BLG 454E Learning From Data

FALL 2024-2025

Assoc. Prof. Yusuf Yaslan

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

Maximum Likelihood Estimation

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

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

- Log likelihood

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

- Maximum likelihood estimator (MLE)

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

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

Examples: Bernoulli (Derivation)

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

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

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

$$\frac{dL(p_0 \mid X)}{dp_0} = \sum_{t=1}^{N} x^t \frac{d}{dp_0}\log(p_0) = \sum_{t=1}^{N}(1 - x^t)\frac{d}{dp_0}\log(1 - p_0)$$

$$= \frac{1}{p_0}\sum_{t=1}^{N} x^t - \sum_{t=1}^{N}(1 - x^t)\frac{1}{1 - p_0} = 0$$

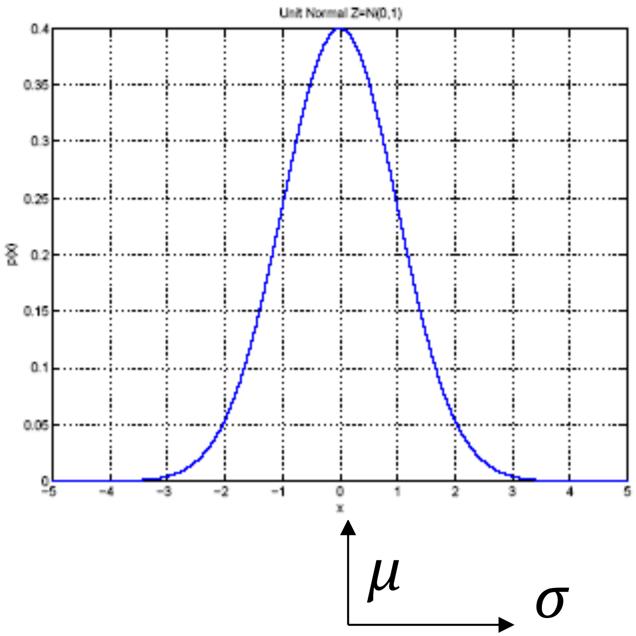
Bernoulli (Derivation)

$$\Rightarrow (1-p_0)\sum_{t=1}^{N} x^t - p_0 \sum_{t=1}^{N} 1 + p_0 \sum_{t=1}^{N} x^t = 0$$

$$\Rightarrow \sum_{t=1}^{N} x^t - p_0 N = 0 \Rightarrow p_0 = \frac{1}{N}\sum_{t=1}^{N} x^t$$

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

Gaussian (Normal) Distribution



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

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

Gaussian (Normal) Distribution

- Given that $\mathcal{X} = \{ x^t \}_t$ with $x^t \sim \mathcal{N}(\mu, \sigma^2)$

$$L(\mu, \sigma | X) = -\frac{N}{2}\log(2\pi) - N log(\sigma) - \frac{\sum_{n=1}^{N}(x^t - \mu)^2}{2\sigma^2}$$

MLE for μ and σ^2:

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

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

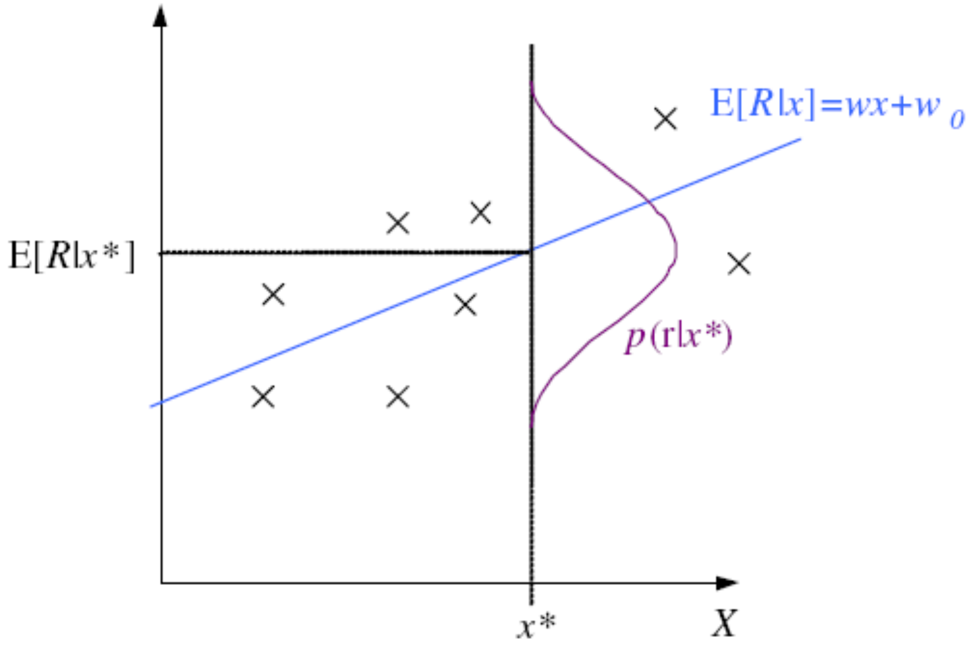
Probabilistic Interpretation of Linear Regression

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

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

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

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



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

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

Regression: From LogL to Error

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

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

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

Linear Regression

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

$$g\left(x^t|w_1,w_0\right)=w_1x^t+w_0$$

Take derivative of E

$$\sum_t r^t = Nw_0 + w_1\sum_t x^t$$

…wrto w0

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

…wrto w1

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

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

Polynomial Regression

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

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

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

Other Error Measures

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

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

Bias and Variance

Let X be a sample from a population specified up to a parameter θ

To evaluate the quality of this estimator we can measure how much it is different from θ
That is $(d\,(X)-\theta)^2$

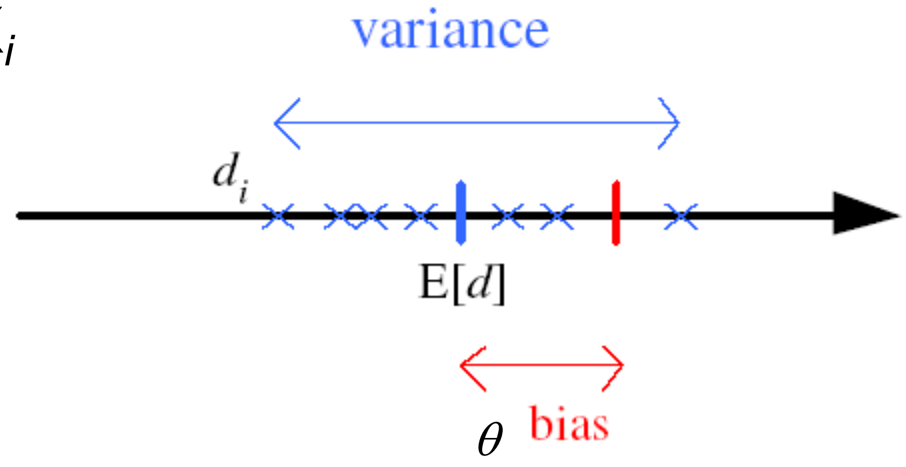
But since it is random variable (it depends on the sample) we need to average over all possible X and consider meas square error of the estimator

Remember the properties of expectation

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:

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

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

Remember the properties of expectation

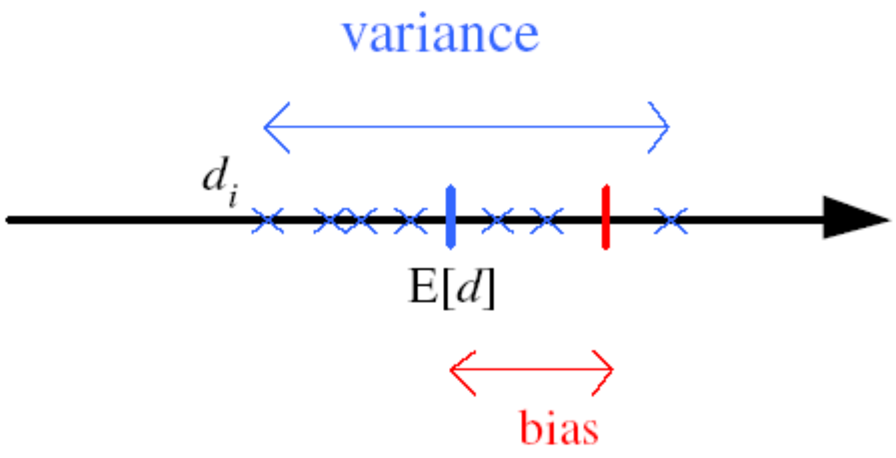
$$= E[(E[d]-\theta)^2]+E\,[(d-E\,[d])^2] +2\,E[(d-E[d])(E[d]-\theta)]$$

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

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

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

$$= \text{Bias}^2 + \text{Variance}$$



Bias and Variance

$$E\left[\left(r-g(x)\right)^2 \mid x\right] = \underbrace{E\left[\left(r-E[r \mid x]\right)^2 \mid x\right]}_{\textit{noise}} + \underbrace{\left(E[r \mid x]-g(x)\right)^2}_{\textit{squared error}}$$

$$E_{\mathcal{X}}\left[\left(E[r \mid x]-g(x)\right)^2 \mid x\right] = \underbrace{\left(E[r \mid x]-E_{\mathcal{X}}[g(x)]\right)^2}_{\textit{bias}} + \underbrace{E_{\mathcal{X}}\left[\left(g(x)-E_{\mathcal{X}}[g(x)]\right)^2\right]}_{\textit{variance}}$$

Estimating Bias and Variance

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

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

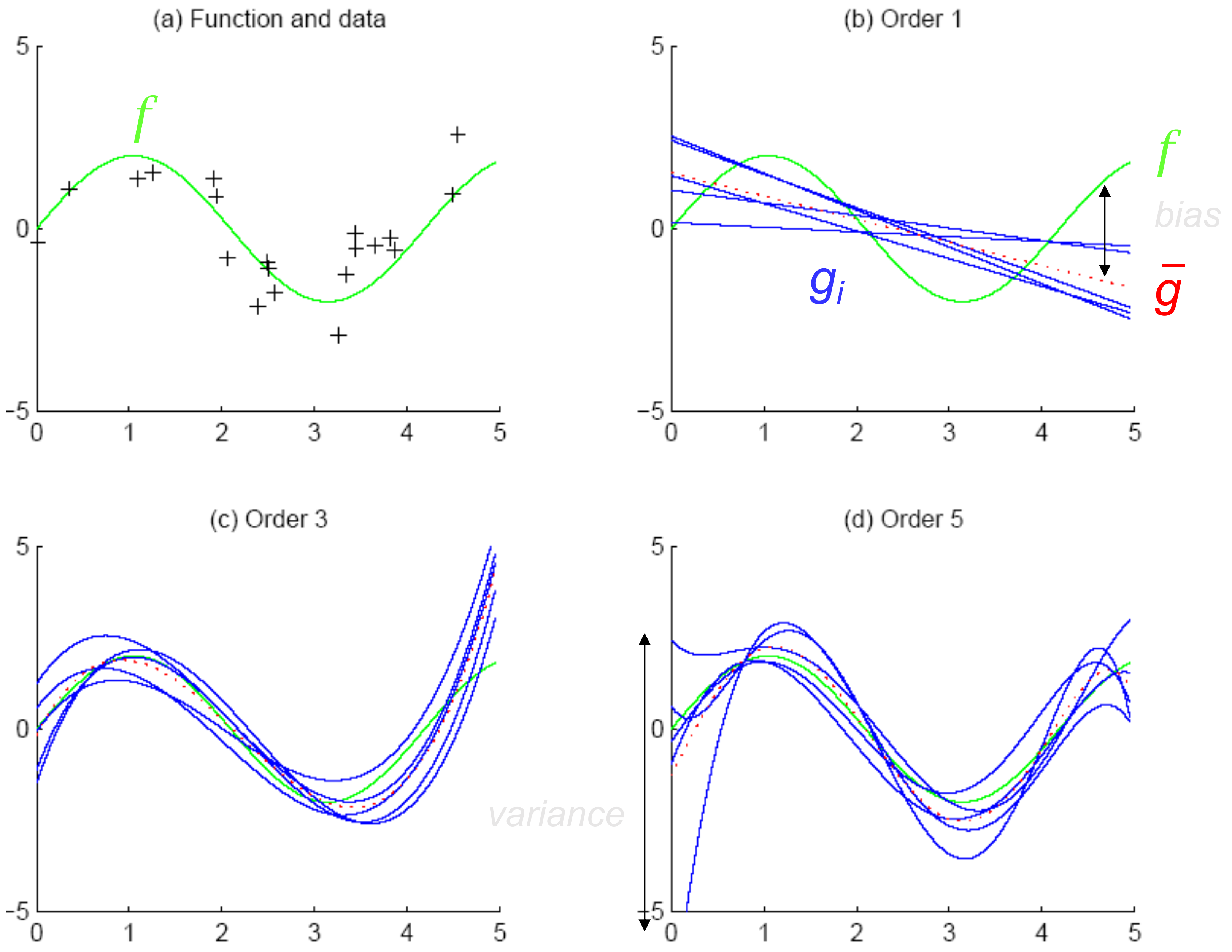
$$\text{Variance}(g) = \frac{1}{NM}\sum_t \sum_i \left[g_i(x^t) - \bar{g}(x^t)\right]^2$$

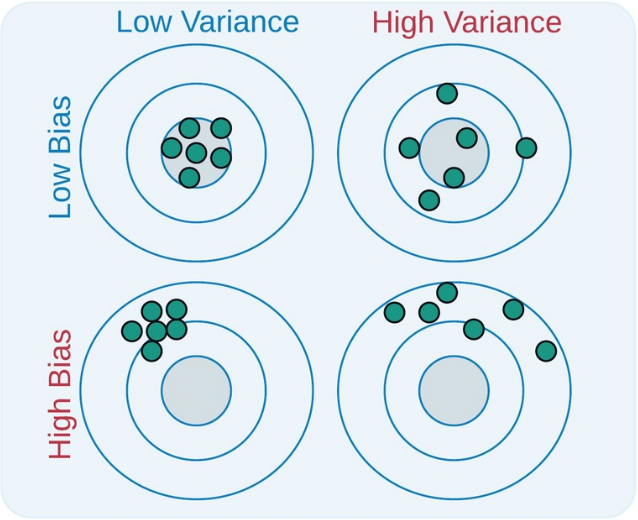
$$\bar{g}(x) = \frac{1}{M}\sum_i g_i(x)$$

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

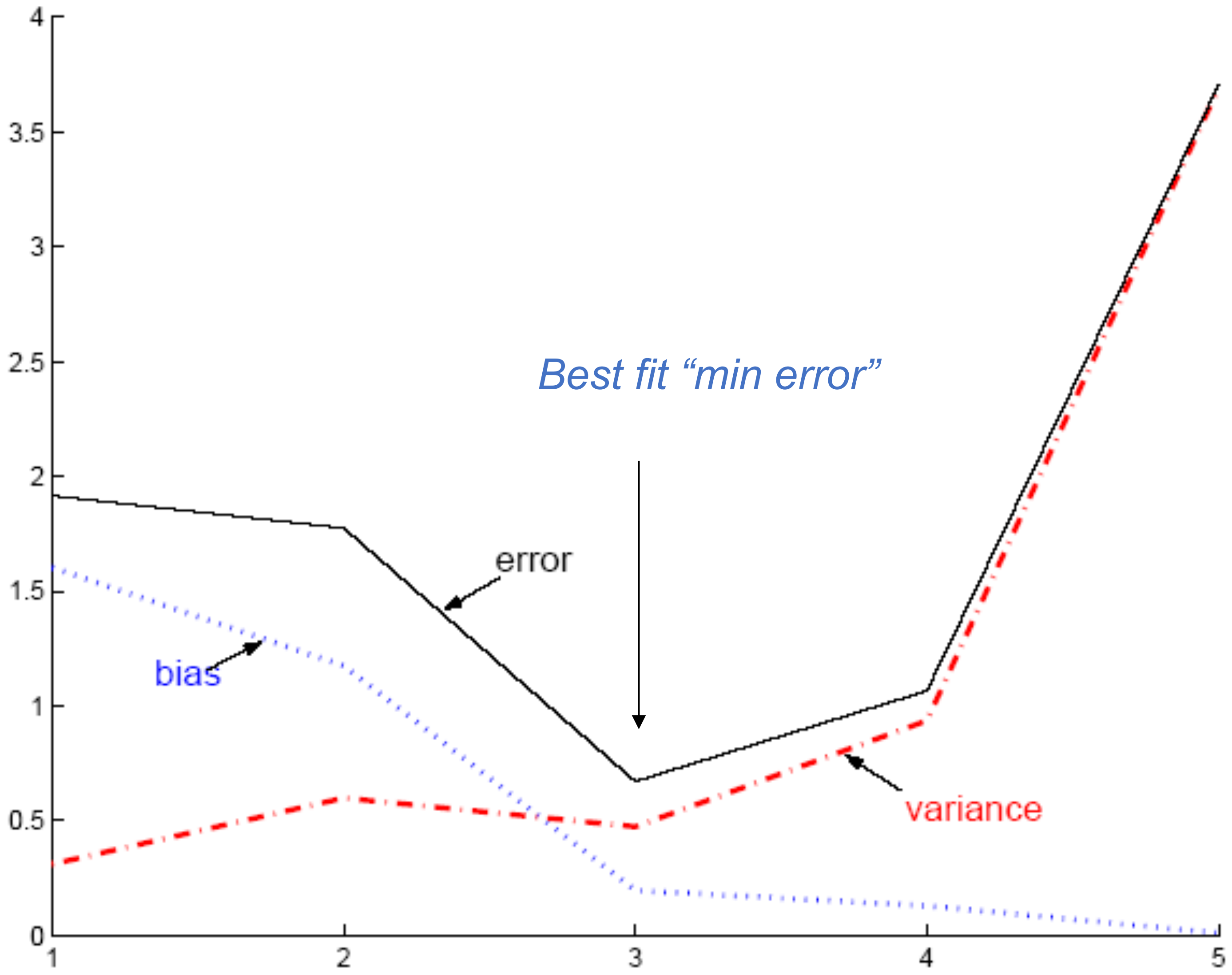
- As we increase complexity,
  bias decreases (a better fit to data) and
  variance increases (fit varies more with data)
- Bias/Variance dilemma: (Geman et al., 1992)

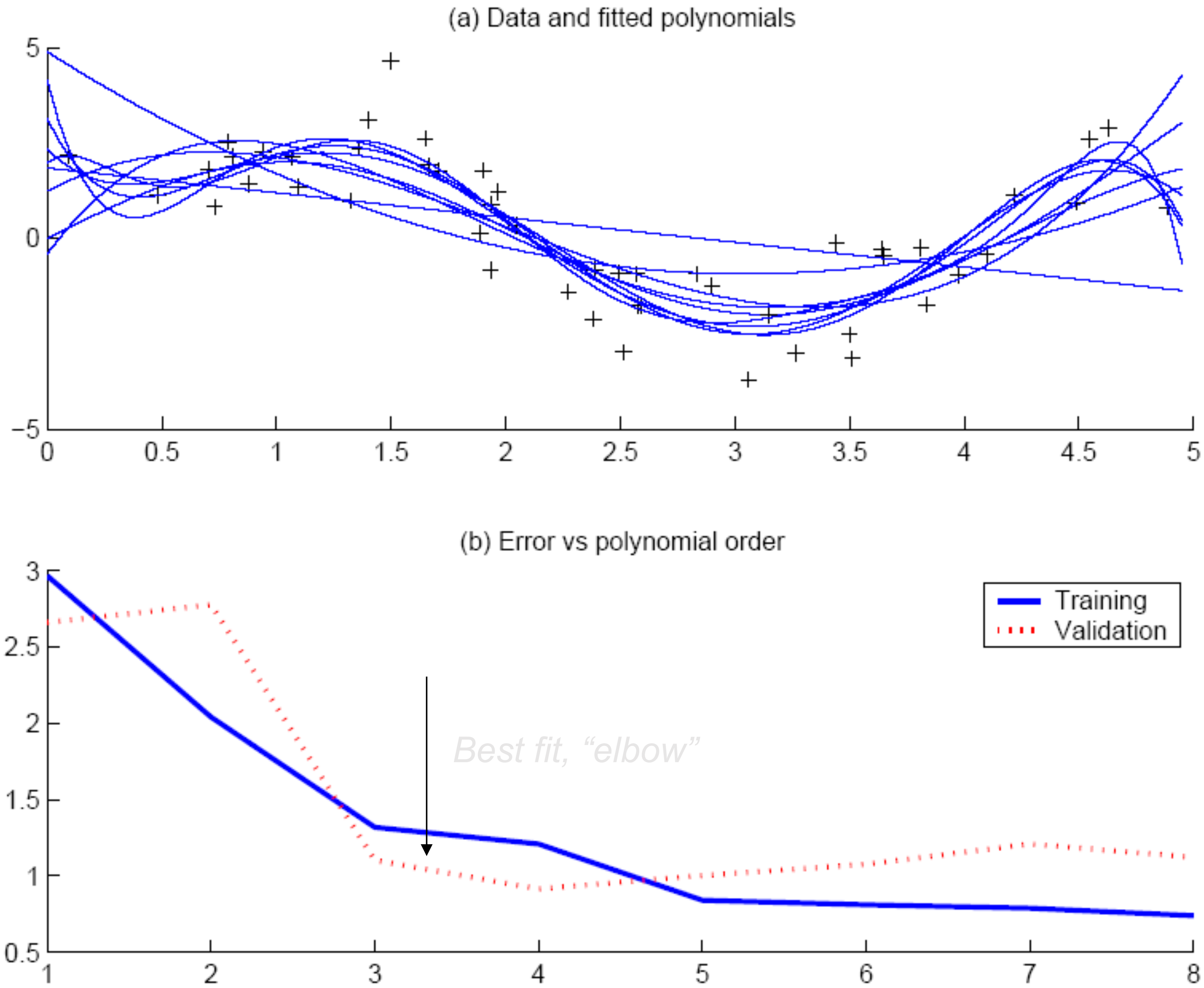




Bias Variance Image is obtained from: https://www.researchgate.net/figure/Visualizing-bias-and-variance-tradeoff-using-a-bulls-eye-diagram_fig3_318432363

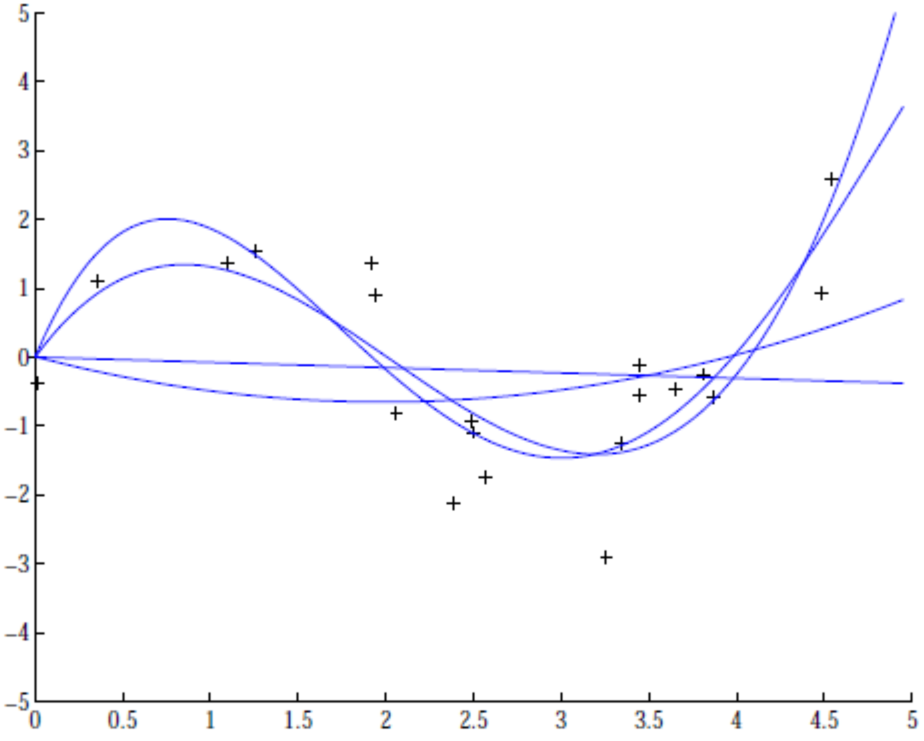
Polynomial Regression





(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
Training
Validation
Best fit, “elbow”
3
2.5
2
1.5
1
0.5
1
2
3
4
5
6
7
8

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]

Idea: Penalize large coefficients

Regularization

- New Cost Function $E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[y^t - g(x^t \mid \mathbf{w})\right]^2 + \lambda\sum_i w_i^2$

- Ridge Regression $R(w) = \|w\|^2 = \sum_i w_i^2$

- LASSO: $R(w) = \|w\|_1 = \sum_i |w_i|$

$$\mathcal{L}(W) = \frac{1}{2}\sum_{i=1}^{N}(y - Xw)^2 + \lambda\sum_i w_i^2 \Rightarrow \frac{1}{2}(y - X\mathrm{w})^T(y - X\mathrm{w}) + \lambda w^T w$$

- $\nabla\, \mathcal{L}$ = - $\frac{2}{2}$ $X^T(y - Xw)$+λw

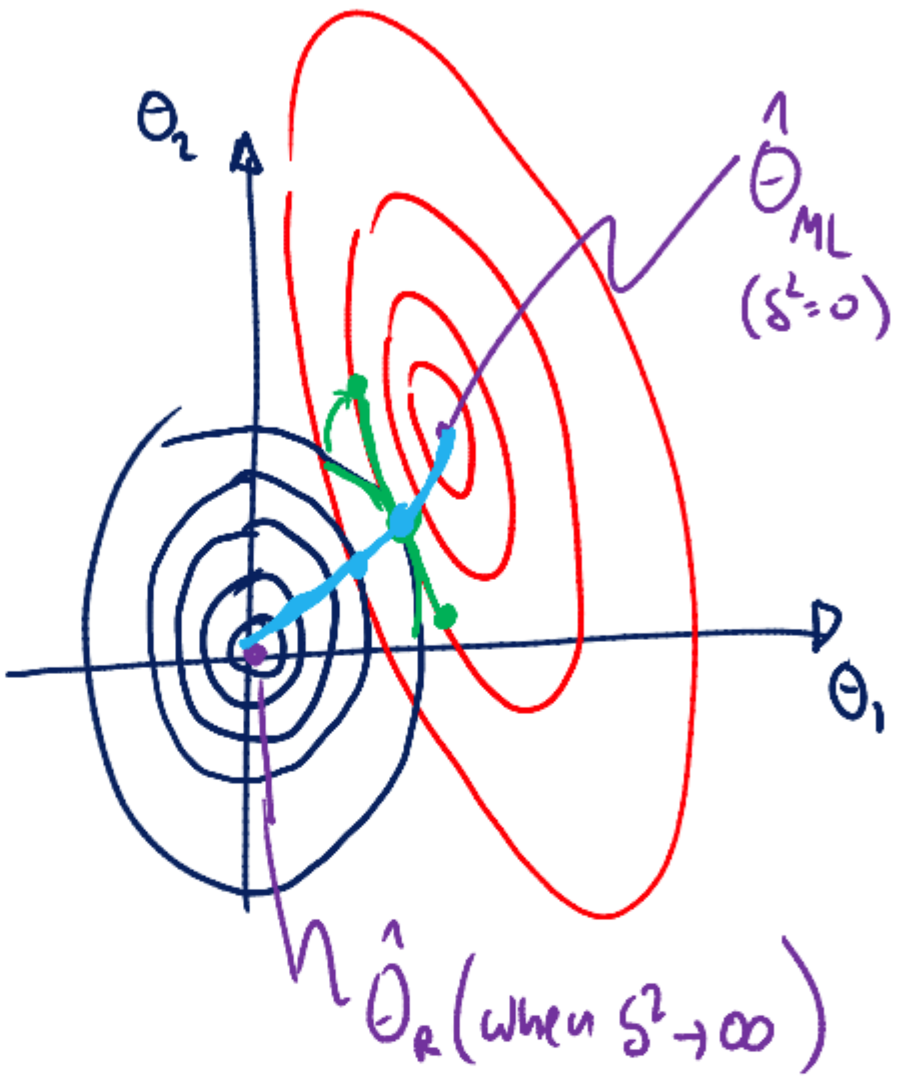
- - $\frac{2}{2}$ $X^T(y - Xw)$ +λw = 0 → $X^T y = X^T X$w +λw →

- $X^T y = (X^T X$ +λI) w

- $\hat{w}$= $(X^T X$+λI$)^{-1} X^T y$

ellipses

$$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \delta^2 \boldsymbol{\theta}^T\boldsymbol{\theta}$$



- Image is obtained from Nando Freitas' lecture notes

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 $\mathcal{X} = \{x^t, r^t\}_{t=1}^{N}$

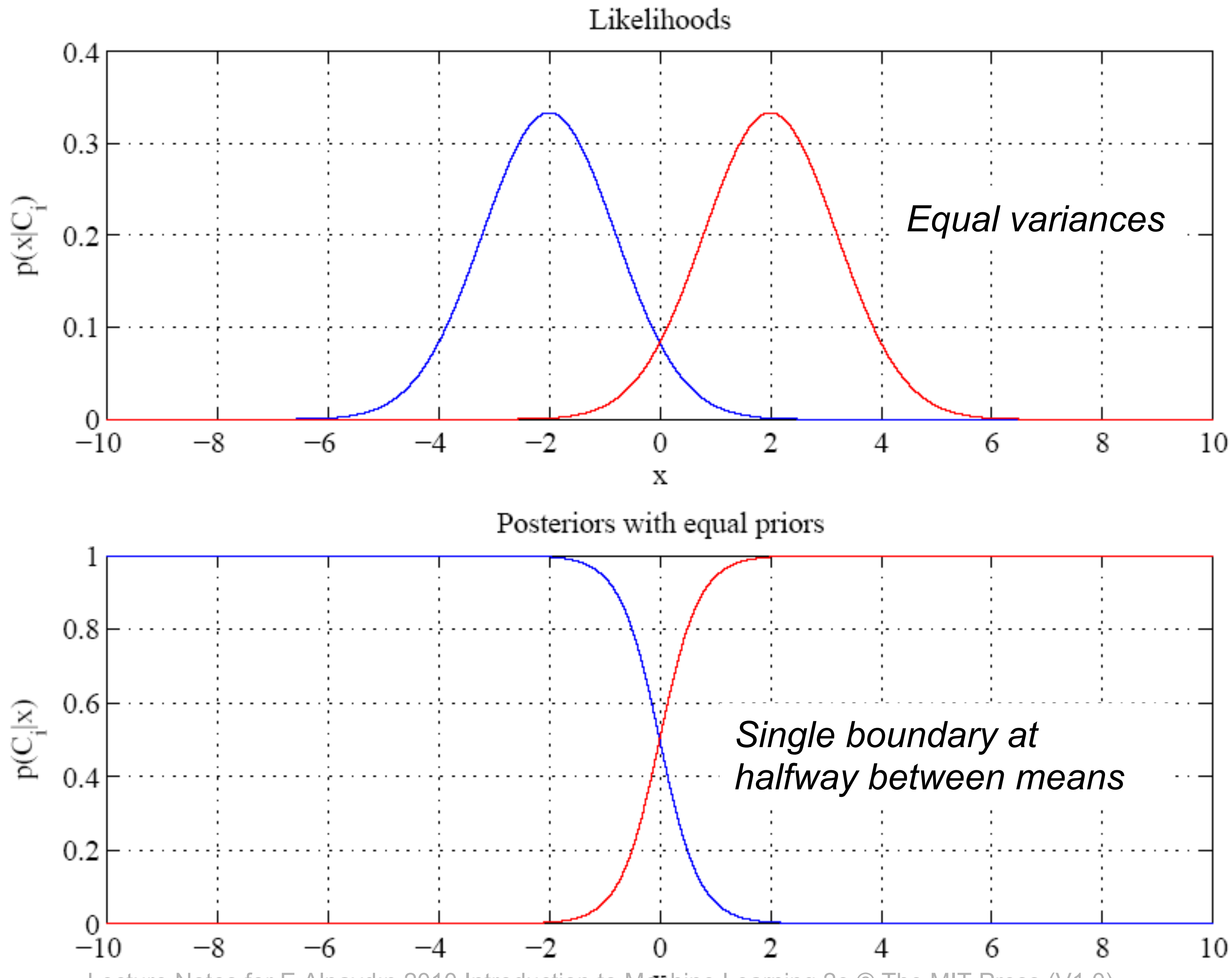
$$x \in \Re \qquad r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j, j \neq i \end{cases}$$

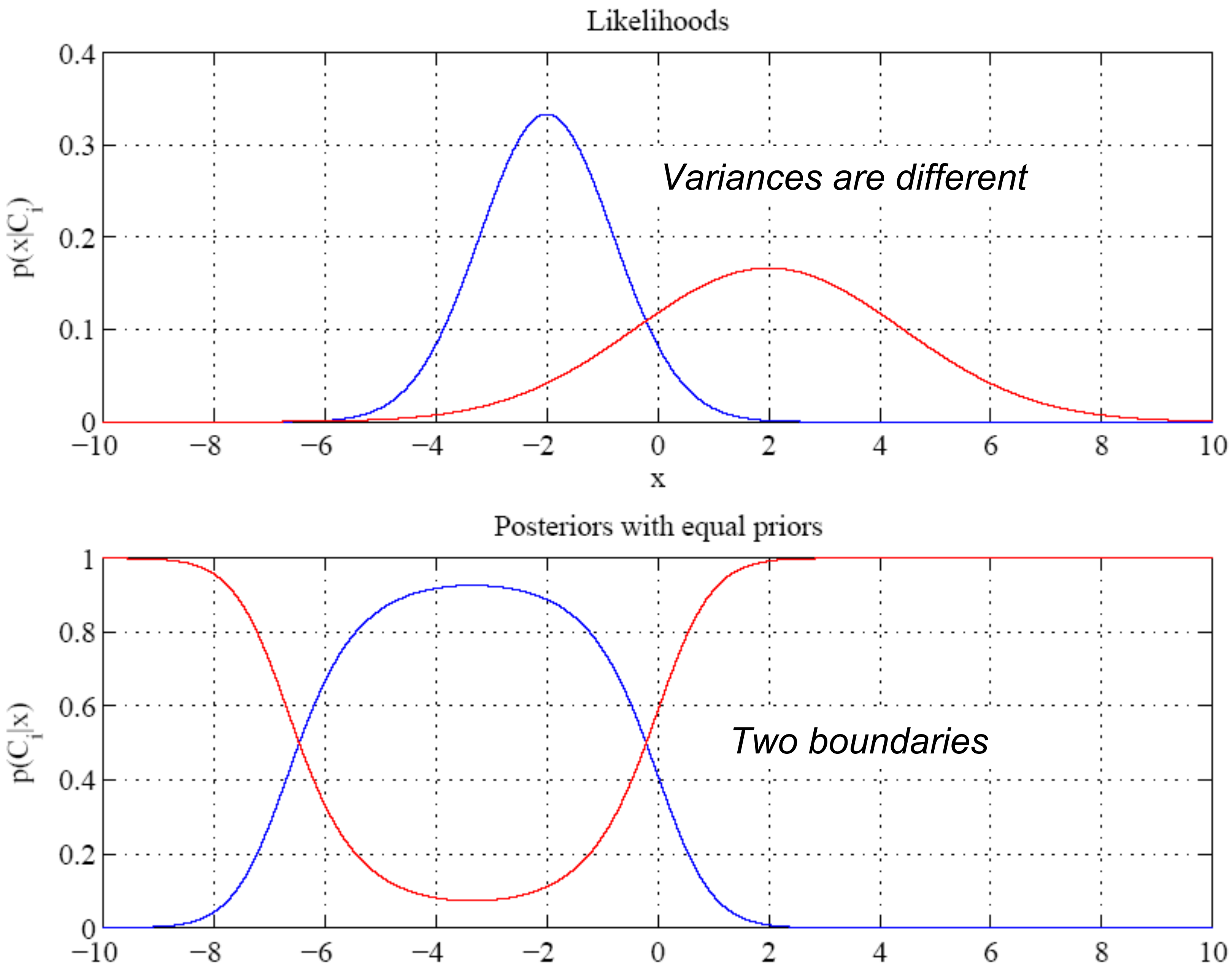
- ML estimates are

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

- Discriminant becomes

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


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


Likelihoods
0.4
0.3
0.2
0.1
0
p(x|C_i)
−10
−8
−6
−4
−2
0
2
4
6
8
10
x
Variances are different
Posteriors with equal priors
1
0.8
0.6
0.4
0.2
0
p(C_i|x)
Two boundaries