APSIPA-2009-Sapporo

Divergence, Signal Decomposition and Information Geometry

Shun-ichi Amari
RIKEN Brain Science Institute

Signal and Information Processing

Signal and information:

Probability distributions and positive arrays

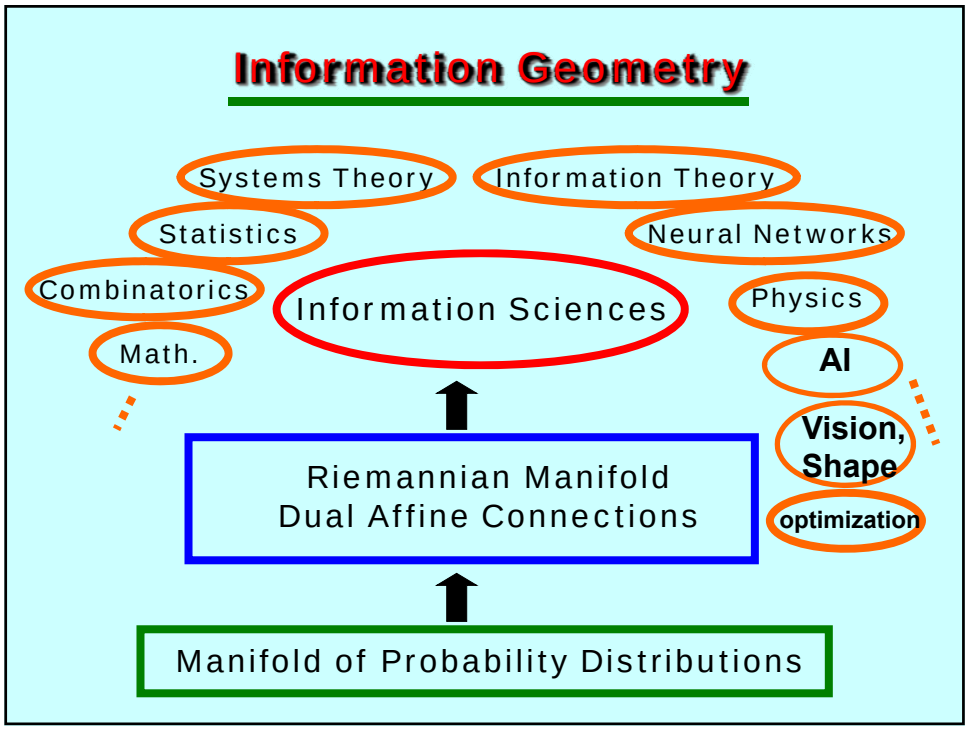
Information Geometry

Divergence and Geometry

decomposition of signals

ICA (Independent Component Analysis)
NMF (Non-Negative Matrix Factorization)
Sparse representation of Signals

LASSO and LARS



Manifold of Probability Distributions:

$$p(x;\theta) = \exp\{\sum \theta_i x_i - \psi(\theta)\}$$

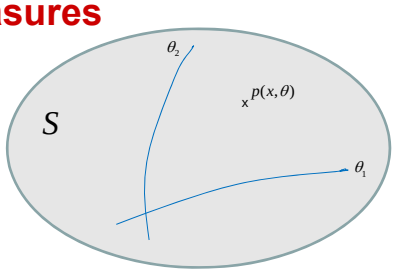
$$p(x) = \sum p_i \delta_i(x): \quad \mathbf{p} = (p_1, ..., p_n); \quad \sum p_i = 1$$

Manifold of positive measures

$$\mathbf{m} = (m_1, \ ..., \ m_n) \quad m_i > 0$$

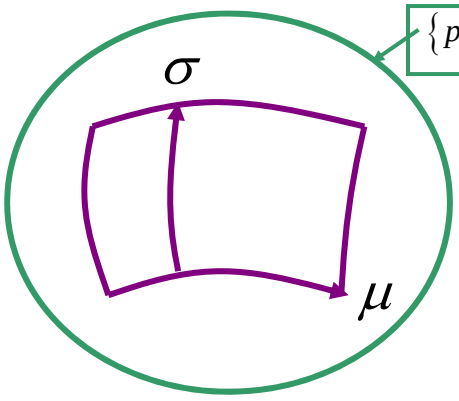
$$A = (A_{ij}), \quad A_{ij} > 0$$

$s(x, y)$: vision;



Information Geometry ?

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



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

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

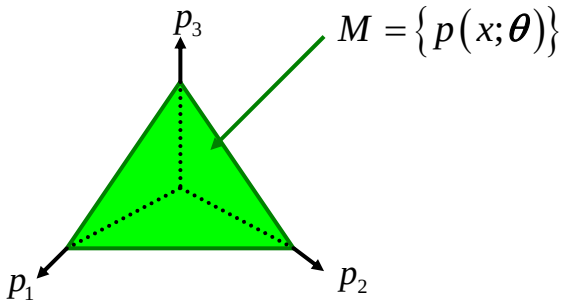
Riemannian metric

Dual affine connections

Manifold of Probability Distributions

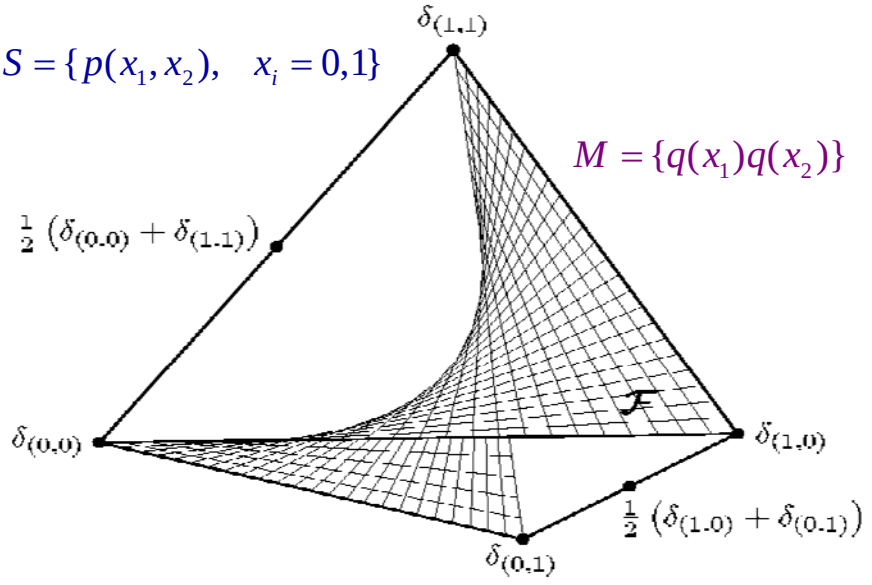
$$x = 1, 2, 3 \qquad \{p(x)\}$$

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



Independent Distributions

$S = \{p(x_1, x_2), \quad x_i = 0,1\}$



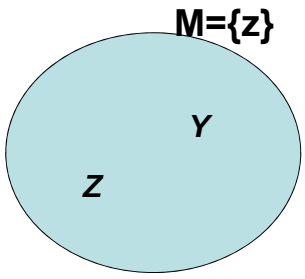
Divergence $D[z:y]$

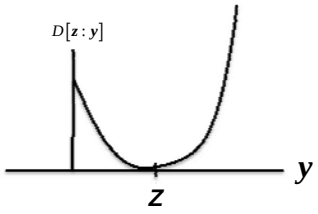
$$D[z:y] \geq 0$$

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

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

positive-definite





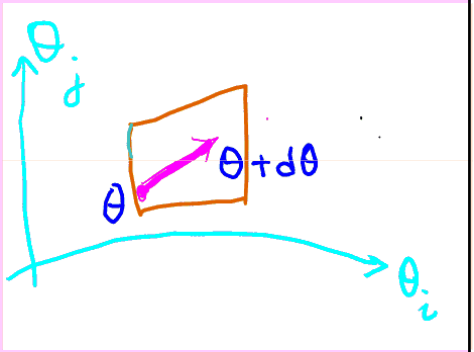
Riemannian Structure

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

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

Euclidean $G = E$

Fisher information



Metric and Connections Induced by Divergence

(Eguchi)

$$g_{ij}(\boldsymbol{z}) = \partial_i \partial_j D[\boldsymbol{z} : \boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{z}}$$

$$\Gamma_{ijk}(\boldsymbol{z}) = -\partial_i \partial_j \partial'_k D[\boldsymbol{z} : \boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{z}}$$

$$\Gamma^*_{ijk}(\boldsymbol{z}) = -\partial'_i \partial'_j \partial_k D[\boldsymbol{z} : \boldsymbol{y}]_{|\boldsymbol{y}=\boldsymbol{z}}$$

$$\partial_i = \frac{\partial}{\partial z_i}, \quad \partial'_i = \frac{\partial}{\partial y_i}$$

Affine Connection

covariant derivative

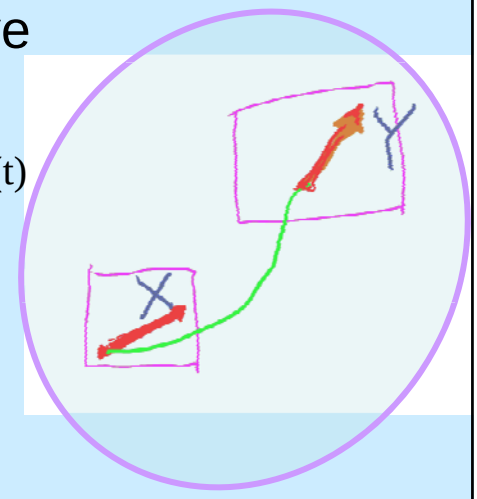
$\nabla_X Y, \quad \Pi_c X = Y$

geodesic $\nabla_{\dot{X}} \dot{X} = 0, \quad X=X(t)$

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

minimal distance

straight line



Duality:

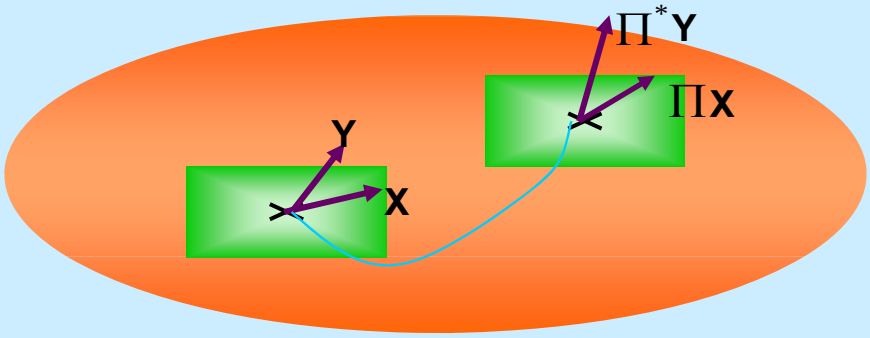
$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma^*_{kji}$$

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

$$\{S, g, T\}$$

Duality

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



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

Dual Affine Connections (∇, ∇^*)

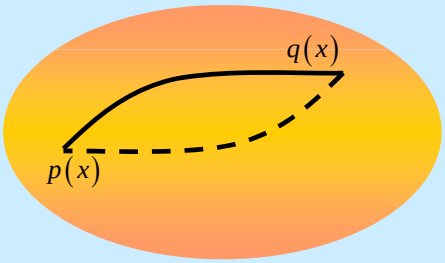
(Π, Π^*)

e-geodesic

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

m-geodesic

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



Invariance

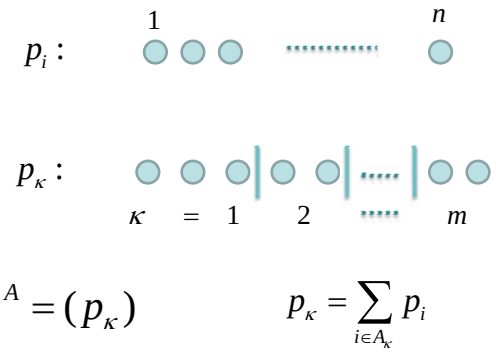
$y = k(x)$: one-to-one

sufficient statistics

$$D\left[p_X(x) : q_X(x)\right] = D\left[p_Y(y) : q_Y(y)\right]$$

Invariance: invariant divergence

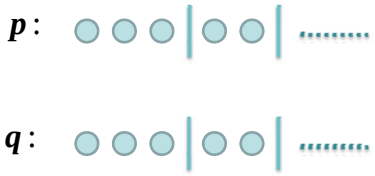
--- characterization of *f*-divergence



$$p^A = (p_\kappa) \qquad p_\kappa = \sum_{i \in A_\kappa} p_i$$

$$D[\boldsymbol{p}:\boldsymbol{q}] \geq D[\boldsymbol{p}^A:\boldsymbol{q}^A]$$
$$D[\boldsymbol{p}:\boldsymbol{q}] = D[\boldsymbol{p}^A:\boldsymbol{q}^A]$$

$$\Leftrightarrow p_i = c_\kappa q_i \ ; i \in A_\kappa$$



Invariance ⇒ *f*-divergence

Csiszar *f*-divergence

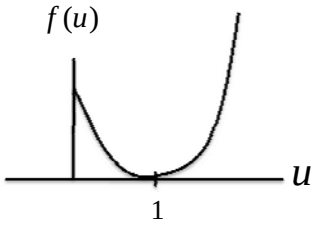
$$D_f[\boldsymbol{p}:\boldsymbol{q}] = \sum p_i f\left(\frac{q_i}{p_i}\right),$$

$$f(u):\ \text{convex}, \qquad f(1) = 0,$$

$$D_{cf}[\boldsymbol{p}:\boldsymbol{q}] = cD_f[\boldsymbol{p}:\boldsymbol{q}]$$

$$\tilde{f}(u) = f(u) - c(u-1)$$

$$f(1) = f'(1) = 0\ ;\ f''(1) = 1$$



Invariant Geometrical Structure:

Fisher information metric

alpha-affine connections : dual

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

$$g_{ij} = E[\frac{\partial}{\partial\theta_i} l(x,\theta) \frac{\partial}{\partial\theta_j} l(x,\theta)]$$

$$T_{ijk} = E[\frac{\partial}{\partial\theta_i} l(x,\theta) \frac{\partial}{\partial\theta_j} l(x,\theta) \frac{\partial}{\partial\theta_k} l(x,\theta)]$$

$$\Gamma_{ijk}^{(\alpha)} = \Gamma_{ijk}^{(0)} \pm \alpha T_{ijk}$$

f divergence of $\tilde{S}$ $\quad \sum \tilde{p}_i > 0$

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

$$D_f[\tilde{\boldsymbol{p}}:\tilde{\boldsymbol{q}}] = 0 \Leftrightarrow \tilde{\boldsymbol{p}} = \tilde{\boldsymbol{q}}$$

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

α Divergence: why?

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

KL-divergence

$$f(u) = u\log u - (u-1)$$

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

α divergence

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

KL-divergence

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

Dually flat manifold

θ-coordinates $\leftrightarrow$ η-coordinates

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

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

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

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

ψ : cumulant generating function

φ : negative entropy

Dually Flat Manifold

1. Potential Functions

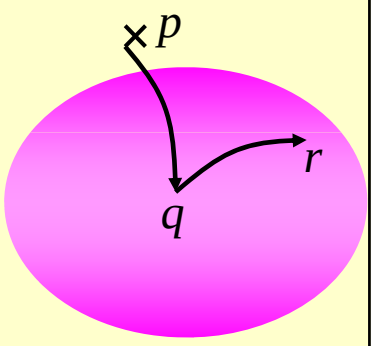
---convex (Bregman divergence, Legendre transformation)

2. Divergence $D[p:q]$

3. Pythagoras Theorem

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

4. Projection Theorem



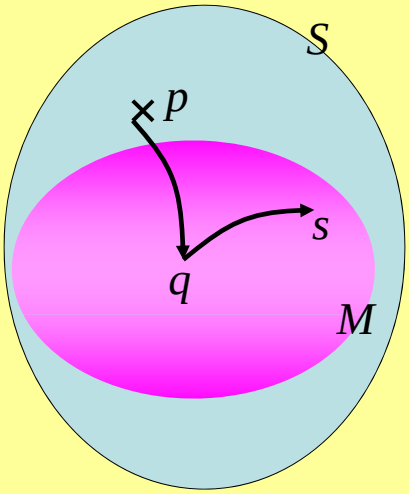
Projection Theorem

$$q=\arg\min_{s\in M} D[p:s]$$

m-geodesic

$$q=\arg\min_{s\in M} D[s:p]$$

e-geodesic



Dually Flat Structure

affine coordinates θ

dual affine coordinates η

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

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

$D(\theta_1 : \theta_2) = \psi(\theta_1) + \varphi(\eta_2) - \theta_1 \cdot \eta_2$

Exponential Family

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

$\exp\{\psi(\theta)\} = \int \exp\{\theta \cdot \boldsymbol{x} + k(\boldsymbol{x}) d\boldsymbol{x}\}$

Mixture Family

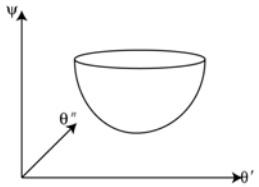
$p(\boldsymbol{x}, \eta) = \sum_{i=0}^{n} \eta_i p_i(\boldsymbol{x})$, $\eta_0 = 1 - \sum_{i=0}^{n} \eta_i$

$\varphi(\eta) = \int p(\boldsymbol{x}, \eta) \log p(\boldsymbol{x}, \eta) d\boldsymbol{x}$

Manifold with Convex Function

S **coordinates** $\boldsymbol{\theta} = (\theta^1, \theta^2, \mathrm{L}\ , \theta^n)$

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



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

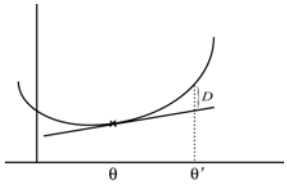
Euclidean $\psi(\boldsymbol{\theta}) = \frac{1}{2} \sum (\theta^i)^2$

mathematical programming, control systems, physics, engineering, economics

Riemannian metric and flatness

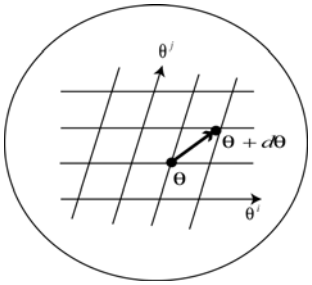
Bregman divergence

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



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

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



flatness θ **:geodesic**

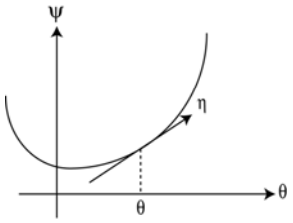
Legendre Transformation

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

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

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

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



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

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

Geometry: Dually Flat Geometry

Riemannian metric $G = (G_{ij})$

$$D_{\psi}[\boldsymbol{p} : \boldsymbol{p} + d\boldsymbol{p}] = \frac{1}{2} \boldsymbol{p}^T G \boldsymbol{p},$$

$$G(\boldsymbol{p}) = \nabla\nabla\psi(\boldsymbol{p})$$
$$G^*(\boldsymbol{p}^*) = \nabla\nabla\psi^*(\boldsymbol{p}^*) = G^{-1}$$

Straightness (affine connection)

$$\boldsymbol{p}(t) = \boldsymbol{a} + \boldsymbol{b}t \qquad : \psi\text{-geodesic}$$
$$\boldsymbol{p}^*(t) = \boldsymbol{a}^* + \boldsymbol{b}^* t \qquad : \psi^*\text{-geodesic}$$

Two flat coordinate systems (θ, η)

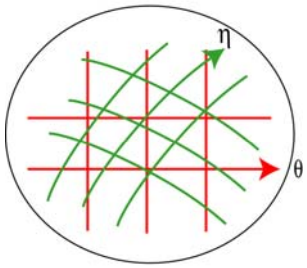
θ : geodesic (e-geodesic)

η : dual geodesic (m-geodesic)

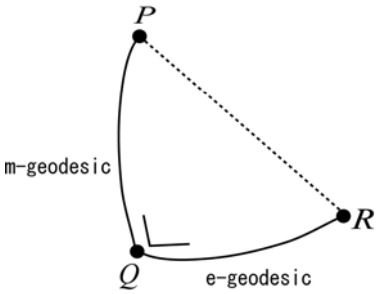
"dually orthogonal"

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

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



Pythagorean Theorem



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

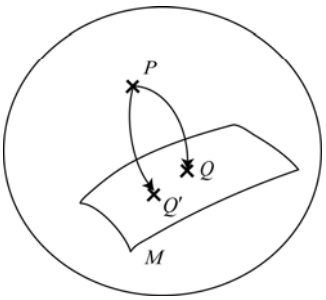
Euclidean space: self-dual $\theta = \eta$

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

Projection Theorem

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

$Q =$ m-geodesic projection of P to M



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

$Q' =$ e-geodesic projection of P to M

divergence

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

invariance

dually flat space

***f*-divergence**

Bregman divergence

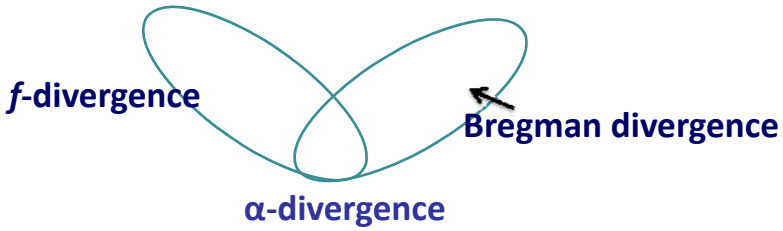
convex functions

Fisher inf metric
Alpha connection

KL-divergence

Space of positive measures : vectors, matrices, arrays

$$\tilde{S} = \{\tilde{\boldsymbol{p}}\}, \quad \tilde{p}_i > 0 : (\sum \tilde{p}_i = 1)$$



α-representation

$$r_i = \begin{cases} p_i^{\frac{1-\alpha}{2}} & , \quad \alpha \neq 1 \\ \log p_i & , \quad \alpha = 1 \end{cases}$$

$$\psi(\boldsymbol{r}) = \sum U(r_i)$$

typical case:

$$U(z) = \begin{cases} z^{\frac{2}{1-\alpha}} & , \quad \alpha \neq 1 \\ e^z & , \quad \alpha = 1 \end{cases}$$

$$U(r_i) = p_i$$

Divergence over α-representation

$$D_\alpha[\tilde{\boldsymbol{p}}:\tilde{\boldsymbol{q}}]=\sum U(r_i)+\sum V(r_i^*)-\sum r_i r_i^*$$

$r_i : \tilde{p}_i^{\frac{1-\alpha}{2}}$ α-geodesic

$r_i^* : \frac{2}{1-\alpha}\tilde{p}_i^{\frac{1+\alpha}{2}}$ $-\alpha$-geodesic

$r_i = \log \tilde{p}_i$ $\alpha = 1$

$r_i^* = p_i$

α divergence

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

$\tilde{S}$: dually flat

S : not dually flat (except $\alpha = \pm 1$)

$$\sum p_i = 1$$

$$\sum r_i^{\frac{2}{1-\alpha}} = 1$$

Integration of Stochastic Evidences

—Information Geometry Approach

Shun-ichi Amari

RIKEN Brain Science Institute

Various Means

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

arithmetic　　geometric　　harmonic

Any other mean?

Generalized mean: f-mean

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

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

scale free　　$m_f(ca,cb) = cm_f(a,b)$

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

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

α-mean : $m_\alpha(p_1(s), p_2(s))$

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

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

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

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

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

α-divergence : positive real numbers

$$D_\alpha[z:y] = \begin{cases} \dfrac{2z}{1+\alpha} + \dfrac{2y}{1-\alpha} - \dfrac{4}{1-\alpha^2} z^{\frac{1-\alpha}{2}} y^{\frac{1+\alpha}{2}}, & \alpha \neq \pm 1 \\ z - y + y\log\dfrac{y}{z}, & \alpha = 1 \\ y - z + z\log\dfrac{z}{y}, & \alpha = -1 \end{cases}$$

α – Family of Distributions

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

mixture family :

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

exponential family :

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

α – Bayes Predictive Distribution

$p(x|\theta);\ \ p(\theta|D)$

$$p_\alpha(x|D) = f_\alpha^{-1}\left[\int f_\alpha\left[p(x|\theta)\right] p(\theta|D) d\theta\right]$$

$\alpha = 1$: predictive distribution

$\alpha = -1$: product of experts

Optimal Property

given data $D=[x_1,\cdots,x_k]\in p(x|\theta_0)$

find $q(x|D)$ that minimizes the $\alpha-$risk

$\min_q \quad R_\alpha\left[p(x|\theta_0):q(x)\right]$

$$R_\alpha[p:q]=\int \pi(\theta)D_\alpha\left[p(x|\theta):q(x;D)\right]p(D|\theta)dDd\theta$$

Theorem

The $\alpha-$predictive distribution minimizes the $\alpha-$ risk.

The Bayes predictive distribution :

The product of experts : $KL\left[q(x|D):p(x|\theta)\right]$

$\alpha=0$ **predictive distribution : Hellinger**

$\alpha=\infty$ **pessimistic ;** $\alpha=-\infty$**: optimistic**

Tsallis q-Entropy

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

$$\ln_q(u) = \begin{cases} \dfrac{1}{1-q}\left(u^{1-q}-1\right), & q \neq 1 \\ \log u, & q = 1 \end{cases}$$

$$\exp_q(u) = \left(\ln_q\right)^{-1}(u)$$

$$H_T = E\left[\ln_q \frac{1}{p(x)}\right]$$

$$D\left[p(x):r(x)\right] = -E_p\left[\ln_q \frac{r(x)}{p(x)}\right]$$

$$= \frac{1}{1-q}\left(1 - \int p(x)^q\, r(x)^{1-q} dx\right)$$

: α-divergence ; α-structure

$$h_q[p] = \int p(x)^q dx$$

$$H_T = \frac{1}{1-q}\left(h_q - 1\right)$$

$$H_R = \frac{1}{1-q}\log h_q$$

$\Rightarrow -h_q(\boldsymbol{p})$: convex

β-structure

q -exponential family

$$p(x,\theta) = \exp_q\left\{\boldsymbol{\theta}\cdot\boldsymbol{x} - \psi(\boldsymbol{\theta})\right\}$$

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

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

$$\eta_i = E_q[x_i]$$

$$\varphi(\boldsymbol{\eta}) = \frac{1}{1-q}\left(\frac{1}{h_q(\boldsymbol{p})} - 1\right) = E_q\left[\ln_q p(x)\right]$$

$$S_n = \left\{\boldsymbol{p}\,\middle|\,\sum_{i=0}^{n} p_i = 1\right\}$$

$$p(x) = \sum_{i=0}^{n} p_i\delta_i(x)$$

$$\theta^i = \frac{1}{1-q}\left(p_i^{1-q} - p_0^{1-q}\right)$$

$$\eta_i = \frac{1}{h_q}p_i^q$$

conformal transformation

q -Fisher information

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

$q-divergence$

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

q-escort probability distribution

$$\hat{p}(x) = \frac{1}{h_q(p)} p(x)^q$$

Escort geometry

$$\hat{g}_{ij}(\theta) = E_q[\partial_i \log \hat{p}(x,\theta) \partial_j \log \hat{p}(x,\theta)]$$

q -Cramer Rao : $p(x,\xi)$

q-unbiased estimator $\hat{\hat{\xi}}$

$$E_q\left[\hat{\hat{\xi}}\right]=\xi$$

$$E_q\left[\left(\xi_i-\hat{\hat{\xi}}_i\right)\left(\xi_j-\hat{\hat{\xi}}_j\right)\right]\geq\left(g_{ij}^{\varepsilon}\right)^{-1}$$

q -maximum likelihood estimator

$$\hat{\hat{\xi}}_q=\arg\max\ \hat{p}\left(x_1,\cdots,x_N;\xi\right)$$

$$=\arg\max\frac{1}{h_q\left(\xi\right)^N}\prod p\left(x_i;\xi\right)^q$$

$\Leftrightarrow$ Bayes MAP with $\pi\left(\xi\right)=h_q\left(\xi\right)^{-N/q}$

q -super-robust estimator (Eguchi)

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

bias-corrected q-estimating function

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

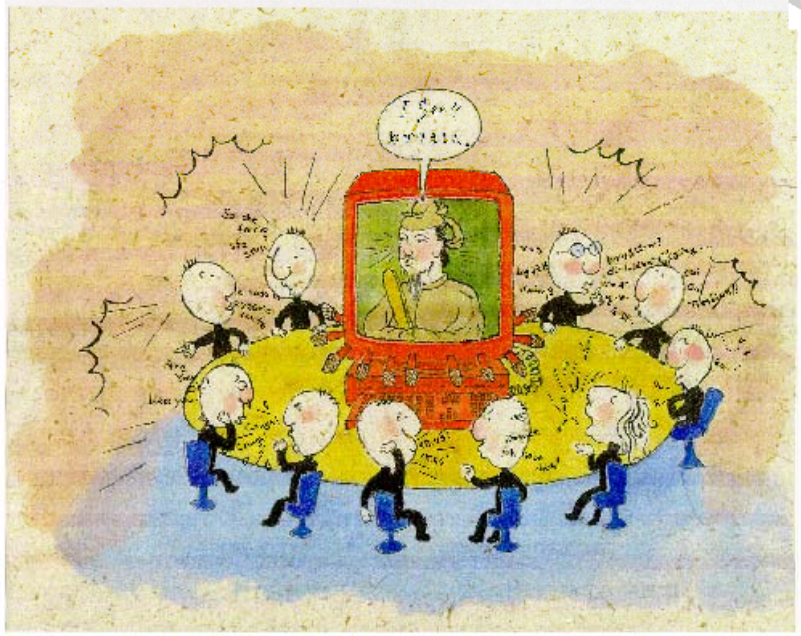
$$c_{q+1} = \frac{1}{q+1}\partial \log h_{q+1}(\xi)$$

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

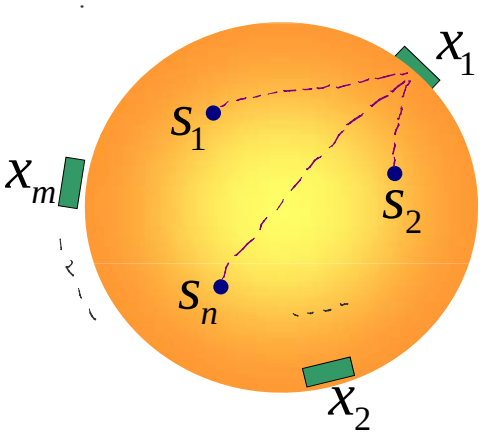
Applications of geometry:

ICA
Independent Component Analysis

聖徳太子型計算機



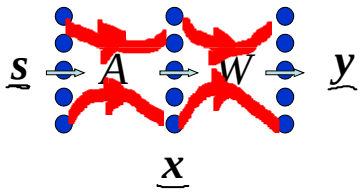
mixture and unmixture of independent signals



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

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

Independent Component Analysis



$$x = As$$

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

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

x=As ⇨ y=Wx : W= A^{-1}

observations: x(1), x(2), …, x(t)

A: unknown matrix

s: unknown

$$r(\mathbf{s}) = r_1(s_1)r_2(s_2)...r_n(s_n)$$

independent distribution

r(s): unknown

$$E[\mathbf{s}] = 0$$

$$\Delta W = -\eta \frac{\partial l(\mathbf{y}, W)}{\partial W}$$

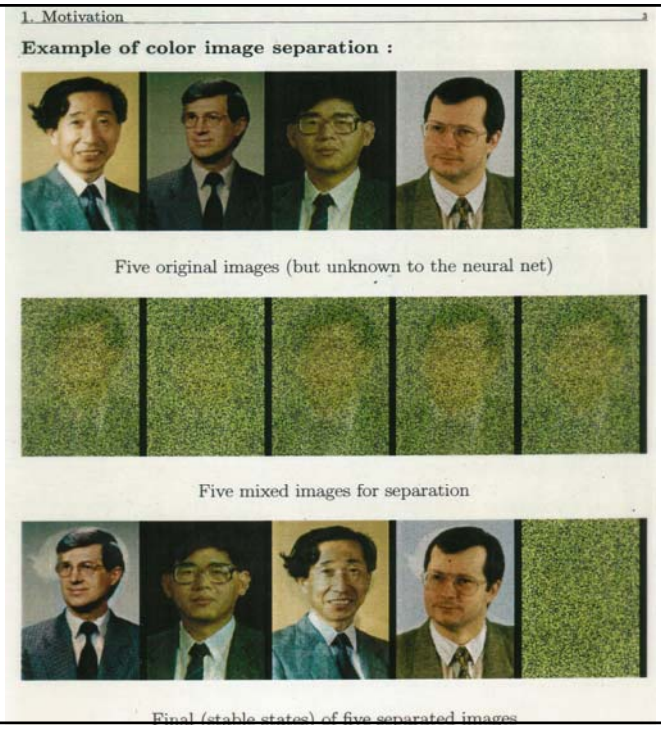
cost function:
degree of non-independenc

Semiparametric Statistical Model

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

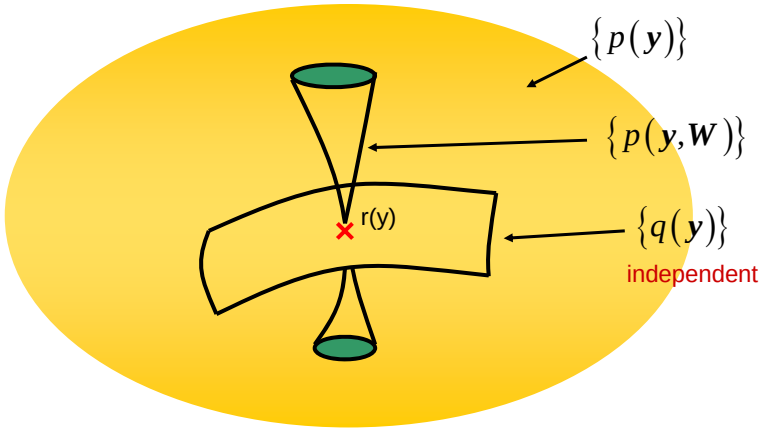
$\mathbf{W} = \mathbf{A}^{-1}$, $r(s)$: unknown $\qquad r = \Pi r_i$

x(1), x(2), …, x(t)



Information Geometry of ICA ---Independent Component Analysis

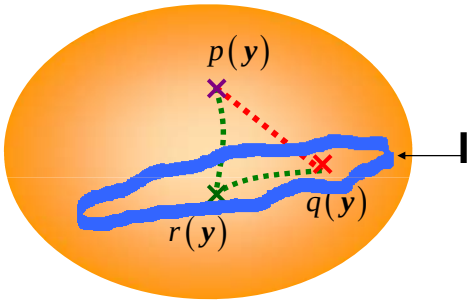
$$x = As \quad \Rightarrow \quad y = Wx = WAs$$



KL-divergence

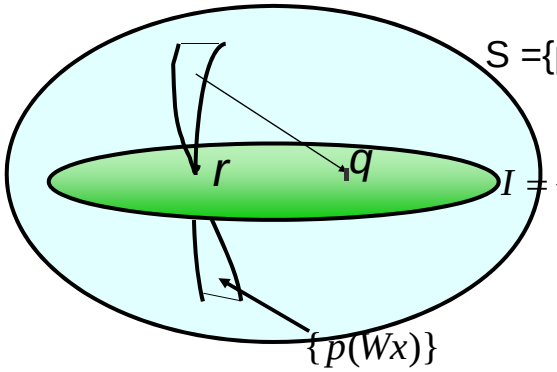
$$D[p(y):q(y)] = \int p(y) \log \frac{p(y)}{q(y)} dy$$

$$= E\left[\log \frac{p(y)}{q(y)}\right]$$



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

Information Geometry of ICA

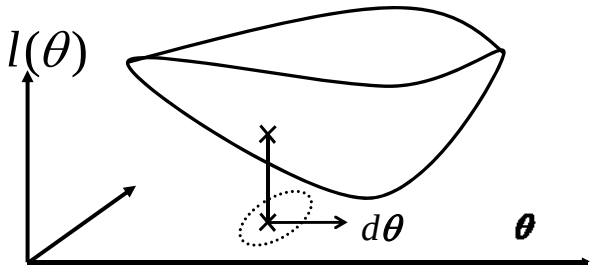


natural gradient
estimating function
stability, efficiency

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

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

Steepest Direction---Natural Gradient



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

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

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

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

Natural Gradient

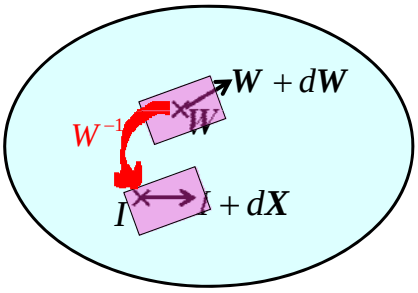
$$\max \qquad dl = l(\theta + d\theta) - l(\theta)$$

$$|d\theta|^2 = \varepsilon$$

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

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

Space of Matrices : Lie group



$$dX = dW W^{-1}$$

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

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

dX : **non-holonomic basis**

Natural Gradient

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

Stabilization—Newton Method

$$\dot{W}_{ij}(t) = \eta_t \frac{1}{\sigma_i^2 \sigma_j^2 k_i k_j - 1} \left[\sigma_i^2 k_j \varphi(y_i) y_j - \varphi(y_j) y_i \right] W$$

$$= \eta_t \{ \alpha \varphi(y_i) y_j - \varphi(y_j) y_i \} W$$

$$\sigma_i^2 = E[y_i^2], \quad \kappa_i = E[\varphi_i'(y_i)]$$

efficient algorithm

adaptive method

Spatial mixing

$$\begin{cases} x_1 = a_1 s_1 + \cdots + a_n s_n \\ x_2 = a_1' s_1 + \cdots + a_n' s_n \end{cases}$$

$$\cdots\cdots$$

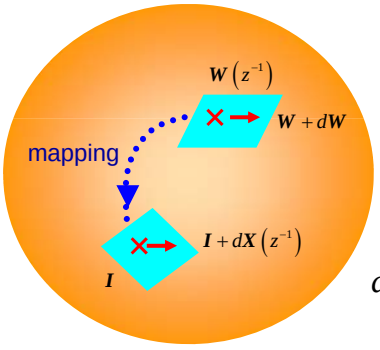
$$\boldsymbol{x} = A\boldsymbol{s}$$

Temporal mixing—convolution

$$x(t) = a_1 s(t) + a_2 s(t-1) + \cdots + a_n s(t-n)$$

$$\boldsymbol{x}(t) = \int \boldsymbol{A}(t-\tau)\boldsymbol{s}(\tau)\, d\tau$$

Manifold of Linear Systems



$$dX(z^{-1}) = dW(z^{-1})W^{-1}(z^{-1})$$

Metric Structure

Lie group
Fisher information

$$\tilde{\nabla} f(\boldsymbol{W}) = \frac{\partial f}{\partial \boldsymbol{W}} \circ W^T(z) W(z^{-1})$$

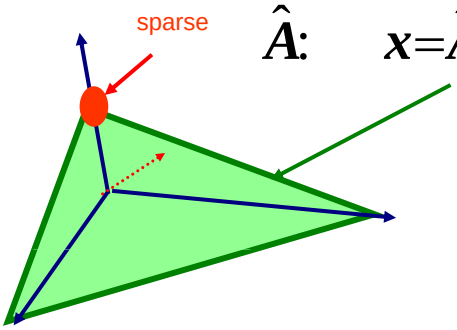
Basis Given: overcomplete case Sparse Solution

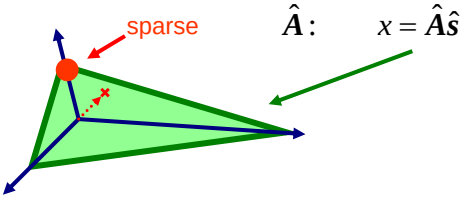
$$\boldsymbol{x} = A\boldsymbol{s} = \sum s_i \boldsymbol{a}_i$$

many solutions

many $s_i \to 0$

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



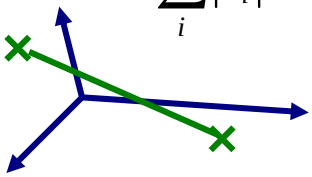


generalized inverse

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

sparse solution

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



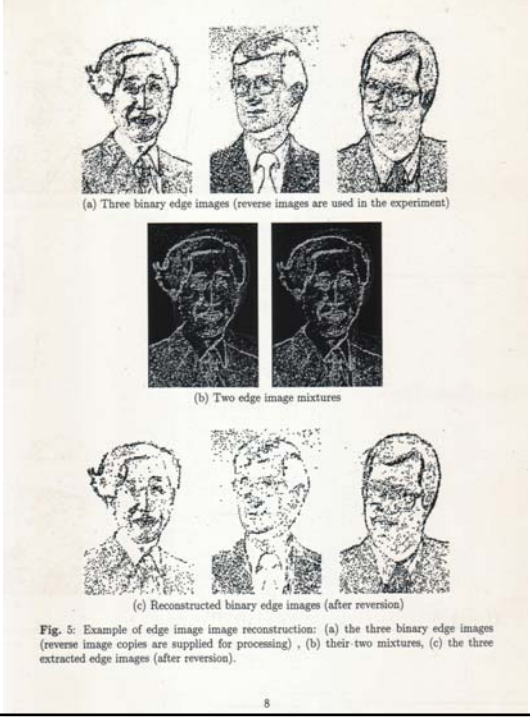
Overcomplete Basis and Sparse Solution

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

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

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

non-linear denoising



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

(b) Two edge image mixtures

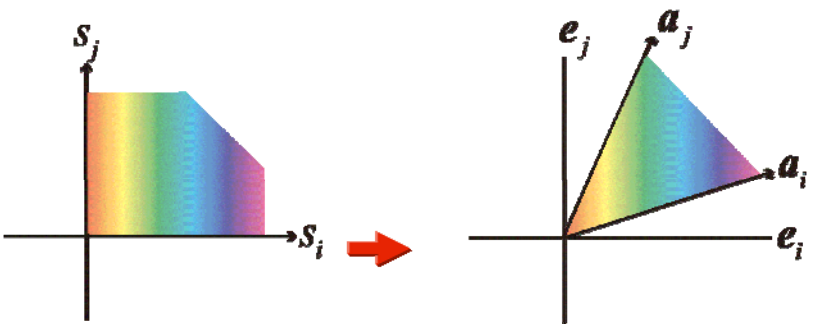
(c) Reconstructed binary edge images (after reversion)

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

8

Non-Negative Matrix Factorization

$$\boldsymbol{s} \geq 0 \ , \qquad \boldsymbol{x} = A\boldsymbol{s}$$



$$\mathbf{x}_t = A\mathbf{s}_t; \quad t = 1, 2, \ldots T$$

$$X = AS$$

All elements: non-negative

$$D[X : AS]$$

Alternative minimization of divergence function

$$a_{ij} \leftarrow a_{ij} \frac{(XS^T)_{ij}}{(ASS^T)_{ij}}; \quad s_{jt} \leftarrow s_{jt} \frac{(A^T X)_{jt}}{(A^T AS)_{jt}}$$

Robust Regression and Minkovskian Gradient

$$y_t = \sum_{i=1}^{k} x_i^t \beta_i + n_t, \quad t = 1, 2, \ldots N$$

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{n} \qquad X = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_N \end{bmatrix}$$

$$p(n) = c \exp\{-\varphi(n)\}, \qquad \varphi(n) = \frac{1}{2} n^2,$$

Maximum likelihood estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_t \varphi(y_t - \boldsymbol{\beta} \cdot \boldsymbol{x}_t),$$

$$\sum_t \varphi'(y_t - \boldsymbol{\beta} \cdot \boldsymbol{x}_t)\boldsymbol{x}_t = 0$$

Euclidean case (Gaussian noise):

$$X^T X \hat{\boldsymbol{\beta}} = X^T \boldsymbol{y} \qquad \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

LASSO: Minimize L(β)

Constraint: $\sum|\beta_i| = c$

$\varphi(n) = \frac{1}{2}n^2$ $\quad L(\boldsymbol{\beta}) = |\boldsymbol{y} - X\boldsymbol{\beta}|^2$

LARS: $\boldsymbol{\beta}$**=0** **increase** $\quad \boldsymbol{\beta}_i : \Delta_i L(\beta) \to \max;$ *then,* β_i *and* β_j;

$\varphi(n) = |n|^\alpha$ **robust**

$$L(\boldsymbol{\beta}) = \sum_t \varphi(y_t - \boldsymbol{\beta}\cdot\mathbf{x}_t)^\alpha$$

$$\sum_t \varphi'(y_t - \boldsymbol{\beta}\cdot\mathbf{x}_t)\mathbf{x}_t$$
$$= \alpha\sum_t (y_t - \boldsymbol{\beta}\cdot\mathbf{x}_t)^{\alpha-1}\mathbf{x}_t = 0$$

$S = \{y\}$ $\quad \varphi(\boldsymbol{y}) = \sum_t \varphi(y_t)$ **convex function**

$M = \{\boldsymbol{y} = X\boldsymbol{\beta}\}$ $\quad M \subset S$: linear subspace
$\phi(\beta)=\phi(\mathrm{X}\beta)$

$$D[\boldsymbol{y}:X\boldsymbol{\beta}]=\varphi(\boldsymbol{y})-\varphi(X\boldsymbol{\beta})-\sum_t\varphi'(\boldsymbol{\beta}\cdot\boldsymbol{x}_t)(y_t-\boldsymbol{\beta}\cdot\boldsymbol{x}_t)$$

$$\tilde{D}[\boldsymbol{y}:X\boldsymbol{\beta}]=\varphi(X\boldsymbol{\beta})-\varphi(\boldsymbol{y})-\sum_t\varphi'(\boldsymbol{y}_t)(\boldsymbol{\beta}\cdot\boldsymbol{x}_t-\boldsymbol{y}_t)$$

$$\varphi_M(\boldsymbol{\beta})=\varphi(X\boldsymbol{\beta})=\sum_t\varphi(\boldsymbol{\beta}\cdot\boldsymbol{x}_t)$$

$$\boldsymbol{\eta}=\nabla\varphi_M=\sum\varphi'(\boldsymbol{\beta}\cdot\boldsymbol{x}_t)\boldsymbol{x}_t$$

1) Minimize $D(\boldsymbol{\beta})=D[\boldsymbol{y}:X\boldsymbol{\beta}]$

2) Minimize $\tilde{D}(\boldsymbol{\beta})=\tilde{D}[\boldsymbol{y}:X\boldsymbol{\beta}]$

$$\nabla L=\sum(y_t-\boldsymbol{\beta}\cdot\boldsymbol{x}_t)^{\alpha-1}\boldsymbol{x}_t,$$

$$\nabla D=\sum_t(\boldsymbol{\beta}\cdot\boldsymbol{x}_t)^{\alpha-2}(y_t-\boldsymbol{\beta}\cdot\boldsymbol{x}_t)\boldsymbol{x}_t,$$

$$\nabla\tilde{D}=\sum_t\left\{y_t^{\alpha-1}-(\boldsymbol{\beta}\cdot\boldsymbol{x}_{t-1})^{\alpha-1}\right\}\boldsymbol{x}_t.$$

Minkovskian gradient

$$\frac{d}{dt}F(\boldsymbol{\beta}+t\boldsymbol{a})\big|_{t=0}=\nabla F\cdot\boldsymbol{a}$$

$$G(\boldsymbol{a}) = \sum g_{ij}(\boldsymbol{\beta}) a_i a_j$$ **Riemannian metric**

$$G(t\boldsymbol{a}) = |t|^p G(\boldsymbol{a}) > 0$$ **Minkovskian metric**

$$G(\boldsymbol{a}) = \sum |a_i|^p . \qquad p > 1$$ **Lp-norm**

$$\frac{\delta}{\delta \boldsymbol{a}} \{ \nabla F \cdot \boldsymbol{a} - \lambda G(\boldsymbol{a}) \} = 0$$

$$\delta G(\boldsymbol{a}) = p(\operatorname{sign} a_i)|a_i|^{p-1}$$

$$f_i = \frac{\partial}{\partial \beta_i} F(\boldsymbol{\beta})$$

$$a_i = c(\operatorname{sgn} f_i)|f_i|^{\frac{1}{p-1}}$$

$$G(\boldsymbol{a}) = \sum g_{ij} a_i a_j , \qquad \tilde{\nabla} F = G^{-1} \nabla f$$ **Natural gradient**

$$\tilde{\nabla} F = \nabla f$$ **Euclidean case**

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

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

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

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

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

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

Try for various p, p->1
Try for various noise function
LASSO and flat geometry