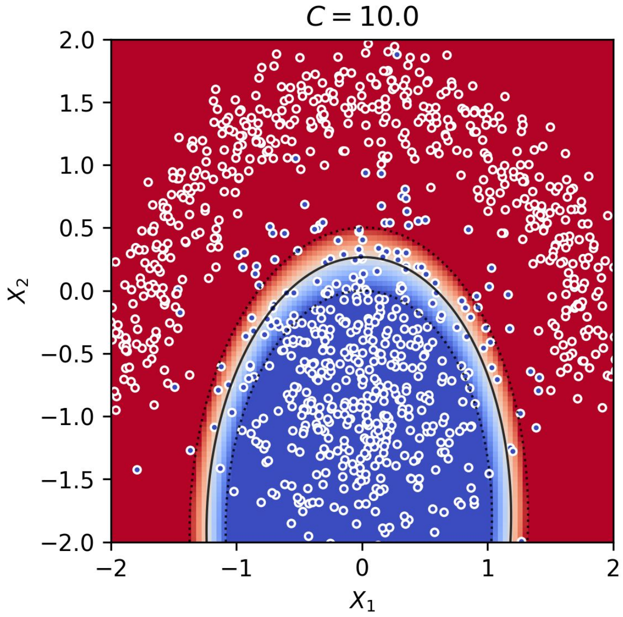
CS 480/680
Introduction to Machine Learning

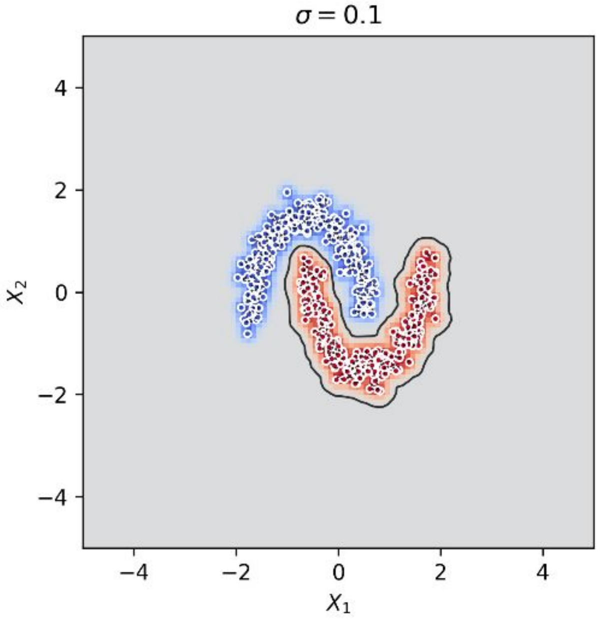
Lecture 9a
Maximum A Posteriori and Bayesian Learning

Kathryn Simone
8 October 2024

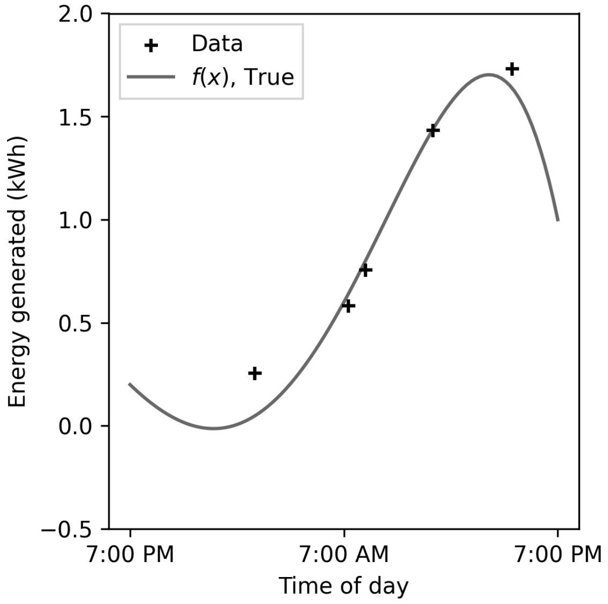
UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

We applied nonlinear basis functions to classification

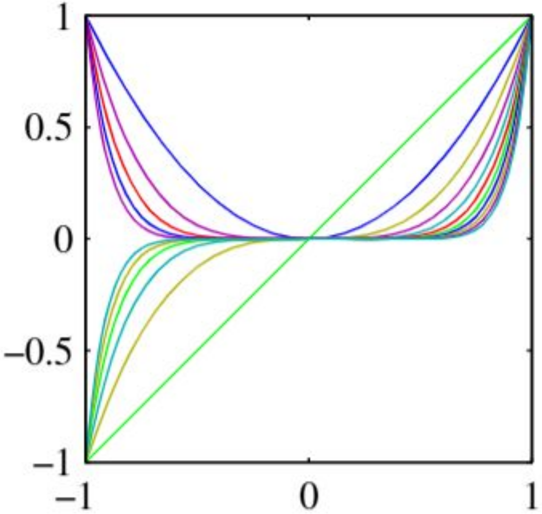




Many regression problems will require nonlinearity

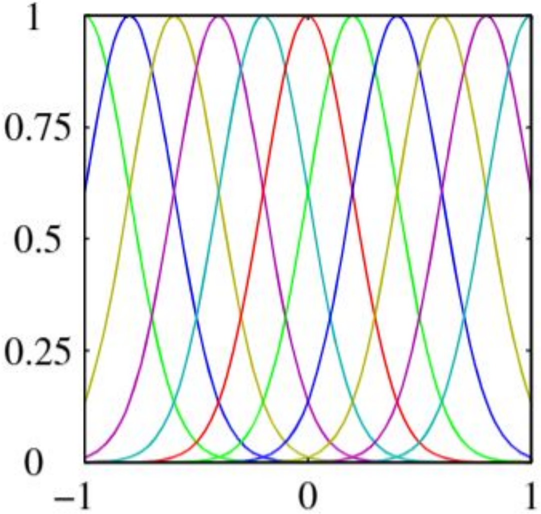


We can use nonlinear basis functions in regression



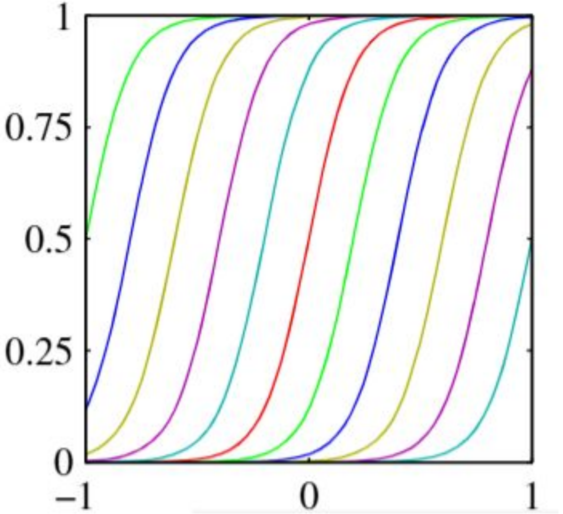
Polynomial:

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



Gaussian:

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



Sigmoidal:

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

Pattern Recognition and Machine Learning, Section 3.1

Linear modelling with nonlinear basis functions

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

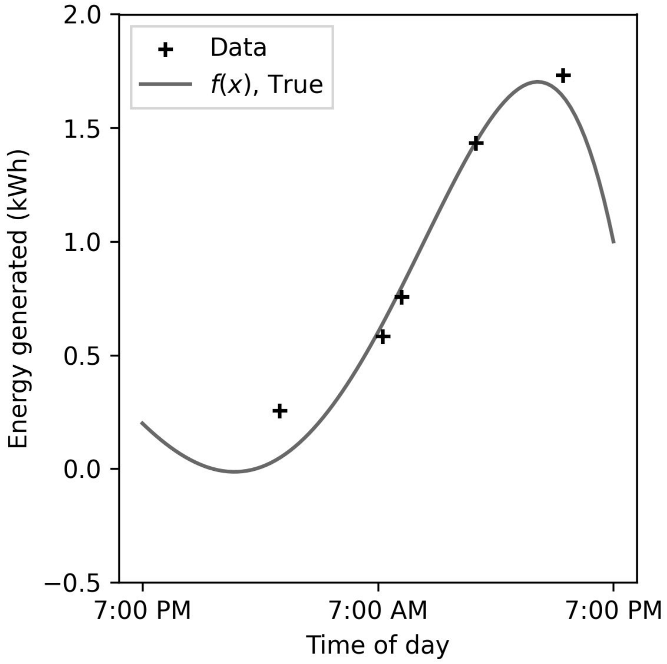
$$\begin{aligned}\hat{y} &= b + \sum_{j=1}^{m} w_j \phi(x_j) \\ &= \sum_{j=0}^{m} w_j \phi(x_j), \text{with } \phi_0(x) = 1 \\ &= \langle w, \phi(x) \rangle\end{aligned}$$

Consider data generated from the model

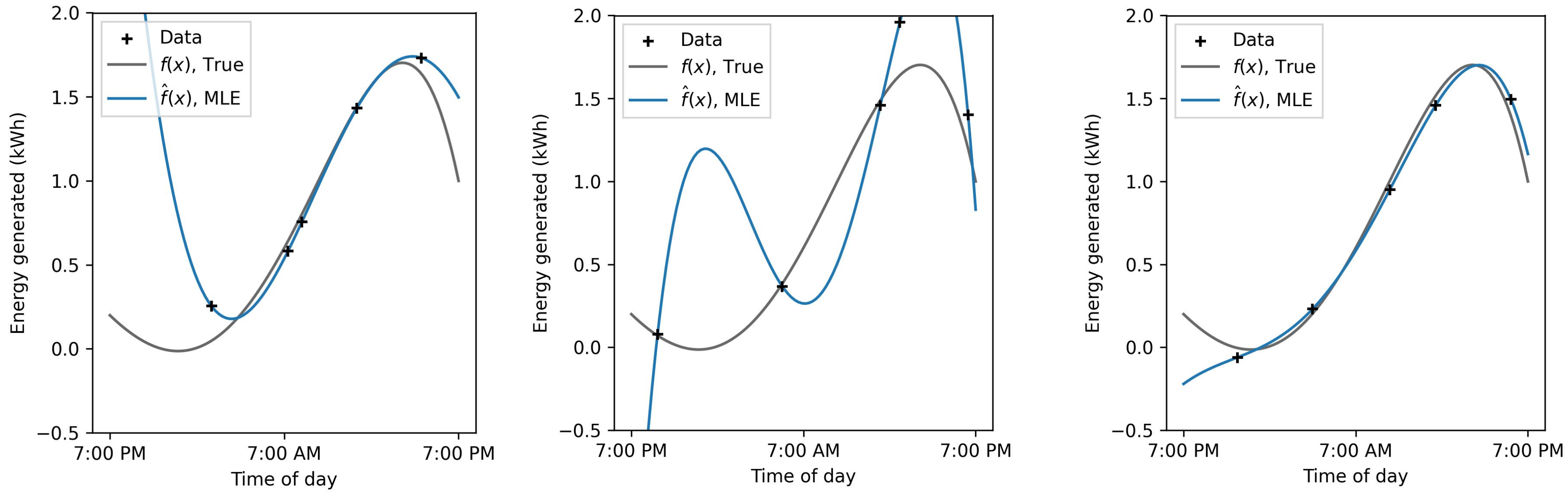
$$y_n = \langle w, \phi(x) \rangle + b + \eta,$$

Where

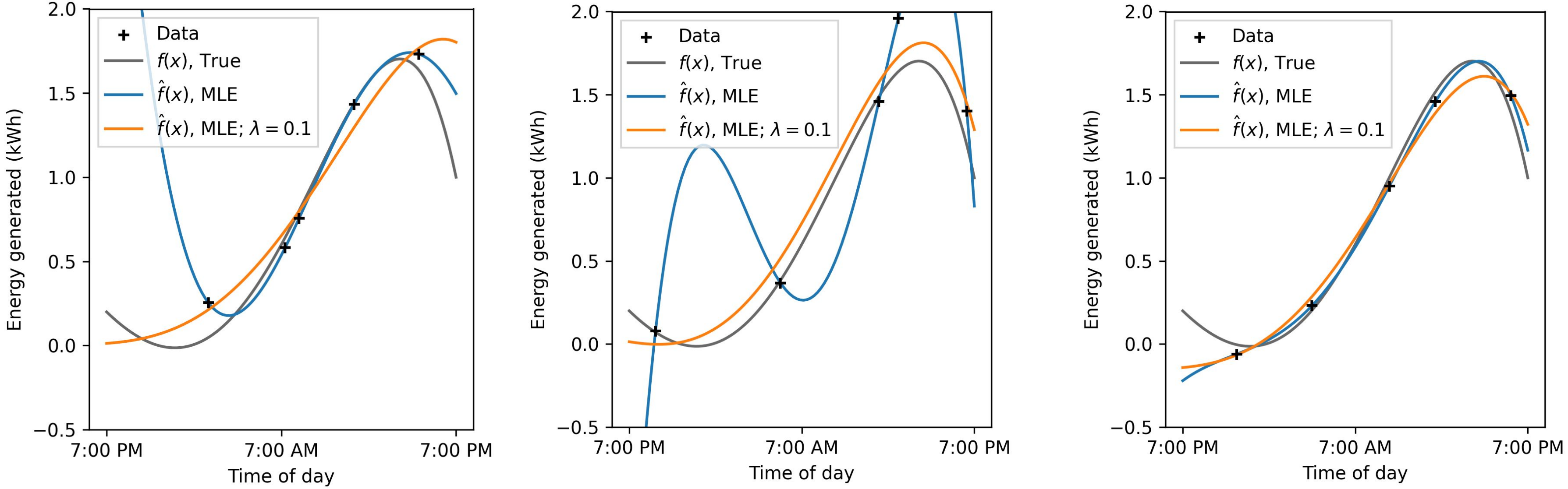
$$\begin{aligned}\phi(x) &= \begin{bmatrix} 1 & x & x^2 & x^3 & x^5 \end{bmatrix} \\ w &= \begin{bmatrix} 0.2 & -1 & 0.9 & 0.7 & -0.2 \end{bmatrix} \\ \eta &\sim \mathcal{N}(0, \sigma^2)\end{aligned}$$



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

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

Where π 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:

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

What is the MLE for π?

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

Representing prior knowledge

Suppose that historical data suggests that π tends to fall around 0.7.

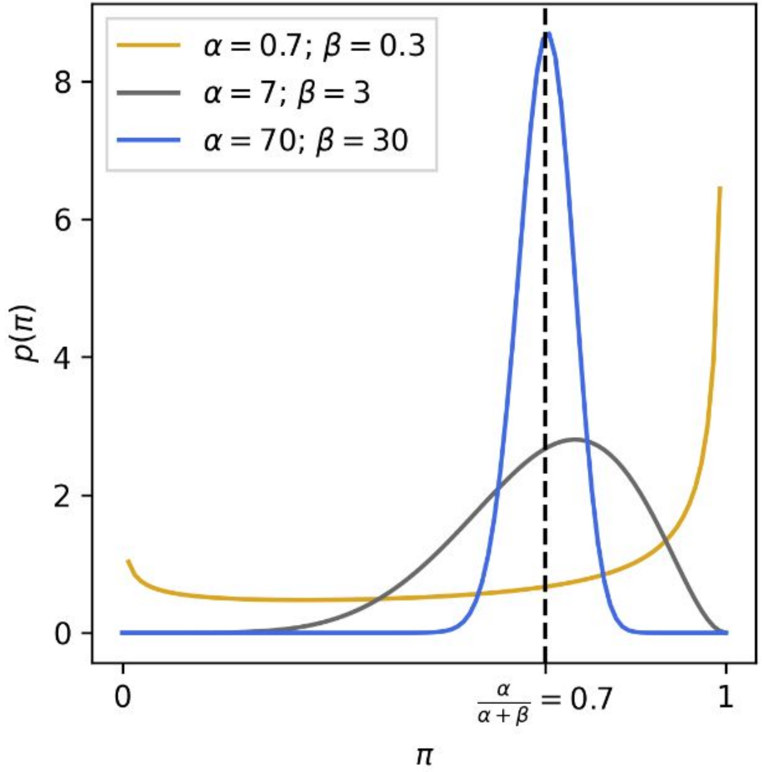
A common choice to represent prior knowledge about π for the Bernoulli model is to use a beta distribution:

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

For which the expectation of π is

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

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



Incorporating prior knowledge

Bayesian update:

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

where:
$p(\pi)$ is the **prior** distribution,
$p(\boldsymbol{y} \mid \pi)$, is the **likelihood**, and
$p(\pi \mid \boldsymbol{y})$ is the **posterior** distribution.

For our Bernouilli-Beta model, we have

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

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

$$p(\pi \mid \boldsymbol{y}) \propto \left(\pi^{k} (1-\pi)^{(n-k)} \right) \left(\pi^{\alpha-1} (1-\pi)^{\beta-1} \right)$$
$$\propto \pi^{k+\alpha-1} (1-\pi)^{n-k+\beta-1}$$

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

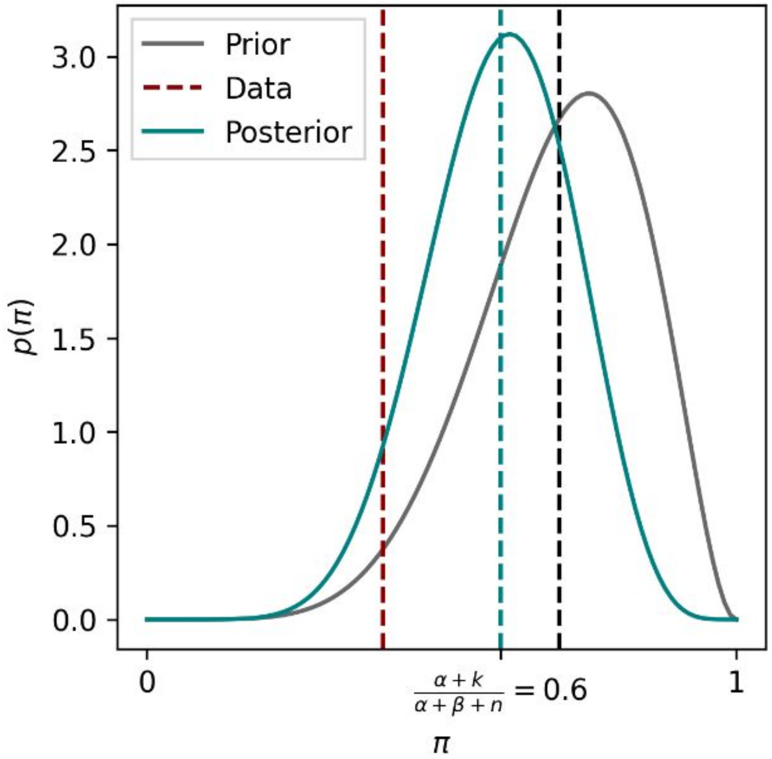
Bayesian update:

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

where:
$p(\pi)$ is the **prior** distribution,
$p(\boldsymbol{y} \mid \pi)$, is the **likelihood**, and
$p(\pi \mid \boldsymbol{y})$ is the **posterior** distribution.

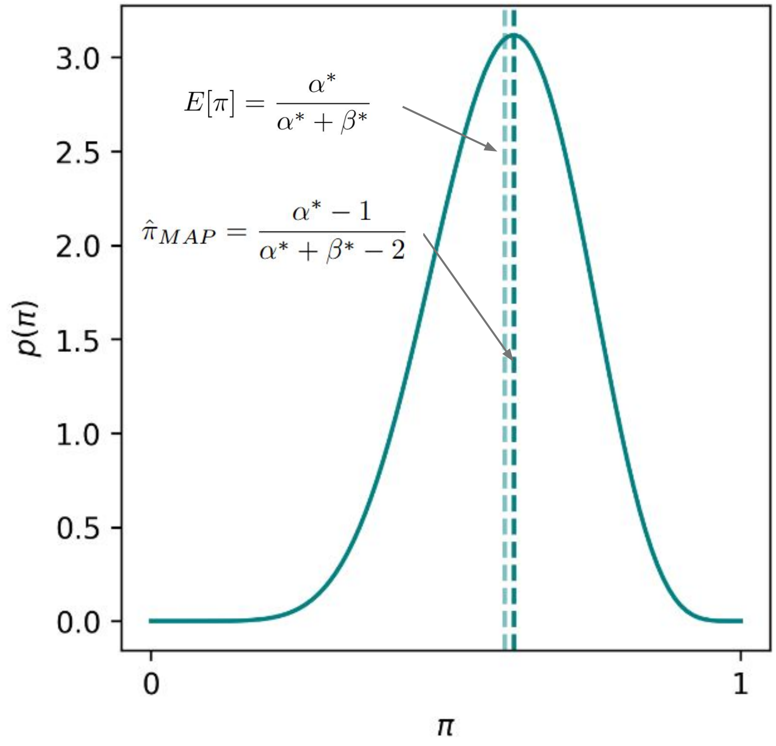
$$\begin{aligned} p(\pi \mid \boldsymbol{y}) &\propto \left(\pi^k (1-\pi)^{(n-k)}\right)\left(\pi^{\alpha-1}(1-\pi)^{\beta-1}\right) \\ &\propto \pi^{k+\alpha-1}(1-\pi)^{n-k+\beta-1} \end{aligned}$$

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



Maximum a posteriori (MAP) estimate

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

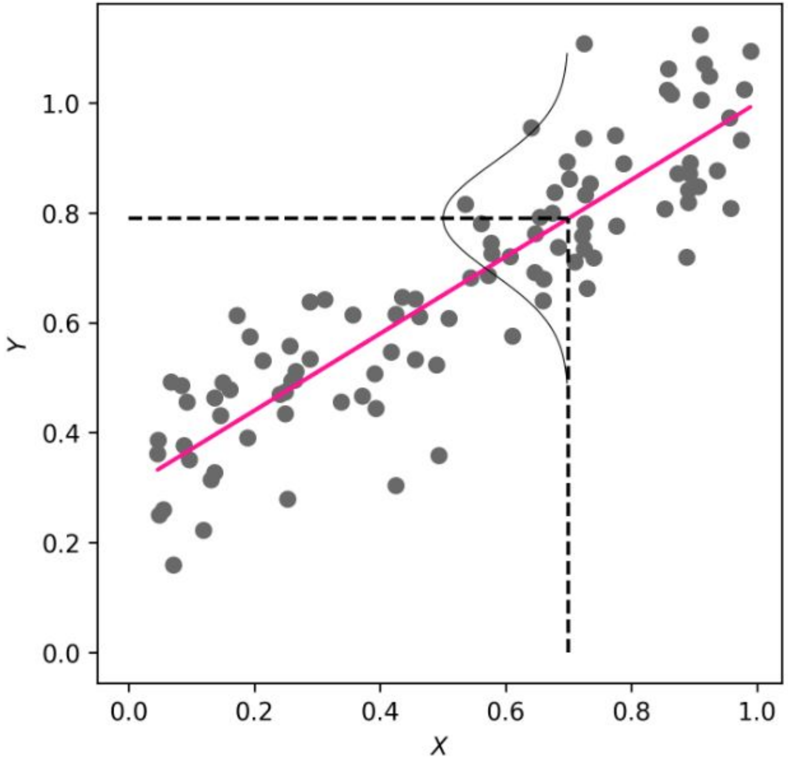


Towards MAP for linear regression: Recall MLE

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

$$\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{x}, w, \sigma^2) &= \mathcal{N}(w^T \boldsymbol{x}, \sigma^2) \\
&= \prod_{i=1}^{n} p(y_i | x_i, w, \sigma^2)
\end{aligned}$$

$$\begin{aligned}
\implies \quad \hat{w} &= \operatorname*{argmax}_w \prod_{i=1}^{n} p(y_i | x_i, w, \sigma^2) \\
&= \operatorname*{argmax}_w \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2} \\
&= \operatorname*{argmax}_w \sum_{i=1}^{n} -(y_i - w^T x_i)^2 \\
&= \operatorname*{argmin}_w \sum_{i=1}^{n} (y_i - w^T x_i)^2
\end{aligned}$$



Incorporating a prior in linear regression with MAP

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

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

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

Incorporating a prior in linear regression with MAP

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

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

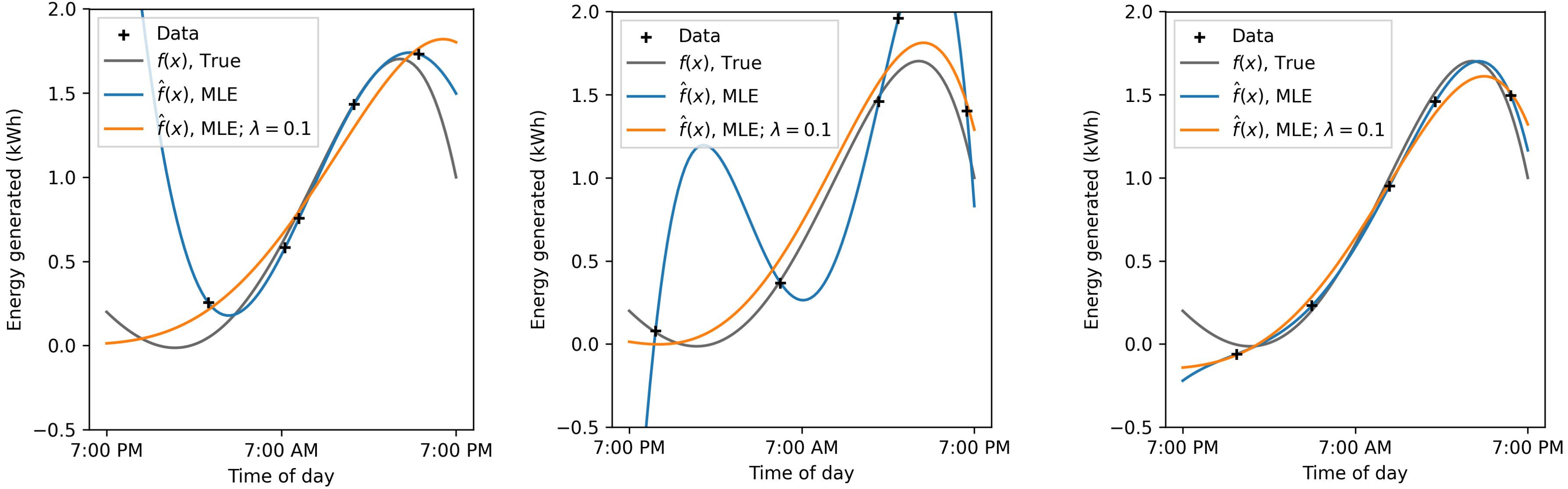
$$= \arg\max_{\boldsymbol{w}} \left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w}\right]$$

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

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

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

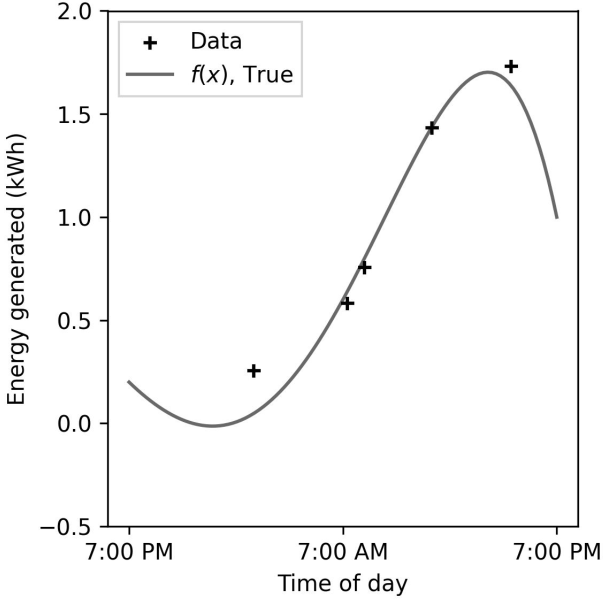
L2-Regularization (Ridge) regression as imposing a prior on $\boldsymbol{w}$

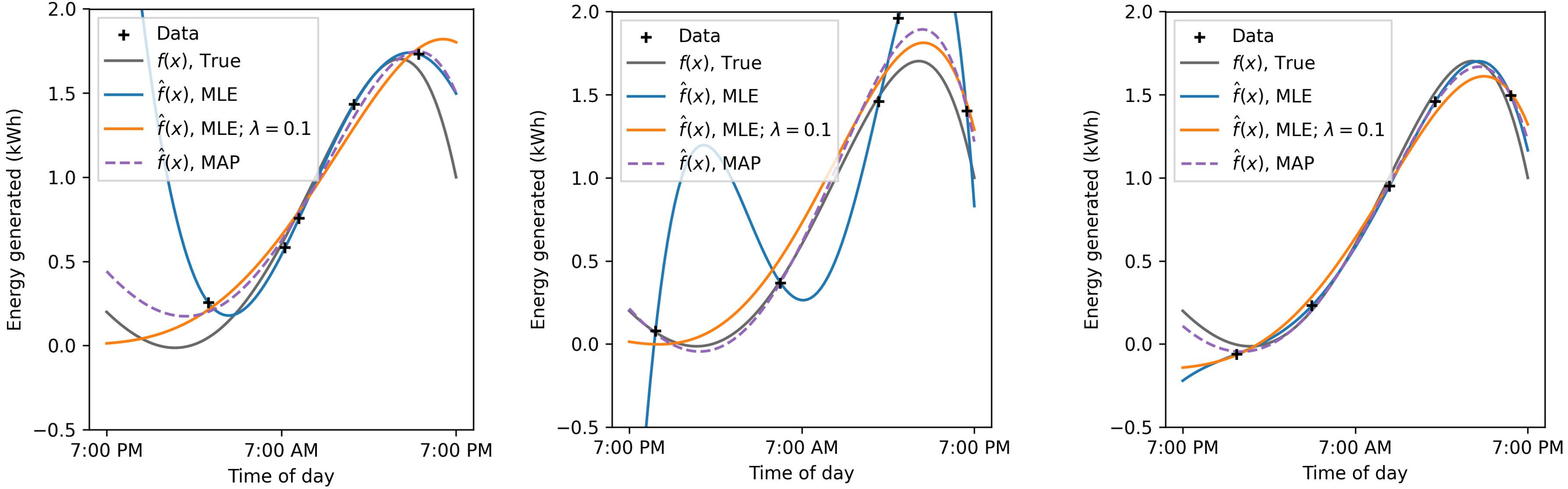


Flexible priors in linear regression with MAP

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma})$$
$$= \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu_0})\right)$$

$$\boldsymbol{\mu_0} = \begin{bmatrix} 0 & -1 & 1. & 0 & 0 \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$




Data
$f(x)$, True
$\hat{f}(x)$, MLE
$\hat{f}(x)$, MLE; $\lambda = 0.1$
$\hat{f}(x)$, MAP
Energy generated (kWh)
Time of day
7:00 PM
7:00 AM
7:00 PM
2.0
1.5
1.0
0.5
0.0
−0.5

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 $\boldsymbol{w}$
- Assumes there exists a true underlying $\boldsymbol{w}$

Bayesian Linear Regression

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

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

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

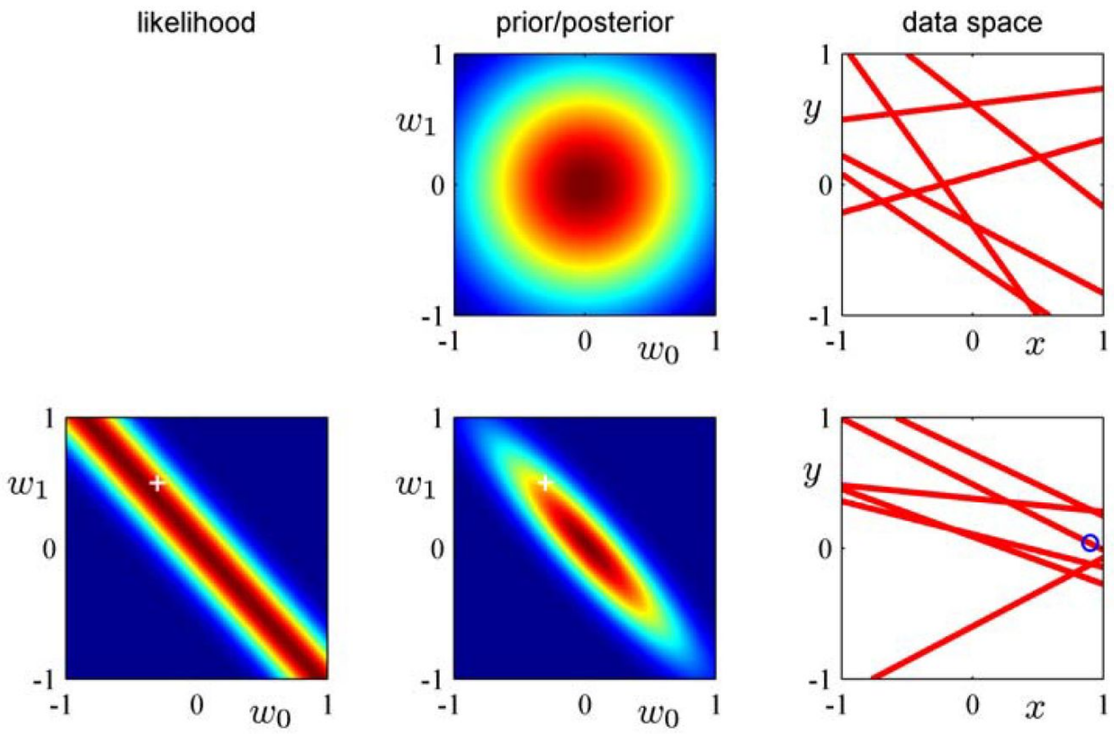
$$
\begin{aligned}
\bar{\boldsymbol{w}} &= \boldsymbol{A}^{-1}\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y} \\
\boldsymbol{A} &= \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \Sigma^{-1}
\end{aligned}
$$

which follows a multivariate normal distribution:

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

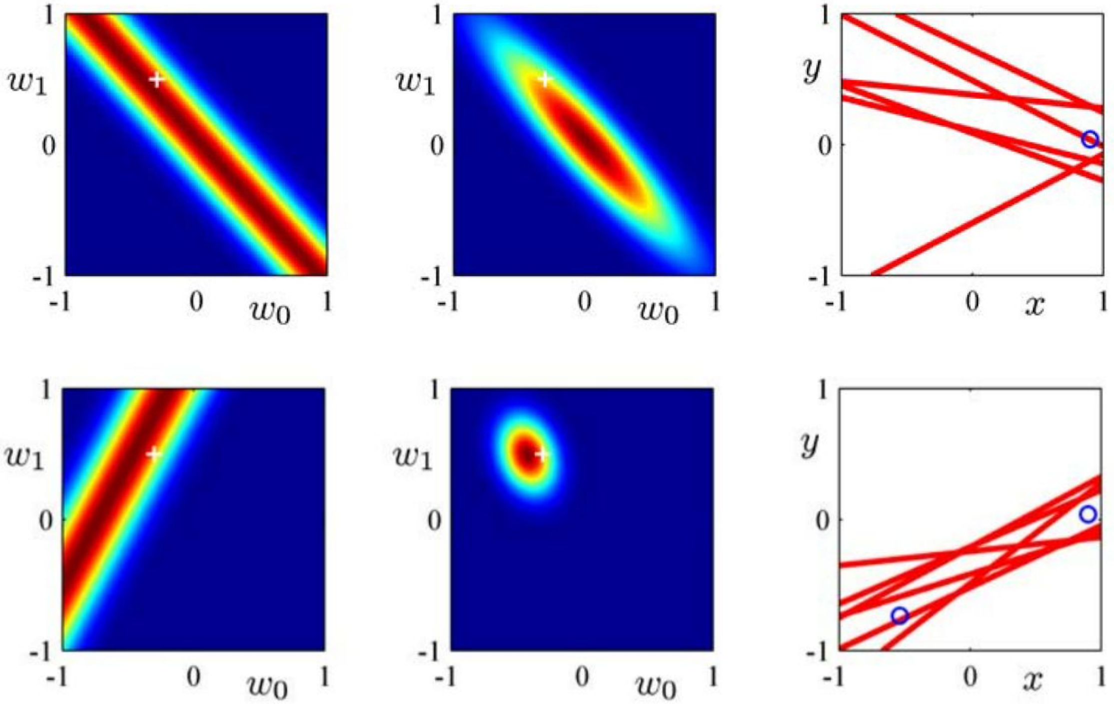
Pattern Recognition and Machine Learning, Section 3.3

Sequential Bayesian update (1/3)



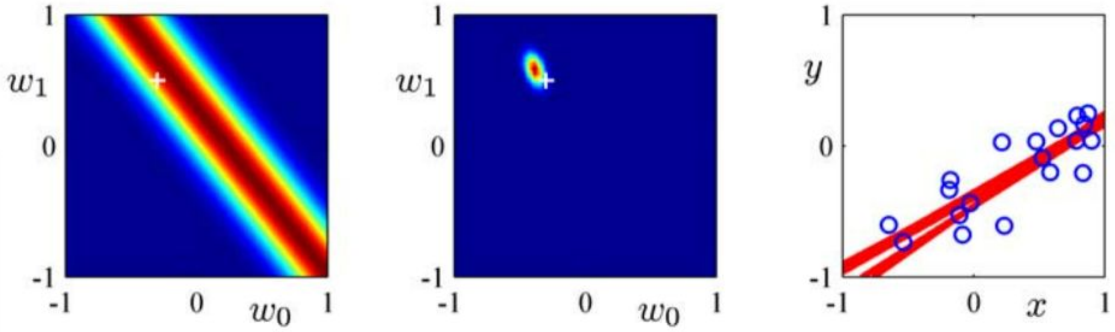
Pattern Recognition and Machine Learning, Section 3.3

Sequential Bayesian update (2/3)



Pattern Recognition and Machine Learning, Section 3.3

Sequential Bayesian update (3/3)



Pattern Recognition and Machine Learning, Section 3.3

Prediction in Bayesian linear regression

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

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

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

Prediction in Bayesian linear regression

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

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

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

Prediction in Bayesian linear regression

$$p(y^* \mid x^*, \boldsymbol{X}, \boldsymbol{y}) = \int_{\boldsymbol{w}} \exp\left(-\frac{(y^* - x^{*T}\boldsymbol{w})^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2}(\boldsymbol{w} - \bar{\boldsymbol{w}})^T \boldsymbol{A}(\boldsymbol{w} - \bar{\boldsymbol{w}})\right) d\boldsymbol{w}$$

$$= \mathcal{N}\left(x^{*T}\bar{\boldsymbol{w}}, \sigma^2 + x^{*T}\boldsymbol{A}^{-1}x^*\right)$$

Errata

- On slide 23, a previous version of the slides referred to $\boldsymbol{A}$ as the covariance matrix of the posterior distribution over the weights, which was incorrect. Matrix $\boldsymbol{A}$ is the precision matrix (inverse of covariance), which now appears correctly on slide 23.