MACHİNE LEARNİNG

Chao Wang (*Frank*)

wangchao@nankai.edu.cn

CHAPTER 4:

PARAMETRİC METHODS

Parametric Estimation

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

  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 $\mathcal{X}$

$l(\vartheta | \mathcal{X}) = p(\mathcal{X} | \vartheta) = \prod_t p(x^t | \vartheta)$

- Log likelihood

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

- Maximum likelihood estimator (MLE)

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

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

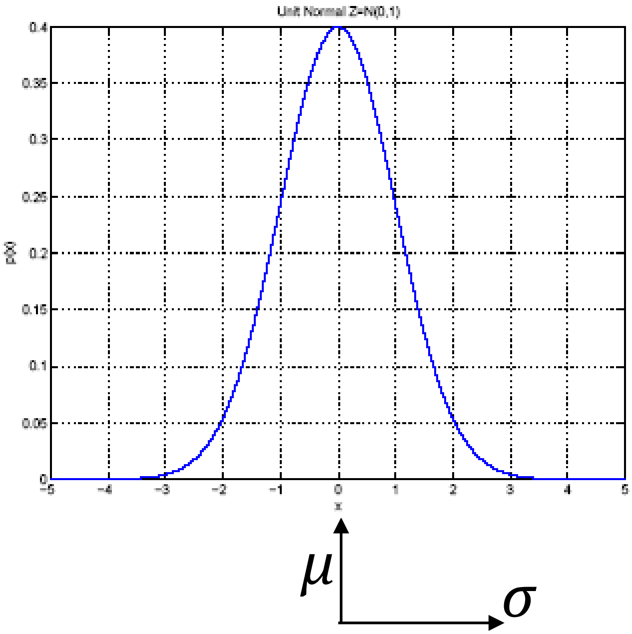
- Multinomial: $K > 2$ states, x_i in {0,1}

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

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

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

Gaussian (Normal) Distribution



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

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

- MLE for μ and σ^2:

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

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

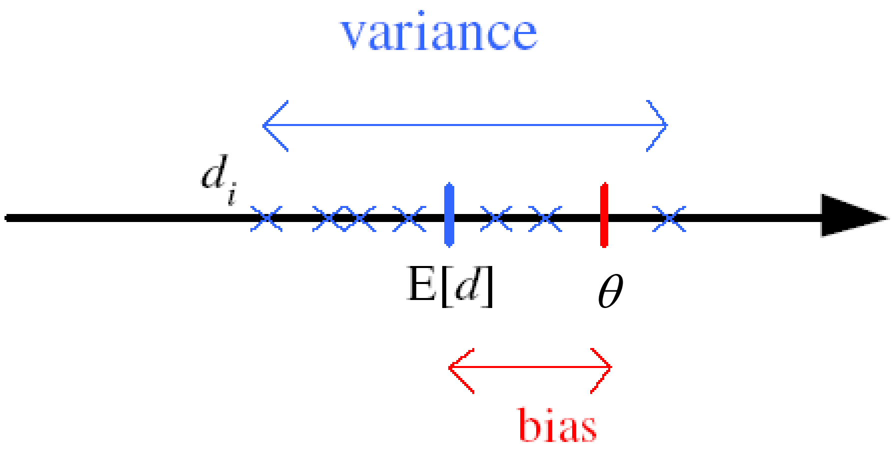
Bias and Variance

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:

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

Bayes' Estimator

- Treat ϑ as a random var with prior $p(\vartheta)$
- Bayes' rule:

  $$p(\vartheta|\mathcal{X}) = p(\mathcal{X}|\vartheta)\, p(\vartheta) \,/\, p(\mathcal{X})$$
- Full: $p(x|\mathcal{X}) = \int p(x|\vartheta)\, p(\vartheta|\mathcal{X})\, d\vartheta$
- Maximum a Posteriori (MAP):

  $$\vartheta_{MAP} = \text{argmax}_\vartheta\, p(\vartheta|\mathcal{X})$$
- Maximum Likelihood (ML):

  $$\vartheta_{ML} = \text{argmax}_\vartheta\, p(\mathcal{X}|\vartheta)$$
- Bayes':

  $$\vartheta_{Bayes'} = E[\vartheta|\mathcal{X}] = \int \vartheta\, p(\vartheta|\mathcal{X})\, d\vartheta$$

Bayes'Estimator: Example

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

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

Parametric Classification

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

or

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

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

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

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

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

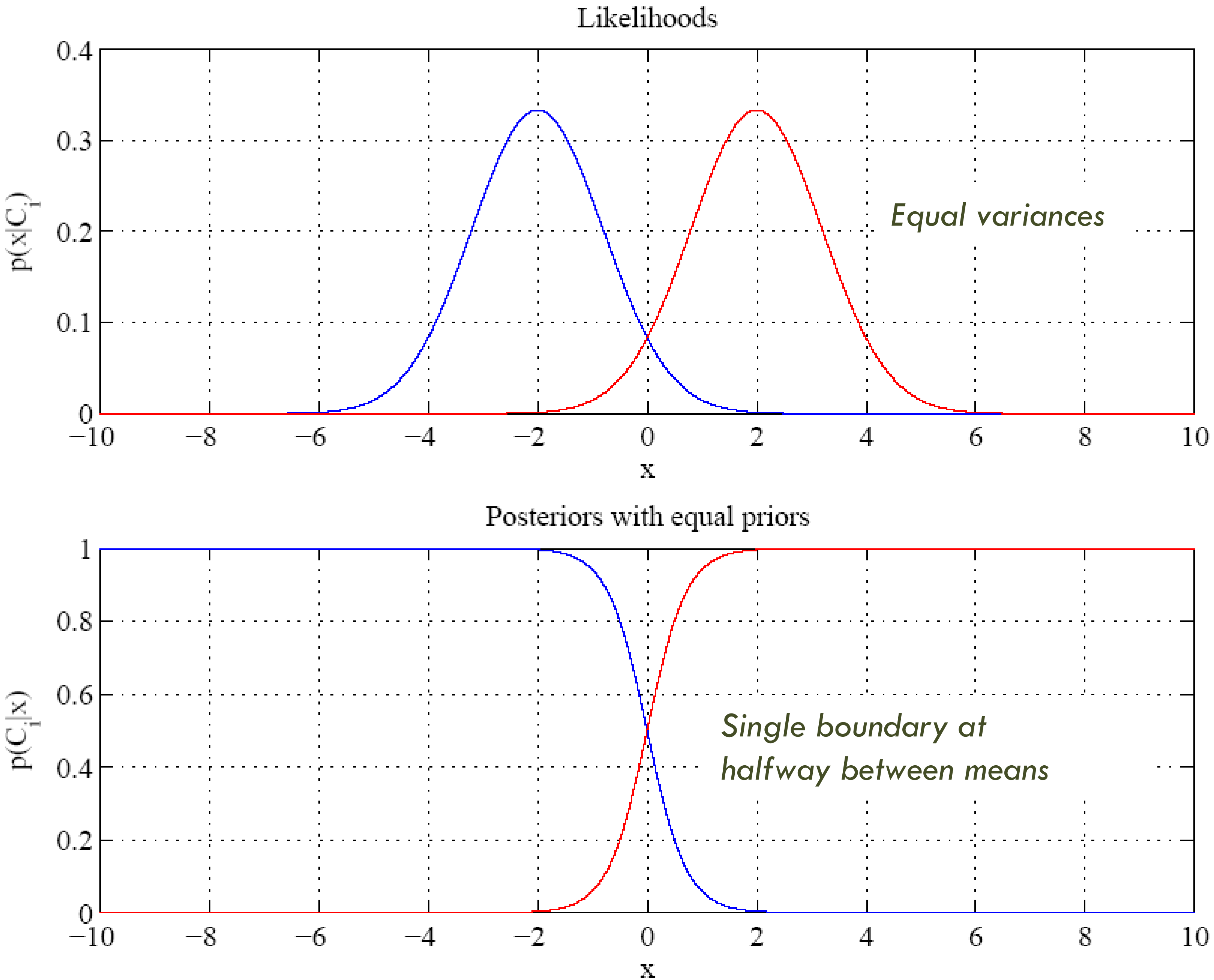
- ML estimates are

$\hat{p}_i$

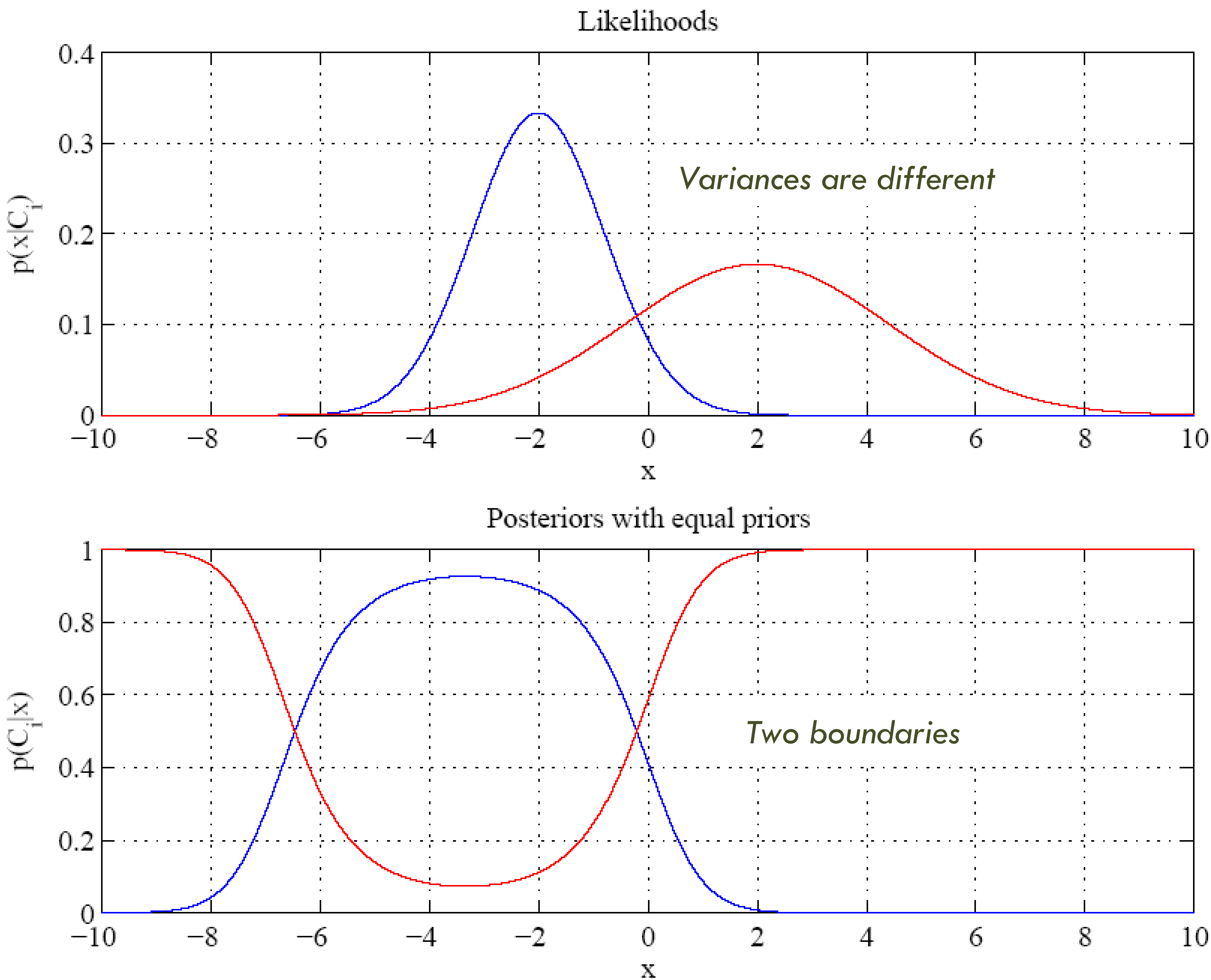
$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t \left(x^t - m_i\right)^2 r_i^t}{\sum_t r_i^t}$$

- Discriminant

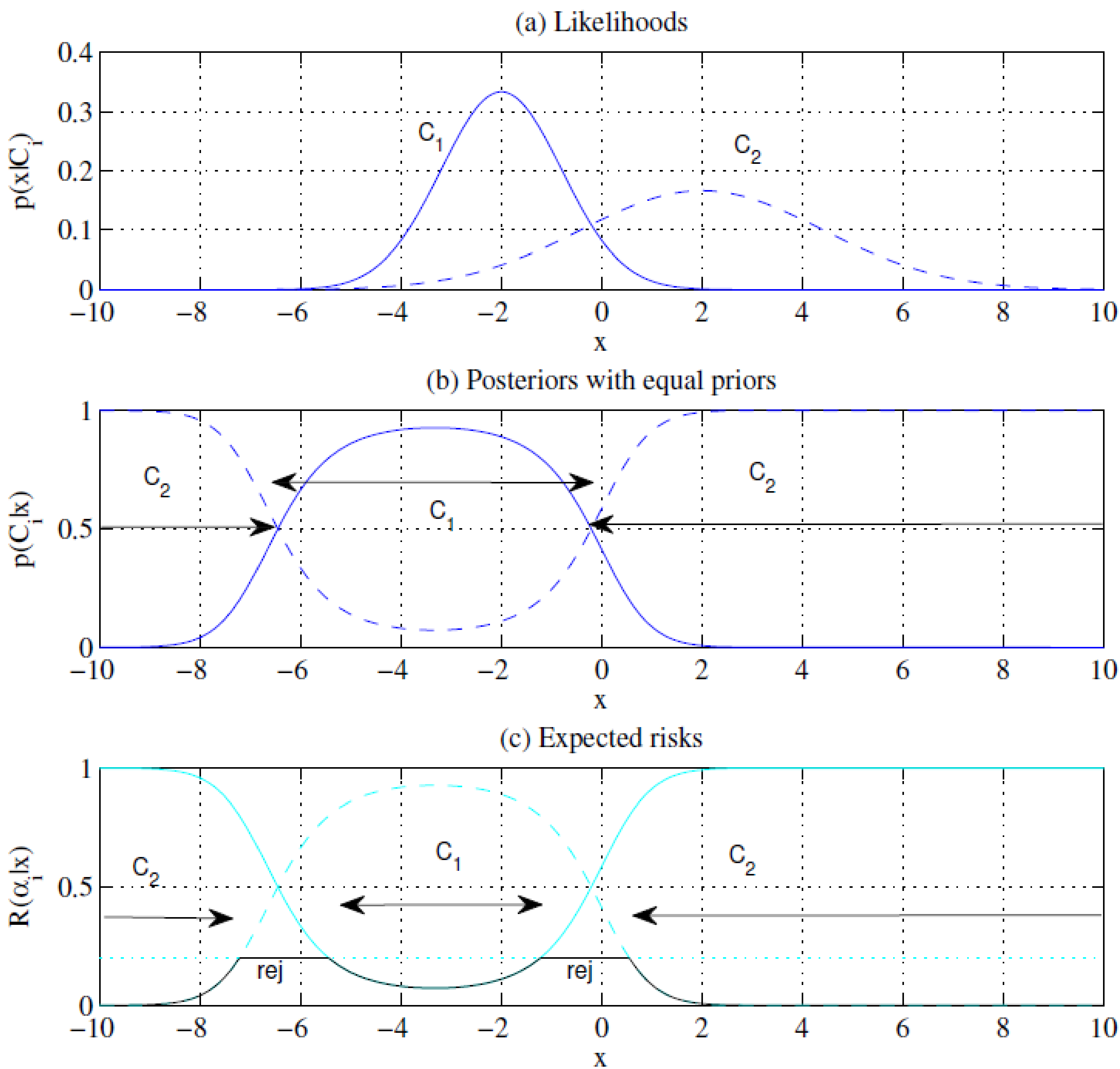
$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{\left(x - m_i\right)^2}{2{s_i}^2} + \log \hat{P}(C_i)$$



Likelihoods
p(x|C_i)
0.4
0.3
0.2
0.1
0
−10
−8
−6
−4
−2
0
2
4
6
8
10
x
Equal variances
Posteriors with equal priors
p(C_i|x)
1
0.8
0.6
0.4
0.2
0
−10
−8
−6
−4
−2
0
2
4
6
8
10
x
Single boundary at
halfway between means



Likelihoods
p(x|C_i)
Variances are different
x
Posteriors with equal priors
p(C_i|x)
Two boundaries
x



(a) Likelihoods
p(x|C_i)
C_1
C_2
x
(b) Posteriors with equal priors
p(C_i|x)
C_2
C_1
C_2
x
(c) Expected risks
R(α_i|x)
C_2
C_1
C_2
rej
rej
x

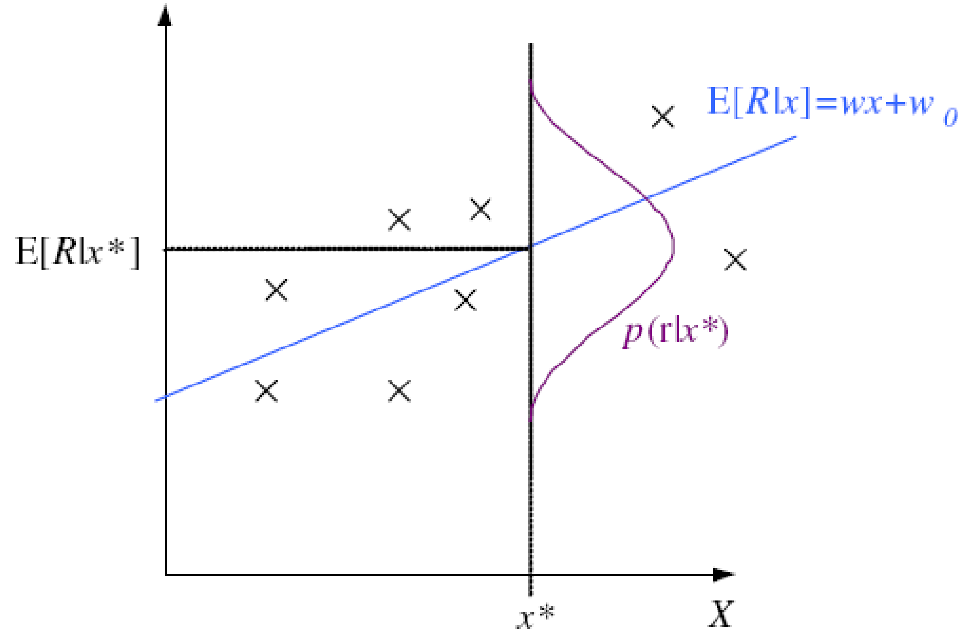
Regression

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

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

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

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



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

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

Regression: From LogL to Error

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

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

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

Linear Regression

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

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

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

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

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

Polynomial Regression

$$g\left(x^t \mid w_k,\ldots,w_2,w_1,w_0\right)=w_k\left(x^t\right)^k+\cdots+w_2\left(x^t\right)^2+w_1x^t+w_0$$

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

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

Other Error Measures

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

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

Bias and Variance

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

noise *squared error*

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

bias *variance*

Estimating Bias and Variance

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

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

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

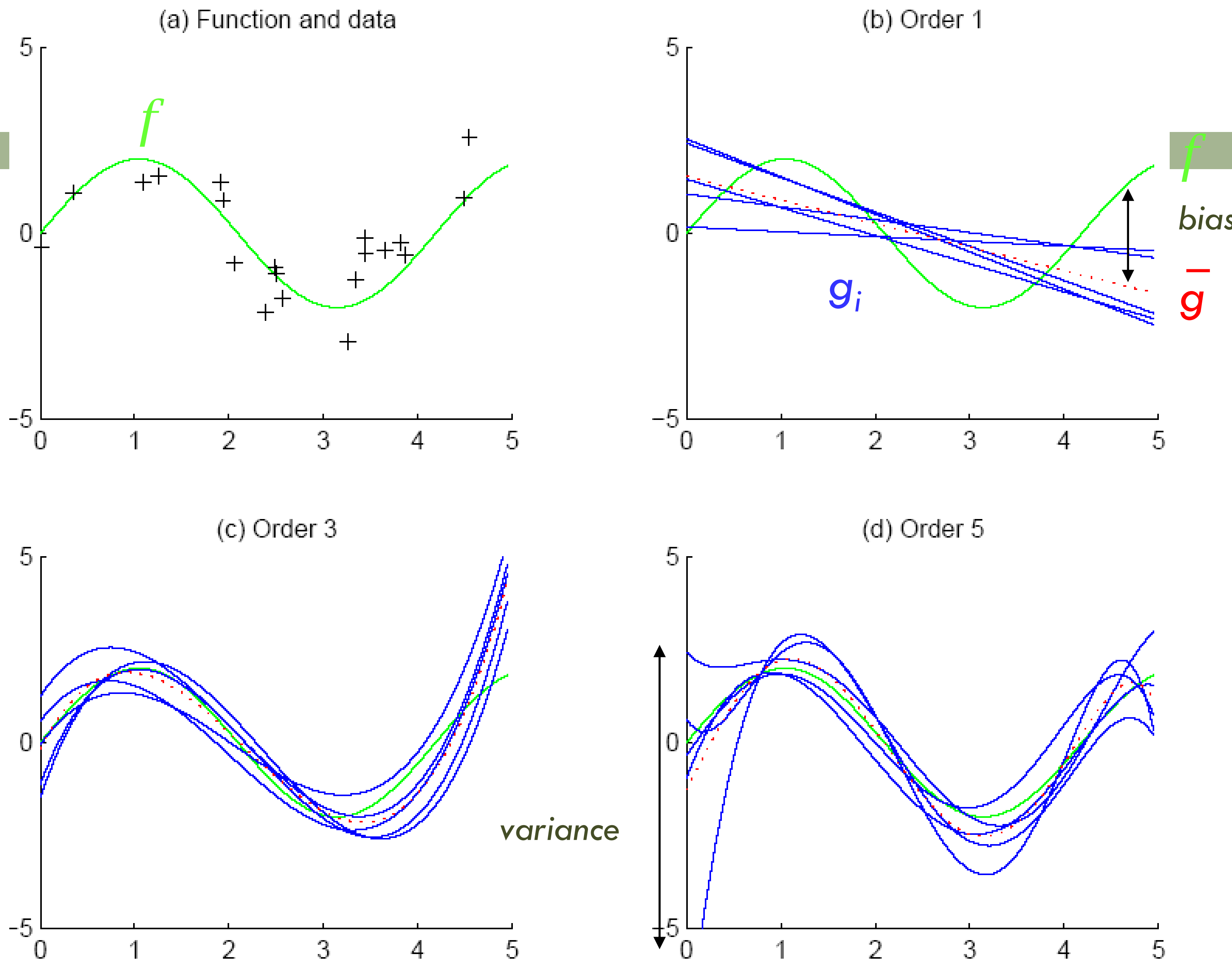
$$\bar{g}(x)=\frac{1}{M}\sum_t g_i(x)$$

Bias/Variance Dilemma

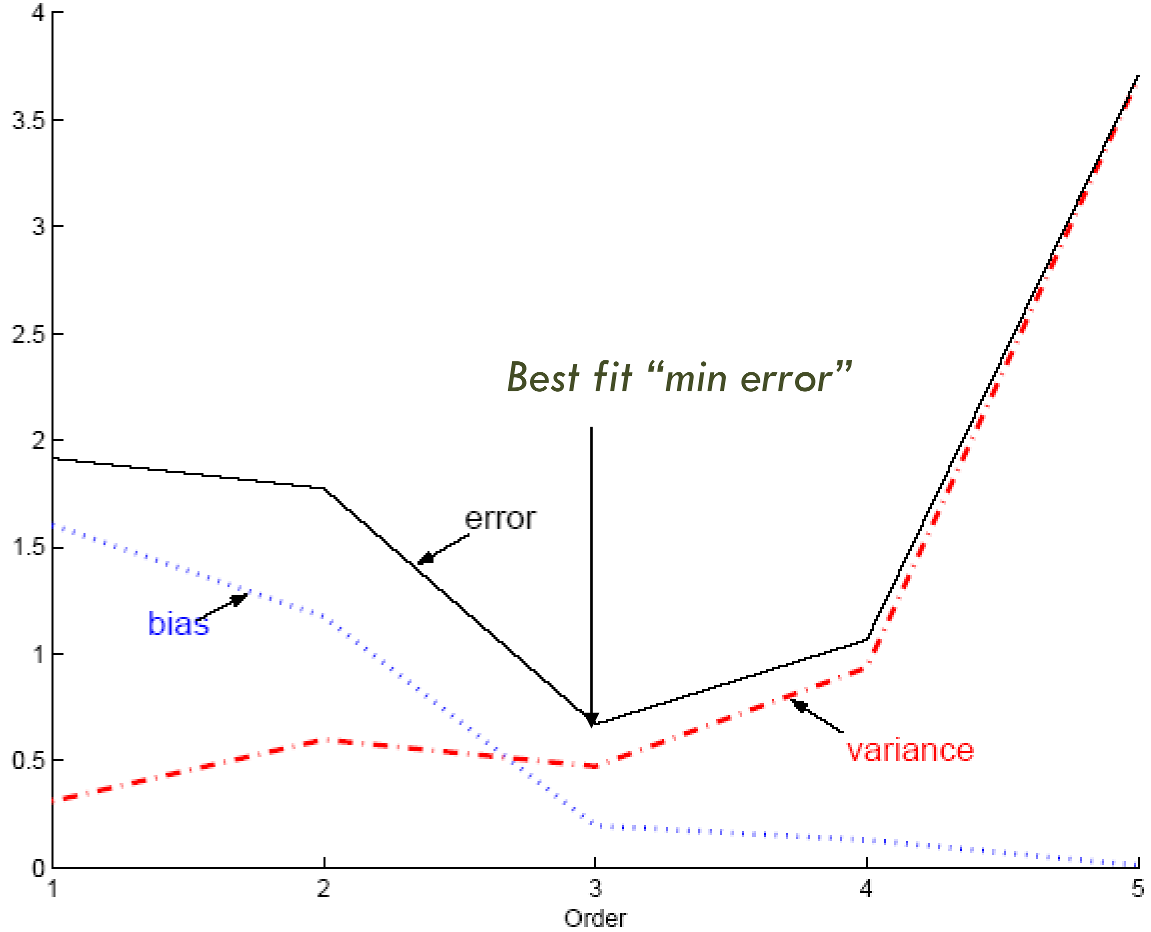
- Example: $g_i(x)=2$ has no variance and high bias

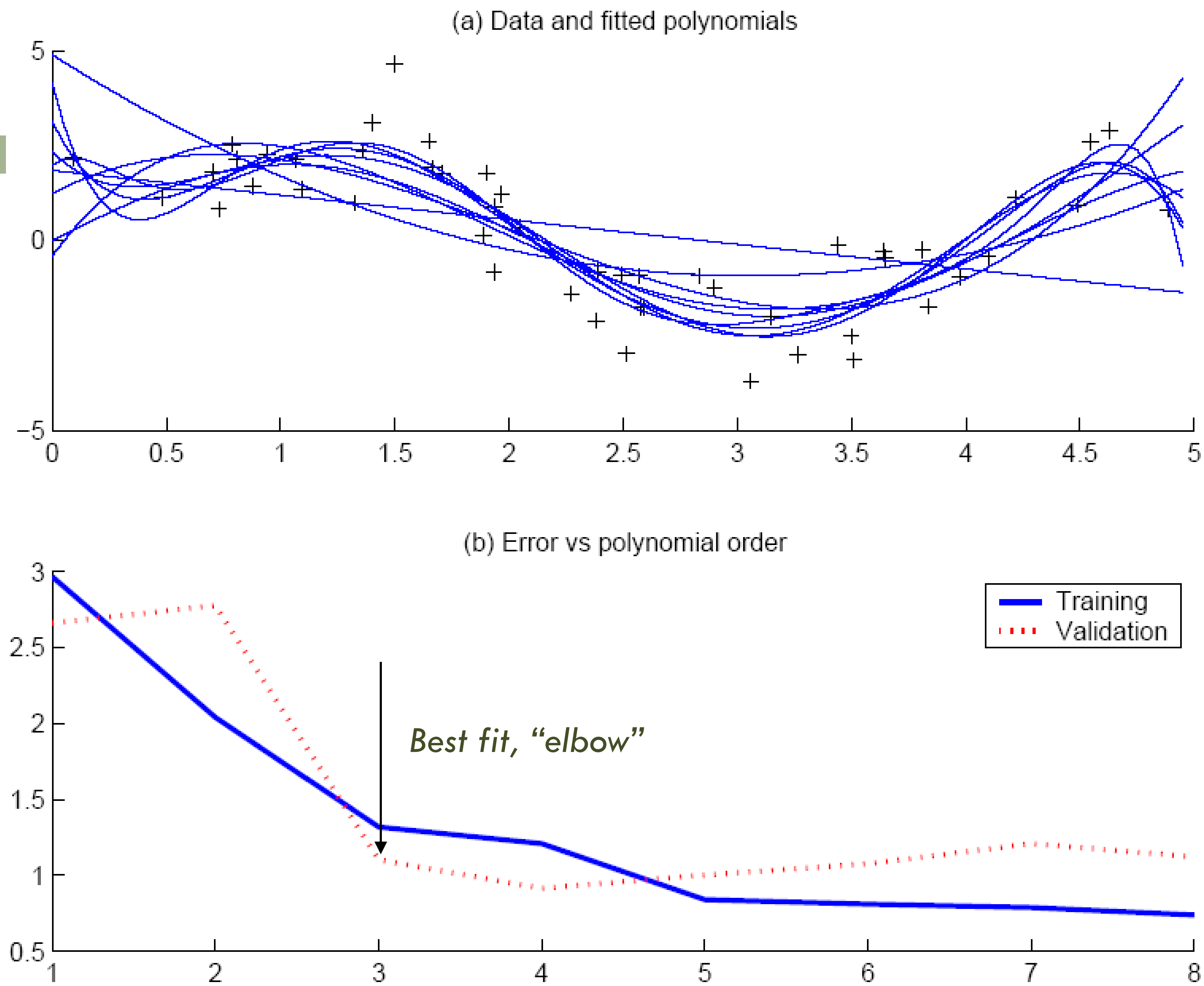
  $g_i(x)= \sum_t r^t_i/N$ has lower bias with variance

- 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)


(a) Function and data
f
(b) Order 1
f
bias
g_i
$\bar{g}$
(c) Order 3
variance
(d) Order 5

Polynomial Regression




(a) Data and fitted polynomials
5
0
−5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
(b) Error vs polynomial order
3
2.5
2
1.5
1
0.5
1
2
3
4
5
6
7
8
Training
Validation
Best fit, "elbow"

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)

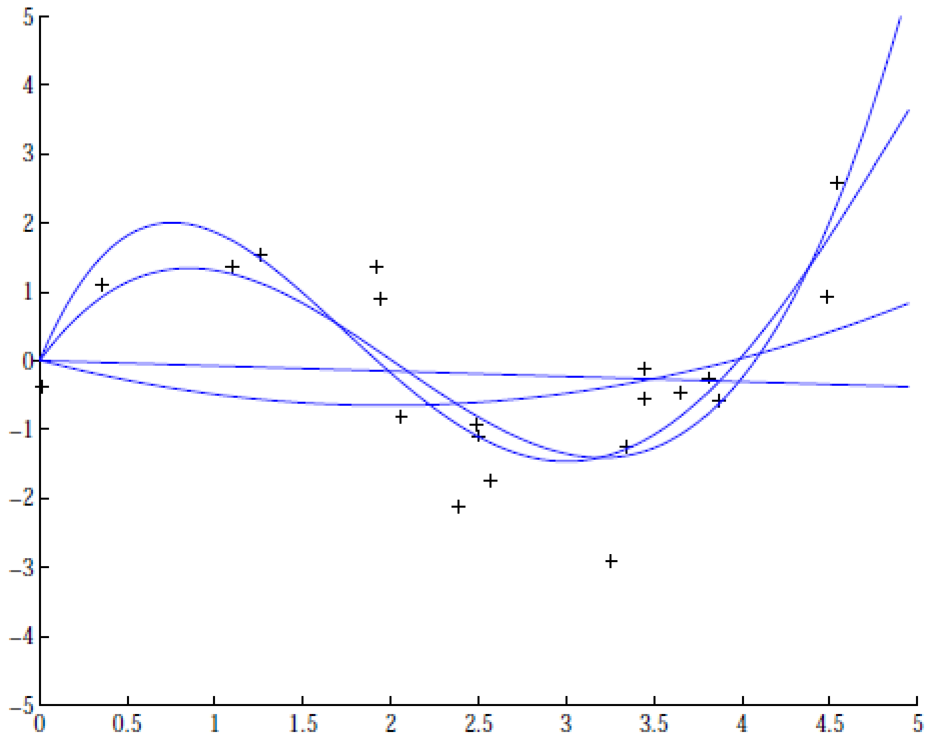
Bayesian Model Selection

- Prior on models, p(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(model|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): $$E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \mathbf{w}\right)\right]^2 + \lambda \sum_i w_i^2$$

Questions?