Clusterização com Divergências de Bregman

Prof. Heudson Mirandola

mirandola@ufrj.br

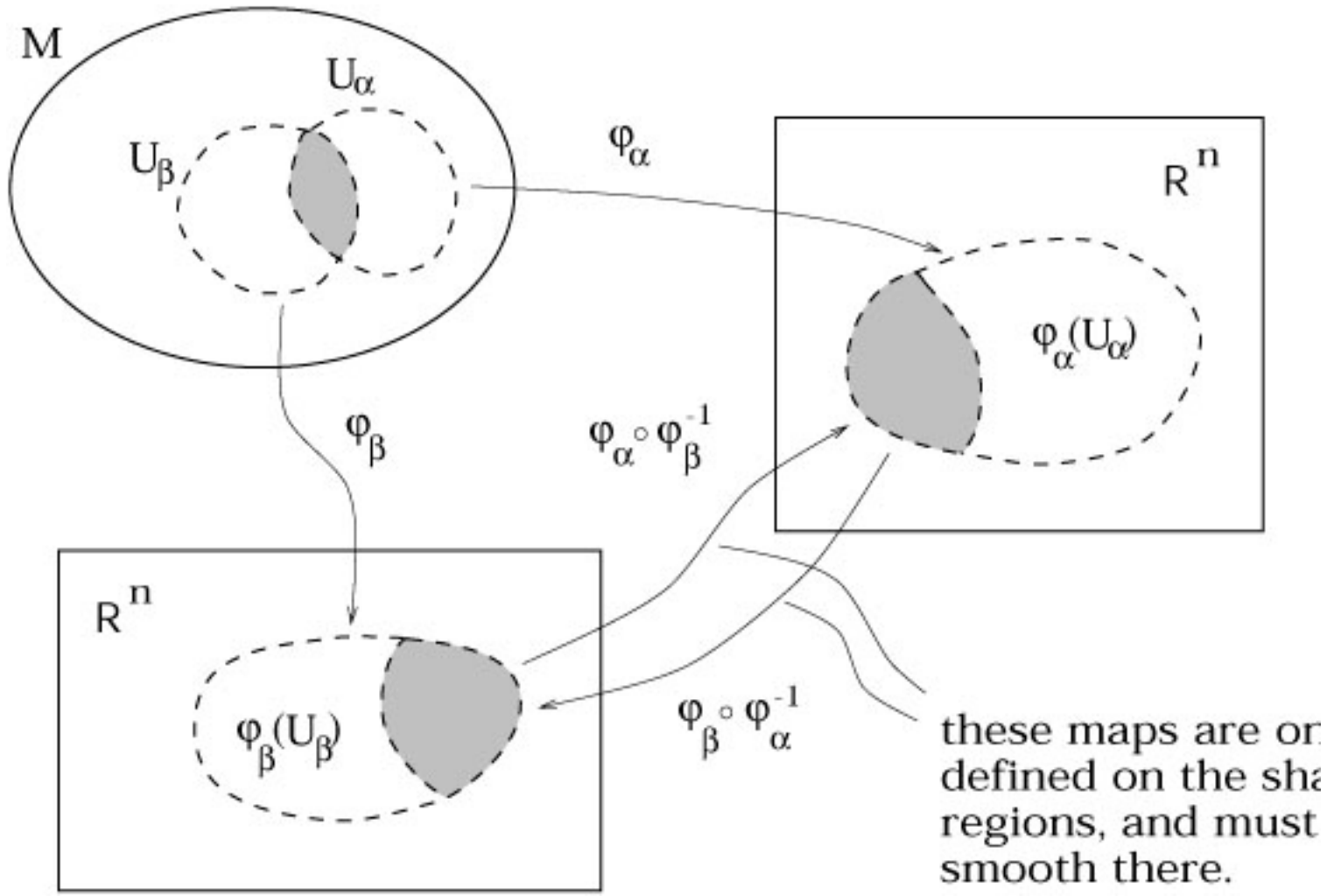
Geometria da Informação:

- Nasce do estudo de invariantes geométricos envolvido na inferência estatística.

- Em IG, tem-se uma métrica Riemanniana e uma conexão afim (em geral não-Riemanniana) sobre uma variedade imersas no espaço das distribuições de probabilidades sobre algum espaço amostral.

- Tais estruturas diferenciáveis podem ser obtidas a partir de uma divergência (que quantifica dissimilaridades entre dois pontos da variedade estatística)

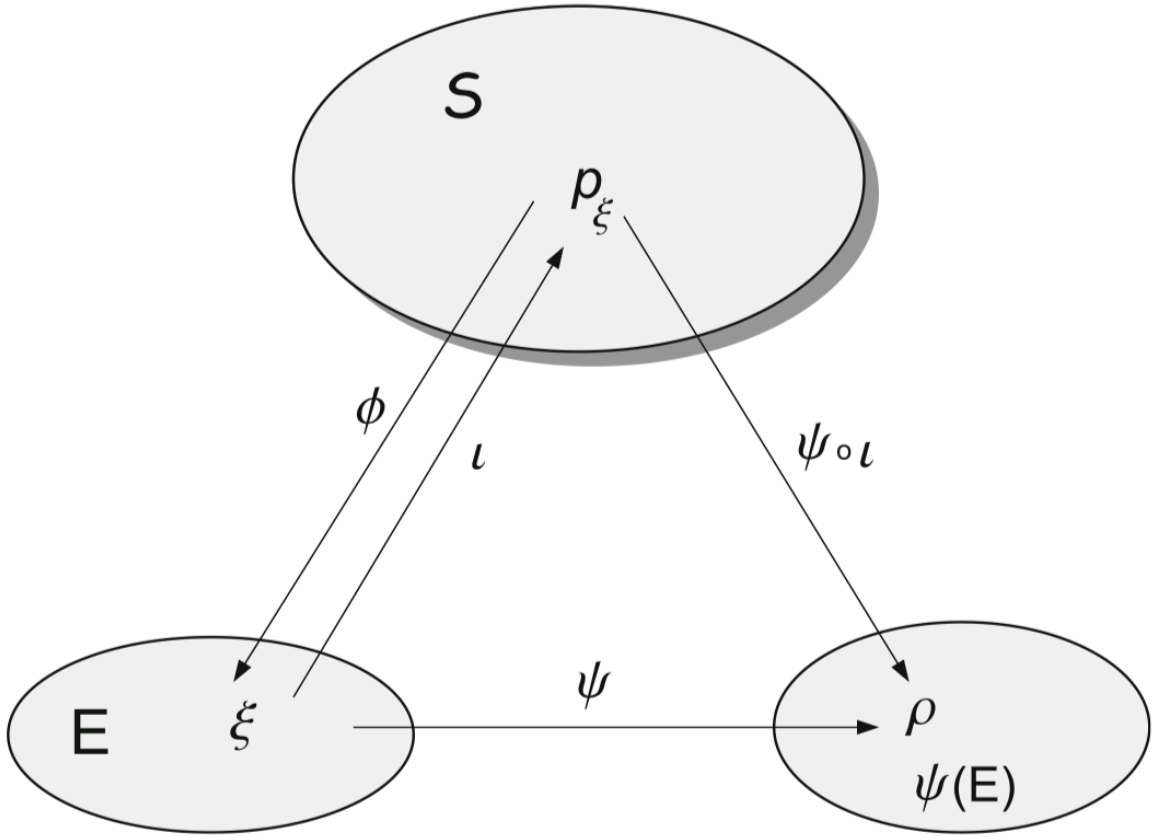
Variedades Diferenciáveis



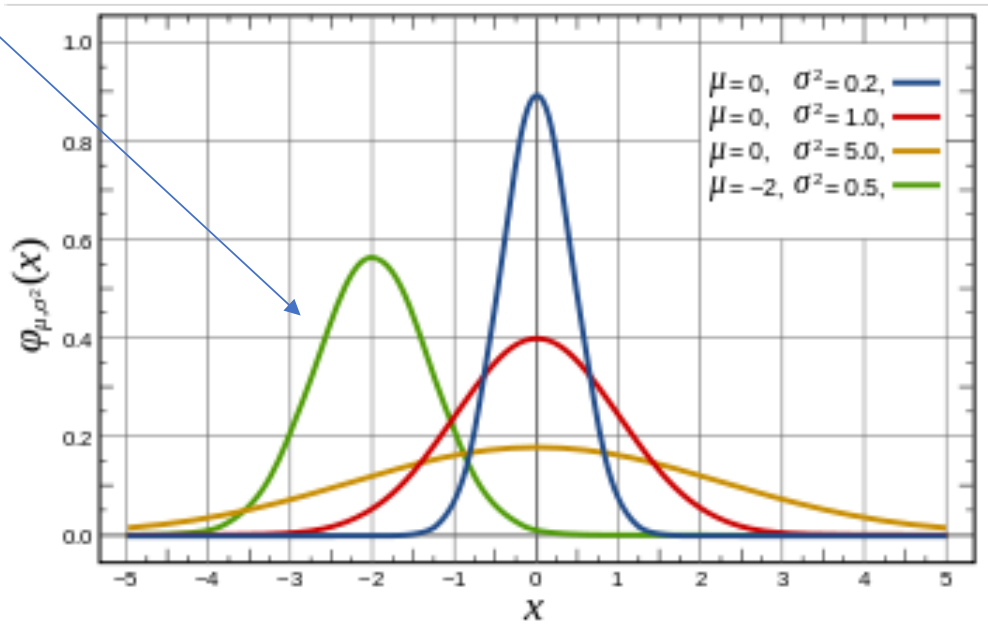
Modelos Estatísticos

$$S = \{p_\xi : \mathcal{X} \to (0, \infty) \mid \xi \in E \text{ aberto de } \mathbb{R}^n \text{ e } \int p_\xi(x) d\mu(x) = 1\}$$

é uma variedade cuja parametrização pode ser dada por $\xi \in E \mapsto p_\xi \in S$



Cada shape é um ponto de S



Dist. Gaussianas

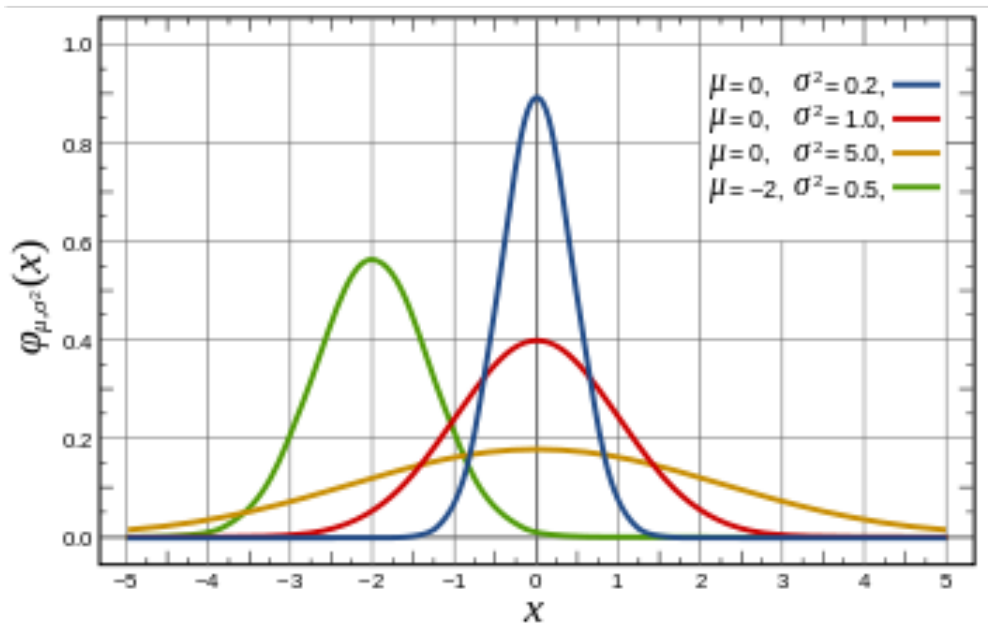
$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Coordenadas:

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

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

$$\xi = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$$

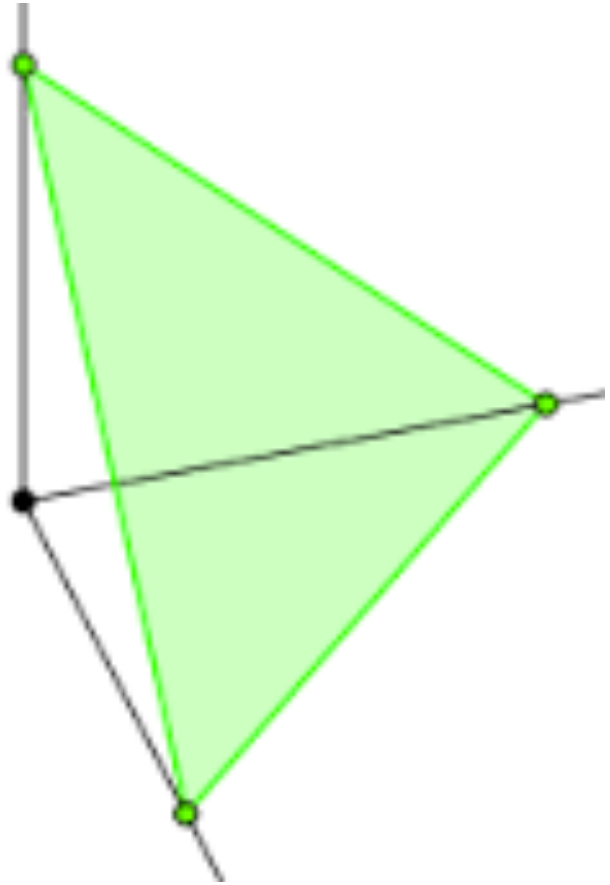


Distribuição Categórica

X variável aleatória com $\mathcal{X} = \{0, 1, \ldots, n\}$

Defina: $p_i = P[X = i]$

$$S_n = \{p = (p_0, \ldots, p_n) \mid p_i > 0 \text{ e } \sum_i p_i = 1\}$$



Coordenadas:
$$\eta = (p_1, \ldots, p_n)$$
$$\theta = (\theta_1, \ldots, \theta_n) \quad \text{com} \quad \theta_i = \log(p_i/p_0)$$

$$p(x) = p_1\delta(x-1) + \ldots + p_k\delta(x-k) + (1 - \sum_i p_i)\delta(x-0)$$

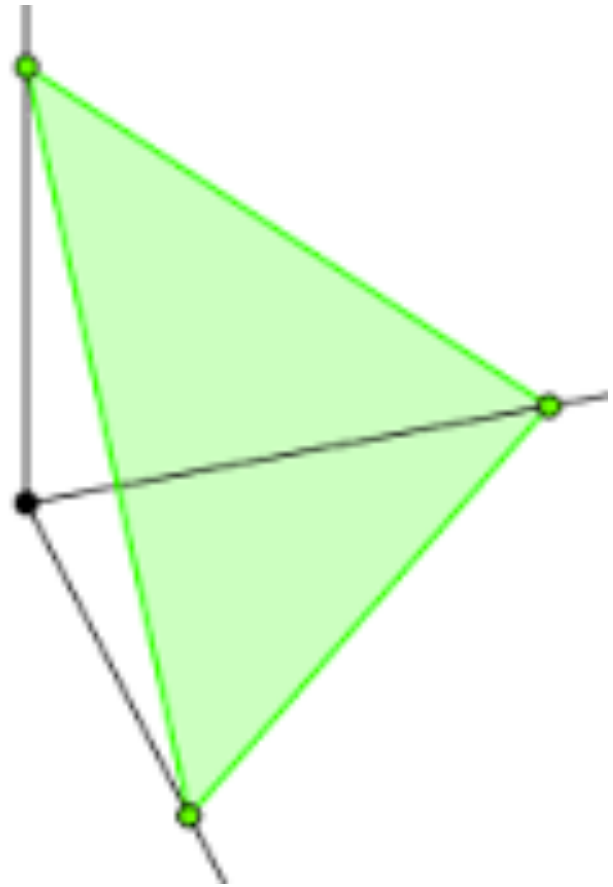
Misturas de Probabilidades

Sejam $P_0, \ldots, P_k$ dist. de probabilidades sobre χ l.i.

$$p_\eta(x) = (1 - \sum \eta_i) P_0(x) + \eta_1 P_1(x) + \ldots + \eta_k P_k(x)$$

$\eta = (\eta_1, \ldots, \eta_k)$ com $\eta_i > 0$ e $\sum_i \eta_i < 1$.

$S_k = \{p_\eta\}$ é um modelo estatístico de dimensão k.



Multinomial: $p(x) = p_1 \delta(x-1) + \ldots + p_k \delta(x-k) + (1 - \sum_i p_i) \delta(x-0)$

Famílias Exponenciais

Sejam $C, F_1, \ldots, F_k : \chi \to \mathbb{R}$ funções, tais que $1, F_1, \ldots, F_k$ sejam l.i.

Considere $p_\theta(x) = e^{C(x)+\langle\theta, F(x)\rangle - \psi(\theta)}$, onde $F = (F_1, \ldots, F_k)$ e $\theta \in E$

$$E = \{\theta \mid \psi(\theta) := \log(\int e^{C(x)+\langle\theta,F(x)\rangle})\} < \infty\}$$

Alguns exemplos

Normal: $N(x \mid \mu, \Sigma) = \frac{e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}$.

Exponencial: $f(x \mid \lambda) = \lambda e^{-\lambda x}$.

Gamma: $f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$.

Poisson: $f(x \mid \lambda) = \frac{\lambda^x}{x!} e^{-x}$.

Beta: $f(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$.

Dirichlet: $f(x_1, \ldots, x_k \mid \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_1+\ldots\alpha_k)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} \Pi_i x_i^{\alpha_i}$.

Bernoulli: $f(x \mid \lambda) = \lambda^x(1-\lambda)^{1-x}$

Binomial (com N fixo): $f(x \mid \lambda) = \binom{N}{x}\lambda^x(1-\lambda)^{N-x}$

Categórica: $P[X = i \mid (p_0, \ldots, p_k)] = p_i$

Famílias Exponenciais

Como $p_\theta(x) = s(F(x);\theta)\, t(x) \;\rightsquigarrow\; F$ é uma estatistica suficiente.
$p_\theta(x|F(x))$ independe de θ.

$$Y = F(X) \text{ v.a. } \rightsquigarrow p_\theta(y) = \int_{F(x)=y} p_\theta(x) d\mu(x) = e^{\langle\theta,y\rangle - \psi(\theta)} \int_{F(x)=y} e^{C(x)} d\mu(x)$$

$$S = \{p_\theta(y) = e^{\langle\theta,y\rangle - \psi(\theta)}\} \quad \text{sobre } (F(\chi), d\mu(y)) \qquad d\mu(y)$$

Familias exponenciais naturais

Coordenadas: θ (coordenadas canônicas)

$\eta = E_{p_\theta}[y] = \nabla\psi(\theta)$ (coordenadas experadas)

Espaço das Densidades Positivas

$$P(\chi) = \{p : \chi \to (0,\infty) \mid \int p(x)d\mu(x) = 1\} \quad \text{(Espaço das dist. de probabilidade positivas sobre } \chi)$$

$$Dens_+(\chi) = \{p : \chi \to (0,\infty) \mid \int p(x)d\mu(x) < \infty\} \quad \text{(Espaço das densidades positivas sobre } \chi)$$

$$\chi = \{0, 1, \ldots, k\} \qquad \rightsquigarrow \qquad P(\chi) = S_k = \{p = (p_0, \ldots, p_k) \mid p_i > 0 \text{ e } \sum p_i = 1\}$$

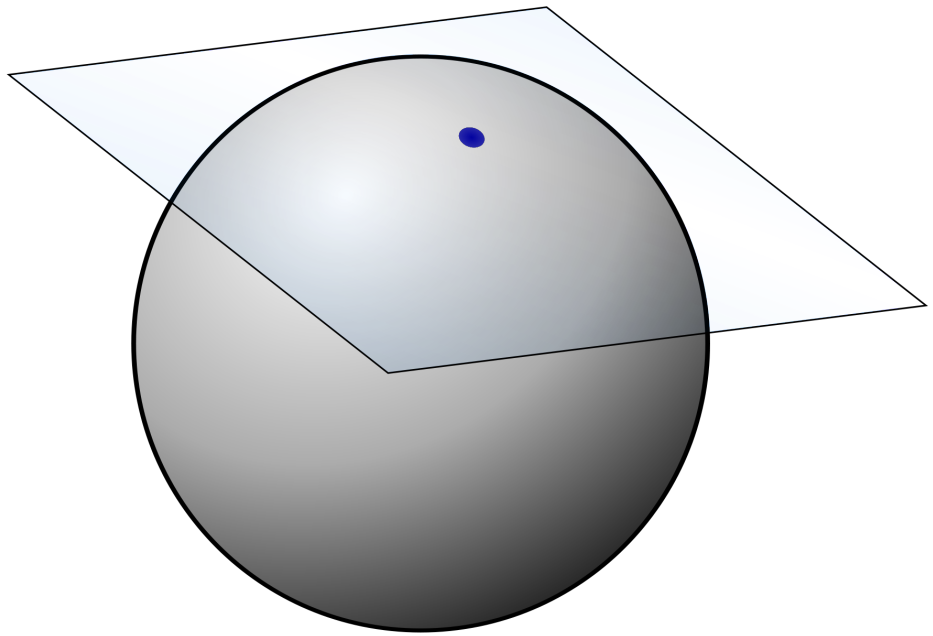
$$Dens_+(\chi) = \mathbb{R}^k_{++} = \{\xi = (m_1, \ldots, m_k) \in \mathbb{R}^k \mid \ m_i > 0\}$$

Se $S = \{p_\xi \mid \xi \in E\} \subset P(\chi)$ é um modelo estatístico, $\tilde{S} = \{\tau p_\xi \mid \tau > 0 \text{ e } \xi \in E\} \subset Dens_+(\chi)$

Modelo estatístico não-normalizado

Métricas Riemannianas

É um produto interno em cada plano tangente: $g_p = \langle \cdot, \cdot \rangle_p : T_pS \times T_pS \to \mathbb{R}$ que varia suavemente com o ponto p.



$\xi = (\xi^1, \ldots, \xi^n)$ coordenadas locais de S

$\partial_i = \frac{\partial}{\partial \xi^i}$, $i = 1, \ldots, n$, campos coordenados

$g_{ij}(p) = \langle \partial_i, \partial_j \rangle_p$ varia suavemente com o ponto p.

$(g_{ij}(p))$ é uma matriz positiva-definida, em cada ponto p.

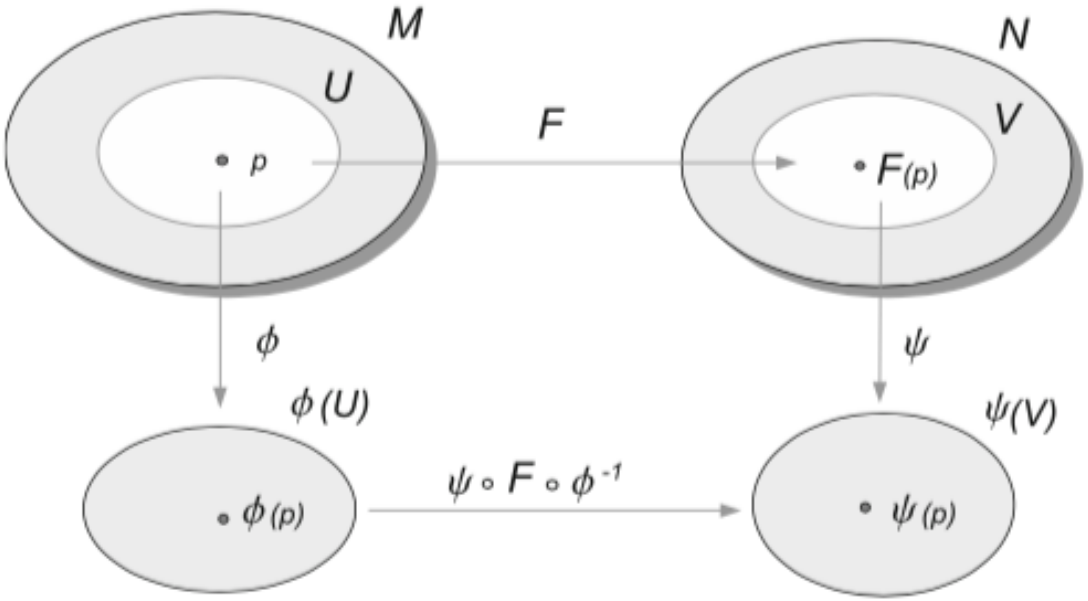
$g = g_{ij}\, d\xi^i d\xi^j$

$u, v \in T_pS$, escrevendo $u = u^i \partial_i$ e $v = v^j \partial_j$

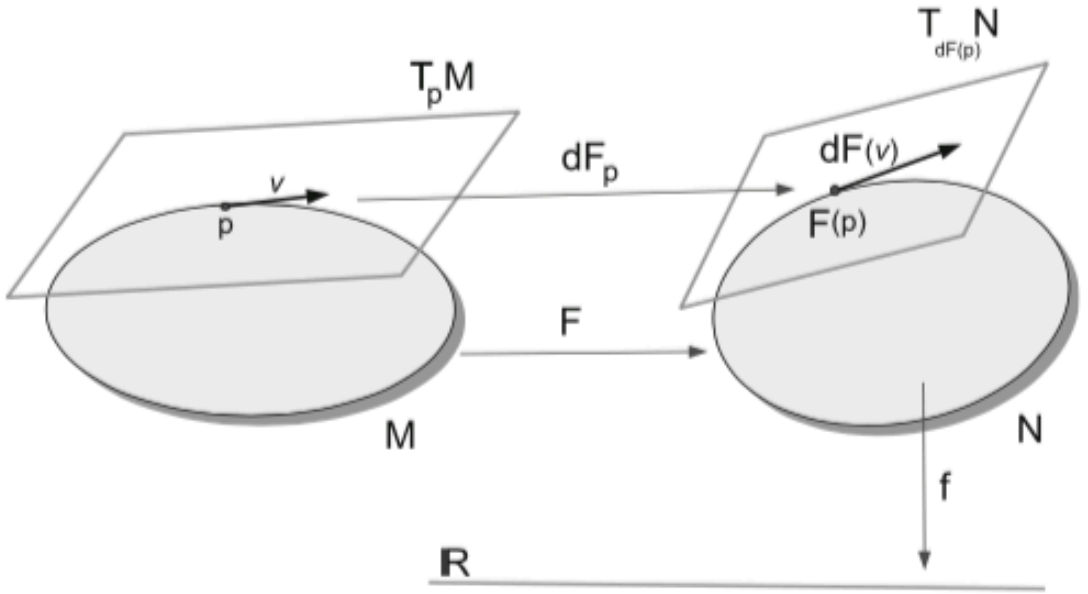
Então, $g(u, v) = \langle u, v \rangle = u^i v^j g_{ij}$

Notação de Einstein: indices consecutivos em cima e em baixo são considerados somados

Aplicações diferenciáveis entre variedades



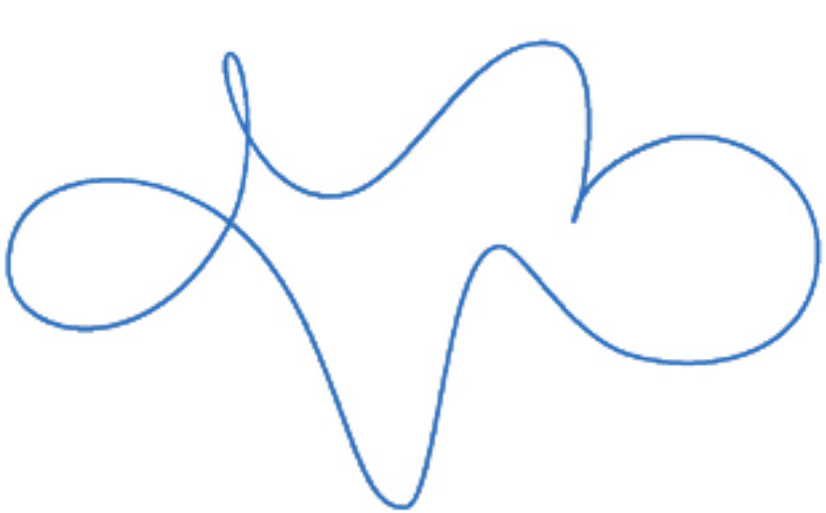
Diferenciabilidade é uma noção local, logo podendo ser definida via cartas locais.

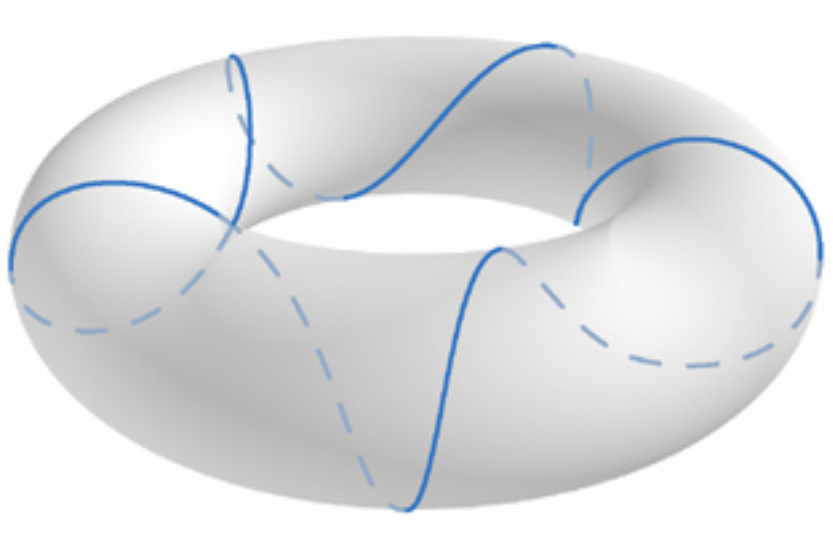


A diferencial é uma aplicação linear entre os planos tangentes, logo não depende das cartas locais

Métricas Riemannianas

É um produto interno em cada plano tangente: $g_p = \langle \cdot, \cdot \rangle_p : T_pS \times T_pS \to \mathbb{R}$ que varia suavemente com o ponto p.





Comp. de uma curva $\alpha : [a, b] \to S$: $|\alpha| = \int_a^b \sqrt{\langle \alpha'(t), \alpha'(t) \rangle}\, dt$

forma elem. de volume de S: $dS = \sqrt{\det g_{ij}}\, d\xi^1 \wedge \ldots \wedge d\xi^n$

independe do sistema de coordenadas escolhido (logo, está globalmente definida)

Volume de um Lebesgue-mensurável: $\mathrm{vol(B)} = \int_{\mathrm{B}} \mathrm{dS}$

Integral de uma função: $\int_S f dS$

Métricas de Informação de Fisher

$S = \{p_\xi(x)\}_{\xi \in E}$ modelo estatístico sobre $\chi = X(\Omega)$

$\xi = (\xi^1, \ldots, \xi^n) \in E$ um sistema de coordenadas

$$g_{ij}(\xi) = E_{p_\xi}[\partial_i(\ln p_\xi)\partial_j(\ln p_\xi)] = \int_\chi \frac{\partial}{\partial \xi^i}(\ln p_\xi)\frac{\partial}{\partial \xi^j}(\ln p_\xi)\, p_\xi(x)\, d\mu(x)$$

(i) é uma matriz positiva-definida em cada ponto $\xi \in E$;

(ii) varia suavemente com ξ

(iii) Se $u = u^i\partial_i, v = v^j\partial_j \in T_pS$ então $g(u,v) = u^i v^j g_{ij} = E_p[\frac{u}{p}\frac{v}{p}]$ não depende do sistema de coordenadas ξ escolhido.

Logo, define uma métrica Riemanniana sobre S.

Priori de Jeffrey: $\pi(\xi) = \frac{dS}{\text{vol(S)}}$.

independe do sist. de coordenadas escolhido

Um pouco da história.

- 1945 – C.R. Rao acreditava ser o primeiro a observar que a matriz de informação de Fisher define uma métrica Riemanniana. No mesmo artigo, ele apresenta o famoso Teorema de Cramer-Rao.

- Recentemente, foi descoberto um artigo de Harold Hotelling (não publicado, mas apresentado no AMS meeting de 1929), mostrando o caráter Riemanniano da matriz de Fisher e provando que a variedade estatística de uma location-scale family possui curvatura constante.

- 1946 – Jeffreys usou a forma elemento de volume desta métrica como uma distribuição a priori sobre o modelo estatístico (logo, deve ser invariante por mudança de coordenadas).

Teorema de Cramer-Rao

$S = \{p_\xi(x)\}_{\xi \in E}$ modelo estatístico sobre $\chi = X(\Omega)$

$$\hat{\xi} = \xi(x_1, \ldots, x_n) : \chi^N \to \mathbb{R}^n \text{ estimador não enviesado de } \xi$$

$$\text{i.e., } E_{p_\xi^N}[\hat{\xi}] = \xi \qquad \text{Então, } V_{p_\xi^N}[\hat{\xi}] \geq \tfrac{1}{N}(g_{ij}(\xi))^{-1}$$

Além disso, vale a igualdade $\iff$ S é uma família exponencial da forma

$$p_\xi(x) = e^{C(x) + \langle \theta, \hat{\xi}(x) \rangle - \psi(\theta)},$$

onde $\hat{\xi}(x) = E_{p_\xi^{N-1}}[\hat{\xi}(x, X_2, \ldots, X_N)]$ e $\xi = \nabla\psi(\theta)$ (coordenadas esperadas).

Neste caso, vale que $g^{ij}(\xi) = g_{ij}(\theta) = Var[\hat{\xi}] = \nabla^2\psi(\theta)$.

Divergência entre dois pontos Sejam P e Q dois pontos numa variedade diferenciável M.

A divergencia $D[P:Q]$ é uma função satisfazendo:

(i) $D[P:Q] \geq 0$;

(ii) $D[P:Q] = 0 \iff P = Q$;

(iii) Se P, Q estão próximos e ξ_P e $\xi_Q = \xi_P + d\xi$ são as coordenadas de P e Q respectivamente, então vale a expansão de Taylor:

$$D[P:Q] = \frac{1}{2} g_{ij}(\xi_P) d\xi^i d\xi^j + O(|d\xi|^3)$$ onde $(g_{ij}(\xi_P))$ é uma matriz positiva-definida.

Em geral

(a) $D[P:Q] \neq D[Q:P]$. Defina $D^*[P:Q] = D[Q:P]$ divergência dual de D;

(b) Em geral, não vale a desigualdade triangular.

Exemplos

Distância Euclideana: $P, Q \in \mathbb{R}^n \quad \rightsquigarrow \quad D[P:Q] = \frac{1}{2}|P-Q|^2 = \frac{1}{2}\delta_{ij}(P^i - Q^i)(Q^j - Q^j)$

Distância de Mahalanobis: $P, Q \in \mathbb{R}^n \quad \rightsquigarrow \quad D[P:Q] = \frac{1}{2}(P-Q)^T A\,(P-Q)$

Divergência de Itakura-Saito: $p, q \in Dens_+(\chi) \quad \rightsquigarrow \quad D[p:q] = \int [\frac{p(x)}{q(x)} - \log(\frac{p(x)}{q(x)}) - 1] d\mu(x)$

Divergência de Kullback-Leibler: $p, q \in P(\chi) \quad \rightsquigarrow \quad KL[p:q] = \int p(x) \log(\frac{p(x)}{q(x)}) d\mu(x)$

I-divergencia: $p, q \in Dens_+(\chi) \quad \rightsquigarrow \quad KL[p:q] = \int p(x) \log(\frac{p(x)}{q(x)}) d\mu(x) - \int (p(x) - q(x)) d\mu(x)$

f-divergencia: $p, q \in P(\chi) \quad \rightsquigarrow \quad D_f[p:q] = \int p(x) f(\frac{q(x)}{p(x)}) d\mu(x)$ ($f : (0, \infty) \to \mathbb{R}$ convexa com $f(1) = 0$.)

Divergência de Kullback-Leibler:

$$p(x), q(x) \in P(\chi) \qquad \rightsquigarrow \qquad KL[p:q] = \int p(x) \log(\frac{p(x)}{q(x)}) d\mu(x)$$

KL não é simétrica Se $p = \mathcal{N}(x \mid 0, 1)$ e $q(x) = \frac{1}{\pi(1+x^2)}$ então $KL[p, q] < \infty$ e $KL[q, p] = \infty$.

Divergencia de Jensen-Shannon: $JS[p:q] = \frac{1}{2}(KL[q:r] + KL[p:r]), \quad$ onde $r = \frac{1}{2}(p+q)$

KL define uma divergência

A) Desigualdade de Gibbs: $KL[p:q] \geq 0$ e vale "$= 0$" $\iff p = q$.

B) Se $p, q \in S = \{p_\xi\}$ é um modelo estatístico (em $P(\chi)$ ou $Dens_+(\chi)$). Escrevendo $\xi_q = \xi_p + d\xi$ então

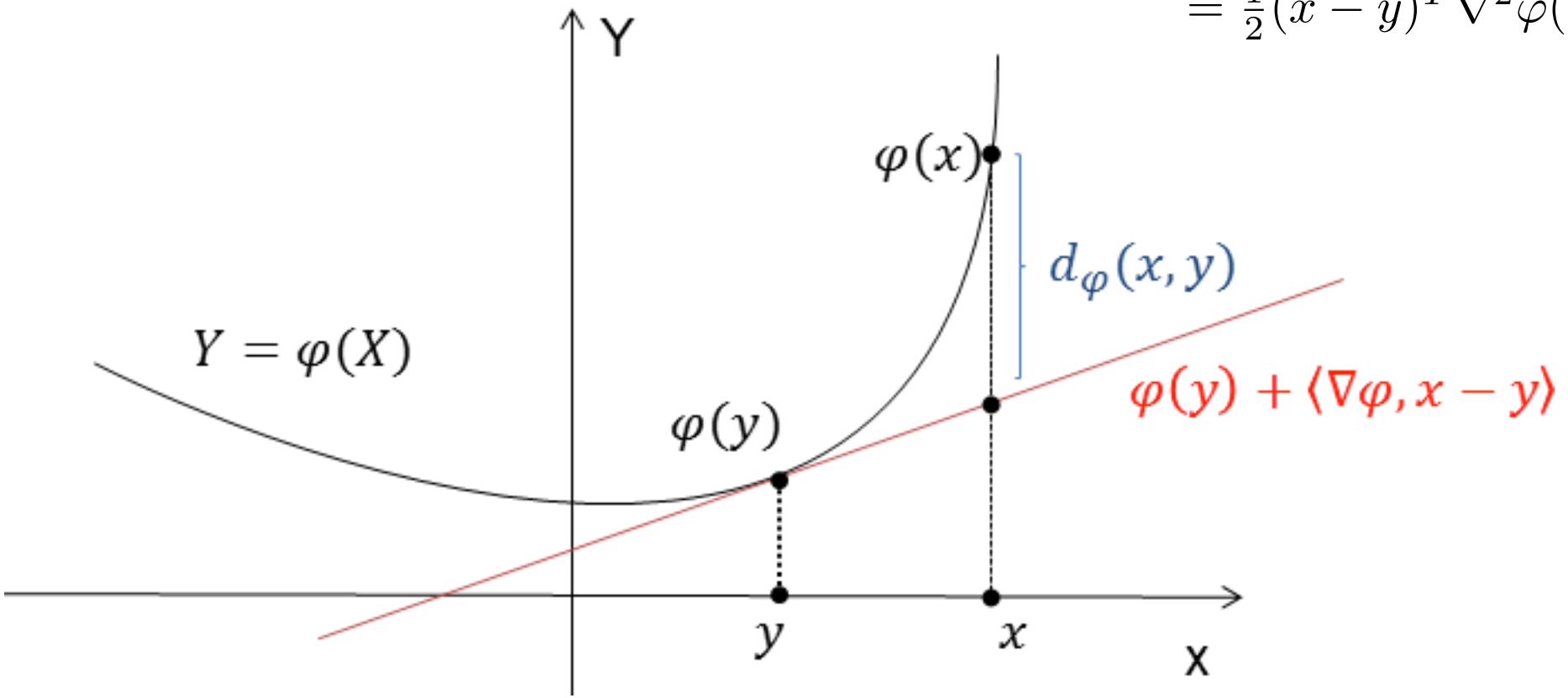
$KL[p:q] = \frac{1}{2} g_{ij}(\xi_p) d\xi^i d\xi^j + O(|d\xi|^3)$, com $g_{ij} = E_p[\partial_i(\ln p_\xi)\partial_j(\ln p_\xi)]$ (Métrica de Fisher)

Divergências de Bregman

Seja $\varphi : \bar{U} \times U \to \mathbb{R}$ estrit. convexa, com $U \subset \mathbb{R}^n$ aberto convexo.

Defina $D_\varphi[x : y] = \varphi(x) - \varphi(y) - \langle \nabla\varphi(y), (x - y)\rangle$

$= \frac{1}{2}(x - y)^T \nabla^2\varphi(y)(x - y) + O(|x - y|^3)$



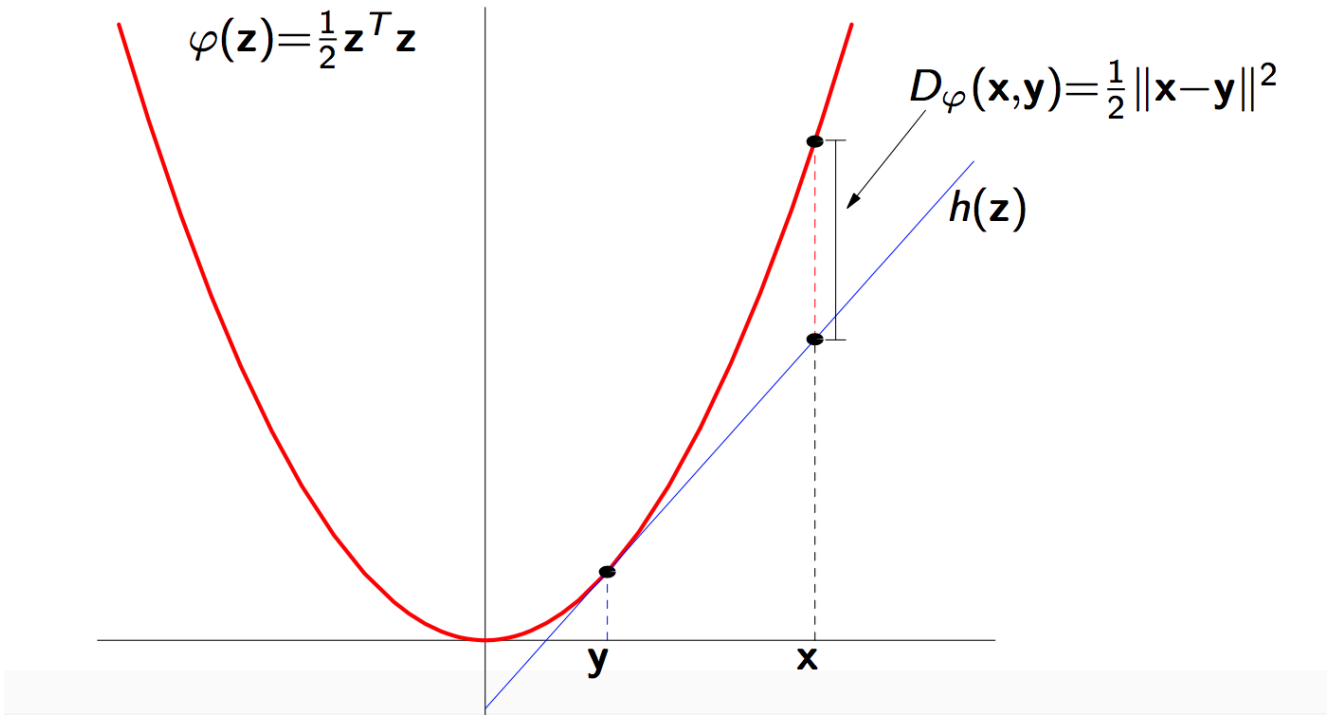
D_φ define uma divergência

Exemplos (Distância Euclideana)

$$\varphi(z) = \frac{1}{2}|z|^2 = \frac{1}{2}\sum_i z_i^2 \quad \rightsquigarrow \quad \nabla\varphi(z) = z \quad \rightsquigarrow \quad D_\varphi[x:y] = \frac{1}{2}[|x|^2 - |y|^2 - 2y^T(x-y)]$$

$$= \frac{1}{2}[|x|^2 + |y|^2 - 2x^T y]$$

$$= \frac{1}{2}|x-y|^2$$



Distância Euclideana $D[x:y] = \frac{1}{2}|x-y|^2$ é um divergência de Bregman

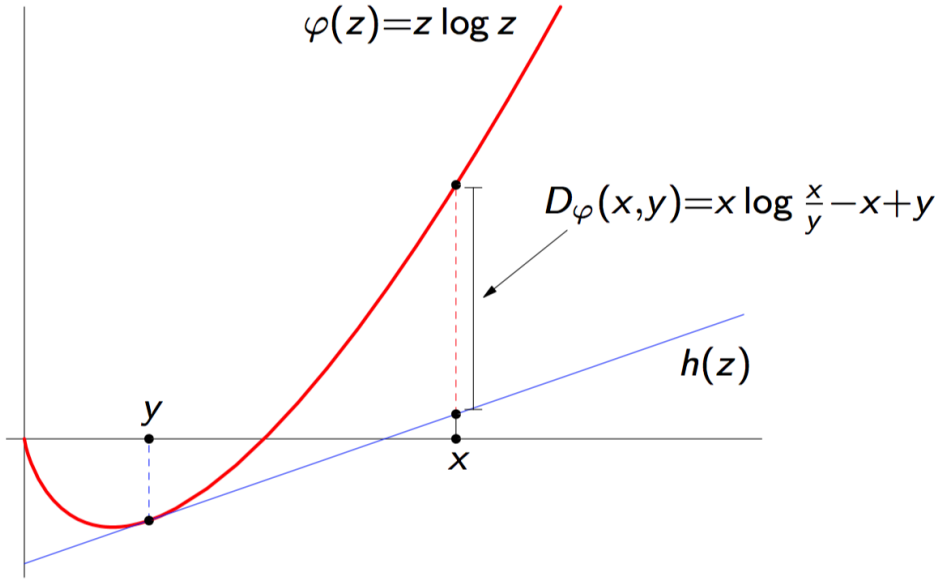
Exemplos (Divergência de Kullback-Leibler)

$$\varphi(z) = \sum z_i \log(z_i) = -H[z]$$

(H = Entropia de Shannon)

$$\varphi_i(z) = \log z_i + 1$$

$$\begin{aligned} D_\varphi[x:y] &= \sum [x^i \log x^i - y^i \log y^i - (\log y_i + 1)(x^i - y^i)] \\ &= \sum [x^i \log x^i - x^i \log y^i - x^i + y^i] \\ &= \sum [x^i \log(\frac{x^i}{y^i}) - (x^i - y^i)] \\ &= KL[x:y] \end{aligned}$$

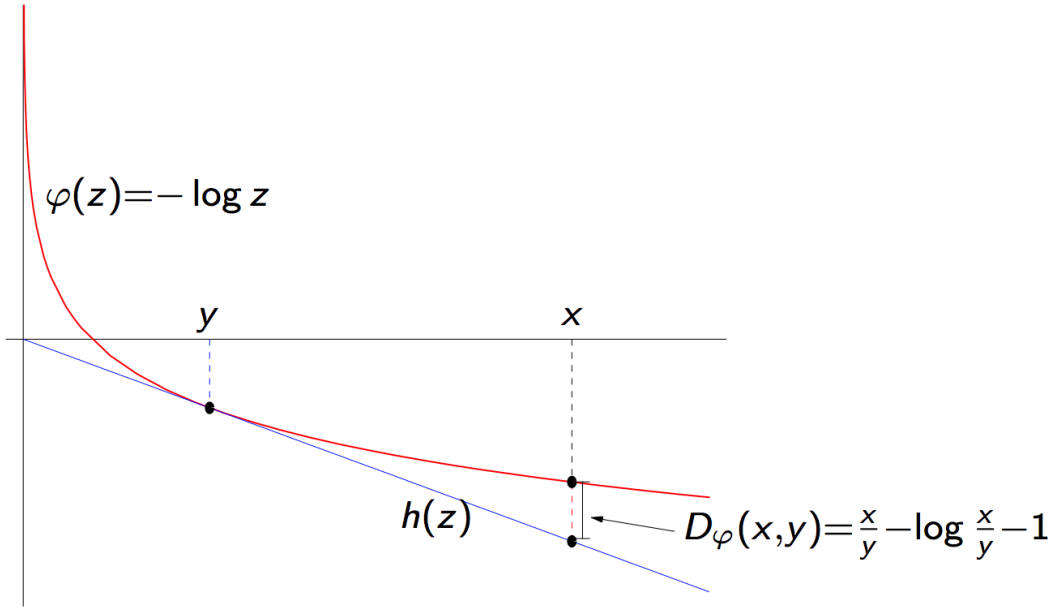


Exemplos (Divergência de Itakura-Saito)

$$\varphi(z) = -\sum \log(z_i) \qquad \varphi_i(z) = -\frac{1}{z_i}$$

(-φ = Entropia de Burg)

$$\begin{aligned} D_\varphi[x:y] &= \sum_i -\log(x_i) + \log(y_i) + \frac{1}{y_i}(x_i - y_i) \\ &= \sum_i \frac{x_i}{y_i} - \log(\frac{x_i}{y_i}) - 1 \\ &= IS[x:y] \end{aligned}$$



Informação de Bregman $I[X] = \min_{\eta} E[D_{\varphi}[X, \eta]] = \min_{\eta} \sum D[x_i : \eta] p(x_i)$

Proposição: $\exists!$ minimizador de I[X] dado por $\eta^* = E[X] = \sum_i x_i p(x_i)$.

Não depende de φ

$$
\begin{aligned}
E[D_{\varphi}[X : \eta]] - E[D_{\varphi}[X : \eta^*]] &= \sum_i (D_{\varphi}[x_i : \eta] - D_{\varphi}[x_i : \eta^*]) p(x_i) \\
&= \varphi(\eta^*) - \varphi(\eta) - \sum_i p(x_i)(\nabla\varphi(\eta)(x_i - \eta)) - \nabla\varphi(\eta^*)(x_i - \eta^*)) \\
&= \varphi(\eta^*) - \varphi(\eta) - (\nabla\varphi(\eta)(\sum_i p(x_i) x_i - \eta)) - \nabla\varphi(\eta^*)(\sum_i p(x_i) x_i - \eta^*)) \\
&= \varphi(\eta^*) - \varphi(\eta) - (\nabla\varphi(\eta)(\eta^* - \eta)) - \nabla\varphi(\eta^*)(\eta^* - \eta^*)) \\
&= D_{\varphi}[\eta^* : \eta]
\end{aligned}
$$

$\rightsquigarrow \quad I[X] = E[D_{\varphi}[X : E[X]]]$

Informação de Bregman

$$I[X] = \min_{\eta} E[D_\varphi[X : \eta]] = E[D_\varphi[X : E[X]]]$$

$$E[X] = \arg\min_{\eta} E[D_\varphi[X : \eta]]$$

Variância é uma informação de Bregman.

$$\varphi(z) = |z|^2 \;\; \rightsquigarrow \;\; D_\varphi[x : y] = |x - y|^2$$

$$\rightsquigarrow \quad I[X] = E[D_\varphi[X : E[X]]] = E[|X - E[X]|^2] = V[X]$$

Informação de Bregman

Informação mutua entre X e Y é uma informação de Bregman.

$$I[X;Y] = KL[p(x,y) : p(x)p(y)] = \sum_{x,y} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})$$

$$= \sum_x p(x) \sum_y p(y|x) \log(\frac{p(y|x)}{p(y)}) = \sum_x p(x) KL[p(y|x) : p(y)]$$

Seja Z_x a v.a. que toma valores $p(y|x) = (p(0|x), \ldots, p(k|x)) \in S_k$ para cada $x \in \chi$, com probab. $p(x)$.

Note que $E[Z_x] = \sum_x p(x) Z_x = \sum_x p(x)p(y|x) = p(y) \in S_k$.

$$= \sum_x p(x) KL[Z_x : E_x[Z_x]] = I[Z_x]$$

Informação de Bregman (associada à divergência de Itakura-Saito)

Seja X a v.a. cujos valores são os vetores $x^{(i)} = (x_1^i, \ldots, x_k^i) \in \mathbb{R}_+^k$,
com $i = 1, \ldots, N$, distribuidos uniformemente

média aritmética $E[X] = \mu = \frac{1}{N} \sum_{i=1}^{N} x^{(i)}$

média geométrica $g = (g_1, \ldots, g_k)$, com $g_j = (\Pi_{i=1}^{N} x_j^{(i)})^{\frac{1}{N}}$.

$$I_\varphi[X] = \sum_i p(x^{(i)}) IS[x^{(i)} : \mu] = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{k} (\frac{x_j^{(i)}}{\mu_j} - \log(\frac{x_j^{(i)}}{\mu_j}) - 1)$$

$$= \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{N} (\frac{x_j^{(i)}}{\mu_j} - \log(\frac{x_j^{(i)}}{\mu_j}) - 1) = \sum_{j=1}^{k} (\frac{\mu_j}{\mu_j} - \frac{1}{N} \sum_{i=1}^{N} \log(\frac{x_j^{(i)}}{\mu_j}) - 1)$$

$$= -\frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{N} \log(\frac{x_j^{(i)}}{\mu_j}) = \sum_{j=1}^{k} (\log(\mu_j) - \frac{1}{N} \sum_{i=1}^{N} \log(x_j^{(i)}) = \sum_{j=1}^{k} \log(\frac{\mu_j}{g_j}).$$

Desigualdade de Jensen e Informação de Bregman

$$E[\varphi(X)] - \varphi(E[X]) = E[D_\varphi[X : E[X]]] = I_\varphi[X]$$

$$\begin{aligned}
E[D_\varphi[X : E[X]]] &= \sum_x p(x) D_\varphi[x : E[X]] \\
&= \sum_x p(x)(\varphi(x) - \varphi(E[X]) - \langle \nabla\varphi(E[X]), x - E[X]\rangle) \\
&= E[\varphi(X)] - \varphi(E[X]) - \langle \nabla\varphi(E[X]), \sum_x p(x)(x - E[X])\rangle) \\
&= E[\varphi(X)] - \varphi(E[X])
\end{aligned}$$

$$E[X] - E[X] = 0$$

Logo, $E[\varphi(X)] \geq \varphi(E[X])$, e vale "$= 0$" $\iff$ X é constante.

Clusterização com Divergências de Bregman

Convex function	Bregman divergence	Algorithm
Squared norm	Squared Loss	KMeans [M'67] MacQueen (1967)
Negative entropy	KL-divergence	Information Theoretic [DMK'03] Dhillon, Mallela and Kumar (2003)
Burg entropy	Itakura-Saito distance	Linde-Buzo-Gray [LBG'80]

Clusterização com Divergências de Bregman

Considere uma partição $\{\chi_h\}_{h=1}^k$ de χ (i.e. subconj disjuntos que cobrem χ)

Em cada cluster, considere $\pi_h = \sum_{x \in \chi_h} p(x)$.

Considere X_h a v.a. que toma valores em χ_h com prob. $p(x|h) = \frac{p(x)}{\pi_h}$

Considere $\mu_h := E[X_h] = \sum_{x \in \chi_h} x\, p(x|h)$ o repres. de cada cluster.

Considere M a v.a. que toma valores em $\mathcal{M} = \{\mu_h\}_{h=1}^k$ com prob. π_h.

Medida de qualidade da clusterização

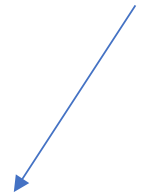
$$E_X[D_\varphi[X : M]] := \sum_{h=1}^{k} \sum_{x \in \chi_h} p(x) D_\varphi[x : \mu_h] = \sum_{h=1}^{k} \pi_h \sum_{x \in \chi_h} p(x|h) D_\varphi[x : \mu_h]$$

$$= \sum_{h=1}^{k} \pi_h I_\varphi[X_h] = E_\pi[I[X_h]] \qquad \text{(Inf. Bregman esperada de um cluster)}$$

Medidas da Qualidade de uma Clusterização

Clusterização ótima: Fixado $\leq k \leq |\chi|$, obter uma partição $\{\chi_h\}_{h=1}^{k}$ de χ tal que $M = \arg\min_{M'} E[D_\varphi[X : M']]$

Considere a seguinte medida de qualidade da clusterização: $L[M] := I_\varphi[X] - I_\varphi[M]$



inf. inter-cluster

Se $k = 1$ então $\pi = \sum_x p(x) = 1$ e o representante $\mu = E[X]$. $\rightsquigarrow$ $I_\varphi[M] = E[X]$

Se $k = |\chi|$ então cada χ_h é unitário $\rightsquigarrow$ $\chi_h = \{x\}$ $\rightsquigarrow$ $D_\varphi[x : \mu_h] = 0$ $\rightsquigarrow$ $I_\varphi[X_h] = 0$

$$I_\varphi[M] = I_\varphi[X]$$

Teorema. $E_\pi[I_\varphi[X_h]] = L_\varphi[M] := I_\varphi[X] - I_\varphi[M]$.

Prova:

$$
\begin{aligned}
I_\varphi[X] &= \sum p(x) D_\varphi[x:\mu] = \sum p(x)(\varphi(x) - \varphi(\mu) - \langle \nabla\varphi(\mu), x-\mu\rangle) \\
&= \sum_h \sum_{x\in\chi_h} p(x)(\varphi(x) - \varphi(\mu_h) - \langle \nabla\varphi(\mu_h), x-\mu_h\rangle) \\
&+ \sum_h \sum_{x\in\chi_h} p(x)(\varphi(\mu_h) - \varphi(\mu) + \langle \nabla\varphi(\mu_h), x-\mu_h\rangle - \langle \nabla\varphi(\mu), x-\mu\rangle) \\
&= \sum_h \sum_{x\in\chi_h} p(x)(D_\varphi[x:\mu_h] + D_\varphi[\mu_h:\mu]) \\
&= E_X[D_\varphi[X:M]] + \sum_h \pi_h D_\varphi[\mu_h:\mu] \\
&= E_\pi[I_\varphi[X_h]] + I_\varphi[M]
\end{aligned}
$$

Algoritmo: Bregman k-means

A) Input: Dados $\chi = \{x_1, \ldots, x_n\} \subset U \subset \mathbb{R}^k$ munido de uma prob. $p(x)$, Divergência de Bregman $D_\varphi[x : y]$, com $x \in \bar{U}$ e $y \in U$, num. de clusters k.

B) Output: Clusterização $\{\chi_h\}_{h=1}^k$ que minimiza loc. $E[D_\varphi[X : M]]$.

Inicio: Escolha $\mu_1, \ldots, \mu_k \in U$ (podendo ser aleatóriamente)

Repita até a convergência:

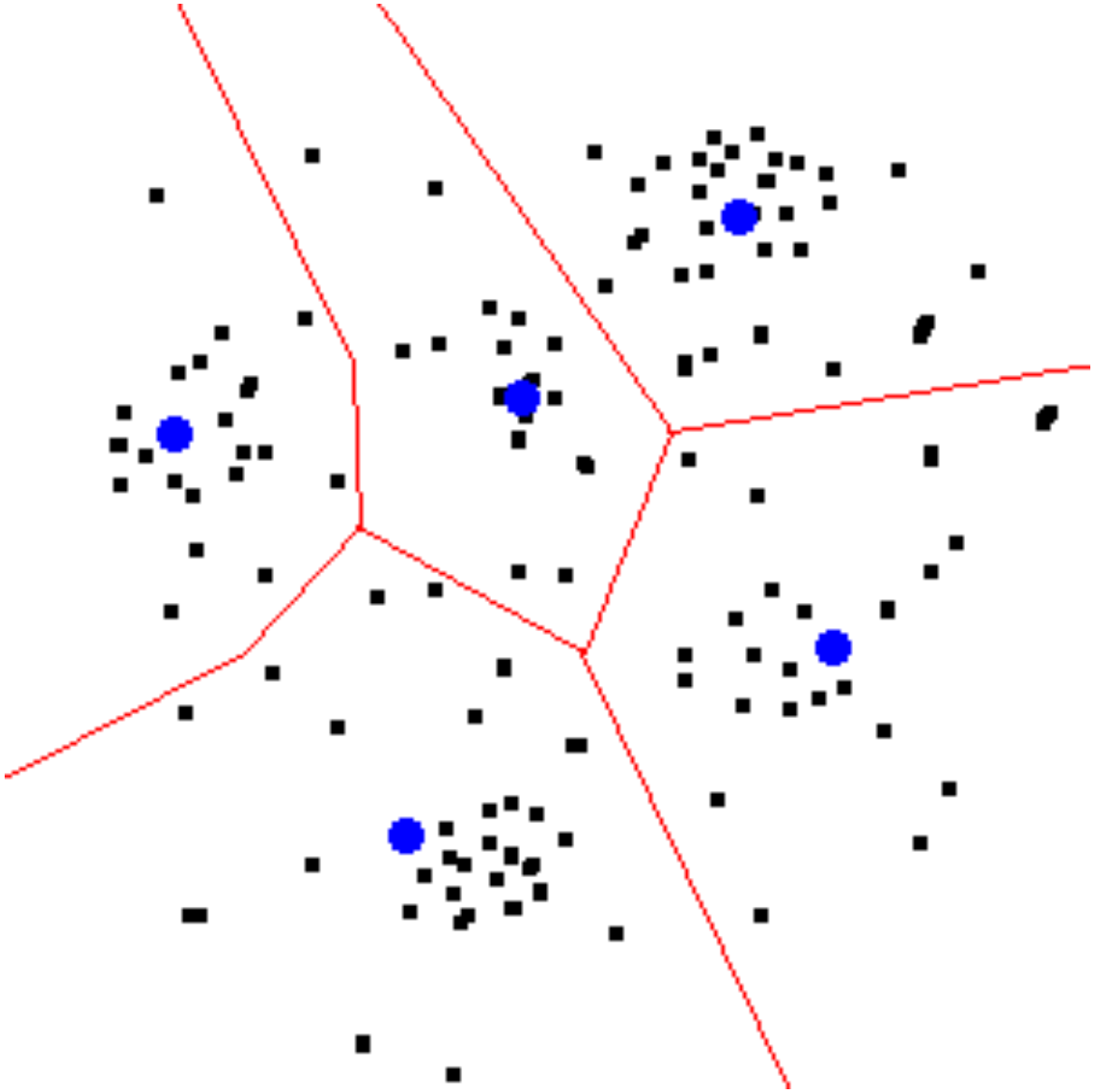
Passo 1: Para cada $x \in \chi$, associe $x \in \chi_h$ com $h = \arg\min_h D_\varphi[x : \mu_h]$.
(Novos clusters foram formados)

(centro mais próximo)

Passo 2: Escolha novos representantes $\mu_h = \sum_{x \in \chi_h} x\, p(x|h)$
(Representantes de cada cluster foram escolhidos)

com $p(x|h) = \frac{p(x)}{\pi_h}$, sendo $\pi_h = \sum_{x \in \chi_h} p(x)$.

Algoritmo: Bregman k-means



Algoritmo: Bregman k-means

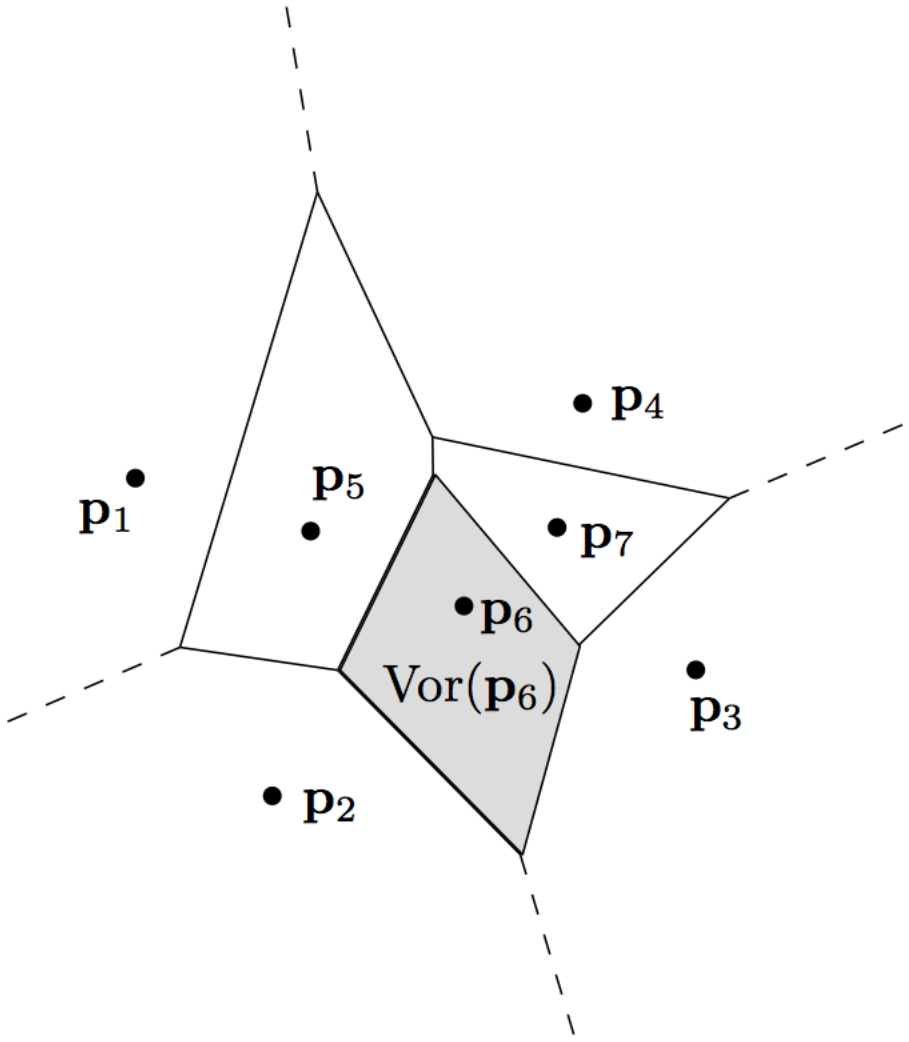
Dados μ_1, μ_2, o conjunto $\{x \mid D_\varphi[x : \mu_1] = D_\varphi[x : \mu_2]\}$ é um hiperplano.

$$
\begin{aligned}
x \in & D_\varphi[x : \mu_1] = D_\varphi[x : \mu_2] \iff \varphi(\mu_1) + \langle \nabla\varphi(\mu_1), x - \mu_1 \rangle = \varphi(\mu_2) + \langle \nabla\varphi(\mu_2), x - \mu_2 \rangle \\
& \iff \langle x, \nabla\varphi(\mu_1) - \nabla\varphi(\mu_2) \rangle = \varphi(\mu_2) - \varphi(\mu_1) + \langle \nabla\varphi(\mu_1), \mu_1 \rangle - \langle \nabla\varphi(\mu_2), \mu_2 \rangle
\end{aligned}
$$

que é um hiperplano.

Assim, as partições geradas pela clusterização são regioes planas, chamadas de diagrama de Voronoi.

Algoritmo: Bregman k-means



https://www.youtube.com/watch?v=cF0hlD_bmdc

https://www.youtube.com/watch?v=KrUFrMOpPEk

Algoritmo: Bregman k-means

Afirmação. $L[M^t]$ decresce monotonicamente a cada iteração (t).

A partir dos representantes $\{\mu_h^{(t)}\}$ os clusteres $\{\chi_h^{(t)}\}$ são formados. Os novos representantes $\{\mu_h^{(t+1)}\}_h$ são dados por $\mu_h^{(t+1)} = \arg\min_\eta \sum_{x \in \chi_h^{(t)}} D_\varphi[x : \eta]$.

$$L_\varphi[M^{(t)}] = E_X[D_\varphi[X : M^{(t)}]] = \sum_h \sum_{x \in \chi_h^{(t)}} p(x) D_\varphi[x : \mu_h^{(t)}]$$

$$\overset{(a)}{\geq} \sum_h \sum_{x \in \chi_h^{(t)}} p(x) D_\varphi[x : \mu_h^{(t+1)}] \overset{(b)}{\geq} \sum_h \sum_{x \in \chi_h^{(t+1)}} p(x) D_\varphi[x : \mu_h^{(t+1)}]$$

(a) ocorre pois $\mu_h^{(t+1)} = \arg\min_\eta \sum_{x \in \chi_h^{(t)}} p(x) D_\varphi[x : \eta]$.

(b) ocorre pois $x \in \chi_h^{(t)}$ não pertencerá a $\chi_h^{(t+1)}$ se x estiver mais próximo do centro $\mu_{h'}^{(t+1)}$ de outro cluster $\chi_{h'}$, logo, na soma de todos os cluster, diminuirá.

Algoritmo: Bregman k-means

Como há apenas uma quantidade finita de possíveis clusterizações $\{\chi_h\}_{h=1}^{k}$ e $L(M^{(t)})$ é monótona não-crescente, o algoritmo deve convergir num **numero finito** de iterações.

A complexidade em cada iteração é Nk, onde $N = |\chi|$. Isso porque, a todo elemento x procura o centro $\{\mu_h\}_{h=1}^{k}$ mais próximo.

No caso de dados mistos, pode-se usar uma divergência de Bregman proveniente de uma combinação linear de funções convexas.

Transformada de Legendre

Seja $\varphi : U \to \mathbb{R}$ função estritamente convexa, definida num aberto convexo $U \subset \mathbb{R}^n$. Considere φ a função $\phi(\eta) = \sup_{\theta' \in U}(\langle \theta', \eta \rangle - \varphi(\theta'))$. Então, valem os seguintes itens abaixo:

(i) Se $\eta = \nabla\varphi(\theta)$ então $\theta = \nabla\phi(\eta)$.

(ii) $\phi(\eta) = \langle \eta, \theta \rangle - \psi(\theta)$, com $\theta = \nabla\phi(\eta)$.

Consequências:

(a) θ e $\eta = \nabla\varphi(\theta)$ definem mudanças de coordenadas;

(b) $\nabla^2\phi(\eta) = (\nabla\varphi(\theta))^{-1}$;

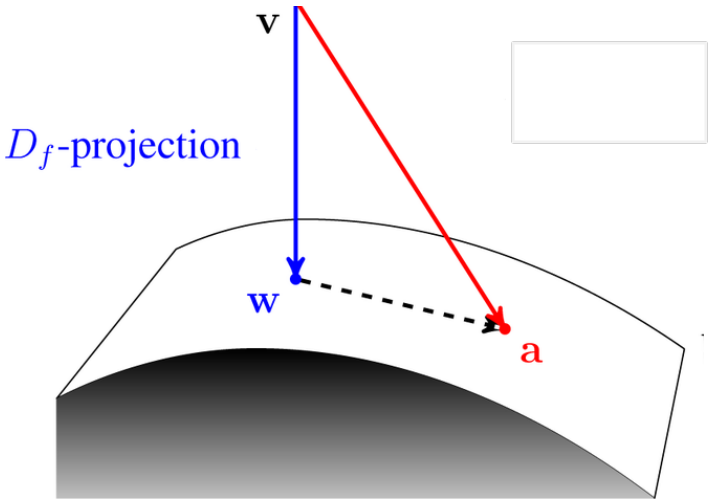
Transformada de Legendre

Consequências: Seja $S = \{p_\theta\}$ um modelo estatístico (normalizado ou não), parametrizado por $\theta \in E$. Se $p, q \in S$ então

(a) $D_\varphi[\theta_p, \theta_q] = D_\phi[\eta_q, \eta_p]$ (Divergencia Dual)

(b) $D_\varphi[\theta_p, \theta_q] = \varphi(\theta_p) + \phi(\eta_q) - \langle \theta_p, \eta_q \rangle$ (Desigualdade de Frenchel)

(c) $D_\varphi[\theta_p : \theta_r] = D_\varphi[\theta_p : \theta_r] + D_\varphi[\theta_r : \theta_q] - \langle \theta_p - \theta_r, \eta_q - \eta_r \rangle$. (Teo. Pitagoreano)



φ-projeção. Seja $r = \arg\min_{s \in C} D_\varphi[p, s]$. Seja $C \subset S$ um subconjunto η-convexo então para todo $q \in C$, vale

$$D_\varphi[p : q] \geq D_\varphi[p : r] + D_\varphi[r : q].$$

Se C é η-afim então vale a igualdade.

Divergencia de Bregman e Famílias exponenciais

Considere $p_\theta(x) = e^{\langle \theta, x\rangle - \psi(\theta)}$, $\theta \in E \subset \mathbb{R}^k$ uma familia exponencial natural.

Temos o seguinte:

(i) $\frac{\partial}{\partial \theta^i} \log(p_\theta(x)) = x_i - \psi_i(\theta) \;\rightsquigarrow\; \nabla\psi(\theta) = E_{p_\theta}[x]$;

(ii) $\frac{\partial^2}{\partial \theta^i \partial \theta^j} \log(p_\theta(x)) = -\psi_{ij}(\theta) \;\rightsquigarrow\; \nabla^2\psi(\theta) = -E_{p_\theta}[\frac{\partial^2}{\partial \theta^i \partial \theta^j} \log(p_\theta(x))]$ (Matriz de Inf. de Fisher)

$\psi(\theta)$ é convexa $\rightsquigarrow$ (conj. de Legendre) $\phi(\eta) = \langle \eta, \theta\rangle - \psi(\theta)$, com $\eta = \nabla\psi(\theta) = E_{p_\theta}[x]$.

$$= E[\langle x, \theta\rangle - \psi(\theta)] = E[\log(p_\theta)]$$

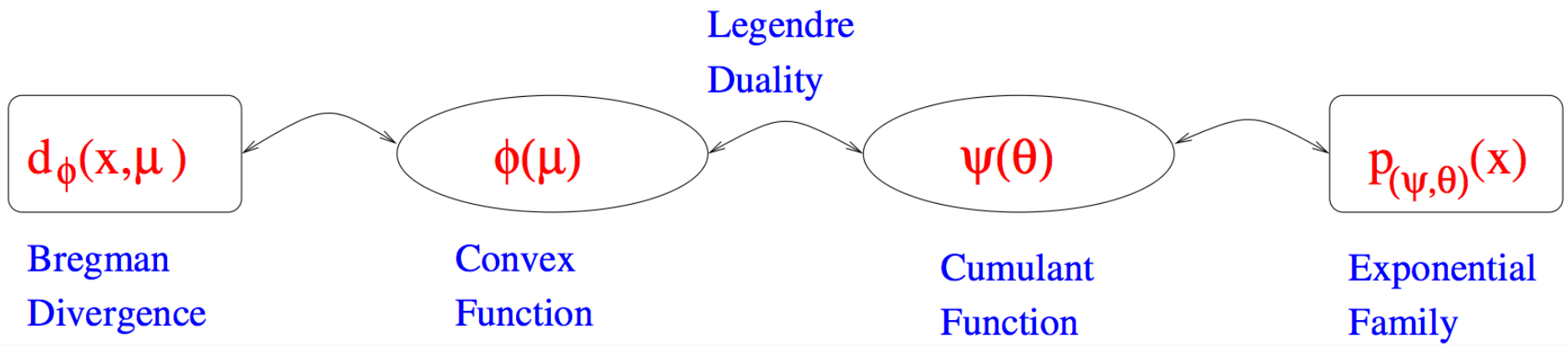
$\rightsquigarrow$ $D_\phi[x : \eta] = \phi(x) + \psi(\theta) - \langle x, \theta\rangle$ $\rightsquigarrow$ $p_\theta(x) = e^{-D_\phi[x:\eta] + \phi(x)}$

(**Teo. (A.Banerjee, 2003)** Existe uma bijeção entre FE e Div. Bregman)

Theorem: For any regular exponential family $p_{(\psi,\boldsymbol{\theta})}$, for all $\mathbf{x} \in \mathrm{dom}(\phi)$,

$$p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) = \exp(-d_\phi(\mathbf{x}, \boldsymbol{\mu}))b_\phi(\mathbf{x}),$$

for a uniquely determined b_ϕ, where $\boldsymbol{\theta}$ is the natural parameter and μ is the expectation parameter



(**Teo. (A.Banerjee, 2003)** Existe uma bijeção entre FE e Div. Bregman)

$$p_\theta(x) = e^{-D_\phi[x:\eta]+\phi(x)}$$

Regular exponential families $\leftrightarrow$ Regular Bregman divergences

Gaussian	$\leftrightarrow$	Squared Loss
Multinomial	$\leftrightarrow$	KL-divergence
Geometric	$\leftrightarrow$	Itakura-Saito distance
Poisson	$\leftrightarrow$	I-divergence

(**Teo. (A.Banerjee, 2003)** Existe uma bijeção entre FE e Div. Bregman)

$$p_\theta(x) = e^{-D_\phi[x:\eta]+\phi(x)}$$

Soft Clustering

- Data modeling with mixture of exponential family distributions
- Solved using Expectation Maximization (EM) algorithm

Maximum log-likelihood $\equiv$ Minimum Bregman divergence

$$\log p_{(\psi,\boldsymbol{\theta})}(\mathbf{x}) \equiv -d_\phi(\mathbf{x},\mu)$$

Bijection implies a Bregman divergence viewpoint

- Efficient algorithm for soft clustering

Bregman Soft k-means Algorithm:

Início: Tome pesos e representantes $\{\pi_h, \mu_h\}$

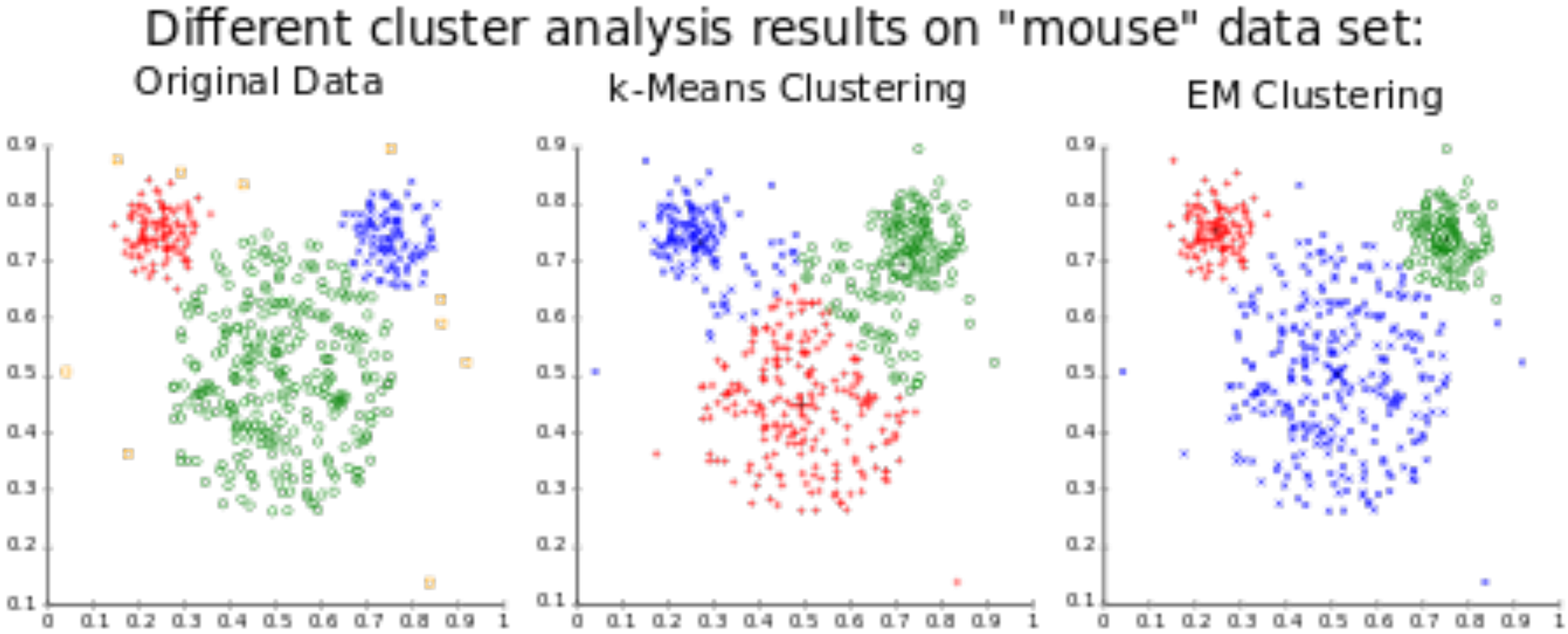
Repita até a convergência:

E-passo: Para todo x e h considere: $p(h|x) = \pi_h \dfrac{e^{-D_\phi[x:\mu_h]}}{Z(x)}$ (constante normalizadora)

M-passo Para todo h:

$$\pi_h = \frac{1}{n}\sum_x p(h|x) \qquad \mu_h = \frac{\sum_x p(h|x)\, x}{\sum_x p(h|x)}$$

Bregman Soft k-means Algorithm:



Bregman Soft k-means Algorithm:

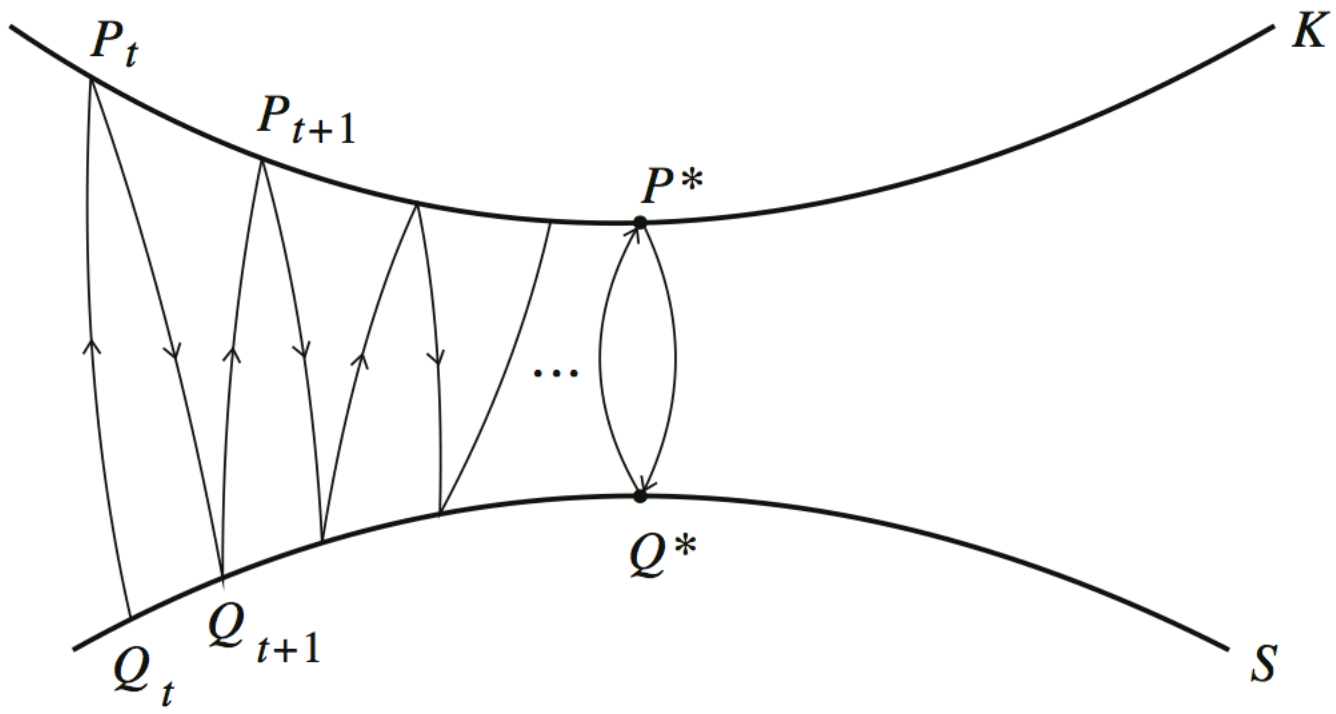


Fig. 1.9 Iterated dual geodesic projections (*em* algorithm)