Lecture 5 - statistical decision theory

Tuesday, September 19, 2023 2:30 PM

today : · MLE ct'ed
· statistical decision theory

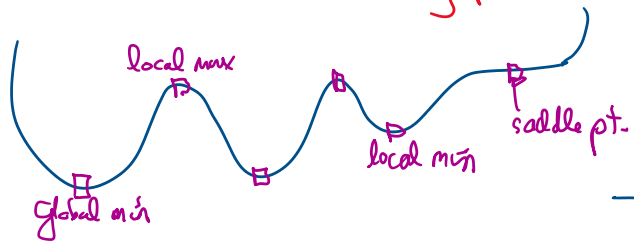
optimization comments about MLE

$\min_{\theta \in \Theta} f(\theta)$ ⊛ $\nabla f(\theta^*) = 0$ (f is differentiable on Θ)

"stationary pts."

is a necessary cond.:

for θ^* to be a local min when θ^* is in interior of Θ



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

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

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

⊛ only local results in general

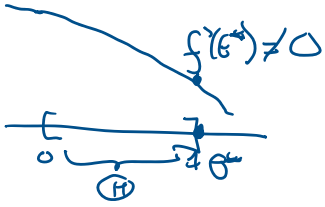
- but if Hessian$(f(\theta)) \succeq 0 \quad \forall \theta \in \Theta$, fct. is said to be "convex"

  then $\nabla f(\theta^*) = 0 \Rightarrow \theta^*$ is a global min

- otherwise, for smooth fct., looking at all zero gradient pts and boundary pts gives you enough information to find global min ⊛⊛

⊛ be careful with boundary cases

ie. $\theta^* \in$ boundary(Θ) e.g.



other example:

↑ example where MLE does not exist

✱ some notes about MLE

- does not always exist [$\theta^* \in$ bd(Θ) but Θ is open] or when "$\theta^* = +\infty$"

  $\Theta =]0,1[$

- if is not rec. unique [eg. mixture models]

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

Example II: multinomial distribution

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

(we could choose $\Omega_{X_i} = \{1, 2, \dots, k\}$

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

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

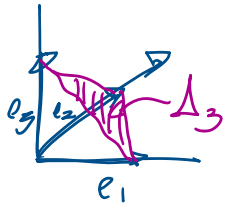
$$e_j = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \leftarrow j^{th} \text{ coordinate}$$

$\pi \in \mathbb{R}^k$

parameter for discrete R.V. $\pi \in \Delta_k$ $\left(\Theta = \Delta_k\right)$

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

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"

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

pmf for X:

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

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

$(x)_j \triangleq n_j$

[vs. x_i of X_i]

← multinomial coeff

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

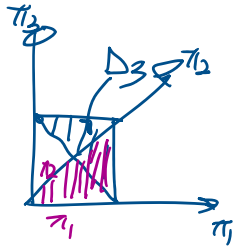
15h29

multinomial MLE

log-likelihood $\ell(\pi) = \log p(x|\pi) = \underbrace{\log \binom{n}{n_1, \dots n_k}}_{\text{constant during MLE} \rightarrow \text{ignore}} + \sum_{j=1}^{k} n_j \log \pi_j$

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

MLE: $\hat{\pi}_{MLE}(x) = \underset{\pi \in \mathbb{R}^k}{\arg\max}\ \ell(\pi)$
s.t. $\pi \in \Delta_k$ } constraint



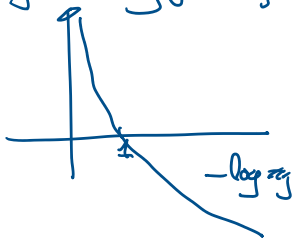
two options:

a) reparametrize problem so that Θ is full dimensional

$\pi_k \triangleq 1 - \sum_{j=1}^{k-1} \pi_j$

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

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



can try unconstrained optimization on $\pi_1, \dots, \pi_{k-1}$ of $\ell(\pi_1, \dots, \pi_{k-1})$

having soln is in the interior of constraint set (and it usually will for log-type problems)

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. $g(\pi) = 0$
$[\, 1 - \sum_{j=1}^{k-1} \pi_j = 0 \,]$
$\triangleq g(\pi)$

$J(\pi, \lambda) \triangleq f(\pi) + \lambda g(\pi)$
λ: Lagrange multiplier

method: look at stationary pts. of $J(\pi, \lambda)$ (0-gradient)

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

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

necessary cond. for local opt.

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

(see Wikipedia)

$\ell(\pi) = \sum_j n_j \log \pi_j$ (strictly concave fct. in π_j)

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

want $g(\pi^*) = 0$ i.e. $\sum_j \pi_j^* = 1$

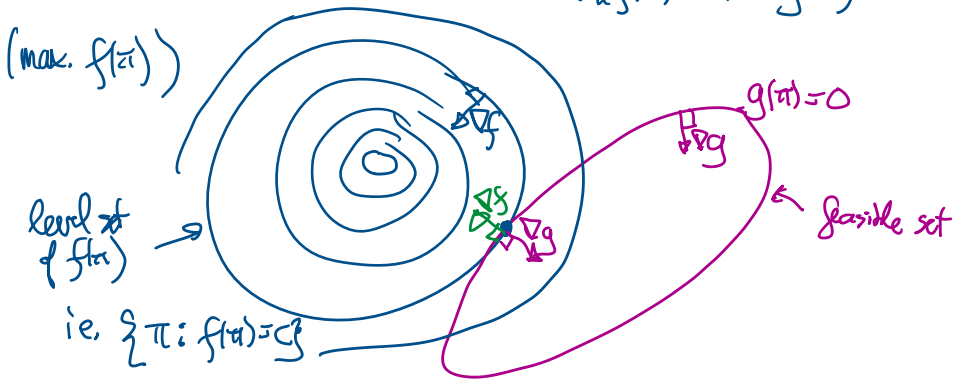
$\sum_j \frac{n_j}{\lambda} = 1$

$\Rightarrow \lambda^* = \sum_j n_j = n$

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

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

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



~~Statistical~~ decision theory

A) Bias-variance decomposition for squared loss

estimator : fct. from data (observation) to parameter

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

likelihood prior

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

$\log p(x|\theta) + \log p(\theta)$

* how do we evaluate those estimators? estimator $\delta : \Omega_X \to \Theta$

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

most standard tool : frequentist risk of an estimator $\boxed{R(\theta, \delta) \triangleq \mathbb{E}_X [L(\theta, \delta(x))]}$

~~frequentist risk~~ $R(\theta, \delta) = \mathbb{E}_X[L(\theta, \delta(X))]$

↳ average over possible data

(statistical) loss fct.

squared loss: $L(\theta, \theta') \triangleq \|\theta - \theta'\|_2^2$ $\quad \hat{\theta} = \delta(X)$

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

$$\mathbb{E}_X[\|\theta - \hat{\theta}\|_2^2] = \mathbb{E}[\|\theta - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \hat{\theta}\|_2^2]$$

($\hat{\theta}$: random)

$$= \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]$

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

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

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

$$R(\theta, \delta) = \mathbb{E}_X[\|\theta - \hat{\theta}\|_2^2] = \|\theta - \mathbb{E}[\hat{\theta}]\|_2^2 + \mathbb{E}[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_2^2]$$

$\|\theta - \mathbb{E}[\hat{\theta}]\|_2^2$: bias^2 $\quad$ $\mathbb{E}[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_2^2]$: variance

bias $\triangleq \|\theta - \mathbb{E}[\hat{\theta}]\|_2$

(freq.) risk for squared loss = bias^2 + variance

bias–variance "tradeoff"

* consistency: informally "do right thing as $n \to \infty$" where n is training set size

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

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

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

$(\hat{\theta}_n \xrightarrow{p} \theta)$