CS 461: Machine Learning
Lecture 6

Dr. Kiri Wagstaff
wkiri@wkiri.com

Plan for Today

- Solution to Midterm
- Solution to Homework 3

- Parametric methods
- Bayes estimation
- Parametric classification

Review from Lecture 5

- Probability
  - Axioms
- Bayesian Learning
  - Classification
  - Bayes's Rule
  - Bayesian Networks
  - Naïve Bayes Classifier
  - Association Rules

Parametric Methods

Chapter 4

Parametric Learning

- Assume: data x comes from a distribution $p(x)$
- Model this distribution by selecting parameters θ
  - e.g., $\mathcal{N}(\mu, \sigma^2)$ where $\theta = \{\mu, \sigma^2\}$

Maximum Likelihood Model

- Likelihood of θ given the sample $\mathcal{X}$

  $l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \Pi_t p(x^t|\theta)$

- Log likelihood

  $\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \Sigma_t \log p(x^t|\theta)$

- Maximum likelihood estimator (MLE)

  $\theta^* = \text{argmax}_\theta \mathcal{L}(\theta|\mathcal{X})$

Example: do you wear glasses?

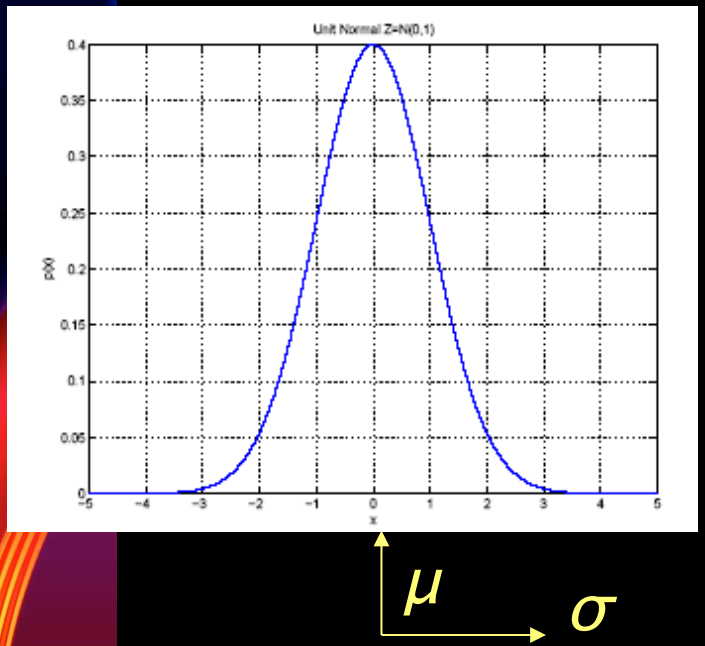
- Bernoulli: Two states, x in {0,1}

  $P(x) = p_o^x (1 - p_o)^{(1-x)}$

  $\mathcal{L}(p_o|\mathcal{X}) = \log \Pi_t p_o^{x^t} (1 - p_o)^{(1-x^t)}$

  MLE: $p_o = \Sigma_t x^t / N$

Gaussian (Normal) Distribution



- $p(x) = \mathcal{N}(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- MLE for μ and σ^2:

$$m = \frac{\sum_t x^t}{N}$$

$$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

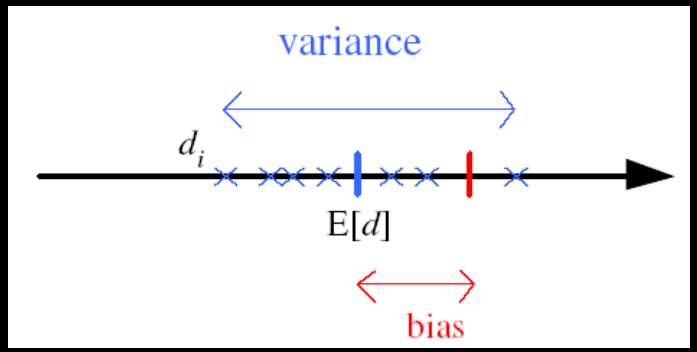
How good is that estimate?

- Let d be the estimate of θ
  - It is also a random variable
  - Technically, it's $d(X)$ since it depends on X
  - $E[d]$ is the expected value of d (regardless of X)

- Bias: $b_\theta(d) = E[d] - \theta$
  - How far off the correct value is it?
- Variance: $Var(d) = E[(d-E[d])^2]$
  - How much does it change with different X?

- Mean square error:

$$r(d,\theta) = E[(d-\theta)^2]$$
$$= (E[d] - \theta)^2 + E[(d-E[d])^2]$$
$$= \text{Bias}^2 + \text{Variance}$$



Bayes Estimator: Using what we already know

- Prior information $p(\theta)$
- Bayes's rule (get posterior):
  $p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta)\, p(\theta) / p(\mathcal{X})$
- Maximum a Posteriori (MAP):
  $\theta_{MAP} = \text{argmax}_\theta\, p(\theta|\mathcal{X})$
- Maximum Likelihood (ML):
  $\theta_{ML} = \text{argmax}_\theta\, p(\mathcal{X}|\theta)$

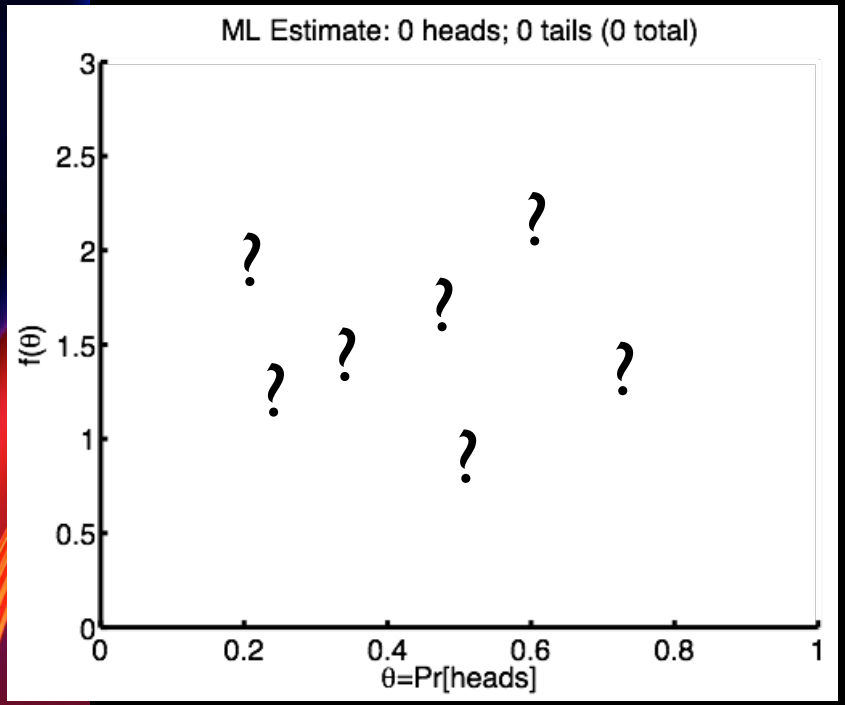
Bayes Estimator: Continuous Example

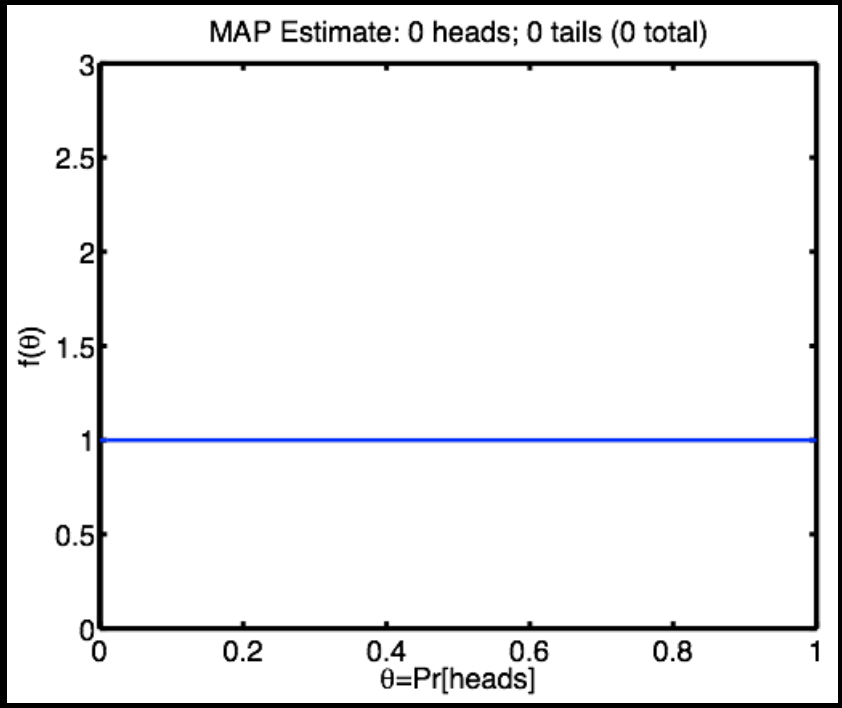
- Assume $x^t \sim \mathcal{N}(\theta, \sigma_0^2)$ and $\theta \sim \mathcal{N}(\mu, \sigma^2)$
- $\theta_{ML} = m$ (sample mean)
- $\theta_{MAP} =$

$$E[\theta \mid \mathcal{X}] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$$

- Estimated mean = weighted average of sample mean m and prior mean μ
  - Weights indicate how much you trust the sample

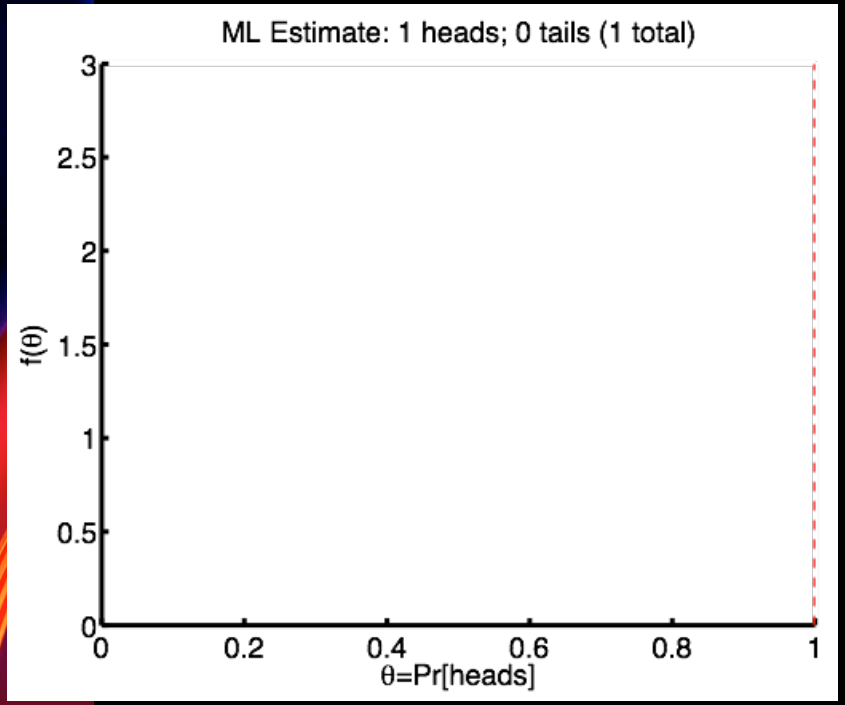
Example: Coin flipping

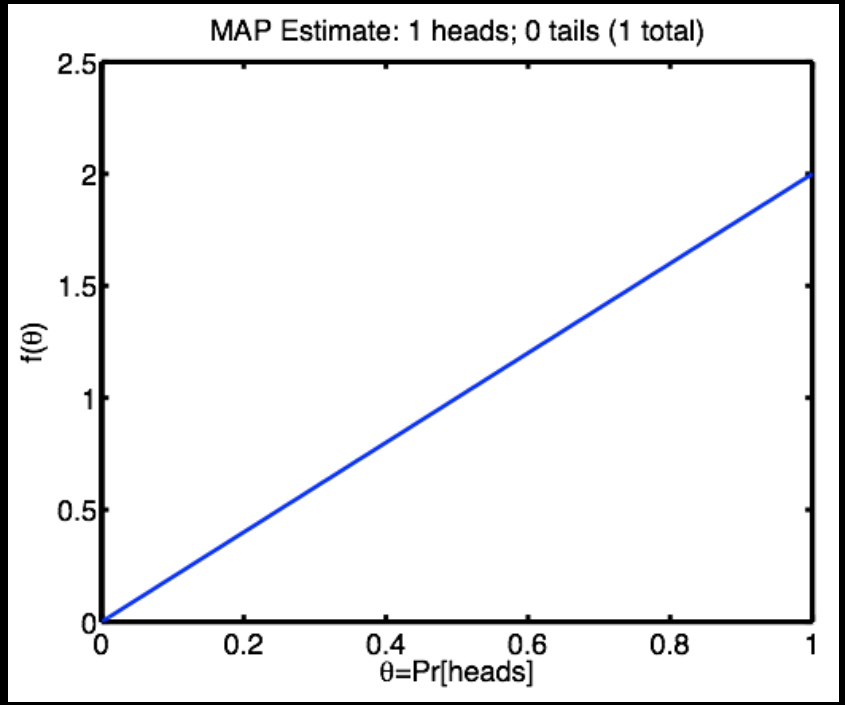




0 flips total

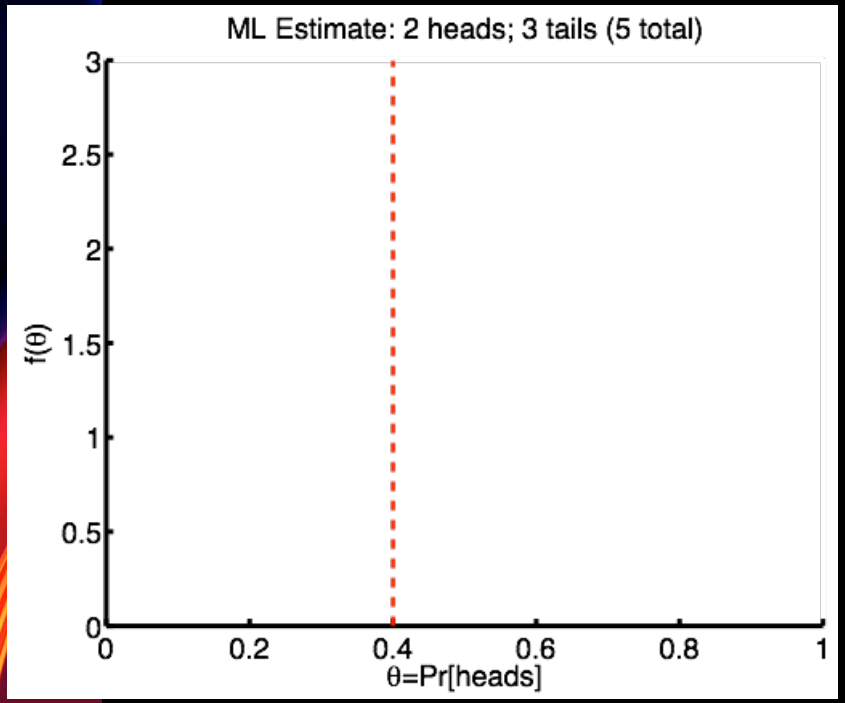
Example: Coin flipping

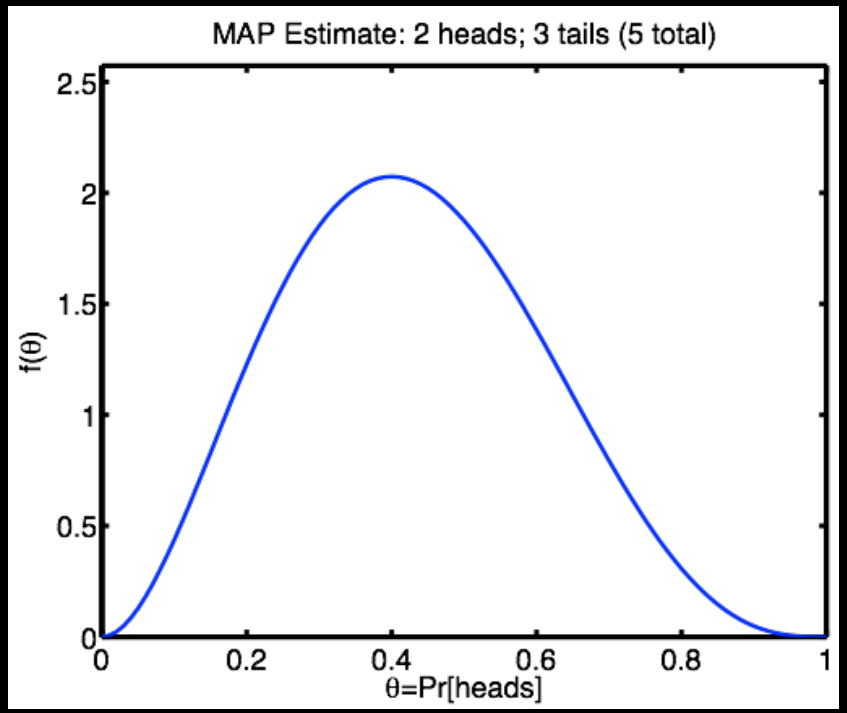




1 flip total

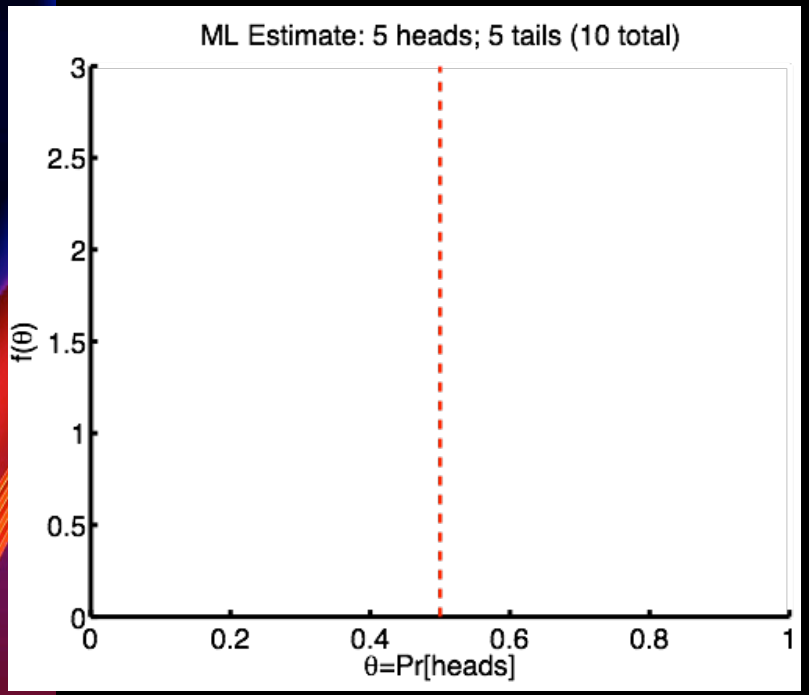
Example: Coin flipping

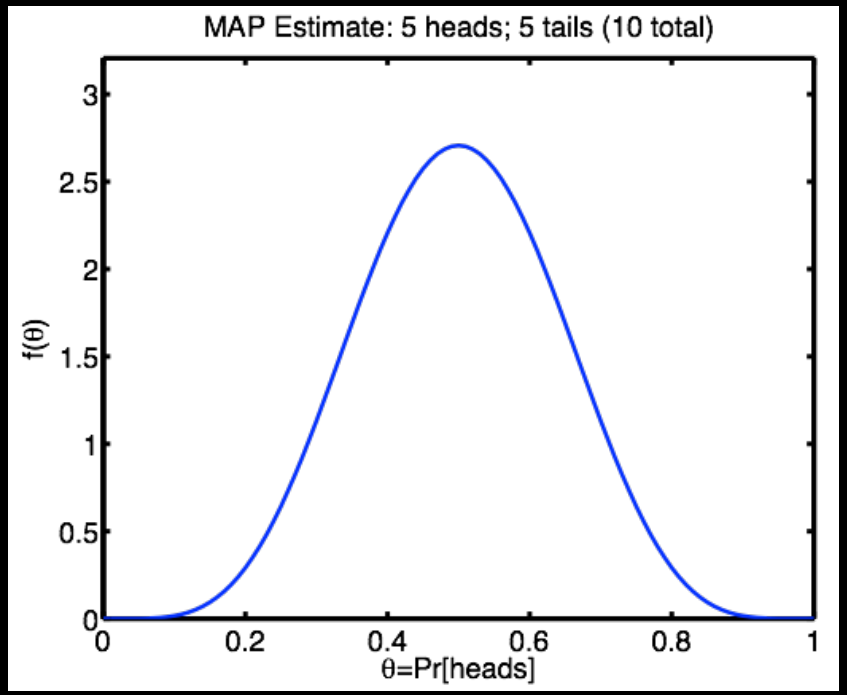




5 flips total

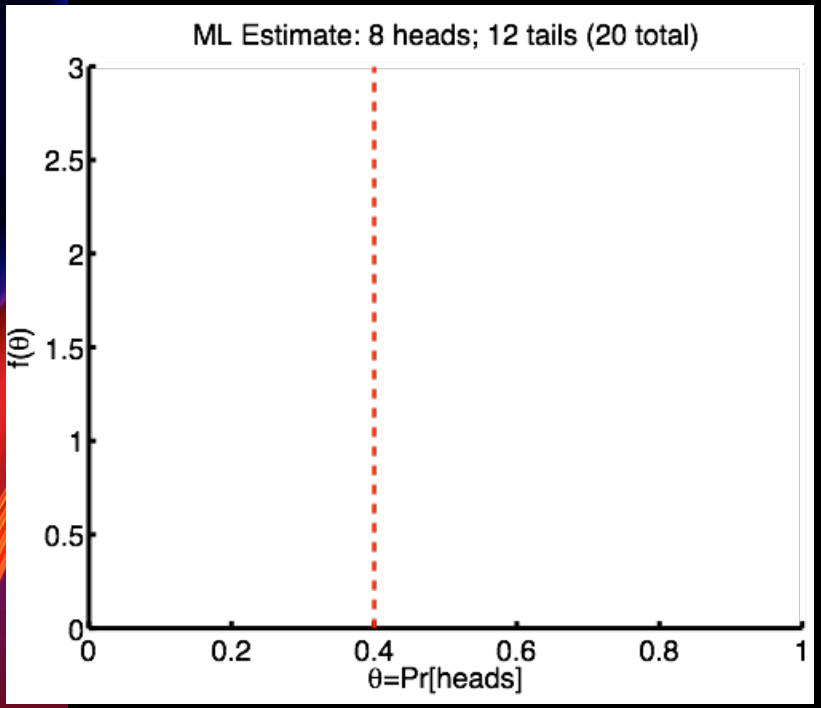
Example: Coin flipping

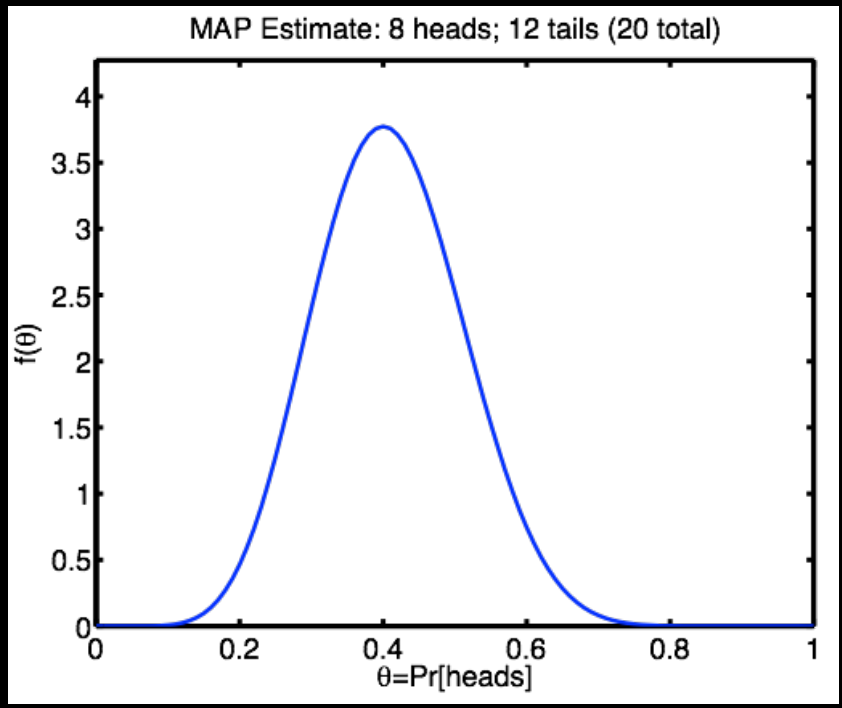




10 flips total

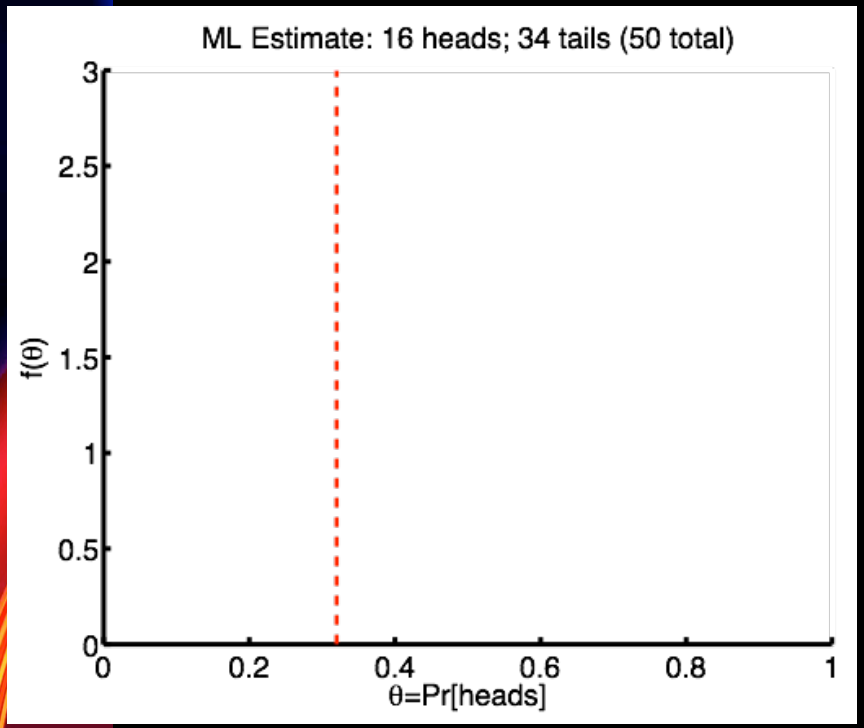
Example: Coin flipping

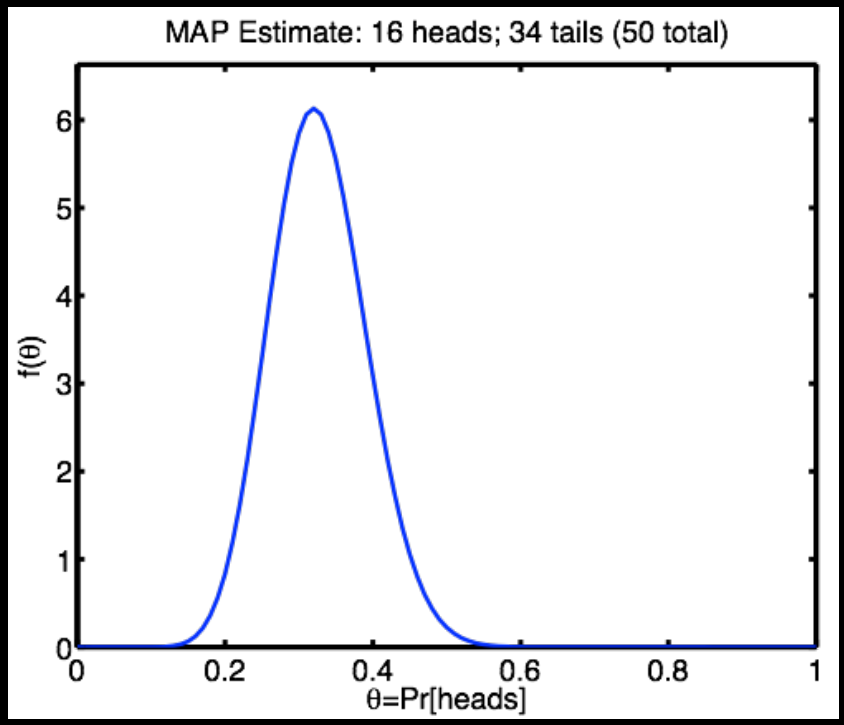




20 flips total

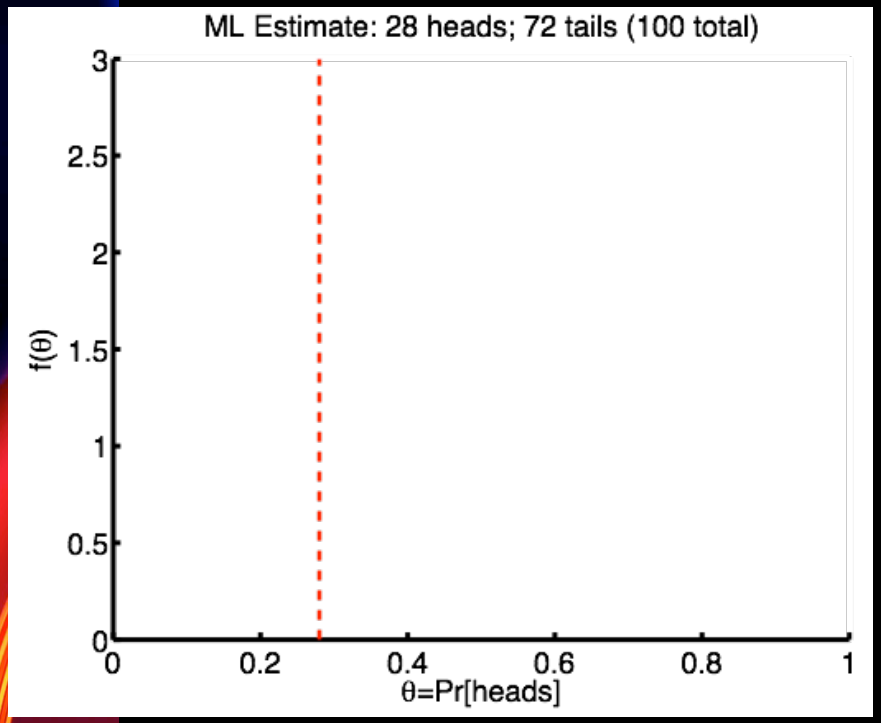
Example: Coin flipping

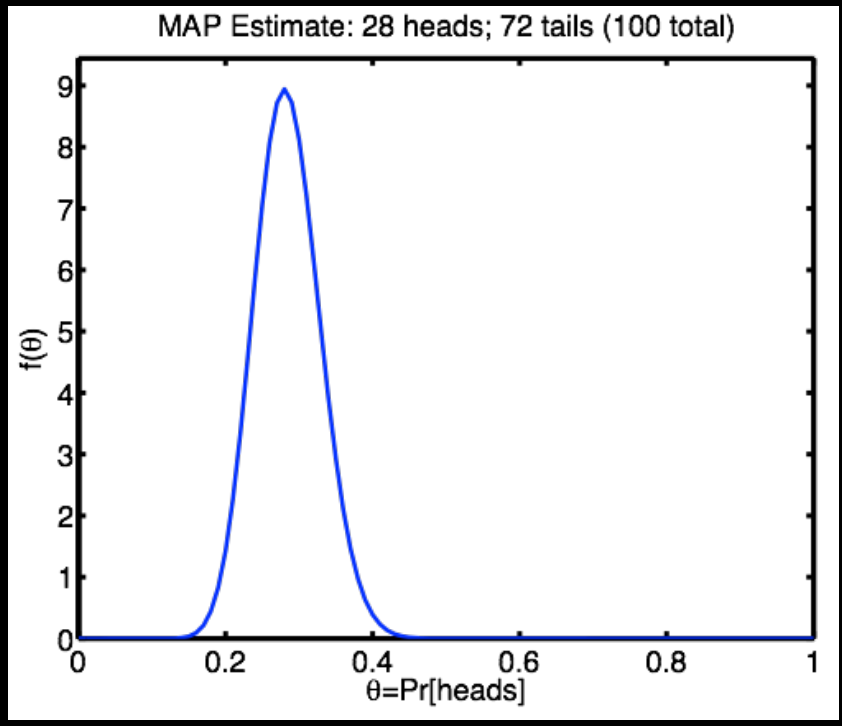




50 flips total

Example: Coin flipping





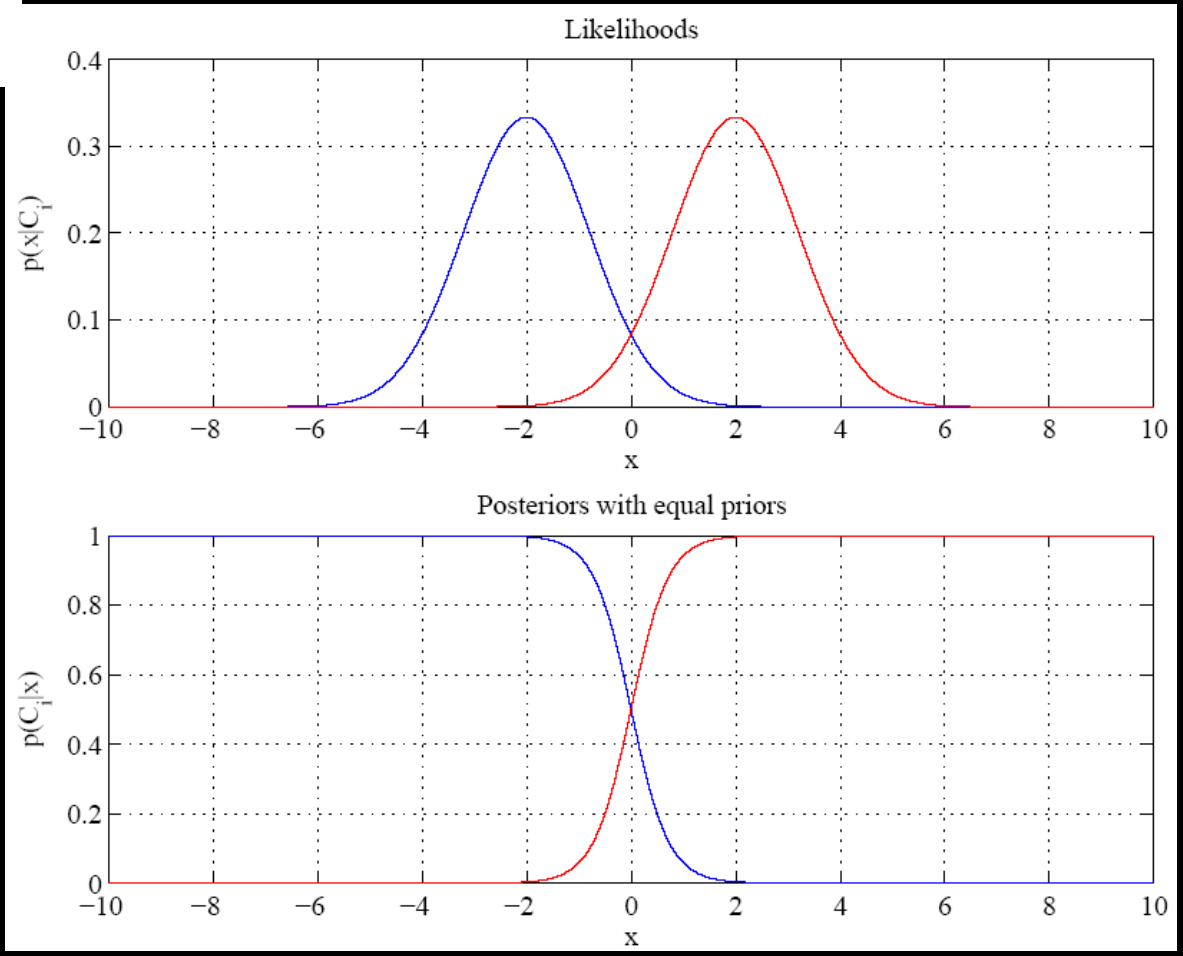
100 flips total

Classification: 2 classes (same variance)

Assume both classes have the same prior

Likelihood = Gaussian with mean, std dev

Posterior = discriminant g(x) normalized by P(x)

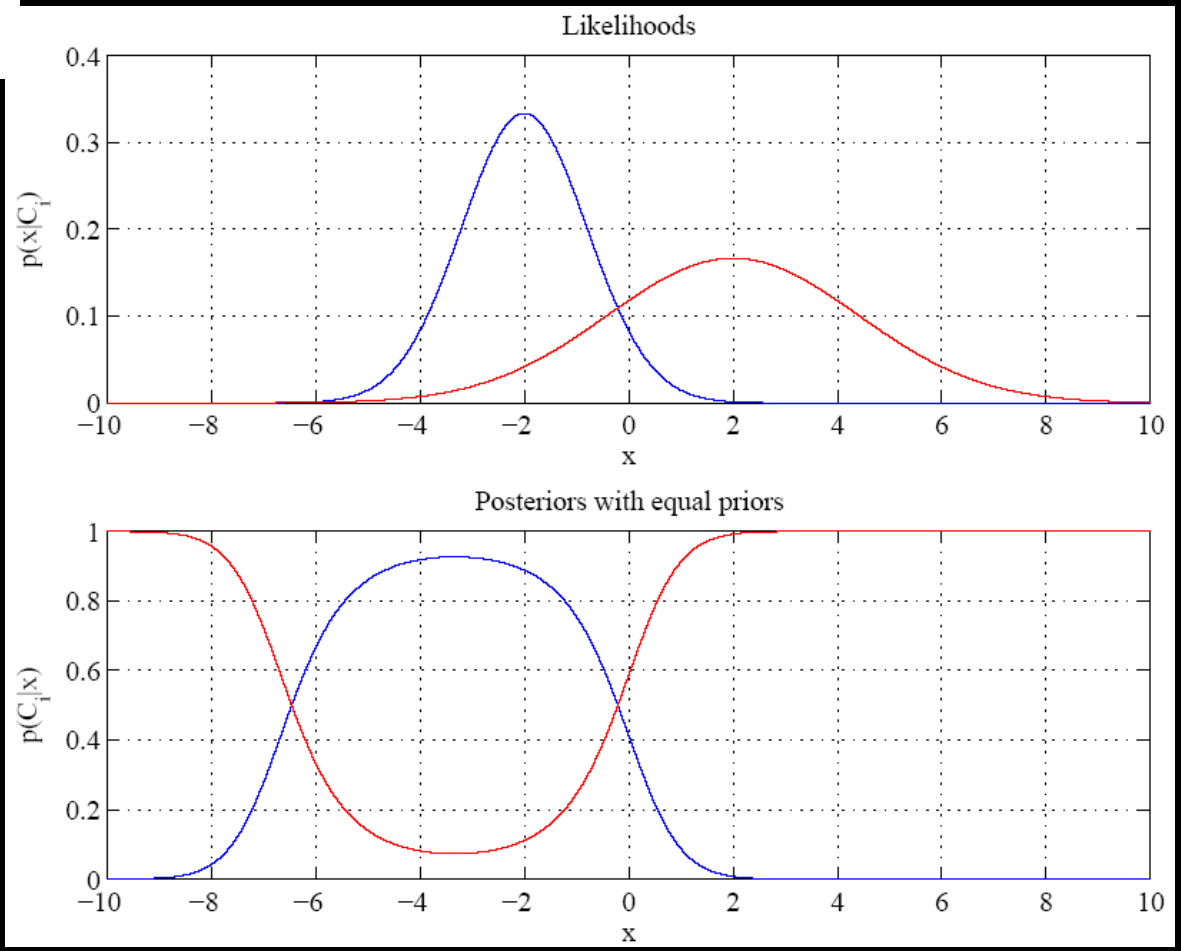


Classification: 2 classes (diff variance)

Assume both classes have the same prior

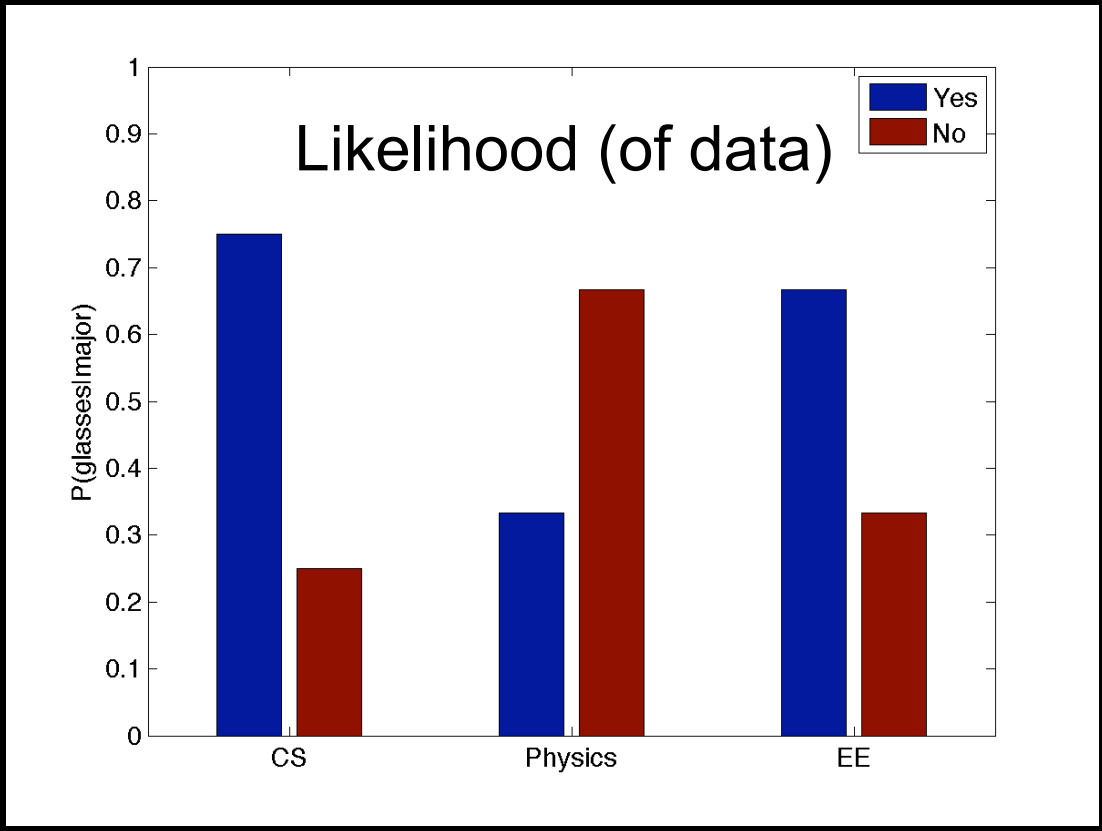
Likelihood = Gaussian with mean, std dev

Posterior = discriminant g(x) normalized by P(x)



Example: Predicting student's major

- From HW 3: Use "glasses" feature: yes or no
  - Discrete distribution

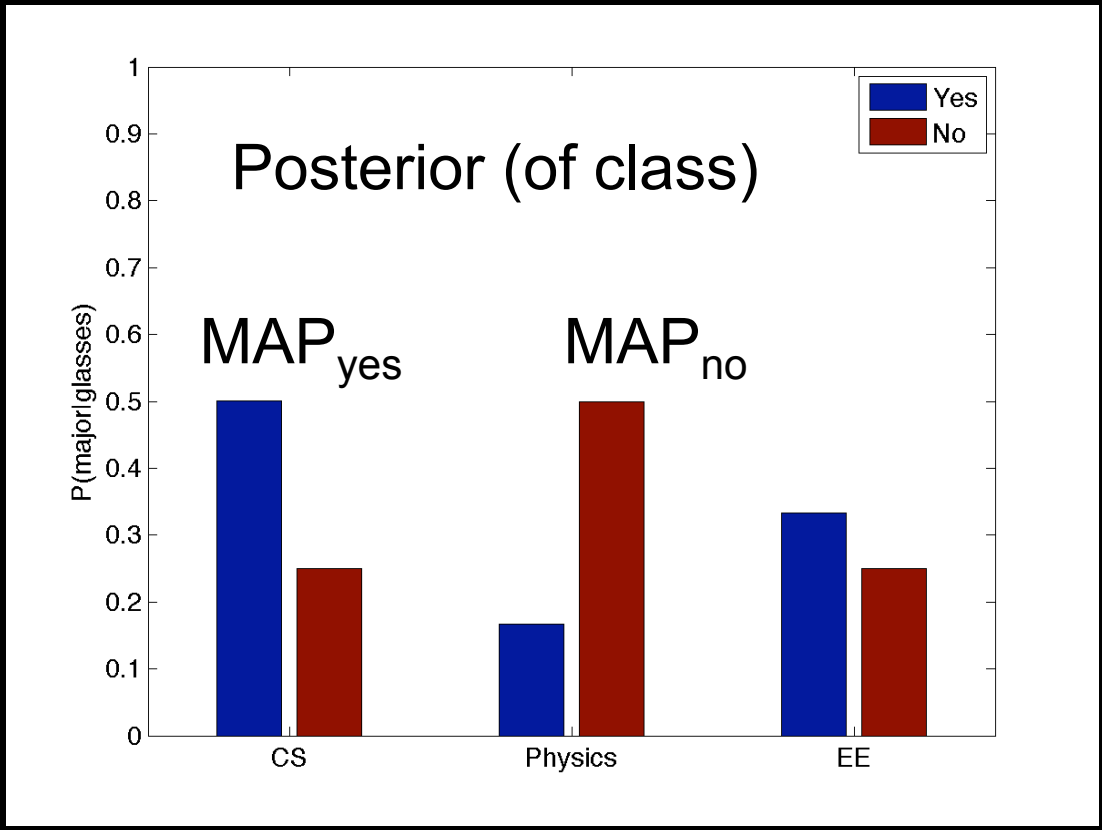


Example: Predicting student's major

Bayes: P(major|glasses) = P(glasses|major) P(major) / P(glasses)

Priors: P(CS) = 0.4, P(Physics) = 0.3, P(EE) = 0.3

Probability of glasses: P(yes) = 0.6, P(no) = 0.4



Summary: Key Points for Today

- Parametric methods
  - Data comes from distribution
  - Bernoulli, Gaussian, and their parameters
  - How good is a parameter estimate? (bias, variance)
- Bayes estimation
  - ML: use the data
  - MAP: use the prior and the data
- Parametric classification
  - Maximize the posterior probability

Next Time

- Parametric Methods
  (read Ch. 4.1-4.5)

- Volunteers for reading questions:
  - Roice, Deidre, Robert