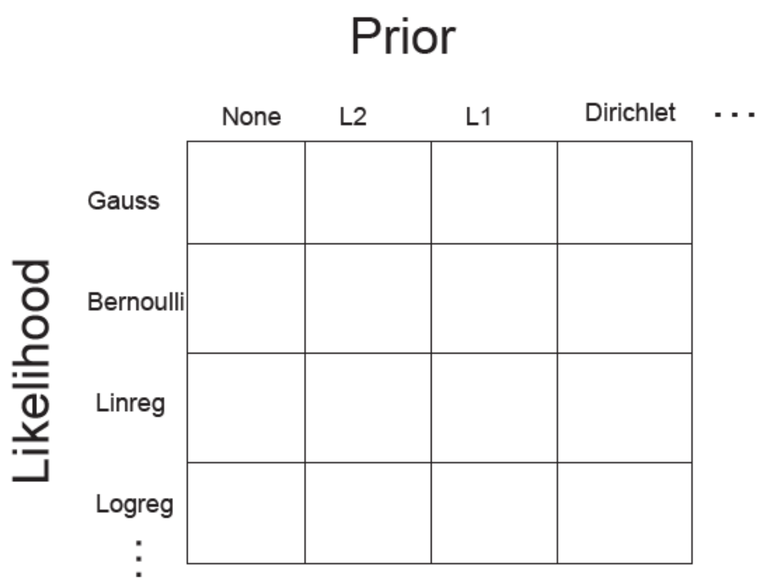
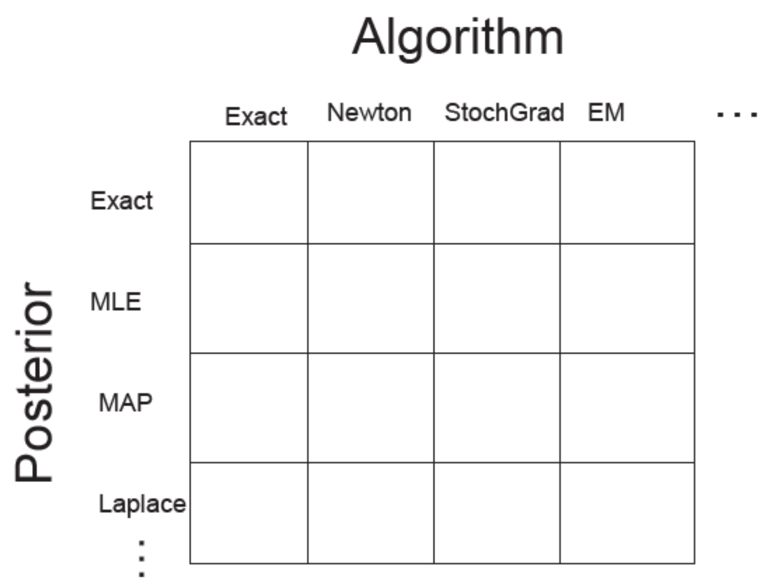
CS540 Machine learning
Lecture 11
Decision theory, model selection

Outline

- Summary so far
- Loss functions
- Bayesian decision theory
- ROC curves
- Bayesian model selection
- Frequentist decision theory
- Frequentist model selection

Models vs algorithms





P(x|theta) scalar x

Likelihood	Prior	Posterior	Algorithm
Bernoulli	None	MLE	Exact §**??**
Bernoulli	Beta	Beta	Exact §**??**
Gauss	None	MLE	Exact §**??**
Gauss	Gauss	Gauss	Exact §**??**
Gauss	NIG	NIG	Exact
Student T	None	MLE	EM §**??**
Beta	NA	NA	NA §**??**
Gamma	NA	NA	NA §**??**

P(x|theta) vector x

Likelihood	Prior	Posterior	Algorithm
MVN	None	MLE	Exact §**??**
MVN	MVN	MVN	Exact
MVN	MVNIW	MVNIW	Exact
Multinomial	None	MLE	Exact §**??**
Multinomial	Dirichlet	Dirichlet	Exact §**??**
Dirichlet	NA	NA	NA §**??**
Wishart	NA	NA	NA §**??**

P(x,y|theta)

Likelihood	Prior	Posterior	Algorithm
GaussClassif	None	MLE	Exact §**??**
GaussClassif	MVNIW	MVNIW	Exact
NB binary	None	MLE	Exact §**??**
NB binary	Beta	Beta	Exact §**??**
NB Gauss	None	MLE	Exact §**??**
NB Gauss	NIG	NIG	Exact §**??**

P(y|x,theta)

Likelihood	Prior	Posterior	Algorithm
Linear regression	None	MLE	QR §**??**, SVD §**??**, LMS
Linear regression	L2	MAP	QR §**??**, SVD §**??**
~~Linear regression~~	~~L1~~	~~MAP~~	~~QP §**??**, CoordDesc §**??**,~~
Linear regression	MVN	MVN	QR/Cholesky §**??**
Linear regression	MVNIG	MVNIG	-
Logistic regression	None	MLE	IRLS §**??**, perceptron §**?**
Logistic regression	L2	MAP	Newton §**??**, BoundOpt §
~~Logistic regression~~	~~L1~~	~~MAP~~	~~BoundOpt §**??**~~
Logistic regression	MVN	LaplaceApprox	Newton §**??**
~~GP regression~~	~~MVN~~	~~MVN~~	~~Exact~~
~~GP classi¬ cation~~	~~MVN~~	~~LaplaceApprox~~	~~-~~

From beliefs to actions

- We have discussed how to compute p(y|x), where y represents the unknown *state of nature* (eg. does the patient have lung cancer, breast cancer or no cancer), and x are some observable features (eg., symptoms)
- We now discuss: what action a should we take (eg. surgery or no surgery) given our beliefs?
-

Loss functions

- Define a loss function L(θ,a), θ=true (unknown) state of nature, a = action

	Surgery	No surgery
No cancer	20	0
Lung cancer	10	50
Breast cancer	10	60

Asymmetric costs

0-1 loss

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	0	1
$y=0$	1	0

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	0	L_{FN}
$y=0$	L_{FP}	0

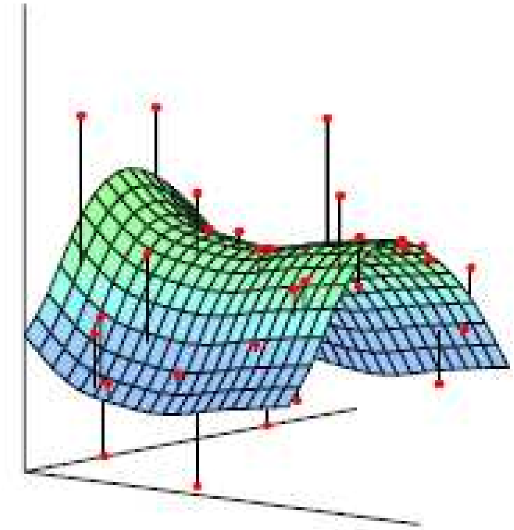
Hypothesis tests

	Accept	Reject
H_0 true	0	L_I
H_1 true	L_{II}	0

Utility = negative loss

More loss functions

- Regression $L(y, \hat{y}) = (y - \hat{y})^2$



- Parameter estimation

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

- Density estimation

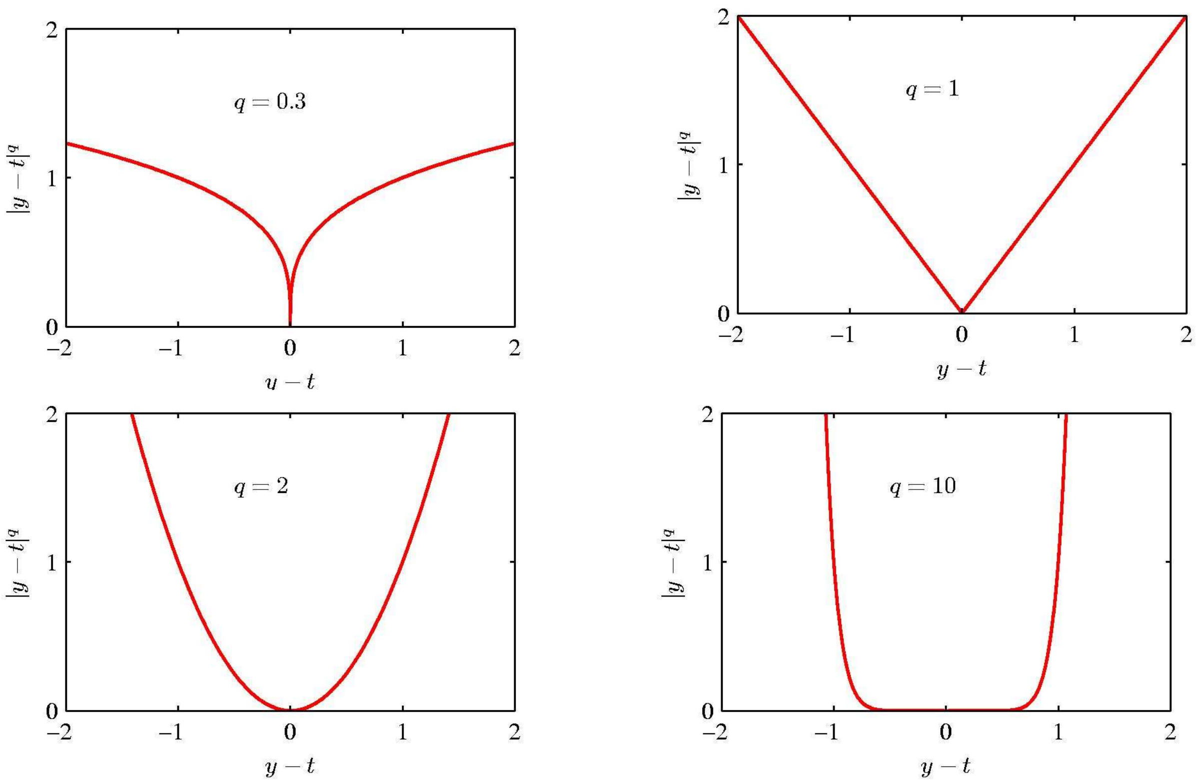
$$L_{KL}(p, q) = \sum_j p(j) \log \frac{p(j)}{q(j)}$$

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = KL(p(\cdot|\boldsymbol{\theta})||p(\cdot|\hat{\boldsymbol{\theta}})) = \int p(y|\boldsymbol{\theta}) \log \frac{p(y|\boldsymbol{\theta})}{p(y|\hat{\boldsymbol{\theta}})} dy$$

Robust loss functions

- Squared error (L2) is sensitive to outliers
- It is common to use L1 instead.
- In general, Lp loss is defined as

$$L_p(y, \hat{y}) = |y - \hat{y}|^p$$



Outline

- Loss functions
- Bayesian decision theory
- Bayesian model selection
- Frequentist decision theory
- Frequentist model selection

Optimal policy

- Minimize posterior expected loss

$$\rho(\mathbf{a}|\mathbf{x},\pi) \stackrel{\text{def}}{=} E_{\boldsymbol{\theta}|\pi,\mathbf{x}}[L(\boldsymbol{\theta},\mathbf{a})] = \int_{\Theta} L(\boldsymbol{\theta},\mathbf{a})p(\boldsymbol{\theta}|\mathbf{x})d\theta$$

- Bayes estimator

$$\delta^{\pi}(\mathbf{x}) = \arg\min_{\mathbf{a}\in\mathcal{A}} \rho(\mathbf{a}|\mathbf{x},\boldsymbol{\pi})$$

L2 loss

- Optimal action is posterior expected mean

$$\begin{aligned} L(\theta, a) &= (\theta - a)^2 \\ \rho(a|\mathbf{x}) &= E_{\theta|\mathbf{x}}[(\theta - a)^2] = E[\theta^2|\mathbf{x}] - 2aE[\theta|\mathbf{x}] + a^2 \\ \frac{\partial}{\partial a}\rho(a|\mathbf{x}) &= -2E[\theta|\mathbf{x}] + 2a = 0 \\ a &= E[\theta|\mathbf{x}] = \int \theta p(\theta|\mathbf{x}) d\theta \end{aligned}$$

$$\hat{y}(\mathbf{x}, \mathcal{D}) = E[y|\mathbf{x}, \mathcal{D}]$$

Minimizing robust loss functions

- For L2 loss, mean $p(y|x)$
- For L1 loss, median $p(y|x)$
- For L0 loss, mode $p(y|x)$

0-1 loss

- Optimal action is most probable class

$$
\begin{aligned}
L(\theta, a) &= 1 - \delta_\theta(a) \\
\rho(a|\mathbf{x}) &= \int p(\theta|\mathbf{x}) d\theta - \int p(\theta|\mathbf{x}) \delta_\theta(a) d\theta \\
&= 1 - p(a|\mathbf{x}) \\
a^*(\mathbf{x}) &= \arg\max_{a \in \mathcal{A}} p(a|\mathbf{x}) \\
\hat{y}(\mathbf{x}, \mathcal{D}) &= \arg\max_{y \in 1:C} p(y|\mathbf{x}, \mathcal{D})
\end{aligned}
$$

Binary classification problems

- Let Y=1 be 'positive' (eg cancer present) and Y=2 be 'negative' (eg cancer absent).
- The loss/ cost matrix has 4 numbers:

state y

action $\hat{y}$

	1	2
1	True positive λ_{11}	False positive λ_{12}
2	False negative λ_{21}	True negative λ_{22}

Optimal strategy for binary classification

- We should pick class/ label/ action 1 if

$$\begin{aligned}
\rho(\alpha_2|\mathbf{x}) &> \rho(\alpha_1|\mathbf{x}) \\
\lambda_{21}p(Y=1|\mathbf{x}) + \lambda_{22}p(Y=2|\mathbf{x}) &> \lambda_{11}p(Y=1|\mathbf{x}) + \lambda_{12}p(Y=2|\mathbf{x}) \\
(\lambda_{21}-\lambda_{11})p(Y=1|\mathbf{x}) &> (\lambda_{12}-\lambda_{22})p(Y=2|\mathbf{x}) \\
\frac{p(Y=1|\mathbf{x})}{p(Y=2|\mathbf{x})} &> \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}
\end{aligned}$$

  where we have assumed λ_{21} (FN) $>\lambda_{11}$ (TP)

- As we vary our loss function, we simply change the optimal threshold θ on the decision rule

$$\delta(x) = 1 \text{ iff } \frac{p(Y=1|x)}{p(Y=2|x)} > \theta$$

Definitions

- Declare x_n to be a positive if $p(y=1|x_n)>\theta$, otherwise declare it to be negative (y=2)

$$\hat{y}_n = 1 \iff p(y=1|x_n) > \theta$$

- Define the number of true positives as

$$TP = \sum_n I(\hat{y}_n = 1 \wedge y_n = 1)$$

- Similarly for FP, TN, FN – all functions of θ

$\hat{y}$ \ y	1	2	
1	TP	FP	$\hat{P}$
2	FN	TN	$\hat{N}$
	P	N	

Performance measures

		Truth		
		1	0	Σ
Estimate	1	TP	FP	$\hat{P} = TP + FP$
	0	FN	TN	$\hat{N} = FN + TN$
	Σ	$P = TP + FN$	$N = FP + TN$	$n = TP + FP + FN + TN$

	$y = 1$	$y = 0$
$\hat{y} = 1$	$TP/\hat{P}$=precision=PPV	$FP/\hat{P}$=FDP
$\hat{y} = 0$	$FN/\hat{N}$	$TN/\hat{N}$=NPV

Normalize along rows P(y|yhat)

Normalize along cols P(yhat|y)

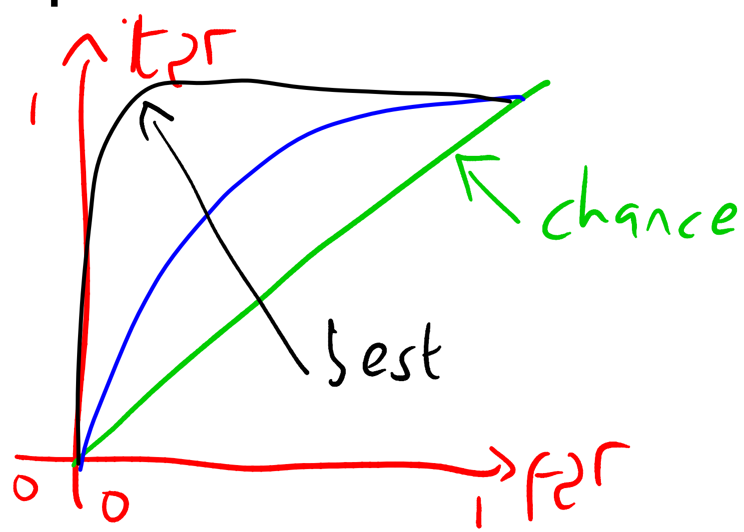
	$y = 1$	$y = 0$
$\hat{y} = 1$	TP/P=TPR=sensitivity=recall	FP/N=FPR
$\hat{y} = 0$	FN/P=FNR	TN/N=TNR=speci┐ty

ROC curves

- The optimal threshold for a binary detection problem depends on the loss function

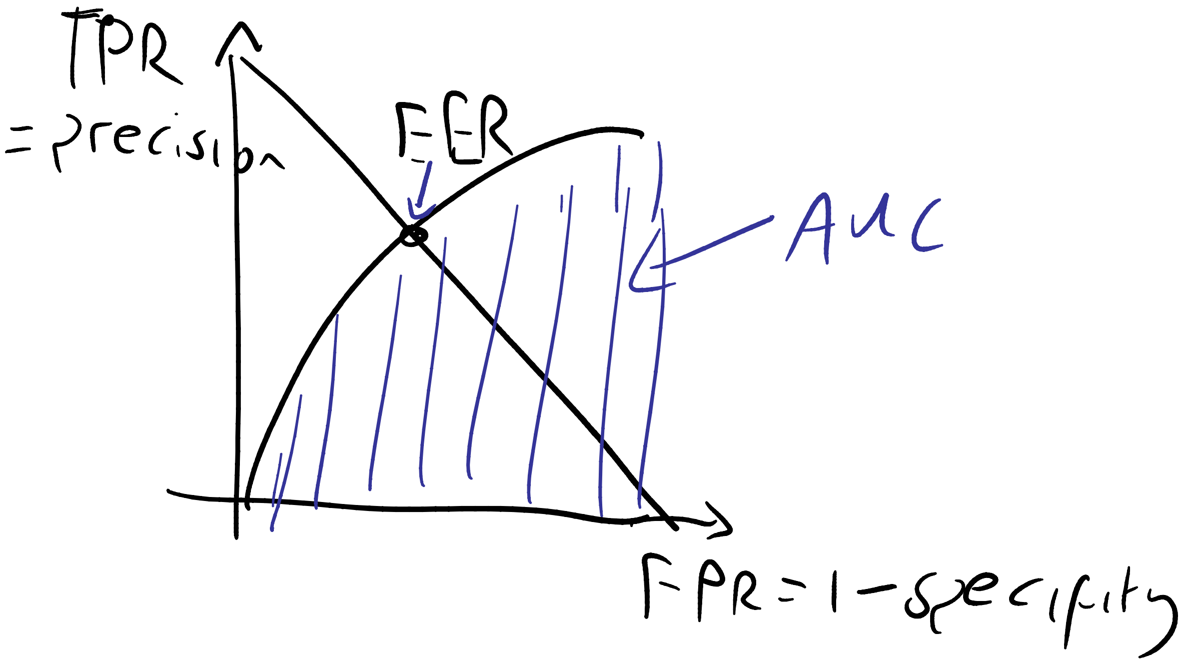
$$\delta(x) = 1 \quad \iff \quad \frac{p(Y=1|\mathbf{x})}{p(Y=2|\mathbf{x})} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$$

- Low threshold will give rise to many false positives (Y=1) and high threshold to many false negatives.
- A receive operating characteristic (ROC) curves plots the true positive rate vs false positive rate as we vary θ



Reducing ROC curve to 1 number

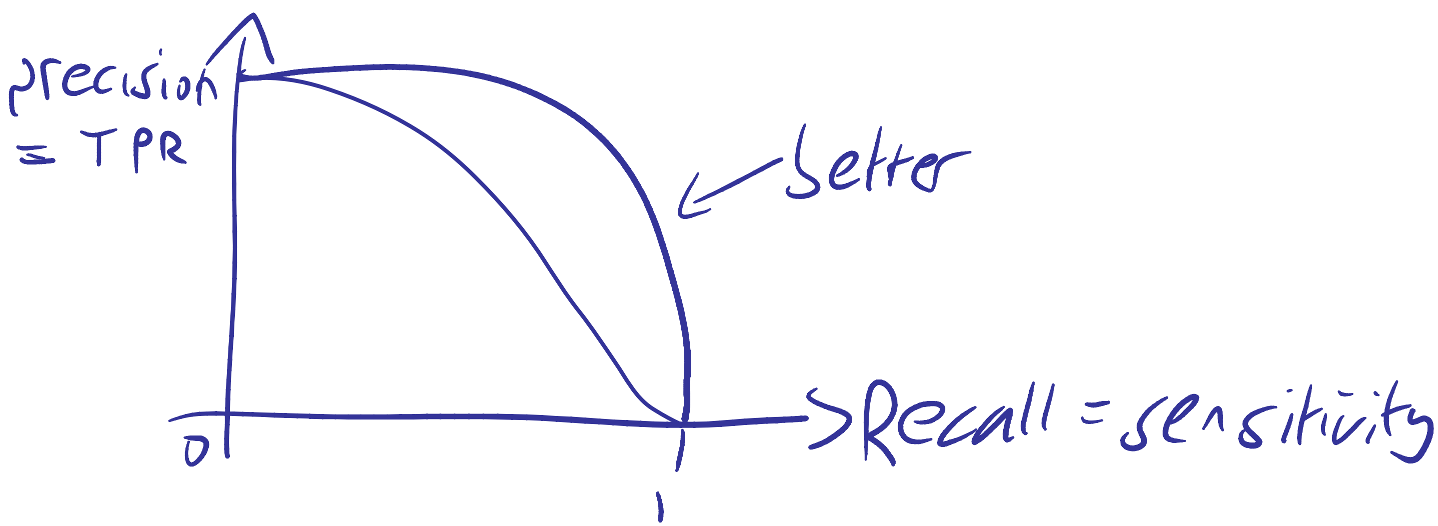
- EER- Equal error rate (precision=specificity)
- AUC - Area under curve



Precision-recall curves

- Useful when notion of "negative" (and hence FPR) is not defined
- Used to evaluate retrieval engines
- Recall = of those that exist, how many did you find?
- Precision = of those that you found, how many correct?
- F-score is geometric mean

$$F = \frac{2}{1/P + 1/R} = \frac{2PR}{R + P}$$

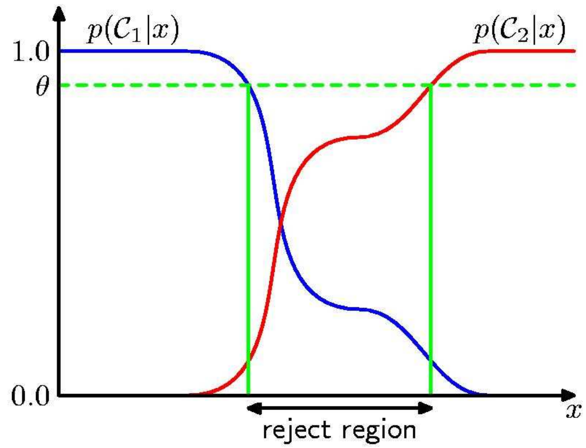


Reject option

- Suppose we can choose between incurring loss λ_s if we make a misclassification (label substitution) error and loss λ_r if we declare the action "don't know"

$$\lambda(\alpha_i|Y=j) = \begin{cases} 0 & \text{if } i=j \text{ and } i,j \in \{1,\ldots,C\} \\ \lambda_r & \text{if } i = C+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

- In HW5, you will show that the optimal action is to pick "don't know" if the most probable class is below a threshold $1-\lambda_r/\lambda_s$



Bishop 1.26

Discriminant functions

- The optimal strategy $\pi(x)$ partitions X into decision regions R_i, defined by discriminant functions $g_i(x)$

$$\pi(x) = \arg\max_i g_i(x)$$

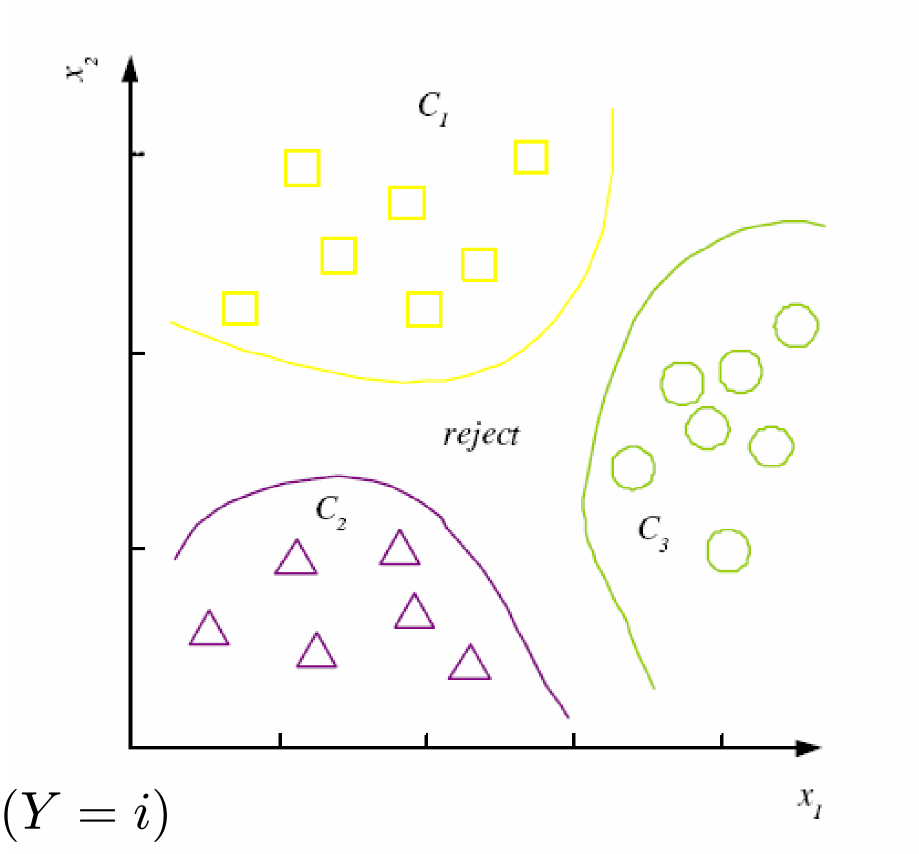
$$R_i = \{x : g_i(x) = \max_k g_k(x)\}$$

In general

$$g_i(x) = -R(a = i|x)$$

But for 0-1 loss we have

$$\begin{aligned} g_i(x) &= p(Y = i|x) \\ &= \log p(Y = i|x) \\ &= \log p(x|Y = i) + \log p(Y = i) \end{aligned}$$

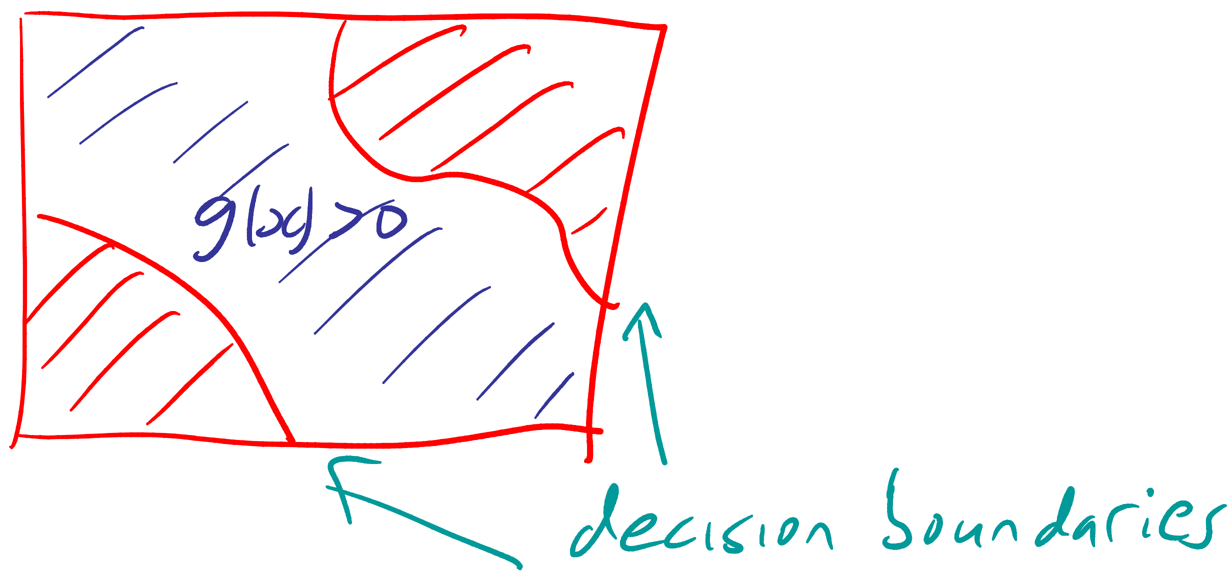


Class prior merely shifts decision boundary by a constant

Binary discriminant functions

- In the 2 class case, we define the discriminant in terms of the log-odds ratio

$$
\begin{aligned}
g(x) &= g_1(x) - g_2(x) \\
&= \log p(Y=1|x) - \log p(Y=2|x) \\
&= \log \frac{p(Y=1|x)}{p(Y=2|x)}
\end{aligned}
$$



Outline

- Loss functions
- Bayesian decision theory
- Bayesian model selection
- Frequentist decision theory
- Frequentist model selection

Bayesian model selection

- 0-1 loss

$$
\begin{aligned}
L(m, \hat{m}) &= I(m \neq \hat{m}) \\
m^* &= \arg \max_{m \in \mathcal{M}} p(m|\mathcal{D})
\end{aligned}
$$

- KL loss

$$
\begin{aligned}
L(p_*, m) &= KL(p_*(y|\mathbf{x}), p(y|m, \mathbf{x}, \mathcal{D})) \\
\rho(m|\mathbf{x}) &= EKL(p_*, p_m) = E[p_* \log p_* - p_* \log p_m] \\
\bar{p} &= Ep_* = \sum_{m \in m} p_m p(m|\mathcal{D}) \\
m^* &= \arg \min_{m \in \mathcal{M}} KL(\bar{p}, p_m)
\end{aligned}
$$

Posterior over models

- Key quantity

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m' \in \mathcal{M}} p(\mathcal{D}|m')p(m')}$$

- Marginal / integrated likelihood

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|m, \boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

Example: is the coin biased?

- Model M0: theta= 0.5

$$p(\mathcal{D}|m_0) = \frac{1}{2}^n$$

- Model M1: theta could be any value in [0,1] (includes 0.5 but with negligible probability)

$$\begin{aligned} p(\mathcal{D}|m_1) &= \int p(\mathcal{D}|\theta)p(\theta)d\theta \\ &= \int [\prod_{i=1}^{n} \text{Ber}(x_i|\theta)] \text{Beta}(\theta|\alpha_0, \alpha_1)d\theta \end{aligned}$$

Computing the marginal likelihood

- For the Beta-Bernoulli model, we know the posterior is Beta(θ|α_1',α_0') so

$$
\begin{aligned}
p(\theta|\mathcal{D}) &= \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} \\
&= \frac{1}{p(\mathcal{D})}\left[\frac{1}{B(\alpha_1,\alpha_0)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\right]\left[\theta^{N_1}(1-\theta)^{N_0}\right] \\
&= \frac{1}{p(\mathcal{D})}\frac{1}{B(\alpha_1,\alpha_0)}\left[\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\theta^{N_1}(1-\theta)^{N_0}\right] \\
&= \frac{1}{B(\alpha'_1,\alpha'_0)}\left[\theta^{\alpha'_1-1}(1-\theta)^{\alpha'_0-1}\right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{p(\mathcal{D})}\frac{1}{B(\alpha_1,\alpha_0)} &= \frac{1}{B(\alpha'_1,\alpha'_0)} \\
p(\mathcal{D}) &= \frac{B(\alpha'_1,\alpha'_0)}{B(\alpha_1,\alpha_0)}
\end{aligned}
$$

ML for Dirichlet-multinomial model

- Normalization constant is

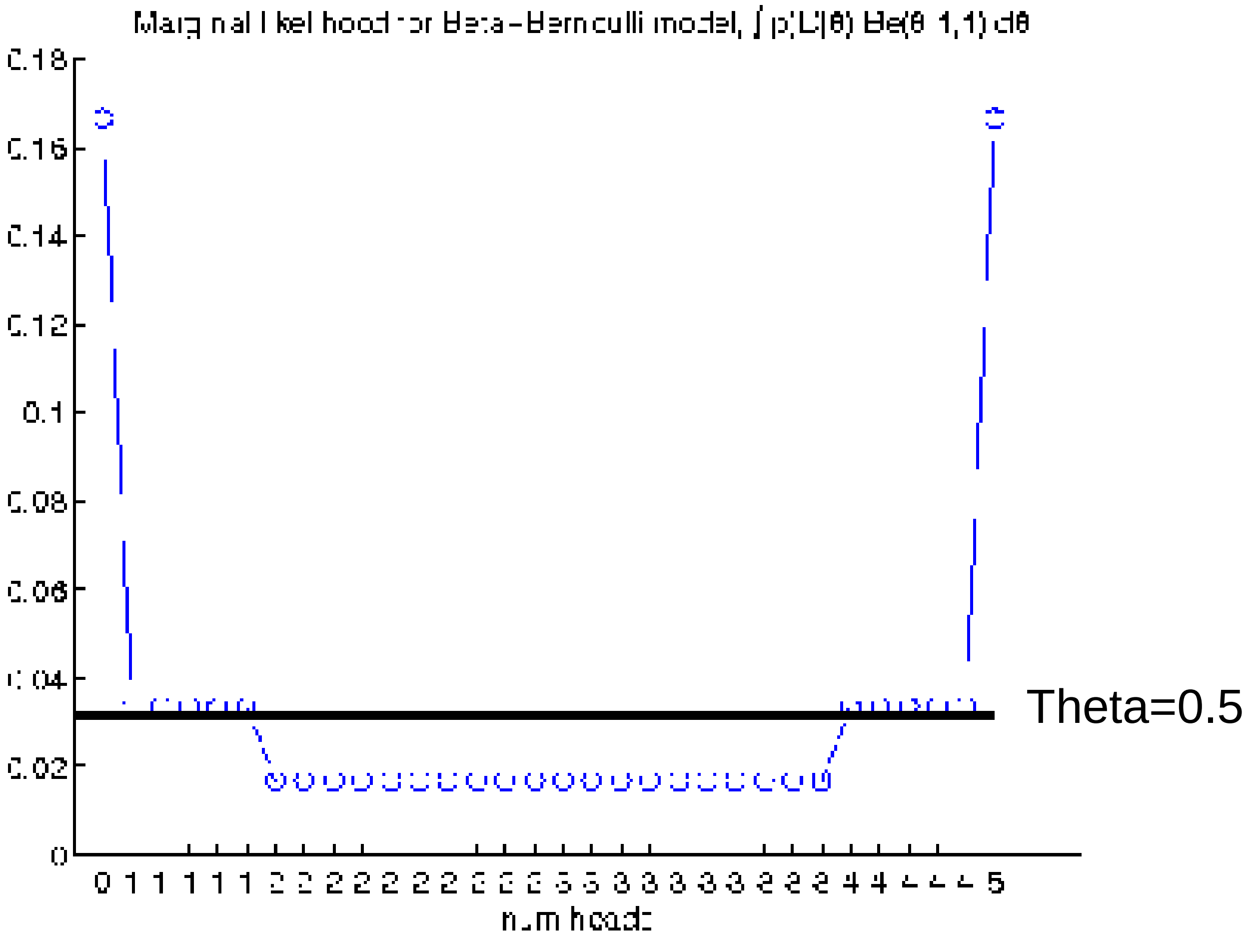
$$Z_{Dir}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}$$

- Hence marg lik is

$$p(\mathcal{D}) = \frac{Z_{Dir}(\mathbf{N} + \boldsymbol{\alpha})}{Z_{Dir}(\boldsymbol{\alpha})} = \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)}$$

ML for biased coin

- P(D|M_1) for $\alpha_0=\alpha_1=1$

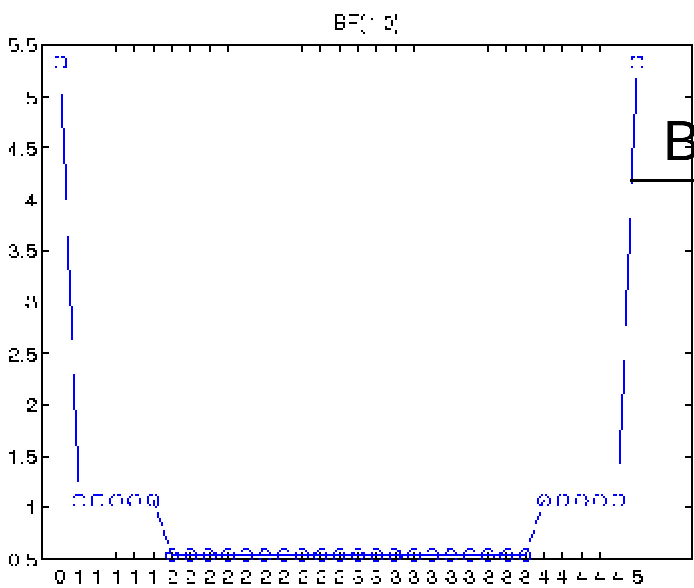


If nheads = 2 or 3, M1 is less likely than M0

Bayes factors

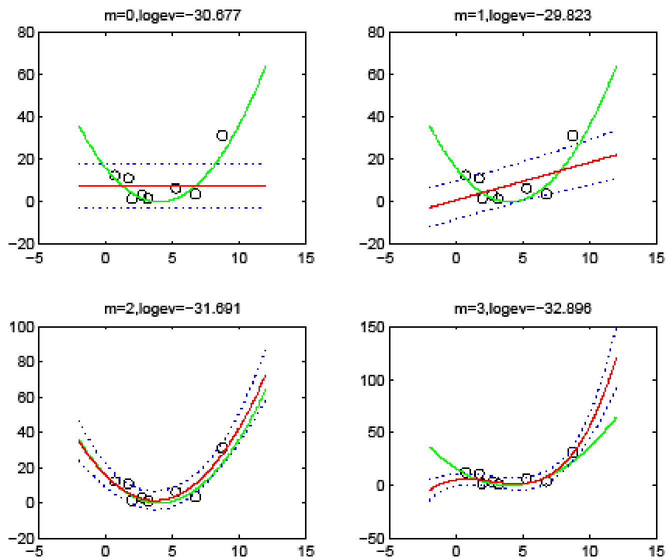
$$BF(M_i, M_j) = \frac{p(\mathcal{D}|M_i)}{p(\mathcal{D}|M_j)} = \frac{p(M_i|\mathcal{D})}{p(M_j|\mathcal{D})} \Big/ \frac{p(M_i)}{p(M_j)}$$

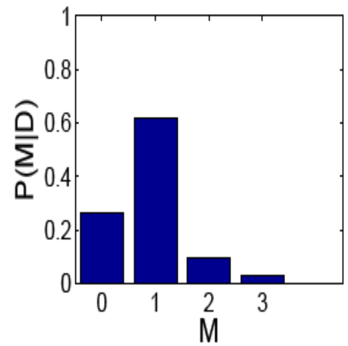
$$BF(M_1, M_0) = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} \frac{1}{0.5^N}$$

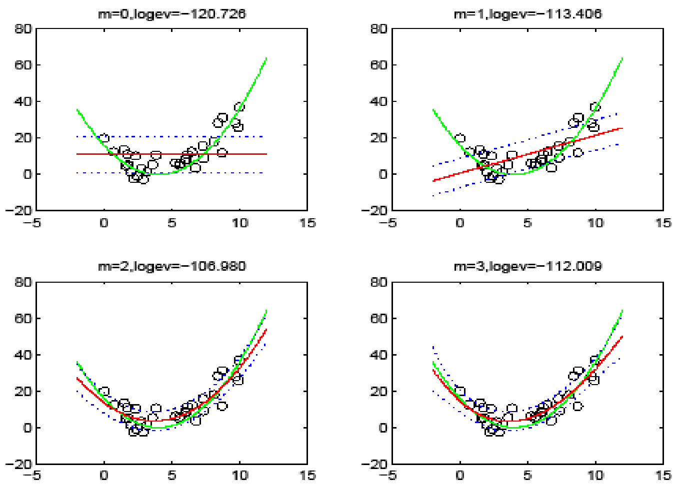


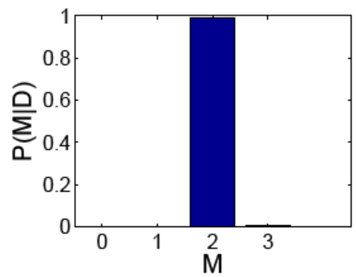
Bayes factor $BF(1,0)$	Interpretation
$B < \frac{1}{10}$	Strong evidence for H_0
$\frac{1}{10} < B < \frac{1}{3}$	Moderate evidence for H_0
$\frac{1}{3} < B < 1$	Weak evidence for H_0
$1 < B < 3$	Weak evidence for H_1
$3 < B < 10$	Moderate evidence for H_1
$B > 10$	Strong evidence for H_1

Polynomial regression



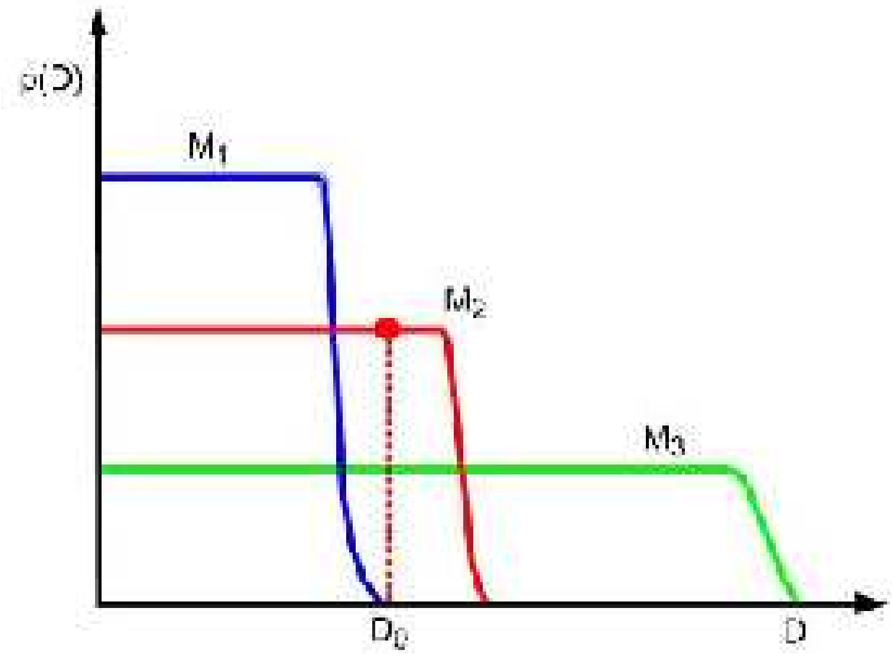






Bayesian Ockham's razor

- Marginal likelihood automatically penalizes complex models due to sum-to-one constraint



BIC

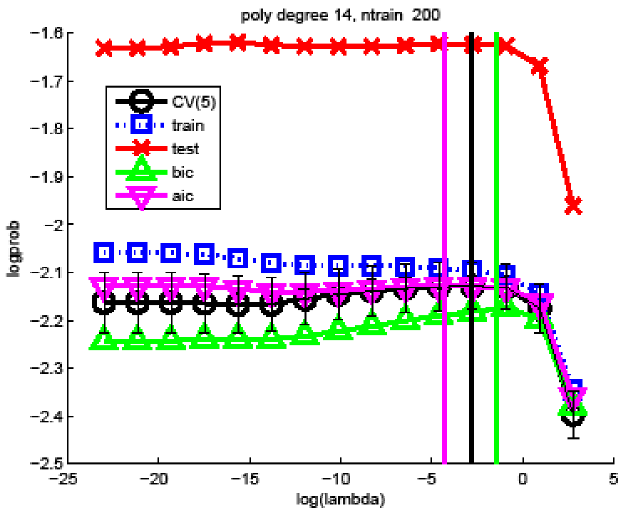
- Computing the marginal likelihood is hard unless we have conjugate priors.
- One popular approach is to make a Laplace approx to the posterior and then approximate the log normalizer

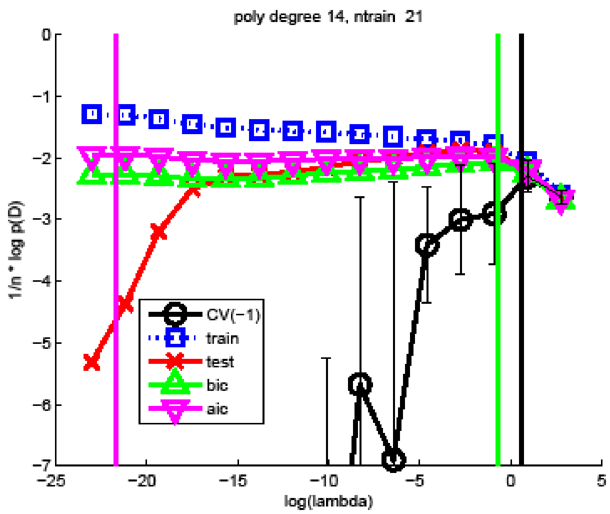
$$
\begin{aligned}
p(\mathcal{D}) &\approx p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{map})p(\hat{\boldsymbol{\theta}}_{map})(2\pi)^{d/2}|\mathbf{C}|^{\frac{1}{2}} \\
\mathbf{C} &= -\mathbf{H}^{-1} \\
|\mathbf{H}| &\approx n^{\text{dof}} \\
\log p(\mathcal{D}) &\approx \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{1}{2}\text{dof} \log n
\end{aligned}
$$

BIC vs CV for ridge

- Define dof in terms of singular values

$$df(\lambda) = \sum_{j=1}^{d} \frac{d_j^2}{d_j^2 + \lambda}$$





Outline

- Loss functions
- Bayesian decision theory
- Bayesian model selection
- Frequentist decision theory
- Frequentist model selection

Frequentist decision theory

- Risk function

$$R(\boldsymbol{\theta}, \delta) = E_{\mathbf{x}|\boldsymbol{\theta}} L(\boldsymbol{\theta}, \delta(\mathbf{x})) = \int_{\mathcal{X}} L(\boldsymbol{\theta}, \delta(\mathbf{x})) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

- Example: L2 loss

$$MSE = E_{\mathcal{D}|\theta_0} (\hat{\theta}(\mathcal{D}) - \theta_0)^2$$

- Assumes that true parameter θ_0 is known, and averages over data

Bias/variance tradeoff

$$\begin{aligned} MSE &= E(\hat{\theta}(\mathcal{D}) - \theta_0)^2 \\ &= E(\hat{\theta}(\mathcal{D}) - \overline{\theta} + \overline{\theta} - \theta_0)^2 \\ &= E(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + 2(\overline{\theta} - \theta_0)E(\hat{\theta}(\mathcal{D}) - \overline{\theta}) + (\overline{\theta} - \theta_0)^2 \\ &= E(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + (\overline{\theta} - \theta_0)^2 \\ &= \text{Var}\ (\hat{\theta}) + \text{bias}^2(\hat{\theta}) \end{aligned}$$

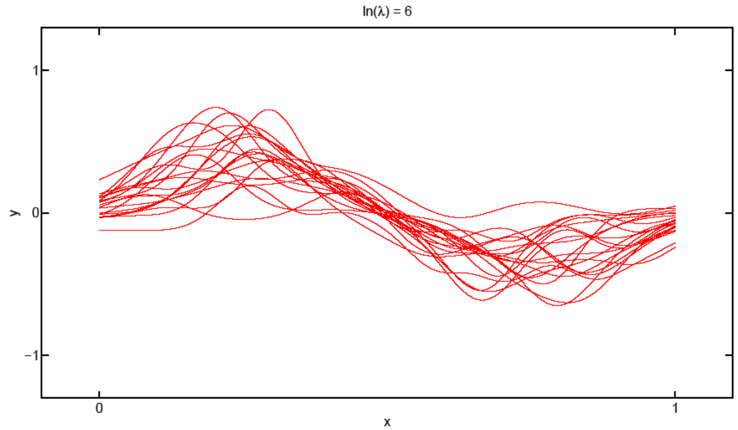
$$\text{bias}^2 \approx \frac{1}{n}\sum_{i=1}^{n}(\overline{y}(\mathbf{x}_i) - f_{true}(\mathbf{x}_i))^2$$

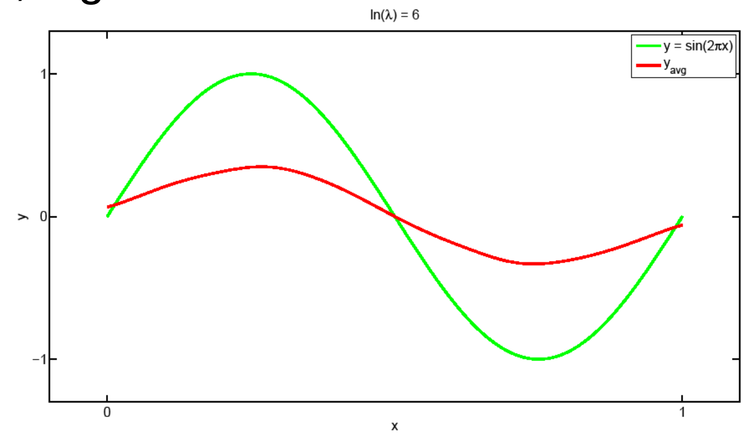
$$\text{var} \approx \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{S}\sum_{s=1}^{S}(y^s(\mathbf{x}_i) - \overline{y}(\mathbf{x}_i))^2\right]$$

Average over S training sets drawn from true dist.

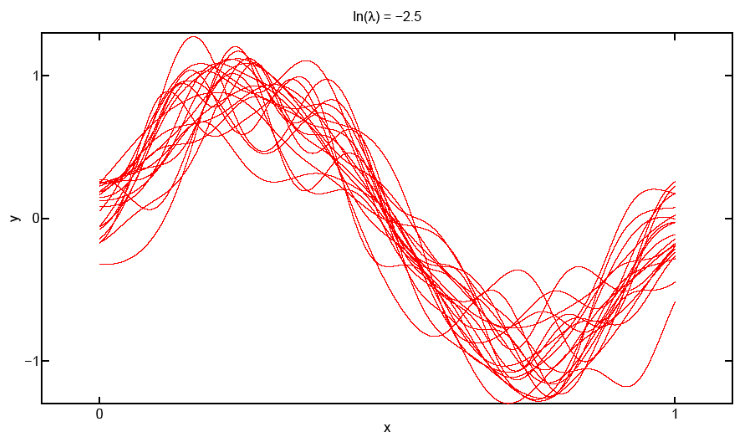
Bias/variance tradeoff

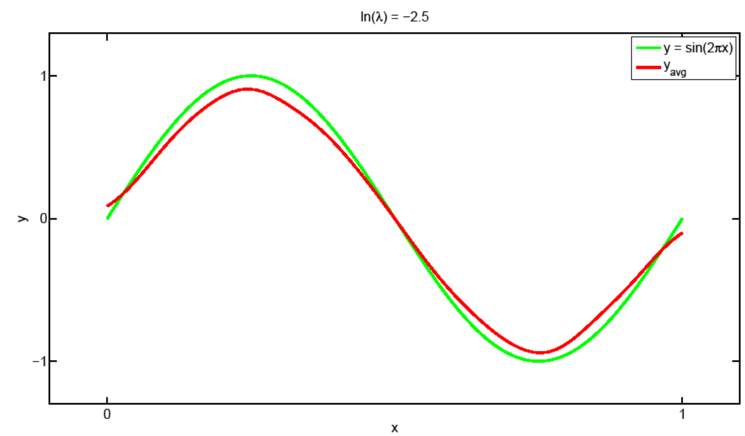
$\lambda = e^{6}$: low variance, high bias



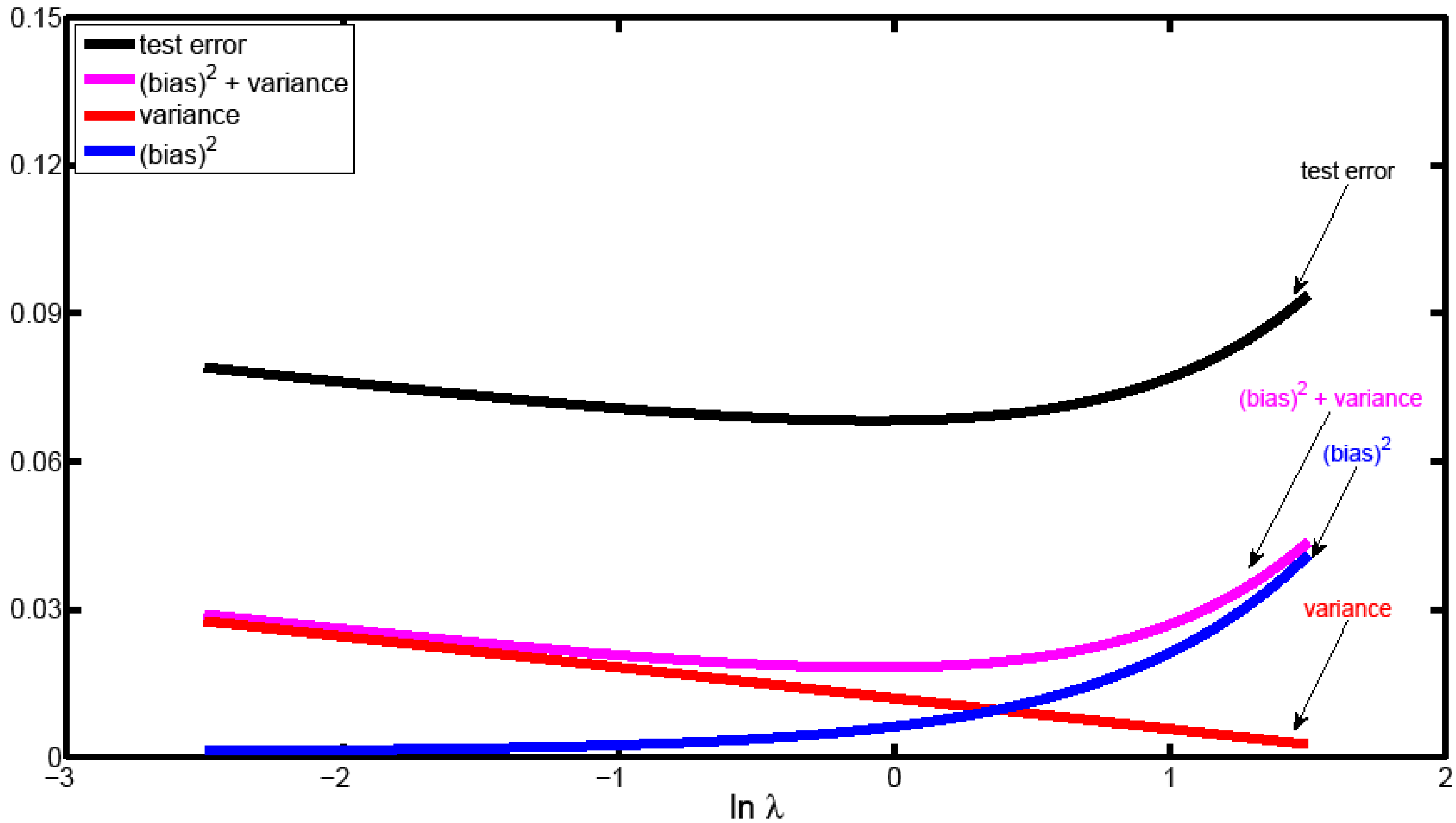


$\lambda = e^{-2.5}$: high variance, low bias





Bias/ variance tradeoff



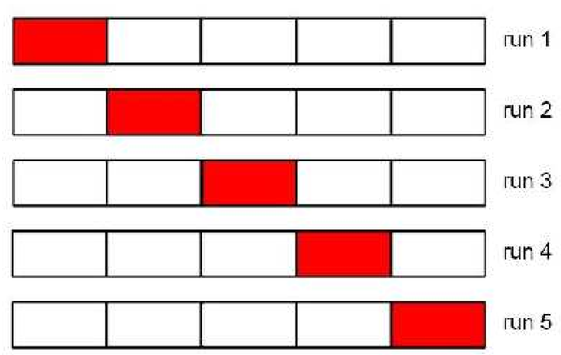
Empirical risk minimization

- Risk for function approximation

$$R(\boldsymbol{\pi}, \hat{f}(\cdot)) = E_{(\mathbf{x},y)\sim\boldsymbol{\pi}} L(y, \hat{f}(\mathbf{x})) = \int p(y, \mathbf{x}|\boldsymbol{\pi}) L(y, \hat{f}(\mathbf{x})) d\mathbf{x} d$$

$$\hat{R}(\hat{f}(\cdot), \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i))$$

- To avoid overly optimistic estimate, can use bootstrap resampling or cross validation

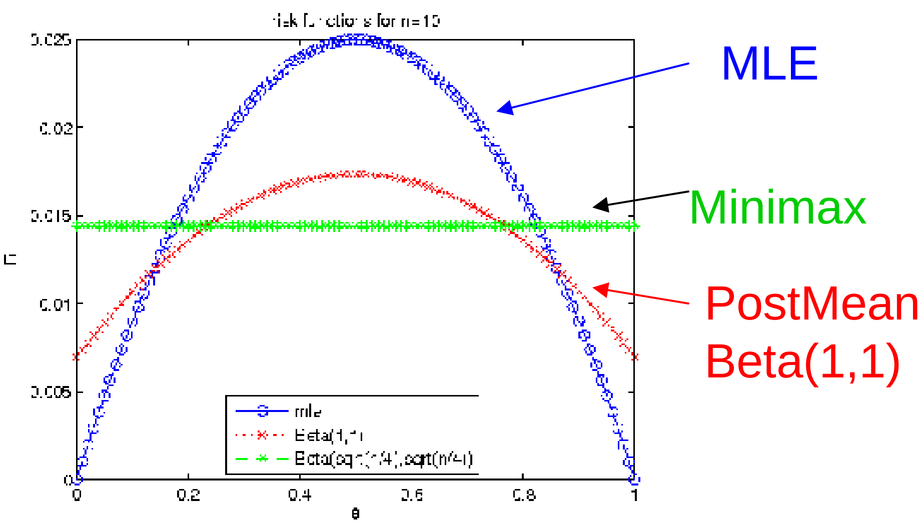


Risk functions for parameter estimation

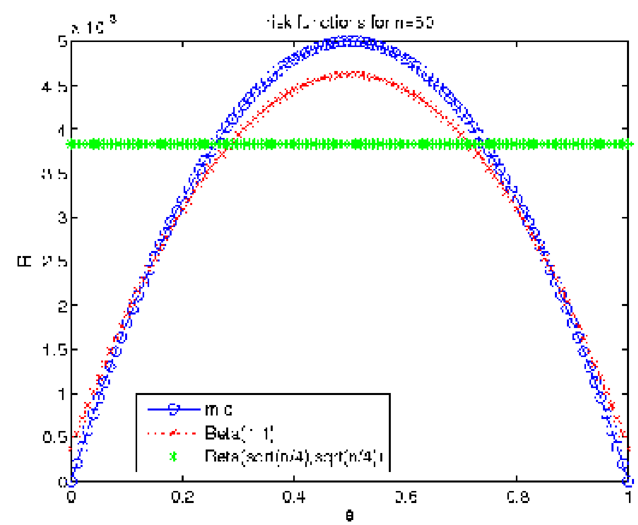
- Risk function depends on unknown theta

$$R(\boldsymbol{\theta}, \delta) = E_{\mathbf{x}|\boldsymbol{\theta}} L(\boldsymbol{\theta}, \delta(\mathbf{x})) = \int_{\mathcal{X}} L(\boldsymbol{\theta}, \delta(\mathbf{x})) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

$$X_i \sim \text{Ber}(\theta), L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$



N=10



N=50

Summarizing risk functions

- Risk function

$$R(\boldsymbol{\theta}, \delta) = E_{\mathbf{x}|\boldsymbol{\theta}} L(\boldsymbol{\theta}, \delta(\mathbf{x})) = \int_{\mathcal{X}} L(\boldsymbol{\theta}, \delta(\mathbf{x})) p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$$

- Minimax risk – very pessimistic

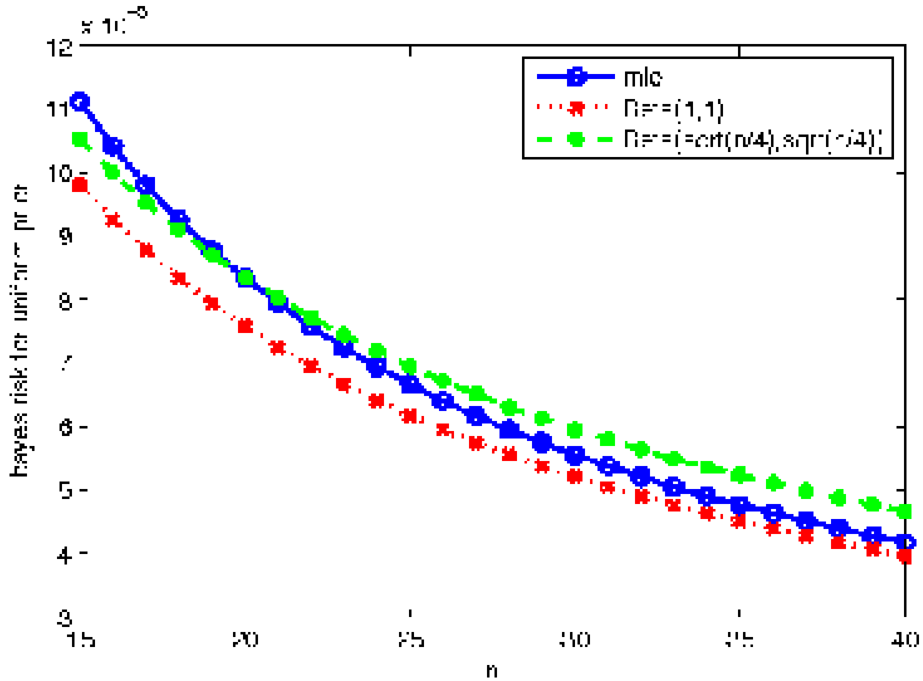
$$R^{max}(\delta) = \max_{\theta \in \Theta} R(\theta, \delta)$$

- Bayes risk – requires a prior over theta

$$R^{\pi}(\delta) = E_{\theta|\pi} R(\theta, \delta) = \int_{\Theta} R(\theta, \delta) \pi(\theta) d\theta$$

Bayes risk vs n

$$X_i \sim \text{Ber}(\theta), L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2, \pi(\theta) = U$$



Bayes meets frequentist

- To minimize the Bayes risk, minimize the posterior expected loss

$$
\begin{aligned}
R^\pi(\delta) &= \int_\Theta \left[\int_\mathcal{X} L(\theta, \delta(\mathbf{x})) p(\mathbf{x}|\theta) d\mathbf{x} \right] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int_\mathcal{X} \int_\Theta L(\theta, \delta(\mathbf{x})) p(\mathbf{x}|\theta) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{x} \\
&= \int_\mathcal{X} \left[\int_\Theta L(\theta, \delta(\mathbf{x})) p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} \right] p(\mathbf{x}) d\mathbf{x} \\
&= \int_\mathcal{X} \rho(\delta(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

To minimize the integral, minimize $\rho(\delta(x)|x)$ for each x.

Bayesian estimators have good frequentist properties.

Outline

- Loss functions
- Bayesian decision theory
- Bayesian model selection
- Frequentist decision theory
- Frequentist model selection

Frequentist model selection

- 0-1 loss: classical hypothesis testing, not covered in this class (similar to, but more complex than, Bayesian case)
- Predictive loss: minimize empirical risk, or CV/ bootstrap approximation thereof

$$\hat{R}(m) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, m(\mathbf{x}_i))$$