CS540 Machine learning
Lecture 3

Review

- Probability models: Gaussian, Binomial, Multinomial, linear regression, logistic regression
- MLE of Gaussian, Binomial, Multinomial

Outline

- Basic concepts
  - Loss functions
  - Estimation vs inference
  - Decision boundaries
  - Overfitting
  - Regularization
  - Model selection
  - Structural error vs approximation error

Loss functions

- Squared error, 0-1 loss

$$\begin{aligned} L(y, \hat{y}) &= (y - \hat{y})^2 \\ L(y, \hat{y}) &= I(y \neq \hat{y}) \end{aligned}$$

- Minimize risk (expected loss, empirical loss)

$$\begin{aligned} R(\hat{f}) &= E_{\mathbf{x},y} L(f(\mathbf{x}), \hat{f}(\mathbf{x})) \\ \hat{R}(\hat{f}) &= \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i)) \end{aligned}$$

Loss functions for density estimation

- Suppose output is \hat{p}(.|x), truth is p(.|x)
- Use KL (Kullback Leibler) loss

$$L(p(y|\mathbf{x}), \hat{p}(y|\mathbf{x})) = KL(p(y|\mathbf{x}), \hat{p}(y|\mathbf{x})) = \sum_y p(y|\mathbf{x}) \log \frac{p(y|\mathbf{x})}{\hat{p}(y|\mathbf{x})}$$

- Risk is expected negative log likelihood

$$R(\hat{p}) = -E_{\mathbf{x}} \sum_y p(y|\mathbf{x}) \log \hat{p}(y|\mathbf{x}) = -E_{\mathbf{x},y} \log \hat{p}(y|\mathbf{x})$$

Estimation vs Inference

- Learning as optimization (frequentist): Given D, Choose $\hat{f}$ to approximate f as closely as possible, so as to minimize (future) expected loss
- Usually compute parameter estimate $\hat{\boldsymbol{\theta}}$
- Learning as inference (Bayesian): given D, compute posterior over functions $p(f|\mathcal{D})$
- Or posterior over parameters

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- In the decision theory chapter, we show that one of the best ways to minimize frequentist risk is to be Bayesian

MAP estimation

- One possible point estimate derived from the posterior is the posterior mode or Maximum A Posterior value

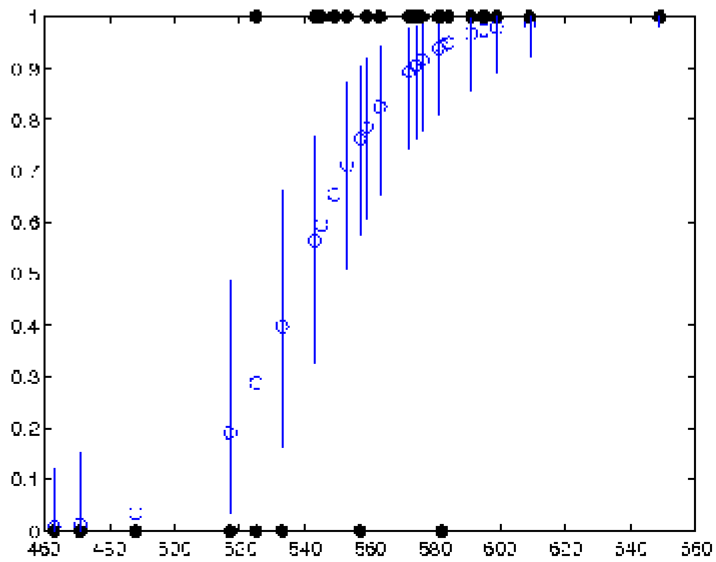
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

- Equivalent to penalized maximum likelihood
- Computing MAP is optimization problem (fast)
- Not strictly Bayesian, since it is a point estimate, not a probability distribution
- We will study Bayesian methods later

Uncertainty in parameter estimates

- Uncertainty in p(θ|D) induces uncertainty in p(y|x,θ)
- Ignoring uncertainty in parameters can cause over confidence

$$p(y=1|x,\mathbf{w}) = \sigma(\mathbf{w}^T[1,x]) = \sigma(w_0 + w_1 x)$$



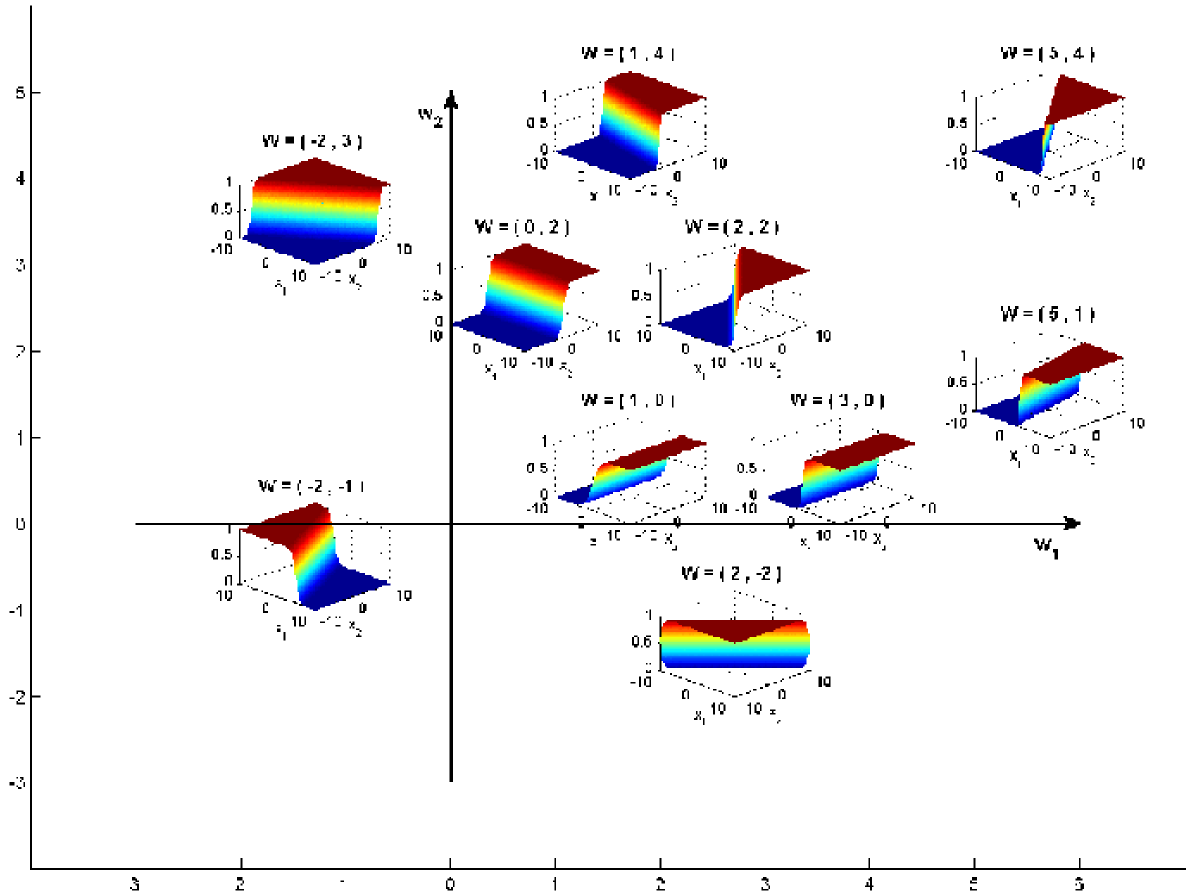
Outline

- Basic concepts
  - Loss functions
  - Estimation vs inference
  - Decision boundaries
  - Overfitting
  - Regularization
  - Model selection
  - Structural error vs approximation error

Decision boundaries

- Logistic regression in 2D

$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x})$$



Decision boundaries

- Decision region and boundary

$$
\begin{aligned}
R_0 &= \{\mathbf{x} : p(y=0|\mathbf{x},\mathbf{w}) > p(y=1|\mathbf{x},\mathbf{w})\} \\
\mathcal{B} &= \{\mathbf{x} : p(y=1|\mathbf{x},\mathbf{w}) = p(y=0|\mathbf{x},\mathbf{w}) = 0.5\} \\
\mathcal{B} &= \{\mathbf{x} : \log \frac{p(y=1|\mathbf{x},\mathbf{w})}{p(y=0|\mathbf{x},\mathbf{w})} = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) = 0
\end{aligned}
$$

Log odds ratio $\log \frac{p(y=1|\mathbf{x},\mathbf{w})}{p(y=0|\mathbf{x},\mathbf{w})} = \log \frac{e^\eta}{1+e^\eta} \frac{1+e^\eta}{1} = \log e^\eta = \eta$

- 2D input

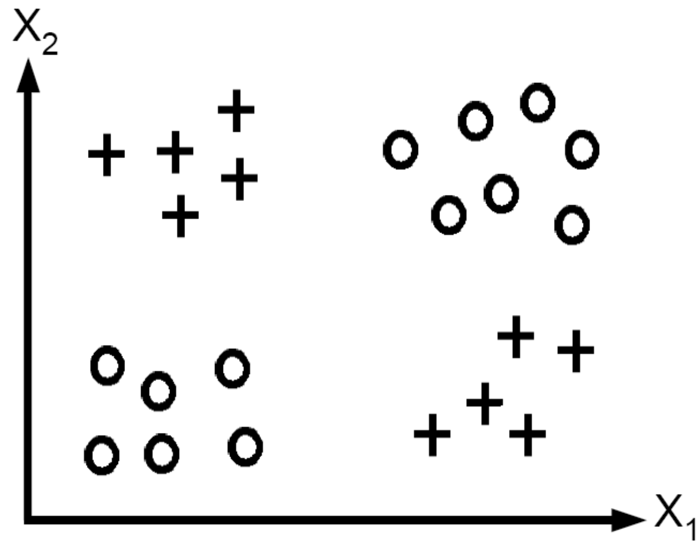
$$\mathcal{B} = \{\mathbf{x} : w_0 + w_1 x_1 + w_2 x_2 = 0\}$$

- 1D input

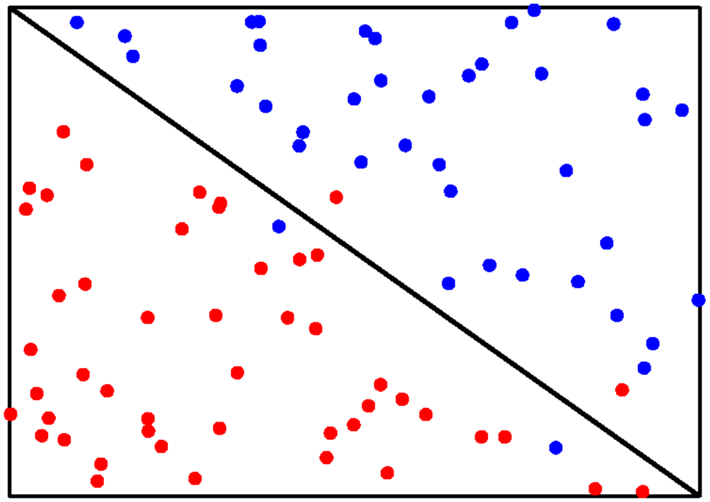
$$\mathcal{B} = \{x : w_0 + w_1 x = 0\} = \{x : x = \frac{-w_0}{w_1} = w^*\}$$

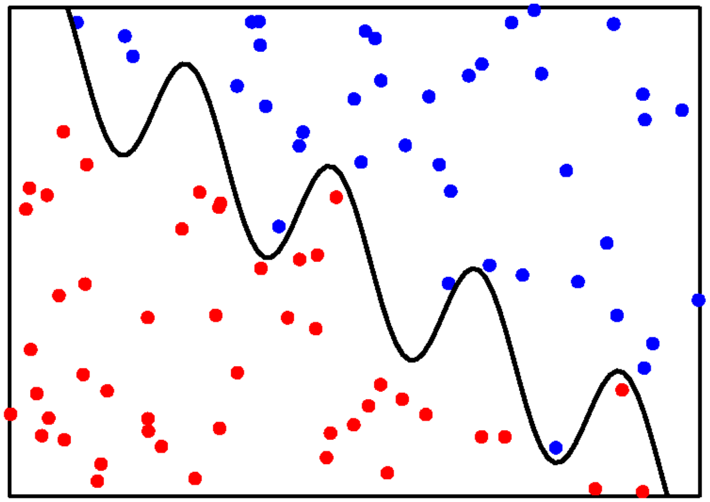
Xor problem

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

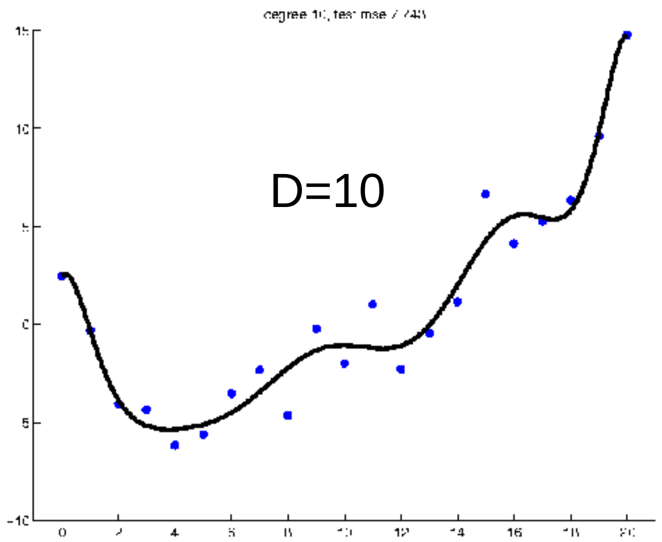


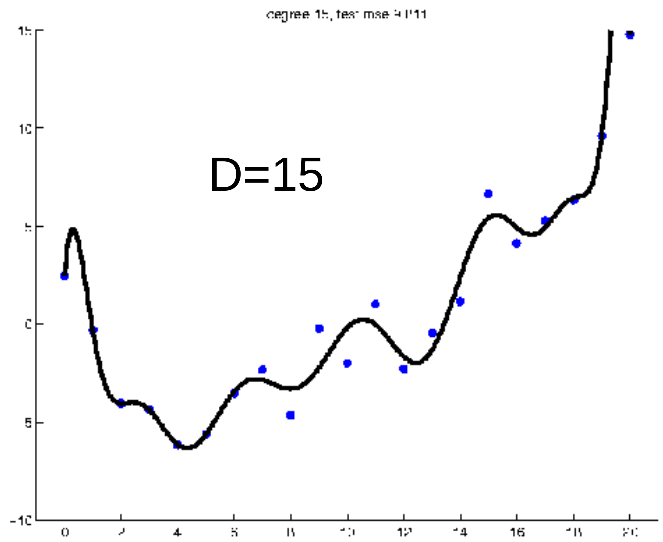
Linearly separable data

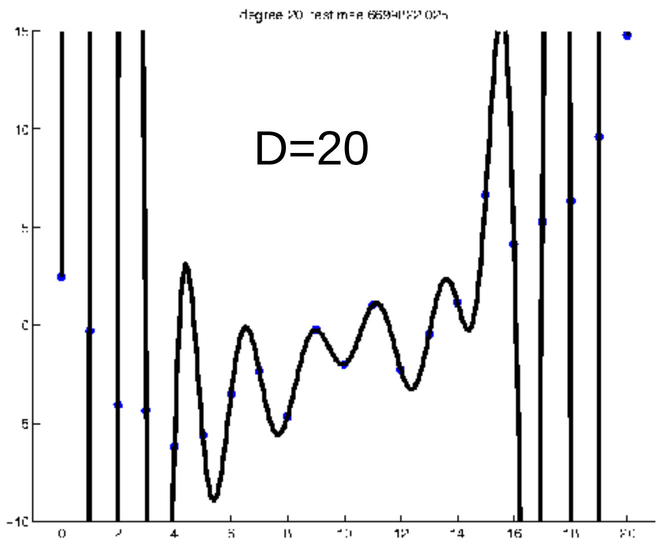


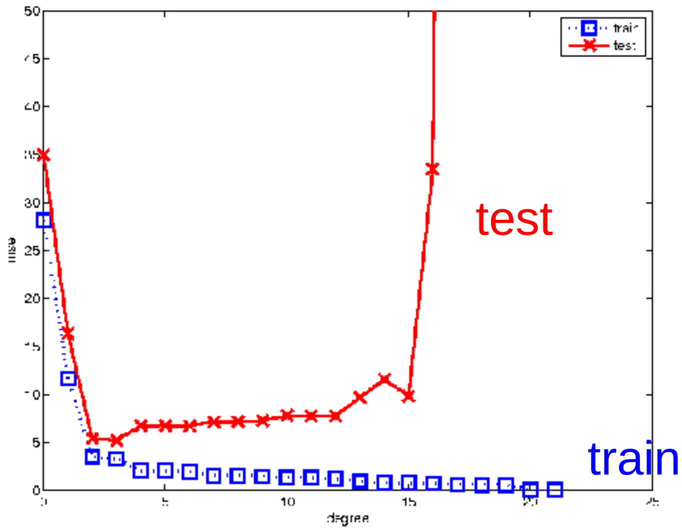


Overfitting









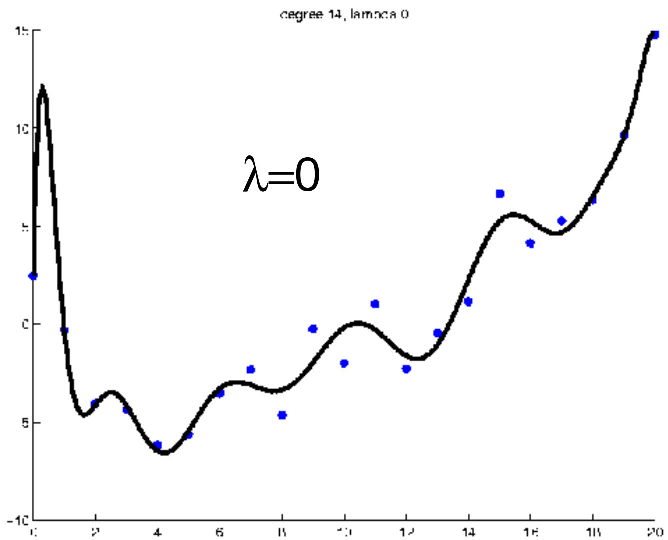
Regularization

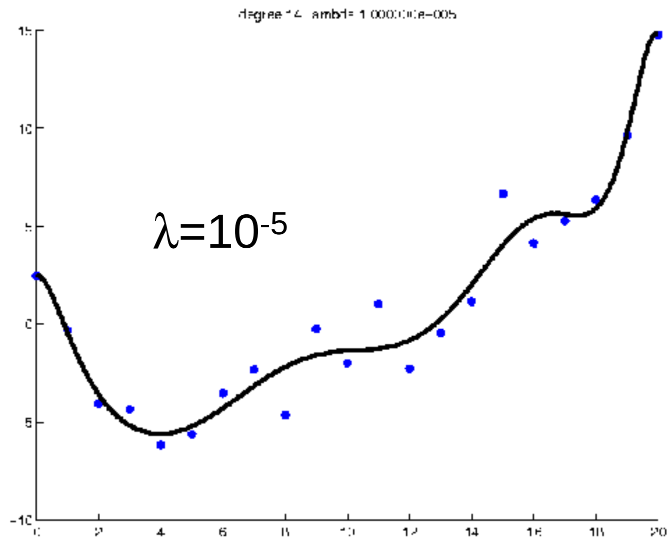
- Minimize penalized negative log likelihood

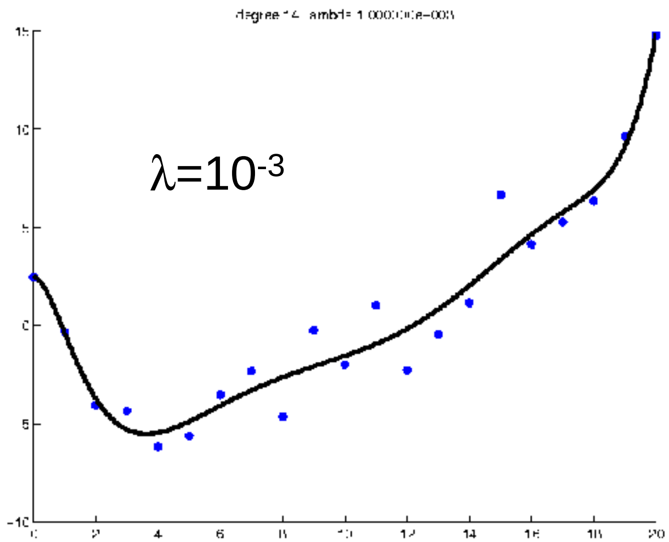
$$-\ell(\mathbf{w}) + \lambda||\mathbf{w}||_2^2$$

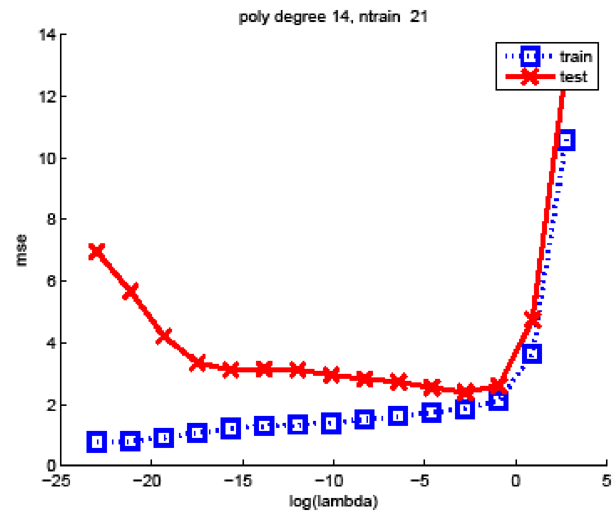
- Weight decay, shrinkage, L2 regularization, ridge regression

Regularization D=14









Why it works

- Coefficients if $\lambda=0$ (MLE)

```
-0.18, 10.57, -110.28, -245.63, 1664.41, 2647.81, -965
27669.94, 19319.66, -41625.65, -16626.90, 31483.81, 54
```

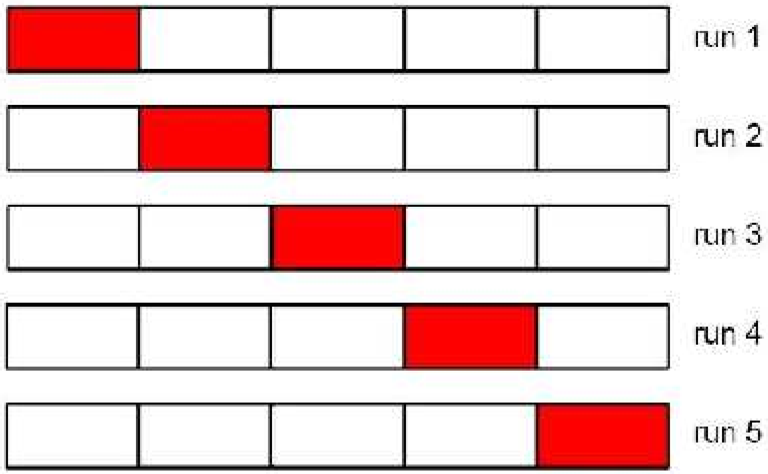
- Coefficients if $\lambda=10^{-3}$

```
-1.54,  5.52,  3.66, 17.04, -2.63, -23.06, -0.37, -8.49
7.92,  5.40,  8.29,  7.75,  1.78,  2.03, -8.42,
```

- Small weights mean the curve is almost linear (same is true for sigmoid function)

Model selection

- Cannot use test set to pick D or λ
- Partition training into train and validation
- If training set is small, use cross validation



Cross validation

- CV estimate of risk (expected loss)

$$
\begin{aligned}
J(\lambda) &= \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f(x_i, D_{-b(i)}, \lambda)\right) \\
J(\lambda) &= \frac{1}{n} \sum_{b=1}^{B} \sum_{i \in b} L\left(y_i, f(x_i, D_{-b}, K)\right)
\end{aligned}
$$

- Leave one out (LOOCV)

$$
J(\lambda) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i, D_{-i}, \lambda))
$$

Standard errors

- CV score is an estimate of the expected loss

$$\begin{aligned} L_i &= L(y_i, f(\mathbf{x}_i, D_{-b(i)}, K)) \\ J(\lambda) &= \overline{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \end{aligned}$$

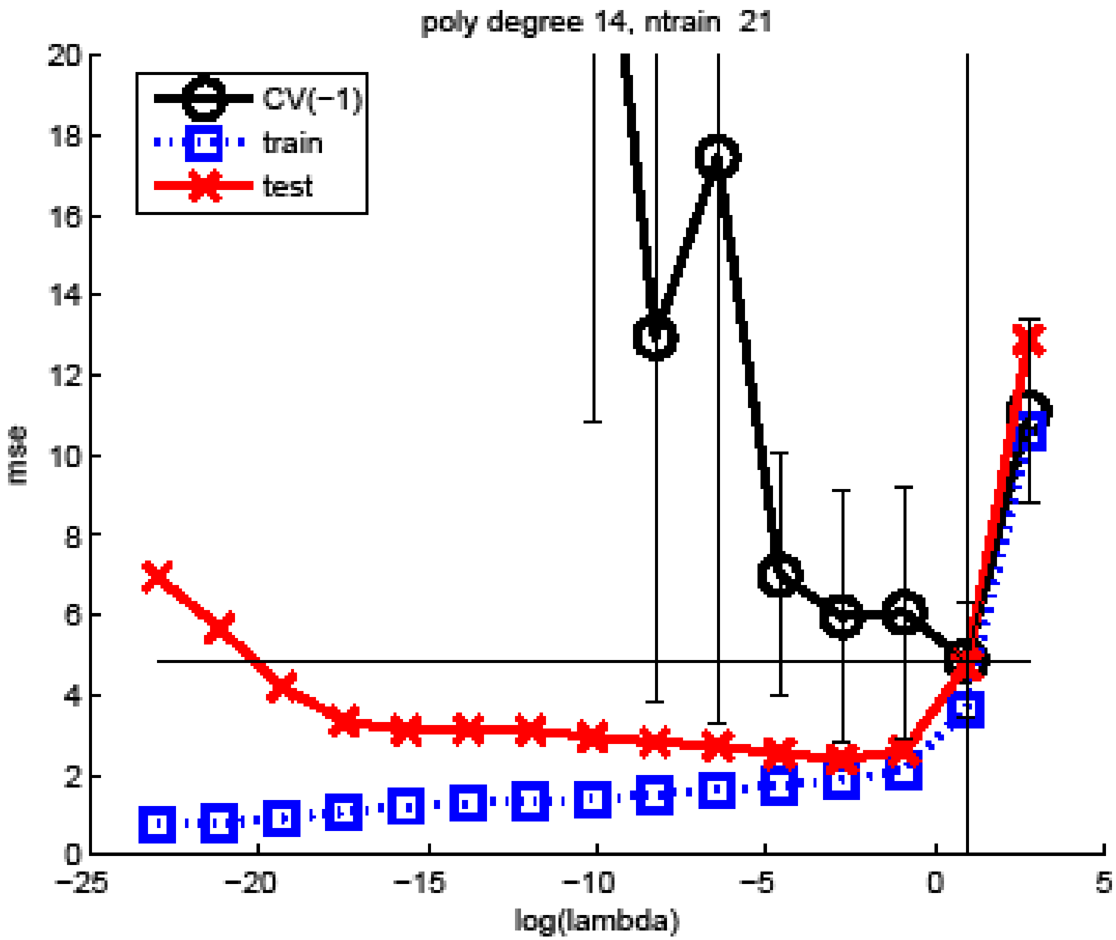
- Uncertainty in the mean can be quantified using the standard error

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(L_i - \overline{L})^2 \\ se &= \frac{\hat{\sigma}}{\sqrt{n}} = \sqrt{\frac{\hat{\sigma}^2}{n}} \end{aligned}$$

- From a Bayesian viewpoint, we can think of this as

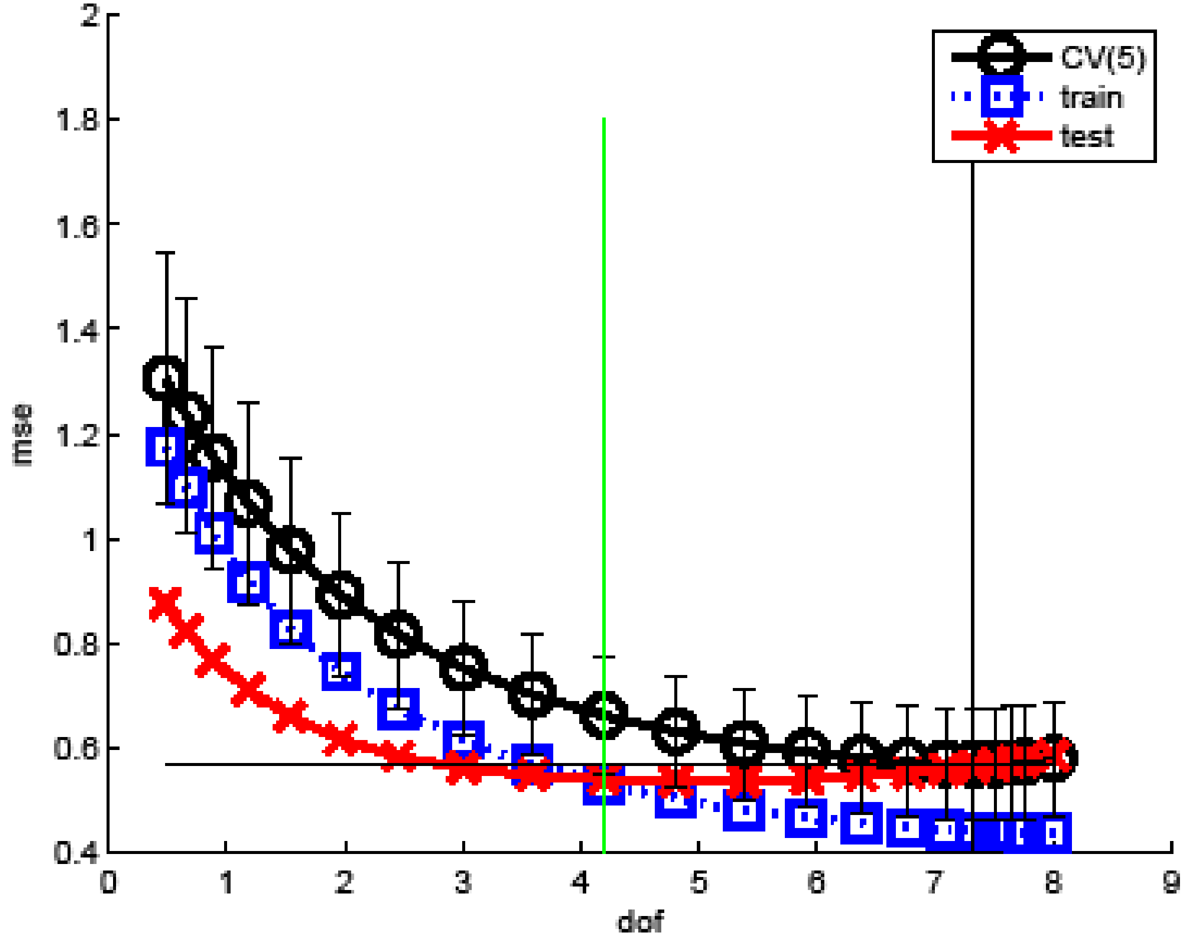
$$\overline{L} = Ep(J(\lambda)|\mathcal{D}), \quad se = \sqrt{\text{Var } p(J(\lambda)|\mathcal{D})}$$

CV with se



poly degree 14, ntrain 21
CV(−1)
train
test
mse
log(lambda)
20
18
16
14
12
10
8
6
4
2
0
−25
−20
−15
−10
−5
0
5

One standard error rule



Penalized likelihood methods

- CV can be slow: have to fit B models
- Instead pick λ to minimize

$$J(\lambda) = -\log p(\mathcal{D}|\hat{\boldsymbol{\theta}}(\lambda)) + C(\hat{\boldsymbol{\theta}}(\lambda))$$

- Eg BIC, AIC – see later
- Or use Empirical Bayes – see later

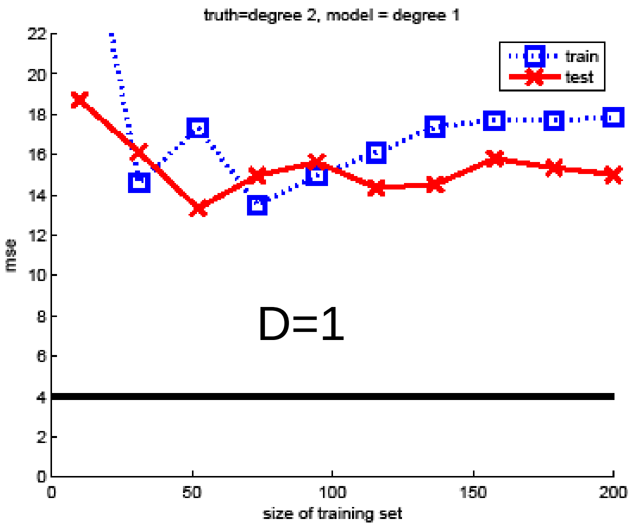
$$J(\lambda) = p(\mathcal{D}|\lambda) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}$$

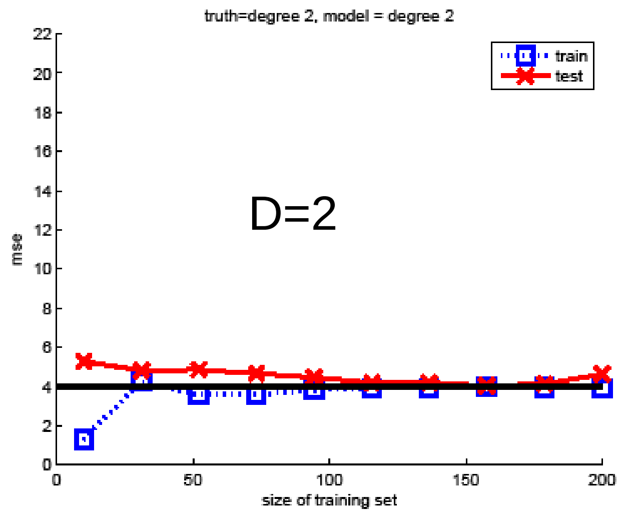
- CV estimate J(λ) requires grid search over λ; other methods can use gradient-based optimization

Outline

- Basic concepts
  - Loss functions
  - Estimation vs inference
  - Decision boundaries
  - Overfitting
  - Regularization
  - Model selection
  - Structural error vs approximation error

Structural error vs approximation error





Truth = D=2
Sigma^2 = 4

Ntest=200