4 WHAT IS BAYESIAN INFERENCE, WHEN DO I USE IT, AND WHY?

WHAT? A PRINC. LEARN. FRAMEWORK (MLE, ERM, ...) BASED ON HAVING A PROB OPINION & UPDATING IT

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)}$$

WHEN? ALWAYS! (COMP. INFEASIBLE, LAZY)

WHY? EASIER:

- TO INTERPRET
- TO INCLUDE PRIOR KNOWLEDGE
- DECISION THEORY (!)
- TO NOT OVERFIT (NO FITTING ⇒ NO OVERFITTING)

⊖ COMPUTATION

⊖ PRIORS

LESS OF A CONCERN

EXAMPLE: COVID-19 NATIONAL SURVEY

$n = 1318$, $y_i \in \{0,1\}$, $\sum_{i=1}^{n} y_i = 41$

$\hat{\theta} = \frac{\sum y_i}{n} = \frac{41}{1318} \approx 0.031 = 3.1\%$ $\times$ QUANTIFY UNCERTAINTY!

- THIS IS A STAT. TASK (QUANT. UNC. & MAKING DECISIONS)

- WE NEED A STAT. MODEL (A PROBABILISTIC MODEL) (WE'LL OPT FOR A PARAMETRIC MODEL)

A PARAMETRIC MODEL FOR SUCCESS RATE

$\theta \in [0,1]$

$$y_i \mid \theta = \begin{cases} 1, & \text{WITH } \theta \text{ PROB.} \\ 0, & \text{WITH } 1-\theta \text{ PROB.} \end{cases}$$

OR

$$P(y_i = k \mid \theta) = \theta^k (1-\theta)^{1-k}$$

OR

$$y_i \mid \theta \sim \text{Bernoulli}(\theta)$$

(LIKELIHOOD)

MAX. LIK. VIEW

- y_i ARE RANDOM SAMPLES FROM DGP (WE COULD SAMPLE MORE)
  ↖ R.V.s
- θ IS AN UNKNOWN (BUT CONSTANT)
- FIND $\hat{\theta}$ THAT MAX. THE LIKELIHOOD

$$\rightarrow \underbrace{\hat{\theta}}_{\text{R.V.}} = \underbrace{\underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} p(y_i|\theta)}_{\text{OPTIMIZATION (FITTING)}} = 3.1\%$$

HOW DO I QUANTIFY UNC.?

$$P(\theta < 0.02 | y) = ?$$ ← THIS QUESTION IS INAPPROPRIATE

$$\hat{\theta} - \theta \xrightarrow{D} N(0, \sigma_\theta^2)$$

THE CONF. INTERVAL

FOR EXAMPLE,

$$CI_{95\%} \approx \Big[\underbrace{\hat{\theta}}_{\text{R.V.}} - 2 \cdot \sigma_\theta \, , \, \underbrace{\hat{\theta}}_{\text{R.V.}} + 2\sigma_\theta\Big]$$

(95% OF SUCH INTERVALS WILL CONTAIN TRUE VALUE θ)

"95% OF THE CI WILL CONTAIN θ" ✗

$\neq$

"θ IS IN CI WITH 95% PROB." ✓

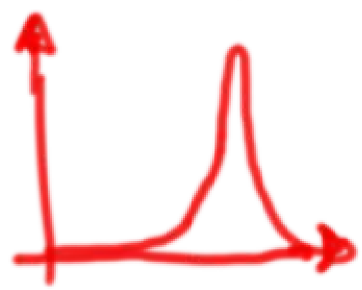
95% OF STUDIES (AT RANDOM) $CI_{95\%} = [0, 1]$

5% -'' - $CI_{95\%} = [-5, -4]$

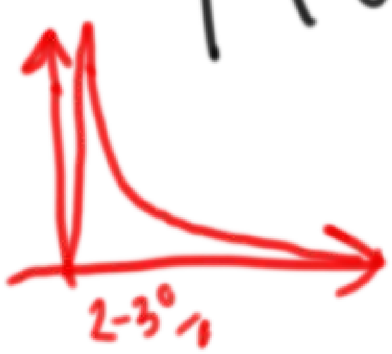
95% CI

BAYESIAN VIEW

- y_i ARE SAMPLES FROM OUR DGP
- θ IS A CONSTANT (UNKNOWN)
  R.V. ~~BE~~ TO REPRESENT MY UNCERTAINTY
- I'M GOING TO UPDATE MY OPINION COND. ON DATA



$$p(\theta) \longrightarrow p(\theta|y)$$



$$\underset{\text{POSTERIOR}}{p(\theta|y)} = \overbrace{\underset{\text{LIKELIHOOD}}{p(y|\theta)}\ \underset{\text{PRIOR}}{p(\theta)}}^{\text{KEY (SHAPE)}} \Big/ \underset{\text{NORMALIZATION}}{p(y)}$$

OUR EXAMPLE

⊖ COMPUTATION

$$y_i \mid \theta \sim_{iid} \text{Bernoulli}(\theta)$$

$$\theta \sim \text{Beta}(\alpha_0, \beta_0)$$

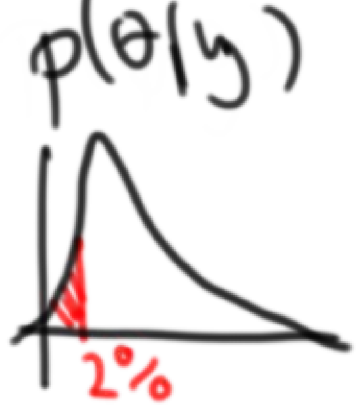
$$\theta \mid y \sim \text{Beta}(\underbrace{\alpha_0 + \sum y_i}_{\#\,\text{POS}}, \beta_0 + n - \sum y_i) \qquad \#\,\text{NEG}$$

$$1)\ \frac{p(\theta)}{\boxed{\int p(\theta)\, d\theta}}$$

ALWAYS DERIVE POST?

NO (RARELY)

A "NICE"



$$P(\theta > 2\% \mid y) = \int_{2\%}^{2\%} p(\theta \mid y)\, d\theta$$

$$P(\theta > 2\%) = \int_{0}^{c} p(\theta)\, d\theta$$

(+) EASIER TO INTERPRET

(+) EASY TO INCL. PRIOR KNOWLEDGE

(+) EASIER DECISION THEORY

(−) COMPUTATION (= TIME & EFFORT)

LINEAR REGRESSION

$$y_i = \beta^T x_i + \varepsilon_i , \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$y_i \mid \beta, \sigma^2, x_i \sim N(\beta^T x_i , \sigma^2)$$

$$p(y \mid \dots) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\underbrace{(y - X\beta)^T(y - X\beta)}_{\text{SUM OF SQ. RESIDUALS}}\right\}$$

PRIOR:

$$p(\beta, \sigma^2) = p(\beta \mid \sigma^2)\, p(\sigma^2)$$

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

$\sigma^2 \in \mathbb{R}$

$$\underbrace{N_k(\mu_\beta, \sigma^2 V_\beta) \times IG(\alpha, \beta)}_{NIG}$$

$$\boxed{p(\beta, \sigma^2 \mid \dots)}$$

REGULARIZATION

$$\beta^* = (\lambda I + X^T X)^{-1} X^T y$$

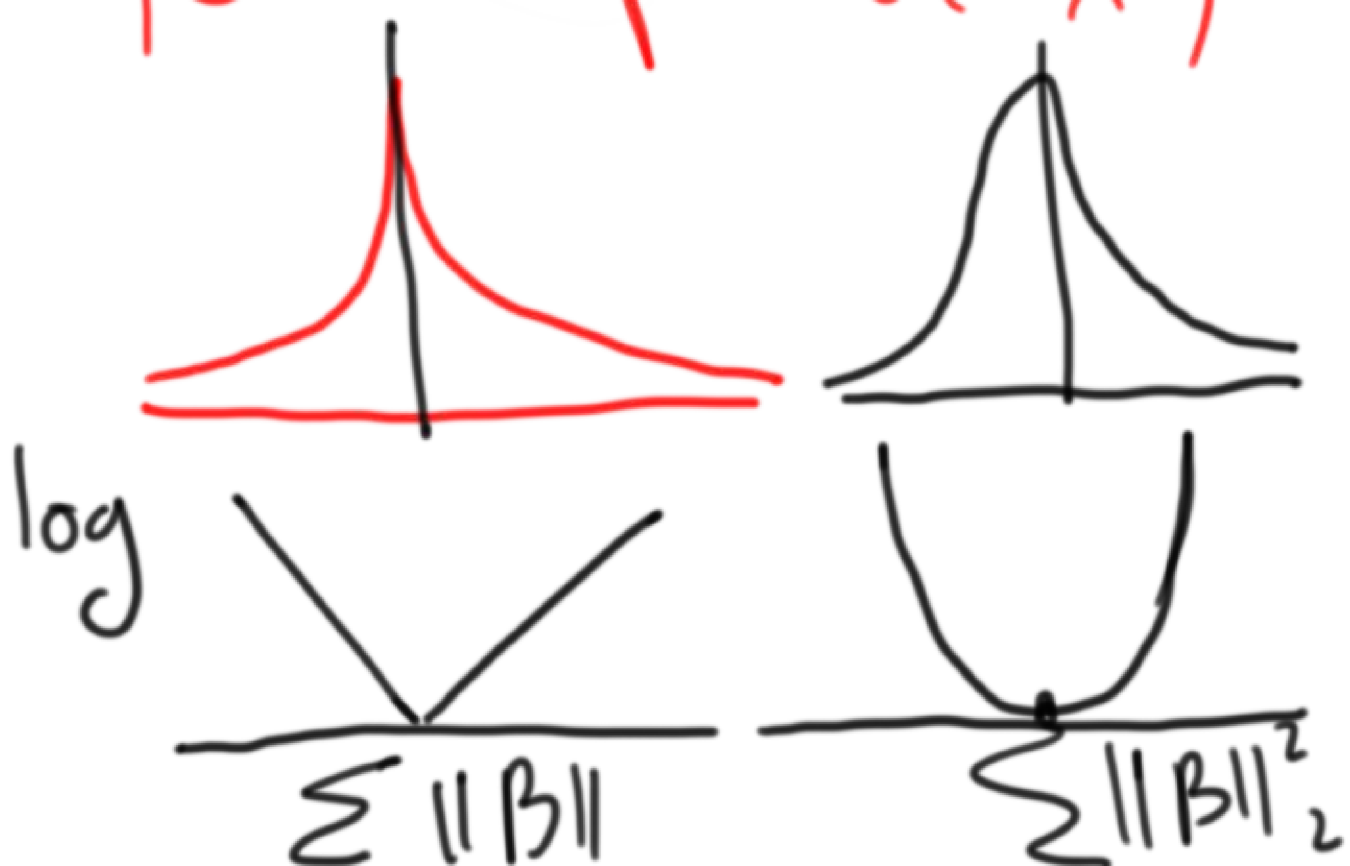
RIDGE REGRESSION

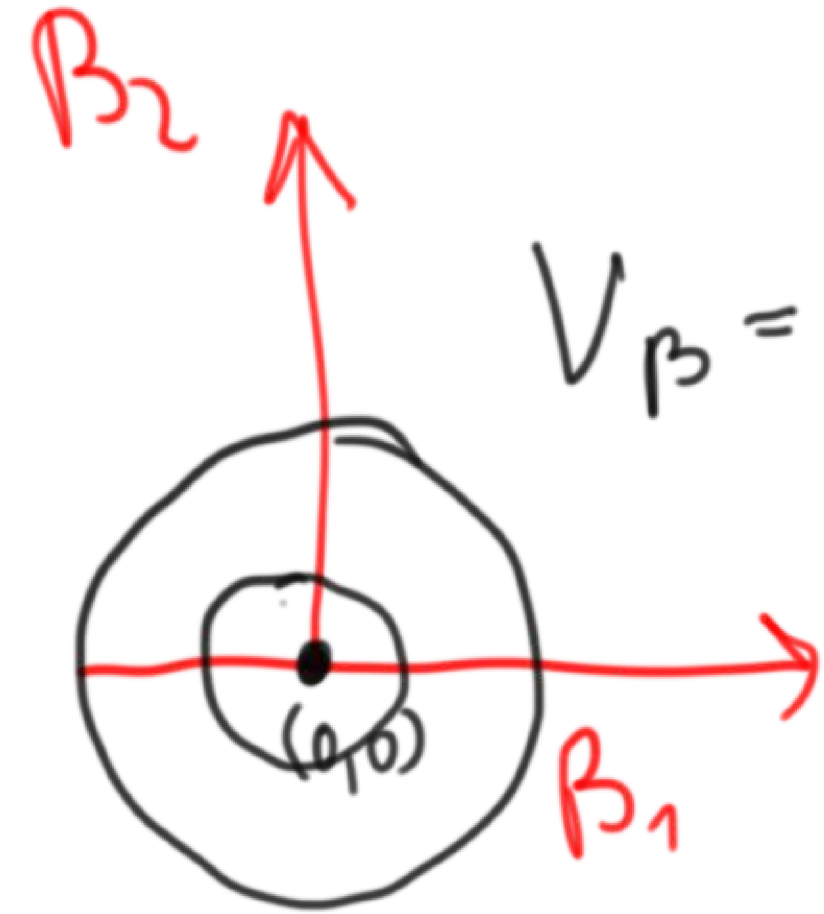
$$\beta_{OLS} = (X^T X)^{-1} X^T y$$

$$\underbrace{\beta|\sigma^2 \sim N_k(\mu_\beta^{=0}, V_\beta)}_{\text{PRIOR ON } \beta}$$

LASSO (L1)

$$\beta \sim \mathrm{Laplace}(1/\lambda)$$





$$V_\beta = \frac{1}{\lambda} \cdot I$$

(+) EASY TO INCLUDE PRIOR KNOWLEDGE (REG.) IS AN EXAMPLE)

(+) NO FITTING $\Rightarrow$ NO OVERFITTING

(−) COMPUTATION

COMPUTATION FOR BAYESIAN INF.

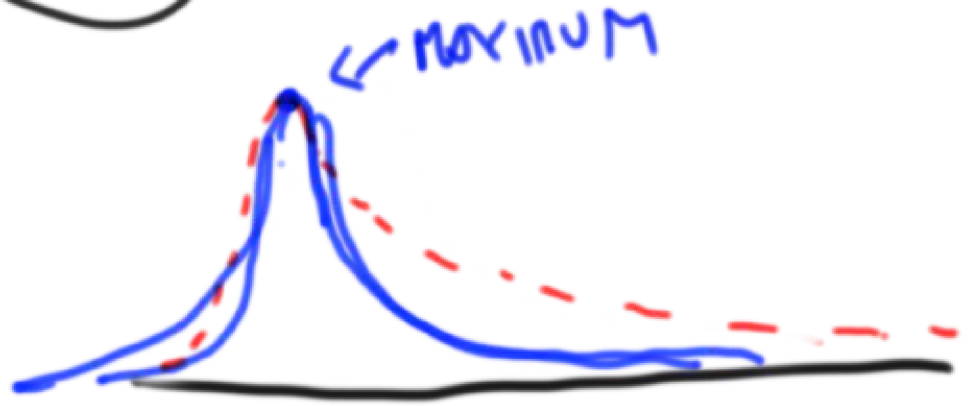
1. "NICE" POSTERIOR (CONJUGATE PRIOR) (RARE)

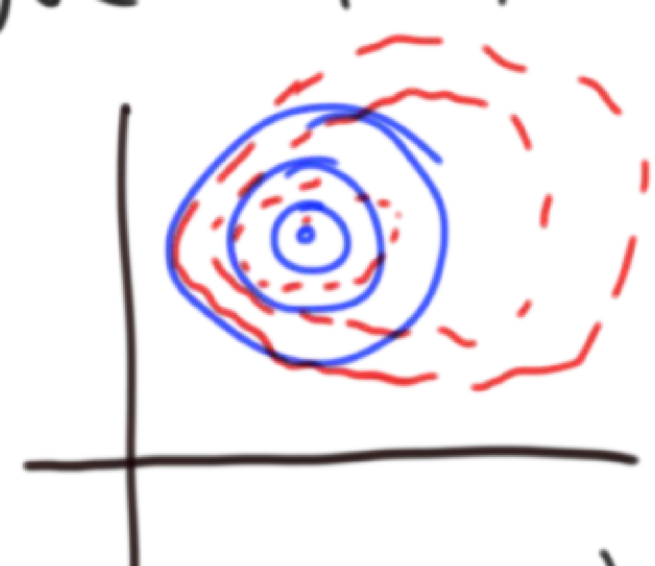
$$\int \theta \, p(\theta|y)\, d\theta$$

(NUMERICAL) $p(\theta > 2\%) = \int p(\theta|y)\, d\theta$

2. MCMC (UNBIASED, COMP. INTENSIVE)

3. STRUCTURAL APPROX. (BIAS)





$\theta_{MAP} = \arg\max \; p(\theta|y)$

$p(y|\theta)$

LAPLACE APPROX

VARIATIONAL INFERENCE