Intro to Bayesian Linear Regression

"Bayesian" means taking into account prior information.

$$p(x) = \sum_{Y} p(X|Y)\, p(Y)$$

posterior ↑ likelihood ↑ prior ↑

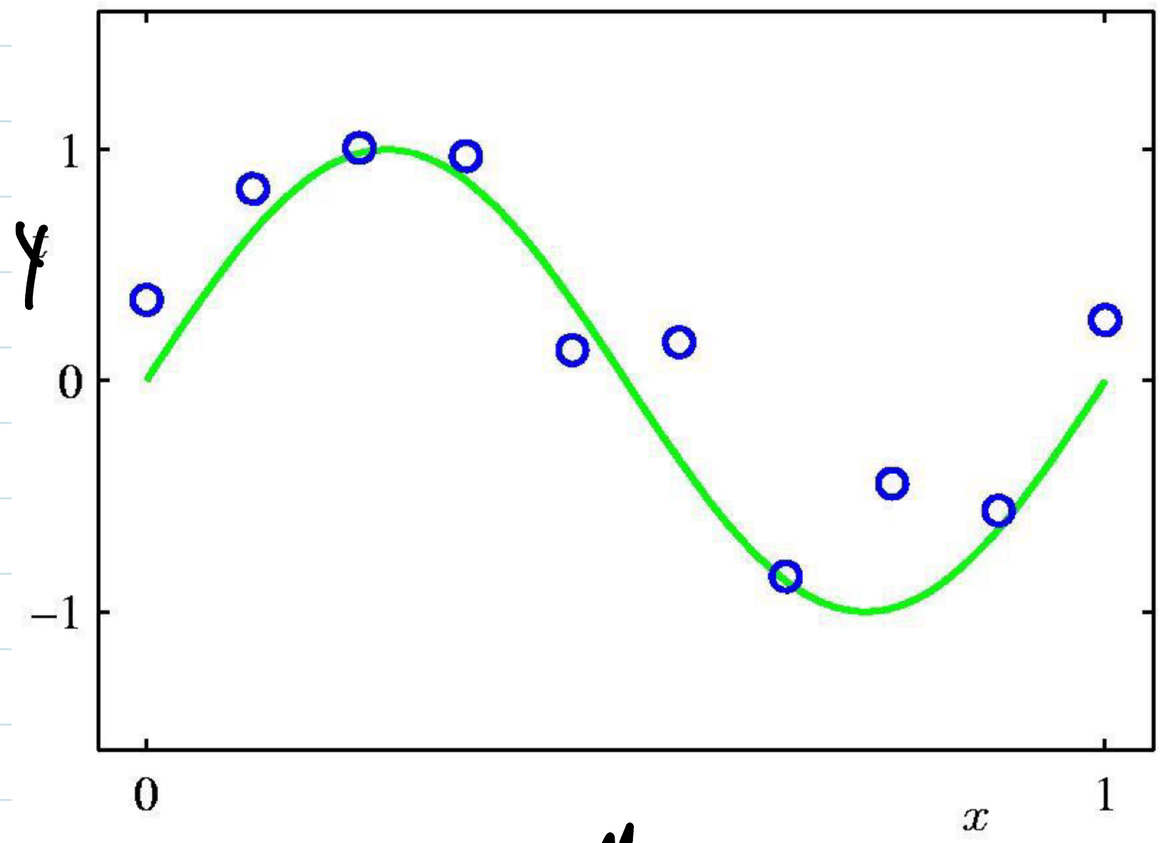
We are going to do regression where we combine data we gather with prior information to compute a predictive distribution.

To get there

- linear basis functions
- Gaussian distributions
- maximum likelihood estimation
- Bayesian linear regression
- pendulum example

Linear Basis Function Models

Example: Polynomial Curve Fitting



$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

$$W = [\, w_0 \;\; w_1 \;\; w_2 \;\; \ldots \;\; w_M \,]$$

Generally $$y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x)$$

x vector of inputs

y output

M # of basis functions

$w \in \mathbb{R}^M$ weights for basis functions (parameters)

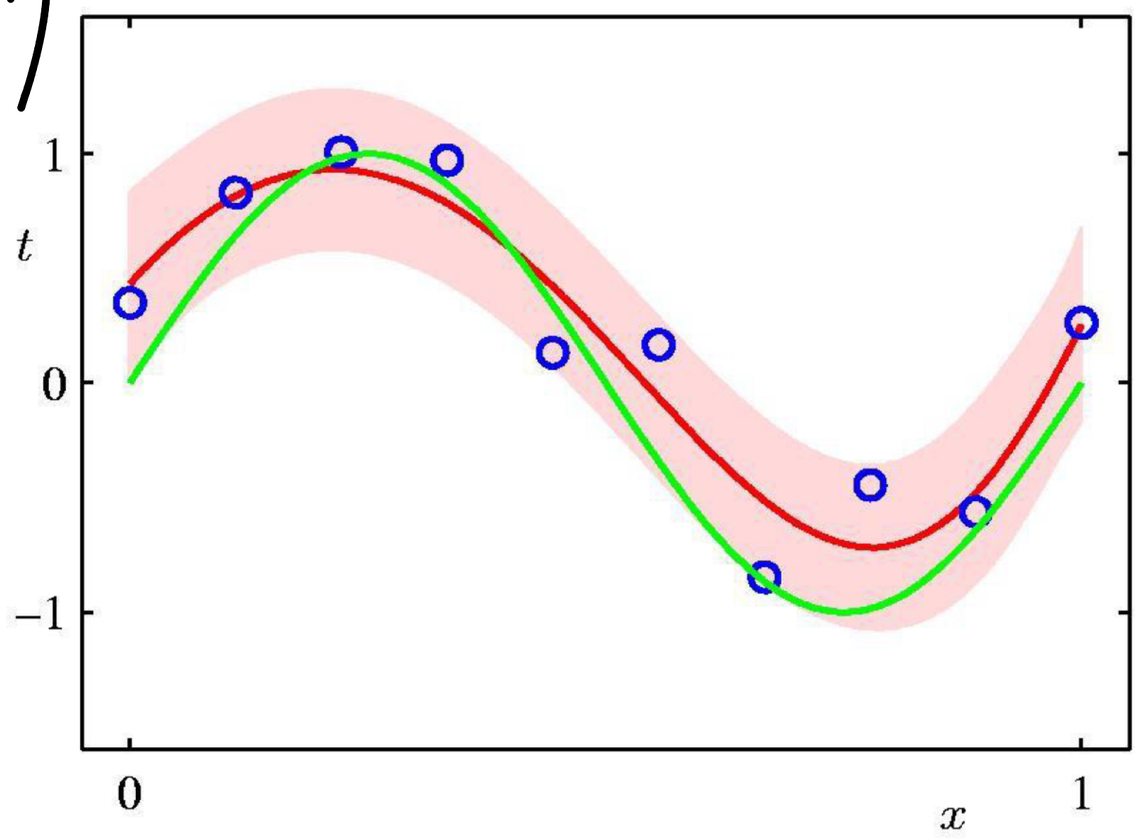
$\phi_j(x)$ is the jth basis function

We want to find

- $p(w)$ predictive distribution of weights
- $p(y|x, w)$ predictive distribution of outputs

Predictive Distribution

$$p(t \mid x, w, \beta) = \mathcal{N}(t \mid y(x, w), \beta^{-1})$$

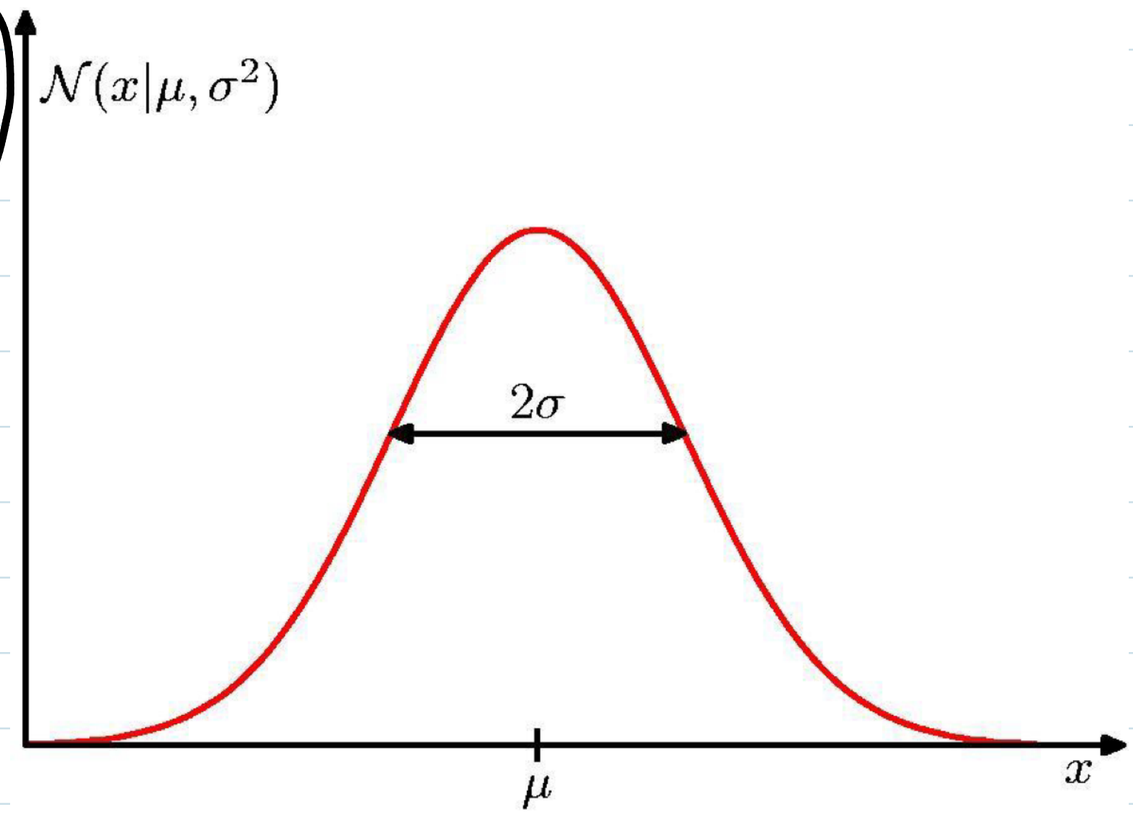


What do these predictive distributions look like?

We will assume they are Gaussians

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

probability that a random variable takes the value x given its mean is μ and variance σ^2



Matlab function

randn( )

Notes:

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\,dx = 1$$

$$E[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)\, x\, dx = \mu$$

↑ expected value of x

$$E[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2)\, x^2\, dx = \mu^2 + \sigma^2$$

$$Var[x] = E[x^2] - E[x]^2 = \sigma^2$$

$$E[f] = \sum_x p(x) f(x) = \int p(x) f(x)\, dx$$

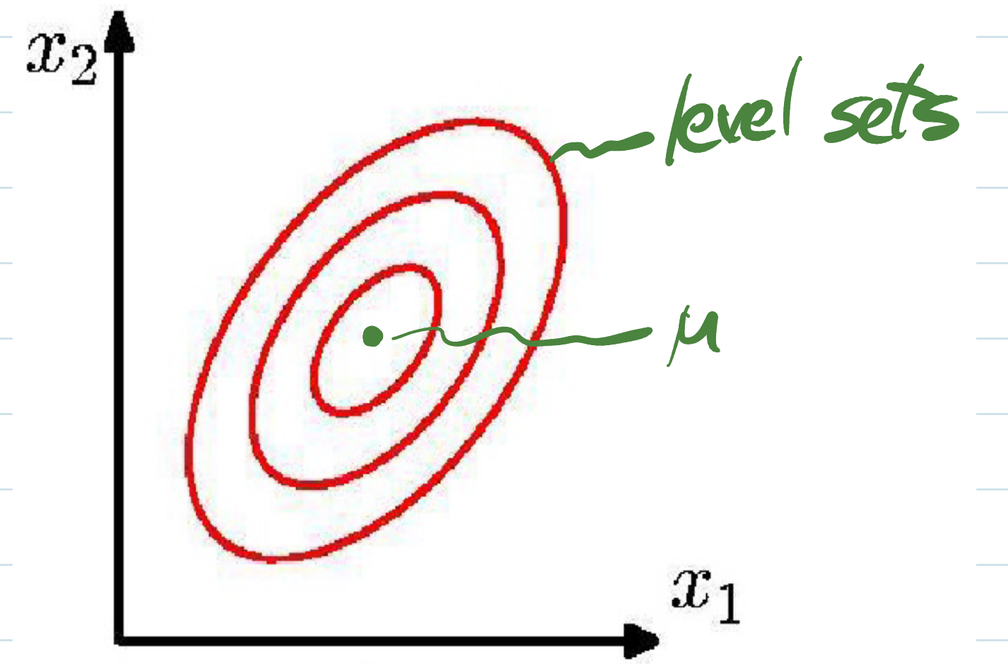
$$\mathrm{Var}[f] = E\left[(f(x) - E[f(x)])^2\right] = E[f(x)^2] - E[f(x)]^2$$

$$\mathrm{Cov}[x, y] = E_{x,y}\left[\{x - E[x]\}\{y - E[y]\}\right]$$
$$= E_{x,y}[xy] - E[x]E[y]$$

Multivariate Gaussian

$$N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \cdots$$

$$\exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

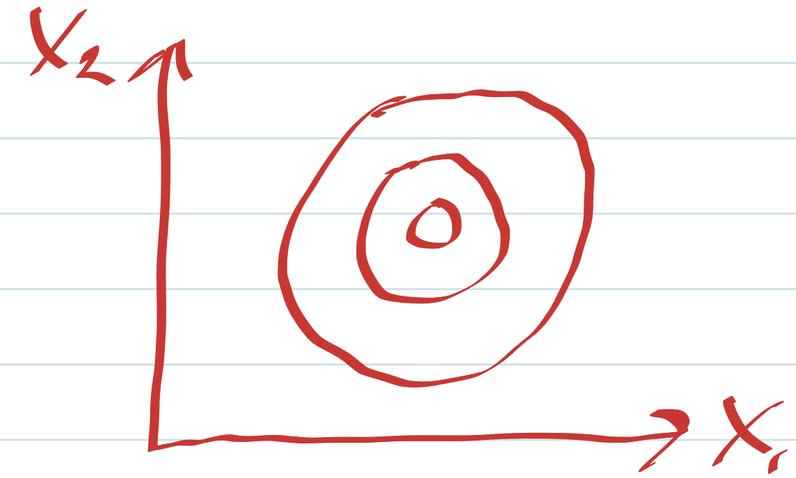


$\mu \in \mathbb{R}^d$ mean vector
d dimension of x

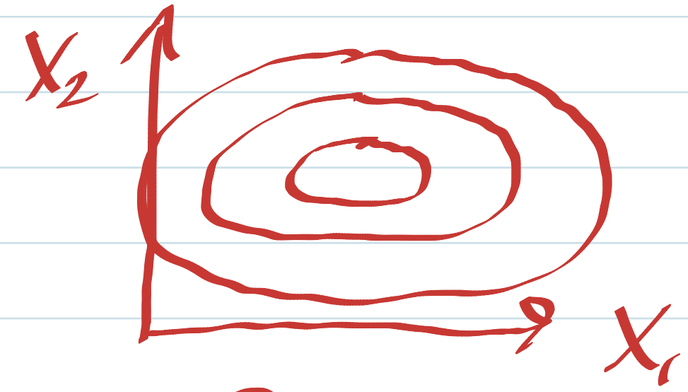
$x \in \mathbb{R}^d$ random variable
$\Sigma \in \mathbb{R}^{d \times d}$ covariance matrix

Σ, the covariance matrix determines the orientation of the level sets

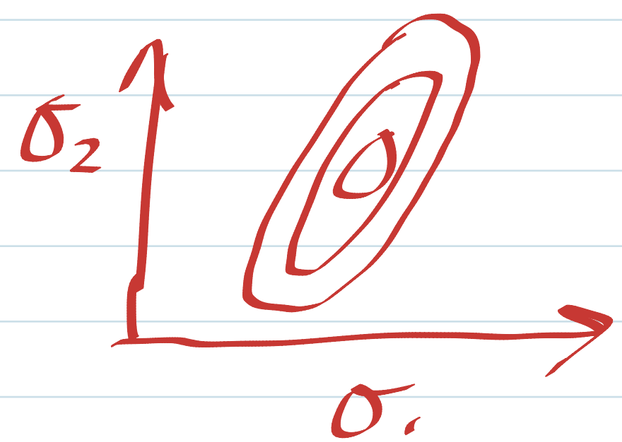
e.g $\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$



$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ $\sigma_1^2 > \sigma_2^2$



$\Sigma = \begin{bmatrix} \sigma_1^2 & \text{Cov}\,\sigma_1\sigma_2 \\ \text{Cov}\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$



Maximum Likelihood and Least Squares

assume we make observations from a deterministic function with added Gaussian noise

$$t = y(x, w) + \epsilon \qquad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

β is the inverse of variance

or a.k.a the confidence

recall

$$y(x, w) = w^T \phi(x)$$

$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1})$$

Given observed inputs $X \in \mathbb{R}^{d \times N}$

$x \in \mathbb{R}^d$ input vector

N # of observations

and given observed targets

$$t = [t_1, t_2, \ldots, t_n]^T$$

The likelihood function is

$$p(t \mid X, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})$$

The probability of observing the set of targets (noisy observations of y) given inputs X and parameters w and β.

Goal: Maximize the likelihood over w and β

maximum likelihood estimation (MLE)

Take the log of the likelihood

$$\ln p(t \mid w, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n \mid w^T \phi(x_n), \beta^{-1})$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{1}{2} \beta \sum_{n=1}^{N} \left[t_n - w^T \phi(x_n) \right]^2$$

sum of squared errors

MLE is the same as minimizing the ss error w.r.t w, β

To maximize the likelihood take the gradient

$$\nabla_w \ln p(t \mid w, \beta) = \beta \sum_{n=1}^{N} \left[t_n - w^T \phi(x_n) \right] \phi(x_n)^T = 0$$

Solve for W

$$W_{ML} = \underbrace{\left(\Phi^T \Phi\right)^{-1} \Phi^T}\, t$$

Moore-Penrose pseudo-inverse of Φ

Where

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \dots & \phi_{M-1}(x_2) \\ & \vdots & & \\ \phi_0(x_N) & \phi_1(x_N) & \dots & \phi_{M-1}(x_N) \end{bmatrix}$$

$\Phi \in \mathbb{R}^{N \times M}$

in Matlab

pinv(Φ)

Multiple Outputs

$$p(\bar{t}\,|\,x, W, \beta) = \mathcal{N}(t\,|\,y(W,x), \beta^{-1}I)$$

$$= \mathcal{N}(t\,|\,W^T\phi(x), \beta^{-1}I)$$

q is the dimension of output vector $\bar{t}$

M is the number of basis function

$W \in \mathbb{R}^{M \times q}$ matrix of parameters

For outputs $T = [\bar{t}_1, \bar{t}_2, \ldots \bar{t}_N]^T$ the log likelihood is

$$\ln p(T\,|\,X, W, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(\bar{t}_n\,|\,W^T\phi(x_n), \beta^{-1}I)$$

$$= \frac{Nq}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\|t_n - W^T\phi(x_n)\|^2$$

maximize w.r.t. W

$$W_{ML} = (\Phi^T \Phi)^{-1} \Phi^T T$$

Bayesian Linear Regression

Define a prior over W

$$p(w) = \mathcal{N}(w | m_0, S_0)$$

m_0 prior mean of parameters

S_0 prior variance

After a bunch of Gaussian math the posterior distribution for w is

$$p(w|t) = \mathcal{N}(w|m_N, S_N)$$

$$m_N = S_N\left(S_0^{-1} m_0 + \beta \Phi^T t\right)$$

posterior mean

$S_0^{-1} m_0$: confidence in prior; m_0: prior mean

$\beta \Phi^T t$: like MLE estimate; β: inverse of observation noise

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$$

posterior confidence

S_0^{-1}: prior confidence

$\beta \Phi^T \Phi$: confidence gained by observing data

A common choice for a prior is

$$p(w) = \mathcal{N}(w \mid 0, \alpha^{-1} I)$$

Zero mean with confidence α

then posterior becomes

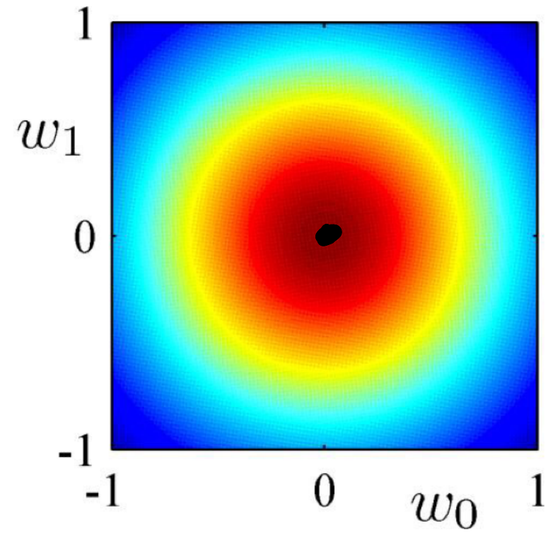
$$m_N = \beta S_N \Phi^T t$$

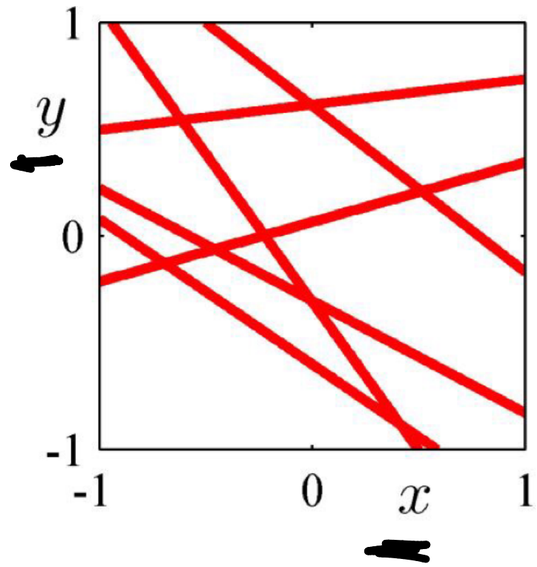
$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

if α is small (low confidence in prior)

$$m_N = \beta \beta^{-1} (\Phi^T \Phi)^{-1} \Phi^T t \Rightarrow \text{MLE}$$

Examples of Bayesian Regression





prior distribution of parameters

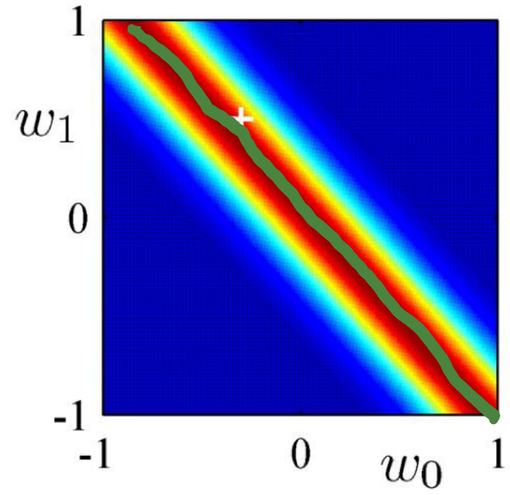
function $y = w_0 + w_1 x$

$$p(w) = \mathcal{N}(w \mid 0, S_0^{-1})$$

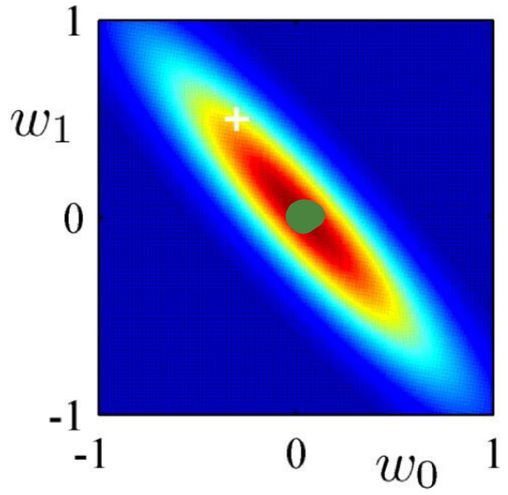
$$m_0 = [0, 0]$$

$$S_0^{-1} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}$$

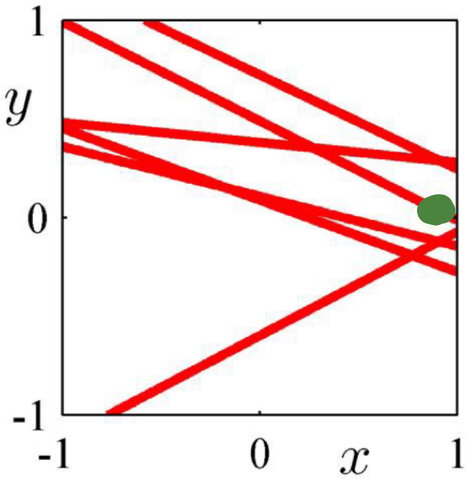
Observe 1 data point



likelihood



posterior



$X_1 \approx 0.9$

$Y_1 \approx 0.1$

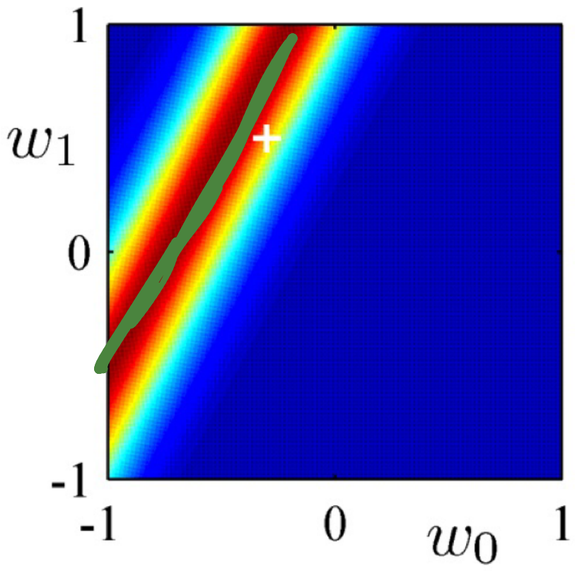
many optimal solutions

MLE

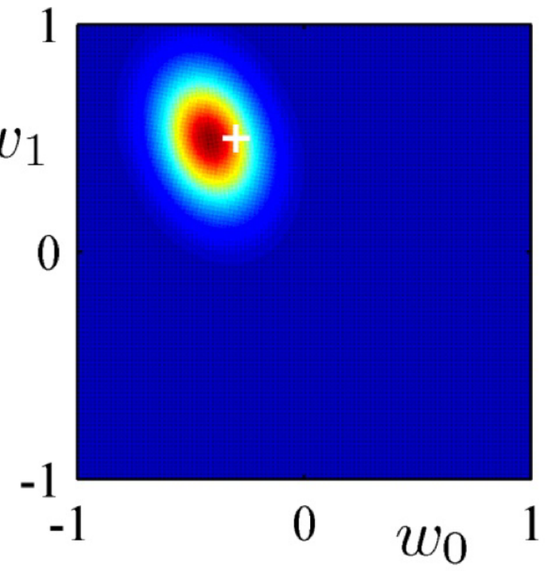
one optimal solution

Bayesian

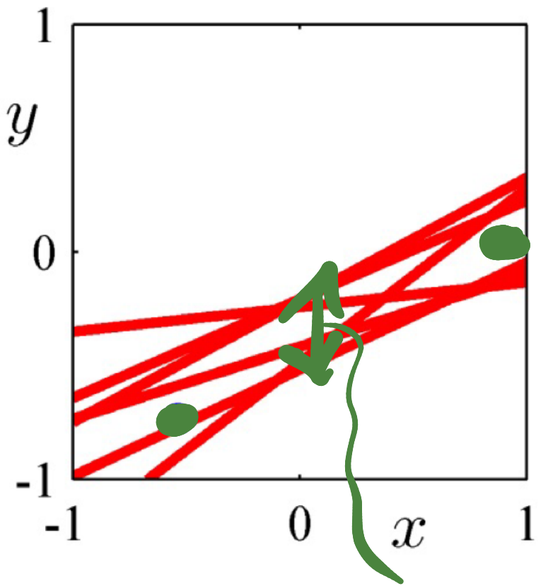
2 data points



likelihood

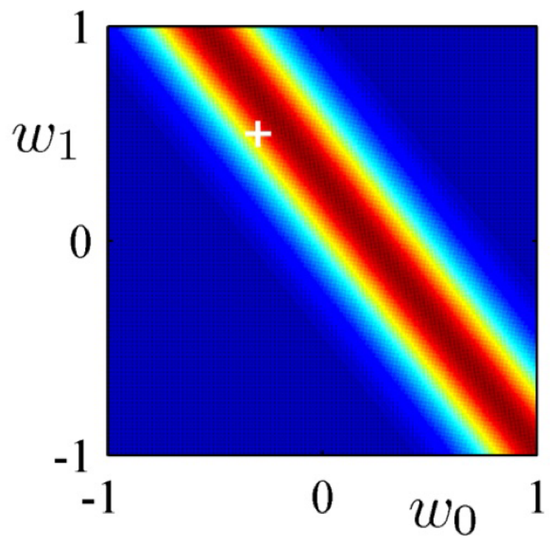


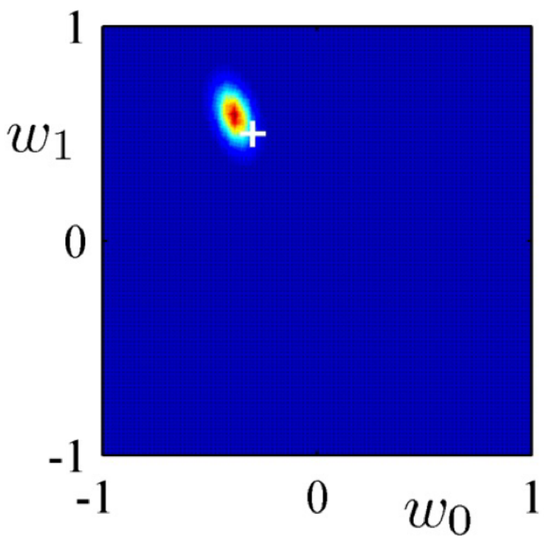
posterior

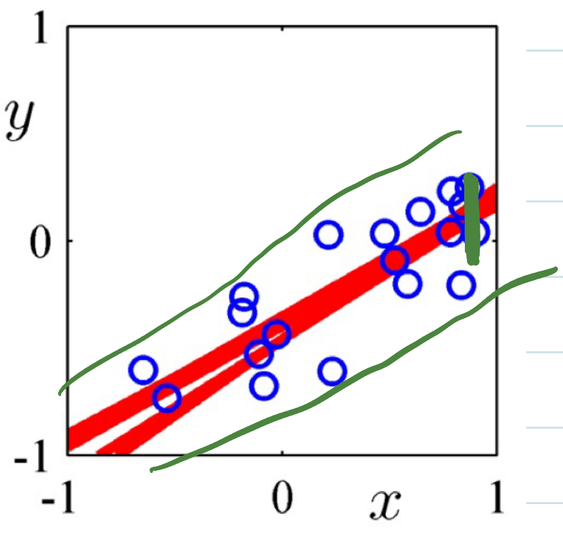


noise?

20 data points







Predictive Distribution

We want to predict t for new values of X

$$p(t \mid \bar{t}, \alpha, \beta) = \int p(t \mid w, \beta)\, p(w \mid \bar{t}, \alpha, \beta)\, dw$$

- t: new output
- $\bar{t}$: vector of previous outputs
- α: confidence in prior
- β: sensor conf
- $p(t \mid w, \beta)$: probability of new t given parameters
- $p(w \mid \bar{t}, \alpha, \beta)$: probability of parameters given observed outputs

$$= N\left(t \mid m_N^T \phi(x), \sigma_N^2(x)\right)$$

- m_N^T: mean values of parameters
- $\phi(x)$: basis functions

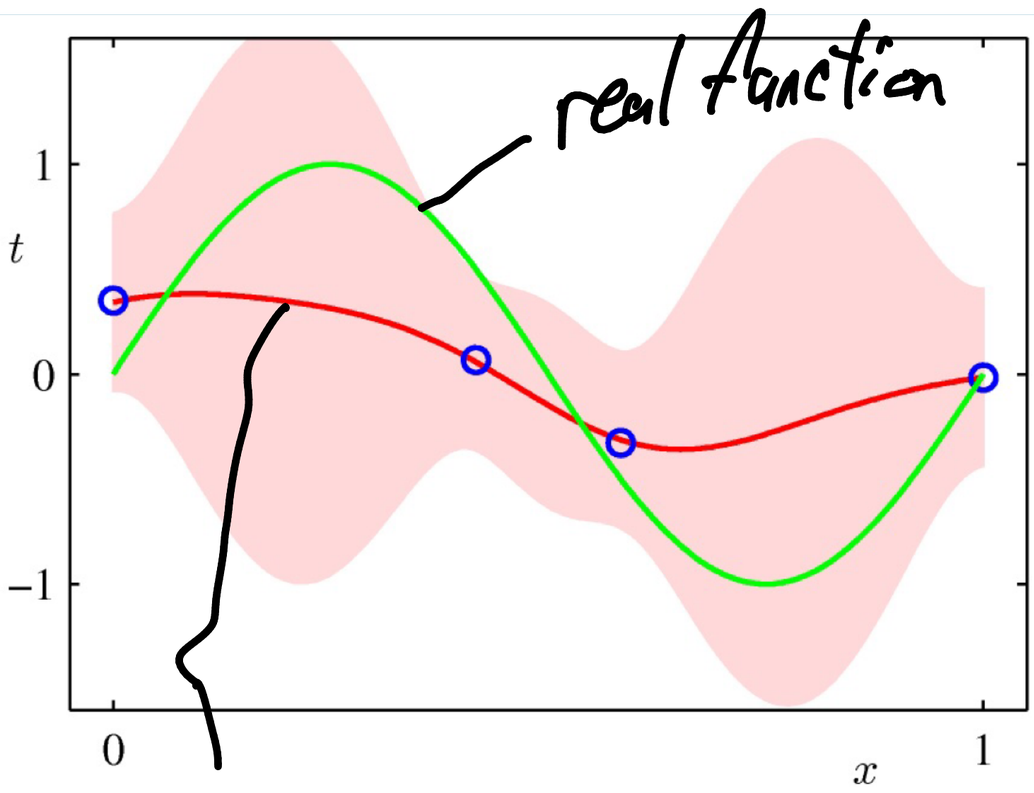
$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$$

- $\frac{1}{\beta}$: sensor Variance
- S_N: parameter Variance

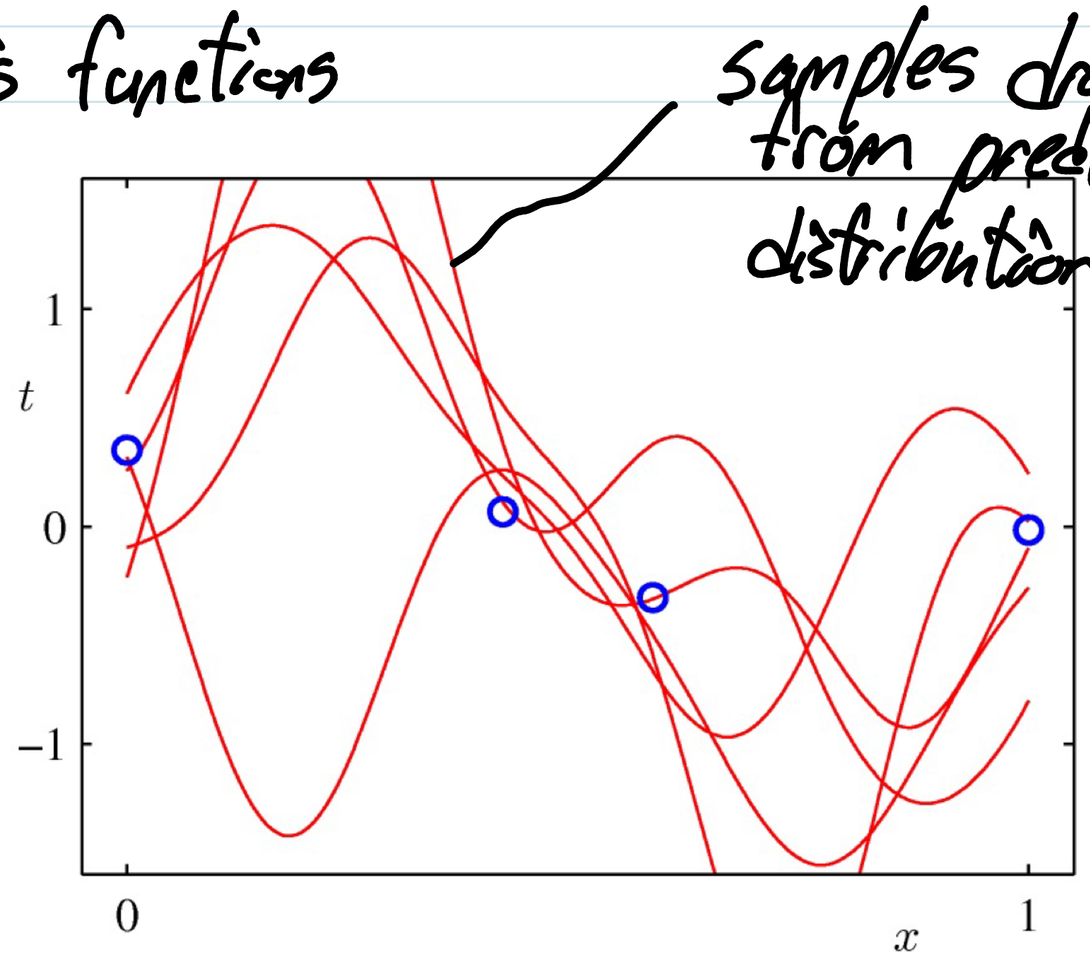
Example:

underlying process is sinusoidal data

4 data points, 9 Gaussian basis functions

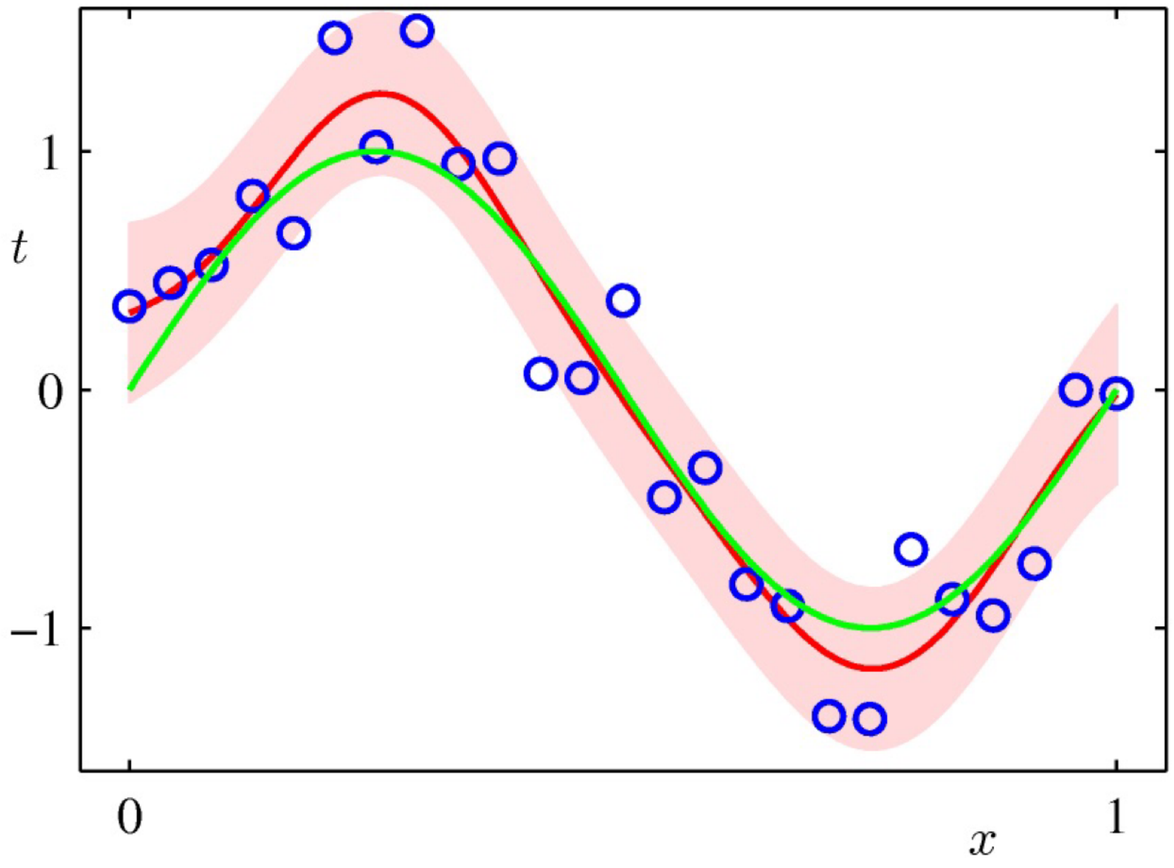


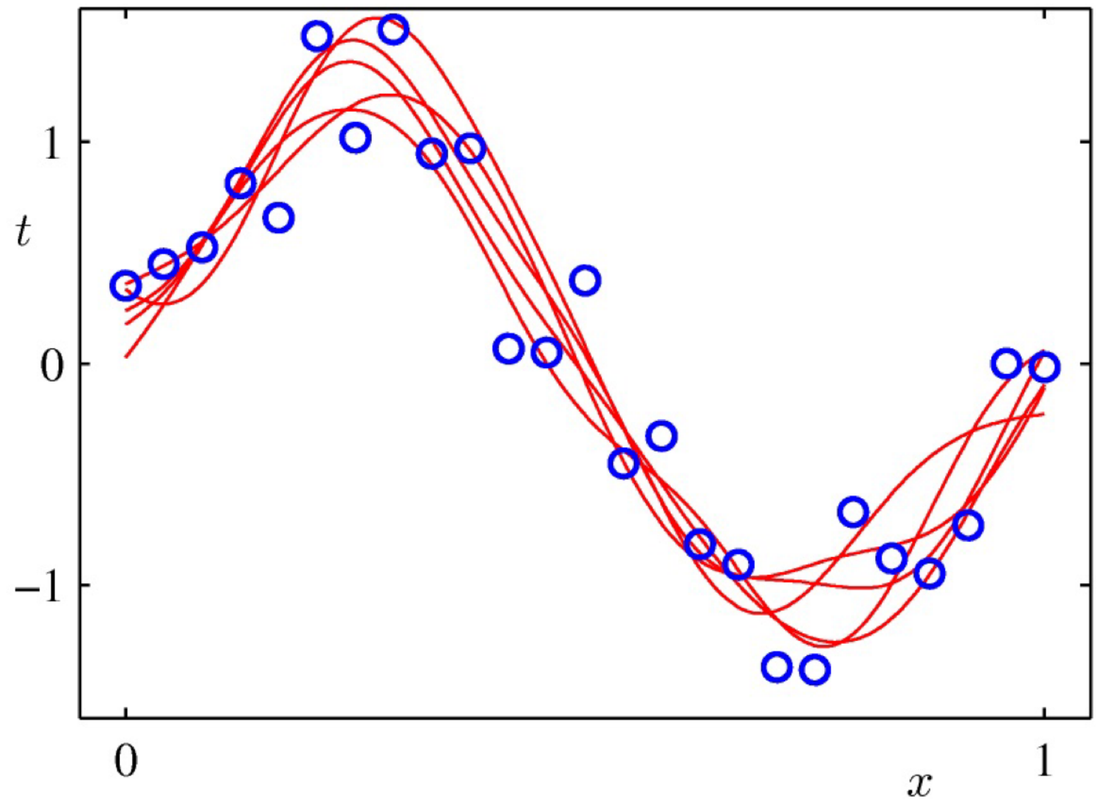
mean. of predictive distribution



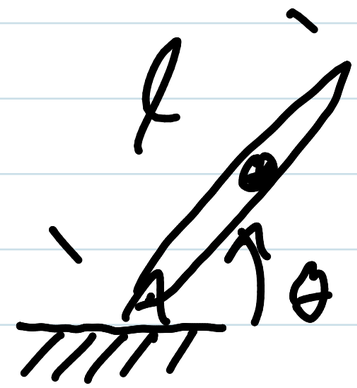
$$t = \sum_{i=1}^{9} c_i \exp\left(-\frac{(x-\mu_i)^2}{\sigma_i^2}\right)$$

Now with 25 data points





Pendulum Example



$I = \frac{ml^2}{12}$

Equation of motion

$$\frac{ml^2}{3}\ddot{\theta} + \frac{mgl\cos\theta}{2} = \tau$$

Goal: find m and l given we can measure $\theta, \ddot{\theta}, \tau$

pick $\ddot{\theta}$ as the output and τ and θ as input

$$X = [\tau \quad \theta]^T \qquad Y = \ddot{\theta}$$

$$\ddot{\theta} = \frac{3\tau}{ml^2} - \frac{3g\cos\theta}{2l}$$

Write model in the form $y = w^T \phi(x)$

$$w = \left[\frac{1}{ml^2} \quad \frac{1}{l} \right]^T$$

$$\phi(x) = \left[3\tau \quad -\frac{3g\cos\theta}{2} \right]^T$$

$$y = w^T \phi(x) = \left[\frac{1}{ml^2} \quad \frac{1}{l} \right] \begin{bmatrix} 3\tau \\ -\frac{3g\cos\theta}{2} \end{bmatrix} = \frac{3\tau}{ml^2} - \frac{3g\cos\theta}{2l}$$

To get started with Bayesian regression

- β^{-1} sensor noise
- m_0 prior mean of parameters
- S_0 prior covariance of parameters

Suppose we move the pendulum around by commanding torques and measure θ and $\ddot{\theta}$. N samples

$$t = [\ddot{\theta}_1 \ \ddot{\theta}_2 \ \dots \ \ddot{\theta}_N]^T$$

$$\Phi = \begin{bmatrix} 3\tau_1 & -3g\cos\theta_1/2 \\ 3\tau_2 & -3g\cos\theta_2/2 \\ \vdots & \\ 3\tau_N & -3g\cos\theta_N/2 \end{bmatrix}$$

posterior confidence $\quad S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$

posterior mean of parameters $\quad m_N = S_N(S_0^{-1} m_0 + \beta \Phi^T t)$

$$= \begin{bmatrix} \frac{1}{ml^2} & \frac{1}{l} \end{bmatrix}^T$$

the answer

or a predictive distribution of $\ddot{\theta}$

$$p(\ddot{\theta} \mid t, \alpha, \beta) = \mathcal{N}(\ddot{\theta} \mid m_N^T \phi(x), \sigma_N^2(x))$$

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

Other things to do!

1) predict τ given θ and $\ddot{\theta}$ by writing the EoM with different basis functions (inverse dynamics)

2) write discrete version of EoM

e.g $\theta_{i+1} = \theta_i + \dot{\theta}_i \Delta t$

$\dot{\theta}_{i+1} = \dot{\theta}_i + \ddot{\theta}_i \Delta t$

function of parameters

$$W = \begin{bmatrix} \frac{1}{ml^2} & \frac{1}{l} \end{bmatrix} = \begin{bmatrix} 1.63 & 1.66 \end{bmatrix}$$