Bayes Rule.

$$P(A|B)P(B) = P(B|A)P(A)$$

or

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

A = Parameters (θ)

B = Data (D)

LIKELIHOOD

$$P(\theta|D) = \frac{P(D|\theta)\,P(\theta)}{P(D)}$$

PRIOR PROBABILITY

(like Gaussian with mean 0)

"ALSO ACTS AS REGULARIZER"

EVIDENCE

POSTERIOR PROBABILITY

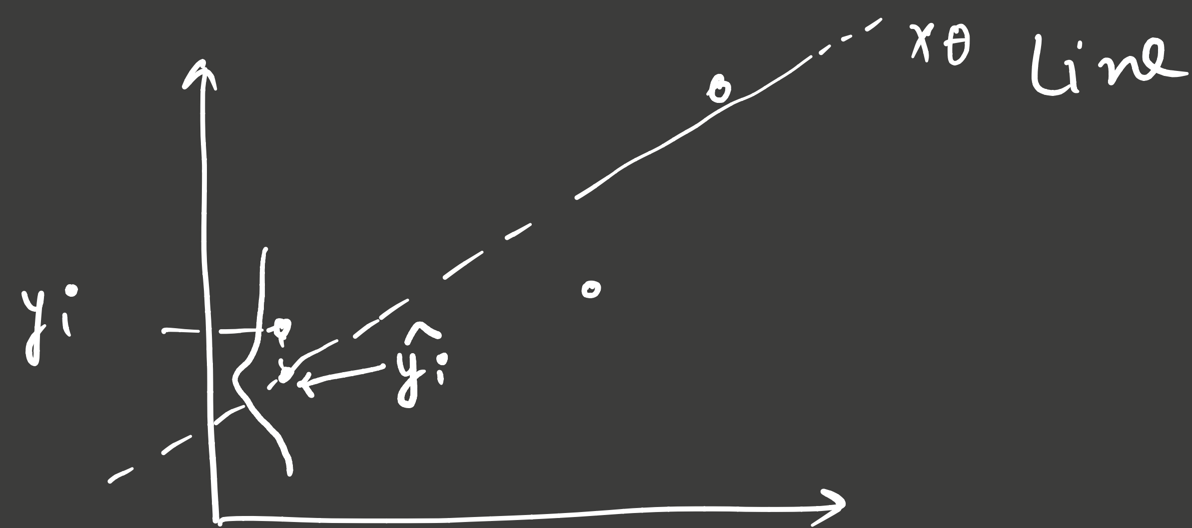
(Prob. after seeing data)

MLE FOR LINEAR REGRESSION

$$\hat{\theta}_{LS} = \underset{\theta}{\arg\min}\ \epsilon^T \epsilon = \underset{\theta}{\arg\min}\ (Y - X\theta)^T (Y - X\theta)$$

LEAST SQUARES SOL^N

1. $\epsilon_p \sim N(0, \sigma^2)$ (0 MEAN, KNOWN VARIANCE)
2. ϵ INDEPENDENT ACROSS OBSERVATIONS
3. $Y \sim N(X\theta, \sigma^2)$ $\{X\theta + N(0, \sigma^2)\}$



Likelihood = $P(\text{Data} \mid \text{Parameters})$

Data $= \langle y_i, x_i \rangle$; Parameters $= \theta, \sigma$

$$p(y \mid X, \theta, \sigma) = p(y_1 \mid x_1, \theta, \sigma^2)\, p(y_2 \mid x_2, \theta, \sigma^2) \ldots$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i, \theta, \sigma^2)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{1}{2\sigma^2}(y_i - x_i\theta)^2}$$

$$\log \text{likelihood} = \sum_{i=1}^{n} \text{Constant} * (y_i - x_i\theta)^2 = \text{Constant} * \sum_{i=1}^{n} \epsilon_i^2$$

$$\boxed{\hat{\theta}_{MLE} = \underset{\theta}{\arg\max}\, (y - X\theta)^T (y - X\theta) = \hat{\theta}_{LS}}$$

UNDER NORMALLY DISTRIBUTED RESIDUAL

MAP FOR LINEAR REGRESSION

1. $\eta \sim N(0, \sigma^2)$
2. η is independent across observations
3. $y \sim N(X\theta, \sigma^2)$
4. θ has Gaussian prior i.e $\theta \sim N(0, \tau^2 I)$ ← Identity

$$P(\theta|D) \propto P(D|\theta)\,P(\theta) \Rightarrow \log P(\theta|D) \propto \log P(D|\theta) + \log P(\theta)$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\arg\max} \{\log P(D|\theta) + \log P(\theta)\}$$

$$= \underset{\theta}{\arg\min}\,(y - X\theta)^T (y - X\theta) + \lambda^2 \theta^T \theta$$

$$\boxed{\text{MAP with GAUSSIAN PRIOR} \Rightarrow \text{RIDGE REGRESSION}}$$

$$\text{PRIOR} \xRightarrow{\text{LEADS TO}} \text{REGULARIZATION}$$

$\theta \sim N(0, \tau^2 I)$: RIDGE REGRESSION

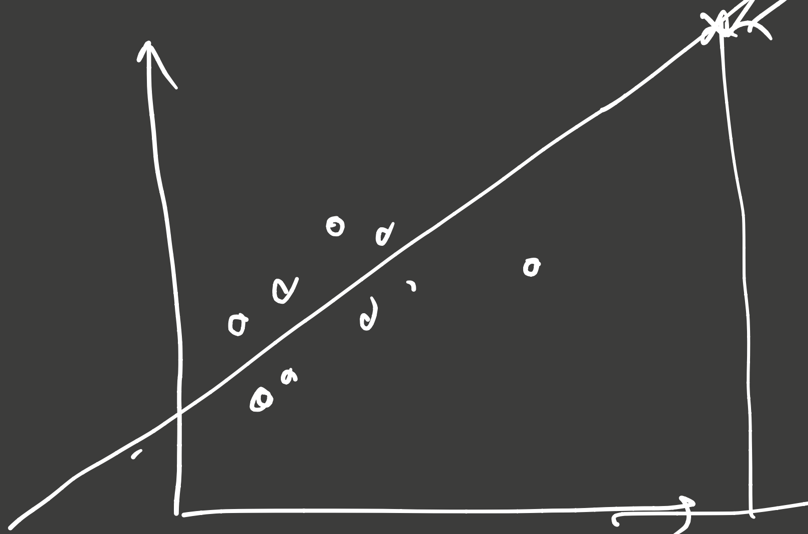
$\theta_i \sim \text{LAPLACE}(0, t)$: LASSO

BAYESIAN LINEAR REGRESSION

* POINT ESTIMATE → NO MEASURE OF UNCERTAINTY

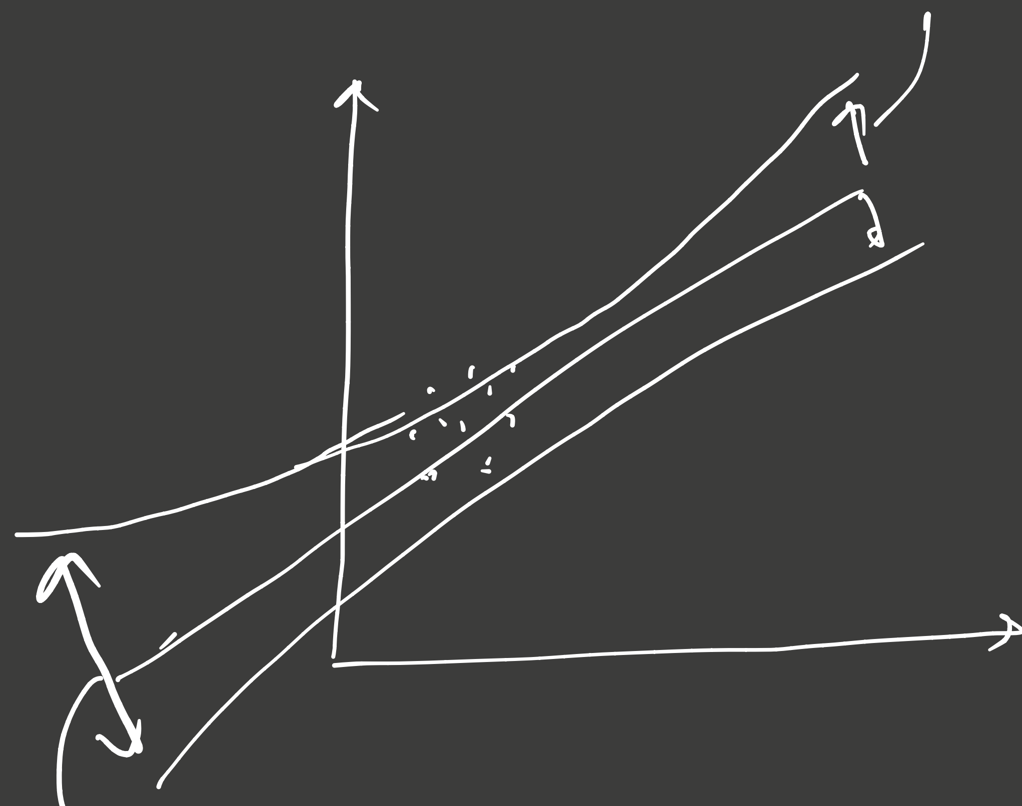
NON "FULLY" BAYESIAN

(MAP; MLE)



TEST PREDICTION

"FULLY" BAYESIAN



UNCERTAINTY

Q: what is $p(y^*|x^*, D)$

$y_i \sim N(x_i \theta, a^{-1})$ where $x_i \in \mathbb{R}^d$

$a^{-1} \rightarrow$ Precision $\left(a = \frac{1}{\sigma^2}\right)$ UNIVARIATE

$\theta \sim N(0, b^{-1} I)$

$b^{-1} \rightarrow$ Precision MULTIVARIATE

$a, b > 0$

ASSUMPTION

a, b are KNOWN

$$p(D|\theta) \propto e^{-a/2 (y-X\theta)^T (y-X\theta)}$$ {likelihood}

$$p(\theta|D) \propto p(D|\theta)\, p(\theta)$$

$$\propto e^{-a/2 (y-X\theta)^T (y-X\theta)} \cdot e^{-b/2\, \theta^T \theta}$$

$$\propto e^{-a/2 (y-X\theta)^T (y-X\theta) - b/2\, \theta^T \theta}$$

Let $Z = a(y-X\theta)^T (y-X\theta) + b\,\theta^T\theta$

$$= a y^T y - 2a\,\theta^T X^T y + \theta^T (a X^T X + bI)\theta$$

Let $Z = (\theta-\mu)^T \Lambda (\theta-\mu)$

$$\theta^T \Lambda \theta - 2\theta^T \Lambda \mu + \text{constant} = a y^T y - 2a\,\theta^T X^T y + \theta^T (a X^T X + bI)\theta$$

$$\Rightarrow \Lambda = a X^T X + b I$$

and $\mu = a \Lambda^{-1} X^T y$

$$p(\theta | D) \sim N(\theta | \mu, \Lambda^{-1})$$

NOW WE WANT

$p(y^* | x^*, D, a, b)$ Predictive distributⁿ,

$\hookrightarrow (x^*, y^*)$ new point

USE ALL θ

$$p(y^* \mid x^*, D, a, b) = \int p(y^* \mid \theta, x^*, a)\, p(\theta \mid D, a, b)\, d\theta$$

$$= \int N(y^* \mid x^{*}\theta, a^{-1})\, N(\theta \mid \mu, \Lambda^{-1})\, d\theta$$

$$\propto e^{-\frac{a}{2}(y^* - x^{*}\theta)^2 - \frac{1}{2}(\theta - \mu)^T \Lambda (\theta - \mu)}\, d\theta$$

F

$$F = (\theta - m)^T L (\theta - m)$$

$$= \theta^T (a x^* x^{*T} + \Lambda)\hat{\theta} - 2\theta^T (x^* y^* a + \Lambda\mu) + a y^2$$

$$L = a x^* x^{*T} + \Lambda$$

$$m = L^{-1} (a y^* x^* + \Lambda \mu)$$

$\vdots$

$$p(y^* \mid x^*, D) = N\left(y^* \mid u, \frac{1}{\lambda}\right)$$

$$u = \mu^T x^*$$

$$\frac{1}{\lambda} = \frac{1}{a} + x^{*T} \Lambda^{-1} x^*$$

GRAPHICAL MODEL FOR BAYESIAN LINEAR REGRESSION

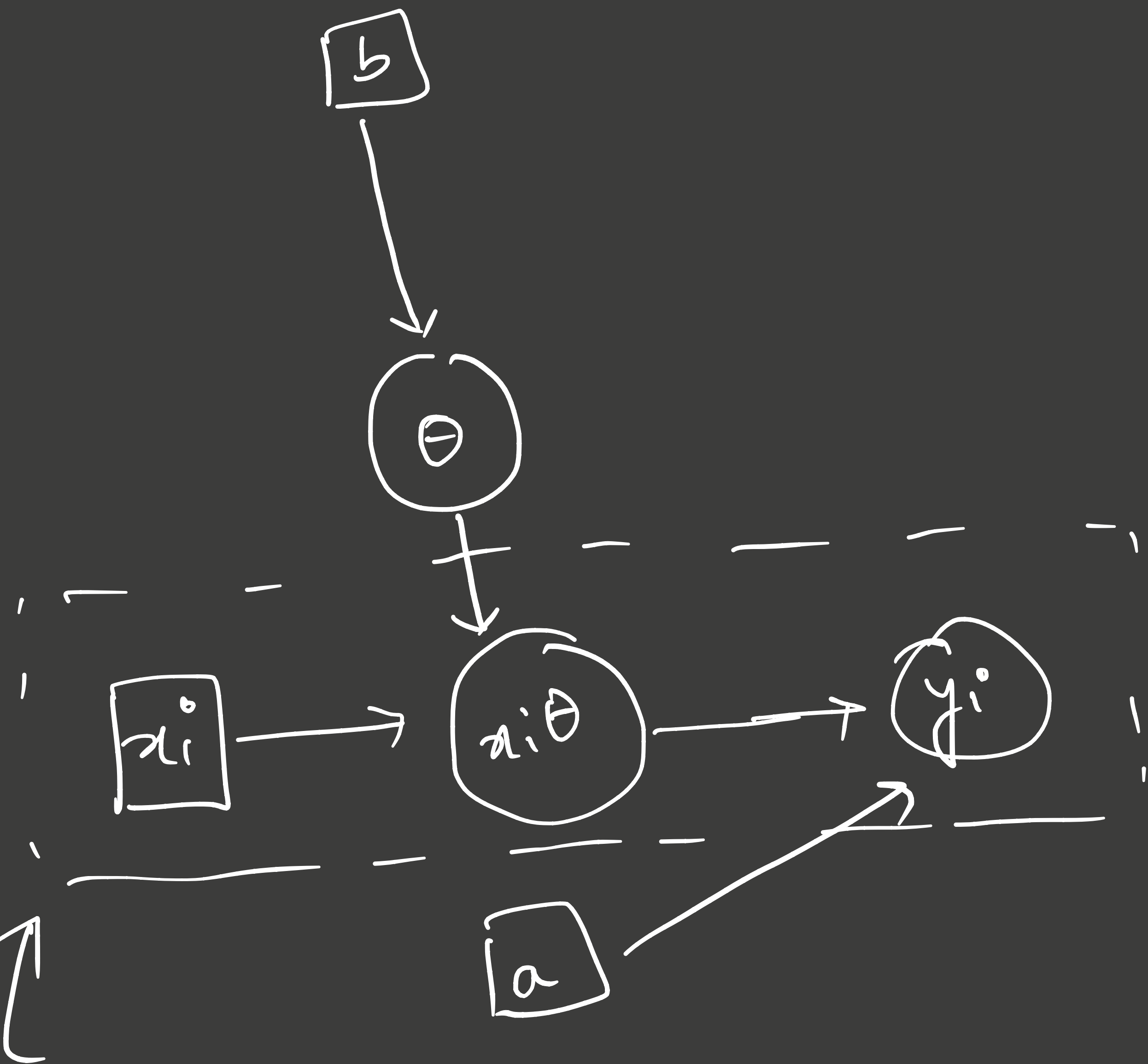


Plate model. (repeated over time | i)