



Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

© The MIT Press, 2004

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml>



CHAPTER 4:

Parametric Methods

Parametric Estimation

- $X = \{ x^t \}_t$ where $x^t \sim p(x)$
- Our model is a specified probability distribution. Learning = estimating its parameters.
- Parametric estimation:

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

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

Density estimation (can then be used for classification, etc.).

Maximum Likelihood Estimation

- Likelihood of θ given the sample X

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

- Log likelihood

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

- Why? Small numbers, converts product to sum, removes exp(?)

- Maximum likelihood estimator (MLE)

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

Examples: Bernoulli/Multinomial

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

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

$$L(p_o | \mathbf{X}) = \log \prod_t p_o^{x_t} (1 - p_o)^{(1-x_t)}$$

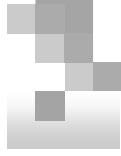
$$\text{MLE: } p_o = \sum_t x_t / N$$

- Multinomial: $K > 2$ states, x_i in $\{0,1\}$

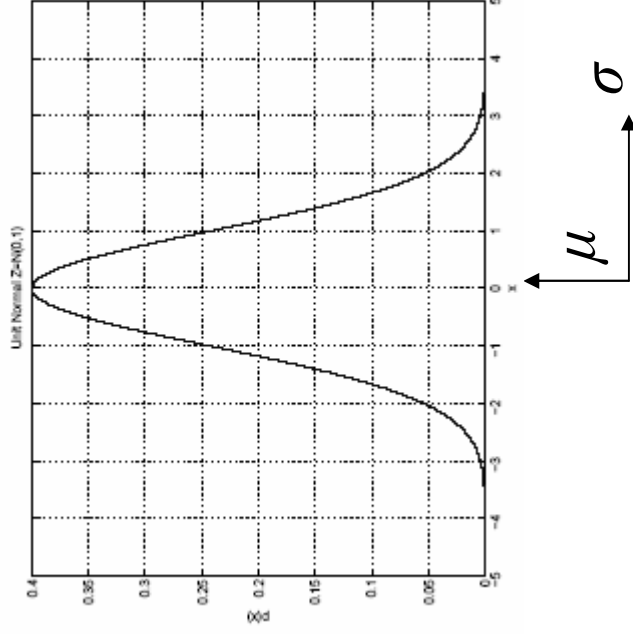
$$P(x_1, x_2, \dots, x_K) = \prod_i p_i^{x_i}$$

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

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



Gaussian (Normal) Distribution



true parameters

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

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

- MLE for μ and σ^2 :

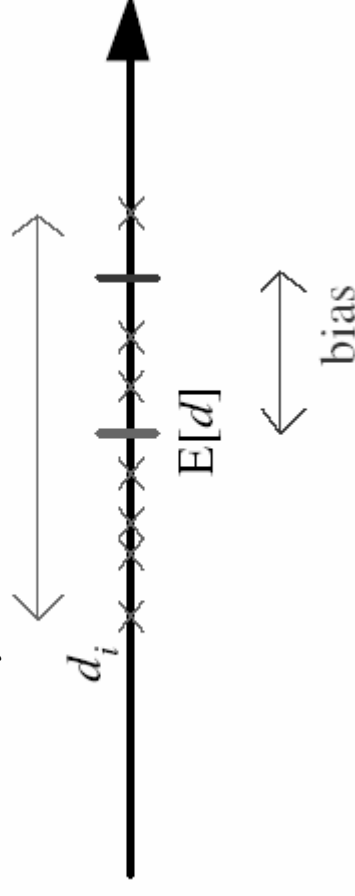
estimated values

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

Bias and Variance

Unknown parameter θ

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




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

$$\text{Variance: } E[(d - E[d])^2]$$

Mean square error:

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

- 
- Bias: how much the expected value of the estimator varies from the correct
 - Variance: variation around the expected value.
 - Examples:
 - Sample mean, m , is an unbiased estimator of the true mean. It's also a consistent estimator, since $\text{Var}(m)$ tends to zero as N tends to infinity.
 - Sample variance (as it turns out) is a biased estimator of the true variance.

Bayes' Estimator

- Treat θ as a random var with prior $p(\theta)$
- Bayes' rule: $p(\theta|X) = p(X|\theta)p(\theta) / p(X)$
- Full: $p(x|X) = \int p(x|\theta)p(\theta|X) d\theta$
 - i.e an average over predictions using all values of θ , weighted by the probability of each θ value.
- Maximum a Posteriori (MAP): $\theta_{\text{MAP}} = \operatorname{argmax}_{\theta} p(\theta|X)$
 - This uses a prior
- Maximum Likelihood (ML): $\theta_{\text{ML}} = \operatorname{argmax}_{\theta} p(X|\theta)$
 - This doesn't have a prior. If prior is flat, MAP == ML.
- Bayes': $\theta_{\text{Bayes'}} = E[\theta|X] = \int \theta p(\theta|X) d\theta$

Bayes' Estimator: Example

- $x^t \sim \mathbf{N}(\theta, \sigma_0^2)$ and $\theta \sim \mathbf{N}(\mu, \sigma^2)$
- $\theta_{\text{ML}} = m$
- $\theta_{\text{MAP}} = \theta_{\text{Bayes}} = \frac{N / \sigma_0^2}{N / \sigma_0^2 + 1 / \sigma^2} m + \frac{1 / \sigma^2}{N / \sigma_0^2 + 1 / \sigma^2} \mu$

Parametric Classification

$$g_i(x) = p(x | C_i)P(C_i)$$

or equivalently

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

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

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

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

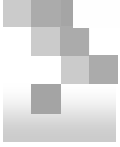
$$x \in \mathcal{X} \quad r_i^t = \begin{cases} 1 & \text{if } x^t \in C_i \\ 0 & \text{if } x^t \in C_j, j \neq i \end{cases}$$

- ML estimates are

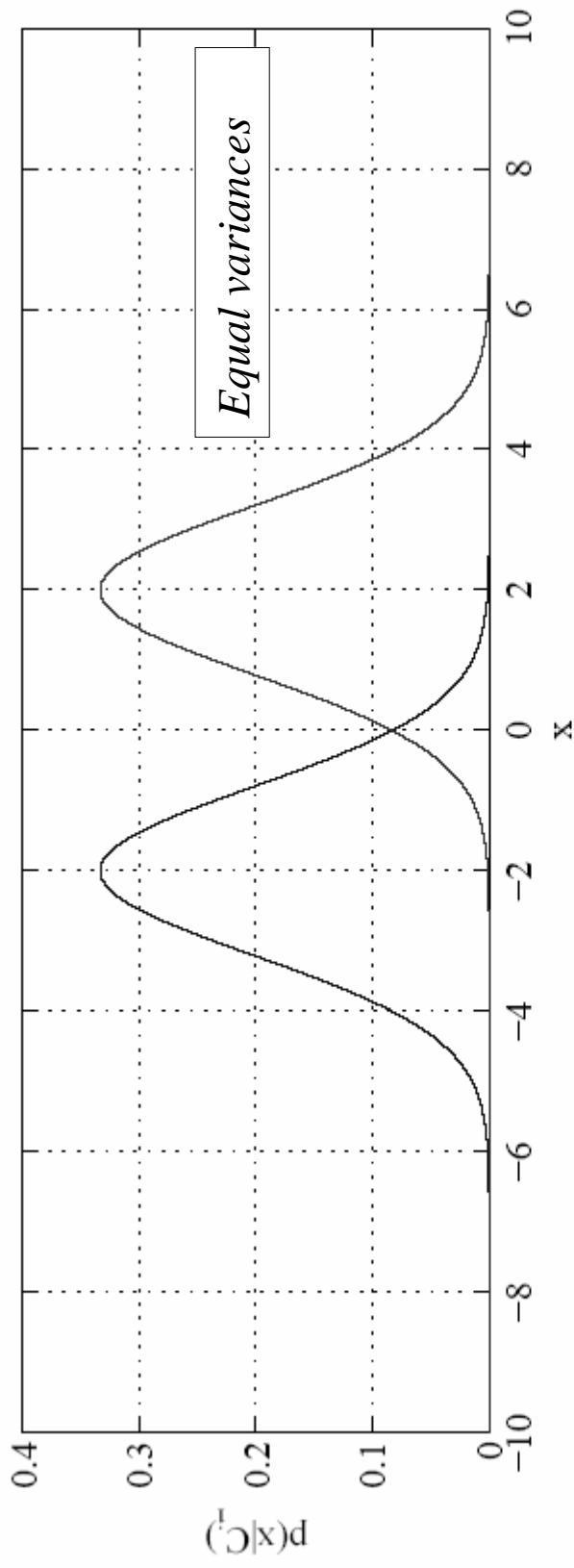
$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

- Discriminant becomes

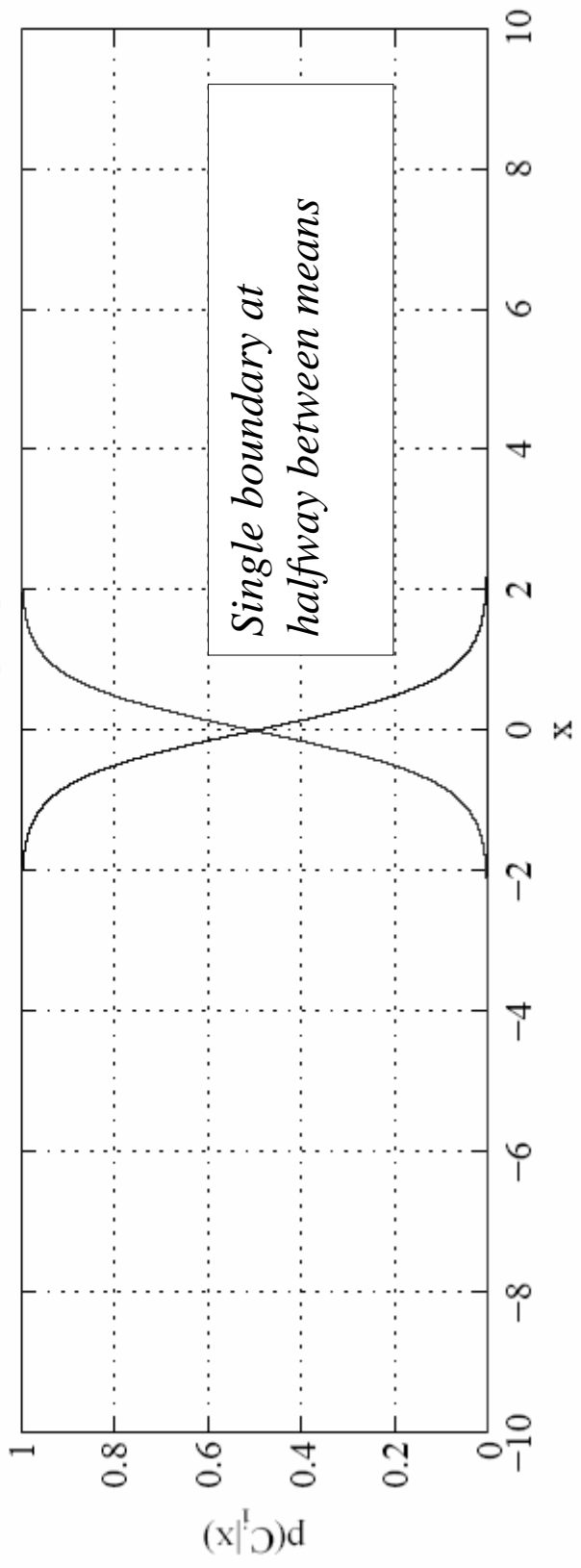
$$g_i(x) = -\frac{1}{2} \log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$



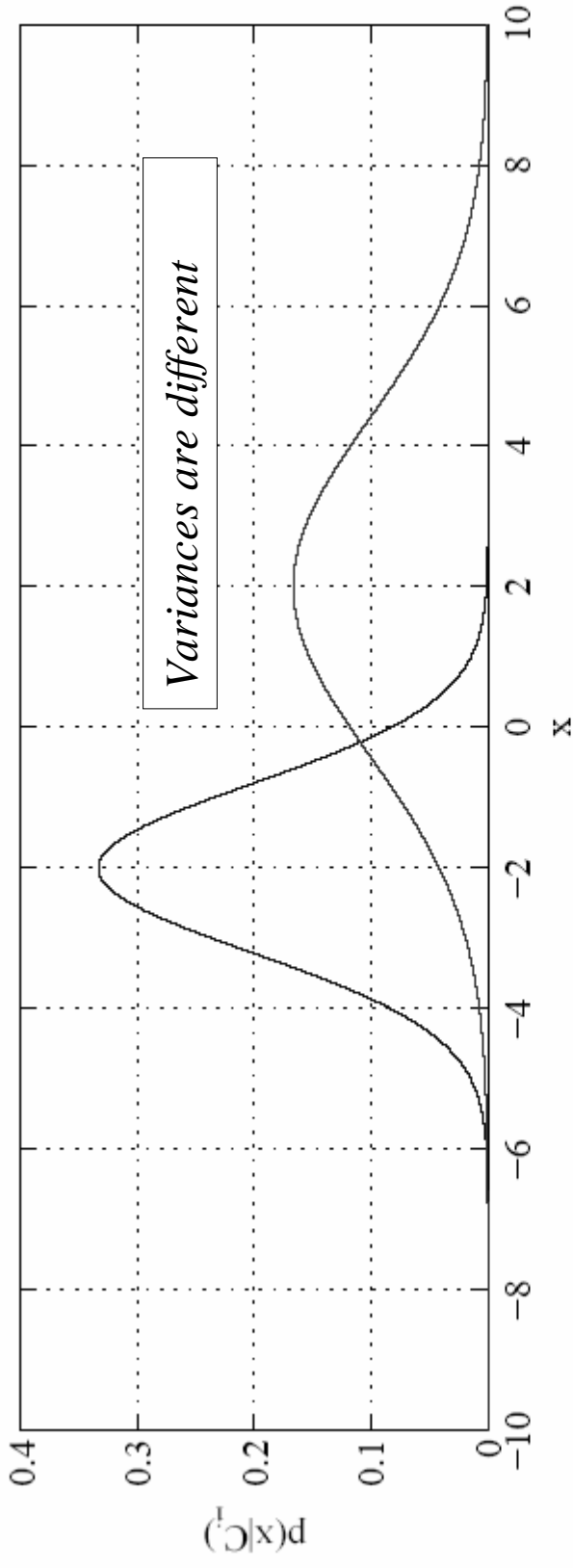
Likelihoods



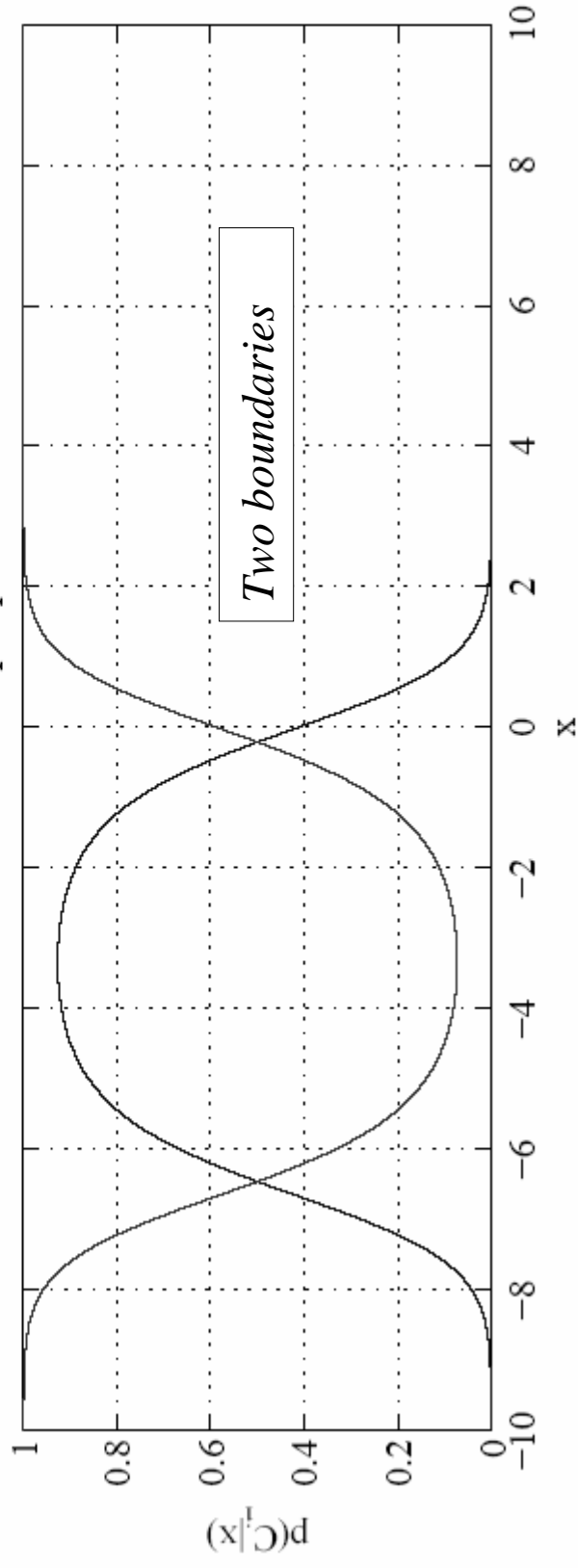
Posteriors with equal priors



Likelihoods



Posteriors with equal priors



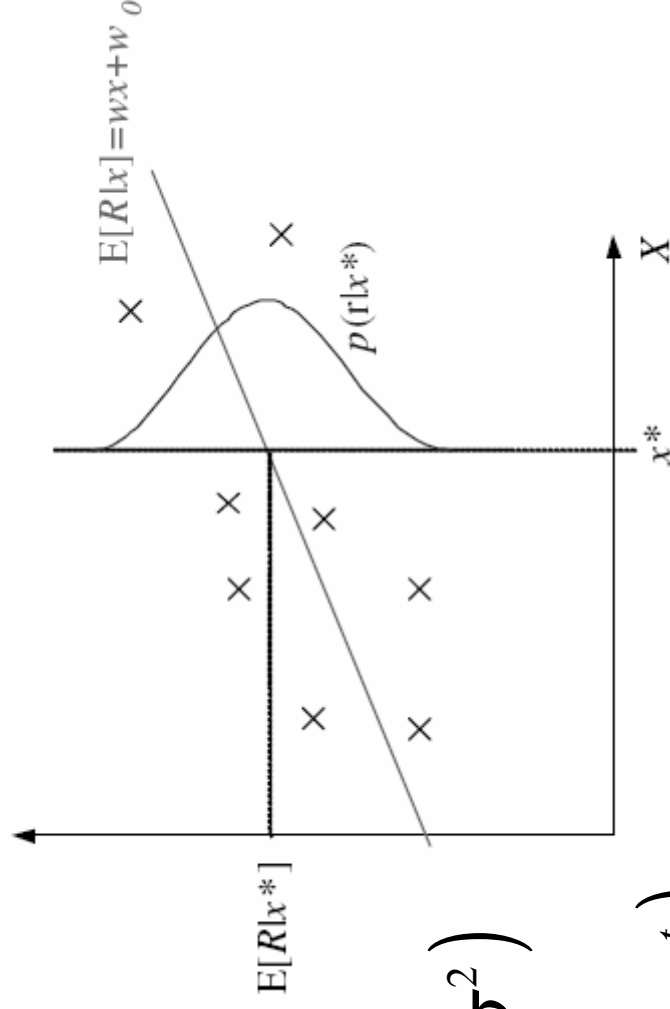
Regression

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

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

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

$$p(r | x) \sim \mathbf{N}(g(x | \theta), \sigma^2)$$



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

We want to learn the parameters of our model via Max. Likelihood

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

Regression: From LogL to Error

$$\begin{aligned}L(\theta | \mathbf{X}) &= \log \prod_{t=1}^N \frac{1}{\sqrt{2\pi\sigma}} \exp \left[-\frac{[r^t - g(x^t | \theta)]^2}{2\sigma^2} \right] \\ &= -N \log \sqrt{2\pi\sigma} - \frac{1}{2\sigma^2} \sum_{t=1}^N [r^t - g(x^t | \theta)]^2 \\ E(\theta | \mathbf{X}) &= \frac{1}{2} \sum_{t=1}^N [r^t - g(x^t | \theta)]^2\end{aligned}$$

Important: so maximizing L is equivalent to minimizing MSE (assuming Gaussian noise)

Linear Regression

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

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

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

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

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

Polynomial Regression

$$g(x^t | w_k, \dots, w_2, w_1, w_0) = w_k (x^t)^k + \dots + w_2 (x^t)^2 + w_1 x^t + w_0$$

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

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

Other Error Measures

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

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

Estimating Bias and Variance

- M samples $X_i = \{x_i^t, r_i^t\}$, $i=1, \dots, M$ are used to fit $g_i(x)$, $i=1, \dots, M$

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

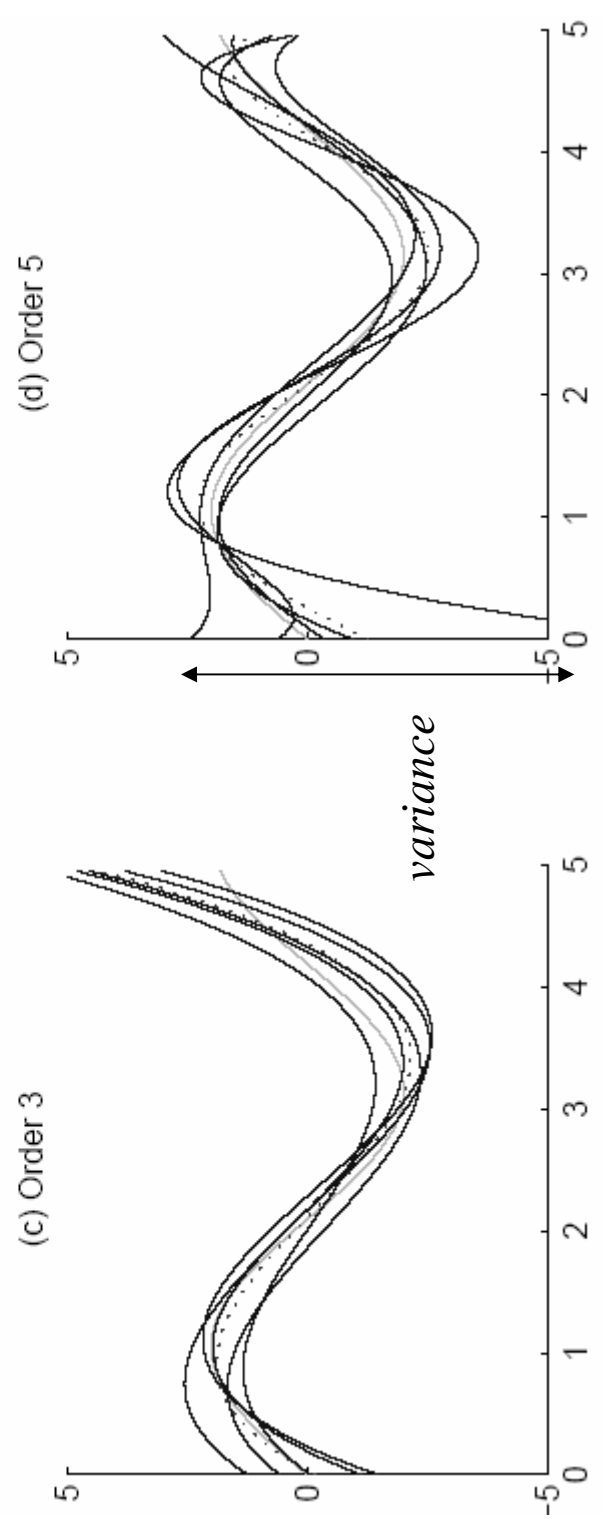
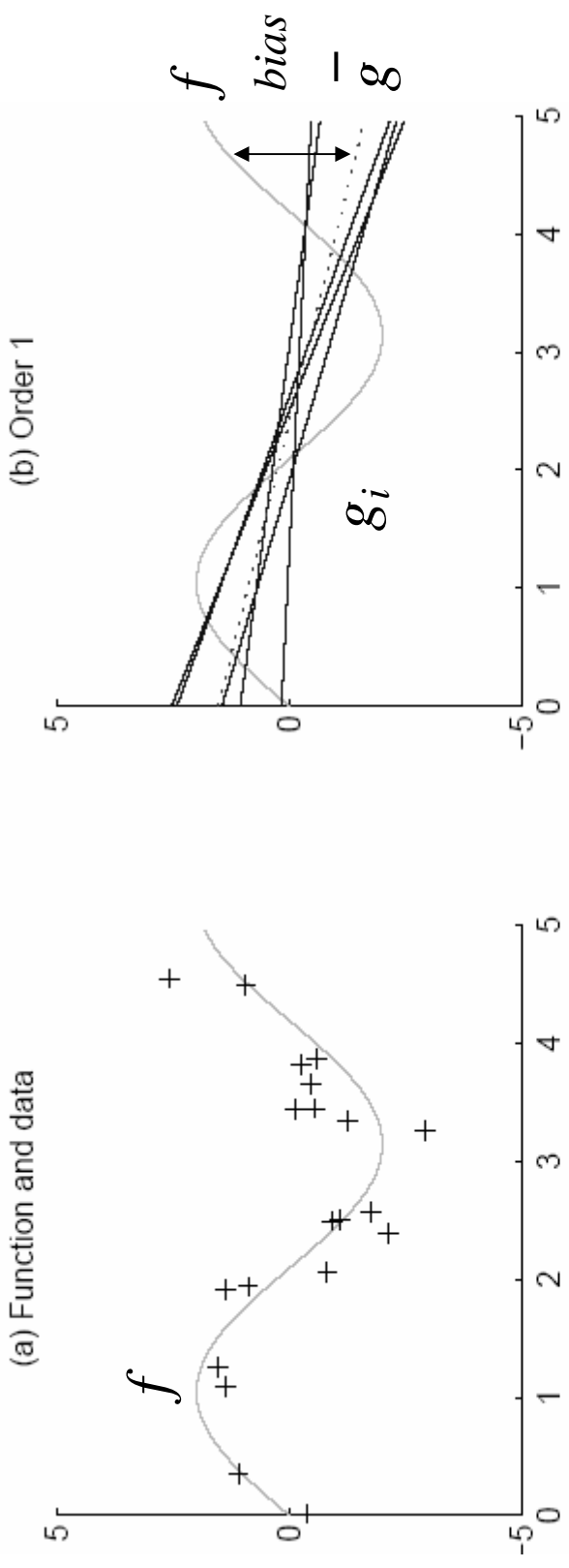
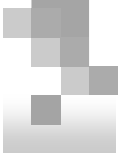
$$\text{Variance}(g) = \frac{1}{NM} \sum_t \sum_i [g_i(x^t) - \bar{g}(x^t)]^2$$

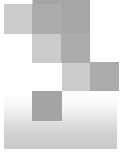
$$\bar{g}(x) = \frac{1}{M} \sum_t g_i(x)$$



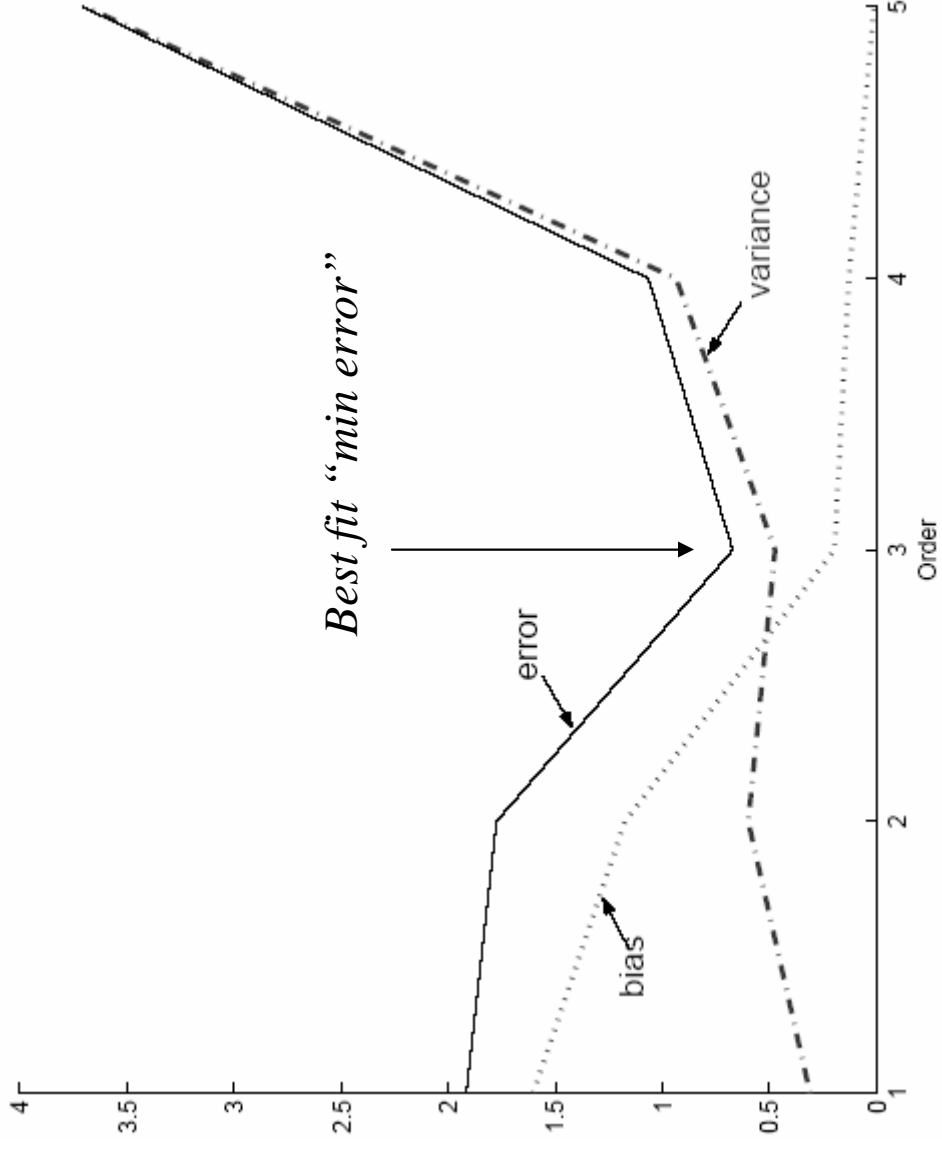
Bias/Variance Dilemma

- Example: $g_i(x)=2$ has no variance and high bias
- $g_i(x)=\sum_t r_t^i/N$ 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)

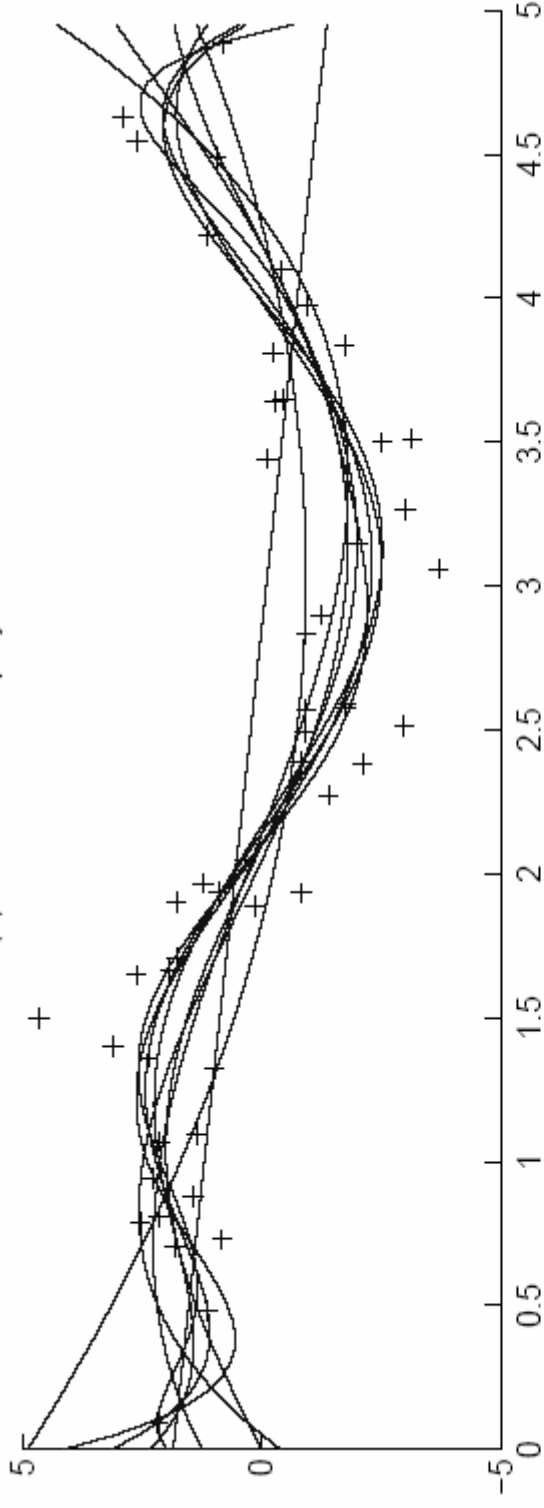




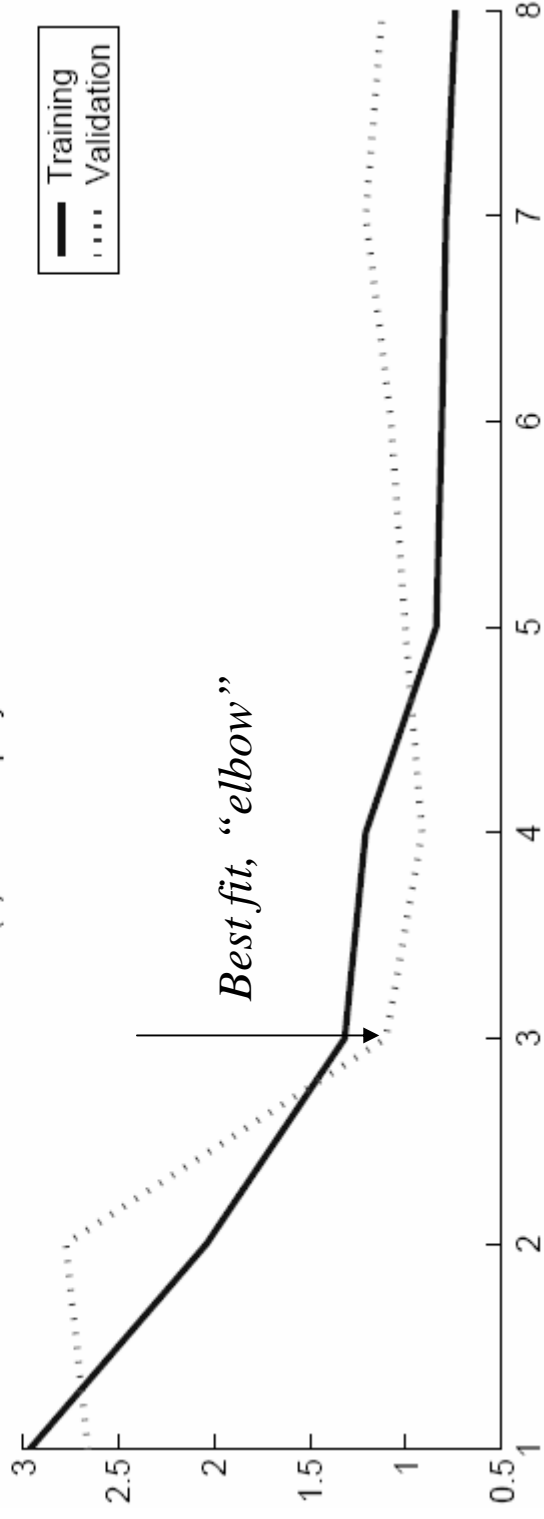
Polynomial Regression



(a) Data and fitted polynomials



(b) Error vs polynomial order





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

Akaike's information criterion (AIC), Bayesian information criterion (BIC)

- 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, $p(\text{model} \mid \text{data})$
- Average over a number of models with high posterior (voting, ensembles: Chapter 15)