2015 Workshop on High-Dimensional Statistical Analysis
Dec.11 (Friday) ~15 (Tuesday)
Humanities and Social Sciences Center, Academia Sinica, Taiwan

Information Geometry and Spontaneous Data Learning

Shinto Eguchi

Institute Statistical Mathematics, Japan

This talk is based on a joint work with

Osamu Komori and Atsumi Ohara, University of Fukui

Outline

- **Short review for Information geometry**
- **Kolmogorov-Nagumo mean**
- **ϕ-path in a function space**
- **Generalized mean and variance**
- **ϕ–divergence geometry**
- ***U*-divergence geometry**
- **Minimum *U*-divergence and density estimation**

A short review of IG

Nonparametric space $\mathscr{F}_\mu = \{f : f(x) \geq 0, \int f(x)d\mu(x) = 1\}$

Space of statistics $\mathscr{S}_\mu = \{s(x) : \int |s(x)|\, d\mu(x) < \infty\}$

where μ is a fixed probability measure.

Information geometry is discussed on the product space $\mathscr{F}_\mu \times \mathscr{S}_\mu$

$$\{\text{model}\} \subseteq \mathscr{F}_\mu$$
$$\text{IG} \uparrow\downarrow$$
$$\{\text{estimators}\} \subseteq \mathscr{S}_\mu$$

A canonical biliear form : $\mathscr{F}_\mu \times \mathscr{S}_\mu \to \mathbb{R}$ defined by

$$\langle f, s \rangle = \int f(x)s(x)\mathrm{d}\mu(x)$$

Bartlett's identity

Parametric model $M = \{ f_\theta(x) \in \mathscr{F}_\mu : \theta = (\theta_1, \cdots, \theta_d) \in \Theta \}$

Bartlett's first identity

$$\left\langle f_\theta, \frac{\partial}{\partial \theta} \log f_\theta \right\rangle \rightarrow \mathrm{E}_{f_\theta} \{ \frac{\partial}{\partial \theta} \log f_\theta(X) \} = 0$$

Bartlett's second identity

$$\mathrm{E}_{f_\theta} \{ - \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \log f_\theta(X) \} = \mathrm{Var}_{f_\theta} \{ \frac{\partial}{\partial \theta} \log f_\theta(X) \}$$

Metric and connections

$$\text{Let} \quad M = \{ f_\theta(x) \in \mathscr{F}_\mu : \theta = (\theta_1, \cdots, \theta_d) \in \Theta \}$$

Information metric $G_{ij}(\theta) = \mathrm{Cov}_{f_\theta} \{ S_i(X, \theta), S_j(X, \theta) \}$

$$\text{where} \quad S_i(x, \theta) = \frac{\partial}{\partial \theta_i} \log f_\theta(x)$$

Mixture connection

$$\Gamma^{(\mathrm{m})}_{ij,k}(\theta) = \int \frac{\partial^2 f_\theta(x)}{\partial \theta_i \partial \theta_i} S_k(x, \theta) \mathrm{d}\mu(x)$$

Exponential connection

$$\Gamma^{(\mathrm{e})}_{ij,k}(\theta) = \mathrm{E}_{f_\theta} \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_\theta(X, \theta) \, S_k(X, \theta) \right\}$$

Rao (1945), Dawid (1975), Amari (1982)

Geodesic curves and surfaces in IG

Let us fix f and g of $\mathscr{F}_\mu$.

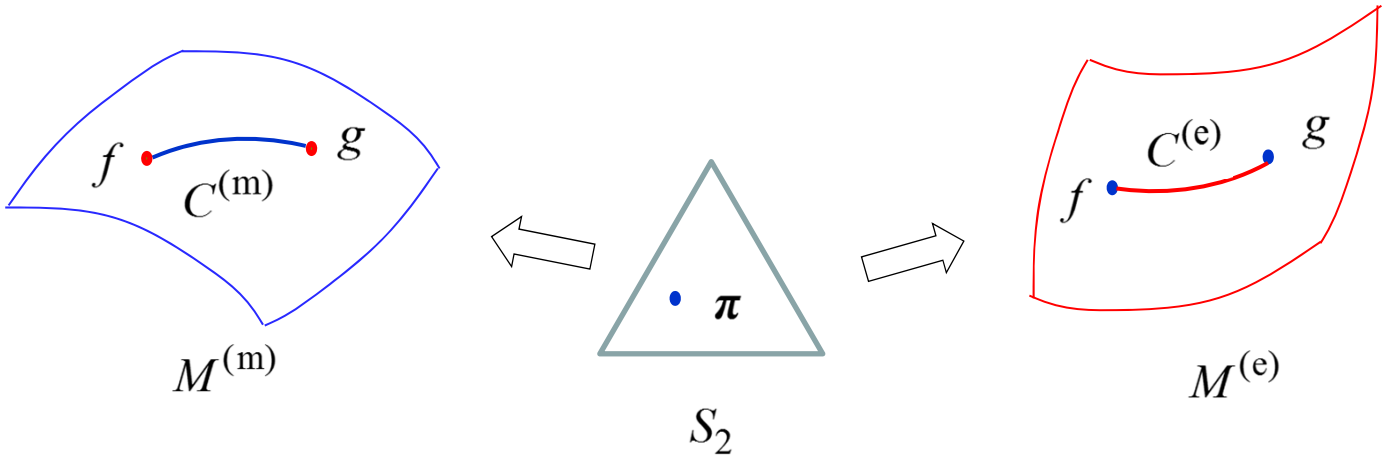
m-geodesic curve $C^{(\mathrm{m})} = \{tf(x) + (1-t)g(x) : t \in (0, 1)\}$

e-geodesic curve $C^{(\mathrm{e})} = \{\exp\{t \log f(x) + (1-t)\log g(x) - \kappa_t\} : t \in (0, 1)\}$

Let us fix $\{f_k\}_{k=1}^K$ in $\mathscr{F}_\mu$.

m-geodesic surface $M^{(\mathrm{m})} = \{\sum_{k=1}^K \pi_k f_k(x) : \boldsymbol{\pi} \in S_K\}$ where S_{K-1} denotes a simplex.

e-geodesic surface $M^{(\mathrm{e})} = \{\exp\{\sum_{k=0}^K \pi_k \log f_k(x) - \kappa(\boldsymbol{\pi})\} : \boldsymbol{\pi} \in S_K\}$



Kullback-Leibler

K-L divergence $$D_0(f,g)=\int f(x)\log\frac{f(x)}{g(x)}d\mu(x)$$

1. KL divergence is the expected log-likelihood ratio

$$D_0(f,g)=\mathrm{E}_f\left\{\log\frac{f(X)}{g(X)}\right\}$$

2. Maximum likelihood is minimum KL divergence. Akaike (1974)

$$L_0(\theta)=\mathrm{E}_{\hat{f}_n}\left\{\log f_\theta(X)\right\}=-D_0(\hat{f}_n,f_\theta)+\text{const}$$

3. KL divergence induces to the m-connection and e-connection

$$\Gamma^{(\mathrm{m})}_{ij,k}(\theta)=-\frac{\partial^3}{\partial\theta_{1i}\partial\theta_{1j}\partial\theta_{2k}}D_0(f_{\theta_1},f_{\theta 2})\Big|_{\substack{\theta_1=\theta\\ \theta_2=\theta}}$$

$$\Gamma^{(\mathrm{e})}_{ij,k}(\theta)=-\frac{\partial^3}{\partial\theta_{1i}\partial\theta_{2j}\partial\theta_{2k}}D_0(f_{\theta_2},f_{\theta 1})\Big|_{\substack{\theta_1=\theta\\ \theta_2=\theta}}$$

Eguchi (1983)

Pythagoras

Thm Let us fix distinct densities f, g and h in $\mathcal{F}_\mu$

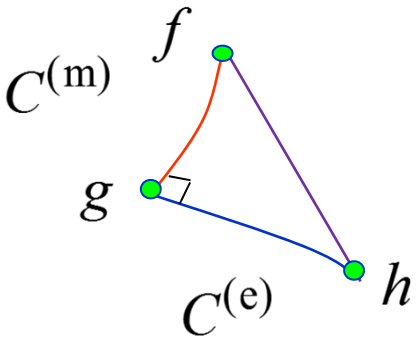
$$C^{(\mathrm{m})} = \{ t f(x) + (1-t) g(x) : t \in (0, 1) \}$$

$$C^{(\mathrm{e})} = \{ \exp\{ t \log h(x) + (1-t) \log g(x) - \kappa_t \} : t \in (0, 1) \}$$

If $C^{(\mathrm{m})}$ and $C^{(\mathrm{e})}$ orthogonally intersects at g, then

$$D_0(f, h) = D_0(f, g) + D_0(g, h)$$

Amari-Nagaoka (2001)



Pf
$$G(\dot{C}^{(\mathrm{m})}, \dot{C}^{(\mathrm{e})}) = \int (f - g)(\log g - \log h - \dot{\kappa}_t) d\mu$$
$$= D_0(f, h) - \{ D_0(f, g) + D_0(g, h) \}$$

Exponential model

Let us fix $\boldsymbol{t} = (t_1, \cdots, t_K)$ where $t_k \in \mathscr{S}_\mu$ for $k = 1, \ldots, K$

Exponential model $$M^{(\mathrm{e})} = \{ f_{\boldsymbol{\theta}}^{(\mathrm{e})}(\boldsymbol{x}) := \exp\{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}) - \kappa(\boldsymbol{\theta})\} : \boldsymbol{\theta} \in \Theta \}$$

Mean parameter $$\boldsymbol{\eta} = \mathrm{E}_{f_{\boldsymbol{\theta}}^{(\mathrm{e})}} \{ \boldsymbol{t}(\boldsymbol{X}) \} = \frac{\partial}{\partial \boldsymbol{\theta}} \kappa(\boldsymbol{\theta})$$

For $\boldsymbol{\theta}$ $$G_{ij}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \kappa(\boldsymbol{\theta}) \qquad \Gamma_{ij,k}^{(\mathrm{e})}(\boldsymbol{\theta}) = 0$$

For $\boldsymbol{\eta}$ $$G_{ij}(\eta) = (G(\theta)^{-1})_{ji} \qquad \Gamma_{ij,k}^{(\mathrm{m})}(\boldsymbol{\eta}) = 0$$

Amari (1982)

Degenerated Bartlett identity $$-\frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \log f_{\boldsymbol{\theta}}^{(\mathrm{e})}(\boldsymbol{x}) = \mathrm{Var}_{f_\theta} \left(\frac{\partial}{\partial \theta} \log f_{\boldsymbol{\theta}}^{(\mathrm{e})} \right)$$

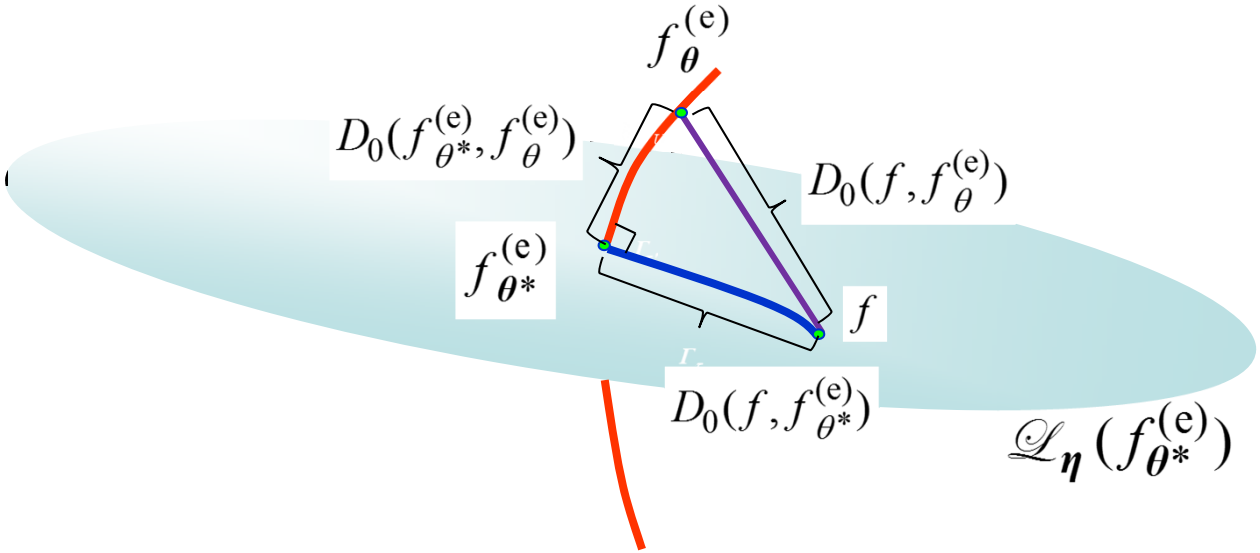
Minimum KL leaf

Exponential model $M^{(\mathrm{e})} = \{ f_{\boldsymbol{\theta}}^{(\mathrm{e})}(\boldsymbol{x}) := \exp\{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{x}) - \kappa(\boldsymbol{\theta})\} : \boldsymbol{\theta} \in \Theta \}$

Mean equal space $\mathscr{L}(g) = \{ f \in \mathscr{F}_{\mu} : \mathrm{E}_f\{\boldsymbol{t}(\boldsymbol{X})\} = \mathrm{E}_g\{\boldsymbol{t}(\boldsymbol{X})\} \}$

$$f \in \mathscr{L}(f_{\theta^*}^{(\mathrm{e})}) \quad \Rightarrow \quad D_0(f, f_{\theta}^{(\mathrm{e})}) = D_0(f, f_{\theta^*}^{(\mathrm{e})}) + D_0(f_{\theta^*}^{(\mathrm{e})}, f_{\theta}^{(\mathrm{e})})$$

$$\text{Hence} \quad D_0(f, f_{\theta^*}) = \min_{\theta \in \Theta} D_0(f, f_{\theta}^{(\mathrm{e})})$$

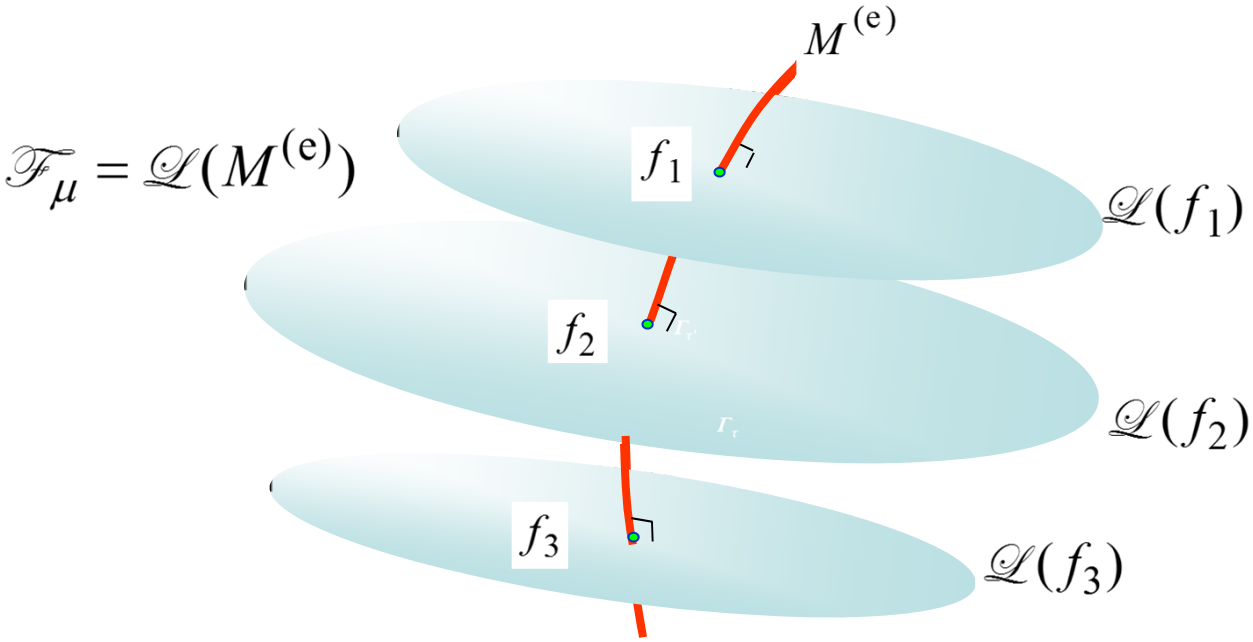


Pythagoras foliation

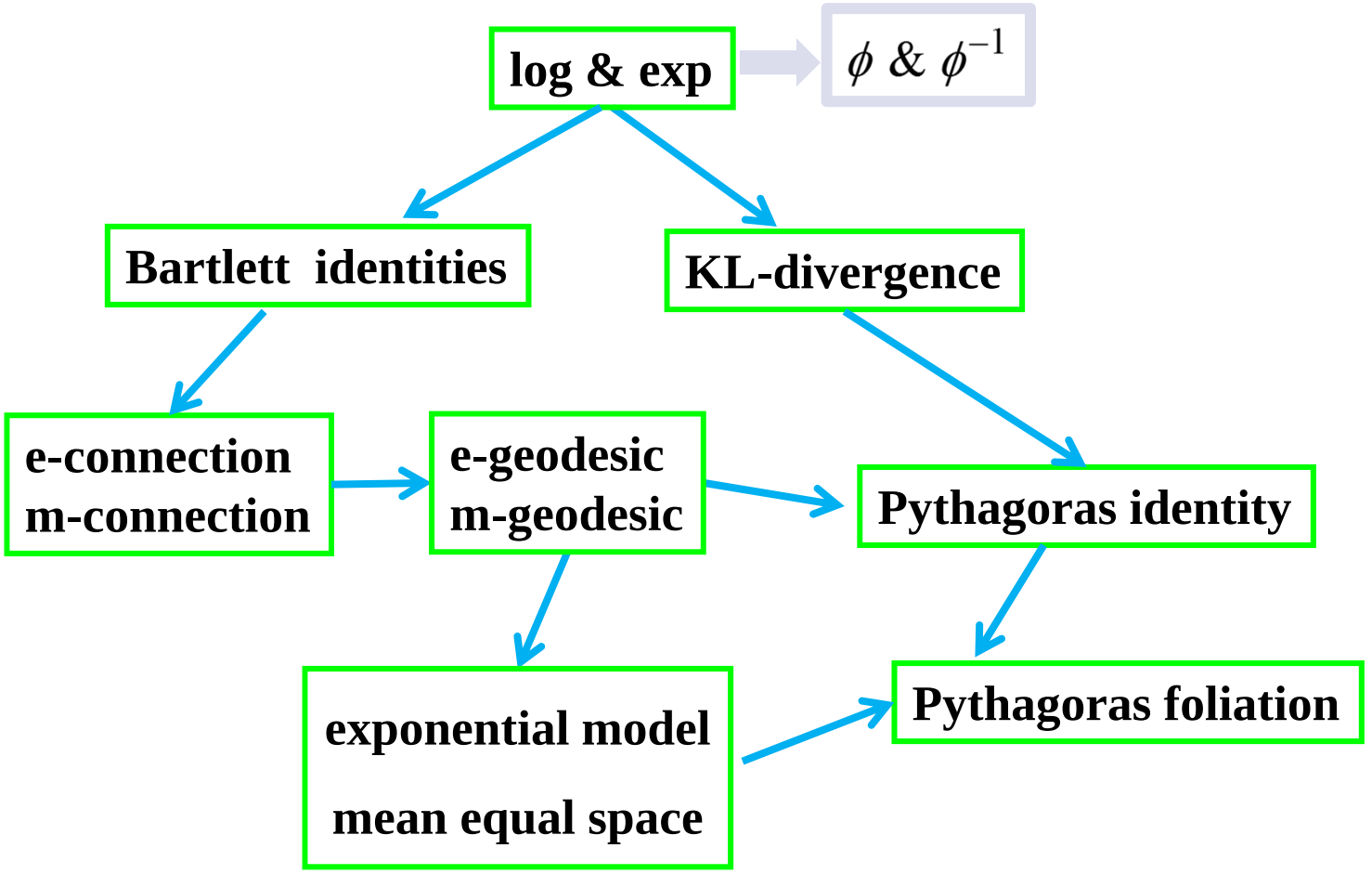
$\{\mathscr{L}(f)\colon f \in M^{(\mathrm{e})}\}$ is a foliation

(i) $\quad f_1 \neq f_2 \Rightarrow \mathscr{L}_{f_1} \cap \mathscr{L}_{f_2} = \phi$

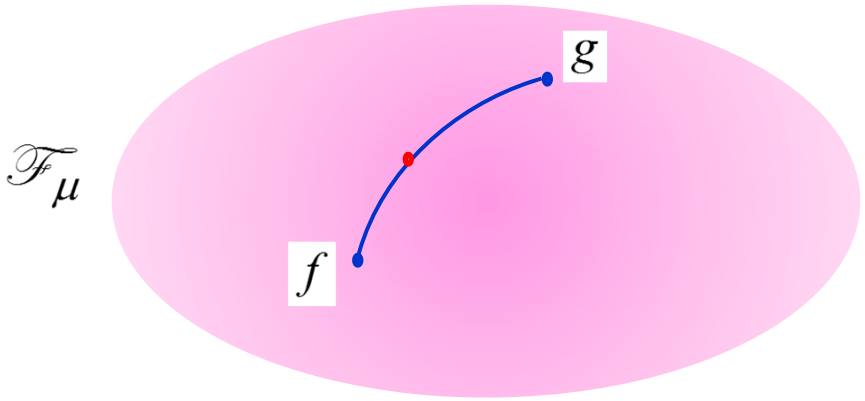
(ii) $\quad \mathscr{F}_\mu = \bigcup_{f \in M^{(\mathrm{e})}} \mathscr{L}(f)$



log + exp


log & exp
ϕ & ϕ^{-1}
Bartlett identities
KL-divergence
e-connection
m-connection
e-geodesic
m-geodesic
Pythagoras identity
exponential model
mean equal space
Pythagoras foliation

Path geometry



{ m-geodesic, e-geodesic , … }

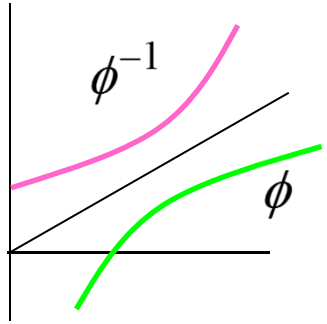
Kolmogorov-Nagumo mean

K-N mean is $\phi^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right)$ for positive numbers $\{x_1,\ldots,x_n\}$

(1) $\phi(s)=s$ sample mean : $\frac{1}{n}\sum_{i=1}^{n}x_i$

(2) $\phi(s)=\log(s)$ geometric mean : $\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log x_i\right)=(x_1\cdots x_n)^{\frac{1}{n}}$

(3) $\phi(s)=s^{-1}$ harmonic mean : $\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$

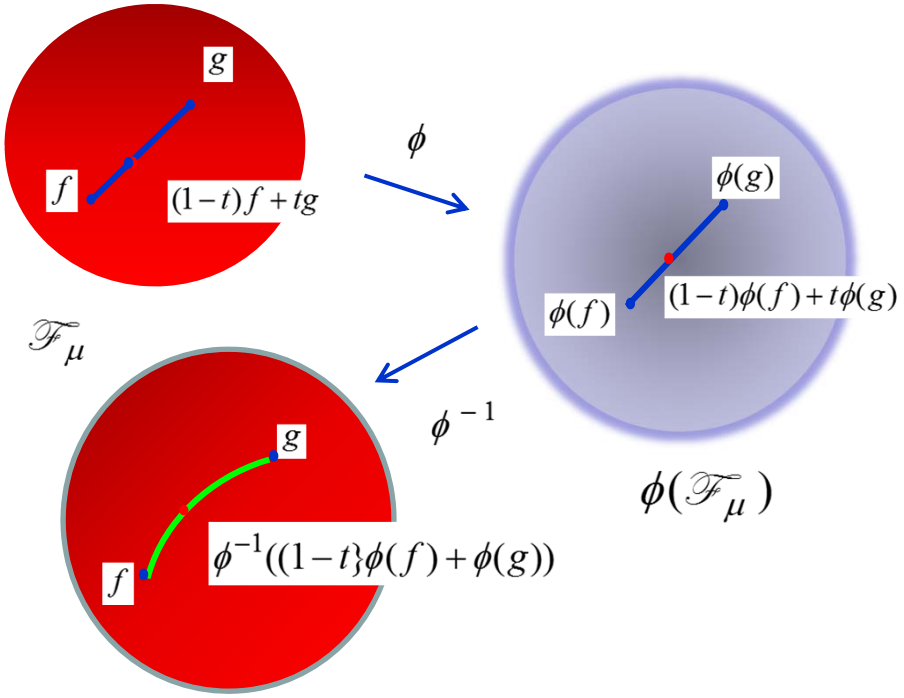


K-N mean in $\mathscr{F}_\mu$

Def . **K-N mean**

$$\phi^{-1}\big((1-t)\,\phi(f(x))+t\phi(g(x))\big), \qquad f,g\in\mathscr{F}_\mu, \quad t\in[0,1]$$

Cf. Naudts (2009)



ϕ-path

Def. **ϕ-path connecting f and g**

$$f_t(x,\phi) = \phi^{-1}((1-t)\phi(f(x)) + t\phi(g(x)) - \kappa_t),$$

where κ_t is a normalizing factor to satisfy $\int \phi^{-1}((1-t)\phi(f) + t\phi(g) - \kappa_t)d\mu = 1.$

Thm Assume that ϕ is strictly increasing and concave.

Then there exists such a normalizing constant κ_t as above.

(Pf) Since ϕ^{-1} is strictly inceasing and convex from the assumption

$$\phi^{-1}((1-t)\phi(f) + t\phi(g)) \le (1-t)f + tg$$

$$\int \phi^{-1}((1-t)\phi(f) + t\phi(g))d\mu \le 1.$$

Further $\lim_{c\to\infty} \phi^{-1}(c) = \infty$

■

Examples of ϕ -path

Exm 0 $\phi(s) = s \quad \Rightarrow \quad f_t(x,\phi) = (1-t)f(x) + tg(x) \quad (\text{m - geodesic})$

Exm 1 $\phi_0(s) = \log s \ \Rightarrow$

$$f_t(x,\phi_0) = \exp((1-t)\log f(x) + t\log g(x) - \kappa_t) \quad (\text{e - geodesic})$$

$$\text{where } \kappa_t = \log \int \exp((1-t)\log f + t\log g)d\mu.$$

Exm 2 $\phi_\beta(s) = \dfrac{s^\beta - 1}{\beta} \Rightarrow \quad f_t(x,\phi_\beta) = \{(1-t)f(x)^\beta + tg(x)^\beta - \kappa_t)^{\frac{1}{\beta}}$

Exm 3 $\phi_\eta(s) = \log(s+\eta) \ \Rightarrow$

$$f_t(x,\phi_\eta) = \exp((1-t)\log(f(x)+\eta) + t\log(g(x)+\eta) - \kappa_t)$$

$$\text{where } \kappa_t = \log\left[\int \exp((1-t)\log(f+\eta) + t\log(g+\eta))d\mu - \eta\right]$$

Identities of ϕ-density

Model $$M = \{f_\theta(x) : \theta \in \Theta\}$$

1st identity $$\int \frac{1}{\phi'(f_\theta)} \frac{\partial}{\partial \theta} \phi(f_\theta) \, \mathrm{d}\mu = 0$$

2nd identity $$-\int \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} \phi(f_\theta) \frac{1}{\phi'(f_\theta)} \mathrm{d}\mu = \int \frac{-\phi''(f_\theta)}{\{\phi'(f_\theta)\}^2} \frac{\partial}{\partial \theta} f_\theta \frac{\partial}{\partial \theta^{\mathrm{T}}} \phi(f_\theta) \, \mathrm{d}\mu$$

because $$\int \frac{\partial}{\partial \theta} f_\theta \, \mathrm{d}\mu = 0 \qquad \int \frac{\partial^2}{\partial \theta \partial \theta^{\mathrm{T}}} f_\theta \, \mathrm{d}\mu = 0$$

Generalized mean and variance

Def

$$\mathrm{E}_f^{(\phi)}\{a(X)\} = \int a(x) f^{(\phi)}(x) d\mu(x)$$

$$\mathrm{Cov}_f^{(\phi)}\{a(X), b(X)\} = \int w^{(\phi)}(f)(a - \mathrm{E}^{(\phi)}a)(b - \mathrm{E}^{(\phi)}b)^{\mathrm{T}} f^{(\phi)} d\mu$$

$$\text{where} \quad f^{(\phi)}(x) = \frac{\dfrac{1}{\phi'(f(x))}}{\displaystyle\int \frac{1}{\phi'(f)} d\mu} \qquad w^{(\phi)}(f) = \frac{-\phi''(f)}{\{\phi'(f)\}^3}$$

Note

$\mathrm{E}_f^{(\phi)}$ is a linear functional on $\mathscr{S}_\mu$

$\mathrm{Cov}_f^{(\phi)}$ is a bilinear positive definite functional on $\mathscr{S}_\mu$

Generalized mean and variance

Exm $\phi(s) = \log s \quad \Rightarrow \quad \mathrm{E}_f^{(\phi)}\{a(X)\} = \int a\, f\, d\mu$

$$\mathrm{Cov}_f^{(\phi)}\{a(X), b(X)\} = \int (a - \mathrm{E}a)(b - \mathrm{E}b)^{\mathrm{T}} f\, d\mu$$

$$\phi(s) = \frac{1}{\beta}(s^{\beta} - 1) \quad \Rightarrow \quad \mathrm{E}_f^{(\phi)}\{a(X)\} = \frac{\int a\, f^{1-\beta}\, d\mu}{\int f^{1-\beta}\, d\mu}$$

$$\mathrm{Cov}_f^{(\phi)}\{a(X), b(X)\} = \frac{\int (a - \mathrm{E}^{(\phi)}a)(b - \mathrm{E}^{(\phi)}b)^{\mathrm{T}} f^{1-2\beta}\, d\mu}{\int f^{1-\beta}\, d\mu}$$

Bartlett Identity

Model

$$M = \{f_\theta(x)\}_{\theta\in\Theta}$$

Bartlett identity

$$\mathrm{E}_{f_\theta}\{\frac{\partial}{\partial\theta}\log f_\theta\} = 0$$

$$\mathrm{E}_{f_\theta}\{-\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\log f_\theta\} = \mathrm{Var}_{f_\theta}(\frac{\partial}{\partial\theta}\log f_\theta)$$

ϕ–Bartlett identities

$$\mathrm{E}^{(\phi)}_{f_\theta}\{\frac{\partial}{\partial\theta_j}\phi(f_\theta)\} = 0$$

$$\mathrm{E}^{(\phi)}_{f}\{-\frac{\partial^2}{\partial\theta\partial\theta^{\mathrm{T}}}\phi(f_\theta)\} = \mathrm{Var}^{(\phi)}_{f}\{\frac{\partial}{\partial\theta}\phi(f_\theta)\}$$

Tangent space of $\mathscr{F}_\mu$

Tangent space $$T_f^{(\phi)} = \{a(x) \in \mathscr{S}_\mu : \mathrm{E}_f^{(\phi)}\{a(X)\} = 0\}$$

Riemannian metric $$\langle a, b \rangle_f^{(\phi)} = \mathrm{Cov}_f^{(\phi)}\{a(X), b(X)\}$$

Expectation gives the tangent space.

Topological properties of $T_f^{(\phi)}$ depend on ϕ.

If ϕ = log, then $T_f^{(\phi)}$ is too large to do statistics on $\mathscr{F}_\mu$

Cf. Pistone (1992)

Parallel transport

Def

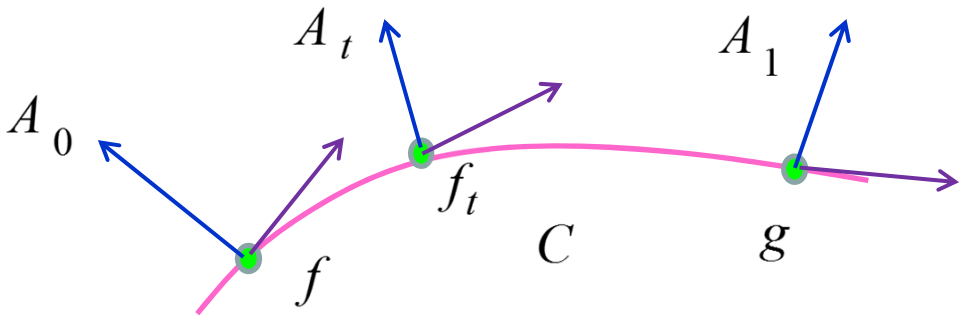
A vector field $\{A_t(x) : t \in [0,1]\}$ **is parallel along a curve** $C = \{f_t : t \in [0,1]\}$

$$\Leftrightarrow \qquad \frac{d}{dt} A_t(x) = \mathrm{Cov}_{f_t}^{(\phi)} \left\{ A_t, \frac{d}{dt} \phi(f_t) \right\}$$

A curve $C = \{f_t : t \in [0,1]\}$ **is ϕ-geodesic**

$$\Leftrightarrow \qquad \frac{d^2}{dt^2} \phi(f_t(x)) = \mathrm{Var}_{f_t}^{(\phi)} \left(\frac{d}{dt} \phi(f_t(X)) \right)$$

Cf. Amari (1982).



ϕ-geodesic

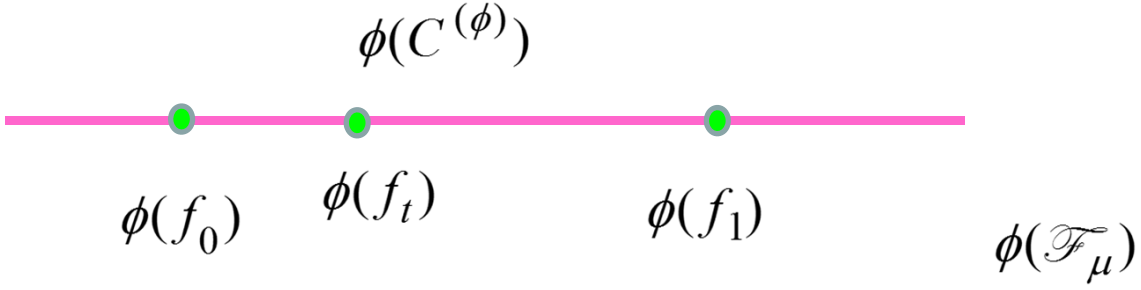
Thm If $C = \{f_t\}_{0 \le t \le 1}$ is the ϕ-geodesic curve then C is the ϕ-path connecting f_0 with f_1.

Proof. By defintion, $\frac{d^2}{dt^2}\phi(f_t(x)) = \text{const in } x \quad (\forall t \in (0, 1))$

which is solved by

$$\phi(f_t(x)) = (1-t)\phi(f_1(x)) + t\phi(f_0(x)) - \kappa_t.$$

■



ϕ-divergence

ϕ-cross entropy $$C^{(\phi)}(f,g) = -\mathrm{E}_f^{(\phi)}\{\phi(g)\}$$

ϕ-entropy $$H^{(\phi)}(f) = C^{(\phi)}(f,f)$$

ϕ-divergence

$$D^{(\phi)}(f,g) = C^{(\phi)}(f,g) - H^{(\phi)}(f) = \mathrm{E}_f^{(\phi)}\{\phi(f) - \phi(g)\}$$

Note: ϕ-divergence is KL-divergence if ϕ = log

Divergence geometry

Def. D is said to be a diveregence measure on $\mathscr{F}_\mu$ if

$$D(f,g) \geq 0 \quad (\forall f, g \in \mathscr{F}_\mu) \text{ with equality iff } f = g \quad (\text{a.e.}\mu)$$

Let $M = \{f_\theta : \theta \in \Theta\}$ be a statistical model.

$$D\,|_{\Theta\times\Theta} \rightarrow (G_{ij}^{(D)}, \Gamma_{ij,k}^{(D)}, {}^*\Gamma_{ij,k}^{(D)}) \text{ on } \Theta$$

with the Riemannian metric on *M* :

$$G_{ij}^{(D)}(\theta) = -\frac{\partial^2}{\partial\theta_{1i}\partial\theta_{2j}} D(f_{\theta 1}, f_{\theta_2})\,|_{\substack{\theta_1=\theta \\ \theta_2=\theta}}$$

the pair of affine connections on *M*:

$$\Gamma_{ij,k}^{(D)}(\theta) = -\frac{\partial^3}{\partial\theta_{1i}\,\partial\theta_{1j}\,\partial\theta_{2k}} D(f_{\theta_1}, f_{\theta_2})\,|_{\substack{\theta_1=\theta \\ \theta_2=\theta}} \quad (\theta \in \Theta)$$

$${}^*\Gamma_{ij,k}^{(D)}(\theta) = -\frac{\partial^3}{\partial\theta_{1i}\,\partial\theta_{1j}\,\partial\theta_{2k}} D(f_{\theta_2}, f_{\theta_1})\,|_{\substack{\theta_1=\theta \\ \theta_2=\theta}} \quad (\theta \in \Theta)$$

ϕ-divergence geometry

The metric

$$G_{ij}^{(\phi)}(\theta) = \int \frac{\partial}{\partial \theta_i} \phi(f_\theta) \frac{\partial}{\partial \theta_j} \frac{1}{\phi'(f_\theta)} d\mu$$

Affine connection pair

$$\Gamma_{ij,k}^{(\phi)}(\theta) = \int \frac{\partial}{\partial \theta_k} \phi(f_\theta) \frac{\partial^2}{\partial \theta_j \partial \theta_k} \frac{1}{\phi'(f_\theta)} d\mu$$

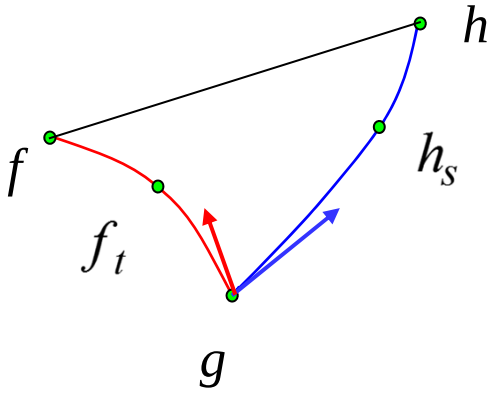
$$^*\Gamma_{ij,k}^{(\phi)}(\theta) = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} \phi(f_\theta) \frac{\partial}{\partial \theta_k} \frac{1}{\phi'(f_\theta)} d\mu$$

ϕ–Pythagorean theorem

Thm **ϕ-geodesic** $f_t(\boldsymbol{x}) = \phi^{-1}((1-t)\phi(g(\boldsymbol{x})) + t\phi(f(\boldsymbol{x})))$

ϕ^*-geodesic $h_s(\boldsymbol{x})$ s.t. $\phi'(h_s) = \dfrac{1}{(1-s)\dfrac{1}{\phi'(g)} + s\dfrac{1}{\phi'(h)}}$

$$G^{(\phi)}(\dot{f}_t, \dot{h}_s)|_{t=0,s=0} = 0 \Rightarrow D_\phi(f,g) = D_\phi(f,g) + D_\phi(g,h)$$



Pf

$$G^{(\phi)}(\dot{f}_t, \dot{h}_s)|_{t=0,s=0}$$

$$= \int \frac{d}{dt}\phi(f_t)\frac{d}{ds}\frac{1}{\phi'(g_s)}d\mu\,|_{t=0,s=0}$$

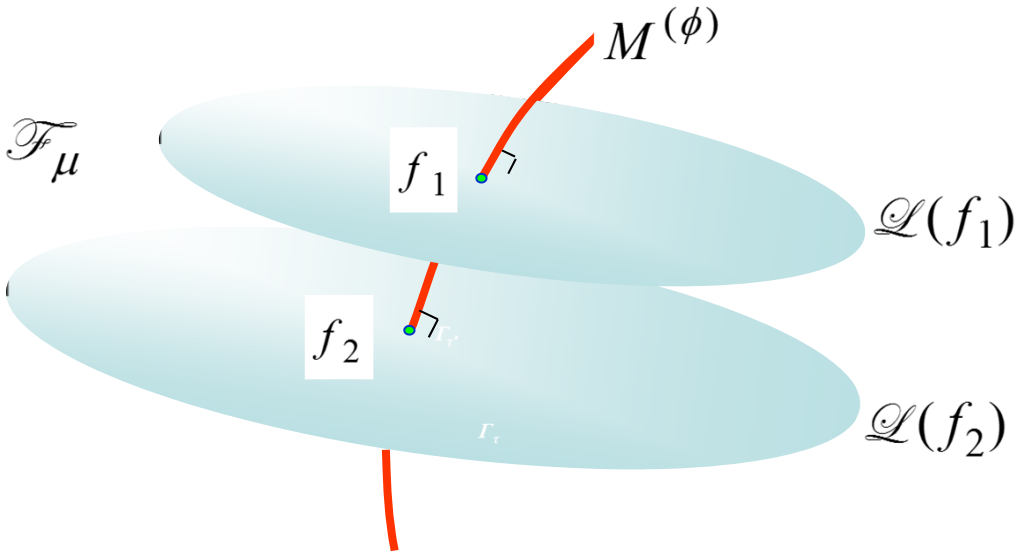
$$= \int \{\phi(f) - \phi(g)\}\{\frac{1}{\phi'(g)} - \frac{1}{\phi'(h)}\}d\mu$$

$$= D_\phi(f,h) - \{D_\phi(f,g) + D_\phi(g,h)\}$$ ■

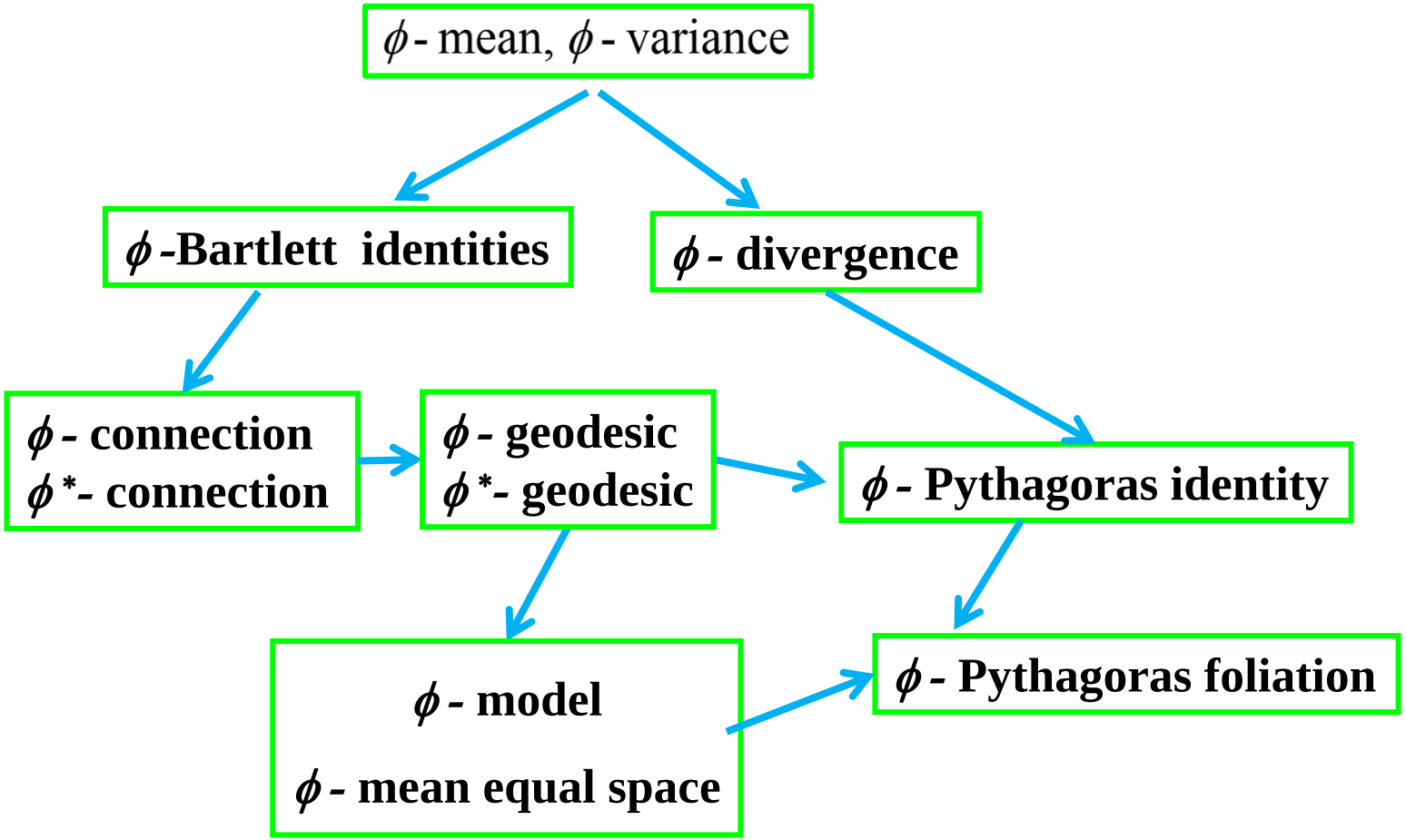
ϕ-Pythagorean foliation

If f satisfies $\mathrm{E}_f^{(\phi)}\{t(X)\} = \mathrm{E}_{f_{\theta^*}^{(\phi)}}^{(\phi)}\{t(X)\}$, then $D_\phi(f, f_\theta^{(\phi)}) = D_\phi(f, f_{\theta^*}^{(\phi)}) + D_\phi(f_{\theta^*}^{(\phi)}, f_\theta^{(\phi)})$

ϕ-mean equal space $\mathscr{L}(g) = \{f : \mathrm{E}_f^{(\phi)}\{t(X)\} = \mathrm{E}_g^{(\phi)}\{t(X)\}\}$



KN-mean



What is a statistical meaning of ϕ-mean and ϕ-variance?

ϕ- independence, ϕ- LLN, ϕ- CLT, ϕ- moment generating function, ...

U-divergence

Let U be a function satisfying $U'(s) = \phi^{-1}(s)$

***U*-cross-entropy** $C_U(f,g) = \int \{-f\,\phi(g) + U(\phi(g))\}d\mu$

***U*-entropy** $H_U(f) = C_U(f,f)$

***U*-divergence** $D_U(f,g) = C_U(f,g) - H_U(f)$

Note $C_U(f,g) \geq H_U(f)$ or $D_U(f,g) \geq 0$

Exm Let $U_\beta(s) = \frac{1}{1+\beta}(1-\beta s)^{\frac{1+\beta}{\beta}}$.

Then power entropy $C_{U_\beta}(f,g) = -\frac{1}{\beta}\int f g^\beta \mathrm{d}\mu + \frac{1}{\beta+1}\int g^{\beta+1}\mathrm{d}\mu$

power divergence $D_{U_\beta}(f,g) = \frac{1}{\beta}\int f(f^\beta - g^\beta)\mathrm{d}\mu + \frac{1}{\beta+1}\int (f^{\beta+1} - g^{\beta+1})\mathrm{d}\mu$

U-divergence geometry

The metric associated with *U*-divergence:

$$G_{ij}^{(U)}(\theta) = \int \frac{\partial}{\partial \theta_i} f_\theta \, \frac{\partial}{\partial \theta_j} \phi(f_\theta) d\mu$$

Affine connections associated with *U*-divergence:

$$\Gamma_{ij,k}^{(U)}(\theta) = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta \, \frac{\partial}{\partial \theta_k} \phi(f_\theta) d\mu$$

$$^{*}\Gamma_{ij,k}^{(U)}(\theta) = \int \frac{\partial}{\partial \theta_k} f_\theta \, \frac{\partial^2}{\partial \theta_j \partial \theta_k} \phi(f_\theta) d\mu$$

Thm

(i) *U*-geodesic is mixture geodesic.

(ii) *U**-geodesic is ϕ geodesic

U-geometry ≠ ϕ-geometry

ϕ-metric on a model M

$$G_{ij}^{(\phi)}(\theta) = \int \frac{\partial}{\partial \theta_i} \phi(f_\theta) \frac{\partial}{\partial \theta_j} \frac{1}{\phi'(f_\theta)} d\mu$$

U-metric on a model M

$$G_{ij}^{(U)}(\theta) = \int \frac{\partial}{\partial \theta_i} f_\theta \frac{\partial}{\partial \theta_j} \phi(f_\theta) d\mu$$

ϕ*-connection

$${}^*\Gamma_{ij,k}^{(\phi)}(\theta) = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} \phi(f_\theta) \frac{\partial}{\partial \theta_k} \frac{1}{\phi'(f_\theta)} d\mu$$

U*-connection

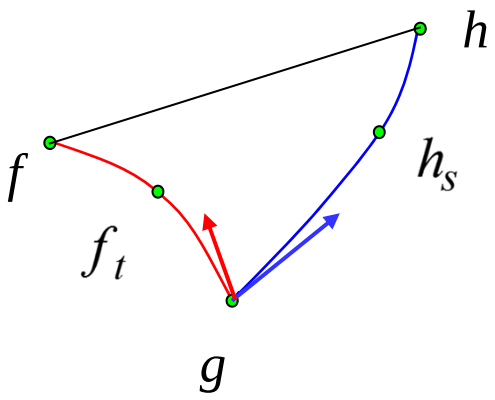
$${}^*\Gamma_{ij,k}^{(U)}(\theta) = \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} \phi(f_\theta) \frac{\partial}{\partial \theta_k} f_\theta \, d\mu$$

Triangle with D_U

Thm **mixture geodesic** $f_t(\boldsymbol{x}) = (1-t)g(\boldsymbol{x}) + t\,f(\boldsymbol{x})$

ϕ-geodesic $h_s(\boldsymbol{x}) = \phi^{-1}\big((1-s)\phi(g(\boldsymbol{x})) + s\phi(h(\boldsymbol{x})) - \kappa_s\big)$

$$G^{(U)}(\dot{f}_t, \dot{h}_s)|_{t=0,s=0} = 0 \Rightarrow D_U(f,g) = D_U(f,g) + D_U(g,h)$$



Pf

$$G^{(U)}(\dot{f}_t, \dot{h}_s)|_{t=0,s=0}$$

$$= \int \frac{d}{dt} f_t \frac{d}{ds} \xi(g_s) d\mu|_{t=0,s=0}$$

$$= \int (f-g)\{\phi(h) - \phi(g)\} d\mu$$

$$= D_U(f,h) - \{D_U(f,g) + D_U(g,h)\}$$

■

U-estimation

Let $g(\boldsymbol{x})$ be a data density function with statistical model $f_{\boldsymbol{\theta}}(\boldsymbol{x})$

U-loss function

$$L_U(\boldsymbol{\theta}) = C_U(f, f_{\boldsymbol{\theta}}) = -\mathrm{E}_f\{\phi(f_{\boldsymbol{\theta}})\} + \int U(\phi(f_{\boldsymbol{\theta}}))\mathrm{d}\mu$$

U-empirical loss function

$$L_U^{\mathrm{emp}}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\phi(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \int U(\phi(f_{\boldsymbol{\theta}}))\mathrm{d}\mu$$

U-estimator for $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}}_U = \arg\min_{\boldsymbol{\theta}\in\Theta} L_U^{\mathrm{emp}}(\boldsymbol{\theta})$$

U-estimator under ϕ-model

ϕ-model

$$M^{(\phi)} = \{ f_\theta^{(\phi)}(\boldsymbol{x}) = \phi^{-1}(\boldsymbol{\theta}^T \boldsymbol{t}(\boldsymbol{x}) - \kappa_\theta) : \boldsymbol{\theta} \in \Theta \}$$

U-empirical loss function

$$L_U(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \phi(f_\theta^{(\phi)}(\boldsymbol{x}_i)) + \int U(\phi(f_\theta^{(\phi)}))d\mu$$

$$= -\{ \boldsymbol{\theta}^{\mathrm{T}} \bar{\boldsymbol{t}} - \Phi(\boldsymbol{\theta}) \}$$

where $\bar{\boldsymbol{t}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{t}(\boldsymbol{x}_i)$, $\Phi(\boldsymbol{\theta}) = \kappa(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{t} - \kappa(\boldsymbol{\theta}))d\mu$

U-estimator for θ

$$\hat{\boldsymbol{\theta}}_U = \underset{\boldsymbol{\theta} \in \Theta}{\arg\,\mathrm{solve}} \{ \frac{\partial}{\partial \boldsymbol{\theta}} \Phi(\boldsymbol{\theta}) = \bar{\boldsymbol{t}} \} = \left(\frac{\partial \Phi}{\partial \boldsymbol{\theta}} \right)^{-1} (\bar{\boldsymbol{t}})$$

U-estimator under ϕ-model has analogy with MLE under exponential model

Potential function Φ

Def We call $\Phi(\boldsymbol{\theta}) = \kappa(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t} - \kappa(\boldsymbol{\theta}))d\mu$

the potential function on ϕ-model

Note

$$\frac{\partial}{\partial \boldsymbol{\theta}}\Phi(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}\kappa(\boldsymbol{\theta}) + \int \{\boldsymbol{t}(\boldsymbol{x}) - \frac{\partial}{\partial \boldsymbol{\theta}}\kappa(\boldsymbol{\theta})\} f_{\boldsymbol{\theta}}^{(U)}(\boldsymbol{x})d\mu(\boldsymbol{x})$$

$$= \mathrm{E}_{f_{\boldsymbol{\theta}}^{(\phi)}}\{\boldsymbol{t}(\boldsymbol{X})\}$$

We define the mean parameter η by

$$\boldsymbol{\eta} = \frac{\partial}{\partial \boldsymbol{\theta}}\Phi(\boldsymbol{\theta})$$

Cf. ϕ–mean parameter $\mathrm{E}_{f_{\boldsymbol{\theta}}}^{(\phi)}\{\boldsymbol{t}(\boldsymbol{X}\} = \frac{\partial}{\partial \boldsymbol{\theta}}\kappa(\boldsymbol{\theta})$

Thm *U*-estimator for η is given by the sample mean

$$\hat{\boldsymbol{\eta}}_U = \bar{\boldsymbol{t}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{t}(\boldsymbol{x}_i),$$

Pythagoras foliation

Thm If a pdf $\hat{f}$ satisfies $\mathrm{E}_{\hat{f}}\{\boldsymbol{t}(\boldsymbol{X})\} = \bar{\boldsymbol{t}}$, then

$$D_U(\hat{f}, f_{\boldsymbol{\theta}}) = D_U(\hat{f}, f_{\hat{\boldsymbol{\theta}}_U}) + D_U(f_{\hat{\boldsymbol{\theta}}_U}, f_{\boldsymbol{\theta}}).$$

Pf Consider

$$\hat{f}_t = t\hat{f} + (1-t) f_{\hat{\boldsymbol{\theta}}_U}^{(U)} \quad , h_s = \phi^{-1}(s\phi(f_{\boldsymbol{\theta}}^{(U)}) + (1-s)\phi(f_{\hat{\boldsymbol{\theta}}_U}^{(U)}))$$

Then

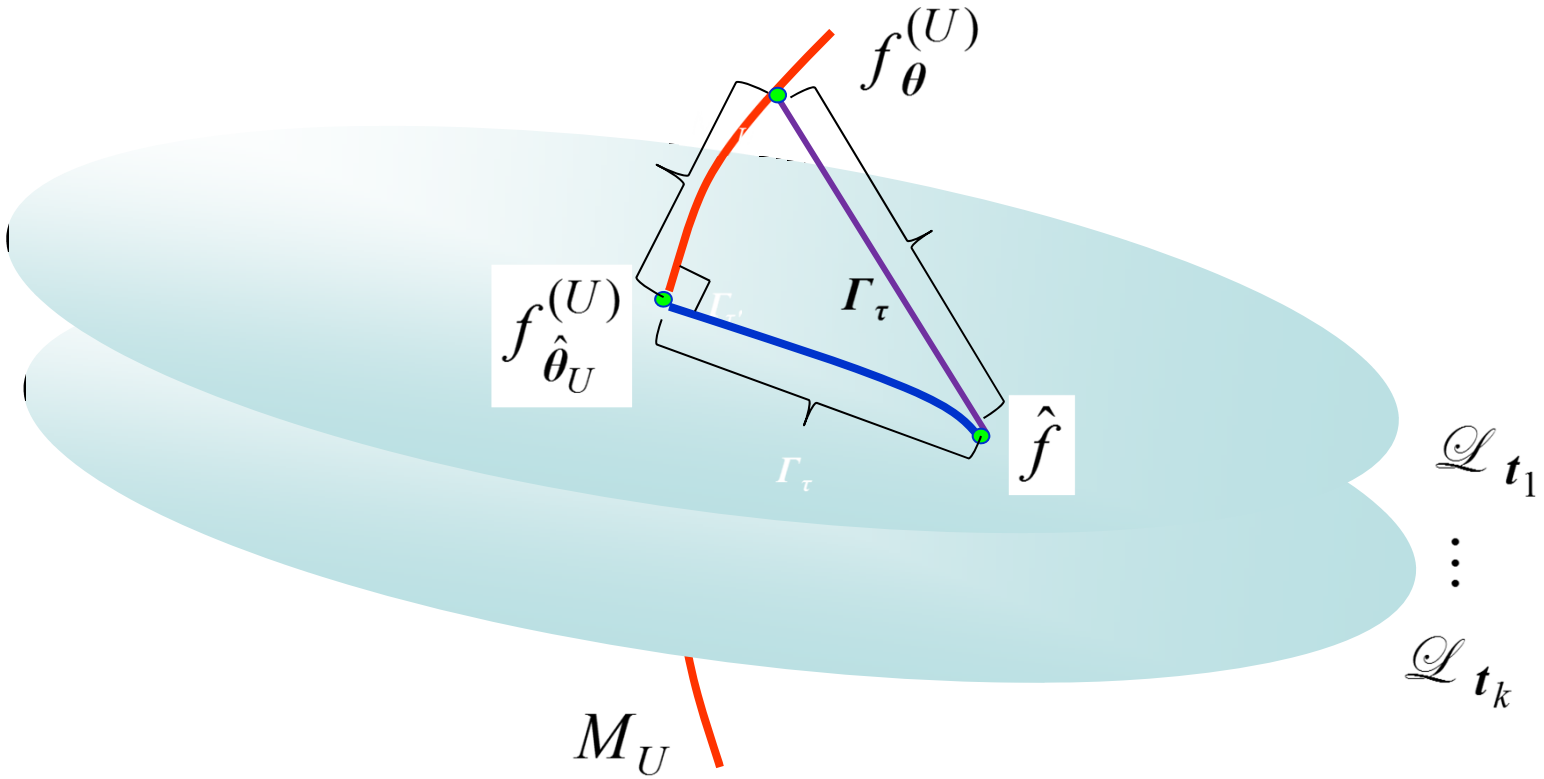
$$G^{(U)}(\dot{\hat{f}}_t, \dot{h}_s) = \int (\hat{f} - f_{\hat{\boldsymbol{\theta}}_U}^{(U)})\{\phi(f_{\hat{\boldsymbol{\theta}}_U}^{(U)}) - \phi(f_{\boldsymbol{\theta}}^{(U)})\} d\mu$$

$$= (\hat{\boldsymbol{\theta}}_U - \boldsymbol{\theta})^{\mathrm{T}} (\mathrm{E}_{\hat{f}}\{\boldsymbol{t}(X)\} - \mathrm{E}_{f_{\hat{\boldsymbol{\theta}}_U}^{(U)}}\{\boldsymbol{t}(X)\}) = 0$$

Pythagoras foliation

$$D_U(\hat{f}, f_{\boldsymbol{\theta}}^{(U)}) = D_U(\hat{f}, f_{\hat{\boldsymbol{\theta}}_U}^{(U)}) + D_U(f_{\hat{\boldsymbol{\theta}}_U}^{(U)}, f_{\boldsymbol{\theta}}^{(U)})$$

$$\mathscr{F}_\mu = \bigcup_{f \in M^{(U)}} \mathscr{L}(f) \quad \text{where} \quad \mathscr{L}(f) = \{\, g : \mathrm{E}_g\{\boldsymbol{t}(\boldsymbol{X})\} = \mathrm{E}_f\{\boldsymbol{t}(\boldsymbol{X})\} \,\}$$



U-Boost learning for density estimation

U-loss function $$L_U(f) = -\frac{1}{n}\sum_{i=1}^{n}\phi(f(\boldsymbol{x}_i)) + \int_{\mathbb{R}^p} U(\phi(f(\boldsymbol{x})))\mathrm{d}\boldsymbol{x}$$

Dictionary of density functions

$$\mathscr{D} = \{\, g_\lambda(\boldsymbol{x}) : g_\lambda(\boldsymbol{x}) \geq 0,\ \int g_\lambda(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1,\ \ \lambda \in \Lambda\}$$

Learning space = ϕ-model

$$\mathscr{D}_\phi^{\,*} = \phi^{-1}\big(\mathrm{co}(\phi(\mathscr{D}))\big) = \{\, \phi^{-1}\Big(\sum_{\lambda\in\Lambda}\pi_\lambda\,\phi(g_\lambda(\boldsymbol{x}))\Big)\}$$

- Let $f(\boldsymbol{x},\boldsymbol{\pi}) = \phi^{-1}\Big(\sum_{\lambda\in\Lambda}\pi_\lambda\,\phi(g_\lambda(\boldsymbol{x}))\Big)$. Then $f(\boldsymbol{x},\underset{(\lambda)}{(0,\cdots,1,\cdots 0)}) = g_\lambda(\boldsymbol{x})$
- $\mathscr{D}_\phi^{\,*} \supseteq \mathscr{D}$

Goal : find $$f^* = \underset{f\in\mathscr{D}_U^{\,*}}{\mathrm{argmin}}\, L_U(f)$$

U-Boost algorithm

(A) Find $f_1 = \arg\min_{g \in \mathscr{D}} L_U(g)$

(B) Update $f_k \longrightarrow f_{k+1} = \xi^{-1}((1-\alpha_{k+1})\xi(f_k) + \alpha_{k+1}\xi(g_{k+1}))$ st

$$(\alpha_{k+1}, g_{k+1}) = \arg\min_{(\alpha, g) \in (0,1) \times \mathscr{D}} L_U(\xi^{-1}((1-\alpha)\xi(f_k) + \alpha\xi(g)))$$

(C) Select K, and $\hat{f} = \xi^{-1}((1-\alpha_K)\xi(f_{K-1}) + \alpha_K \xi(g_K))$

Example 2. Power entropy $f^*(x) = \left(\sum_k \pi_k \, g_k(x)^\beta\right)^{\frac{1}{\beta}}$

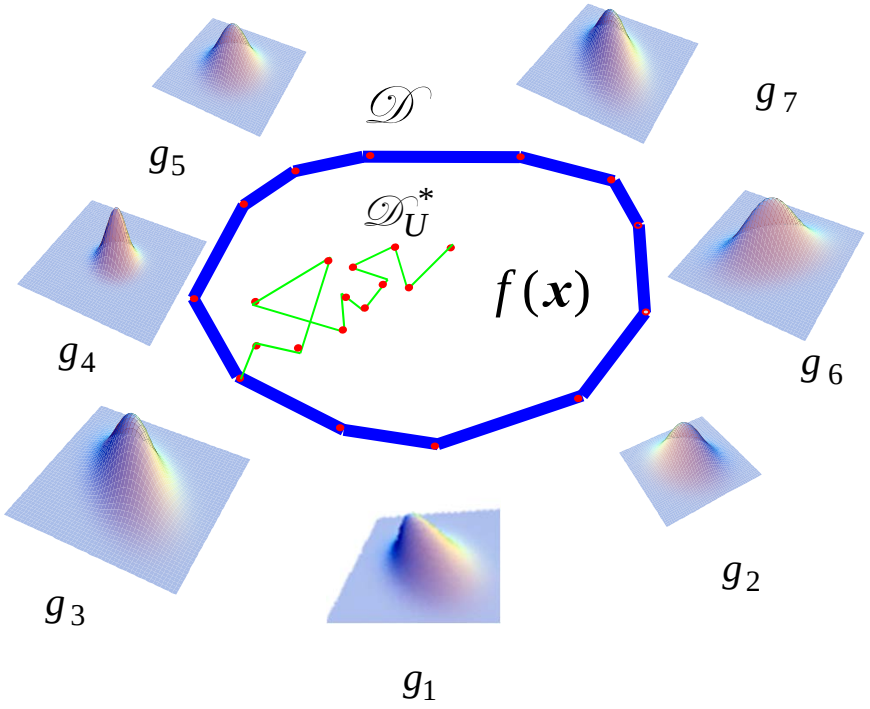
If $\beta = 0$, $f^*(\boldsymbol{x}) = \exp\left(\sum_k \pi_k \, \log g_k(\boldsymbol{x})\right) = \prod_k g_k(\boldsymbol{x})^{\pi_k}$ Friedman et al (1984)

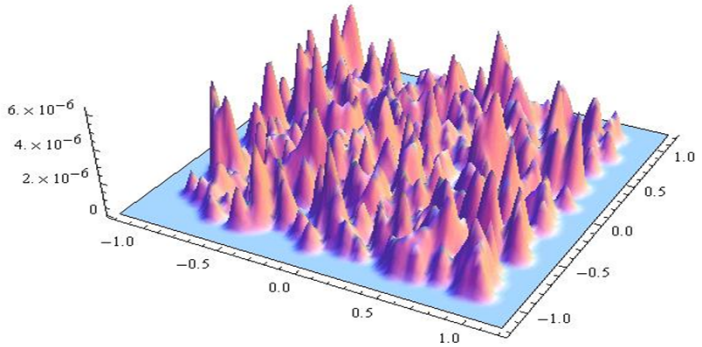
If $\beta = 1$, $f^*(\boldsymbol{x}) = \sum_k \pi_k \, g_k(\boldsymbol{x})$ Klemela (2007)

Inner step in the convex hull

$$\mathscr{D} = \{t(x,\lambda) : \lambda \in \Lambda\} \longrightarrow \mathscr{D}_U^{\,*} = \phi^{-1}(\mathrm{co}\,\phi(\mathscr{D}))$$

Goal : $f^* = \underset{f \in \mathscr{D}_U^{\,*}}{\mathrm{argmin}}\, L_U(f)$ $\quad f^*(\boldsymbol{x}) = \phi^{-1}(\hat{\pi}_1 \phi(\hat{f}_1(\boldsymbol{x})) + \cdots + \hat{\pi}_{\hat{k}} \phi(\hat{f}_{\hat{k}}(\boldsymbol{x})))$





$f(\boldsymbol{x})$

Non-asymptotic bound

Theorem. Assume that a data distribution has a density $g(\boldsymbol{x})$ and that

$$\text{(A)} \qquad \sup_{(\psi,\xi,\varphi)\in \mathrm{co}(\xi(\mathscr{D}))\times\mathscr{D}\times\mathscr{D}} \int U''(\psi)\{\phi(\xi)-\phi(\varphi)\}^2 \le b_U$$

Then we have

$$\mathbb{E}_g D_U(g, \hat{f}_K) \le \mathrm{FA}(g, \mathscr{D}_U^*) + \mathrm{EE}(g, \mathscr{D}) + \mathbb{E}(K, \mathscr{D}),$$

where

$$\mathrm{FA}(g, \mathscr{D}_U^*) = \inf_{f \in \mathscr{D}_U^*} D_U(g, f) \qquad \text{(Functional approximation)}$$

$$\mathrm{EE}(g, \mathscr{D}) = 2\,\mathbb{E}_g\Big\{ \sup_{f\in\mathscr{D}} \Big| \frac{1}{n}\sum_{i=1}^{n} \phi(f(\boldsymbol{x}_i)) - \mathbb{E}_p \phi(f) \Big| \Big\} \qquad \text{(Estimation error)}$$

$$\mathrm{IE}(K, \mathscr{D}) = \frac{c^2 {b_U}^2}{K+c-1} \quad (c\text{:step-lengthconstant}) \qquad \text{(Iteration effect)}$$

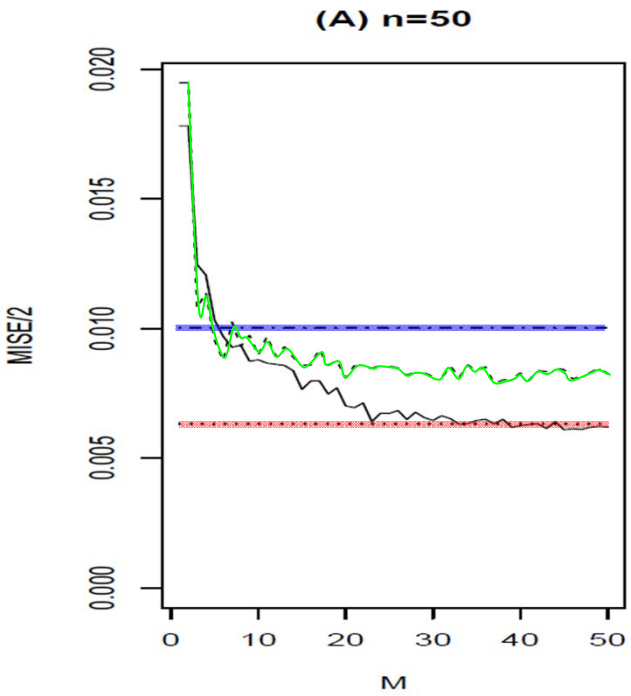
Remark. Trade between $\mathrm{FA}(p, \mathscr{D}_U^*)$ and $\mathrm{EE}(p, \mathscr{D})$

β-**Boost** (β=1)

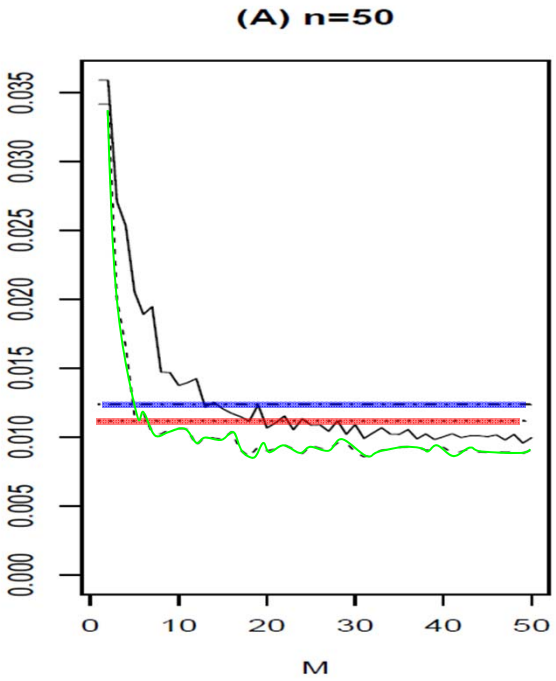
β-**Boost** (β=1/2)

KDE

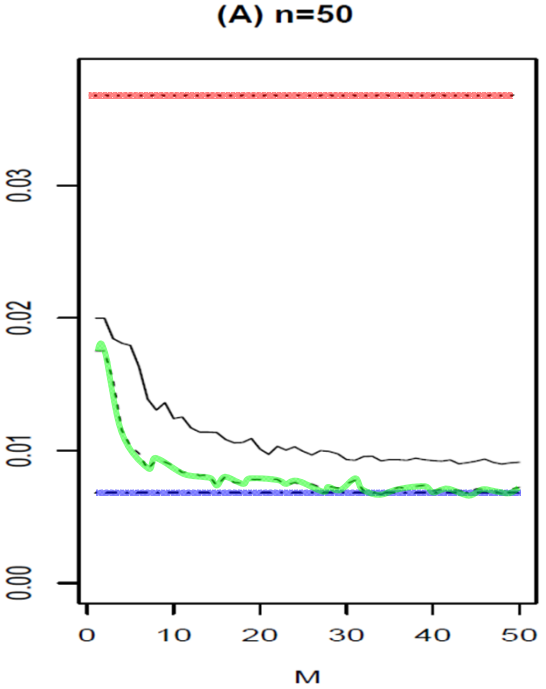
RSDE (Girolami-He, 2004)



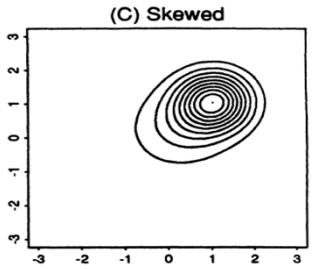
C (skewed-unimodal)

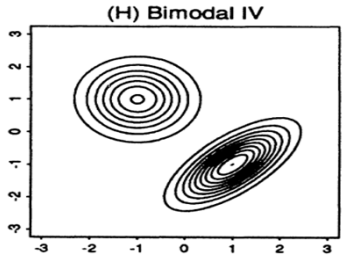


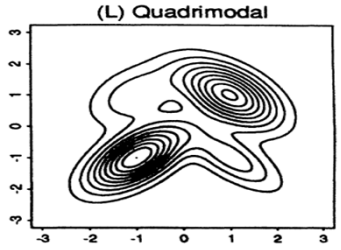
H (bimodal)



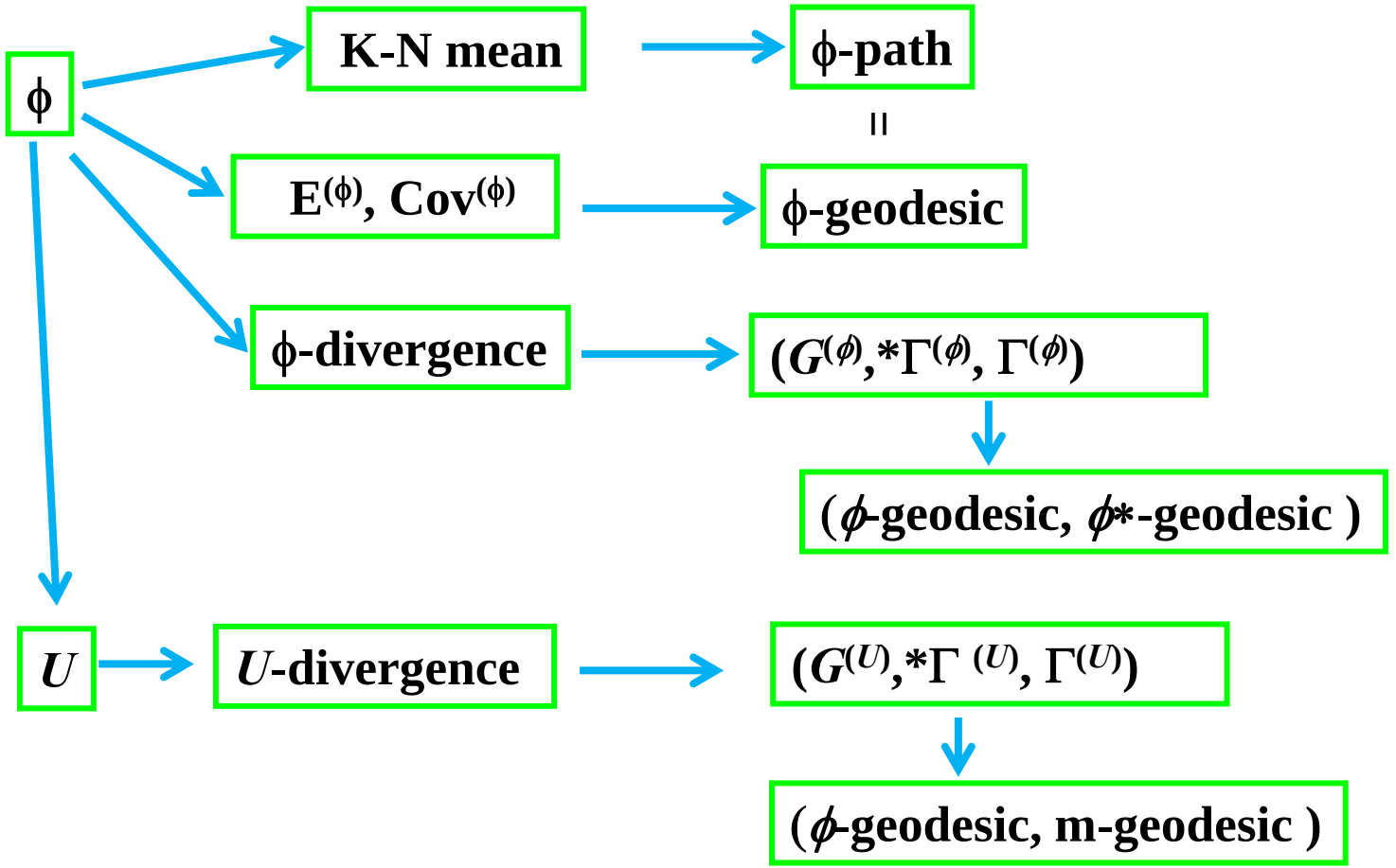
L (quadrimodal)







Conclusion


ϕ
K-N mean
ϕ-path
=
E(ϕ), Cov(ϕ)
ϕ-geodesic
ϕ-divergence
(G(ϕ),*Γ(ϕ), Γ(ϕ))
(ϕ-geodesic, ϕ*-geodesic)
U
U-divergence
(G(U),*Γ (U), Γ(U))
(ϕ-geodesic, m-geodesic)
ϕ-path
$\mathscr{D}_U^{\ *} = \phi^{-1}(\mathrm{co}\phi(\mathscr{D}))$

Future problems

Tangent space

$$T_f^{(\phi)} = \{a(x) \in H_f^{(\phi)} : \mathrm{E}_f^{(\phi)}\{a(X)\} = 0\}$$

where $\langle a, b \rangle_f^{(\phi)} = \mathrm{Cov}_f^{(\phi)}\{a(X), b(X)\}$

Path space

$$\mathscr{P}_f^{(\phi)} = \{P^{(\phi)}(f, g) : g \in \mathscr{F}_\Lambda\}$$

where $P^{(\phi)}(f, g) = \{\phi^{-1}((1-t)\phi(f) + t\phi(g) - \kappa_t) : t \in [0,1]\}$

Future problems

ϕ-mean, ϕ-variance, ϕ-divergence

These are natural ideas from IG view point

We can build ϕ-efficiency in estimation theory, but

What is a statistical meaning of ϕ-mean and ϕ-variance?

Can we define a random sample of ϕ-version?

ϕ - independence, ϕ - LLN, ϕ - CLT, ϕ - moment generating function, ...

Cf. q - independec.....

Thank you