CHAPTER 4:

Parametric Methods

Parametric Estimation

- STATISTIC – any value calculated from a sample
- X $= \{ x^t \}_t$ where $x^t \sim p(x)$

  – the task of estimating $p(x)$
- Parametric estimation:

  $p(x \mid \theta)$ – pde defined up to parameters θ

  Assume a form for $p(x \mid \theta)$ and estimate θ, its sufficient statistics, using X

  e.g., N (μ, σ^2) where $\theta = \{ \mu, \sigma^2 \}$

Maximum Likelihood Estimation

- Likelihood of θ given the sample X

  $$l(\theta|X) = p(X|\theta) = \prod_t p(x^t|\theta)$$

- Want to find θ that make X most likely to be drawn

- Log likelihood

  $$L(\theta|X) = \log l(\theta|X) = \sum_t \log p(x^t|\theta)$$

  - Replaces a product with a sum

- Maximum likelihood estimator (MLE)

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

Examples: Bernoulli/Multinomial

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

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

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

  MLE (by solving dL/dp): $p_o = \sum_t x^t / N$

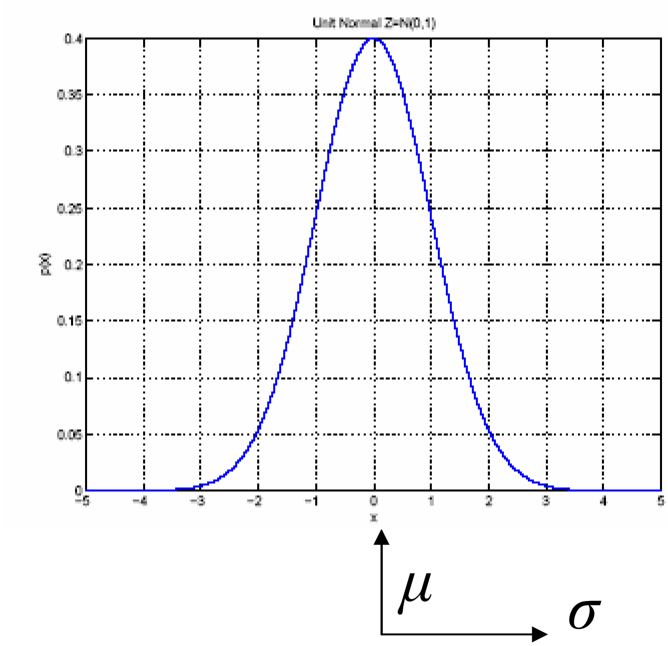
- Multinomial: K>2 mutually exclusive states, x_i in {0,1} ; x_i =1 for state i

  $P(x_1,x_2,...,x_K) = \prod_i p_i^{x_i}$

  $\mathcal{L}(p_1,p_2,...,p_K|\mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t}$

  MLE: $p_i = \sum_t x_i^t / N$

Gaussian (Normal) Distribution



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

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

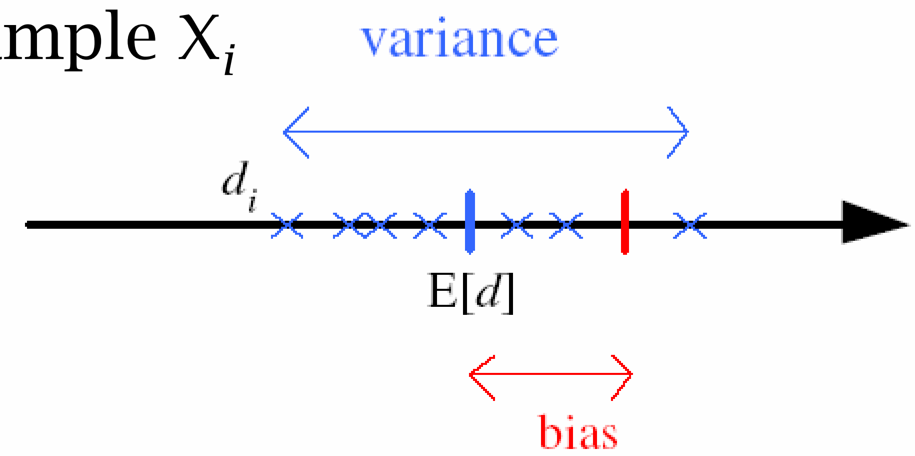
- MLE for μ and σ^2:

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

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

Bias and Variance

Unknown parameter θ
Estimator $d_i = d\,(\mathsf{X}_i)$ on sample X_i



Bias: $b_\theta(d) = E\,[d] - \theta$
Variance: $E\,[(d - E\,[d])^2]$

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

What makes a good estimator?

Bayes' Estimator

- Treat θ as a random var with prior $p(\theta)$
- Bayes' rule: $p(\theta|X) = p(X|\theta)\, p(\theta) \,/\, p(X)$

  $p(\theta)$ – prior density estimate (before looking at the sample)

  $p(\theta|X)$ – posterior density of θ, after looking at the sample
- Estimate density at x: $p(x|X) = \int p(x|\theta)\, p(\theta|X)\, d\theta$

  – integral may be difficult to evaluate, can be reduced to a point
- Maximum a Posteriori estimate (MAP):

$$\theta_{\mathrm{MAP}} = \mathrm{argmax}_\theta\, p(\theta|X)$$

  for flat prior density, MAP is equivalent to ML estimate!
- Maximum Likelihood (ML): $\theta_{\mathrm{ML}} = \mathrm{argmax}_\theta\, p(X|\theta)$

  another possibility is to use Bayes' estimator – expected value of posterior density
- Bayes': $\theta_{\mathrm{Bayes'}} = E[\theta|X] = \int \theta\, p(\theta|X)\, d\theta$

Bayes' Estimator: Example

- $x^t \sim \mathcal{N}(\theta, \sigma_o^2)$ and $\theta \sim \mathcal{N}(\mu, \sigma^2)$
- $\theta_{ML} = m$
- $\theta_{MAP} = \theta_{Bayes'} =$

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

Bayes' estimator produces a weighted average of the prior mean μ and the sample mean *m* with weights inversely proportional to their variances.

Parametric Classification

According to Bayes' rule,
posterior probability of class Ci

$$P(C_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C_i)P(C_i)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} p(\boldsymbol{x} \mid C_k)P(C_k)}$$

Classification can be performed using this discriminant function:

$$g_i(x) = p(x \mid C_i)P(C_i)$$

or equivalently

$$g_i(x) = \log p(x \mid C_i) + \log P(C_i)$$

Parametric Classification

If we assume that $p(x \mid C_i)$ is Gaussian:

$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

then the discriminant function

$$g_i(x) = \log p(x \mid C_i) + \log P(C_i)$$

becomes

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

- Selling K different cars, choice affected only by income *x*

- Proportion of customers who buy car *i* $P(C_i)$

- If income of customers who buy car *i* has Gaussian distribution then $p(x \mid C_i) \sim N(\mu_i, \sigma_i^2)$

- We do not know $P(C_i)$ and $p(x \mid C_i)$ and estimate them from a sample

$$x \in \Re \qquad \mathcal{X} = \{x^t, r^t\}_{t=1}^N \qquad r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j, j \neq i \end{cases}$$

ML estimates are

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

- Discriminant becomes

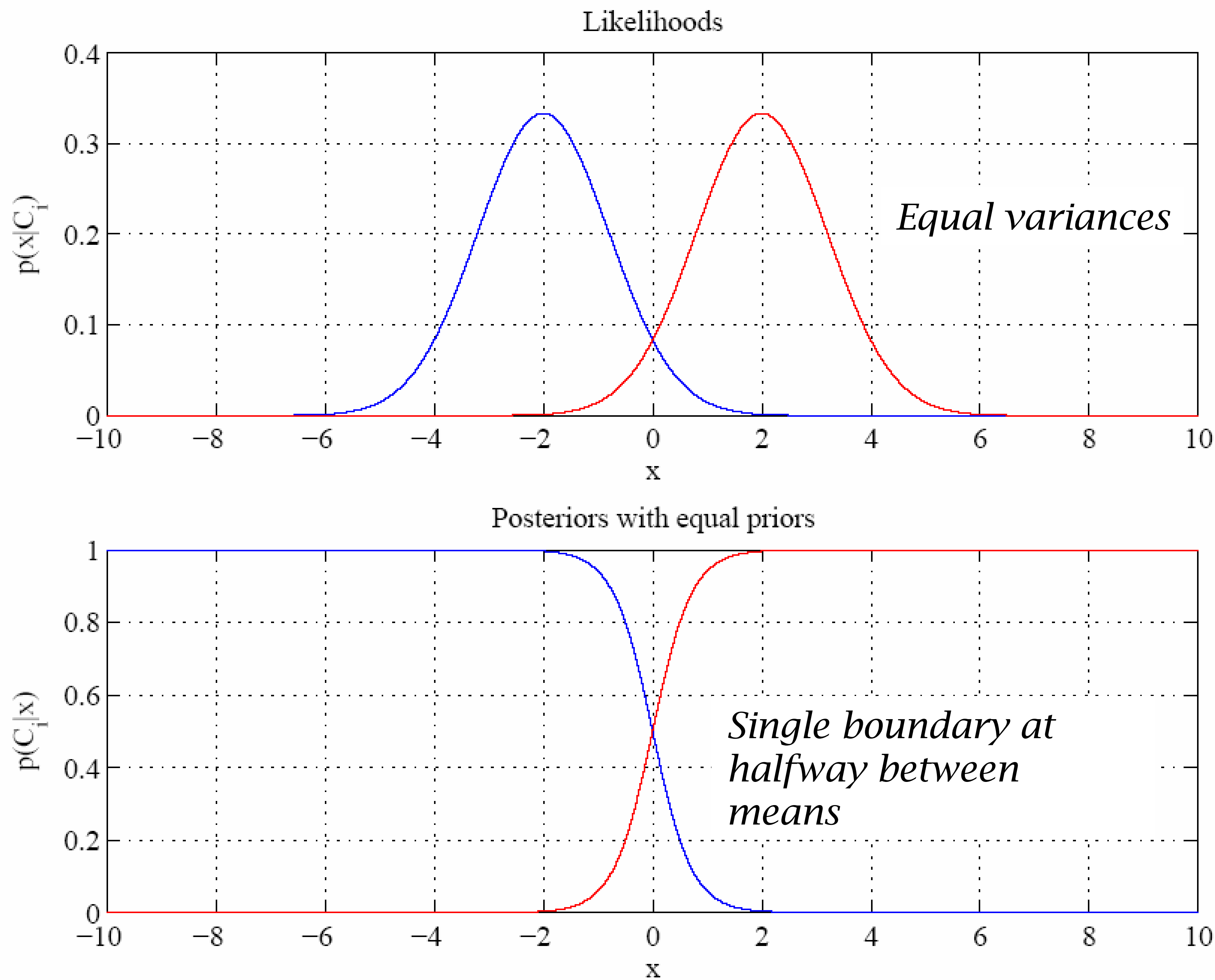
$$g_i(x) = \cancel{-\frac{1}{2}\log 2\pi} \cancel{- \log s_i} - \frac{(x - m_i)^2}{2\cancel{s_i^2}} \cancel{+ \log \hat{P}(C_i)}$$

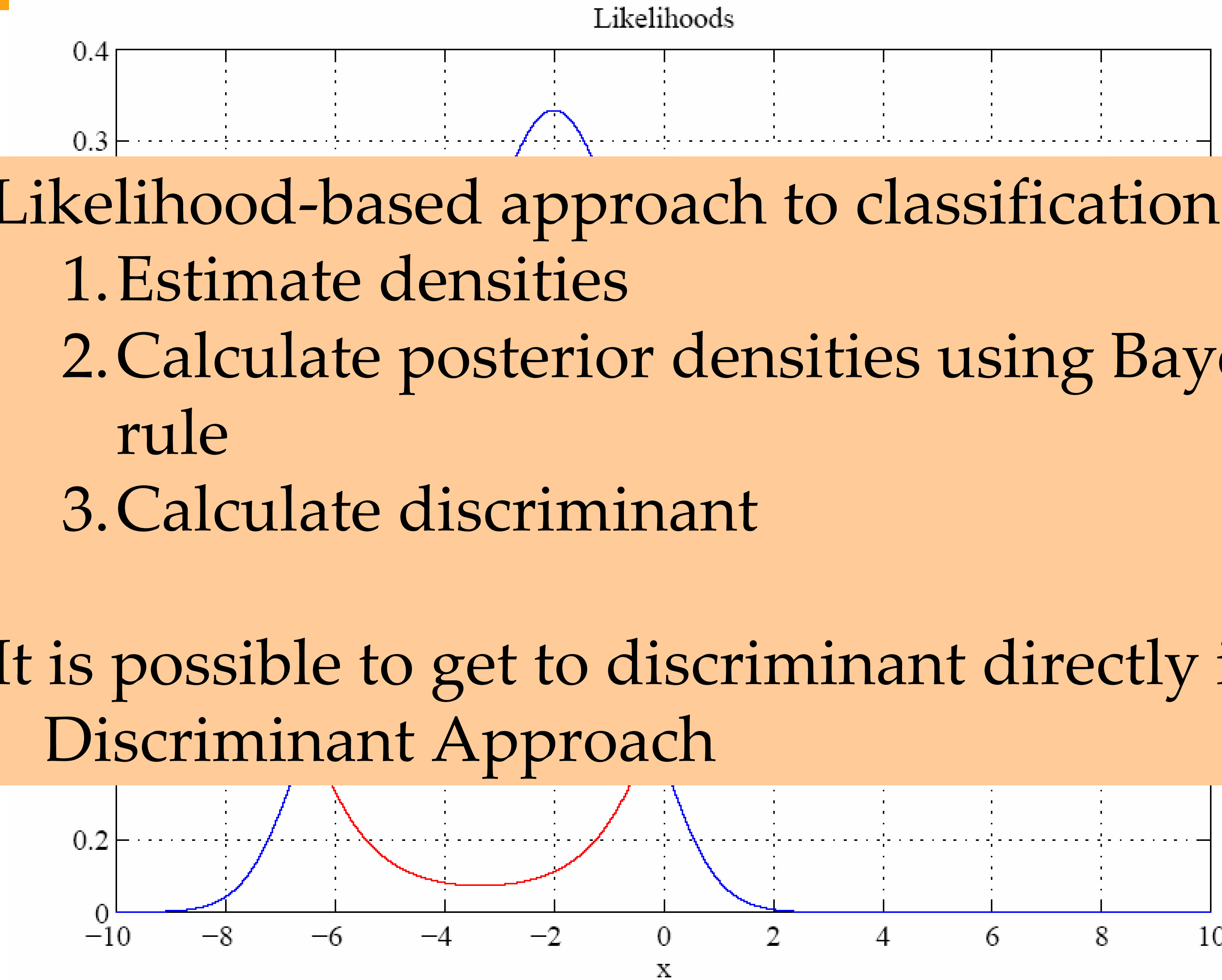
For case of equal variances

For case of equal priors

- Then $g_i(x) = -(x - m_i)^2$

- Chose Ci if $|x - m_i| = \min_k |x - m_k|$


Likelihoods
Equal variances
p(x|C_i)
x
Posteriors with equal priors
Single boundary at halfway between means
p(C_i|x)
x



Likelihood-based approach to classification:

1. Estimate densities
2. Calculate posterior densities using Bayes' rule
3. Calculate discriminant

It is possible to get to discriminant directly in Discriminant Approach

Regression

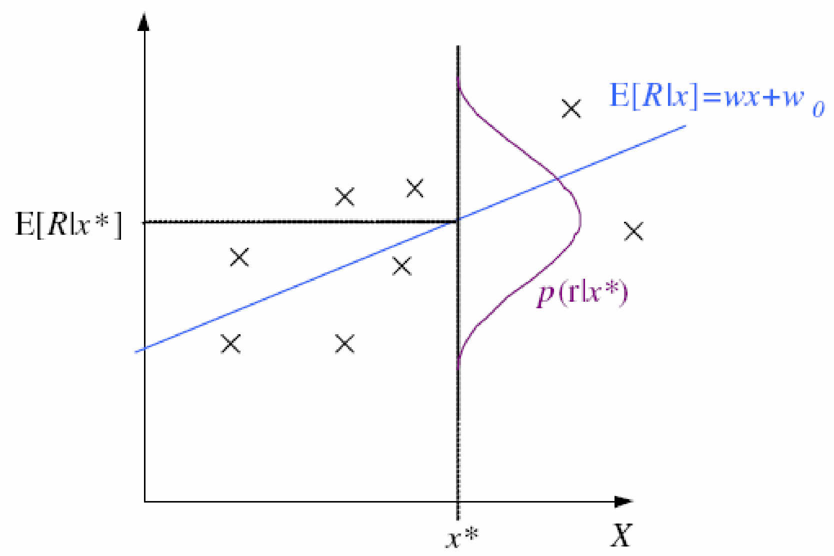
Dependent variable

$$r = f(x) + \varepsilon$$

estimator for $f(x)$: $g(x \mid \theta)$

assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

then $p(r \mid x) \sim \mathcal{N}(g(x \mid \theta), \sigma^2)$



Use maximum likelihood approach to learn parameters θ

Pairs (x^t, r^t) are drawn from unknown pdf: $p(x,r) = p(r|x)\,p(x)$

where $p(r|x)$ is probability of the output given the input and $p(x)$ is the input density. Then log likelihood is

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} p(x^t, r^t) = \log \prod_{t=1}^{N} p(r^t \mid x^t) + \log \prod_{t=1}^{N} p(x^t)$$

Regression

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} p\left(r^t \mid x^t\right) + \log \prod_{t=1}^{N} p\left(x^t\right)$$

We can ignore the second term since it does not depend on the estimator

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{2\sigma^2}\right] = -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

First term is independent of θ and can be dropped,
as the variance factor. Maximizing the likelihood is equivalent to minimizing the following function

$$E(\theta \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

- Least square estimates

$$E(\theta \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g(x^t \mid \theta)\right]^2$$

Linear Regression

$$g(x^t \mid w_1, w_0) = w_1 x^t + w_0$$

Taking derivatives of the sum of squared errors

$$\sum_t r^t = N w_0 + w_1 \sum_t x^t$$

$$\sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2$$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

$$\boldsymbol{w} = \mathbf{A}^{-1}\boldsymbol{y}$$

Polynomial Regression

$$g\left(x^t \mid w_k, \ldots, w_2, w_1, w_0\right) = w_k\left(x^t\right)^k + \cdots + w_2\left(x^t\right)^2 + w_1 x^t + w_0$$

$$\mathbf{D} = \begin{bmatrix} 1 & x^1 & \left(x^1\right)^2 & \cdots & \left(x^1\right)^k \\ 1 & x^2 & \left(x^2\right)^2 & \cdots & \left(x^2\right)^k \\ \vdots & & & & \\ 1 & x^N & \left(x^N\right)^2 & \cdots & \left(x^N\right)^2 \end{bmatrix} \quad \boldsymbol{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

$$\boldsymbol{w} = \left(\mathbf{D}^T\mathbf{D}\right)^{-1}\mathbf{D}^T\boldsymbol{r}$$

Other Error Measures

- Square Error: $$E(\theta \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2$$
- Relative Square Error: $$E(\theta \mid \mathcal{X}) = \frac{\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{\sum_{t=1}^{N}\left[r^t - \bar{r}\right]^2}$$
- Absolute Error: $E(\theta|\mathcal{X}) = \sum_t |r^t - g(x^t|\theta)|$
- ε–sensitive Error:

$$E(\theta|\mathcal{X}) = \sum_t 1(|r^t - g(x^t|\theta)| > \varepsilon)\ (|r^t - g(x^t|\theta)| - \varepsilon)$$

Bias and Variance

Expected error given regression model g(.) and Bayes' estimator

$$E\left[(r-g(x))^2 \mid x\right] = E\left[(r-E[r \mid x])^2 \mid x\right] + (E[r \mid x] - g(x))^2$$

noise *squared error*

First term – variance of *r* given *x* – does not depend on g(.) or X

This is variance of noise!

- Second term – quantifies deviation of *g(x)* from *E(r|x)*

Depends on the estimator and training set!

To estimate goodness of *g(.)* we need to average over possible samples. Expected value of squared error over samples X of size N

$$E_X\left[(E[r \mid x] - g(x))^2 \mid x\right] = (E[r \mid x] - E_X[g(x)])^2 + E_X\left[(g(x) - E_X[g(x)])^2\right]$$

bias *variance*

Estimating Bias and Variance

- Example:
- M samples $\mathcal{X}_i=\{x^t_i, r^t_i\}$, $i=1,\ldots,M$ are used to fit $g_i(x)$, $i=1,\ldots,M$

$$\text{Bias}^2(g)=\frac{1}{N}\sum_t\left[\overline{g}(x^t)-f(x^t)\right]^2$$

$$\text{Variance}(g)=\frac{1}{NM}\sum_t\sum_i\left[g_i(x^t)-\overline{g}(x^t)\right]^2$$

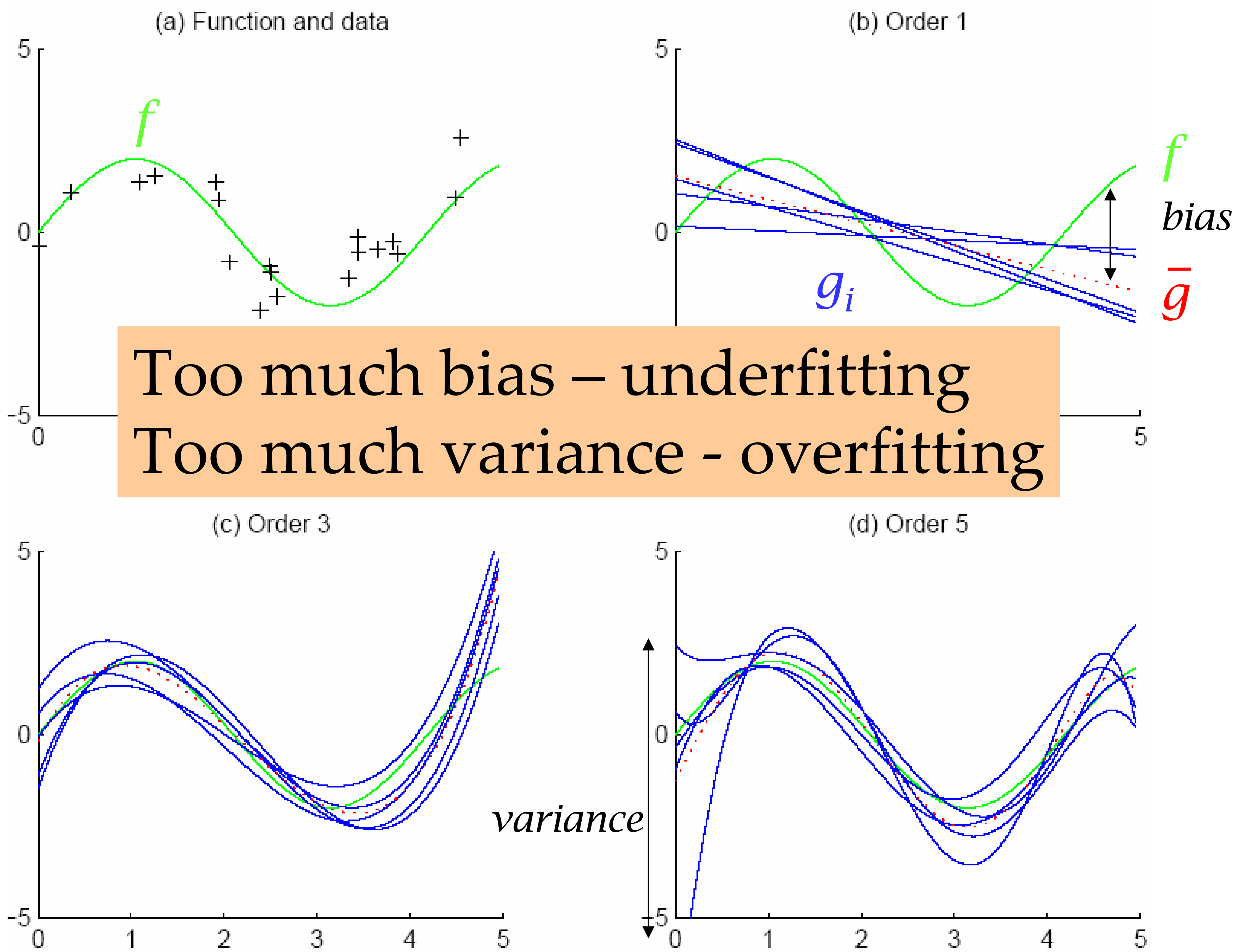
$$\overline{g}(x)=\frac{1}{M}\sum_t g_i(x)$$

Can we do this for practical application?

Bias/Variance Dilemma

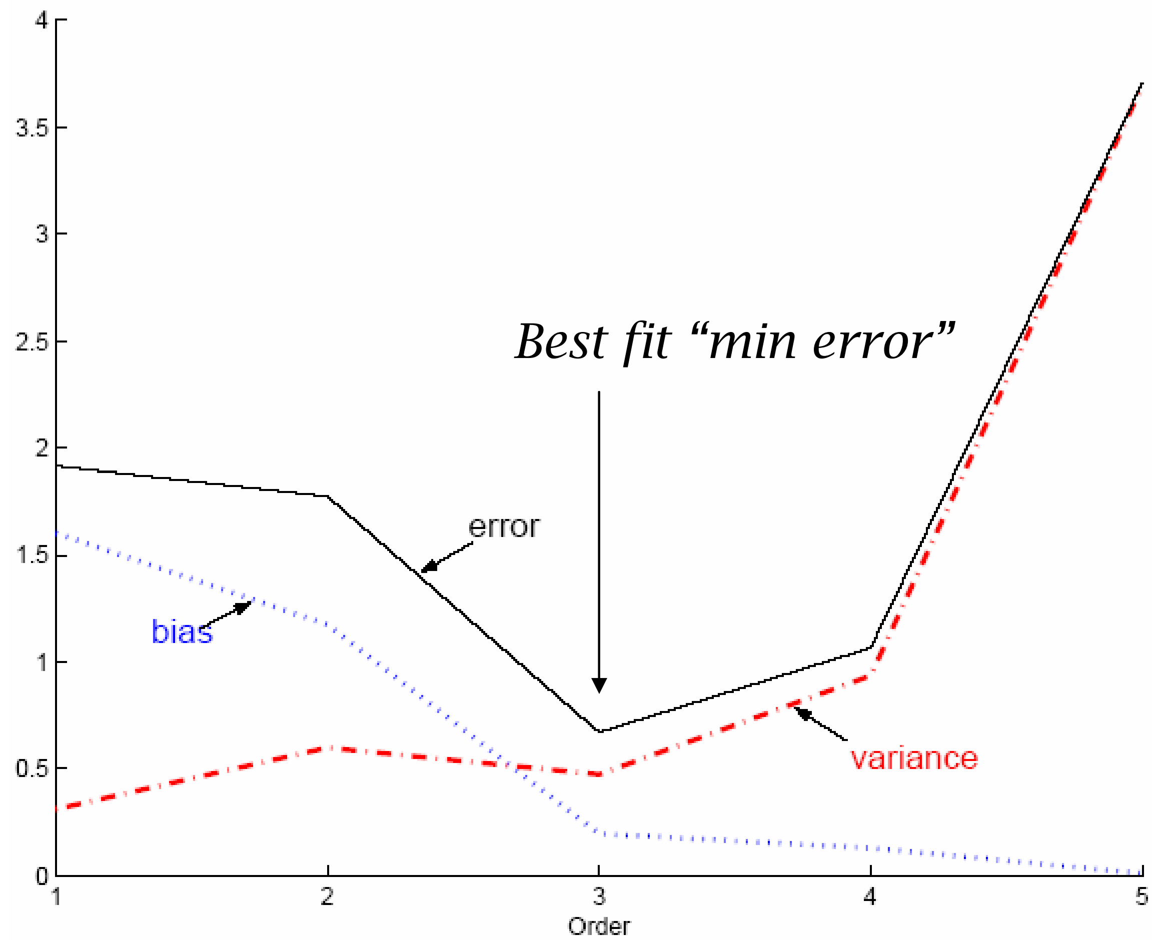
- Example:
  Constant fit: $g_i(x)=2$ has no variance and high bias
  Average fit: $g_i(x)= \sum_t r^t_i/N$ has lower bias with variance

- As we increase complexity,
  bias decreases (a better fit to data) and
  variance increases (fit varies more with data)
- Bias/Variance dilemma: (Geman et al., 1992)

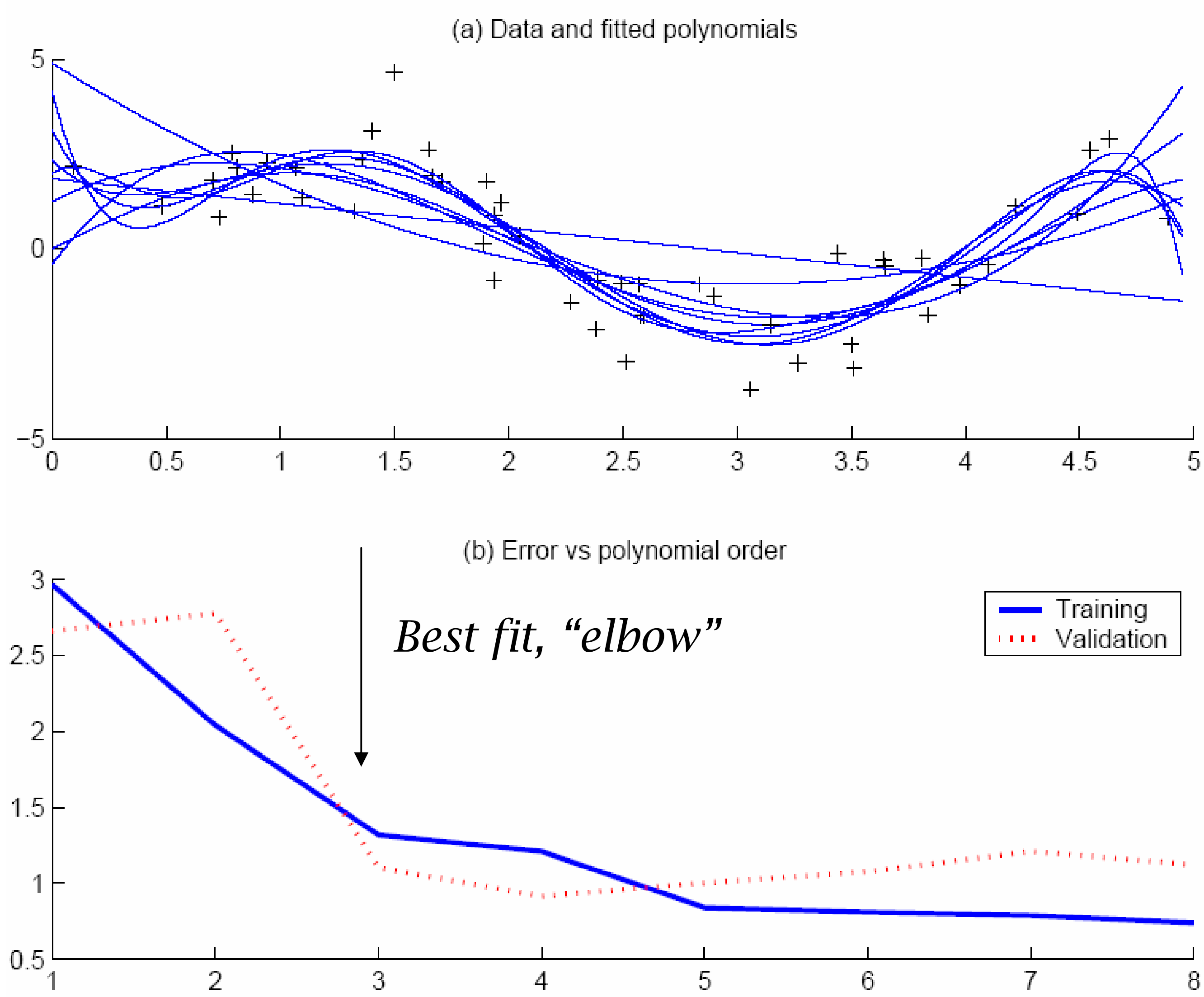


Too much bias – underfitting
Too much variance - overfitting

Polynomial Regression



Model selection with cross-validation



Model Selection

- Cross-validation: Measure generalization accuracy by testing on data unused during training
- Regularization: Penalize complex models

  E'=error on data + λ model complexity

- Minimum description length (MDL): Kolmogorov complexity, shortest description of data
- Structural risk minimization (SRM)

Bayesian Model Selection

- Prior on models, $p(\text{model})$

$$p(\text{model} \mid \text{data}) = \frac{p(\text{data} \mid \text{model})\, p(\text{model})}{p(\text{data})}$$

- Regularization, when prior favors simpler models
- Bayes, MAP of the posterior, $p(\text{model}|\text{data})$
- Average over a number of models with high posterior (voting, ensembles: Chapter 15)