Linear Regression

Alice Oh
alice.oh@kaist.edu

Linear Regression

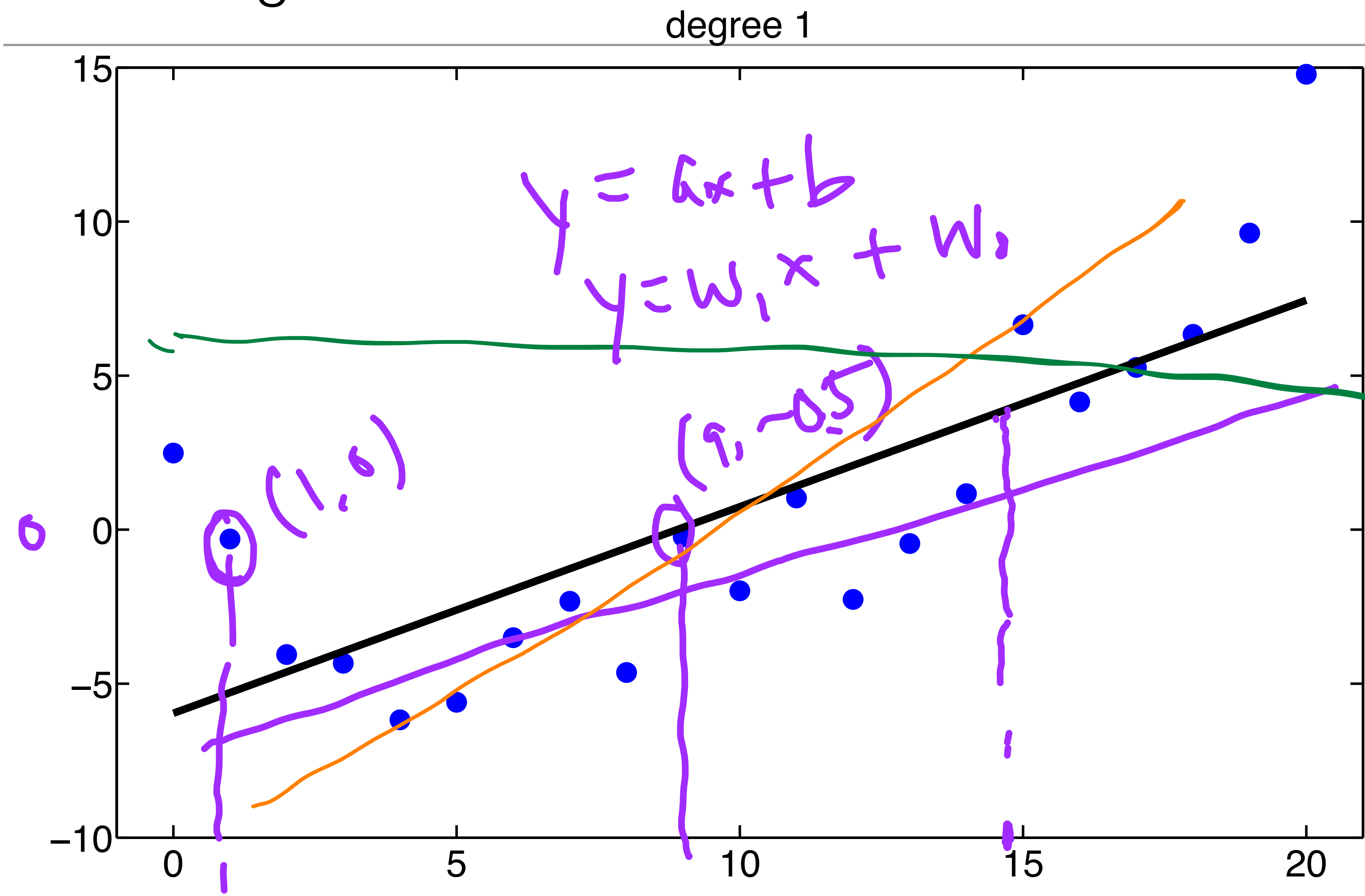


Figure from Kevin Murphy's book -- Machine Learning: A Probabilistic Perspective

Linear Regression

- Response (real number) is a linear function of the inputs

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \epsilon$$

- Assume that \epsilon (the residual error) has a Gaussian distribution

$$p(y|\mathbf{x}, \theta) = \mathcal{N}(y|\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$$

where

$$\mu(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = w_0 + w_1 x$$

Modeling non-linear relationships

- Simply take

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

and replace x with some non-linear function of the inputs

$\phi(x) = x^2$

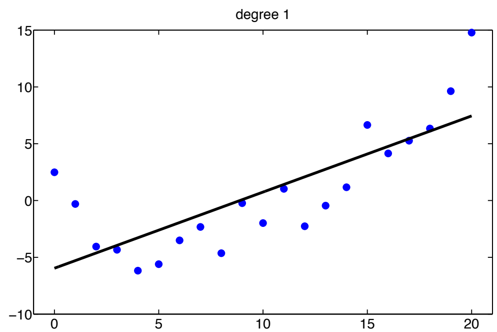
$$p(y|\mathbf{x}, \theta) = \mathcal{N}(y|\mathbf{w}^T\boxed{\phi(\mathbf{x}),}\sigma^2)$$

$2x + 3x^2$

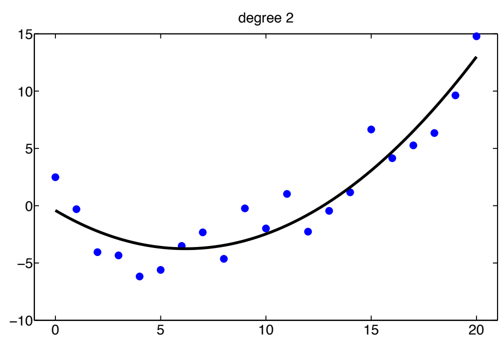
This is called the basis function expansion.

(But the model is still called linear regression because it is linear in the parameters w)

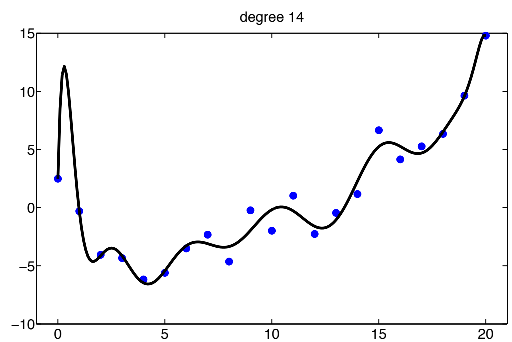
Polynomial Regression



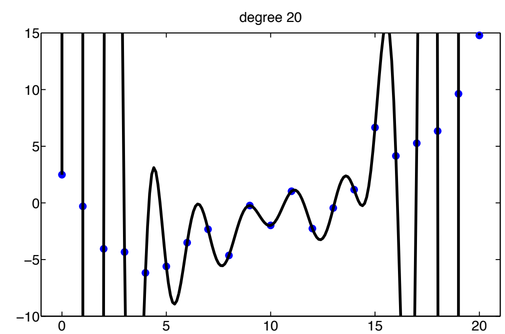
x



x^2



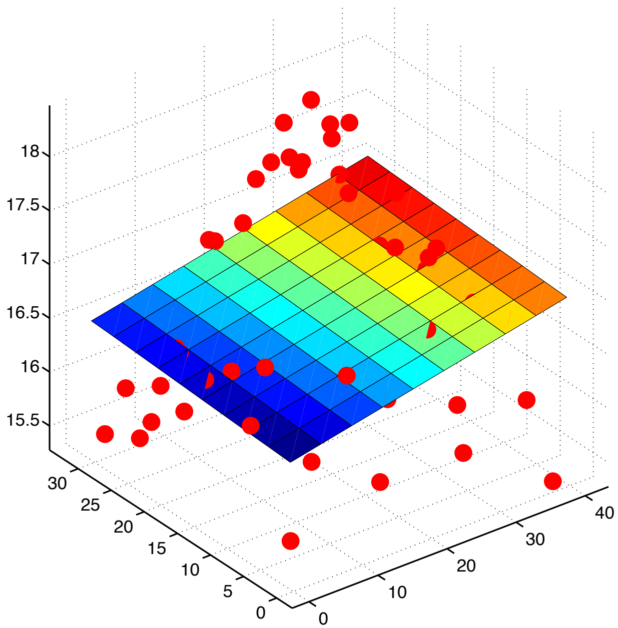
x^{14}



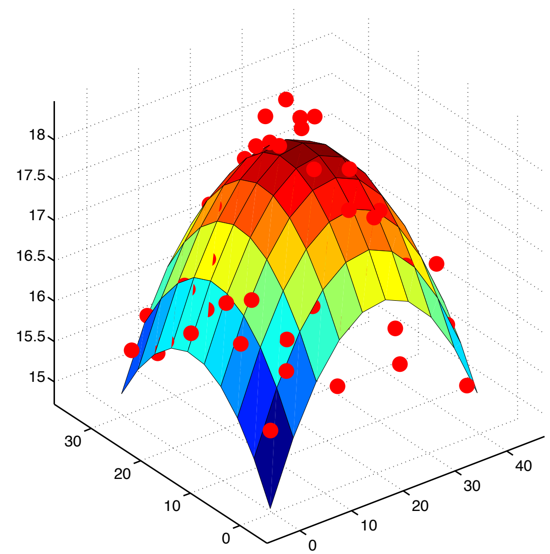
x^{20}

Figures from Kevin Murphy's book -- Machine Learning: A Probabilistic Perspective

Multivariate linear regression



$$w_0 + w_1 x_1 + w_2 x_2$$



$$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2$$

Figures from Kevin Murphy's book -- Machine Learning: A Probabilistic Perspective

Maximum likelihood estimation

- Using MLE, arguments \theta can be computed by

$$\arg\max \log p(D|\theta)$$

$$= \sum_{i=1}^{N} log p(y_i|\mathbf{x}_i, \theta)$$

$y = w x + w_0$

3 2

- If we plug in the Gaussian formulation

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

and put it into the log likelihood above, we get

$$= \sum_{i=1}^{N} log[(\frac{1}{2\pi\sigma^2})^{\frac{1}{2}} exp(-\frac{1}{2\sigma^2}(y_i - \mathbf{w}^T\mathbf{x}_i)^2)]$$

$$= -\frac{N}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

Residual Sum of Squares

- Log likelihood

$$-\frac{N}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

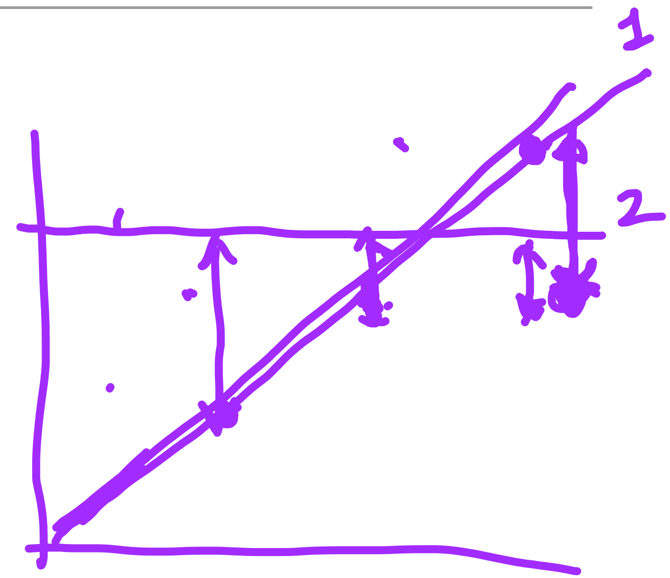
- Negative log likelihood (NLL)

$$\frac{N}{2}log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2$$

- To minimize NLL, we minimize this term

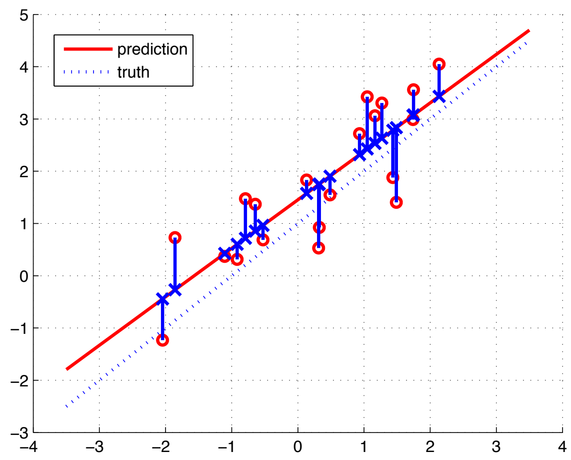
$$\boxed{\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2}$$

called residual sum of squares (RSS)



Least Squares

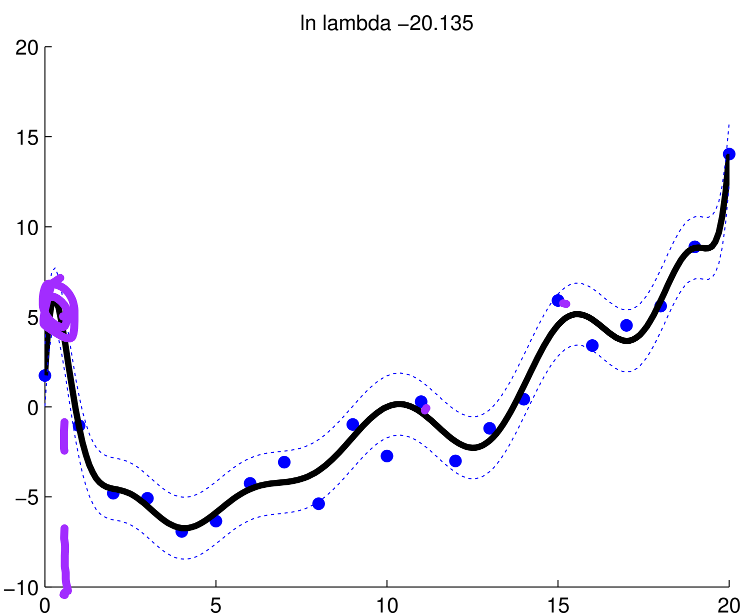
- MLE for w is the one that minimizes the RSS



Ridge Regression

Regularization

- MLE can overfit
  - For linear regression, this means the weights can become large



6.560, -36.934, -109.255,
543.452, 1022.561,
-3046.224, -3768.031,
8524.540 …

- We can encourage the weights to be small by putting a zero-mean Gaussian prior on the weights – l_2 Regularization!

Figure from Murphy's MLPP book

Ridge Regression

- Zero-mean Gaussian prior on the weights

$$p(\mathbf{w}) = \prod_j \mathcal{N}(w_j|0, \tau^2)$$

- MAP estimation problem

$$\text{argmax } \textstyle\sum_{i=1}^{N} log\mathcal{N}(y_i|w_0 + \mathbf{w}^T\mathbf{x}_i, \sigma^2) + \textstyle\sum_{j=1}^{D} log\mathcal{N}(w_j|0, \tau^2)$$

- Compare with the MLE problem

$$\arg\max \log p(D|\theta) = \sum_{i=1}^{N} log p(y_i|\mathbf{x}_i, \theta)$$

- To solve the MAP estimation, minimize

$$J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2 + \lambda||\mathbf{w}||_2^2$$

$$\text{where } \lambda = \tfrac{\sigma^2}{\tau^2} \text{ and } ||\mathbf{w}||_2^2 = \textstyle\sum_j w_j^2 = \mathbf{w}^T\mathbf{w}$$

Ridge Regression

$$J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \mathbf{w}^T\mathbf{x}_i))^2 + \boxed{\lambda||\mathbf{w}||_2^2}$$

$w_1 = 105$

$w_2 = 1300$

- Compare with NLL before

$$\boxed{\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2}$$

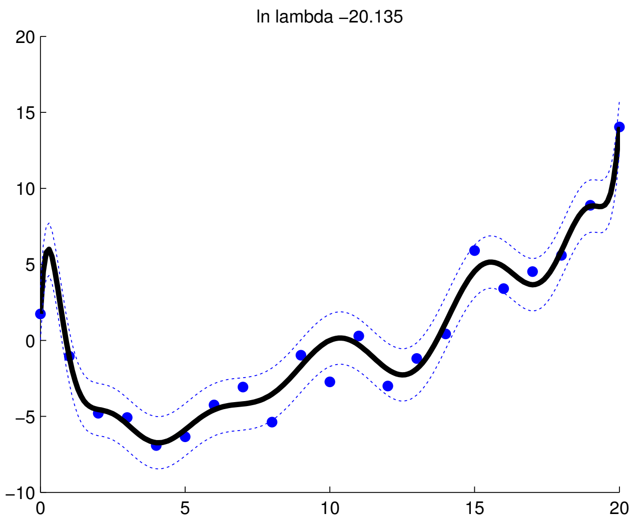
- So the first term of ridge regression is same as NLL, and the second term is the complexity penalty (when \lamba > 0)

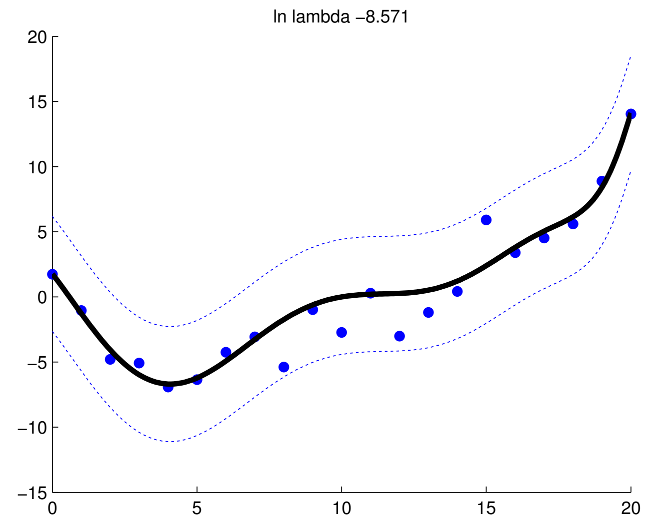
- Corresponding solution is

$$\hat{\mathbf{w}}_{ridge} = (\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$ (from Murphy's MLPP book Section 7.5.1)

$$\hat{\mathbf{w}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$ (from Murphy's MLPP book Section 7.3.1)

Ridge Regression



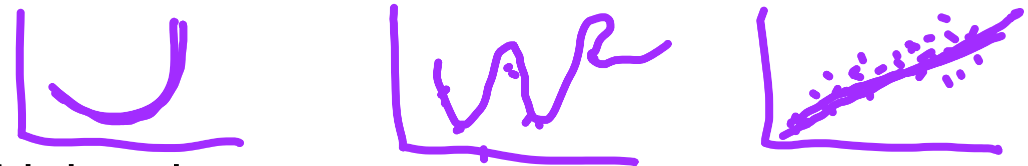


$$\hat{\mathbf{w}}_{ridge} = (\lambda \mathbf{I}_D + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\hat{\mathbf{w}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

What happens when \lambda is 0?
What happens as \lambda increases?
What happens when \lambda is infinity?

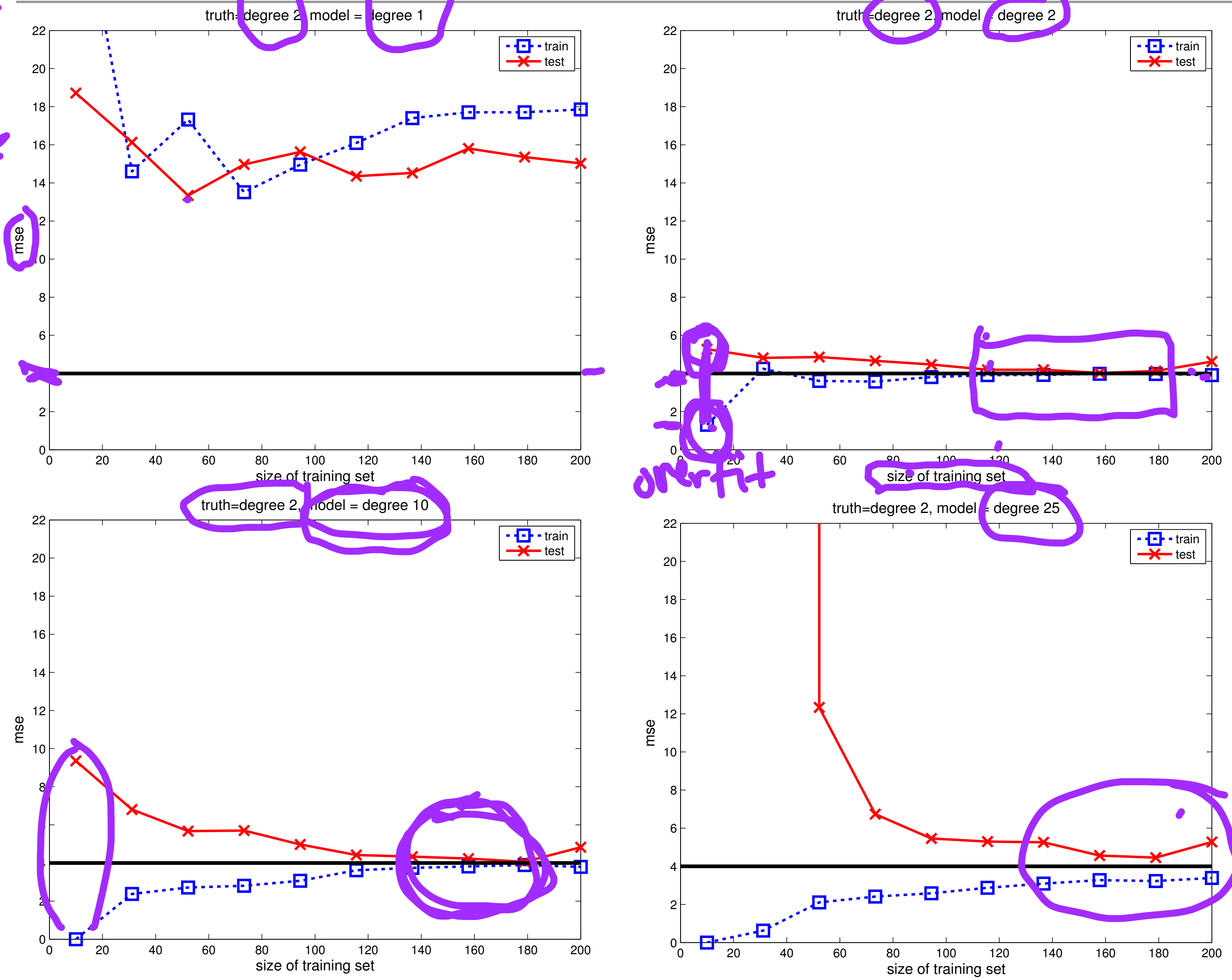
Figures from Murphy's MLPP book



Regularization effects of big data (Murphy MLPP Section 7.5.4)

x^2

x



$\sum |w|^2$

ℓ_2 norm

ℓ_2 reg