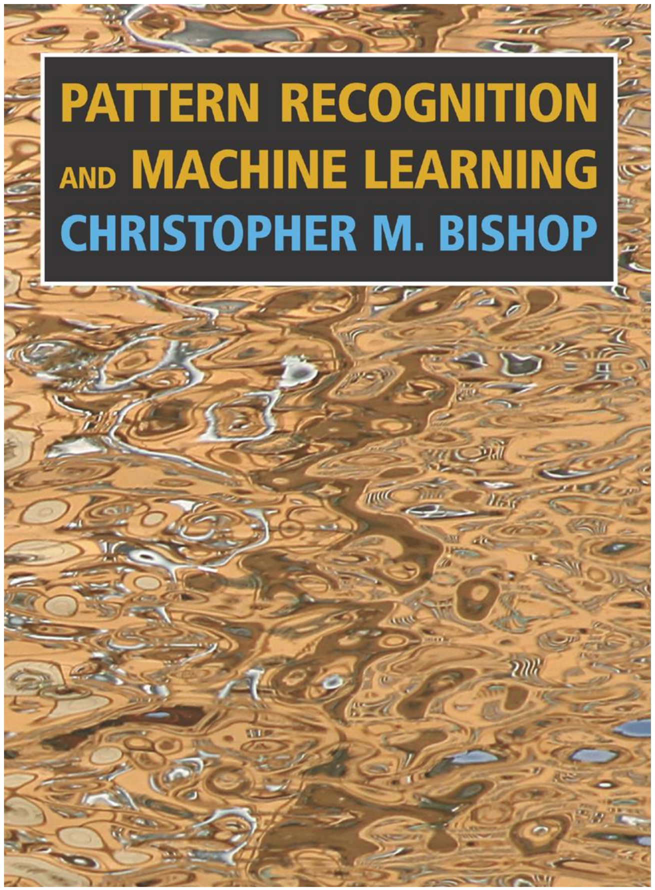
Probabilistic Model Selection in Regression

Song Liu (song.liu@bristol.ac.uk)

Reference



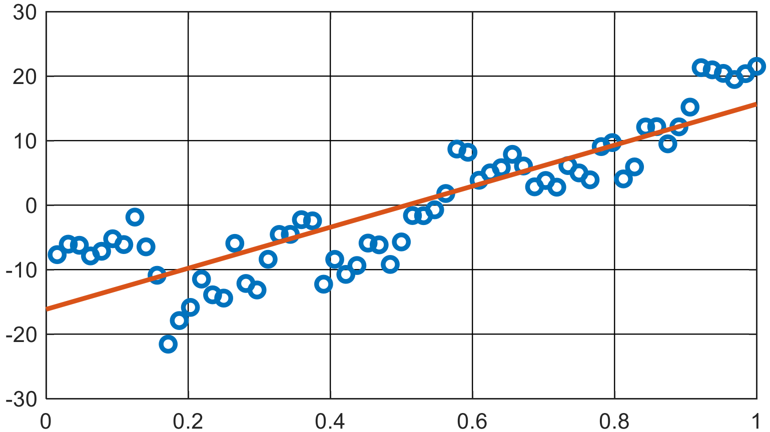
Today's class *roughly* follows Chapter 3.4-3.52.

Pattern Recognition and Machine Learning

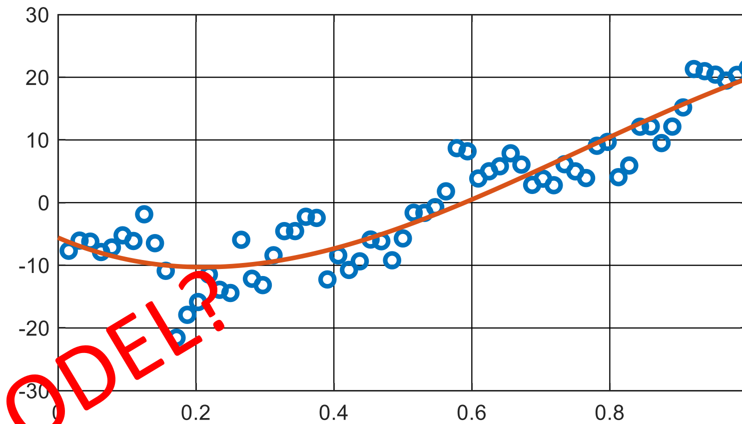
Christopher Bishop, 2006

Apple Stock Price Jul – Sep

Linear Model

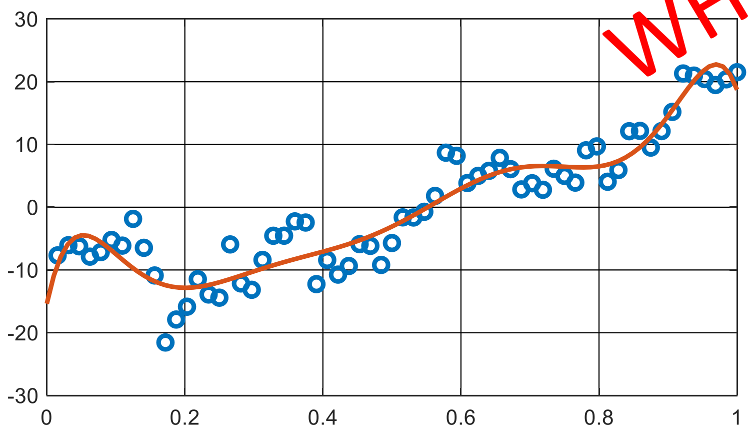


Polynomial

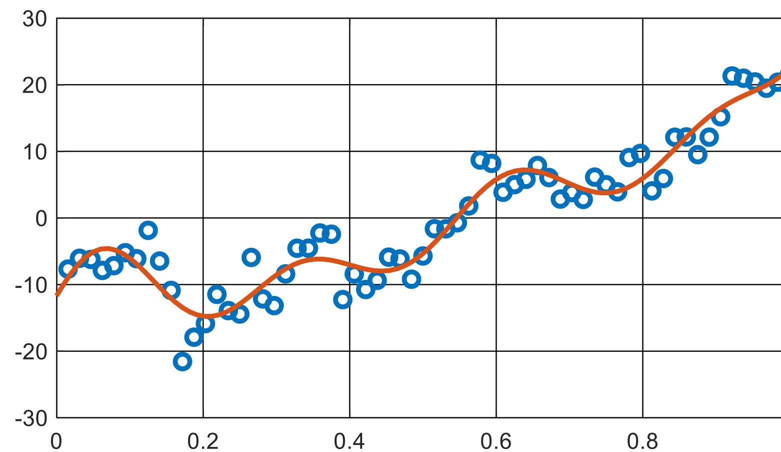


WHICH MODEL?

Trigonometric



RBF basis



Frequentist Model Selection

- We want to minimize the expected (squared) loss:
- $\sum_{i=1..n} \mathbb{E}_D\left[[y - f(\boldsymbol{x}_i; \boldsymbol{w}\,)]^2 | \boldsymbol{x}_i\right]$, over our dataset D.
- $\mathbb{E}_D\left[[y - f(\boldsymbol{x}_i; \boldsymbol{w})]^2 | \boldsymbol{x}_i\right]$ is **minimized**
  - when bias and variance is balanced
- This cannot be done in practice as, $\mathbb{E}_D\left[[y - f(\boldsymbol{x}_i; \boldsymbol{w})]^2 | \boldsymbol{x}_i\right]$ cannot be calculated.
  - We cannot generate different D easily!
- Use out sample error (approximated by testing error):
  - $\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D\left[[y - f(\boldsymbol{x}_i; \boldsymbol{w})]^2 | \boldsymbol{x}\right]\right] \approx \frac{1}{n'} \sum_{i \in D'} [y - f(\boldsymbol{x}_i; \boldsymbol{w}_{\mathrm{LS}})]^2$

Frequentist Model Selection

- There are issues regarding this model selection approach.
- This frequentist approach requires us to hold out sample during training.
    - We lose information in part of our dataset.
    - CV helps, but **calculation is heavy.**
    - Our dataset **may not be IID.**

- How would we select a model if we adopt a probabilistic view?

Probabilistic Model “Selection”

$P(m|D)$

- Build uncertainty of models using **priors over models:**
- Let $m \in \{m_1 \ldots m_K\}$,

1 2 3 . . 5

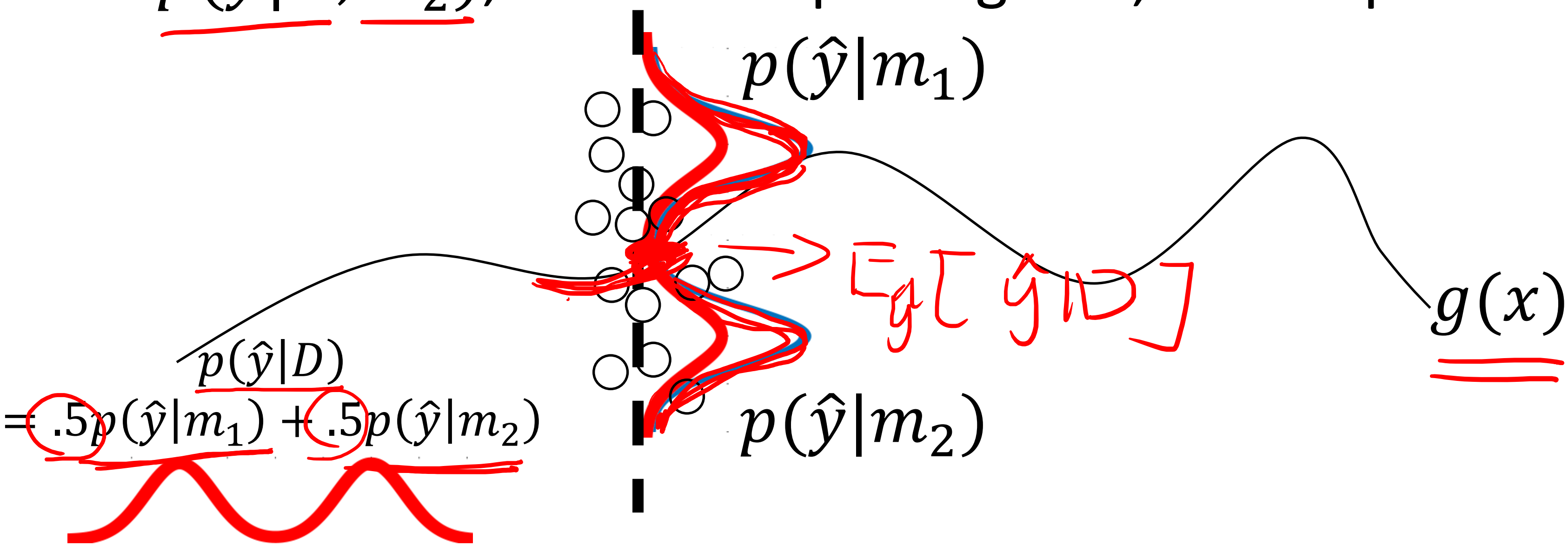
- If we choose $p(m)$ as a model prior.

uniform

- Then we can write **posterior of** model using Bayes rule:
- $\boxed{p(m|D)} \propto p(D|m)p(m)$
- This express the preference over models given D.
- How do we choose a model for prediction?

Probabilistic Model Average

- Bayesians never choose, they marginalize:
- $p(\hat{y}|D) = \sum_{m \in \{\dots\}} p(\hat{y}|D, m)\ p(m|D)$
  - a weighted sum
  - If $p(\hat{y}|D, m_1)$ gives a different prediction than $p(\hat{y}|D, m_2)$, instead of picking one, we keep both.



Probabilistic Model Average

- $p(\hat{y}|D) = \sum_{m\in\{\dots\}} p(\hat{y}|D,m)p(m|D)$
  - Probabilistic model sel.: Using the most probable models given by $p(m|D)$ to approx. $p(\hat{y}|D)$.

- In comparison with frequentist model sel.

$$\hat{m} = \underset{m}{\operatorname{argmin}} \sum_{i=1..n} \mathbb{E}_D\left[[y - f(\boldsymbol{x}_i; \boldsymbol{w}, m)]^2 | \boldsymbol{x}_i\right]$$

- We can see:
- Frequentist minimizes, Bayesian marginalizes.

Probabilistic Model Selection

- $p(\hat{y}|D) = \sum_{m \in \{\dots\}} p(\hat{y}|D, m) p(m|D)$
- How can you calculate $p(m|D)$?

$$\boxed{p(m|D)} \propto \underbrace{p(D|m)}_{model\ evidence} \underbrace{p(m)}_{prior}$$

Model Evidence

- Suppose your model m is governed a set of parameters $\mathbf{w}$
- Then $p(D|m) = \int p(D|\mathbf{w}, m)p(\mathbf{w}|m)d\mathbf{w}$

$m = 2$

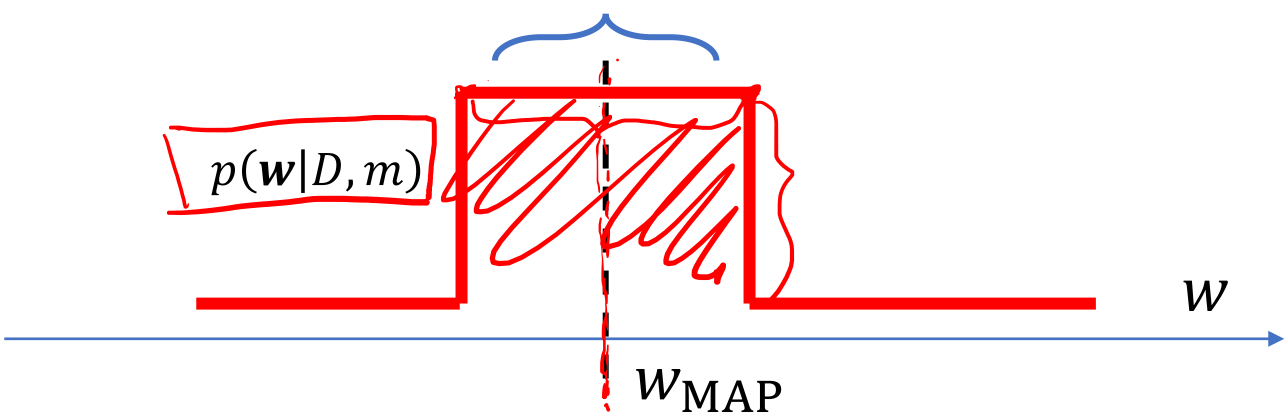
w_1 w_2 w_3

- **Note:** model evidence is the normalizer of para. posterior
- $p(\mathbf{w}|D, m) = \dfrac{p(D|\mathbf{w}, m)p(\mathbf{w}|m)}{p(D|m)}$

Model Evidence Approximation

- Let us consider the simplest approximation of
- $\boxed{p(D|m)} = \int p(D|\mathbf{w}, m)p(\mathbf{w}|m)dw$
- Note: $p(\mathbf{w}|D, m) \propto p(D|\mathbf{w}, m)p(\mathbf{w}|m)$
- Suppose $p(\mathbf{w}|D, m)$ plateaus at w_{MAP}

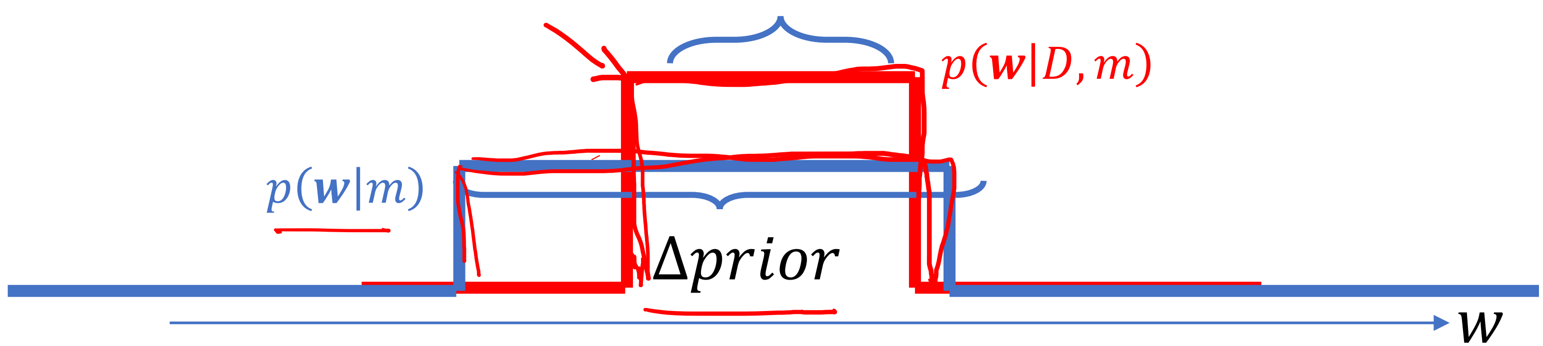
$\Delta posterior$



Model Evidence Approximation

- Then $\int p(D|\mathbf{w}, m)p(\mathbf{w}|m)d\mathbf{w}$
  $$\approx p(D|\mathbf{w}_{\mathrm{MAP}}, m)p(\mathbf{w}_{\mathrm{MAP}}|m) \cdot \Delta posterior$$
- as $\int f(x)\mathrm{d}x \approx f(x_0) \cdot \Delta x$, if f can be approx. by a step function with "length" Δx peaks at x_0
- If $p(\mathbf{w}|m) = \boxed{\frac{1}{\Delta \mathrm{prior}}}$ is flat as well, then
- $p(D|m) \approx p(D|w_{\mathrm{MAP}}, m) \boxed{\frac{\Delta posterior}{\Delta prior}}$

$\boxed{\Delta posterior}$



Model Evidence Approximation

- $\log p(D|m) \approx \log p(D|w_{\text{MAP}}, m) + \log \boxed{\frac{\Delta posterior}{\Delta prior}}$
- As posterior is always sharper than prior, $\boxed{\frac{\Delta posterior}{\Delta prior}} < 1$.

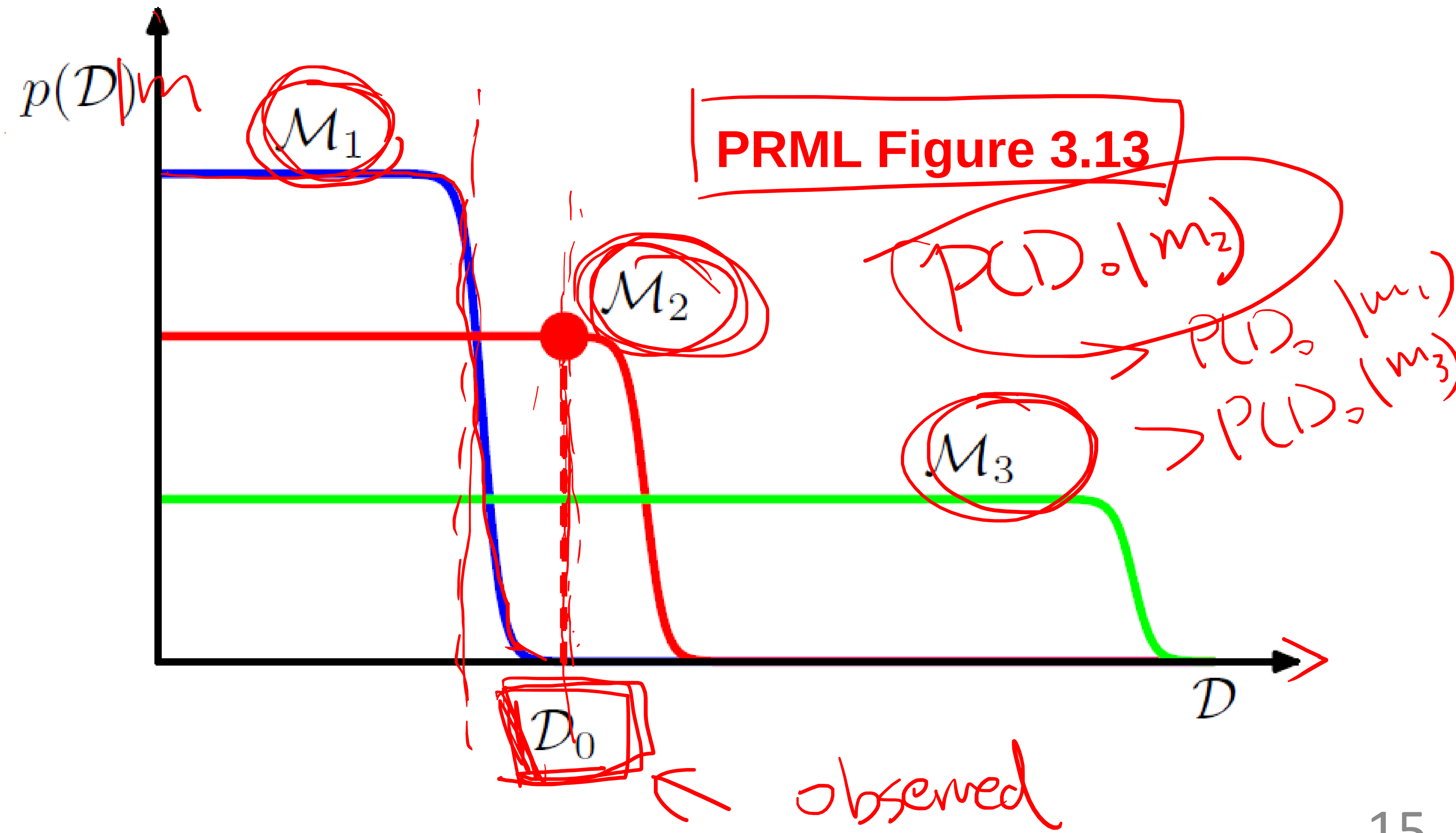
  P(W|D)   P(W)

- The second term is always negative. In fact, **the sharper** our posterior is, **more negative** it is.
- Trade-off is made between $\log p(D|w_{\text{MAP}}, m)$ and $\log \frac{\Delta posterior}{\Delta prior}$

Model Evidence Approximation

- Now, analyze a model with b parameters:
- Assuming $\frac{\Delta posterior}{\Delta prior}$ is the same for all w_i
- $\log p(D|m) \approx \log p(D|w_{\text{MAP}}, m) + b \log \frac{\Delta posterior}{\Delta prior}$
- Why? Prove this.
- If too many parameters in a model, $b \log \frac{\Delta posterior}{\Delta prior}$ decreases!
  - $\log p(D|w_{\text{MAP}}, m)$ increases (why?).
- Model evidence prefers intermediate model complexity.

Model Evidence Prefers Intermediate Model Complexity



Tuning Hyper Parameters

- In most cases, we select a model by selecting a hyper parameter, such as regularization parameter, the degree of the polynomial transform, etc.

λ 6

- Can probabilistic model selection help us determine a hyper parameter?

Tuning Hyper Parameters

λ is hyper Para

P(m|D)

- We would like to calculate the predictive distribution:
- $p(\hat{y}|D) = \int p(\hat{y}|D,\alpha)\boxed{p(\alpha|D)}d\alpha$

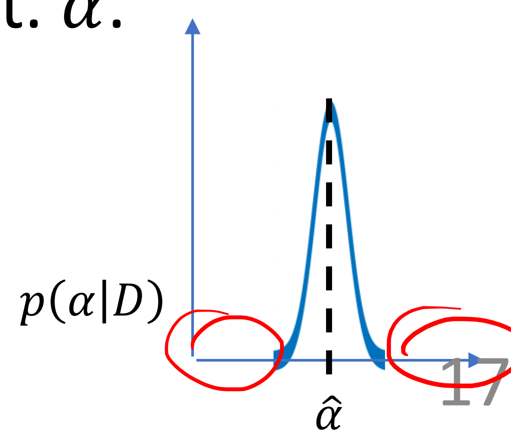
$$= \int\int p(\hat{y}|\mathbf{w},\alpha)p(\mathbf{w}|D,\alpha)\,p(\alpha|D)d\mathbf{w}d\alpha$$

- However, integral w.r.t. α may not be easy ("intractable").

- If $p(\alpha|D)$ is super "pointy" at $\hat{\alpha}$, we only need to use one parameter to approximate the integral w.r.t. α.
- $\int\int p(\hat{y}|\mathbf{w},\alpha)p(\mathbf{w}|D,\alpha)p(\alpha|D)d\mathbf{w}d\alpha \approx$

$$\int p(\hat{y}|\mathbf{w},\hat{\alpha})p(\mathbf{w}|D,\hat{\alpha})d\mathbf{w}$$

α



Model Evidence Approximation with Hyper Parameters

- To find $\hat{\alpha}$ at the peak, we need to maximize $p(\alpha|D)$
- $p(\alpha|D) \propto \underbrace{p(D|\alpha)}_{\text{Model Evidence!}} p(\alpha) = p(\alpha)\int p(D|\mathbf{w},\alpha)p(\mathbf{w}|\alpha)d\mathbf{w}$
- If $p(\alpha)$ is relatively flat, we just
- $\hat{\alpha} := \underset{\alpha}{\operatorname{argmax}} \int p(D|\mathbf{w},\alpha)p(\mathbf{w}|\alpha)d\mathbf{w}$
- "Marginalized Likelihood Maximization"
- Or "Evidence Approximation"

Example: Linear Regression

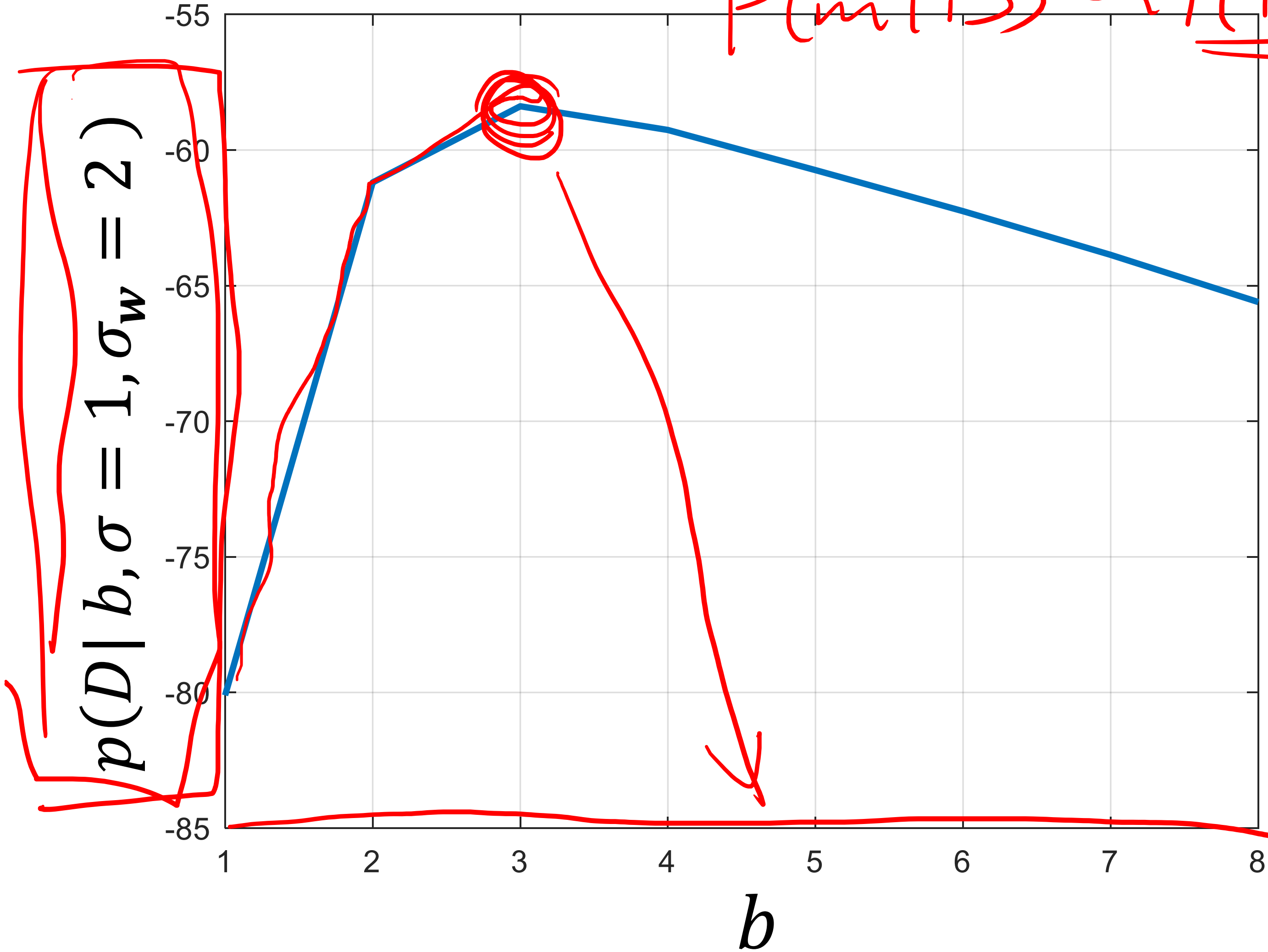
- Suppose we have a likelihood model:
- $p(\boldsymbol{y}_1 \ldots \boldsymbol{y}_n | \boldsymbol{x}_1 \ldots \boldsymbol{x}_n; \boldsymbol{w}, b) := \prod_{i \in D} N_{y_i}(\langle \boldsymbol{w}, \boldsymbol{\phi}_b(\boldsymbol{x}_i) \rangle, \sigma^2 \boldsymbol{I})$
- $p(\boldsymbol{w}) := N_{\boldsymbol{w}}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$

- Marginalized Likelihood
- $p(\boldsymbol{y}_1 \ldots \boldsymbol{y}_n | \boldsymbol{x}_1 \ldots \boldsymbol{x}_n;\ b, \sigma, \sigma_{\boldsymbol{w}})$
  $= \int p(\boldsymbol{y}_1 \ldots \boldsymbol{y}_n | \boldsymbol{x}_1 \ldots \boldsymbol{x}_n; \boldsymbol{w}, b, \sigma, \sigma_{\boldsymbol{w}}) p(\boldsymbol{w}) d\boldsymbol{w}$
  $\boxed{= N_{\boldsymbol{y}}(\boldsymbol{0}, \boldsymbol{\sigma}_w^2 \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \sigma^2 \boldsymbol{I})}$

hint: use Gaussian identity!

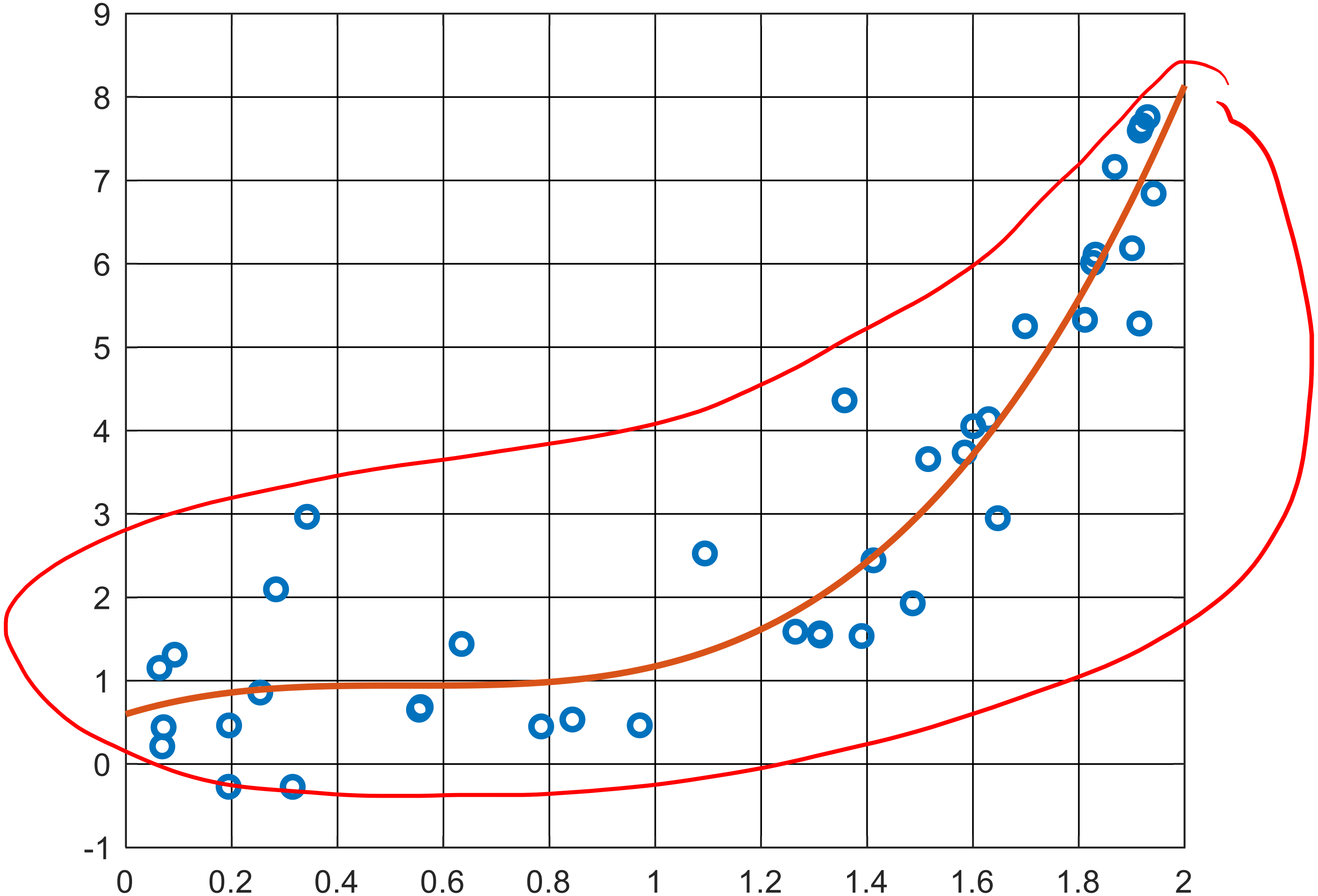
Example: Linear Regression

$P(m \mid D) \propto \underline{P(D \mid m)} \cdot P(m)$



Example: Linear Regression

$$f(x; w_{\mathrm{LS}}), \; b = 3$$



Conclusion

- We introduced probabilistic model selection.
- The principle: Integrate over models w.r.t. model posterior.
- $p(m|D) \propto p(D|m)p(m)$
- Approximation using flat posterior and prior of $\boldsymbol{w}$.
  - $p(D|m)$ decreases as b increase.
- Approximation using marginalized likelihood.
  - Allows us to select hyper-parmeters

Homework

- Prove statement on page 14.
- Prove statement on page 19.

- Read PRML 3.52

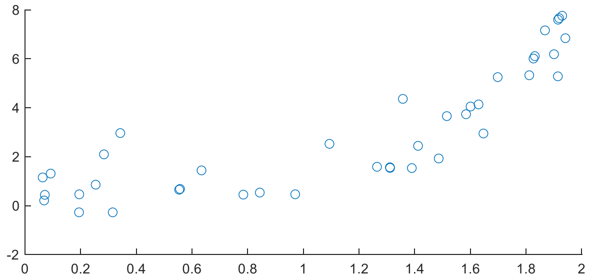
Computing Lab

- Implementing least square regression with different choices of kernels:
- Linear kernel,
- Polynomial kernel,
- RBF kernel.

- Apply it on prostate cancer dataset. What choice of kernel/kernel parameters minimizes the CV error?

Computing Lab

- Generate, $x \sim U(0,2)$, $y =$ exp(1.5x−1) $+\ \epsilon, \epsilon \sim N(0,1)$,



- Select number of basis using marginalized likelihood for different basis:
- Polynomial basis
- Trigonometric basis
- RBF basis

(Fix σ and $\sigma_{\boldsymbol{w}}$)