Métodos geométricos aplicados a ciências da informação

Fábio C. C. Meneghetti

Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas

22 de fevereiro de 2022

Parte I

Motivação

História

- Claude Shannon (1916–2001): entropia como medida de informação de uma variável aleatória: $H(X) = -\sum_i p_i \log p_i$.

História

- Claude Shannon (1916–2001): entropia como medida de informação de uma variável aleatória: $H(X) = -\sum_i p_i \log p_i$.
- Fisher (1890–1962): medida de informação sobre um parâmetro θ que uma variável aleatória carrega: $I(\theta) = \int_{\mathcal{X}} p_\theta(x) \left(\log \frac{\mathrm{d} p_\theta}{\mathrm{d}\theta}(x)\right)^2 \mathrm{d}x$

História

- Claude Shannon (1916–2001): entropia como medida de informação de uma variável aleatória: $H(X) = -\sum_i p_i \log p_i$.
- Fisher (1890–1962): medida de informação sobre um parâmetro θ que uma variável aleatória carrega: $I(\theta) = \int_{\mathcal{X}} p_\theta(x) \left(\log \frac{\mathrm{d} p_\theta}{\mathrm{d}\theta}(x)\right)^2 \mathrm{d}x$
- C. R. Rao (1920–): a informação de Fisher sobre múltiplos parâmetros $(\theta_1, \ldots, \theta_d)$ é uma métrica Riemanniana!

História

- Claude Shannon (1916–2001): entropia como medida de informação de uma variável aleatória: $H(X) = -\sum_i p_i \log p_i$.
- Fisher (1890–1962): medida de informação sobre um parâmetro θ que uma variável aleatória carrega: $I(\theta) = \int_{\mathcal{X}} p_\theta(x) \left(\log \frac{\mathrm{d} p_\theta}{\mathrm{d}\theta}(x)\right)^2 \mathrm{d}x$
- C. R. Rao (1920–): a informação de Fisher sobre múltiplos parâmetros $(\theta_1, \ldots, \theta_d)$ é uma métrica Riemanniana!
    - *métrica da informação de Fisher*

História

- Claude Shannon (1916–2001): entropia como medida de informação de uma variável aleatória: $H(X) = -\sum_i p_i \log p_i$.
- Fisher (1890–1962): medida de informação sobre um parâmetro θ que uma variável aleatória carrega: $I(\theta) = \int_{\mathcal{X}} p_\theta(x) \left(\log \frac{\mathrm{d}p_\theta}{\mathrm{d}\theta}(x)\right)^2 \mathrm{d}x$
- C. R. Rao (1920–): a informação de Fisher sobre múltiplos parâmetros $(\theta_1, \ldots, \theta_d)$ é uma métrica Riemanniana!
    - *métrica da informação de Fisher*
    - ela induz uma geometria sobre as distribuições de probabilidade

História

- Claude Shannon (1916–2001): entropia como medida de informação de uma variável aleatória: $H(X) = -\sum_i p_i \log p_i$.
- Fisher (1890–1962): medida de informação sobre um parâmetro θ que uma variável aleatória carrega: $I(\theta) = \int_{\mathcal{X}} p_\theta(x) \left(\log \frac{\mathrm{d} p_\theta}{\mathrm{d}\theta}(x)\right)^2 \mathrm{d}x$
- C. R. Rao (1920–): a informação de Fisher sobre múltiplos parâmetros $(\theta_1, \ldots, \theta_d)$ é uma métrica Riemanniana!
    - *métrica da informação de Fisher*
    - ela induz uma geometria sobre as distribuições de probabilidade
    - isso nos permite falar sobre distâncias e curvaturas no espaço das distribuições

História

- Claude Shannon (1916–2001): entropia como medida de informação de uma variável aleatória: $H(X) = -\sum_i p_i \log p_i$.
- Fisher (1890–1962): medida de informação sobre um parâmetro θ que uma variável aleatória carrega: $I(\theta) = \int_{\mathcal{X}} p_\theta(x) \left(\log \frac{\mathrm{d}p_\theta}{\mathrm{d}\theta}(x)\right)^2 \mathrm{d}x$
- C. R. Rao (1920–): a informação de Fisher sobre múltiplos parâmetros $(\theta_1, \ldots, \theta_d)$ é uma métrica Riemanniana!
  - *métrica da informação de Fisher*
  - ela induz uma geometria sobre as distribuições de probabilidade
  - isso nos permite falar sobre distâncias e curvaturas no espaço das distribuições
  - área de pesquisa: *geometria da informação*

Por exemplo:

- no caso das distribuições gaussianas univariadas...

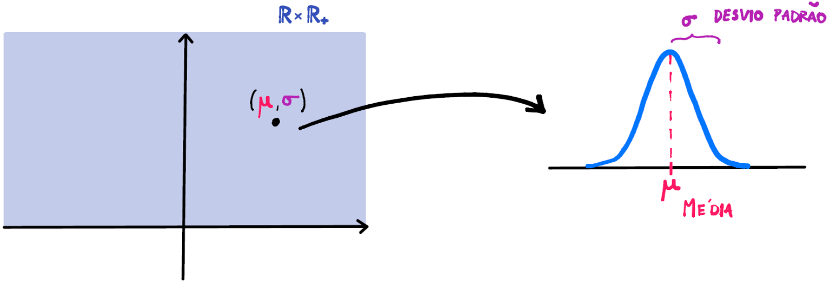
$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|x-\mu|^2}{2\sigma^2}\right),$$

Por exemplo:

- no caso das distribuições gaussianas univariadas...

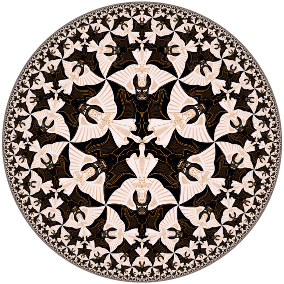
$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{|x-\mu|^2}{2\sigma^2}\right),$$

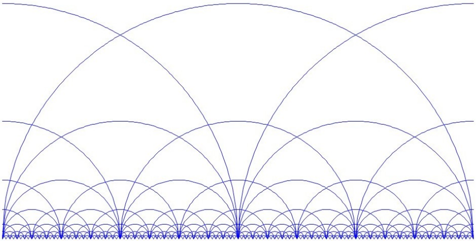
o espaço de parâmetros é



- com a métrica da informação de Fisher, obtemos uma geometria hiperbólica! (versão deformada do meio-plano de Poincaré)

- com a métrica da informação de Fisher, obtemos uma geometria hiperbólica! (versão deformada do meio-plano de Poincaré)





Parte II

Teoria

Distribuições de probabilidade

- distribuições de probabilidade

Distribuições de probabilidade

- distribuições de probabilidade
  - discretas: $X \in \{x_1, \ldots, x_n\}$, $\mathbb{P}[X = x_i] = p_i$, $\sum_i p_i = 1$.

Distribuições de probabilidade

- distribuições de probabilidade
  - discretas: $X \in \{x_1, \ldots, x_n\}$, $\mathbb{P}[X = x_i] = p_i$, $\sum_i p_i = 1$.
  - contínuas: $X \in \mathcal{X}$, $\mathbb{P}[X \in A] = \int_A p(x)\,\mathrm{d}x$.

Distribuições de probabilidade

- distribuições de probabilidade
  - discretas: $X \in \{x_1, \ldots, x_n\}$, $\mathbb{P}[X = x_i] = p_i$, $\sum_i p_i = 1$.
  - contínuas: $X \in \mathcal{X}$, $\mathbb{P}[X \in A] = \int_A p(x)\,\mathrm{d}x$.
- de forma mais geral, temos um espaço de probabilidade $(\mathcal{X}, \mathcal{F}, \mathbb{P})$, e uma medida σ-finita dominante μ.

Distribuições de probabilidade

- distribuições de probabilidade
  - discretas: $X \in \{x_1, \ldots, x_n\}$, $\mathbb{P}[X = x_i] = p_i$, $\sum_i p_i = 1$.
  - contínuas: $X \in \mathcal{X}$, $\mathbb{P}[X \in A] = \int_A p(x)\,\mathrm{d}x$.
- de forma mais geral, temos um espaço de probabilidade $(\mathcal{X}, \mathcal{F}, \mathbb{P})$, e uma medida σ-finita dominante μ.
  - A *função densidade* é a derivada de Radon-Nikodym $p(x) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}(x)$, $p\colon \mathcal{X} \to \mathbb{R}_+$ que satisfaz $\mathbb{P}(A) = \int_A p(x)\,\mathrm{d}\mu(x)$.

Distribuições de probabilidade

- distribuições de probabilidade
  - discretas: $X \in \{x_1, \ldots, x_n\}$, $\mathbb{P}[X = x_i] = p_i$, $\sum_i p_i = 1$.
  - contínuas: $X \in \mathcal{X}$, $\mathbb{P}[X \in A] = \int_A p(x)\,\mathrm{d}x$.
- de forma mais geral, temos um espaço de probabilidade $(\mathcal{X}, \mathcal{F}, \mathbb{P})$, e uma medida σ-finita dominante μ.
  - A *função densidade* é a derivada de Radon-Nikodym $p(x) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}(x)$, $p\colon \mathcal{X} \to \mathbb{R}_+$ que satisfaz $\mathbb{P}(A) = \int_A p(x)\,\mathrm{d}\mu(x)$.
  - $\mathcal{X}$ enumerável e μ_c medida de contagem $\implies$ distribuição discreta, $p =$ função massa, $\int_{x\in A} = \sum_{x \in A}$

Distribuições de probabilidade

- distribuições de probabilidade
  - discretas: $X \in \{x_1, \dots, x_n\}$, $\mathbb{P}[X = x_i] = p_i$, $\sum_i p_i = 1$.
  - contínuas: $X \in \mathcal{X}$, $\mathbb{P}[X \in A] = \int_A p(x)\,\mathrm{d}x$.
- de forma mais geral, temos um espaço de probabilidade $(\mathcal{X}, \mathcal{F}, \mathbb{P})$, e uma medida σ-finita dominante μ.
  - A *função densidade* é a derivada de Radon-Nikodym $p(x) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mu}(x)$, $p \colon \mathcal{X} \to \mathbb{R}_+$ que satisfaz $\mathbb{P}(A) = \int_A p(x)\,\mathrm{d}\mu(x)$.
  - $\mathcal{X}$ enumerável e μ_c medida de contagem $\implies$ distribuição discreta, $p =$ função massa, $\int_{x \in A} = \sum_{x \in A}$
  - $\mathcal{X} \subset \mathbb{R}^n$ e $\mu_{\mathcal{L}}$ medida de Lebesgue $\implies$ distribuição contínua, $p =$ função densidade

Modelos estatísticos

- um *modelo estatístico* é uma família parametrizada de distribuições $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ no espaço $(\mathcal{X}, \mathcal{F})$.

Modelos estatísticos

- um *modelo estatístico* é uma família parametrizada de distribuições $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ no espaço $(\mathcal{X}, \mathcal{F})$.
  - $\Theta \subset \mathbb{R}^d$ conjunto aberto de parâmetros.

Modelos estatísticos

- um *modelo estatístico* é uma família parametrizada de distribuições $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ no espaço $(\mathcal{X}, \mathcal{F})$.
  - $\Theta \subset \mathbb{R}^d$ conjunto aberto de parâmetros.
  - na vida real costuma-se tomar como modelo a família de funções densidade $\left\{p_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu} : \theta \in \Theta\right\}$.

Modelos estatísticos

- um *modelo estatístico* é uma família parametrizada de distribuições $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ no espaço $(\mathcal{X}, \mathcal{F})$.
  - $\Theta \subset \mathbb{R}^d$ conjunto aberto de parâmetros.
  - na vida real costuma-se tomar como modelo a família de funções densidade $\left\{p_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu} : \theta \in \Theta\right\}$.
  - vamos considerar modelos estatísticos *regulares*:

Modelos estatísticos

- um *modelo estatístico* é uma família parametrizada de distribuições $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ no espaço $(\mathcal{X}, \mathcal{F})$.
    - $\Theta \subset \mathbb{R}^d$ conjunto aberto de parâmetros.
    - na vida real costuma-se tomar como modelo a família de funções densidade $\left\{p_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu} : \theta \in \Theta\right\}$.
    - vamos considerar modelos estatísticos *regulares*:
        - a função $\theta \mapsto p_\theta$ é C^∞

Modelos estatísticos

- um *modelo estatístico* é uma família parametrizada de distribuições $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ no espaço $(\mathcal{X}, \mathcal{F})$.
  - $\Theta \subset \mathbb{R}^d$ conjunto aberto de parâmetros.
  - na vida real costuma-se tomar como modelo a família de funções densidade $\left\{p_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu} : \theta \in \Theta\right\}$.
  - vamos considerar modelos estatísticos *regulares*:
    - a função $\theta \mapsto p_\theta$ é C^∞
    - $p_\theta(x) > 0$ para todo x, θ

Métrica Riemanniana

Vamos agora adicionar uma estrutura Riemanniana ao espaço $\Theta \simeq \mathcal{P}$.

Métrica Riemanniana

Vamos agora adicionar uma estrutura Riemanniana ao espaço $\Theta \simeq \mathcal{P}$.

- denote por $\ell_\theta(x) := \log p_\theta(x)$ a função log-probabilidade

Métrica Riemanniana

Vamos agora adicionar uma estrutura Riemanniana ao espaço $\Theta \simeq \mathcal{P}$.

- denote por $\ell_\theta(x) := \log p_\theta(x)$ a função log-probabilidade
- a *métrica de Fisher* é a métrica Riemanniana (produto interno em $T_\theta\Theta$) dada por

$$g_\theta(V, W) := \int_{\mathcal{X}} p_\theta(x) \frac{\partial \ell_\theta(x)}{\partial V} \frac{\partial \ell_\theta(x)}{\partial W} \, \mathrm{d}\mu(x).$$

Métrica Riemanniana

Vamos agora adicionar uma estrutura Riemanniana ao espaço $\Theta \simeq \mathcal{P}$.

- denote por $\ell_\theta(x) := \log p_\theta(x)$ a função log-probabilidade
- a *métrica de Fisher* é a métrica Riemanniana (produto interno em $T_\theta\Theta$) dada por

$$g_\theta(V, W) := \int_{\mathcal{X}} p_\theta(x) \frac{\partial \ell_\theta(x)}{\partial V} \frac{\partial \ell_\theta(x)}{\partial W} \, \mathrm{d}\mu(x).$$

- na base coordenada local $\left\{ e_i = \frac{\partial}{\partial \theta_i}\Big|_\theta \right\}_i$, a matriz da métrica é chamada de *matriz de Fisher* $I(\theta)$, com elementos

$$g_{ij}(\theta) := g_\theta(e_i, e_j) = \int_{\mathcal{X}} p_\theta(x) \frac{\partial \ell_\theta(x)}{\partial \theta_i} \frac{\partial \ell_\theta(x)}{\partial \theta_j} \, \mathrm{d}\mu(x)$$

Métrica Riemanniana

Vamos agora adicionar uma estrutura Riemanniana ao espaço $\Theta \simeq \mathcal{P}$.

- denote por $\ell_\theta(x) := \log p_\theta(x)$ a função log-probabilidade
- a *métrica de Fisher* é a métrica Riemanniana (produto interno em $T_\theta\Theta$) dada por

$$g_\theta(V, W) := \int_{\mathcal{X}} p_\theta(x) \frac{\partial \ell_\theta(x)}{\partial V} \frac{\partial \ell_\theta(x)}{\partial W} \, \mathrm{d}\mu(x).$$

- na base coordenada local $\left\{ e_i = \frac{\partial}{\partial \theta_i}\Big|_\theta \right\}_i$, a matriz da métrica é chamada de *matriz de Fisher* $I(\theta)$, com elementos

$$g_{ij}(\theta) := g_\theta(e_i, e_j) = \int_{\mathcal{X}} p_\theta(x) \frac{\partial \ell_\theta(x)}{\partial \theta_i} \frac{\partial \ell_\theta(x)}{\partial \theta_j} \, \mathrm{d}\mu(x)$$

  - $I(\theta)$ é simétrica e positiva-definida.

Interpretações da métrica de Fisher

- Em estatística

Interpretações da métrica de Fisher

- Em estatística
  - O *escore* de θ em x é o gradiente da função log-probabilidade: $s_\theta(x) = \nabla_\theta \log p_\theta(x) = \left(\frac{\partial \ell_\theta}{\partial \theta_1}, \ldots, \frac{\partial \ell_\theta}{\partial \theta_d}\right)^\top (x)$
    - mede sensibilidade a mudanças nos parâmetros da função log-probabilidade

Interpretações da métrica de Fisher

- Em estatística
  - O *escore* de θ em x é o gradiente da função log-probabilidade: $s_\theta(x) = \nabla_\theta \log p_\theta(x) = \left(\frac{\partial \ell_\theta}{\partial \theta_1}, \ldots, \frac{\partial \ell_\theta}{\partial \theta_d}\right)^\top(x)$
    - mede sensibilidade a mudanças nos parâmetros da função log-probabilidade
  - A matriz de Fisher é a matriz de covariância do escore: $I(\theta) = cov(s_\theta, s_\theta) = \mathbb{E}[s_\theta \cdot s_\theta^\top]$
    - é um limitante inferior para a covariância de um estimador não-enviesado $\hat{\theta}$ (*Limitante de Cramér-Rao*):

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

- em teoria da informação

- em teoria da informação
  - entropia relativa (divergência de Kullback-Leibler): $D_{\mathsf{KL}}(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}\mu(x)$

- em teoria da informação
  - entropia relativa (divergência de Kullback-Leibler): $D_{\mathrm{KL}}(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}\mu(x)$
    - diz quantos bits (na verdade nats), em média, são necessários para codificar p usando um código otimizado para codificar q

- em teoria da informação
  - entropia relativa (divergência de Kullback-Leibler): $D_{\mathsf{KL}}(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}\mu(x)$
    - diz quantos bits (na verdade nats), em média, são necessários para codificar p usando um código otimizado para codificar q
  - a matriz de Fisher é o Hessiano diagonal da entropia relativa:

$$g_{ij}(\theta_0) = \frac{\partial^2}{\partial\theta_i \partial\theta_j} D_{\mathsf{KL}}(p_{\theta_0} \| p_\theta) \Big|_{\theta=\theta_0}$$

Unicidade

- uma *estatística* é um mapa mensurável $\kappa\colon \mathcal{X} \to \mathcal{Y}$

Unicidade

- uma *estatística* é um mapa mensurável $\kappa\colon \mathcal{X} \to \mathcal{Y}$
- induz uma família de distribuições de probabilidade *empurradas* $\kappa_*\mathbb{P}_\theta$ em $\mathcal{Y}$: $\kappa_*\mathbb{P}_\theta(A) := \mathbb{P}_\theta(\kappa^{-1}(A))$

Unicidade

- uma *estatística* é um mapa mensurável $\kappa\colon \mathcal{X} \to \mathcal{Y}$
- induz uma família de distribuições de probabilidade *empurradas* $\kappa_* \mathbb{P}_\theta$ em $\mathcal{Y}$: $\kappa_* \mathbb{P}_\theta(A) := \mathbb{P}_\theta(\kappa^{-1}(A))$
  - por sua vez, obtemos novas funções densidade $\tilde{p}_\theta = \frac{\mathrm{d}\kappa_* \mathbb{P}_\theta}{\mathrm{d}\mu}$

Unicidade

- uma *estatística* é um mapa mensurável $\kappa\colon \mathcal{X} \to \mathcal{Y}$
- induz uma família de distribuições de probabilidade *empurradas* $\kappa_* \mathbb{P}_\theta$ em $\mathcal{Y}$: $\kappa_* \mathbb{P}_\theta(A) := \mathbb{P}_\theta(\kappa^{-1}(A))$
  - por sua vez, obtemos novas funções densidade $\tilde{p}_\theta = \frac{\mathrm{d}\kappa_* \mathbb{P}_\theta}{\mathrm{d}\mu}$
- a estatística κ é dita *suficiente* se $p_\theta(x) = \tilde{p}_\theta(\kappa(x))h(x)$ para alguma função h independente de θ.

Unicidade

- uma *estatística* é um mapa mensurável $\kappa \colon \mathcal{X} \to \mathcal{Y}$
- induz uma família de distribuições de probabilidade *empurradas* $\kappa_* \mathbb{P}_\theta$ em $\mathcal{Y}$: $\kappa_* \mathbb{P}_\theta(A) := \mathbb{P}_\theta(\kappa^{-1}(A))$
    - por sua vez, obtemos novas funções densidade $\tilde{p}_\theta = \frac{\mathrm{d}\kappa_* \mathbb{P}_\theta}{\mathrm{d}\mu}$
- a estatística κ é dita *suficiente* se $p_\theta(x) = \tilde{p}_\theta(\kappa(x))h(x)$ para alguma função h independente de θ.

Teorema (Chentsov)

A métrica de Fisher é a única métrica Riemanniana, a menos de uma constante, invariante por estatísticas suficientes.

Geodésicas

- geodésicas são “linhas retas” nas variedades Riemannianas

Geodésicas

- geodésicas são “linhas retas” nas variedades Riemannianas
- dados dois pontos $p_\theta, p_{\theta'}$, a curva $\gamma\colon [0,1] \to \mathcal{P}$ ligando-os, que minimiza comprimento, é um segmento de geodésica

Geodésicas

- geodésicas são "linhas retas" nas variedades Riemannianas
- dados dois pontos $p_\theta, p_{\theta'}$, a curva $\gamma\colon [0,1] \to \mathcal{P}$ ligando-os, que minimiza comprimento, é um segumento de geodésica
    - o comprimento é dado por $\ell(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\,\mathrm{d}t$

Geodésicas

- geodésicas são "linhas retas" nas variedades Riemannianas
- dados dois pontos $p_\theta, p_{\theta'}$, a curva $\gamma\colon [0,1] \to \mathcal{P}$ ligando-os, que minimiza comprimento, é um segumento de geodésica
  - o comprimento é dado por $\ell(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, \mathrm{d}t$
  - essa noção define uma distância *geodésica*, chamada *distância de Fisher-Rao*[1] na geometria da informação:

$$d_{\mathrm{FR}}(p_\theta, p_\theta') = \min_\gamma \left\{ \ell(\gamma) : \gamma(0) = p_\theta,\ \gamma(1) = p_\theta' \right\}$$

[1] quando a variedade é completa e conexa por caminhos

Geodésicas

- geodésicas são "linhas retas" nas variedades Riemannianas
- dados dois pontos $p_\theta, p_{\theta'}$, a curva $\gamma\colon [0,1] \to \mathcal{P}$ ligando-os, que minimiza comprimento, é um segumento de geodésica
  - o comprimento é dado por $\ell(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, \mathrm{d}t$
  - essa noção define uma distância *geodésica*, chamada *distância de Fisher-Rao*[1] na geometria da informação:

$$d_{\mathsf{FR}}(p_\theta, p'_\theta) = \min_\gamma \left\{ \ell(\gamma) : \gamma(0) = p_\theta,\ \gamma(1) = p'_\theta \right\}$$

- formalmente, são curvas que têm derivada covariante zero: $\nabla_{\dot\gamma}\dot\gamma \equiv 0$

[1] quando a variedade é completa e conexa por caminhos

Geodésicas

- geodésicas são "linhas retas" nas variedades Riemannianas
- dados dois pontos $p_\theta, p_{\theta'}$, a curva $\gamma\colon [0,1] \to \mathcal{P}$ ligando-os, que minimiza comprimento, é um segumento de geodésica
  - o comprimento é dado por $\ell(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\,\mathrm{d}t$
  - essa noção define uma distância *geodésica*, chamada *distância de Fisher-Rao*[1] na geometria da informação:

$$d_{\mathsf{FR}}(p_\theta, p'_\theta) = \min_{\gamma} \left\{ \ell(\gamma) : \gamma(0) = p_\theta,\ \gamma(1) = p'_\theta \right\}$$

- formalmente, são curvas que têm derivada covariante zero: $\nabla_{\dot\gamma}\dot\gamma \equiv 0$
  - ∇ é a conexão dada pelos símbolos de Christoffel $\Gamma_{ij,k}(\theta) = \mathbb{E}_{p_\theta}[(\partial_i\partial_j\ell_\theta + \frac{1}{2}\partial_i\ell_\theta\partial_j\ell_\theta)\,\partial_k\ell_\theta]$

[1] quando a variedade é completa e conexa por caminhos

Famílias exponenciais

- uma família exponencial $\{p_\theta : \theta \in \Theta\}$ com *parâmetros naturais* $\theta \in \Theta \subset \mathbb{R}^d$ é dada por

$$p_\theta(x) := \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right), \quad x \in \mathcal{X}$$

Famílias exponenciais

- uma família exponencial $\{p_\theta : \theta \in \Theta\}$ com *parâmetros naturais* $\theta \in \Theta \subset \mathbb{R}^d$ é dada por

$$p_\theta(x) := \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right), \quad x \in \mathcal{X}$$

  - $t(x)$ estatística suficiente

Famílias exponenciais

- uma família exponencial $\{p_\theta : \theta \in \Theta\}$ com *parâmetros naturais* $\theta \in \Theta \subset \mathbb{R}^d$ é dada por

$$p_\theta(x) := \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right), \quad x \in \mathcal{X}$$

  - $t(x)$ estatística suficiente
  - $F(\theta)$ função estritamente convexa

Famílias exponenciais

- uma família exponencial $\{p_\theta : \theta \in \Theta\}$ com *parâmetros naturais* $\theta \in \Theta \subset \mathbb{R}^d$ é dada por

$$p_\theta(x) := \exp\big(\langle t(x), \theta\rangle - F(\theta) + k(x)\big)\,, \quad x \in \mathcal{X}$$

  - $t(x)$ estatística suficiente
  - $F(\theta)$ função estritamente convexa
  - $k(x)$ qualquer

Famílias exponenciais

- uma família exponencial $\{p_\theta : \theta \in \Theta\}$ com *parâmetros naturais* $\theta \in \Theta \subset \mathbb{R}^d$ é dada por

$$p_\theta(x) := \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right), \quad x \in \mathcal{X}$$

  - $t(x)$ estatística suficiente
  - $F(\theta)$ função estritamente convexa
  - $k(x)$ qualquer
- há uma expressão simples para a matriz de Fisher: $g_{ij}(\theta) = \frac{\partial^2 F(\theta)}{\partial\theta_i \partial\theta_j}$

Exemplos

Famílias exponenciais englobam muitos casos

Exemplos

Famílias exponenciais englobam muitos casos

- distribuições normais: $t(x) = (x, x^2)$, $(\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2})$, $F = \frac{-\theta_1^2}{4\theta_2} + \frac{1}{2} \log \frac{-\pi}{\theta_2}$, $k(x) = 0$

Exemplos

Famílias exponenciais englobam muitos casos

- distribuições normais: $t(x) = (x, x^2)$, $(\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2})$, $F = \frac{-\theta_1^2}{4\theta_2} + \frac{1}{2}\log\frac{-\pi}{\theta_2}$, $k(x) = 0$
- distribuições poisson: $t(x) = x$, $k(x) = x!$, $\theta = \log\lambda$, $F(\theta) = \lambda = e^\theta$

Exemplos

Famílias exponenciais englobam muitos casos

- distribuições normais: $t(x) = (x, x^2)$, $(\theta_1, \theta_2) = (\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2})$, $F = \frac{-\theta_1^2}{4\theta_2} + \frac{1}{2} \log \frac{-\pi}{\theta_2}$, $k(x) = 0$
- distribuições poisson: $t(x) = x$, $k(x) = x!$, $\theta = \log \lambda$, $F(\theta) = \lambda = e^\theta$
- gama, beta, exponencial, etc.

Parâmetros duais

- famílias exponenciais têm parametrizações duais: $\eta = \nabla_\theta F(\theta)$

Parâmetros duais

- famílias exponenciais têm parametrizações duais: $\eta = \nabla_\theta F(\theta)$
- é possível voltar para os parâmetros naturais via $\theta = \nabla_\eta F^*(\eta)$

Parâmetros duais

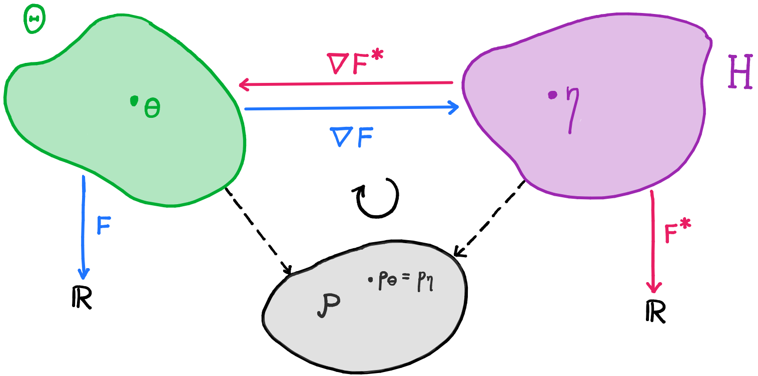
- famílias exponenciais têm parametrizações duais: $\eta = \nabla_\theta F(\theta)$
- é possível voltar para os parâmetros naturais via $\theta = \nabla_\eta F^*(\eta)$
  - onde $F^*(\eta) := \langle \theta, \eta \rangle - F(\theta)$ é a transformada de Legendre

Parâmetros duais

- famílias exponenciais têm parametrizações duais: $\eta = \nabla_\theta F(\theta)$
- é possível voltar para os parâmetros naturais via $\theta = \nabla_\eta F^*(\eta)$
  - onde $F^*(\eta) := \langle \theta, \eta \rangle - F(\theta)$ é a transformada de Legendre
  - $F^*(\eta) = \int_{\mathcal{X}} p_\theta(x) \ell_\theta(x) \, \mathrm{d}\mu(x)$ é a entropia de Shannon negativa

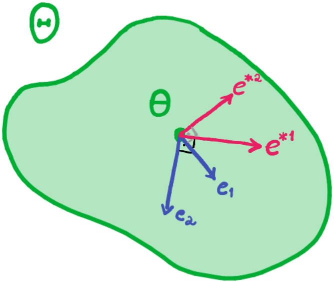
Parâmetros duais

- famílias exponenciais têm parametrizações duais: $\eta = \nabla_\theta F(\theta)$
- é possível voltar para os parâmetros naturais via $\theta = \nabla_\eta F^*(\eta)$
    - onde $F^*(\eta) := \langle \theta, \eta \rangle - F(\theta)$ é a transformada de Legendre
    - $F^*(\eta) = \int_{\mathcal{X}} p_\theta(x) \ell_\theta(x) \, d\mu(x)$ é a entropia de Shannon negativa

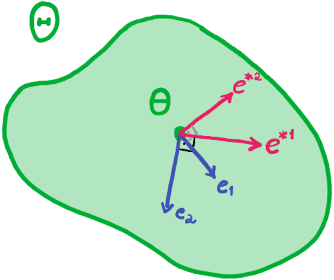


- as parametrizações θ e η de fato são duais, no sentido que:
  - $e_i = \frac{\partial}{\partial \theta_i}, \; e^{*j} = \frac{\partial}{\partial \eta_j} \implies g(e_i, e^{*j}) = \delta_{ij}$.

- as parametrizações θ e η de fato são duais, no sentido que:
  - $e_i = \frac{\partial}{\partial \theta_i}, \ e^{*j} = \frac{\partial}{\partial \eta_j} \implies g(e_i, e^{*j}) = \delta_{ij}.$

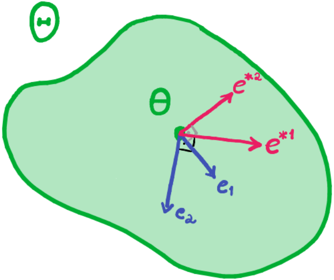


- as parametrizações θ e η de fato são duais, no sentido que:
  - $e_i = \frac{\partial}{\partial \theta_i}$, $e^{*j} = \frac{\partial}{\partial \eta_j} \implies g(e_i, e^{*j}) = \delta_{ij}$.



- temos que $g_{ij}(\theta) = \frac{\partial \eta_i}{\partial \theta_j}$, e $g^{*ij}(\eta) = \frac{\partial \theta_i}{\partial \eta_j}$.

- as parametrizações θ e η de fato são duais, no sentido que:
  - $e_i = \frac{\partial}{\partial \theta_i}, \ e^{*j} = \frac{\partial}{\partial \eta_j} \implies g(e_i, e^{*j}) = \delta_{ij}$.



- temos que $g_{ij}(\theta) = \frac{\partial \eta_i}{\partial \theta_j}$, e $g^{*ij}(\eta) = \frac{\partial \theta_i}{\partial \eta_j}$.
- **Observação:** outras famílias, como as misturas, também têm parâmetros duais

Geodésicas duais

- as parametrizações duais induzem duas geometrias *dualmente planas*: a (e)-geometria e a (m)-geometria
  - são simplesmente as geometrias planas das parametrizações θ e η

Geodésicas duais

- as parametrizações duais induzem duas geometrias *dualmente planas*: a (e)-geometria e a (m)-geometria
  - são simplesmente as geometrias planas das parametrizações θ e η
  - uma *(e)-geodésica* é uma reta nos parâmetros θ

Geodésicas duais

- as parametrizações duais induzem duas geometrias *dualmente planas*: a (e)-geometria e a (m)-geometria
  - são simplesmente as geometrias planas das parametrizações θ e η
  - uma *(e)-geodésica* é uma reta nos parâmetros θ
  - uma *(m)-geodésica* é uma reta nos parâmetros η

Geodésicas duais

- as parametrizações duais induzem duas geometrias *dualmente planas*: a (e)-geometria e a (m)-geometria
  - são simplesmente as geometrias planas das parametrizações θ e η
  - uma *(e)-geodésica* é uma reta nos parâmetros θ
  - uma *(m)-geodésica* é uma reta nos parâmetros η
- as geometrias (e portanto as geodésicas) duais são descritas por uma família conexões ∇^α, $\alpha \in \mathbb{R}$, com coeficientes

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathbb{E}_{p_\theta}\left[\left(\partial_i\partial_j\ell_\theta + \tfrac{1-\alpha}{2}\partial_i\ell_\theta\partial_j\ell_\theta\right)\partial_k\ell_\theta\right]$$

Geodésicas duais

- as parametrizações duais induzem duas geometrias *dualmente planas*: a (e)-geometria e a (m)-geometria
  - são simplesmente as geometrias planas das parametrizações θ e η
  - uma *(e)-geodésica* é uma reta nos parâmetros θ
  - uma *(m)-geodésica* é uma reta nos parâmetros η
- as geometrias (e portanto as geodésicas) duais são descritas por uma família conexões ∇^{α}, $\alpha \in \mathbb{R}$, com coeficientes

$$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathbb{E}_{p_\theta}\left[\left(\partial_i\partial_j\ell_\theta + \tfrac{1-\alpha}{2}\partial_i\ell_\theta\partial_j\ell_\theta\right)\partial_k\ell_\theta\right]$$

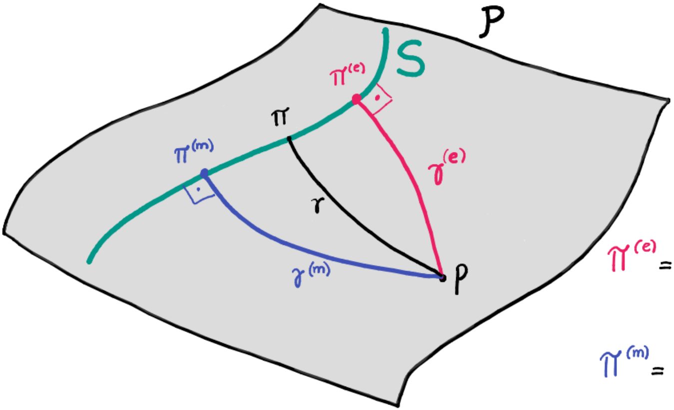
  - $\alpha = 0 \implies$ geodésicas usuais
  - $\alpha = 1 \implies$ (e)-geodésicas
  - $\alpha = -1 \implies$ (m)-geodésicas

Projeções de informação

- as (e)-projeções e (m)-projeções ortogonais em uma subvariedade S minimizam as entropias relativas duais $D_{\mathrm{KL}}(p\|q)$ e $D_{\mathrm{KL}}(q\|p)$

Projeções de informação

- as (e)-projeções e (m)-projeções ortogonais em uma subvariedade S minimizam as entropias relativas duais $D_{\mathrm{KL}}(p\|q)$ e $D_{\mathrm{KL}}(q\|p)$



$$\pi^{(e)} = \underset{q\in S}{\mathrm{argmin}}\, D_{KL}(q\|p)$$

$$\pi^{(m)} = \underset{q\in S}{\mathrm{argmin}}\, D_{KL}(p\|q)$$

Distribuições discretas

- os pontos são distribuições discretas $p\colon \mathcal{X} \to [0,1]$
  - $\mathcal{X} = \{x_1, \ldots, x_{d+1}\}$
  - $p(x_i) = p_i \in (0,1), \quad \sum_i p_i = 1$

Distribuições discretas

- os pontos são distribuições discretas $p\colon \mathcal{X} \to [0,1]$
  - $\mathcal{X} = \{x_1, \dots, x_{d+1}\}$
  - $p(x_i) = p_i \in (0,1), \quad \sum_i p_i = 1$
- essa variedade pode ser identificada com o interior do *simplexo padrão*

$$\mathring{\mathbb{A}}^d = \left\{ p \in \mathbb{R}^{d+1} \,\middle|\, 0 < p_j < 1, \quad \sum_i p_i = 1 \right\}$$

Distribuições discretas

- os pontos são distribuições discretas $p\colon \mathcal{X} \to [0,1]$
  - $\mathcal{X} = \{x_1, \dots, x_{d+1}\}$
  - $p(x_i) = p_i \in (0,1), \quad \sum_i p_i = 1$
- essa variedade pode ser identificada com o interior do *simplexo padrão*

$$\mathring{\Delta}^d = \left\{ p \in \mathbb{R}^{d+1} \;\middle|\; 0 < p_j < 1, \quad \sum_i p_i = 1 \right\}$$

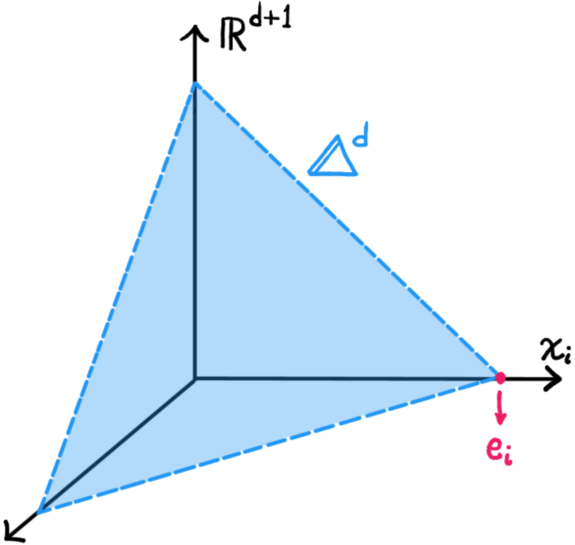
- parametrização
  - domínio $\Theta = \left\{ (p_1, \dots, p_d) \in \mathbb{R}^d_+ \mid \sum_i p_i < 1 \right\}$

Distribuições discretas

- os pontos são distribuições discretas $p \colon \mathcal{X} \to [0,1]$
  - $\mathcal{X} = \{x_1, \ldots, x_{d+1}\}$
  - $p(x_i) = p_i \in (0,1), \quad \sum_i p_i = 1$
- essa variedade pode ser identificada com o interior do *simplexo padrão*

$$\mathring{\Delta}^d = \left\{ p \in \mathbb{R}^{d+1} \;\middle|\; 0 < p_j < 1, \quad \sum_i p_i = 1 \right\}$$

- parametrização
  - domínio $\Theta = \left\{ (p_1, \ldots, p_d) \in \mathbb{R}^d_+ \;\middle|\; \sum_i p_i < 1 \right\}$
  - $\phi(p_1, \ldots, p_d) = (p_1, \ldots, p_d, p_{d+1}), \quad$ com $p_{d+1} = 1 - \sum_{i=1}^d p_i$



$\mathbb{R}^{d+1}$
Δ^d
x_i
e_i

- matriz de Fisher:

$$g_{ij}(p) = \frac{1}{p_{d+1}} + \frac{\delta_{ij}}{p_i}$$

- matriz de Fisher:

$$g_{ij}(p) = \frac{1}{p_{d+1}} + \frac{\delta_{ij}}{p_i}$$

- para calcular a *distância de Fisher-Rao*, fazemos uma reparametrização $z_i = 2\sqrt{p_i}$,

- matriz de Fisher:

$$g_{ij}(p) = \frac{1}{p_{d+1}} + \frac{\delta_{ij}}{p_i}$$

- para calcular a *distância de Fisher-Rao*, fazemos uma reparametrização $z_i = 2\sqrt{p_i}$,
    - que leva pontos $p = (p_1, \dots, p_{d+1}) \in \mathbb{\Delta}^d$ em pontos z no setor positivo da esfera

$$\mathbb{S}^d_{2,+} = \left\{ z \in \mathbb{R}^{d+1}_+ \;\middle|\; \sum_{i=1}^{d+1} z_i^2 = 4 \right\}$$

- matriz de Fisher:

$$g_{ij}(p) = \frac{1}{p_{d+1}} + \frac{\delta_{ij}}{p_i}$$

- para calcular a *distância de Fisher-Rao*, fazemos uma reparametrização $z_i = 2\sqrt{p_i}$,
  - que leva pontos $p = (p_1, \ldots, p_{d+1}) \in \mathbb{\Delta}^d$ em pontos z no setor positivo da esfera

$$\mathbb{S}^d_{2,+} = \left\{ z \in \mathbb{R}^{d+1}_+ \;\middle|\; \sum_{i=1}^{d+1} z_i^2 = 4 \right\}$$

  - nessa nova parametrização, a métrica de Fisher é a métrica esférica usual de $\mathbb{S}^d_{2,+} \subset \mathbb{R}^{d+1}$: $\quad g_{ij}(z) = \left\langle \frac{\partial z}{\partial p_i}, \frac{\partial z}{\partial p_j} \right\rangle$

- matriz de Fisher:

$$g_{ij}(p) = \frac{1}{p_{d+1}} + \frac{\delta_{ij}}{p_i}$$

- para calcular a *distância de Fisher-Rao*, fazemos uma reparametrização $z_i = 2\sqrt{p_i}$,
    - que leva pontos $p = (p_1, \ldots, p_{d+1}) \in \mathbb{A}^d$ em pontos z no setor positivo da esfera

$$\mathbb{S}^d_{2,+} = \left\{ z \in \mathbb{R}^{d+1}_+ \,\middle|\, \sum_{i=1}^{d+1} z_i^2 = 4 \right\}$$

    - nessa nova parametrização, a métrica de Fisher é a métrica esférica usual de $\mathbb{S}^d_{2,+} \subset \mathbb{R}^{d+1}$: $\quad g_{ij}(z) = \left\langle \frac{\partial z}{\partial p_i}, \frac{\partial z}{\partial p_j} \right\rangle$
    - portanto a distância de Fisher-Rao entre p e q pode facilmente ser calculada como o comprimento do arco ligando z_p e z_q, que equivale a

- matriz de Fisher:

$$g_{ij}(p) = \frac{1}{p_{d+1}} + \frac{\delta_{ij}}{p_i}$$

- para calcular a *distância de Fisher-Rao*, fazemos uma reparametrização $z_i = 2\sqrt{p_i}$,
  - que leva pontos $p = (p_1, \dots, p_{d+1}) \in \triangle^d$ em pontos z no setor positivo da esfera

$$\mathbb{S}^d_{2,+} = \left\{ z \in \mathbb{R}^{d+1}_+ \;\middle|\; \sum_{i=1}^{d+1} z_i^2 = 4 \right\}$$

  - nessa nova parametrização, a métrica de Fisher é a métrica esférica usual de $\mathbb{S}^d_{2,+} \subset \mathbb{R}^{d+1}$: $g_{ij}(z) = \left\langle \frac{\partial z}{\partial p_i}, \frac{\partial z}{\partial p_j} \right\rangle$
  - portanto a distância de Fisher-Rao entre p e q pode facilmente ser calculada como o comprimento do arco ligando z_p e z_q, que equivale a

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

- um fato interessante é que o *comprimento da corda* ligando z_p e z_q fornece uma boa aproximação:

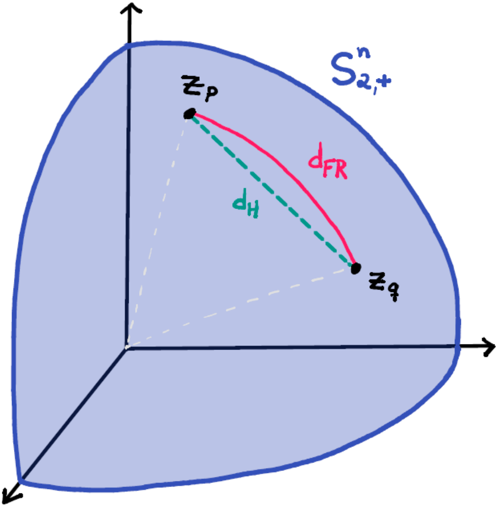
- um fato interessante é que o *comprimento da corda* ligando z_p e z_q fornece uma boa aproximação:

$$\left\| z_p - z_q \right\| = 2 \left(\sum_{i=1}^{d+1} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2 \right)^{1/2}$$

- um fato interessante é que o *comprimento da corda* ligando z_p e z_q fornece uma boa aproximação:

$$\left\| z_p - z_q \right\| = 2 \left(\sum_{i=1}^{d+1} \left(\sqrt{p_i} - \sqrt{q_i} \right)^2 \right)^{1/2}$$

- essa distância, sem o fator 2, é chamada *distância de Hellinger* d_{H}


$S^n_{2,+}$
z_p
d_{FR}
d_H
z_q

Parte III

Aplicações

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:
- temos uma família parametrizada de funções $\left\{f_\theta\colon \mathcal{X} \to \mathbb{R}^K\right\}_{\theta\in\Theta}$

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:
- temos uma família parametrizada de funções $\left\{f_\theta \colon \mathcal{X} \to \mathbb{R}^K\right\}_{\theta\in\Theta}$
  - $\mathcal{X} \subset \mathbb{R}^n$ é o espaço dos *vetores de característica* (ex: imagens)

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:
- temos uma família parametrizada de funções $\left\{ f_\theta \colon \mathcal{X} \to \mathbb{R}^K \right\}_{\theta \in \Theta}$
  - $\mathcal{X} \subset \mathbb{R}^n$ é o espaço dos *vetores de característica* (ex: imagens)
  - K é o número de classes (ex: cachorro, gato, etc.)

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:
- temos uma família parametrizada de funções $\left\{ f_\theta \colon \mathcal{X} \to \mathbb{R}^K \right\}_{\theta \in \Theta}$
  - $\mathcal{X} \subset \mathbb{R}^n$ é o espaço dos *vetores de característica* (ex: imagens)
  - K é o número de classes (ex: cachorro, gato, etc.)
  - $z = f_\theta(x)$ é chamado vetor *escore*

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:
- temos uma família parametrizada de funções $\left\{ f_\theta \colon \mathcal{X} \to \mathbb{R}^K \right\}_{\theta \in \Theta}$
  - $\mathcal{X} \subset \mathbb{R}^n$ é o espaço dos *vetores de característica* (ex: imagens)
  - K é o número de classes (ex: cachorro, gato, etc.)
  - $z = f_\theta(x)$ é chamado vetor *escore*
  - $\theta \in \Theta$ são o *parâmetros da máquina* (geralmente dados por uma rede neural)

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:
- temos uma família parametrizada de funções $\left\{ f_\theta \colon \mathcal{X} \to \mathbb{R}^K \right\}_{\theta \in \Theta}$
  - $\mathcal{X} \subset \mathbb{R}^n$ é o espaço dos *vetores de característica* (ex: imagens)
  - K é o número de classes (ex: cachorro, gato, etc.)
  - $z = f_\theta(x)$ é chamado vetor *escore*
  - $\theta \in \Theta$ são o *parâmetros da máquina* (geralmente dados por uma rede neural)
- transformamos o vetor escore em um vetor de probabilidades através da função *softmax* $\sigma \colon \mathbb{R}^K \to \mathring{\Delta}^{K-1}$ dada em coordenadas por

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{i=1}^K e^{z_i}}$$

Funções perda para classificadores

- *problemas de classificação* em aprendizado de máquina:
- temos uma família parametrizada de funções $\left\{ f_\theta \colon \mathcal{X} \to \mathbb{R}^K \right\}_{\theta \in \Theta}$
  - $\mathcal{X} \subset \mathbb{R}^n$ é o espaço dos *vetores de característica* (ex: imagens)
  - K é o número de classes (ex: cachorro, gato, etc.)
  - $z = f_\theta(x)$ é chamado vetor *escore*
  - $\theta \in \Theta$ são o *parâmetros da máquina* (geralmente dados por uma rede neural)
- transformamos o vetor escore em um vetor de probabilidades através da função *softmax* $\sigma \colon \mathbb{R}^K \to \mathring{\Delta}^{K-1}$ dada em coordenadas por

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{i=1}^K e^{z_i}}$$

  - podemos interpretar $\sigma(z)_i$ como a probabilidade do vetor pertencer à classe i

- treinamento supervisionado: somos fornecidos com um *conjunto de treinamento* $\{(x_i, y_i)\}_{i=1}^{m} \subset \mathcal{X} \times \{1, \ldots, K\}$

- treinamento supervisionado: somos fornecidos com um *conjunto de treinamento* $\{(x_i, y_i)\}_{i=1}^{m} \subset \mathcal{X} \times \{1, \ldots, K\}$
- tomamos uma função perda $\mathcal{L}\colon \mathbb{A}^{K-1} \times \mathbb{A}^{K-1} \to \mathbb{R}_+$

- treinamento supervisionado: somos fornecidos com um *conjunto de treinamento* $\{(x_i, y_i)\}_{i=1}^{m} \subset \mathcal{X} \times \{1, \ldots, K\}$
- tomamos uma função perda $\mathcal{L} \colon \mathbb{\Delta}^{K-1} \times \mathbb{\Delta}^{K-1} \to \mathbb{R}_+$
- o problema de aprendizado de máquina consiste em minimizar a perda média do conjunto de treinamento:

$$\min_{\theta \in \Theta} \quad \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(\sigma \circ f_\theta(x_i), e_{y_i}\right)$$

- treinamento supervisionado: somos fornecidos com um *conjunto de treinamento* $\{(x_i, y_i)\}_{i=1}^{m} \subset \mathcal{X} \times \{1, \ldots, K\}$
- tomamos uma função perda $\mathcal{L}\colon \mathbb{\Delta}^{K-1} \times \mathbb{\Delta}^{K-1} \to \mathbb{R}_+$
- o problema de aprendizado de máquina consiste em minimizar a perda média do conjunto de treinamento:

$$\min_{\theta \in \Theta} \quad \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}\left(\sigma \circ f_\theta(x_i), e_{y_i}\right)$$

  - isso costuma ser feito através do método do gradiente

- funções perda mais usadas:
  - entropia cruzada: $h^{\times}(p, q) = \sum_i p_i \log \frac{1}{q_i}$

- funções perda mais usadas:
  - entropia cruzada: $h^{\times}(p,q) = \sum_i p_i \log \frac{1}{q_i}$
  - perda quadrática: $\mathcal{L}(p,q) = \|p-q\|_2^2$

- funções perda mais usadas:
  - entropia cruzada: $h^{\times}(p,q) = \sum_i p_i \log \frac{1}{q_i}$
  - perda quadrática: $\mathcal{L}(p,q) = \|p - q\|_2^2$
- nossa proposta: usar as perdas geométrico-informacionais no simplexo, dadas pelo quadrado das distâncias apresentadas:

- funções perda mais usadas:
  - entropia cruzada: $h^{\times}(p,q) = \sum_i p_i \log \frac{1}{q_i}$
  - perda quadrática: $\mathcal{L}(p,q) = \|p-q\|_2^2$
- nossa proposta: usar as perdas geométrico-informacionais no simplexo, dadas pelo quadrado das distâncias apresentadas:
  - $4L_{\mathrm{SFR}} = d^2_{\mathrm{FR}}(p,q) = 4\,\mathrm{arccos}\left(\sum_{i=1}^K \sqrt{p_i q_i}\right)^2$
  - $L_{\mathrm{SH}} = d^2_{\mathrm{H}}(p,q) = \sum_{i=1}^K \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$

- funções perda mais usadas:
  - entropia cruzada: $h^{\times}(p, q) = \sum_i p_i \log \frac{1}{q_i}$
  - perda quadrática: $\mathcal{L}(p, q) = \|p - q\|_2^2$
- nossa proposta: usar as perdas geométrico-informacionais no simplexo, dadas pelo quadrado das distâncias apresentadas:
  - $4L_{\mathrm{SFR}} = d_{\mathrm{FR}}^2(p, q) = 4 \arccos\left(\sum_{i=1}^K \sqrt{p_i q_i}\right)^2$
  - $L_{\mathrm{SH}} = d_{\mathrm{H}}^2(p, q) = \sum_{i=1}^K \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$
- este é um trabalho em conjunto com H.K. Miyamoto e S.I.R. Costa, submetido para o ISIT 2022 (*International Symposium on Information Theory*)[2]

[2] Henrique K. Miyamoto, Fábio C. C. Meneghetti e Sueli I. R. Costa. "Information-Geometric Loss Functions for Learning". Em: ISIT 2022. 2022.

- observamos que existem relações assintóticas e desigualdades entre as perdas que introduzimos e a perda da entropia cruzada

- observamos que existem relações assintóticas e desigualdades entre as perdas que introduzimos e a perda da entropia cruzada

asymptotic relations between different loss functions.

Proposition 1. *Let L_{CE}, L_{SFR} and L_{SH} be the cross-entropy loss, the squared Fisher-Rao loss, and the squared Hellinger loss, as defined in* (7), (8) *and* (9) *respectively. Then we have:*

1) $L_{\mathrm{SFR}}(y, f(\boldsymbol{x})) = L_{\mathrm{SH}}(y, f(\boldsymbol{x})) + O\big(L^2_{\mathrm{SH}}(y, f(\boldsymbol{x}))\big)$;

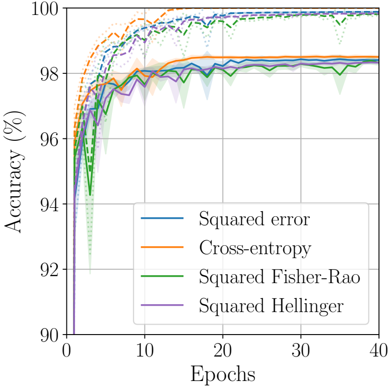
2) $L_{\mathrm{SFR}}(y, f(\boldsymbol{x})) = L_{\mathrm{CE}}(y, f(\boldsymbol{x})) + O\big(L^2_{\mathrm{CE}}(y, f(\boldsymbol{x}))\big)$.

Moreover, we have the inequality chain:

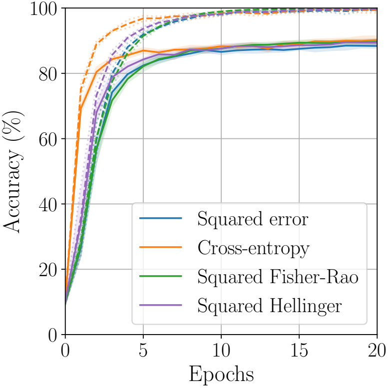
3) $L_{\mathrm{SH}}(y, f(\boldsymbol{x})) \leq L_{\mathrm{SFR}}(y, f(\boldsymbol{x})) \leq L_{\mathrm{CE}}(y, f(\boldsymbol{x}))$.

Proof: 1) is a direct consequence of (5). For 2), isolate

Resultados



(a) MNIST

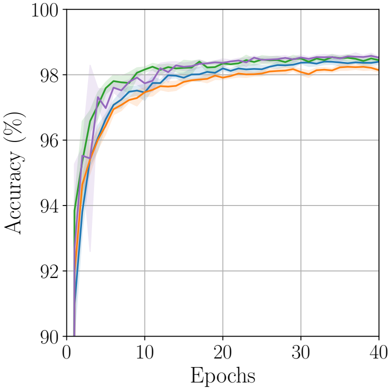


(b) Banco de dados sintético

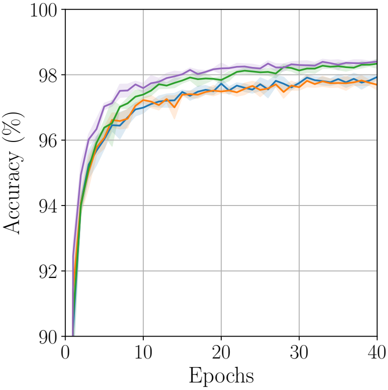
Figura: Acurácia dos aprendizados com diferentes funções perda.

Resultados com ruído

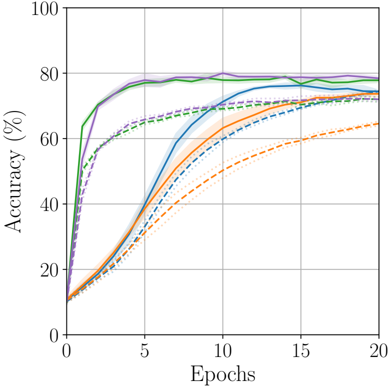
(alguns rótulos do conjunto de treinamento recebem a classe errada)



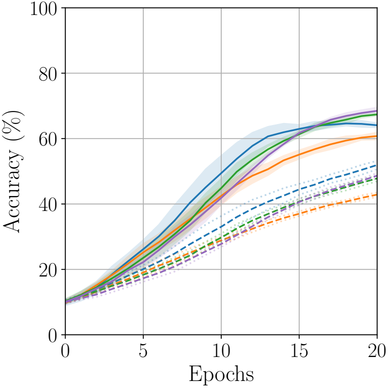
(a) MNIST, $\eta = 0.3$



(b) MNIST, $\eta = 0.5$



(a) Sintético, $\eta = 0.3$



(b) Sintético, $\eta = 0.5$

- possivelmente uma das razões das perdas de Fisher-Rao e Hellinger terem boa performance no caso ruidoso seja por elas serem limitadas

- possivelmente uma das razões das perdas de Fisher-Rao e Hellinger terem boa performance no caso ruidoso seja por elas serem limitadas
- este é um trabalho em andamento

Distribuições de probabilidade enroladas no toro

- um reticulado posto-completo é um subconjunto de $\mathbb{R}^n$ gerado por combinações lineares inteiras de uma base $\{b_1, \ldots, b_n\}$

Distribuições de probabilidade enroladas no toro

- um reticulado posto-completo é um subconjunto de $\mathbb{R}^n$ gerado por combinações lineares inteiras de uma base $\{b_1, \ldots, b_n\}$
- definimos o toro enrolado por Λ como

$$\mathbb{T}_\Lambda := \mathbb{R}^n/\Lambda = \left\{ [\![x]\!]_\Lambda = x + \Lambda \;\middle|\; x \in \mathbb{R}^n \right\}$$

Distribuições de probabilidade enroladas no toro

- um reticulado posto-completo é um subconjunto de $\mathbb{R}^n$ gerado por combinações lineares inteiras de uma base $\{b_1, \ldots, b_n\}$
- definimos o toro enrolado por Λ como

$$\mathbb{T}_\Lambda := \mathbb{R}^n/\Lambda = \left\{ [\![x]\!]_\Lambda = x + \Lambda \;\middle|\; x \in \mathbb{R}^n \right\}$$

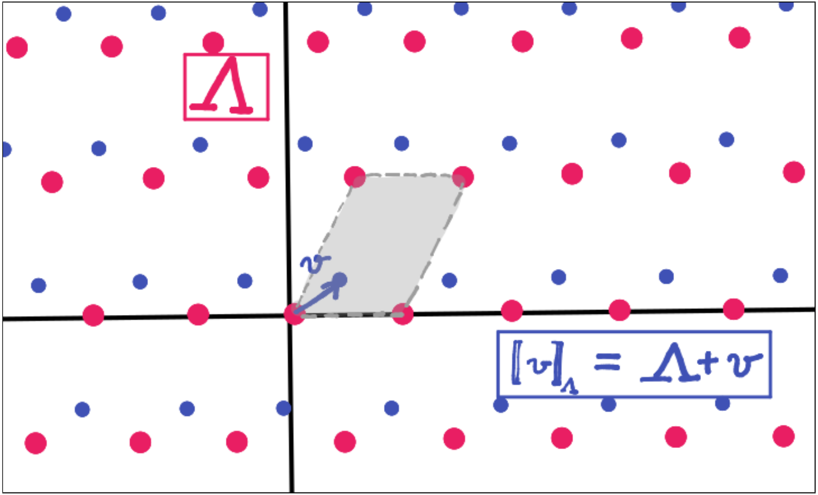
- temos uma projeção canônica $\pi_\Lambda \colon \mathbb{R}^n \to \mathbb{T}_\Lambda, \;\; \pi_\Lambda(x) = [\![x]\!]_\Lambda$

Distribuições de probabilidade enroladas no toro

- um reticulado posto-completo é um subconjunto de $\mathbb{R}^n$ gerado por combinações lineares inteiras de uma base $\{b_1, \ldots, b_n\}$
- definimos o toro enrolado por Λ como

$$\mathbb{T}_\Lambda := \mathbb{R}^n/\Lambda = \left\{ [\![x]\!]_\Lambda = x + \Lambda \;\middle|\; x \in \mathbb{R}^n \right\}$$

- temos uma projeção canônica $\pi_\Lambda \colon \mathbb{R}^n \to \mathbb{T}_\Lambda, \;\; \pi_\Lambda(x) = [\![x]\!]_\Lambda$
- dada uma distribuição de probabilidade $\mathbb{P}$ em $\mathbb{R}^n$, podemos enrolá-la no toro com um empurro via π_Λ, isto é, $\mathbb{P}_\Lambda := (\pi_\Lambda)_*\mathbb{P}$


Λ
v
$[v]_\Lambda = \Lambda + v$

- se uma distribuição $\mathbb{P}$ em $\mathbb{R}^n$ tem densidade $p(x)$, $x \in \mathbb{R}^n$, então a distribuição enrolada tem densidade

$$p_\Lambda(x) = \sum_{\lambda \in \Lambda} p(x + \lambda)$$

sobre $\mathbb{T}_\Lambda$ ou uma região fundamental (ex: região de Voronoi)

- se uma distribuição $\mathbb{P}$ em $\mathbb{R}^n$ tem densidade $p(x)$, $x \in \mathbb{R}^n$, então a distribuição enrolada tem densidade

$$p_\Lambda(x) = \sum_{\lambda \in \Lambda} p(x + \lambda)$$

sobre $\mathbb{T}_\Lambda$ ou uma região fundamental (ex: região de Voronoi)

- assim, a partir de um modelo estatístico $\{p_\theta\}_{\theta \in \Theta}$ em $\mathbb{R}^n$ obtemos um modelo estatístico $\left\{p_{\theta;\Lambda}\right\}_{\theta \in \Theta}$ em $\mathbb{T}_\Lambda$.

- se uma distribuição $\mathbb{P}$ em $\mathbb{R}^n$ tem densidade $p(x)$, $x \in \mathbb{R}^n$, então a distribuição enrolada tem densidade

$$p_\Lambda(x) = \sum_{\lambda \in \Lambda} p(x + \lambda)$$

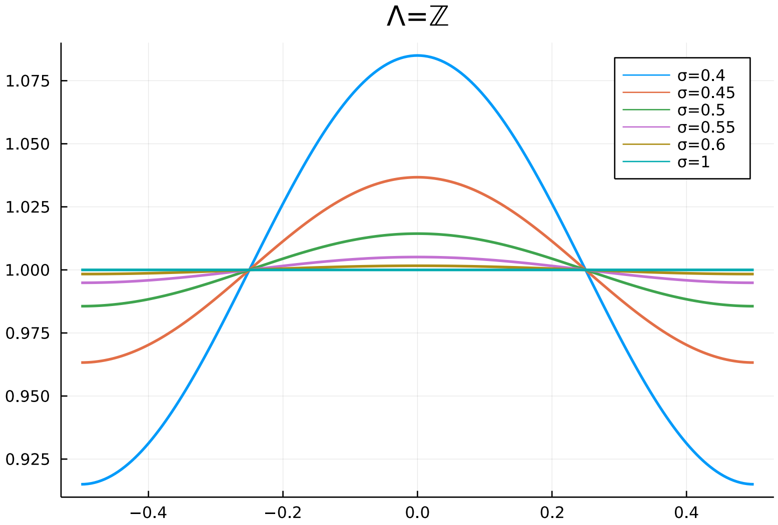
sobre $\mathbb{T}_\Lambda$ ou uma região fundamental (ex: região de Voronoi)

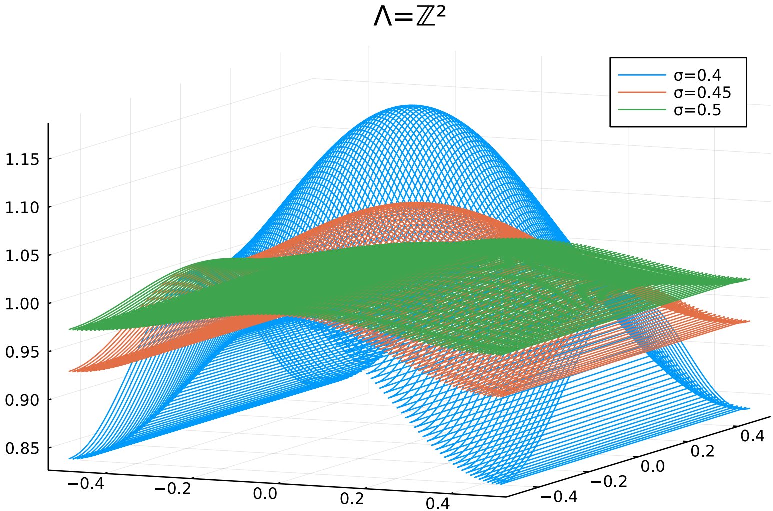
- assim, a partir de um modelo estatístico $\{p_\theta\}_{\theta \in \Theta}$ em $\mathbb{R}^n$ obtemos um modelo estatístico $\left\{p_{\theta;\Lambda}\right\}_{\theta \in \Theta}$ em $\mathbb{T}_\Lambda$.
- ex: gaussianas multivariadas

$$p_{\mu,K;\Lambda}(x) = \frac{1}{\sqrt{(2\pi)^d \left|\det K\right|}} \sum_{\lambda \in \Lambda} e^{-\frac{1}{2}(x+\lambda-\mu)^\top K^{-1}(x+\lambda-\mu)}$$

- uma propriedade central dessas distribuições é que para variância crescente elas se aproximam da distribuição uniforme $\mathcal{U}(x) = \frac{1}{|\det \Lambda|}$

- uma propriedade central dessas distribuições é que para variância crescente elas se aproximam da distribuição uniforme $\mathcal{U}(x) = \frac{1}{|\det \Lambda|}$




Λ=ℤ²
σ=0.4
σ=0.45
σ=0.5
1.15
1.10
1.05
1.00
0.95
0.90
0.85
−0.4
−0.2
0.0
0.2
0.4
0.4
0.2
0.0
−0.2
−0.4

- gostaríamos de estudar a geometria dessas distribuições

- gostaríamos de estudar a geometria dessas distribuições
    - em termos da geometria de Fisher-Rao

- gostaríamos de estudar a geometria dessas distribuições
  - em termos da geometria de Fisher-Rao
  - e em termos de medidas de divergência (Kulback-Leibler, f-divergências, normas L^p, etc.)

- gostaríamos de estudar a geometria dessas distribuições
  - em termos da geometria de Fisher-Rao
  - e em termos de medidas de divergência (Kulback-Leibler, f-divergências, normas L^p, etc.)
- já existe alguma pesquisa sobre distribuições desse tipo em termos da geometria de Wasserstein [3]

[3] Anton Mallasto e Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions". Em: 38th MaxEnt. 2018

- gostaríamos de estudar a geometria dessas distribuições
  - em termos da geometria de Fisher-Rao
  - e em termos de medidas de divergência (Kulback-Leibler, f-divergências, normas L^p, etc.)
- já existe alguma pesquisa sobre distribuições desse tipo em termos da geometria de Wasserstein [3]
- nossa motivação: o fator de achatamento (*flatness factor*) é a distância L^∞ entre uma distribuição $\mathbb{P}_\Lambda$ e a uniforme

[3] Anton Mallasto e Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions". Em: 38th MaxEnt. 2018

- gostaríamos de estudar a geometria dessas distribuições
  - em termos da geometria de Fisher-Rao
  - e em termos de medidas de divergência (Kulback-Leibler, f-divergências, normas L^p, etc.)
- já existe alguma pesquisa sobre distribuições desse tipo em termos da geometria de Wasserstein [3]
- nossa motivação: o fator de achatamento (*flatness factor*) é a distância L^∞ entre uma distribuição $\mathbb{P}_\Lambda$ e a uniforme
  - ele é um parâmetro importante para construir códigos que atingem capacidade no canal AWGN e para garantir segredo no canal Wiretap

[3] Anton Mallasto e Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions". Em: 38th MaxEnt. 2018

- gostaríamos de estudar a geometria dessas distribuições
  - em termos da geometria de Fisher-Rao
  - e em termos de medidas de divergência (Kulback-Leibler, f-divergências, normas L^p, etc.)
- já existe alguma pesquisa sobre distribuições desse tipo em termos da geometria de Wasserstein [3]
- nossa motivação: o fator de achatamento (*flatness factor*) é a distância L^∞ entre uma distribuição $\mathbb{P}_\Lambda$ e a uniforme
  - ele é um parâmetro importante para construir códigos que atingem capacidade no canal AWGN e para garantir segredo no canal Wiretap
  - queremos entender se o fator de achatamento medido com outras divergências também tem comportamento interessante

[3] Anton Mallasto e Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions". Em: 38th MaxEnt. 2018

- gostaríamos de estudar a geometria dessas distribuições
  - em termos da geometria de Fisher-Rao
  - e em termos de medidas de divergência (Kulback-Leibler, f-divergências, normas L^p, etc.)
- já existe alguma pesquisa sobre distribuições desse tipo em termos da geometria de Wasserstein [3]
- nossa motivação: o fator de achatamento (*flatness factor*) é a distância L^∞ entre uma distribuição $\mathbb{P}_\Lambda$ e a uniforme
  - ele é um parâmetro importante para construir códigos que atingem capacidade no canal AWGN e para garantir segredo no canal Wiretap
  - queremos entender se o fator de achatamento medido com outras divergências também tem comportamento interessante
  - este tema está diretamente conectado ao mestrado do aluno [4]

[3] Anton Mallasto e Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions". Em: 38th MaxEnt. 2018

[4] Fábio C. C. Meneghetti. "Reticulados: um estudo de alguns parâmetros relevantes para aplicações em criptografia". 2020

Parte IV

Futuro

Futuro

- queremos continuar estudando alguns aspectos teóricos

Futuro

- queremos continuar estudando alguns aspectos teóricos
  - a relação entre a estrutura dualmente plana de Amari (M, g, ∇, ∇^*) e as geometrias simplética e Kähler

Futuro

- queremos continuar estudando alguns aspectos teóricos
  - a relação entre a estrutura dualmente plana de Amari (M, g, ∇, ∇^*) e as geometrias simplética e Kähler
  - extensões da geometria da informação para espaços de dimensão infinita (ex: estrutura de Pistone-Sempi)

Futuro

- queremos continuar estudando alguns aspectos teóricos
  - a relação entre a estrutura dualmente plana de Amari (M, g, ∇, ∇^*) e as geometrias simplética e Kähler
  - extensões da geometria da informação para espaços de dimensão infinita (ex: estrutura de Pistone-Sempi)
- entender se há relação entre nossa proposta de funções perda, e as α-Divergências, e também com o método do gradiente natural

Futuro

- queremos continuar estudando alguns aspectos teóricos
  - a relação entre a estrutura dualmente plana de Amari (M, g, ∇, ∇^*) e as geometrias simplética e Kähler
  - extensões da geometria da informação para espaços de dimensão infinita (ex: estrutura de Pistone-Sempi)
- entender se há relação entre nossa proposta de funções perda, e as α-Divergências, e também com o método do gradiente natural
- formalizar a teoria das distribuições gaussianas enroladas no toro

Futuro

- queremos continuar estudando alguns aspectos teóricos
  - a relação entre a estrutura dualmente plana de Amari (M, g, ∇, ∇^*) e as geometrias simplética e Kähler
  - extensões da geometria da informação para espaços de dimensão infinita (ex: estrutura de Pistone-Sempi)
- entender se há relação entre nossa proposta de funções perda, e as α-Divergências, e também com o método do gradiente natural
- formalizar a teoria das distribuições gaussianas enroladas no toro
  - procurar relações entre reticulados diferentes (ex: se $\Lambda = B \cdot \mathbb{Z}^n$, então $p_{\mu,K;\Lambda}(x) = \frac{1}{\det \Lambda} p_{\tilde{\mu},\tilde{K};\mathbb{Z}^n}(B^{-1}x)$, $\tilde{\mu} = B^{-1}\mu$, $\tilde{K} = B^{-1}KB^{-t}$)

Futuro

- queremos continuar estudando alguns aspectos teóricos
  - a relação entre a estrutura dualmente plana de Amari (M, g, ∇, ∇^*) e as geometrias simplética e Kähler
  - extensões da geometria da informação para espaços de dimensão infinita (ex: estrutura de Pistone-Sempi)
- entender se há relação entre nossa proposta de funções perda, e as α-Divergências, e também com o método do gradiente natural
- formalizar a teoria das distribuições gaussianas enroladas no toro
  - procurar relações entre reticulados diferentes (ex: se $\Lambda = B \cdot \mathbb{Z}^n$, então $p_{\mu,K;\Lambda}(x) = \frac{1}{\det \Lambda} p_{\tilde{\mu},\tilde{K};\mathbb{Z}^n}(B^{-1}x)$, $\tilde{\mu} = B^{-1}\mu$, $\tilde{K} = B^{-1}KB^{-t}$)
  - reticulados duais parecem ter relação com a transformada de Fourier da distribuição

Livros

[1] Shun'ichi Amari e Hiroshi Nagaoka. *Methods of information geometry*. Trad. por Daishi Harada. Translations of mathematical monographs. American Mathematical Society, 2007.

[3] Nihat Ay, Jürgen Jost, Hông Vân Lê e Lorenz Schwachhöfer. *Information Geometry*. Springer International Publishing, 2017.

[4] Ovidiu Calin e Constantin Udriște. *Geometric Modeling in Probability and Statistics*. Springer International Publishing, 2014.

Artigos

[2] Colin Atkinson e Ann F. S. Mitchell. "Rao's Distance Measure". Em: *Sankhyā: The Indian Journal of Statistics, Series A* (1981).

[6] Ahmet Demirkaya, Jiasi Chen e Samet Oymak. "Exploring the Role of Loss Functions in Multiclass Classification". Em: 54th CISS. 2020.

[7] Anton Mallasto e Aasa Feragen. "Optimal Transport Distance between Wrapped Gaussian Distributions". Em: 38th MaxEnt. 2018.

[8] Fábio C. C. Meneghetti. "Reticulados: um estudo de alguns parâmetros relevantes para aplicações em criptografia". 2020.

[10] Frank Nielsen. "An Elementary Introduction to Information Geometry". Em: *Entropy* (2020).

[12] Julianna Pinele, João E. Strapasson e Sueli I. R. Costa. "The Fisher-Rao Distance between Multivariate Normal Distributions: Special Cases, Bounds and Applications". Em: *Entropy* (2020).

[5] Alberto Cena e Giovanni Pistone. "Exponential statistical manifold". Em: *Annals of the Institute of Statistical Mathematics* (2006).

[11] Tomonori Noda. "Sympletic Structures on Statistical Manifolds". Em: *J. Aust. Math. Soc.* (2011).

[13] Rui F. Vigelis, Luiza H. F. De Andrade e Charles C. Cavalcante. "Properties of a Generalized Divergence Related to Tsallis Generalized Divergence". Em: *IEEE Transactions on Information Theory* (2020).