Geometria da informação e algumas de suas aplicações

Fábio C. C. Meneghetti

IMECC (Unicamp)

31 de março de 2022

Probabilidades

- uma variável aleatória X pode ser vista como uma variável que assume valores em $\mathcal{X} = \{x_1, \ldots, x_{n+1}\}$ com probabilidades $p_i = P[X = x_i]$.

Probabilidades

- uma variável aleatória X pode ser vista como uma variável que assume valores em $\mathcal{X} = \{x_1, \ldots, x_{n+1}\}$ com probabilidades $p_i = P[X = x_i]$.
- necessariamente $p_1 + \cdots + p_{n+1} = 1$, isto é, a probabilidade total é 1.

Probabilidades

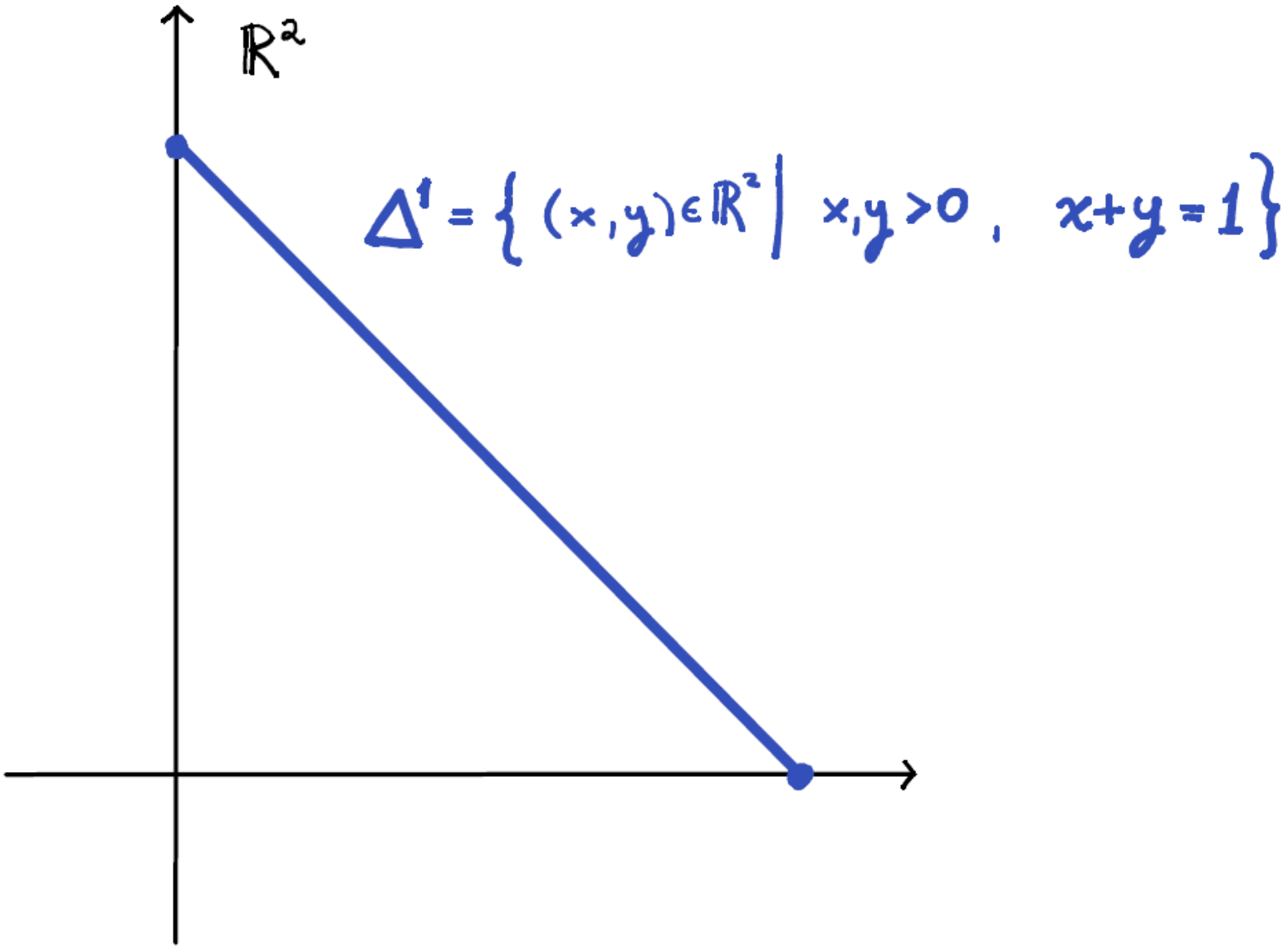
- uma variável aleatória X pode ser vista como uma variável que assume valores em $\mathcal{X} = \{x_1, \ldots, x_{n+1}\}$ com probabilidades $p_i = P[X = x_i]$.
- necessariamente $p_1 + \cdots + p_{n+1} = 1$, isto é, a probabilidade total é 1.
- **ex:** um dado honesto é uma variável aleatória que assume valores em $\{1, 2, 3, 4, 5, 6\}$, com probabilidades todas $p_i = 1/6$ para todo i. Esses valores poderiam ser diferentes se o dado fosse viciado.

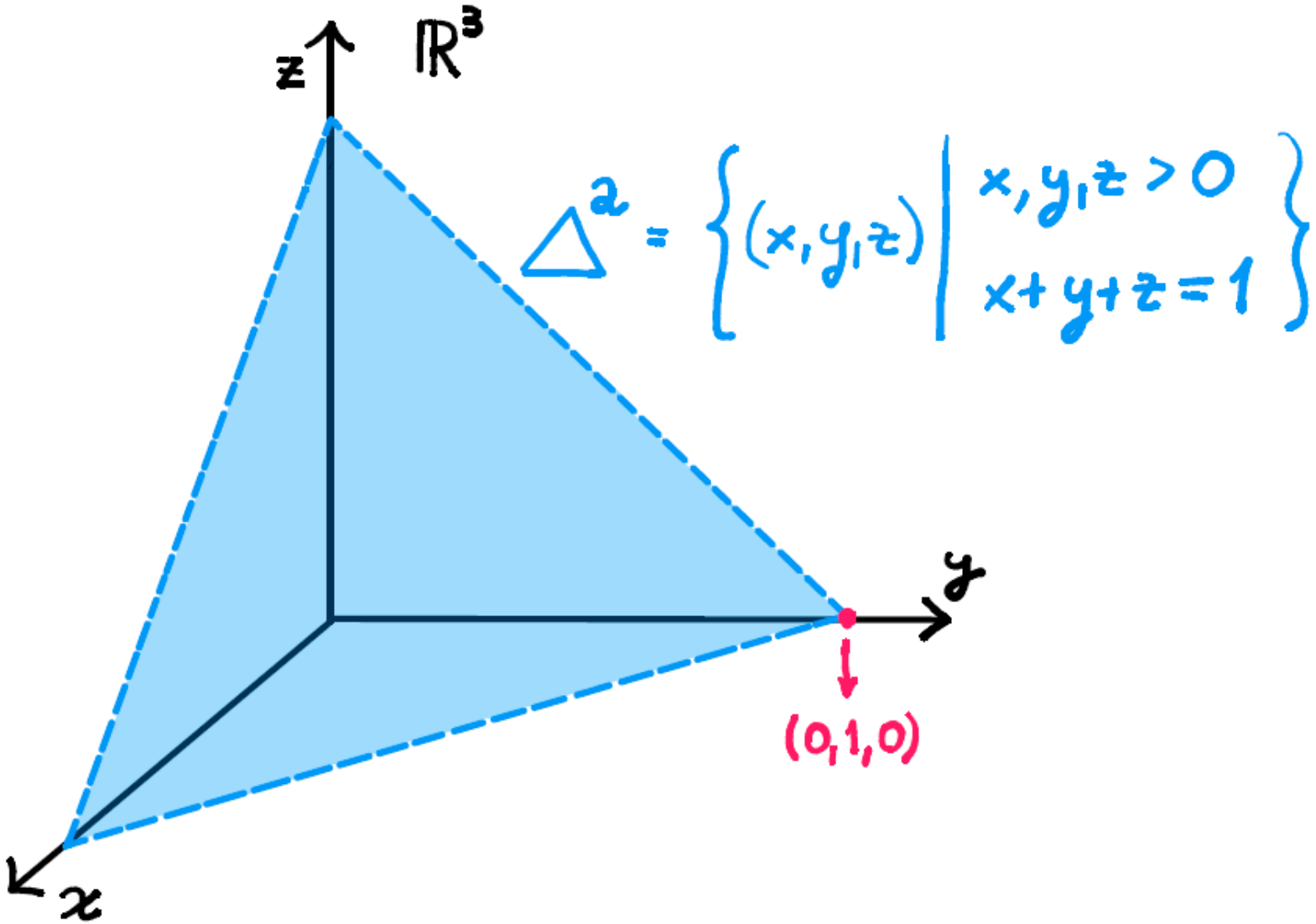
Probabilidades

- uma variável aleatória X pode ser vista como uma variável que assume valores em $\mathcal{X} = \{x_1, \ldots, x_{n+1}\}$ com probabilidades $p_i = P[X = x_i]$.
- necessariamente $p_1 + \cdots + p_{n+1} = 1$, isto é, a probabilidade total é 1.
- **ex:** um dado honesto é uma variável aleatória que assume valores em $\{1, 2, 3, 4, 5, 6\}$, com probabilidades todas $p_i = 1/6$ para todo i. Esses valores poderiam ser diferentes se o dado fosse viciado.
- o expaço que descreve todas as possíveis distribuições de variáveis aleatórias tomando $n + 1$ valores é

$$\Delta^n = \left\{ p = (p_1, \ldots, p_{n+1}) \in \mathbb{R}^{n+1} \;\middle|\; 0 \leq p_i \leq 1,\ p_1 + \cdots + p_{n+1} = 1 \right\}$$

Esse espaço é chamado de *simplexo padrão*.




z
$\mathbb{R}^3$
$\Delta^2 = \left\{ (x,y,z) \;\middle|\; \begin{array}{l} x,y,z > 0 \\ x+y+z = 1 \end{array} \right\}$
y
(0,1,0)
x

- o simplexo Δ^n pode ser parametrizado (por exemplo) pela função

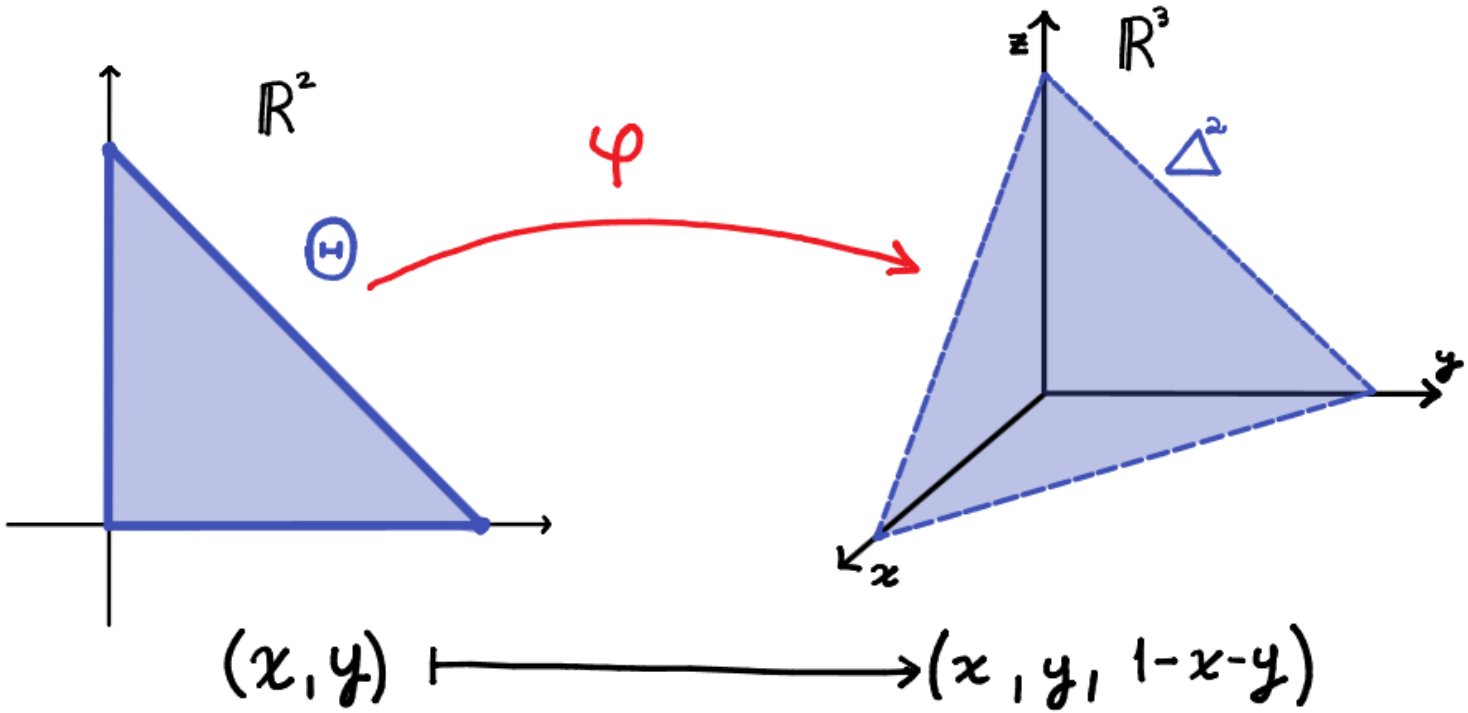
$$\varphi(p_1, \ldots, p_n) = (p_1, \ldots, p_n, p_{n+1}), \qquad p_{n+1} = 1 - \sum_{i=1}^{n} p_i,$$

onde $\varphi\colon \Theta \to \Delta^n$ é definida sobre
$\Theta = \{\theta = (p_1, \ldots, p_n) \in \mathbb{R}^n \mid 0 \leq p_1 + \cdots + p_n \leq 1\}$

- o simplexo Δ^n pode ser parametrizado (por exemplo) pela função

$$\varphi(p_1, \ldots, p_n) = (p_1, \ldots, p_n, p_{n+1}), \qquad p_{n+1} = 1 - \sum_{i=1}^{n} p_i,$$

onde $\varphi\colon \Theta \to \Delta^n$ é definida sobre
$\Theta = \{\theta = (p_1, \ldots, p_n) \in \mathbb{R}^n \mid 0 \leq p_1 + \cdots + p_n \leq 1\}$



Informação de Fisher

- queremos estudar a *geometria* natural desse espaço Δ^n

Informação de Fisher

- queremos estudar a *geometria* natural desse espaço Δ^n
- ideia central da geometria da informação: introduzimos uma *métrica Riemanniana* chamada métrica da informação de Fisher, que fornece essa geometria

Informação de Fisher

- queremos estudar a *geometria* natural desse espaço Δ^n
- ideia central da geometria da informação: introduzimos uma *métrica Riemanniana* chamada métrica da informação de Fisher, que fornece essa geometria
- sejam $\mathcal{X} = \{x_1, \ldots, x_n, x_{n+1}\}$ o espaço de X, $\varphi(\theta)$ uma parametrização de Δ^n e $p_\theta(x)$ a função massa de probabilidade de X dada por $p_\theta(x_i) = p_i$.

Informação de Fisher

- queremos estudar a *geometria* natural desse espaço Δ^n
- ideia central da geometria da informação: introduzimos uma *métrica Riemanniana* chamada métrica da informação de Fisher, que fornece essa geometria
- sejam $\mathcal{X} = \{x_1, \ldots, x_n, x_{n+1}\}$ o espaço de X, $\varphi(\theta)$ uma parametrização de Δ^n e $p_\theta(x)$ a função massa de probabilidade de X dada por $p_\theta(x_i) = p_i$.
- a matriz da informação de Fisher com respeito a $\varphi(\theta)$ é a matriz $I(\theta) = [g_{ij}(\theta)]$ dada por

$$g_{ij}(\theta) = \sum_{i=1}^{n} p_\theta(x) \frac{\partial \log p_\theta(x)}{\partial \theta_i} \frac{\partial \log p_\theta(x)}{\partial \theta_j}$$

- Por exemplo, se usarmos a parametrização

$$\varphi(p_1, \ldots, p_n) = (p_1, \ldots, p_n, p_{n+1}), \qquad p_{n+1} = 1 - \sum_{i=1}^{n} p_i,$$

já mostrada anteriormente, então obtemos a matriz de Fisher dada por $g_{ij}(p_1, \ldots, p_n) = \frac{1}{p_{n+1}} + \frac{\delta_{ij}}{p_i}$

- Por exemplo, se usarmos a parametrização

$$\varphi(p_1,\ldots,p_n) = (p_1,\ldots,p_n,p_{n+1}), \qquad p_{n+1} = 1 - \textstyle\sum_{i=1}^{n} p_i,$$

já mostrada anteriormente, então obtemos a matriz de Fisher dada por $g_{ij}(p_1,\ldots,p_n) = \frac{1}{p_{n+1}} + \frac{\delta_{ij}}{p_i}$

$$I(\theta) = \begin{bmatrix} \frac{1}{p_1} + \frac{1}{p_{n+1}} & \frac{1}{p_{n+1}} & \cdots & \frac{1}{p_{n+1}} \\ \frac{1}{p_{n+1}} & \frac{1}{p_2} + \frac{1}{p_{n+1}} & \cdots & \frac{1}{p_{n+1}} \\ \vdots & \vdots & \ddots & \frac{1}{p_{n+1}} \\ \frac{1}{p_{n+1}} & \frac{1}{p_{n+1}} & \cdots & \frac{1}{p_n} + \frac{1}{p_{n+1}} \end{bmatrix}$$

- isto é, $I(\theta) = \frac{1}{p_{n+1}} I_{n\times n} + \operatorname{diag}(\frac{1}{p_1},\ldots,\frac{1}{p_n})$

Para quê serve essa métrica?

- uma métrica Riemanniana, como a métrica de Fisher, nos permite definir produto interno e norma:

$$\langle v, w \rangle_{I(\theta)} = v^\top I(\theta) w, \qquad \|v\|_{I(\theta)} = \sqrt{v^\top I(\theta) v}$$

para vetores tangentes, de um jeito que é invariante por reparametrização.

Para quê serve essa métrica?

- uma métrica Riemanniana, como a métrica de Fisher, nos permite definir produto interno e norma:

$$\langle v, w \rangle_{I(\theta)} = v^\top I(\theta) w, \qquad \|v\|_{I(\theta)} = \sqrt{v^\top I(\theta) v}$$

  para vetores tangentes, de um jeito que é invariante por reparametrização.

- Com isso, conseguimos também falar de ângulos: $\angle(v, w) = \arccos \frac{\langle v, w \rangle_{I(\theta)}}{\|v\|_{I(\theta)} \|w\|_{I(\theta)}}$.

Para quê serve essa métrica?

- uma métrica Riemanniana, como a métrica de Fisher, nos permite definir produto interno e norma:

$$\langle v, w \rangle_{I(\theta)} = v^\top I(\theta) w, \qquad \|v\|_{I(\theta)} = \sqrt{v^\top I(\theta) v}$$

  para vetores tangentes, de um jeito que é invariante por reparametrização.

- Com isso, conseguimos também falar de ângulos: $\angle(v, w) = \arccos \frac{\langle v, w \rangle_{I(\theta)}}{\|v\|_{I(\theta)} \|w\|_{I(\theta)}}$.

- Com todas essas propriedades, podemos também calcular o *comprimento de uma curva* na métrica da informação!

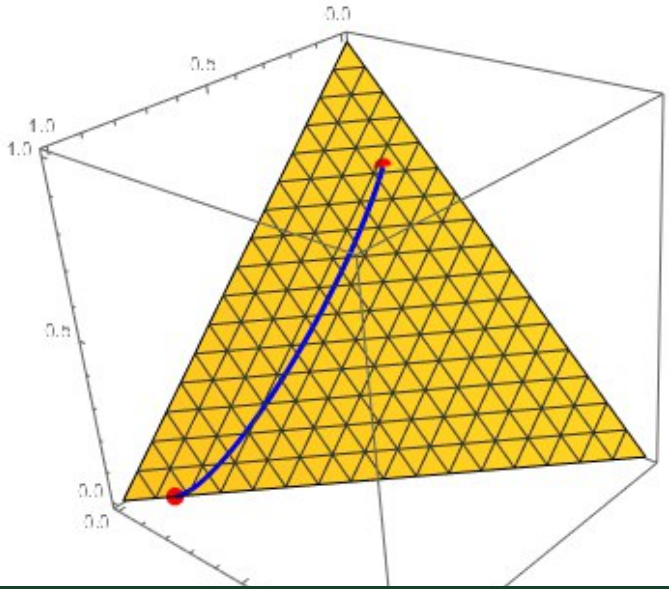
- o comprimento de uma curva $\gamma\colon [0,1] \to \Delta^n$ é (assim como no espaço euclideano) dado pela integral do tamanho vetor velocidade:

$$\ell(\gamma) = \int_0^1 \left\|\gamma'(t)\right\|_{I(\theta(t))} \mathrm{d}t$$

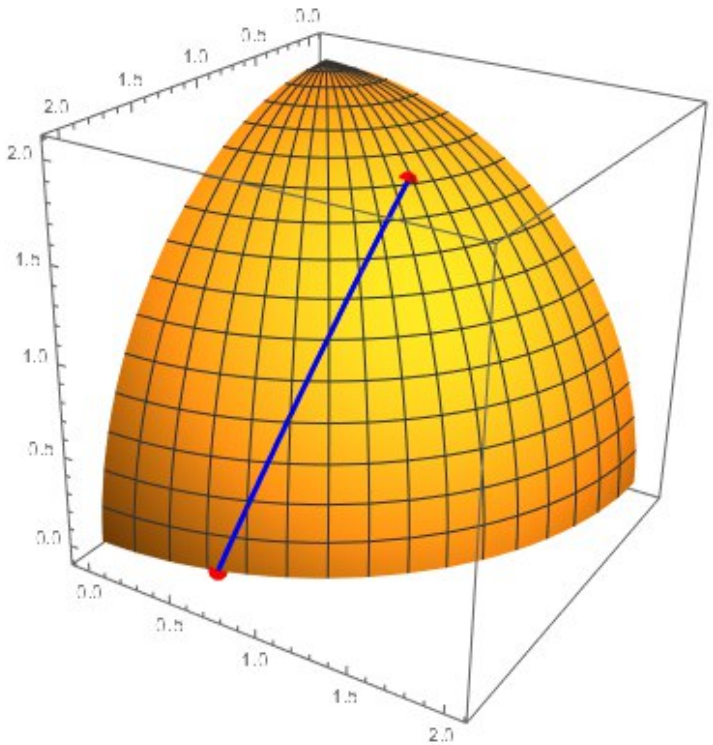
- o comprimento de uma curva $\gamma\colon [0,1] \to \Delta^n$ é (assim como no espaço euclideano) dado pela integral do tamanho vetor velocidade:

$$\ell(\gamma) = \int_0^1 \left\| \gamma'(t) \right\|_{I(\theta(t))} \mathrm{d}t$$

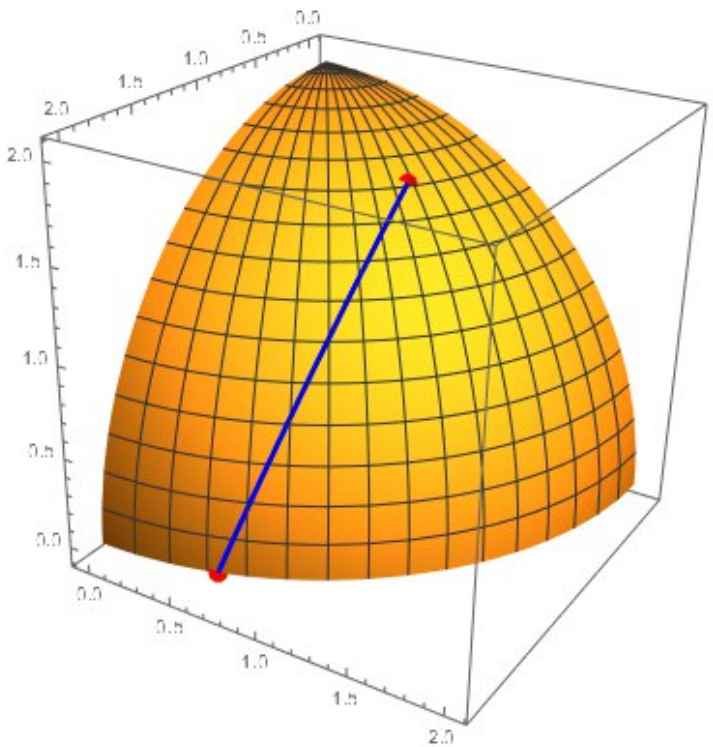
- dados dois pontos $p = (p_1, \ldots, p_{n+1})$ e $q = (q_1, \ldots, q_{n+1})$ em Δ^n, existem diversas curvas ligando p e q. Aquela de menor comprimento é chamada o *segmento de geodésica* ligando p e q. As geodésicas de uma variedade dão noção de "linha reta" naquela geometria



- quando "enchemos" esse triângulo para virar um quadrante (octante) de esfera, através do mapa $z_i = 2\sqrt{p_i}$, a métrica torna-se a métrica esférica, cuja geometria é bem conhecida



- quando "enchemos" esse triângulo para virar um quadrante (octante) de esfera, através do mapa $z_i = 2\sqrt{p_i}$, a métrica torna-se a métrica esférica, cuja geometria é bem conhecida



- na geometria esférica, as geodésicas são os *grandes círculos* e seu comprimento é dado por 2α (o dobro do ângulo entre os dois vetores)

- tendo a geometria de Fisher, podemos calcular a distância entre dois pontos $p, q \in \Delta^n$ como o comprimento da geodésica ligando esses pontos! No caso do simplexo, podemos calcular essa distância:

$$d_{\mathrm{FR}}(p, q) = 2 \arccos \left(\sum_{i=1}^{n+1} \sqrt{p_i q_i} \right)$$

- tendo a geometria de Fisher, podemos calcular a distância entre dois pontos $p, q \in \Delta^n$ como o comprimento da geodésica ligando esses pontos! No caso do simplexo, podemos calcular essa distância:

$$d_{\mathrm{FR}}(p, q) = 2 \arccos \left(\sum_{i=1}^{n+1} \sqrt{p_i q_i} \right)$$

- essa distância é chamada *distância de Fisher-Rao*, e é a distância natural da geometria de Fisher.

Porque essa métrica?

- uma estatística suficiente é um mapa $\kappa\colon \mathcal{X} \to \mathcal{Y}$ que preserva a estrutura estatística, e satisfaz

$$p_\theta(x) = \tilde{p}_\theta(k(x)) \cdot h(x)$$

Porque essa métrica?

- uma estatística suficiente é um mapa $\kappa\colon \mathcal{X} \to \mathcal{Y}$ que preserva a estrutura estatística, e satisfaz

$$p_\theta(x) = \tilde{p}_\theta(k(x)) \cdot h(x)$$

- a métrica de Fisher é invariante por estatísticas suficientes!

Porque essa métrica?

- uma estatística suficiente é um mapa $\kappa\colon \mathcal{X} \to \mathcal{Y}$ que preserva a estrutura estatística, e satisfaz

$$p_\theta(x) = \tilde{p}_\theta(k(x)) \cdot h(x)$$

- a métrica de Fisher é invariante por estatísticas suficientes!
- Teorema de Chentsov: a métrica de Fisher é a única métrica Riemanniana (a menos de uma constante) invariante por estatísticas suficientes!

Porque essa métrica?

- uma estatística suficiente é um mapa $\kappa\colon \mathcal{X} \to \mathcal{Y}$ que preserva a estrutura estatística, e satisfaz

$$p_\theta(x) = \tilde{p}_\theta(k(x)) \cdot h(x)$$

- a métrica de Fisher é invariante por estatísticas suficientes!
- Teorema de Chentsov: a métrica de Fisher é a única métrica Riemanniana (a menos de uma constante) invariante por estatísticas suficientes!
  - este teorema nos mostra que, se queremos estudar a estrutura estatística dos espaços de probabilidades, a métrica de Fisher é a escolha natural

Importância na estatística

- um *estimador* do parâmetro θ é uma estimativa $\hat{\theta}(X)$ de θ usado, na ausência de conhecimento deste

Importância na estatística

- um *estimador* do parâmetro θ é uma estimativa $\hat{\theta}(X)$ de θ usado, na ausência de conhecimento deste
- dizemos que um estimador é *não-enviesado* se $\mathrm{E}[\hat{\theta}(X) - \theta] = 0$

Importância na estatística

- um *estimador* do parâmetro θ é uma estimativa $\hat{\theta}(X)$ de θ usado, na ausência de conhecimento deste
- dizemos que um estimador é *não-enviesado* se $\mathrm{E}[\hat{\theta}(X) - \theta] = 0$
- Limitante de Cramér-Rao: a informação de Fisher (invertida) é um limitante inferior para a variância de um estimador não-enviesado:

$$\mathrm{V}(\hat{\theta}) \geq I(\theta)^{-1}$$

Importância na estatística

- um *estimador* do parâmetro θ é uma estimativa $\hat{\theta}(X)$ de θ usado, na ausência de conhecimento deste
- dizemos que um estimador é *não-enviesado* se $\mathrm{E}[\hat{\theta}(X) - \theta] = 0$
- Limitante de Cramér-Rao: a informação de Fisher (invertida) é um limitante inferior para a variância de um estimador não-enviesado:

$$\mathrm{V}(\hat{\theta}) \geq I(\theta)^{-1}$$

  - isso se generaliza para o caso multiparâmetros, trocando variância por matriz de covariância, e $\geq$ pela ordem de Loewner ($A \geq B$ se $A - B$ é positiva definida)

Distribuições normais

- toda a construção que fizemos também pode ser extendida para distribuições de probabilidade contínuas.

Distribuições normais

- toda a construção que fizemos também pode ser extendida para distribuições de probabilidade contínuas.
- um dos principais exemplos é a família de distribuições normais (ou gaussianas), com média $\mu \in \mathbb{R}$ e desvio padrão $\sigma > 0$:

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

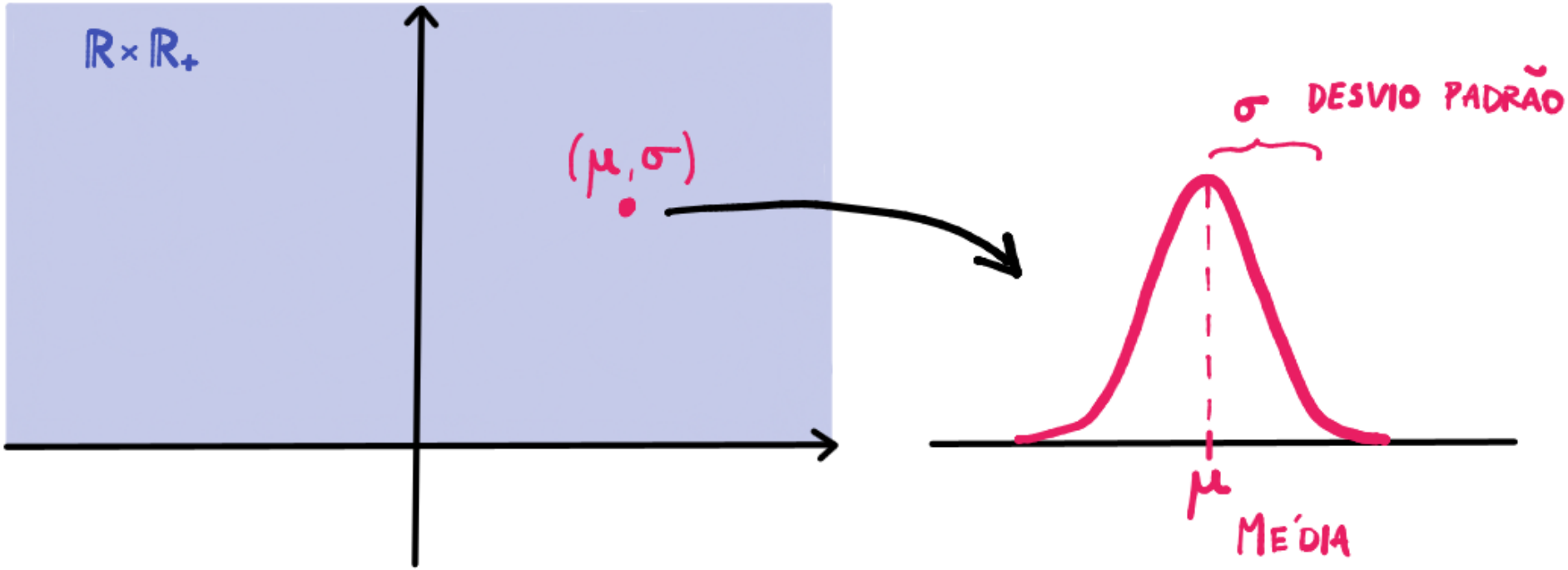
Distribuições normais

- toda a construção que fizemos também pode ser extendida para distribuições de probabilidade contínuas.
- um dos principais exemplos é a família de distribuições normais (ou gaussianas), com média $\mu \in \mathbb{R}$ e desvio padrão $\sigma > 0$:

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

- o espaço de parâmetros é o meio-plano

$$\mathbb{R} \times \mathbb{R}_+ = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$$


$\mathbb{R} \times \mathbb{R}_+$
(μ, σ)
σ DESVIO PADRÃO
μ
MÉDIA

- denotando os parâmetros por $\mu = \theta_1$ e $\sigma = \theta_2$, a métrica de Fisher é a matriz 2×2 dada por

$$g_{ij}(\mu, \sigma) = \int_{-\infty}^{\infty} g_{\mu,\sigma}(x) \frac{\partial \log g_{\mu,\sigma}(x)}{\partial \theta_i} \frac{\partial \log g_{\mu,\sigma}(x)}{\partial \theta_j} \, \mathrm{d}x$$

- denotando os parâmetros por $\mu = \theta_1$ e $\sigma = \theta_2$, a métrica de Fisher é a matriz 2×2 dada por

$$g_{ij}(\mu, \sigma) = \int_{-\infty}^{\infty} g_{\mu,\sigma}(x) \frac{\partial \log g_{\mu,\sigma}(x)}{\partial \theta_i} \frac{\partial \log g_{\mu,\sigma}(x)}{\partial \theta_j} \, \mathrm{d}x$$

- fazendo as contas, ela é dada por

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^2} \end{bmatrix}$$

- denotando os parâmetros por $\mu = \theta_1$ e $\sigma = \theta_2$, a métrica de Fisher é a matriz 2×2 dada por

$$g_{ij}(\mu, \sigma) = \int_{-\infty}^{\infty} g_{\mu,\sigma}(x) \frac{\partial \log g_{\mu,\sigma}(x)}{\partial \theta_i} \frac{\partial \log g_{\mu,\sigma}(x)}{\partial \theta_j} \, \mathrm{d}x$$

- fazendo as contas, ela é dada por

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^2} \end{bmatrix}$$

- a geometria definida por essa métrica é chamada de *geometria hiperbólica*

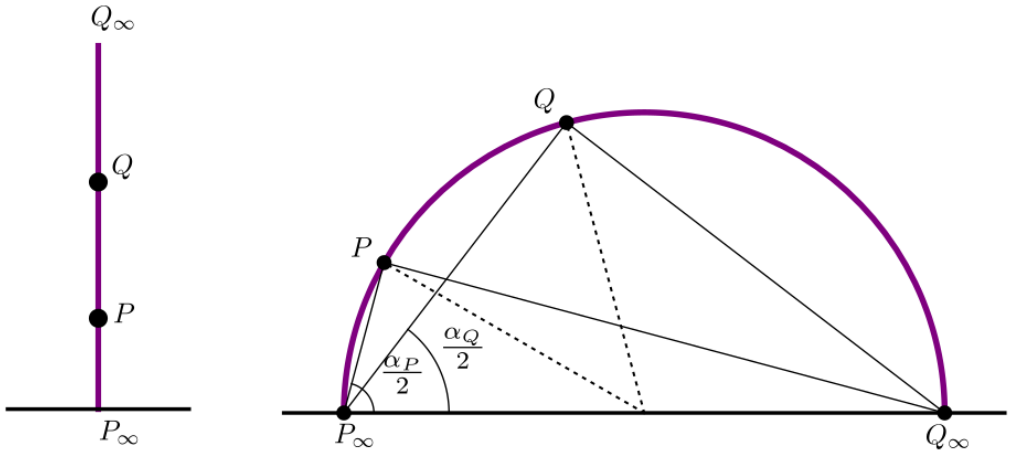


Fig. 3. Elements to compute the distance $d_H(P, Q)$, in case the points $P, Q \in \mathbb{H}^2$ are vertically aligned (left) or not (right).

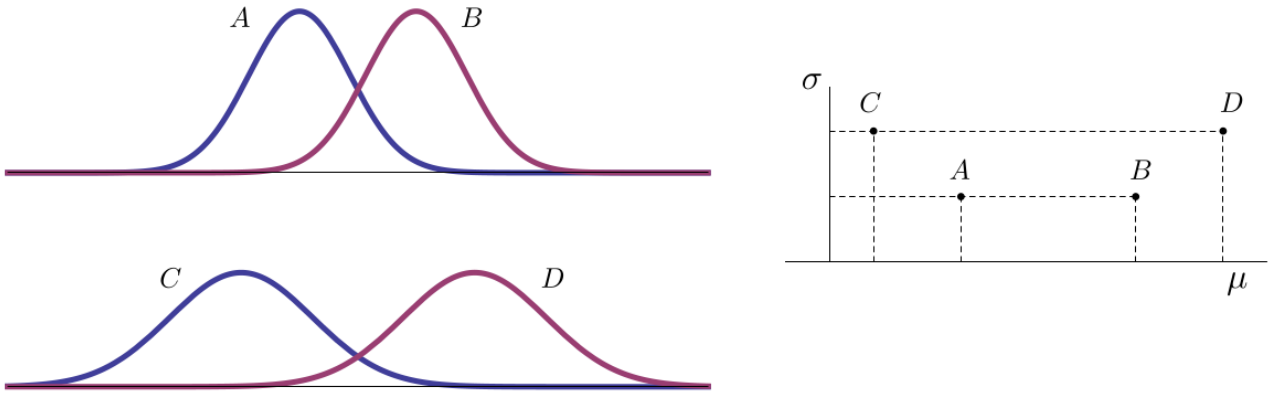


Fig. 4. Equidistant pairs in Fisher metric: $d_H(A, B) = d_F(C, D) = 2.37687$, where $A = (1.5, 0.75)$, $B = (3.5, 0.75)$ and $C = (0.5, 1.5)$, $D = (4.5, 1.5)$.

Divergência de Kullback-Leibler

- uma importante medida de dissimilaridade entre distribuições é a divergêngia de Kullback-Leibler (também chamada entropia relativa), dada por

$$\mathrm{D}_{\mathrm{KL}}(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

  no caso contínuo,

Divergência de Kullback-Leibler

- uma importante medida de dissimilaridade entre distribuições é a divergêngia de Kullback-Leibler (também chamada entropia relativa), dada por

$$\mathrm{D}_{\mathrm{KL}}(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

no caso contínuo, e

$$\mathrm{D}_{\mathrm{KL}}(p\|q) = \sum_{i=1}^{n+1} p_i \log \frac{p_i}{q_i}$$

no caso discreto.

- na teoria da informação, a entropia relativa D_{KL} nos diz a quantidade média de símbolos binários (0's ou 1's) necessários para codificar p usando um código otimizado para codificar q

- na teoria da informação, a entropia relativa D_{KL} nos diz a quantidade média de símbolos binários (0's ou 1's) necessários para codificar p usando um código otimizado para codificar q
- D_{KL} não é uma distância (é assimétrica), mas satisfaz $D_{KL} \geq 0$

- na teoria da informação, a entropia relativa $\mathrm{D_{KL}}$ nos diz a quantidade média de símbolos binários (0's ou 1's) necessários para codificar p usando um código otimizado para codificar q
- $\mathrm{D_{KL}}$ não é uma distância (é assimétrica), mas satisfaz $\mathrm{D_{KL}} \geq 0$
- existe relação com a métrica de Fisher:

$$g_{ij}(\theta) = \left. \frac{\partial^2 \mathrm{D_{KL}}(p_\theta \| p_\eta)}{\partial \eta_i \partial \eta_j} \right|_{\eta=\theta}$$

- na teoria da informação, a entropia relativa $\mathrm{D_{KL}}$ nos diz a quantidade média de símbolos binários (0's ou 1's) necessários para codificar p usando um código otimizado para codificar q
- $\mathrm{D_{KL}}$ não é uma distância (é assimétrica), mas satisfaz $\mathrm{D_{KL}} \geq 0$
- existe relação com a métrica de Fisher:

$$g_{ij}(\theta) = \left. \frac{\partial^2 \mathrm{D_{KL}}(p_\theta \| p_\eta)}{\partial \eta_i \partial \eta_j} \right|_{\eta=\theta}$$

- essencialmente, isso significa que a métrica de Fisher fornece uma aproximação de ordem 2 para a entropia relativa

Aplicações a aprendizado de máquina

- uma importante aplicação é o método do *gradiente natural*

Aplicações a aprendizado de máquina

- uma importante aplicação é o método do *gradiente natural*
- no problema de aprendizado temos

Aplicações a aprendizado de máquina

- uma importante aplicação é o método do *gradiente natural*
- no problema de aprendizado temos
  - um conjunto de dados de treinamento $\left\{(x_i, y_i)\right\}_{i=1}^{m}$

Aplicações a aprendizado de máquina

- uma importante aplicação é o método do *gradiente natural*
- no problema de aprendizado temos
  - um conjunto de dados de treinamento $\left\{(x_i, y_i)\right\}_{i=1}^{m}$
  - uma família parametrizada de funções $f_\theta(x) = y$ (geralmente dadas por uma rede neural)

Aplicações a aprendizado de máquina

- uma importante aplicação é o método do *gradiente natural*
- no problema de aprendizado temos
  - um conjunto de dados de treinamento $\left\{(x_i, y_i)\right\}_{i=1}^{m}$
  - uma família parametrizada de funções $f_\theta(x) = y$ (geralmente dadas por uma rede neural)
- tomamos uma função perda L, e consideramos o problema de encontrar θ que minimize $J(\theta) = \frac{1}{m}\sum_i L(f_\theta(x_i), y_i)$

Aplicações a aprendizado de máquina

- uma importante aplicação é o método do *gradiente natural*
- no problema de aprendizado temos
  - um conjunto de dados de treinamento $\{(x_i, y_i)\}_{i=1}^{m}$
  - uma família parametrizada de funções $f_\theta(x) = y$ (geralmente dadas por uma rede neural)
- tomamos uma função perda L, e consideramos o problema de encontrar θ que minimize $J(\theta) = \frac{1}{m}\sum_i L(f_\theta(x_i), y_i)$
  - isso costuma ser feito através de um método iterativo, usando a *descida por gradiente*:

$$\theta_N = \theta_{N-1} - \nabla_\theta J(\theta_{N-1})$$

- em geral, $\{f_\theta\}$ forma uma variedade estatística parametrizada por θ

- em geral, $\{f_\theta\}$ forma uma variedade estatística parametrizada por θ
- podemos, então, substituir o gradiente usual $\nabla_\theta J$ pelo *gradiente natural* que considera a estrutura geométrica estatística:

$$\tilde{\nabla}_\theta J(\theta) := I(\theta)^{-1} \nabla_\theta J(\theta)$$

- em geral, $\{f_\theta\}$ forma uma variedade estatística parametrizada por θ
- podemos, então, substituir o gradiente usual $\nabla_\theta J$ pelo *gradiente natural* que considera a estrutura geométrica estatística:

$$\tilde{\nabla}_\theta J(\theta) := I(\theta)^{-1} \nabla_\theta J(\theta)$$

- de fato, o gradiente natural funciona de forma eficiente, e aparenta evitar o "efeito platô" que ocorre com o gradiente usual

- em geral, $\{f_\theta\}$ forma uma variedade estatística parametrizada por θ
- podemos, então, substituir o gradiente usual $\nabla_\theta J$ pelo *gradiente natural* que considera a estrutura geométrica estatística:

$$\tilde{\nabla}_\theta J(\theta) := I(\theta)^{-1} \nabla_\theta J(\theta)$$

- de fato, o gradiente natural funciona de forma eficiente, e aparenta evitar o "efeito platô" que ocorre com o gradiente usual
- Referência: Shun-ichi Amari. *"Natural Gradient Works Efficiently in Learning"*. Em: Neural Computation 10.2 (1998)

Algumas referências

- AMARI, S. Information Geometry and Its Applications. Tokyo: Springer Japan, 2016. v. 194
- CALIN, O.; UDRIȘTE, C. Geometric Modeling in Probability and Statistics. Cham: Springer International Publishing, 2014.
- COSTA, S. I. R.; SANTOS, S. A.; STRAPASSON, J. E. Fisher information distance: A geometrical reading. Discrete Applied Mathematics, v. 197, p. 59–69, 2015.
- NIELSEN, F. An Elementary Introduction to Information Geometry. Entropy, v. 22, n. 10, p. 1100, out. 2020.