

COMP90051

Statistical Machine Learning

Workshop Week 8

Xudong Han

https://github.com/HanXudong/COMP90051_Workshops

Bayesian Regression

- Frequentist V.S. Bayesian
- Bayesian regression with known variance
- Bayesian model selection
- Bayesian regression with unknown variance

Frequentist V.S. Bayesian

- Frequentist
Maximum Likelihood Estimation(MLE)
Generally reduces to minimizing the negative log-likelihood. Returns a point-estimate.

$$\theta_{MLE} = \operatorname{argmax}_{\theta} p(X|\theta) = \operatorname{argmax}_{\theta} \prod_i^n p(x_i|\theta) = \operatorname{argmax}_{\theta} \sum_i^n \log p(x_i|\theta)$$

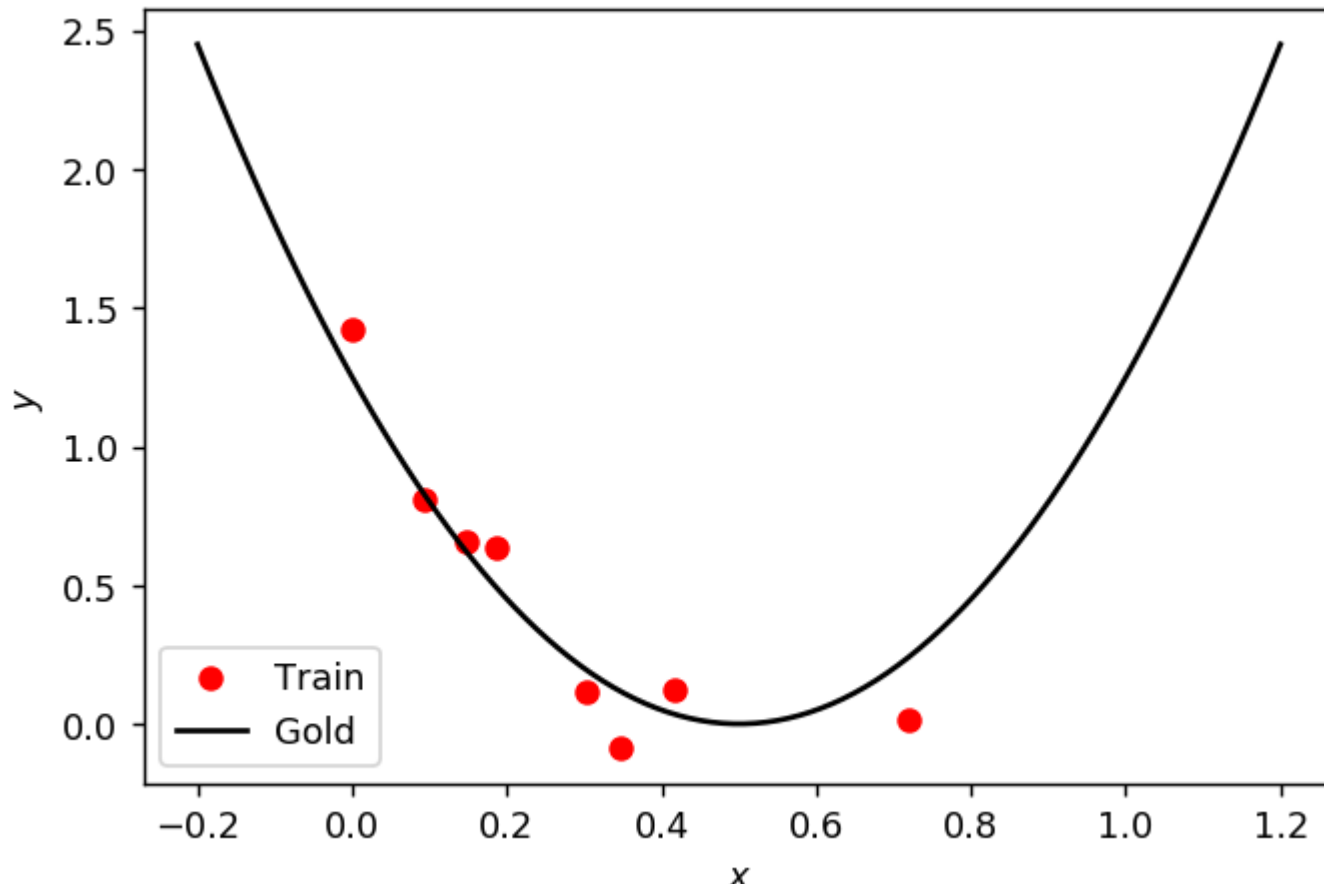
- Bayesian:

$$p(X|\theta) = \frac{\prod_i^n p(\theta|x_i)p(\theta)}{\int d\theta \prod_i^n p(\theta|x_i)p(\theta)}$$

1. Regression data set

$$x \sim \text{Uniform}[0, 1]$$

$$y|x, \sigma^2 \sim \text{Normal} \left[5 \left(x - \frac{1}{2} \right)^2, \sigma^2 \right]$$



Polynomial basis functions

Since the relationship between \mathbf{y} and \mathbf{x} is non-linear, we'll apply polynomial basis expansion to degree d .

$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^d \end{bmatrix}$$

2. Bayesian regression with known variance

- Prior

$$W|\gamma \sim \text{Normal}(\mathbf{0}, \gamma^2 I_m)$$

- Likelihood

$$p(y|X, W, \sigma) = \prod_{i=1}^n p(y_i|X_i, W, \sigma)$$

Since $y_i|X_i, W, \sigma \sim \text{Normal}(X_i^T W, \sigma^2)$,

$$y|X, W, \sigma \sim \text{Normal}(Xw, \sigma^2 I_n)$$

Bayesian regression with known variance

Given this formulation, the next step is to solve for the posterior over \mathbf{w}

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma, \gamma) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma)p(\mathbf{w}|\gamma)}{p(\mathbf{y}|\mathbf{X}, \sigma)}$$

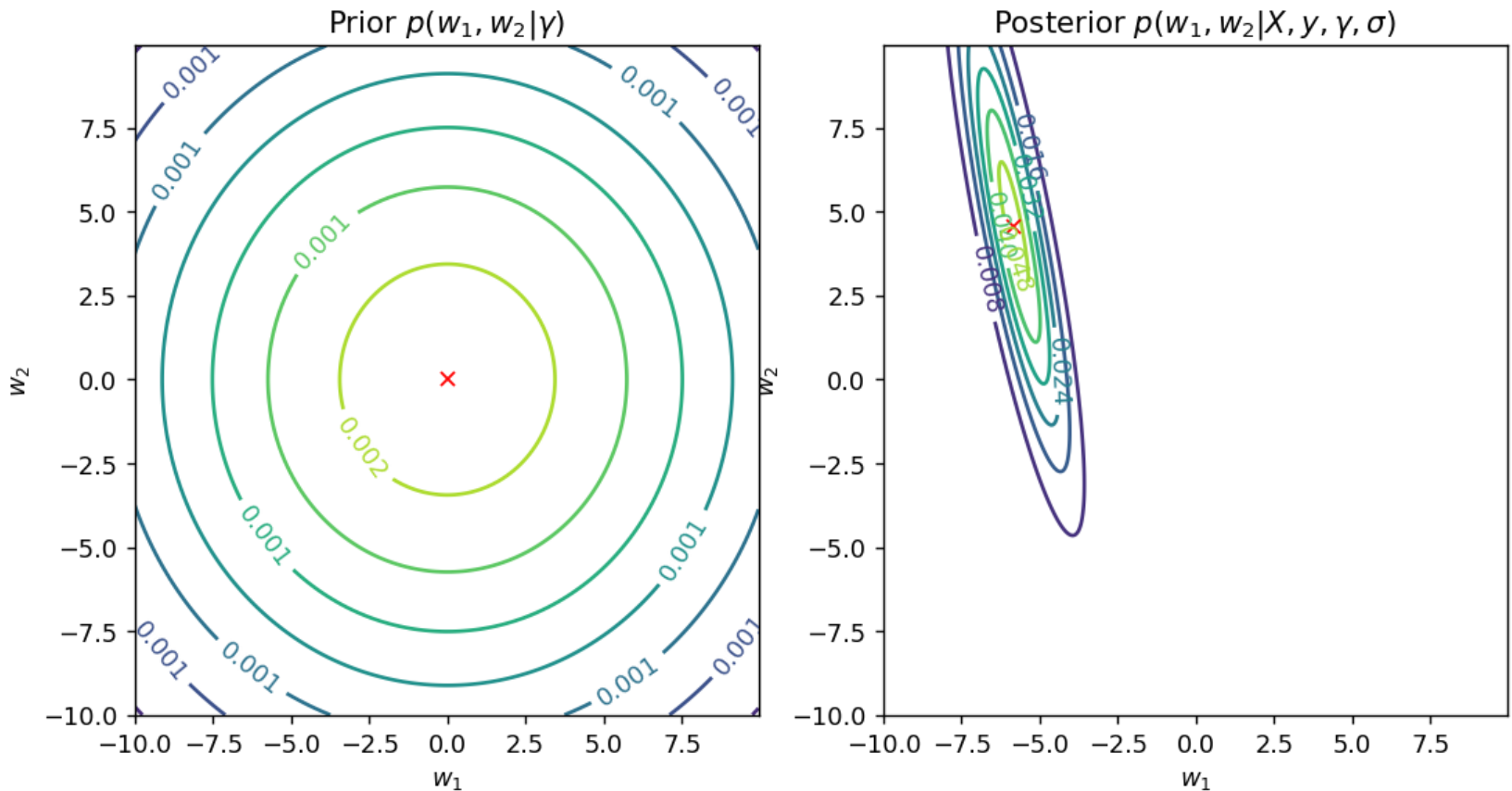
where $\mathbf{X} \in \mathbb{R}^{n \times m}$ is the feature matrix and $\mathbf{y} \in \mathbb{R}^n$ is the vector of target values for each instance.

In lectures, we derived the following solution:

$$\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma, \gamma \sim \text{Normal}(\mathbf{w}_N, \mathbf{V}_N)$$

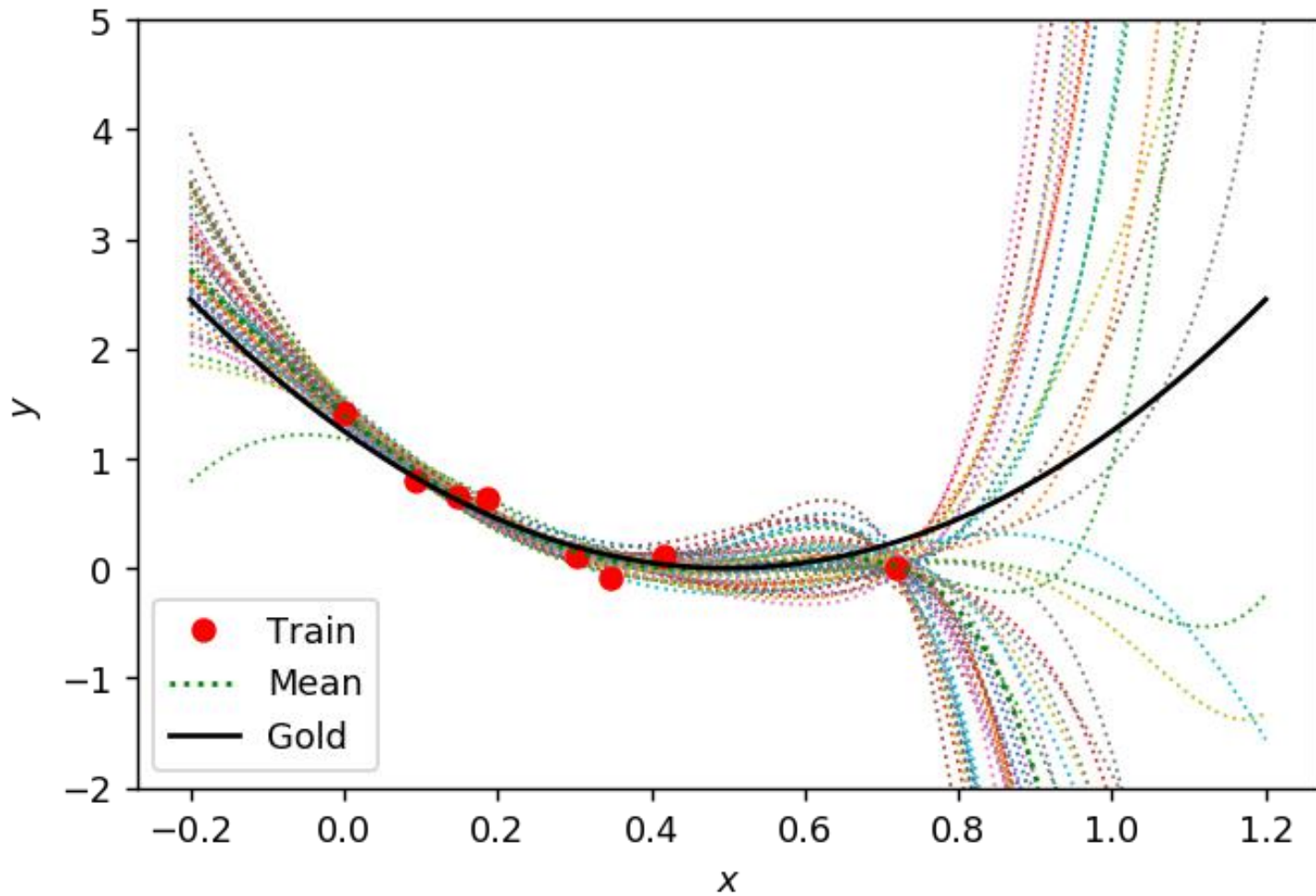
$$\text{where } \mathbf{V}_N = \sigma^2 \left(\mathbf{X}^T \mathbf{X} + \frac{\sigma^2}{\gamma^2} \mathbf{I}_m \right)^{-1} \text{ and } \mathbf{w}_N = \frac{1}{\sigma^2} \mathbf{V}_N \mathbf{X}^T \mathbf{y}.$$

plot the prior and posterior over w_1, w_2

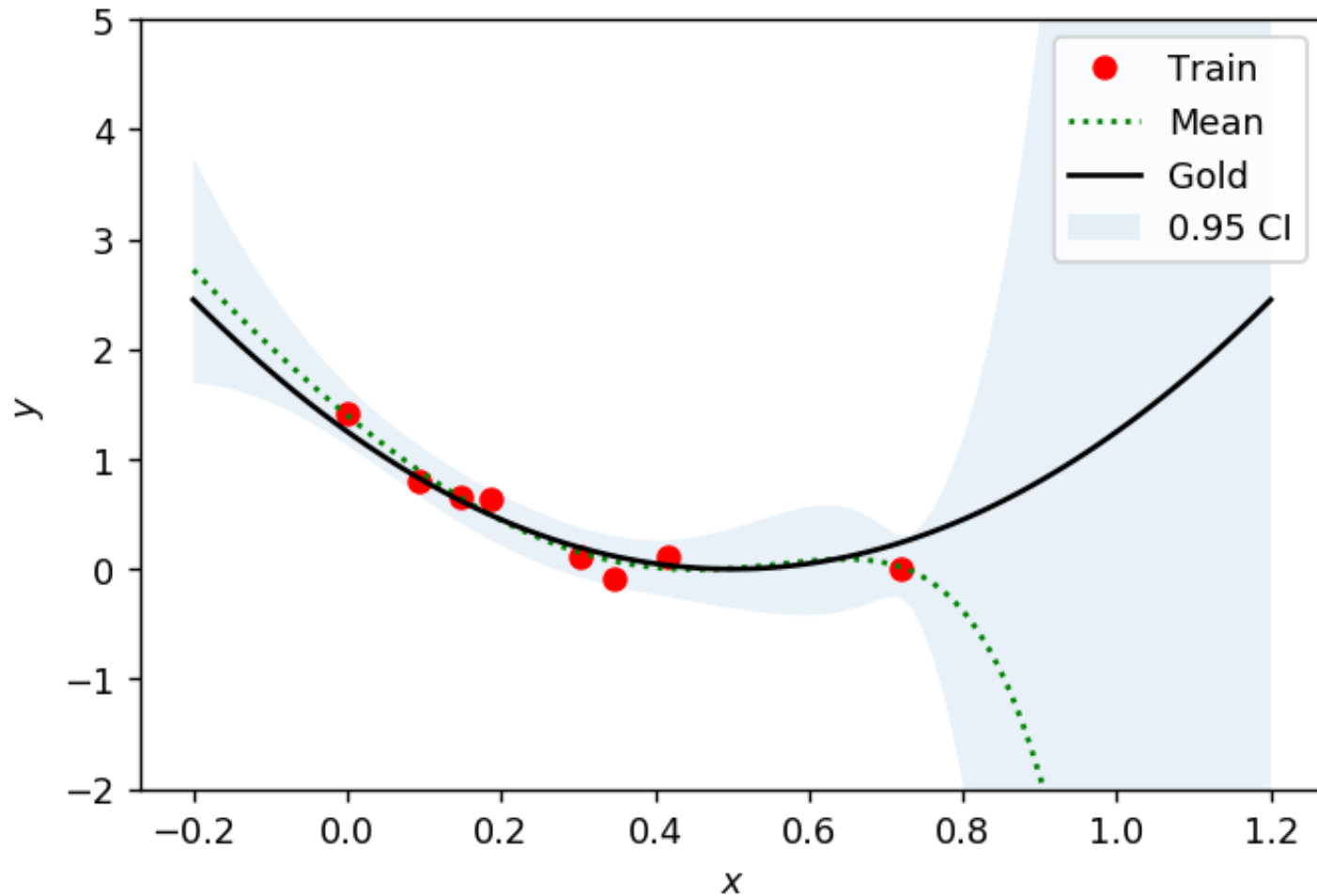


Discussion question: Can you explain why the prior and the posterior are so different? How is this related to the dataset? Why are the ellipses in the posterior not aligned to the axes? *You might want to change the parameter indices from 0,1 to other pairs to get a better idea of the full posterior.*

Bayesian inference



Bayesian inference



Bayesian model selection

