ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 14:
Bayesian Olympics, Marginal Likelihood Model Selection

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

9 October 2017 1

Return (again) to the Olympics 100m

The Bayesian treatment…

- First, the model:

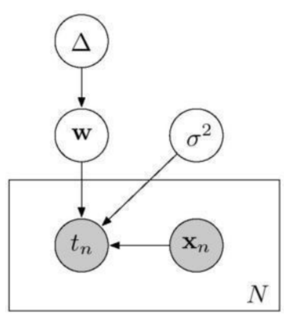
$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_k x_n^k + \epsilon_n$$

k[th]-order polynomial (Ch 1)
Gaussian distributed noise (Ch 2)

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \qquad \mathbf{t} = \mathbf{X}^\top \mathbf{w} + \epsilon$$

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta) p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)} \\ &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}. \\ &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2) p(\mathbf{w}|\Delta)\, d\mathbf{w}} \end{aligned}$$



Predictions, Likelihood & Prior

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)\ d\mathbf{w}}$$

Can use $p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)$ to make predictions:

$$p(t_{new}|\mathbf{x}_{new},\mathbf{X},\mathbf{t},\sigma^2,\Delta) = \int p(t_{new}|\mathbf{x}_{new},\mathbf{w},\sigma^2)p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)\ d\mathbf{w}$$

$$p(t_{new} < 9.5|\mathbf{x}_{new},\mathbf{X},\mathbf{t},\sigma^2,\Delta) = \int p(t_{new} < 9.5|\mathbf{x}_{new},\mathbf{w},\sigma^2)p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)\ d\mathbf{w}$$

The Likelihood:

$$p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}_N)$$ analogous to binomial likelihood in coin example

The Prior:

Want an exact posterior, so want prior that is conjugate to the Gaussian likelihood

$$p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$$ analogous to beta prior in coin example

3

The Posterior

We know that a Gaussian prior over the mean is ***conjugate*** with a Gaussian likelihood, so the posterior is Gaussian!
Our goal is therefore to multiply the two and manipulate the prior and likelihood to get them into a single Gaussian form.

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) = \mathcal{N}(\mu_{\mathbf{w}},\boldsymbol{\Sigma}_{\mathbf{w}})$$

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) &\propto p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)\\
&= \frac{1}{(2\pi)^{N/2}|\sigma^2\mathbf{I}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{t}-\mathbf{X}\mathbf{w})\right)\\
&\quad\times\frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_0|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\mathbf{w})\right)\exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right)\\
&= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\mathbf{w})+(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right)\right\}.\\
&\propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^{\mathsf{T}}\mathbf{X}\mathbf{w}+\frac{1}{\sigma^2}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w}+\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}-2\boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\}
\end{aligned}$$

Ignore any term that doesn't involve **w**

4

The Posterior

$$\propto \exp\left\{-\frac{1}{2}\left(\overset{\text{linear}}{-\frac{2}{\sigma^2}\mathbf{t}^\top\mathbf{X}\mathbf{w}} + \overset{\text{quadratic}}{\frac{1}{\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}} + \overset{\text{quadratic}}{\mathbf{w}^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w}} - \overset{\text{linear}}{2\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w}}\right)\right\}$$

The form we want…

$$\begin{aligned} p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) &= \mathcal{N}(\boldsymbol{\mu}_\mathbf{w},\boldsymbol{\Sigma}_\mathbf{w}) \\ &\propto \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})\right) \\ &\propto \exp\left\{-\frac{1}{2}\left(\underbrace{\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w}}_{\text{quadratic}} - \underbrace{2\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w}}_{\text{linear}}\right)\right\} \end{aligned}$$

(again, only include terms with **w**)

Combine the quadratic terms to isolate $\boldsymbol{\Sigma}_\mathbf{w}$ …

$$\begin{aligned} \mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} &= \frac{1}{\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} + \mathbf{w}^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\ &= \mathbf{w}^\top\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)\mathbf{w} \end{aligned}$$

$$\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

Combine the linear terms to isolate $\boldsymbol{\mu}_\mathbf{w}$ …

$$\begin{aligned} -2\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} &= -\frac{2}{\sigma^2}\mathbf{t}^\top\mathbf{X}\mathbf{w} - 2\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\ \boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} &= \frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X}\mathbf{w} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w} \\ \boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1} &= \frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1} \\ \boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\boldsymbol{\Sigma}_\mathbf{w} &= \left(\frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_\mathbf{w} \\ \boldsymbol{\mu}_\mathbf{w}^\top &= \left(\frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_\mathbf{w} \\ \boldsymbol{\Sigma}_\mathbf{w}^\top &= \boldsymbol{\Sigma}_\mathbf{w} \end{aligned}$$

$$\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

5

The Posterior

In summary:

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w},\boldsymbol{\Sigma}_\mathbf{w})$$

$$\begin{aligned} \boldsymbol{\Sigma}_\mathbf{w} &= \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \\ \boldsymbol{\mu}_\mathbf{w} &= \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) \end{aligned}$$

$$\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + N\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{t}$$

$$\boldsymbol{\mu}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

Given that the posterior is a Gaussian, the single most likely value of **w** is the ***mean*** of the posterior, $\mu_\mathbf{w}$

Since this also happens to be the **mode** (of a Gaussian), it is the **maximum a posteriori** (MAP) estimate of **w** … and is the maximum of the joint posterior probability of **w** and **t**:

$$p(\mathbf{w},\mathbf{t}|\mathbf{X},\sigma^2,\Delta)$$

Recall that the squared loss considered in Chapter 1 is very similar to the Gaussian likelihood.

Computing the most likely posterior (when the prior over the mean is a zero-mean Gaussian) is equivalent to using regularized least squares!

Can help provide intuition about effect of prior: the (inverse) of prior covariance controls the amount of regularization.

6

Consider 1st-Order Polynomial

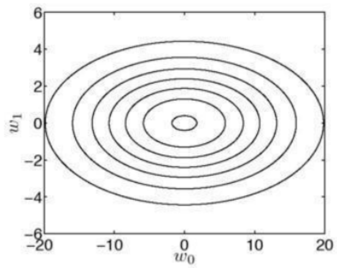
- With a 1st-order polynomial, we can visualize the two-dimensions of the parameters **w**.
- Choose our priors:

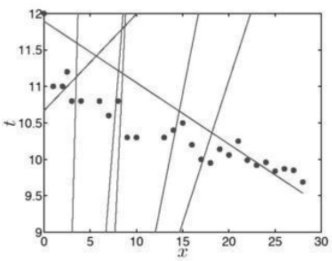
variance for w_0 (intercept term)

$$\mu_\mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{\Sigma_0} = \begin{bmatrix} 100 & 0 \\ 0 & 5 \end{bmatrix}$$

variance for w_1 (slope term)

Zero's indicate prior independence…
Does not preclude posterior dependence!





Can sample **w** from prior

7

Prior and Posterior as we add Data

Posterior over **w**:

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$$

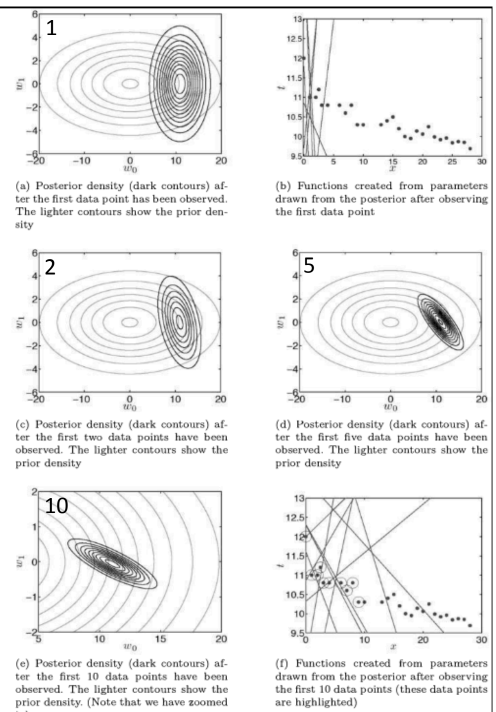
$$\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

assume $\sigma^2 = 10$
(actually pretty high, but for illustration)

1: Lots of info about intercept,
No info about slope

2, 5, 10: Get's more dense!
Starts to tilt: dependency between $\mathbf{w}_0$ and $\mathbf{w}_1$
(based entirely on evidence)



(a) Posterior density (dark contours) after the first data point has been observed. The lighter contours show the prior density

(b) Functions created from parameters drawn from the posterior after observing the first data point

(c) Posterior density (dark contours) after the first two data points have been observed. The lighter contours show the prior density

(d) Posterior density (dark contours) after the first five data points have been observed. The lighter contours show the prior density

(e) Posterior density (dark contours) after the first 10 data points have been observed. The lighter contours show the prior density. (Note that we have zoomed in)

(f) Functions created from parameters drawn from the posterior after observing the first 10 data points (these data points are highlighted)

Prior and Posterior as we add Data

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$$

$$\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

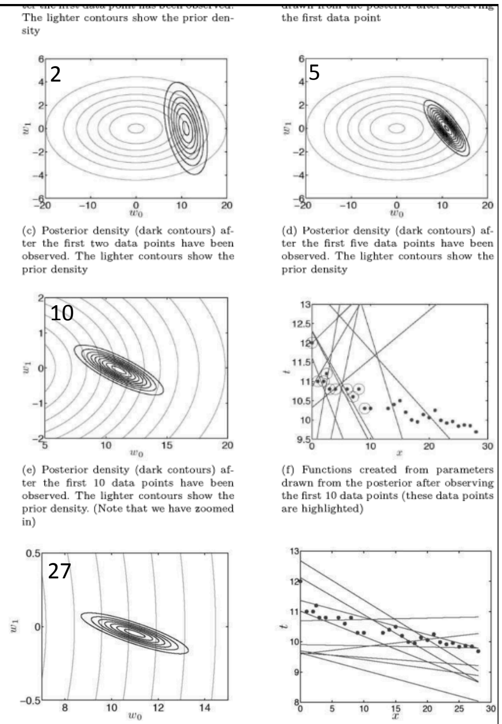
assume $\sigma^2 = 10$

(actually pretty high, but for illustration)

1: Lots of info about intercept,
No info about slope

2, 5, 10: Get's more dense!
Starts to tilt: dependency between $\mathbf{w}_0$ and $\mathbf{w}_1$
(based entirely on evidence)

27: Posterior distribution now
Still lots of variance in sample due to $\sigma^2 = 10$



Prior and Posterior as we add Data

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$$

$$\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

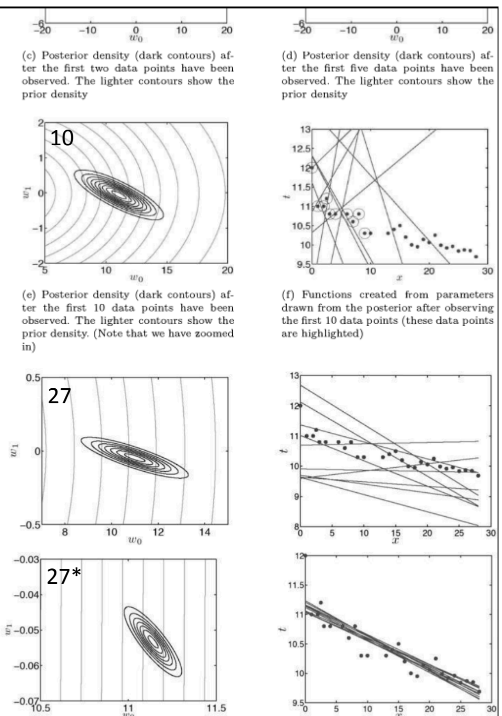
assume $\sigma^2 = 10$

(actually pretty high, but for illustration)

1: Lots of info about intercept,
No info about slope

2, 5, 10: Get's more dense!
Starts to tilt: dependency between $\mathbf{w}_0$ and $\mathbf{w}_1$
(based entirely on evidence)

27*: Very tight posterior distribution
Now adjust to $\sigma^2 = 0.05$



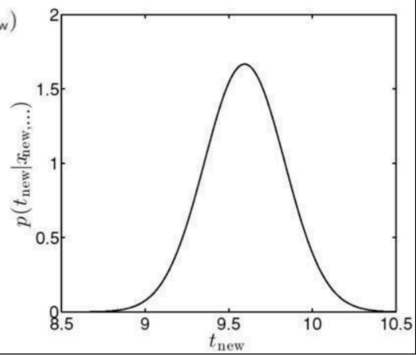
Making Predictions

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathbf{E}_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2)} \left\{ p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) \right\}$$
$$= \int p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2)\ d\mathbf{w}$$

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathsf{T}}\mathbf{w}, \sigma^2)$$

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathcal{N}(\mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\mu}_{\mathbf{w}}, \sigma^2 + \mathbf{x}_{\text{new}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{w}}\mathbf{x}_{\text{new}})$$
$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$
$$\boldsymbol{\mu}_{\mathbf{w}} = \boldsymbol{\Sigma}_{\mathbf{w}}\left(\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathcal{N}(9.5951, 0.0572)$$



Back to Model Selection

- Recall in Chapter 1 we used Cross-Validation to estimate the generalization error of different orders of polynomial model, and selected the model order with the lowest loss.
- We also used Marginal Likelihood to choose among different *prior* densities.
- We can also use Marginal Likelihood to choose *models*

12

Marginal Likelihood for Model Selection

Marginal Likelihood for our Gaussian Model

$$p(\mathbf{t}|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\, d\mathbf{w}$$
$$= \mathcal{N}(\mathbf{X}\boldsymbol{\mu}_0, \sigma^2 \mathbf{I}_N + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^\mathsf{T})$$

Just as in the simulated experiment in Ch 1, generate data from a 3rd-order polynomial

Then compute the marginal likelihood for models from 1st to 7th order

For each model, use Gaussian prior on **w** with zero mean and an identity covariance matrix

$$\boldsymbol{\mu}_0 = [0,0]^\mathsf{T},\ \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

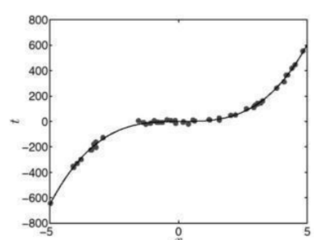
First-order model

$$\boldsymbol{\mu}_0 = [0,0,0,0,0]^\mathsf{T},\ \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

4th-order model

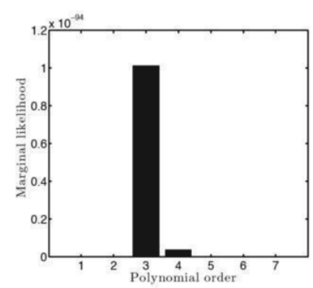
13

Results of Simulation



$$t = 5x^3 - x^2 + x$$

+ Gaussian noise: mean = 0, var = 150



Marginal likelihood for models 1st through 7th order

Plug in relevant prior and evaluate the density at **t**

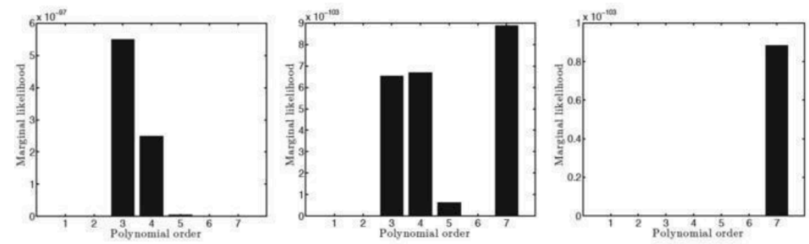
Advantages:
- Very clear peak
- Don't have to compute CV over multiple datasets
- Get to use ***all*** of the data

Disadvantage:
- Calculating marginal likelihood is generally very hard

14

Results Depend on Priors

define $\boldsymbol{\Sigma}_0 = \sigma_0^2 \mathbf{I}$ and vary σ_0^2



(a) $\sigma_0^2 = 0.7$ (b) $\sigma_0^2 = 0.4$ (c) $\sigma_0^2 = 0.3$

By decreasing, we're saying parameters have to take smaller and smaller values

To fit our model well, one of the parameters needs to be 5: $t = 5x^3 - x^2 + x$

By decreasing σ_0^2, 5 becomes less likely and higher order models with lower parameter values become more likely.

When we talk about a **model**, we mean the order of polynomial ***AND*** the prior specification

15