

# CS 480/680

# Introduction to Machine Learning

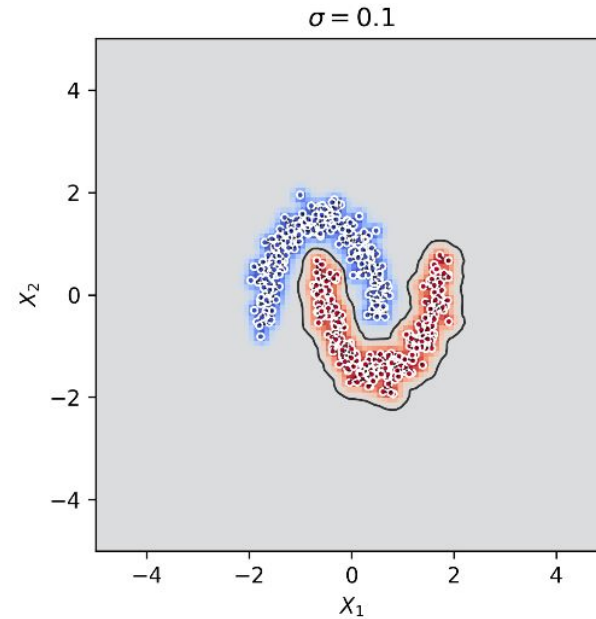
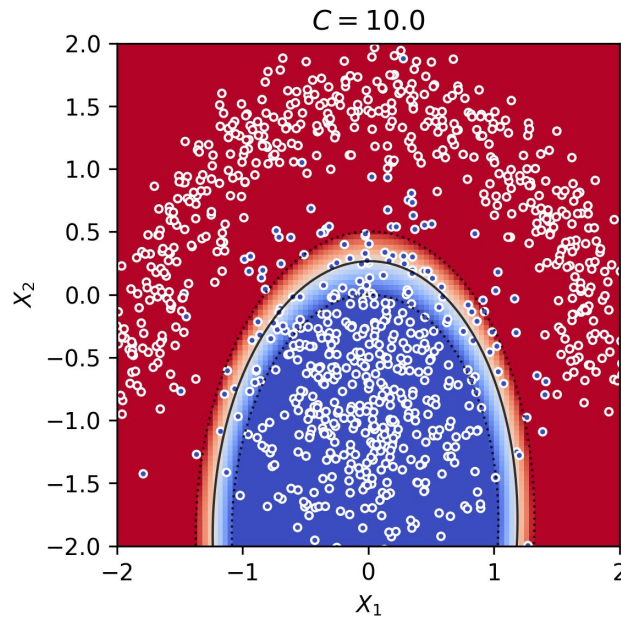
## Lecture 9a

## Maximum A Posteriori and Bayesian Learning

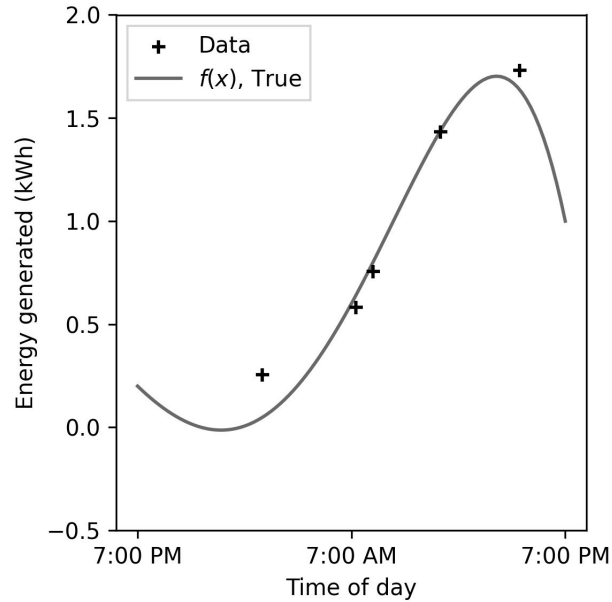
Kathryn Simone

8 October 2024

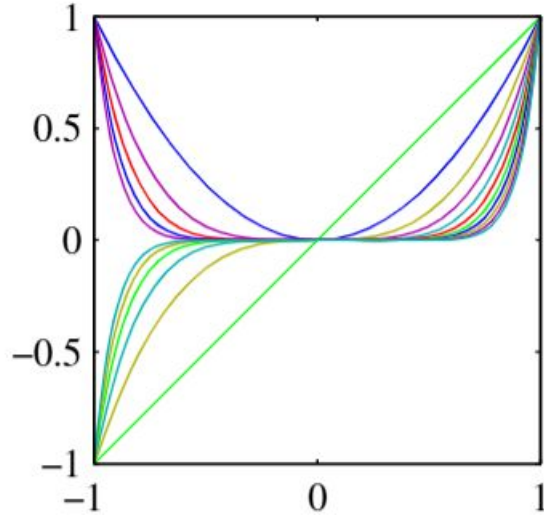
# We applied nonlinear basis functions to classification



# Many regression problems will require nonlinearity

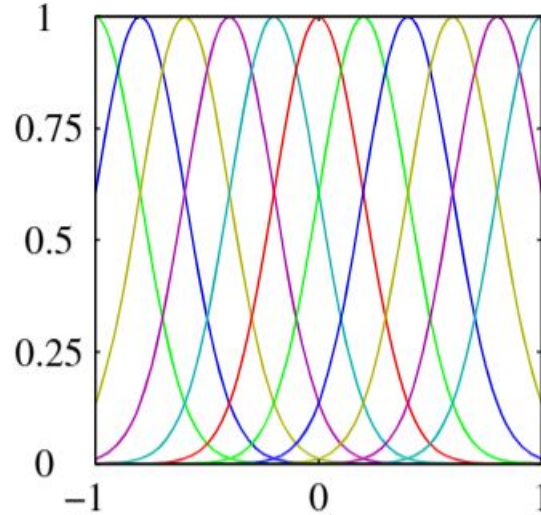


# We can use nonlinear basis functions in regression



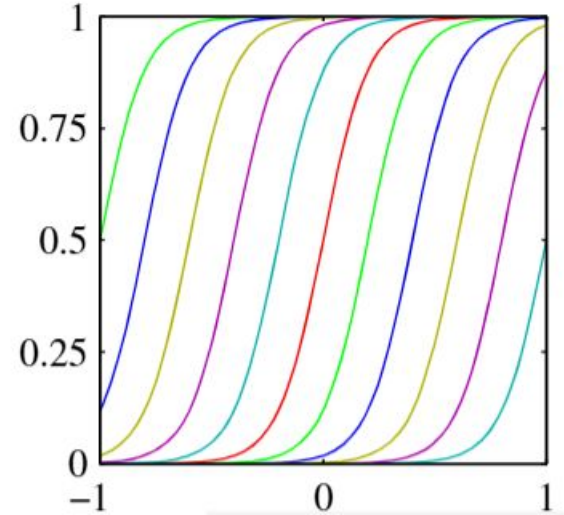
Polynomial:

$$\{\phi_j(\mathbf{x})\} = \{x^j\}$$



Gaussian:

$$\{\phi_j(\mathbf{x})\} = \left\{ e^{-\frac{(x-\mu_j)^2}{2\sigma^2}} \right\}$$



Sigmoidal:

$$\{\phi_j(\mathbf{x})\} = \left\{ \frac{1}{1 + e^{\frac{-(x-\mu_j)}{\sigma}}} \right\}$$

# Linear modelling with nonlinear basis functions

For a dataset of  $n$  pairs  $(x_i, y_i)$ ,  $x_i \in \mathbb{R}^d$ ,  $y_i \in \mathbb{R}$ , we consider the class of models defined by linear combinations of  $m$  fixed nonlinear basis functions of the input features:

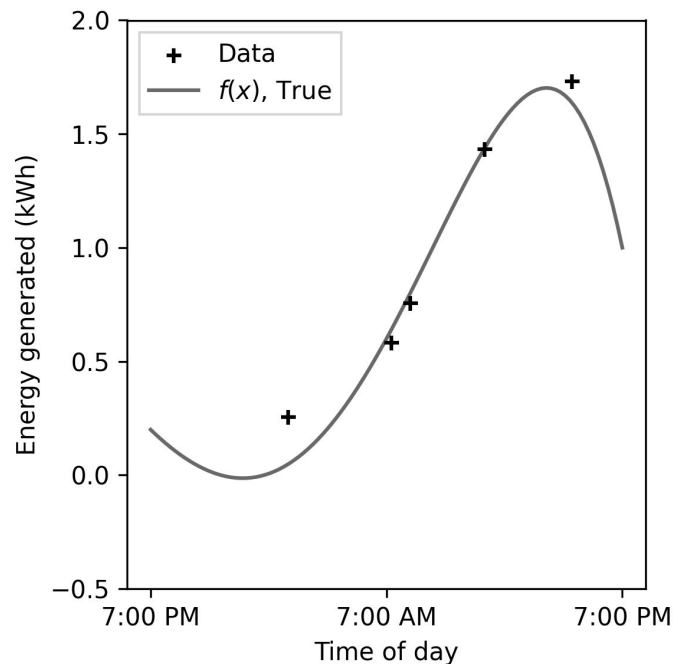
$$\begin{aligned}\hat{y} &= b + \sum_{j=1}^m w_j \phi(x_j) \\ &= \sum_{j=0}^m w_j \phi(x_j), \text{ with } \phi_0(x) = 1 \\ &= \langle w, \phi(x) \rangle\end{aligned}$$

Consider data generated from the model

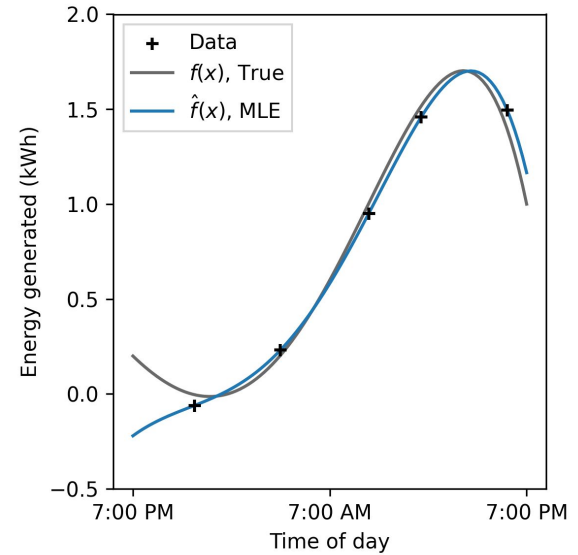
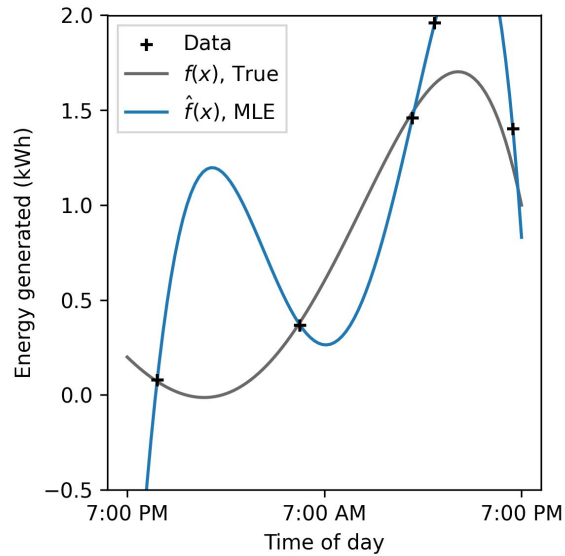
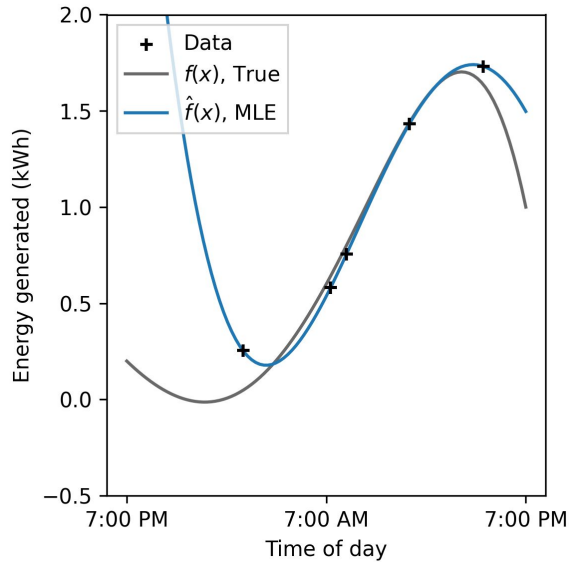
$$y_n = \langle w, \phi(x) \rangle + b + \eta,$$

Where

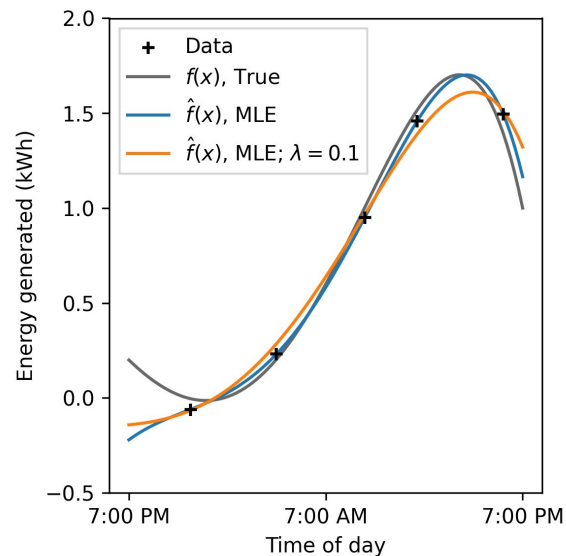
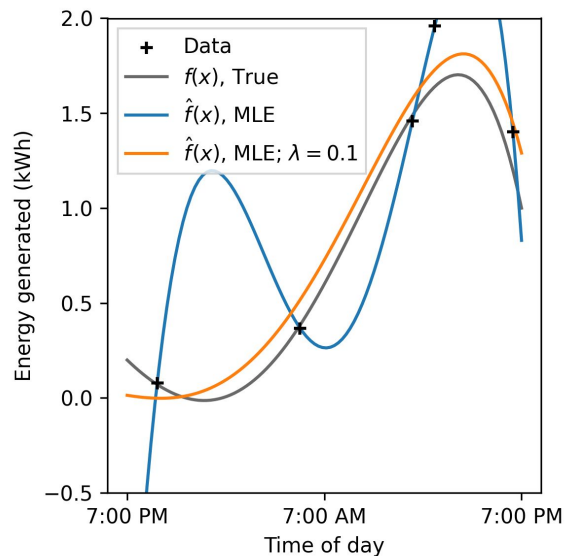
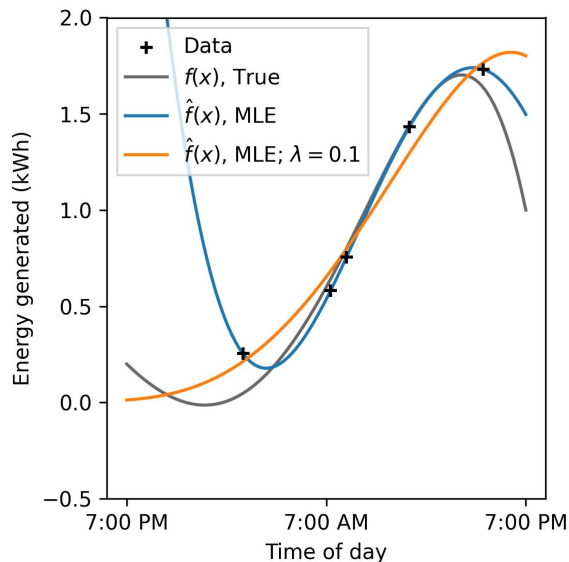
$$\begin{aligned}\phi(x) &= [1 \quad x \quad x^2 \quad x^3 \quad x^5] \\ w &= [0.2 \quad -1 \quad 0.9 \quad 0.7 \quad -0.2] \\ \eta &\sim \mathcal{N}(0, \sigma^2)\end{aligned}$$



# MLE may be prone to overfitting



# Ridge regression alleviates overfitting...



... but, could we do better?

# Key Questions

9a

- I. How can we incorporate prior knowledge into a model?
- II. How can we account for uncertainty in parameters?

9b

- III. What if we don't even know the structure of a model?



# Key Questions

**I. How can we incorporate prior knowledge into a model?**

II. How can we account for uncertainty in parameters?

# Motivating example: Windy day or not?



# Revisiting maximum likelihood estimation

We model each individual outcome  $y_i$  as a Bernoulli random variable,

$$y_i \sim \text{Bernoulli}(\pi).$$

where  $i = 1, 2, \dots, n$ , and the outcomes are independently and identically distributed (i.i.d). The likelihood function for a set of observations  $y = \{y_1, y_2, \dots, y_n\}$  defined as:

$$\begin{aligned}\mathcal{L}(\pi \mid \mathbf{y}) &= \prod_{i=1}^n \pi^{y_i} (1 - \pi)^{(1-y_i)} \\ &= p(\mathbf{y} \mid \pi)\end{aligned}$$

Where  $\pi$  is the probability of a windy day (i.e.  $y_i = 1$ ) for each realization of the Bernoulli random variable  $y_i$ .

Suppose we have data for five days in October:

$$\mathbf{y} = \{1, 0, 1, 0, 0\}$$

What is the MLE for  $\pi$ ?

$$\begin{aligned}\log \mathcal{L}(\pi \mid \mathbf{y}) &= \log(\pi) \sum_{i=1}^n y_i + \log(1 - \pi) \sum_{i=1}^n (1 - y_i) \\ \frac{\partial \log \mathcal{L}}{\partial \pi} &= \frac{1}{\pi} \sum_{i=1}^n y_i - \frac{1}{1 - \pi} \sum_{i=1}^n (1 - y_i) = 0 \\ \implies \hat{\pi}_{MLE} &= \frac{1}{n} \sum_{i=1}^n y_i \\ &= \frac{1}{5} (1 + 0 + 1 + 0 + 0) = 0.4\end{aligned}$$

# Representing prior knowledge

Suppose that historical data suggests that  $\pi$  tends to fall around 0.7.

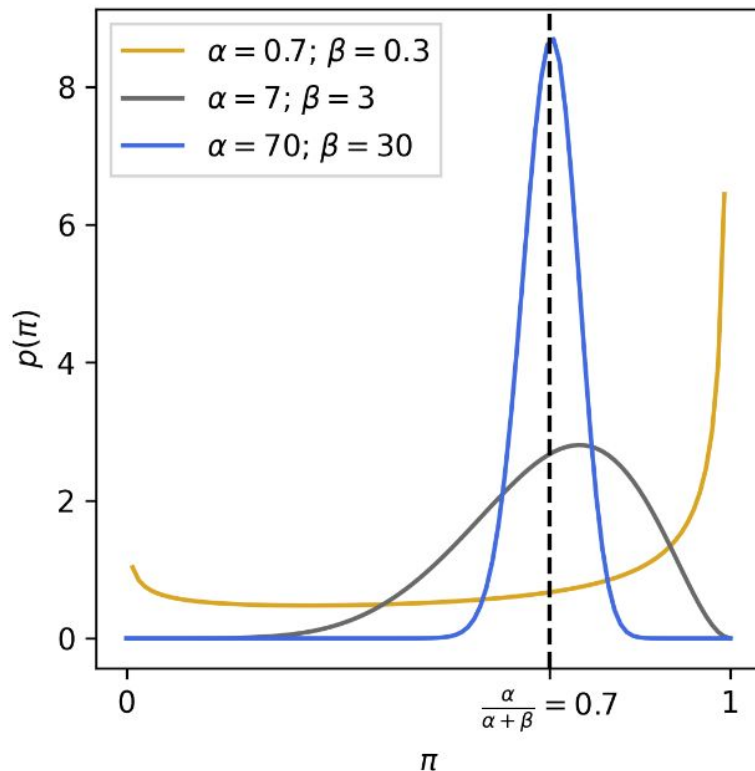
A common choice to represent prior knowledge about  $\pi$  for the Bernoulli model is to use a beta distribution:

$$\begin{aligned}\pi &\sim \text{Beta}(\alpha, \beta) \\ &= \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{\int \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi} \\ p(\pi) &\propto \pi^{\alpha-1}(1-\pi)^{\beta-1}\end{aligned}$$

For which the expectation of  $\pi$  is

$$E[\pi] = \frac{\alpha}{\alpha + \beta}$$

We can capture both our knowledge and uncertainty about  $\pi$  by letting  $\alpha = 7$ , and  $\beta = 3$ .



# Incorporating prior knowledge

Bayesian update:

$$p(\pi \mid \mathbf{y}) \propto p(\mathbf{y} \mid \pi)p(\pi)$$

where:

$p(\pi)$  is the **prior** distribution,

$p(\mathbf{y} \mid \pi)$ , is the **likelihood**, and

$p(\pi \mid \mathbf{y})$  is the **posterior** distribution.

For our Bernoulli-Beta model, we have

$$\begin{aligned} p(\pi \mid \mathbf{y}) &\propto \text{Bernoulli}(\mathbf{y} \mid \pi)\text{Beta}(\alpha, \beta) \\ &\propto \left( \prod_{i=1}^n \pi^{y_i} (1 - \pi)^{(1-y_i)} \right) \left( \pi^{\alpha-1} (1 - \pi)^{\beta-1} \right) \end{aligned}$$

We can therefore rewrite  $p(\pi \mid \mathbf{y})$  as

$$\begin{aligned} p(\pi \mid \mathbf{y}) &\propto \left( \pi^k (1 - \pi)^{(n-k)} \right) \left( \pi^{\alpha-1} (1 - \pi)^{\beta-1} \right) \\ &\propto \pi^{k+\alpha-1} (1 - \pi)^{n-k+\beta-1} \end{aligned}$$

The likelihood can be simplified by introducing a Binomial random variable

$$b_n \sim \sum_{i=1}^n y_i,$$

for which the probability of  $k$  windy days out of  $n$  is

$$\Pr[b_n = k \mid \pi] = \frac{\pi^k (1 - \pi)^{n-k}}{n! / (k!(n-k)!)}$$

# Incorporating prior knowledge

Bayesian update:

$$p(\pi \mid \mathbf{y}) \propto p(\mathbf{y} \mid \pi)p(\pi)$$

where:

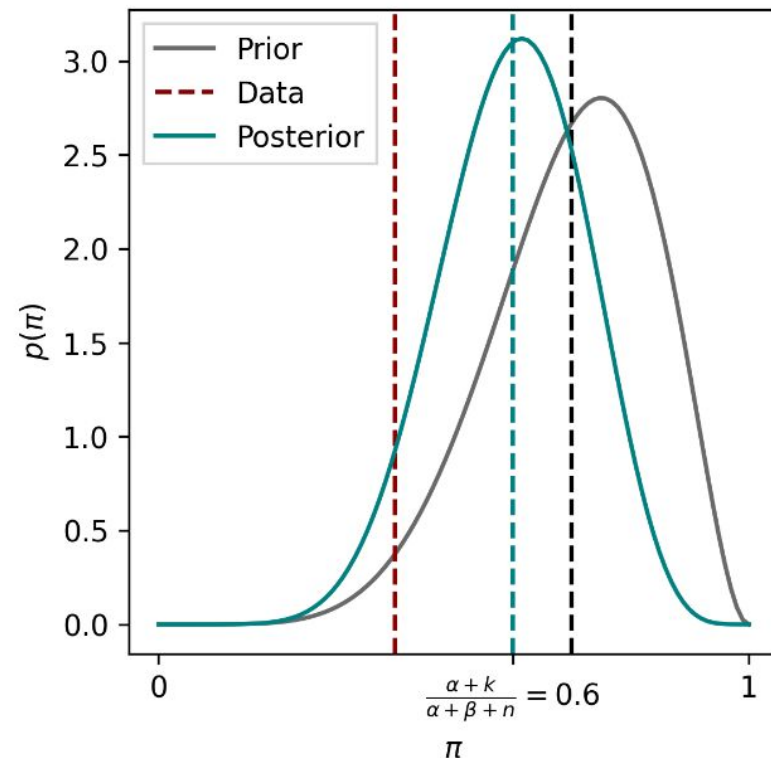
$p(\pi)$  is the **prior** distribution,

$p(\mathbf{y} \mid \pi)$ , is the **likelihood**, and

$p(\pi \mid \mathbf{y})$  is the **posterior** distribution.

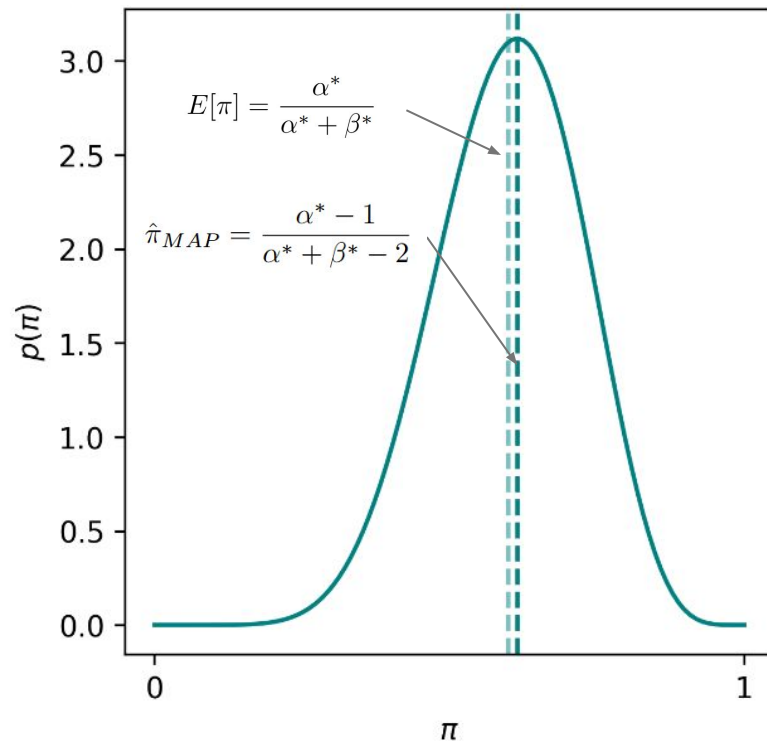
$$\begin{aligned} p(\pi \mid \mathbf{y}) &\propto \left( \pi^k (1 - \pi)^{(n-k)} \right) \left( \pi^{\alpha-1} (1 - \pi)^{\beta-1} \right) \\ &\propto \pi^{k+\alpha-1} (1 - \pi)^{n-k+\beta-1} \end{aligned}$$

$$\implies \pi \mid \mathbf{y} = \text{Beta}(k + \alpha, n - k + \beta)$$



# Maximum a posteriori (MAP) estimate

$$\begin{aligned}\hat{\pi}_{MAP} &= \operatorname{argmax}_{\pi} \pi^{k+\alpha-1} (1-\pi)^{n-k+\beta-1} \\ &= \frac{\alpha + k - 1}{\alpha + \beta + n - 2}\end{aligned}$$



# Towards MAP for linear regression: Recall MLE

$$Y \sim \mathcal{N}(w^T X, \sigma^2)$$

$$y_i = w^T x_i + \mathcal{N}(0, \sigma^2)$$

$$p(\mathbf{y} \mid \mathbf{x}, w, \sigma^2) = \mathcal{N}(w^T \mathbf{x}, \sigma^2)$$

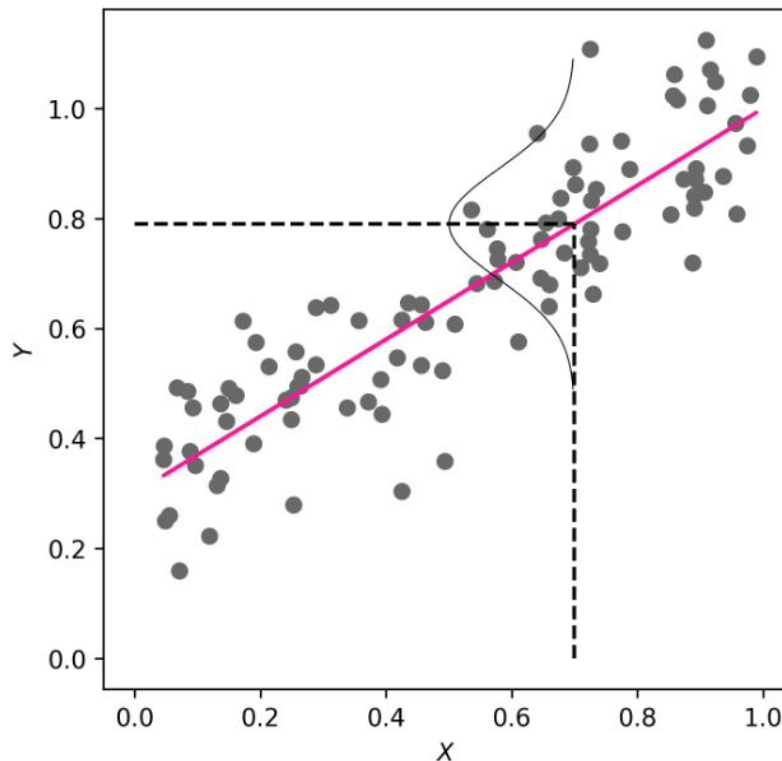
$$= \prod_{i=1}^n p(y_i | x_i, w, \sigma^2)$$

$$\Rightarrow \hat{w} = \operatorname{argmax}_w \prod_{i=1}^n p(y_i | x_i, w, \sigma^2)$$

$$= \operatorname{argmax}_w \prod_{i=1}^n \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2} (y_i - w^T x_i)^2}$$

$$= \operatorname{argmax}_w \sum_{i=1}^n -(y_i - w^T x_i)^2$$

$$= \operatorname{argmin}_w \sum_{i=1}^n (y_i - w^T x_i)^2$$





# Incorporating a prior in linear regression with MAP

$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}, \sigma^2) \propto p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w})$$
$$\implies \hat{\mathbf{w}} = \arg \max_{\mathbf{w}} p(\mathbf{w} \mid \mathbf{X}, \mathbf{y})$$


$$p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma})$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\boldsymbol{\Sigma})}} \exp \left( -\frac{1}{2} \mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w} \right)$$

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp \left( -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right)$$

# Incorporating a prior in linear regression with MAP

$$\hat{\mathbf{w}} = \arg \max_{\mathbf{w}} p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w})$$

$$= \arg \max_{\mathbf{w}} \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp \left( -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right) \cdot \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} \mathbf{w}^T \Sigma^{-1} \mathbf{w} \right)$$

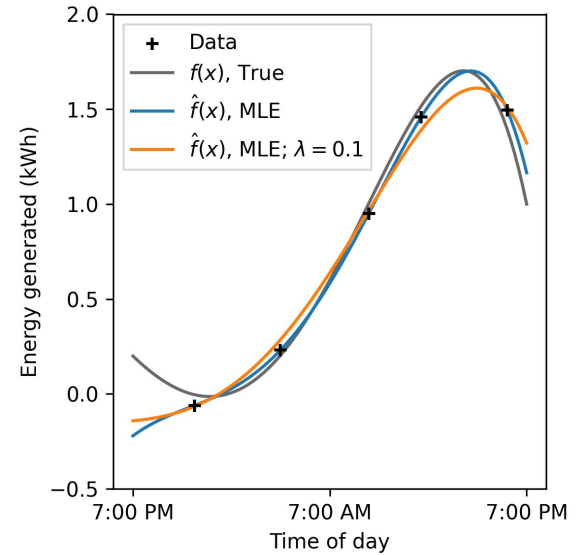
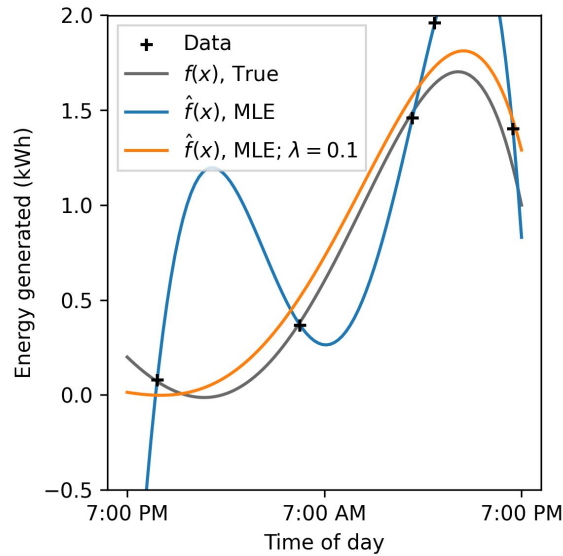
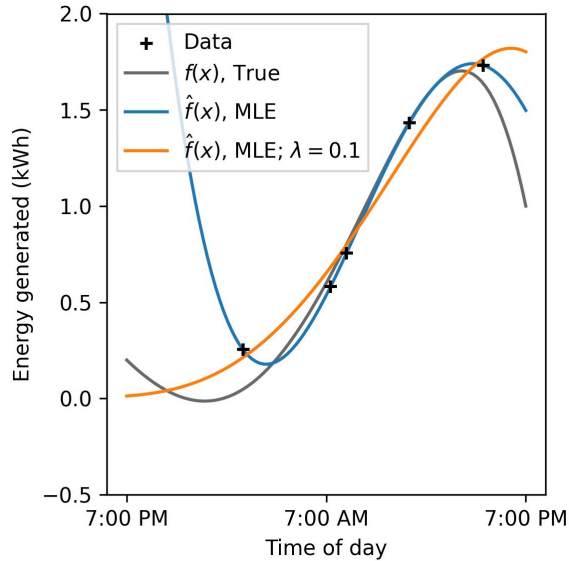

$$= \arg \max_{\mathbf{w}} \left[ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) - \frac{1}{2} \mathbf{w}^T \Sigma^{-1} \mathbf{w} \right]$$

$$= \arg \min_{\mathbf{w}} \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \mathbf{w}^T \Sigma^{-1} \mathbf{w}$$

if we let  $\Sigma^{-1} = \lambda \mathbf{I}$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

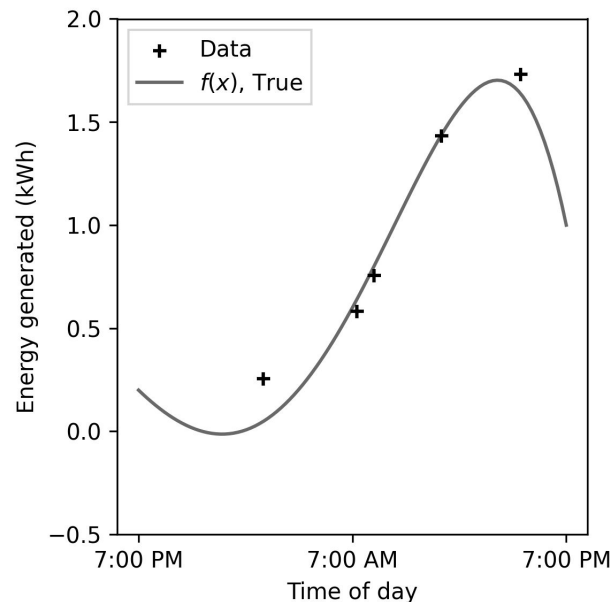
# L2-Regularization (Ridge) regression as imposing a prior on $w$

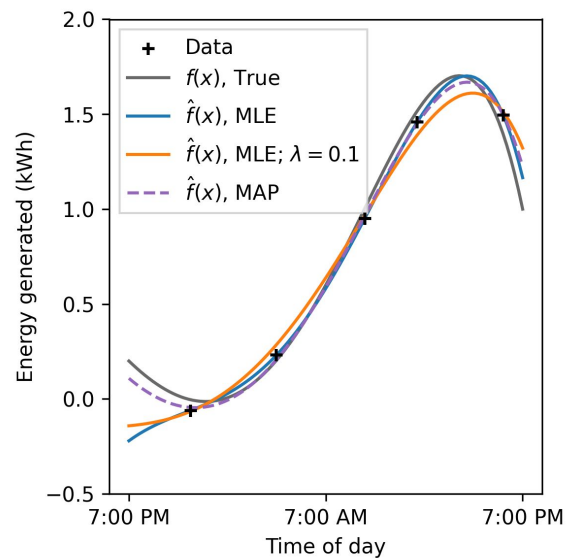
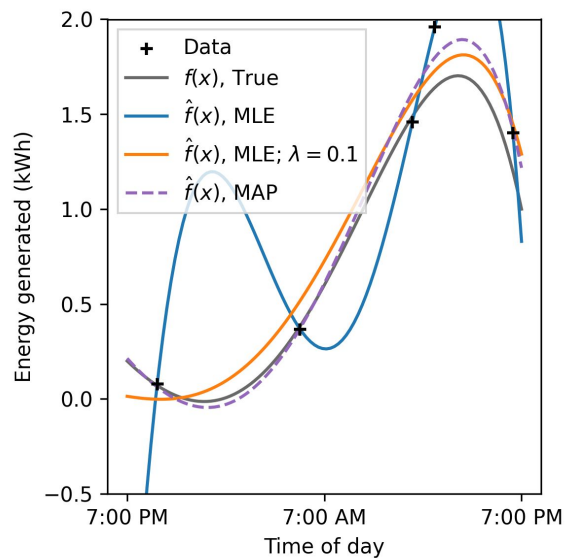
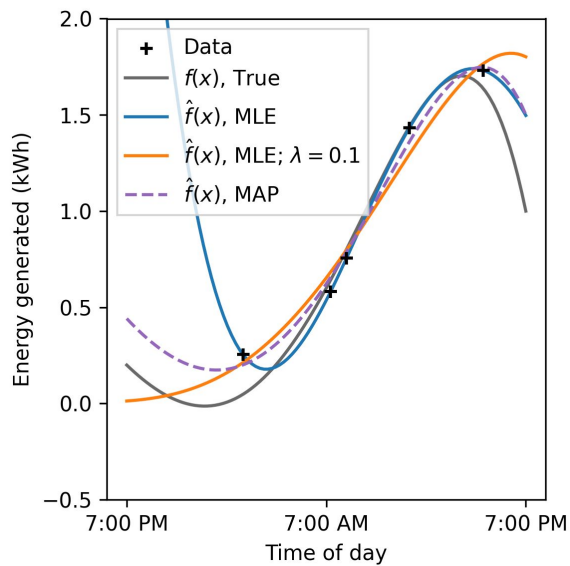


# Flexible priors in linear regression with MAP

$$p(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\boldsymbol{\Sigma})}} \exp \left( -\frac{1}{2} (\mathbf{w} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\mathbf{w} - \boldsymbol{\mu}_0) \right)$$

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0 & -1 & 1 & 0 & 0 \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$





# Key Questions

I. How do you incorporate prior knowledge into a model?

**II. How can we account for uncertainty in parameters?**

# Bayesian linear regression

## Classical Linear Regression

- OLS, MLE, and MAP produce point estimates for  $\mathbf{w}$
- Assumes there exists a true underlying  $\mathbf{w}$

## Bayesian Linear Regression

- Computes a weighted average prediction over the posterior distribution of  $\mathbf{w}$

→

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) &\propto \exp\left(-\frac{1}{2}\mathbf{w}^T \Sigma^{-1} \mathbf{w}\right) \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_n (y_n - \mathbf{w}^T \mathbf{x}_n)^2\right) \\ &\propto \exp\left(-\frac{1}{2} \left( \mathbf{w}^T \Sigma^{-1} \mathbf{w} + \frac{1}{\sigma^2} \sum_n (y_n - \mathbf{w}^T \mathbf{x}_n)^2 \right)\right) \\ &\propto \exp\left(-\frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A} (\mathbf{w} - \bar{\mathbf{w}})\right) \end{aligned}$$

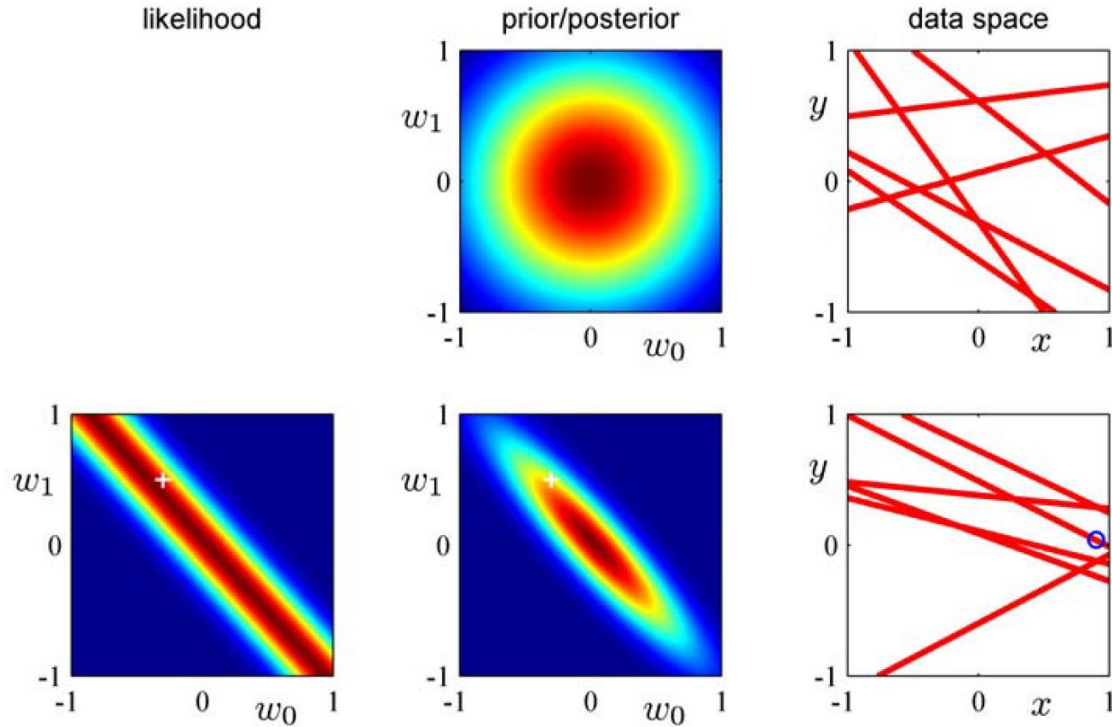
The mean  $\bar{\mathbf{w}}$  and covariance matrix  $\mathbf{A}$  of the posterior are given by:

$$\begin{aligned} \bar{\mathbf{w}} &= \mathbf{A}^{-1} \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{y} \\ \mathbf{A} &= \sigma^{-2} \mathbf{X}^T \mathbf{X} + \Sigma^{-1} \end{aligned}$$

which follows a multivariate normal distribution:

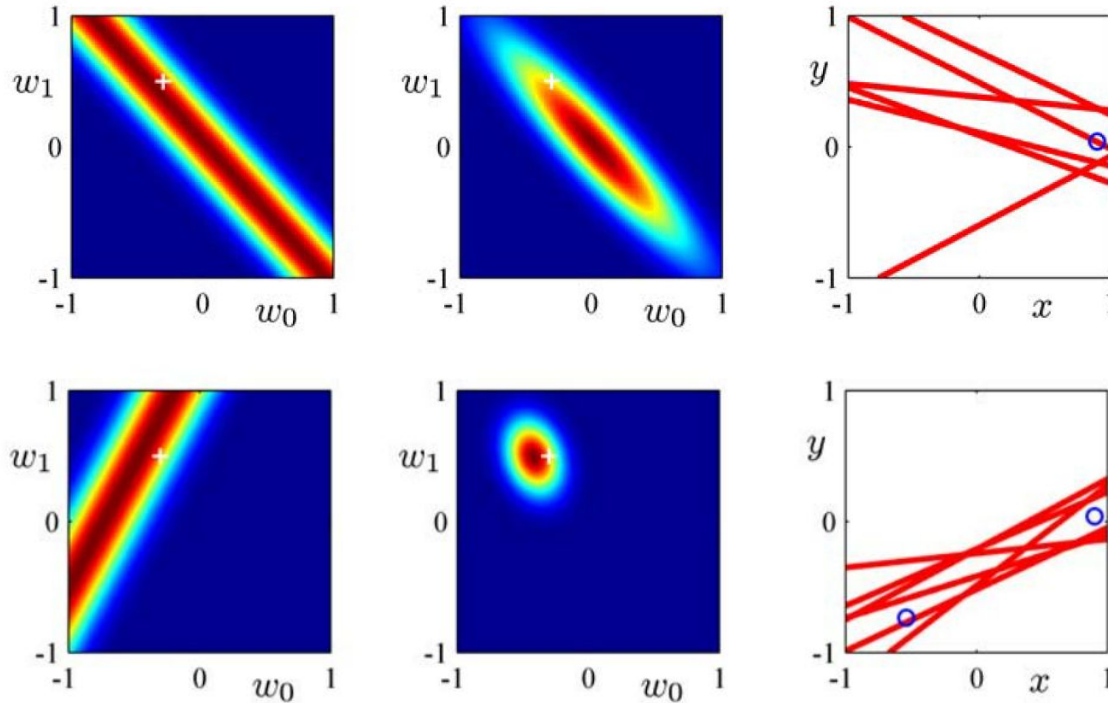
$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) = \mathcal{N}(\bar{\mathbf{w}}, \mathbf{A}^{-1})$$

# Sequential Bayesian update (1/3)

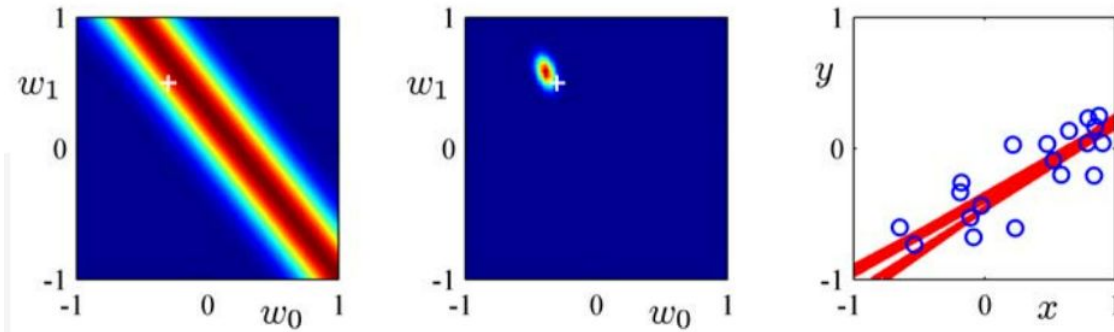




# Sequential Bayesian update (2/3)



# Sequential Bayesian update (3/3)



# Prediction in Bayesian linear regression

$$p(y^* | \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int_{\mathbf{w}} p(y^* | \mathbf{x}^*, \mathbf{w}) p(\mathbf{w} | \mathbf{X}, \mathbf{y}) d\mathbf{w}$$

$p(y^* | \mathbf{x}^*, \mathbf{w})$  is the **likelihood** of  $y^*$ , given the input  $\mathbf{x}^*$  and the weight vector  $\mathbf{w}$ .

$p(\mathbf{w} | \mathbf{X}, \mathbf{y})$  is the **posterior distribution** of  $\mathbf{w}$  given the training data.

# Prediction in Bayesian linear regression

$$p(y^* | \mathbf{x}^*, \mathbf{X}, \mathbf{y}) = \int_{\mathbf{w}} p(y^* | \mathbf{x}^*, \mathbf{w}) p(\mathbf{w} | \mathbf{X}, \mathbf{y}) d\mathbf{w}$$

$$p(y^* | \mathbf{x}^*, \mathbf{w}) = \exp \left( -\frac{(y^* - \mathbf{x}^{*T} \mathbf{w})^2}{2\sigma^2} \right)$$

$$p(\mathbf{w} | \mathbf{X}, \mathbf{y}) = \exp \left( -\frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A} (\mathbf{w} - \bar{\mathbf{w}}) \right)$$

# Prediction in Bayesian linear regression

$$\begin{aligned} p(y^* \mid x^*, \mathbf{X}, \mathbf{y}) &= \int_{\mathbf{w}} \exp \left( -\frac{(y^* - x^{*T} \mathbf{w})^2}{2\sigma^2} \right) \exp \left( -\frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A} (\mathbf{w} - \bar{\mathbf{w}}) \right) d\mathbf{w} \\ &= \mathcal{N} \left( x^{*T} \bar{\mathbf{w}}, \sigma^2 + x^{*T} \mathbf{A}^{-1} x^* \right) \end{aligned}$$