Bayesian Update

→ Given new data, old posterior becomes new prior.

→ Say we go through n examples ⟶ n posteriors

$\boxed{\text{Value w/ highest posterior = Maximum a Posteriori (MAP)}}$

P(D) is again product of likelihoods and priors

$$P(D) = \int P(D|H)\, P(H)\, dH$$ } often intractable

solution

Conjugate Priors

convenience

prior, posterior from same distribution ⟶ simplified calculations

[no need for computing product of likelihood and prior]

} product of likelihood and prior gives same distribution as prior with changed parameters ⟶ posterior follows known distribution, integral becomes easy; known normalising constant

↓ why?

god knows

CONJUGATE DISTRIBUTIONS

Beta posterior	Gamma posterior	Gaussian posterior
Beta * Bernoulli	Gamma * Poisson	Normal * Normal
Beta * Binomial	Gamma * exp.	
Beta * -ve binomial		
Beta * Geometric		

Bayes' Theorem

$$\underbrace{P(H|D)}_{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}} \cdot \overbrace{P(H)}^{\text{prior}}}{\underbrace{P(D)}_{\text{normalising factor}}}$$

likelihood → likelihood of data given our hypothesis

prior → likeliness of hypothesis before looking at data

$$P(D) = \sum_i P(D|H_i)\, P(H_i)$$ } total probability of seeing data over all hypotheses

$P(D|H_i)$ ↓ if H_i is true, would I see this data?

$P(H_i)$ ↓ is H_i true?

$\boxed{\text{posterior} \propto \text{likelihood} \cdot \text{prior}}$

Bayesian Linear Regression

- Frequentist linear regression: entirely on data

  $\hat{\beta} = (X^T X)^{-1} X^T Y$

- Small dataset → overfitting

  ∴ Stochastic approach → Bayesian

  ↓

  distribution of possible values

Note:

- Fewer datapoints → greater uncertainty → greater variation
- All datapoints → Frequentist ≈ Bayes

  Why?

  Likelihoods wash out priors

  $$P(H|D) \propto P(D|H)\, P(H)$$

  $$\propto \underbrace{\left[\prod_{i=1}^{N} P(D_i|H)\right]}_{\text{assuming iid}} P(H)$$

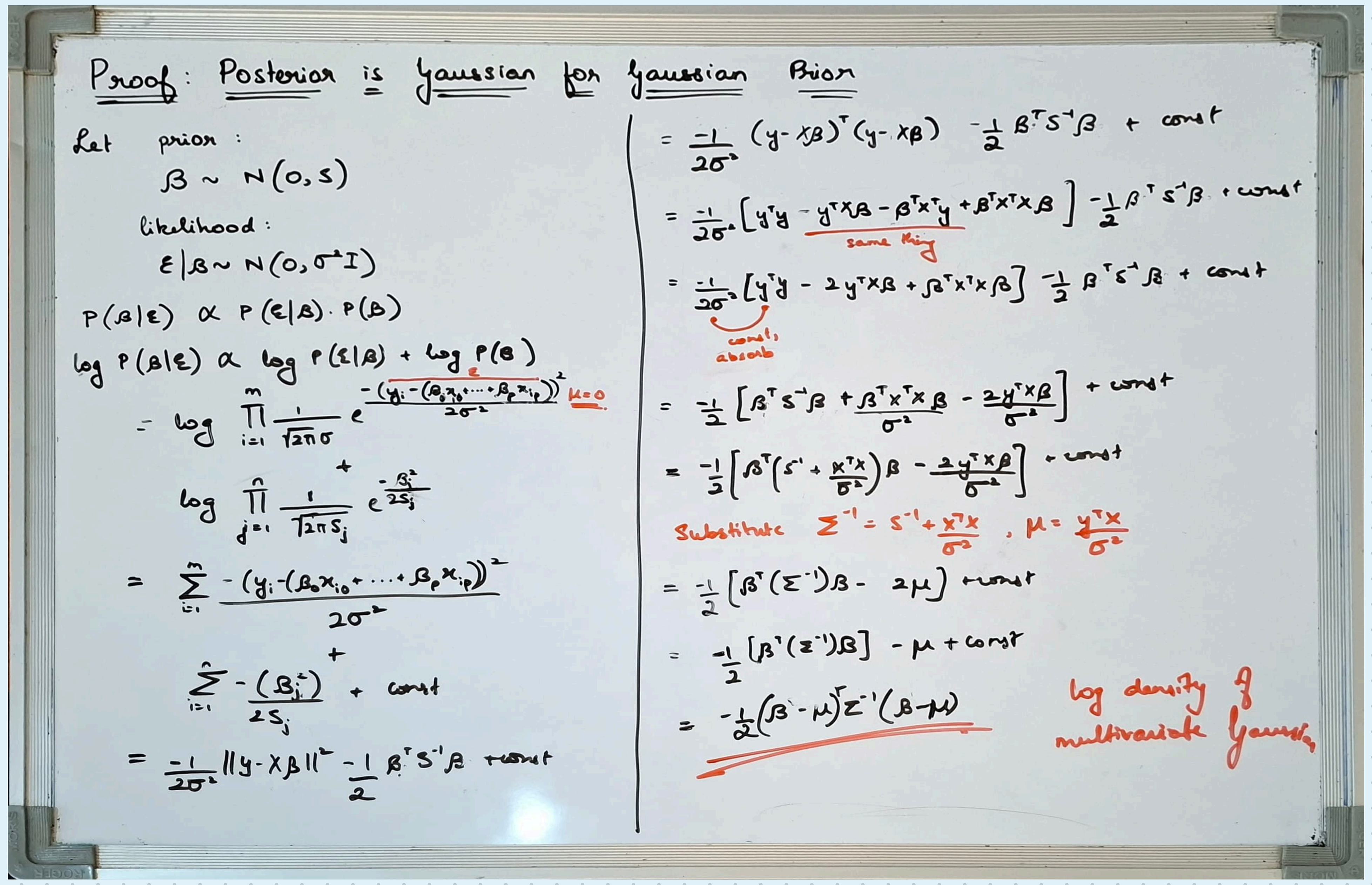
Extremely efficient for tiny dataset

Well suited for online learning

No prior knowledge of dataset needed.

Inference takes time

Not worth it if large amt of data present



Proof: Posterior is Gaussian for Gaussian Prior

Let prior:

$\beta \sim N(0, S)$

likelihood:

$\varepsilon|\beta \sim N(0, \sigma^2 I)$

$P(\beta|\varepsilon) \propto P(\varepsilon|\beta) \cdot P(\beta)$

$\log P(\beta|\varepsilon) \propto \log P(\varepsilon|\beta) + \log P(\beta)$

$= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - (\beta_0 x_{i0} + \dots + \beta_p x_{ip}))^2}{2\sigma^2}}$ $\mu = 0$

$+ \log \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi S_j}} e^{-\frac{\beta_j^2}{2S_j}}$

$= \sum_{i=1}^{m} \frac{-(y_i - (\beta_0 x_{i0} + \dots + \beta_p x_{ip}))^2}{2\sigma^2} + \sum_{i=1}^{n} -\frac{(\beta_j^2)}{2S_j} + \text{const}$

$= \frac{-1}{2\sigma^2} \|y - X\beta\|^2 - \frac{1}{2} \beta^T S^{-1} \beta + \text{const}$

$= \frac{-1}{2\sigma^2} (y - X\beta)^T (y - X\beta) - \frac{1}{2} \beta^T S^{-1} \beta + \text{const}$

$= \frac{-1}{2\sigma^2} [y^T y \underbrace{- y^T X\beta - \beta^T X^T y}_{\text{same thing}} + \beta^T X^T X \beta] - \frac{1}{2} \beta^T S^{-1} \beta + \text{const}$

$= \frac{-1}{2\sigma^2} [y^T y - 2y^T X\beta + \beta^T X^T X \beta] - \frac{1}{2} \beta^T S^{-1} \beta + \text{const}$

const, absorb

$= \frac{-1}{2} \left[\beta^T S^{-1} \beta + \frac{\beta^T X^T X \beta}{\sigma^2} - \frac{2y^T X\beta}{\sigma^2}\right] + \text{const}$

$= \frac{-1}{2} \left[\beta^T \left(S^{-1} + \frac{X^T X}{\sigma^2}\right) \beta - \frac{2y^T X\beta}{\sigma^2}\right] + \text{const}$

Substitute $\Sigma^{-1} = S^{-1} + \frac{X^T X}{\sigma^2}$, $\mu = \frac{y^T X}{\sigma^2}$

$= \frac{-1}{2} [\beta^T (\Sigma^{-1}) \beta - 2\mu] + \text{const}$

$= \frac{-1}{2} [\beta^T (\Sigma^{-1}) \beta] - \mu + \text{const}$

$= -\frac{1}{2} (\beta - \mu)^T \Sigma^{-1} (\beta - \mu)$

log density of multivariate Gaussian