筑波大学講義

情報幾何の展開

甘利俊一
理化学研究所　栄誉研究員
東京大学　名誉教授
帝京大学特任教授

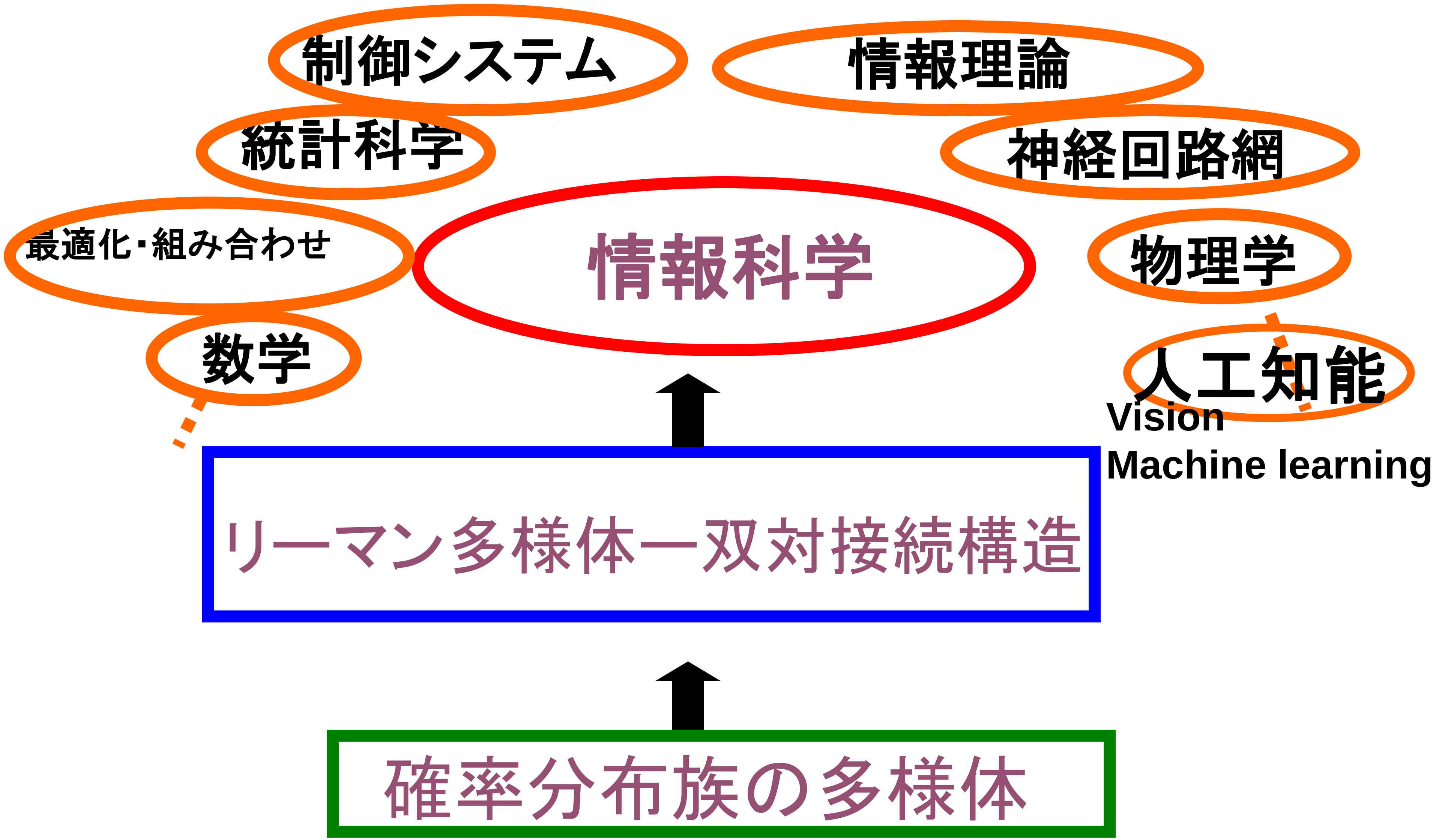
1. 統計推論の情報幾何：不変性
2. 双対平坦空間：凸解析とLegendre変換
3. 多層神経回路の統計神経力学
4. Wasserstein距離の情報幾何

情報幾何

-- 確率分布族のなす多様体

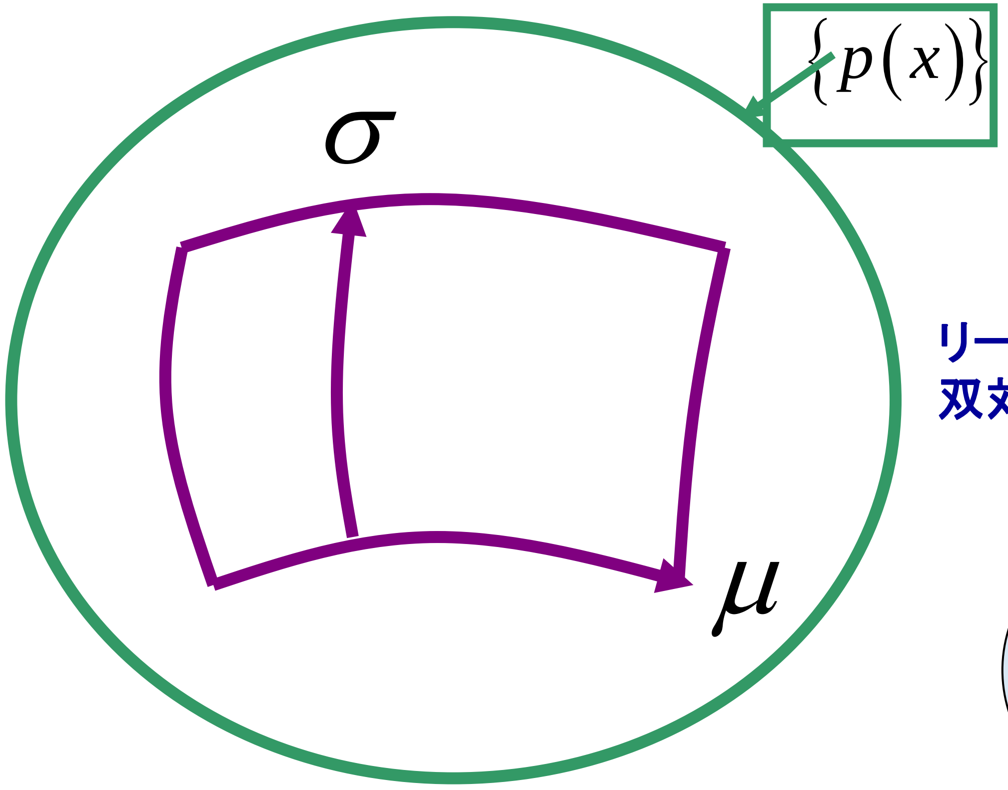
$$M = \{ p(\mathbf{x}) \}$$

情報幾何


制御システム
情報理論
統計科学
神経回路網
最適化・組み合わせ
情報科学
物理学
数学
人工知能
Vision
Machine learning
リーマン多様体ー双対接続構造
確率分布族の多様体

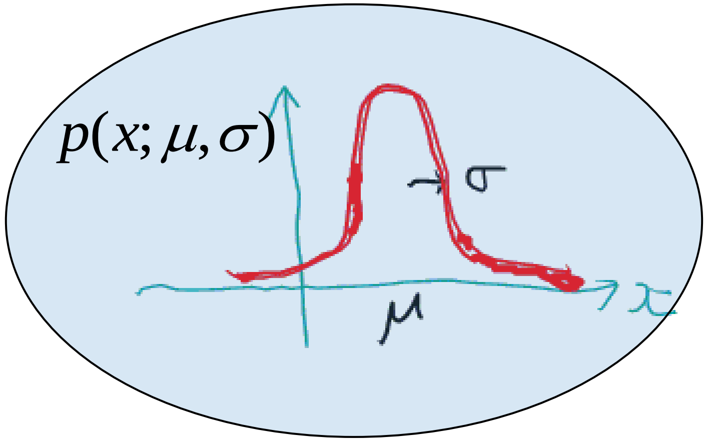
情報幾何とは ？

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



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

リーマン幾何
双対アファイン接続

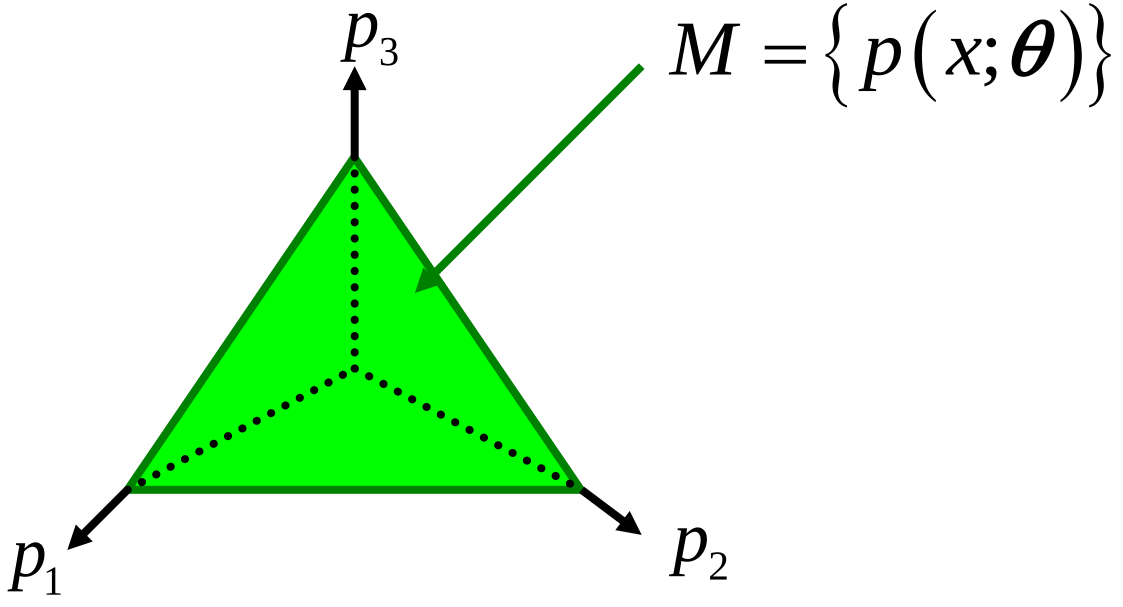


離散確率分布(三つ目さいころ)

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

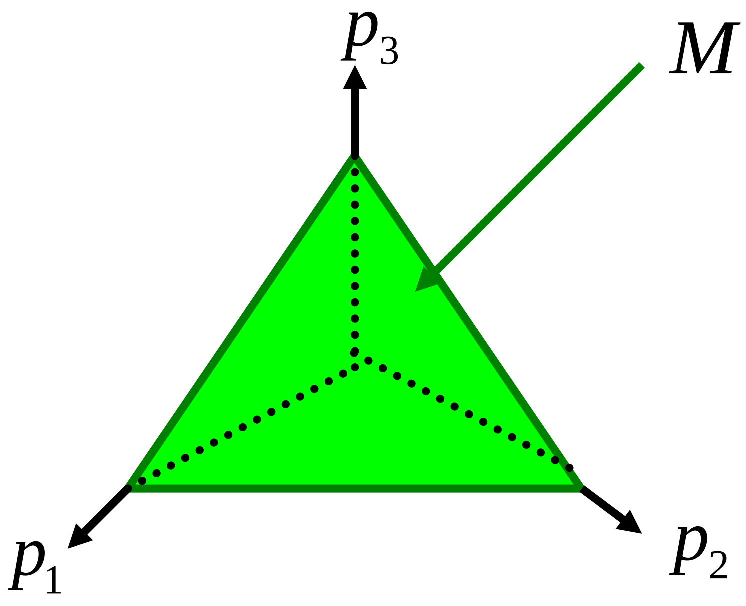
n = 3

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



確率分布族のつくる多様体(座標系)

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



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

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

$$\boldsymbol{\theta} = (p_1, p_2)$$

$$\boldsymbol{\theta} = \left(\log \frac{p_1}{p_3}, \log \frac{p_2}{p_3} \right)$$

不変性の原理:　$S=\{p(x,\boldsymbol{\theta})\}$

1. **パラメータのとり方によらない**

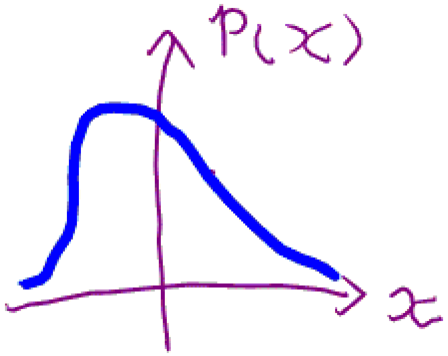
$$\xi=\xi(\theta),\quad \bar{p}(x,\xi)\qquad D=\sum\theta_i^2\neq\sum\xi_i^2$$

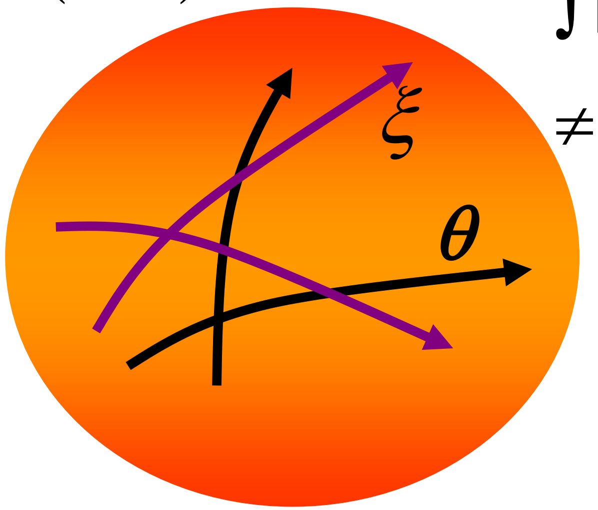
2. **確率変数の表示スケールによらない**

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

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

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





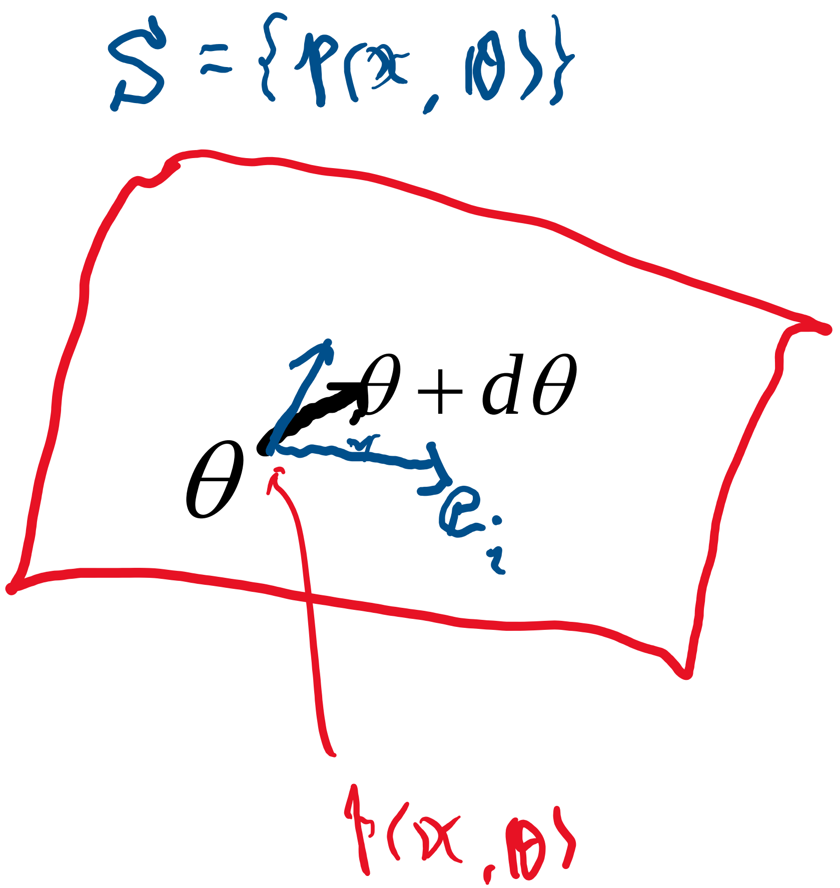
確率分布空間の接空間

Spanned by scores

$$d\boldsymbol{\theta} = \sum d\theta^i \mathbf{e}_i$$

$$g_{ij}(\boldsymbol{\theta}) = < \mathbf{e}_i, \mathbf{e}_j >$$

$$\mathbf{e}_i = \frac{\partial}{\partial \theta^i} \approx \frac{\partial}{\partial \theta^i} \log p(x, \theta)$$



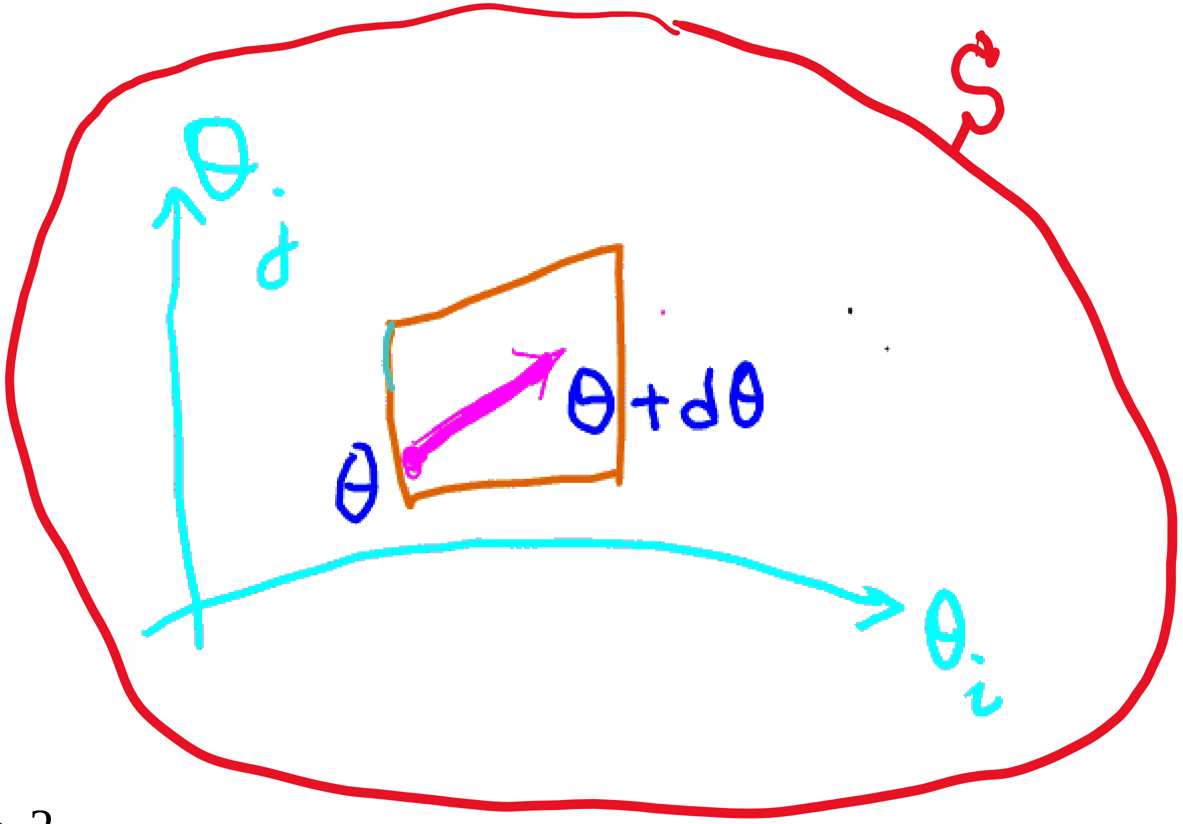
リーマン構造

$$ds^2 = \langle d\boldsymbol{\theta}, d\boldsymbol{\theta} \rangle = \sum g_{ij}(\theta) d\theta^i d\theta^j$$

$$= d\theta^T G(\theta) d\theta$$

$$G(\theta) = (g_{ij}) = \langle d\mathbf{e}_i, d\mathbf{e}_j \rangle$$

Euclidean $G = E$ $\quad ds^2 = \sum (d\theta^i)^2$



リーマン計量とアファイン接続
双対接続　{M, G, ∇, ∇*}

Fisher情報行列

$g = (g_{ij})$

共変微分

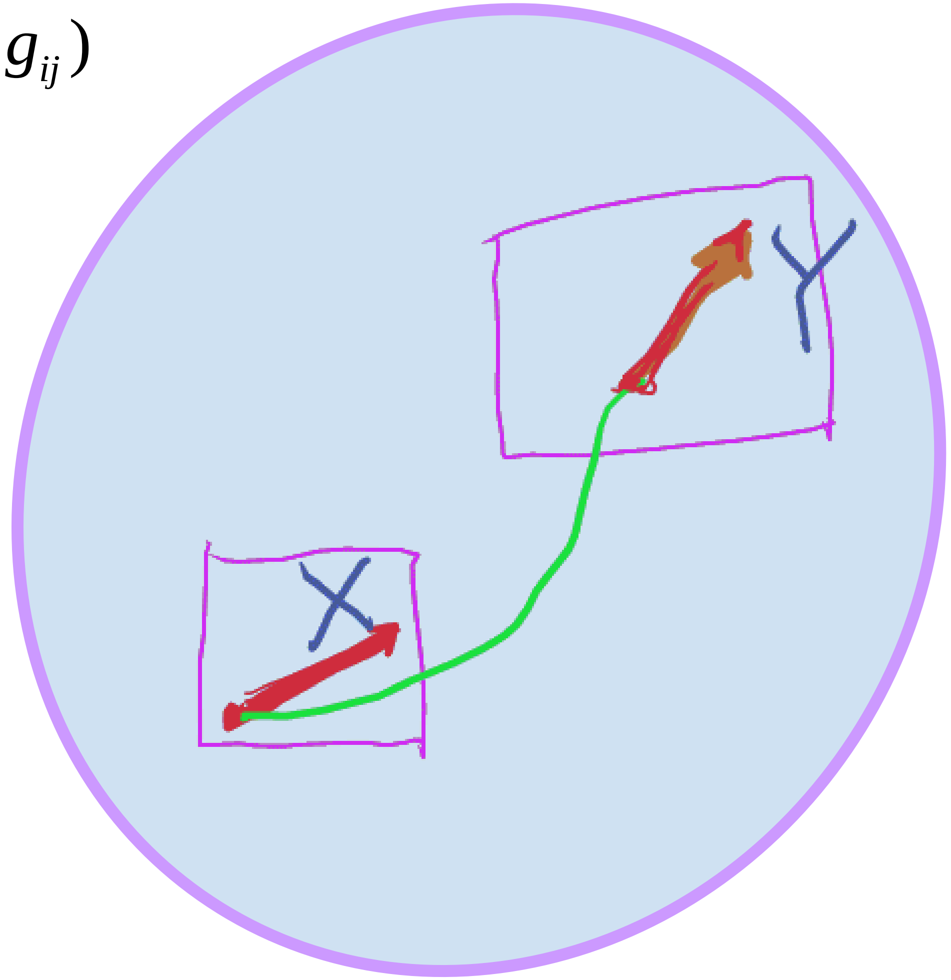
$\nabla_X Y$

$\Pi_c X = Y$

測地線　$\Pi\dot{X} = \dot{X}$　$X = X(t)$

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

最短距離：まっすぐ



二つのアファイン接続 (∇, ∇^*)

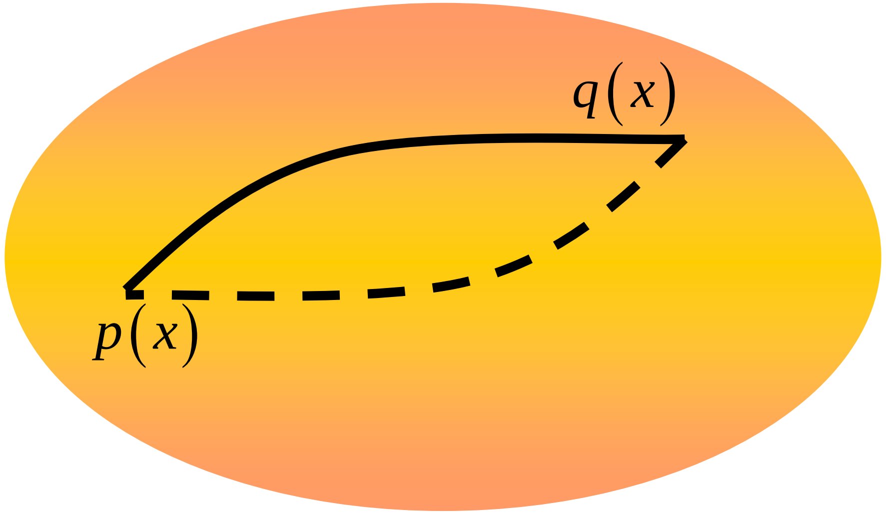
(Π, Π^*)

e-geodesic

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

m-geodesic

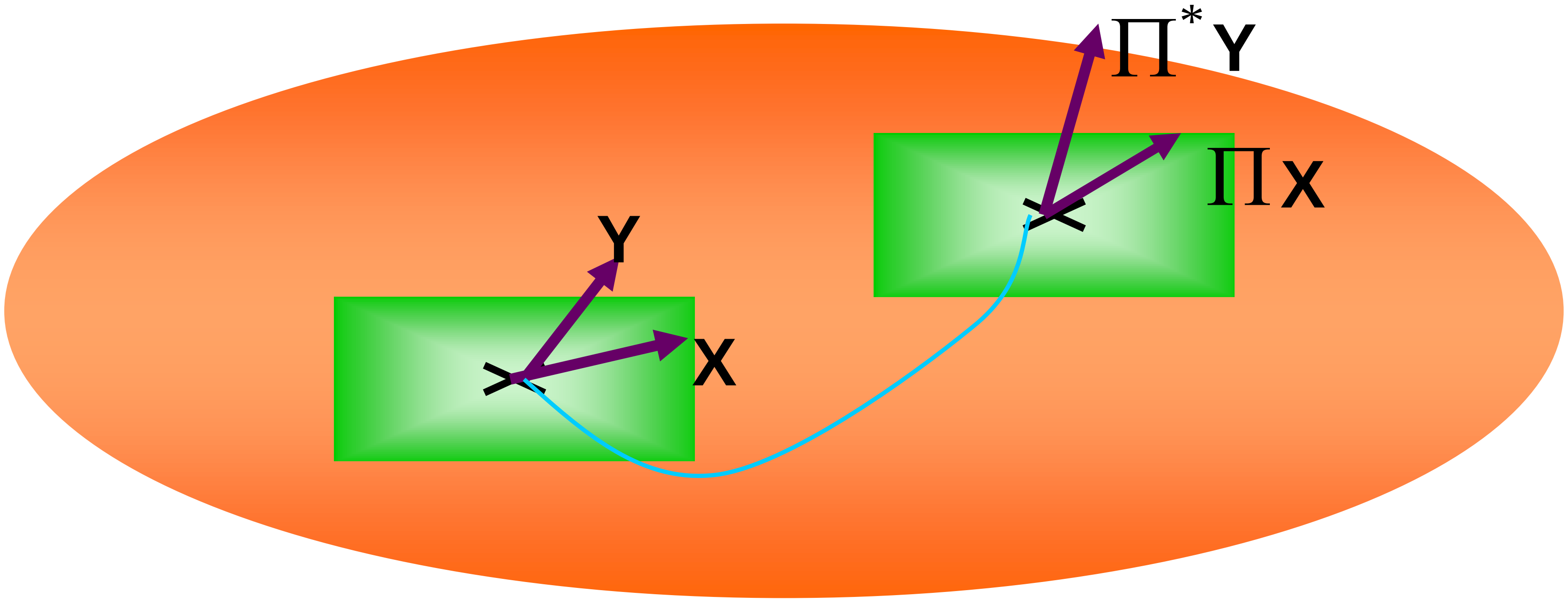
$$r(x,t) = tp(x) + (1-t)q(t)$$



双対接続:二つのアファイン接続

$$X\langle Y,Z\rangle = \langle \nabla_X Y, Z\rangle + \langle \nabla_X Z, Y\rangle$$

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



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

指数型分布族： 双対平坦空間

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

$\psi(\theta)$： convex function, free-energy

Gaussian:

$e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x \\ x^2 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta^1 \\ \theta^2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu^2}{\sigma^2} \end{pmatrix}$$

$$\theta \cdot x = -\frac{(x-\mu)^2}{2\sigma^2} + c$$

entropy ・ $-\varphi(\eta) = -\int p(x,\theta) \log p(x,\theta)\, dx$

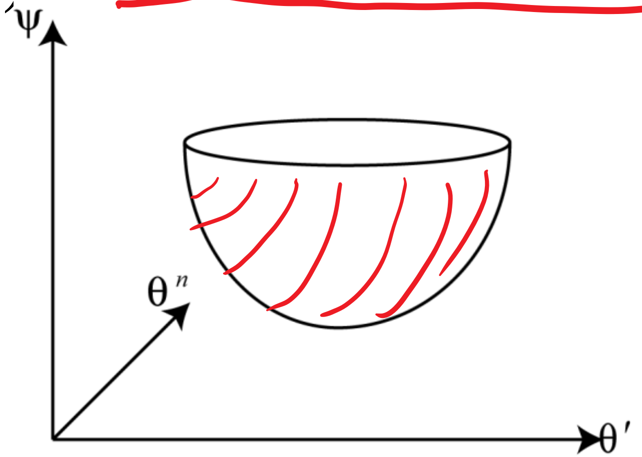
natural parameter : $\theta = \frac{\partial}{\partial \eta}\varphi(\eta)$

expectation parameter : $\eta = E[x] = \frac{\partial}{\partial \theta}\psi(\theta)$

凸関数、凸解析——双対平坦

S ： 座標系 $\boldsymbol{\theta} = \left(\theta^1, \theta^2, \ldots, \theta^n\right)$

$\psi\left(\boldsymbol{\theta}\right)$: 凸関数 function



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

negative entropy

energy

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

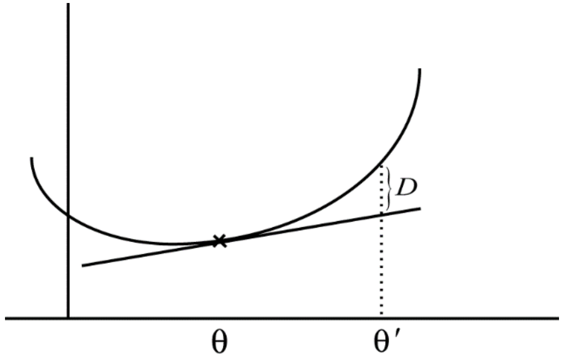
リーマン計量と平坦性 $\{S,\ \psi(\theta),\ \theta\}$

convex : $\tilde{\theta} = A\theta + c$

affine structure

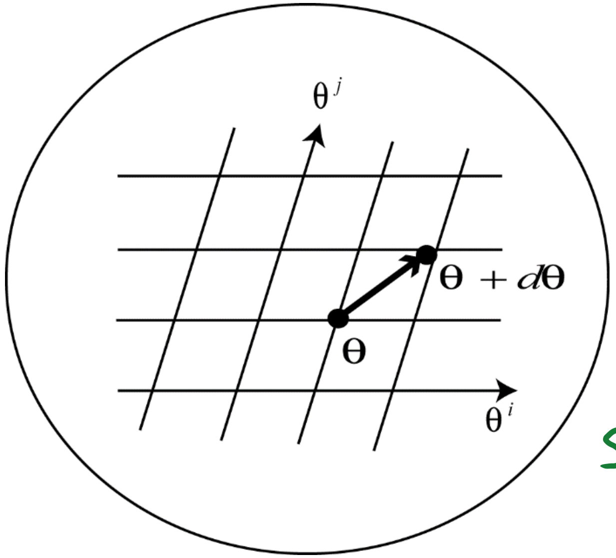
Bregman divergence

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



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

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



straight line

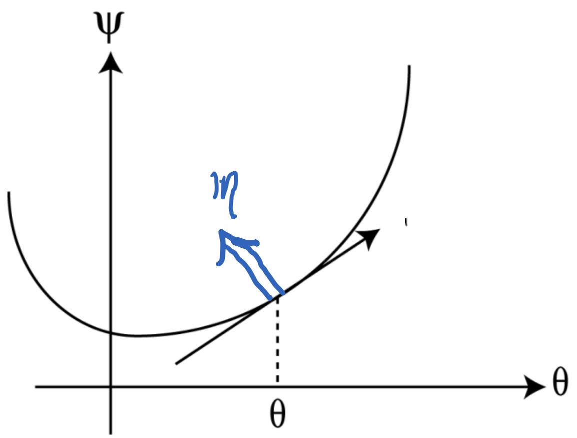
$\theta(t) = ta + b$

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

Legendre 変換

dual coordinates θ, η

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



$\psi(\theta)$ $\boldsymbol{\theta} \leftrightarrow \boldsymbol{\eta}$ $\varphi(\eta)$

one-to-one

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

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

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

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

: proof easy

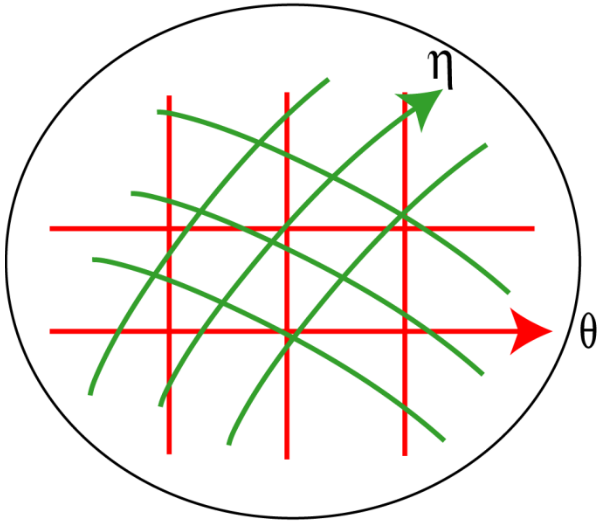
$$D(\boldsymbol{\theta}, \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}') - \boldsymbol{\theta} \cdot \boldsymbol{\eta}'$$

双対平坦空間の双対アファイン座標

$\theta = (\theta_1, \cdots, \theta_n)$

$\eta = (\eta_1, \cdots, \eta_n)$

$\eta = \eta(\theta) \longleftrightarrow \theta = \theta(\eta)$



one-to-one

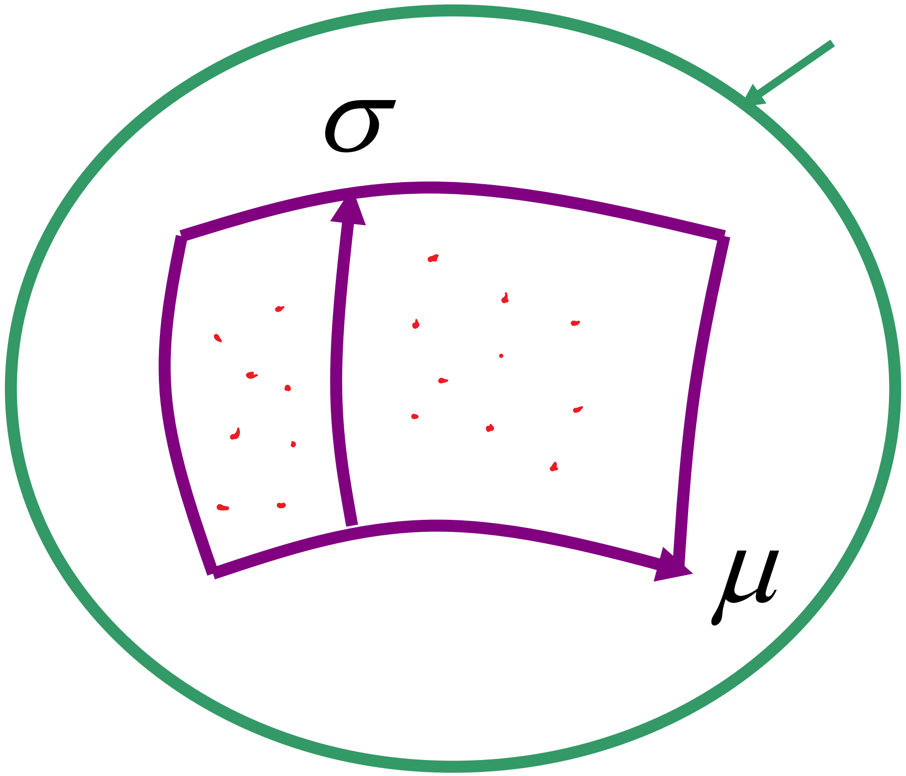
differentiable

Gaussian distributions

$\xi = (\mu, \sigma^2),$

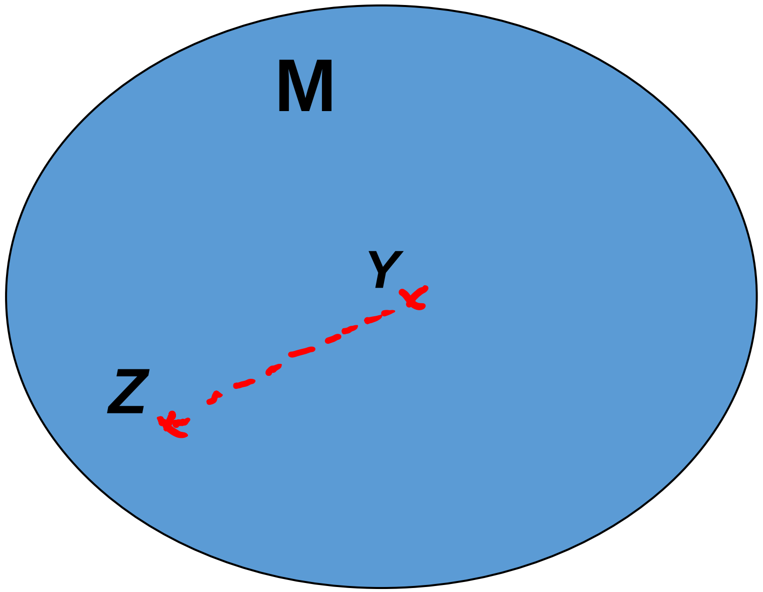
$\theta = \left(-\frac{1}{2\sigma^2}, \frac{\mu^2}{\sigma^2}\right),$

$\eta = (\mu, \mu^2 + \sigma^2)$



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

Divergence: $D[z:y]$



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

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

Not necessarily symmetric

D[z : y] = D[y : z]

$$D[\mathbf{z}:\mathbf{z}+d\mathbf{z}] = \frac{1}{2}\sum g_{ij}\, dz_i\, dz_j$$

positive-definite $G = (g_{ij})$

Taylor expansion

$$D(z : z + dz) = \frac{1}{2}\sum g_{ij}\, dz_i\, dz_j + \frac{1}{6}\sum K_{ijk}\, dz_i\, dz_j\, dz_k$$

$$+ \cdots$$

双対平坦空間

θ-coordinates $\leftrightarrow$ η-coordinates

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

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

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

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

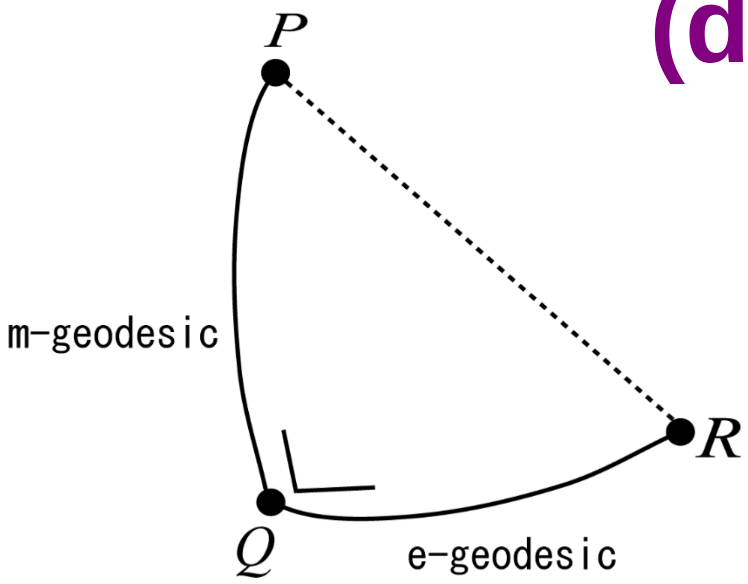
ψ : cumulant generating function

φ : negative entropy

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

拡張ピタゴラスの定理

(dually flat manifold)



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

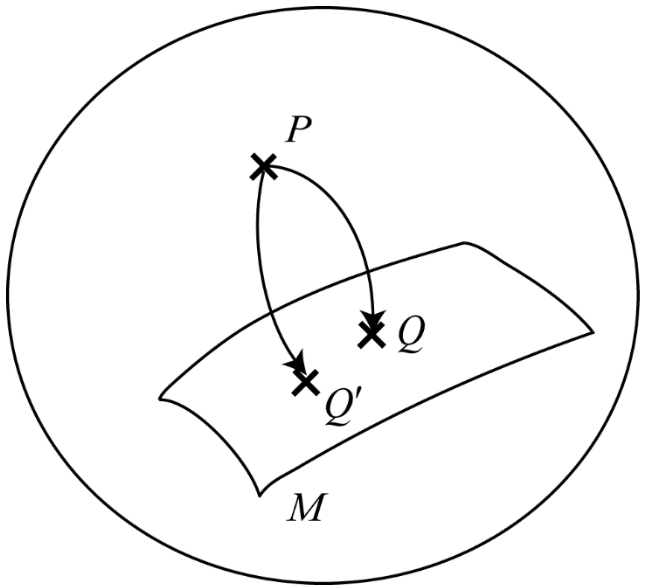
proof

ユークリッド空間：自己双対

$$\boldsymbol{\theta}=\boldsymbol{\eta}$$

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

射影定理



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

Q = m-射影　P から M

unique when M is e-flat

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

Q' = e-射影　P から M

unique when M is m-flat

双対平坦幾何

Convex function – Bregman divergence– exponential family – Dually flat Riemannian divergence

$$\psi(\theta) \Longrightarrow D_\psi[\theta : \theta'] \Longrightarrow \{\theta, \eta\}, \quad g = \nabla\nabla\psi$$

Dually flat R-manifold – convex function – canonical divergence KL-divergence

$$\{\theta, \eta\} \quad g = \frac{\partial \eta}{\partial \theta} \Longrightarrow \psi(\theta) \Longrightarrow D_\psi[\theta : \theta'] = D_{KL}[p(x) : q(x)]$$

一般にダイバージェンスは計量と接続を与える

リーマン計量

(Eguchi)

$$g_{ij}(\mathbf{z}) = \partial_i \partial_j D[\mathbf{z}:\mathbf{y}]_{|\mathbf{y}=\mathbf{z}} : \quad D[\mathbf{z}:\mathbf{y}] = \frac{1}{2} g_{ij}(\mathbf{z})(z_i - y_i)(z_j - y_j)$$

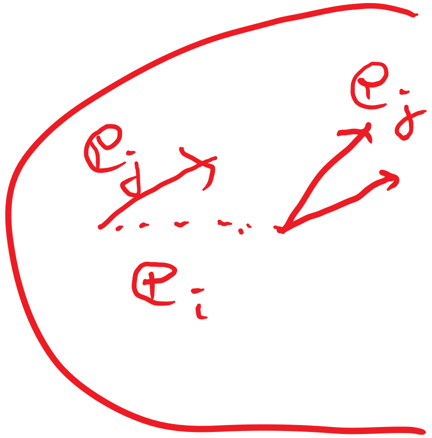
二つのアファイン接続

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

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

$$\Gamma_{ijk}(\mathbf{z}) = -\partial_i \partial_j \partial'_k D[\mathbf{z}:\mathbf{y}]_{|\mathbf{y}=\mathbf{z}} \qquad \partial_i = \frac{\partial}{\partial z_i}, \quad \partial'_i = \frac{\partial}{\partial y_i}$$

$$\Gamma^*_{ijk}(\mathbf{z}) = -\partial'_i \partial'_j \partial_k D[\mathbf{z}:\mathbf{y}]_{|\mathbf{y}=\mathbf{z}} \qquad T = \Gamma^* - \Gamma$$



不変でないダイバージェンス

Wasserstein距離

q-ダイバージェンス ↔α-ダイバージェンス

$$D_\alpha[p:q] = \frac{4}{1-\alpha^2}\sum\left(1 - p_i^{\frac{1-\alpha}{2}} q_i^{\frac{1+\alpha}{2}}\right)$$

$$D_q[p:q] = \frac{4}{1-q}\sum\left(1 - p_i^{q} q_i^{1-q}\right); \quad q = 2\alpha - 1$$

projectively-dually flat

divergence

(n > 1)

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

invariance

dually flat space

invariant divergence

F-divergence
Fisher inf metric
Alpha connection

KL-divergence

Flat divergence

convex functions
Bregman

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

$q-$esponential family

$$\log_q(u)=\frac{1}{1-q}(u^{1-q}-1)$$

$$\log\{p(x,\theta)\}=\theta\cdot\mathbf{x}-\psi_q(\theta)$$

$\psi_q(\theta):$ convex function

双対平坦空間(非不変)
Pythagorusの定理

$$D_q{}^*(p:q)=\frac{1}{h_q(\theta)}D_q{}^*(p:q):\quad h_q=\sum p_i^q$$

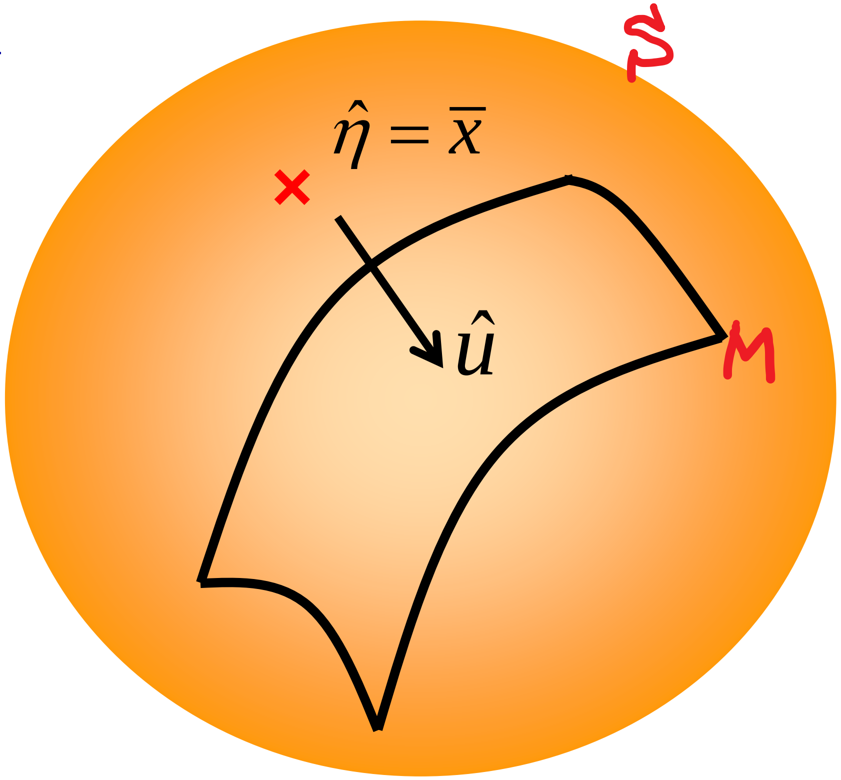
統計学への応用： 曲指数型分布族:

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

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

$$p(D,u) = \exp\left\{\theta(u)\cdot \bar{\mathbf{x}} - \psi\left(\theta(u)\right)\right\}$$

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



統計学への応用

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

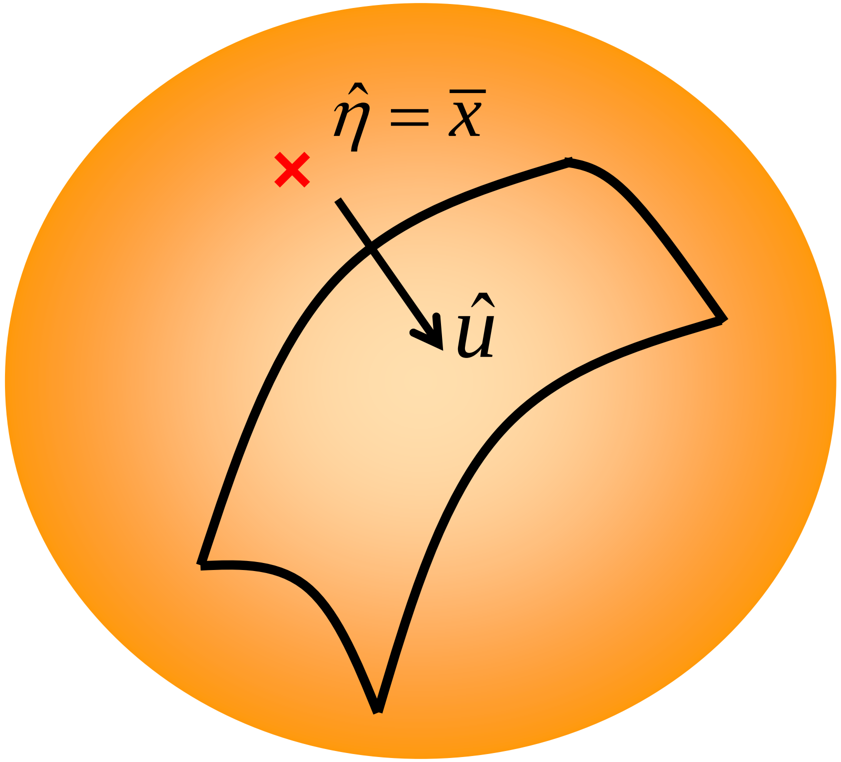
$p(x,u)$ □ $x_{1,} x_2, \ldots x_n$

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

$\hat{u}(x_1, \ldots, x_n)$　　:**推定**

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

$H_0 : u = u_0$　　:**検定**



推定誤差

$$e = \hat{\eta} - \eta = \frac{1}{N}\sum x_i - E[x], \quad \tilde{e} = \sqrt{N}\, e$$

($\bar{x}$ → $\frac{1}{N}\sum x_i$, η → $E[x]$)

Cramer-Rao bound

$$E[e_i e_j] \geq \frac{1}{N} g_{ij}$$

$$E[\tilde{e}_i] = 0$$

$$E[\tilde{e}_i \tilde{e}_j] = g_{ij}$$

$$E[\tilde{e}_i \tilde{e}_j \tilde{e}_k] = \frac{1}{\sqrt{N}} T_{ijk}$$

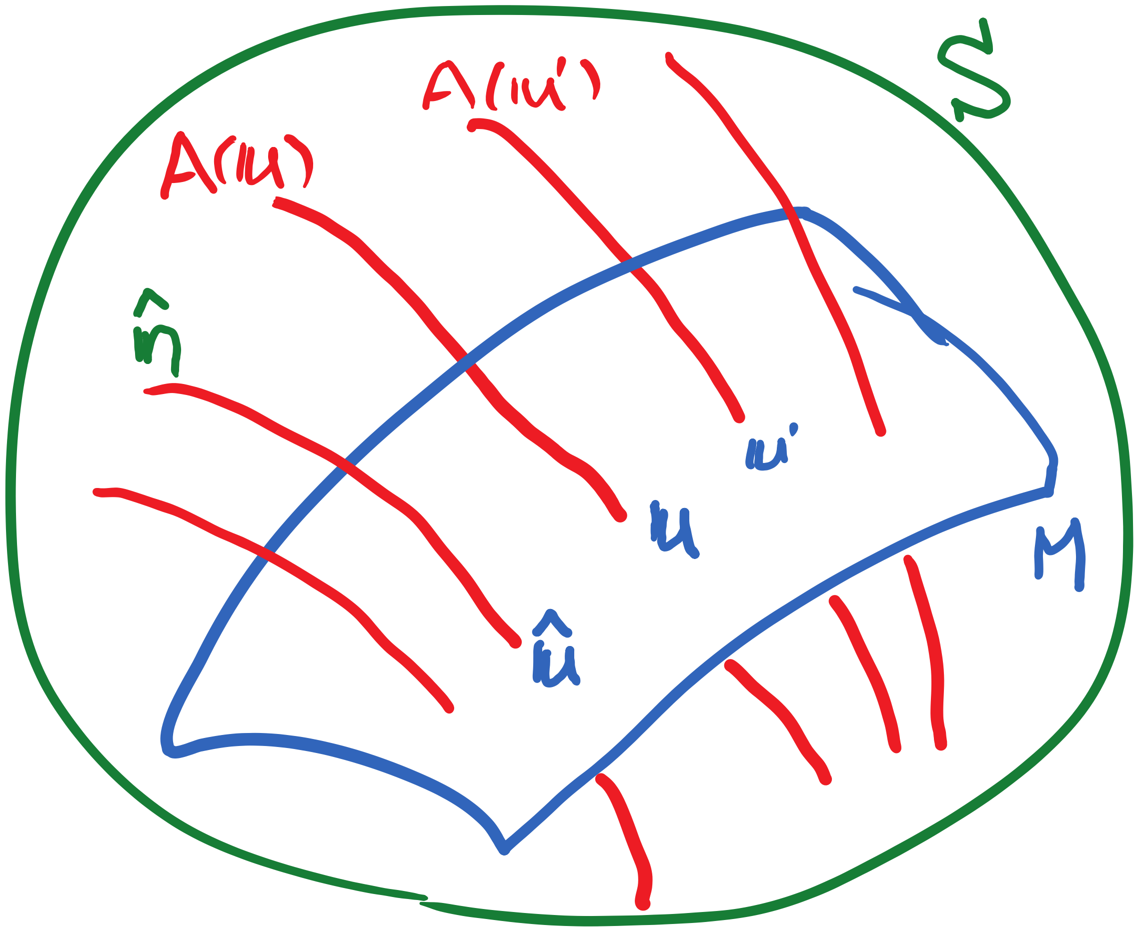
$$E[\tilde{e}_i \tilde{e}_j \tilde{e}_k \tilde{e}_l] = \frac{1}{N} S_{ijkl}$$

(g_{ij}: ∇∇ψ, T_{ijk}: ∇∇∇ψ, S_{ijkl}: ∇∇∇∇ψ)

補助多様体族

推定量 --- $\hat{u} = f(\hat{\eta})$

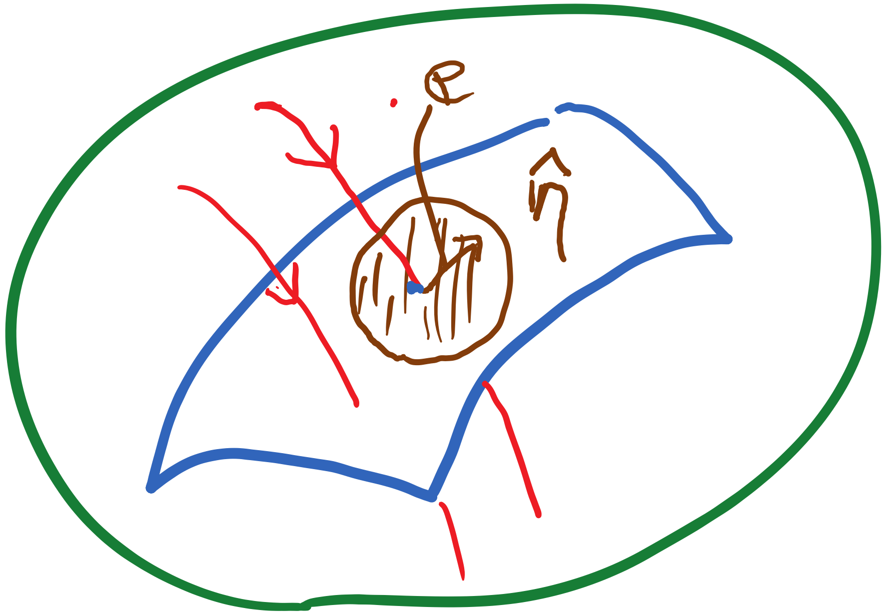
Ancillary family $A(u)$



最尤推定は一致性を持ち有効

$\hat{u}_{mle}$: $\min KL[\hat{\eta} : \eta(u)]$

m-projection of $\hat{\eta}$ to M



Efficient estimator --- orthogonal projection

誤差の高次漸近理論

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

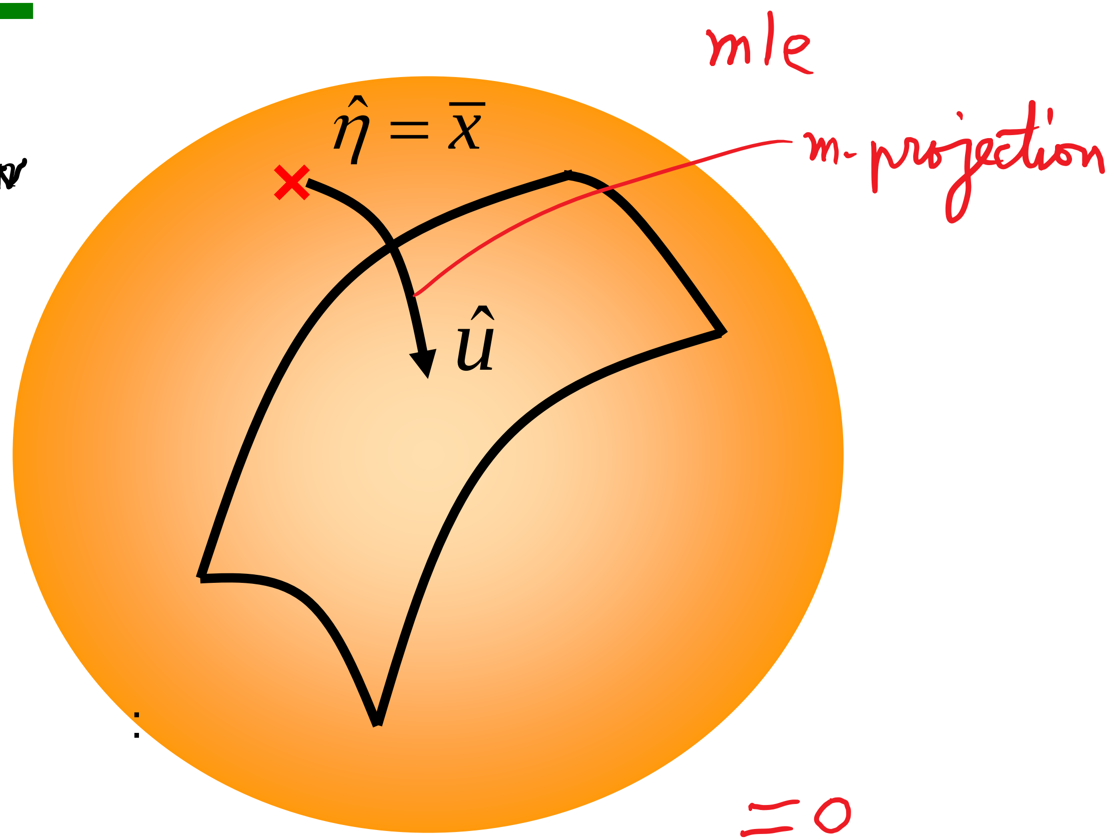
$\hat{\mathrm{u}} = \mathrm{u}(x_1, \cdots, x_n)$

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

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

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

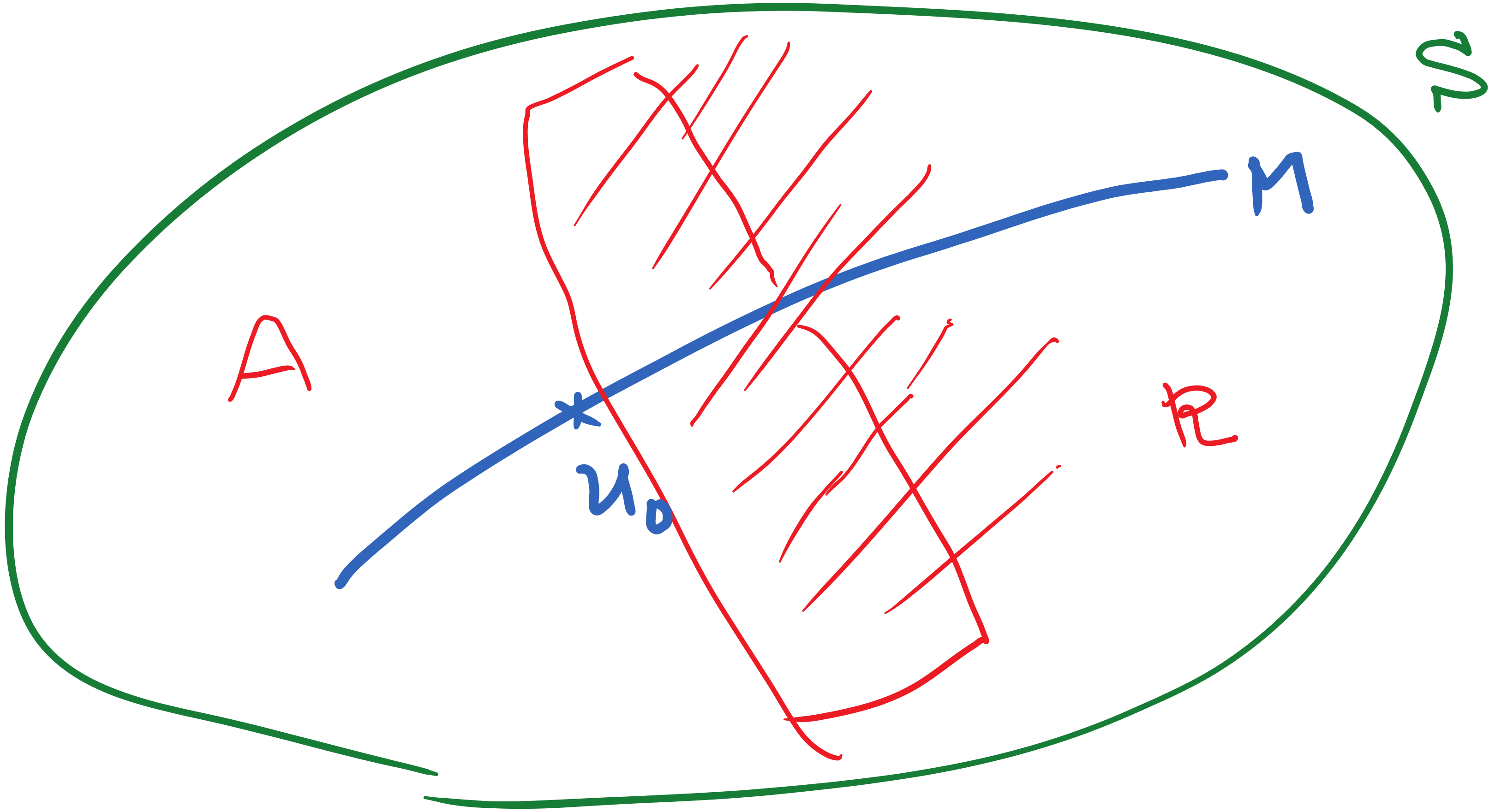
$G_2 = H_M^{(e)^2} + H_A^{(m)^2} + \Gamma^{(m)^2}$ quadratic approximation



仮説検定

$H_0 : u = u_0$

$H_1 : u = u_1$



Neyman-Scott問題：　局外母数のある問題

Estimation with nuisance parameter

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

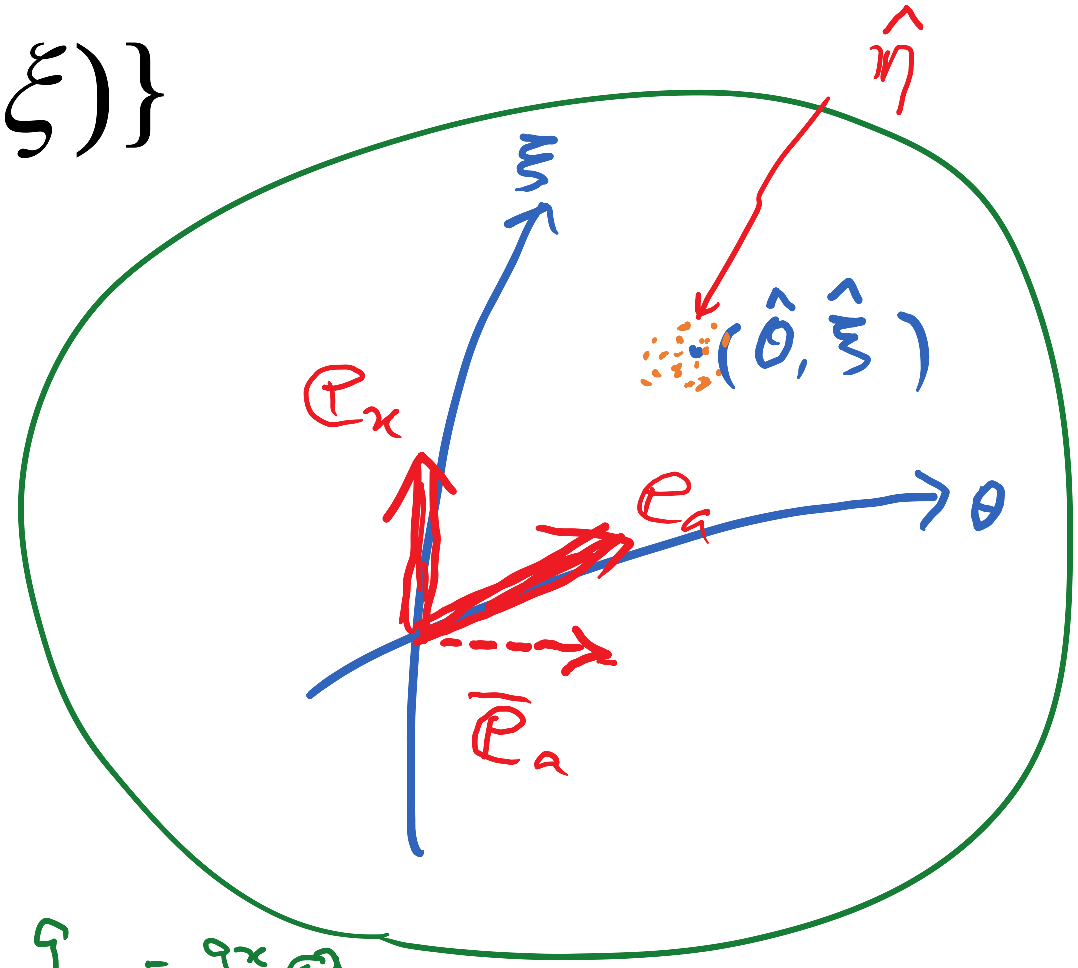
Efficient score

$$\partial_a l = \frac{\partial}{\partial \theta^a} \log p(x, \theta, \xi) = e_a$$

$$\partial_\kappa l = \frac{\partial}{\partial \xi^\kappa} \log p(x, \theta, \xi) = e_\kappa$$

$$\bar{\partial}_a l = \partial_a l - g_{a\kappa} g^{\kappa\lambda} \partial_\lambda l$$

$$\langle e_a, e_b \rangle = \frac{1}{N} \left(\bar{g}_{ab}\right)^{-1}$$



$$\bar{g}_{ab} = g_{ab} - g_a^{\kappa} g_b^{\lambda} g_{\kappa\lambda}$$

Neyman-Scott問題：　無限個の局外母数

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

$$x_1 \sim p(x, \theta, \xi_1)$$

$$x_2 \sim p(x, \theta, \xi_2)$$

$$\cdots$$

$$x_N \sim p(x, \theta, \xi_N)$$

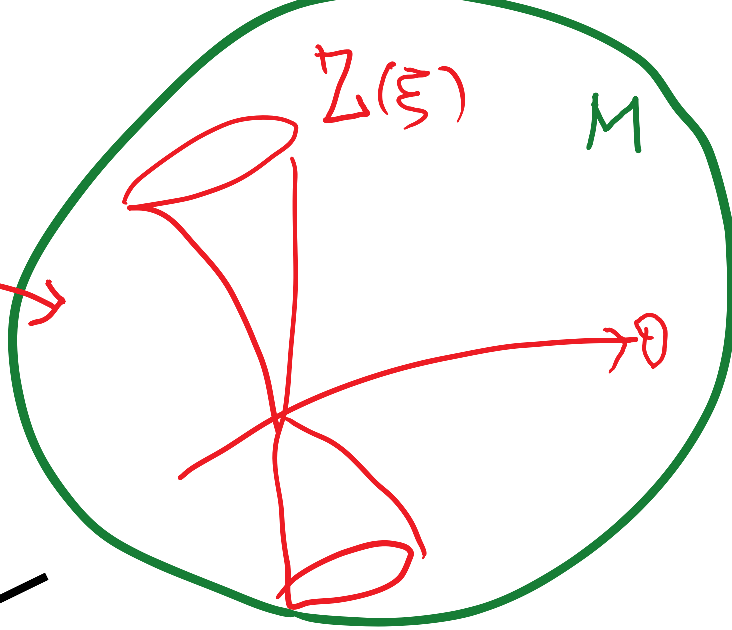
θ : parameter of interest

ξ : nuisance parameter

Semiparametric 統計モデル:比例定数の推定

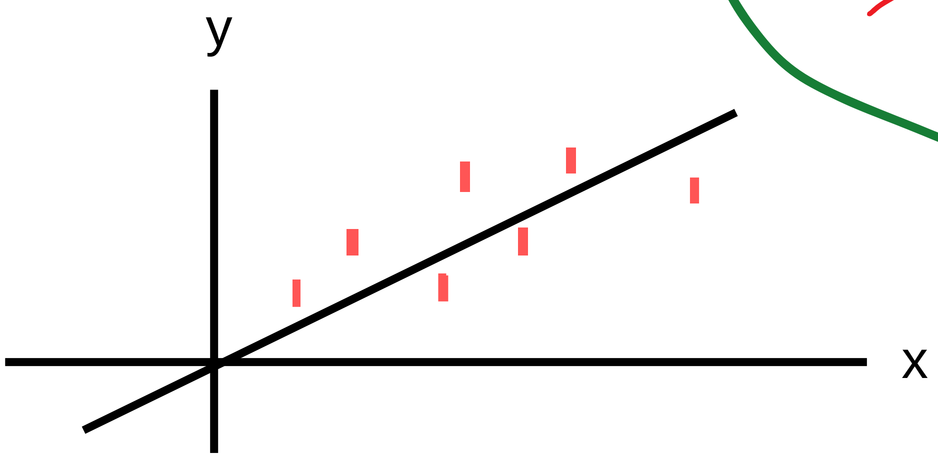
$$M = \{ p(x, \theta, Z) \}$$

$$\xi \sim Z(\xi)$$ 関数自由度の未知母数



linear relation $\mathbf{x} = (x, y)$

$$y = \theta x$$



$$\begin{cases} y_i = \theta \xi_i + \varepsilon_i \\ x_i = \xi_i + \varepsilon_i' \end{cases} \qquad p(x, y; \theta, Z) = \int p(x, y; \xi, \theta) Z(\xi)\, d\xi$$

mle, least square, total least square

統計 Model

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

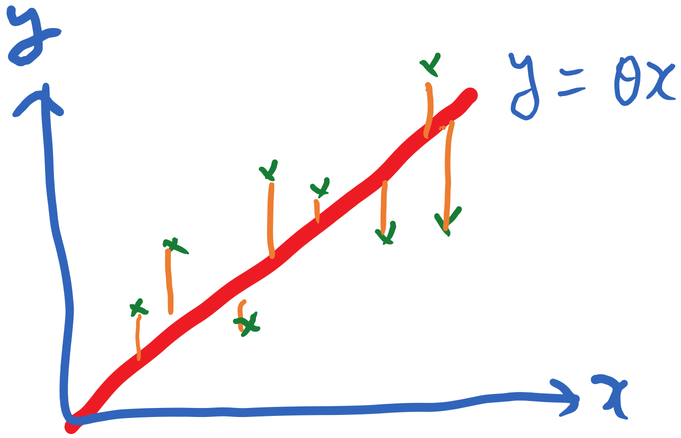
$$\prod p\left(x_i, y_i \middle| \theta, \xi_i\right) \; : \; \theta, \xi_1, \cdots, \xi_n$$

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

——— semiparametric

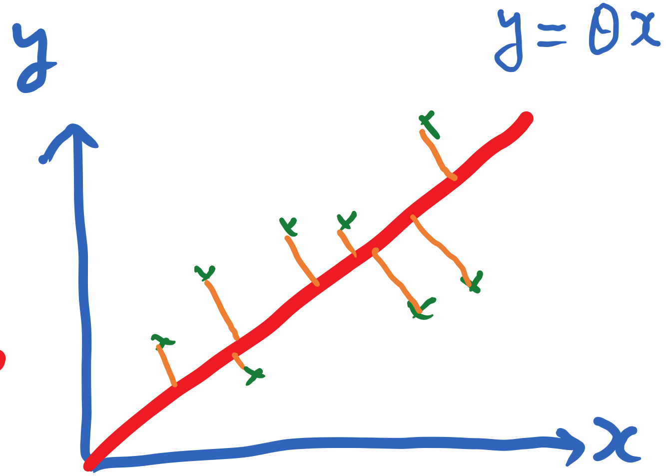
最小二乗法は良いか？

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



$$\frac{1}{n}\sum \frac{y_i}{x_i}$$: averag

$$\frac{\sum y_i}{\sum x_i}$$: gross average



mle, TLS

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

Neyman-Scott

セミパラ統計モデル

$$x_1, x_2, \cdots \square\ p(x, \theta, Z)$$

推定関数

$$f(x, \theta) \qquad E_{\theta, Z}\left[f(x, \theta)\right] = 0$$

$$E_{\theta', Z}\left[f(x, \theta)\right] \neq 0 \qquad \theta' \neq \theta$$

推定方程式

$$\sum f(x_i, \theta) = 0 \quad \Rightarrow \hat{\theta}$$

$$\frac{1}{N}\sum f(x_i, \theta) \Rightarrow E_{\theta, Z}\left[f(x, \theta)\right]$$

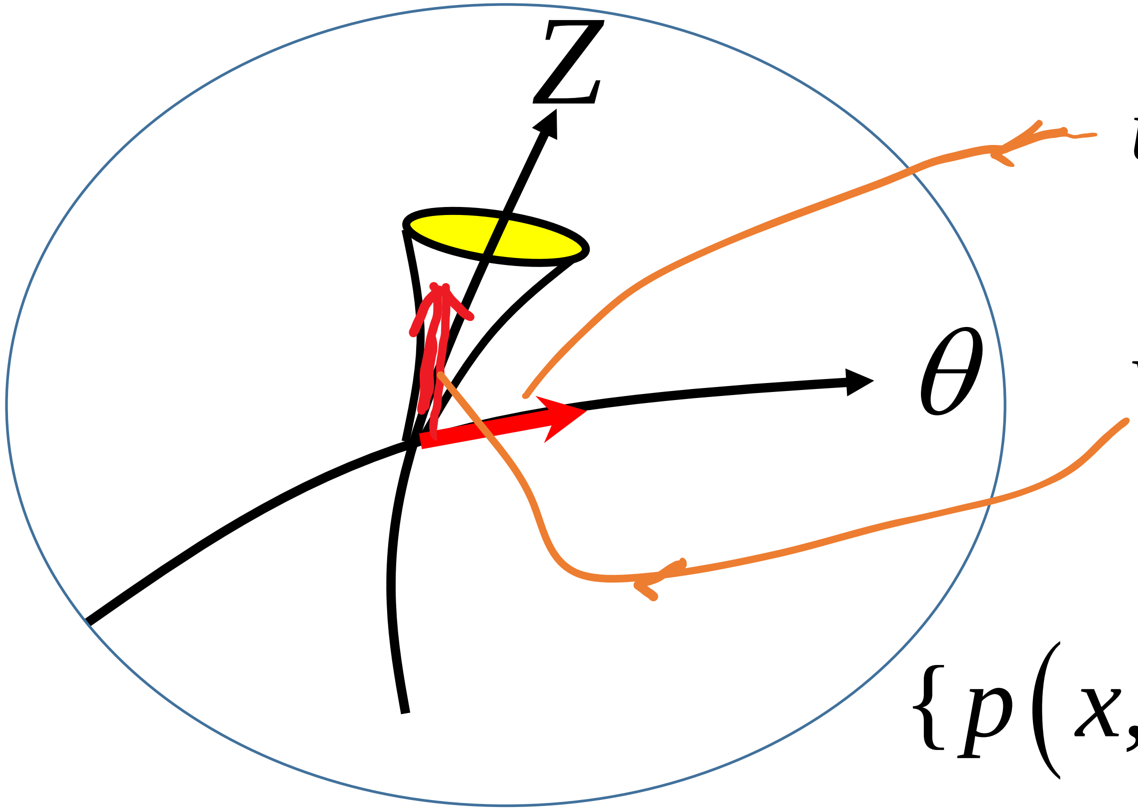
推定関数

$$E_{\theta,Z}\left[f\left(x,\theta\right)\right]=0: \text{ unbiased}$$

$$\sum_{i=1}^{n} f\left(x_i,\hat{\theta}\right)=0: \ \hat{\theta}=\theta+e$$

$$E\left[\left(\hat{\theta}-\theta\right)^2\right]=\frac{1}{n}\frac{E\left[f^2\right]}{E\left[\left(\partial_\theta f\right)^2\right]}$$

Fiber Bundle



$$u(x, y; \theta, Z) = \frac{\partial}{\partial \theta} \log p$$

$$v(x, y; \theta, Z) = \frac{\partial}{\partial Z} \log p$$

$$\{p(x, y | \theta, Z)\}$$

Estimating Function

$$f(x,\theta)$$

$$e-\text{invariant}: E_{\theta,Z}\left[f(x,\theta)\right]=0$$

$$\prod_{z}^{e\,z'} f(x,\theta)=f$$

$$T_\theta = T_\theta^I \oplus T_\theta^N \oplus T_\theta^A$$

$$\int p(x,\theta,\xi)\,Z(\xi)\,f(x,\theta)\,dx\,d\xi$$

$$m-\text{orthogonality}: \langle v, f\rangle = 0$$

$$\left\langle \prod_{z}^{m\,z'} v, f\right\rangle = 0$$

$$\langle \delta Z, f\rangle = 0$$

$u^I(x,\theta,z)$: optimal estimating function

Efficient Score

$$\partial_\theta l = \partial_\theta \log p(x, \theta, \xi)$$

$$\bar{\partial}_\theta l = \partial_\theta l - g_{a\varkappa}\, g^{\varkappa\lambda}\, \partial_{\xi^\lambda} l$$

$$\dot{l}^E(x, \theta, Z) = \int \bar{\partial}_\theta l(x, \theta, \xi)\, Z(\xi)\, d\xi$$

orthogonal

$$f(x, \theta) = \dot{l}^E(x, \theta, Z_0) + a(x)$$

$$\sum f\left(x_i, y_i; \theta\right) = 0$$

$$f\left(x, y; \theta\right) = \left(x + \theta y + c\right)\left(y - \theta x\right)$$

$$c = \frac{\bar{\xi}\sigma^2}{\overline{\xi^2} - \left(\bar{\xi}\right)^2} \qquad \left(\begin{array}{l} \bar{\xi} = 1 \\ \overline{\xi^2} = 2 \end{array}\right.$$

$$c = 0: \quad V = \frac{1}{n}\frac{\left(2 + \sigma^2\right)\sigma^2}{4} \qquad : \frac{3}{4}$$

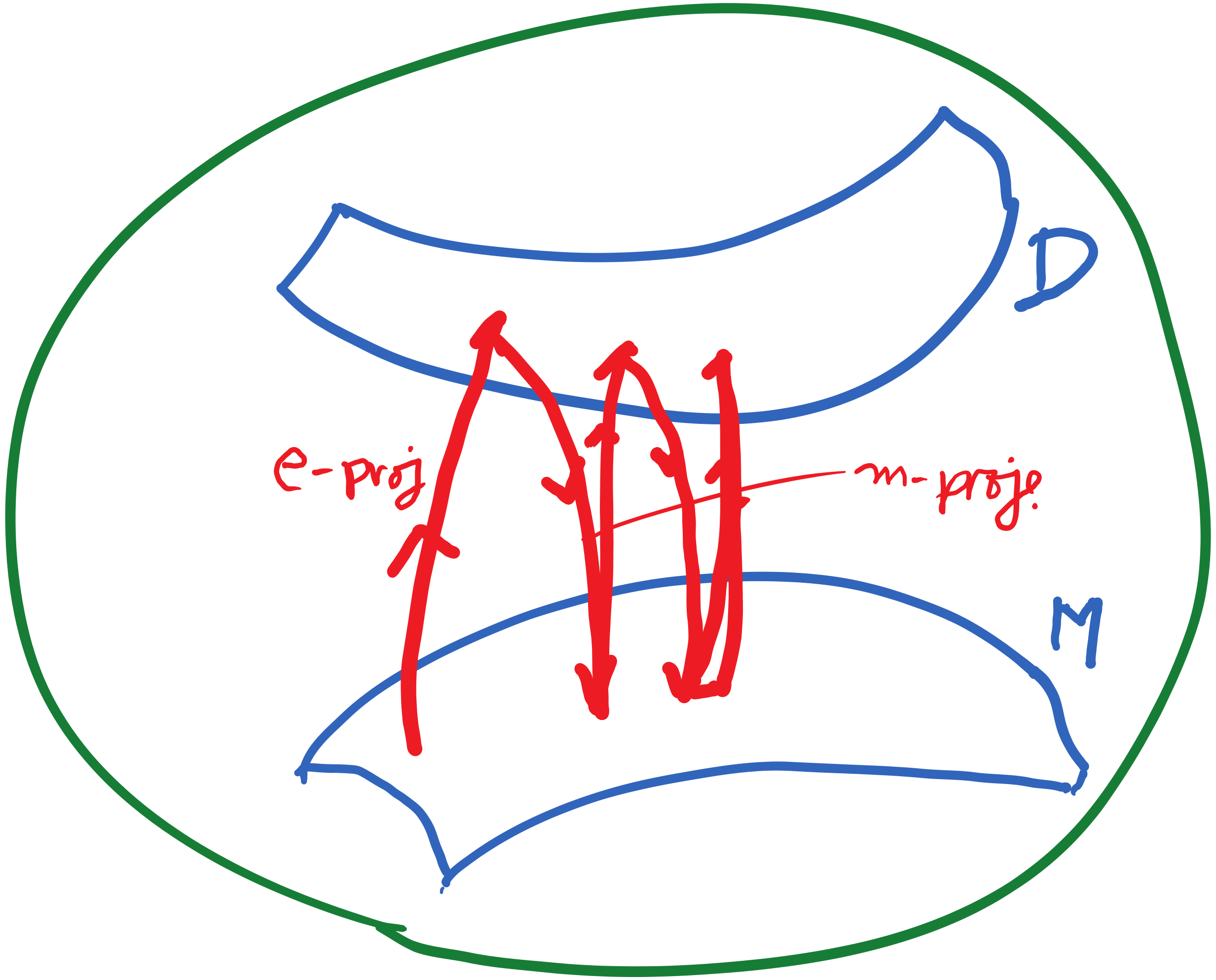
$$c = 1: \quad V = \frac{1}{n}\left(1 - \frac{1}{\sigma^2 + 2}\right)\sigma^2 \qquad : \frac{2}{3}$$

$$c = \infty: \quad v = \frac{1}{n}\sigma^2 \qquad : 1$$

em-algorithm EM-algorithm
Variational Bayes

$$\min D_{KL}[q(x) : p(x)]$$

$$q(x) \in D, \quad p(x) \in M$$



EM algorithm

hidden variables

observe — hidden

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

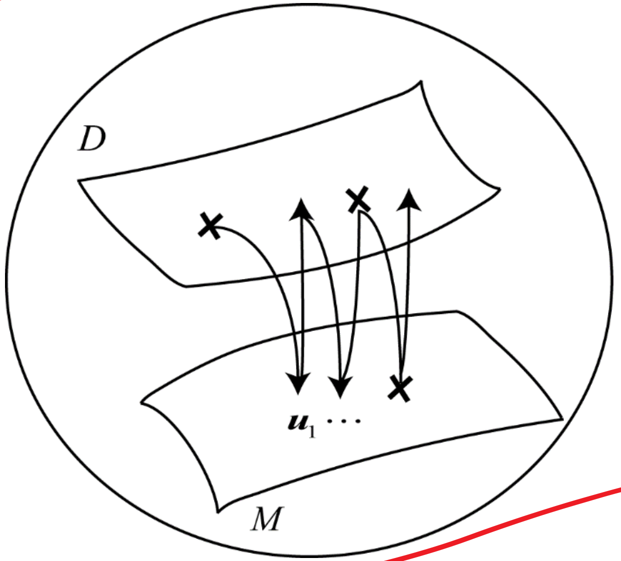
$$D = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$$

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

$$D_M = \{q(\boldsymbol{x}, \boldsymbol{y}) \mid q(\boldsymbol{x}) = q_D(\boldsymbol{x})\}$$

$$q_D = \frac{1}{N} \sum \delta(x - x_i)$$

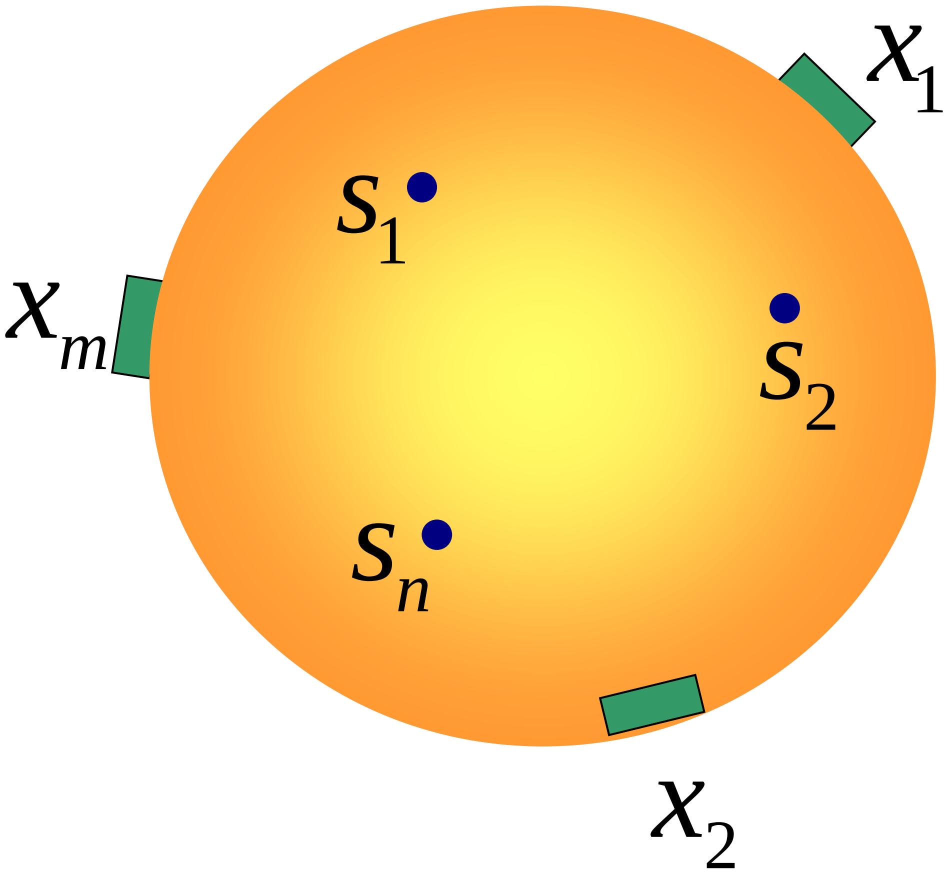
$$q(x, y) = q_D(x)\, r(y \mid x)$$



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

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

信号の混合と分解



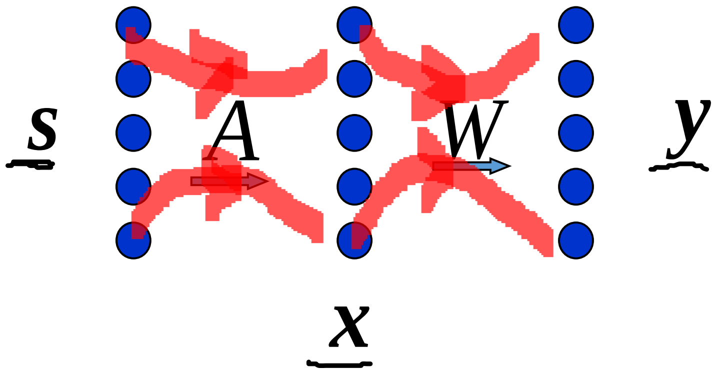
生体情報解析

カクテルパーティ効果

移動体通信

画像解析

独立成分分析

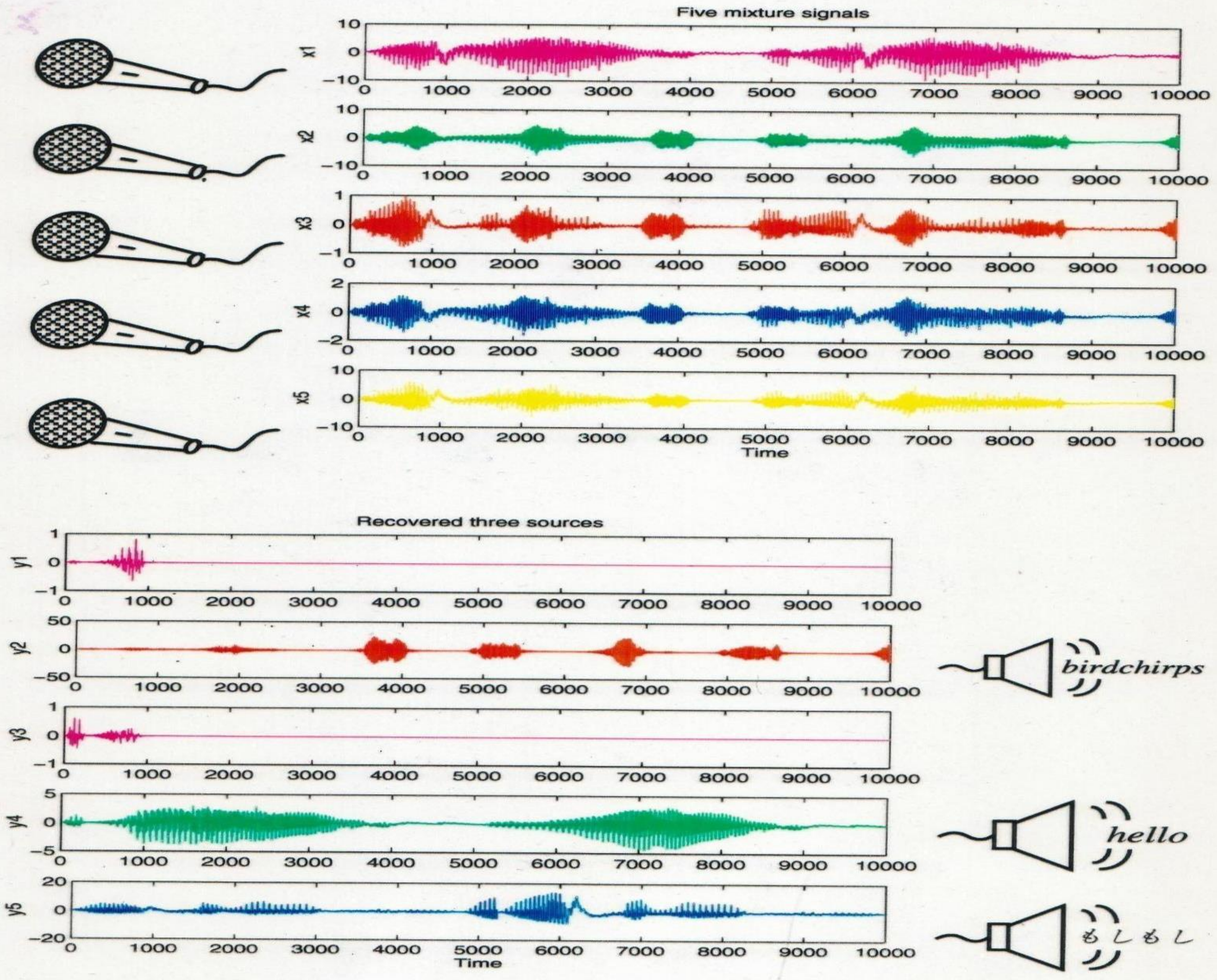


$$\mathbf{x} = A\mathbf{s} \qquad x_i = \sum A_{ij} s_j$$

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

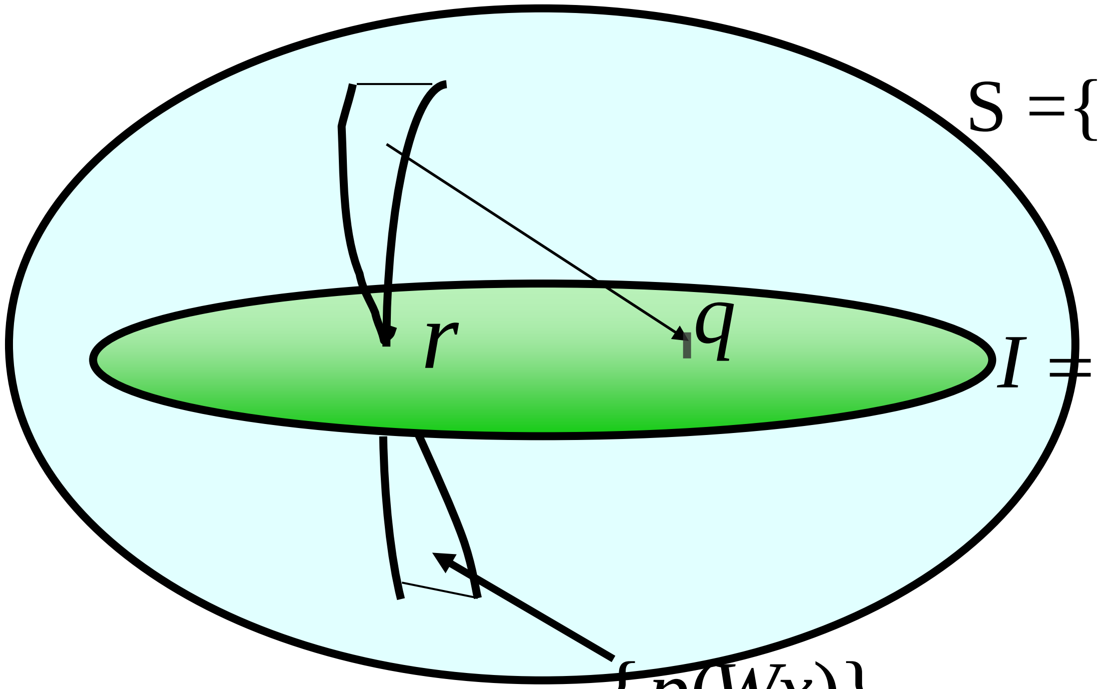
観測信号：　x(1), x(2), …, x(t)
復元信号：　s(1), s(2),　…, s(t)

Cocktail party experiment



- 5 microphones (sensors) and only 3 speakers

情報幾何による評価関数



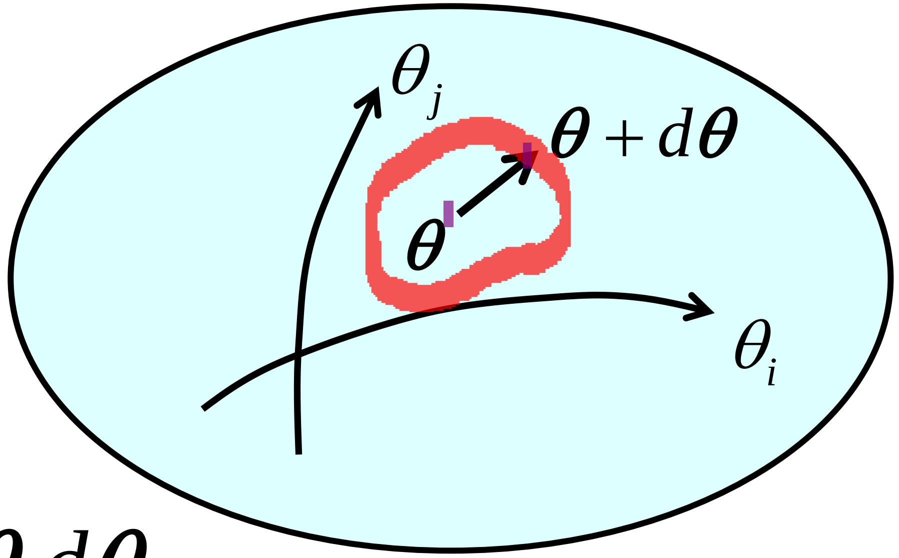
$$l(\mathbf{W}) = KL[p(\mathbf{y};\mathbf{W}) : q(\mathbf{y})]$$

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

行列Wの空間：Gl(n)
リーマン空間

$$\theta \to \mathbf{W}$$

$$ds^2 = |d\boldsymbol{\theta}|^2$$
$$= \sum g_{ij}(\boldsymbol{\theta}) d\boldsymbol{\theta}_i d\boldsymbol{\theta}_j$$
$$= d\boldsymbol{\theta}^T G(\boldsymbol{\theta}) d\boldsymbol{\theta}$$



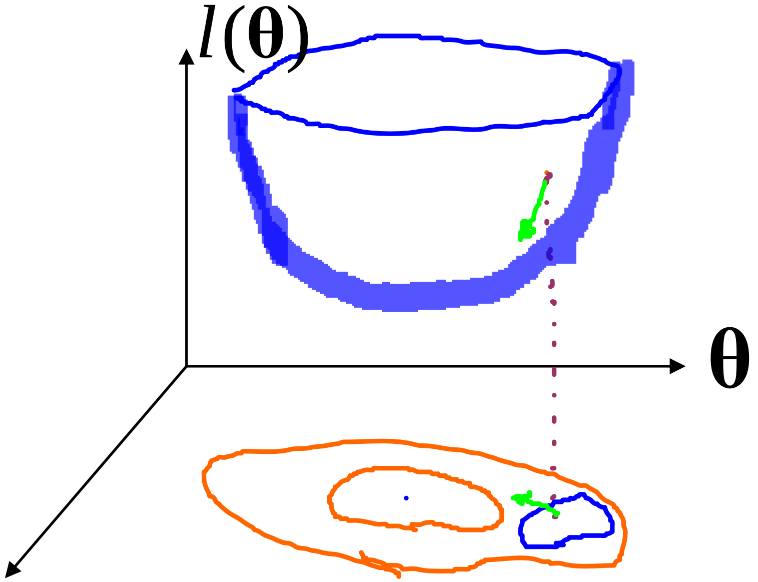
Euclid: G= I

自然勾配（Natural Gradient）

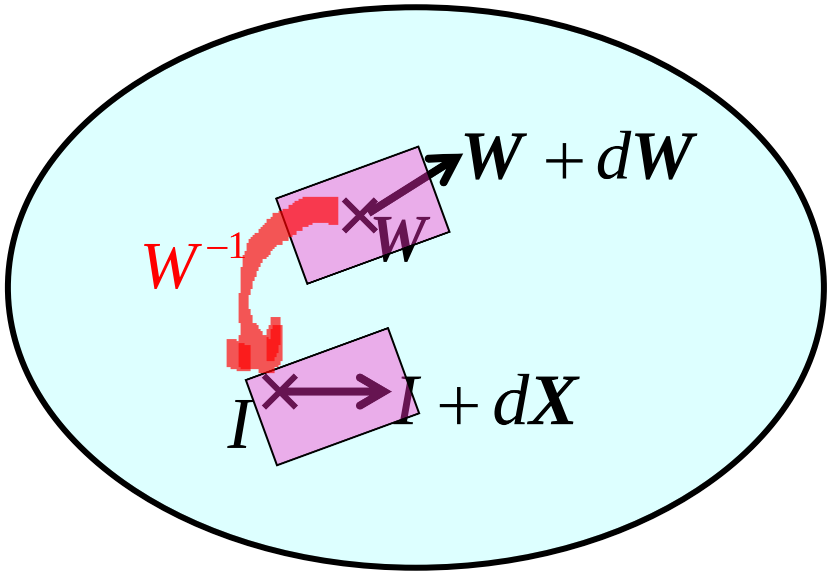
max $\quad dl = l(\boldsymbol{\theta} + d\boldsymbol{\theta}) - l(\boldsymbol{\theta})$

$$|d\boldsymbol{\theta}|^2 = \varepsilon$$

$$\tilde{\nabla} l = G^{-1}(\boldsymbol{\theta}) \nabla l$$



行列の空間: Lie群



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

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

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

dX :   **non-holonomic basis**

自然勾配

$$\Delta \boldsymbol{W} = -\eta G^{-1} \frac{\partial l\left(\boldsymbol{y}, \boldsymbol{W}\right)}{\partial \boldsymbol{W}}$$

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

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

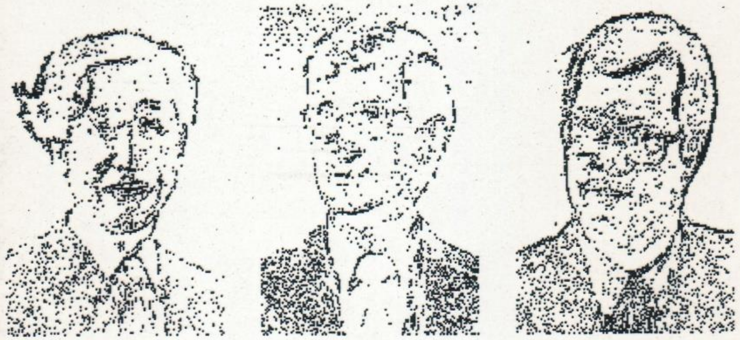
ICAから派生したもの

$$x = As$$

非負行列分解
スパース信号解析

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

(b) Two edge image mixtures



(c) Reconstructed binary edge images (after reversion)

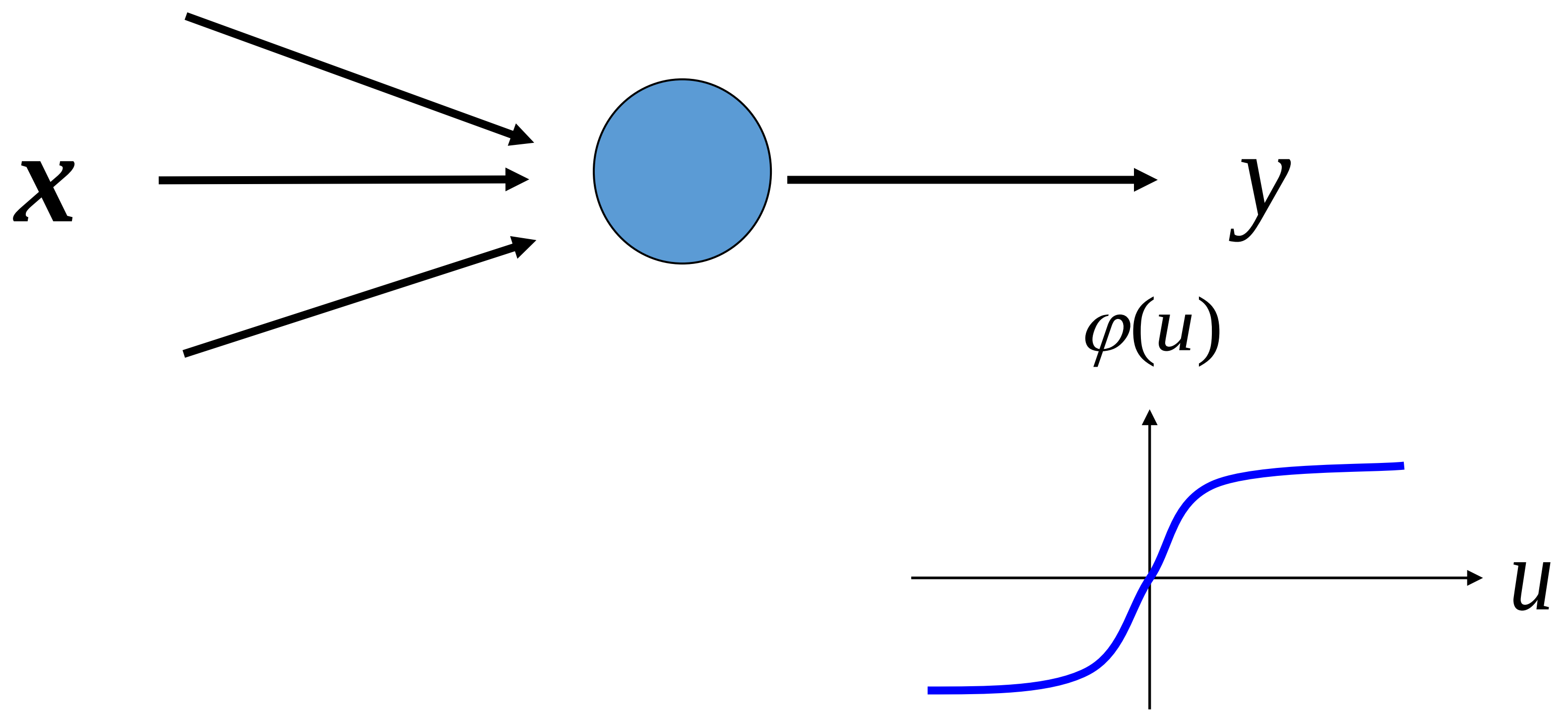
Fig. 5: Example of edge image image reconstruction: (a) the three binary edge images (reverse image copies are supplied for processing) , (b) their two mixtures, (c) the three extracted edge images (after reversion).

多層パーセプトロンの情報幾何

Natural Gradient and Singularities

数理ニューロン

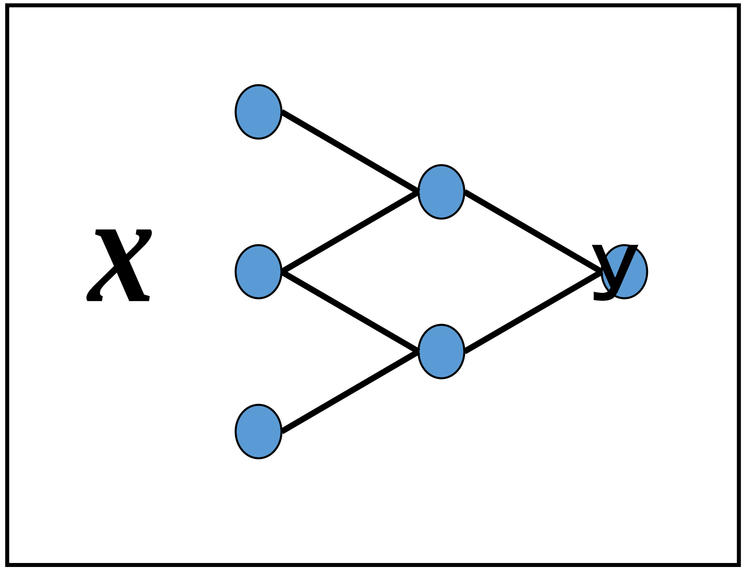
$$y = \varphi\left(\sum w_i x_i - h\right) = \varphi(\mathbf{w} \cdot \mathbf{x})$$



多層パーセプトロン

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

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



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

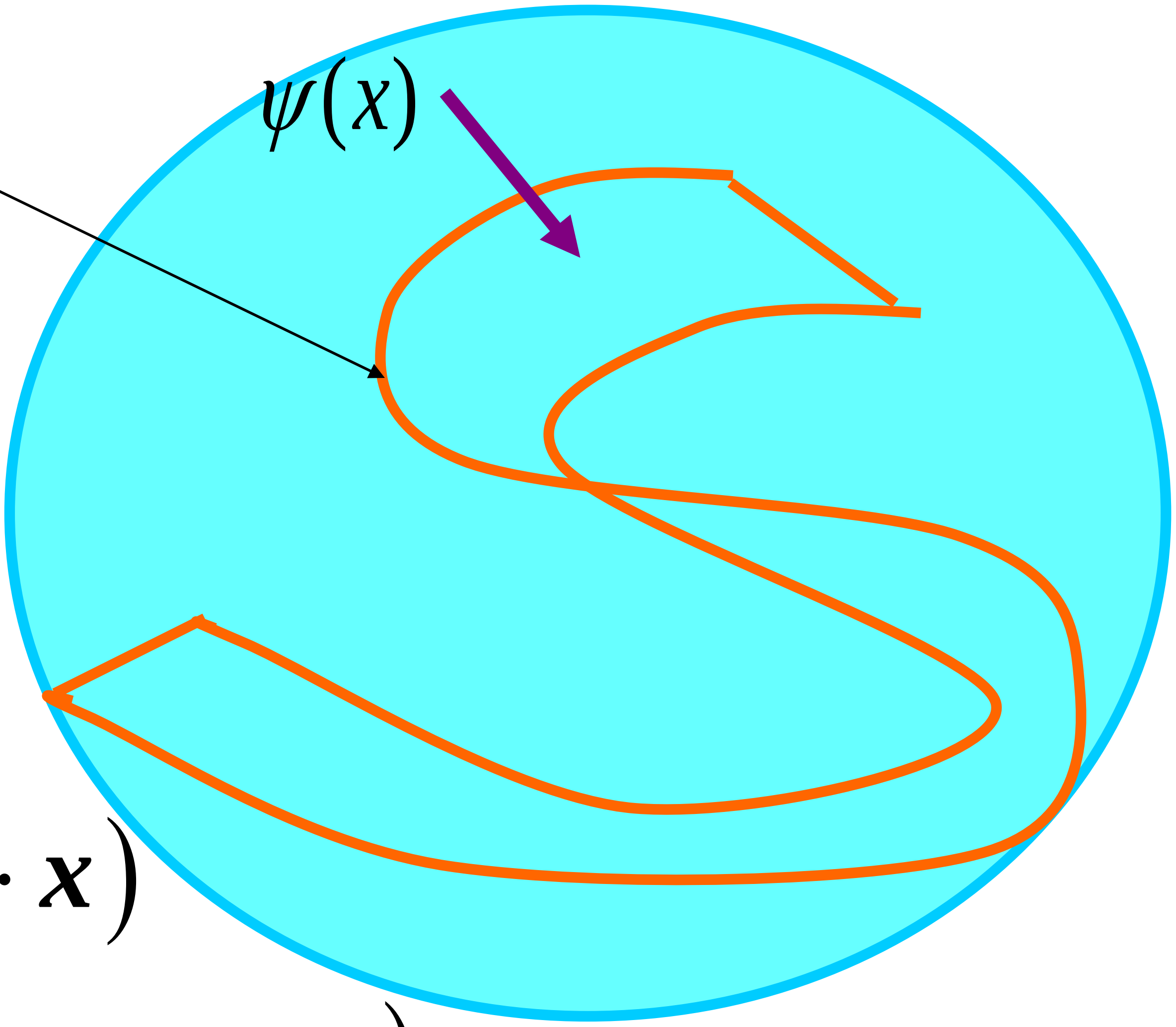
$$f(\boldsymbol{x},\boldsymbol{\theta}) = \sum v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x})$$

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

多層パーセプトロンと神経多様体

neuromanifold

space of functions



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

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

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

例題からの学習

$$\psi(\boldsymbol{x}) \approx f\left(x, \hat{\theta}\right)$$

多数の例題 $\cdots$ $D = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$

learning ; estimation

Backpropagation ---確率降下学習

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

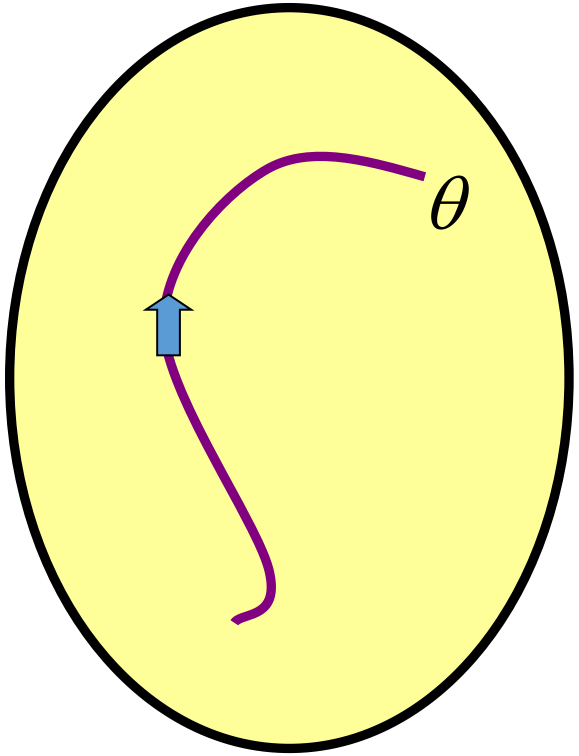
$$y = f(x, \theta) + n$$

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

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

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

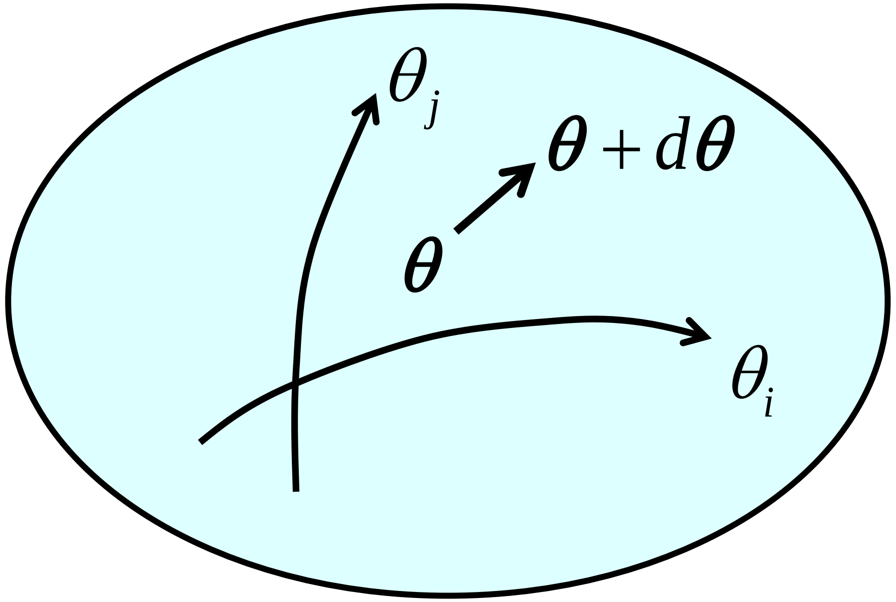
$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum v_i \varphi(\boldsymbol{w}_i \cdot \boldsymbol{x})$$



計量: 実はリーマン空間であった

$$g_{ij}(\theta) = E[\frac{\partial \log p(y \mid x;\theta) \partial \log p(y \mid x;\theta)}{\partial \theta_i \partial \theta_j}]$$

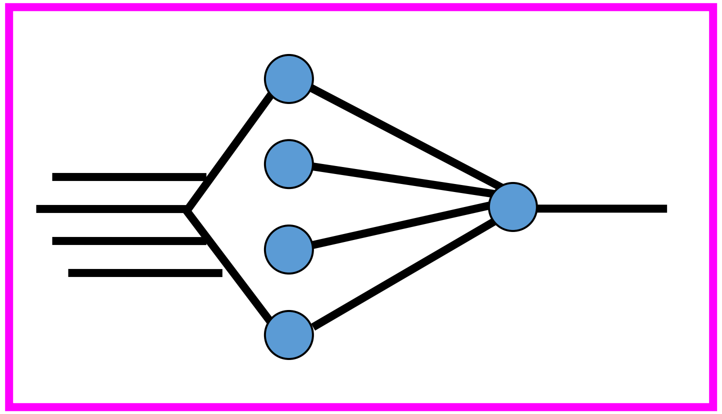
$$ds^2 = \left| d\boldsymbol{\theta} \right|^2$$
$$= \sum g_{ij}(\boldsymbol{\theta}) d\boldsymbol{\theta}_i d\boldsymbol{\theta}_j$$
$$= d\boldsymbol{\theta}^T G(\boldsymbol{\theta}) d\boldsymbol{\theta}$$



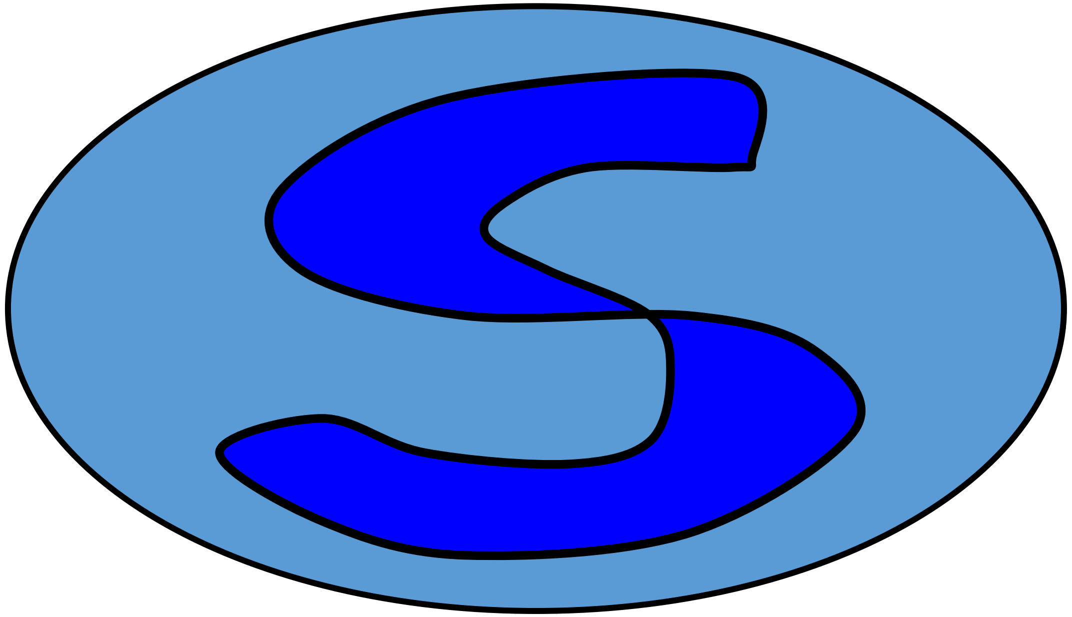
Topology: Neuromanifold

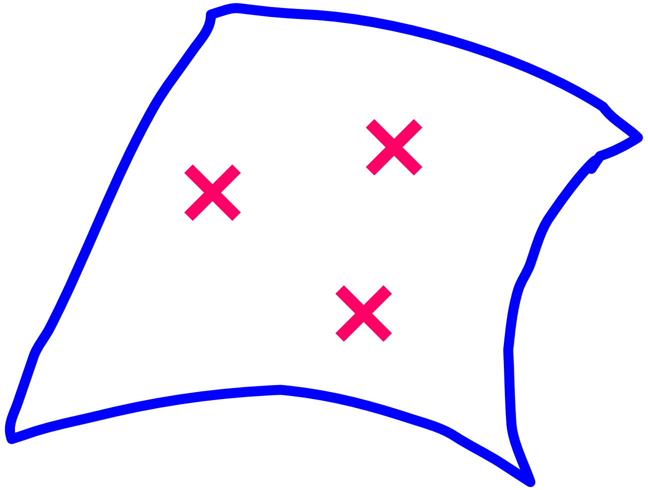
- **Metrical structure**

- **Topological structure**

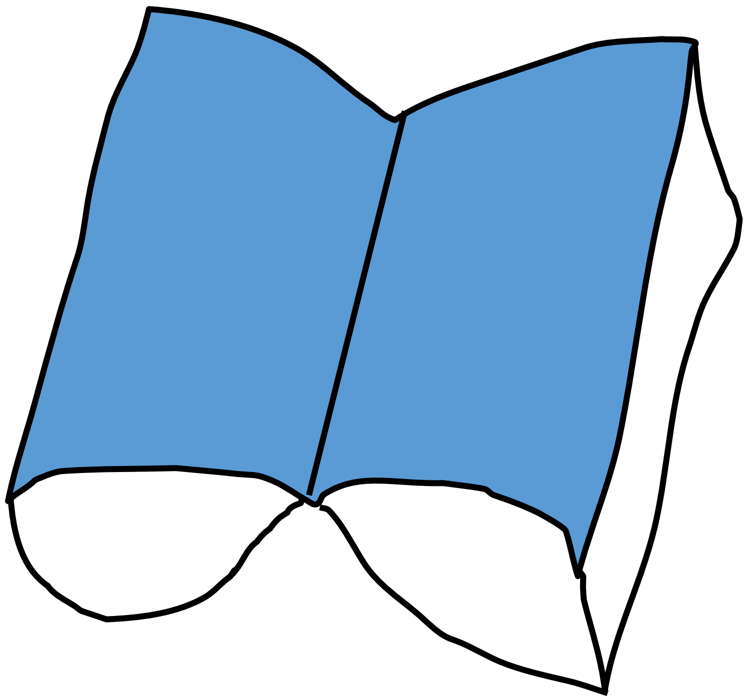


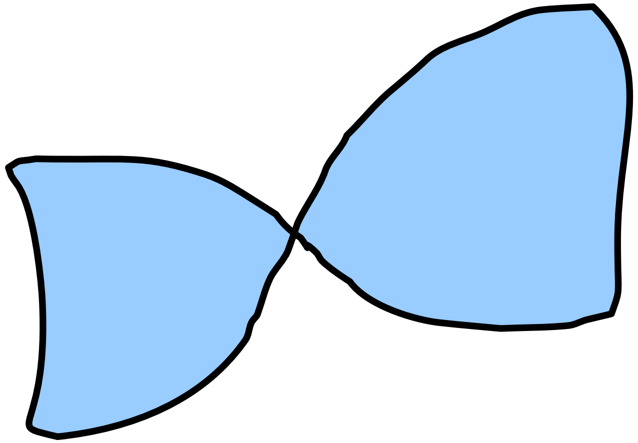
θ





singularities





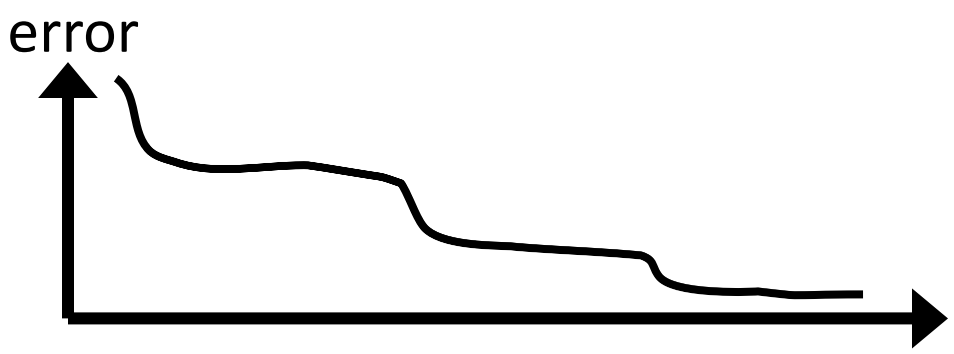
Backprop の問題点

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

- **slow convergence----plateau---saddle**
- **local minima**

MLP学習の欠陥

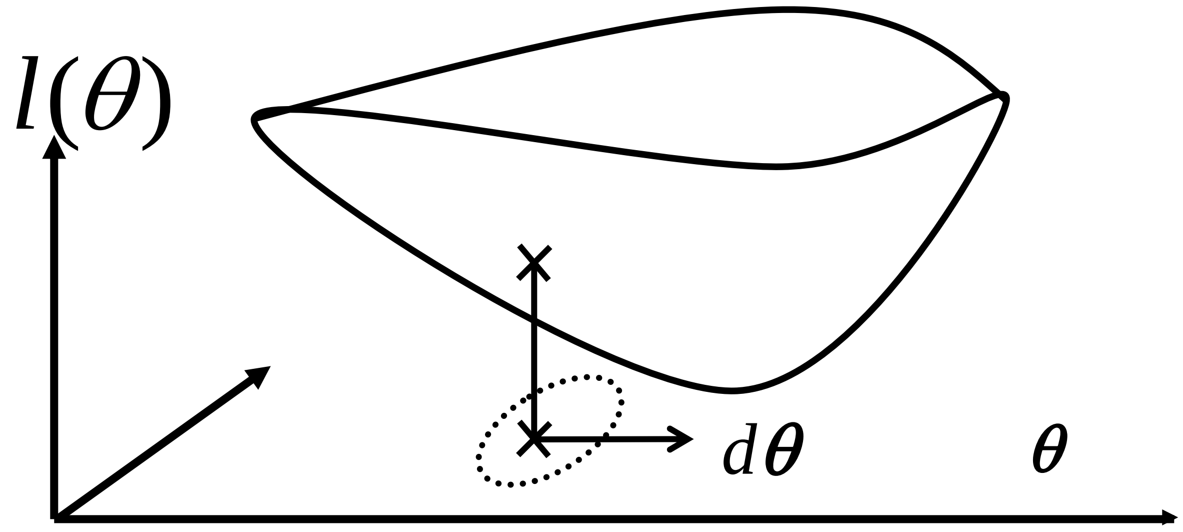
slow convergence : plateau



local minima

→ Boosting, Bagging, SVM

最急降下方向--- Natural Gradient



$$\nabla l = \left(\frac{\partial l}{\partial \theta_1}, \cdots, \frac{\partial l}{\partial \theta_n} \right)$$

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

$$\nabla l = G^{-1}(\boldsymbol{\theta}) \nabla l$$

$$|d\boldsymbol{\theta}|^2 = d\boldsymbol{\theta}^T G d\boldsymbol{\theta} = \sum G_{ij} d\theta^i d\theta^j$$

自然勾配学習 Natural Gradient

$$\max \quad dl = l(\boldsymbol{\theta} + d\boldsymbol{\theta}) - l(\boldsymbol{\theta}) = \nabla l \cdot d\theta$$

$$\text{under } |d\boldsymbol{\theta}|^2 = \sum g_{ij} d\theta_i d\theta_j = \varepsilon^2$$

$$d\theta \approx \nabla l = G^{-1}(\boldsymbol{\theta}) \nabla l$$

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

MLPの情報幾何

Natural Gradient Learning :
S. Amari ; H.Y. Park

$$\Delta\boldsymbol{\theta} = -\eta G^{-1}(\boldsymbol{\theta})\frac{\partial l}{\partial \boldsymbol{\theta}}$$

Adaptive natural gradient learning

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

Landscape of error at singularity

Milner attractor

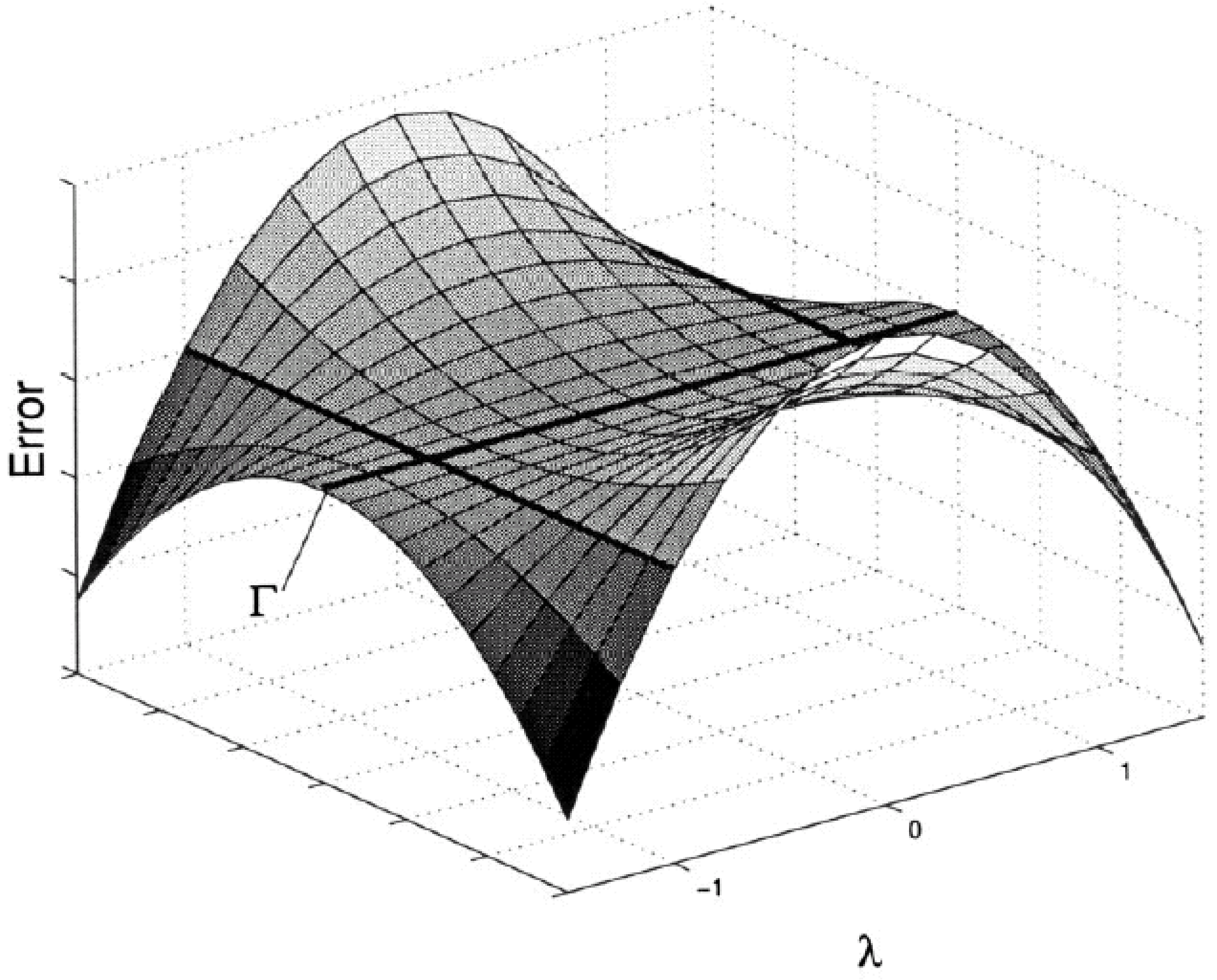
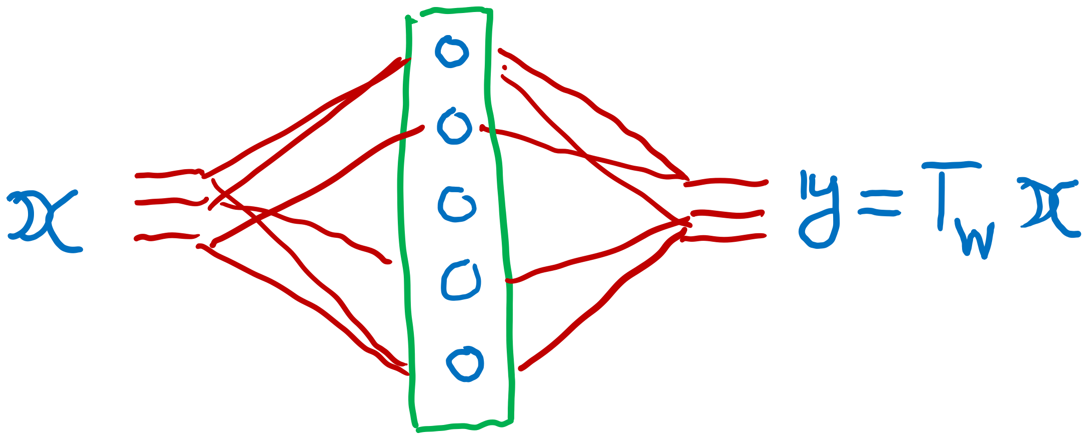


Fig. 5. Critical set with local minima and plateaus.

統計神経力学

Rozonoer (1969)
Amari (1971; 1974)
Amari et al (2013)
Toyoizumi et al (2015)
Poole, …, Ganguli (2016)
Schoenholz et al (2017)



$$w_{ij} \sim N(0,\ 1)$$

巨視的振舞い

ほとんどすべての（典型的）回路に共通

巨視変数

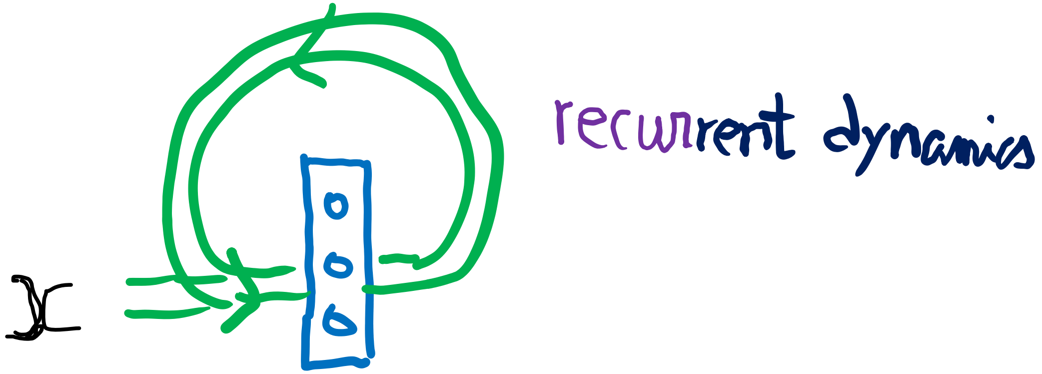
活動度 : $A = \frac{1}{n}\sum x_i^{\,2}$

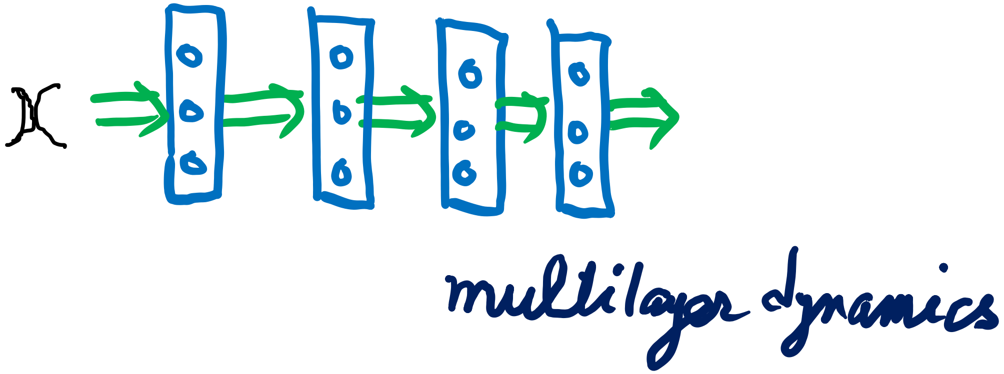
距離・計量: $D = D[\mathbf{x} : \mathbf{x}']$

曲率 :

$$A_{l+1} = F(A_l)$$

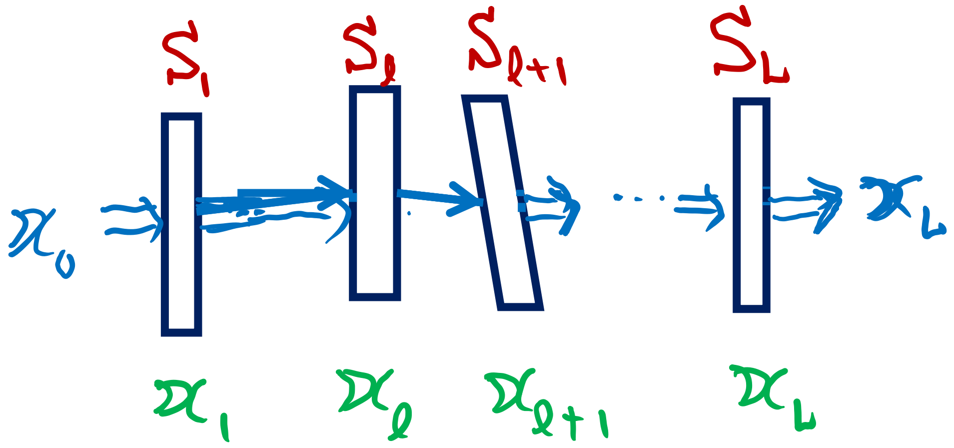
$$D_{l+1} = K(D_l)$$





深層回路

$$\underset{l+1}{x_i} = \varphi\left(\sum w_{ij}\, \underset{l}{x_i} + w_{0i}\right)$$



$$\underset{l}{A} = \frac{1}{n_l}\sum \underset{l}{x_i}^{\,2}$$

$$w_{ij} \sim N(0,\ \sigma^2/\sqrt{n})$$

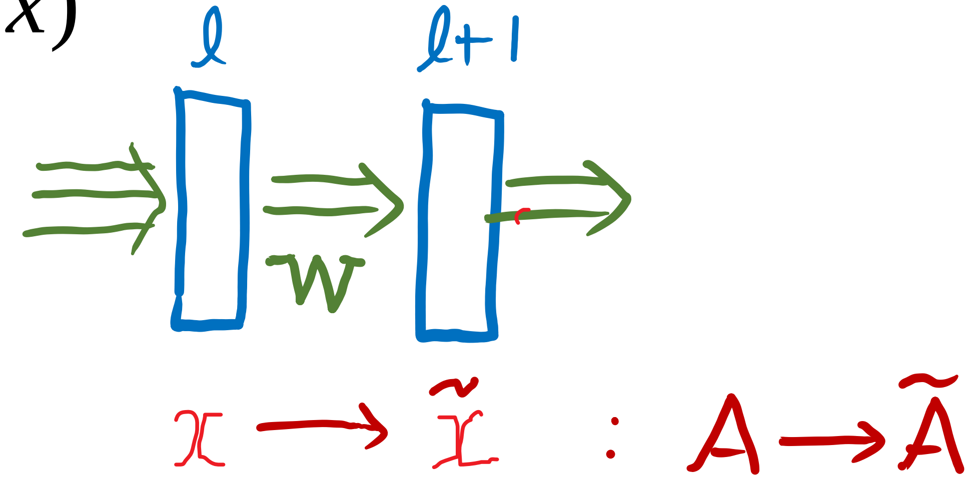
$$\underset{l+1}{A} = F(\underset{l}{A})$$

$$w_{0i} = b \ \sim N(0,\ \sigma_b^2)$$

Dynamics of Activity: law of large numbers

$$\tilde{x}_\alpha = \varphi(\sum w_{\alpha k} x_k + b_\alpha) = \varphi(u_\alpha) : \tilde{x} = \phi(Wx)$$

$$u_\alpha \sim N(0, A)$$



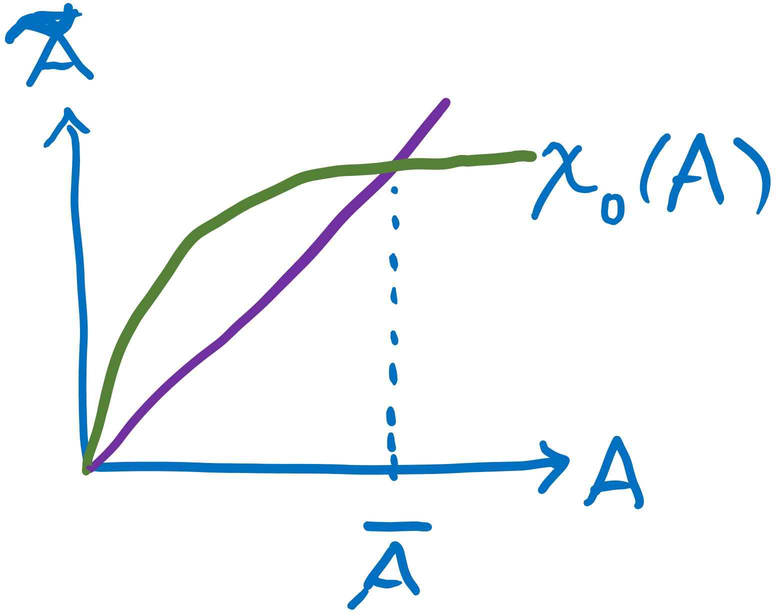
$$\tilde{A} = \frac{1}{n_{l+1}} \sum (\tilde{x}_\alpha)^2 = E[\varphi(u_\alpha)^2] = \chi_0(A)$$

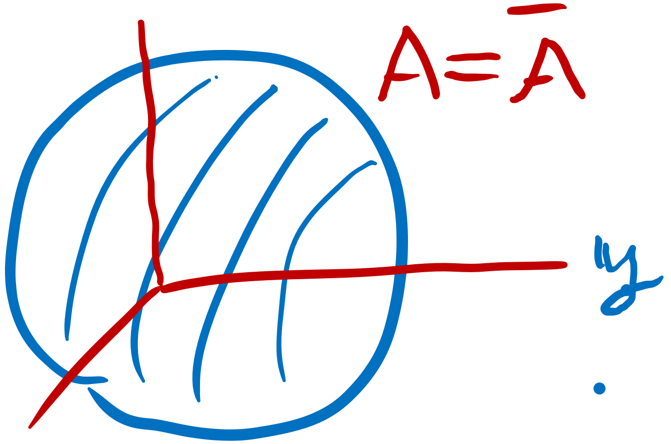
$$\chi_0(A) = \int \varphi^2(\sqrt{A}v) Dv \qquad v \sim N(0,1)$$

$$\chi_0'(0) > 1$$

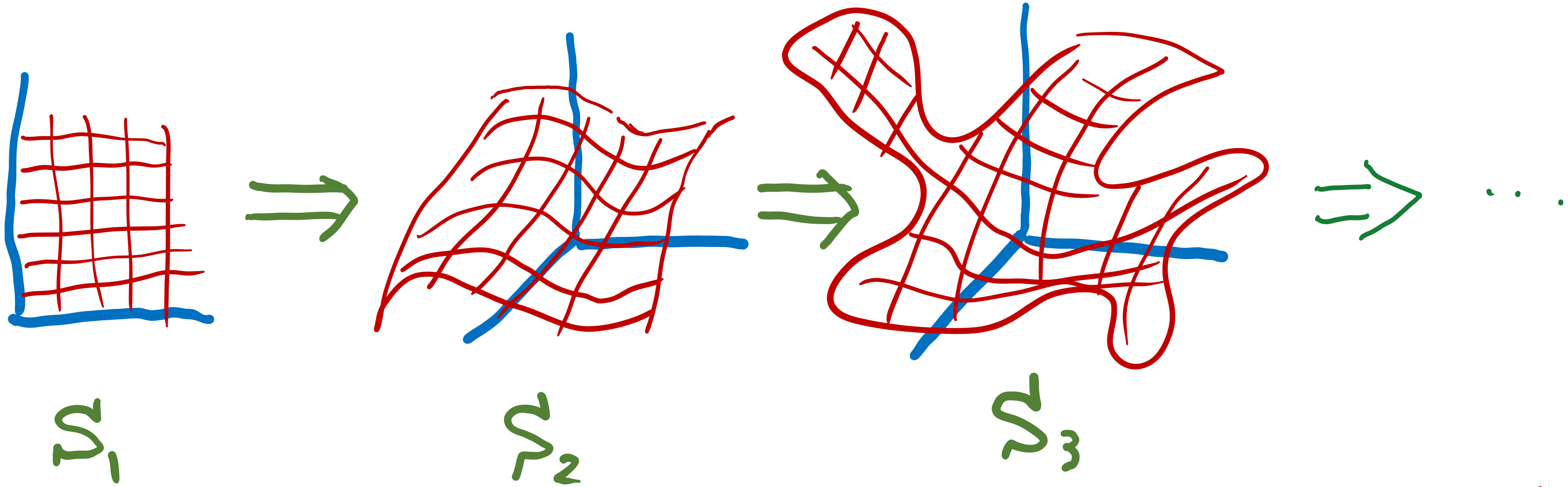
$$\bar{A} = \chi_0(\bar{A})$$

$$\sum x_i^2 \to \text{converge}$$



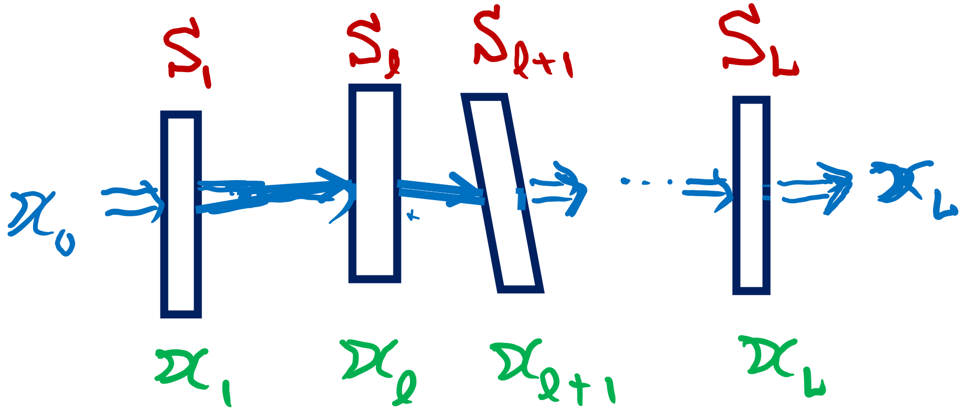


引き戻し計量（リーマン計量・距離・曲率）

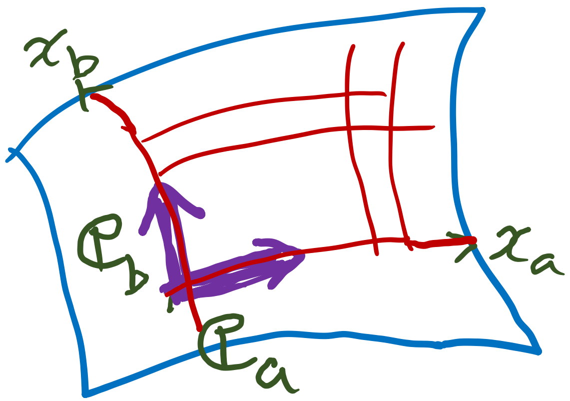


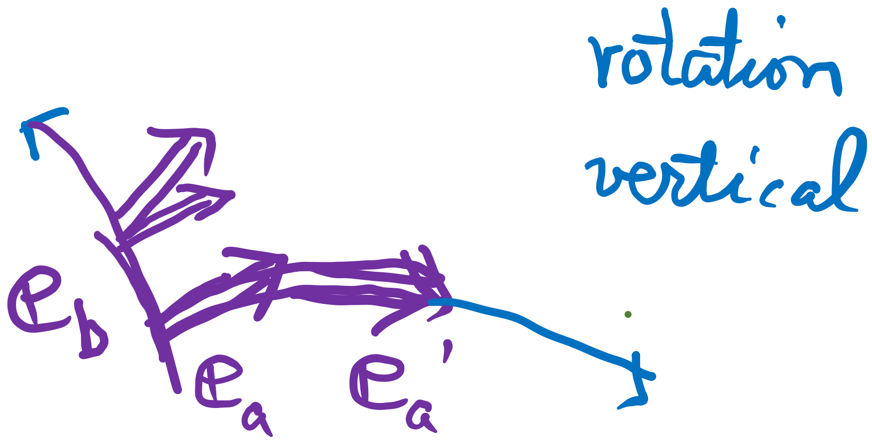
$$ds^2 = \sum g^l{}_{ab}\, dx^a dx^b = \frac{1}{n_l} d\mathbf{x}^l \cdot d\mathbf{x}^l$$

$$g^l{}_{ab} = \mathbf{e}^l{}_a \cdot \mathbf{e}^l{}_b$$

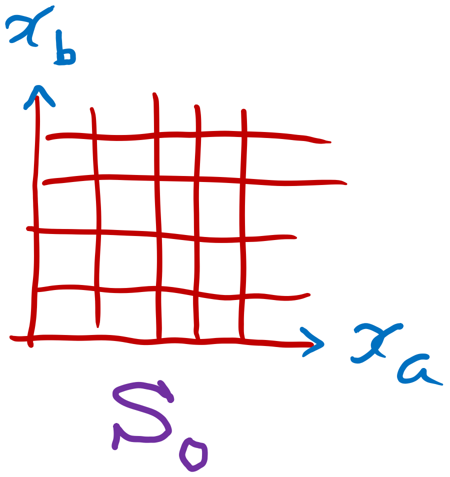


曲率

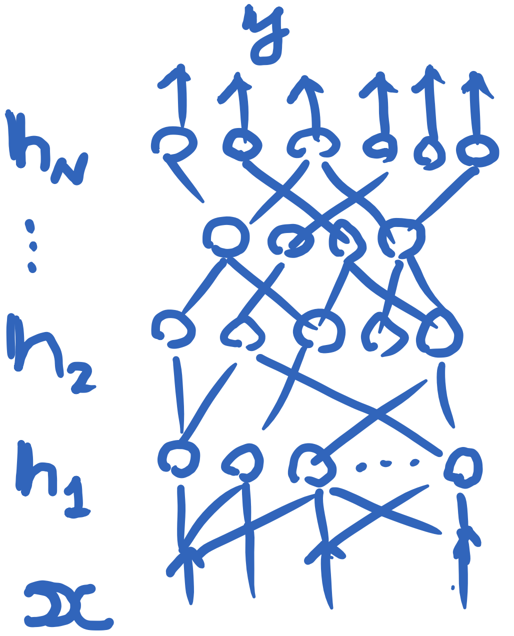


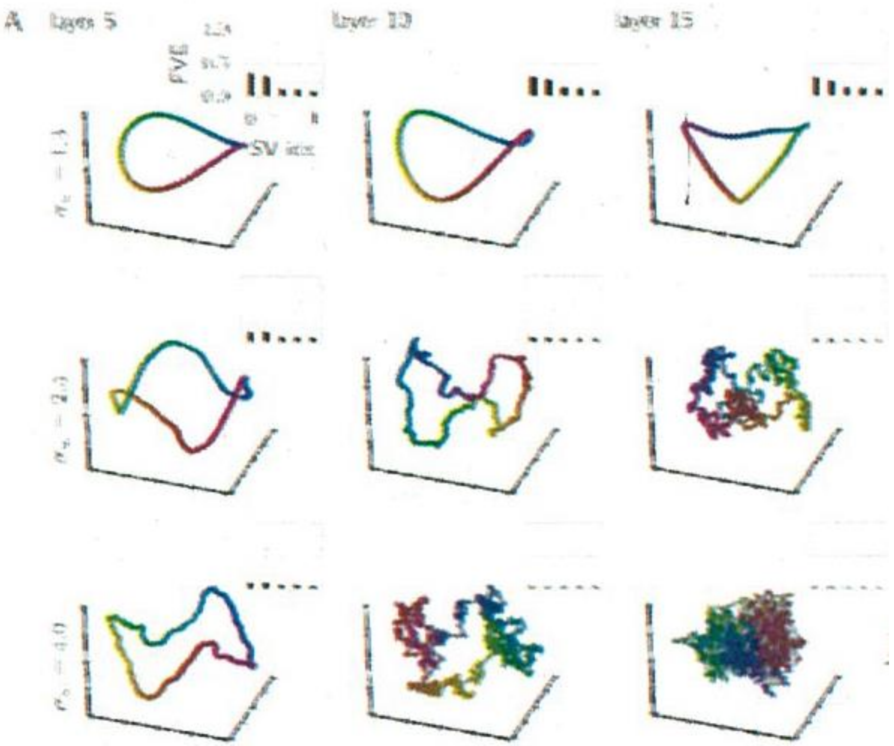


$$\overset{\ell}{H}_{abi} = \nabla_a \overset{\ell}{\mathbf{e}}_b$$



Poole et al (2016)
Random deep neural networks





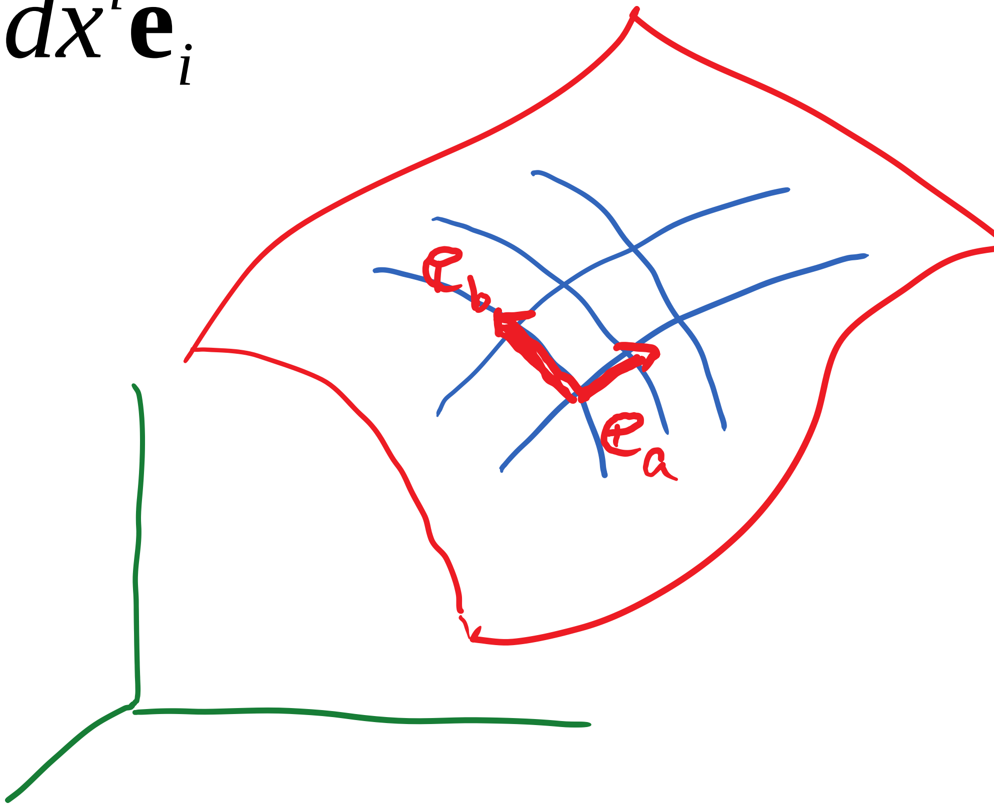
Basis vectors

$$dx_{i_l} = \sum \varphi'\left(u_{i_l}\right) W^{i_l}_{i_{l-1}} dx_{i_{l-1}} = \sum B^{i_l}_{i_{l-1}} dx_{i_{l-1}}$$

$$d\overset{l}{\boldsymbol{x}} = \overset{l}{B}\, d\overset{l-1}{\boldsymbol{x}} = \overset{l}{B}\dots\overset{m}{B}\, d\overset{m-1}{\boldsymbol{x}} \qquad d\mathbf{x} = \sum dx^i \mathbf{e}_i$$

$$B^{i_l}_{i_{l-1}} = \varphi'\left(u_{i_l}\right) W^{i_l}_{i_{l-1}}$$

$$\overset{l}{\boldsymbol{e}}_a = \overset{l}{B}\,\overset{l-1}{\boldsymbol{e}}_a = \overset{l}{B}\dots\overset{m}{B}\,\overset{m-1}{\boldsymbol{e}}_a$$

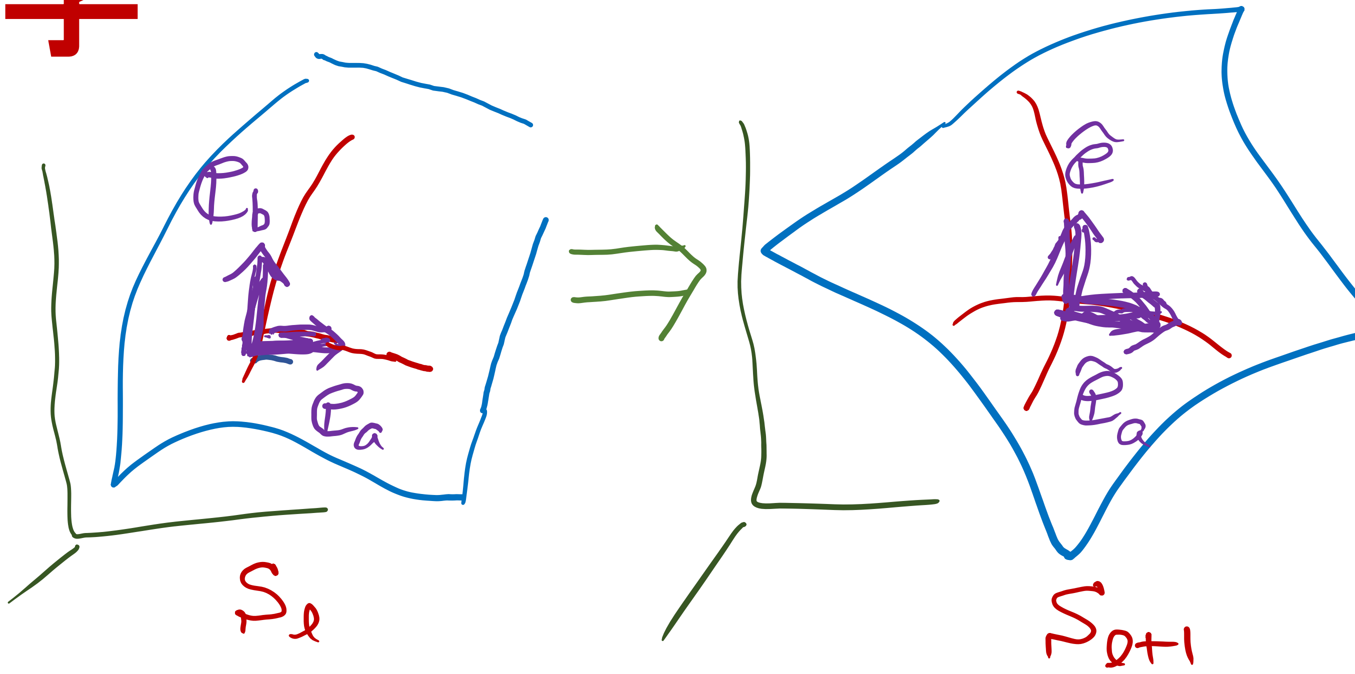


リーマン計量の力学

$$\tilde{y}_\alpha = \varphi\left(\sum w_{\alpha k} y_k + b_\alpha\right) = \varphi(u_\alpha)$$

$$d\tilde{y}_\alpha = \sum B_k^\alpha dy_k \qquad \tilde{\mathbf{e}}_a = B\mathbf{e}_a$$

$$ds^2 = \sum g_{ij} dy^i dy^j = <\mathbf{dy}, \mathbf{dy}>$$



$$B = (B_k^\alpha) = (\varphi'(u_\alpha) w_k^\alpha)$$

$$<\tilde{\mathbf{e}}_\alpha, \tilde{\mathbf{e}}_\beta> = \sum B_k^\alpha B_j^\alpha <\mathbf{e}_k, \mathbf{e}_j> = \chi_1 \delta_\alpha^k \delta_\beta^j g_{jk}$$

$$\mathrm{E}[\varphi'(u_\alpha))^2 w_k^\alpha w_j^\alpha] = \mathrm{E}[\varphi'(u_\alpha))^2]\mathrm{E}[w_k^\alpha w_j^\alpha]$$

平均場近似

$$\chi_1(A) = \int \sigma^2 \{\varphi'(\sqrt{A}v)\}^2 Dv = \frac{1}{2\pi} \frac{\sigma^2 A + \sigma_b^2}{\sqrt{1 + 2(\sigma^2 A + \sigma_b^2)}}$$

Mectric

Law of large numbers

$$\overset{l}{g}_{ab} = \left\langle \overset{l}{\boldsymbol{e}}_a, \overset{l}{\boldsymbol{e}}_b \right\rangle = BB \overset{l-1}{g}_{ab}$$

$$ds^2 = \sum \overset{l}{g}_{ab} \, d\overset{l}{x}_a \, d\overset{l}{x}_b$$

$$BB = \sum_{i_l} W^{i_l}_{i_{l-1}} W^{i_l}_{i'_{l-1}} \varphi'\left(u_{i_l}\right)^2 \approx \sigma_l^2 E\left[\varphi'^2\right] \delta_{i_{l-1} i'_{l-1}}$$

$$\chi_1 = \sigma_l^2 E\left[\varphi'\left(u_{i_l}\right)^2\right]$$

Mectric

Law of large numbers

$$\overset{l}{g}_{ab} = \left\langle \overset{l}{\boldsymbol{e}}_a , \overset{l}{\boldsymbol{e}}_b \right\rangle = BB \, \overset{l-1}{g}_{ab}$$

$$ds^2 = \sum \overset{l}{g}_{ab} \, d \overset{l}{x}_a \, d \overset{l}{x}_b$$

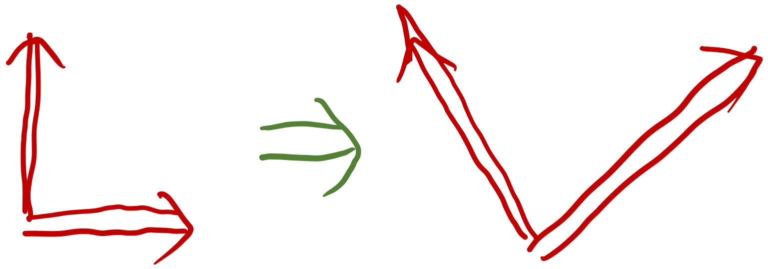
$$BB = \sum_{i_l} W^{i_l}_{i_{l-1}} \, W^{i_l}_{i'_{l-1}} \varphi'\left(u_{i_l}\right)^2 \approx \sigma_l^2 E\left[\varphi'^2\right] \delta_{i_{l-1} \, i'_{l-1}}$$

$$\chi_1 = \sigma_l^2 E\left[\varphi'\left(u_{i_l}\right)^2\right]$$

$$\tilde{g}_{ab} = \chi_1(A) g_{ab}$$

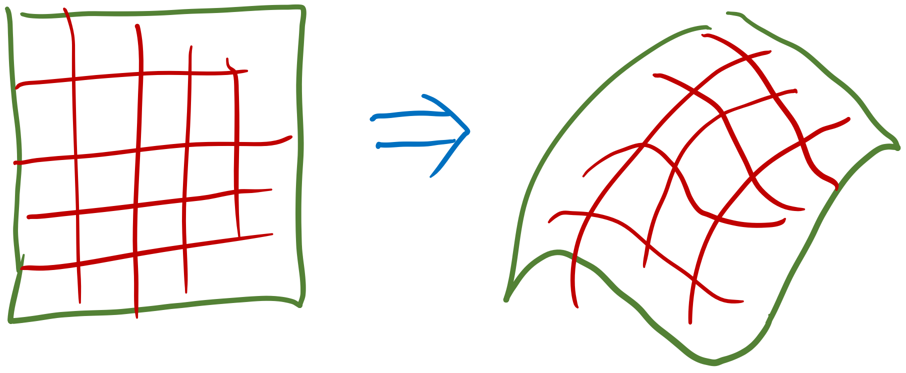
conformal transformation!

回転,拡大・縮小



$$\bar{\chi}_1 = \bar{\chi}_1(\bar{A}) > 1:$$

拡大(カオス、Lyapnov指数)

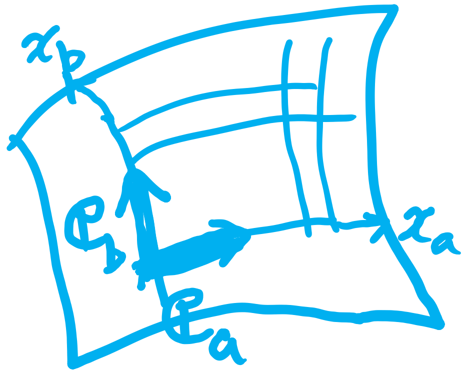


$$\Rightarrow g^{l}{}_{ab} = \Pi \chi_1 (A^s) \delta_{ab}$$

$$\overset{l}{g}_{ab}(x) = \left(\prod \chi_1(x)\right) g_{ab}(x)$$

conformal geometry

曲率の力学



rotation

vertical

$$\tilde{H}_{ab}^{\alpha} = \nabla_a \tilde{\mathbf{e}}_b^l = \partial_a \partial_b \tilde{y}^\alpha$$

$$= \varphi''(u_\alpha)(\mathbf{w}\cdot\mathbf{e}_a)(\mathbf{w}\cdot\mathbf{e}_b) + \varphi'(\mathbf{w}\cdot\partial_a \mathbf{e}_b)$$

$$\tilde{\mathbf{H}}_{ab} = \mathbf{H}_{ab}^{\perp} + \mathbf{H}_{ab}^{\parallel}$$

Euler-Schouten曲率
Affine connection

$$\tilde{H}_{ab}^2 = |\tilde{\mathbf{H}}_{ab}|^2$$

curvature & distortion

$$\overset{l}{\boldsymbol{H}}_{ab} = \nabla_a \overset{l}{\boldsymbol{e}}_b = \nabla_a \left(B \overset{l-1}{\boldsymbol{e}}_b \right) = B \nabla_a \overset{l-1}{\boldsymbol{e}}_b + (\nabla_a B) \overset{l-1}{\boldsymbol{e}}_b$$

$$\left| \overset{l}{\boldsymbol{H}}_{ab} \right|^2 = \chi_1 \left| \overset{l-1}{\boldsymbol{H}}_{ab} \right|^2 + \frac{1}{n\chi_1^2} \left(1 + 2\delta_{ab} \right) \chi_2$$

$$\chi_2 = \sigma^2 E\left[\varphi''(u)^2 \right]$$

$$\chi_2(A) = \int \varphi''(\sqrt{A}v)^2\, Dv$$

$$\overset{l+1}{H}{}^{2}_{ab} = \frac{1}{n\chi_1^2}\chi_2(\overset{l}{A})(2\delta_{ab}+1) + \chi_1(\overset{l}{A})\overset{l}{H}{}^{2}_{ab}$$

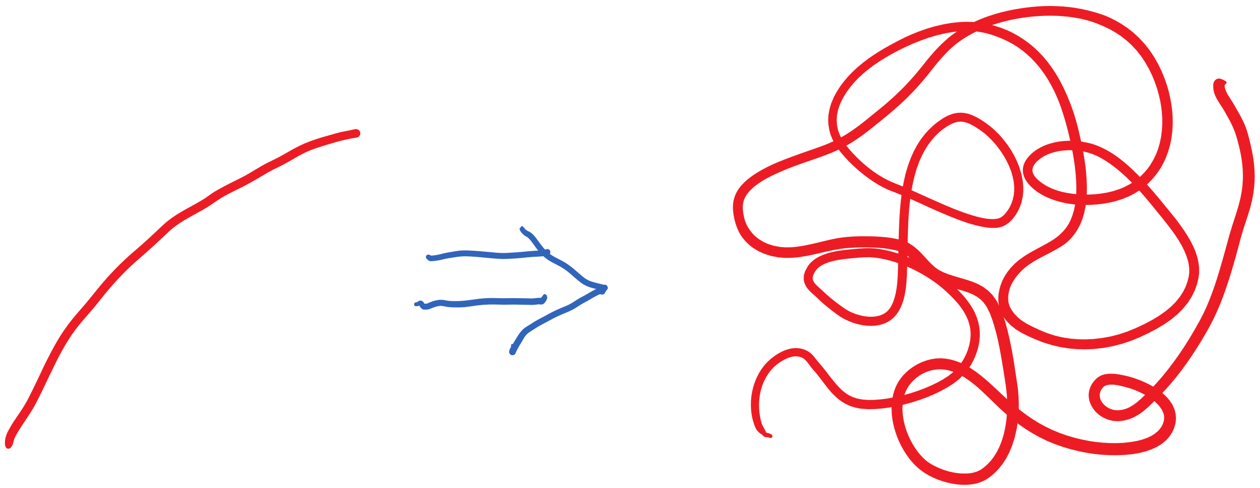
$$\chi_1 > 1$$

exponential expansion!   creation is small!

scalar curvature & distortion

$$\overset{l}{\gamma}{}^2 = \frac{1}{\chi_1}\overset{l-1}{\gamma}{}^2 + \frac{3}{n}\frac{\chi_2^2}{\chi_1^2} \qquad \gamma^2 = H_{ab}{}^{i} H_{cd}{}^{j} g^{ac} g^{bd} \delta_{ij}$$

$$\overset{l}{\gamma}{}^2 \to \frac{3\chi_2}{n\chi_1\left(\chi_1 - 1\right)} \quad : \quad \to \infty,\ \chi_1 \leq 1$$

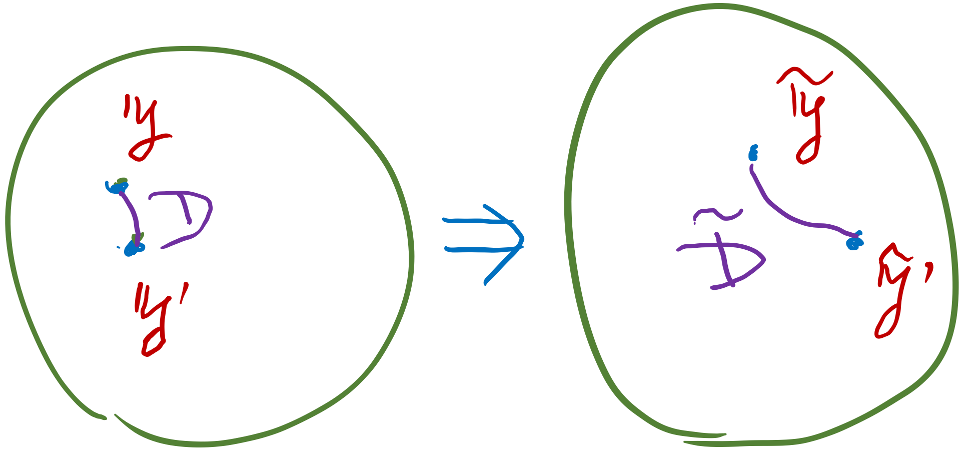


距離法則 (Amari, 1974)

$$D(x,x') = \frac{1}{n}\sum (x_i - x_i')^2$$

$$C(x,x') = \frac{1}{n} x \cdot x' = \sum x_i x_i'$$

$$D = A + A' - 2C$$



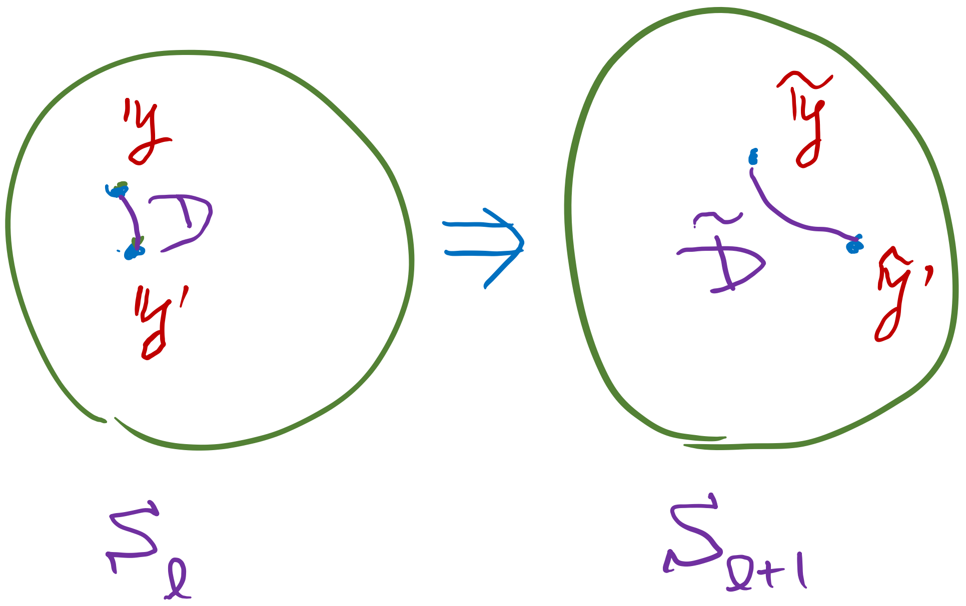
S_l S_{l+1}

Dynamics of Distance (Amari, 1974)

$$D(x,x') = \frac{1}{n}\sum (x_i - x_i')^2$$

$$C(x,x') = \frac{1}{n} x \cdot x' = \sum x_i x_i'$$

$$D = A + A' - 2C$$



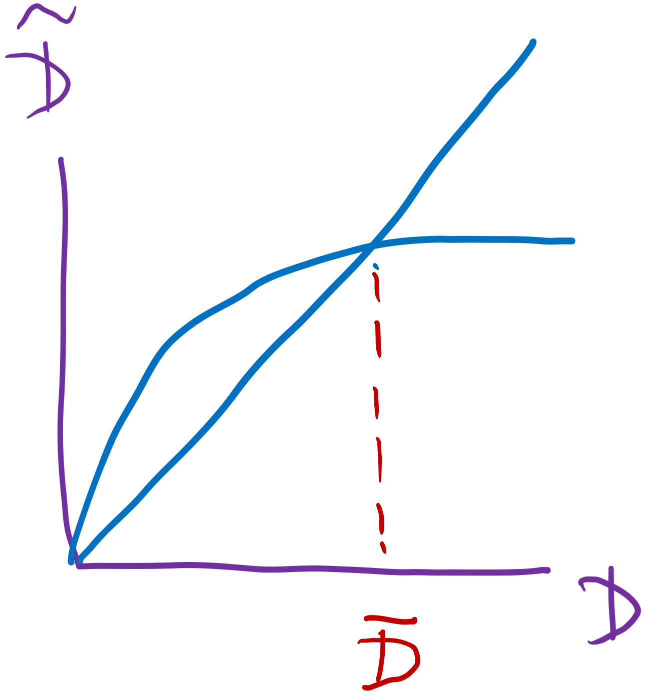
$$u_\alpha = \sum w_{\alpha k} y_k \qquad \sim \mathrm{N}(0, \mathrm{V})$$

$$u'_\alpha = \sum w_{\alpha k} y'_k \qquad \mathrm{V} = \begin{pmatrix} A & C \\ C & A' \end{pmatrix}$$

$$\tilde{C} = \mathrm{E}[\varphi(\sqrt{A - C}\varepsilon + \sqrt{C}\nu)\varphi(\sqrt{A' - C}\varepsilon + \sqrt{C}\nu)]$$

$$D_{l+1} = K(D_l)$$

$$\frac{d\tilde{D}}{dD} = \chi_1 > 1$$



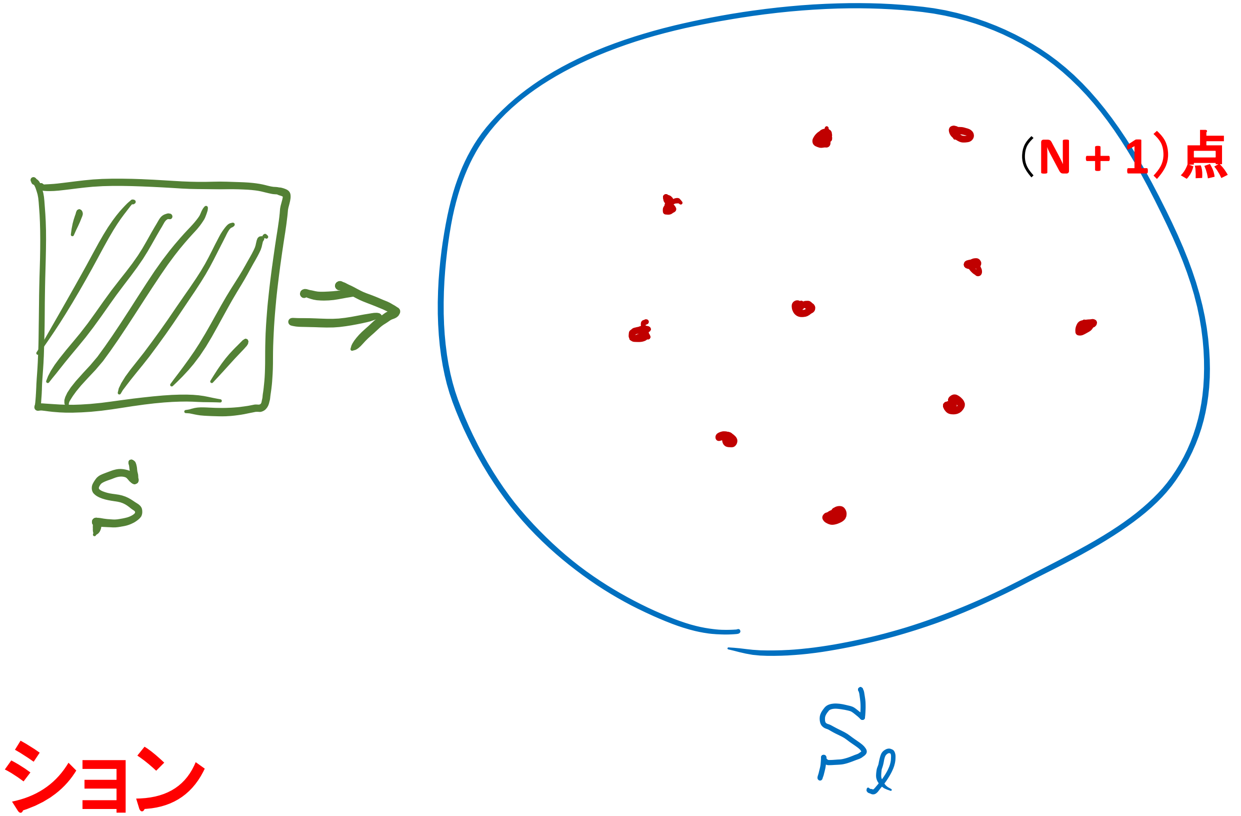
本当か! $n \to \infty; l \to \infty$

equidistance

$$D(\mathbf{x}_l, \mathbf{x}'_l) \to \bar{D}$$

$$\bar{D} = \xi(\bar{D})$$

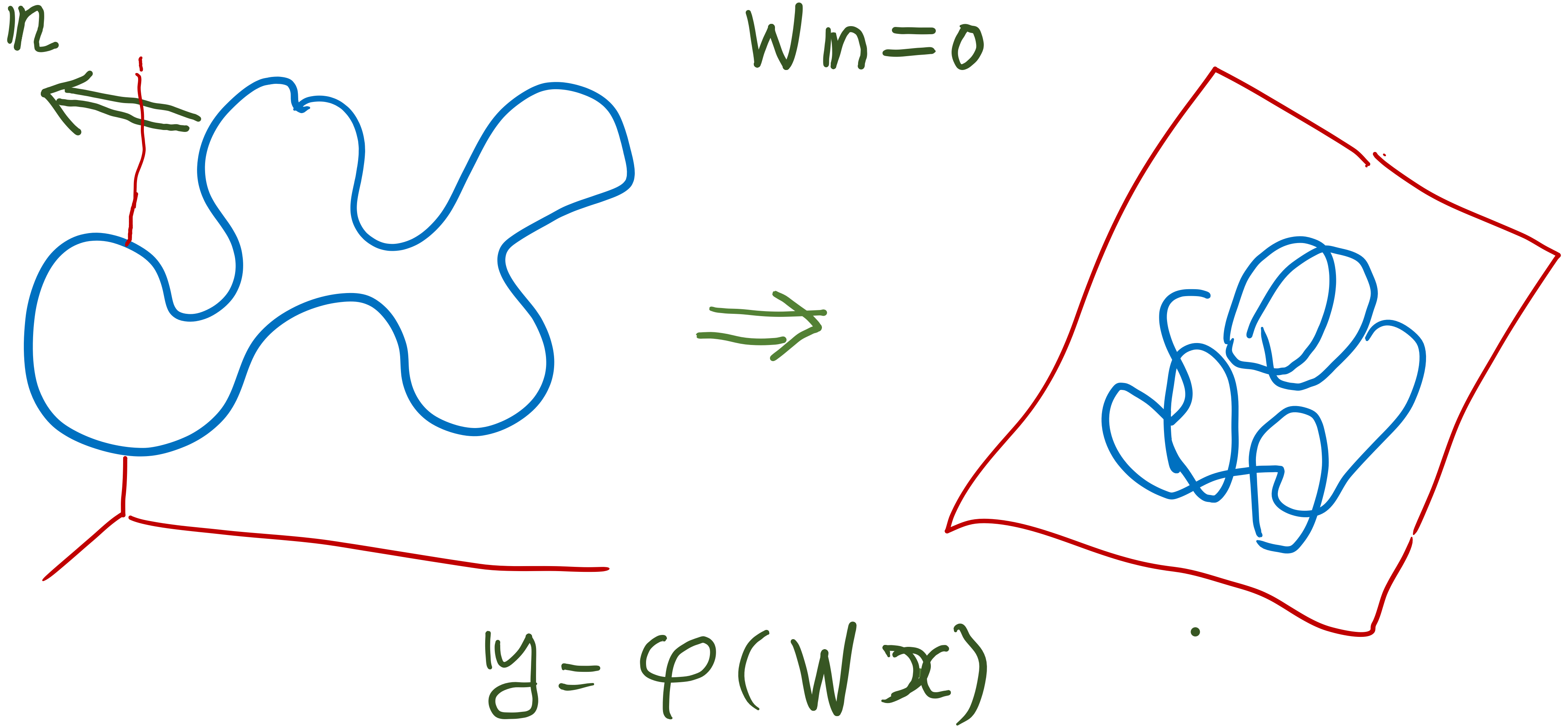
等距離



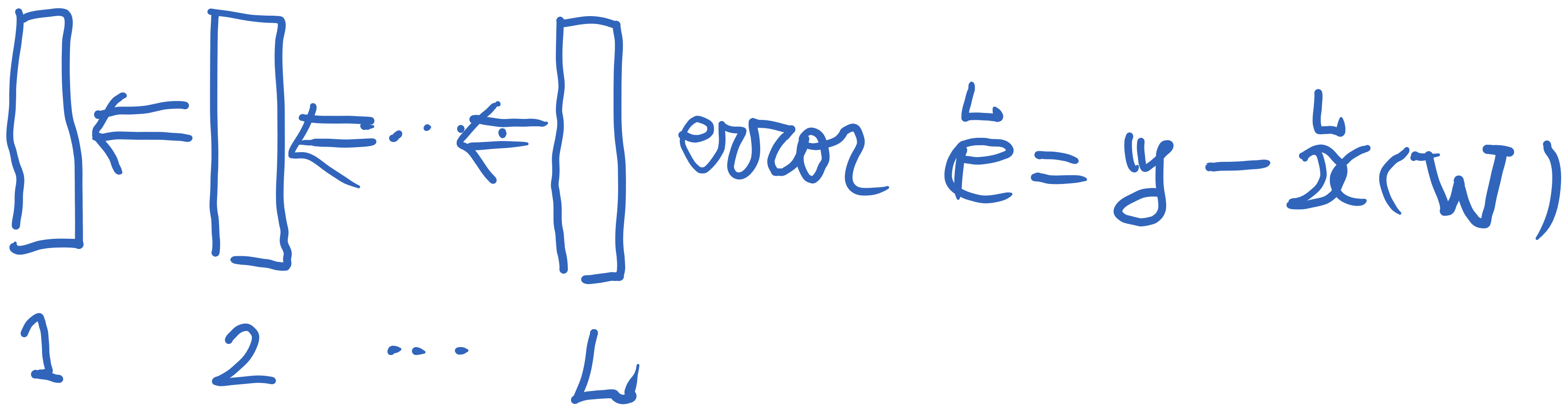
フラストレーション

フラクタル

$n_{l+1} < n_l$ **次元の縮小**



Fisher 情報行列と逆向き情報伝播



$$l(x,W) = \frac{1}{2}\left|y - \varphi(x;W)\right|^2 = \left|e(x,y)\right|^2$$

確率 model： 深層回路の多様体

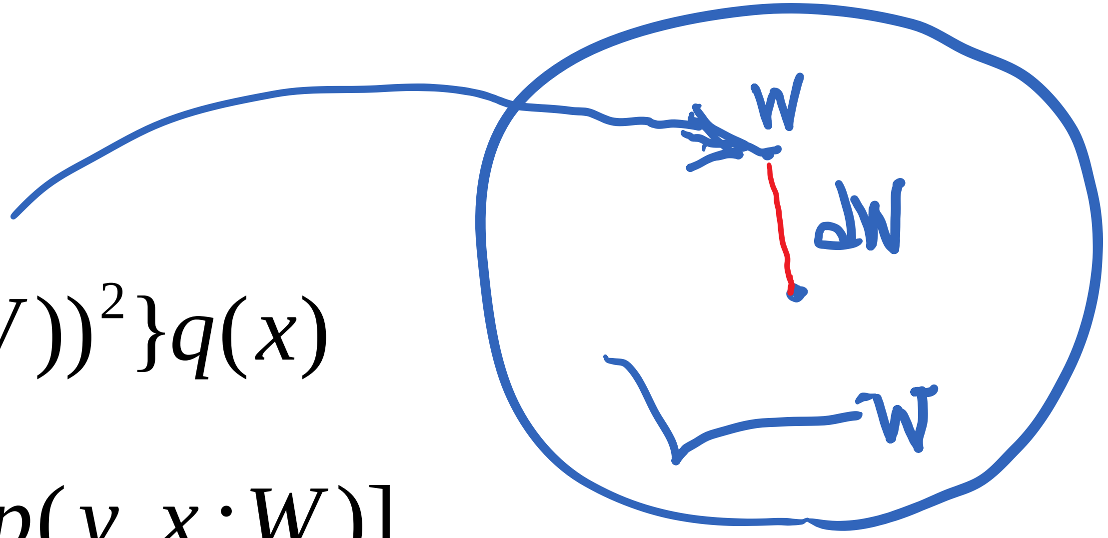
$$y = \varphi(u) + \varepsilon; \quad \varepsilon \sim N(0,1)$$

$$p(y, x : W) = c \exp\{-\frac{1}{2}(y - \varphi(x; W))^2\} q(x)$$

$$G = E_x[\nabla_W \log p(y, x : W) \nabla_W \log p(y, x : W)]$$

$$ds^2 = dWGdW$$

Fisher information



Riemannian

Natural Gradient

$$\max \qquad dl = l(\boldsymbol{\theta} + d\boldsymbol{\theta}) - l(\boldsymbol{\theta})$$

$$|d\boldsymbol{\theta}|^2 = \varepsilon \qquad \mathrm{KL}[\mathrm{p}(\mathrm{x}, \theta) : \mathrm{p}(\mathrm{x}, \theta + d\theta)] = \varepsilon$$

$$\nabla l = G^{-1}(\boldsymbol{\theta}) \nabla l$$

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

Fisher information

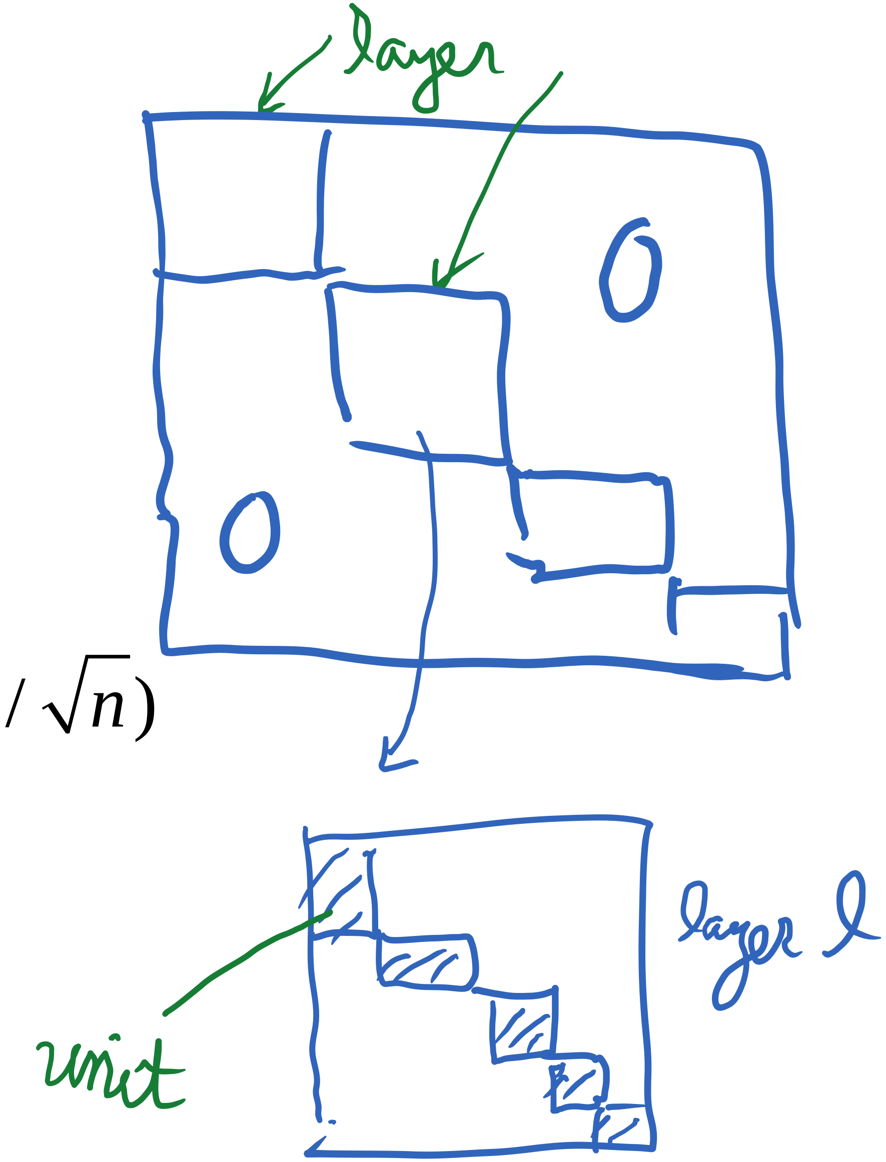
$$G = E_x\left[\frac{\partial\varphi}{\partial W_m}\frac{\partial\varphi}{\partial W_l}\right]$$

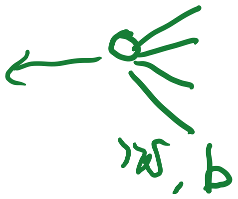
$$\frac{\partial\varphi^l}{\partial W_m} = \varphi' W\frac{\partial\varphi^{l-1}}{\partial W_m} = B\frac{\partial\varphi^{l-1}}{\partial W_m} = BB\ldots B\frac{\partial\varphi^{m+1}}{\partial W_m}$$

$$G\left(W_l, W_m\right) = \prod \chi_1 E_x\left[\varphi'\left(\overset{l}{\boldsymbol{w}}_i\right)^2 \overset{l-1}{\boldsymbol{x}}\ \overset{l-1}{\boldsymbol{x}}\right] + O_p(1/\sqrt{n})$$

$$G\left(W_l, W_m\right) = 0 \sim O_p(1/\sqrt{n}), \qquad l \neq m$$

$$G\left(\overset{l}{\boldsymbol{w}}_i, \overset{l}{\boldsymbol{w}}_j\right) = 0 \sim O_p(1/\sqrt{n}), \quad i \neq j$$

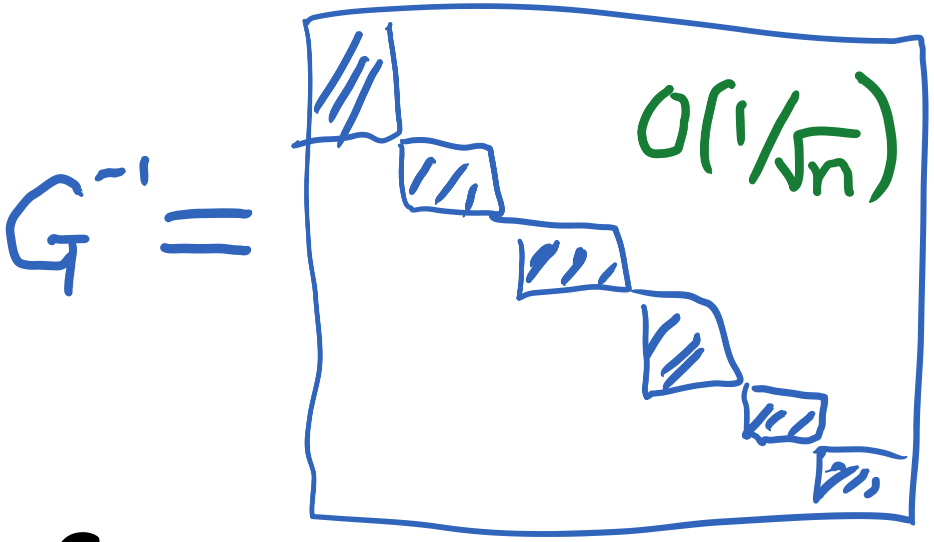




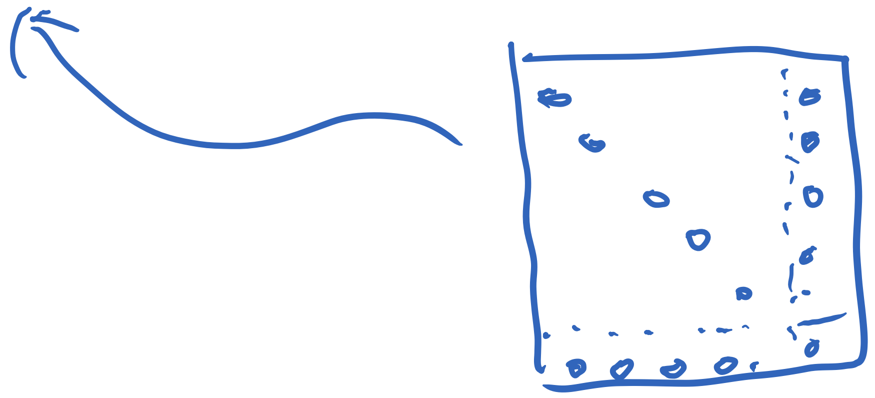
Y. Ollivier

Unitwise natural gradient

$$\Delta W = -\eta G^{-1} \nabla_W l$$

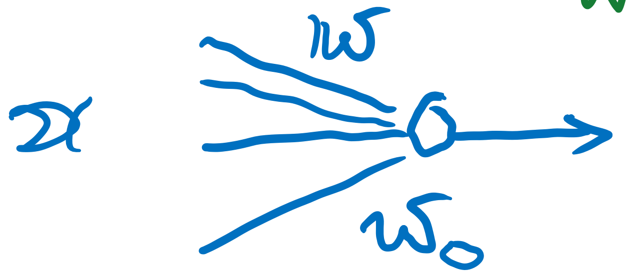


Y. Ollivier; Marceau-Caron



一個のニューロンの情報行列

$\tilde{w} = (w, w_0)$



$$y = \varphi\left(\mathbf{w} \cdot \mathbf{x} + w_0\right)$$

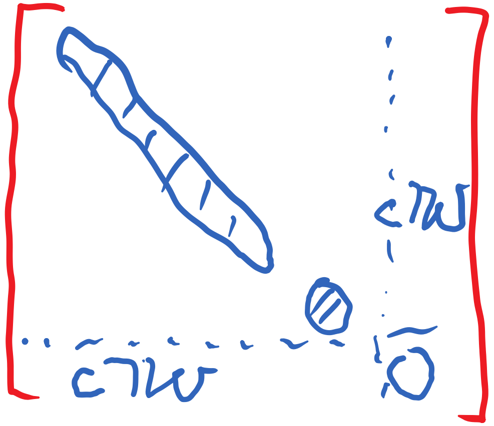
$$G = \mathrm{E}_{\mathbf{x}}\left[\partial_{\mathbf{w}}\varphi \; \partial_{\mathbf{w}}\varphi\right] = \mathrm{E}_x[(\varphi')^2 \mathbf{x}\mathbf{x}]$$

$G(\tilde{w}, \tilde{w})$

$$\{\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_n\} \Rightarrow \{\mathbf{e}_1{}^*, \mathbf{e}_2{}^*, \ldots, \mathbf{e}_n{}^*\}: \text{ortho-normal basis}$$

$$\mathbf{e}_n{}^* = \frac{\mathbf{w}}{w},$$

Input x: independent and identically distributed, 0-mean



$$G = A\mathbf{I} + \frac{B}{\boldsymbol{w}^2}\boldsymbol{ww} + \frac{C}{\boldsymbol{w}}\left(\boldsymbol{wb} + \boldsymbol{bw}\right)$$

$$\boldsymbol{b} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}'$$

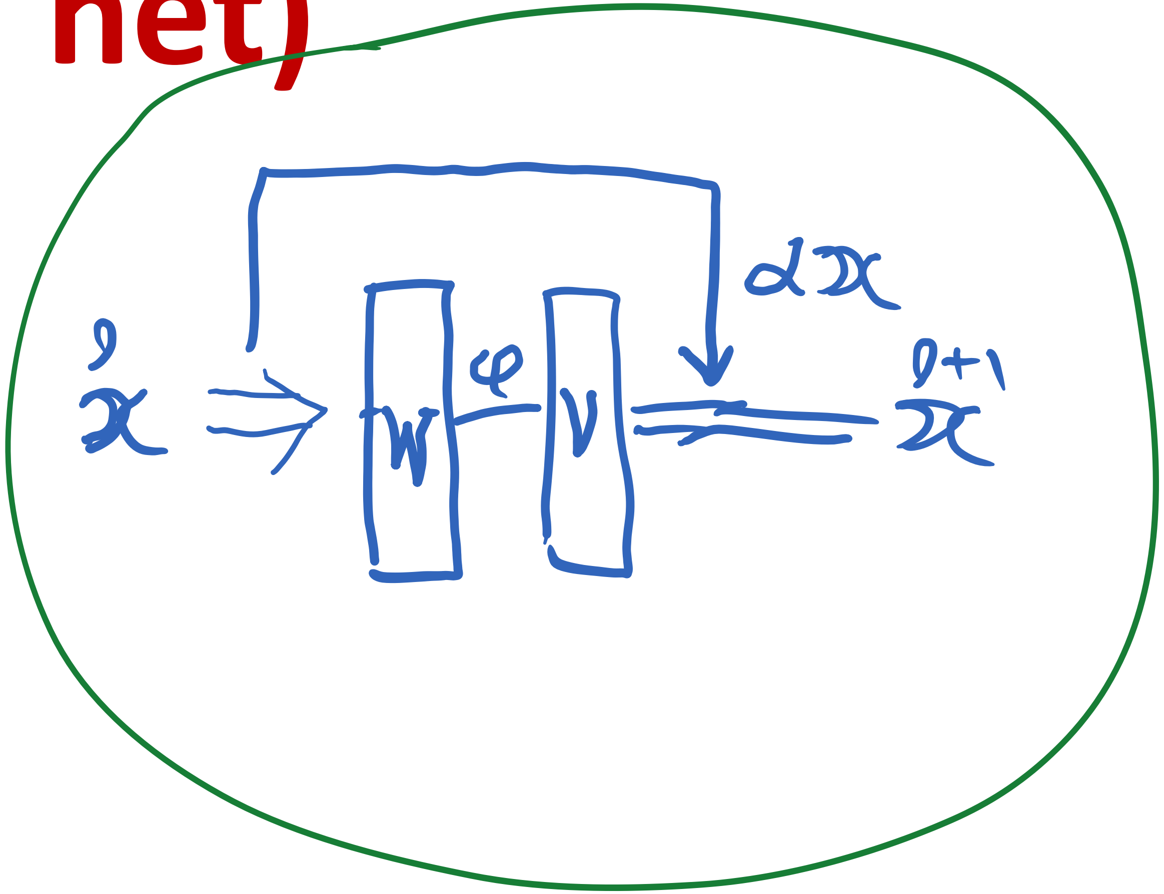
A, B, C ⏟ (w, b)

G^{-1} : similar form

Resnet (residual net)

$$\overset{l}{\boldsymbol{x}} = V\varphi\left(W\overset{l-1}{\boldsymbol{x}}\right) + \alpha\overset{l-1}{\boldsymbol{x}}$$

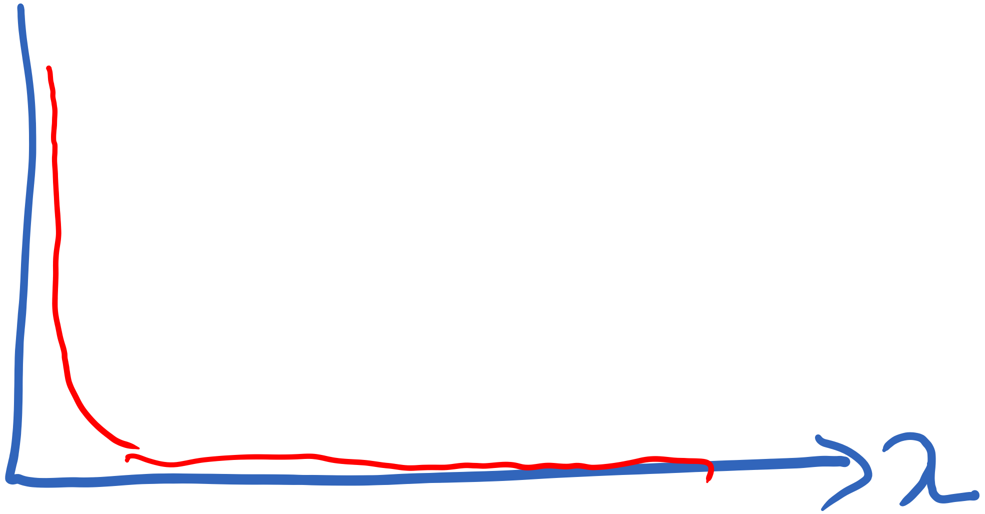
$$\chi_1 \to \sigma_v^2 \chi_1 + \alpha^2$$



Karakida theory

eigenvalues of G

$$\frac{1}{P}\sum \lambda_i = \frac{1}{n}, \quad \frac{1}{P}\sum \lambda_i^2 = O(1)$$



distorted Riemannian metric

Wasserstein Distance の情報幾何

Shun-ichi Amari
RIKEN Brain Science Institute
R. Karakida. M. Oizumi

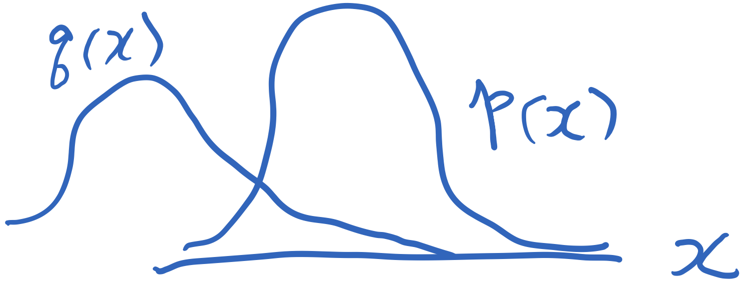
Base space X; X上のパターン: p(x)

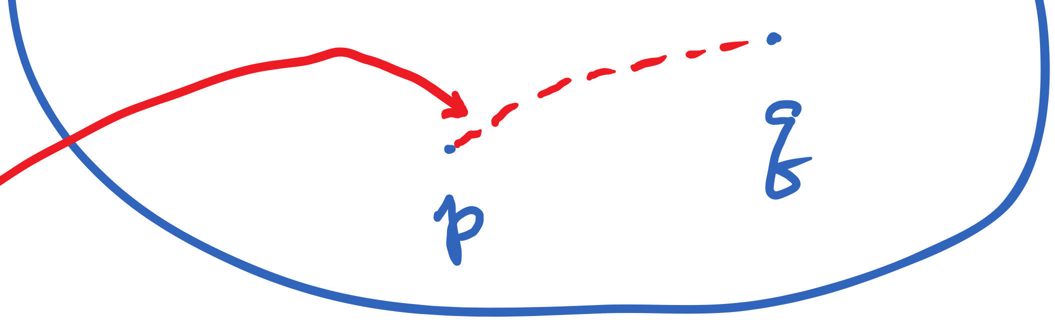
picture X = (x, y); Boltzmann machine X = $\{0, 1\}^n$ x

probability distributions p(x)

Geometry of Distributions S = {p(x)}

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





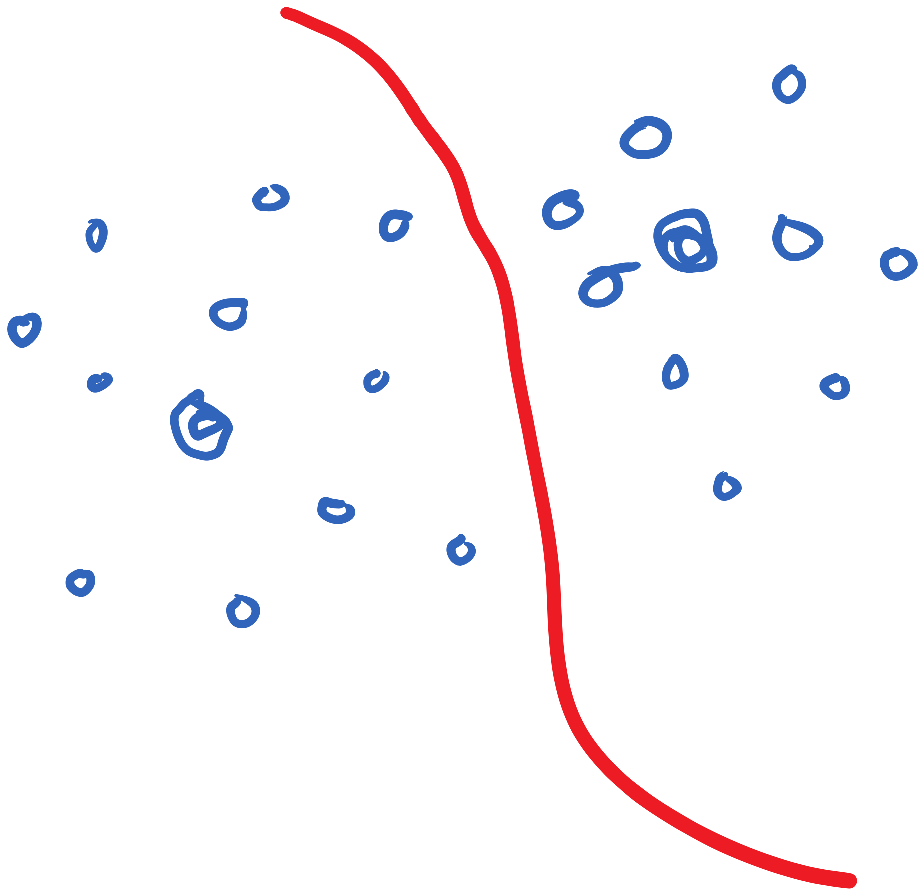
ダイバージェンス

クラスタリング

Patパターン認識

機械学習

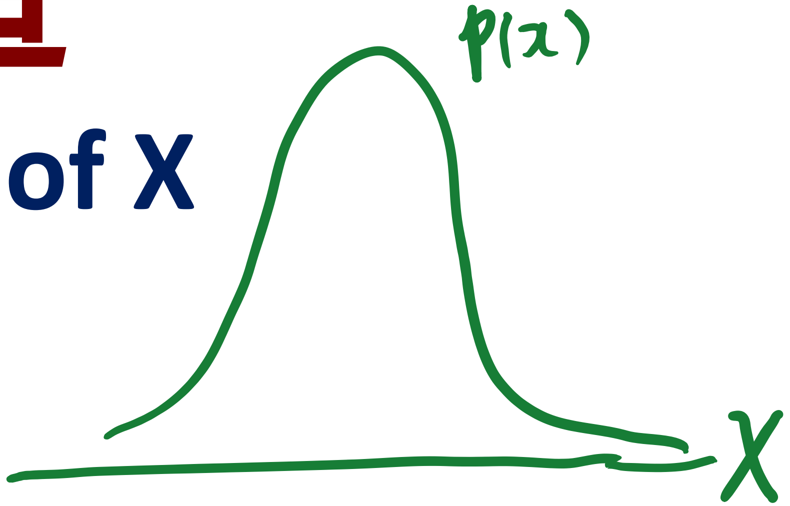
統計的推論



情報幾何と不変構造

invariant under transformations of X

p(x) ~ p(y)



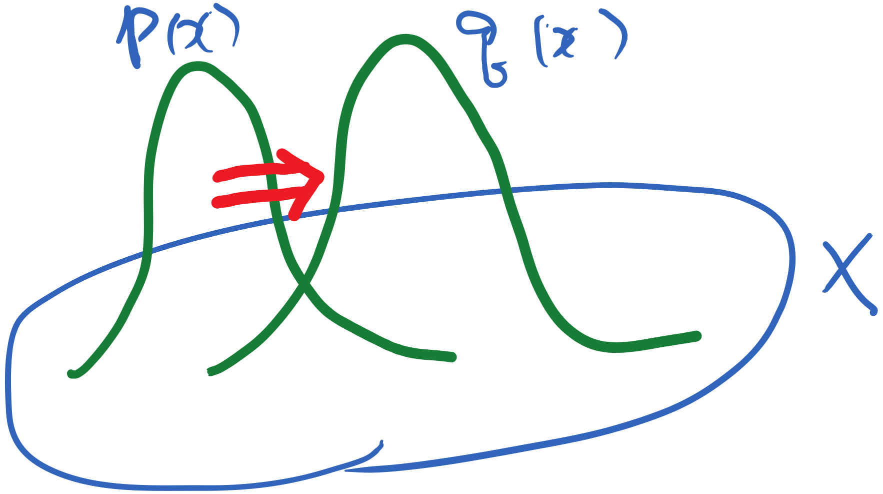
Fisher Information:

Affine connections: α-connections

Duality: Dually coupled Riemannian manifold

Wasserstein 距離

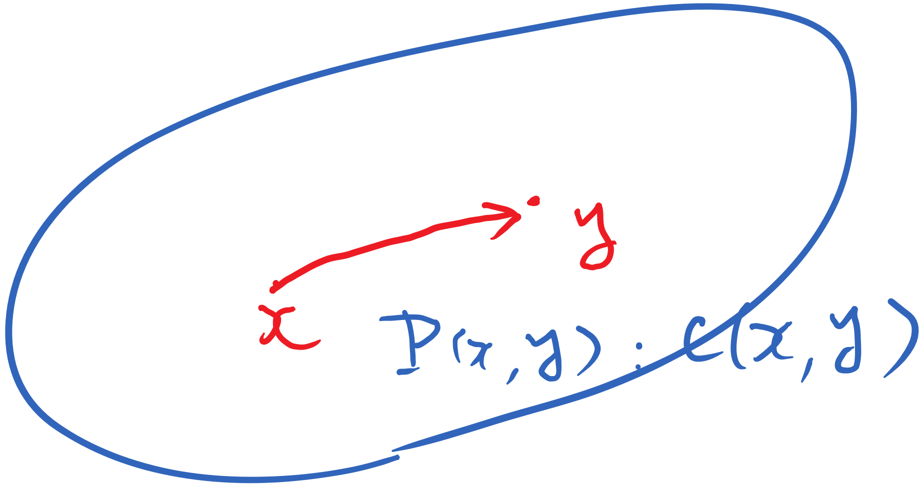
—Monge, Kantrovich



輸送問題 p(x)→q(x)

cost c(x, y) = metric over X

輸送計画 P(x, y)



minimize <c, P> =

$$\int c(x,y)P(x,y)\,dx\,dy$$

線形計画問題

minimize <c, P> $= \int c(x,y) P(x,y)\, dx\, dy$

under constraints

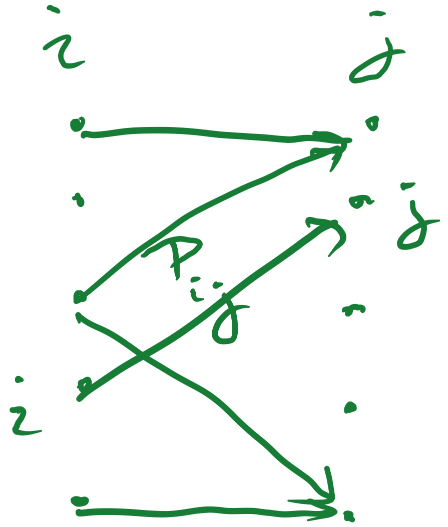
$\int P(x, y)dy = p(x)$

$\int P(x, y)dx = q(y)$

Discrete case : i→j

minimize $\langle c, P \rangle = \sum c_{ij} P_{ij}$

constraints $\sum_j P_{ij} = p_i$, $\sum_i P_{ij} = q_j$



Entropy-正則化輸送問題

Marco Cuturi

$$\min_{P} \; F = \langle c, P \rangle \; - \; \lambda H[P(x, y)]$$

$\lambda \to 0$ Wasserstein

$\lambda \to \infty$ entropy term $H[P(x, y)]$

$P(x, y) = p(x)q(y)$ --- $KL[p(x): q(x)]$

information geometry

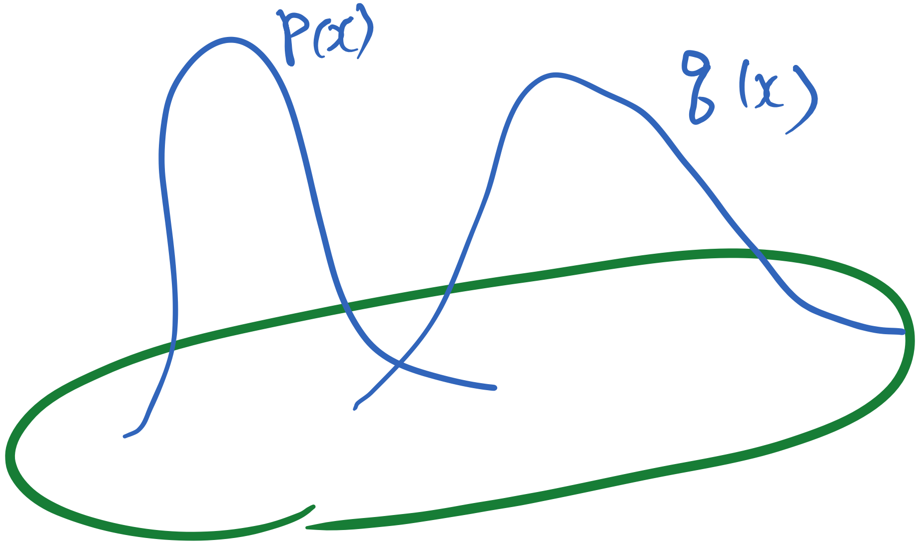
パターンp(x) と q(x)の距離?

KL-divergence
Hellinger distance
Wasserstein distance

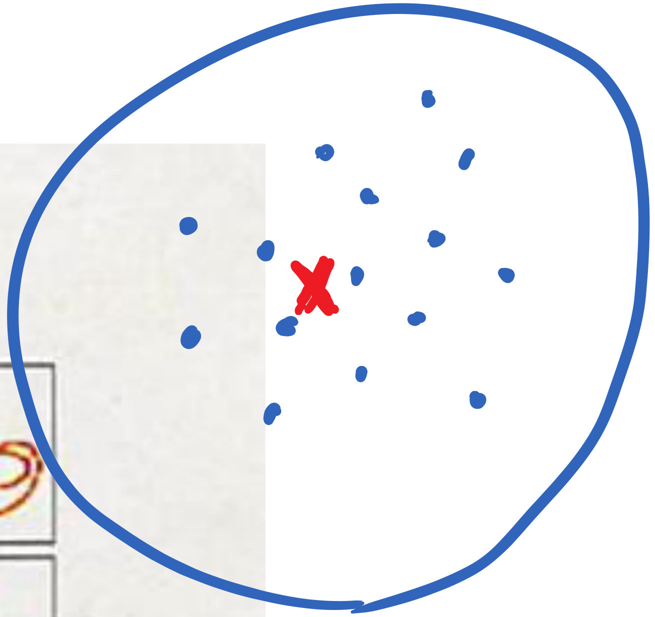
$$D_{KL} = \int p(x) \log \frac{p(x)}{q(x)} dx$$

$$D_{Hell} = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$$

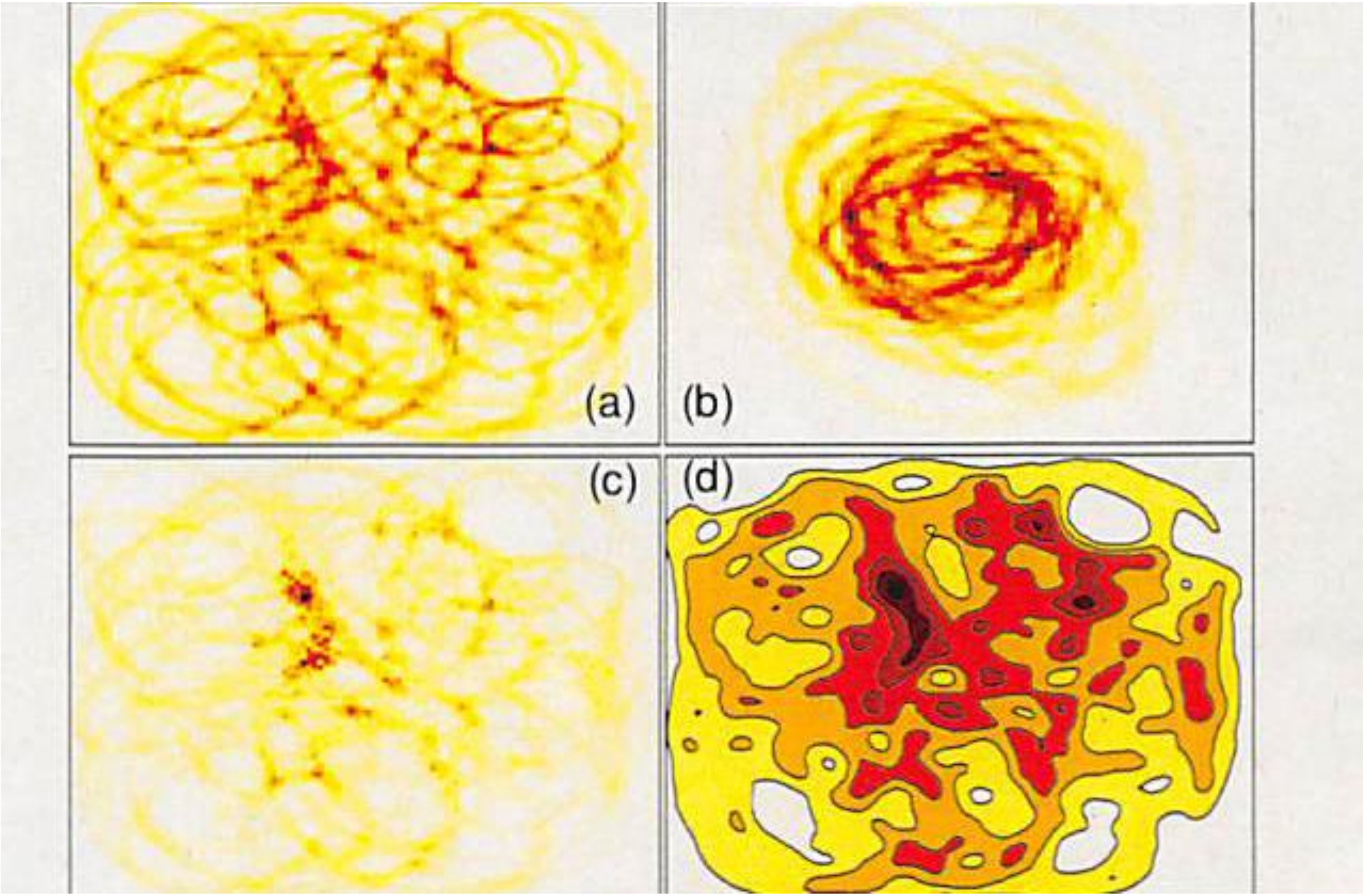
$$D_{Wass} = \int c(x,y) P(x,y) dxdy$$

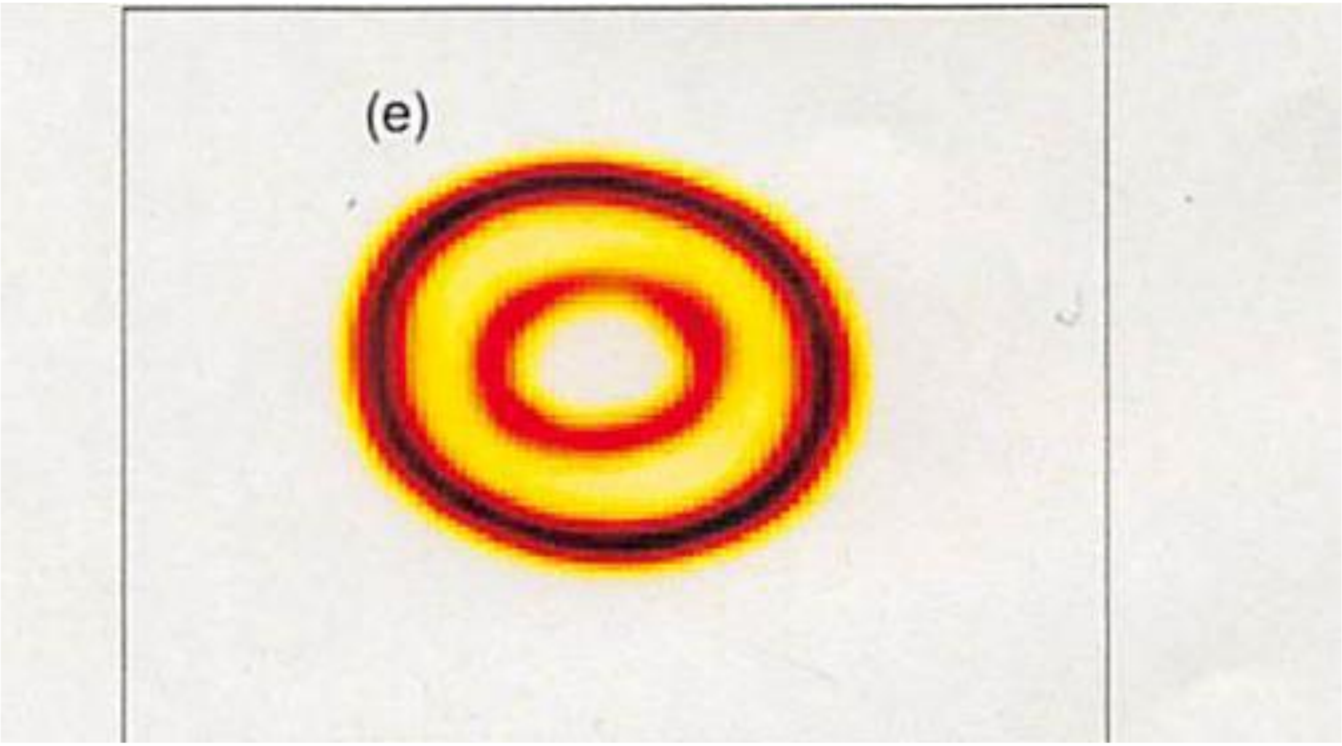


Cuturi: cluster center of double circles



Surprising Results!! Cuturi





D Wass

Discrete Case

Minimize F

$$F_\lambda(\mathbf{P}) = \frac{1}{1+\lambda}\langle \mathbf{c}, \mathbf{P}\rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P})$$

constraints

$$c(\mathbf{P}) = \langle \mathbf{c}, \mathbf{P}\rangle = \sum c_{ij} P_{ij}.$$

$$\sum_j P_{ij} = p_i, \quad \sum_i P_{ij} = q_j, \quad \sum_{ij} P_{ij} = 1.$$

Optimal Transportation Plan

$$L_\lambda(\mathbf{P}) = \frac{1}{1+\lambda}\langle \mathbf{c}, \mathbf{P}\rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P}) - \sum_{i,j} (\alpha_i + \beta_j) P_{ij}.$$

$$\frac{\partial}{\partial P_{ij}} L_\lambda(\mathbf{P}) = \frac{1}{1+\lambda} c_{ij} + \frac{\lambda}{1+\lambda} \log P_{ij} - \alpha_i - \beta_j + \frac{\lambda}{1+\lambda}.$$

$$P_{ij} = \exp\left\{ -\frac{c_{ij}}{\lambda} + \frac{1+\lambda}{\lambda} (\alpha_i + \beta_j + 1) \right\}.$$

$$K_{ij} = \exp\left\{-\frac{c_{ij}}{\lambda}\right\},$$

$$a_i = \exp\left(\frac{1+\lambda}{\lambda}\alpha_i\right) \quad b_j = \exp\left(\frac{1+\lambda}{\lambda}\beta_j\right),$$

the optimal solution is written as

$$P^*_{ij} = a_i b_j K_{ij},$$

Exponential Family of Optimal Transportation Plans

$$P(x) = \sum_{i,j=1}^{n} P_{ij}\delta_{ij}(x). \qquad \theta^{ij} = \log \frac{P_{ij}}{P_{nn}}, \quad \boldsymbol{\theta} = \left(\theta^{ij}\right),$$

$$P(x, \boldsymbol{\theta}) = \exp\left\{\sum_{i,j} \theta^{ij}\delta_{ij}(x) + \log P_{nn}\right\}$$

$$P(x, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \exp\left[\sum_{i,j}\left\{\frac{\lambda+1}{\lambda}(\alpha_i + \beta_j) - \frac{c_{ij}}{\lambda}\right\}\delta_{ij}(x) - \frac{(\lambda+1)}{\lambda}\psi\right]$$

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\lambda}{1+\lambda}\log\sum_{i,j}\exp\left\{\frac{\lambda+1}{\lambda}(\alpha_i + \beta_j) - \frac{1}{\lambda}(c_{ij})\right\}$$

$$\theta^{ij} = \frac{1+\lambda}{\lambda}(\alpha_i + \beta_j) - \frac{c_{ij}}{\lambda}$$

$$\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{1+\lambda}\langle \mathbf{c}, \mathbf{P}\rangle + \frac{\lambda}{1+\lambda}\sum_{i,j} P_{ij}\left\{\frac{1+\lambda}{\lambda}(\alpha_i + \beta_j) - \frac{c_{ij}}{\lambda} - \frac{(1+\lambda)}{\lambda}\psi_\lambda\right\} \quad (32)$$

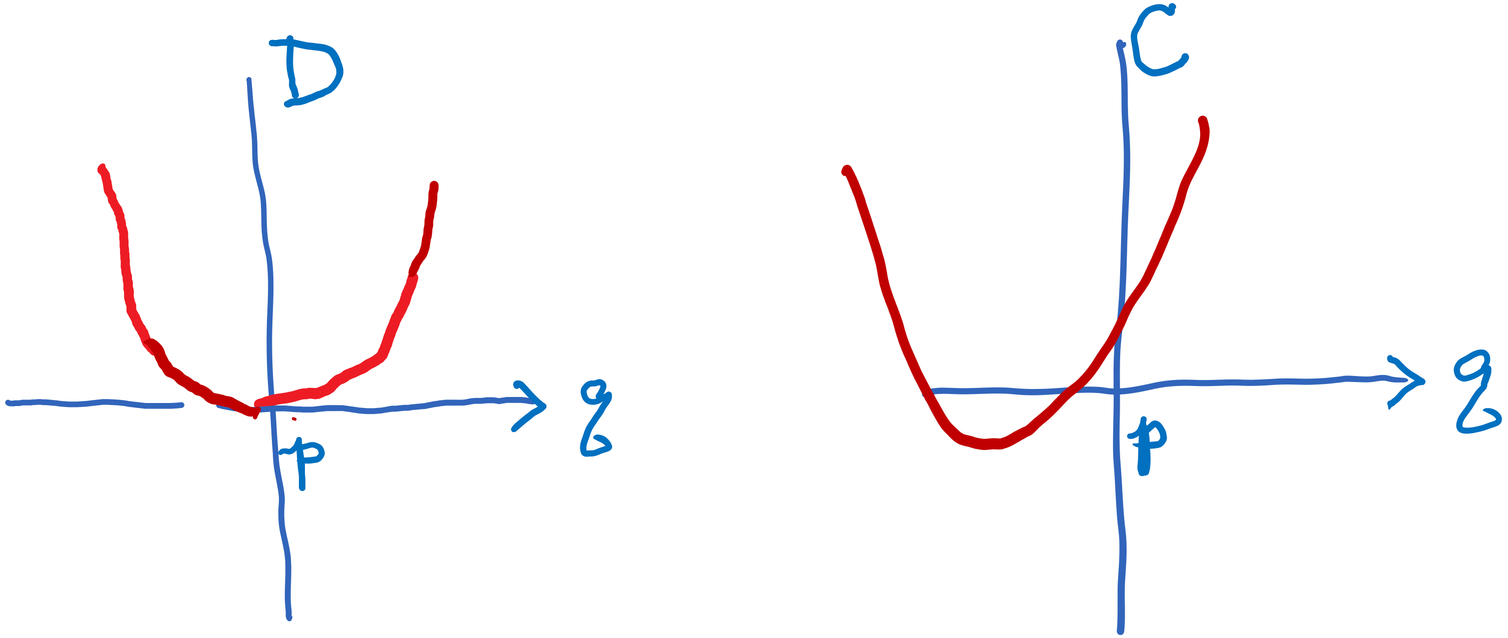
$$= \boldsymbol{p}\cdot\boldsymbol{\alpha} + \boldsymbol{q}\cdot\boldsymbol{\beta} - \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}). \quad (33)$$

$$\psi_\lambda(\boldsymbol{\theta}) + \varphi_\lambda(\boldsymbol{\eta}) = \boldsymbol{\theta}\cdot\boldsymbol{\eta}, \qquad \boldsymbol{\eta} = (\boldsymbol{p}, \boldsymbol{q})^T, \boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^T$$

C関数には問題あり

$$C_\lambda(p,q) \Rightarrow D_\lambda(p,q)$$

$q = p$ is not the minimizer of $C_\lambda(p,q)$



新しいダイバージェンス：その幾何学

$$D_\lambda(p:q) = C_\lambda(p:K_\lambda q) - C_\lambda(p:K_\lambda p)$$

K_λ: diffusion operator

$$\tilde{D}_\lambda(p:q) = C_\lambda(p:q) - \frac{1}{2}\{C_\lambda(p:p) - C_\lambda(q:q)\}$$

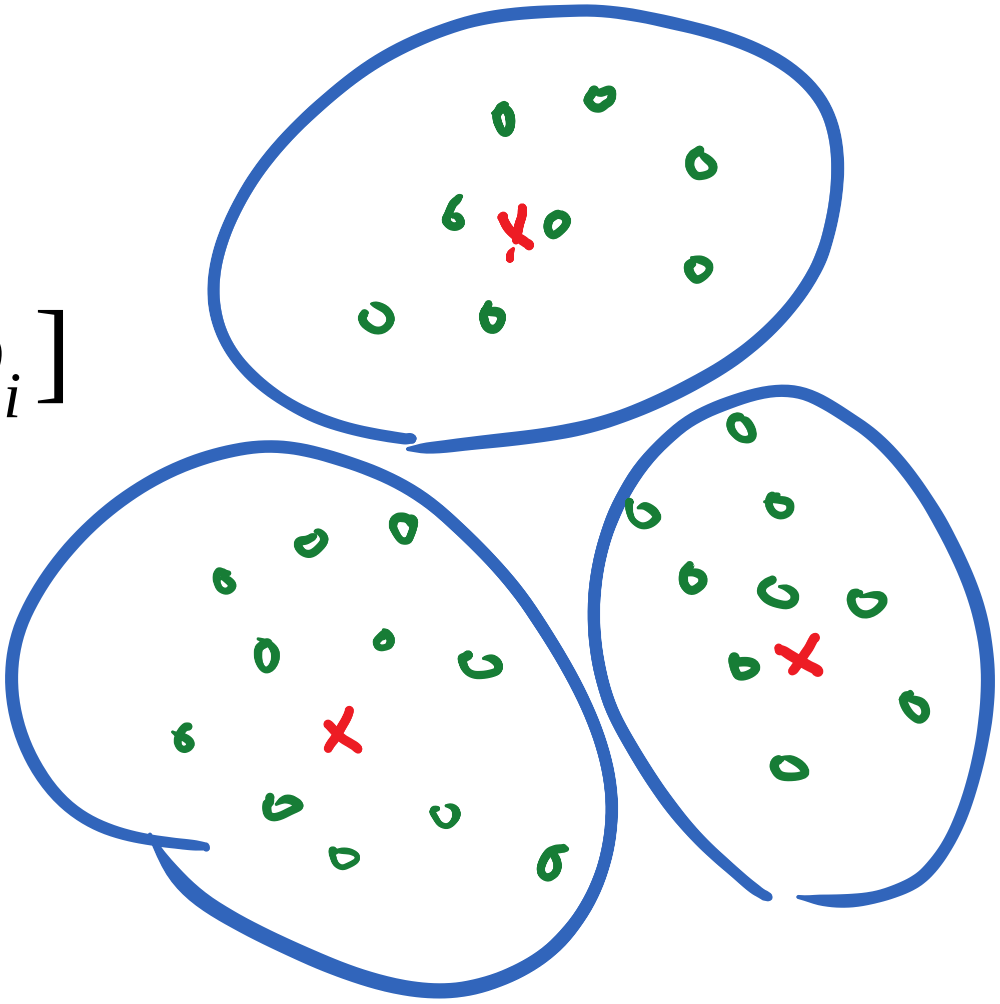
エントロピー制約の一般化

qーエントロピー

$$\min \ F = \langle c, P \rangle - \lambda H[P(x, y)]$$

clustering

$$p_{center} = \arg\min_{p} \sum D[p : p_i]$$



W-GAN

$$D[p_r : p_g] = E_{pr}[\log D(x)] + E_{pg}[1 - \log D(G(z))]$$

$$\mathrm{KL}[p_r : p_m] + \mathrm{KL}[p_g : p_m]$$

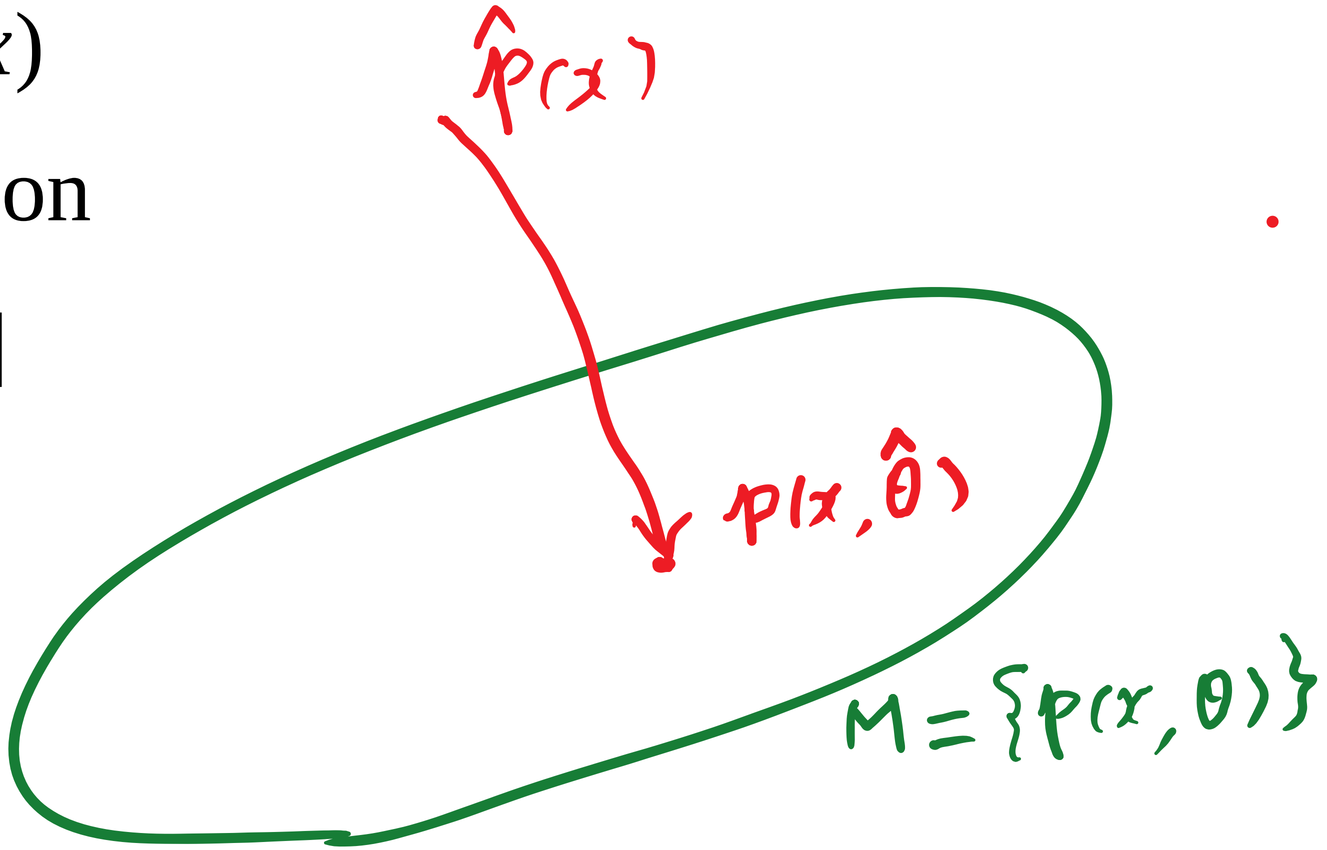
$$\mathrm{D}_{Wass}$$

W-statistics vs likelihood statistics

$$p(x,\theta): \quad x_1, x_2, \ldots, x_N \rightarrow \hat{p}(x)$$

empirical distribution

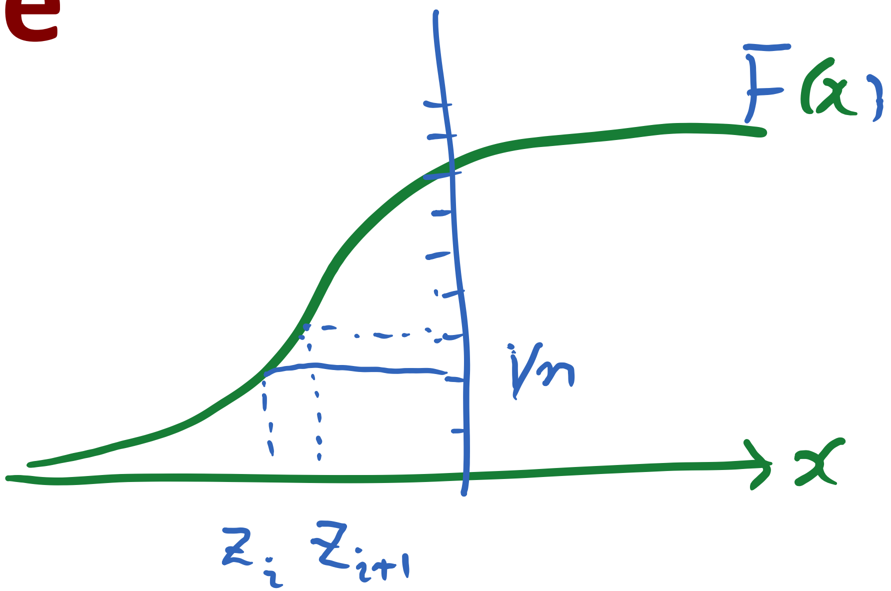
$$\hat{\theta} = \operatorname{argmin} D[\hat{p}(x) : p(x,\theta)]$$



Estimation : Gaussian case

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

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



$$KL: \quad \bar{x} = \frac{1}{N}\sum x_i; \quad \hat{\sigma}^2 = \frac{1}{N}\sum (x_i - \mu)^2$$

$$D_W: \quad \bar{x} = \frac{1}{N}\sum x_i; \quad \hat{\sigma} = \frac{1}{N}\sum z_i x_i$$

$$z_i = F^{-1}(\frac{i}{N})$$

W-statistics: λ = 0,   X = R

Model: p(x, ξ) $$x_1 \le x_2 \le \cdots \le x_n$$

Observed data

Partition points $$z_i(\xi) = P(\frac{i}{n}) = \int_{-\infty}^{i/n} p(x,\xi)dx$$

Optimal transport plan

$$x_i \to z_i(\xi)$$

Cost

$$C(\xi) = \tfrac{1}{n}\sum_i |x_i - z_i(\xi)|^2$$

Estimating equation

$$\sum_i \{x_i - z_i(\xi)\}\partial_\xi z_i(\xi) = 0$$

Consistency and efficiency

$$\lim_{n\to\infty} \mathrm{E}[\hat{\xi}] = \xi$$

$$\mathrm{V}[\hat{\xi}] = \tfrac{1}{n} G^{-1} H G^{-1}$$

$$G(\xi) = \int \partial_{\xi} P(x,\xi) \{\partial_{\xi} P(x,\xi)\}^{T} dx$$

$$= \partial_{\xi} \partial_{\xi} C(\xi', \xi)_{|\xi'=\xi}$$

$$H(\xi) = \int P^{2}(x,\xi) \partial_{\xi} P(x,\xi) \{\partial_{\xi} P(x,\xi)\}^{T} dx$$

Information Geometry of Sinkhorn Algorithm

Obtaining a and b in P* = ca b K

$M_p = \{P_{ij} \mid \sum_j P_{ij} = p_i\}$

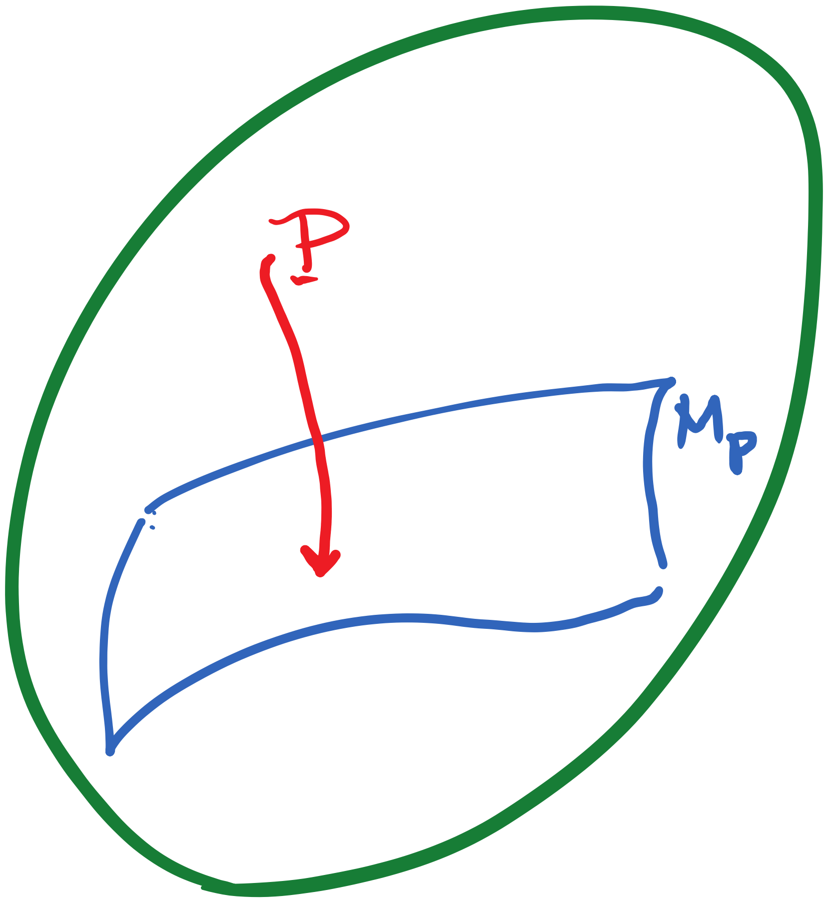
e-projection of P to M :

$$M_p = \{P_{ij} \mid \sum_j P_{ij} = p_i\}$$

$$M_q = \{P_{ij} \mid \sum_i P_{ij} = q_j\}$$

$$P_{ij} \to a_i P_{ij} \qquad a_i = \frac{p_i}{\sum_j P_{ij}}$$

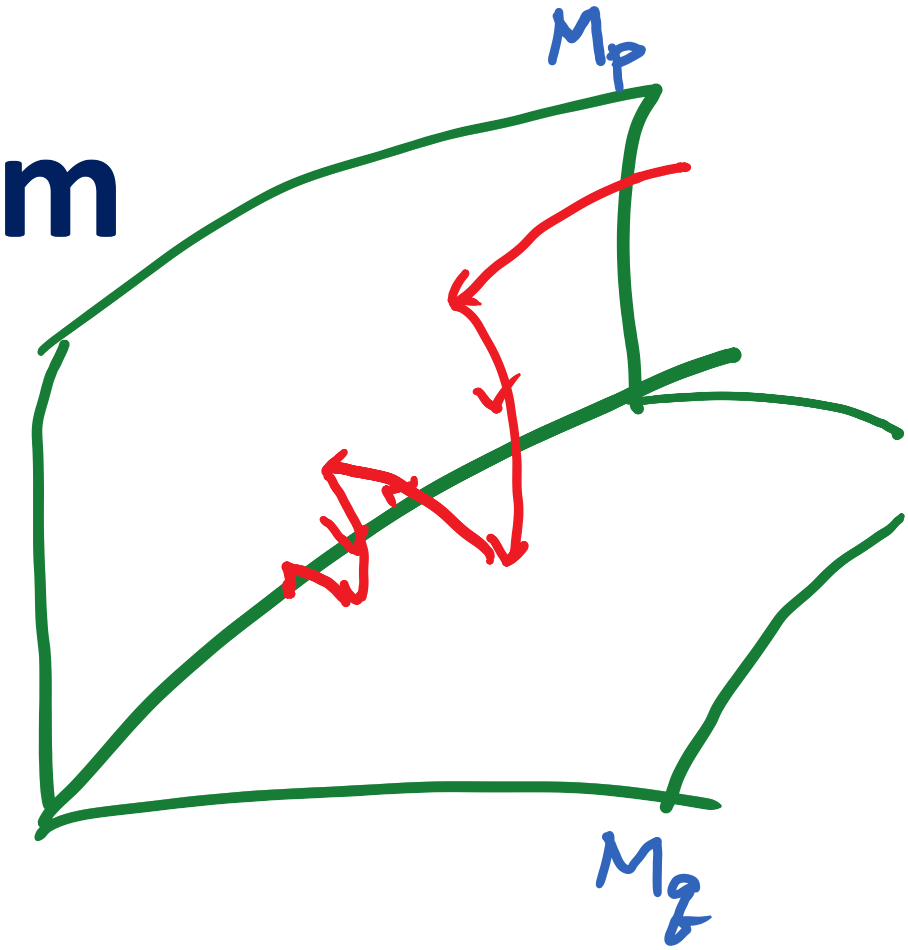
$$P_{ij} \to b_j P_{ij} \qquad b_j = \frac{q_j}{\sum_i P_{ij}}$$



Iterative Algorithm

e-projection to M
e-projection to M

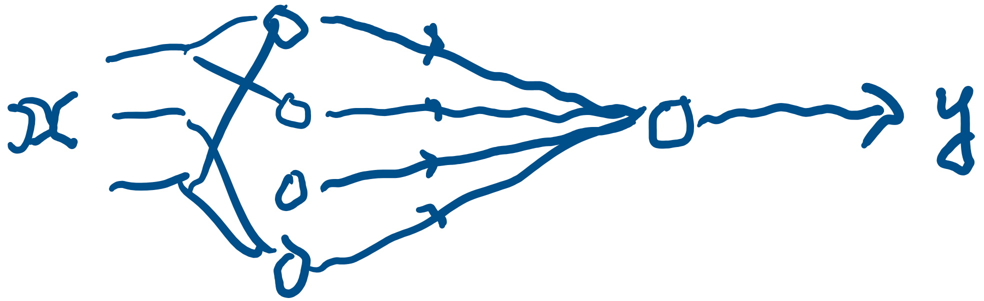
easy to solve (not LP)



3層パーセプトロン学習のW幾何

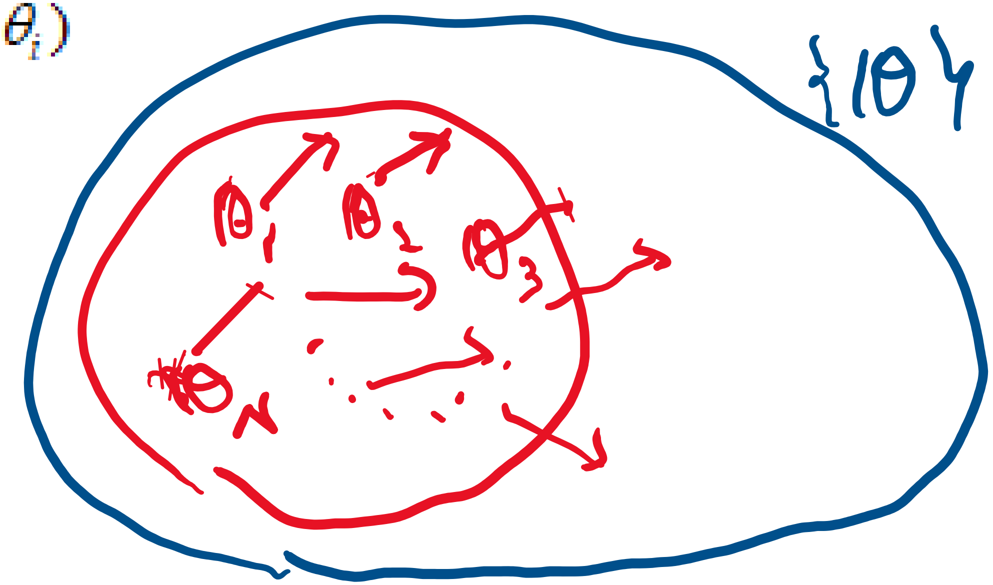
$$y = f(\mathbf{x},\Theta) = \frac{1}{N}\Sigma f(\mathbf{x},\theta_i);$$

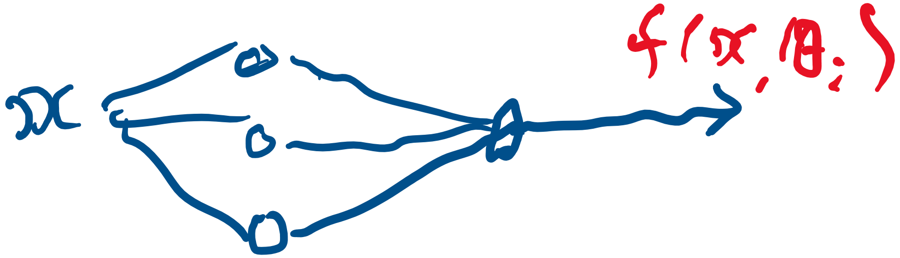
$$f(\mathbf{x},\theta_i) = v_i\phi(\mathbf{w}_i\cdot\mathbf{x}+b_i)$$



$$l = \frac{1}{2}\{y - f(\mathbf{x},\Theta)\}^2 = \frac{1}{2}y^2 - \frac{1}{N}\sum y f(\mathbf{x},\theta_i) + \frac{1}{2N^2}\sum f(\mathbf{x},\theta_i)f(\mathbf{x},\theta_j)$$

$$\dot{\theta}_i = \frac{\partial l}{\partial\theta_i} = v_i(\mathbf{x},y,\theta_i)$$





$$l = \frac{1}{2}\{y - f(\mathbf{x},\Theta)\}^2 = \frac{1}{2}y^2 - \frac{1}{N}\sum y f(\mathbf{x},\theta_i) + \frac{1}{2N^2}\sum f(\mathbf{x},\theta_i) f(\mathbf{x},\theta_j)$$

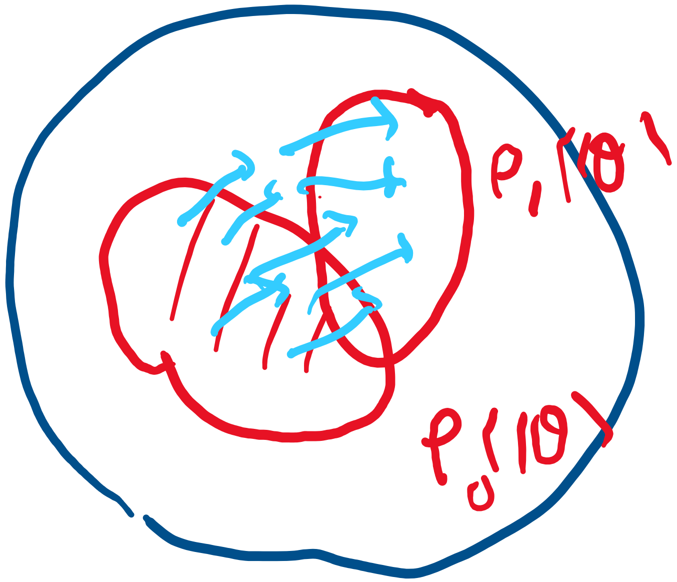
$$\rho(\theta) = \frac{1}{N}\sum \delta(\theta - \theta_i)$$

$$V(\theta) =< y f(\mathbf{x},\theta_i >_{\rho}; \quad U(\theta) = \frac{1}{2} < f(\mathbf{x},\theta) f(\mathbf{x},\theta') >_{\rho,\rho'}$$

$$\Psi(\theta) = V(\theta) + < U(\theta,\theta' >_{\rho'}$$

$$v_i(\mathbf{x},y,\Theta) = \nabla\Psi(\Theta)$$

$$\dot{\rho}_t(\theta) = \eta\nabla_{\theta}\cdot\{\rho_t(\theta)\nabla\Psi(\Theta)\}$$



深層学習とW幾何　　拡散モデル

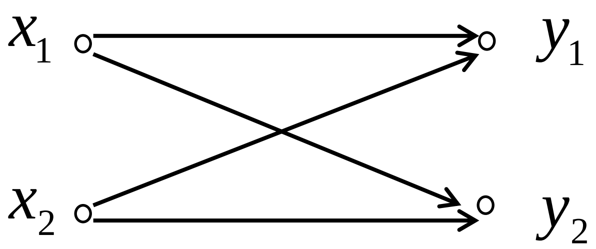
3層パーセプトロン学習のW幾何

アファイン変換モデルのW幾何

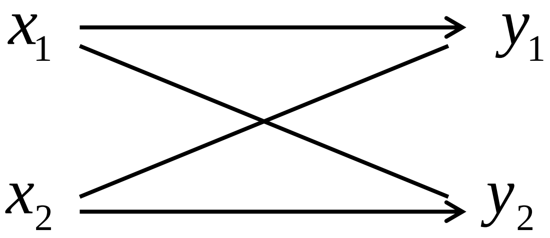
Computational Information Geometry for Image and Signal Processing
Edinburgh -- 2015

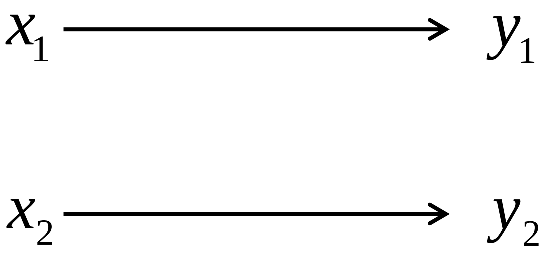
Information Integration and Complexity of Systems

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



Shun-ichi Amari (*RIKEN Brain Science Institute)*
Masafumi Oizumi (RIKEN BSI, Monash U.)
Naotsugu Tsuchiya (Monash U.)





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

split model: $S_S = \{q(\mathbf{x}, \mathbf{y})\}$

$$q(\mathbf{y} \mid \mathbf{x}) = \Pi q(y_i \mid x_i)$$

measure of interaction : N. Ay

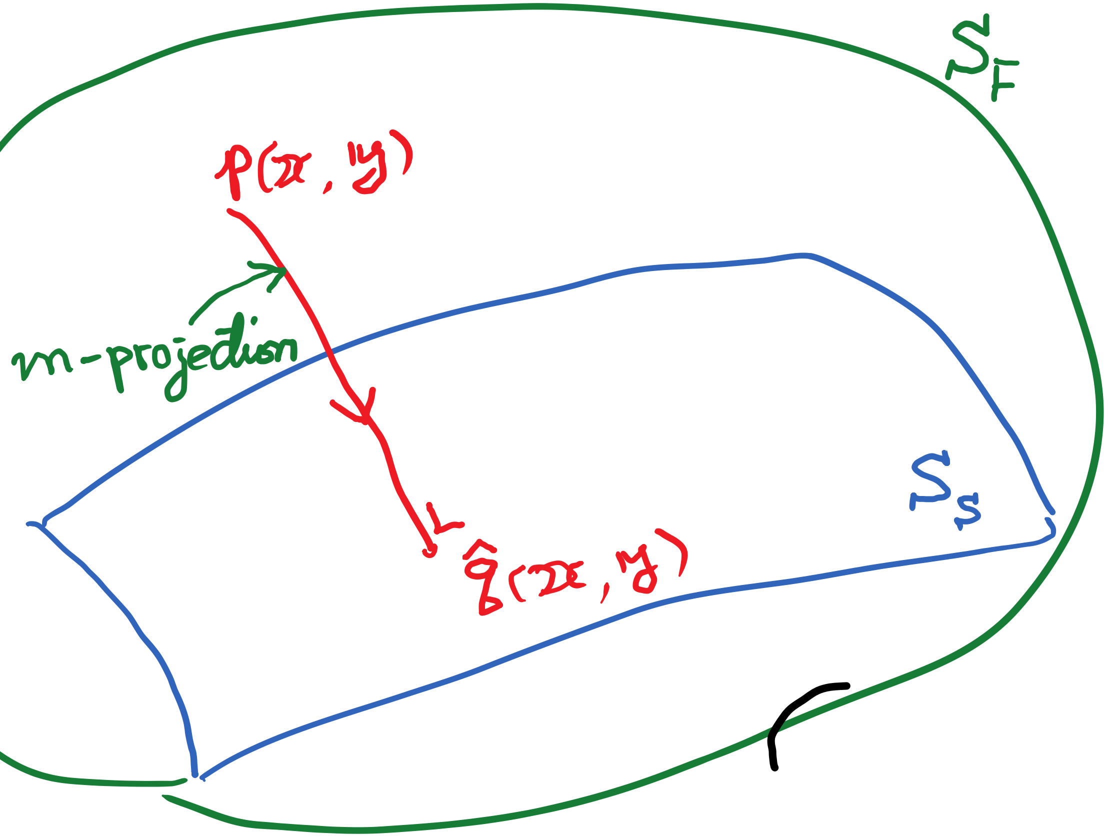
information integration : Tononi

Barrett and Seth

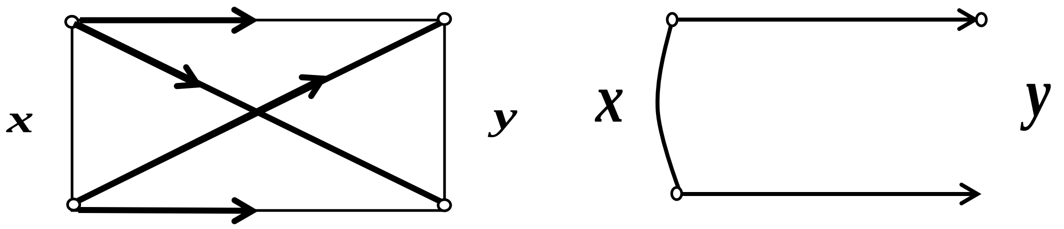
Measure of information integration, or system complexity Φ

Information Geometry N. Ay

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



Split Model S_H : Ay, Barrett & Seth



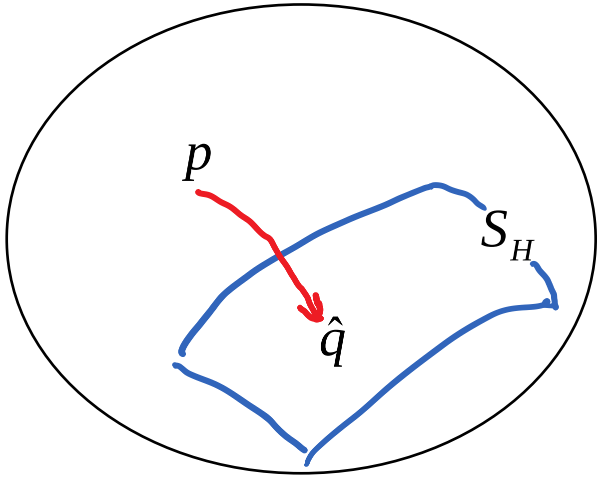
$$q(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{x})\prod q(y_i|x_i)$$

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

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

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

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



Split Model S_G

graphical model

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

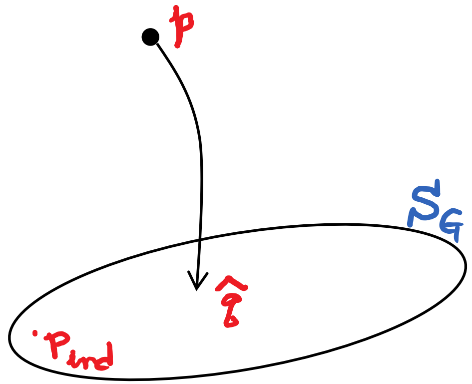
$$\theta_{12}^{XY} = \theta_{21}^{XY} = 0$$

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

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

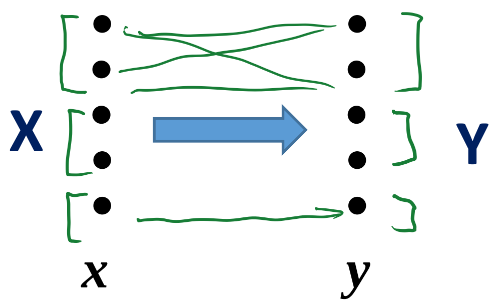
$$\hat{q}_X(\boldsymbol{x}) = p_Y(\boldsymbol{x}), \qquad \hat{q}_Y(\boldsymbol{y}) = p_Y(\boldsymbol{y})$$

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



Hierarchy: transfer entropy

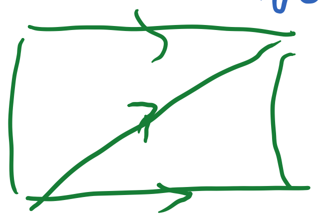
Partition of X



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