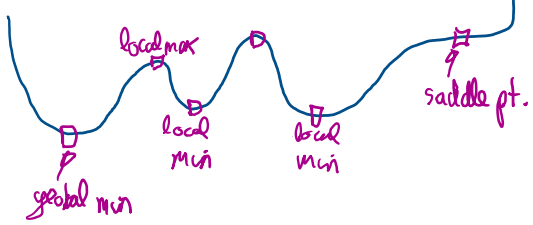
Lecture 5 - MLE

Thursday, September 19, 2024 1:57 PM

today: • MLE cf'ed
• statistical decision theory

optimization comments about MLE

$\min_{\theta \in \Theta} f(\theta)$



(*) $Df(\theta^*) = 0$ (θ^* is stationary point)

(if f is differentiable on Θ is a necessary cond. for θ^* to be a local min when θ^* is in interior of Θ

→ also check that $\text{Hessian}(f)(\theta^*) \succ 0$ for a local min

$H \succ 0 \Leftrightarrow u^T H u > 0 \quad \forall u \neq 0 \in \mathbb{R}^q$

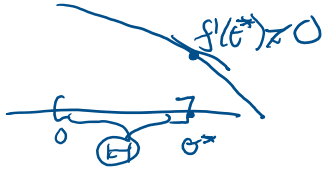
$(f''(\theta^*) > 0)$

(*) only local results in general

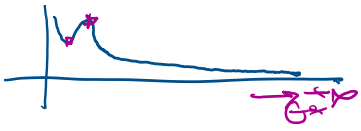
- but if $\text{Hessian}(f(\theta)) \succeq 0 \quad \forall \theta \in \Theta$, fct. is said to be "convex"
  then $Df(\theta^*) = 0 \Rightarrow \theta^*$ is a global min
- otherwise, for smooth fct., looking at zero gradient pts & boundary points give you enough info to find global min (**)

(*) be careful with boundary cases

ie. $\theta^* \in \text{boundary}(\Theta)$

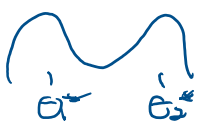


other example:



* some notes about MLE

- does not always exist [$\theta^* \in \text{bd}(\Theta)$ but Θ is open] or when "$\theta^* = \pm\infty$"
- if it is not rec unique [ex. mixture models $]0,1[$]



- is not "admissible" in general [see later]
  ∃ strictly "better" estimator

example II: multinomial distribution

suppose X_i is a discrete R.V. on k choices "multinoulli"

(one could choose $\Omega_{X_i} = \{1, 2, 3, \dots, k\}$)

but instead, convenient to encode the k possibilities using unit basis in $\mathbb{R}^k$

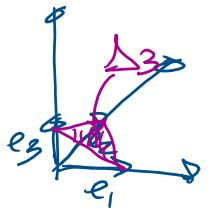
i.e. $\Omega_{X_i} = \{e_1, e_2, \dots, e_k\}$ where $e_j \in \mathbb{R}^k$ "one hot encoding"

$e_j = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ ← j^{th} coordinate

parameter for discrete R.V. $\pi \in \mathbb{R}^k$ $\quad (\Theta = \Delta_k)$

$\pi \in \Delta_k \qquad \Delta_k = \{\pi \in \mathbb{R}^k : \pi_j \geq 0 ; \sum_{j=1}^k \pi_j = 1\}$

probability simplex on k choices



we will write $X_i \sim \text{Mult}(\pi) = \text{Mult}(1, \pi)$ (π: parameter)

(*) consider $X_i \overset{iid}{\sim} \text{Mult}(\pi)$

then $X = \sum_{i=1}^n X_i \sim \text{Mult}(n, \pi)$

↳ "multinomial distribution"

$x_9 \to \in \mathbb{R}^k$

$X \in \mathbb{N}^k \qquad \Omega_X = \{(n_1, n_2, \dots, n_k) : n_j \in \mathbb{N} ; \sum_{j=1}^k n_j = n\}$

proof for X: $\quad x = (n_1, \dots, n_k)$ $\quad (x)_j \triangleq n_j$

$$p(x|\pi) = \binom{n}{(x)_1 (x)_2 \dots (x)_k} \prod_{j=1}^k \pi_j^{(x)_j}$$

multinomial coeff

$$\binom{n}{n_1 n_2 \dots n_k} \triangleq \frac{n!}{n_1! n_2! \dots n_k!}$$

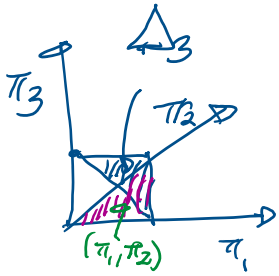
15h15 **multinomial MLE**

log-likelihood $\quad \ell(\pi) = \log p(x|\pi) = \log \binom{n}{n_1, \dots n_k} + \sum_{j=1}^k n_j \log \pi_j$

$x = (n_1, \dots, n_k)$

constant → ignore

MLE $\quad \hat{\pi}_{MLE}(x) = \underset{\substack{\pi \in \mathbb{R}^k \\ \text{s.t. } \pi \in \Delta_k}}{\text{argmax}} \ell(\pi)$ } constraint



two options:

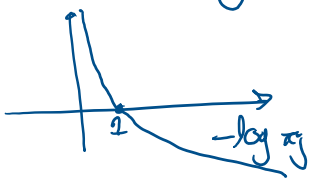
a) reparameterize problem so that Θ is full dimensional

$\pi_K = 1 - \sum_{j=1}^{K-1} \pi_j$

↳ $\pi_1, \dots, \pi_{K-1} \in [0,1]$ with constraint $\sum_{j=1}^{K-1} \pi_j \leq 1$

↳ $\pi_1, \dots, \pi_{K-1} \in [0,1]$ with constraint $\sum_{j=1}^{K-1} \pi_j \leq 1$

have magic that $\log \pi_j$ acts as a barrier fct. away from $\pi_j = 0$



can try unconstrained opt. on $\pi_1, \dots, \pi_{K-1}$

of $l(\pi_1, \dots, \pi_{K-1}, 1 - \sum_{j=1}^{K-1} \pi_j)$

hoping soln is in interior of constraint set

(and it is usually case for log-type problem)

b) use Lagrange multiplier approach to handle equality constraint on Δ_K

[and still ignoring $\pi_j \in [0,1]$]

$\max f(\pi)$

s.t. $g(\pi) = 0$

$[g(\pi) \triangleq 1 - \sum_{j=1}^{K} \pi_j]$.

$J(\pi, \lambda) \triangleq f(\pi) + \lambda g(\pi)$

Lagrange multiplier

method: look at stationary pt. of $J(\pi, \lambda)$

(0-gradient)

necessary cond. for local opt.

(check "bordered Hessian" to get local or max)

(see Wikipedia)

ie. $\nabla_\pi J(\pi, \lambda) = 0$

$\nabla_\lambda J(\pi, \lambda) = 0 \Rightarrow g(\pi) = 0$

$l(\pi) = \sum_j n_j \log \pi_j$

(strictly concave fct. in π_j)

want $\frac{\partial J}{\partial \pi_j} = 0 \rightarrow \frac{n_j}{\pi_j} - \lambda = 0 \Rightarrow \pi_j^* = \frac{n_j}{\lambda}$ } scaling constant

want $g(\pi^*) = 0$; ie. $\sum_j \pi_j^* = 1$

ie. $\sum_j \frac{n_j}{\lambda} = 1$

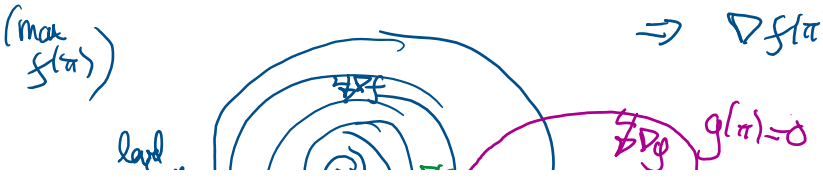
$\Rightarrow \lambda^* = \sum_{j=1}^{n} n_j = n$

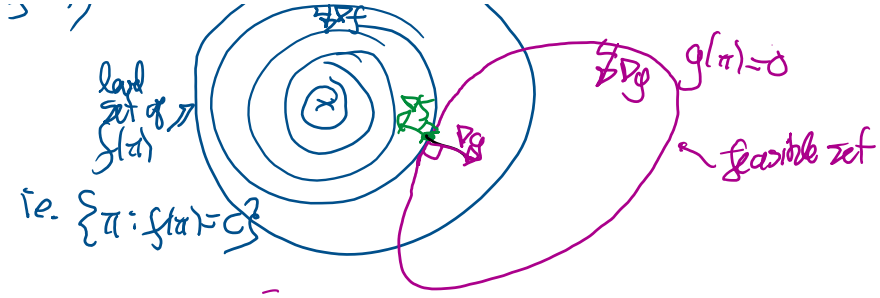
notice: $\pi_j^* \in [0,1]$ $\boxed{\pi_j^* = \frac{n_j}{n}}$ MLE for multinomial

$\nabla_\pi L(\pi, \lambda) = 0 \Rightarrow \nabla_\pi f(\pi) + \lambda \nabla_\pi g(\pi) = 0$

$\Rightarrow \nabla f(\pi) = -\lambda \nabla g(\pi)$

(max $f(\pi)$)





Statistical decision theory

A) Bias-variance decomposition for squared loss

estimator : fct. from data (observation) to parameter

$$\text{MLE}: \hat{\theta}_{MLE}(x) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(x|\theta)$$

$$\text{MAP}: \hat{\theta}_{MAP}(x) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(\theta|x) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \underbrace{p(x|\theta)}_{\text{likelihood}}\ \underbrace{p(\theta)}_{\text{prior}}$$

$$\log p(x|\theta) + \underbrace{\log p(\theta)}_{\text{regularizer}}$$

* how do we evaluate these estimators? estimator $\delta : \Omega_x \to \Theta$

$$\hat{\theta} = \delta(x)$$

most standard fct: frequentist risk of an estimator

$$R(\theta, \delta) \triangleq \mathbb{E}_x[L(\theta, \delta(x))]$$

- average over possible data / training set
- "statistical loss fct."

squared loss, $L(\theta, \theta') \triangleq \|\theta - \theta'\|_2^2$

$$\mathbb{E}_x[\|\theta - \hat{\theta}\|_2^2] = \mathbb{E}[\|\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta}\|_2^2] \qquad \hat{\theta} = \delta(x)$$

$\hat{\theta}$: random; $\mathbb{E}\hat{\theta}$: constant

$$\|a\|_2^2 = \langle a, a \rangle$$

$$= \mathbb{E}[\|\theta - \mathbb{E}\hat{\theta}\|^2] + \mathbb{E}[\|\hat{\theta} - \mathbb{E}\hat{\theta}\|^2] + 2\mathbb{E}[\langle \theta - \mathbb{E}\hat{\theta}, \mathbb{E}\hat{\theta} - \hat{\theta} \rangle]$$

$$\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b \rangle$$

$\theta - \mathbb{E}\hat{\theta}$: constant

$$= 2\langle \theta - \mathbb{E}\hat{\theta}, \mathbb{E}(\mathbb{E}\hat{\theta} - \hat{\theta}) \rangle$$

$\mathbb{E}\hat{\theta} - \mathbb{E}\hat{\theta} = 0$

$$R(\theta, \delta) = \mathbb{E}_x[\|\theta - \hat{\theta}\|_2^2] = \underbrace{\|\theta - \mathbb{E}\hat{\theta}\|_2^2}_{\text{bias}^2} + \underbrace{\mathbb{E}[\|\hat{\theta} - \mathbb{E}\hat{\theta}\|_2^2]}_{\text{variance}}$$

$$\text{bias} \triangleq \|\theta - \mathbb{E}(\hat{\theta})\|_2$$

(frequentist) risk for squared loss = bias^2 + variance. "bias-variance"

$\text{bias} \triangleq \|\theta - F(\theta')\|_2$

(freq.) risk for squared loss = $\text{bias}^2 + \text{variance}$.

"bias-variance tradeoff"

(*) Consistency: informally "do right thing as $n \to \infty$" where n is the training set size

$\hat{\theta}_n \xrightarrow{P} \theta$

$X \sim (X_i)_{i=1}^n$

$\hat{\theta}_n$ (data of size)

assignment: if $\text{bias}(\hat{\theta}_n) \xrightarrow{n\to\infty} 0$ and $\text{variance}(\hat{\theta}_n) \xrightarrow{n\to\infty} 0$ $\Rightarrow R(\theta, \hat{\theta}_n) \xrightarrow{n\to\infty} 0 \Rightarrow (\hat{\theta}_n \xrightarrow{P} \theta)$

i.e. $\hat{\theta}_n$ is consistent