Universidade Federal do Rio de Janeiro
Instituto de Matemática

Programa de Pós-Graduação em Matemática

Geometria da Informação: Teorema Pitagoreano e Aplicações

Gil dos Santos Navarro

Dissertação de Mestrado

UFRJ
Rio de Janeiro - 2017

Universidade Federal do Rio de Janeiro
Instituto de Matemática

Gil dos Santos Navarro

Geometria da Informação: Teorema Pitagoreano e Aplicações

Trabalho apresentado ao Programa de Programa de Pós-Graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio de Janeiro como requisito parcial para obtenção do grau de Mestre em Matemática.

Orientador: *Prof. Heudson Mirandola*

UFRJ
Rio de Janeiro - 2017

Geometria da Informação: Teorema Pitagoreano e Aplicações

Gil dos Santos Navarro

Orientador: Heudson Tosta Mirandola

Dissertação de Mestrado submetida ao Programa de Pós-Graduação do Instituto de Matemática da Universidade Federal do Rio de Janeiro - UFRJ como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

Aprovada por:

(Presidente) Prof. Heudson Tosta Mirandola - IM/UFRJ

Prof. Fábio Antonio Tavares Ramos - IM/UFRJ

Prof. César Javier Niche Mazzeo - IM/UFRJ

Prof. Hugo Tremonte de Carvalho - DME/UFRJ

Agradecimentos

Agradeço a todos que me ajudaram de alguma forma na realização deste trabalho. Em especial meus pais, Sandra e Roberto, por toda a educação e carinho que me deram. À Adriene, por todo o apoio e companheirismo. Aos meus irmãos Leon e Nino, todos os amigos e família, simplesmente por existirem. Em especial ao Wellington, por todas as horas de trabalho. Agradeço ao meu orientador Heudson pela ajuda, atenção e por todas discussões produtivas que me ajudaram a concluir essa dissertação.

Resumo

Um dos objetivos de geometria da informação é revisitar métodos consagrados na estatística com ferramentas de geometria diferencial. Neste trabalho apresentamos a linguagem básica da Geometria da Informação, como a matriz de informação de Fisher, as α-conexões e as funções de divergência. Em seguida, apresentamos um análogo ao Teorema de Pitágoras para as divergências de Bregman e concluímos com algumas aplicações deste teorema.

Palavras-chave: Geometria da Informação, Divergência de Bregman, Matriz de Informação de Fisher, Conexões Duais, α-conexões, Teorema Pitagoreano, Teorema da Projeção

Abstract

Information Geometry aims to see statistics with a geometric point of view. The main goal is to reach a deep understanding of how well and efficient some of the statistics tools really are. In this work, we present some basic definitions such as the Fisher Information Matrix, α - connections and divergence functions. Next, we present the Pythagorean Theorem for Bregman divergences and will conclude with some applications of this theorem.

Keywords: Information Geometry, Bregman Divergence, Fisher Information matrix, Dual Connections, α-Connections, Pythagorean Theorem, Projection Theorem

Sumário

CAPÍTULO 1

Preliminares de Probabilidade e Estatística.

1.1 Espaços de probabilidade e variáveis aleatórias

Definição 1. *Seja Ω um conjunto qualquer e Σ uma família de subconjuntos de Ω. Chamamos Σ de σ-algebra quando satisfaz as seguntes condições:*

(i) $\emptyset$ *e* Ω *pertencem a* Σ.

(ii) Se A pertence a Σ, então o complementar A^c pertence a Σ.

(iii) Se $(A_n)_{n\in\mathbb{N}}$ é uma sequência enumerável de conjuntos em Σ, então a união $\cup_{n=1}^{\infty} A_n$ pertence a Σ.

O par (Ω,Σ) é chamado de espaço mensurável. Um elemento de Σ é chamado de conjunto Σ-mensurável.

Definição 2. *Uma função $P : \Sigma \to \mathbb{R}\cup\{\pm\infty\}$ é chamada de medida de probabilidade sobre Σ quando satisfaz as seguintes propriedades:*

(i) $P(A) \geq 0$ *para todo* $A \in \Sigma$.

(ii) $P(\Omega) = 1$.

(iii) Se (A_n) é uma sequência de conjuntos disjuntos em Σ, então $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

A tupla (Ω,Σ,P) é chamada de espaço de probabilidade. Neste caso, um conjunto $E \in \Sigma$ é chamado de evento de Ω.

Definição 3. *Seja $B \in \Sigma$ um evento tal que $P(B) > 0$. Definimos a probabilidade condicional do evento A dado B por*

$$P(A \mid B) = \frac{P(A\cap B)}{P(B)}$$

Note $P(\cdot \mid B)$ define uma nova medida de probabilidade em Ω, e em B (com a σ-algebra relativa). Dizemos que A e B são eventos independentes se $P(A\cap B) = P(A)P(B)$. Neste caso, segue-se diretamente que $P(A|B) = P(A)$ e $P(B|A) = P(B)$.

Proposição 1.1 (Teorema da Probabilidade Total)**.** *Se $\{B_n\}_n$ é uma partição de Ω (i.e., uma família disjunta de eventos que cobrem Ω), então vale*

$$P(A) = \sum_n P(A \mid B_n)P(B_n),$$

para qualquer evento A.

Demonstração.

$$P(A) = P(A \cap (\cup_n B_n)) = P(\cup_n (A \cap B_n)) = \sum_n P(A \cap B_n) = \sum_n P(A \mid B_n) P(B_n).$$

Teorema 1.2. *(Teorema de Bayes) Se A e B são dois eventos com $P(A) > 0$, então vale que*

$$P(B \mid A) = \frac{P(A \mid B) P(B)}{P(A)}.$$

Definição 4. *Uma variável aleatória é uma função $X : \Omega \to \mathscr{X} = X(\Omega) \subset \mathbb{R}^n$ mensurável, ou seja, para todo Boreliano $B \subset \mathbb{R}^n$, tem-se $X^{-1}(B) \in \Sigma$. Assim, se $X : \Omega \to \mathscr{X} = X(\Omega) \subset \mathbb{R}^n$ é uma variável aleatória então cada uma de suas componentes X^j de $X = (X^1, \ldots, X^n)$ é também uma variável aleatória.*

Uma variável aleatória $X = (X_1, \ldots, X_n)$ induz de maneira natural uma medida de probabilidade na σ-álgebra induzida pela σ-algebra de Borel. Basta definir $\mu(B) = P[X \in B]$, onde a notação $[X \in B]$ denota o conjunto $X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$. Definimos a *função de probabilidade acumulada* de X por $F(x_1, \ldots, x_n) = P[X_1 \leq x_1, \ldots, X_n \leq x_n]$. No caso de um vetor aleatório $X = (X_1, \ldots, X_n)$, com $n \geq 2$, a medida μ é chamada de medida de probabilidade conjunta de $X_1, \ldots, X_n$. Sobre uma função mensurável $f : \mathscr{X} \to \mathbb{R}^k$ define-se a integral $\int f d\mu$ de f com respeito à medida de probabilidade μ.

Abaixo, definimos os dois principais tipos de variáveis aleatórias, a saber as variáveis aleatórias discretas e contínuas.

Definição 5. *Uma variável aleatória é dita discreta quando a imagem $\mathscr{X}$ de X é um conjunto enumerável. A função $P(x) = P[X = x]$ é chamada de função massa de probabilidade de X. A integral de uma função mensurável $f : \mathscr{X} \to \mathbb{R}^k$ é dado por $\int f d\mu = \sum_{x \in \mathscr{X}} f(x)$.*

Definição 6. *Uma variável aleatória X é dita contínua quando existe uma função não-negativa $p : \mathbb{R} \to [0, +\infty)$ de modo que $\int_{-\infty}^{\infty} p(x)dx = 1$ e, para todo $a < b$ tem-se,*

$$P[X \in (a,b)] = \int_a^b p(x)dx.$$

A função p é chamada de função densidade de probabilidade de X. Na verdade, pelo Teorema de Radon-Nikodym, isto é equivalente a dizer que a medida $\mu(B) = P[X \in B]$ é absolutamente contínua em relação à medida de Lesbegue. Convém observar que a densidade de probabilidade p não é uma probabilidade; em geral $p(x)$ pode assumir valores maiores do que 1. A integral de uma função mensurável $f : \mathscr{X} \to \mathbb{R}^k$ é dada por $\int f d\mu = \int f(x)p(x)dx$.

Vejamos também alguns modelos simples de distribuições de probabilidades.

Exemplo 1. *Fixado um parâmetro $\lambda \in (0,1)$, dizemos que uma variável aleatória X possui distribuição de Bernoulli se a imagem $\mathscr{X} = \{0,1\}$, e*

$$P(x) = \lambda^x (1-\lambda)^{1-x}, \quad x \in \{0,1\}.$$

Escrevemos $X \sim Ber(\lambda)$.

Exemplo 2. *Dizemos que uma variável aleatória discreta X com $\mathscr{X} = \{0,1,\ldots,\}$ tem distribuição de Poisson com parâmetro $\lambda > 0$ quando a função de probabilidade de X é dada por*

$$P(x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

Escrevemos $X \sim Pois(\lambda)$.

Exemplo 3. *Dizemos que uma variável aleatória $T > 0$ tem distribuição exponencial com parâmetro $\beta > 0$ e escrevemos $T \sim \exp(\beta)$ se sua função densidade de probabilidade é dada por*

$$p(t;\beta) = \frac{1}{\beta}e^{-t/\beta}$$

Claramente, $f(t;\beta) > 0$, para todo $t > 0$ e é imediato verificar que $\int_0^\infty p(t;\beta)dt = 1$.

.

Exemplo 4. *Sejam $\alpha,\beta > 0$. Dizemos que uma variável aleatória $X \geq 0$ tem distribuição gama com parâmetros α e β se a densidade de probabilidade é dada por*

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

Onde $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dx$.

Exemplo 5. *Sejam $\alpha,\beta > 0$. Dizemos que uma variável aleatória $0 < X < 1$ tem distribuição beta com parâmetros α e β se a densidade de probabilidade é dada por*

$$p(x \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

Exemplo 6. *Dizemos que uma variável aleatória X tem distribuição normal univariada com parâmetros $\mu \in (-\infty,\infty)$ e $\sigma^2 > 0$ quando sua distribuição de probabilidade é dada por*

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ para todo } -\infty < x < \infty.$$

Escrevemos $X \sim N(\mu,\sigma)$.

Exemplo 7. *Seja $\mu \in \mathbb{R}^n$ e Σ uma matriz simétrica positiva-definida. Dizemos que um vetor aleatório $X = (X^1,\ldots,X^n)$ tem distribuição normal multivariada com parâmetros μ e Σ, se a densidade de probabilidade é dada por*

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right\}$$

Chamamos μ de vetor de médias e Σ de matriz de covariâncias. Escrevemos $X \sim N_n(\mu,\Sigma)$.

1.2 Distribuições multivariadas

Nesta seção, falaremos de conceitos importantes em distribuições multivariadas, tais como independência, distribuições marginais, distribuições condicionais e amostras i.i.d. Começaremos considerando um vetor aleatório $X = (X^1, \ldots, X^n) : \Omega \to \mathscr{X} = X(\Omega) \subset \mathbb{R}^n$ e uma distribuição de probabilidade conjunta $f(x^1, \ldots, x^n)$ associada à $X = (X^1, \ldots, X^n)$.

Definição 7. *Dizemos que dois vetores aleatórios $X : \Omega \to \mathscr{X} \subset \mathbb{R}^n$ e $Y : \Omega \to \mathscr{Y} \subset \mathbb{R}^m$ são independentes se, para quaisquer conjuntos Borelianos $B_1 \subset \mathbb{R}^n$ e $B_2 \subset \mathbb{R}^m$, vale o seguinte:*

$$P\Big[[X \in B_1] \cap [Y \in B_2]\Big] = P[X \in B_1, Y \in B_2] = P[X \in B_1]P[Y \in B_2].$$

Note que se as componentes de um vetor aleatório $X = (X^1, \ldots, X^n)$ são variáveis aleatórias independentes então a distribuição conjunta acumulada satisfaz:

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n] = \prod_{i=1}^{n} P[X_i \leq x_i],$$

para quaisquer números reais $x_1, \ldots, x_n$. Sabemos que se os eventos $E_1, \ldots, E_n \in \Sigma$ são independentes então

$$P\Big[\bigcap_{i=1}^{n} E_i\Big] = \prod_{i=1}^{n} P[E_i]. \tag{1.1}$$

Em termos de variáveis aleatórias, a equação (1.1) acima é equivalente a dizer que as funções indicadoras $X_i = \mathbb{I}_{E_i}$ são variáveis aleatórias independentes. Aqui, dado $E \in \Sigma$, definimos a função indicadora $\mathbb{I}_E(w) = 1$ se $w \in E$ e 0 caso contrário.

Definição 8. *Seja $f(x,y)$ a função de probabilidade conjunta de (X,Y). Definimos a função de probabilidade marginal de X por*

$$f(x) = \int f(x,y)dy. \tag{1.2}$$

No caso em que Y é uma variável aleatória discreta, a integral se reduz à soma: $f(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f(x,y)$. No caso em que a distribuição marginal $f(y) > 0$, define-se a probabilidade condicional de X dado Y por

$$f(x \mid y) = \frac{f(x,y)}{f(y)}. \tag{1.3}$$

Das definições (1.2) e (1.3) acima, segue-se a seguinte versão do Teorema da Probabilidade Total:

$$f(x) = \int f(x \mid y) f(y) dy.$$

Além disso, se $X_1, \ldots, X_n$ são variáveis aleatórias independentes então a distribuição conjunta é o produto das distribuições marginais, ou seja, $f(x_1, \ldots, x_n) = f(x_1) \ldots f(x_n)$. Alé disso, o Teorema de Bayes é dada por: $f(x \mid y) = f(y \mid x) f(x)/f(y)$, onde $f(x)$ e $f(y)$ são as distribuições marginais.

Definição 9. *Diremos que as variáveis aleatórias $X_1,\ldots,X_n$ são independentes e igualmente distribuidas (i.i.d.) se elas forem independentes e as suas respectivas distribuições marginais f_{X_i} definirem uma mesma função f. Também diremos que uma amostra $\{x_1,\ldots,x_n\}$ é i.i.d. se o vetor $(x_1,\ldots,x_n)$ é gerado a partir de um vetor aleatório $X=(X_1,\ldots,X_n)$ cujas componentes são variáveis aleatórias i.i.d. Na prática, obtém-se amostras i.i.d. a partir de visualizações independentes de uma mesma variável aleatória, ou seja, por sucessivas realiações independentes de um mesmo experimento.*

1.3 Esperança e variância de uma variável aleatória

Definição 10. *Seja $X:\Omega\to\mathscr{X}=X(\Omega)\subset\mathbb{R}^n$ um vetor aleatório. Definimos a esperança de X pela integral*

$$E[X]=\int X(w)\,dP(w). \tag{1.4}$$

Como exemplo simples, consideremos $\Omega=\{HH,HT,TH,TT\}$ como sendo o resultado de dois lançamentos independentes de moedas fiéis (ou seja, os resultados H ('Cara') e T ('Coroa') são equiprováveis). Como os lançamentos são independentes, segue-se que a probabilidade de qualquer resultado de Ω é 1/4. Considere a variável aleatória, $X(w)=$ numero de 'Caras' em w. Segue que

$$\begin{aligned} E[X] &= X(HH)P(HH)+X(HT)P(HT)+X(TH)P(TH)+X(TT)P(TT)\\ &= \frac{1}{4}(2+1+1+0)=1. \end{aligned}$$

Se X é uma variável aleatória discreta, então como a imagem $\mathscr{X}=X(\Omega)$ é enumerável e

$$E[X]=\int X(w)dP(w)=\sum_x\sum_{w\in[X=x]}X(w)P(w)=\sum_x xP([X=x]). \tag{1.5}$$

Em geral, vale o seguinte

$$E[X]=\int x d\mu(x), \tag{1.6}$$

onde μ denota a distribuição de probabilidade sobre $\mathscr{X}$ induzida por X. Se $X\geq 0$ é uma variável aleatória (unidimensional) não-negativa então segue-se que X é limite de uma sequência monótona não-decrescente de variáveis aleatórias discretas não-negativas $X_n\geq 0$ (isso vem do fato de que X, sendo uma função mensurárel não-negativa, é limite de uma sequência monótona não-decrescente de funções simples). Como $E[X_n]=\int x d\mu_n(x)=\sum xP[X_n=x]$, segue-se do teorema da convergência monótona que $E[X]=\lim E[X_n]=\int x d\mu(x)$. O caso geral $X:\Omega\to\mathscr{X}\subset\mathbb{R}^n$ segue-se decompondo cada componente X^j de $X=(X^1,\ldots,X^n)$ em suas partes positiva $(X^j)^+=\max\{0,X^j\}$ e negativa $(X^j)^-=\max\{0,-X^j\}$.

Propriedades importantes da esperança. Os seguintes itens seguem diretamente das equações (1.4) e (1.3).

(i) Se $X = c$ (isto é, $X(\omega) = c$ para todo $\omega \in \Omega$), então $E[X] = c$.

(ii) Se $X_1, \ldots, X_n$ são variáveis aleatórias e $a_1, \ldots, a_n$ são constantes, então

$$E[\sum_i a_i X_i] = \sum_i a_i E[X_i].$$

(iii) Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes, então

$$E[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} E[X_i].$$

Exemplo 8. *Se $X \sim Ber(\lambda)$, então $E[X] = 0P[X = 0] + 1P[X = 1] = \lambda$.*

Exemplo 9. *Se $X \sim Poi(\lambda)$, então*

$$E[X] = \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} = \lambda e^{-} \lambda \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda.$$

Exemplo 10. *se $X \sim N(\mu, \sigma^2)$, então*

$$E[X] = \int x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \mu.$$

Teorema 1.3 (Lei do Estatístico Preguiçoso)**.** *Seja $X : \Omega \to \chi \subset \mathbb{R}^n$ uma variável aleatória e $f : \mathscr{X} \to \mathbb{R}^k$ uma função mensurável. Considere a variável aleatória $Y = f(X)$. Temos que a esperança $E[Y] = \int y d\mu(y)$ pode ser calculada por:*

$$E[Y] = \int f(x) d\mu(x).$$

Demonstração. Vamos considerar um caso simples onde X é uma variável aleatória discreta. Escreva $y = f(x)$. Assim,

$$\begin{aligned} E[Y] &= \sum_y y P[Y = y] = \sum_y y P[f(X) = y] = \sum_y y \sum_{x \in f^{-1}(y)} P[X = x] \\ &= \sum_y \sum_{x \in f^{-1}(y)} f(x) P[X = x] = \sum_{x \in \chi} f(x) P[X = x]. \end{aligned} \tag{1.7}$$

□

Definição 11. *Seja X uma variável aleatória com esperança $\mu = E[X]$. A variância de X (denotada por $V[X]$ ou σ_X^2 ou σ^2), quando existir, é definida por*

$$V[X] = E[(X - \mu)^2] = \int (x - \mu)^2 d\mu(x). \tag{1.8}$$

Propriedades importantes da variância.

(V_1) Se $X = c$, então $V[X] = 0$.

(V_2) $V(X) = E(X^2) - \mu^2$.

(V_3) Se a e b são constantes então $V(aX + b) = a^2 V(X)$.

(E_a) Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes e $a_1, \ldots, a_n$ são constantes, então

$$V\left[\sum_i a_i X_i\right] = \sum_i a_i^2 V[X_i].$$

Exemplo 11. *Se $X \sim Ber(\lambda)$, então $E[X] = E[X^2] = 0P([X = 0]) + 1P([X = 1]) = \lambda$, donde $V[X] = E(X^2) - E[X]^2 = \lambda - \lambda^2 = \lambda(1-\lambda)$. Se $X \sim Bin(n, \lambda)$*

$$p(k) = \binom{n}{k} \lambda^k (1-\lambda)^{n-k}$$

então podemos escrever $X = X_1 + \ldots + X_n$, onde $X_1, \ldots, X_n$ são variáveis aleatórias independentes distribuidas por $Ber(\lambda)$. Temos assim que $E[X] = E[X_1] + \ldots + E[X_n] = n\lambda$ e $V[X] = V[X_1] + \ldots + V[X_n] = n\lambda(1-\lambda)$.

CAPÍTULO 2

Variedades diferenciáveis e modelos estatísticos

2.1 Variedades diferenciáveis e modelos estatísticos

Definição 12. *Um conjunto S e uma família de aplicações biunívocas $\xi_a : U_a \subset \mathbb{R}^n \to S$, de abertos U_a de $\mathbb{R}^n$ é uma variedade n-dimensional se*

(i) $\bigcup_a \xi_a(U_a) = S$;

(ii) Para todo par a, b com $\xi_a(U_a) \cap \xi_b(U_b) = W \neq \emptyset$, os conjuntos $\xi_a^{-1}(W)$ e $\xi_b^{-1}(W)$ são abertos de $\mathbb{R}^n$ e as aplicações $\xi_b^{-1} \circ \xi_a$ são diferenciáveis (vide figura 2.1);

(iii) A família $\{(U_a, \xi_a)\}$ (chamada de atlas de S) é maximal em relação a (i) e (ii)

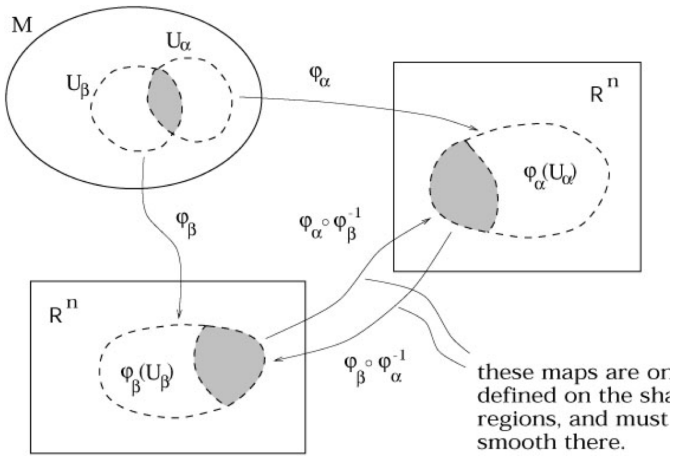

Figura 2.1 Variedade Diferenciável

Neste trabalho, apresentaremos principalmente variedades cujos pontos são distribuições de probabilidades. Uma simplificação que se mostrará bastante útil é que para esses conjuntos consideraremos apenas parametrizações globais.

Definição 13. *Seja $\mathscr{X}$ a imagem de uma variável aleatória X. Um modelo estatístico n-dimensional $S = \{p_\xi\}$ é uma família de distribuições de probabilidades sobre $\mathscr{X}$, globalmente parametrizadas por um aberto $\Theta \subset \mathbb{R}^n$. Ou seja,*

$$S = \{p_\xi(x) = p(x;\xi) \mid x \in \mathscr{X} \text{ e } \xi \in \Theta\},$$

é uma variedade n-dimensional cujo atlas $\{(\Theta, \psi)\}$ é unitário e é dado por $\psi : \xi \in \Theta \mapsto p_\xi \in S$.

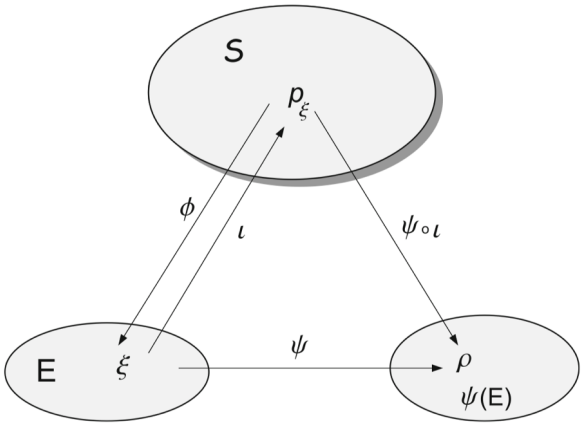

Figura 2.2 Modelo Estatístico

Fixado um atlas global $\{\Theta, \psi\}$ de um modelo estatístico S, segue-se (vide figura acima) que qualquer outro atlas global $\{(\tilde{\Theta}, \varphi\}$ é da forma $\psi = \varphi \circ \phi$, sendo $\phi : \Theta \to \tilde{\Theta}$ um difeomorfismo. O difeomorfismo $\phi : \xi \in \Theta \mapsto \eta \in \tilde{\Theta}$ é chamado de mudança de parametrização de S.

Definição 14. *Diremos que um modelo estatístico $S = \{p_\xi\}$ é regular se satisfaz as condições abaixo.*

(i) O suporte de cada p_ξ não depende de ξ, ou seja, o conjunto

$$\text{supp}(p_\xi) = \{x \in \mathscr{X} | p_\xi(x) > 0\}$$

é constante em relação a ξ.

(ii) Podemos trocar livremente a ordem de integração e derivação. Por exemplo,

$$\int \partial_i p(x,\xi) dx = \partial_i \int p(x,\xi) dx = 0. \tag{2.1}$$

(iii) Fixado $x \in \mathscr{X}$, as funções $\xi \in \Theta \mapsto p(x,\xi)$ são suaves, ou seja, admitem derivadas parciais (em relação a ξ) em todas as ordens.

(iv) Para cada $p \in S$ as derivadas parciais $\frac{\partial}{\partial \xi^i} p(x;\xi)$, com $i = 1,2,...,n$, são linearmente independentes (LI) como funções de x.

Assim, os vetores coordenados $\partial_i = \psi_* e_i$ (veja definição de vetor tangente na próxima seção) da parametrização $\psi : \xi \mapsto p_\xi$ podem ser representadas pelas derivadas parciais $\partial_i = \frac{\partial p_\xi}{\partial \xi^i}$, com $i = 1, \ldots, n$. Propriedade importantes sobre vetores tangentes e mudanças de coordenadas serão dadas na próxima seção.

Observe que S é um subconjunto de $P(\mathscr{X}) = \{p : \mathscr{X} \to (0,\infty) \mid \int p(x) d\mu(x) = 1\}$. Sem muitos detalhes, se $\mathscr{X}$ é infinito, então $P(\mathscr{X})$ é uma variedade diferenciável de dimensão infinita, cujo espaço tangente em cada ponto pode ser representado por $T_p P(\mathscr{X}) = \{\alpha : \mathscr{X} \to \mathbb{R} \mid \int \alpha(x) d\mu(x) = 0\}$. Assim, modelos estatísticos são subvariedade imersas em $P(\mathscr{X})$. No caso em que o $\mathscr{X}$ é finito, veremos que $P(\mathscr{X})$ é um modelo estatístico de dimensão $|\mathscr{X}| - 1$.

Exemplo 12. *(Modelo finito ou Distribuições categóricas $P(\mathscr{X})$ com $\mathscr{X} = \{0,\dots,n\}$)*
Dada uma distribuição de probabilidade $p(x)$ sobre $\mathscr{X} = \{0,1,\dots,n\}$, usando que $p(j) > 0$, para todo j, e $\sum_{j=0}^{n} p(j) = 1$, considere o vetor $\eta = (\eta_1,\dots,\eta_n)$, com $\eta_j = p(j)$. Temos que η pertence ao simplexo $S_n = \{\eta = (\eta_1,\dots,\eta_n) \mid \eta_i > 0 \text{ e } \sum_{i=1}^{n} \eta_i < 1\}$. Alem disso, a aplicação:

$$\eta \in S_n \mapsto p(x),$$

com $p(i) = \eta_i$, com $i = 1,\dots,n$ e $p(0) = 1 - \sum_j \eta_j$, define uma parametrização de $P(\mathscr{X})$.

Exemplo 13. *(Modelo Normal). A distribuição normal de média $\mu \in \mathbb{R}$ e desvio padrão $\sigma > 0$ é dada por*

$$p_\xi = p(x,\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.2}$$

Variando $(\mu,\sigma) \in E = \mathbb{R} \times (0,\infty)$, obtemos um modelo estatístico sobre $\mathscr{X} = \mathbb{R}$.

Exemplo 14. *(Distribuição Gama) A distribuição Gama é dada por*

$$p(x,\xi) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0, \tag{2.3}$$

com $\xi = (\alpha,\beta) \in E = (0,\infty) \times (0,\infty)$. Segue-se que $S = \{p_\xi\}_{\xi \in E}$ é um modelo estatístico de dimensão 2 sobre $\mathscr{X} = (0,\infty)$.

Exemplo 15. *(Distribuição de Poisson) A distribuição de Poisson é dada por*

$$p(x,\xi) = e^{-\xi} \frac{\xi^x}{x!}$$

com $x \in \mathscr{X} = \{0,1,\dots\}$ e $\xi \in E = (0,\infty)$.

Exemplo 16. *(Variedade de medidas positivas finitas). Muitas vezes é conveniente considerar o espaço das densidades positivas e finitas sobre $\mathscr{X}$. Ou seja, ao invés de $P(\mathscr{X})$, considere*

$$Dens_+(\mathscr{X}) = \{p : \mathscr{X} \to (0,\infty) \mid \int p(x) d\mu(x) < \infty\}.$$

Se $\mathscr{X} = \{1,\dots,n\}$ então $Dens_+(\mathscr{X}) = \mathbb{R}^n_+ = \{m = (m_1,\dots,m_n) \mid m_i > 0\}$ é um aberto de $\mathbb{R}^n$, donde é uma variedade n-dimensional parametrizada. Note que $Dens_+(\mathscr{X}) = \{q = \alpha p \mid \alpha > 0 \text{ e } p \in P(\mathscr{X})\}$. Assim, $P(\mathscr{X}) \subset Dens_+(\mathscr{X})$ possui codimensão 1.

2.2 Espaço tangente

Nesta seção, introduziremos a noção de aplicações diferenciáveis, vetor tangente e espaço tangente de uma variedade diferenciável S. Por fim, mostraremos que modelos estatísticos são subvariedades imersas em $P(\mathscr{X})$ (ou em $Dens_+(\mathscr{X})$ para modelos não-normalizados).

Definição 15. *Sejam M e N variedades diferenciáveis de dimensão m e n, respectivamente. Uma aplicação $F : M \to N$ é diferenciável em $p \in M$ se, dada uma parametrização $\psi : V \subset N \to \psi(V) \subset \mathbb{R}^n$ de $f(p)$, existe uma parametrização $\phi : U \subset M \to \phi(U) \subset \mathbb{R}^m$ de p tal que $F(U) \subset V$ e a aplicação*

$$\psi \circ F \circ \phi^{-1} : \phi(U) \to \mathbb{R}^m$$

é diferenciável em $\phi(p)$ (vide Figura 2.3). A aplicação F é diferenciável em um aberto de M se for diferenciável em todos os pontos desse aberto. Naturalmente, a definição dada acima independe da escolha de parametrizações.

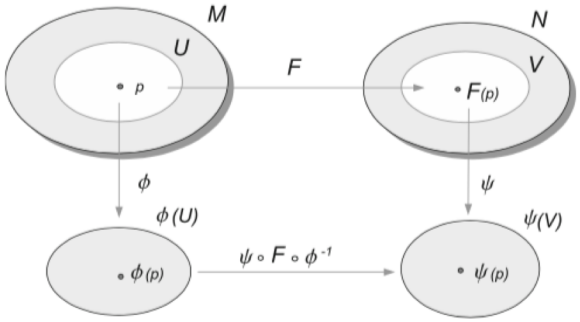

Figura 2.3 Aplicações diferenciáveis

Uma curva em M é simplesmente uma aplicação diferenciável $\gamma : (-\varepsilon, \varepsilon) \to M$.

Definição 16. *Sejam γ uma curva em M com $\gamma(0) = p$ e $f : M \to \mathbb{R}$ uma função diferenciável em p. O vetor tangente à curva γ em $t = 0$ é a aplicação*

$$\gamma'(0) : f \mapsto \frac{d(f \circ \gamma)}{dt}\Big|_{t=0}. \tag{2.4}$$

Um vetor tangente em p é o vetor tangente em $t = 0$ de alguma curva γ em M com $\gamma(0) = p$.

Vamos calcular a expressão do vetor $\gamma'(0)$ a partir de uma parametrização $\phi : U \subset M \to \mathbb{R}^m$. Considere $\phi^i(t) := \phi^i(\gamma(t))$ e $f(\gamma(t)) = f^*(\phi^1(t), \ldots, \phi^n(t))$, onde $f^*(\phi^1, \ldots, \phi^n) = f \circ \phi^{-1}$. Por abuso de notação, escrevemos $f(\phi^1, \ldots, \phi^n) = f^*(\phi^1, \ldots, \phi^n)$. Neste caso, diremos que f está expressada em termos de coordenadas locais. Segue-se que

$$\gamma'(0)f = \frac{d}{dt} f(\gamma(t))\Big|_{t=0} = \frac{\partial f}{\partial \phi^i}(\phi(0)) \frac{d\phi^i(0)}{dt} = \frac{d\phi^i(0)}{dt} \left(\frac{\partial}{\partial \phi^i}\right)_p f, \tag{2.5}$$

onde $(\frac{\partial}{\partial \phi^i})_p$ denota o vetor tangente $\frac{\partial}{\partial \phi^i}(p) : f \mapsto \frac{\partial f}{\partial \phi^i}(p)$. Assim, o vetor tangente $\gamma'(0)$ pode ser expresso como combinação linear dos vetores coordenados $\partial_i = \frac{\partial f}{\partial \phi^i}$, ou seja,

$$\gamma'(0) = \frac{d\phi^i}{dt}\Big|_{t=0} \left(\frac{\partial}{\partial \phi^i}\right)_p. \tag{2.6}$$

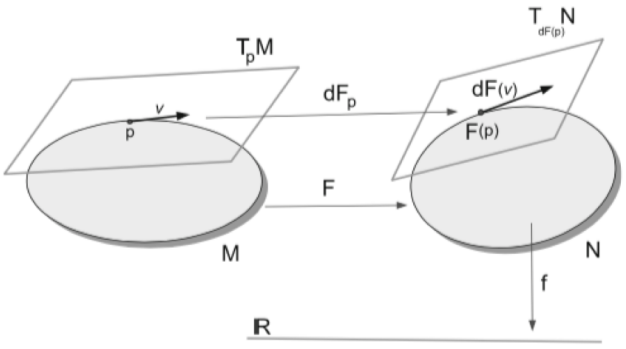

Figura 2.4 Vetores tangentes

Da definição (16), a noção de vetor tangente independe do sistema de coordenadas. Já da equação (2.6), temos que vetores tangentes são gerados como combinação linear dos vetores coordenados ∂_i. Assim, o espaço tangente de S em p, dado pelo conjunto dos vetores tangentes no ponto p,

$$T_pS = \{\gamma'(0) \mid \gamma\colon (-\varepsilon,\varepsilon) \to S \text{ diferenciável com } \gamma(0) = p\}$$

é um espaço vetorial de dimensão n, visto que $T_pS = \text{span}\{(\partial_1)_p,\ldots,(\partial_n)_p\}$ e os vetores coordenados ∂_i são linearmente independentes em cada ponto p.

Agora, dado um outro sistema de coordenadas $[\rho_j]$, usando que a mudança de cartas é um difeomorfismo, podemos escrever $\rho_j = \rho_j(\phi^1,\ldots,\phi^n)$ (em termos das coordenadas locais ϕ). Assim,

$$\frac{\partial f}{\partial \phi^i} = \frac{\partial \rho_j}{\partial \phi^i}\frac{\partial f}{\partial \rho_j}. \tag{2.7}$$

Logo, os vetores coordenados das parametrizações $[\phi^i]$ e $[\rho_j]$ relacionam-se por

$$\frac{\partial}{\partial \phi^i} = \frac{\partial \rho_j}{\partial \phi^i}\frac{\partial}{\partial \rho_j}. \tag{2.8}$$

Proposição 2.1. *Seja $F : M \to N$ uma aplicação diferenciável entre duas variedades M e N. Dados $p \in M$ e $v \in T_pM$, considere uma curva diferenciável $\alpha : (-\varepsilon,\varepsilon) \to M$ tal que $\alpha(0) = p$ e $\alpha'(0) = v$. Considere a curva $\beta(t) = F(\alpha(t))$. A aplicação $dF_p : T_pM \to T_{F(p)}N$ dada por $dF_pv = \beta'(0)$ é linear e independe da escolha de α.*

Demonstração. Sejam ϕ e ψ sistemas de coordenadas locais de M e N, respectivamente. Em termos de coordendas locais, $F(\phi_1,\ldots,\phi_m) = (\psi_1,\ldots,\psi_n)$, onde cada coordenada $\psi_i = \psi_i(\phi_1,\ldots,\phi_m)$. Seja $\alpha(t) = (\phi_1(t),\ldots,\phi_m(t))$ em coordendas locais. Assim, em coordendas locais, $\beta(t) = F(\alpha(t))$ se escreve por

$$\beta(t) = (\psi_1(\phi_1(t),\ldots,\phi_n(t)),\ldots,\psi_m(\phi_1(t),\ldots,\phi_n(t))).$$

Portanto,

$$\beta'(0) = \left(\frac{\partial \psi_1}{\partial \phi_j}\phi_j'(0),\ldots,\frac{\partial \psi_m}{\partial \phi_j}\phi_j'(0)\right)$$

Logo, $\beta'(0)$ independe da escolha da curva α, apenas de $\alpha'(0)$. Além disso, $\beta'(0) = dF_p\alpha'(0)$, onde $dF_p : T_pM \to T_{F(p)}N$ é uma aplicação linear, cuja representação matricial, fixada as bases coordenadas $\{\frac{\partial}{\partial\phi}\}$ e $\{\frac{\partial}{\partial\psi_j}\}$, é dada por $(\frac{\partial\psi_i}{\partial\phi_j})$. Logo, dF_p é uma aplicação linear.

Definição 17. *A aplicação dF_p definida acima é chamada de diferencial de F em p.*

Definição 18. *Seja $F : M \to N$ uma aplicação diferenciável entre duas variedades M e N. Dizemos que F é uma imersão se a diferencial dF_p for injetiva em cada ponto $p \in M$. Dizemos também que F é um mergulho se for F uma imersão injetiva e, considerando $F(M) \subset N$ com a topologia relativa, temos que F é um homeomorfismo sobre $F(M)$. Se $M \subset N$ e a inclusão $i : M \to N$ é um mergulho, dizemos que M é uma subvariedade de N.*

Quando $\mathscr{X}$ é finito, podemos identifica-lo ao conjunto $\{0,1,\ldots,n\}$. Assim, vale o seguinte

Teorema 2.2. *Modelos estatísticos regulares sobre $\mathscr{X}$ são subvariedades imersas em $P(\mathscr{X})$.*

Demonstração. Seja $S = \{p_\xi\}$, com $\xi \in E \subset \mathbb{R}^k$, um modelo estatístico regular sobre $\mathscr{X}$. Temos que provar que a aplicação inclusão: $\Psi : p_\xi \in S \mapsto p_\xi \in P(\mathscr{X})$ é uma imersão injetiva. A injetividade é obvia. Para provar que Ψ é uma imersão, temos que mostrar que, vista em termos de coordenadas locais, Ψ possui derivada injetiva. As coordenadas de p_ξ, visto como elemento de $P(\mathscr{X})$ são dadas por $\eta = (\eta_1,\ldots,\eta_n) \in S_n$ com $\eta_j = p_\xi(j)$. Assim, em termos de coordenadas locais, $\Psi(\xi) = \eta$. Afirmamos que os vetores $\Psi_*\partial_i = \frac{\partial}{\partial\xi^j}\eta$, com $j = 1,\ldots,k$, são vetores linearmente independentes. De fato, se $\alpha^j\frac{\partial}{\partial\xi^j}\eta = 0$ é uma combinação linear nula, então, cada coordenadas é também nula, ou seja, $\alpha^j\frac{\partial}{\partial\xi^j}\eta_i = 0$, com $i = 1,\ldots,n$. Como $p_\xi(0) = 1 - \sum_{j=1}^n \eta_j$ e $p_\xi(i) = \eta_i$, segue-se que $\alpha^j\frac{\partial}{\partial\xi^j}p_\xi = 0$. Como S é um modelo estatístico regular, segue-se que as funções $\partial_j = \frac{\partial}{\partial\xi^j}p_\xi$ são linearmente independentes, donde cada $\alpha^j = 0$. Portanto, a aplicação Ψ possui derivada Ψ_* injetiva. Assim, Ψ é uma imersão. □

2.3 Campos de vetores

Definição 19. *Seja S uma variedade. Uma aplicação $X : p \mapsto X_p$ que associa a cada ponto $p \in S$ um vetor tangente $X_p \in T_pS$ é chamada campo de vetores.*

Exemplo 17. *Dado um sistema de coordenadas $[\xi^i]$ em uma variedade S, temos n campos de vetores definidos por $\partial_i : p \mapsto (\partial_i)_p$, $i = 1,\ldots,n$.*

Dado um campo de vetores (ou campo vetorial) X e um sistema de coordenadas $[\xi^i]$, para cada ponto $p \in S$ escrevemos $X_p = X_p^i(\partial_i)_p$, ou seja, para cada $p \in S$, existem n escalares $X_1^p,\ldots,X_n^p$ que determinam X_p unicamente. Definimos as componentes de um campo de vetores X em relação a um sistema de coordenadas $[\xi^i]$ pelas funções $X^i : p \mapsto X_p^i$. Escrevemos então $X = X^i(\partial_i)_p$. Dado outro sistema de coordenadas $[\rho_j]$, da mesma forma temos $X = X_j^*\left(\partial^j\right)_p$. Segue então que

$$X_j^* = X^i\frac{\partial\rho_j}{\partial\xi^i} \quad e \quad X^i = X_j^*\frac{\partial\xi^i}{\partial\rho_j} \tag{2.9}$$

Se as componentes de um campo de vetores X em relação a um sistema de coordenadas é C^∞, então as componentes são C^∞ em relação a qualquer outro sistema de coordenadas. Quando isto ocorre, chamamos tal campo de vetores de um campo de vetores C^∞. Neste trabalho, consideraremos apenas tais campos. Denotaremos por $\tau(S)$ ou simplesmente τ o conjunto de todos os campos vetoriais em S. É fácil ver que as funções

(i) $X+Y : p \mapsto X_p + Y_p$

(ii) $cX : p \mapsto cX_p, c \in \mathbb{R}$

também são campos vetoriais. Temos também que se $f : S \to \mathbb{R}$ é uma função C^∞, a aplicação $fX : p \mapsto f(p)X$ também pertence a τ.
Podemos pensar em um campo de vetores como uma aplicação $X : D \to D$ do conjunto das funções C^∞ de S em $\mathbb{R}$ do seguinte modo:

$$(Xf)(p) = X_p^i \frac{\partial f}{\partial \xi^i}(p), f \in D \tag{2.10}$$

Com esta interpretação de X, podemos considerar os iterados de campos vetoriais. Sejam X e Y campos de vetores em S e $f \in D$. As funções $X(Yf)$ e $Y(Xf)$ em geral não são campos vetoriais, entretanto temos o seguinte resultado:

Teorema 2.3. *Sejam X e Y campos de vetores em S. então existe um único campo de vetores Z tal que, para todo $f \in D$, $Zf = (XY - YX)f$. Este campo de vetores é chamado de colchete de X e Y, e denotado por $[X,Y]$.*

Demonstração. Escreva $X = X^i\partial_i$ e $Y = Y^j\partial_j$. Então, $XYf = X(Y^j\partial_j f) = X^i\partial_i(Y^j\partial_j f) = X^i\partial_i Y^j\partial_j f + X^iY^j\partial_i\partial_j f$. Da mesma forma, $YXf = Y^j\partial_j X^i\partial_i f + Y^jX^i\partial_j\partial_i f$. Pelo Teorema de Schwarz, $[\partial_i,\partial_j]f = \partial_i\partial_j f - \partial_j\partial_i f = \frac{\partial^2 f}{\partial\xi^i\partial\xi^j} - \frac{\partial^2 f}{\partial\xi^j\partial\xi^i} = 0$. Assim,

$$Zf = [X,Y]f = (X^i\partial_i Y^j - Y^i\partial_i X^j)\partial_j f. \tag{2.11}$$

Assim, $Z = [X,Y] = (X^i\partial_i Y^j - Y^i\partial_i X^j)\partial_j$, define de fato um campo de vetores.

2.4 Métricas Riemannianas e métrica de Fisher

Definição 20. *Seja $S = \{p_\xi\}$ um modelo estatístico n-dimensional e $p_\xi \in S$. A matriz de informação de Fisher de S no ponto p é a matriz $[g_{ij}(\xi)]$ definida por*

$$g_{ij}(\xi) \stackrel{def}{=} E_{p_\xi}[\partial_i\ell_\xi\,\partial_j\ell_\xi] = \int \partial_i\ell_\xi(x)\partial_j\ell_\xi(x)p_\xi(x)\,dx. \tag{2.12}$$

onde $\ell_\xi(x) = \ln p_\xi(x)$ e $\partial_i\ell_\xi(x) = \frac{\partial}{\partial\xi^i}\ell_\xi$.

Também podemos representar a matriz de informação de Fisher de outras maneiras:

(1) $g_{ij} = 4\int \partial_i(\sqrt{p_\xi(x)})\partial_j(\sqrt{p_\xi(x)})dx.$

Isto segue-se diretamente da expressão $g_{ij}(\xi) = \int \frac{\partial_i p_\xi}{p_\xi}\frac{\partial_j p_\xi}{p_\xi}p_\xi(x)dx = \int \frac{\partial_i p_\xi}{\sqrt{p_\xi}}\frac{\partial_j p_\xi}{\sqrt{p_\xi}}\,dx.$

(2) $g_{ij} = -E_{p_\xi}[\partial_i\partial_i\ell_\xi]$.

De fato, como $E_{p_\xi}[\partial_j\ell_\xi] = \int(\partial_j\ell_\xi)p_\xi = \int \partial_j p_\xi = \partial_j\int p_\xi = \partial_j(1) = 0$, segue-se que

$$\begin{aligned} 0 &= \partial_i\int \partial_j\ell_\xi p_\xi = \int(\partial_i\partial_j\ell_\xi)p_\xi + \int \partial_j\ell_\xi\partial_i p_\xi \\ &= \int(\partial_i\partial_j\ell_\xi)p_\xi + \int \partial_j\ell_\xi\partial_i\ell_\xi p_\xi \\ &= E_{p_\xi}[\partial_i\partial_j\ell_\xi] + g_{ij}. \end{aligned}$$

É imediato verificar que a matriz de informação de Fisher é simétrica e suave sobre ξ. Note também que ela é positiva-semidefinida pois, para todo $c \in \mathbb{R}^n$,

$$c^t[g_{ij}(\xi)]c = c^i c^j g_{ij}(\xi) = \int\{c^i\partial_i\ell(x,\xi)\}^2 p(x,\xi)dx \geq 0. \tag{2.13}$$

Veremos também que a matriz de informação de Fisher define uma métrica Riemanniana sobre um modelo estatístico regular.

Definição 21. *Uma métrica Riemanniana em uma variedade diferenciável S é uma função diferenciável de S que associa a cada ponto $p \in S$ um produto interno $\langle,\rangle_p$ no espaço tangente T_pS. Em outras palavras, se $\xi : U \subset \mathbb{R}^n \to S$ é um sistema de coordenadas locais em torno de p, e $\{\partial_1,\ldots,\partial_n\}$ são os campos coordenados de ξ (isto é, $\partial_i(\xi(q)) = d\xi_q(e_i)$), então as funções*

$$g_{ij} : q \in U \mapsto g_{ij}(q) = \langle\partial_i(\xi(q)),\partial_j(\xi(q))\rangle_{\xi(q)}$$

são diferenciáveis em $\xi(U)$. Além disso, para todo $q \in U$, a matriz $(g_{ij}(q))$ é simétrica, positiva-definida e se $u = u^i\partial_i(\xi(q))$ e $v = v^j\partial_j(\xi(q))$ são vetores tangentes em $\xi(q)$, o produto interno $\langle u,v\rangle_{\xi(q)} = u^i v^j g_{ij}(q)$ independe da escolha do sistema de coordenadas ξ.

Teorema 2.4. *Seja S uma variedade estatística regular. Então a matriz de informação de Fisher define uma metrica Riemanniana em S.*

Demonstração. Fixemos um sistema de coordenadas em S, ou seja, $S = \{p_\xi\}$, com ξ definida num espaço de parâmetros $E \subset \mathbb{R}^n$. Primeiro, vamos provar que $[g_{ij}(\xi)]$ é uma matriz positiva-definida. De fato, de (2.13) segue-se que a matriz de informação de Fisher é positiva-semidefinida. Alem disso, mantendo-se a mesma notação acima, se $c \in \mathbb{R}^n$ satisfaz $c^t[g_{ij}(\xi)]c = 0$ então, por continuidade, $c^i\partial_i(x,\xi) = 0$, para todo x. Como $\partial_i\ell_\xi = \frac{\partial_i p_\xi}{p_\xi}$, da hipótese de regularidade, temos que as funções $\{\partial_1\ell_\xi,\ldots,\partial_n\ell_\xi\}$ são linearmente independentes (como funções de x). Assim, temos que $c = 0$. Portanto, a matriz $[g_{ij}(\xi)]$ é positiva-definida.

Agora, dado $p = p_\xi$, considere vetores tangentes $u, v \in T_pS$ e escreva $u = u^i\partial_i$ e $v = v^j\partial_j$ em termos de coordenadas locais. Assim, da definição de vetor tangente, temos que

$$u(\ell_\xi) = u^i\partial_i(\ell_\xi) = u^i\frac{\partial_i p_\xi}{p_\xi} = \frac{u(p)}{p}. \tag{2.14}$$

Logo, $u^i\partial_i(\ell_\xi) = \frac{u(p)}{p}$ depende apenas do ponto p e do vetor tangente u. Assim,

$$\begin{aligned} u^iv^jg_{ij}(\xi) &= u^iv^jE_p[\partial_i\ell_\xi\partial_j\ell_\xi] = E_p[(u^i\partial_i\ell_\xi)(v^j\partial_j\ell_\xi)] = E_p[u(\ell_\xi)v(\ell_\xi)] \\ &= E_p[\frac{u(p)}{p}\frac{v(p)}{p}], \end{aligned}$$

depende apenas do ponto p e dos vetores tangentes $u, v \in T_pS$. Como $g_{ij}(\xi)$ é uma matriz positiva-definida que varia suavemente com o parâmetro ξ, segue-se que $g(u,v)_p = u^iv^jg_{ij}(\xi)$ define uma métrica Riemanniana em S. □

Exemplo 18. *Seja $S = \{p_\xi = N(\cdot \mid \mu, \sigma^2) \mid \mu \in \mathbb{R} \text{ e } \sigma > 0\}$ o modelo de distribuiçoes normais com parametrização $\xi = (\mu, \sigma) \in \mathbb{R} \times (0, +\infty)$. Vamos calcular a matriz de informação de Fisher de S.*

Observemos que $\ell_\xi = \ln p_\xi = -\frac{(x-\mu)^2}{2\sigma^2} - \ln\sqrt{2\pi}\,\sigma$. Logo, derivando em relação a μ e σ,

(i) $\frac{\partial\ell_\xi}{\partial\mu} = \frac{x-\mu}{\sigma^2}$ e $\frac{\partial\ell_\xi}{\partial\sigma} = \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma}$.

(ii) $\frac{\partial^2\ell_\xi}{\partial\mu^2} = -\frac{1}{\sigma^2}$, $\frac{\partial^2\ell_\xi}{\partial\mu\partial\sigma} = -\frac{2(x-\mu)}{\sigma^3}$ e $\frac{\partial^2\ell_\xi}{\partial\sigma^2} = -\frac{3(x-\mu)^2}{\sigma^4} + \frac{1}{\sigma^2}$.

Usando que $\mu = E_{p_\xi}[x]$ e $\sigma^2 = E_{p_\xi}[(x-\mu)^2]$ e usando que $g_{ij} = -E_{p_\xi}[\partial_i\partial_j\ell_\xi]$, temos

(a) $g_{11} = -E_{p_\xi}[-\frac{1}{\sigma^2}] = \frac{1}{\sigma^2}$.

(b) $g_{12} = g_{21} = -E_{p_\xi}[-\frac{2(x-\mu)}{\sigma^3}] = 0$.

(c) $g_{22} = -E_{p_\xi}[-\frac{3(x-\mu)^2}{\sigma^4} + \frac{1}{\sigma^2}] = \frac{3}{\sigma^2} - \frac{1}{\sigma^2} = \frac{2}{\sigma^2}$.

Assim, S é uma superfície cuja métrica Riemanniana $ds^2 = g_{ij}d\xi^i \otimes d\xi^j$ é dada por:

$$ds^2 = \frac{1}{\sigma^2}(d\mu^2 + 2d\sigma^2). \tag{2.15}$$

Esse modelo é muito parecido com o plano de hiperbólico (modelo do semiplano de Poincaré), dado por $\mathbb{H}^2 = \{(\mu, \sigma) \mid \mu \in \mathbb{R} \text{ e } \sigma > 0\}$, com a seguinte métrica Riemanniana:

$$ds^2 = \frac{1}{\sigma^2}(d\mu^2 + d\sigma^2). \tag{2.16}$$

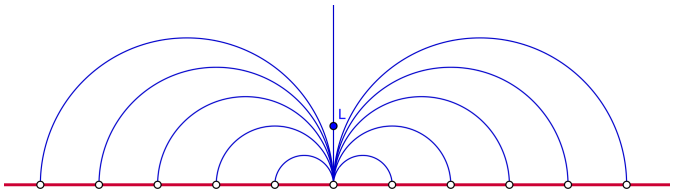

Figura 2.5 Plano Hiperbólico

*A métrica ds^2 de $\mathbb{H}^2$, dada por (*2.16), é completa, possui curvatura Gaussiana constante -1 e as geodésicas são as retas verticais $\mu =$ constante, ou os semicírculos que interseptam ortogonalmente o eixo μ (veja Figura 2.5).

Já no nosso caso, a métrica de Fisher ds^2 de S, dada por (2.15), é completa, possui curvatura Gaussiana constante $-1/2$ e as geodesicas são as retas verticais $\mu =$ constante, ou elipses que interseptam o eixo μ ortogonalmente.

Exemplo 19. *Vimos que o modelo das distribuições categóricas $P(\mathscr{X})$ com $\mathscr{X} = \{0,1,\ldots,n\}$ pode ser parametrizado pelo simplexo $S_n = \{\eta = (\eta_1,\ldots,\eta_n) \mid \eta_i > 0 \text{ e } \sum \eta_i = 1\}$. Ou seja, cada $p_\eta \in P(\mathscr{X})$ é dado por $p_\eta(i) = \eta_i$, com $i = 1,\ldots,n$ e $p_\eta(0) = 1 - \sum \eta_i$. Defina*

$$\xi_j = 2\sqrt{p_\eta(j)},$$

com $j = 0,\ldots,n$. Segue-se que $\xi_j > 0$, para todo j, e $\|\xi\|^2 = \sum_{j=0}^n \xi_j^2 = 4$. Assim, os pontos $\xi = (\xi_0,\ldots,\xi_n) \in S_+^n(2) = S^n(2) \cap \mathbb{R}_+^{n+1}$ (porção da esfera de raio 2 de coordenadas positivas).

Sendo $S_+^n(2)$ um aberto da esfera $S^n(2)$, segue-se que $S_+^n(2)$ é uma variedade de curvatura seccional constante $1/2$. Além disso, $S_+^n(2)$ pode ser parametrizada por $\bar{\xi} = (\xi_1,\ldots,\xi_n)$.

Um vetor tangente a S_+^n no ponto ξ é dada pela derivada $\xi'(0) = (\xi_0'(0),\ldots,\xi_n'(0))$, sendo $\xi(t)$ a curva em $S_+^n(2)$ dada por $\xi_k(t) = 2\sqrt{p_{\eta(t)}(k)}$, com $\eta(t) \in S_n$, $t \in (-\varepsilon,\varepsilon)$, e $\eta(0) = \eta$. Assim, derivando em $t = 0$,

$$\xi_k'(0) = \frac{1}{\sqrt{p_\eta(k)}} p_{\eta'(0)}(k),$$

com $k = 0,\ldots,n$. Logo, os campos coordenados de $S_+^n(2)$ são dados por

$$\frac{\partial}{\partial \bar{\xi}_i} = \Big[\frac{1}{\sqrt{p_\eta(k)}} \partial_i p_\eta(k)\Big]_{k=0,\ldots,n},$$

com $i = 1,\ldots,n$.

A métrica Riemanniana de $S_+^n(2)$ é induzida pelo produto escalar usual de $\mathbb{R}^{n+1}$. Assim,

$$\Big\langle \frac{\partial}{\partial \bar{\xi}_i}, \frac{\partial}{\partial \bar{\xi}_j} \Big\rangle = 4 \sum_k \partial_i(\sqrt{p_\eta}(k)) \partial_j(\sqrt{p_\eta}(k)) = g_{ij}(\eta).$$

Isso mostra que a metrica de Fisher de $P(\chi)$ coincide com a métrica usual de $S_+^n(2)$, logo $P(\chi)$ possui curvatura seccional constante positiva $1/2$ (veja Figura 2.6).

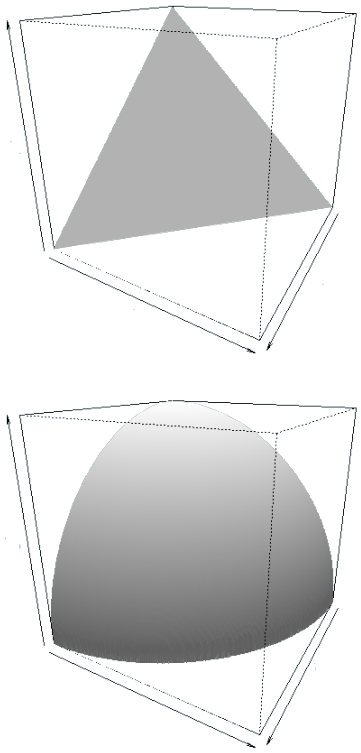

Figura 2.6 Na figura acima vemos o simplexo S_n e, em seguida, vemos o modelo das distribuições categóricas $P(\chi)$, com $\chi = \{0,1,\ldots,n\}$, visto como a porção positiva da esfera.

Vejamos algumas outras propriedades importantes da matriz de informação de Fisher G.

Aditividade. Sejam $S_1 = \{p_\xi(x) \mid x \in \chi_1\}_{\xi\in E}$ e $S_2 = \{q_\xi(y) \mid y \in \chi_2\}_{\xi\in E}$ dois modelos estatisticos regulares possuindo em comum o mesmo espaço de parâmetros E. Segue-se que

$$S = \{r_\xi(x,y) = p_\xi(x)q_\xi(y)\}$$

também define um modelo estatístico regular. Consideremos $G(\xi), G_1(\xi), G_2(\xi)$ as matrizes de informação de Fisher de S, S_1 e S_2, respectivamente. Temos

$$G(\xi) = G_1(\xi) + G_2(\xi).$$

De fato, como $\ln r_\xi(x,y) = \ln p_\xi(x) + \ln q_\xi(y)$, tem-se $\partial_i\partial_j \ln r_\xi(x,y) = \partial_i\partial_j \ln p_\xi(x) + \partial_i\partial_j \ln q_\xi(y)$. Assim,

$$\begin{aligned} E_r[\partial_i\partial_j \ln r_\xi(x,y)] &= \iint \partial_i\partial_j \ln r_\xi(x,y)p(x)q(y)dxdy \\ &= \int \partial_i\partial_j \ln p_\xi(x)p(x) \int q(y) + \int \partial_i\partial_j \ln q_\xi(y)q(y) \int p(x) \\ &= E_{p_\xi}[\partial_i\partial_j \ln p_\xi] + E_{q_\xi}[\partial_i\partial_j \ln q_\xi]. \end{aligned}$$

Usando (2), temos que $G(\xi) = G_1(\xi) + G_2(\xi)$.

Convexidade. Sejam $S_1 = \{p_1(x;\xi) \mid x \in \chi\}_{\xi \in E}$ e $S_2 = \{p_2(x;\xi) \mid x \in \chi\}_{\xi \in E}$ dois modelos estatisticos regulares possuindo em comum o mesmo espaço amostral $\mathscr{X}$ e de parâmetros E. Fixado $\lambda \in [0,1]$, segue-se que

$$S_\lambda = \{p_\lambda(x;\xi) = (1-\lambda)p_1(x;\xi) + \lambda p_2(x;\xi)\},$$

define um modelo estatístico regular. As matrizes de informação de Fisher G_1, G_2, G_λ, de S_1, S_2 e S_λ, respectivamente, satisfazem:

$$G_\lambda(\xi) \preceq (1-\lambda)G_1(\xi) + \lambda G_2(\xi), \tag{2.17}$$

onde " $\preceq$ " significa que a diferença entre $G_\lambda(\xi)$ e $(1-\lambda)G_1(\xi)+\lambda G_2(\xi)$ é negativa-semidefinida. A desigualdade (2.17) será provada mais adiante (vide Teorema 4.8), como consequência da convexidade da Divergência de Kullback-Leibler (ou mais geralmente, das f-divergência).

Monotonicidade e invariância por estatísticas suficientes. Uma estatística é uma função mensurável $F : \mathscr{X} \to \mathscr{Y}$ entre espaços amostrais. A grosso modo, F comprime os dados de uma variável aleatória X e espera-se poder inferir a distribuição de probabilidade $p(x)$ relacionada a X simplesmente olhando para os resultados da variável aleatória $Y = F(X)$. Mostraremos que a matriz de informação de Fisher capta essa perda de informação no sentido de que a diferença da matriz original com a induzida pela estatística é positiva-semidefinida. E as matrizes não se alteram se, e somente se, a estatística é suficiente. Para isso, vamos precisar de algumas definições. Uma distribuição de probabilidade $p(x)$ sobre $\mathscr{X}$ induz uma distribuição de probabilidade $q(y)$ sobre $\mathscr{Y}$ da seguinte forma: Se $B \subset \mathscr{Y}$ é mensurável então, a probabilidade $P(B) = \int_B q(y)d\mu(y)$ é dada por $P(B) = \int_{F^{-1}(B)} p(x)d\mu(x)$. Chamaremos $q(y)$ a distribuição de probabilidade induzida de p pela estatística F.

Se F é um difeomorfismo entre abertos $\mathscr{X}$ e $\mathscr{Y}$ de $\mathbb{R}^n$, e $d\mu(x), d\mu(y)$ são as correspondentes medidas de Lebesgue, então pelo teorema da mudança de variáveis, $\int_B q(y)d\mu(y) = \int_{F^{-1}(B)} q(F(x))|JF(x)|d\mu(x)$, onde $JF(x)$ é o determinante Jacobiano. O mesmo vale se $\mathscr{X}$ e $\mathscr{Y}$ são variedades Riemannianas e $d\mu(x), d\mu(y)$ são as correspondentes formas elemento de volume. Assim, vale o seguinte:

$$p(x) = q(F(x))|JF(x)|. \tag{2.18}$$

Mais geralmente, apenas supondo F uma estatística, escreva $p(y \mid x) = \delta(y - F(x))$, onde δ é a função de Kronecker, dada por $\delta(z) = 1$ se $z = 0$ e $\delta(z) = 0$ se $z \neq 0$. Assim,

$$p(x \mid y) = p(y \mid x)\frac{p(x)}{q(y)} = \delta(y - F(x))\frac{p(x)}{q(F(x))}. \tag{2.19}$$

Logo,

$$p(x) = q(F(x))p(x \mid F(x)). \tag{2.20}$$

Definição 22. *Seja $F : \mathscr{X} \to \mathscr{Y}$ uma estatística entre espaços amostrais. Dado um modelo estatístico $S = \{p_\xi(x)\}$ sobre $\mathscr{X}$, considere o modelo estatístico induzido $S_F = \{q_\xi(y)\}$ sobre $\mathscr{Y}$. Diremos que F é uma estatística suficiente para S se a probabilidade condicional $p_\xi(x \mid F(x))$ não depender do parâmetro ξ.*

Por (2.18), difeomorfismos entre espaços amostrais (sendo estas variedades Riemannianas) são casos particulares de estatísticas suficientes. Considere também o seguinte exemplo.

Exemplo 20. *Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes e igualmente distribuidas pela distribuição de Bernoulli $\text{Ber}(x \mid \xi) = \xi^x(1-\xi)^{1-x}$, com $x \in \{0,1\}$ e $\xi \in (0,1)$. Considere a estatística $F : \mathscr{X}^n = \{0,1\}^n \to \mathscr{Y} = \{0,1,\ldots,n\}$ dada por $F(x_1,\ldots,x_n) = x_1 + \ldots + x_n$. A variável aleatória $X = (X_1,\ldots,X_n)$ é distribuida pela distribuição Multinomial $\text{Mult}(x \mid \xi) = \Pi_{i=1}^n \xi^{x_i}(1-\xi)^{1-x_i} = \xi^{F(x)}(1-\xi)^{n-F(x)}$, com $x = (x_1,\ldots,x_n) \in \{0,1\}^n$ e a variável aleatória $Y = F(X_1,\ldots,X_n) = X_1 + \ldots + X_n$ é distribuida pela distribuição binominal $\text{Bin}(y \mid n,\xi) = \binom{n}{y}\xi^y(1-\xi)^{n-y}$. Usando (2.20), a probabilidade condicional $p_\xi(x \mid F(x))$ é dada por*

$$p_\xi(x \mid F(x)) = \frac{\text{Mult}(x \mid \xi)}{\text{Bin}(F(x) \mid n,\xi)} = \frac{\xi^{F(x)}(1-\xi)^{n-F(x)}}{\binom{n}{F(x)}\xi^{F(x)}(1-\xi)^{n-F(x)}} = \frac{1}{\binom{n}{F(x)}},$$

que independe de ξ. Logo, obtemos que F é uma estatística suficiente para o modelo estatístico $S = \{\text{Ber}(x \mid \xi)\}$.

O teorema abaixo dá uma condição necessária e suficiente para que uma estatística seja suficiente.

Teorema 2.5 (Decomposição de Fisher-Neyman)**.** *Uma estatística $F : \mathscr{X} \to \mathscr{Y}$ é suficiente para um modelo estatístico $S = \{p_\xi(x)\}_{\xi \in E}$ se, e somente se, existem funções $s(y;\xi)$ e $t(x)$ tais que vale a seguinte decomposição:*

$$p_\xi(x) = s(F(x);\xi)t(x), \tag{2.21}$$

para todo $x \in \mathscr{X}$ e $\xi \in E$.

Note que a condição "necessária" do Teorema 2.5 é obtida a partir de (2.20), juntamente com a hipótese de F ser uma estatística suficiente.

A monotonicidade da matriz de informação de Fisher é descrita no seguinte teorema:

Teorema 2.6 (Monotonicidade). *Seja $F : \mathscr{X} \to \mathscr{Y}$ uma estatística em espaços amostrais. Seja $G(\xi) = (g_{ij}(\xi))$ a matriz de informação de Fisher de um modelo estatístico $S = \{p_\xi\}$ sobre $\mathscr{X}$. Seja $G^F = (g^F_{ij})$ a matriz de informação de Fisher do modelo induzido $S_F = \{q_\xi(y)\}$. Temos que $G(\xi) \succeq G_F(\xi)$, no sentido de que a matriz $\Delta G(\xi) = G(\xi) - G_F(\xi)$ é positiva-semidefinida. Além disso, a igualdade vale se, e somente se F é uma estatística suficiente.*

Demonstração. Seja $B \subset \mathscr{Y}$ mensurável

$$\begin{aligned}\int_B \partial_i \log q_\xi(y) q_\xi(y) dy &= \int_B \partial_i q_\xi(y) dy = \partial_i \int_B q_\xi(y) dy \\ &= \partial_i \int_{F^{-1}(B)} p_\xi(x) dx = \int_{F^{-1}(B)} \partial_i \log p_\xi(x) p_\xi(x) dx \\ &= \int_{F^{-1}(B)} \partial_i \log p_\xi(x) \int_{\mathscr{Y}} p_\xi(x \mid y) q_\xi(y) dy dx \\ &= \int_B \int_{\chi} \partial_i \log p_\xi(x) p_\xi(x \mid y) q_\xi(y) dx dy.\end{aligned}$$

Para a ultima igualdade, usamos que $p(x \mid y) = 0$ se $x \in F^{-1}(B)$ e $y \notin B$ ou se $y \in B$ e $x \notin F^{-1}(B)$. Isso segue do fato de que $p(x \mid y)$ é um multiplo de $p(y \mid x) = \delta(y - F(x))$ (veja (2.19)). Logo, como B é arbitrário,

$$\partial_i \log q_\xi(y) = E[\partial_i \log p_\xi(X) \mid y]. \tag{2.22}$$

Por outro lado, por (2.20), temos que $p_\xi(x) = q_\xi(F(x)) r_\xi(x)$, onde $r_\xi(x) = p(x \mid F(x))$. Assim,

$$\partial_i \log p_\xi(x) = \partial_i \log q_\xi(F(x)) + \partial_i \log r_\xi(x).$$

Daí temos

$$\begin{aligned}E[\partial_i \log p_\xi(x) \mid F(x)] &= E[\partial_i \log q_\xi(F(x)) \mid F(x)] + E[\partial_i \log r_\xi(x) \mid F(x)] \\ &= \partial_i \log q_\xi(F(x)) + E[\partial_i \log r_\xi(x) \mid F(x)].\end{aligned}$$

Por (2.22), segue-se que $\partial_i \log q_\xi(F(x)) = E[\partial_i \log p_\xi(X) \mid F(x)]$, donde

$$E[\partial_i \log r_\xi(x) \mid F(x)] = 0. \tag{2.23}$$

Denotando $\ell_\xi(x) = \log p_\xi(x)$, por (2.22), temos que matriz de covariancia,

$$\begin{aligned}cov[\partial_i \ell_\xi \partial_j \ell_\xi \mid y] &= E[(\partial_i \ell_\xi - E[\partial_i \ell_\xi \mid y])(\partial_j \ell_\xi - E[\partial_j \ell_\xi \mid y]) \mid y] \\ &= E[\partial_i \ell_\xi \partial_j \ell_\xi \mid y] - E[\partial_i \ell_\xi \mid y] E[\partial_j \ell_\xi \mid y] \\ &= E[\partial_i \ell_\xi \partial_j \ell_\xi \mid y] - \partial_i \log q_\xi(y) \partial_j \log q_\xi(y).\end{aligned}$$

Assim,

$$\begin{aligned}E^y[cov[\partial_i \ell_\xi \partial_j \ell_\xi \mid y]] &= E^y[E^x[\partial_i \ell_\xi \partial_j \ell_\xi \mid y]] - E[\partial_i \log q_\xi(y) \partial_j \log q_\xi(y)] \\ &= E^x[\partial_i \ell_\xi \partial_j \ell_\xi] - E^y[\partial_i \log q_\xi(y) \partial_j \log q_\xi(y)] \\ &= g_{ij}(\xi) - g^F_{ij}(\xi).\end{aligned} \tag{2.24}$$

Aqui, para evitar embiguidades, denotamos E^x, E^y as esperanças nas variáveis x e y, respectivamente. Como matrizes covariantes são positiva-semidefinidas, segue-se que $\Delta G(\xi) = G(\xi) - G^F(\xi)$ é uma matriz positiva-semidefinida.

Além disso, se $\Delta G(\xi) = 0$ então, $cov[\partial_i \ell_\xi \partial_j \ell_\xi \mid y] = 0$, para todo y. Em particular,

$$0 = \partial_i \ell_\xi(x) - E[\partial_i \ell_\xi \mid F(x)] = \partial_i \ell_\xi(x) - \partial_i \log q_\xi(F(x)) = \partial_i \log r_\xi(x),$$

para todo i. Como $r_\xi(x) = p_\xi(x \mid F(x))$, segue-se que $p_\xi(x \mid F(x))$ independe de ξ. Logo, F é uma estatística suficiente. □

Provaremos uma versão mais geral desse teorema mais adiante, na seção 4.2.

CAPÍTULO 3

Conexões

3.1 Conexões afins e α-conexões

Definição 23. *Sejam S uma variedade diferenciável, X, Y, Z campos de vetores e $f, g \in F$. Uma conexão afim em S é uma aplicação $\nabla : (X,Y) \mapsto \nabla_X Y$ e que satisfaz:*

(i) $\nabla_{fX+gY} Z = f\nabla_X Y + g\nabla_Y Z$;

(ii) $\nabla_X (Y+Z) = \nabla_X Y + \nabla_X Z$;

(iii) $\nabla_X (fY) = f\nabla_X Y + X(f)Y$.

Vamos escrever $\nabla_X Y$ em termos de coordenadas locais $[\xi^i]$. Escrevendo $X = X^i \partial_i$, $Y = Y^j \partial_j$ e $\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k \partial_k$, temos

$$\nabla_X Y = \nabla_X (Y^j \partial_j) = X(Y^j)\partial_j + X^i Y^j \nabla_{\partial_i} \partial_j = \{X(Y^k) + X^i Y^j \Gamma_{ij}^k\}\partial_k \tag{3.1}$$

Os coeficientes Γ_{ij}^k são chamados de símbolos de Christoffel em relação ao sistema de coordenadas. De (3.1), vemos que em cada ponto $p \in S$, a conexão $(\nabla_X Y)(p)$ depende apenas de $X(p)$ e de Y ao longo de uma curva tangente a X proximo ao ponto p.

Definição 24. *Uma conexão ∇ em uma variedade diferenciável S é simétrica se satisfaz*

$$\nabla_X Y - \nabla_Y X = [X,Y]. \tag{3.2}$$

Aqui, $[X,Y] = XY - YX$ é o campo colchete entre X e Y (veja Teorema 2.3).

Por (2.11) e (3.1), temos que $[X,Y] = \{X^i \partial_i Y^j - Y^i \partial_i X^j\}\partial_j$. Assim,

$$\begin{aligned}\nabla_X Y - \nabla_Y X - [X,Y] &= X^i\{\partial_i Y^k + Y^j \Gamma_{ij}^k\}\partial_k - Y^i\{\partial_i X^k + X^j \Gamma_{ij}^k\}\partial_k - \{X^i \partial_i Y^k - Y^i \partial_i X^k\}\partial_k \\ &= \{X^i Y^j \Gamma_{ij}^k - X^j Y^i \Gamma_{ij}^k\}\partial_k = X^i Y^j \{\Gamma_{ij}^k - \Gamma_{ji}^k\}\partial_k.\end{aligned}$$

Portanto, temos o seguinte:

Proposição 3.1. *Uma conexão ∇ em uma variedade S é simétrica se, e somente se $\Gamma_{ij}^k = \Gamma_{ji}^k$.*

Proposição 3.2. *Seja S uma variedade diferenciável com uma conexão afim ∇. Então existe uma única correspondência que associa a um campo vetorial V ao longo de uma curva γ um outro campo vetorial (que denotaremos por $\frac{DV}{dt}$ e chamaremos de derivada covariante de V ao longo de γ), tal que:*

(i) $\frac{D}{dt}(V+W) = \frac{DV}{dt} + \frac{DW}{dt}$;

(ii) $\frac{D}{dt}(fV) = \frac{df}{dt} + f\frac{DV}{dt}$, *onde* V *é um campo vetorial e* $f : I \to \mathbb{R}$ *é diferenciável;*

(iii) Se $V(t) = Y(\gamma(t))$, *para algum* $Y \in \tau(S)$, *então* $\frac{DV}{dt} = \nabla_{\frac{d\gamma}{dt}} Y$.

Demonstração. Supondo a existência, a expressão de $\frac{DV}{dt}$ em um sistema de coordenadas $[\xi^i]$ é dada por

$$\frac{DV}{dt} = \left\{ \frac{dv^k}{dt} + \frac{d\gamma^i}{dt}\frac{dv^j}{dt}\Gamma^k_{ij} \right\} \partial_k,$$

o que prova a unicidade. Para provar a existência, basta definir $\frac{DV}{dt}$ como a equação acima e verificar que satisfaz $(i), (ii)$ e (iii)

Definição 25. *Seja* S *uma variedade diferenciável com uma conexão* ∇. *Um campo vetorial* V *ao longo de uma curva* $\gamma : I \to S$ *é chamado transporte paralelo se* $\frac{DV}{dt} = 0$, *para todo* $t \in I$.

Proposição 3.3. *Seja* S *uma variedade diferenciável e* ∇ *uma conexão em* S. *Seja* $\gamma : I \to \mathbb{R}$ *uma curva em* S *e* V_0 *um vetor tangente a* S *em* $c(t_0) \in S$. *Então existe um único transporte paralelo* V *ao longo de* γ *satisfazendo* $V(t_0) = V_0$.

Demonstração. Supondo inicialmente a existência e que γ está contida em uma parametrização, a expressão da derivada covariante é

$$\frac{DV}{dt} = \left\{ \frac{dv^k}{dt} + \frac{d\gamma^i}{dt}\frac{dv^j}{dt}\Gamma^k_{ij} \right\} \partial_k = 0.$$

Logo, temos um sistema de equações diferenciais

$$\frac{dv^k}{dt} + \frac{d\gamma^i}{dt}\frac{dv^j}{dt}\Gamma^k_{ij} = 0.$$

que possui solução única dada a condição inicial $v^k(t_0) = v^k_0$. □

Vimos que modelos estatísticos regulares admitem, de modo natural, uma métrica Riemanniana, a saber, a métrica de Fisher. Essa métrica possui a importante propriedade de ser, a menos de uma constante, a única métrica Riemanniana (sobre modelos estatísticos) que é" invariante por estatísticas suficientes. Vamos exibir uma família de conexões sobre modelos estatísticos, as α-conexões. Tais conexões também possuem a propriedade de serem as únicas conexões afins e simétricas invariantes por estatísticas suficientes (veja Cĕncov [11] e N. Ay, J. Jost, H. V. Lê and L. Schwachhöfer [4]).

Definição 26. *Seja* $S = \{p_\xi\}$ *um modelo estatístico munido com a métrica de Fisher* g. *Dado* $\alpha \in \mathbb{R}$, *defina a* α*-conexão afim* $\nabla^{(\alpha)}$, *cujos símbolos de Christoffel são dados da seguinte forma:*

$$\Gamma^{(\alpha)}_{ij,k}(\xi) = E_\xi \left[\left(\partial_i\partial_j\ell_\xi + \frac{1-\alpha}{2}\partial_i\ell_\xi\partial_j\ell_\xi \right)(\partial_k\ell_\xi) \right]. \tag{3.3}$$

Em outras palavras, as funções $\Gamma^{(\alpha)}_{ij,k}$ *definem* $\nabla^{(\alpha)}$ *da seguinte forma*

$$\langle \nabla^{(\alpha)}_{\partial_i} \partial_j, \partial_k \rangle = \Gamma^{(\alpha)}_{ij,k}.$$

É imediato verificar que a α-conexão é simétrica. Além disso, vale que, para todo $\beta \in \mathbb{R}$,

$$\Gamma^{(\beta)}_{ij,k} = \Gamma^{(\alpha)}_{ij,k} + \frac{\alpha-\beta}{2} T_{ijk}$$

onde $T_{ijk} \stackrel{def}{=} E_\xi[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi]$.

3.2 Conexões Riemannianas

Definição 27. *Seja S uma variedade com uma conexão ∇ e uma métria Riemanniana g e sejam X, Y, Z campos vetoriais em S. Dizemos que uma conexão é compatível com a métrica $g = \langle\ ,\ \rangle$ se vale*

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle. \tag{3.4}$$

Proposição 3.4. *Em coordenadas locais, a expressão (*3.4) é equivalente a

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}. \tag{3.5}$$

Demonstração. É imediato que (3.4) implica em (3.5), basta tomar $X = \partial_i, Y = \partial_j$ e $Z = \partial_k$. Reciprocamente, escreva $X = X^i\partial_i, Y = Y^j\partial_j$ e $Z = Z^k\partial_k$. Basta observar que

$$\begin{aligned}\langle \nabla_Z X, Y\rangle &= Z^k Y^j \langle \nabla_{\partial_k} X^i \partial_i, \partial_j\rangle = Z^k Y^j X^i \langle \nabla_{\partial_k}\partial_i, \partial_j\rangle + Z^k Y^j (\partial_k X_i) g_{ij} \\ &= Z^k Y^j X^i \Gamma_{ki,j} + Z^k Y^j (\partial_k X^i) g_{ij}.\end{aligned}$$

Da mesma forma, temos $\langle X, \nabla_Z Y\rangle = Z^k Y^j X^i \Gamma_{kj,i} + Z^k X^i (\partial_k Y^j) g_{ij}$. Assim, usando que $\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}$, temos

$$\begin{aligned}Z\langle X, Y\rangle = Z^k \partial_k \langle X^i\partial_i, Y^j\partial_j\rangle &= Z^k \partial_k (X^i Y^j g_{ij}) \\ &= Z^k (\partial_k X^i) Y^j g_{ij} + Z^k X^i (\partial_k Y^j) g_{ij} + Z^k X^i Y^j (\partial_k g_{ij}) \\ &= Z^k (\partial_k X^i) Y^j g_{ij} + Z^k X^i (\partial_k Y^j) g_{ij} + Z^k X^i Y^j (\Gamma_{ki,j} + \Gamma_{kj,i}) \\ &= \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle,\end{aligned}$$

□

Definição 28. *Seja S uma variedade com uma métrica g. Dizemos que uma conexão ∇ em S é Riemanniana (ou de Levi-Civita) se ∇ é simétrica e compatível com a métrica g.*

Teorema 3.5. *Uma variedade Riemanniana (S, g) admite uma única conexão ∇ de Levi-Civita.*

Demonstração. Supondo a existência, vale:

$$Z\langle X,Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle; \tag{3.6}$$

$$X\langle Y,Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle; \tag{3.7}$$

$$Y\langle Z,X\rangle = \langle \nabla_Y Z, X\rangle + \langle Z, \nabla_Y X\rangle. \tag{3.8}$$

Temos também

$$\begin{aligned}\langle \nabla_X Y, Z\rangle + \langle Z, \nabla_Y X\rangle =& \langle \nabla_X Y, Z\rangle - \langle Z, \nabla_Y X\rangle + 2\langle Z, \nabla_Y X\rangle \\ =& \langle [X,Y], Z\rangle + 2\langle Z, \nabla_Y X\rangle.\end{aligned}$$

Daí, somando (3.7) e (3.8) e subtraindo (3.6), temos

$$\begin{aligned}X\langle Y,Z\rangle + Y\langle Z,X\rangle - Z\langle X,Y\rangle =& \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle + \langle \nabla_Y Z, X\rangle + \\ & \langle Z, \nabla_Y X\rangle - \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle.\end{aligned}$$

Logo,

$$\begin{aligned}X\langle Y,Z\rangle + Y\langle Z,X\rangle - Z\langle X,Y\rangle =& 2\langle Z, \nabla_Y X\rangle + \langle [X,Y], Z\rangle \\ & + \langle [X,Z], Y\rangle + \langle [Y,Z], X\rangle.\end{aligned}$$

Daí,

$$\begin{aligned}\langle Z, \nabla_Y X\rangle =& \frac{1}{2}\{X\langle Y,Z\rangle + Y\langle Z,X\rangle - Z\langle X,Y\rangle \\ & - \langle [X,Y], Z\rangle - \langle [X,Z], Y\rangle - \langle [Y,Z], X\rangle\}.\end{aligned} \tag{3.9}$$

O que mostra que ∇ é única. A identidade (3.9) é conhecida por Formula de Koszul. Para mostrar a existência, basta definir ∇ pela equação acima e verificar que é simétrica e compatível com a métrica g. □

Proposição 3.6. *Os símbolos de Christoffel de uma conexão Riemanniana ∇ são dados por*

$$\Gamma_{ij}^k = \frac{1}{2}\{\partial_i g_{jm} + \partial_j g_{mi} - \partial_m g_{ij}\} g^{mk}. \tag{3.10}$$

Demonstração. De (3.9), temos

$$\langle \Gamma_{ij}^l \partial_l, \partial_m\rangle = \Gamma_{ij}^l g_{lk} = \frac{1}{2}\{\partial_i g_{jm} + \partial_j g_{mi} - \partial_m g_{ij}\}. \tag{3.11}$$

De onde, multiplicando pela matriz inversa g^{mk}, segue o resultado. □

Teorema 3.7. *Seja S uma variedade estatística com g a métrica de Fisher. Então a 0-conexão é a conexão de Levi-Civita em relação a g.*

Demonstração.

$$\begin{aligned}\partial_k g_{ij} &= \partial_k E[\partial_i \ell \partial_j \ell] \\ &= E[(\partial_k \partial_i \ell)(\partial_j \ell)] + E[(\partial_i \ell)(\partial_k \partial_j \ell)] + E[(\partial_i \ell)(\partial_j \ell)(\partial_k \ell)] \\ &= E\left[\left(\partial_k \partial_i \ell + \frac{(1-\alpha)}{2}\partial_k \ell \partial_i \ell\right)(\partial_j \ell)\right] + E\left[\left(\partial_k \partial_j \ell + \frac{(1+\alpha)}{2}\partial_k \ell \partial_j \ell\right)(\partial_i \ell)\right] \\ &= \Gamma_{ki,j}^{(\alpha)} + \Gamma_{kj,i}^{(-\alpha)}.\end{aligned}$$

Daí vemos que a condição de compatibilidade (3.5) é satisfeita se, e somente se, $\alpha = 0$. □

3.3 Conexões duais

Definição 29. *Seja S uma variedade munida de uma métrica Riemanniana $g = \langle , \rangle$ e sejam ∇ e ∇^* duas conexões em S. Diremos que ∇ e ∇^* são conexões duais se, para todos campos vetoriais $X, Y, Z \in \tau(S)$, vale*

$$Z\langle X,Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla^*_Z Y\rangle \tag{3.12}$$

Analogamente à Proposição 3.5, mostra-se que a expressão em (3.12) é equivalente, em coordenadas locais, a $\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma^*_{kj,i}$.

Proposição 3.8. *Dada uma métrica g e uma conexão em S, existe uma única conexão dual ∇^* de ∇ em relação a g*

Demonstração. Basta definir ∇^* por $\langle X, \nabla^*_Z Y\rangle = Z\langle X,Y\rangle - \langle \nabla_Z X, Y\rangle$. A unicidade decorre imediatamente. □

A conexão de Levi-Civita é auto-dual, isto é, satisfaz a condição de dualidade com $\nabla^* = \nabla$. logo mais apresentaremos conexões que não são auto-duais.

Proposição 3.9. *Seja (S, g, ∇, ∇^*) uma estrutura dual e suponha ∇ e ∇^* simétricas. Então a conexão $\tilde{\nabla} = \frac{\nabla + \nabla^*}{2}$ é a conexão de Levi-Civita em relação à métrica g.*

Demonstração. É simples verificar que $\tilde{\nabla}$ é uma conexão compatível com a métrica. Para isto, basta somar as duas equações abaixo

$$\frac{1}{2} Z\langle X,Y\rangle = \frac{1}{2}\left(\langle \nabla_Z X, Y\rangle + \langle X, \nabla^*_Z Y\rangle\right)$$

e

$$\frac{1}{2} Z\langle Y,X\rangle = \frac{1}{2}\left(\langle \nabla_Z Y, X\rangle + \langle Y, \nabla^*_Z X\rangle\right),$$

Agora, suponha que ∇ e ∇^* são simétricas. Temos que

$$\begin{aligned}
\tilde{\nabla}_X Y - \tilde{\nabla}_Y X &= \frac{1}{2}(\nabla_X Y + \nabla^*_X Y) - \frac{1}{2}(\nabla_Y X + \nabla^*_Y X) \\
&= \frac{1}{2}(\nabla_X Y - \nabla_Y X) + \frac{1}{2}(\nabla^*_X Y - \nabla^*_Y X) \\
&= \frac{1}{2}[X,Y] + \frac{1}{2}[X,Y].
\end{aligned}$$

Assim, segue-se que $\tilde{\nabla}$ é simétrica. Conclui-se então que $\tilde{\nabla}$ é a conexão de Levi-Civita. □

Teorema 3.10. *Seja S um modelo estatístico. Então a α-conexão e a $(-\alpha)$-conexão são duais em relação à Métrica de Fisher.*

Demonstração.

$$\begin{aligned}\partial_k g_{ij} =& \partial_k E[\partial_i l_\xi \partial_j l_\xi] \\ =& E[(\partial_k\partial_i \ell_\xi)(\partial_j \ell_\xi)] + E[(\partial_i \ell_\xi)(\partial_k\partial_j \ell_\xi)] + E[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi] \\ =& E[(\partial_k\partial_i \ell_\xi)(\partial_j \ell_\xi)] + E[(\partial_i \ell_\xi)(\partial_k\partial_j \ell_\xi)] + \frac{1-\alpha}{2} E[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi] + \frac{1+\alpha}{2} E[\partial_i \ell_\xi \partial_j \ell_\xi \partial_k \ell_\xi] \\ =& \Gamma^{(\alpha)}_{ki,j} + \Gamma^{(-\alpha)}_{kj,i}\end{aligned}$$

□

Teorema 3.11. *Denotando R e R^* respectivamente a curvatura com respeito ∇ e ∇^*, então para todos campos de vetores $X,Y,Z,W \in \tau(S)$ temos que $\langle R^*(X,Y)Z,W\rangle = -\langle R(X,Y)W,Z\rangle$. Em particular $R = 0$ se, e somente se, $R^* = 0$.*

Demonstração.

$$\begin{aligned}\langle R^*(X,Y)Z,W\rangle &= \langle \nabla^*_Y\nabla^*_X Z,W\rangle - \langle \nabla^*_X\nabla^*_Y Z,W\rangle + \langle \nabla^*_{[X,Y]}Z,W\rangle \\ &= Y\langle \nabla^*_X Z,W\rangle - \langle \nabla^*_X Z,\nabla_Y W\rangle - (X\langle \nabla^*_Y Z,W\rangle - \langle \nabla^*_Y Z,\nabla_X W\rangle) + \langle \nabla^*_{[X,Y]}Z,W\rangle \\ &= YX\langle Z,W\rangle - Y\langle Z,\nabla_X W\rangle - \langle \nabla^*_X Z,\nabla_Y W\rangle - XY\langle Z,W\rangle + X\langle Z,\nabla_Y W\rangle \\ &\quad + \langle \nabla^*_Y Z,\nabla_X W\rangle + \langle \nabla^*_{[X,Y]}Z,W\rangle \\ &= -(Y\langle Z,\nabla_X W\rangle - \langle \nabla^*_Y Z,\nabla_X W\rangle) + (X\langle Z,\nabla_Y W\rangle - \langle \nabla^*_X Z,\nabla_Y W\rangle) \\ &\quad -([X,Y]\langle Z,W\rangle - \langle \nabla^*_{[X,Y]}Z,W\rangle) \\ &= -\langle \nabla_Y\nabla_X W,Z\rangle + \langle \nabla_X\nabla_Y W,Z\rangle - \langle \nabla_{[X,Y]}W,Z\rangle \\ &= -\langle R(X,Y)W,Z\rangle.\end{aligned}$$

3.4 Geodésicas

Definição 30. *Seja S uma variedade diferenciável munida de uma conexão afim ∇. Dizemos que uma curva diferenciável $\gamma : I \to S$ é uma geodésica se*

$$\frac{D\gamma'}{dt} = \nabla_{\gamma'}\gamma' = 0 \tag{3.13}$$

para todo $t \in I$. Em outras palavras, uma geodésica é uma curva que cuja derivada é um transporte paralelo.

Em termos de coordenadas locais $[\xi^i]$, escreva $\gamma(t) = [\gamma^1(t),\ldots,\gamma^n(t)]$. Temos que $\gamma'(t) = \frac{d\gamma^i}{dt}\partial_i$, logo

$$\begin{aligned}\frac{D\gamma'}{dt} = \nabla_{\gamma'}\gamma' = \nabla_{\gamma'}\Big(\frac{d\gamma^i}{dt}\partial_i\Big) = \frac{d^2\gamma^i}{dt^2}\partial_i + \frac{d\gamma^i}{dt}\nabla_{\gamma'}\partial_i = \frac{d^2\gamma^k}{dt^2}\partial_k + \frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}\nabla_{\partial_j}\partial_i \\ = \Big(\frac{d^2\gamma^k}{dt^2} + \Gamma^k_{ij}\frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}\Big)\partial_k.\end{aligned}$$

Logo, γ é uma geodésica se, e somente se

$$\frac{d^2\gamma^k}{dt^2} + \Gamma^k_{ij}\frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt} = 0. \tag{3.14}$$

Portanto, usando o teorema de existência e unicidade de soluções de sistema de equações diferenciais, temos o resultado abaixo:

Teorema 3.12. *Seja S uma variedade diferenciável munida de uma conexão afim ∇. Então, para cada $p \in S$, $v \in T_pS$ e $t_0 \in \mathbb{R}$, existem um intervalo $I \subset \mathbb{R}$ contendo t_0 e uma única geodésica $\gamma : I \to S$ tal que $\gamma(t_0) = p$ e $\gamma'(t_0) = v$.*

Se (S, g) é uma variedade Riemanniana munida de uma conexão Riemaniana ∇, então

$$\frac{d}{dt}\langle\gamma', \gamma'\rangle = 2\langle\frac{D\gamma'}{dt}, \gamma'\rangle = 0.$$

Isto é, $|\gamma'|$ é constante. Logo, as geodésicas são parametrizadas proporcionalmente ao comprimento de arco, i.é., se $\gamma : [0,b] \to S$ é uma geodésica, então para todo $s \in [0,b]$, temos que o comprimento $L(\gamma|_{[0,s]}) = \int_0^s |\gamma'(t)|dt = cs$, para alguma constante c. Dizemos que a geodésica γ está normalizada se $|\gamma'| = 1$. É claro que toda geodesica pode ser renormalizada. Para isso, basta reparametrizar $\tilde{\gamma}(t) = \gamma(at)$, para algum $a \in \mathbb{R}$.

Já para conexões não-Riemannianas, essa propriedade não é valida em geral. No entanto, se $\gamma(t)$ e $\gamma^*(t)$ são geodesicas com respeito as conexões duais ∇ e ∇^*, temos

$$\frac{d}{dt}\langle\gamma', \gamma^{*\prime}\rangle = \langle\frac{D\gamma'}{dt}, \gamma'\rangle + \langle\gamma', \frac{D\gamma^{*\prime}}{dt}\rangle = 0.$$

Assim, $\langle\gamma', \gamma^{*\prime}\rangle$ é constante.

Exemplo 21. *(Geodésicas do $\mathbb{R}^n$) Como a métrica do $\mathbb{R}^n$ é dada por $g_{ij} = \langle e_i, e_j\rangle = \delta_{ij}$ e a conexão Riemanniana é dada (via seus simbolos de Christoffel) por*

$$\Gamma^k_{ij} = \frac{1}{2}\{\partial_i g_{jm} + \partial_j g_{im} - \partial_m g_{ij}\}g^{km} = 0$$

*Temos que a equação das geodésicas (*3.14) é dada por

$$\frac{d^2\gamma^k}{dt^2} = 0 \tag{3.15}$$

Daí, $\gamma(t) = ct + v_0$, para alguns c e v_0 em $\mathbb{R}^n$. Isto é, as geodésicas do $\mathbb{R}^n$ são retas em $\mathbb{R}^n$.

Note que (3.15) é válido sempre que $\Gamma_{ij,k} = 0$. Dizemos nesse caso que as coordenadas $[\xi^i]$ são ∇-afim. Veremos mais adiante que famílias exponenciais satisfazem $\Gamma^{(1)}_{ij,k} = 0$, com respeito ao sistema de coordenadas naturais. E para as famílias misturas temos $\Gamma^{(-1)}_{ij,k} = 0$ com respeito aos parâmetros da mistura.

3.5 Conexões planas

Definição 31. *Seja $X \in \tau(S)$ um campo vetorial em S. Se para quaquer curva γ em S, $X_\gamma : t \mapsto X_{\gamma(t)}$ é paralelo ao longo de γ em relação a uma conexão ∇, dizemos que X é paralelo em S (em relação à conexão ∇).*

Usando a unicidade do transporte paralelo, verifica-se facilmente que para qualquer curva γ, ligando dois pontos $p, q \in S$, o transporte paralelo X_p em q ao longo de γ é dado por $\prod_\gamma X_p = X_q$. Equivalentemente, note que um campo de vetores X é paralelo se, e somente se $\nabla_Y X = 0$ para todo $Y \in \tau(S)$.

Definição 32. *Seja S uma variedade S com uma conexão afim ∇. Dizemos que um sistema de coordenadas $[\xi^i]$ é afim com respeito à ∇ (ou simplesmente ∇-afim), se os campos coordenados $\partial_1, \ldots, \partial_n$ são paralelos com respeito à ∇. É imediato que $[\xi^i]$ é ∇-afim se, e somente se, $\nabla_{\partial_i}\partial_j = 0$, ou equivalentemente, os símbolos de Christoffel Γ^k_{ij} são identicamente nulos.*

Por outro lado, uma conexão afim ∇ é plana se admitir um sistema de coordenadas ∇-afim.
Obs. Mesmo sem demonstração vale a pena comentar que uma conexão simétrica é plana se, e somente se, o tensor curvatura, $R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z$, é identicamente nulo.

Proposição 3.13. *Seja $[\xi^i]$ um sistema de coordenadas para uma variedade S com uma conexão ∇. São equivalentes:*

(i) *$[\xi^i]$ é um sistema de coordenadas afim para ∇;*

(ii) *$\nabla_{\partial_i}\partial_j = 0$, para todos i, j;*

(iii) *Os símbolos de Christoffel Γ^k_{ij} são nulos.*

Vamos ver agora uma condição necessária e suficiente para outro sistema de coordenadas $[\rho_j]$ ser afim. Os símbolos de Christoffel nesse sistema são dados por

$$\tilde{\Gamma}^t_{rs} = \frac{\partial^2 \xi^k}{\partial \rho_r \partial \rho_s}\frac{\partial \rho_t}{\partial \xi^k}.$$

Daí,

$$\tilde{\Gamma}^t_{rs} = 0 \iff \frac{\partial^2 \xi^k}{\partial \rho_r \partial \rho_s} = 0.$$

Daí, tiramos que $\xi(p) = A\rho(p) + B$, onde A é uma matriz regular $n \times n$ e $B \in \mathbb{R}^n$. Isto é, sistemas de coorenadas afins são preservados por transformações afins.

3.6 Famílias exponenciais e famílias misturas

Dizemos que um modelo estatístico $S = \{p_\theta | \theta \in \Theta\}$ é uma família exponencial de dimensão n se cada distribuição de probabilidade $p \in S$ pode ser expressada por

$$p(x,\theta) = e^{C(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)},$$

onde C e F_i são funções reais de $\mathscr{X}$, ψ uma função real de Θ e o conjunto $\{1, F_1, \ldots, F_n\}$ é linearmente independente como funções de x. O sistema de coordenadas $[\theta]$ é chamado de sistema de coordenadas naturais (ou canônicos).

Há uma vasta lista de modelos estatísticos que são famílias exponenciais, incluindo distribuições normais, gamma, beta, Dirichlet, Bernoulli, Poisson, dentre outros. Famílias misturas (veremos a seguir), em geral, não são famílias exponenciais. Um outro exemplo de modelo estatístico que não é da família exponencial é o modelo de Weibull, cujas funções de distribuição de probabilidades são dadas por

$$w(x,\xi) = k\tau(kx)^{\tau-1}e^{(kx)^{\tau-1}} \quad k,\tau > 0.$$

Veja Khadiga Arwini [3] para mais detalhes da geometria desta distribuição. A função $\psi : \theta \to \mathbb{R}$ é dada por

$$\psi(\theta) = \ln \int e^{C(x)+\theta^i F_i(x)} dx,$$

e é chamada de função cumulante.

Exemplo 22. *Modelos finitos $P(\mathscr{X})$, com $\mathscr{X} = \{0,\ldots,n\}$, são famílias exponenciais. De fato, como $p(x) = \sum_{i=0}^{n} p_i \delta(x-i)$, temos*

$$\begin{aligned}\log p(x) &= \sum_{i=0}^{n} \log p_i\, \delta(x-i) = \sum_{i=1}^{n} \log p_i\, \delta(x-i) + \log p_0\, \delta(x-0) \\ &= \sum_{i=1}^{n} \left(\log \frac{p_i}{p_0}\right) \delta(x-i) + \log p_0.\end{aligned}$$

Daí, $C(x) = 0$; $\theta^j = \log \frac{p_j}{p_0}$ e $F_j(x) = \delta(x-j)$, com $j = 1,\ldots,n$; e $\psi(\theta) = -\log p_0$. Agora, como $p_j = p_0 e^{\theta^j}$, com $j = 1,\ldots,n$, temos que $p_0 = 1 - \sum_{j=1}^{n} p_j = 1 - p_0 \sum_{j=1}^{n} e^{\theta^j}$. Donde, $p_0 = (1+\sum_{j=1}^{n} e^{\theta^j})^{-1}$. Logo, $\psi(\theta) = -\log p_0 = \log\left(1+\sum e^{\theta^i}\right)$.

Exemplo 23. *No modelo das distribuições normais $S = \{N(x;\mu,\sigma^2) \mid \mu \in \mathbb{R} \text{ e } \sigma > 0\}$, temos*

(i) $C(x) = 0$, $F_1(x) = x$, $F_2(x) = x^2$.

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

(iii) $\psi(\theta) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\ln(-\frac{\pi}{\theta^2})$.

Geometricamente, famílias exponenciais são 1-planas com respeito ao sistema de coordenadas naturais. De fato, como $\ell_\theta(x) = \ln p_\theta(x) = C(x) + \theta^i F_i(x) - \psi(\theta)$. Assim,

$$\Gamma^{(1)}_{ij,k} = \int \partial_i\partial_j \ell_\theta \partial_k \ell_\theta dx = -\partial_i\partial_j\psi(\theta) \int \partial_k \ell_\theta dx = 0. \tag{3.16}$$

Chamamos a 1-conexão de conexão exponencial ou e-conexão e escreveremos $\nabla^{(1)} = \nabla^{(e)}$.

Vamos apresentar agora um tipo de modelo estatístico que deixa a (-1)-conexão plana. Famílias misturas são modelos estatísticos $S = \{p_\theta\}$ cujas distribuições de probabilidades $p_\theta(x)$ podem ser escritos da forma

$$p_\theta(x) = C(x) + \sum \theta^i F_i(x). \tag{3.17}$$

Misturas de probablidades são distribuições de probabilidades definidas por

$$p(x) = \sum_{i=1}^{n} \alpha^i p_i(x), \tag{3.18}$$

onde $\alpha^0, \ldots, \alpha^n \in [0,1]$ são tais que $\sum \alpha^i = 1$ e $p_0(x), \ldots, p_n(x)$ são distribuições de probabilidades. É claro que mistura de probabilidades são famílias misturas, para isto basta escrever $p(x) = (1 - \sum_{i=1}^{n} \alpha^i) p_0(x) + \alpha^i p_i(x)$. O sistema de coordenadas $[\theta^i]$ em (3.17) é chamado de sistema de coordenadas misturadas (ou parâmetros misturados). Como $\partial_i \ell_\theta = \frac{F_i(x)}{p(x,\theta)}$ e $\partial_i \partial_j \ell_\theta = -\frac{F_i(x)F_j(x)}{p(x,\theta)^2}$, temos que $\partial_i \partial_j \ell_\theta + \partial_i \ell_\theta \partial_j \ell_\theta = 0$. Assim, os coeficientes da conexão $\nabla^{(-1)}$ satisfazem

$$\Gamma^{(-1)}_{ij,k} = \int (\partial_i \partial_j \ell_\theta + \partial_i \ell_\theta \partial_j \ell_\theta) \partial_k \ell_\theta dx = 0. \tag{3.19}$$

Logo, o sistema de coordenadas misturadas $[\theta]$ é (-1)-afim. Chamamos a conexão $\nabla^{(-1)}$ de conexão mistura ou m-conexão.

No modelo finito $P(\chi)$, com $\chi = \{0, \ldots, n\}$, temos $p(x) = \sum_{i=0}^{n} p_i \delta(x-i)$. Logo $P(\chi)$ é uma mistura de probabilidades. As coordenadas da mistura são $\eta_i = p_i$ com $i = 1, \ldots, n$. Em Exemplo 22, também mostramos que $P(\chi)$ é uma família exponencial.

3.7 Conexões induzidas sobre subvariedades

Seja (S, g, ∇) uma variedade Riemanniana munida uma conexão $\bar{\nabla}$ e seja $f : M \to S$ uma imersão de uma variedade diferenciável M em S. Nesse caso, dizemos que M é uma subvariedade imersa em S. A métrica induzida de S em M é definida por

$$\langle X, Y \rangle = \langle f_* X, f_* Y \rangle \tag{3.20}$$

para quaisquer $X, Y \in \tau(M)$. Aqui, f_* denota a diferencial $(f_* X)(p) = df_p(X(p))$, para todo $p \in M$. Se M é munida da métrica induzida (3.20), dizemos que a imersão é isométrica. Como exemplo, famílias exponenciais curvadas são simplesmente subvariedades imersas em uma família exponencial. Induzido em ambas variedades a métrica de Fisher, temos que a subvariedade está isometricamente imersa.

A conexão ∇ induzida sobre uma variedade isométricamente imersa em M é dada pela projeção ortogonal

$$\langle \nabla_X Y, Z \rangle = \langle \bar{\nabla}_{f_* X} f_* Y, f_* Z \rangle,$$

para quaisquer campos de vetores X, Y e $Z \in \tau(M)$. Ou seja,

$$\nabla_X Y = (f_*)^{-1} \Big((\bar{\nabla}_{f_* X} f_* Y)^T \Big). \tag{3.21}$$

Como f é uma imersão, temos que, em cada ponto $p \in M$, a derivada $(f_*)(p) = df_p$ é um isomorfismo de T_pM sobre sua imagem $f_*(T_pM) = df_p(T_pM)$. Por abuso de notação, vamos identificar f_*X com X. Assim, $\nabla = \bar{\nabla}^T$ é a projeção ortogonal de $\bar{\nabla}$ sobre M. Vê-se que ∇ define as hipóteses de uma conexão afim sobre M (veja Definição 23).

Sejam $[\xi^i]$ e $[u^a]$ sistemas de coordenadas de S e M, respectivamente. Considere os campos de vetores $X = X^a\partial_a, Y = Y^b\partial_b \in \tau(M)$. Como ∇ define uma conexão afim sobre M temos

$$\begin{aligned}\nabla_X Y &= \nabla_X(Y^b\partial_b) = X(Y^b)\partial_b + Y^b\nabla_X\partial_b = X(Y^b)\partial_b + X^aY^b\nabla_{\partial_a}\partial_b \\ &= X(Y^b)\partial_b + X^aY^b\Gamma_{ab,c}g^{cd}\partial_d,\end{aligned}$$

onde $\Gamma_{ab,c}\langle\nabla_{\partial_a}\partial_b, \partial_c\rangle$ são os simbolos de Christoffel da conexão ∇. Usando que $\nabla = \bar{\nabla}^T$, podemos escrever $\Gamma_{ab,c}$ em termos da conexão do ambiente $\bar{\nabla}$ da seguinte forma:

$$\begin{aligned}\Gamma_{ab,c} &= \langle\nabla_{\partial_a}\partial_b, \partial_c\rangle = \langle\bar{\nabla}_{\partial_a}\Big(\frac{\partial\xi^j}{\partial u^b}\partial_j\Big), \partial_c\rangle = \frac{\partial^2\xi^j}{\partial u^b\partial u_b}\langle\partial_j, \partial_c\rangle + \frac{\partial\xi^j}{\partial u^b}\langle\bar{\nabla}_{\partial_a}\partial_j, \partial_c\rangle \\ &= \frac{\partial^2\xi^j}{\partial u^b\partial u_b}\langle\partial_j, \partial_c\rangle + \frac{\partial\xi^j}{\partial u^b}\frac{\partial\xi^i}{\partial u^a}\langle\bar{\nabla}_{\partial_i}\partial_j, \partial_c\rangle \\ &= \frac{\partial^2\xi^j}{\partial u^b\partial u_b}\frac{\partial\xi^k}{\partial u^c}g_{jk} + \frac{\partial\xi^i}{\partial u^a}\frac{\partial\xi^j}{\partial u^b}\frac{\partial\xi^k}{\partial u^c}\bar{\Gamma}_{ij,k}.\end{aligned}$$

Assim, uma subvariedade imersa $f : M \to S$ herda do ambiente S tanto a métrica g quanto a conexão $\nabla = \bar{\nabla}^T$.

A projeção ortogonal $B(X,Y) = (\bar{\nabla}_{f_*X}f_*Y)^N$ é chamada de *segunda forma fundamental* da imersão f. Usando (3.21), segue-se a formula de Weingartein,

$$\bar{\nabla}_{f_*X}f_*Y = f_*(\nabla_X Y) + B(X,Y). \tag{3.22}$$

Assumindo que $\bar{\nabla}$ é simétrica temos que B é simétrica e tensorial (i.e., $B(X,Y)(p)$ depende apenas de $X(p)$ e $Y(p)$). De fato,

$$B(X,Y) - B(Y,X) = (\bar{\nabla}_{f_*X}f_*Y)^\perp - (\bar{\nabla}_{f_*Y}f_*X)^\perp = [f_*X, f_*Y]^\perp = 0.$$

Assim, como $(\nabla_X Y)(p)$ depende de X apenas no ponto $X(p)$ e como $B(X,Y) = B(Y,X)$, temos que $B(X,Y)(p)$ depende apenas de $X(p)$ e $Y(p)$. Logo, B é tensorial.

3.8 Subvariedades totalmente geodésicas

Definição 33. *Uma subvariedade M isometricamente imersa em S é dita totalmente geodésica se a imagem por f de geodésicas de M são geodésicas de S.*

Por exemplo, vimos que as geodésicas do espaço Euclideano são retas $\gamma(t) = vt + v_0$, com $v, v_0 \in \mathbb{R}^n$. Assim, uma subvariedade m-dimensional M é totalmente geodesica se, e somente se, é parte de um m-plano.

Em termos de conexões,

Proposição 3.14. *Sejam $(S, g = \langle , \rangle, \bar{\nabla})$ uma variedade diferenciável n-dimensional munida de uma conexão $\bar{\nabla}$ e seja M uma subvariedade m-dimensional em S. Então, M é uma subvariedade totalmente geodésica se, e somente se, $\bar{\nabla}_X Y \in \tau(M)$, para quaisquer $X, Y \in \tau(M)$. Em outras palavras, M é totalmente geodésica se, e só se, a segunda forma fundamental da imersão é identicamente nula.*

Teorema 3.15. *Uma subvariedade isometricamente imersa M em uma família exponencial (analog. familia mistura) S é uma família exponencial (analog. familia mistura) se, e somente se M é e-totamente geodésica (analog. m-totalmente geodésica) em S.*

Demonstração. Suponha que M é uma família exponencial e seja $f : M \to S$ uma imersão isométrica injetiva de $M = \{q_u(x)\}$ sobre a família exponencial $S = \{p_\xi(x)\}$. Escreva $\log q_u(x) = D(x) + u^a G_a(x) - \varphi(u)$, sendo $1, G_1, \ldots, G_m$ funções linearmente independentes. Assim, $\bar{M} = \{\bar{q}_u = f \circ q_u\}$ é uma família exponencial curvada da forma

$$\log \bar{q}_u(x) = D(x) + u^a G_a(x) - \varphi(u) = C(x) + \xi^i(u) F_i(x) - \psi(\xi(u)).$$

Derivando-se ambos lados duas vezes em relação a u^a e u^b, temos

$$-\partial_a \partial_b \varphi(u) = \frac{\partial^2 \xi^i}{\partial u^a u^b} F_i(x) - \partial_a \partial_b \psi(\xi(u)).$$

Temos assim, $\frac{\partial^2 \xi^i}{\partial u^a u^b} F_i(x) + \partial_a \partial_b (\varphi(u) - \psi(\xi(u))) = 0$. Usando que $1, F_1, \ldots, F_n$ são l.i., segue-se que $\frac{\partial^2 \xi^i}{\partial u^a u^b} = 0 = \partial_a \partial_b (\varphi(u) - \psi(\xi(u)))$, para quaisquer $i = 1, \ldots, n$ e $a = 1, \ldots, m$. Assim,

$$\xi^i(u) = A^i_a u^a + b^i \quad \text{e} \quad \psi(\xi(u)) - \varphi(u) = cu + d, \tag{3.23}$$

com A^i_a, b^i, c, d constantes. Como $\bar{q}_u(x) = f(q_u(x)) = p_{\xi(u)}(x)$, usando (3.23), os campos coordenados de $M = \{q_u\}$ satisfazem

$$\partial_a \bar{q}_u = f_* \partial_a q_u = \frac{\partial p_\xi}{\partial \xi^i} \frac{\partial \xi^i}{\partial u^a} = \frac{\partial \xi^i}{\partial u^a} \partial_i p_\xi = A^i_a \partial_i.$$

Como S é e-plana, segue-se que os vetores coordenados ∂_i são e-paralelos. Assim,

$$\langle \nabla^{(e)}_{f_* \partial_a} f_* \partial_b, \partial_k \rangle = \sum_{i,j} A^a_i A^b_j \langle \nabla^{(e)}_{\partial_i} \partial_j, \partial_k \rangle = 0,$$

para todos $a, b = 1, \ldots, n$ e $k = 1, \ldots, m$. Segue-se que M é e-totalmente geodésica.

Agora, suponha que M é e-totalmente geodésica. Então, a segunda forma fundamental

$$B(\partial_a, \partial_b) = (\nabla^{(e)}_{f_* \partial_a} f_* \partial_b)^N = 0.$$

Logo, dada uma curva $\alpha(t) \in M$ ligando dois pontos e um e-transporte paralelo $V(t)$ ao longo de α, tem-se que $f_*(V(t))$ é um e-transporte de S ao longo de $\beta(t) = f(\alpha(t))$. Como S é e-plana, o e-transporte paralelo não depende da curva ligando os pontos extremos. Assim, o e-transporte

paralelo sobre M também não depende da curva. Segue-se que M possui e-curvatura $R_M^{(e)} = 0$. Como a conexão $\nabla^{(e)}$ é simétrica, tem-se que M admite coordenadas e-afins $u = [u^a]$. Usando que ∂_i são e-paralelos, $f_*(\partial_a) = \frac{\partial \xi^i}{\partial u^a}\partial_i$, e $\nabla^{(e)}_{f_*\partial_a} f_*\partial_b = \Gamma^{(e)c}_{ab} f_*\partial_c + B(\partial_a, \partial_b) = 0$, segue-se que

$$0 \;=\; \nabla^{(e)}_{f_*\partial_a} f_*\partial_b = \nabla^{(e)}_{f_*\partial_a}(\frac{\partial \xi^i}{\partial u^b}\partial_i) = \frac{\partial^2 \xi^i}{\partial u^a \partial u^b}\partial_i + \frac{\partial \xi}{\partial u^b}\nabla^{(e)}_{f_*\partial u^a}\partial_i = \frac{\partial^2 \xi^i}{\partial u^a \partial u^b}\partial_i.$$

Logo, $\frac{\partial^2 \xi^i}{\partial u^a \partial u^b} = 0$, para quaisquer i, a, b. Temos que $\xi^i(u) = \sum_a A^i_a u^a + b^i$, sendo A^i_a e b^i constantes. Como S é uma família exponencial, segue-se que

$$\begin{aligned} \log \bar{q}_u(x) &= p(x;\xi(u)) = C(x) + \xi^i(u)F_i(x) - \psi(\xi(u)) \\ &= C(x) + b^i F_i(x) + \sum_a (\sum_i A^i_a F_i(x))u^a - \psi(\xi(u)). \end{aligned}$$

Donde, $\bar{M} = \{\bar{q}_u\}$ é uma família exponencial. Isto implica que $M = \{q_u\}$ é também uma família exponencial.

Da mesma forma prova-se que uma subvariedade isometricamente imersa M de uma família mistura é uma família mistura se, e somente se M é m-totalmente geodésica em S. □

CAPÍTULO 4

Divergência e Teorema Pitagoreano

Um dos pontos principais do estudo de geometria da informação é sem duvida a noção de divergências. Com ela é possível medir dissimilaridades e hierarquias em dados. Veremos nos próximos capítulos que algumas medidas usuais de posição e dispersão podem ser traduzidos e generalizados de um modo simples e natural ao contexto de divergências.

Veremos como divergências induzem métricas Riemannianas e conexões afins sobre modelos estatísticos. Isso será bastante útil tanto ao Teorema Pitagoreano (que veremos a seguir) quanto ao estudo de convergência no algoritmo EM, que será feita no último capitulo.

4.1 Geometria induzida por uma divergência

Em geometria, é comum medir a distância entre dois pontos p e q em uma variedade diferenciável S (não necessariamente um modelo estatístico). Uma distância em S é uma aplicação $d : S \times S \to [0,\infty)$ que satisfaz as seguintes condições:

(i) $d[p : q] \geq 0$ e vale a igualdade se e só se $p = q$;

(ii) d é simétrica: $d[p : q] = d[q : p]$;

(iii) d satisfaz a desigualdade triangular: $d[p : q] \leq D[p : r] + D[r : q]$, para todo r.

No entanto, em análise de dados, a noção de proximidade pode ter mais a ver com as noções de similaridade ou hierarquia do que simplesmente distâncias em si. Porém, a relação de similaridade é obviamente assimétrica. Por exemplo, em análise de dados, poderíamos nos deparar facilmente com o seguinte:

"Coreia do Norte"é similar a "China"mais do que "China"é similar a "Coreia do Norte".

Ou ainda,

"Bayes"está para "Matemática"mais do que "Matemática"está para "Bayes".

Nesta seção, apresentaremos a noção de divergência, que é uma medida de dissimilaridade entre dois pontos de um modelo estatístico S. Veremos também que uma divergência induz de maneira natural uma métrica Riemanniana sobre S, além de duas conexões afins duais.

Definição 34. *Sejam S uma variedade diferenciável e considere dois pontos $p, q \in S$. Uma divergência é uma função suave $D : S \times S \to \mathbb{R}$ satisfazendo os seguintes itens:*

(i) $D[p:q] \geq 0$;

(ii) $D[p:q] = 0$ *se e somente se* $p = q$;

(iii) *Se* p *e* q *são próximos (isto é,* p *e* q *estão num mesmo sistema de coordenadas locais, e tais coordenadas* ξ_p *e* $\xi_q = \xi_p + d\xi$ *estão próximas), então vale o seguinte:*

$$D[\xi_p : \xi_p + d\xi] = \frac{1}{2} g_{ij}(\xi_p) d\xi^i d\xi^j + o(||d\xi||^2), \tag{4.1}$$

onde a matriz $G = [g_{ij}]$ *é simétrica e positiva-definida.*

Em todo este texto, estamos sempre supondo que a variedade diferenciável S é parametrizada globalmente. Isso simplifica a definição de p e q serem pontos próximos. Além disso, frequentemente escreveremos $D[p:q] = D[\xi_p : \xi_q]$, referenciando pontos de S com suas coordenadas correspondentes. Para mais detalhes sobre como se definir uma divergência D em uma variedade que não dispõe de uma parametrização global, veja [9].

Definiremos as derivadas parciais em relação às primeiras e segundas coordenadas, além das coordenadas mistas, do seguinte modo:

$$\begin{aligned}
D[p : (\partial_i)_q] &= \partial_{\xi_q^i} D[\xi_p : \xi_q]. \\
D[(\partial_i)_p : q] &= \partial_{\xi_p^i} D[\xi_p : \xi_q] \\
D[(\partial_i)_p : (\partial_j)_q] &= \partial_{\xi_p^i} \partial_{\xi_q^j} D[\xi_p : \xi_q] \\
D[(\partial_i \partial_j)_p : (\partial_k)_q] &= \partial_{\xi_p^i} \partial_{\xi_p^j} \partial_{\xi_q^k} D[\xi_p : \xi_q],
\end{aligned}$$

e assim por diante.

Observe que o ponto de mínimo de D é atingido quando $p = q$. Portanto, vale que

$$D[p : (\partial_i)_p] = D[(\partial_i)_p : p] = 0. \tag{4.2}$$

Isto também se verifica derivando a expressão (4.1). De fato, $D[p : (\partial_i)_p] = g_{ij}(p) d\xi^j \,|_{q=p} = 0$. Pela linearidade da derivada, vale, de uma forma mais geral, que a derivada em relação ao campo de vetores $X = X^i \partial_i$, satisfaz:

$$D[p : X(p)] = X^i(p) D[(\partial_i)_p : p] = 0.$$

Vamos definir as seguintes funções:

$$\begin{aligned}
D[\cdot : \partial_i] &: p \mapsto D[p : (\partial_i)_p]; \\
D[\partial_i : \partial_j] &: p \mapsto D[(\partial_i)_p : (\partial_j)_p]; \\
D[\partial_i \partial_j : \partial_k] &: p \mapsto D[(\partial_i \partial_j)_p : (\partial_k)_p].
\end{aligned}$$

Observe que, derivando duas vezes a expressão (4.2), obtem-se

$$D[p : (\partial_i \partial_j)_p] = g_{ij}(p). \tag{4.3}$$

Como g_{ij} é simétrica e positiva-definida, tem-se de (4.3) e da proposição abaixo, que D define de modo natural uma única métrica Riemanniana sobre S.

$$g_{ij}^{(D)} = D[\cdot : \partial_i \partial_j]. \tag{4.4}$$

Proposição 4.1. *Seja D uma divergência em uma variedade diferenciável S. Então,*

$$g_{ij}^{(D)} = D[\,\cdot\,:\partial_i\partial_j] = D[\partial_i\partial_j:\,\cdot\,] = -D[\partial_i:\partial_j]. \tag{4.5}$$

Demonstração. Segue diretamente derivando duas vezes a expressão (4.1) e usando (4.2). □

A expansão de Taylor de D de ordem 3 é dada por

$$D[p:q] = \frac{1}{2}g_{ij}^{(D)}(p)d\xi^i d\xi^j + \frac{1}{6}h_{ijk}^{(D)}(p)d\xi^i d\xi^j d\xi^k + o(||d\xi||^3), \tag{4.6}$$

onde $h_{ijk} = D[\partial_i\partial_j\partial_k:\,\cdot\,]$. Observe que uma divergência define uma conexão simétrica $\nabla^{(D)}$ por

$$\Gamma_{ij,k}^{(D)} = \langle\nabla_{\partial_i}^{(D)}\partial_j, \partial_k\rangle = -D[\partial_i\partial_j:\partial_k]. \tag{4.7}$$

Não há dificuldade em se verificar que (4.7) define de fato uma conexão afim (vide Definição 23 no Capítulo 3.1).

Proposição 4.2. *A divergencia dual $D^*[p:q] := D[q:p]$ induz a mesma metrica sobre S. Além disso, a conexão afim de D^* coincide com a conexão dual de D, ou seja $\nabla^{(D^*)} = (\nabla^{(D)})^*$.*

Demonstração. Pela Proposição 4.1, temos que $g_{ij}^* = D^*[\cdot:\partial_i\partial_j] = D[\,\partial_i\partial_j:\,\cdot\,] = g_{ij}$. Agora, derivando g,

$$\begin{aligned}\partial_k g_{ij} &= \partial_k(-D[\partial_i:\partial_j]) = -D[\partial_k\partial_i:\partial_j] - D[\partial_i:\partial_k\partial_j] = \langle\nabla_{\partial_k}\partial_i,\partial_j\rangle + \langle\partial_i,\nabla_{\partial_k}^*\partial_j\rangle\\ &= \Gamma_{ki,j} + \Gamma_{kj,i}^*.\end{aligned}$$

A proposição está provada. □

Note também que a função $h_{ijk}^D(p)$, que aparece na expansão de Taylor (4.6), também pode ser usada para se definir a conexão afim $\nabla^{(D)}$. De fato,

$$h_{ijk}^D = D[\partial_i\partial_j\partial_k:p] = \partial_k D[\partial_i:\partial_j] - D[\partial_j\partial_k:\partial_i] = -\partial_k g_{ij} + \Gamma_{jk,i}^D,$$

donde $\Gamma_{jk,i}^D = h_{ijk}^D + \partial_k g_{ij}$. Como $g_{ij}^D = g_{ij}^{D^*}$, concluimos que as expansões de Taylor de D e D^* diferem exatamente na derivada de ordem 3, a menos que a conexão $\nabla^{(D)}$ seja Riemanniana. O ponto importante é que reciproca também é verdadeira

Proposição 4.3. *Dados uma métrica Riemanniana g_{ij} e uma conexão afim simetrica ∇, podemos definir localmente uma divergencia D que satisfaça $g = g^{(D)}$ e $\nabla = \nabla^{(D)}$.*

Demonstração. Basta definir, em termos de cartas locais,

$$D[p:q] = \frac{1}{2}g_{ij}d\xi^i d\xi^j + \frac{1}{6}h_{ijk}d\xi^i d\xi^j d\xi^k,$$

onde $h_{ijk} = D[\partial_i\partial_j\partial_k:p] = -\partial_k g_{ij} + \Gamma_{jk,i}^D$. □

Para mais detalhes sobre como a tupla (g,∇,∇^*) é induzida por uma divergência no sentido global, veja [13].

Exemplo 24. *(Divergência Euclideana) A divergência Euclideana em uma variedade S é dada por*

$$D[p:q] = \frac{1}{2}\sum_i (\xi^i(p) - \xi^i(q))^2$$

Neste caso, D é simétrica e

$$D[(\partial_i)_p : q] = \xi^i(p) - \xi^i(q) \quad e \quad D[\partial_i \partial_j \,||\, \cdot\,] = \delta^i_j. \tag{4.8}$$

Isto é, $G = [g_{ij}]$ é a matriz identidade.

Exemplo 25. *(Divergência de Kullback-Leibler) A divergência de Kullback-Leibler, definida sobre o espaço das distribuições de probabilidade $P(\chi)$, é dada por*

$$D_{KL}[p:q] = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

Note que $D_{KL} \geq 0$, pois, usando que $-\log$ é uma função convexa, pela desigualdade de Jensen, vale que

$$\int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx = -\int p(x) \log\left(\frac{q(x)}{p(x)}\right) dx \geq -\log\left(\int p(x)\frac{q(x)}{p(x)} dx\right) = 0.$$

Considere um modelo estatístico $S \subset P(\chi)$, veremos que a divergencia de Kullback-Leibler induz a metrica de Fisher sobre S. De fato, considere pontos p e q sobre S de parametros ξ_p e ξ_q, respectivamente. Temos

$$D_{KL}[p : (\partial_i)_q] = -\int p(x)\frac{\partial_i q(x)}{q(x)} dx,$$

logo

$$D_{KL}[(\partial_j)_p : (\partial_i)_q] = -\int \partial_j p(x)\frac{\partial_i q(x)}{q(x)} dx. \tag{4.9}$$

Aplicando em $p = q$:

$$\begin{aligned} -D_{KL}[\partial_j : \partial_i] &= \int \partial_j p(x)\frac{\partial_i p(x)}{p(x)} dx = \int p(x)\partial_j \log p(x) \partial_i \log p(x) dx \\ &= E_p[\partial_i \log p \partial_j \log p]. \end{aligned}$$

Donde $g_{ij} = -D_{KL}[\partial_i : \partial_j]$ coincide com a metrica de Fisher em S.

Sobre a conexão induzida por D_{KL}, veremos que $\nabla^{(D_{KL})}$ coincide com a m-conexão. Com efeito, derivando (4.9) na variável p,

$$D_{KL}[(\partial_k\partial_i)_p : (\partial_j)_q] = -\int (\partial_k\partial_i p(x))\frac{\partial_j q(x)}{q(x)} dx. \tag{4.10}$$

Logo, fazendo $p = q$, os símbolos de Christoffell da conexão induzida são dados por

$$\Gamma^{(D)}_{ij,k} = -D[\partial_i\partial_j : \partial_k] = \int \partial_i\partial_j p(x)\frac{\partial_k p(x)}{p(x)} dx. \tag{4.11}$$

Como

$$\partial_i\partial_j p(x) = \{\partial_i\partial_j \ln p(x) + \partial_j \ln p(x)\partial_i \ln p(x)\}p(x), \tag{4.12}$$

escrevendo $\ell_\xi = \ln p_\xi$, vale que

$$\Gamma^{(D)}_{ij,k} = E_\xi[(\partial_i\partial_j\ell_\xi + \partial_i\ell_\xi\partial_j\ell_\xi)\partial_k\ell_\xi] = \Gamma^{(-1)}_{ij,k}. \tag{4.13}$$

Em particular, a e-conexão coincide com a conexão dual $(\nabla^D)^* = \nabla^{D^*}$. Veremos mais adiante que, sobre famílias exponenciais, a divergencia dual da divergência de Kullback-Leibler é a divergência de Bregman proveniente da função cumulante.

Outra característica importante é a assimetria da divergência de Kullback-Leibler. Consideremos $p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ (distribuição normal padrão) e $q(x) = \frac{1}{\pi(1+x^2)}$ (distribuição de Cauchy). Então,

$$\begin{aligned} D_{KL}[p:q] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \log\left(\frac{\pi e^{-\frac{x^2}{2}}(1+x^2)}{\sqrt{2\pi}}\right) dx \\ &= \log\sqrt{\frac{\pi}{2}} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(-\frac{x^2}{2} + \log(1+x^2)\right)e^{-\frac{x^2}{2}}dx. \end{aligned}$$

Note que $-x^2 \le \log(1+x^2) \le x^2$, portanto

$$|\, -\frac{x^2}{2} + \log(1+x^2)\,| \le \frac{x^2}{2} \tag{4.14}$$

De onde segue que $D_{KL} < \infty$. Por outro lado,

$$\begin{aligned} D_{KL}[q:p] &= \int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} \log(\left(\sqrt{\frac{2}{\pi}}\frac{1}{(1+x^2)e^{-\frac{x^2}{2}}}\right) dx \\ &= \log\sqrt{\frac{2}{\pi}} - 2\int_0^{\infty} \frac{1}{\pi(1+x^2)}\left(-\frac{x^2}{2} + \log(1+x^2)\right) \end{aligned}$$

Como $\frac{1}{\pi(1+x^2)}\left(\frac{-x^2}{2} + \log(1+x^2)\right) = -\frac{1}{2\pi} + o(1)$, quando $x \to \infty$, segue-se que a integral diverge, donde $D_{KL}[q:p] = +\infty$.

Alguns outros exemplos de divergências são o seguinte:

Exemplo 26. *(Divergencia de Itakura-Saito) A divergência de Itakura-Saito, definida sobre* $\mathbb{R}^n_{++} = \{x = (x^1,\ldots,x^n) \mid x^i > 0,$ *para todo* $i\}$*, é dada por*

$$D_{IS}[x:y] = \sum_{i=1}^{n}\left(\frac{x^i}{y^i} - \log\frac{x^i}{y^i} - 1\right).$$

Exemplo 27. *(I-divergência) A I-divergência, também definida sobre $\mathbb{R}^n_{++}$, é definida por*

$$D_I[x:y] = \sum_{i=1}^{n} \left(x^i \log \frac{x^i}{y^i} - x^i + y^i \right)$$

Exemplo 28. *(Divergência χ^2) Seja S_n o modelo finito. A divergência χ^2 é definida por*

$$D[p:q] = \frac{1}{2} \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i}$$

4.2 f-divergências

A f-divergência é uma classe de divergências introduzida por Csiszár [12]. A definição é a seguinte:

Definição 35. *Seja $f : [0,\infty) \to \mathbb{R}$ uma função convexa tal que $f(1) = 0$. Seja S um modelo estatístico sobre χ e sejam $p, q \in S$. Então a função $D_f : S \times S \to \mathbb{R}$ definida por*

$$D_f[p:q] = \int p(x) f\Big(\frac{q(x)}{p(x)}\Big) dx. \tag{4.15}$$

é chamada f-divergência sobre S.

Proposição 4.4. *D_f é uma divergência e induz em um modelo estatístico S a métrica de Fisher multiplicado por uma constante $c > 0$*

Demonstração. Note que $D_f[p:q] \geq 0$, pois pela desigualdade de Jensen, vale

$$\int p(x) f\Big(\frac{q(x)}{p(x)}\Big) dx \geq f\Big(\int p(x) \frac{q(x)}{p(x)} dx \Big) = f\Big(\int q(x) dx \Big) = f(1) = 0. \tag{4.16}$$

Além disso, vale a igualdade, se e somente se, $\frac{q(x)}{p(x)}$ é constante. Como $\int p(x)dx = 1 = \int q(x)dx$, segue que $k = 1$ e $p = q$.

Para mostrar que D_f é uma divergência e induz a métrica de Fisher, consideremos $p, q \in S$, de coordenadas ξ_p e ξ_q, respectivamente. Assim,

$$D[p:(\partial_i)_q] = \int p(x) f'\Big(\frac{q(x)}{p(x)}\Big) \frac{\partial_i q(x)}{p(x)} dx = \int \partial_i q(x) f'\left(\frac{q(x)}{p(x)}\right) dx$$
$$D[p:(\partial_j\partial_i)_q] = \int \partial_j\partial_i q(x) f'\Big(\frac{q(x)}{p(x)}\Big) dx + \int \partial_i q(x) f''\Big(\frac{q(x)}{p(x)} \frac{\partial_j q(x)}{p(x)}\Big) dx.$$

Aplicando em $p = q$ e usando que $f(1) = 0$, temos

$$D[p:(\partial_j\partial_i)_p] = f''(1) \int \frac{1}{p(x)} \partial_j p_\xi(x) \partial_i p_\xi(x) = f''(1) \int p(x) \partial_i \log p_\xi(x) \partial_j \log p_\xi(x) dx.$$

Que é a expressão da métrica de Fisher escalada pela constante $c = f''(1) > 0$, pois f é convexa. Assim, a expansão de Taylor de grau 2 de D_f, é dada por

$$D_f[\xi_p : \xi_p + d\xi] = \frac{1}{2} f''(1) g_{ij}(\xi_p) d\xi^i d\xi^j + O(|d\xi|^3),$$

o que mostra a condição (iii) da definição (34). □

No capítulo 2, vimos o teorema da monotonicidade. Vamos mostrar este resultado segue-se de um fato mais geral, que é a monotonicidade da f-divergência sobre transição de probabilidades. Mais especificamente, seja $\{k(y|x) \geq 0 \mid x \in \mathscr{X}, y \in \mathscr{Y}\}$ uma transição de distribuições de probabilidades, ou seja, $k(y \mid x) \geq 0$ e $\int k(y|x)dy = 1$ para todo $x \in \mathscr{X}$. Como exemplo, seja $F : \mathscr{X} \to \mathscr{Y}$ uma estatística e defina $k(y \mid x) = \delta(y - F(x))$. Note que $k(y|x)$ leva distribuições de probabilidades sobre $\mathscr{X}$ em distribuições de probabilidade sobre $\mathscr{Y}$. Para isso, dado $p(x) \in \mathscr{P}(\mathscr{X})$, defina $p(y|x) = k(y|x)$, assim $p_k(y) = \int_{\mathscr{X}} k(y|x)p(x)dx$. Se $S = \{p(x;\xi)\}$ é um modelo estatístico sobre $\mathscr{X}$, teremos (induzido pela transição de probabilidades $\{k(y|x)\}$) um modelo estatístico $S_k = \{p_k(y;\xi)\}$ sobre $\mathscr{Y}$. Primeiro, vamos provar a seguinte proposição:

Proposição 4.5. *Nas condições acima, temos o seguinte:*

$$D_{KL}[p(x) : q(x)] = D_{KL}[p_k(y) : q_k(y)] + E_{p_k(y)}\Big[D_{KL}[p(x \mid y) : q(x \mid y)]\Big]. \tag{4.17}$$

Demonstração. Como $\frac{p(x)}{q(x)} = \frac{q(x|y)q_k(y)}{k(y|x)} \frac{k(y|x)}{p(x|y)p_k(y)} = \frac{p(x|y)p_k(y)}{q(x|y)q_k(y)}$, temos que $\frac{p(x|y)}{q(x|y)} = \frac{p(x)}{q(x)} \frac{q_k(y)}{p_k(y)}$. Assim,

$$\begin{aligned}
E_{p_k(y)}\Big[D_{KL}[p(x \mid y) : q(x \mid y)]\Big] &= \int p_k(y) \int p(x \mid y) \log\Big(\frac{p(x \mid y)}{q(x \mid y)}\Big) dxdy \\
&= \iint p_k(y)p(x \mid y) \log\Big(\frac{p(x)}{q(x)} \frac{q_k(y)}{p_k(y)}\Big) dxdy \\
&= \iint p_k(y)p(x \mid y) \log\Big(\frac{p(x)}{q(x)}\Big) + \iint p_k(y)p(x \mid y) \log\Big(\frac{q_k(y)}{p_k(y)}\Big) \\
&= \iint p_k(y)p(x \mid y) \log\Big(\frac{p(x)}{q(x)}\Big) + \iint p(x)p(y \mid x) \log\Big(\frac{q_k(y)}{p_k(y)}\Big) \\
&= \int p(x) \log\Big(\frac{p(x)}{q(x)}\Big) + \int p_k(y) \log\Big(\frac{q_k(y)}{p_k(y)}\Big) \\
&= D_{KL}[p(x) : q(x)] - D_{KL}[p_k(y) : q_k(y)].
\end{aligned}$$

Proposição 4.5 está provada. □

Pela Proposição 4.1, derivando (4.17) duas vezes em relação a q e, depois, tomando $q = p$, temos também o seguinte resultado, que generaliza Teorema 2.6,

Teorema 4.6. *Sejam $S = \{p(x;\xi)\}$ um modelo estatístico e $S_k = \{p_k(y;\xi)\}$ o modelo induzido pela transição de probabilidade $\{k(y|x)\}$. As métricas de Fisher G, G_k de S e S_k, respectivamente, satisfazem: $G \succeq G_k$. Mais geralmente,*

$$g_{ij}(\xi) = g^k_{ij}(\xi) + E_{p_k(y)}[g_{ij}(\xi \mid y)], \tag{4.18}$$

onde $g_{ij}(\xi \mid y) = E\left[\partial_i \log p(x|y;\xi)\partial_j \log p(x|y;\xi) \mid y\right]$ *é a matriz de informação de Fisher do modelo estatístico* $S_y = \{p(x \mid y;\xi)\}$. *Assim, a igualdade* $G = G_k$ *vale se, e somente se, as probabilidades condicionais* $p(x|y;\xi)$ *não dependem do parâmetro* ξ.

Uma das principais propriedades das f-divergências é monotonicidade. Mais especificamente, provamos o seguinte:

Teorema 4.7 (Monotonicidade)**.** *Seja* S *um modelo estatístico e* $\{k(y|x) \geq 0; x \in \chi, y \in Y\}$ *uma transição de distribuições de probabilidades. Então vale que*

$$D_f[p:q] \geq D_f[p_k:q_k]. \tag{4.19}$$

A igualdade vale se e somente se as probabilidades condicionais $p(x|y;\xi)$ *não dependem de* ξ.

Demonstração. Sejam $p, q \in S$. Então

$$\begin{aligned} D_f[p:q] &= \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx = \int \left(\int p(x \mid y) p_k(y) dy\right) f\left(\frac{q(x)}{p(x)}\right) dx \\ &= \iint p(x \mid y) p_k(y) f\left(\frac{q(x)}{p(x)}\right) dy dx = \int p_k(y) \int p(x \mid y) f\left(\frac{q(x)}{p(x)}\right) dx dy \\ &\geq \int p_k(y) f\left(\int p(x \mid y) \frac{q(x)}{p(x)} dx\right) dy, \end{aligned} \tag{4.20}$$

pela desigualdade de Jensen aplicada à distribuição $p(x \mid y)$. Como

$$\frac{q(x)}{p(x)} = \frac{q(x \mid y) q_k(y)}{p(x \mid y) p_k(y)}, \tag{4.21}$$

temos que

$$\begin{aligned} D_f[p:q] &\geq \int p_k(y) f\left(\int p(x \mid y) \frac{q(x)}{p(x)} dx\right) dy = \int p_k(y) f\left(\int q(x \mid y) \frac{q_k(y)}{p_k(y)} dx\right) \\ &= \int p_k(y) f\left(\frac{q_k(y)}{p_k(y)}\right) dy = D_f[p_k:q_k]. \end{aligned}$$

Usando (4.21), note que a igualdade vale se, e somente se $\frac{q(x|y)}{p(x|y)} = C(y)$. Como $\int p(x \mid y) dx = 1 = \int q(x \mid y) dx$, segue que $p(x \mid y) = q(x \mid y)$. Pela arbitrariedade de $p, q \in S$, temos que a probabilidade condicional $p(x|y;\xi)$ não depende do parâmetro ξ. O teorema está provado. □

Uma questão bastante interessante (conjecturada em alguns textos) é sobre a recíproca: será que divergencias que satisfaçam o teorema da monotonicidade são f-divergencias?

Mais adiante, em Teorema 5.4 (Seção 5.1), voltaremos à demonstração deste teorema e mostraremos que a diferença em (4.19) é dada por uma informação de Bregman condicional esperada. Divergência e informação de Bregman serão os assuntos das próximas seções.

Agora, vamos provar a propriedade de convexidade da métrica de Fisher (como havíamos prometido em (2.17), como consequência da f-divergencia. Sejam $S_1 = \{p_1(x;\xi) \mid x \in \chi\}_{\xi \in E}$ e

$S_2 = \{p_2(x;\xi) \mid x \in \chi\}_{\xi \in E}$ dois modelos estatisticos possuindo em comum os mesmos espaços amostral χ e de parâmetros E. Fixado $\lambda \in [0,1]$, segue-se que

$$S_\lambda = \{p_\lambda(x;\xi) = \lambda p_1(x;\xi) + (1-\lambda)p_2(x;\xi)\},$$

define um modelo estatístico. Temos o seguinte resultado:

Teorema 4.8 (Convexidade)**.** *Na notação acima, a f-divergência é convexa no seguinte sentido:*

$$D_f[p_\lambda(x;\xi) : q_\lambda(x;\xi)] \leq \lambda D_f[p_1 : q_1] + (1-\lambda)D_f[p_2 : q_2],$$

para todo $p_1, q_1 \in S_1$ e $p_2, q_2 \in S_2$. Além disso, as matrizes de informação de Fisher G_1, G_2, G_λ, de S_1, S_2 e S_λ, respectivamente, satisfazem a condição de convexidade:

$$G_\lambda(\xi) \preceq \lambda G_1(\xi) + (1-\lambda)G_2(\xi).$$

Demonstração. Escreva

$$\begin{aligned}\frac{q_\lambda}{p_\lambda} &= \frac{\lambda q_1 + (1-\lambda)q_2}{\lambda p_1 + (1-\lambda)p_2} \\ &= \frac{\lambda p_1}{\lambda p_1 + (1-\lambda)p_2}\frac{q_1}{p_1} + \frac{(1-\lambda)p_2}{\lambda p_1 + (1-\lambda)p_2}\frac{q_2}{p_2} \\ &= \frac{\lambda p_1}{p_\lambda}\frac{q_1}{p_1} + \frac{(1-\lambda)p_2}{p_\lambda}\frac{q_2}{p_2}.\end{aligned}$$

Como f é convexa e $\frac{\lambda p_1}{p_\lambda} + \frac{(1-\lambda)p_2}{p_\lambda} = 1$, temos

$$f\left(\frac{q_\lambda}{p_\lambda}\right) \leq \frac{\lambda p_1}{p_\lambda} f\left(\frac{q_1}{p_1}\right) + \frac{(1-\lambda)p_2}{p_\lambda} f\left(\frac{q_2}{p_2}\right).$$

Assim, multiplicando ambos lados por p_λ e integrando, temos:

$$\begin{aligned}D_f[p_\lambda : q_\lambda] &= \int p_\lambda f\left(\frac{q_\lambda}{p_\lambda}\right) \\ &\leq \lambda \int p_1 f(\left(\frac{q_1}{p_1}\right) + (1-\lambda)\int p_2 f\left(\frac{q_2}{p_2}\right) \\ &= \lambda D_f[p_1 : q_1] + (1-\lambda)D_f[p_2 : q_2]. \qquad (4.22)\end{aligned}$$

A primeira parte do teorema está provada.

Agora, vamos provar a convexidade para a matriz de Fisher. De fato, sabemos que:

$$\begin{aligned}D_f[p_\lambda : q_\lambda]|_{q_\lambda = p_\lambda} &= 0 \\ D_f[p_\lambda : \partial_i]|_{q_\lambda = p_\lambda} &= 0 \\ D_f[p_\lambda : \partial_i\partial_j]|_{q_\lambda = p_\lambda} &= f''(1)g_{ij}(p_\lambda).\end{aligned}$$

Por (4.22), segue-se que a derivada segunda respeita a desigualdade (visto que as derivadas anteriores se anulam quando $q_\lambda = p_\lambda$). Como $f''(1) > 0$, temos que as matrizes de Fisher,

$$G_\lambda(\xi) \preceq \lambda G_1(\xi) + (1-\lambda)G_2(\xi).$$

O teorema está provado. □

Proposição 4.9. *Seja D_f uma f-divergência. Então vale que*

(i) $D_f = D_g$, *onde* $g(u) = f(u) + c(u-1),\ c \in \mathbb{R}$.

(ii) $D_f^* = D_{f^*},$ *onde* $f^*(u) = uf\left(\frac{1}{u}\right)$

Demonstração. (i) Como

$$\int p(x)c\left(\frac{q(x)}{p(x)} - 1\right)dx = c\left\{\int q(x)dx - \int p(x)dx\right\} = 0, \tag{4.23}$$

vale que

$$D_g(p \,||\, q) = \int p(x) f\left(\frac{p(x)}{q(x)}\right)dx = D_f(p \,||\, q). \tag{4.24}$$

(ii) Temos

$$\begin{aligned} D_{f^*} = \int p(x) f^*\left(\frac{q(x)}{p(x)}\right) &= \int p(x)\frac{q(x)}{p(x)} f\left(\frac{p(x)}{q(x)}\right) \\ &= \int q(x) f\left(\frac{p(x)}{q(x)}\right) = D_f^*(p \,||\, q). \end{aligned}$$

.

Exemplo 29. *(Divergência de Kullback-Leibler) A divergência de Kullback-Leibler é definida a partir de* $f(u) = -\log u$

Exemplo 30. *(α-divergência)*

$$f(u) = \frac{4}{1-\alpha^2}\left(1 - u^{\frac{1+\alpha}{2}}\right) - \frac{2}{1-\alpha}(u-1), \tag{4.25}$$

$$D_\alpha[p:q] = \frac{4}{1-\alpha^2}\int\left\{1 - p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}}\right\}dx. \tag{4.26}$$

Vale comentar que a α-divergência pertence tanto à classe das f-divergências como à classe das divergências de Bregman [14], a qual será apresentada a seguir.

4.3 Divergência de Bregman

Definição 36. *(***Divergência de Bregman***) Sejam $S = \{p_\xi\}$, $\xi \in E \subset \mathbb{R}^n$ aberto, uma variedade diferenciável globalmente parametrizada e $\psi : U \subset \mathbb{R}^n \to \mathbb{R}$ uma função estritamente convexa, definida sobre um aberto U contendo E. A função $D_\psi : S \times S \to \mathbb{R}$ definida por*

$$D_\psi[p:q] = D_\psi[\xi_p : \xi_q] = \psi(\xi_p) - \psi(\xi_q) - \nabla\psi(\xi_q)\cdot(\xi_p - \xi_q). \tag{4.27}$$

é chamada Divergência de Bregman.

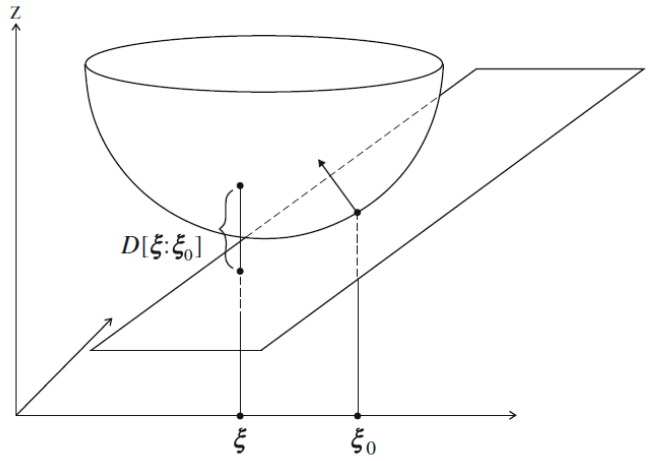

Figura 4.1 Divergência de Bregman de ξ a ξ_0.

A expressão em (4.27) representa a diferença entre os pontos $(\xi_p, \psi(\xi_p))$ do gráfico de ψ e (ξ_p, z) pertencente ao plano tangente de ψ que passa por $(\xi_q, \psi(\xi_q))$ (vide figura (4.1)). Vamos mostrar que D_ψ é uma divergência: como ψ é estritamente convexa, o gráfico de ψ está sempre acima de quaisquer de seus planos tangentes, donde $D_\psi[p:q] \geq 0$ e vale "$=0$" se, e somente se, $p = q$. Além disso, usando a expansão de Taylor de ordem 2 da função ψ no ponto ξ_q,

$$\psi(\xi_p) = \psi(\xi_q) + \partial_i\psi(\xi_q)(\xi_p^i - \xi_q^i) + \frac{1}{2}\partial_i\partial_j\psi(\xi_q)(\xi_p^i - \xi_q^i)(\xi_p^j - \xi_q^j) + O(\|\xi_p - \xi_q\|^3).$$

Usando (4.27), segue que

$$D_\psi[\xi_p : \xi_q] = \frac{1}{2}\partial_i\partial_j\psi(\xi_q)(\xi_p^i - \xi_q^i)(\xi_p^j - \xi_q^j) + O(\|\xi_p - \xi_q\|^3).$$

Como ψ é estritamente convexa, segue-se que $g_{ij}^{(D)} = \partial_i\partial_j\psi$ é simétrica e positiva-definida, o que mostra que D_ψ é de fato uma divergência.

Exemplo 31. *Seja $\psi : \mathbb{R}^n \to \mathbb{R}$ a seguinte função convexa*

$$\psi(\xi) = \frac{1}{2}\|\xi\|^2 = \frac{1}{2}\sum(\xi^i)^2.$$

O gradiente de ψ é dado por $\nabla\psi(\xi) = \xi$. Assim, a divergência de Bregman $D_\psi[\xi : \xi_0]$ é dada por

$$\begin{aligned} D_\psi[\xi : \xi_0] &= \psi(\xi) - \psi(\xi_0) - \nabla\psi(\xi_0)\cdot(\xi - \xi_0) \\ &= \frac{1}{2}[\|\xi\|^2 - \|\xi_0\|^2 - \xi_0\cdot(\xi - \xi_0)] \\ &= \frac{1}{2}\|\xi - \xi_0\|^2, \end{aligned}$$

que é a distância Euclideana apresentada no exemplo 24.

Exemplo 32. *Recordemos que o modelo finito $S_n = P(\chi)$, com $\chi = \{0,\dots,n\}$, é dado por*

$$S_n = \{p_\eta(x) = (1-\sum_i^n \eta_i)\delta(x-0) + \sum_i^n \eta_i \delta(x-i) \mid \eta_j > 0 \ \text{ e } \ \sum_i \eta_i < 1\}.$$

Afirmamos que a entropia negativa $\varphi(\eta) = -H[p_\eta] = \sum_{x=0}^n p_\eta(x)\log p_\eta(x)$ é uma função convexa. De fato, como

$$\varphi(\eta) = \sum_{i=0}^n p_i \log p_i = (1-\sum_{i=1}^n \eta_i)\log(1-\sum_{i=1}^n \eta_i) + \sum_{i=1}^n \eta_i \log \eta_i,$$

temos

$$\frac{\partial \varphi}{\partial \eta_j} = -\log(1-\sum_{i=1}^n \eta_i) - 1 + \log \eta_j + 1 = \log \frac{p_j}{p_0}$$
$$\frac{\partial^2 \varphi}{\partial \eta_i \partial \eta_j} = \frac{1}{1-\sum_{i=1}^n \eta_i}\delta_{ij} + \frac{1}{\eta_j}\delta_{ij} = (\frac{1}{p_0} + \frac{1}{p_j})\delta_{ij}.$$

Aqui, estamos usando que as coordenadas de um ponto $p \in P(\chi)$ são dadas por $\eta^p = (p_1,\dots,p_n)$. Assim, φ é uma função convexa. Além disso, a divergência de Bregman associada a φ entre dois pontos $p,q \in P(\chi)$, de coordenadas $\eta^p = (p_1,\dots,p_n)$ e $\eta^q = (q_1,\dots,q_n)$, respectivamente, é dada por

$$\begin{aligned} D_\varphi[p:q] &= \varphi(\eta^p) - \varphi(\eta^q) - \sum_{j=1}^n \frac{\partial \varphi}{\partial \eta_i^q}(\eta_j^p - \eta_j^q) \\ &= \sum_{i=0}^n p_i \log p_i - \sum_{i=0}^n q_i \log q_i - \sum_{j=0}^n \log(\frac{q_j}{q_0})(p_j - q_j) \\ &= \sum_i p_i \log(\frac{p_i}{q_i}) + \sum_{j=0}^n (p_j - q_j)\log q_0 = \sum_{i=0}^n p_i \log(\frac{p_i}{q_i}) \\ &= D_{KL}[p:q]. \end{aligned}$$

Isto é, D_φ coincide com a divergência de Kullback-Leibler.

Exemplo 33. *Seja $S = \{p_\theta(x) = e^{C(x)+\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta)}\}$ uma família exponencial. Afirmamos que a função cumulante $\psi(\theta)$ é convexa. De fato, primeiro, observemos que*

$$\partial_i \ln p_\theta(x) = F_i(x) - \partial_i \psi(\theta).$$

Daí, como $E_p[\partial_i \ln p_\theta] = 0$, temos $\partial_i \psi(\theta) = \int F_i(x) p_\theta(x) dx = E[F_i(X)]$. Assim,

$$\begin{aligned} \partial_i \partial_j \psi(\theta) &= \int F_i(x) \partial_j p_\theta(x) dx = E[F_i\ \partial_j \ln p_\theta] = E[F_i\ (F_j - \partial_j \psi(\theta))] \\ &= E[F_i F_j] - E[F_i]E[F_j] = Cov(F_i, F_j). \end{aligned}$$

Como a matriz de covariância é sempre simétrica e positiva-definida, segue-se ψ é convexa.

Agora vamos mostrar que a divergencia de Kullback-Leibler coincide com a divergencia de Bregman dual $D^*_\psi[p,q]$. De fato,

$$\begin{aligned} D_{KL}[q:p] &= \int q(x)\log\frac{q(x)}{p(x)}dx = \int q(x)[F_i(x)(\theta^i(q)-\theta^i(p))-\psi(q)+\psi(p)]dx \\ &= \psi(p)-\psi(q)+E_q[F_i](\theta^i(q)-\theta^i(p)) \\ &= \psi(p)-\psi(q)-\partial_i\psi(q)(\theta^i(p)-\theta^i(q)) \\ &= D_\psi[p:q]. \end{aligned} \tag{4.28}$$

4.4 Transformada de Legendre

Sejam S uma variedade globalmente parametrizada e $\xi : E \subset \mathbb{R}^n \to S$ um sistema global de coordenadas. Seja $\psi : U \subset \mathbb{R}^n \to \mathbb{R}$ uma função estritamente convexa definida num aberto U contendo E.

Lema 4.10. *O gradiente de ψ,*

$$\xi^* = \nabla\psi(\xi) \tag{4.29}$$

é um difeomorfismo de U sobre um aberto $V \subset \mathbb{R}^n$.

Demonstração. Primeiro, observe que a aplicação $\xi \mapsto \xi^*$ é injetiva. Para isso, note que o vetor normal ao gráfico de ψ no ponto $(\xi, \psi(\xi)$ é dado por $N(\xi) = (-\nabla\psi(\xi), 1) = (-\xi^*, 1)$. Como o gráfico de ψ é bordo de um corpo convexo (vide Figura 4.1) tem-se que $N(\xi)$ é injetiva, donde $\xi \mapsto \xi^*$ é injetiva. Como a matriz Hessiana de ψ é positiva-definida, temos que $\xi \mapsto \xi^*$ é um difeomorfismo local. Pela injetividade de ξ^*, segue-se que é um difeomorfismo. □

A transformada de Legendre $\psi^* : V \to \mathbb{R}$ é definida por

$$\psi^*(\xi^*) = \max_{\xi' \in U}\{\xi' \cdot \xi^* - \psi(\xi')\} \tag{4.30}$$

Proposição 4.11. *A transformada de Legendre $\psi^*(\xi^*)$ satisfaz*

$$\psi^*(\xi^*) = \xi \cdot \xi^* - \psi(\xi), \tag{4.31}$$

onde ξ é dada por $\xi^* = \nabla\psi(\xi)$. Além disso, ψ^* satisfaz os seguintes itens:

(i) $\xi = \nabla\psi^*(\xi^*)$;

(ii) A Hessiana $\nabla^2\psi^*(\xi^*) = (\nabla^2\psi(\xi))^{-1}$. *Em particular, ψ^* é estritamente convexa.*

Demonstração. Derivando a função $F(\xi') = \xi' \cdot \xi^* - \psi(\xi')$ temos $\frac{\partial F}{\partial \xi'_i}(\xi') = \xi^*_i - \psi_i(\xi')$. Assim, $\nabla F(\xi') = 0$ se, e somente se, $\nabla\psi(\xi') = \xi^*$. Por Lema 4.10, temos que $\xi' = \xi$. Alem disso, a Hessiana $F_{ij}(\xi') = -\psi_{ij}(\xi')$ é negativa-definida, donde $\xi = \arg\max_\xi F(\xi')$. Assim, $\psi^*(\xi^*) = \max_{\xi'} F(\xi') = F(\xi) = \xi \cdot \xi^* - \psi(\xi)$.

Agora, usando (4.29), a derivada parcial $\frac{\partial \xi_i^*}{\partial \xi^j} = \psi_{ij}(\xi)$, donde a matriz inversa $\frac{\partial \xi^i}{\partial \xi_j^*} = \psi^{ij}(\xi)$. Aqui, $(\psi^{ij}(\xi)) = (\psi_{ij}(\xi))^{-1}$ denota a matriz inversa de $\nabla^2\psi(\xi) = (\psi_{ij}(\xi))$. Assim, a derivada parcial de ψ^* é dada por

$$\begin{aligned}\frac{\partial \psi^*}{\partial \xi_j^*} &= \frac{\partial \xi^i}{\partial \xi_j^*}\xi_i^* + \xi^j - \frac{\partial \psi}{\partial \xi_j^*} = \frac{\partial \xi^i}{\partial \xi_j^*}\xi_i^* + \xi^j - \frac{\partial \psi}{\partial \xi^k}\frac{\partial \xi^k}{\partial \xi_j^*} \\ &= \psi^{ij}(\xi)\xi_i^* + \xi^j - \psi_k(\xi)\psi^{kj}(\xi) = \psi^{ij}(\xi)\xi_i^* + \xi^j - \xi_k^*\psi^{kj}(\xi) \\ &= \xi^j.\end{aligned}$$

Logo, o gradiente $\nabla\psi^*(\xi^*) = \xi$. Além disso, $\frac{\partial^2 \psi^*}{\partial \xi_i^* \partial \xi_j^*} = \frac{\partial \xi^j}{\partial \xi_i^*} = \psi^{ij}(\xi)$. Donde, a hessiana $\nabla^2\psi^*(\xi^*) = (\nabla^2\psi(\xi))^{-1}$. Proposição 4.11 está provada. □

Chamamos ψ e ψ^* de funções potenciais.

Proposição 4.12. *Seja D_ψ uma divergência de Bregman, a divergência de Bregman da função ψ^* é a divergência dual de D_ψ, isto é, $D_\psi^* = D_{\psi^*}$. Em particular, $D_{KL}[p:q] = D_{\psi^*}[p:q]$.*

Demonstração. De fato,

$$\begin{aligned}D_{\psi^*}[p:q] &= \psi^*(\xi_p^*) - \psi^*(\xi_q^*) - \nabla\psi^*(\xi_q^*)\cdot(\xi_p^* - \xi_q^*) \\ &= \xi_p\cdot\xi_p^* - \psi(\xi_p) - \xi_q\cdot\xi_q^* + \psi(\xi_q) - \xi_q\cdot(\xi_p^* - \xi_q^*) \\ &= \psi(\xi_q) - \psi(\xi_p) - \xi_p^*\cdot(\xi_q - \xi_p) \\ &= \psi(\xi_q) - \psi(\xi_p) - \nabla\psi(\xi_p)\cdot(\xi_q - \xi_p) \\ &= D_\psi[q:p] = D_\psi^*[p:q].\end{aligned}$$

Em particular, temos que $D_{KL}[p:q] = D_\psi[q:p] = D_{\psi^*}[p:q]$. □

Teorema 4.13. *Se D_ψ é uma divergência de Bregman, então vale que*

$$D_\psi[p:q] = \psi(\xi_p) + \psi^*(\xi_q^*) - \xi_p\xi_q^*.$$

Donde segue-se a desigualdade de Frenchel:

$$\psi(\xi_p) + \psi^*(\xi_q^*) - \xi_p\xi_q^* \geq 0,$$

e vale a igualdade se, e somente se, $p = q$.

Demonstração. Basta substituir

$$\psi^*(\xi_q^*) = \xi_q\xi_q^* - \psi(\xi_q) \quad \text{e} \quad \nabla\psi(\xi_q) = \xi_q^*$$

na equação (4.27). □

Teorema 4.14. *Sejam S uma variedade diferenciável globalmente parametrizada, e $\xi = [\xi^i] : E \subset \mathbb{R}^n \to S$ um sistema de coordenadas globais definida sobre um aberto E de $\mathbb{R}^n$. Seja $\psi : U \to \mathbb{R}$ uma função estritamente convexa definida sobre um aberto U contendo E. Sobre S, considere a métrica $\langle X, Y\rangle = -D_\psi[X, Y]$ e a conexão $\nabla_X Y = -D_\psi[XY : Z]$, com $X, Y \in \tau(S)$. Então, (S, g, ∇, ∇^*) é dualmente plana com respeito aos sistemas de coordenadas duais $\xi = [\xi^i]$ e $[\xi_i^* = \psi_i(\xi)]$. Em outras palavras, os seguintes itens são válidos:*

(i) $g_{ij}(\xi) = \langle \partial_{\xi^i}, \partial_{\xi^j}\rangle$ *e* $g_{ij}(\xi^*) = \langle \partial_{\xi_i^*}, \partial_{\xi_j^*}\rangle$ *satisfazem* $g^{ij}(\xi) = g_{ij}(\xi^*)$;

(ii) $\xi = [\xi^i]$ *é* ∇*-afim, isto é,* $\nabla_{\partial_i}\partial_j = 0$, *para todo* i, j;

(iii) $[\xi_i^* = \psi_i(\xi)]$ *é* ∇^**-afim. Além isso,* $\partial_{\xi_i^*} = g^{ij}(\xi)\partial_{\xi^j} = \partial^j$. *Assim,* $\nabla^*_{\partial^i}\partial^j = 0$.

Demonstração. Da Proposição 4.11, a métrica g satisfaz $g_{ij}(\xi) = -D_\psi[\partial_i : \partial_j] = \psi_{ij}(\xi) = \frac{\partial \xi_i^*}{\partial \xi^j}$. Do Lema 4.29, $\xi^* = \psi(\xi)$ é uma mundança de coordenadas. Assim, sendo $\partial_i = \partial_{\xi^i}$ e $\partial_i^* = \partial_{\xi_i^*}$ os campos coordenados correspondentes, temos

$$\partial_{\xi_i^*} = \partial_{\xi^k}\frac{\partial \xi^k}{\partial \xi_i^*} = g^{ik}\partial_k = \partial^i.$$

Logo, $g^*_{ij}(\xi^*) = \langle \partial_{\xi_i^*}, \partial_{\xi_j^*}\rangle = \langle \partial^i, \partial^j\rangle = g^{ij}(\xi)$.

Agora, do Teorema 4.13, dados $p, q \in S$, de coordenadas ξ_p, ξ_q, respectivamente, temos

$$D_\psi[p : q] = \psi(\xi_p) + \psi^*(\xi_q^*) - \xi_p \cdot \xi_q^*.$$

Assim,

$$\begin{aligned} D_\psi[\partial_{\xi_p^i} : q] &= \psi_i(\xi_p) - (\xi_q^*)_i \\ D_\psi[\partial_{\xi_p^i}\partial_{\xi_p^j} : q] &= \psi_{ij}(\xi_p) \\ \Gamma_{ij,k}(p) &= -D_\psi[\partial_{\xi_p^i}\partial_{\xi_p^j} : \partial_{\xi_q^k}]|_{q=p} = 0. \end{aligned}$$

Analogamente, prova-se que $\nabla^*_{\partial^i}\partial^j = 0$. O teorema está provado. □

Exemplo 34. *Considere uma família exponencial natural da forma*

$$S = \left\{ p_\theta(y) = e^{\theta^i y_i - \psi(\theta)} \right\}.$$

Por um argumento simples, toda família exponencial pode ser escrita dessa forma. Vimos no Exemplo 33 que ψ é uma função convexa, $\nabla\psi(\theta) = E_{p_\theta}[y]$ e a divergência de Bregman $D_\psi[p, q] = D_{KL}[q, p]$. Além disso, a métrica de Fisher $g_{ij} = \psi_{ij}(\theta) = -D_\psi[\partial_i, \partial_j]$. As coordenadas duais são dadas por $\theta^* = \nabla\psi(\theta) = E_p[y]$ e a transformada de Legendre de ψ é dada por

$$\begin{aligned} \psi^*(\theta^*) &= \theta \cdot \theta^* - \psi(\theta) \\ &= \theta \cdot E_{p_\theta}[y] - \psi(\theta) = E_{p_\theta}[\theta \cdot y - \psi(\theta)] \\ &= E_{p_\theta}[\ln p_\theta] = -H[p_\theta]. \end{aligned}$$

Ou seja, a transformada de Legendre coincide com a entropia negativa de p_θ. Em particular, a função cumulante ψ pode ser dada por $\psi(\theta) = \theta \cdot E_{p_\theta}[y] + H[p_\theta]$. O campos coordenados $\partial_i = \frac{\partial}{\partial \theta^i}$ e $\partial_j^* = \frac{\partial}{\partial \theta_j^*}$ satisfazem $g(\partial_i, \partial_j^*) = \delta_{ij}$. Do Teorema 4.14, a conexão induzida

$$\nabla_X Y = -D_\psi[XY : Z] = -D_{KL}[Z : XY]$$

é plana, logo $\nabla_X Y$ coincide com a e-conexão, $\nabla_X^{(1)} Y$, e também a conexão dual

$$\nabla_X^* Y = -D_\psi[Z : XY] = -D_{KL}[XY : Z]$$

coincide com a m-conexão, $\nabla_X^{(-1)} Y$.

4.5 O teorema Pitagoreano e o teorema da projeção

Teorema 4.15. *Sejam D_ψ uma divergência de Bregman e p, q, r in S. Então vale que*

$$D_\psi[p : q] + D_\psi[q : r] - D_\psi[p : r] = (\theta_p - \theta_q) \cdot (\eta_r - \eta_q),$$

onde $[\theta]$ e $[\eta]$ são sistemas de coordenadas duais em S.

Demonstração. Suponha que D_ψ é uma divergência de Bregman e seja φ a transformada de Legendre de ψ. Então,

$$D_\psi[p : q] = \psi(\theta_p) + \varphi(\eta_q) - \theta_p \cdot \eta_q$$
$$D_\psi[q : r] = \psi(\theta_q) + \varphi(\eta_r) - \theta_q \cdot \eta_r$$
$$-D_\psi[p : r] = -\psi(\theta_p) - \varphi(\eta_r) + \theta_p \eta_r.$$

Daí, somando-se as três equações acima,

$$D_\psi[p : q] + D_\psi[q : r] - D_\psi[p : r] = \psi(\theta_q) + \varphi(\eta_q) - \theta_p \cdot \eta_q - \theta_q \cdot \eta_r + \theta_p \cdot \eta_r.$$

Pelo Teorema (4.13), vale

$$\psi(\theta_q) + \varphi(\eta_q) = \theta_q \cdot \eta_q.$$

Logo,

$$D_\psi[p : q] + D_\psi[q : r] - D_\psi[p : r] = (\theta_p - \theta_q) \cdot (\eta_r - \eta_q).$$

□

Teorema 4.16. *(**Teorema Pitagoreano**) Sejam p, q e r três pontos de um espaço dualmente plano (S, g, ∇, ∇^*) induzido por uma divergência de Bregman D_ψ. Seja γ a ∇-geodésica ligando p a q e seja γ^* a ∇^*-geodésica ligado q a r. Se γ e γ^* se interceptam ortogonalmente em q. Então, vale o seguinte:*

$$D_\psi[p : q] + D_\psi[q : r] = D_\psi[p : r].$$

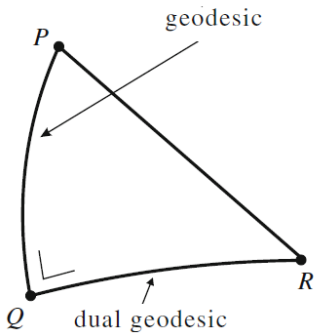

Figura 4.2 Teorema Pitagoreano

Demonstração. Como (S, g, ∇, ∇^*) é dualmente plana, as equações das geodésicas em relação a cada sistema de coordenada tem segunda derivada zero. Logo, podem ser parametrizadas como segmentos de retas:

$$\theta(t) = (1-t)\theta_p + t\theta_q;$$

$$\eta(t) = (1-t)\eta_q + t\eta_r.$$

Assim, os vetores tangentes são dados por

$$\gamma'(t) = \frac{d}{dt}\theta^i(t)\partial_i = (\theta^i_q - \theta^i_p)\partial_i,$$

$$\gamma^{*\prime}(t) = \frac{d}{dt}\eta_i(t)\partial^i = (\eta^r_i - \eta^q_i)\partial^i.$$

Donde, usando que as geodésicas intersectam-se ortogonalmente no ponto q, segue-se que

$$\begin{aligned} 0 = \langle \gamma'(1), \gamma^{*\prime}(0) \rangle &= \left\langle \frac{d}{dt}\theta^i(t)\partial_i \, , \, \frac{d}{dt}\eta_i(t)\partial^i \right\rangle = \left\langle (\theta^i_p - \theta^i_q)\partial_i \, , (\eta^q_i - \eta^r_i)\partial^i \right\rangle \\ &= (\theta_p - \theta_q)\cdot(\eta_r - \eta_q)\langle \partial_i, \partial^i \rangle = (\theta_p - \theta_q)\cdot(\eta_r - \eta_q). \end{aligned} \tag{4.32}$$

Logo, de (4.32) e usando Teorema 4.15, vale que

$$D_\psi[p:q] + D_\psi[q:r] = D_\psi[p:r],$$

como queríamos provar. □

Corolário 4.17. *(**Teorema da Projeção**) Seja (S, g, ∇, ∇^*) induzido por uma divergência de Bregman D_ψ e sejam $p \in S$ e M uma subvariedade em S. Se $q = \arg\min_{r\in M} D_\psi[p, r]$ (respectivamente, $q^* = \arg\min_{r\in M} D_\psi[r, p]$), então, a ∇-geodésica ligando p a q (respectivamente, a ∇-geodesica ligando p a q^*) é ortogonal a M. Além disso, se M é ∇^*-totalmente geodésica (resp. ∇-totalmente geodésica), então a recíproca também é verdadeira.*

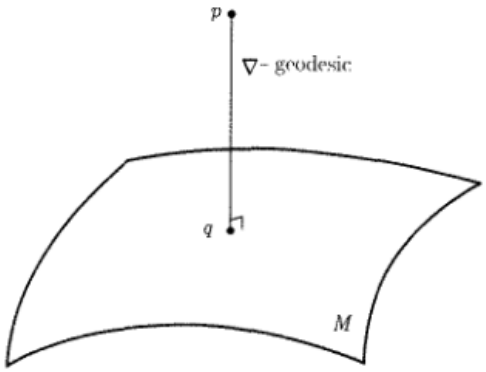

Figura 4.3 Teorema da Projeção

Demonstração. Suponha que $q = \arg\min_{r\in M} D_\psi[p,r]$ e seja $[u_a]$ um sistema de coordenadas para M. Então

$$\begin{aligned} 0 = D[p:(\partial_a)_q] &= D[p:(\partial_a\eta_i\partial^i)_q] = (\partial_a\eta_i)_q D[p:(\partial^i)_q] \\ &= (\partial_a\eta_i)_q(\theta^i_q-\theta^i_p) = (\partial_a\eta_i)_q(\theta^i_q-\theta^i_p)\langle\partial^i,\partial_i\rangle \\ &= \langle(\partial_a\eta_i\partial^i)_q,(\theta^i_q-\theta^i_p)\partial_i\rangle = \left\langle (\partial_a)_q\ ,\ \{\theta^i_q-\theta^i_p\}(\partial_i)_q \right\rangle, \end{aligned}$$

de onde segue que a ∇-geodésica que liga p a q é ortogonal a M. O caso em que $q^* = \arg\min_{r\in M} D_\psi[r,p]$ é totalmente análogo. Suponha agora que M seja ∇^*-totalmente geodésica e que a ∇-geodésica ligando p a q é ortogonal a M no ponto q. Pelo Teorema (4.16), vale que

$$D_\psi[p:q]+D_\psi[q:r] = D_\psi[p:r],$$

para todo r $\in$ M logo, $D_\psi[p:q] \le D_\psi[p:r]$ e, da equação acima, vale a igualdade se, e somente se $D_\psi[q:r]=0$, isto é, $q=r$. □

Definição 37. *Nas hipóteses do Teorema da Projeção (Corolário 4.17),* $q = \arg\min_{r\in M} D_\psi[p:r]$ *será chamado de* ∇*-projeção sobre* M.

Exemplo 35. *Se o ambiente S é uma família exponencial então a função cumulante $\psi(\theta)$ é convexa, donde define uma divergência de Bregman D_ψ. Sabemos que a divergência de Kullback-Leibler é a dual de D_ψ, ou seja, $D_{KL}[p:q] = D_\psi[q:p]$, para todo $p,q \in S$ (vide* (4.28) no Exemplo 33). Além disso, considere a tupla (S,g,∇^e,∇^m), onde g é a métrica de Fisher, $\nabla^{(e)} = \nabla^{(1)}$ é a e-conexão (ou 1-conexão) e $\nabla^{(m)}$ é a m-conexão (ou (-1)-conexão). Temos também que (S,g,∇^e,∇^m) é dualmente plana e induzida por D_ψ.

Agora, sejam $p \in S$ e M uma família exponencial curvada (ou, em outras palavras, uma subvariedade imersa em S), do teorema da projeção, segue-se que:

(i) e-projeção: $q^{(e)} = \arg\min_{r\in M} D_\psi[p,r] = \arg\min_{r\in M} D_{KL}[r,p]$;

(ii) m-projeção: $q^{(m)} = \arg\min_{r \in M} D_{\psi}[r, p] = \arg\min_{r \in M} D_{KL}[p, r]$.

Vale ressaltar que os pontos $q^{(e)}$ e $q^{(m)}$ podem, em geral, diferir. Esse exemplo será fundamental na nossa formulação do EM-algoritmo.

CAPÍTULO 5

Aplicações do Teorema Pitagoreano

Neste capítulo, vamos falar de algumas aplicações do teorema Pitagoreano (e do teorema da projeção). Na seção abaixo, falaremos da informação de Bregman, que é uma medida de dispersão baseada nas divergências de Bregman e que unifica, por exemplo, variância e informação mútua. Um ponto importante a se destacar nessa seção é que o erro em se aplicar o Desigualdade de Jensen é dado exatamente pela informação de Bregman.

5.1 Informação de Bregman

Definição 38. *Seja $\psi : U \to \mathbb{R}$ uma função convexa definida sobre um aberto convexo $U \subset \mathbb{R}^n$. Seja X uma variável aleatória cuja imagem $\mathscr{X}$ está contida em U. Seja D_ψ a divergência de Bregman induzida por ψ. A informação de Bregman é definida por*

$$I_\psi[X] = \min_\eta E[D_\psi[X : \eta]] = \min_\eta \int D_\psi[x : \eta] p(x) dx. \tag{5.1}$$

Proposição 5.1. *Existe um único minimizador de $I_\psi[X]$, dado por $\eta^* = E[X] = \int x p(x)$.*

Demonstração. Como U é convexa, temos que $\eta^* = E[X]$ pertence a U. Para $\eta \in U$, temos

$$\begin{aligned} E[D_\psi[X : \eta]] - E[D_\psi[X : \eta^*]] &= \int \Big(D_\psi[x : \eta] - D_\psi[x : \eta^*] \Big) p(x) \\ &= \psi(\eta^*) - \psi(\eta) + \int \Big(\nabla\psi(\eta^*)(x - \eta^*) - \nabla\psi(\eta)(x - \eta) \Big) p(x) \\ &= \psi(\eta^*) - \psi(\eta) + \nabla\psi(\eta^*)(E[X] - \eta^*) - \nabla\psi(\eta)(E[X] - \eta) \\ &= \psi(\eta^*) - \psi(\eta) - \nabla\psi(\eta)(\eta^* - \eta) \\ &= D_\psi[\eta^* : \eta] \geq 0. \end{aligned}$$

e a igualdade vale se, e somente se $\eta = \eta^*$. □

Um ponto importante a destacar é que $\eta^* = \arg\min_\eta E[D_\psi[X : \eta]] = E[X]$ independe de ψ. Vale a pena também ressaltar (mesmo sem demonstração) que a recíproca também é verdadeira, ou seja, se uma divergência D satisfaz $E[X] = \arg\min_\eta E[D[X : \eta]]$ então D é uma divergência de Bregman (veja Banerjee et al. [6]).

Exemplo 36. *A variância de uma variável aleatória X é a informação de Bregman relativa à função $\psi(z) = |z|^2$. De fato,*

$$I_\psi[X] = E[D_\psi[X : E[X]]] = E[|X - E[X]|^2] = V[X].$$

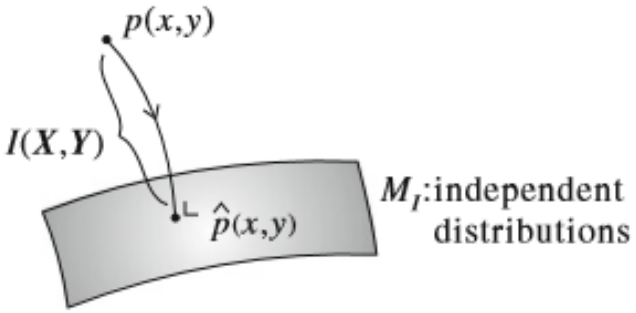

Figura 5.1 A Informação Mútua é a m-projeção de $p(x,y)$ em M

Exemplo 37 (Informação mútua)**.** *A informação mútua entre duas variáveis aleatórias X e Y, definida por*

$$I[X,Y] = D_{KL}[p(x,y) : p(x)p(y)].$$

Afirmamos que $I[X,Y]$ é uma informação de Bregman. De fato, primeiramente, considere a variável aleatória $Z : x \in \chi \mapsto Z_x = \{p(y \mid x)\}_{y\in\mathscr{Y}}$, com probabilidade $p(x)$. Em outras palavras, para cada $x \in \chi$, a variável aleatória Z assume o valor $Z_x = p(\cdot \mid x) \in P(\mathscr{Y})$ com probabilidade $p(x)$. A média $E[Z] = \int Z_x p(x) = \int p(\cdot \mid x)p(x) = p(\cdot) \in P(\mathscr{Y})$. Assim,

$$\begin{aligned} I[X,Y] = D_{KL}[p(x,y) : p(x)p(y)] &= \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dxdy \\ &= \int p(x)\int p(y \mid x)\log\frac{p(y \mid x)}{p(y)}\,dy\,dx = \int p(x)\,D_{KL}[p(y \mid x) : p(y)]\,dx \\ &= E[D_{KL}[Z : E[Z]] = I_{D_{KL}}[Z]. \end{aligned}$$

Proposição 5.2. *A informação mútua $I[X,Y]$ é a m-projeção de $p(x,y)$ sobre o espaço das distribuições de probabilidade independentes de sobre $\mathscr{X}\times\mathscr{Y}$, via a divergência de Kullback-Leibler. Em outras palavras,*

$$I[X : Y] = \min_{\substack{q(x,y)\in P(\mathscr{X}\times\mathscr{Y}) \\ q(x,y)=q(x)q(y)}} D_{KL}[p(x,y) : q(x,y)].$$

Demonstração. Considere a variedade $S = \{p(x,y)\}$, formada por todas as distribuições conjuntas de X e Y como sendo uma família exponencial. Seja M a subvariedade em S formada por todas as distribuições independentes de X e Y, isto é,

$$M = \{q \in S \mid q(x,y) = q(x)q(y)\},$$

onde $q(x), q(y)$ denotam as distribuições marginais sobre x e y respectivamente. Note que M é e-totalmente geodésica. De fato, sejam $q_1, q_2 \in M$ e $\lambda \in [0,1]$. A curva $q_\lambda(x,y) = q_1(x,y)^{1-\lambda}q_2(x,y)^\lambda$ é uma e-geodesica em S, pois $\nabla^{(e)}_{\frac{\partial}{\partial\lambda}}\frac{\partial}{\partial\lambda} = \frac{\partial^2}{\partial\lambda^2}\log q_\lambda = 0$. Além disso,

$q_\lambda \in M$, pois usando que $q_1(x,y) = q_1(x)q_1(y)$ e $q_2(x,y) = q_2(x)q_2(y)$, segue-se diretamente que $q_\lambda(x,y) = (q_1(x)^{1-\lambda}q_2(x)^\lambda)(q_1(y)^{1-\lambda}q_2(y)^\lambda)$. Donde, M é e-totalmente geodésica. Assim, do teorema da projeção, segue-se que $\hat{p} = \arg\min_{q\in M} D_{KL}[p:q]$ é atingido na m-projeção de p em M. Vamos mostrar que $\hat{p}(x,y) = p(x)p(y)$, ou seja, $\hat{p}$ é dado pelo produto das distribuições marginais de $p(x,y)$. De fato, como $D_{KL}[p:\hat{p}]$ é mínimo,

$$\int p(x,y)\log\frac{p(x,y)}{\hat{p}(x,y)}dxdy = D_{KL}[p:\hat{p}] \leq D_{KL}[p:q] = \int p(x,y)\log\frac{p(x,y)}{q(x,y)}dxdy,$$

para qualquer $q \in M$. Daí,

$$\int p(x,y)\log\hat{p}(x,y)dxdy \;\geq \int p(x,y)\log q(x,y)dxdy. \tag{5.2}$$

Mas $q(x,y) = q(x)q(y)$ e $\hat{p}(x,y) = \hat{p}(x)\hat{p}(y)$, logo

$$\begin{aligned}\int p(x,y)\log q(x,y)dxdy &= \int p(x,y)\log q(x)dxdy \;+ \int p(x,y)\log q(y)dxdy \\ &= \int p(x)\log q(x)dx + \int p(y)\log q(y)dy. \end{aligned}\tag{5.3}$$

Da mesma forma, a equação acima também vale para $\hat{p}$, ou seja,

$$\begin{aligned}\int p(x,y)\log\hat{p}(x,y)dxdy &= \int p(x,y)\log\hat{p}(x)dxdy \;+ \int p(x,y)\log\hat{p}(y)dxdy \\ &= \int p(x)\log\hat{p}(x)dx + \int p(y)\log\hat{p}(y)dy.\end{aligned}$$

Tome $q(x,y) = p(x)p(y) \in M$. Usando (5.3) e (5.2), temos que

$$\begin{aligned}0 &\geq \int p(x,y)\log q(x,y) - \int p(x,y)\log\hat{p}(x,y) \\ &= \int p(x)\log p(x) + \int p(y)\log p(y) - \int p(x)\log\hat{p}(x) - \int p(y)\log\hat{p}(y) \\ &= D_{KL}[p(x):\hat{p}(x)] + D_{KL}[p(y):\hat{p}(y)].\end{aligned}$$

Portanto, como $D_{KL} \geq 0$, segue-se que $\hat{p}(x) = p(x)$ e $\hat{p}(y) = p(y)$. Concluímos que $\hat{p} = \arg\min_{q\in M} D_{KL}[p:q]$ satisfaz

$$\begin{aligned}\min_{q(x,y)=q(x)q(y)} D_{KL}[p(x,y):q(x,y)] &= D_{KL}[p(x,y):\hat{p}(x,y)] = D_{KL}[p(x,y):p(x)p(y)] \\ &= I[X:Y],\end{aligned}$$

como queríamos provar. □

A desigualdade de Jensen diz que se $\psi : U \subset \mathbb{R}^n \to \mathbb{R}$ é uma função convexa definida sobre um aberto convexo U e X é uma variável aleatória cuja imagem χ está contida em U, então vale a seguinte desigualdade:

$$E[\psi(X)] \geq \psi(E[X]). \tag{5.4}$$

Além disso, vale a igualdade se, e somente se, X é constante. Veremos que a diferença $E[\psi(X)] - \psi(E[X])$ é exatamente a informação de Bregman $I_\psi[X]$.

Teorema 5.3. *Nas condições dadas acima, temos que* $I_\psi[X] = E[\psi(X)] - \psi(E[X])$*, de onde segue-se imediatamente a desigualdade de Jensen.*

Demonstração.

$$\begin{aligned} I_\psi[X] &= E[D_\psi[X : E[X]]] = \int D_\psi[x : E[X]]p(x)\,dx \\ &= \int \big(\psi(x) - \psi(E[X]) - \nabla\psi(E[X])\cdot(x - E[X])\big)p(x)\,dx \\ &= E[\psi(X)] - \psi(E[X]) - \nabla\psi(E[X])\cdot(E[X] - E[X]) \\ &= E[\psi(X)] - \psi(E[X]) \end{aligned}$$

e a igualdade vale se, e somente se, X é constante.

Como havíamos dito anteriormente, vamos voltar a prova do Teorema 4.7 a fim de provar que a diferença em (4.19) é dado por uma informação de Bregman.

Teorema 5.4 (Monotonicidade para f-divergências)**.** *Seja* $f : (0,\infty) \to \mathbb{R}$ *uma função convexa com* $f(1) = 0$ *e considere* D_f *como a* f*-divergência sobre um modelo estatístico* $S = \{p(x;\xi)\}$*. Considere uma transição de probabilidade* $\{k(y \mid x)\}$ *e seja* $S_k = \{p_k(y \mid \xi)\}$ *o modelo estatístico induzido por* k*. Temos então,*

$$D_f[p : q] = D_f[p_k : q_k] + E_{p_k(y)}\Big[I_f\big[\frac{q(x)}{p(x)} \mid y\big]\Big]. \tag{5.5}$$

onde $I_f[\cdot \mid y]$ *é a informação de Bregman induzida por* f *condicionada à* y*.*

Note que a f-divergência D_f não coincide com a divergência de Bregman BD_f. Por exemplo, se $f(z) = -\log(z)$, com $z > 0$, então $D_f[p : q]$, com $p, q \in P(\chi)$, coincide com a divergência de Kullback-Leibler $D_{KL}[p : q]$. Já a Divergência de Bregman $BD_f[z : w]$, com $z, w > 0$, coincide com a Divergência de Itakura-Saito $D_{IS}[z : w] = \frac{z}{w} - \log(\frac{z}{w}) - 1$.

Demonstração. Sejam $p, q \in S$. Então, usando o Teorema 5.3,

$$\begin{aligned} D_f[p : q] &= \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx = \int \left(\int p(x \mid y)p_k(y)dy\right) f\left(\frac{q(x)}{p(x)}\right) dx \\ &= \iint p(x \mid y)p_k(y) f\left(\frac{q(x)}{p(x)}\right) dydx = \int p_k(y) \int p(x \mid y) f\left(\frac{q(x)}{p(x)}\right) dxdy \\ &= \int p_k(y) f\left(\int p(x \mid y)\frac{q(x)}{p(x)}dx\right) dy + \int p_k(y) I_f[\frac{q(x)}{p(x)} \mid y], \end{aligned}$$

onde $I_f\big[Z[X] \mid y\big]$ denota a informação de Bregman da variável aleatória $Z[X]$ com probabilidade $p(x \mid y)$. Por definição, $p(y \mid x) = q(y \mid x) = k(y \mid x)$. Assim,

$$\frac{q(x)}{p(x)} = \frac{q(x \mid y)q_k(y)}{k(y \mid x)} \frac{k(y \mid x)}{p(x \mid y)p_k(y)} = \frac{q(x \mid y)q_k(y)}{p(x \mid y)p_k(y)}. \tag{5.6}$$

Logo,

$$\begin{aligned}\int p_k(y) f\left(\int p(x \mid y)\frac{q(x)}{p(x)}dx\right)dy &= \int p_k(y) f\left(\int q(x \mid y)\frac{q_k(y)}{p_k(y)}dx\right)\\ &= \int p_k(y) f\left(\frac{q_k(y)}{p_k(y)}\right)dy\\ &= D_f[p_k : q_k].\end{aligned}$$

Portanto,

$$D_f[p : q] = D_f[p_k : q_k] + E_{p_k}\Big[I_f[\frac{q(x)}{p(x)} \mid y]\Big].$$

O teorema está provado. □

Observação. *A respeito do Teorema 5.4, a informação de Bregman $I_f[\frac{q(x)}{p(x)} \mid y]$ pode ser escrita da seguinte forma:*

$$I_f[\frac{q(x)}{p(x)} \mid y] = E\Big[BD_f[\frac{q(x)}{p(x)} : \frac{q_k(y)}{p_k(y)}] \mid y\Big]. \tag{5.7}$$

De fato, note que a média

$$E[\frac{q(x)}{p(x)} \mid y] = \int \frac{q(x \mid y)q_k(y)}{p(x \mid y)p_k(y)}p(x \mid y)dx = \frac{q_k(y)}{p_k(y)}\int q(x \mid y)dx = \frac{q_k(y)}{p_k(y)}.$$

Assim, denotando por BD_f a divergência de Bregman induzida por f, temos

$$\begin{aligned}I_f[\frac{q(x)}{p(x)} \mid y] &= \int p(x \mid y)BD_f\Big[\frac{q(x)}{p(x)} : E[\frac{q(x)}{p(x)} \mid y]\Big] = \int p(x \mid y)BD_f\Big[\frac{q(x)}{p(x)} : \frac{q_k(y)}{p_k(y)}\Big]\\ &= E^x\Big[BD_f[\frac{q(x)}{p(x)} : \frac{q_k(y)}{p_k(y)}] \mid y\Big],\end{aligned}$$

*onde E^x denota a esperança na variável x. Portanto, (*5.5) se reescreve por

$$D_f[p : q] = D_f[p_k : q_k] + E^y\Big[E^x\Big[BD_f[\frac{q(x)}{p(x)} : \frac{q_k(y)}{p_k(y)}] \mid y\Big]\Big]. \tag{5.8}$$

Proposição 4.5 segue-se como Corolário do Teorema 5.4. De fato, considere novamente $f(z) = -\ln(z)$, com $z > 0$. Vimos que a f-divergência, $D_f[p : q] = D_{KL}[p : q]$ e a divergência de Bregman, $BD_f[z : w] = D_{IS}[z : w]$. Assim,

$$\begin{aligned}BD_f\Big[\frac{q(x)}{p(x)} : \frac{q_k(y)}{p_k(y)}\Big] &= \frac{q(x)}{p(x)}\frac{p_k(y)}{q_k(y)} - \ln\Big(\frac{q(x)}{p(x)}\frac{p_k(y)}{q_k(y)}\Big) - 1\\ &= \frac{q(x \mid y)}{p(x \mid y)} - \ln\Big(\frac{q(x \mid y)}{p(x \mid y)}\Big) - 1,\end{aligned}$$

donde,

$$\int p(x \mid y)BD_f\Big[\frac{q(x)}{p(x)} : \frac{q_k(y)}{p_k(y)}\Big] = -\int p(x \mid y)\ln\Big(\frac{q(x \mid y)}{p(x \mid y)}\Big) = D_{KL}[p(x \mid y) : q(x \mid y)].$$

Usando (5.7) e o Teorema 5.4, Proposição 4.5 segue-se. □

Teorema 5.5 (Princípio da máxima entropia)**.** *Seja M um modelo estatístico sobre um espaço amostral finito $\chi = \{0,1,\ldots,n\}$. Então, a distribuição em M que maximiza a entropia*

$$p = \arg\max_{q \in M} H[q]$$

coincide com a e-projeção da distribuição uniforme $p_0(x) = \frac{1}{n+1}$ sobre M.

Demonstração. Relembrando, a entropia $H[q] = \sum_x q(x)\log(1/q(x)) = -\sum_x q(x)\log(q(x))$. Assim,

$$\begin{aligned} D_{KL}[q : p_0] &= \sum_x q(x)\log(\frac{q(x)}{p_0(x)}) = \sum_x q(x)\log q(x) - \sum_x q(x)\log(1/(n+1)) \\ &= -H[q] + \log(n+1). \end{aligned}$$

Portanto, $\arg\max_{q\in M} H[q] = \arg\min_{q\in M} D_{KL}[q : p_0]$, que é a e-projeção de p_0 sobre M. □

Assim, uma interpretação geométrica da distribuição que maximiza a entropia no modelo estatístico é que ela possui a informação mais vaga o possível, no sentido de ser a distribuição no modelo estatístico mais próxima da distribuição uniforme.

Exemplo 38. *Seja $\chi = \{0,\ldots,n\}$ e considere o vetor aleatório $c = (c_1,\ldots,c_k) : \chi \to \mathbb{R}^k$. Fixado um vetor $a \in \mathbb{R}^k$, considere a subvariedade M dada por*

$$M_{n-k}(a) = \{q \in P(\chi) \mid E_q[c] = \sum_x q(x)c(x) = a\}.$$

Note que $M_{n-k}(a)$ é m-totalmente geodésica, visto que $M_{n-k}(a)$ é invariante por combinações lineares convexas. Logo, do teorema da projeção, existe uma única e-projeção da distribuição uniforme $p_0(x) = 1/(n+1)$ sobre $M_{n-k}(a)$ (veja Figura 5.2). Donde, a máxima entropia é dada de modo único.

5.2 Estimadores de máxima verossimilhança

Considere uma variável aleatória X cuja distribuição $p_X(x)$ pertença a um modelo estatístico $M = \{p_\xi\}$. Gostaríamos de estimar o parâmetro ξ. Para isto, considere N observações independentes geradas a partir da v.a. X. O estimador de máxima verossimilhança $\hat{\xi} = \hat{\xi}(x_1,\ldots,x_N)$ é definido por

$$\hat{\xi} = \arg\max_{\xi} \prod_{i=1}^{n} p_\xi(x_i).$$

Vale a seguinte interpretação geométrica de $\hat{\xi}$.

Proposição 5.6. *Sendo $\hat{\xi} = \hat{\xi}(x_1,\ldots,x_N)$ o estimador de máxima verossimilhança, a distribuição $p_{\hat{\xi}}(x) \in M$ é a m-projeção da distribuição empírica $\bar{q}(x) = \frac{1}{N}\sum_i \delta(x - x_i)$ sobre M.*

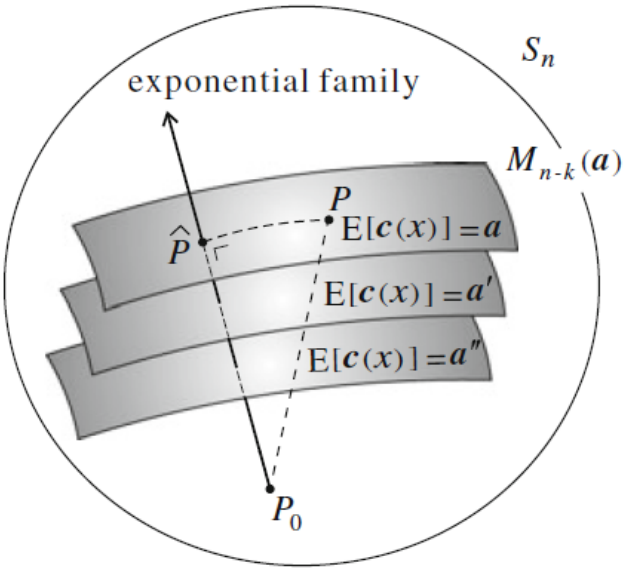

Figura 5.2 Princípio da Máxima Entropia

Em outras palavras, $p_{\hat{\xi}}(x) \in M$ é, dentre todas as distribuiçoes de M, aquela que é mais próxima da distribuição empírica $\bar{q}(x) = \frac{1}{N}\sum_i \delta(x - x_i)$, via a divergência de Kullback-Leibler.

Demonstração. Como log é uma função crescente, segue que

$$\hat{\xi} = \arg\max_{\xi} \log\Big(\prod_{i=1}^{n} p_\xi(x_i)\Big) = \arg\max_{\xi} \sum_{i=1}^{n} \log p_\xi(x_i).$$

Por outro lado, usando que $\int \delta(x - x_i) f(x) dx = f(x_i)$, temos que

$$\begin{aligned} D_{KL}[\bar{q}(x) : p_\xi(x)] &= \int \bar{q}(x) \log(\frac{\bar{q}(x)}{p_\xi(x)}) dx = \frac{1}{N}\sum_{i=1}^{N} \int \delta(x - x_i) \log(\frac{\bar{q}(x)}{p_\xi(x)}) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log(\frac{\bar{q}(x_i)}{p_\xi(x_i)}) = -\log N - \frac{1}{N}\sum_{i=1}^{N} \log(p_\xi(x_i)). \end{aligned}$$

Assim,

$$\arg\min_{\xi} D_{KL}[\bar{q} : p_\xi] = \arg\max_{\xi} \sum_{x} \log(p_\xi(x)) = \hat{\xi}(x_1, \ldots, x_N). \tag{5.9}$$

A proposição está provada. □

Exemplo 39. *Consideramos $S = \{p_\theta(x) = p(x \mid \theta)\}$ uma família exponencial da forma*

$$p(x \mid \theta) = e^{C(x) + \theta \cdot F(x) - \psi(\theta)},$$

com $\Theta = \{\theta\}$ (espaço de parâmetros) sendo um aberto de $\mathbb{R}^n$ e $F = (F_1, \ldots, F_n) : \chi \to \mathbb{R}^n$ uma aplicação sobre $\mathbb{R}^n$. Assim, dados $\{x_1, \ldots, x_N\}$ gerados a partir de uma v.a. X, cuja distribuição $p_X(x) \in S$, vamos calcular o estimador de máxima verossimilhança. Observe que

$$\frac{\partial}{\partial \theta^j} \sum_{i=1}^{N} \log(p(x_i \mid \theta)) = \sum_{i=1}^{N} F_j(x_i) - N \frac{\partial \psi(\theta)}{\partial \theta^j} = \sum_{i=1}^{N} F_j(x_i) - N\eta_j,$$

onde $\eta_j = \frac{\partial \psi(\theta)}{\partial \theta^j} = E_{p_\theta}[F_j(x)]$ são as coordenadas duais (ou coordenadas esperadas) de $S = \{p_\theta\}$. Assim, o estimador de máxima verossimilhança $\hat{\eta} = \hat{\eta}(x_1, \ldots, x_n)$, em termos das coordenadas duais, é dado por

$$\hat{\eta} = \frac{1}{N} \sum_{i=1}^{N} F_i(x).$$

Por exemplo, no modelo normal $p_\theta(x) = N(x \mid \mu, \sigma^2)$, temos que $F(x) = (x, x^2)$ e as coordenadas esperadas $\eta = (\eta_1, \eta_2)$ de $p_\theta = N(x \mid \mu, \sigma^2)$ são dadas por $\eta_1 = E_p[x] = \mu$ e $\eta_2 = E_p[x^2] = \mu^2 + \sigma^2$. Assim, o estimador de máxima verossimilhança $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2)$ é dado por

$$\hat{\eta}_1(x) = \frac{1}{N} \sum_{i=1}^{N} x_i = \hat{\mu}$$

$$\hat{\eta}_2(x) = \frac{1}{N} \sum_{i=1}^{N} x_i^2 = \hat{\mu}^2 + \hat{\sigma}^2,$$

sendo $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$ a média amostral e $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$ a variância amostral.

CAPÍTULO 6

O Algoritmo *EM*

6.1 Algoritmo *em*

Nesta seção, apresentamos o algoritmo *em*, cujo objetivo é minimizar uma divergência de Bregman entre duas subvariedades de um ambiente dualmente plano. Mostraremos que o algoritmo *EM* é um caso particular do algoritmo *em*, quando o ambiente é uma família exponencial e a divergência de Bregman escolhida D_ψ é dada pela dual da divergencia de Kullback-Leibler, $D_{KL} = D_\psi^*$. Isso fornecerá uma interpretação geométrica do algoritmo *EM*.

Seja (M, g, ∇, ∇^*) uma variedade dualmente plana induzida por uma divergência de Bregman D_ψ e sejam S e K subvariedades de M. Considere o problema de achar pontos $p^* \in S$ e $q^* \in K$ que satisfaçam

$$D_\psi[p^* : q^*] = D_\psi[S, K] := \min_{p \in S, q \in K} D_\psi[p : q].$$

Nosso objetivo é descrever um algoritmo que seja capaz de encontrar p^* e q^*. Considere θ e η as coordenadas duais de M, ou seja, θ sendo as coordenadas de ψ e $\eta = \nabla\psi(\theta)$ sendo as coordenadas da transformada de Legendre ψ^*. O algoritmo *em* consiste no seguinte:

Passo zero: Inicialize tomando $p_0 \in S$ (arbitrariamente, por exemplo).

Repita até "convergir":

e-**Passo:** Tome $q_0 = \arg\min_{q \in K} D_\psi[p_0 : q]$ (∇-projeção sobre K).

m-**Passo:** Tome $p_1 = \arg\min_{p \in S} D_\psi[p : q_0]$ (∇^*-projeção sobre S).

(A): Atualize $p_0 \leftarrow p_1$.

Nenhuma convergência é, a priori, garantida. Na prática, espera-se que o algoritmo convirja para mínimos locais de $D_\psi[S : K]$. Para garantir esta convergência, vamos assumir as seguintes hipóteses de convexidades nos *e*-passo e *m*-passo:

Hipóteses de convexidade no *em*-algoritmo:

Convexidade no *e*-Passo: Para a ∇-projeção, $q_0 = \arg\min_{q \in K} D_\psi[p_0 : q]$, assuma que a ∇-geodésica ligando p_0 a q_0 e a subvariedade K estão em lados opostos do plano tangente a $T_{q_0}K$ (em η-coordenadas).

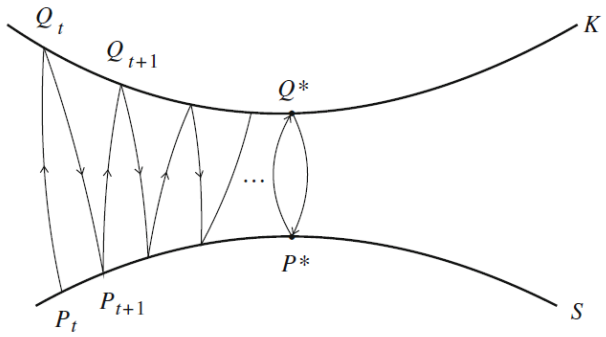

Figura 6.1 Algoritmo *em*

Convexidade m-Passo: Para a ∇^*-projeção, $p_1 = \arg\min_{p\in S} D_\psi[p : q_0]$, assuma que a ∇^*-geodésica ligando q_0 a p_1 e a subvariedade S estão em lados opostos do plano tangente a $T_{p_1}S$ (em θ-coordenadas).

Prova da convergência do algoritmo em: Vamos assumir as hipóteses de convexidades nos e,m-passos. Sejam $p_n, q_n, p_{n+1}, q_{n+1}$ as e,m-projeções sucessivas obtidas através do algoritmo *em* (veja Figura 6.1). Do Teorema Pitagoreano, vale que

$$D_\psi[p_n : p_{n+1}] + D_\psi[p_{n+1} : q_n] - D[p_n : q_n] = \langle \theta(p_n) - \theta(p_{n+1}), \eta(q_n) - \eta(p_{n+1})\rangle. \quad (6.1)$$

Pela hipótese de convexidade do m-passo, a ∇^*-geodésica $\gamma^* : (1-t)\eta(p_{n+1}) + t\eta(q_n)$ (em η-coordenadas) que liga p_{n+1} a q_n e a subvariedade S, estão em lados opostos com respeito ao plano tangente $T_{p_{n+1}}S$ (em θ-coordenadas). Isto implica que $\langle \gamma^{*\prime}(0), \gamma'(0)\rangle \leq 0$, para qualquer ∇-geodésica γ partindo de p_{n+1} ligando qualquer ponto de S, em particular o ponto p_n. Assim, as geodésicas

$$\gamma : \theta(t) = (1-t)\theta(p_{n+1}) + t\theta(p_n)$$
$$\gamma^* : \eta(t)(1-t)\eta(p_{n+1}) + t\eta(q_n).$$

satisfazem $\langle \gamma'(0), \gamma^{*\prime}(0)\rangle = \langle \theta(p_n) - \theta(p_{n+1}), \eta(q_n) - \eta(p_{n+1})\rangle \leq 0$. Logo, de (6.1), segue-se que

$$D_\psi[p_{n+1} : q_n] \leq D[p_n : q_n]. \quad (6.2)$$

Similarmente, do Teorema Pitagoreano,

$$D_\psi[p_{n+1} : q_{n+1}] + D_\psi[q_{n+1} : q_n] - D_\psi[p_{n+1} : q_n] = \langle \theta(p_{n+1}) - \theta(q_{n+1}), \eta(q_n) - \eta(q_{n+1})\rangle. \quad (6.3)$$

Da hipótese de convexidade do e-passo, a ∇-geodesica $\gamma : (1-t)\theta(q_{n+1}) + t\theta(p_{n+1})$ e a subvariedade K estão em lados opostos, com respeito ao plano tangente $T_{q_{n+1}}K$ (em η-coordenadas). Ou seja, dado $q \in K$ a ∇^*-geodesica $\gamma^* : (1-t)\eta(q_{n+1}) + t\eta(q)$ satisfaz $\langle \gamma'(0), \gamma^{*\prime}(0)\rangle \leq 0$. Em particular, aplicando no ponto q_n, temos $\langle \gamma'(0), \gamma^{*\prime}(0)\rangle = \langle \theta(p_{n+1}) - \theta(q_{n+1}), \eta(q_n) - \eta(q_{n+1})\rangle \leq 0$. Assim, de (6.3), tem-se

$$D_\psi[p_{n+1} : q_{n+1}] \leq D_\psi[p_{n+1} : q_n]. \quad (6.4)$$

Portanto, de (6.2) e (6.4), segue-se que

$$D_\psi[p_n : q_n] \geq D_\psi[p_{n+1} : q_{n+1}]. \tag{6.5}$$

Donde $D_\psi[p_n : q_n]$ é uma sequência monónota não-crescente e limitada, donde converge. Além disso, a convergência $D_\psi[p^* : q^*] = \lim D_\psi[p_n : q_n]$ deve satisfazer: $q^* = \arg\min D_\psi[p^* : K]$ (∇-projeção de p^* em K) e $p^* = \arg\min D_\psi[S : q^*]$ (∇^*-projeção de q^* em S).

6.2 Algoritmo EM

Vamos iniciar esta seção com o seguinte exemplo.

Exemplo 40. *Digamos que uma pesquisa foi feita numa determinada cidade, onde uma grande amostra de alturas de indivíduos, todos de uma mesma idade, foi coletada. No entanto, na pesquisa, por algum erro de comunicação, não perguntaram os sexos correspondentes. Será que, apenas com os dados coletados de alturas, $\{h_1, \ldots, h_N\}$, é possível estimar a proporção entre homens e mulheres, e as respectivas médias e variâncias?*

É razoável supor que as alturas entre pessoas do mesmo sexo e idade, são distribuídos por uma distribuição normal $N(h \mid \mu, \sigma^2)$, desse modo a distribuição de probabilidade das alturas é dado por uma mistura de probabilidades:

$$\begin{aligned} p(h) &= p(h \mid s = M)p(M) + p(h \mid s = F)p(F) \\ &= N(h \mid \mu_1, \sigma_1^2)p(M) + N(h \mid \mu_2, \sigma_2^2)p(F), \end{aligned}$$

sendo $s \in \{M, F\}$ a variável aleatória categórica conforme o sexo "masculino"ou "feminino", $p(M), p(F)$ são as proporções entre homens e mulheres, e $(\mu_1, \sigma_1), (\mu_2, \sigma_2)$ são as respectivas médias e desvio padrão entre homens e mulheres (veja Figura 6.2). No entanto, a variável aleatória "sexo"não foi observada.

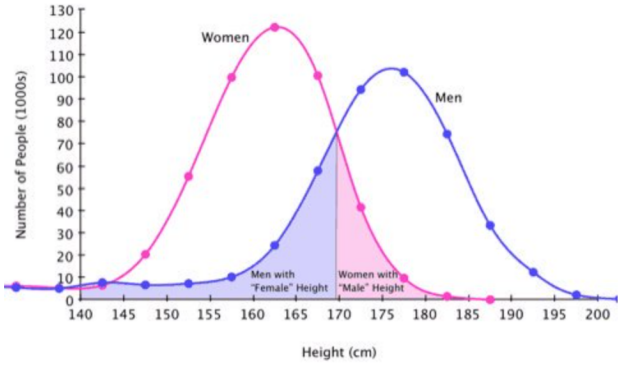

Figura 6.2 Exemplo de distribuição de alturas.

Veremos que o algoritmo EM será capaz de estimar os respectivos pesos $p(M)$, $p(F)$; e parâmetros $(\mu_1, \sigma_1), (\mu_2, \sigma_2)$, maximizando uma certa verossimilhança. Além disso, para um novo dado de altura h^*, seremos capazes estimar a probabilidade do sexo correpondente ser "masculino"ou "feminino".

Seja $X = (Y,H)$ uma variável aleatória tal que Y é totalmente observada, mas H não. Gostaríamos de, pelo menos, poder estimar a probabilidade marginal $p(y)$. Em geral, H é usada como uma variável aleatória latente, cuja probabilidade $p(y,h)$ pertença a um modelo estatístico $S = \{p_\xi(y,h)\}$ mais simples do que o modelo estatístico que contem a probabilidade marginal $p(y)$. Note que $p(y)$ está contida na mistura de probabilidades

$$p_\xi(y) = \int_{\mathscr{H}} p_\xi(y,h)d\mu(h) = \int_{\mathscr{H}} p_\xi(y \mid h)p_\xi(h)d\mu(h), \tag{6.6}$$

onde $d\mu(h)$ denota a medida σ-finita sobre $\mathscr{H}$ e $p_\xi(h)$ denota a probabilidade marginal de $p_\xi(y,h)$ sobre $\mathscr{H}$. O objetivo do algoritmo *EM* é estimar a máxima verossimilhança marginal

$$\hat{\xi} = \arg\max_\xi p(D \mid \xi),$$

onde $D = \{y_1,\ldots,y_N\}$ são os dados e $p(D \mid \xi) = \prod p_\xi(y_i)$ é a função de verossimilhança.

Seja $\bar{q}(y) = \frac{1}{N}\sum_{i=1}^N \delta(y-y_i)$ a distribuição empírica dos dados. Vimos em (5.9) que o estimador de máxima verossilhança $\hat{\xi}$ é dado minimizando-se a divergência de Kullback-Leibler:

$$\hat{\xi} = \arg\min_\xi D_{KL}[\bar{q}(y) : p_\xi(y)].$$

No entanto, a distribuição $p_\xi(y)$, dado pela integral (6.6), não possui, em geral, uma forma fechada. Logo, estimar a maxima verossimilhança de $p_\xi(y)$ pode ser um problema bem difícil. Assim, vamos trabalhar com a varíavel $x = (y,h)$, ignorando o fato de h não ter sido observada.

Como temos apenas a distribuição empírica de Y, $\bar{q}(y)$, qualquer candidato a distribuição empírica de $X = (Y,H)$ deve pertencer à seguinte subvariedade

$$K = \{\bar{q}(y,h) = q(h \mid y)\bar{q}(y) \mid q(h \mid y) > 0 \text{ e } \int_{\mathscr{H}} q(h \mid y)dh = 1\},$$

a qual chamaremos de *subvariedade observada*. Note que K é m-totalmente geodésica. De fato, sejam $q_1, q_2 \in K$. Sabemos que a m-geodésica (do ambiente) ligando q_1 a q_2 é dada por $q_t = (1-t)q_1 + tq_2$, com $t \in [0,1]$. Além disso, como $q_i(y,h) = q_i(h \mid y)\bar{q}(y)$, $i = 1,2$, segue-se que, para todo $t \in [0,1]$,

$$\begin{aligned}(1-t)q_1(y,h) + tq_2(y,h) &= (1-t)q_1(h \mid y)\bar{q}(y) + tq_2(h \mid y)\bar{q}(y) \\ &= [(1-t)q_1(h \mid y) + tq_2(h \mid y)]\bar{q}(y) \in K,\end{aligned}$$

pois

$$\int (1-t)q_1(h \mid y) + tq_2(h \mid y)dh = 1.$$

Assumindo que K e $S = \{p_\xi(y,h)\}$ são subvariedades de uma família exponencial (o que sempre é verdade, mesmo que $\mathscr{X} = \mathscr{Y} \times \mathscr{H}$ seja infinito), aplica-se o algoritmo *em* com o objetivo de minimizar a divergência de Bregman D_ψ, induzida do modelo do ambiente, entre S e K. Como a divergência de Kullback-Leibler satisfaz $D_{KL} = D_\psi^*$, equivalentemente, queremos encontrar $p_{\hat{\xi}} \in S$ e $\hat{q} \in K$ satisfaça:

$$D_{KL}[\hat{q}, p_{\hat{\xi}}] = D_{KL}[K : S] = \min_{\bar{q}\in K, p_\xi \in S} D_{KL}[\bar{q} : p_\xi]. \tag{6.7}$$

Veremos que $\hat{\xi}$ é um estimador de máxima verossimilhança de $p_\xi(y)$.

Teorema 6.1. *Se $\hat{q}(y,h) = \bar{q}(y)q(h \mid y) \in K$ e $p_{\hat{\xi}}(y,h) \in S$ minimizam $D_{KL}[K:S]$, então $\hat{\hat{\xi}}$ é o estimador de máxima verossimilhança do modelo $M' = \{p_\xi(y)\}$ e $q(h \mid y) = p_{\hat{\xi}}(h \mid y)$.*

Demonstração. Vamos calcular $D_{KL}[\bar{q}(y,h) : p_\xi(y,h)]$.

$$\begin{aligned} D_{KL}[\bar{q}(y,h) : p_\xi(y,h)] &= \int \bar{q}(x) \log \frac{\bar{q}(x)}{p_\xi(x)} dx = \iint q(h \mid y)\bar{q}(y) \log \frac{q(h \mid y)\bar{q}(y)}{p_\xi(h \mid y)p_\xi(y)} dydh. \\ &= \iint q(h \mid y)\bar{q}(y) \log \frac{q(h \mid y)}{p_\xi(h \mid y)} dydh + \iint q(h \mid y)\bar{q}(y) \log \frac{\bar{q}(y)}{p_\xi(y)} dydh \\ &= E_{\bar{q}(y)}\Big[D_{KL}[q(h \mid y) : p_\xi(h \mid y)]\Big] + D_{KL}[\bar{q}(y) : p_\xi(y)]. \end{aligned} \tag{6.8}$$

O último termo da equação (6.8) segue-se do fato de que $\int q(h \mid y)dh = 1$. Note que (6.8) também segue-se diretamente da Proposição 4.5.

Considere $\hat{\hat{\xi}}$ um estimador de máxima verossimilhança de $M' = \{p_\xi(y)\}$ e considere também $\bar{q}_0(y,h) = \bar{q}(y)p_{\hat{\xi}}(h \mid y) \in K$. Usando que $\min_\xi D_{KL}[\bar{q}(y) : p_\xi(y)] = D_{KL}[\bar{q}(y) : p_{\hat{\xi}}(y)]$, da equação (6.8), para todo $\bar{q}(y,h) \in K$ e $p_\xi(y,h) \in S$,

$$\begin{aligned} D_{KL}[\bar{q}(y,h) : p_\xi(y,h)] &= E_{\bar{q}(y)}\Big[D_{KL}[q(h \mid y) : p_\xi(h \mid y)]\Big] + D_{KL}[\bar{q}(y) : p_\xi(y)] \\ &\geq 0 + D_{KL}[\bar{q}(y) : p_{\hat{\xi}}(y)] \\ &= E_{\bar{q}(y)}\Big[D_{KL}[p_{\hat{\xi}}(h \mid y) : p_{\hat{\xi}}(h \mid y)]\Big] + D_{KL}[\bar{q}(y) : p_{\hat{\xi}}(y)] \\ &= D_{KL}[\bar{q}_0(y,h), p_{\hat{\xi}}(y,h)]. \end{aligned}$$

Donde $D_{KL}[\bar{q}_0(y,h), p_{\hat{\xi}}(y,h)] = D_{KL}[K,S]$ e a igualdade ocorre (igualando-se as duas primeiras linhas da inequação acima) se, e somente se, $q(h \mid y) = p_\xi(h \mid y)$, sendo ξ um estimador de máxima verossimilhança. O teorema está provado. □

Como já vimos, K é m-totalmente geodésica. Em particular, K satisfaz a condição de m-convexidade (das hipóteses do *em*-algoritmo). Além disso, usando (6.8), podemos também garantir uma expressão simples para a e-projeção.

Teorema 6.2 (E-passo)**.** *A e-projeção de $p_\xi(y,h) \in S$ sobre K, dada por*

$$\hat{q} = \arg\min_{q \in K} D_{KL}[q(y,h) : p_\xi(y,h)],$$

satisfaz $\hat{q}(h \mid y) = p_\xi(h \mid y)$.

Demonstração. Em (6.8), como $D_{KL}[\bar{q}(y) : p_\xi(y)]$ não depende de h, segue-se a e-projeção de $p_\xi(y,h)$ em K é dada por

$$\hat{q}(y,h) = \arg\min_{q \in K} D_{KL}[q(y,h) : p_\xi(y,h)] = \arg\min_{q \in K} E_{\bar{q}(y)}\Big[D_{KL}[q(h \mid y) : p_\xi(h \mid y)]\Big].$$

A última igualdade é minimizada simplesmente tomando-se $\hat{q}(h \mid y) = p_\xi(h \mid y)$. □

Agora, vamos calcular a m-projeção sobre S.

Teorema 6.3 (*M*-passo)**.** *A m-projeção de $q(y,h) \in K$ sobre S, dada por*

$$p_{\hat{\xi}}(y,h) = \arg\min_{\xi} D_{KL}[q(y,h) : p_{\xi}(y,h)],$$

satisfaz $\hat{\xi} = \arg\max_{\xi} L[\xi \mid q]$*, com*

$$L[\xi \mid q] = \sum_{i=1}^{N} \int_{\mathscr{H}} q(h \mid y_i) \log p(y_i,h) d\mu(h).$$

Veremos um caso particular importante do *EM*-algoritmo (o soft k-means), onde o M-passo assume uma expressão fechada simples.

Demonstração. Fixado $\bar{q}(x) = \bar{q}(y,h) \in K$, note que

$$\begin{aligned} D_{KL}[\bar{q}(y,h) : p_{\xi}(y,h)] &= \int \bar{q}(x) \log \bar{q}(x) d\mu(x) - \int \bar{q}(x) \log p_{\xi}(x) d\mu(x) \\ &= \int \bar{q}(x) \log \bar{q}(x) d\mu(x) - \frac{1}{N} L[\xi \mid \bar{q}], \end{aligned} \tag{6.9}$$

com

$$\begin{aligned} \frac{1}{N} L[\xi \mid \bar{q}] &= \int \bar{q}(x) \log p_{\xi}(x) d\mu(x) = \iint \bar{q}(y) q(h \mid y) \log p_{\xi}(y,h) d\mu(h) d\mu(y) \\ &= \int \bar{q}(y) \int q(h \mid y) \log p_{\xi}(y,h) d\mu(h) d\mu(y) \\ &= \frac{1}{N} \sum_{i=1}^{N} \int q(h \mid y_i) \log p_{\xi}(y_i,h) d\mu(h). \end{aligned}$$

Por (6.9), temos que $p_{\hat{\xi}}(y,h) = \arg\min_{\xi} D_{KL}[q(y,h) : p_{\xi}(y,h)]$ satisfaz $\hat{\xi} = \arg\max_{\xi} L[\xi \mid \bar{q}]$. O teorema está provado. □

É importante observar que a monotonicidade, dada por (6.5), ou seja

$$D_{KL}[q_n : p_{\xi_n}] \geq D_{KL}[q_{n+1} : p_{\xi_{n+1}}], \tag{6.10}$$

é garantida sob as hipóteses de S e K serem e-convexa e m-convexa, respectivamente. A m-convexidade de K é verdadeira, visto que ele é m-totalmente geodésico. Já a e-convexidade de S não é, em geral, satisfeita. Alguns algoritmos (k-means ++, por exemplo) buscam obter um chute inicial ξ_0 perto do ponto ótimo $\hat{\xi}$, onde espera-se que a e-convexidade seja satisfeita, garantindo a convergência.

6.3 Soft k-means

Um caso particular importante do *EM*-algoritmo é o soft k-means, onde $p(y)$ é uma mistura de probabilidades,

$$p(y) = \sum_{h=1}^{k} p(y \mid h)\, \pi_h,$$

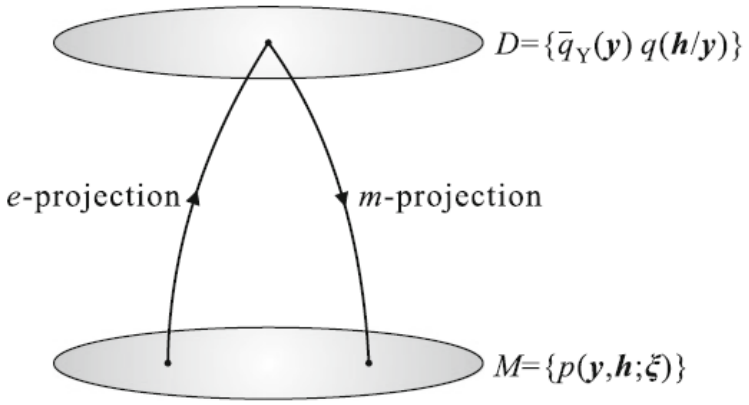

Figura 6.3 Algoritmo *EM*

com $\pi_h > 0$ e $\sum_{h=1}^{k} \pi_h = 1$, e $p(y \mid h)$ pertencendo a uma família mesma família exponencial Σ, ou seja,

$$p(y \mid h) = p(y;\theta_h) = e^{C(y)+\theta_h \cdot F(y)-\psi(\theta_h)}.$$

Assim, $H : h \in \{1,\ldots,k\}$ é uma variável aleatória latente com $p(y,h)$ pertencente ao seguinte modelo estatístico

$$S = \{p_\xi(y,h) = p(y;\theta_h)\pi_h \mid \xi = [\pi_{h'},\theta_{h'}]_{h'=1}^{k}\}.$$

Outra interpretação da variável latente $H = \{h\}$, é a de que ela representa clusters dos dados gerados pela variável aleatória $Y = \{y\}$ e cada dado y pertencerá ao cluster h com uma certa probabilidade $q(h \mid y)$. Essa abordagem difere do hard k-means, onde os cluster separam completamente os dados, via seus representantes mais próximos. Note que isso pode ser visto como um caso particular do soft k-means, tomando $q(h \mid y) = \delta(h - h(y))$, onde $h(y) = \arg\min_{h'} D[y : \mu_{h'}])$ é o cluster C_h no qual y está mais próximo de seu representante μ_h.

Nosso objetivo será dar uma forma fechada para o M-passo, dado pelo Teorema 6.3. Considere a mesma notação do EM-algoritmo. Dado $q(y,h) \in K$, usando que a integral $\int_{\mathscr{H}}$ é dada pela soma $\sum_{h=1}^{k}$, temos que

$$L[\xi \mid q] = \sum_{i=1}^{N}\sum_{h=1}^{k} q(h \mid y_i)\log p_\xi(y_i,h) = \sum_{i=1}^{N}\sum_{h=1}^{k} q(h \mid y_i)(\log p(y_i;\theta_h) + \log \pi_h), \qquad (6.11)$$

sendo $\xi = [\pi_{h'},\theta_{h'}]_{h'=1}^{k}$.

Primeiro, vamos derivar L em relação a $\pi = [\pi_h]_{h=1}^{k}$. Note que isso não é tão simples pois os parâmetros não são livres, visto que $\pi_h > 0$ satisfaz $\pi_1 + \ldots + \pi_k = 1$. Se $t \mapsto \pi_t = [\pi_h(t)]_{h=1}^{k}$ é uma curva no simplexo $S_k = \{\pi = [\pi_h]_{h=1}^{k} \mid \pi_h > 0 \text{ e } \pi_1 + \ldots + \pi_k = 1\}$, a derivada satisfaz $\sum_{h=1}^{k} \pi_h'(t) = 0$, ou seja $\alpha = \frac{d\pi_t}{dt}$, é dada por $\alpha_h = \pi_h'(t)$ com $\langle \alpha, (1,\ldots,1)\rangle = 0$, ou seja, o plano tangente $T_\pi S_k = (1,\ldots,1)^\perp$ (complemento ortogonal do vetor $(1,\ldots,1)$ em $\mathbb{R}^k$). Assim, seja

$\alpha \in T_\pi S_k$, temos que a derivada $L_*\alpha$ de $L[\xi \mid q]$ aplicada ao vetor α é dado por

$$L_*\alpha = \sum_{i=1}^{N}\sum_{h=1}^{k} q(h \mid y_i)\frac{1}{\pi_h}\alpha_h = \left\langle \left[\sum_{i=1}^{N} q(h \mid y_i)\frac{1}{\pi_h}\right]_{h=1}^{k}, \alpha \right\rangle.$$

Portanto, $L_*\alpha = 0$, para todo $\alpha \in T_\pi S_k$ se, e somente se, o vetor $Z = \left[\sum_{i=1}^{N} q(h \mid y_i)\frac{1}{\pi_h}\right]_{h=1}^{k}$ é ortogonal ao plano tangente $T_\pi S_k$. E isso, é equivalente a dizer que $Z = c(1,\ldots,1)$, para alguma constante c. Logo, $\sum_{i=1}^{N} q(h \mid y_i)\frac{1}{\pi_h} = c$, para $h = 1,\ldots,k$. Multiplicando π_h em ambos lados e somando em h, temos

$$c = \sum_{h=1}^{k} c\,\pi_h = \sum_{h=1}^{k}\sum_{i=1}^{N} q(h \mid y_i) = \sum_{i=1}^{N}\sum_{h=1}^{k} q(h \mid y_i) = N.$$

Assim,

$$\pi_h = \frac{1}{N}\sum_{i=1}^{N} q(h \mid y_i). \tag{6.12}$$

Agora, vamos derivar em relação a $\theta = [\theta_h]$. Fixado h, note que $\log p(y;\theta_h) = C(y) + \theta_h \cdot F(y) - \psi(\theta_h)$. Assim, tomando-se uma coordenada θ_h^i, usando (6.11), temos

$$\frac{\partial L}{\partial \theta_h^i} = \sum_{i=1}^{N} q(h \mid y_i)\left[F_i(y_i) - \frac{\partial \psi(\theta_h)}{\partial \theta_h^i}\right] = \sum_{i=1}^{N} q(h \mid y_i)\left[F_i(y_i) - \eta_i^h\right]$$

onde $\eta^h = \nabla\psi(\theta_h) = E_{p(y;\theta_h)}[F(y)]$, são as coordenadas esperadas da família exponencial Σ. Assim, se $\frac{\partial L}{\partial \theta_h^i} = 0$, para todo θ_h, temos que

$$\eta^h = \frac{\sum_{i=1}^{N} q(h \mid y_i)F_i(y_i)}{\sum_{i=1}^{N} q(h \mid y_i)}. \tag{6.13}$$

Portanto, usando (6.12) e (6.13), o M-passo consiste no seguinte: dado

$$q(h \mid y) = p_{\xi^n}(h \mid y) = \frac{p_\xi^n(y;\theta_h)\pi_h^n}{\sum_{h'=1}^{k} p_\xi^n(y;\theta_{h'})\pi_{h'}^n},$$

1. Atualizar os novos pesos: $(\pi_h)^{n+1} = \frac{1}{N}\sum_{i=1}^{N} q(h \mid y_i)$;
2. Atualizar os novos parâmetros: $(\eta^h)^{n+1} = \frac{\sum_{i=1}^{N} q(h|y_i)F_i(y_i)}{\sum_{i=1}^{N} q(h|y_i)}$.

É importante observar que os parâmetros obtidos pelo M-passo são os parâmetros esperados η^h da família exponencial Σ. Em muitos casos, a distribuição $p(y;\theta_h)$ já é dada em termos desses parâmetros η^h. Por exemplo, a distribuição normal pode ser dada em termos da média μ e do segundo momento $\mu^2 + \sigma^2$.

Tal como no algoritmo hard k-means, o soft k-means pode ser expressado diretamente por divergências de Bregman. Isso decorre de um teorema dado por Banerjee [5], no qual prova-se a existência de uma bijeção entre famílias exponenciais e divergências de Bregman, satisfazendo

$$p(y;\theta) = e^{-D_\varphi[y:\eta]+k(y)},$$

onde φ é a Transformada de Legendre de ψ e $k(y)$ é uma função.

Referências Bibliográficas

[1] AMARI, Shun-Ichi. *Information geometry and its Applications*, Applied Mathematics Sciences, Springer, 2016.

[2] AMARI, Shun-Ichi e NAGAOKA, Hiroshi. *Methods of Information Geometry*. Translations of Mathematical Monographs, Vol.191, Am. Math. Soc., 2000.

[3] ARWINI, Khadiga A. e DODSON, Christopher T.J.. *Information Geometry - Near Randomness and Near Independence*. Lecture Notes in Mathematics, 2008.

[4] AY, N., JOST, J., LÊ, V., and SCHWACHHÖFER, L., Information Geometry and Sufficient Statistics. Probability Theory and Related Fields, v. 162, 327 – 364, 2015.

[5] Banerjee, A. Dhillon, I. Ghosh, J., *Clustering with Bregman Divergences*. Journal of Machine Learning Research v.6, 1705-1749, 2005.

[6] Banerjee, A., Guo, X., and Wang, H., *On the optimality of conditional expectation as a Bregman predictor*. IEEE Trans. Inf. Theory, vol. 51, 7, 2005.

[7] BIEZUNER, Rodney Josué. Notas de aula: Geometria Riemanniana. Disponível online.

[8] BUSSAD, Wilton DE O. e MORETTIN, Pedro A. Estatística Básica, 2013, oitava edição

[9] CALIN, Ovidiu e UDRISTE, Constantin. Geometric Modeling in Probability and Statistics, Springer, 2014.

[10] do CARMO, Manfredo Perdigão. Geometria Riemanniana. 2011, 5^a edição, Projeto Euclides, 2011.

[11] CĔNCOV, N. N. C, Statistical Decision Rules and Optimal Inference, AMS, 1982 (originally published in Russian, Nauka, 1972)

[12] CSISZÁR, I., "Eine Informationstheoretische Ungleichung und ihre An- wendung auf den Beweis der Ergodizität on Markoffschen Ketten,"Publ. Math. Inst. Hungar. Acad. Sci., ser. A, vol. 8, pp. 84 – 108, 1963.

[13] Matumoto, Takao. Any statistical manifold has a contrast function—on the C^3-functions taking the minimum at the diagonal of the product manifold. Hiroshima Math. J. 23 (1993), no. 2, 327–332.

[14] AMARI, S.-I., α -Divergence Is Unique, Belonging to Both f-Divergence and Bregman Divergence Classes, IEEE Transactions on Information Theory, vol. 55, no. 11, pp. 4925-4931, Nov. 2009.doi: 10.1109/TIT.2009.2030485

[15] SHERN, S.S., SHEN, H. W. e LAM, K. S., *Lectures on Differential Geometry*. Series on University Mathematics, Vol. 1, 1999.

[16] WASSERMAN, Larry. *All of Statistics: A Concise Couse in Statistical Inference*. Springer Text in Statistics, Springer 2004.

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis).