Lecture 5 - scribbles

Tuesday, September 15, 2020 14:20

today: • MLE
• statistical decision theory

MLE example I: binomial

n coin flips $\quad \Omega_X = \{0, \dots, n\}$

$X \sim \text{Bin}(n, \theta) \quad p(x;\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

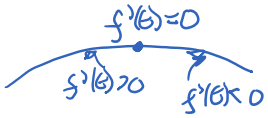
trick: to maximize $\log L(\theta)$ instead $L(\theta)$
$\triangleq \ell(\theta)$ log likelihood

justification: $\log(\cdot)$ is strictly increasing

i.e. $a < b \Leftrightarrow \log a < \log b \quad (\forall a, b > 0)$

$\Rightarrow \underset{\theta \in \Theta}{\text{argmax}} \log p(x;\theta) = \underset{\theta \in \Theta}{\text{argmax}}\, p(x;\theta)$

$\log p(x;\theta) = \log \binom{n}{x} + x \log\theta + (n-x)\log(1-\theta) = \ell(\theta)$



constant w.r.t. θ

look for θ s.t. $\frac{\partial \ell}{\partial \theta} = 0$

usual $\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$

$x(1-\theta) - (n-x)\theta = 0$

$\boxed{\theta^* = \frac{x}{n}}$ used often as solution in optimization

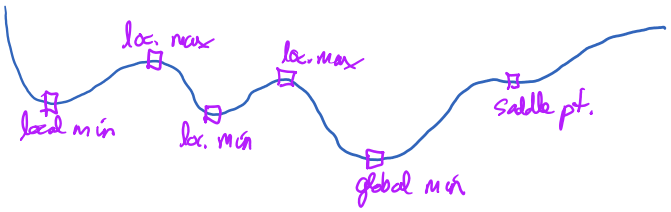
hence $\boxed{\hat{\theta}_{ML}(x) = \frac{x}{n}}$ i.e. relative frequency

some optimization comments

$\min_{\theta \in \Theta} f(\theta) \quad \theta^* \in \Theta \quad \nabla f(\theta^*) = 0$ ("stationary pts.")

(if f is diff. on Θ), is a necessary condition for θ^* being a local min when θ^* is in the interior of Θ

→ also need to check Hessian$(f)(\theta^*) > 0$ for a local min ($f''(\theta^*) > 0$)



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

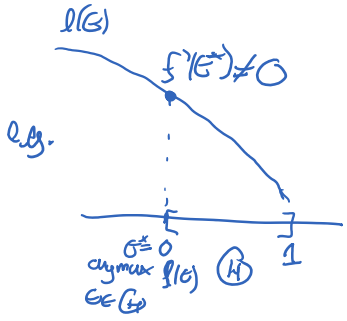
(B) only local result in general

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

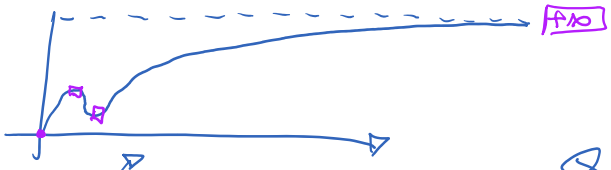
  Cond in this $Df(\theta^*) = 0 \Rightarrow$ sufficient for θ^* to be global min

- otherwise, for smooth fct., looking at zero gradient pts and boundary pts give you enough information to find global optima

(A) be careful with boundary cases
i.e. $\theta^* \in$ boundary(Θ)



another example:



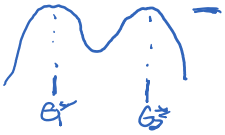
example where the MLE does not exist

* Some notes about MLE

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

  $\Theta =]0,1[$

- is not nec. unique [i.e. multiple maximas] e.g. mixture models



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

example II : multinomial distribution

Suppose X_i is discrete R.V. on k choices "multinoulli"

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

but instead, convenient to encode with unit basis in $\mathbb{R}^k$ "one hot encoding"

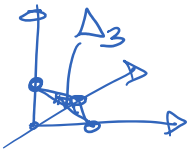
i.e. $\Omega_{X_i} = \{e_1, \dots, e_k\}$ where $e_j \in \mathbb{R}^k$

$e_j = (0, \dots, 0, 1, 0, \dots, 0)$ — j^{th} coordinate

parameter for discrete R.V. : $\pi \in \Delta_k$ $\quad (\Theta = \Delta_k)$

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

probability simplex on k choices



we will write $X_i \sim \text{Mult}(\pi)$ "multinoulli"

(π : parameter)

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

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

"multinomial distribution"

$X \in \mathbb{N}^k$

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

pmf for X :

$$p(x|\pi) = \binom{n}{(x)_1, \dots, (x)_k} \prod_{j=1}^k \pi_j^{(x)_j} \qquad x = (n_1, \dots, n_k)$$

multinomial coeff. $\quad \binom{n}{n_1, \dots, n_k} \triangleq \frac{n!}{n_1! \cdots n_k!}$

15h37

multinomial MLE :

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

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

MLE : $\hat{\pi}_{MLE}(x) = \text{argmax}_{\pi} \ell(\pi)$

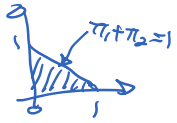
$\pi \in \mathbb{R}^K$
s.t. $\pi \in \Delta_K$ } constraint

two options:

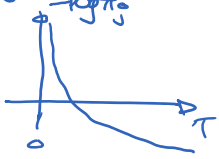
a) reparameterize problem so that Θ is full dimensional

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

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



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



can try unconstrained opt. on $\pi_1, \dots, \pi_{K-1}$ of $l(\pi_1, \dots, \pi_{K-1})$
hoping soln is in the interior of constraint set (and it usually will)

b) use Lagrange multiplier approach to handle equality constraints on Δ_K
[and still ignore $\pi_j \in [0,1]$]

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

$\sum_{j=1}^K \pi_j = 1 \Rightarrow 1 - \sum_{j=1}^K \pi_j = 0 \triangleq g(\pi)$

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

λ: Lagrange multiplier

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

i.e. $\nabla_\pi J(\pi, \lambda) = 0$ (necessary for local opt.; check "bordered Hessian" to get local min or max)

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

$\Rightarrow g(\pi) = 0$

(see Wikipedia)

$l(\pi) = \sum_{j=1}^K n_j \log \pi_j$ (strictly concave fct. in π_j)

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

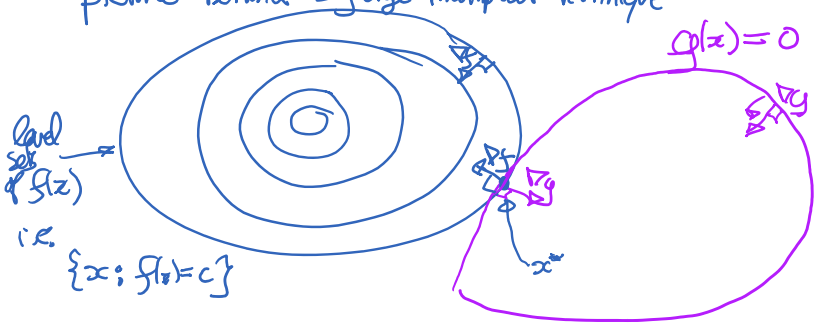
want $g(\pi^*) = 0$ i.e. $\sum_j \pi_j^* = 1 \Rightarrow \sum_j \frac{n_j}{\lambda} = 1$

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

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

$\boxed{\pi_j^* = \frac{n_j}{n}}$ MLE for multinomial

picture behind Lagrange multiplier technique



$\nabla f(x^*) = \lambda \nabla g(x^*)$

i.e. ∇f & ∇g are parallel

Statistical decision theory

A) Bias-variance decomposition for squared loss

estimator: function from data (observation) to parameters

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

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

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

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

most standard tool: frequentist risk of an estimator

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

(L: (statistical) loss fct.; expectation: over possible data)

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

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

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

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

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

$$2 \langle \theta - \mathbb{E}[\hat{\theta}], \mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}] \rangle$$

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

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

risk for squared loss = bias^2 + variance

bias-variance decomposition ("tradeoff")

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

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

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

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

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

variance $(\hat{\theta}_n) \to 0$

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