CHAPTER 4:

PARAMETRİC METHODS

Parametric Estimation

- $X = \{ x^t \}_t$ where $x^t \sim p(x)$

"distributed as" probability density function

- Parametric estimation:

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

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

Remember that a probability density function is defined as:

$$p(x_0) \equiv \lim_{\varepsilon \to 0} P(x_0 \le x < x_0 + \varepsilon)$$

Maximum Likelihood Estimation

- Likelihood of θ given the sample X

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

- Log likelihood

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

- Maximum likelihood estimator (MLE)

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

Examples: Bernoulli/Multinomial

- Bernoulli: Two states, failure/success, x in {0,1} (e.g, coin)

$$P(x) = p_o^{\,x}(1-p_o)^{(1-x)}$$ p_o *is the parameter for probability of success*

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

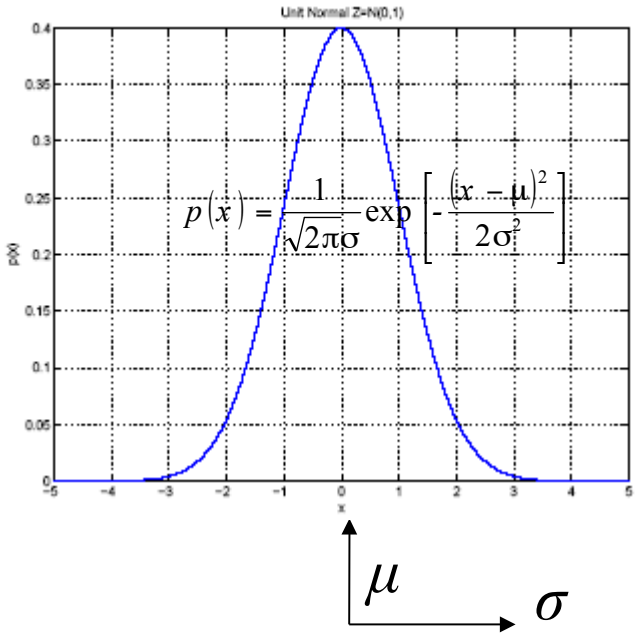
MLE: $p_o = \sum_t x^t / N$

- Multinomial: K>2 states, x_i in {0,1} (e.g., die with 6 faces)

$$P(x_1,x_2,\ldots,x_K) = \prod_i p_i^{x_i}$$ p_i *is the parameter for probability of state i success*

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

Gaussian (Normal) Distribution



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

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

- MLE for μ and σ^2:

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

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

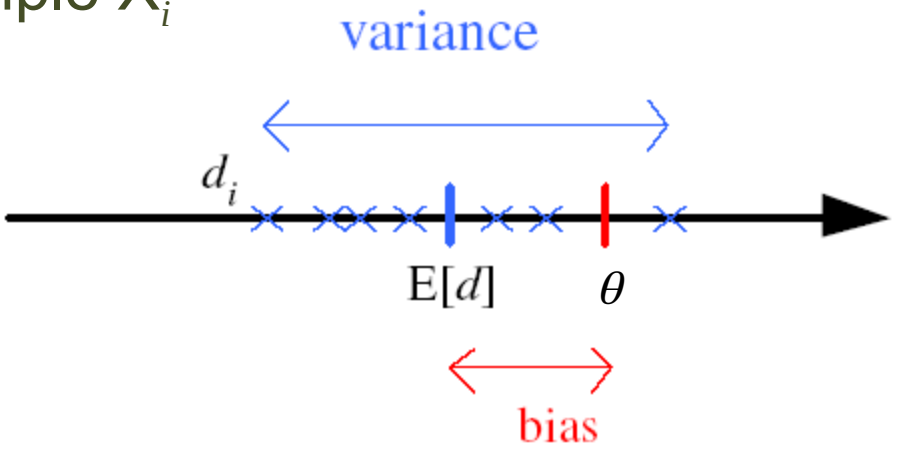
Bias and Variance

Unknown parameter θ

Estimator $d_i = d(X_i)$ on sample X_i

Bias: $b_\theta(d) = E[d] - \theta$

Variance: $E[(d - E[d])^2]$



Mean square error:

$r(d,\theta) = E[(d-\theta)^2]$

$= (E[d] - \theta)^2 + E[(d - E[d])^2]$

$=$ Bias^2 + Variance

the smaller the bias and the variance, the better

Bayes' Estimator

- Treat θ as a random var with prior $p(\theta)$

this prior may come from domain expertise, without looking at X

- Bayes' rule: $p(\theta|X) = p(X|\theta)\, p(\theta) / p(X)$

posterior density of θ ↘ likelihood density ↘

- Full: $p(x|X) = \int p(x|\theta)\, p(\theta|X)\, d\theta$
- Maximum a Posteriori (MAP): $\theta_{MAP} = \text{argmax}_\theta\, p(\theta|X)$
- Maximum Likelihood (ML): $\theta_{ML} = \text{argmax}_\theta\, p(X|\theta)$
- Bayes': $\theta_{Bayes'} = E[\theta|X] = \int \theta\, p(\theta|X)\, d\theta$

Comparing ML, MAP, and Bayes'

Let Θ be the set of all possible solutions θ's

- **Maximum a Posteriori (MAP):** Selects the θ that satisfies $\theta_{MAP} = \text{argmax}_\theta\ p(\theta|X)$
- **Maximum Likelihood (ML):** Selects the θ that satisfies $\theta_{ML} = \text{argmax}_\theta\ p(X|\theta)$
- **Bayes':** Constructs the "weighted average" over all θ's in Θ: $\theta_{Bayes'} = E[\theta|X] = \int \theta\ p(\theta|X)\ d\theta$

Note that if the θ's in Θ are uniformly distributed then $\theta_{MAP}=\theta_{ML}$

since $\theta_{MAP} = \text{argmax}_\theta\ p(\theta|X)$

$= \text{argmax}_\theta\ p(X|\theta)\ p(\theta)/p(X) = \text{argmax}_\theta\ p(X|\theta) = \theta_{ML}$

Bayes' Estimator: Example

- If $x^t \sim N(\theta, \sigma_o^2)$ and $\theta \sim N(\mu, \sigma^2)$

  where σ_o^2, μ, and σ^2 are known

- then:
  - $\theta_{ML} = m$ *sample mean*
  - $\theta_{MAP} = \theta_{Bayes'} =$

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

Note: if the θ's in Θ are normally distributed and $p(X|\theta)$ is normal, then $p(\theta|X)$ is normal and $\theta_{MAP}=\theta_{Bayes'}$

Parametric Classification

Remember from Chapter 3, where g_i is a discriminant function

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

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

or

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

If $p(x \mid C_i)$ are Gaussian, then: (here "log" is natural log)

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

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

- Given the sample $\mathbf{X} = \{x^t, r^t\}_{t=1}^N$

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

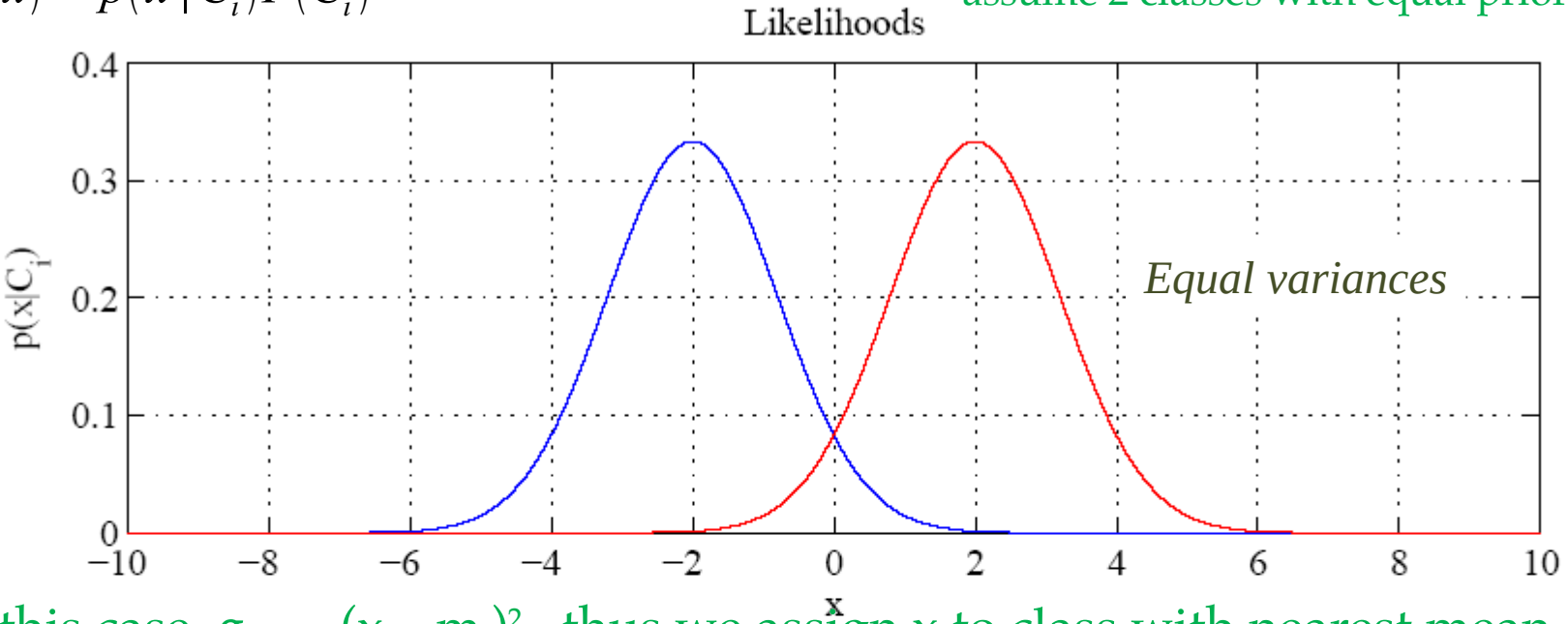
- ML estimates are

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

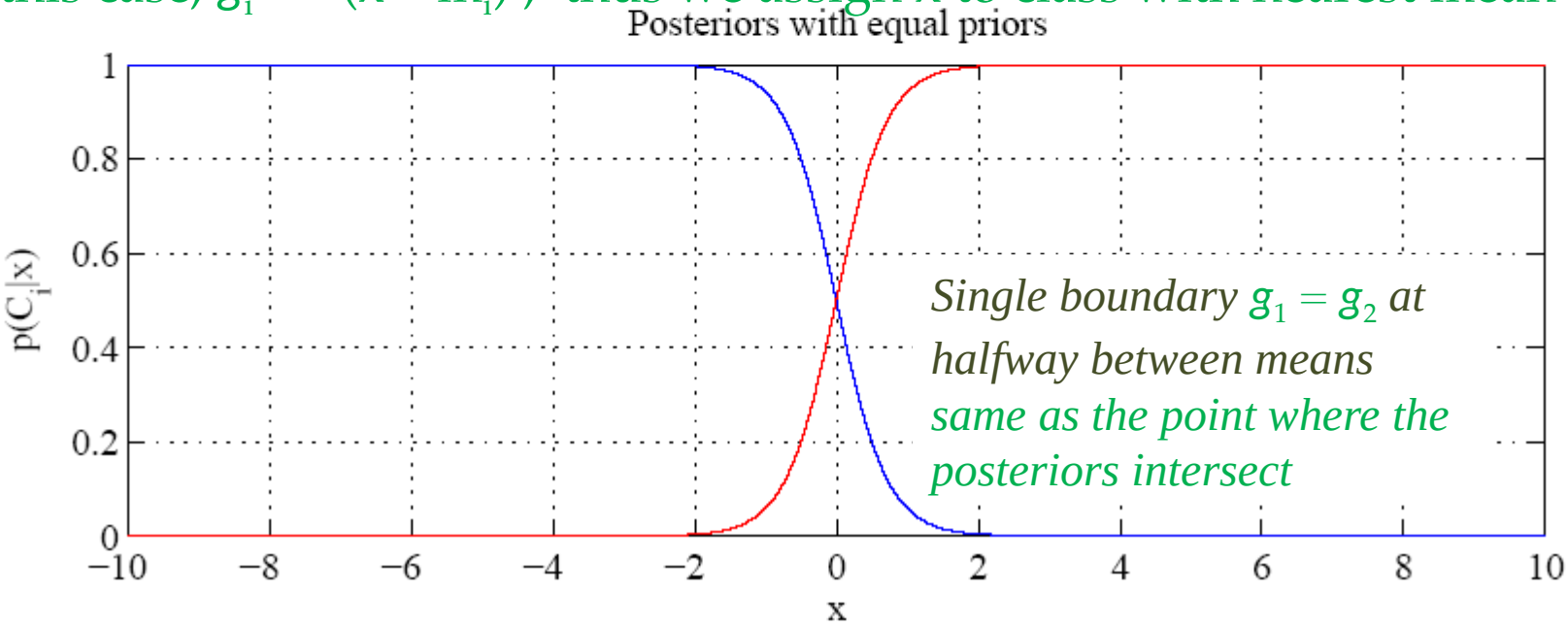
- Discriminant $g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$

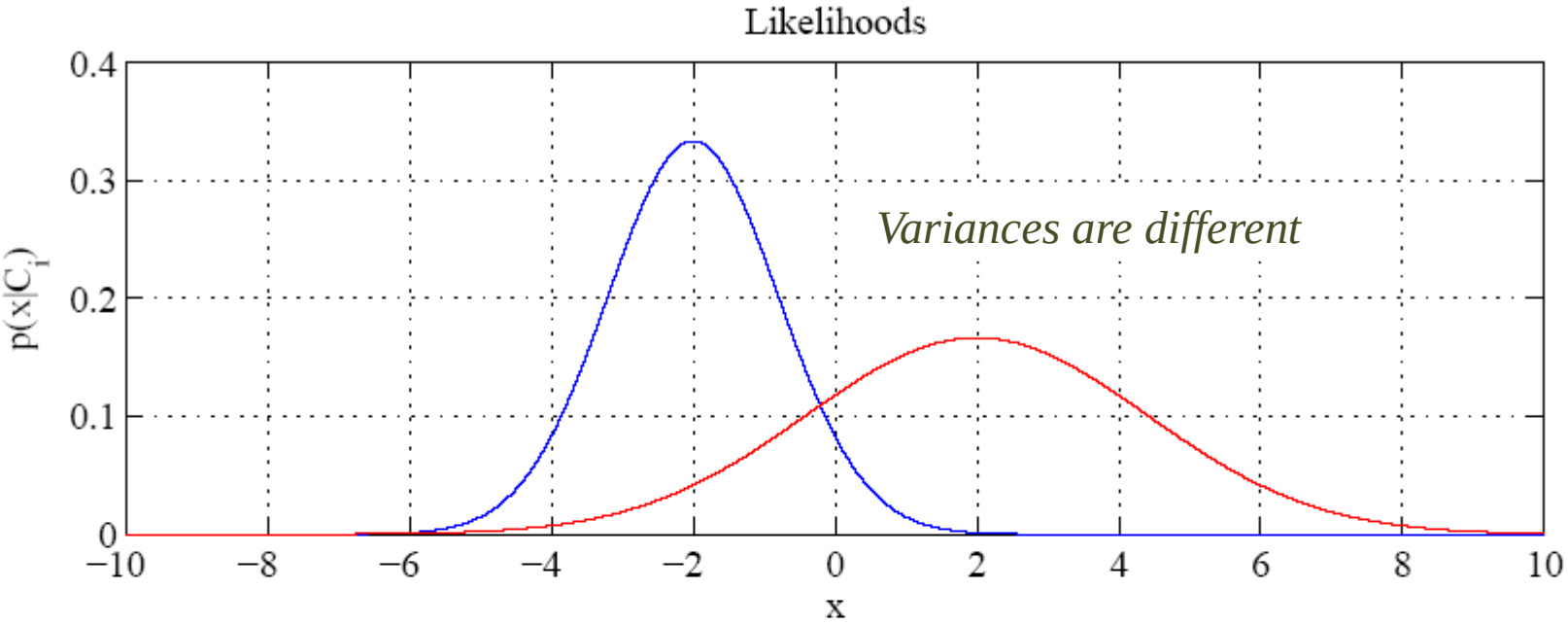
$$g_i(x) = p(x \mid C_i)P(C_i)$$

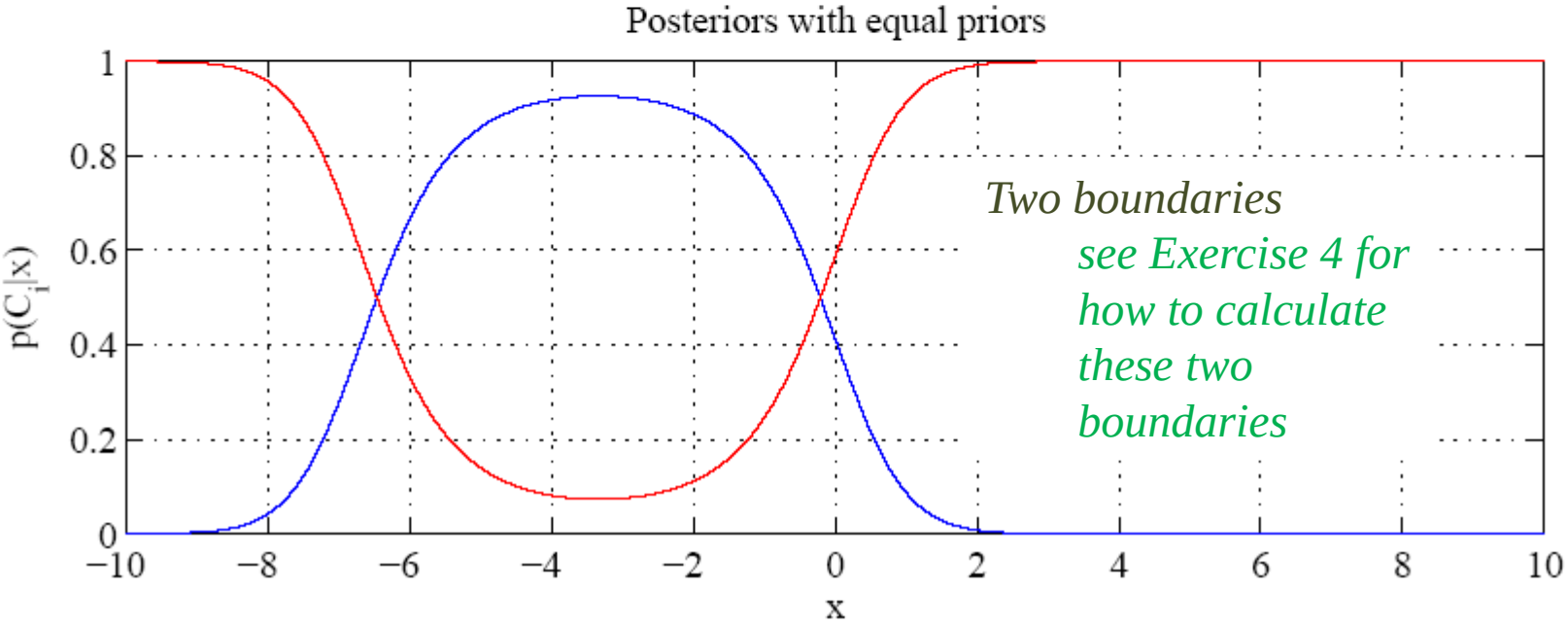
assume 2 classes with equal priors P(Ci)

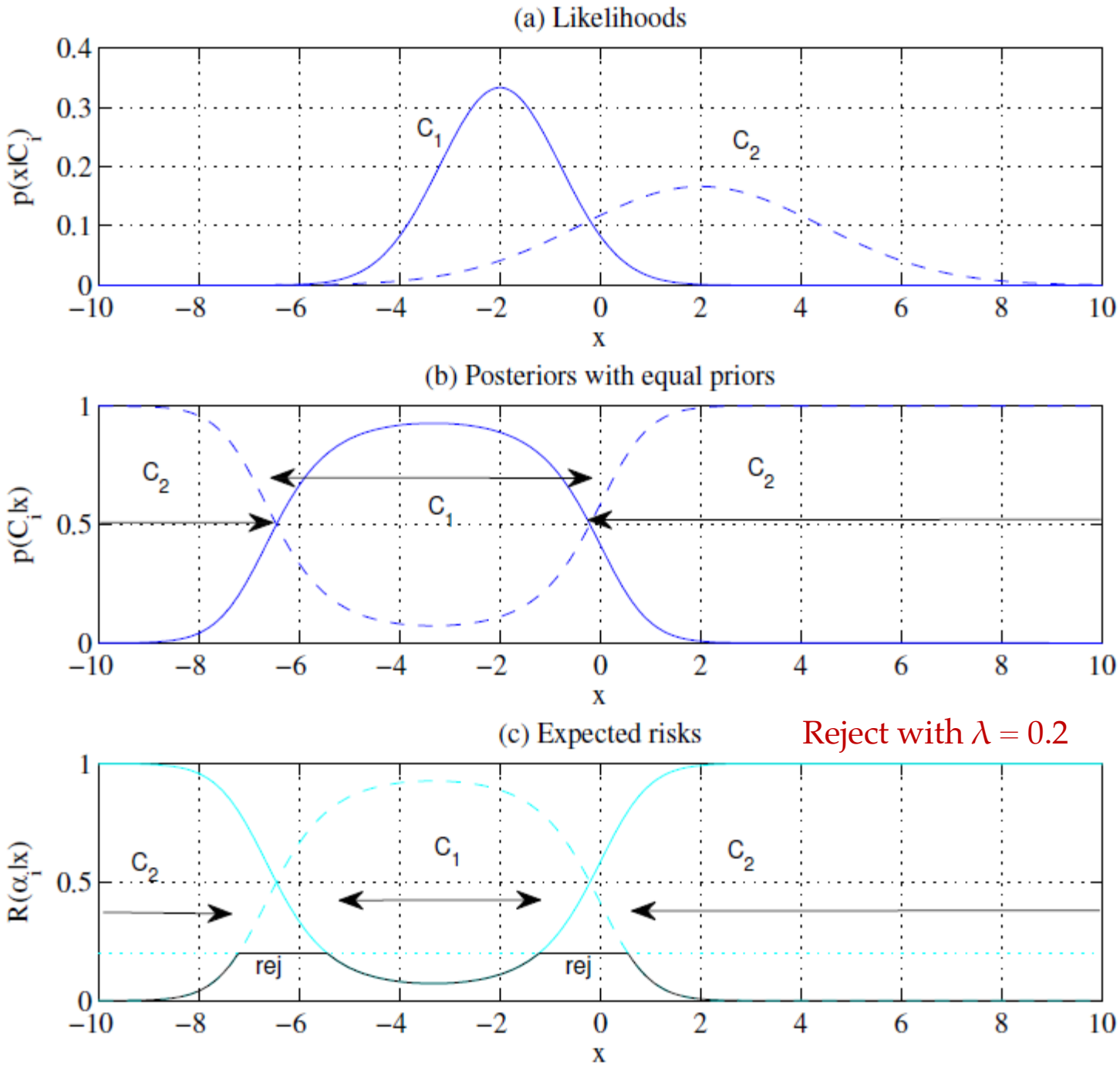


In this case, $g_i = -(x - m_i)^2$, thus we assign x to class with nearest mean




Likelihoods
p(x|C_i)
x
Variances are different


Posteriors with equal priors
p(C_i|x)
x
Two boundaries
see Exercise 4 for how to calculate these two boundaries


(a) Likelihoods
p(x|C_i)
0.4
0.3
0.2
0.1
0
C_1
C_2
−10
−8
−6
−4
−2
0
2
4
6
8
10
x
(b) Posteriors with equal priors
p(C_i|x)
1
0.5
0
C_2
C_1
C_2
x
(c) Expected risks
Reject with λ = 0.2
R(α_i|x)
1
0.5
0
C_2
C_1
C_2
rej
rej
x

Parametric Classification – Notes

- Note that the equations for parametric classification derived in the last few slides assume the data follows a Gaussian distribution

- Hence, you need to determine whether or not your univariate data X follows a Gaussian distribution

  using a statistical test for normality (e.g., Shapiro–Wilk test, Kolmogorov–Smirnov test, Lilliefors test, …)

  before you consider applying these equations

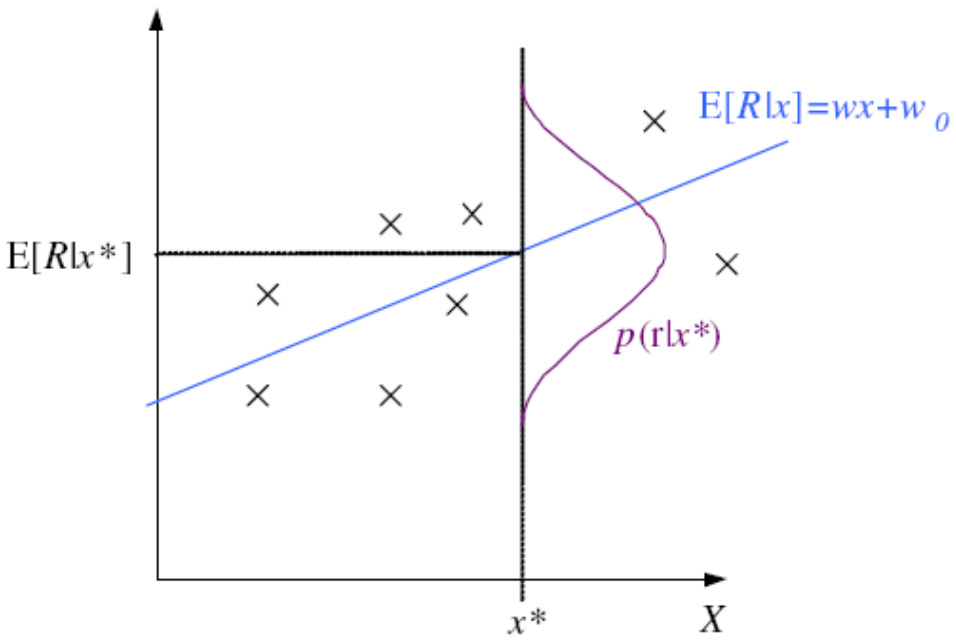
Regression

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

$$\text{estimator} : g(x \mid \theta)$$

$$\varepsilon \sim \mathsf{N}\left(0, \sigma^2\right)$$

$$p(r \mid x) \sim \mathsf{N}\left(g(x \mid \theta), \sigma^2\right)$$



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

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

second term can be ignored

Regression: From LogL to Error

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

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

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

Maximizing the log likelihood above is the same as minimizing the Error.

θ that minimizes Error is called “least squares estimate”

Trick: When maximizing a likelihood l that contains exponents, instead minimize error E = - log l

Linear Regression

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

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

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

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

$$\mathbf{A}\mathbf{w} = \mathbf{y} \quad \text{and so} \quad \mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$$

Polynomial Regression

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

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

$$\mathbf{A}\mathbf{w} = \mathbf{y} \quad \text{where} \quad \mathbf{A} = \left(\mathbf{D}^T\mathbf{D}\right) \text{ and } \mathbf{y} = \mathbf{D}^T\mathbf{r}$$

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

Maximum Likelihood and Least Squares

Taken from Tom Mitchell's Machine Learning textbook:

"… under certain assumptions (*) any learning algorithm that minimizes the squared error between the output hypothesis predictions and the training data will output a maximum likelihood hypothesis."

(*): Assumption:

"the observed training target values are generated by adding random noise to the true target value, where the random noise is drawn independently for each example from a Normal distribution with zero mean."

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_{RSE}):

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

- Absolute Error:

$$E(\theta \mid X) = \sum_t |r^t - g(x^t \mid \theta)|$$

- ε-sensitive Error:

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

- R^2 : Coefficient of Determination: $R^2 = 1 - E_{RSE}$ *(see next slide)*

Coefficient of Determination: R^2

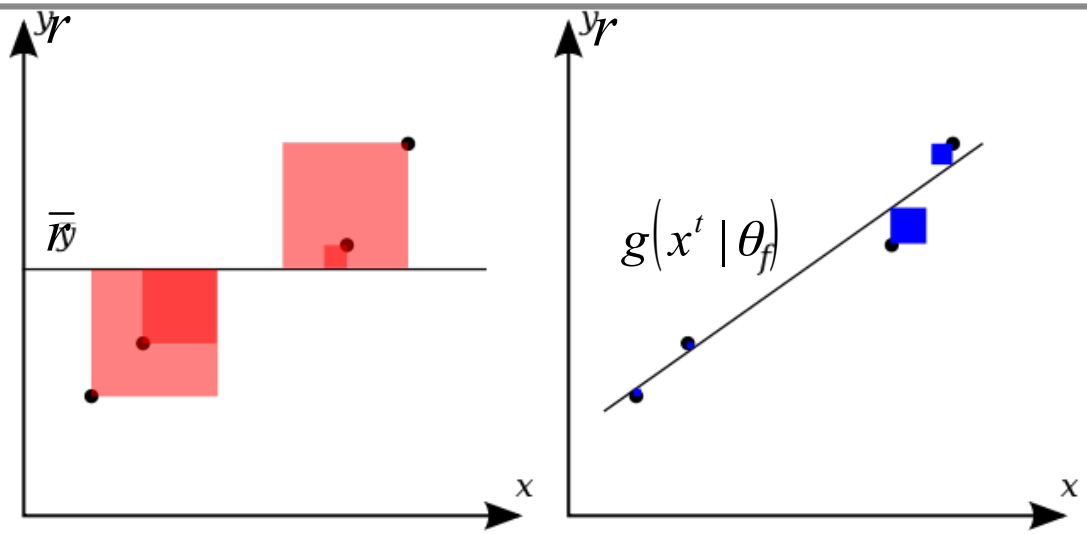


Figure adapted from Wikipedia (Sept. 2015)

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

The closer R^2 is to 1 the better as this means that the *g(.)* estimates (on the right graph) fit the data well in comparison to the simple estimate given by the average value (on the

Bias and Variance

- Given $X = \{x^t, r^t\}$ drawn from unknown pdf $p(x,r)$
  Expected Square Error of the $g(.)$ estimate at x:

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

noise
variance of noise added
does not depend on g(.) or X

squared error
g(x) deviation from E[r|x]
depend on g(.) and X

it may be that for a sample X, g(.) is a very good fit, and for another sample it is not

- Given samples X's, all of size N drawn from the same joint density p(x,r). Expected value, averaged over X's:

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

bias
how much g(.) is wrong
disregarding effect of varying samples

variance
how much g(.) fluctuates
as samples varies

Estimating Bias and Variance

- M samples $X_i=\{x^t_i, r^t_i\}$, $i=1,\ldots,M$ are used to fit $g_i(x)$, $i=1,\ldots,M$

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

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

$$\text{where } \bar{g}(x) = \frac{1}{M}\sum_t g_i(x)$$

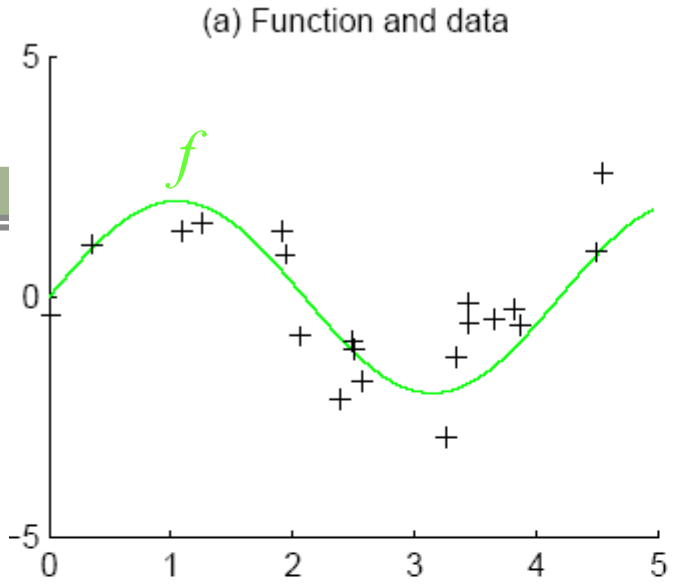
Bias/Variance Dilemma

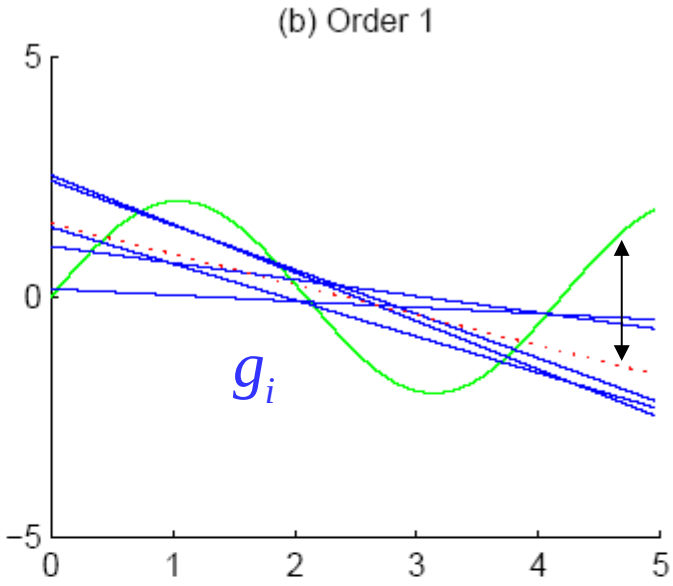
- Examples:
  - $g_i(x)$=2 has no variance and high bias
  - $g_i(x)= \sum_t r^t_i$/N has lower bias but higher variance

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

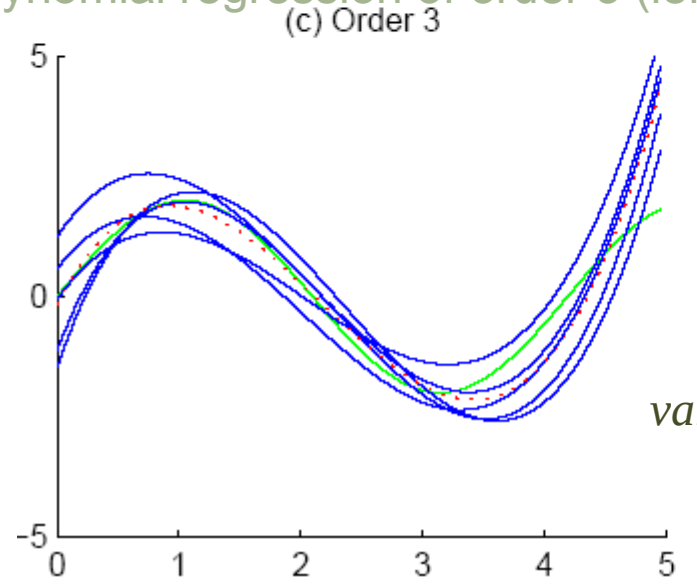
added noise ~ N(0,1)

A linear regression for each of 5 samples

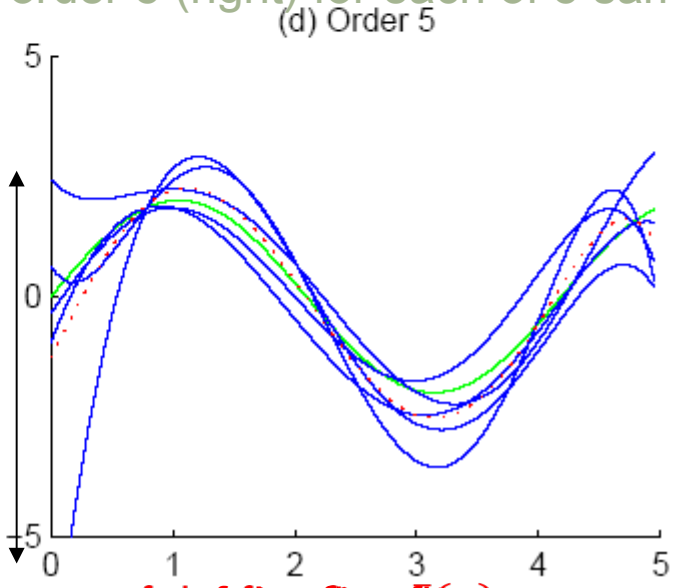




Polynomial regression of order 3 (left) and order 5 (right) for each of 5 samples

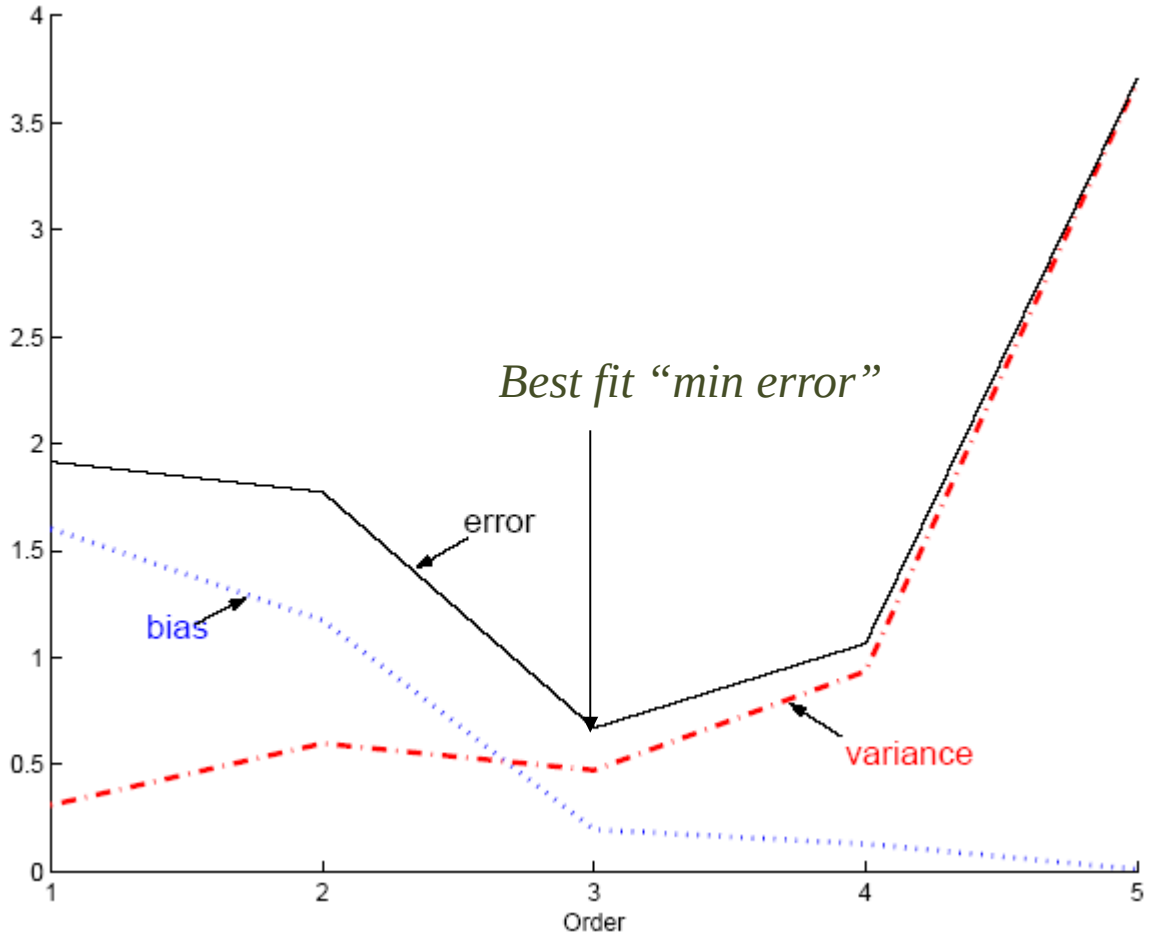


variance

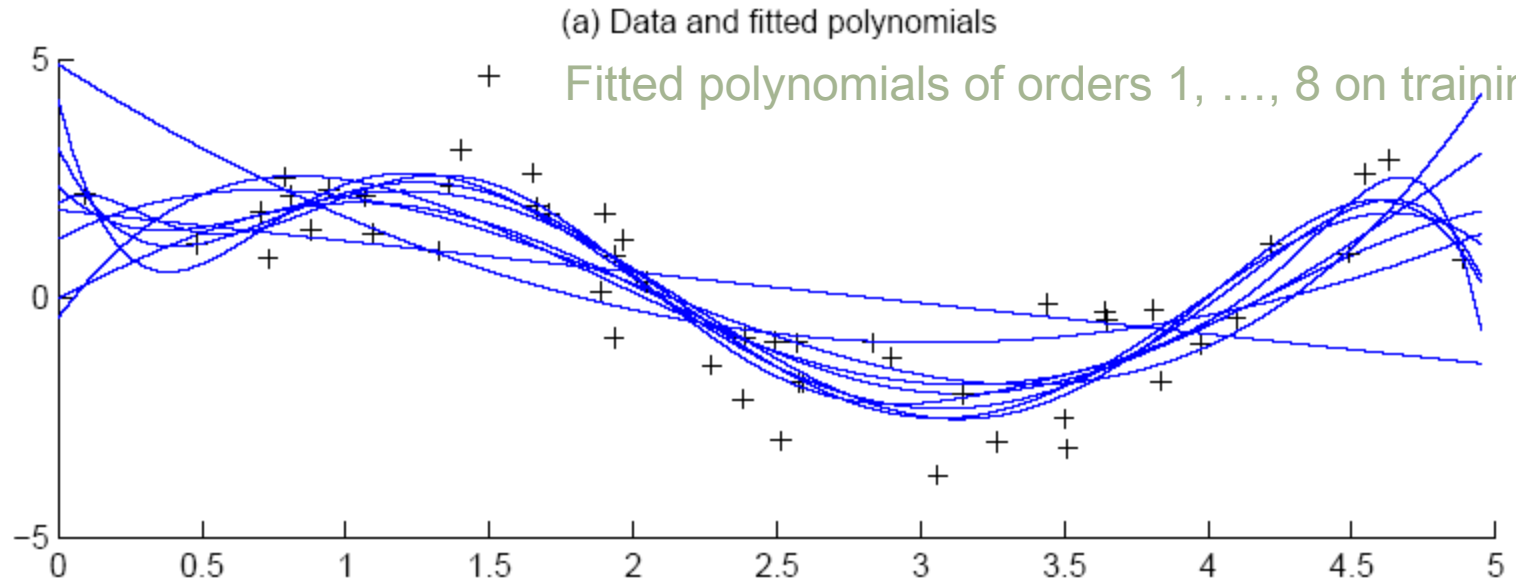


dotted red line in each plot is the average of the five fits, $\bar{g}(x)$

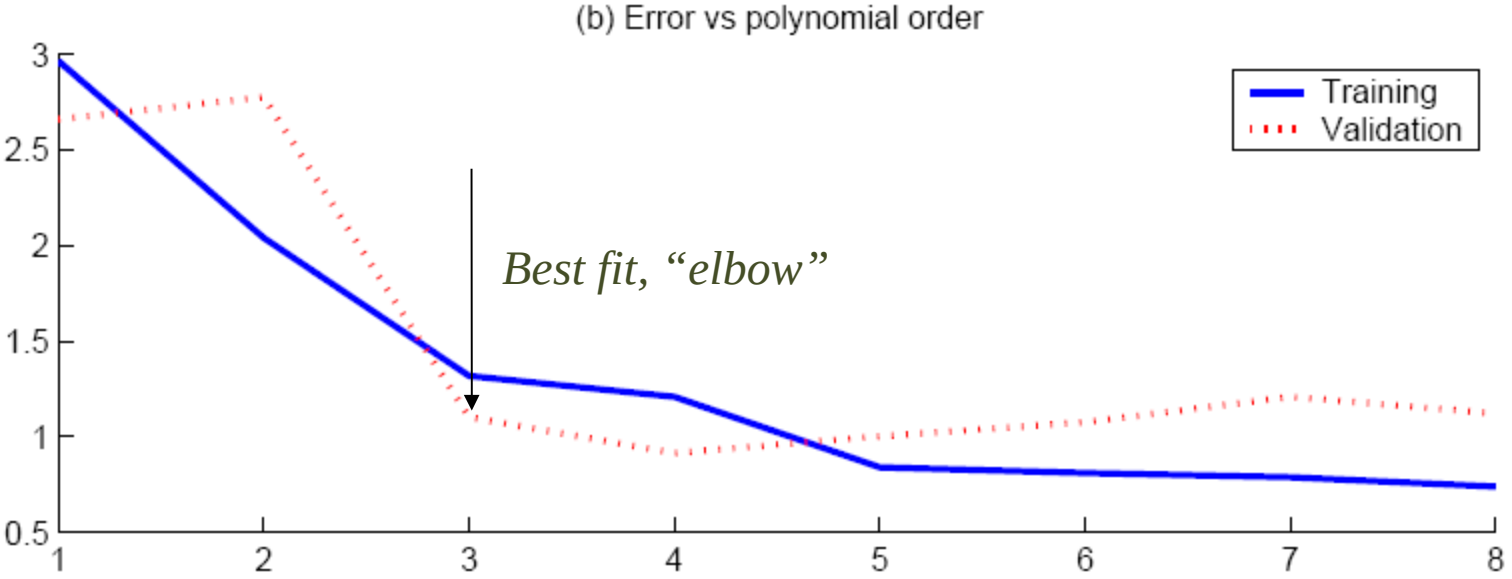
Polynomial Regression



Same settings as plots on the previous slide, but using 100 models instead of 5



Fitted polynomials of orders 1, …, 8 on training set



Settings as plots on previous slides, but using training and validation sets (50 instances eac

Model Selection (1 of 2 slides)
Methods to fine-tune model complexity

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

  E'=error on data + λ model complexity (where λ is a penalty weight)

  the lower the better
- Other measures of "goodness of fit" with complexity penalty:
  - Akaike's information criterion (AIC)

    AIC ≡ $log\ p(X|\theta_{ML},M) - k(M)$
  - Bayesian information criterion (BIC)

    BIC ≡ $log\ p(X|\theta_{ML},M) - k(M)\ log(N)/2$

    where: M is a model

    $log\ p(X|\theta_{ML},M)$: is the *log likelihood of M where M's parameters* θ_{ML} *have been estimated using maximum likelihood*

    $k(M)$ = number of adjustable parameters in θ_{ML} of the model M

    N= size of sample X

  For both AIC and BIC, the higher the better

Model Selection (2 of 2 slides)

Methods to fine-tune model complexity

- Minimum description length (MDL): Kolmogorov complexity, shortest description of data

  Given a dataset X,

  MDL (M) = Description length of model M +

  Description length of data in X not correctly described by M

  the lower the better

- Structural risk minimization (SRM)
  - Uses a set of models and their complexities (measured usually using their number of free parameters or their VC-dimension)
  - Selects the simplest model in terms of order and best in terms of empirical error on the data

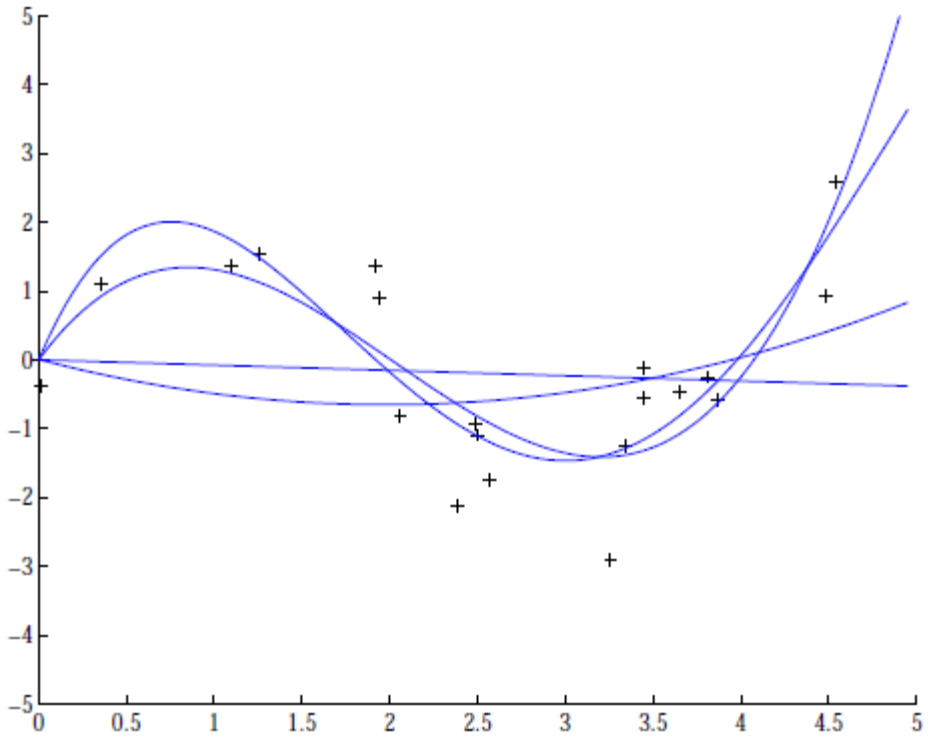
Bayesian Model Selection

- Used when we have Prior on models, $p(\text{model})$

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

- 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 17)

Regression example



Coefficients increase in magnitude as order increases:
1: [-0.0769, 0.0016]
2: [0.1682, -0.6657, 0.0080]
3: [0.4238, -2.5778, 3.4675, -0.0002
4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]

Regularization (L2): $E(\mathbf{w}|\mathbf{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t|\mathbf{w}\right)\right]^2 + \lambda\sum_i w_i^2$