Chapter 4 : Information Geometry

I Fisher information

Why "Inform" geometry"?

↳ about distances in the space of probability measures variations

$p \to p+\delta p$

p deformation χ p'

$d(p, p')$?

↳ $KL(p \| p')$

↳ motivation:

ML problem: $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots\}$

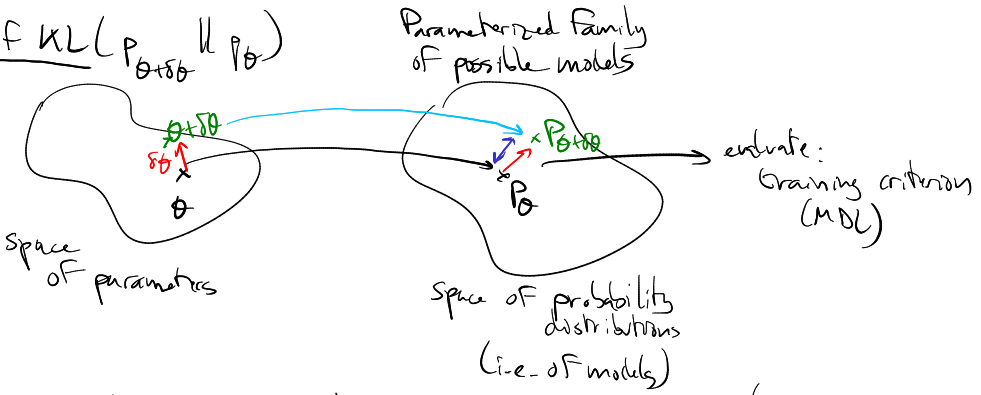
↳ to fit a f^θ $y = f_\theta(x)$ → non-deterministic: $p_\theta(y|x)$

finding the best parameters θ → optimize θ

i.e. the best p_θ

↳ ∇ descent → consider $p_{\theta+\delta\theta}$

Second order of $KL(p_{\theta+\delta\theta} \| p_\theta)$



$KL(p_{\theta+\delta\theta} \| p_\theta)$: infinitesimal distance between p_θ & $p_{\theta+\delta\theta}$ ($\delta\theta$: small parameter variation)

$$= \int_{x \in \mathcal{X}} p_{\theta+\delta\theta}(x) \log \frac{p_{\theta+\delta\theta}(x)}{p_\theta(x)} dx$$

Tricks:

- $\log(1+z) = z - \frac{1}{2}z^2 + O(z^3)$
- $\frac{d}{dz} \log f(z) = \frac{1}{f(z)} \frac{df}{dz}$
- $\int_{x \in \mathcal{X}} p_\theta(x) = 1 \quad \forall \theta$ $\Rightarrow \frac{d^k}{d\theta^k}(1) = 0$ $\Rightarrow \int_{x \in \mathcal{X}} \frac{d^k}{d\theta^k} p_\theta(x)\, dx = 0$

$$\underbrace{p_{\theta+\delta\theta}(x)}_{\in \mathbb{R}} = \underbrace{p_\theta(x)}_{\in \mathbb{R}} + \frac{dp_\theta(x)}{d\theta} \delta\theta + \frac{1}{2} \delta\theta \frac{d^2 p_\theta(x)}{d\theta^2} \delta\theta + O(\delta\theta^3)$$

(matrix of size $|\theta| \times |\theta|$; vector $|\theta|$)

$$\frac{p_{\theta+\delta\theta}(x)}{p_\theta(x)} = 1 + \frac{1}{p_\theta(x)} \frac{dp_\theta(x)}{d\theta} \delta\theta + \frac{1}{2} \delta\theta \frac{1}{p_\theta(x)} \frac{d^2 p_\theta(x)}{d\theta^2} \delta\theta + O(\delta\theta^3)$$

$= z$

$$\log(\,u\,) = \frac{1}{p_\theta(x)} \frac{dp_\theta(x)}{d\theta} \delta\theta + \frac{1}{2} \delta\theta \frac{1}{p_\theta(x)} \frac{d^2 p_\theta(x)}{d\theta^2} \delta\theta - \frac{1}{2} \delta\theta \frac{1}{p_\theta(x)^2} \frac{dp_\theta(x)}{d\theta} \frac{dp_\theta(x)^T}{d\theta} \delta\theta + O(\delta\theta^3)$$

$$KL = \mathop{\mathbb{E}}_{x \sim p_{\theta+\delta\theta}}[u] = \mathop{\mathbb{E}}_{x \sim p_\theta}[u] + \int_{x \in \mathcal{X}} \frac{1}{p_\theta(x)} \delta\theta \frac{dp_\theta(x)}{d\theta} \frac{dp_\theta(x)^T}{d\theta} \delta\theta \, dx + O(\delta\theta^3)$$

$$= \left(\int_{x \in \chi} \frac{p_\theta}{p_\theta} \frac{dp_\theta(x)}{d\theta} dx \right) \delta\theta + 0 - \frac{1}{2} \delta\theta \left(\int_{x \in \chi} \frac{1}{p_\theta(x)} \frac{dp_\theta(x)}{d\theta} \frac{dp_\theta(x)^T}{d\theta} dx \right) \delta\theta + O(\delta\theta^3)$$

$+ 1(\dots)$ same stuff

$$KL(p_{\theta+\delta\theta} \| p_\theta) = \frac{1}{2} \delta\theta \left[\mathop{\mathbb{E}}_{x \sim p_\theta} \left[\frac{d \log p_\theta(x)}{d\theta} \frac{d \log p_\theta(x)}{d\theta}^T \right] \right] \delta\theta + O(\delta\theta^3)$$

matrix of size $|\theta| \times |\theta|$

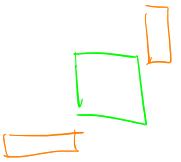
vector of size $|\theta|$

Second order quantity: $O(\delta\theta^2)$

Optimize $KL(q \| p)$ w.r.t q:

$q = p$ is the global optimum

$\Rightarrow \left. \frac{\partial KL(q\|p)}{\partial q} \right|_{q=p} = 0$

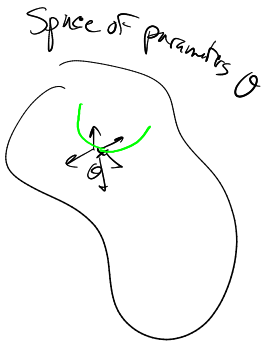


Fisher information matrix (metric) $I(\theta)$

Other formula for Fisher information $I(\theta)$

$$I(\theta) = \mathop{\mathbb{E}}_{x \sim p_\theta} \left[\frac{d^2}{d\theta^2} (-\log p_\theta(x)) \right]$$

average of — 2nd derivative of the encoding cost w.r.t θ (curvature) — encoding cost of x for p_θ



Equivalence between the formulas

$\log p_\theta(x)$

$$\frac{d}{d\theta} \log p_\theta(x) = \frac{1}{p_\theta(x)} \frac{dp_\theta(x)}{d\theta}$$

$$\frac{d^2}{d\theta^2}(\dots) = \frac{d}{d\theta}(\dots) = \frac{-1}{p_\theta(x)^2} \frac{dp_\theta(x)}{d\theta} \times \frac{dp_\theta(x)^T}{d\theta} + \frac{1}{p_\theta(x)} \frac{d^2 p_\theta(x)}{d\theta^2}$$

$$I(\theta) = \mathop{\mathbb{E}}_{x \sim p_\theta}[-u] = \mathop{\mathbb{E}}_{x \sim p_\theta} \left[\frac{d \log p_\theta(x)}{d\theta} \frac{d \log p_\theta(x)}{d\theta}^T \right] + \int_{x \in \chi} \frac{d^2 p_\theta(x)}{d\theta^2} dx \quad (= 0)$$

same vector

symmetric rank 1 (before integration) matrix $|\theta| \times |\theta|$

Vector of size $|\theta|$

In practice: $I(\theta) \simeq \frac{1}{|D|} \sum_{i \in D} \vec{v}_i \otimes \vec{v}_i$ with $\vec{v}_i = \frac{d \log p_\theta(x_i)}{d\theta}$

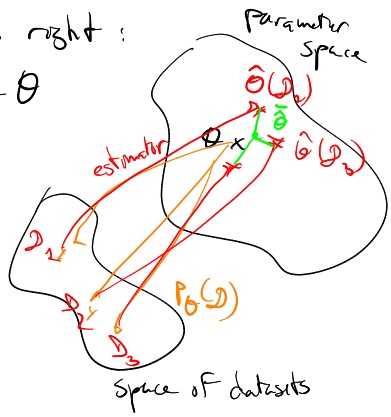
number of points in dataset

Cramer-Rao bound (theorem by Amari) → Cover&Thomas ch 11.10

- given observations $\mathcal{D} \sim (x_i)$, guess the parameter θ s.t. $p_\theta(x)$ is the highest (best fit)
  - ↳ estimator
- an estimator ($\hat{\theta}$) is said to be unbiased if on average it is right:

$$\mathbb{E}_{\mathcal{D}\sim p_\theta}[\hat{\theta}(\mathcal{D})] = \theta$$

- question: average of the estimator
  but what about its variance?
  ↳ is $|\theta - \hat{\theta}|$ typically large or small?



Theorem Cramer-Rao inequality

In dimension 1:

$$\text{variance}(\text{any unbiased estimator}) \geq \frac{1}{J(\theta)} \quad \in \mathbb{R}$$

In dim > 1:

$$\text{covariance}(\quad\text{“}\quad\text{”}\quad) \geq I(\theta)^{-1} \quad \text{(in terms of matrix)}$$

$J = n I$ ↑ number of samples

Proof: online material

II Parameter precision — Yann Ollivier → Jérémy Bensadon

model μ_θ → encode θ up to which precision?

$\theta = \theta^* + \varepsilon$ (precision)

$\theta^* = 0,32085\ldots$

$\theta = 0,32$

(K) cost to describe θ (and the model etc): $C(\varepsilon) = \underbrace{-\log \varepsilon}_{\text{cost to describe } \theta \text{ up to } \varepsilon} - \log p_\theta(x) = -\log \varepsilon - \log p_{\theta^*}(x) + \frac{1}{2}\varepsilon^2 \underbrace{J(\theta^*)}_{\frac{d^2}{d\theta^2}(-\log p_\theta)} + O(\varepsilon^3)$

Best ε? $\frac{\partial}{\partial \varepsilon}(\ldots) = 0$ $\quad\Rightarrow\quad -\frac{1}{\varepsilon} + \varepsilon J(\theta^*) = 0$

$$\boxed{\varepsilon = \frac{1}{\sqrt{J(\theta^*)}}}$$ optimal precision!

$J(\theta^*) = n\, I(\theta^*)$

$$\varepsilon^* \simeq \frac{1}{\sqrt{n}} \frac{1}{\sqrt{I(\theta)}}$$

↑ classical statistics

III Misc

Justification of BIC:

BIC: $K(\mu) := \frac{1}{2}$ number of parameters $\times \log$(number of observations)

↳ encode model with optimal precision

↳ precision: $\varepsilon = \frac{1}{\sqrt{n}} \frac{1}{\sqrt{I(\theta)}}$

↳ encode the parameters: for each parameter, length of encoding: $-\log \varepsilon = +\frac{1}{2}\log n \quad (+\frac{1}{2}\log I(\theta))$ constant

↳ $\mathcal{D}/\mu$ × number of parameters

(Green eqn pp of η)

→ strong limits with natural gradient

→ Jeffrey's prior