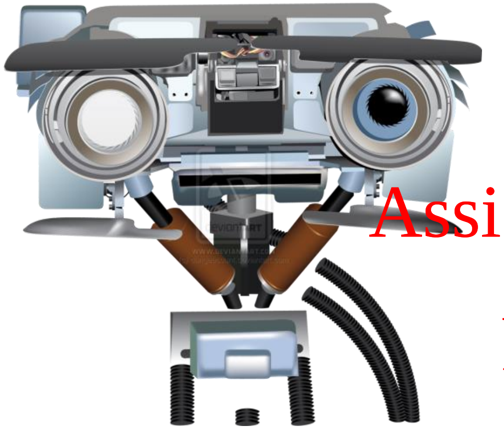
Announcements

Reminder: ps4 self-grading form out, due Friday 10/30

- pset 5 out Thursday 10/29, due 11/5 (1 week)
- Midterm grades will go up by Monday (don't discuss it yet)
- My Thursday office hours moved to 11am
- Lab this week – probabilistic models, ipython notebook examples



Bayesian Methods

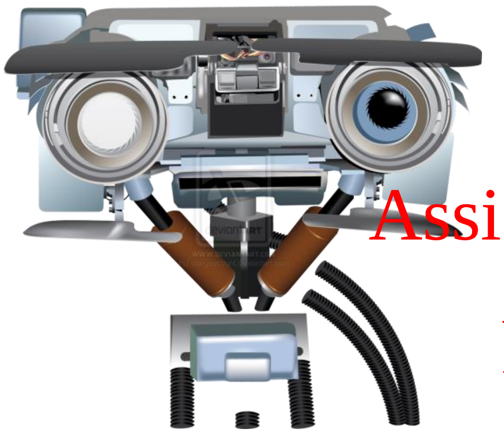
CS542 Machine Learning

Bayesian Methods

- Before, we derived cost functions from maximum likelihood, then added regularization terms to these cost functions

- Can we derive regularization directly from probabilistic principles?

- Yes! Use Bayesian methods

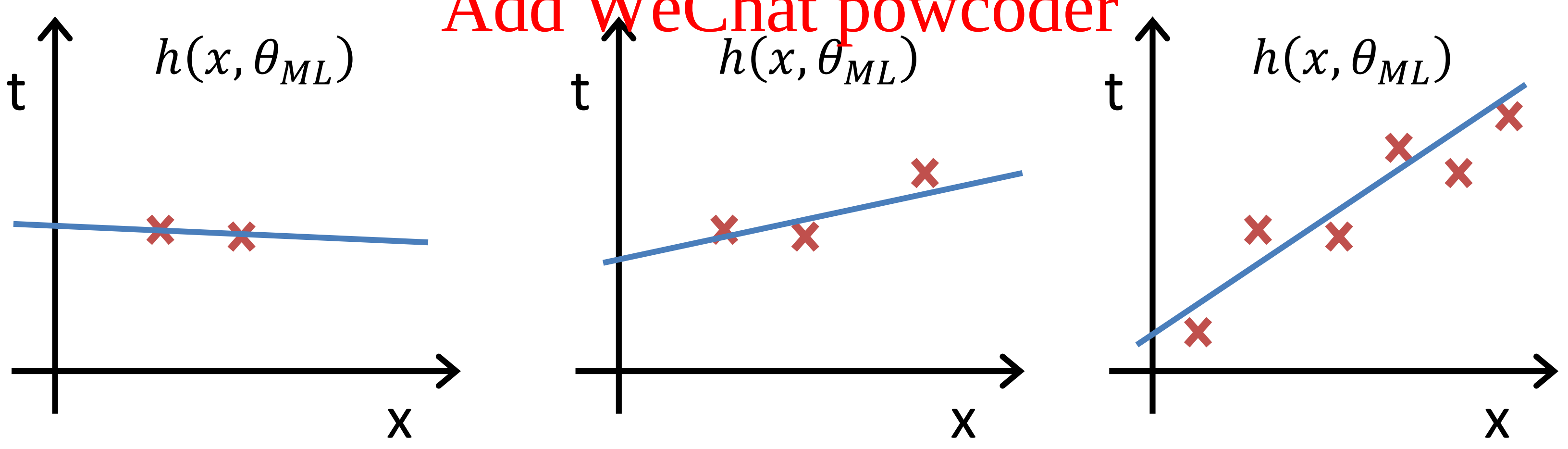


Bayesian Methods

Motivation

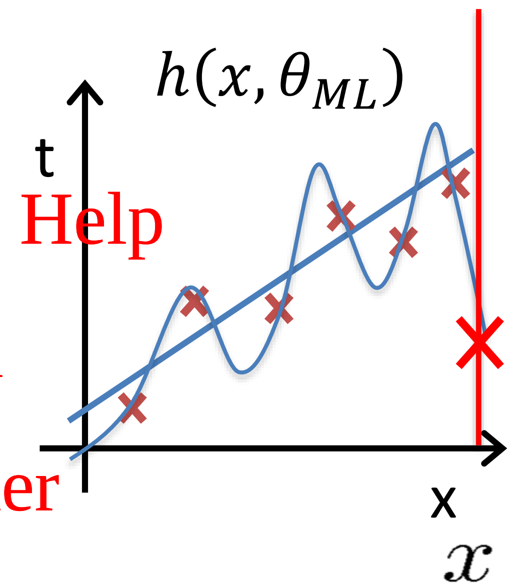
Problem with Maximum Likelihood: Bias

- ML estimates are biased
- Especially a problem for small number of samples, or high input dimensionality
- Suppose we sample 2,3,6 points from the same dataset, use ML to fit regression parameters



Problem with Maximum Likelihood: Overfitting

- ML estimates cannot be used to choose complexity of model
  - E.g. suppose we want to estimate the number of basis functions
  - Choose K=1?
  - Or K=15?



- ML will always choose K that best fits training data (in this case, K=15)

- Solution: use a Bayesian method--define a prior distribution over the parameters (results in regularization)

Bayesian vs. Frequentist

Frequentist: maximize data likelihood

$$p(D|model) = p(D|\theta)$$

Bayesian: treat θ as random variable, maximize posterior

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

$p(D|\theta)$ is the data likelihood, $p(\theta)$ is the prior over the model parameters

Bayesian Method

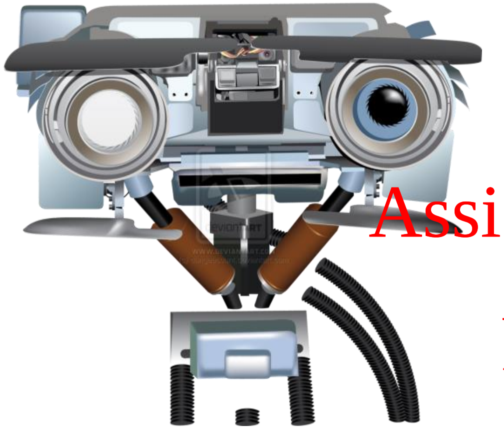
Treat θ as random variable, maximize posterior

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Likelihood $p(D|\theta)$ is the same as before, as in Maximum Likelihood

Prior $p(\theta)$ is a new distribution we model; specifies which parameters are more likely *a priori*, before seeing any data

$p(D)$ does not depend on θ, constant when choosing θ with the highest posterior probability



Prior over Model Parameters

Intuition

Will he score?

Score! Score! Miss Score!

Your estimate of $\theta = p(score)$?

Will he score?

Score! Score! Miss Score!

- Prior information: player= LeBron James
- Your estimate of $\theta = p(score)$?
- Prior $p(\theta)$ reflects prior knowledge, *e.g.*, $\theta \approx 1$

Prior Distribution

Prior distributions $p(\theta)$ are probability distributions of model parameters based on some a priori knowledge about the parameters.

Prior distributions are independent of the observed data.

Coin Toss Example

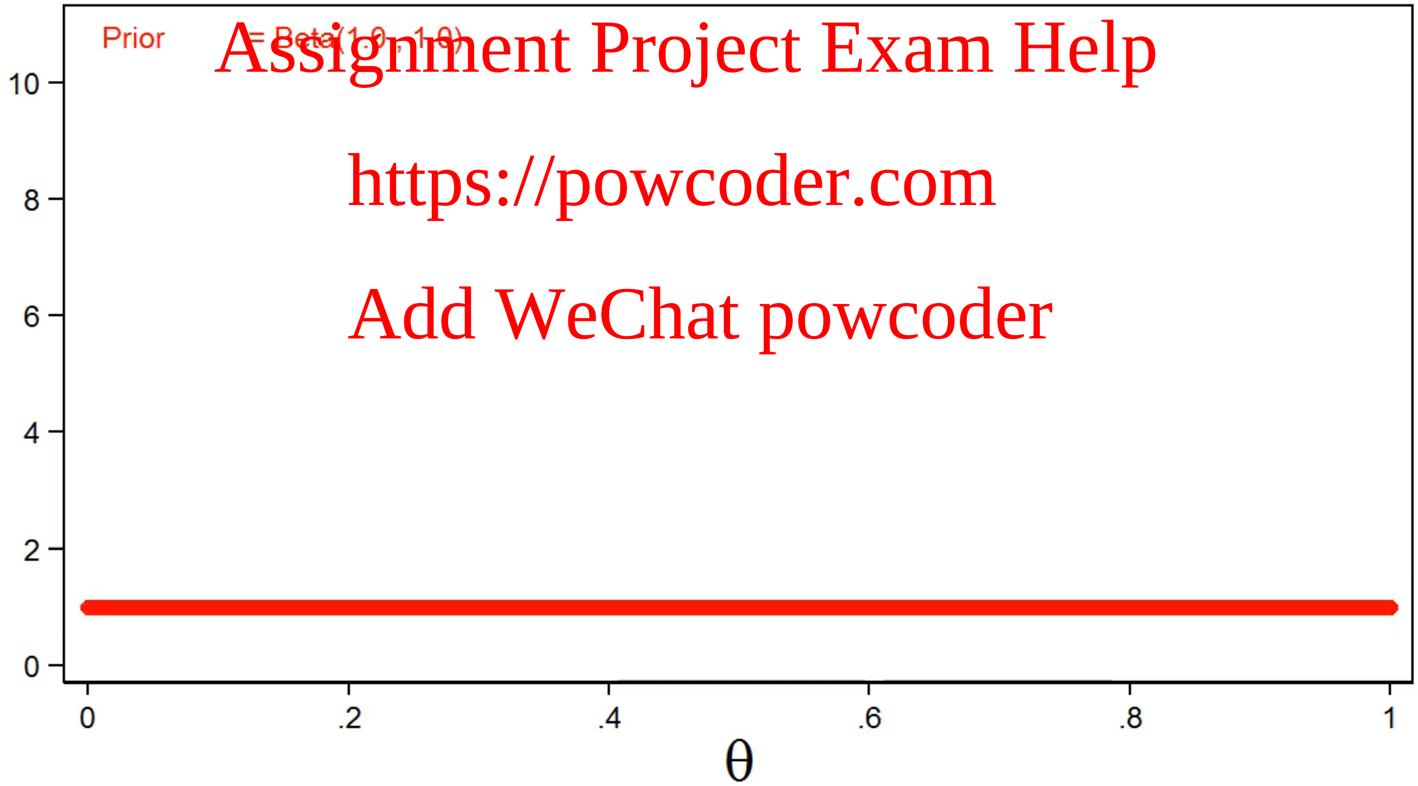
What is the probability of heads (θ)?

Beta Prior for θ

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}$$

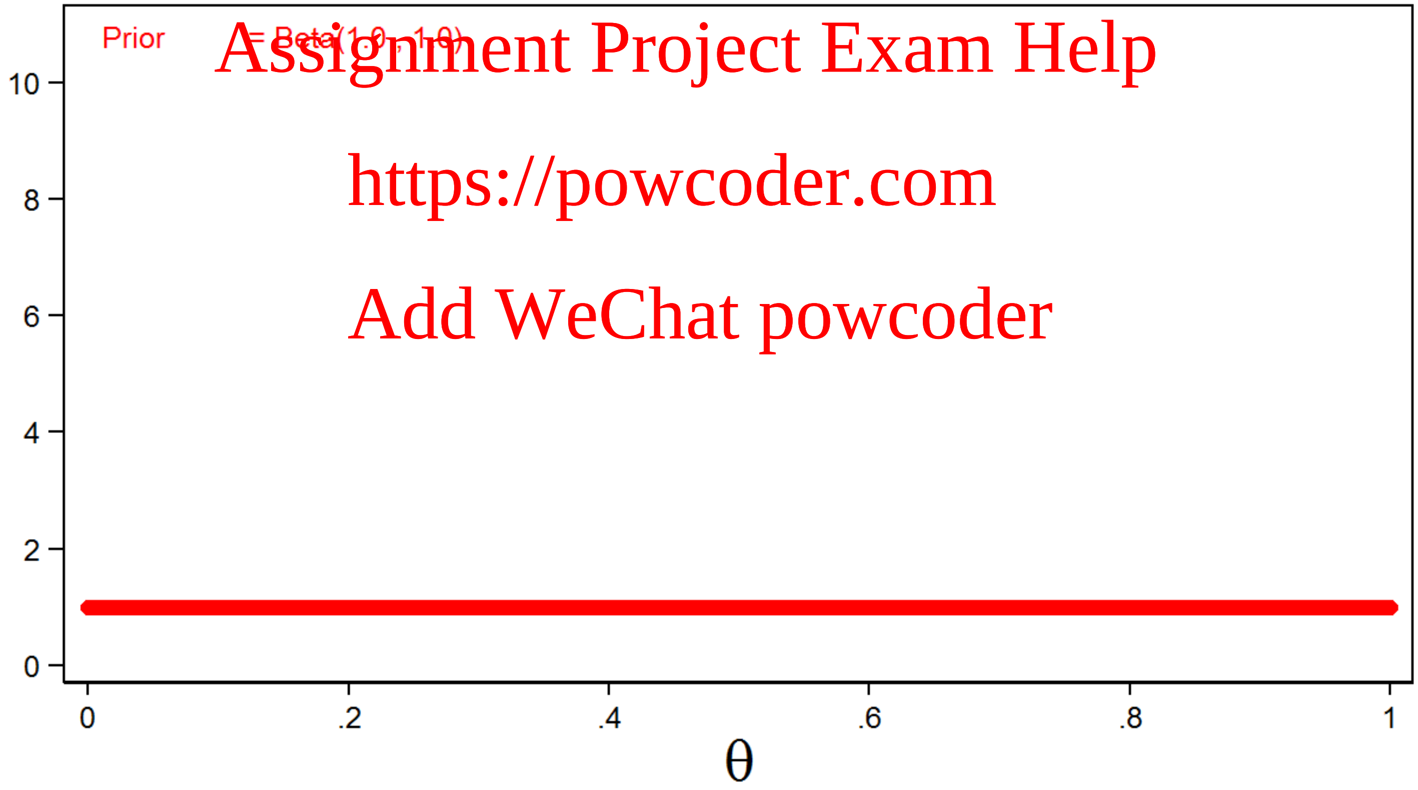
Beta Prior for θ

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}$$



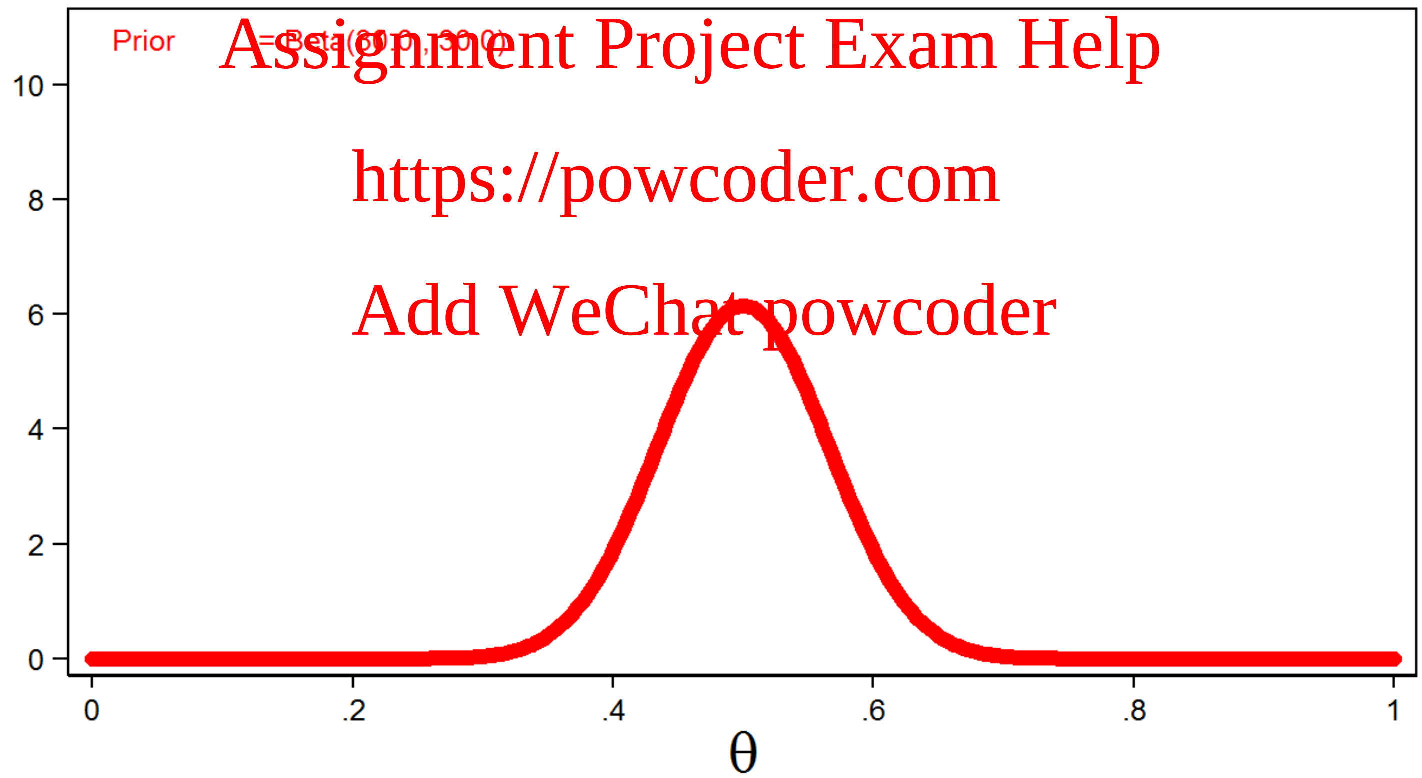
Uninformative Prior

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}$$



Informative Prior

$$P(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{(\alpha-1)} (1-\theta)^{(\beta-1)}$$



Coin Toss Experiment

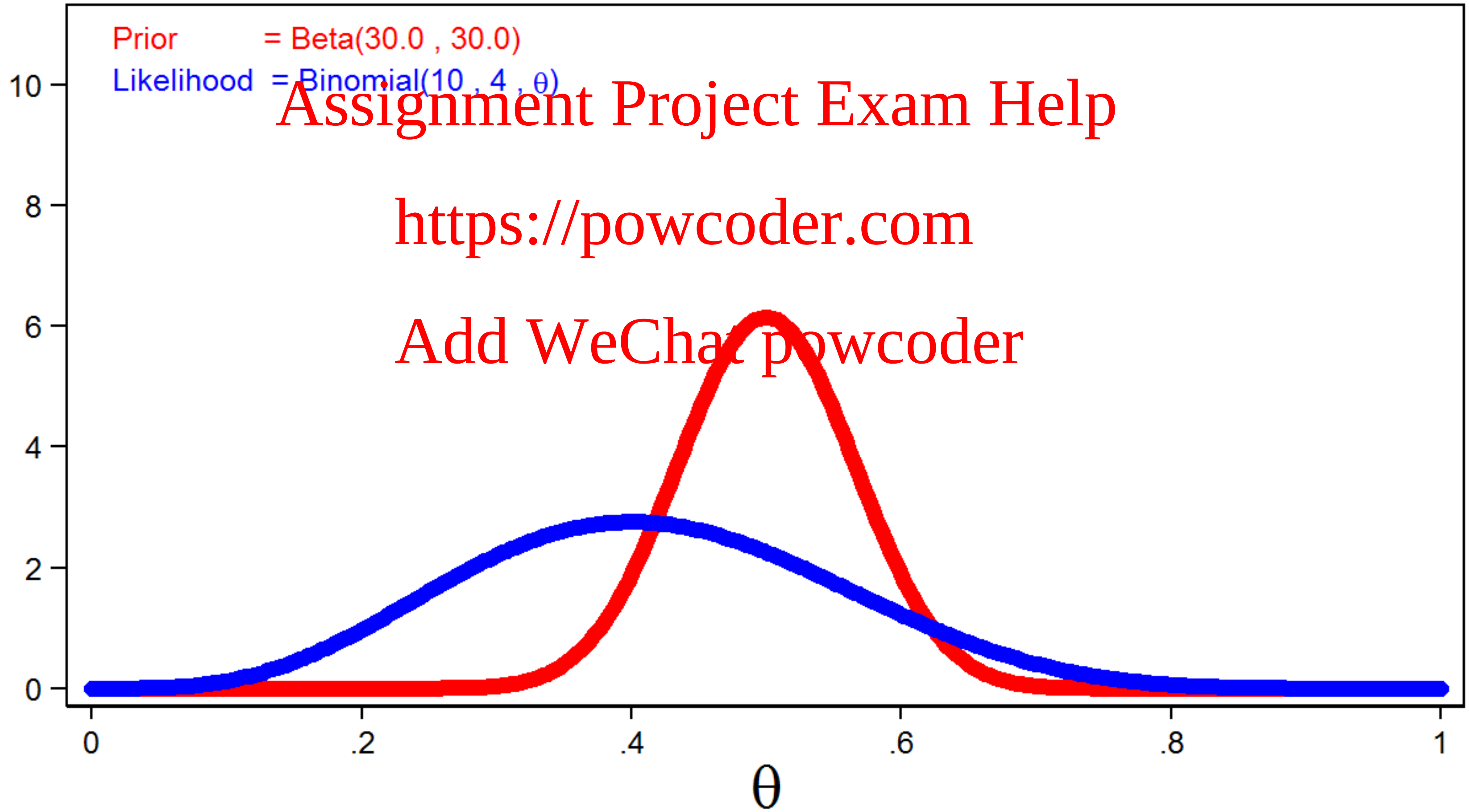
- $n = 10$ coin tosses
- $y = 4$ number of heads

Likelihood Function for the Data

$$P(y|\theta) = Binomial(n, \theta) = \binom{n}{y} \theta^{y} (1 - \theta)^{(n-y)}$$

Prior and Likelihood

$$P(y|\theta) = Binomial(n,\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{(n-y)}$$



Posterior Distribution

$$Posterior = Prior \times Likelihood$$

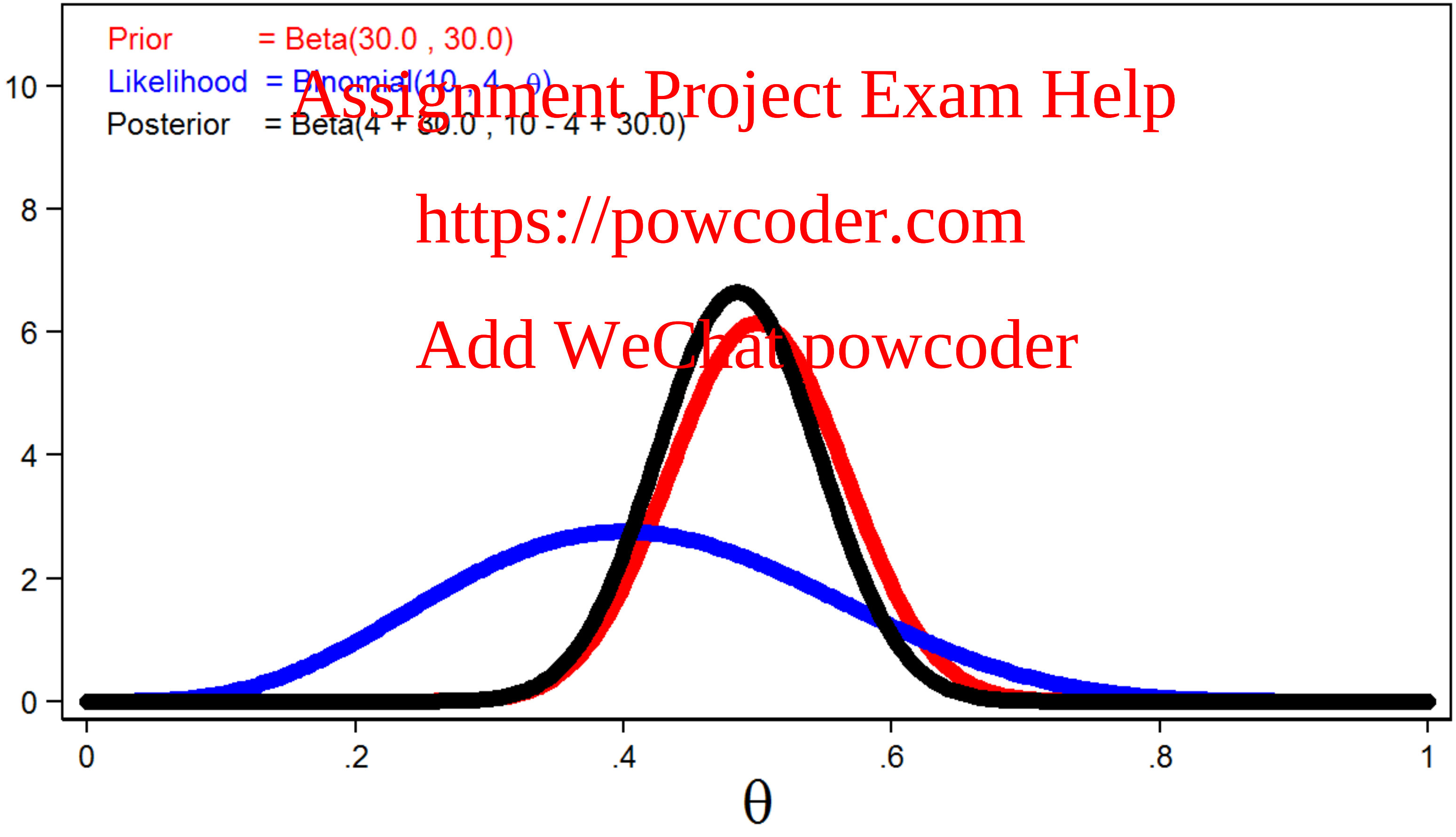
$$P(\theta|y) = P(\theta)P(y|\theta)$$

$$P(\theta|y) = Beta(\alpha, \beta) \times Binomial(n, \theta)$$

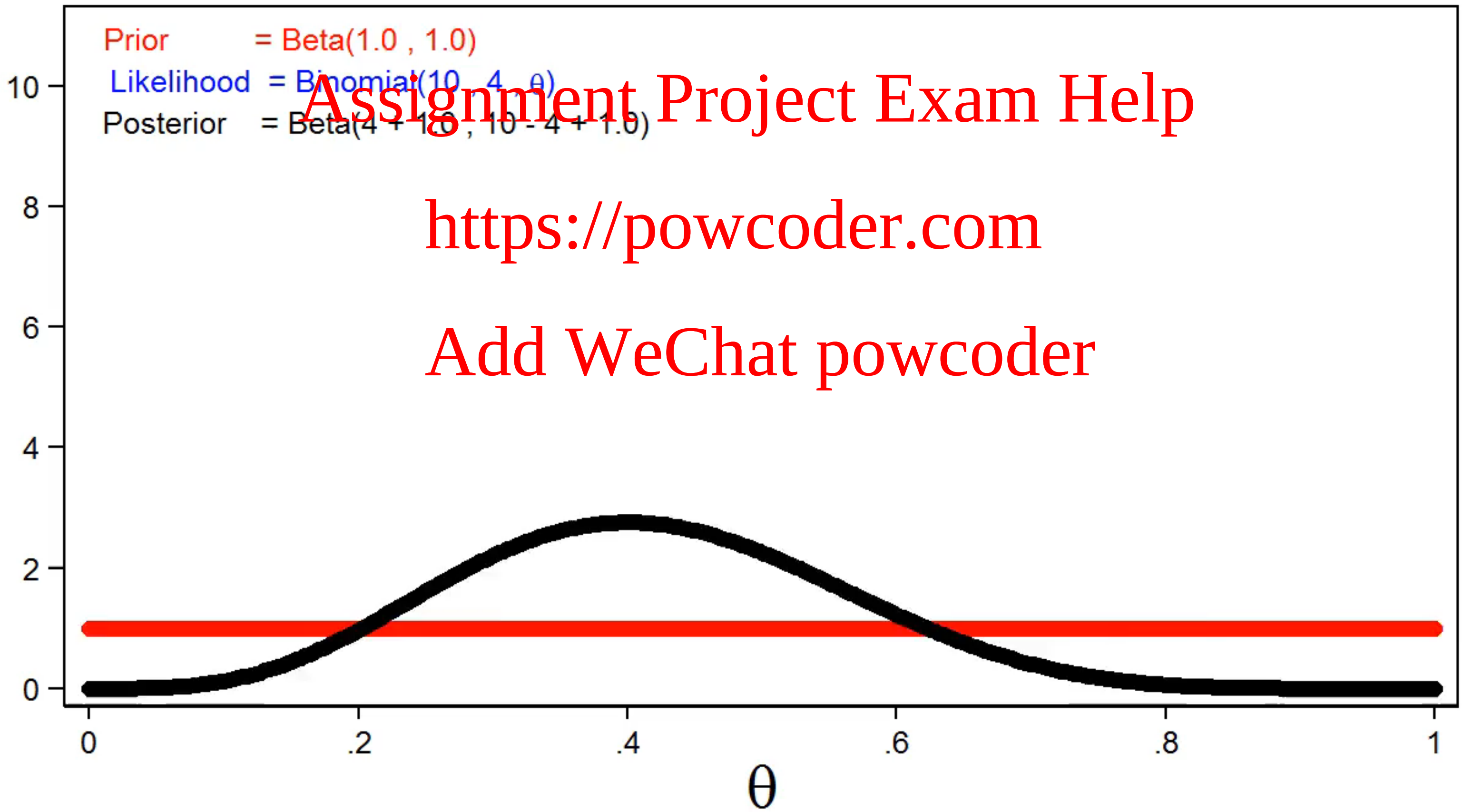
$$= Beta(y + \alpha, n - y + \beta)$$

This is why we chose the Beta distribution as our prior, posterior is also a Beta distribution: conjugate prior.

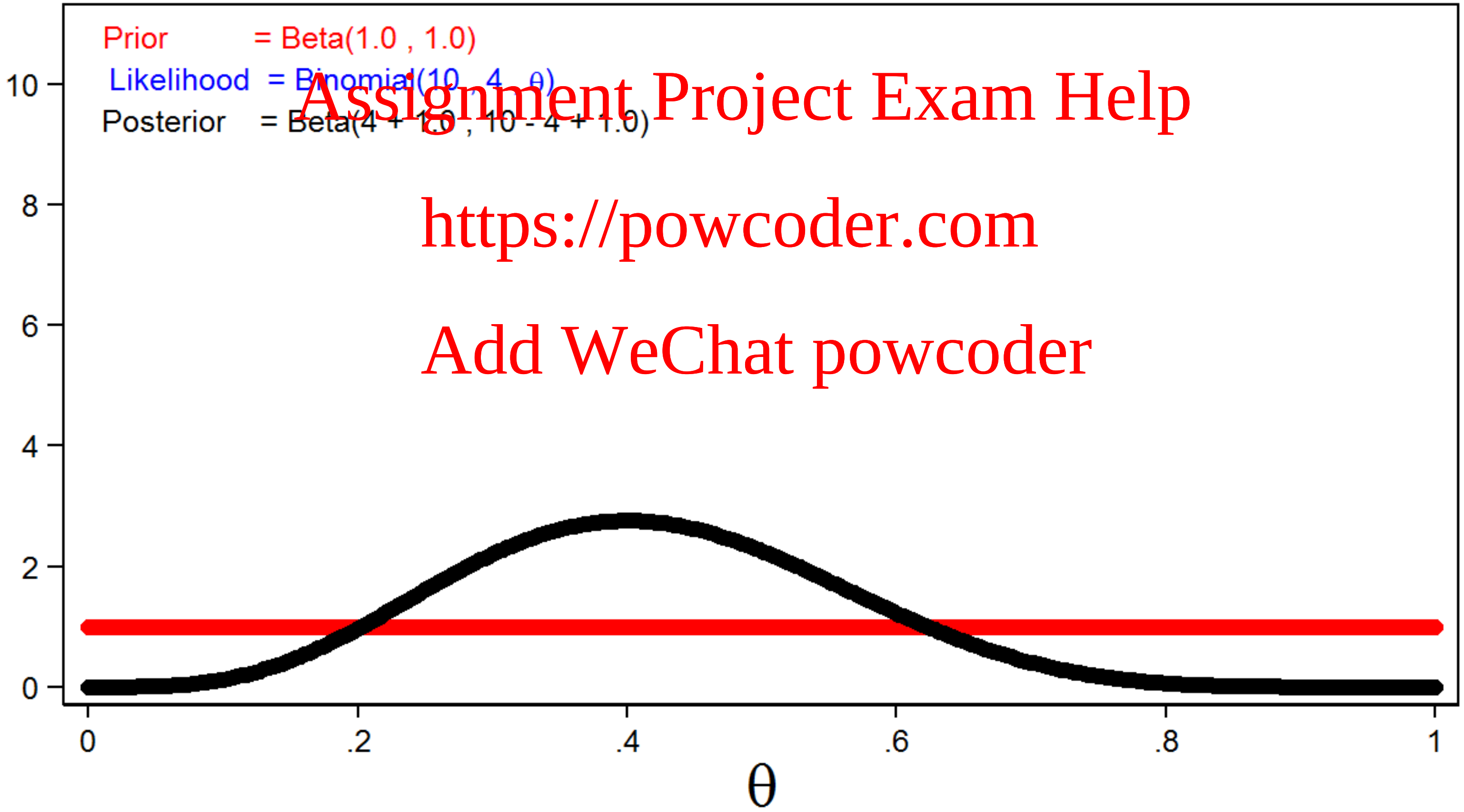
Posterior Distribution

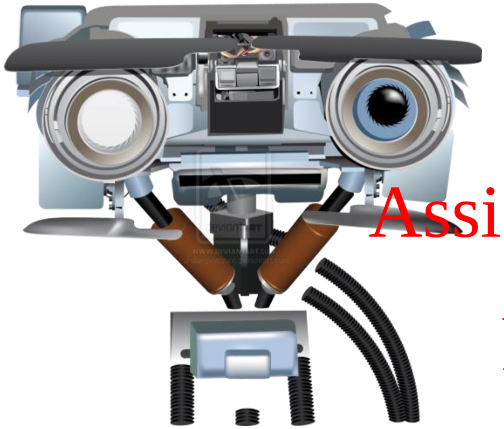


Effect of Informative Prior



Effect of Uninformative Prior





Bayesian Linear Regression

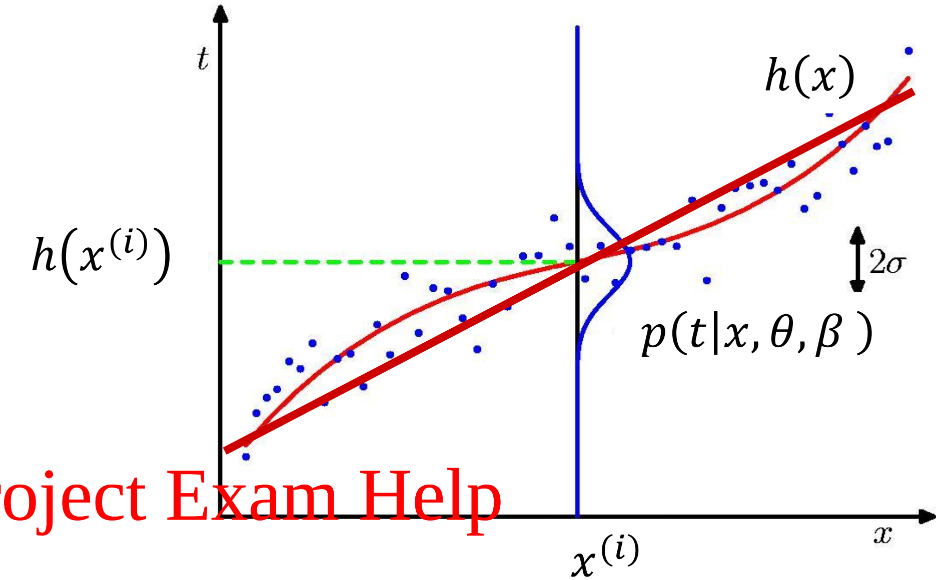
Bayesian Linear Regression

Let's now apply the Bayesian method to linear regression.

To do that, we must treat parameter θ as a random variable, design a prior over it.

First, review maximum likelihood for linear regression.

ML for Linear Regression



$t = y + \epsilon = h(x) + \epsilon$

Noise $\epsilon \sim N(\epsilon|0, \beta^{-1})$,

where $\beta = \frac{1}{\sigma^2}$, $h(x) = \theta^T x$

Probability of one data point

$$p(t|x, \theta, \beta\ ) = N(t|h(x), \beta^{-1})$$

$$p(\boldsymbol{t}|\boldsymbol{x}, \theta, \beta) = \prod_{i=1}^{m} N(t^{(i)}|h(x^{(i)}),\ \beta^{-1})$$

Likelihood function

Maximum likelihood solution

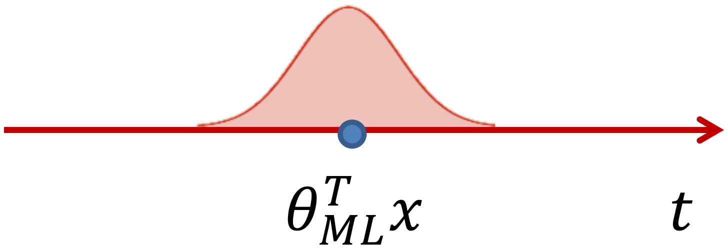
$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{t}|\boldsymbol{x}, \theta, \beta) \qquad \beta_{ML} = \underset{\beta}{\operatorname{argmax}}\, p(\boldsymbol{t}|\boldsymbol{x}, \theta, \beta)$$

What is β useful for?

- Recall: we assumed observations *t* are Gaussian given $h(x)$
- β allows us to write down distribution over *t*, given new *x*, called predictive distribution

$$p(t|x, \theta_{ML}, \beta_{ML})$$

$$= N(t|\theta_{ML}^T x, \beta_{ML}^{-1})$$

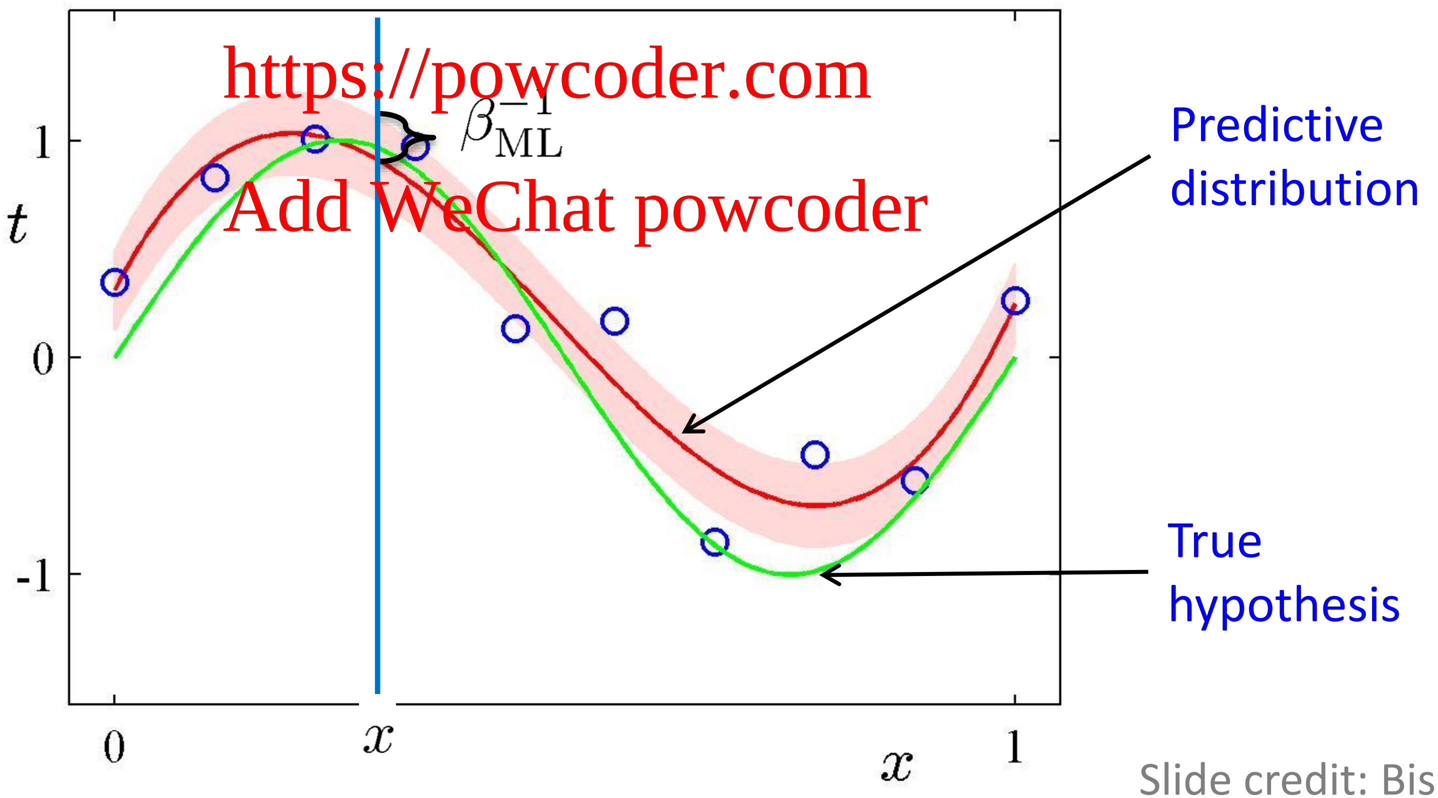


β_{ML}^{-1} is the variance of this distribution

Predictive Distribution

Given a new input point x, we can now compute a distribution over the output t:

$$p(t|x,\theta_{ML},\beta_{ML}) = N(t|\theta_{ML}^T x,\beta_{ML}^{-1})$$



Slide credit: Bishop

Define a distribution over parameters

- Define prior distribution over θ as

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$

- Combining this with the likelihood function and using results for marginal and conditional Gaussian distributions[1], gives the posterior

$$p(\boldsymbol{\theta}|\boldsymbol{t}) = N(\boldsymbol{\theta}|\boldsymbol{m}_N, \boldsymbol{S}_N)$$

- where

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \beta\boldsymbol{X}^T\boldsymbol{t})$$

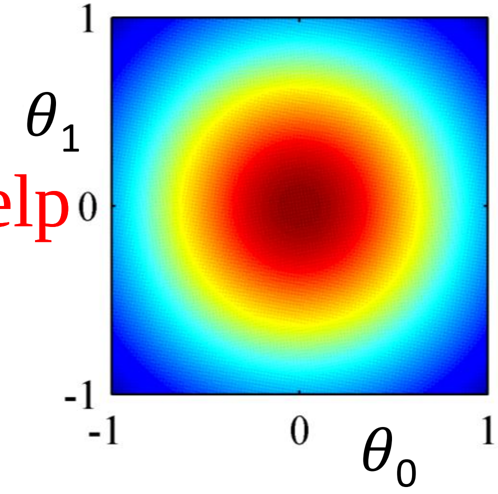
$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \beta\boldsymbol{X}^T\boldsymbol{X}$$

[1]see Bishop 2.3.3

A common choice for prior

- A common choice for the prior is

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\mathbf{0}, \alpha^{-1}\boldsymbol{I})$$



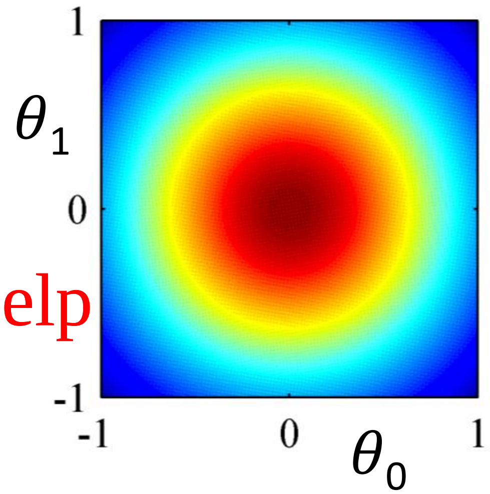
- for which

$$\boldsymbol{m}_N = \beta \boldsymbol{S}_N \boldsymbol{X}^T \boldsymbol{t}$$

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{X}^T \boldsymbol{X}$$

Slide credit: Bishop

Intuition: prefer θ to be simple



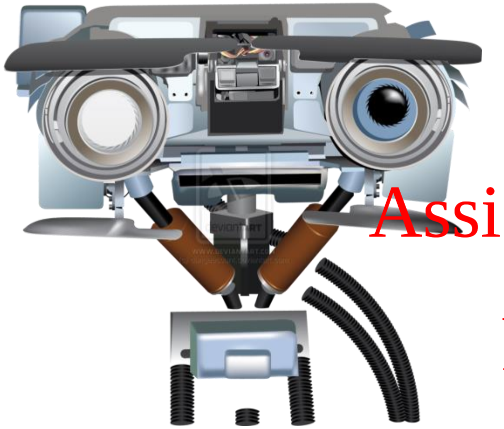
For a linear model for regression, $\theta^T x$

What do we mean by θ being simple?

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\mathbf{0}, \alpha^{-1}\boldsymbol{I})$$

Namely, put a prior on θ, which captures our belief that θ is around zero, i.e., resulting in a simple model for prediction.

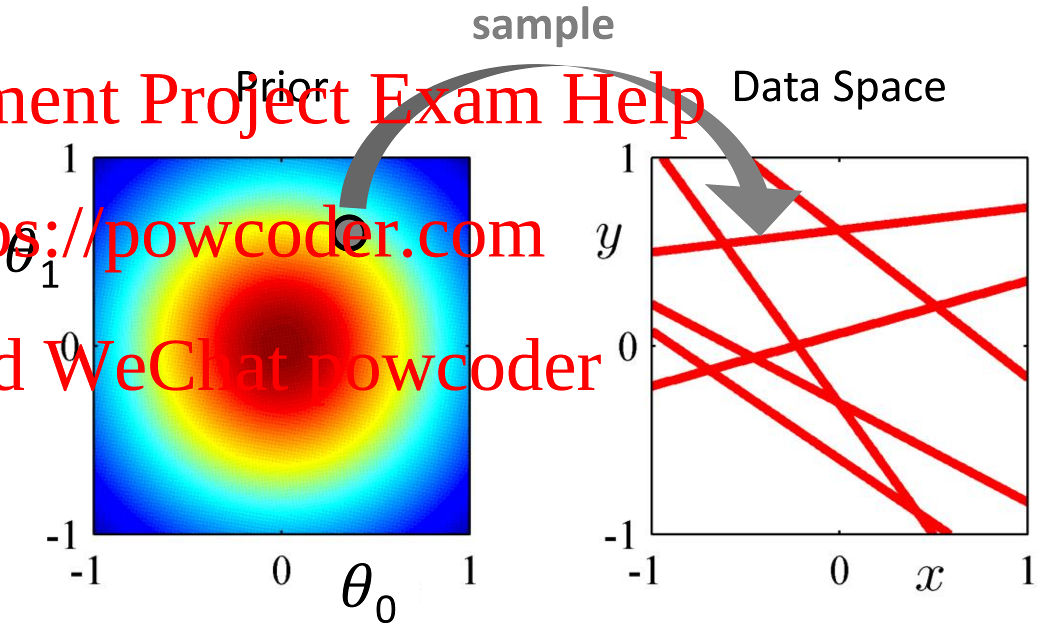
This Bayesian way of thinking is to regard θ as a random variable, and we will use the observed data D to update our prior belief on θ



Bayesian Linear Regression Example

Bayesian Linear Regression Example

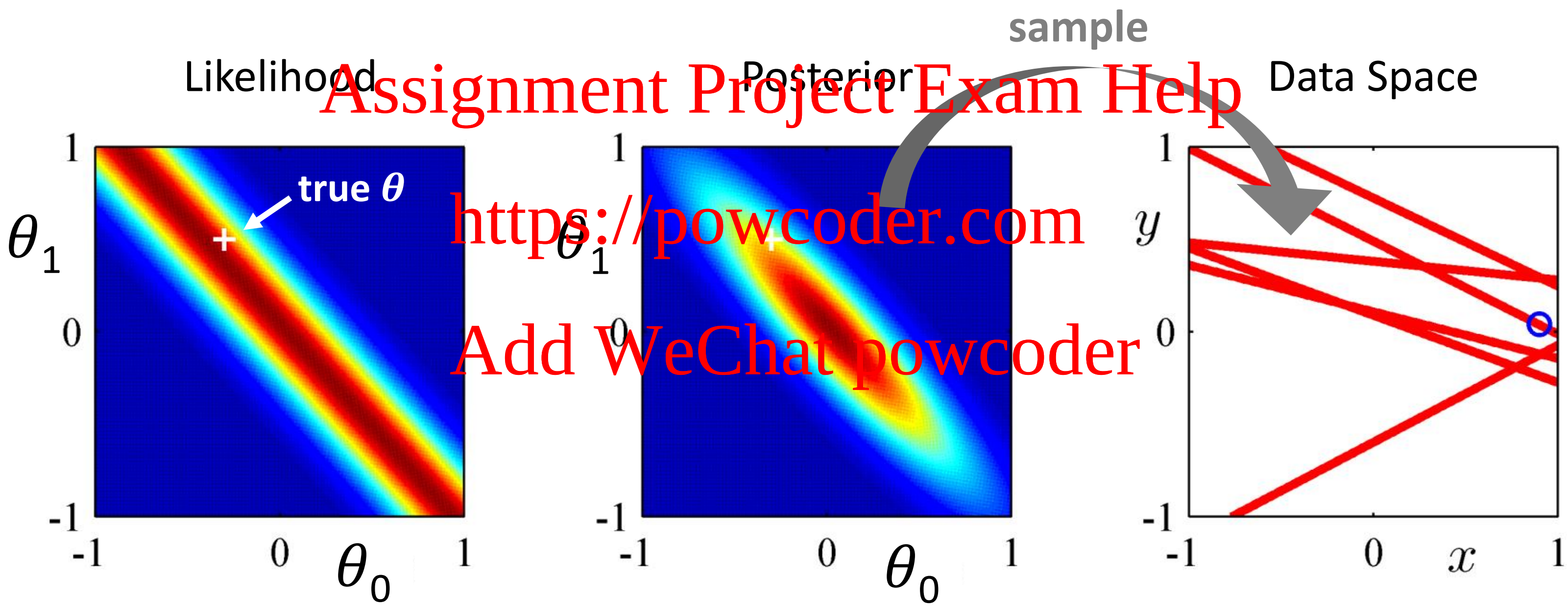
0 data points observed



Slide credit: Bishop

Bayesian Linear Regression Example

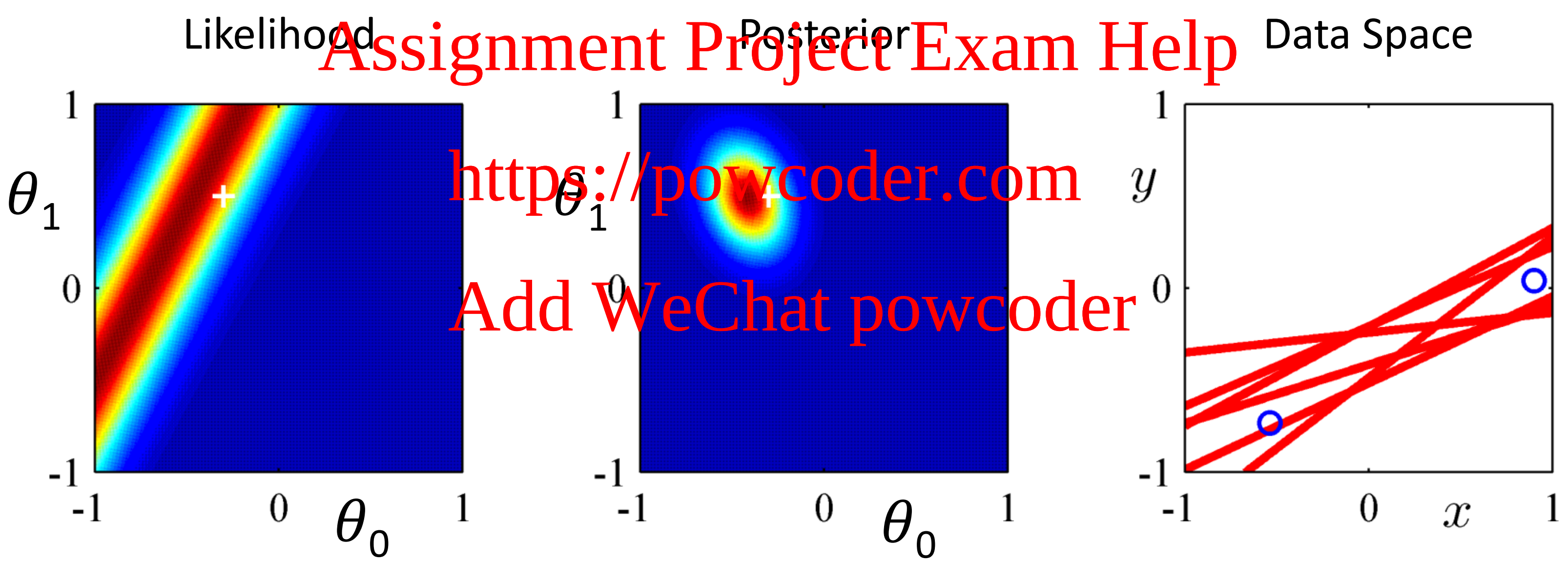
1 data point observed



Slide credit: Bishop

Bayesian Linear Regression Example

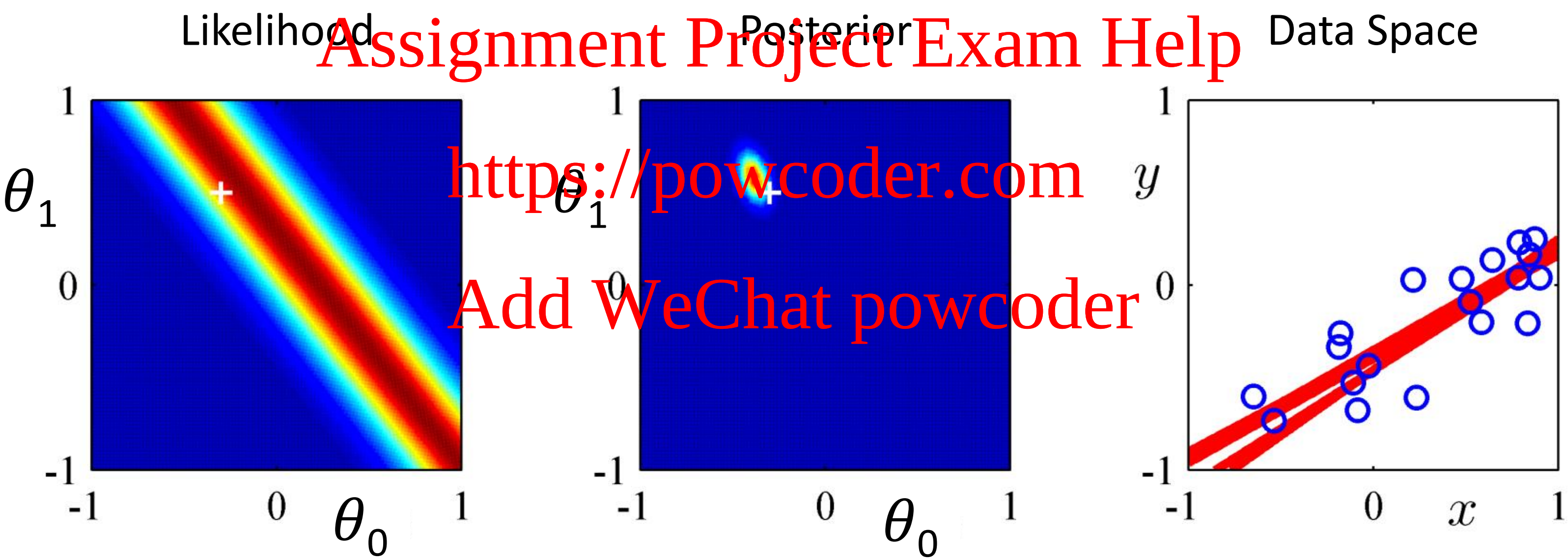
2 data points observed



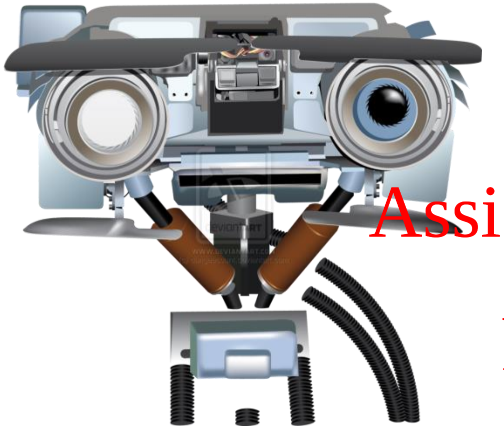
Slide credit: Bishop

Bayesian Linear Regression Example

20 data points observed



Slide credit: Bishop

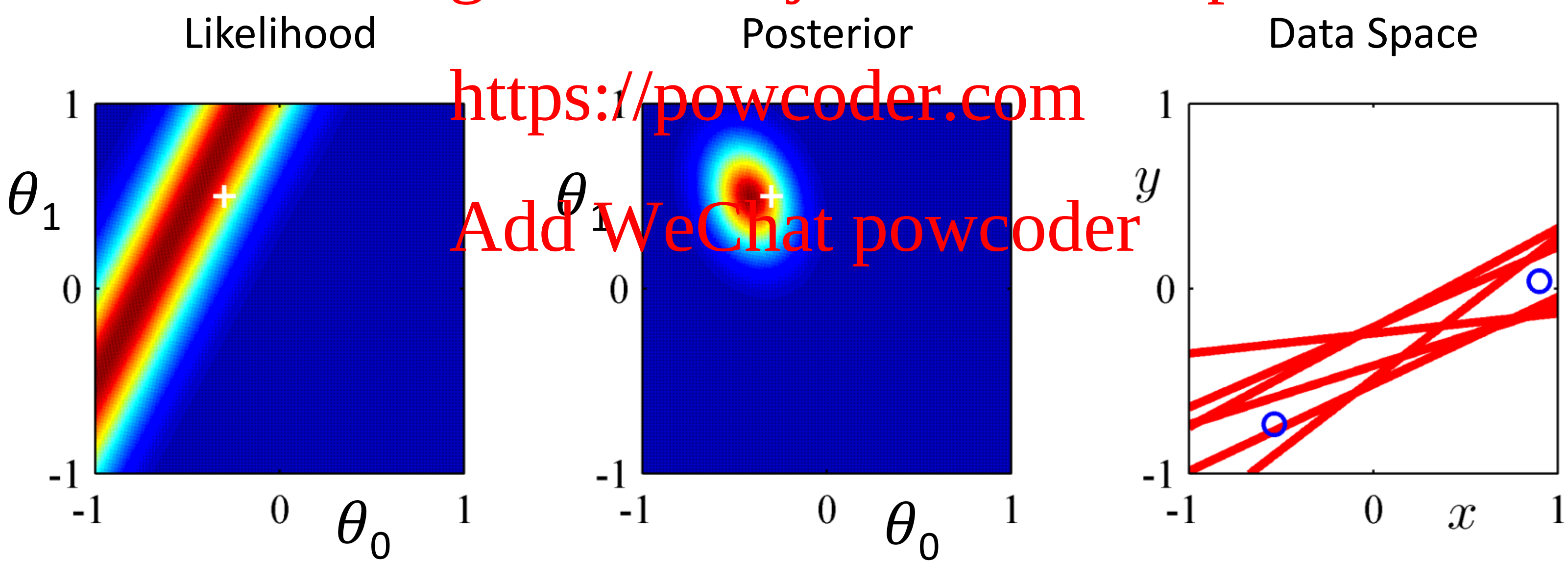


Bayesian Linear Regression

Prediction

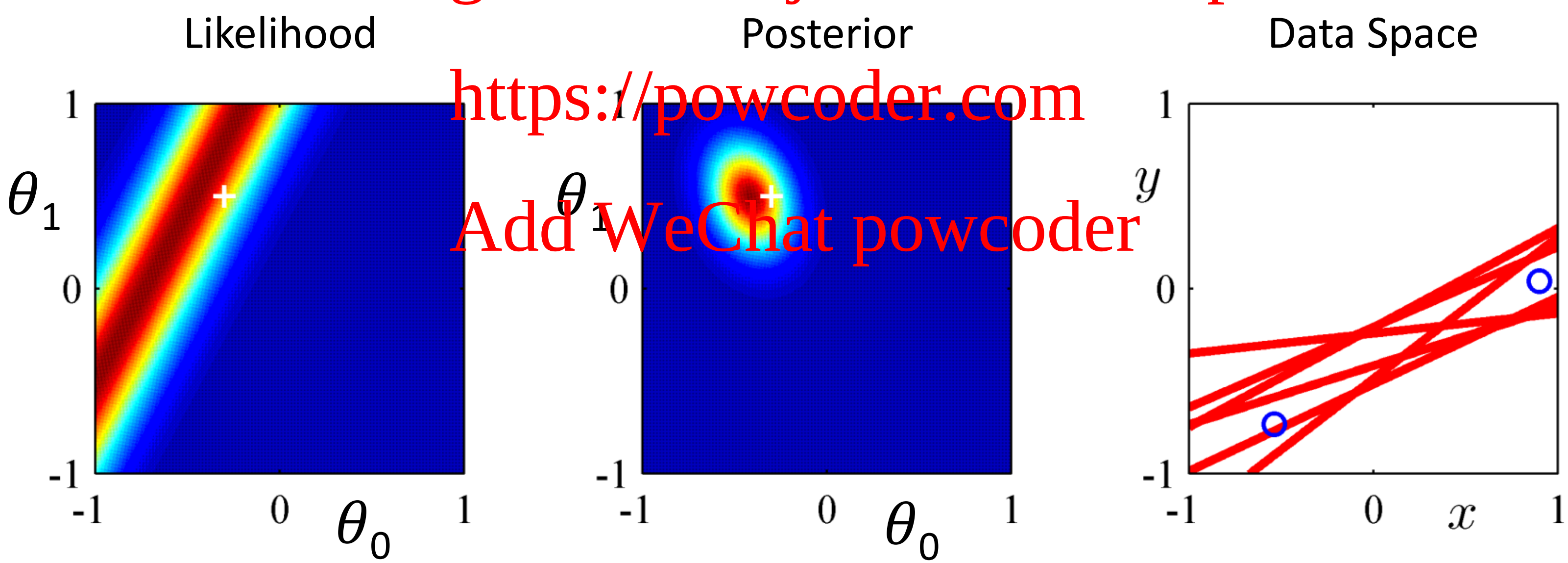
Prediction

- Now that we have a Bayesian model, how do we use it to make predictions for new data points?



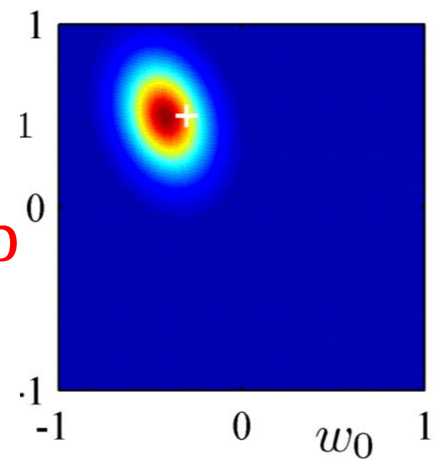
Prediction

- One way is to maximize the posterior to get an estimate of $\boldsymbol{\theta}_*$
- Then, plug $\boldsymbol{\theta}_*$ into the predictive distribution
- This is known as the maximum a posteriori estimate



Maximum A Posteriori (MAP)

Output the parameter that maximizes its posterior distribution given the data



$$\boldsymbol{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{t})$$

Recall: for our prior $p(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\mathbf{0}, \alpha^{-1}\boldsymbol{I})$,

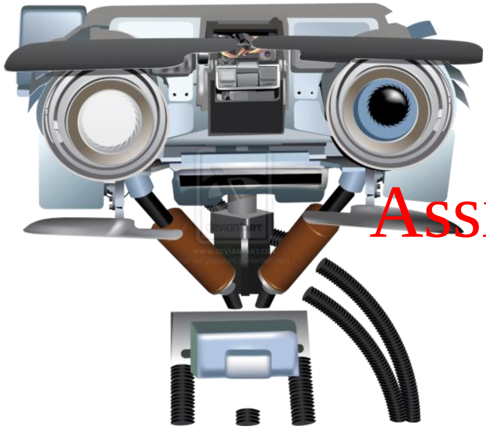
the posterior is $p(\boldsymbol{\theta}|\boldsymbol{t}) = N(\boldsymbol{\theta}|\boldsymbol{m}_N, \boldsymbol{S}_N)$,

where $\boldsymbol{m}_N = \beta \boldsymbol{S}_N \boldsymbol{X}^T \boldsymbol{t}, \quad \boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{X}^T \boldsymbol{X}.$

Therefore, $\boldsymbol{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{t}) = \left(\boldsymbol{X}^T\boldsymbol{X} + \frac{\boldsymbol{\alpha}}{\boldsymbol{\beta}}\boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}$

Same as solution to regularized regression with $||\boldsymbol{\theta}||^2$ term.

Note, this is the mode of the distribution



Bayesian Regression

Connection to Regularized Linear Regression

Maximizing posterior leads to regularized cost function

Joint likelihood of both training data and parameter

$$\log p(\mathcal{D}, \boldsymbol{\theta}) = \sum_n \log p(y_n | \boldsymbol{x}_n, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

$$= -\frac{\sum_n (\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_n - y_n)^2}{2\beta^{-2}} - \sum_d \frac{1}{2\alpha^{-2}} \theta_d^2 + \text{const}$$

where β^{-2} is the noise variance and α^{-2} is the prior covariance parameter

Maximum a posterior (MAP) estimate: we seek to maximize

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\theta} \log p(\boldsymbol{\theta} | \mathcal{D}) \propto \log p(\mathcal{D}, \boldsymbol{\theta})$$

that is, the most likely θ *conditioning* on observed training data $\mathcal{D}$.

Maximizing posterior leads to regularized cost function

Can re-write the optimization in the same form as the **regularized linear regression** cost:

$$L(\boldsymbol{\theta}) = \sum(\boldsymbol{\theta}^T \boldsymbol{x}_n - y_n)^2 + \lambda||\boldsymbol{\theta}||^2$$

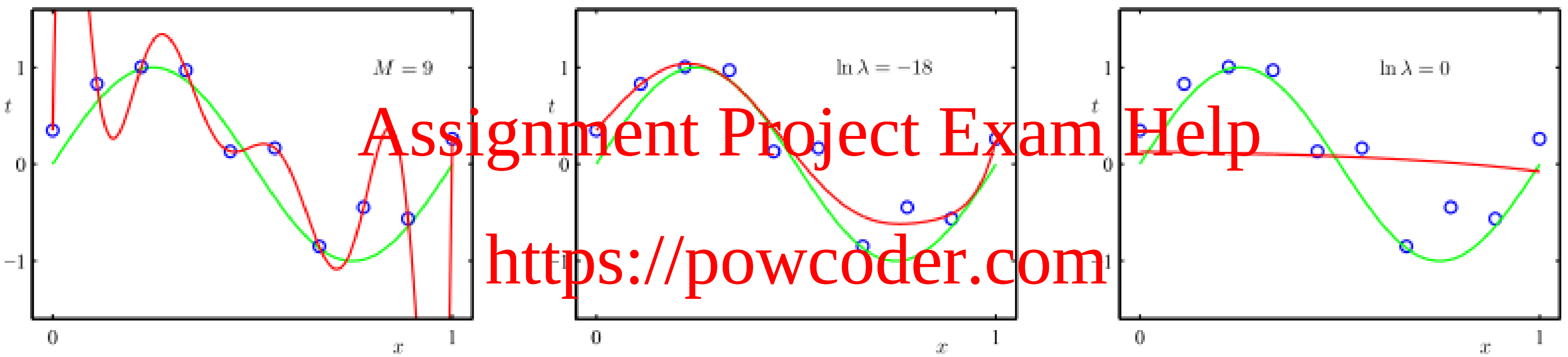
where $\lambda = \beta^{-2}/\alpha^{-2}$ corresponds to the regularization hyperparameter.

- Intuitively, as $\lambda \to +\infty$, then $\beta^{-2} \gg \alpha^{-2}$. That is, the variance of noise if far greater than what our prior model can allow for θ. In this case, our prior would be more accurate than what data can tell us, so we are getting a simple model, where $\theta_{MAP} \to 0$.

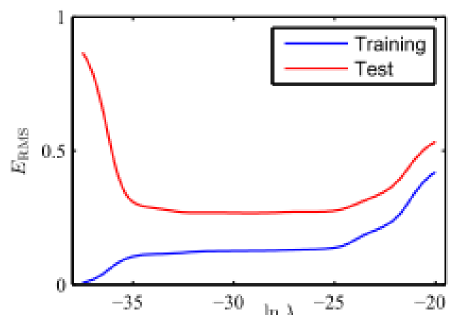
- If $\lambda \to 0$, then $\beta^{-2} \ll \alpha^{-2}$, and we trust our data more, so the MAP solution approaches the maximum likelihood solution, i.e. $\theta_{MAP} \to \theta_{ML}$.

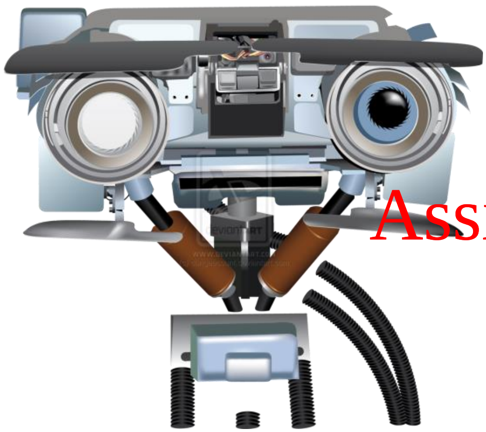
Effect of lambda

Overfitting is reduced from complex model to simpler one with the help of increasing regularizers



λ vs. residual error shows the difference of the model performance on training and testing dataset



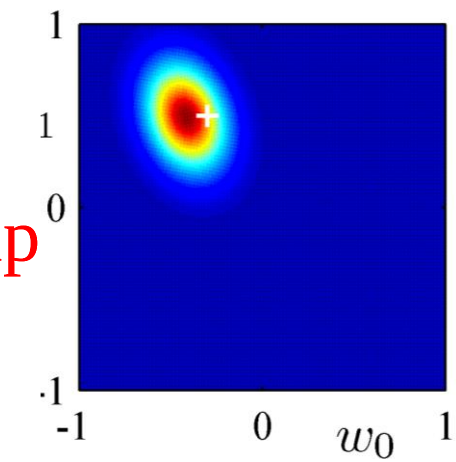


Bayesian Predictive Distribution

Maximum A Posteriori (MAP)

- Output the parameter that maximizes its posterior distribution given the data

$$\boldsymbol{\theta}_{\boldsymbol{MAP}} = \underset{\theta}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\boldsymbol{t})$$



- Note, this is the mode of the distribution

- However, sometimes we may want to hedge our bets and average (integrate) over all possible parameters, e.g. if the posterior is multi-modal

Bayesian Predictive Distribution

- Predict t for new values of x by integrating over $\boldsymbol{\theta}$:

$$p(t|x, \boldsymbol{t}, \alpha, \beta) = \int p(t|\boldsymbol{\theta}, \beta) p(\boldsymbol{\theta}|\boldsymbol{t}, \alpha, \beta) d\theta$$

$$= N(t|m_N^T x, \sigma_N^2(x))$$

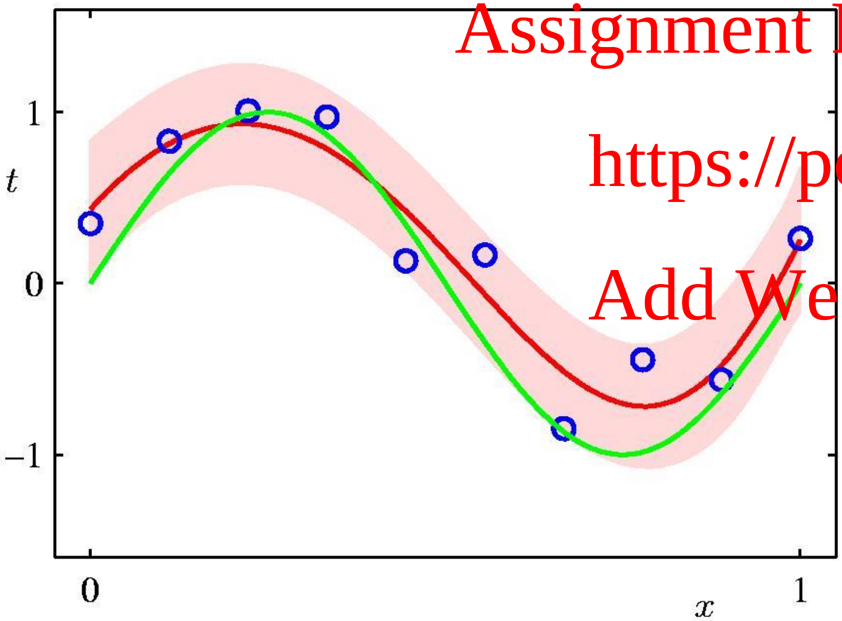
- where

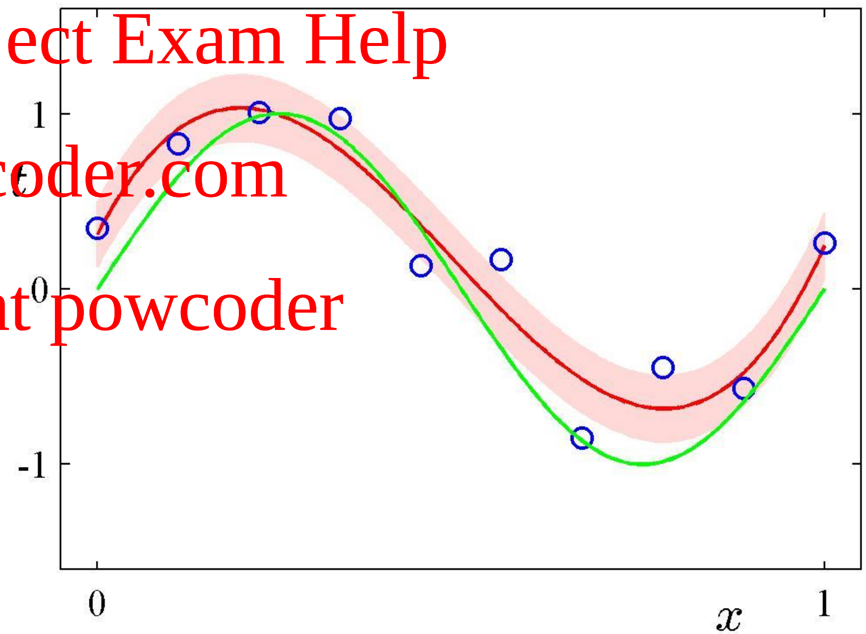
$$\sigma_N^2(x) = \frac{1}{\beta} + x^T S_N x$$

What does it look like?

Compare to Maximum Likelihood:

$$p(t|x,\boldsymbol{x},\boldsymbol{t}) = N(t|m_N^T x, \sigma_N^2) \qquad p(t|x,\theta_{ML},\beta_{ML}) = N(t|\theta_{ML}^T x, \beta_{ML}^{-1})$$





Slide credit: Bishop

Next Class

Support Vector Machines I

maximum margin methods; support vector machines; primal vs dual SVM formulation; Hinge loss vs. cross-entropy loss

Reading: Bishop Ch 7.1.1-7.1.2