BLG 454E Learning From Data

FALL 2022-2023

Assoc. Prof. Yusuf Yaslan

Bias Variance

Linear Regression

$$E(q|\mathcal{X}) = \frac{1}{2}\overset{N}{\underset{t=1}{\text{å}}} \begin{matrix}\text{é}\\ \text{ë}\end{matrix} r^t \text{ - } g\left(x^t|q\right)\begin{matrix}\text{ù}\\ \text{û}\end{matrix}^{2}$$

Take derivative of E

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

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

…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

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

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

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

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: $E\ (\vartheta\ |\mathrm{X}) = \sum_t |r^t - g(x^t|\ \vartheta)|$
- ε-sensitive Error:

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

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

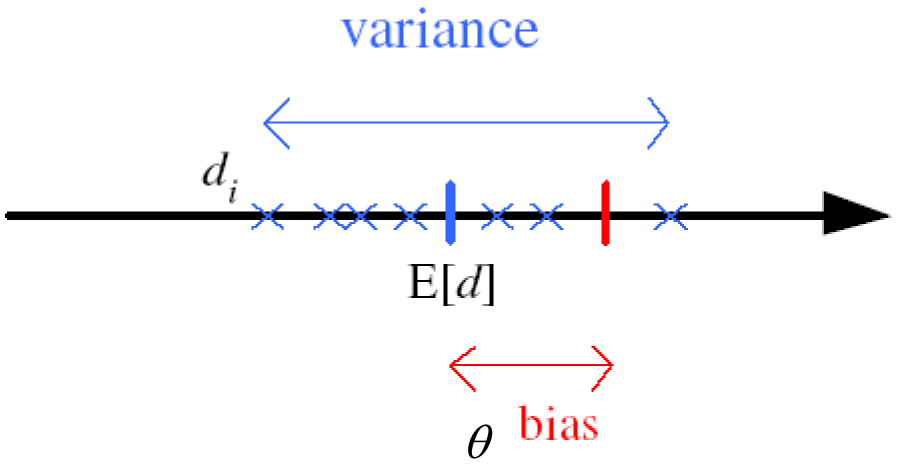
Bias and Variance

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

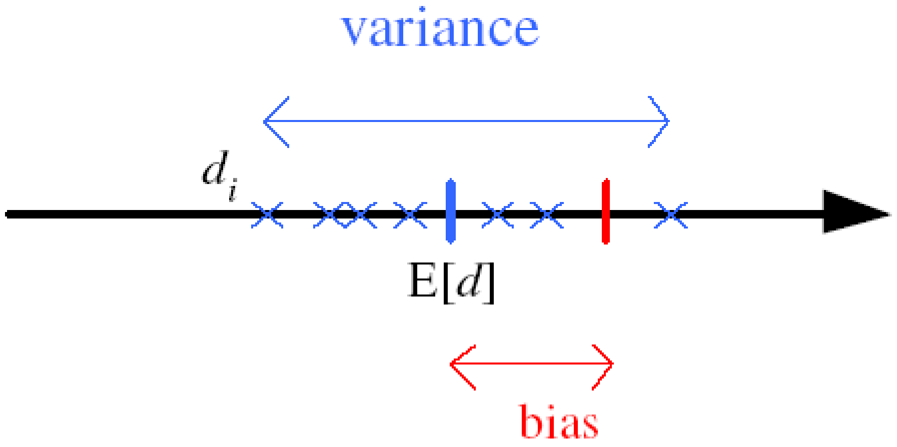
$$= E[(E[d]-\theta)^2]+E\,[(d-E\,[d])^2]\ +2\ E[(d-E[d])(E[d]-\theta)]$$

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

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

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

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



Bias and Variance

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

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

Estimating Bias and Variance

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

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

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

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

Bias/Variance Dilemma

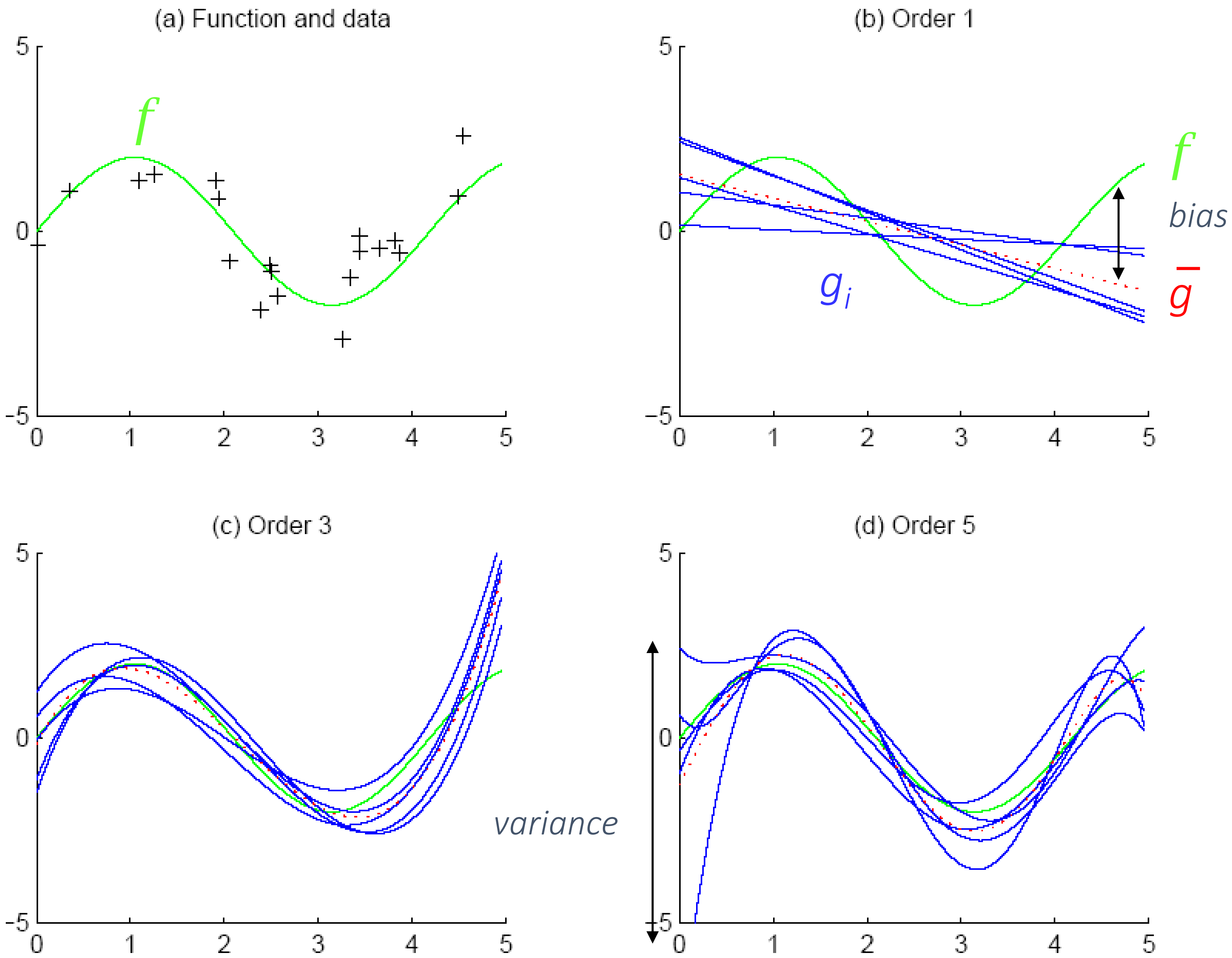
- Example: $g_i(x)=2$ has no variance and high bias

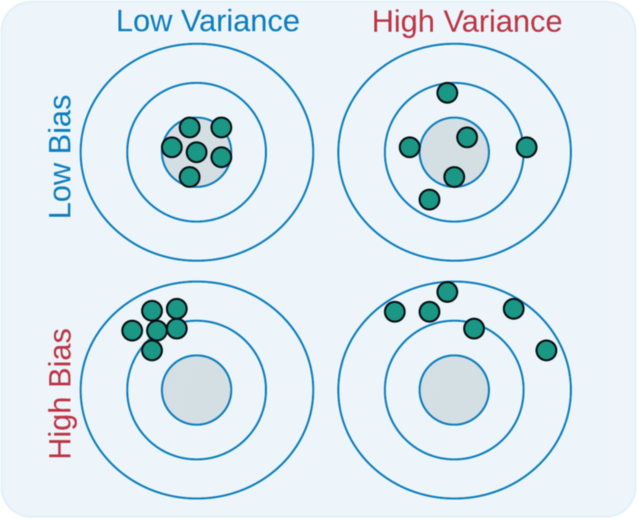
  $g_i(x)= \sum_t r^t_i/N$ has lower bias with variance

- As we increase complexity,

  bias decreases (a better fit to data) and

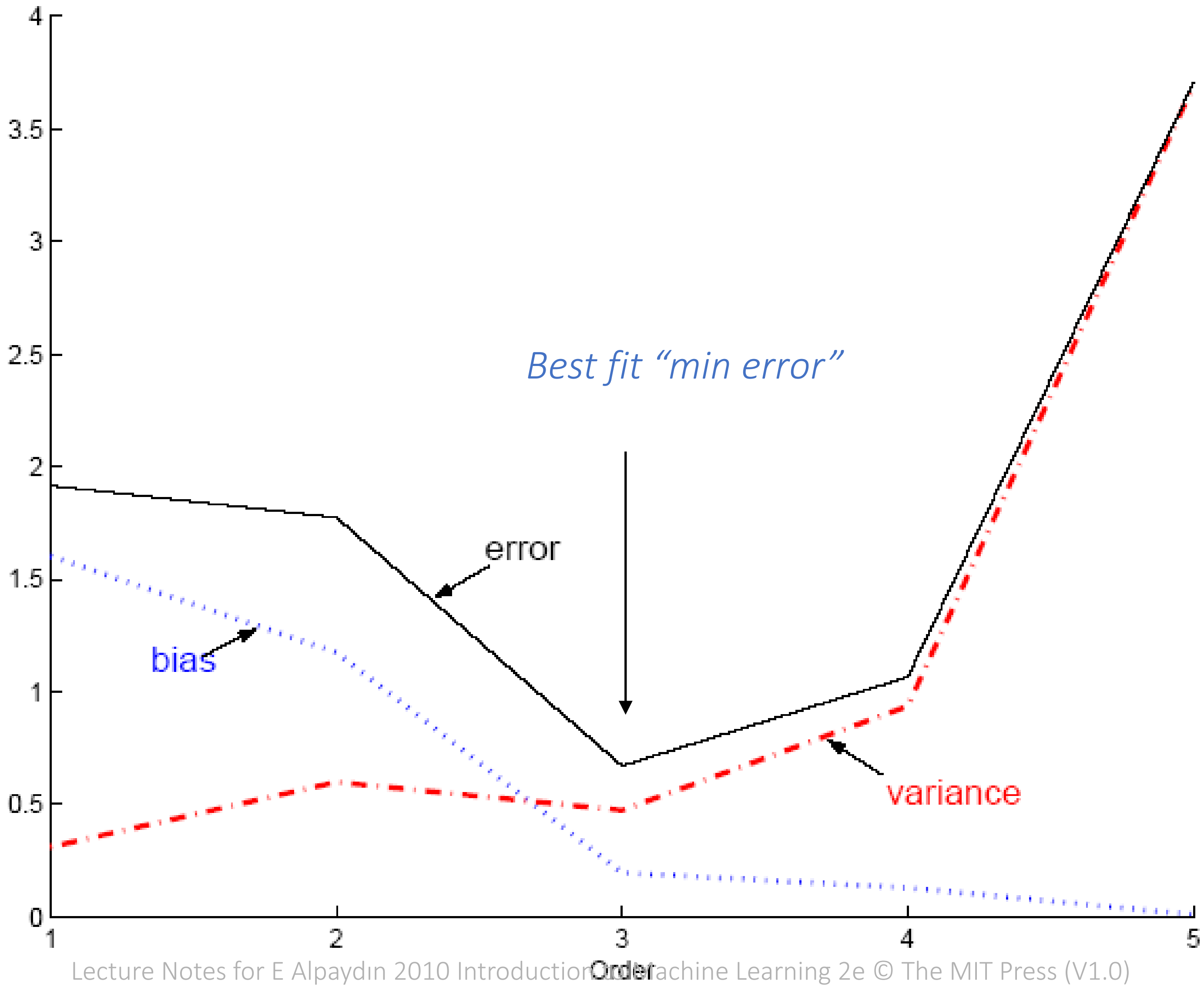
  variance increases (fit varies more with data)
- Bias/Variance dilemma: (Geman et al., 1992)



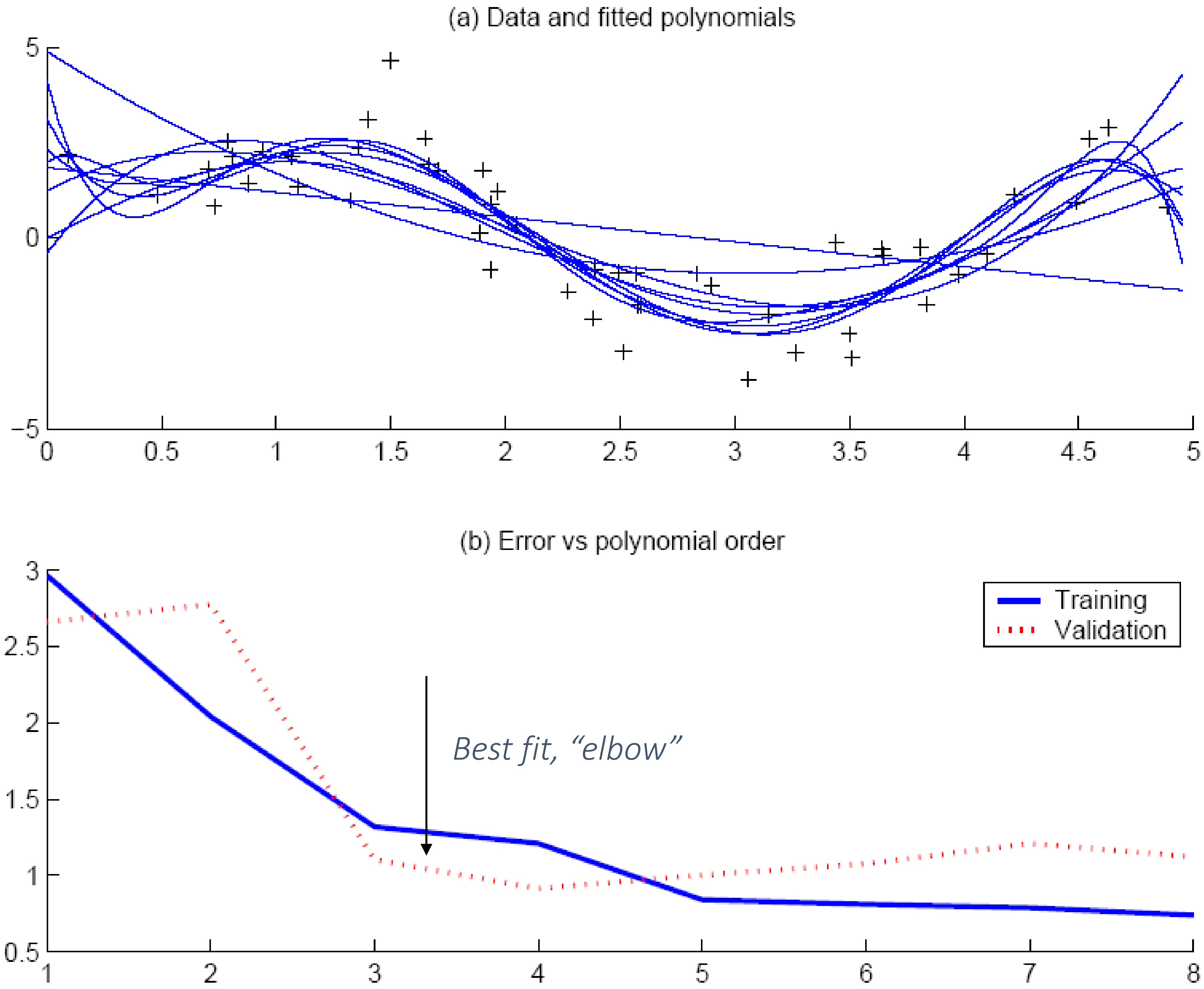


Bias Variance Image is obtained from: https://www.researchgate.net/figure/Visualizing-bias-and-variance-tradeoff-using-a-bulls-eye-diagram_fig3_318432363

Polynomial Regression

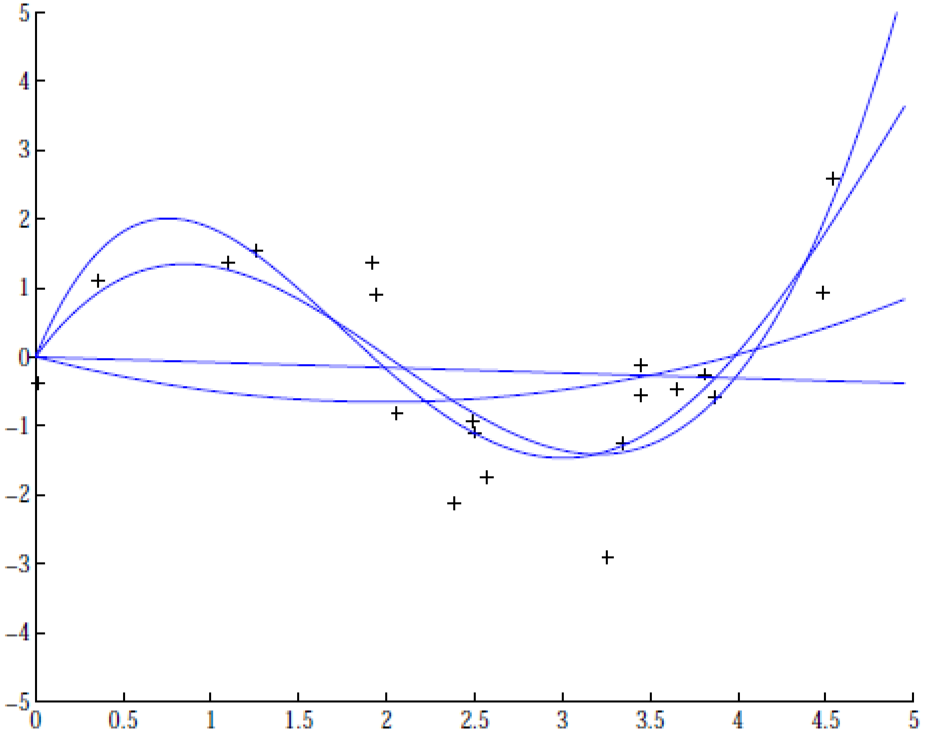


Best fit "min error"



Best fit, “elbow”

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 $E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[y^t - g\left(x^t \mid \mathbf{w}\right)\right]^2 + \lambda \sum_i w_i^2$
- Ridge Regression $R(w) = \|w\|^2 = \sum_i w_i^2$
- LASSO: $R(w) = \|w\|_1 = \sum_i \left|w_i\right|$

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

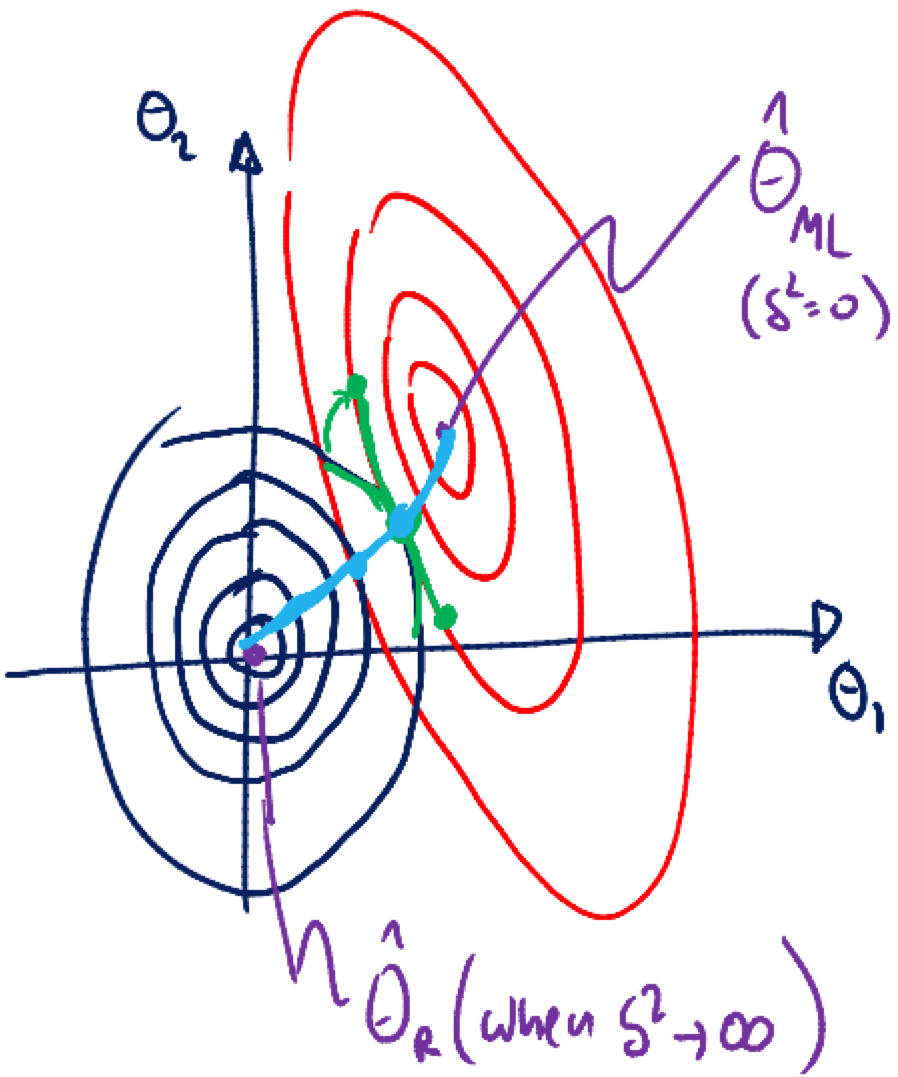
- $\nabla\ \mathcal{L}$ = - $\frac{2}{2}\ X^T(y - Xw)$+λw

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

- $X^T y = (X^T X$ +λI) w

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

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



- Image is obtained from Nando Freitas’ lecture notes