



HOKKAIDO UNIVERSITY

Title	Divergence, Signal Decomposition and Information Geometry
Author(s)	Amari, Shun-ichi
Description	APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. 4-7 October 2009. Sapporo, Japan. Plenary Session 2, Keynote Speech #2 (6 October 2009)
Citation	Proceedings : APSIPA ASC 2009 : Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference
Issue Date	2009-10-04
Doc URL	https://hdl.handle.net/2115/39701
Type	conference paper
File Information	ks_material02.pdf



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

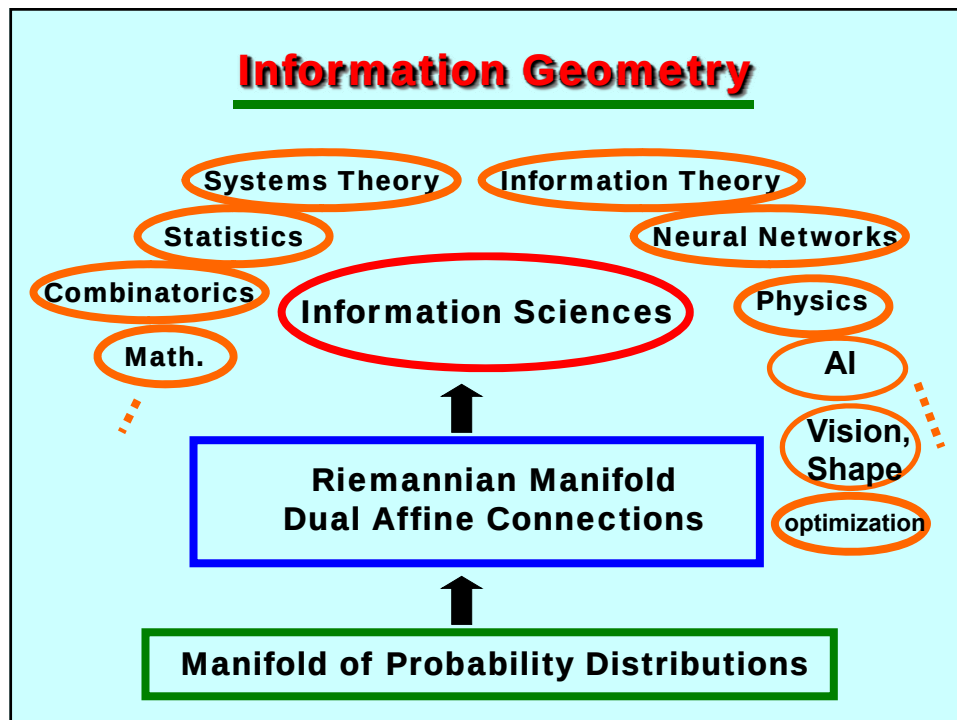
Information Geometry

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:

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

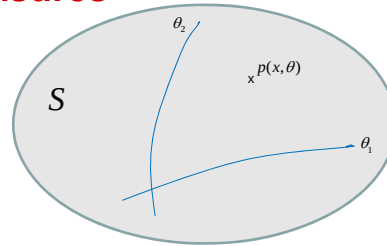
$$p(x) = \sum p_i \delta_i(x): \mathbf{p} = (p_1, \dots, p_n); \sum p_i = 1$$

Manifold of positive measures

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

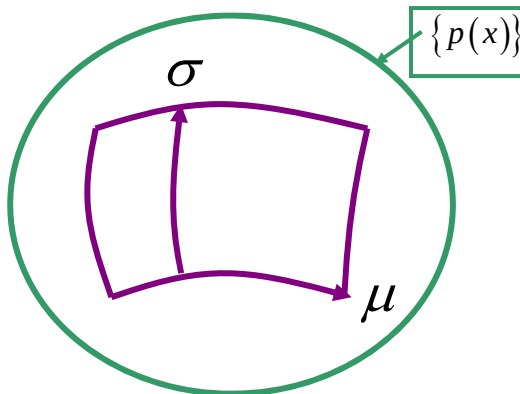
$$A = (A_{ij}), \quad A_{ij} > 0$$

$$s(x, y): \text{vision};$$



Information Geometry ?

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



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

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

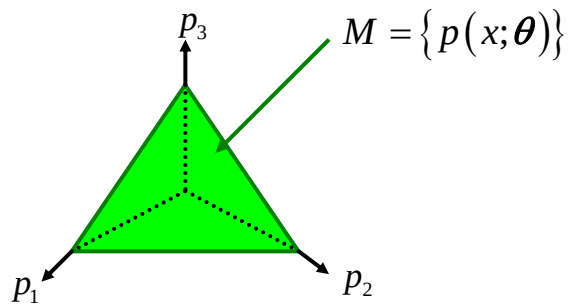
Riemannian metric

Dual affine connections

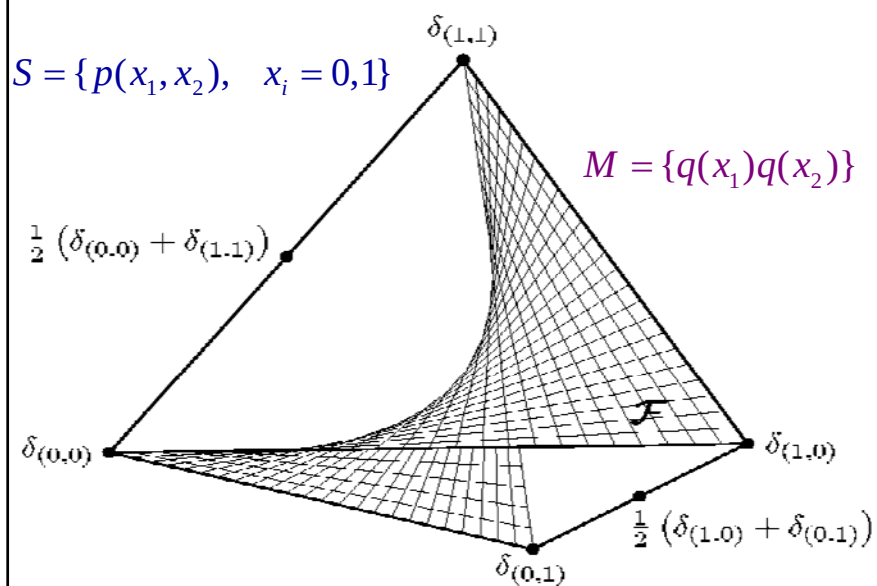
Manifold of Probability Distributions

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

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



Independent Distributions



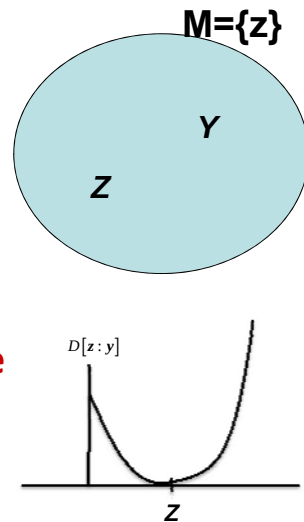
Divergence $D[z:y]$

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

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

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

positive-definite



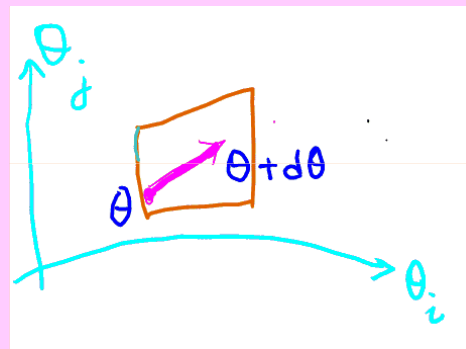
Riemannian Structure

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

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

$$\text{Euclidean } G = E$$

Fisher information



Metric and Connections Induced by Divergence (Eguchi)

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

$$\Gamma_{ijk}(\mathbf{z}) = -\partial_i \partial_j \partial_k D[\mathbf{z} : \mathbf{y}]|_{\mathbf{y}=\mathbf{z}}$$

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

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

Affine Connection

covariant derivative

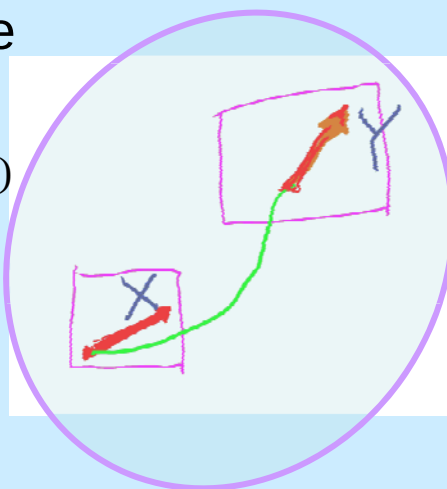
$$\nabla_X Y, \quad \Pi_c X = Y$$

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

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

minimal distance

straight line



Duality:

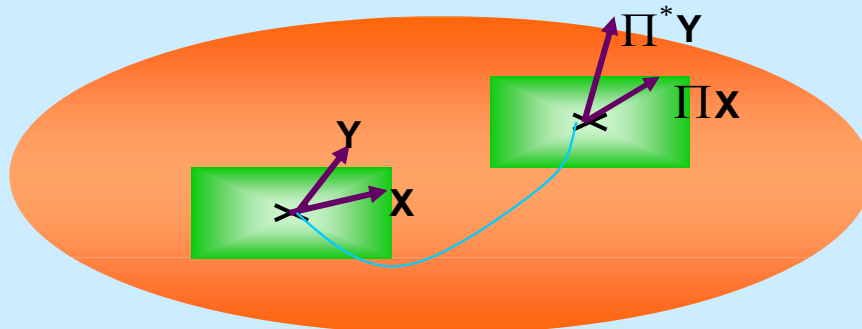
$$\partial_k g_{ij} = \Gamma_{kij} + \Gamma_{kji}^*$$

$$\Gamma_{ijk} = \Gamma_{ijk}^* - \Gamma_{ikj}$$

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

Duality

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



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

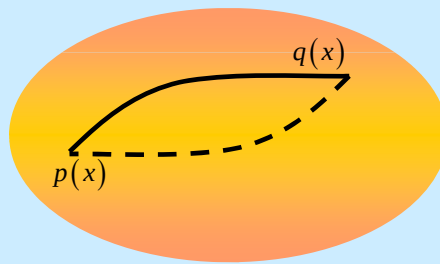
Dual Affine Connections (∇, ∇^*)

e-geodesic

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

m-geodesic

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



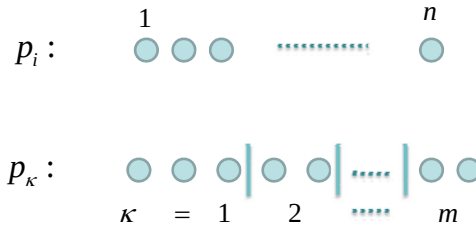
Invariance

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

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

Invariance: invariant divergence

--- characterization of f -divergence



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

$$D[p : q] \geq D[p^A : q^A]$$

$$D[p : q] = D[p^A : q^A]$$

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



Invariance $\Rightarrow f$ -divergence

Csiszar f -divergence

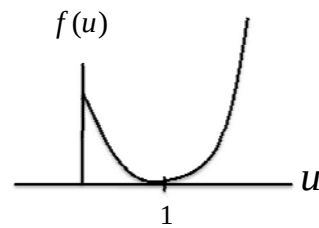
$$D_f[\mathbf{p}:\mathbf{q}] = \sum p_i f\left(\frac{q_i}{p_i}\right),$$

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

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

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

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



Invariant Geometrical Structure:

Fisher information metric

alpha-affine connections : dual

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

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

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

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

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

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

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

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

α Divergence: why?

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

KL-divergence

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

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

α divergence

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

KL-divergence

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

α

Dually flat manifold

θ -coordinates $\leftrightarrow$ η -coordinates

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

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

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

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

ψ : cumulant generating function

φ : negative entropy

Dually Flat Manifold

1. Potential Functions

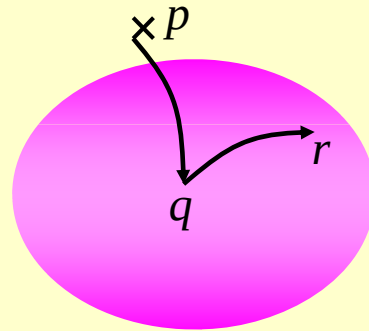
---convex (Bregman divergence, Legendre transformation)

2. Divergence $D[p:q]$

3. Pythagoras Theorem

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

4. Projection Theorem



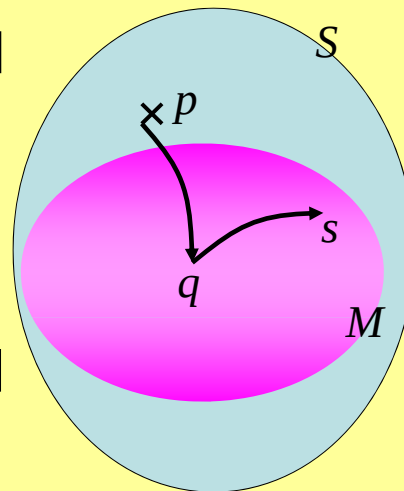
Projection Theorem

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

m-geodesic

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

e-geodesic



Dually Flat Structure

affine coordinates θ

dual affine coordinates η

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

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

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

Exponential Family

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

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

Mixture Family

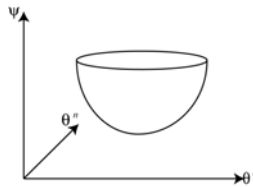
$$p(x, \eta) = \sum_{i=0}^n \eta_i p_i(x), \quad \eta_0 = 1 - \sum_{i=0}^n \eta_i$$

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

Manifold with Convex Function

S coordinates $\theta = (\theta^1, \theta^2, \dots, \theta^n)$

$\psi(\theta)$: convex function



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

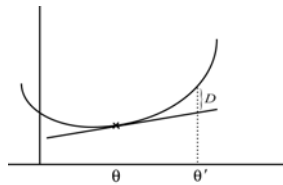
Euclidean $\psi(\theta) = \frac{1}{2} \sum (\theta^i)^2$

mathematical programming, control systems, physics, engineering, economics

Riemannian metric and flatness

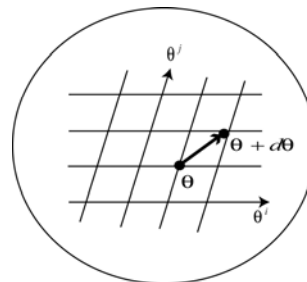
Bregman divergence

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



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

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



flatness θ : geodesic

Legendre Transformation

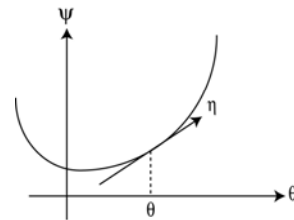
$$\eta_i = \partial_i \psi(\theta)$$

$$\theta \leftrightarrow \eta \quad \text{one-to-one}$$

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

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

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



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

Geometry: Dually Flat Geometry

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

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

$$G(\mathbf{p}) = \nabla \nabla \psi(\mathbf{p})$$

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

Straightness (affine connection)

$$\mathbf{p}(t) = \mathbf{a} + \mathbf{b}t \quad : \psi\text{-geodesic}$$

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

Two flat coordinate systems (θ, η)

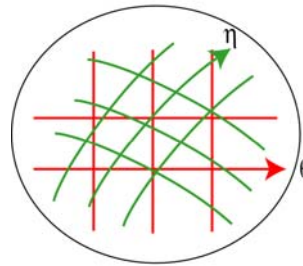
θ : geodesic (e-geodesic)

η : dual geodesic (m-geodesic)

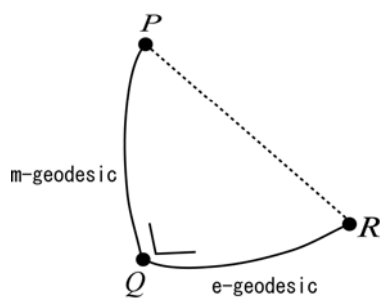
“dually orthogonal”

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

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



Pythagorean Theorem



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

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

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

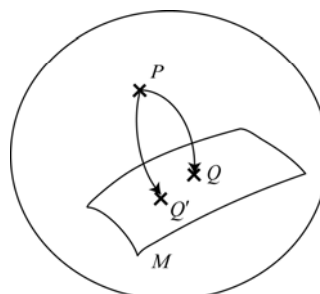
Projection Theorem

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

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

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

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



divergence

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

invariance

dually flat space

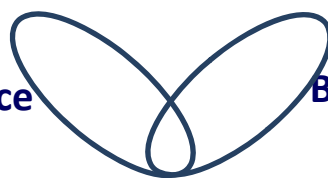
f -divergence

Bregman divergence

Fisher inf metric
Alpha connection

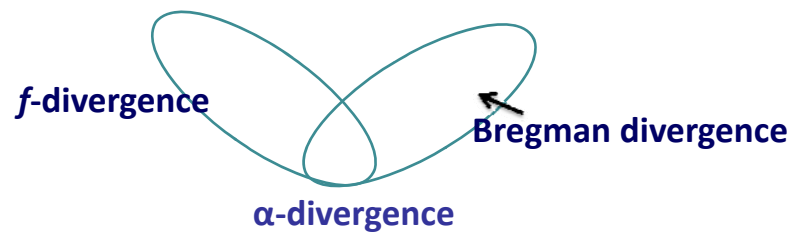
KL-divergence

convex functions



Space of positive measures : vectors, matrices, arrays

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



α-representation

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

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

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

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

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

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

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

$$r_i^* = p_i$$

α divergence

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

$\tilde{S}$: 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

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

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 : $m_\alpha(p_1(s), p_2(s))$

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

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

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

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

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

α -divergence : positive real numbers

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

α – Family of Distributions

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

mixture family :

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

exponential family :

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



α – Bayes Predictive Distribution

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

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

$\alpha = 1$: predictive distribution

$\alpha = -1$: product of experts

Optimal Property

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

find $q(x|D)$ **that minimizes the α – risk**

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

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

Theorem

The α – predictive distribution minimizes
the α – risk.

The Bayes predictive distribution :

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

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

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

Tsallis q -Entropy

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

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

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

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

$$\begin{aligned} D[p(x):r(x)] &= -E_p \left[\ln_q \frac{r(x)}{p(x)} \right] \\ &= \frac{1}{1-q} \left(1 - \int p(x)^q r(x)^{1-q} dx \right) \\ &\quad : \alpha\text{-divergence ; } \alpha\text{-structure} \end{aligned}$$

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

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

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

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

β -structure

q -exponential family

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

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

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

$$\eta_i = E_q[x_i]$$

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

$$S_n = \left\{ \mathbf{p} \left| \sum_{i=0}^n p_i = 1 \right. \right\}$$

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

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

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

conformal transformation

q -Fisher information

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

q -divergence

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

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 $\hat{\xi}$

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

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

q -maximum likelihood estimator

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

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

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

q -super-robust estimator (Eguchi)

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

bias-corrected q -estimating function

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

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

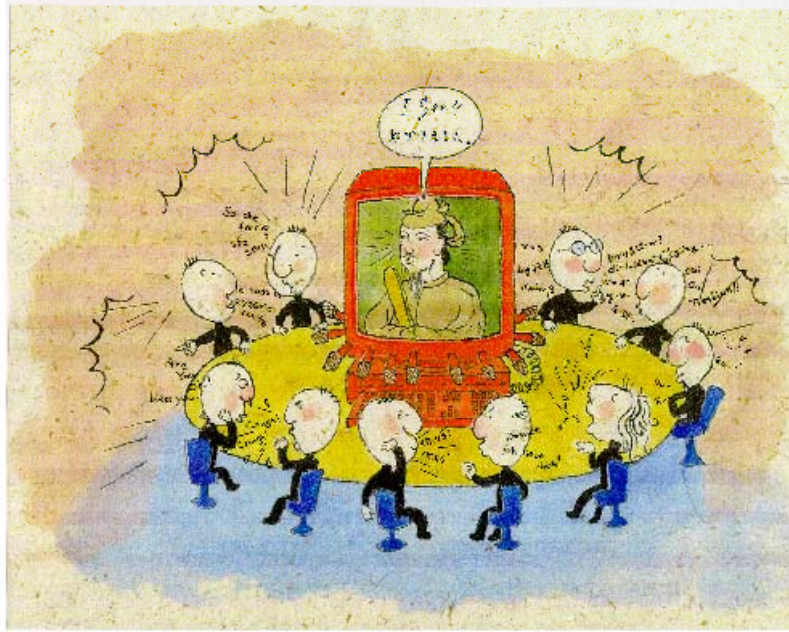
$$\sum_{i=0}^N s_q(x_i, \xi) = 0 \quad \Leftrightarrow \quad \max \frac{1}{h_{q+1}} \sum \hat{p}(x_i, \xi)$$

Applications of geometry:

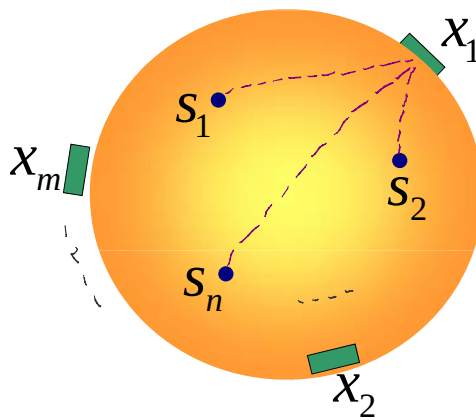
ICA

Independent Component Analysis

聖德太子型計算機



mixture and unmixture
of independent signals



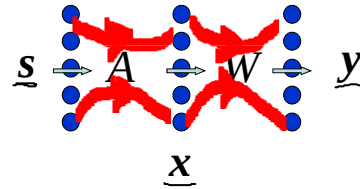
$$x_i = \sum_{j=1}^n A_{ij} s_j$$

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

Independent Component Analysis

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

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



observations: $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)$

recover: $\mathbf{s}(1), \mathbf{s}(2), \dots, \mathbf{s}(t)$

$$\mathbf{x} = \mathbf{A}\mathbf{s} \Rightarrow \mathbf{y} = \mathbf{W}\mathbf{x} : \mathbf{W} = \mathbf{A}^{-1}$$

observations: $\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(t)$

A: unknown matrix

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

independent distribution

r(s): unknown $E[\mathbf{s}] = 0$

$$\Delta \mathbf{W} = -\eta \frac{\partial l(\mathbf{y}, \mathbf{W})}{\partial \mathbf{W}} \quad \text{cost function:}$$

degree of non-independence

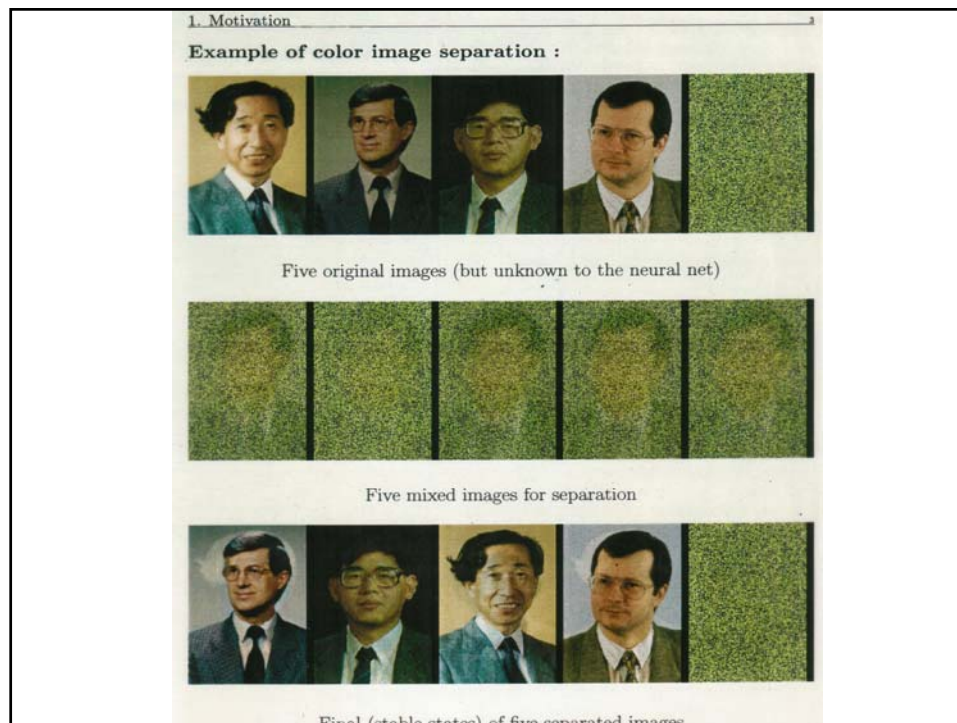
Semiparametric Statistical Model

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

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

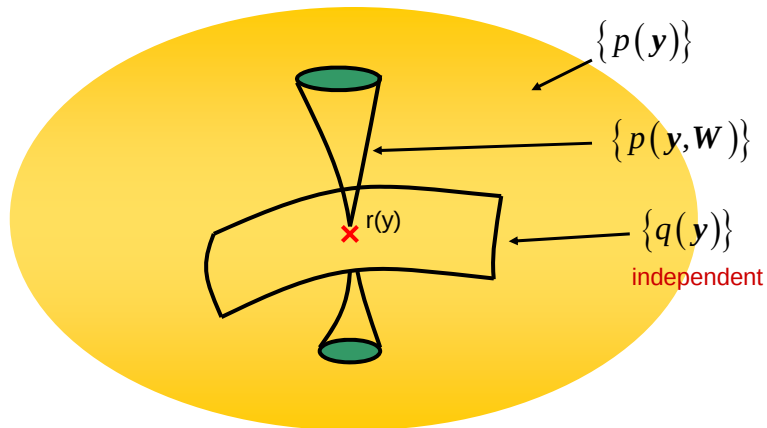
unknown

$$x(1), x(2), \dots, x(t)$$



Information Geometry of ICA ---Independent Component Analysis

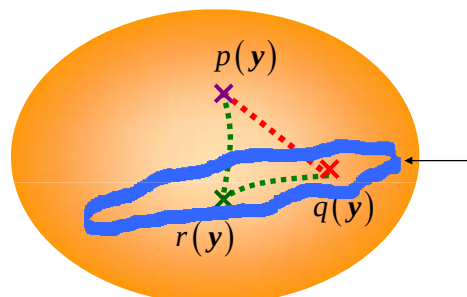
$$\mathbf{x} = \mathbf{A}\mathbf{s} \quad \Rightarrow \quad \mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}$$



KL-divergence

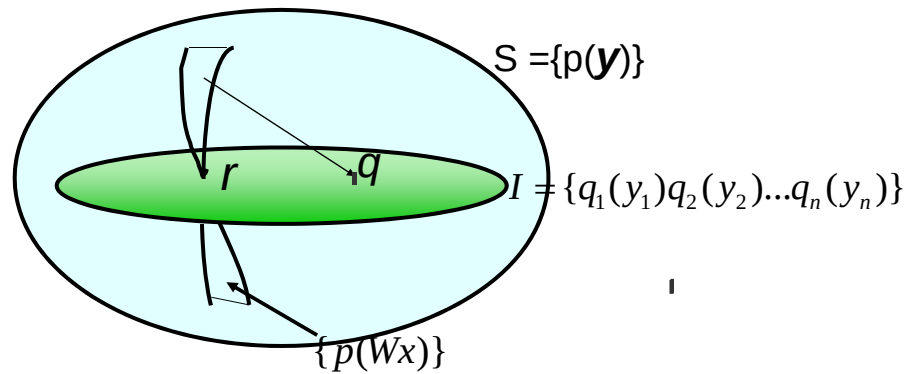
$$D[p(y):q(y)] = \int p(y) \log \frac{p(y)}{q(y)} dy$$

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



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

Information Geometry of ICA

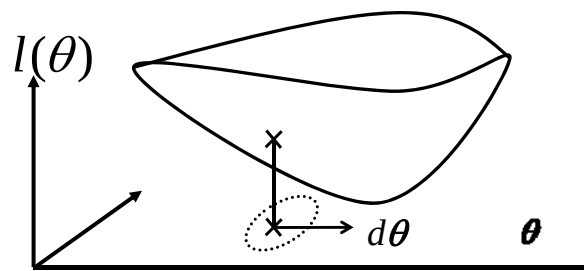


natural gradient
estimating function
stability, efficiency

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

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

Steepest Direction---Natural Gradient



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

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

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

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

Natural Gradient

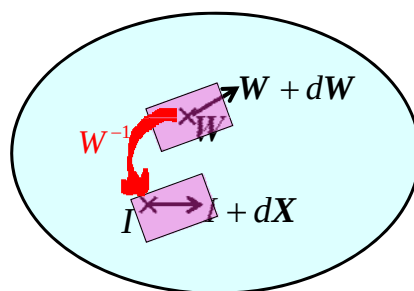
$$\max \quad dl = l(\boldsymbol{\theta} + d\boldsymbol{\theta}) - l(\boldsymbol{\theta})$$

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

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

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

Space of Matrices : Lie group



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

$$|dW|^2 = \text{tr}(dX dX^T) = \text{tr}(dW W^{-1} W^{-T} dW^T)$$

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

dX : non-holonomic basis

Natural Gradient

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

Stabilization—Newton Method

$$\begin{aligned} \dot{\mathbf{W}}_{ij}(t) &= \eta_t \frac{1}{\sigma_i^2 \sigma_j^2 \kappa_i \kappa_j - 1} \left[\sigma_i^2 \kappa_j \varphi(y_i) y_j - \varphi(y_j) y_i \right] \mathbf{W} \\ &= \eta_t \{ \alpha \varphi(y_i) y_j - \varphi(y_j) y_i \} \mathbf{W} \end{aligned}$$

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



efficient algorithm

adaptive method

Spatial mixing

$$\begin{cases} x_1 = a_1 s_1 + \cdots + a_n s_n \\ x_2 = a'_1 s_1 + \cdots + a'_n s_n \\ \dots\dots \end{cases}$$

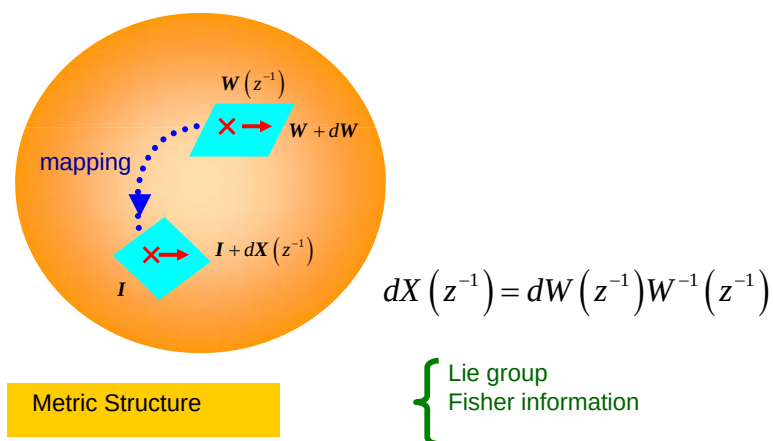
$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

Temporal mixing—convolution

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

$$x(t) = \int A(t-\tau)s(\tau)d\tau$$

Manifold of Linear Systems



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

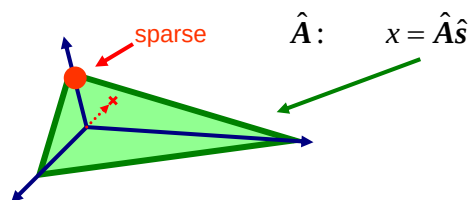
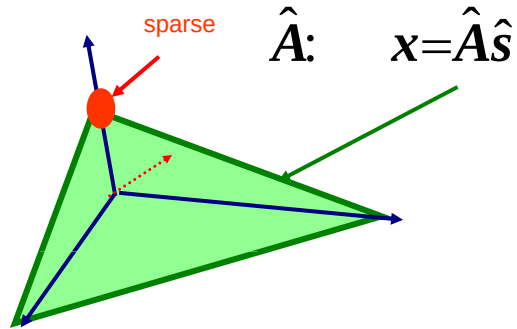
Basis Given: overcomplete case Sparse Solution

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

many solutions

many $s_i \rightarrow 0$

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

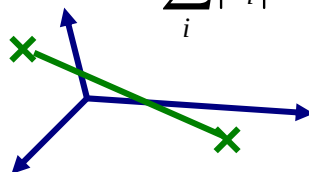


generalized inverse

$$L_2\text{-norm: } \min \sum |\hat{s}_i|^2$$

sparse solution

$$L_1\text{-norm: } \min \sum |\hat{s}_i|$$



Overcomplete Basis and Sparse Solution

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

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

$$\min \|\mathbf{A}\mathbf{s} - \mathbf{x}\|_p + \alpha \|\mathbf{s}\|_p,$$

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)

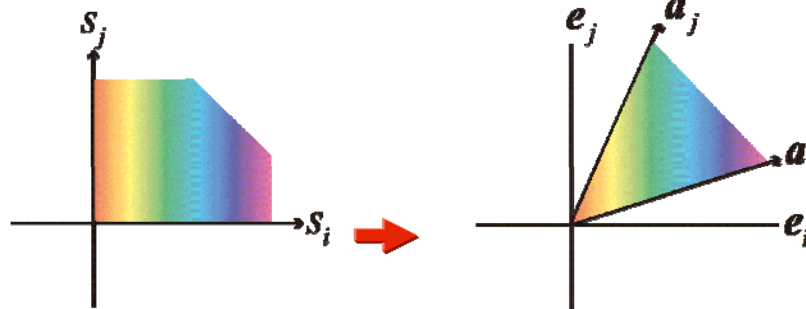
Fig. 5: Example of edge image image reconstruction: (a) the three binary edge images (reverse image copies are supplied for processing), (b) their two mixtures, (c) the three extracted edge images (after reversion).



Non-Negative Matrix Factorization

$$\mathbf{s} \geq 0,$$

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



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

$$\mathbf{X} = \mathbf{A}\mathbf{S}$$

All elements: non-negative

$$D[\mathbf{X} : \mathbf{A}\mathbf{S}]$$

**Alternative minimization
of divergence function**

$$a_{ij} \leftarrow a_{ij} \frac{(\mathbf{X}\mathbf{S}^T)_{ij}}{(\mathbf{A}\mathbf{S}\mathbf{S}^T)_{ij}}; \quad s_{jt} \leftarrow s_{jt} \frac{(\mathbf{A}^T \mathbf{X})_{jt}}{(\mathbf{A}^T \mathbf{A}\mathbf{S})_{jt}}$$

Robust Regression and Minkovskian Gradient

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

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

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

Maximum likelihood estimator

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

$$\sum_t \varphi'(y_t - \beta \cdot x_t) x_t = 0$$

Euclidean case (Gaussian noise):

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

LASSO: Minimize $L(\beta)$

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

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

LARS: $\beta = 0$ increase $\beta_i : \Delta_i L(\beta) \rightarrow \max;$
then, β_i and β_j ;

$$\varphi(n) = |n|^\alpha \quad \text{robust}$$

$$L(\beta) = \sum_t \varphi(y_t - \beta \cdot x_t)^\alpha$$

$$\begin{aligned} \sum_t \varphi'(y_t - \beta \cdot x_t) x_t \\ = \alpha \sum_t (y_t - \beta \cdot x_t)^{\alpha-1} x_t = 0 \end{aligned}$$

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

$$\begin{aligned} M = \{y = X\beta\} \quad M \subset S : \text{linear subspace} \\ \phi(\beta) = \phi(X\beta) \end{aligned}$$

$$D[y: X\beta] = \varphi(y) - \varphi(X\beta) - \sum_t \varphi'(\beta \cdot x_t)(y_t - \beta \cdot x_t)$$

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

$$\varphi_M(\beta) = \varphi(X\beta) = \sum_t \varphi(\beta \cdot x_t)$$

$$\eta = \nabla \varphi_M = \sum_t \varphi'(\beta \cdot x_t) x_t$$

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

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

$$\nabla L = \sum_t (y_t - \beta \cdot x_t)^{\alpha-1} x_t,$$

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

$$\nabla \tilde{D} = \sum_t \left\{ y_t^{\alpha-1} - (\beta \cdot x_{t-1})^{\alpha-1} \right\} x_t.$$

Minkovskian gradient

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

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

$$G(t\mathbf{a}) = |t|^p G(\mathbf{a}) > 0 \quad \text{Minkovskian metric}$$

$$G(\mathbf{a}) = \sum |a_i|^p, \quad p > 1 \quad \text{Lp-norm}$$

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

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

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

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

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

$$\tilde{\nabla} F = \nabla f \quad \text{Euclidean case}$$

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

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

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

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

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

$$\beta_{t+1} = \beta_t - \eta \tilde{\nabla} F \quad \text{LASSO}$$

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