{\xi'=\xi}$$

$$\Gamma^{*}_{ijk} = -\partial'_i \partial'_j \partial_k D\left[\xi : \xi'\right]_{\xi'=\xi}$$

$$\partial_i = \frac{\partial}{\partial \xi^i}; \qquad\qquad \partial'_j = \frac{\partial}{\partial \xi'^{j}}$$

Dual geometry

$$\left\{M, g, \nabla, \nabla^*\right\}$$

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

$$\left\{M, g, T\right\}, \qquad T_{ijk} = \Gamma^*_{ijk} - \Gamma_{ijk}$$

$$\Gamma^{\pm\alpha}_{ijk} = \Gamma^{o}_{ijk} \mp \frac{\alpha}{2} T_{ijk}$$

Γ^{o} : **Levi-Civita connection**

Dual geometry $\Rightarrow$ canonical divergence

$$M: \text{ dually flat } \quad : \exists\, \psi(\boldsymbol{\theta}), \varphi(\boldsymbol{\eta})$$

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

Bregman divergence

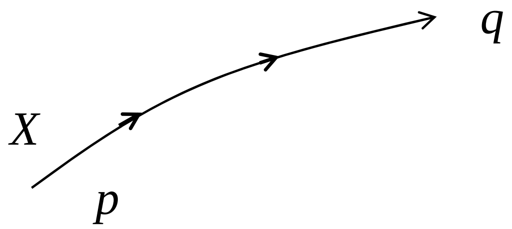
Exponential map : $\xi(t)$ geodesic

$$\nabla_{\dot{\xi}}\dot{\xi} = 0$$

$$\xi(0) = p$$

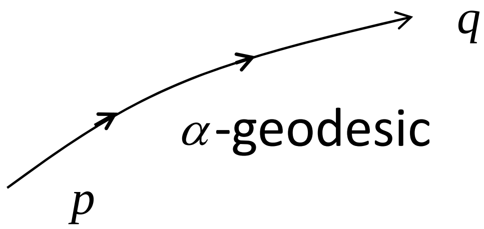
$$\xi(1) = q$$

$$\dot{\xi}(0) = X = \log_p q$$



Exponential map divergence

$$D[p:q] = \|X(p:q)\|^2$$



α-**divergence**

$$D_\alpha[p:q] = \|X_\alpha(p:q)\|^2$$

Theorem 1. **Exponential map divergence induces** $\alpha = -3$ **geometry**

Theorem 2. $\alpha = -\frac{1}{3}$ **exponential map divergence recovers the original geometry**

Standard divergence: $D_{\mathrm{stan}}\left[p : q\right] = \left\| X_{-1/3}\left(p, q\right)\right\|^2$

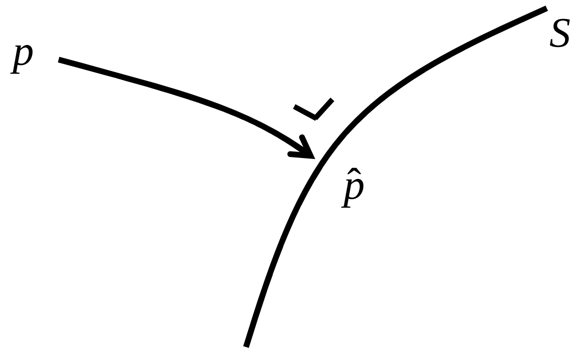
$$D[p:q] = \int_0^1 < X_t(q,p), \dot{\boldsymbol{\xi}}_{q,p}(t) > dt$$

$$= \int_0^1 t \,\|\, \dot{\boldsymbol{\xi}}_{q,p}(t) \,\|^2 \, dt$$

$$\int_0^1 w(t) \,\|\, \dot{\boldsymbol{\xi}}_{q,p}(t) \,\|^2 \, dt$$

Divergence and projection

$$\hat{p} = \arg\min_{q \in S} D[p:q]$$



projection theorem:

$$X = c \ \mathrm{grad}_q \ D[p:q]$$

IEEE ISIT-2011 Sankt Petersburg

Data Compression in Multiterminal Statistical Inference

Shun-ichi Amari
RIKEN Brain Science Institute

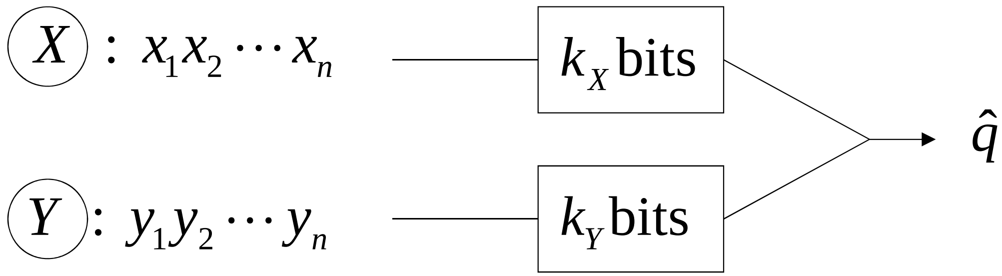
A long standing problem

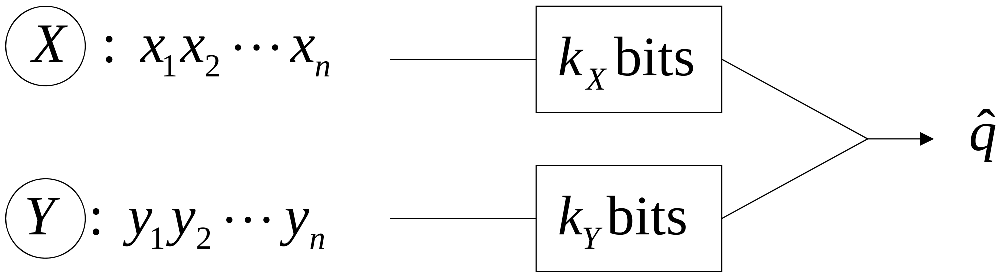
T.Berger; Csiszar, Ahlswede, Burnashev, Han, Amari

correlated sources X, Y

data compression and statistical inference

$$p(x, y; q) \; ; \; iid$$

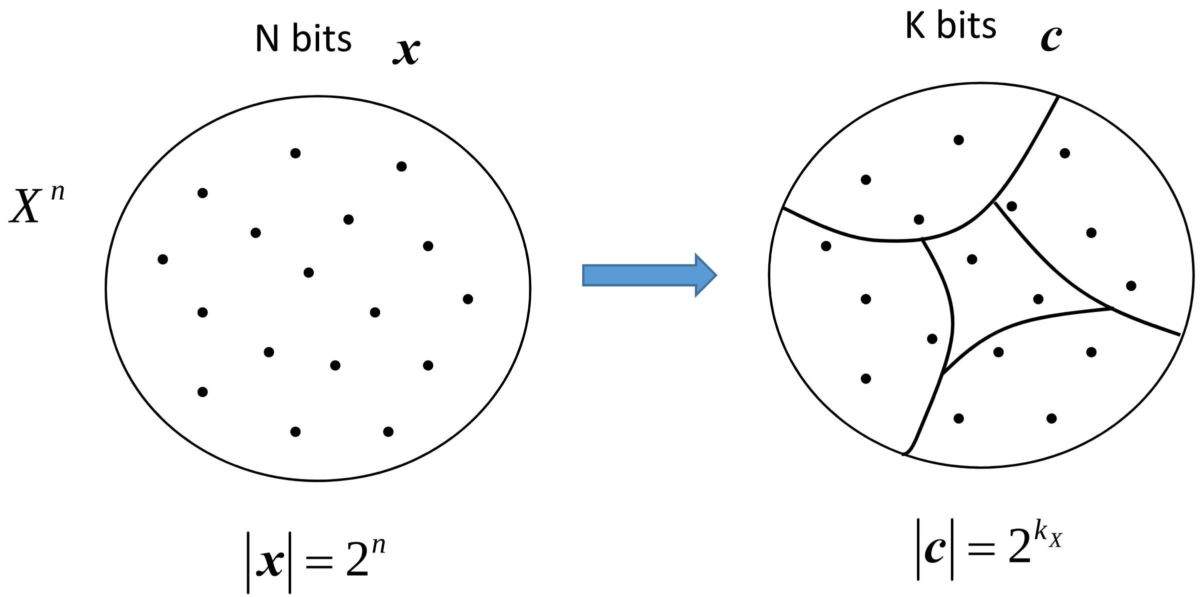




$p(x, y; q)$; *iid* $\qquad \text{Prob}\{x = y\} = q$

$$\text{binary} : x, y = 0,1 \; ; \; \text{Prob}\{x = 1\} = \text{Prob}\{y = 1\} = \frac{1}{2}$$

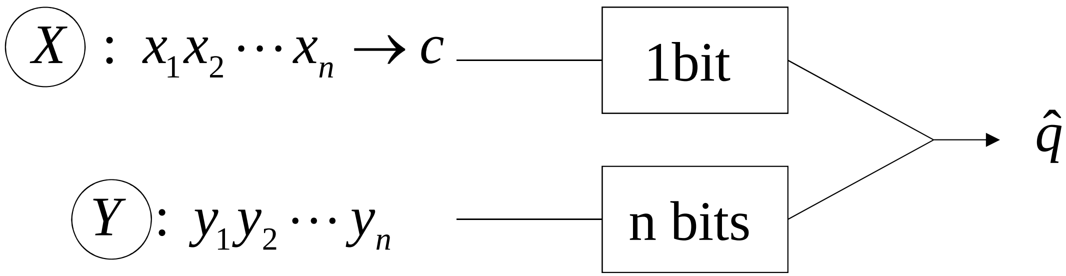
Encoding : data compression



One-bit helper case

$k_X = 1, \quad k_Y = n$ $\qquad$ $c = \mathrm{sgn}(\boldsymbol{a} \cdot \boldsymbol{x})$

$c \qquad \boldsymbol{y}$



Is single-bit encoding optimal?

It is optimal

when $q = 1/2$ **(** x, y **independent),**

but not for general q .

Fisher information: $k_X = 1,\ k_Y = n$

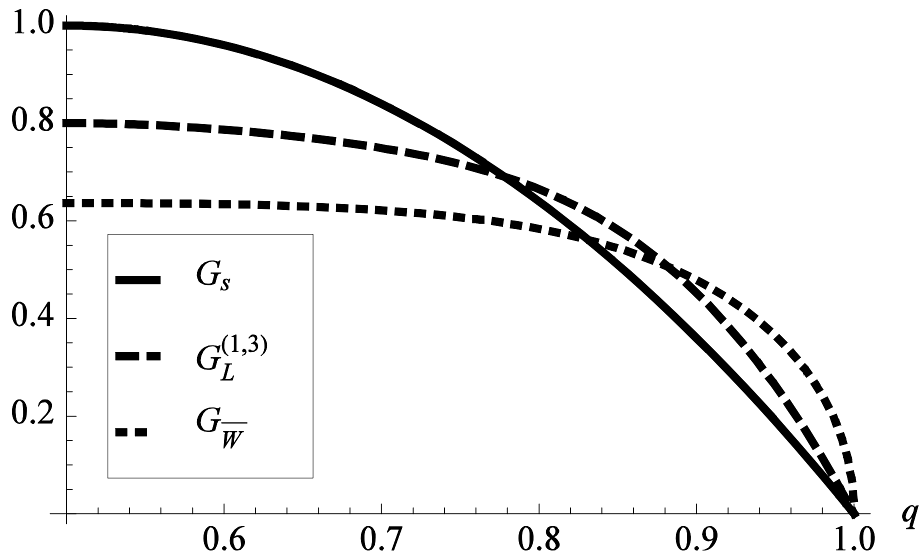
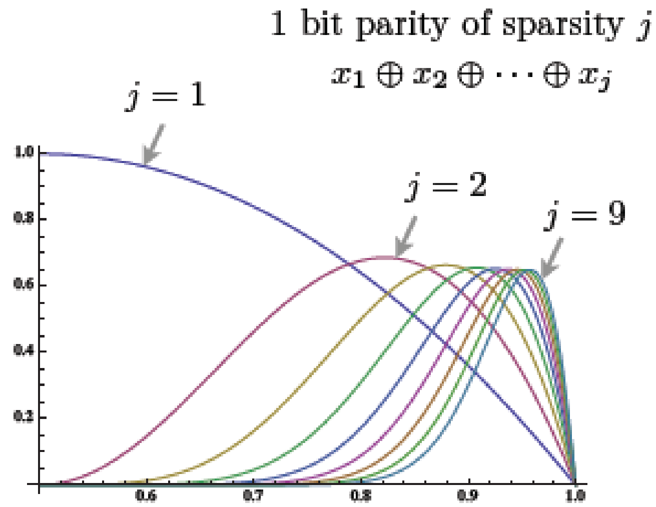


Fig.2

Kingo Kobayashi: parity encoding

$$x_1 \oplus x_2 \oplus \cdots \oplus x_s$$

in progress !



Information Geometry and Transportation Problem (Wasserstein distance)

entropic relaxation : min <c, p> - a{-H(p)} dual

New Paper

S. Pal and T-K L. Wong,
Exponentially concave function and a new information geometry

Portfolio theory, transportation problem and information geometry
(dually projectively flat)