THE UNIVERSITY OF SYDNEY

Advanced Machine Learning

(COMP 5328)

Week 8 Tutorial:
Learning with Noisy Data: On the Robustness of Surrogate Loss Functions

Anjin Liu
anjin.liu@sydney.edu.au

https://github.com/Anjin-Liu/usyd_comp_5328_tutorial_S22025

Tutorial Contents

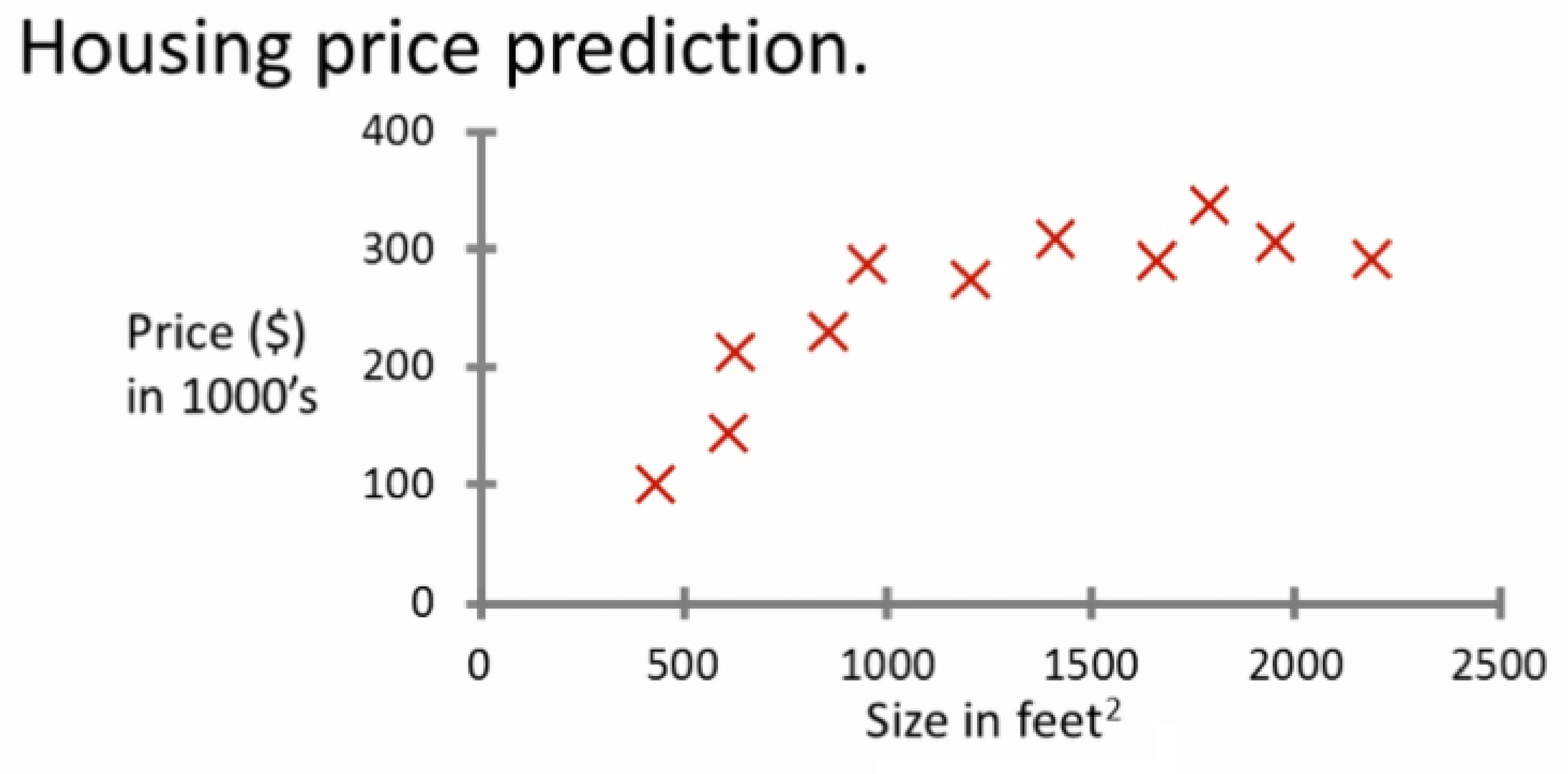
THE UNIVERSITY OF SYDNEY

- Review (20min):
  - Lecture 6: Learning with Noisy Data: On the Robustness of Surrogate Loss Functions
- Tutorial exercise & QA (40min):

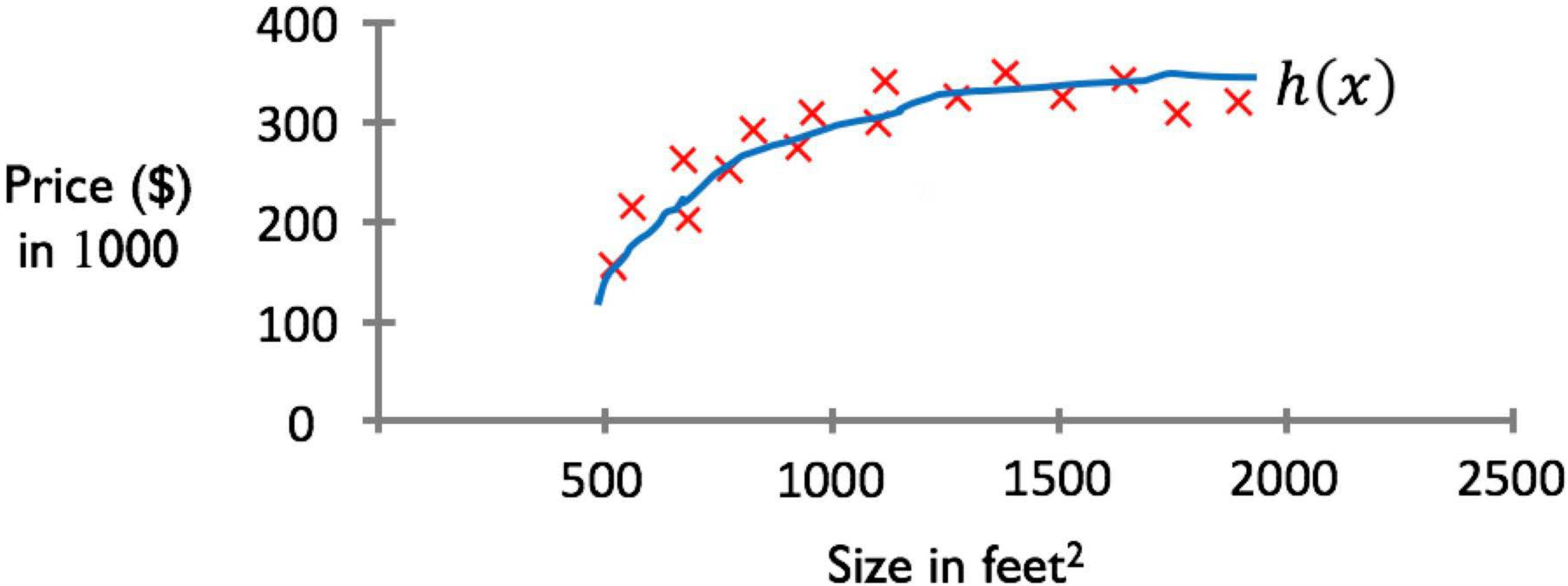
Key points

- Linear Regression with Gaussian Noise
- Maximum Likelihood Estimation (MLE)
- Maximum A Posterior (MAP)
- Bias and variance
- Robustness of surrogate loss functions

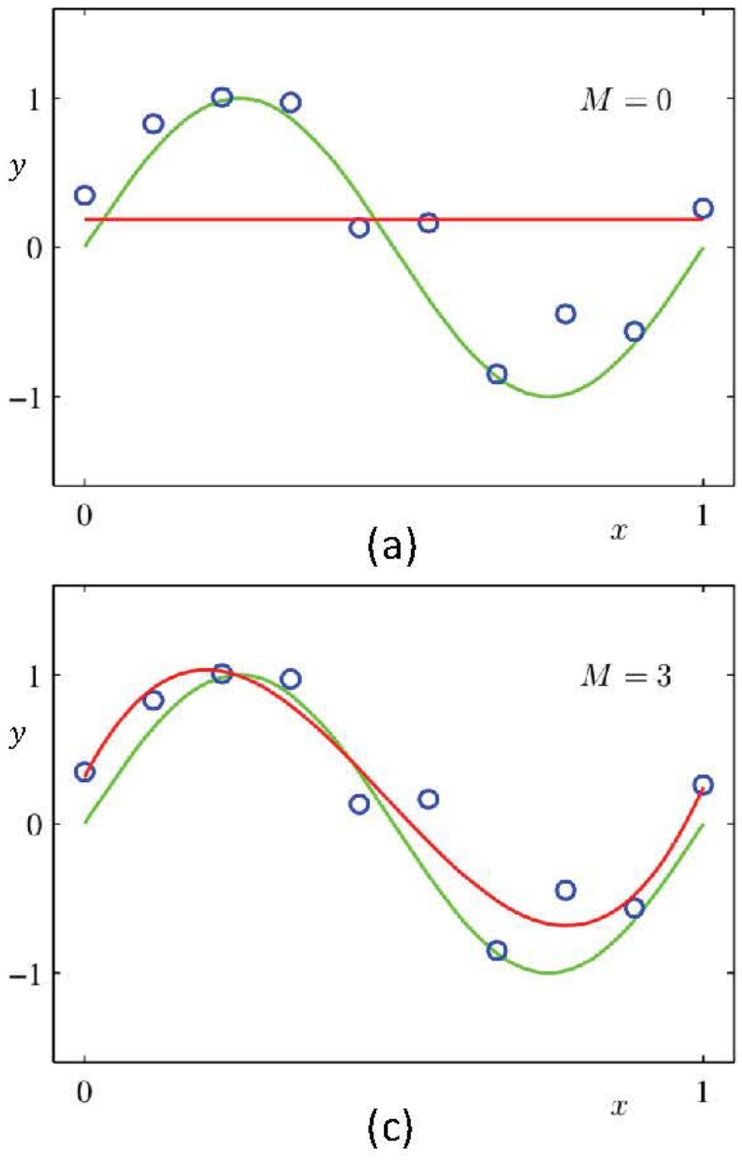
Linear regression



Linear regression



Linear regression



- True target: $\sin(2\pi x)$ with small Gaussian noises.
- $h(x) = w_0 + w_1 x + \cdots + w_M x^M$
- $R_S(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h(x_i)\right)^2$

$$\theta = \{w_0, \ldots, w_M\}$$

Bishop's book: "Pattern Recognition and Machine Learning"

Bayes' rule

THE UNIVERSITY OF SYDNEY

$$p(\theta|S) = \frac{p(S|\theta)p(\theta)}{p(S)}$$

Likelihood, prior, and posterior

THE UNIVERSITY OF SYDNEY

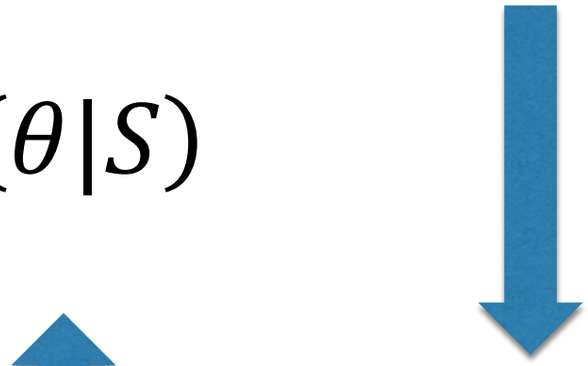
- **Likelihood** of θ is the probability of observing the data given θ. Given the data sample S, we denote the likelihood as $p(S|\theta)$.

- **Prior** of θ is a distribution that describes any prior beliefs of θ. We denote the prior as $p(\theta)$.

- **Posterior** of θ is proportional to the likelihood times the prior. We denote the posterior as $p(\theta|S)$.

$$\theta = \{w_0, \ldots, w_M\}$$

Posterior $P(\theta|S)$



Solution 1

Solution 2

$P(S|\theta)$ Likelihood

$$S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

Maximum Likelihood Estimation (MLE)

THE UNIVERSITY OF SYDNEY

According to the i.i.d. assumption the likelihood function is rewritten as

$$p(S|\theta) = \prod_{i=1}^{n} p(x_i, y_i|\theta).$$

Sometimes, we also define the likelihood as follows

$$p(S|\theta) = \prod_{i=1}^{n} p(y_i|x_i, \theta).$$

Maximum Likelihood: Find the value of θ maximising the likelihood $p(S|\theta)$, i.e., it is the value of θ that makes the observed data the "most probable".

Maximum A Posterior (MAP)

THE UNIVERSITY OF SYDNEY

Bayes' rule

$$P(\theta|S) = \frac{P(S|\theta)P(\theta)}{P(S)}$$

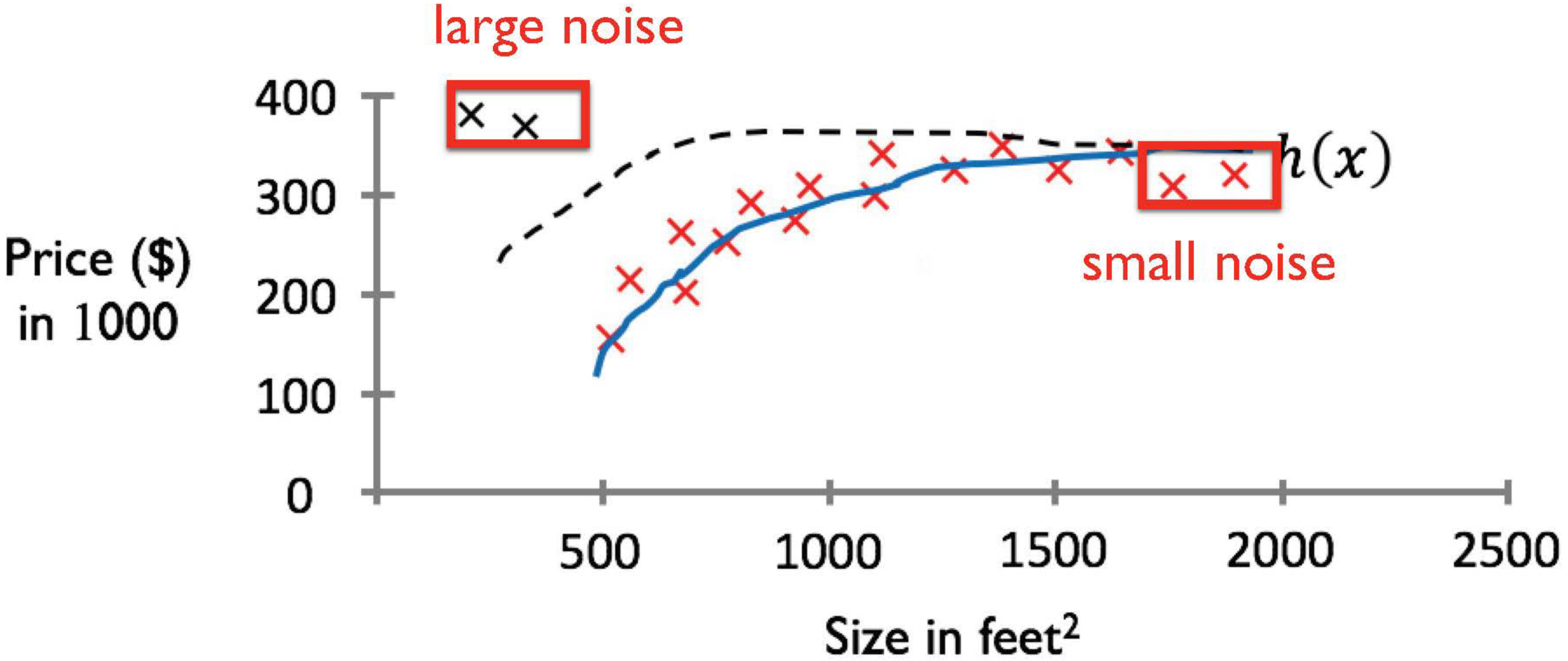
$$P(\theta|S) \propto P(S|\theta)P(\theta)$$

$$\arg\max_{\theta} P(\theta|S) = \arg\max_{\theta} P(S|\theta)P(\theta)$$

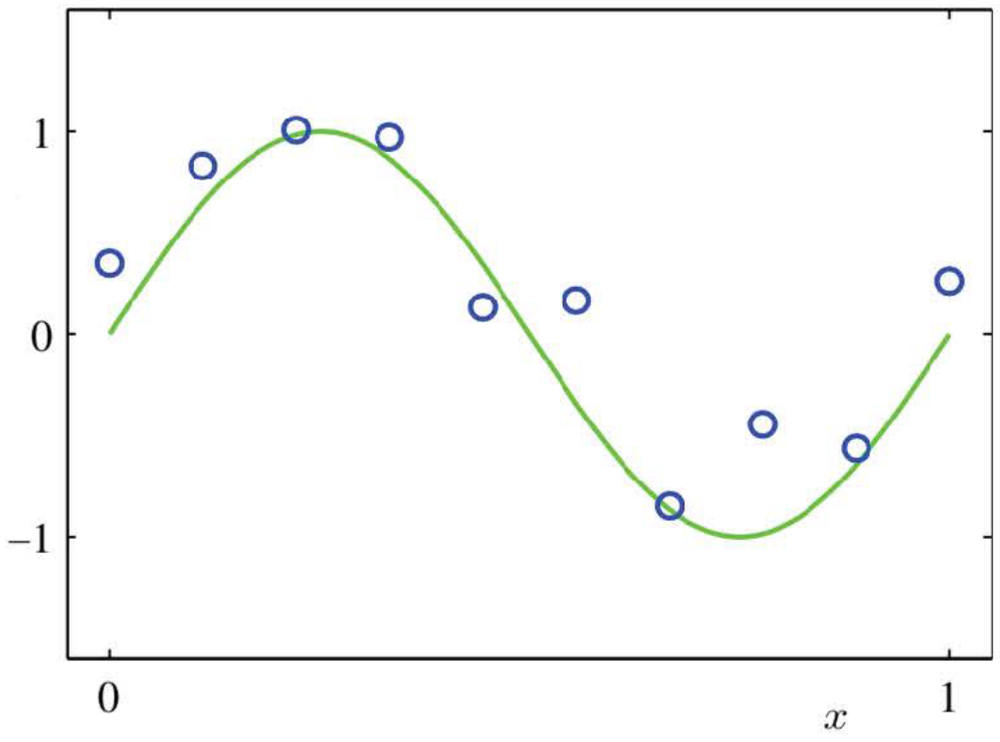
$$\arg\min_{\theta}(-\log P(\theta|S)) = \arg\min_{\theta}(-\log P(S|\theta) - \log P(\theta))$$

Maximum Posterior $P(\theta|S)$: the observed data makes the value of θ the "most probable".

Small noise vs large noise



Linear regression with Gaussian noise

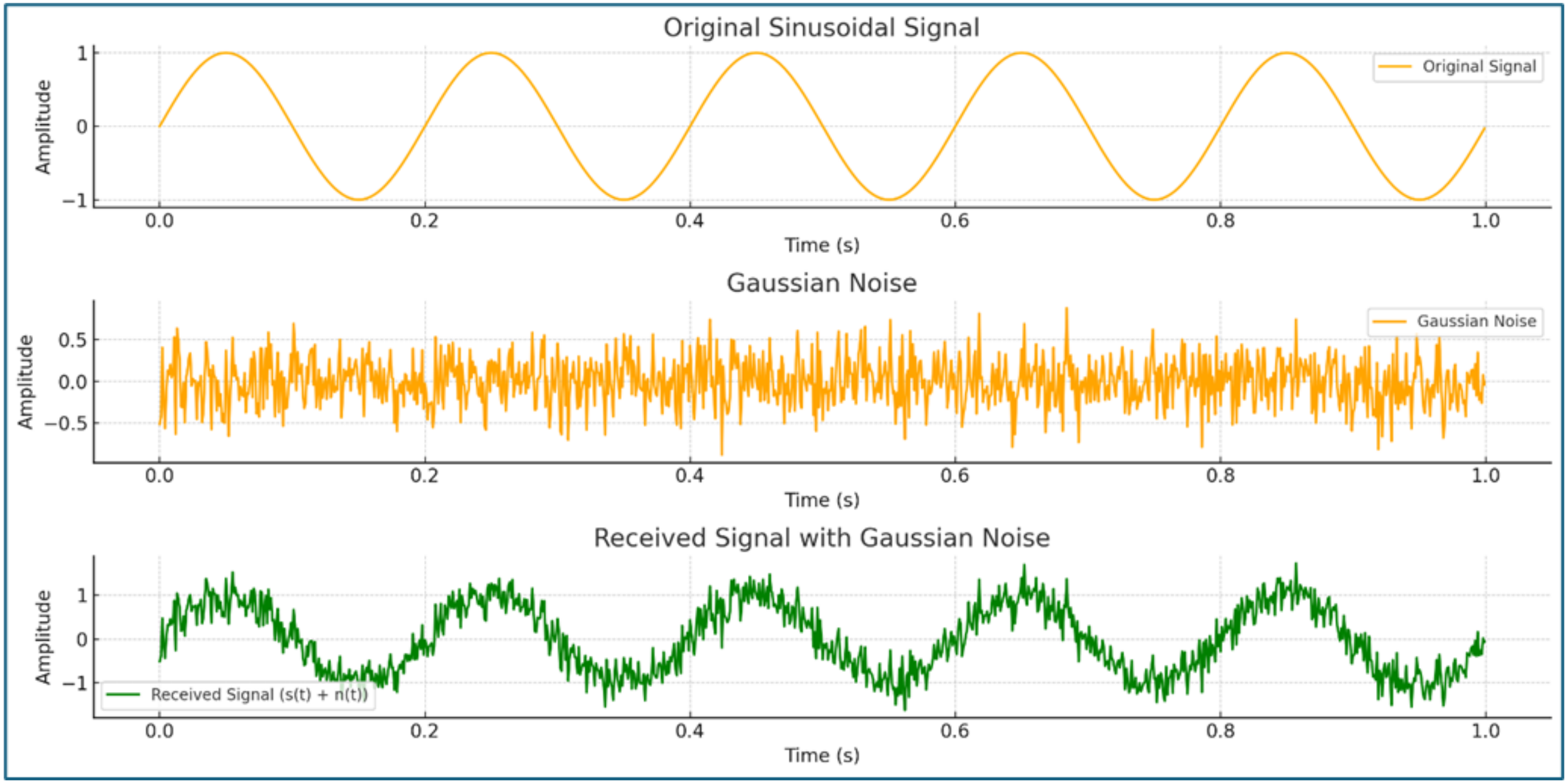


- $y = h(x) + \epsilon.$
- $\epsilon \sim \mathcal{N}(0, \beta^{-1})$
- $y|x \sim \mathcal{N}(h(x), \beta^{-1})$

Bishop's book: "Pattern Recognition and Machine Learning"

$$h(x) = \sin(x)$$
$$y = h(x) + \epsilon$$
$$\epsilon \sim N(0, \beta^{-1})$$
$$y|x \sim N(h(x), \beta^{-1})$$

THE UNIVERSITY OF SYDNEY



Modelling Noisy Observations

Lets assume data is from a deterministic function with additive Gaussian noise.

$$y = h(x) + \epsilon \qquad \text{with} \qquad \epsilon \sim \mathcal{N}(0, \beta^{-1})$$

Equivalently,

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

Given the training data: $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

The expression of the likelihood the model is:

$$p(S|X, h, \beta^{-1}) = \prod_{i=1}^{n} \mathcal{N}(y_i|h(x_i), \beta^{-1})$$

Maximum Likelihood

THE UNIVERSITY OF SYDNEY

$$p(S|X, h, \beta^{-1}) = \prod_{i=1}^{n} \mathcal{N}(y_i|h(x_i), \beta^{-1})$$

PDF $\quad \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$= \prod_{i=1}^{n} \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta(y_i - h(x_i))^2}{2}\right)$$

$$= \left(\frac{\beta}{2\pi}\right)^{n/2} \prod_{i=1}^{n} \exp\left(-\frac{\beta(y_i - h(x_i))^2}{2}\right)$$

$$-\ln(\cdot) = -\ln(\cdot)$$

$$-\ln p(S|X, h, \beta^{-1}) = -\frac{n}{2}\ln\beta + \frac{n}{2}\ln(2\pi) + \frac{\beta}{2}\sum_{i=1}^{n}(y_i - h(x_i))^2$$

$$= -\frac{n}{2}\ln\beta + \frac{n}{2}\ln(2\pi) + \frac{\beta}{2} n R_S(h)$$

Maximum A Posterior (MAP)

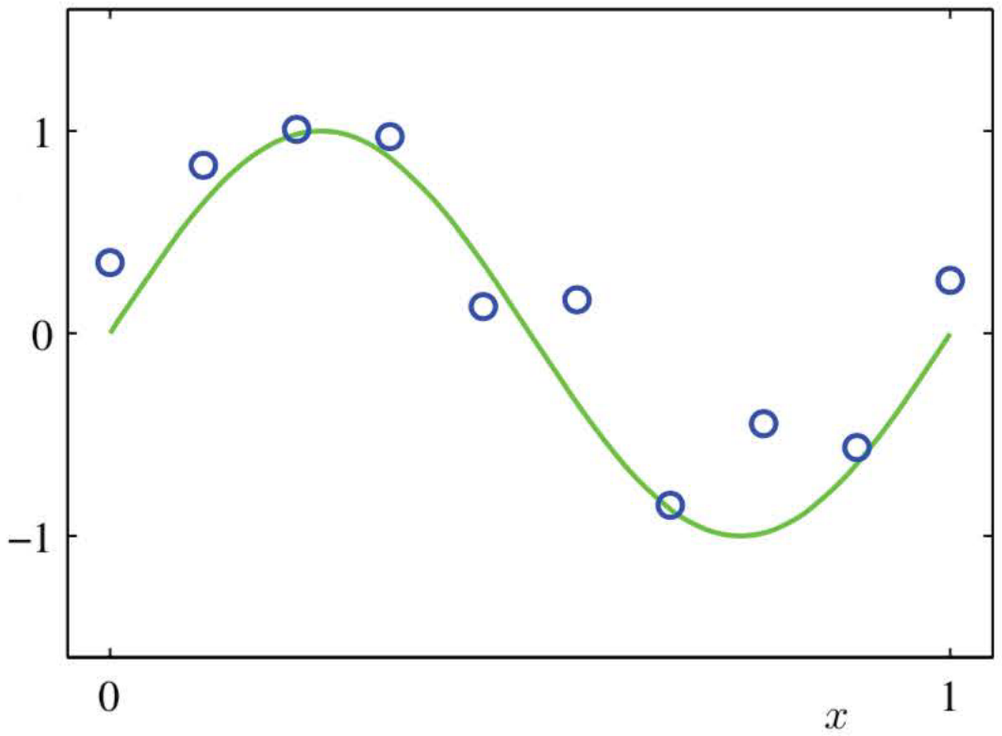
THE UNIVERSITY OF SYDNEY

Because of Bayes' rule

$$\arg\max_{\theta} p(\theta|S) = \arg\max_{\theta} p(S|\theta)p(\theta)$$

$$-\ln(\cdot) = -\ln(\cdot)$$

$$\begin{aligned}
\arg\min_{h}(-\ln p(h|S,\beta^{-1})) &= \arg\min_{h}(-\ln(p(S|X,h,\beta^{-1})p(h))) \\
&= \arg\min_{h}(-\ln p(S|X,h,\beta^{-1}) - \ln p(h)) \\
&= \arg\min_{h}(-\frac{n}{2}\ln\beta + \frac{n}{2}\ln(2\pi) + \frac{\beta}{2}nR_S(h) - \ln p(h))
\end{aligned}$$



$$h(x) = w_0 + w_1 x + \ldots + w_9 x^9$$

Assuming the prior distribution:

$$p(h) = \prod_{i=0}^{9} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau w_i^2}{2}\right)$$

Then, we have

$$\arg\min_h(-\ln p(h|S, \beta^{-1})) = \arg\min_h(-\frac{n}{2}\ln\beta + \frac{n}{2}\ln(2\pi) + \frac{\beta}{2} n R_S(h)$$

$$- 5\ln\tau + 5\ln(2\pi) + \frac{\tau}{2} n \sum_{i=0}^{9} w_i^2)$$

Minimising above equals

$$\min R_S(h) + \lambda \sum_{i=0}^{9} w_i^2 = R_S(h) + \lambda \|w\|_2^2, \quad \lambda = \frac{\tau}{\beta}$$

Bishop's book: "Pattern Recognition and Machine Learning"

ridge regression (L2 penalty)

Bayesian Linear Regression

The full BLR model =

(linear hypothesis class) +

(Gaussian likelihood for noise) +

(Gaussian prior for weights)

$$\theta = \{w_0, \dots, w_M\}$$

Laplacian regression (Least absolute deviation)

- Assumed noise: $\epsilon = Y - h(X)$

Laplacian distribution: $p(\epsilon|X, Y, h, \sigma) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}|\epsilon|}{\sigma}\right)$

Likelihood for one example:

$$p(y_i|x_i, h, b) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}|y_i - h(x_i)|}{\sigma}\right)$$

Laplacian regression (Least absolute deviation)

THE UNIVERSITY OF SYDNEY

- Assumed noise: $\epsilon = Y - h(X)$

Laplacian distribution: $p(\epsilon|X, Y, h, \sigma) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}|\epsilon|}{\sigma}\right)$

The likelihood function is

$$p(S|X, h, b) = \left(\frac{1}{\sqrt{2}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{\sqrt{2}|y_i - h(x_i)|}{\sigma}\right)$$

The negative log-likelihood function is

$$-\ln p(S|X, h, b) = n \ln(\sqrt{2}\sigma) + \frac{\sqrt{2}}{\sigma} \boxed{\sum_{i=1}^{n} |y_i - h(x_i)|}$$

Cauchy regression

THE UNIVERSITY OF SYDNEY

- Assumed noise: $\epsilon = Y - h(X)$

Cauchy distribution: $p(\epsilon|X, Y, h, \gamma) = \frac{1}{\pi\gamma\left(1 + \left(\frac{\epsilon}{\gamma}\right)^2\right)}$

Likelihood for one example:

$$p(y_i|x_i, h, \gamma) = \frac{1}{\pi\gamma\left(1 + \left(\frac{y_i - h(x_i)}{\gamma}\right)^2\right)}$$

Cauchy regression

THE UNIVERSITY OF SYDNEY

- Assumed noise: $\epsilon = Y - h(X)$

Cauchy distribution: $p(\epsilon|X, Y, h, \gamma) = \dfrac{1}{\pi\gamma\left(1 + \left(\frac{\epsilon}{\gamma}\right)^2\right)}$

The likelihood function is

$$p(S|X, h, \gamma) = \left(\frac{1}{\pi\gamma}\right)^n \prod_{i=1}^{n} \frac{1}{1 + \left(\frac{y_i - h(x_i)}{\gamma}\right)^2}$$

The negative log-likelihood function is

$$-\ln p(S|X, h, \gamma) = n\ln(\pi\gamma) + \sum_{i=1}^{n} \ln\left(1 + \left(\frac{y_i - h(x_i)}{\gamma}\right)^2\right)$$

Heavy-tailed shrinkage

THE UNIVERSITY OF SYDNEY

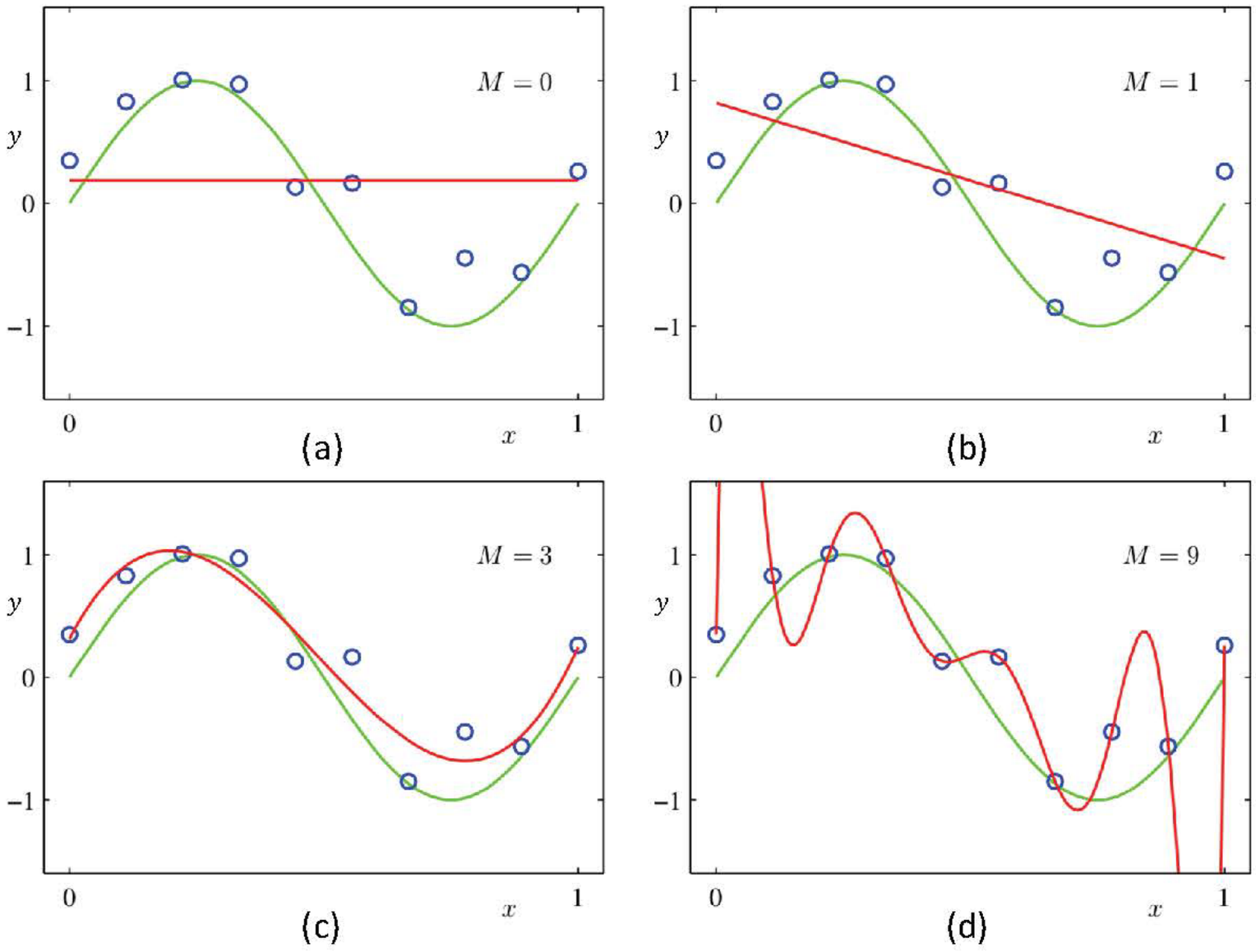
Bias and variance

Underfitting and

THE UNIVERSITY OF SYDNEY

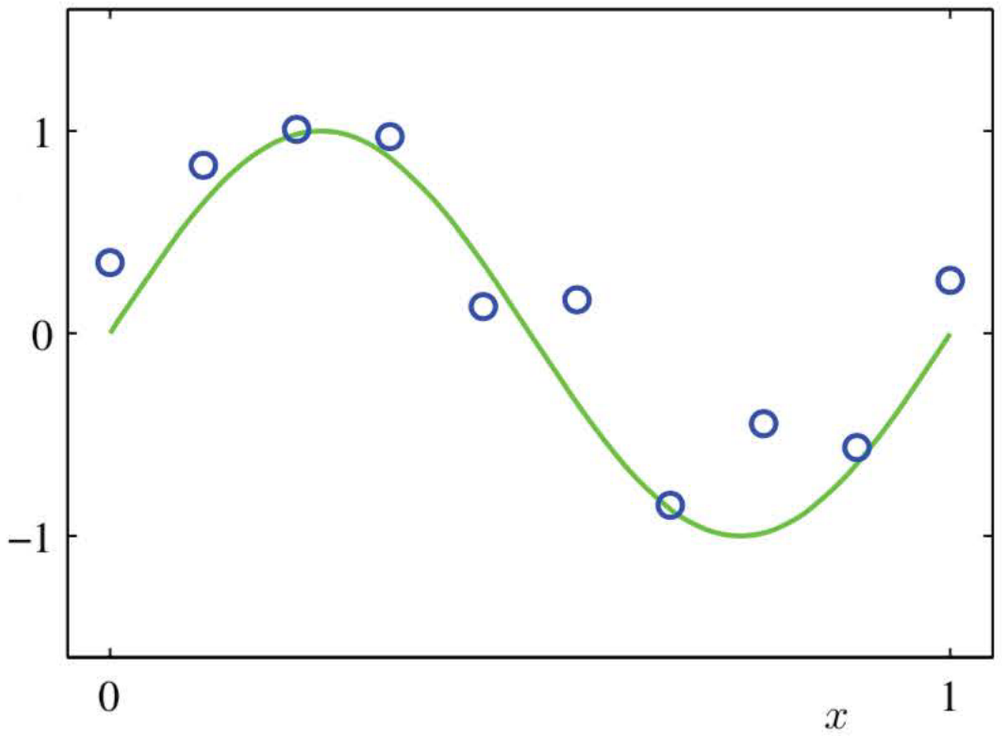
- **Underfitting** is a phenomenon that the learned model does not fit the training data well; that is, large empirical risk.
- **Overfitting** is a phenomenon that the learned model fits the training data very well but it cannot generalise well to unseen examples drawn from the same distribution; that is, large difference between training and test errors.

Linear regression



- True target: $\sin(2\pi x)$ with small Gaussian noises.
- $h(x) = w_0 + w_1 x + \cdots + w_M x^M$
- $R_S(h) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - h(x_i)\big)^2$

Bishop's book: "Pattern Recognition and Machine Learning"



$$h(x) = w_0 + w_1 x + \ldots + w_9 x^9$$

Assuming the prior distribution:

$$p(h) = \prod_{i=0}^{9} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau w_i^2}{2}\right)$$

Then, we have

$$\arg\min_h(-\ln p(h|S, \beta^{-1})) = \arg\min_h(-\frac{n}{2}\ln\beta + \frac{n}{2}\ln(2\pi) + \frac{\beta}{2} n R_S(h)$$

$$- 5\ln\tau + 5\ln(2\pi) + \frac{\tau}{2} n \sum_{i=0}^{9} w_i^2)$$

Minimising above equals

$$\min R_S(h) + \lambda \sum_{i=0}^{9} w_i^2 = R_S(h) + \lambda \|w\|_2^2, \quad \lambda = \frac{\tau}{\beta}$$

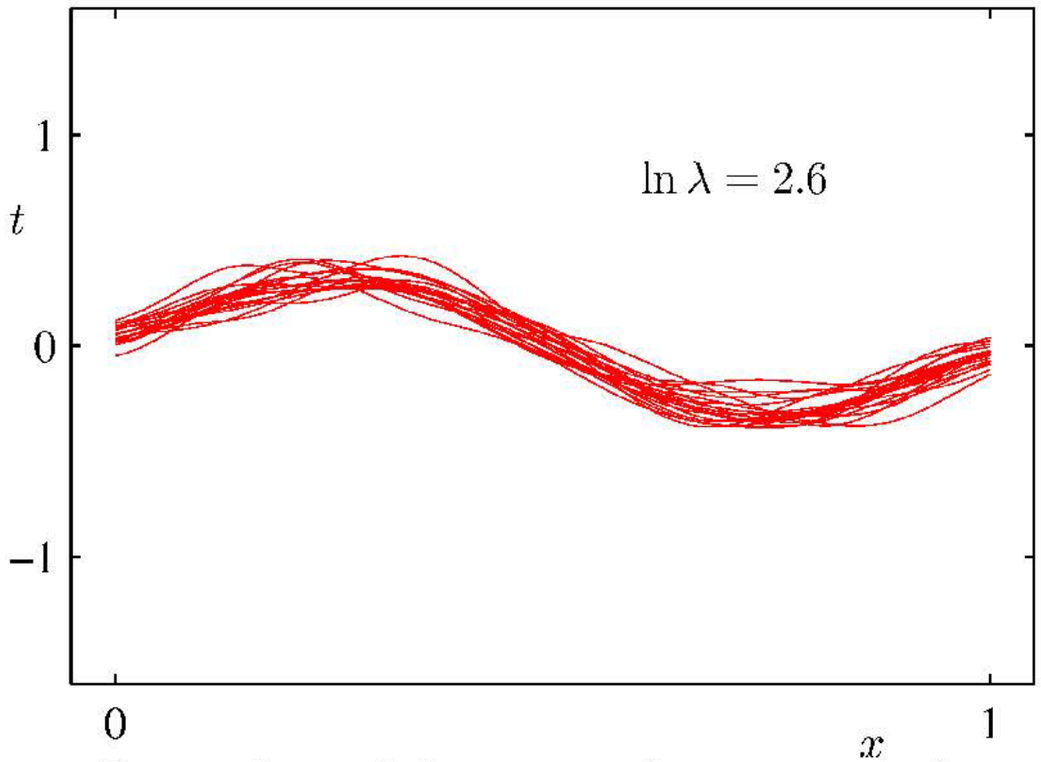
Bishop's book: "Pattern Recognition and Machine Learning"

ridge regression (L2 penalty)

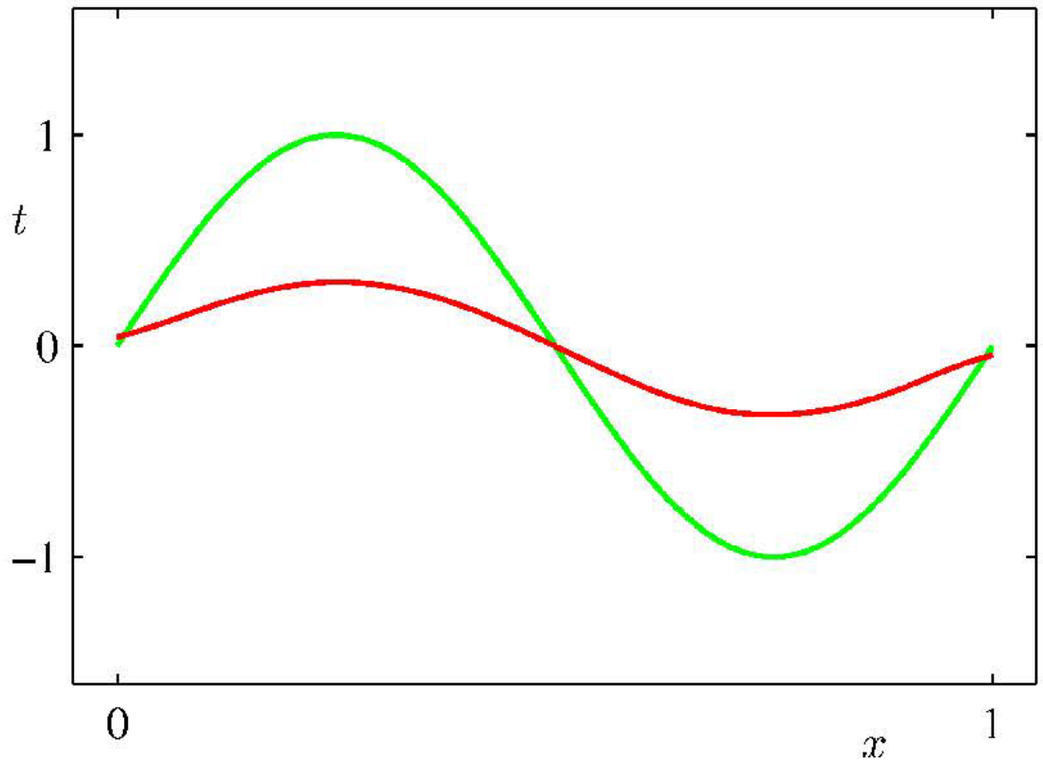
Bias-Variance Visualisation

THE UNIVERSITY OF SYDNEY

20 datasets with varying regularisation parameter.



Result of fitting the model to each dataset.

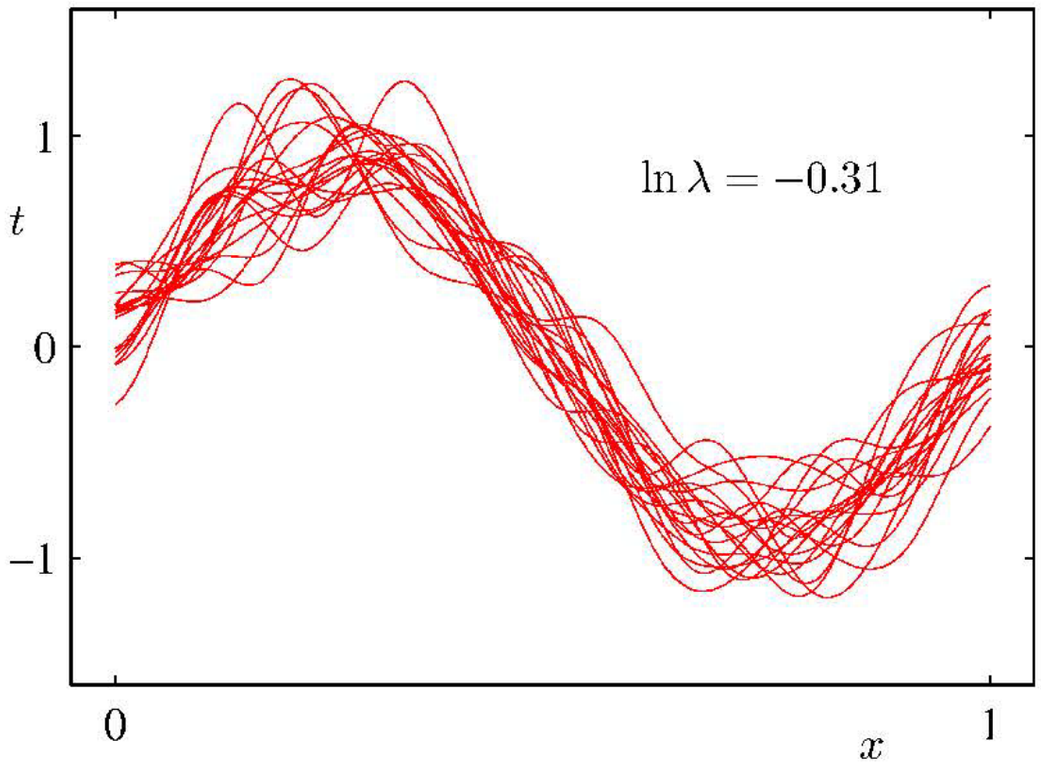


Average of the fits.

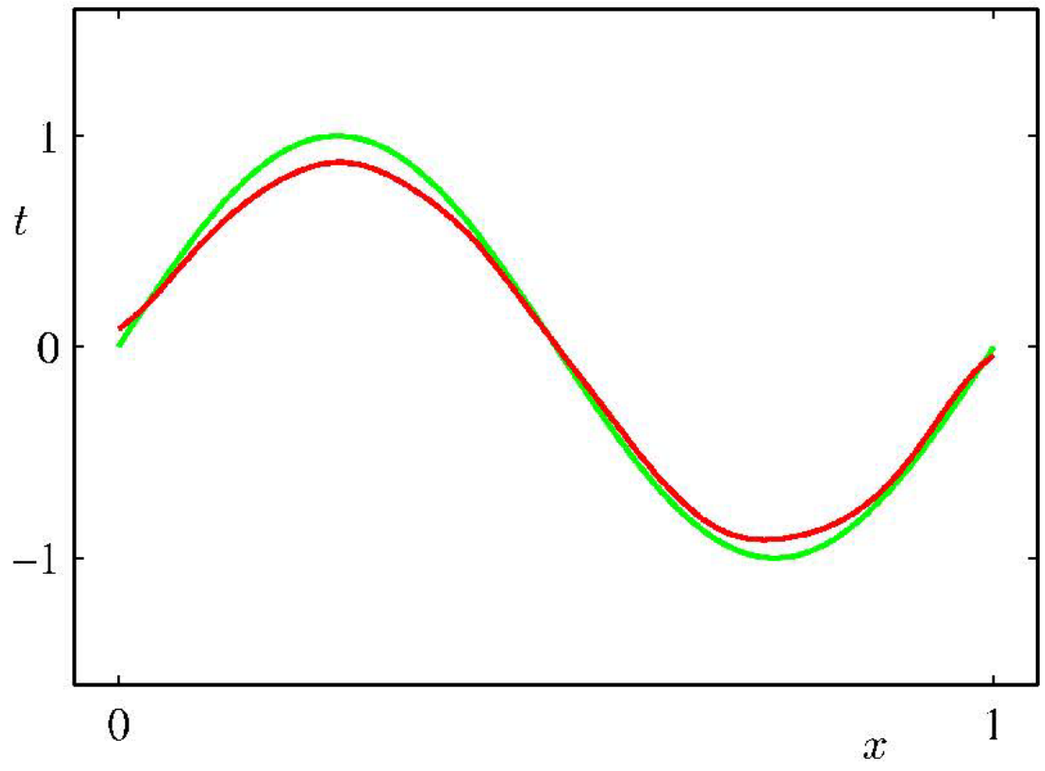
Bias-Variance Visualisation

THE UNIVERSITY OF SYDNEY

20 datasets with varying regularisation parameter.



Result of fitting the model to each dataset.

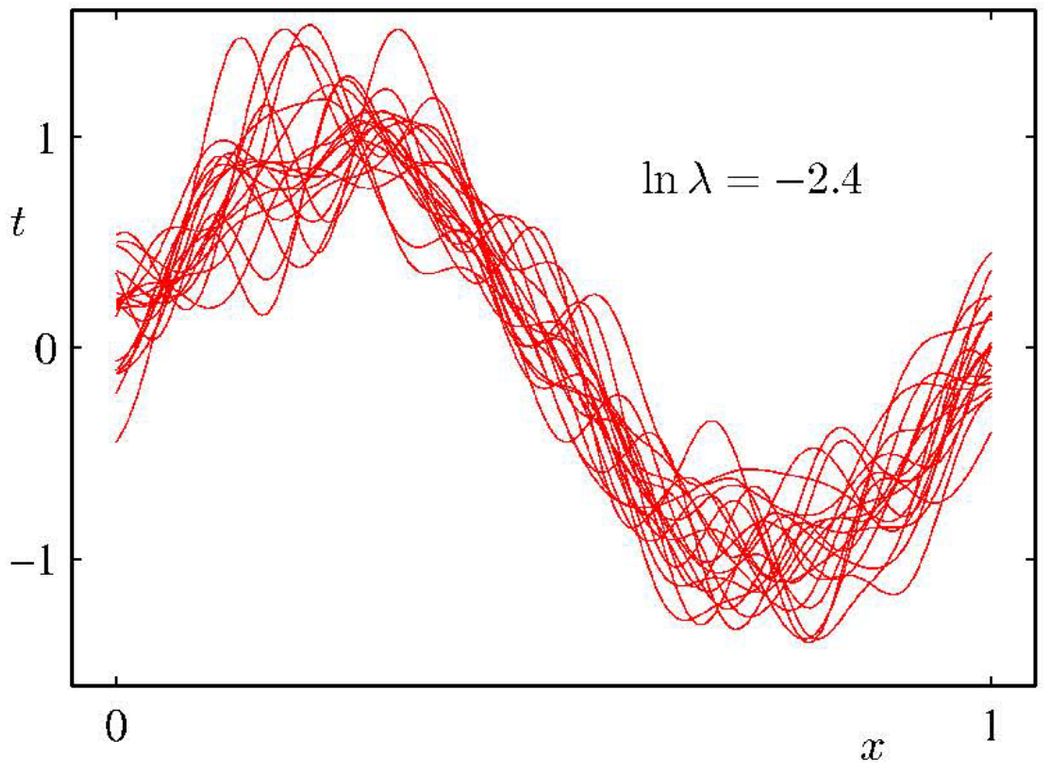


Average of the fits.

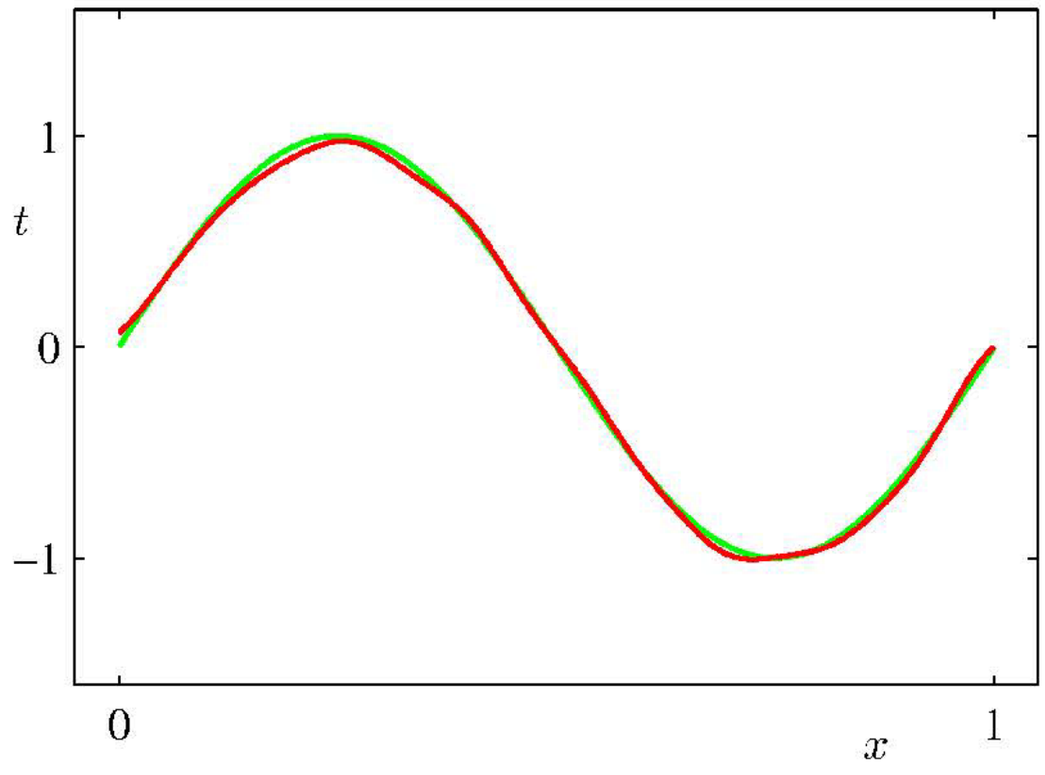
Bias-Variance Visualisation

THE UNIVERSITY OF SYDNEY

20 datasets with varying regularisation parameter.



Result of fitting the model to each dataset.

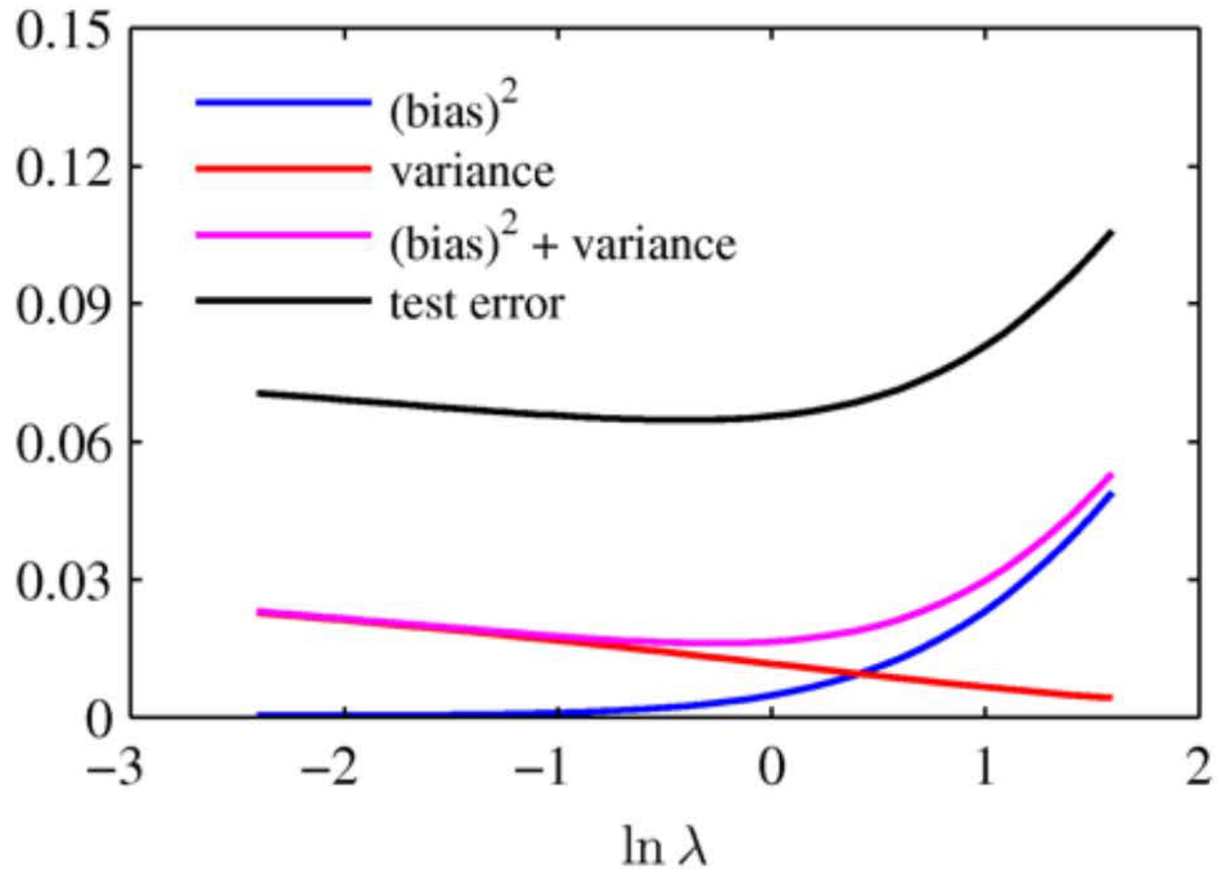


Average of the fits.

The Bias-Variance Trade Off

THE UNIVERSITY OF SYDNEY

From these plots, we note that an over-regularised model (large λ) will have a high bias, while an under-regularised model (small λ) will have a high variance.



Why the above phenomenon happens?

Bias-Variance vs Under-over fitting

THE UNIVERSITY OF SYDNEY

- High variance implies overfitting.
- High bias implies underfitting.

Avoid overfitting

THE UNIVERSITY OF SYDNEY

- Reducing the complexity of the predefined hypothesis class

- Increasing training sample size

Why those two methods help?

Avoid overfitting

THE UNIVERSITY OF SYDNEY

- **Reducing hypothesis complexity**: the hypothesis fits the training data too well because it is too complex.
- **Increasing sample size**: According to the law of large numbers, with more training examples, the empirical risk is closer to the expected risk. Increasing sample size will be helpful to learning the best hypothesis.

THE UNIVERSITY OF SYDNEY

Compare robustness among surrogate loss functions

Surrogate loss functions

THE UNIVERSITY OF SYDNEY

- Least squares loss: $\ell(X, Y, h) = (Y - h(X))^2$
- Absolute loss: $\ell(X, Y, h) = |Y - h(X)|$
- Cauchy loss: $\ell(X, Y, h) = \ln\left(1 + \left(\frac{(Y - h(X))}{\gamma}\right)^2\right)$
- Correntropy loss (Welsch loss):

$$\ell(X, Y, h) = \left(1 - \exp\left(-\left(\frac{Y - h(X)}{\sigma}\right)^2\right)\right)$$

Surrogate loss function robustness

THE UNIVERSITY OF SYDNEY

- Least squares loss: $g'(1) = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - h(x_i))(-h(x_i))$
- Absolute loss: $g'(1) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|y_i - h(x_i)|}(y_i - h(x_i))(-h(x_i))$
- Cauchy loss:

$$g'(1) = \frac{1}{n}\sum_{i=1}^{n} \frac{2}{\gamma^2 + (y_i - h(x_i))^2}(y_i - h(x_i))(-h(x_i))$$

- Correntropy loss (Welsch loss):

$$g'(1) = \frac{1}{n}\sum_{i=1}^{n} \frac{2}{\sigma^2 \exp\left(\frac{y_i - h(x_i)}{\sigma}\right)^2}(y_i - h(x_i))(-h(x_i))$$

Surrogate loss function

THE UNIVERSITY OF SYDNEY

In other words, to minimise *f(h)*, we should find an *h* such that

$$g'(1) = 0.$$

- All $g'(1)$ w.r.t. the above four loss functions has the term $c_i = (y_i - h(x_i))(-h(x_i))$, we treat them as the bases of contribution to optimising the empirical risks. We can see that different surrogate loss functions assign different weights to the bases. A surrogate loss function is more robust if it assign a smaller weights to the bases as the error (or noise) is going bigger.

Key points

- Linear Regression with Gaussian Noise
- Maximum Likelihood Estimation (MLE)
- Maximum A Posterior (MAP)
- Bias and variance
- Robustness of surrogate loss functions