Machine Learning

Model Performance: Key Concepts in Machine Learning

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir

https://www.aparat.com/mehran.safayani

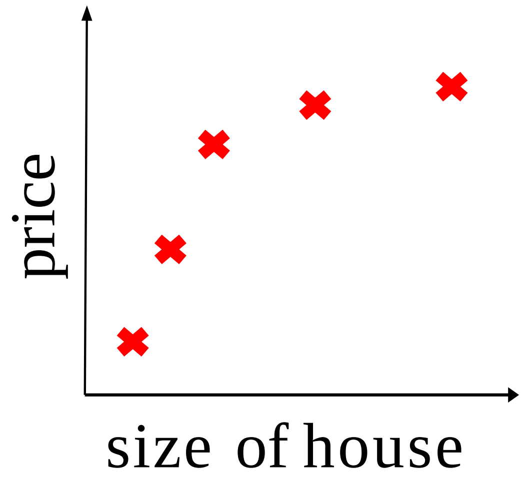
https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

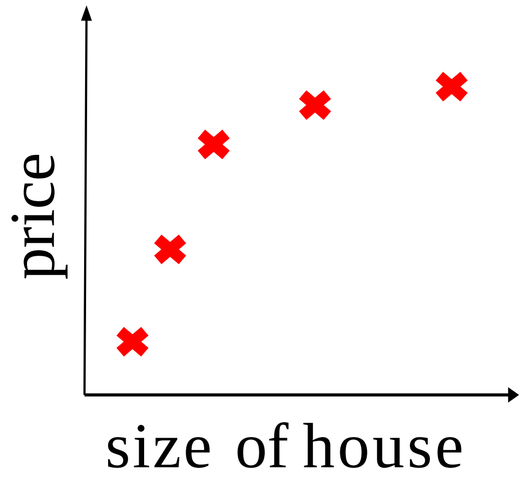
Contents

- Underfitting, Overfitting
- Model Selection
- Train-validation test split
- Regularization: ridge and lasso regression
- Bias-Variance Tradeoff
- K-fold Cross validation

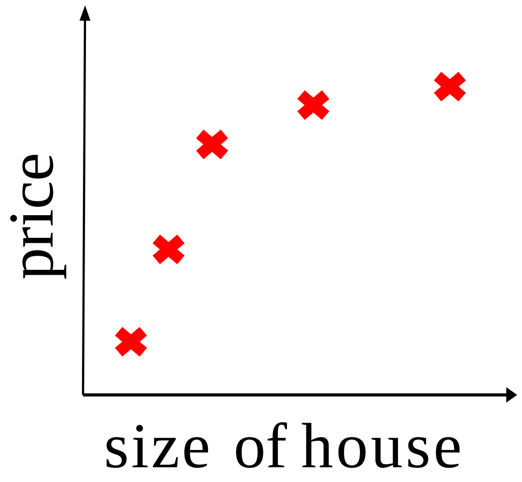
Example: Linear regression (housing prices)



$\theta_0 + \theta_1 x$



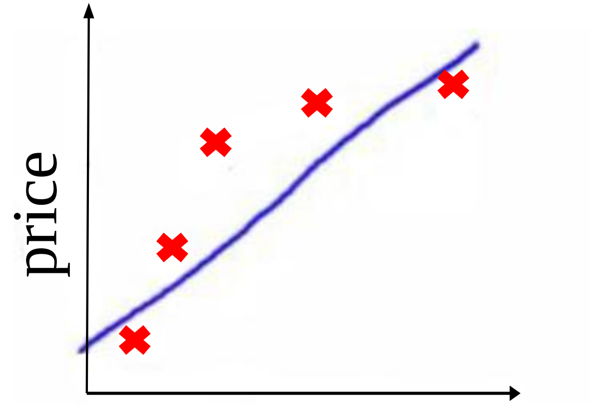
$\theta_0 + \theta_1 x + \theta_2 x^2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

The slides are modified, based on original slides by [Andrew NG, Stanford university

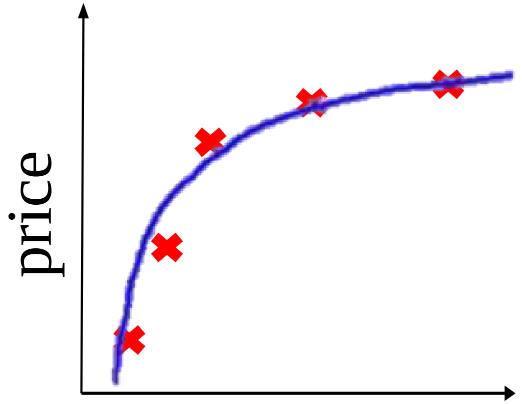
Example: Linear regression (housing prices)



size of house

$\theta_0 + \theta_1 x$

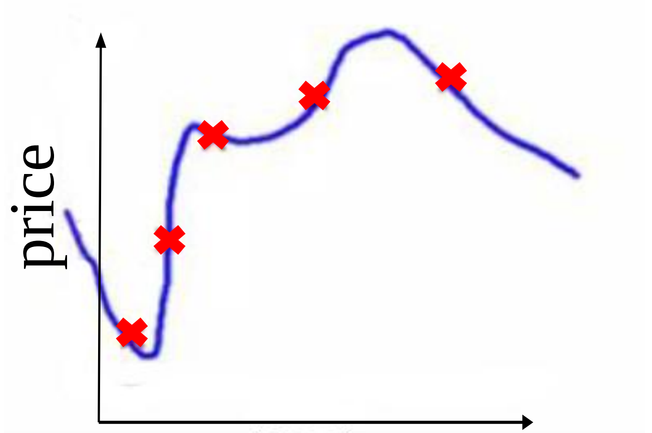
"Underfit" **"High bias"**



size of house

$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"



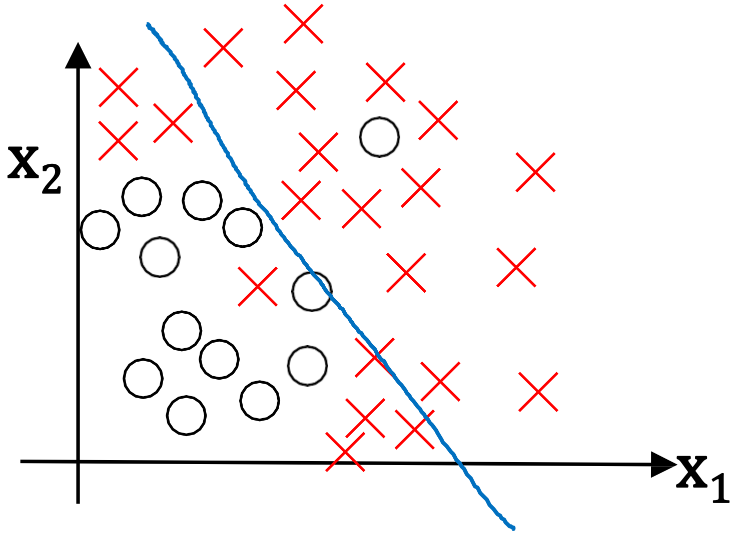
size of house

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

"overfit" **"High variance"**

Overfitting: If we have too many features, the learned hypothesis may fit the training set very well ($J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \approx 0$), but fail to generalize to new examples (predict prices on new examples).

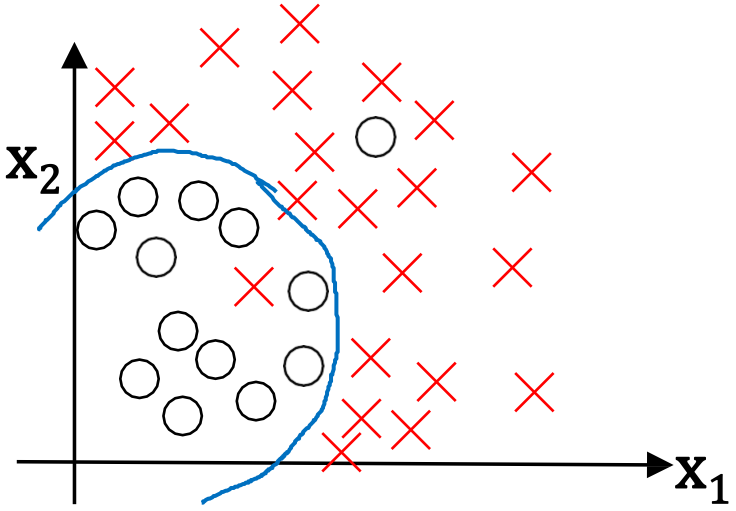
Example: Logistic regression



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

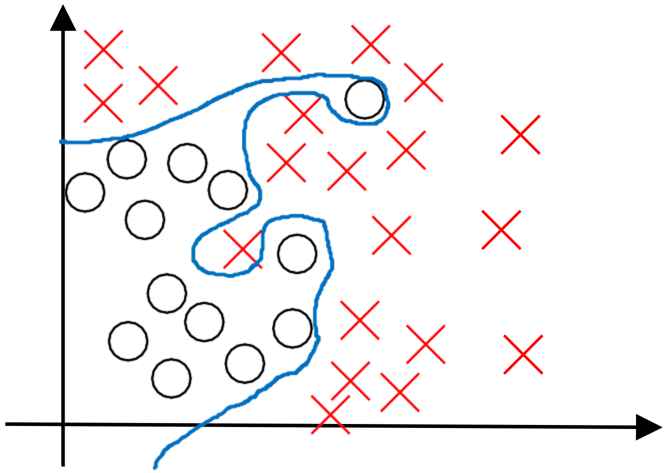
high bias

underfitting



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$

"just right"



$$g\begin{pmatrix} \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 \\ +\theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \cdots \end{pmatrix}$$

high variance

Overfitting

Addressing overfitting:

x_1 = size of house

x_2 = no. of bedrooms

x_3 = no. of floors

x_4 = age of house

x_5 = average income in neighborhood

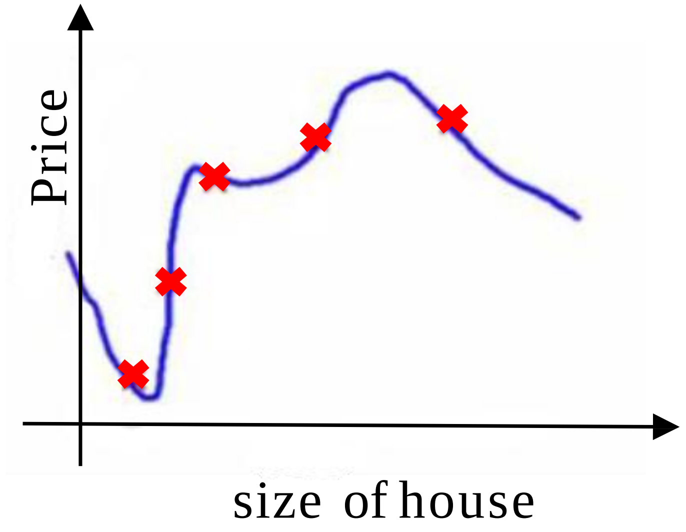
x_6 = kitchen size

.

.

.

x_{100}



Evaluating your hypothesis

- Dataset:

Size	Price
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243

60% Training set

20% Cross validation set (cv)

20% Test set

$$\left(x^{(1)}, y^{(1)}\right)$$
$$\left(x^{(2)}, y^{(2)}\right)$$
$$\vdots$$
$$\left(x^{(m)}, y^{(m)}\right)$$

$$\left(x_{cv}{}^{(1)}, y_{cv}{}^{(1)}\right)$$
$$\left(x_{cv}{}^{(2)}, y_{cv}{}^{(2)}\right)$$
$$\vdots$$
$$\left(x_{cv}{}^{(m_{cv})}, y_{cv}{}^{(m_{cv})}\right)$$

$$\left(x_{test}{}^{(1)}, y_{test}{}^{(1)}\right)$$
$$\left(x_{test}{}^{(2)}, y_{test}{}^{(2)}\right)$$
$$\vdots$$
$$\left(x_{test}{}^{(m_{test})}, y_{test}{}^{(m_{test})}\right)$$

Train/validation/test error

- Training error:

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

- Cross Validation error:

$$J_{cv}(\theta) = \frac{1}{2mcv}\sum_{i=1}^{m_{cv}}\left(h_\theta\left(x_{cv}{}^{(i)}\right) - y_{cv}{}^{(i)}\right)^2$$

- Test error:

$$J_{test}(\theta) = \frac{1}{2mtest}\sum_{i=1}^{m_{test}}\left(h_\theta\left(x_{test}{}^{(i)}\right) - y_{test}{}^{(i)}\right)^2$$

Model Selection

$h_{\theta_1}(x) = \theta_0 + \theta_1 \mathrm{x}$ $\qquad J_{cv}(\theta^1)$

$h_{\theta_2}(x) = \theta_0 + \theta_1 \mathrm{x} + \theta_2\, x^2$ $\qquad J_{cv}(\theta^2)$

.

.

.

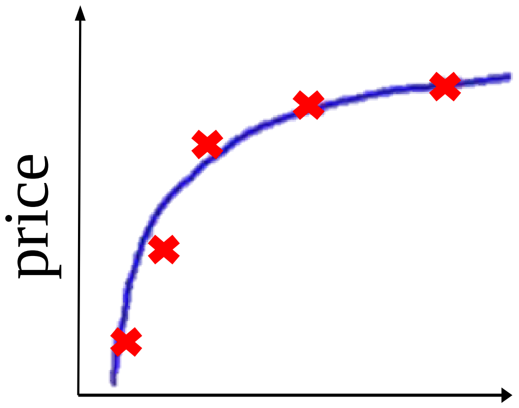
$h_{\theta_{10}}(x) = \theta_0 + \theta_1 \mathrm{x} + \ldots + \theta_{10}\, x^{10}$ $\qquad J_{cv}(\theta^{10})$

$$i^* = \underset{\mathrm{i}}{\operatorname{argmin}}\ J_{cv}(\theta^{\mathrm{i}})$$

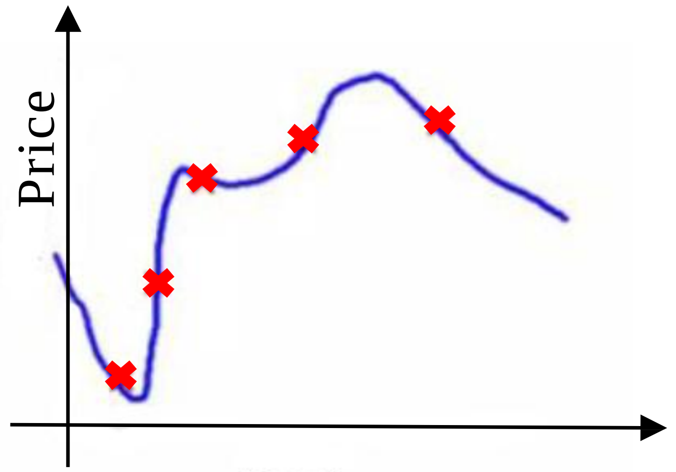
$$\Downarrow$$

$$J_{test}(\theta^{i^*})$$

Regularization Intuition



$\theta_0 + \theta_1 x + \theta_2 x^2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

- Suppose we penalize and make θ_3, θ_4 really small.

$$\min_\theta \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \underbrace{1000\theta_3^2 + 1000\theta_4^2}_{\theta_3 \approx 0 \quad \theta_4 \approx 0}$$

Regularization

- Small values for parameters $\theta_0, \theta_1, \ldots ,\theta_n$
  - "Simpler" hypothesis
  - Less prone to overfitting
- Housing:
  - Features: x_1, x_2, … , x_{100}
  - Parameters: $\theta_0, \theta_1, \theta_2, \ldots, \theta_{100}$

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$$

$$\boldsymbol{\theta_0, \theta_1, \theta_2, \ldots, \theta_{100}}$$

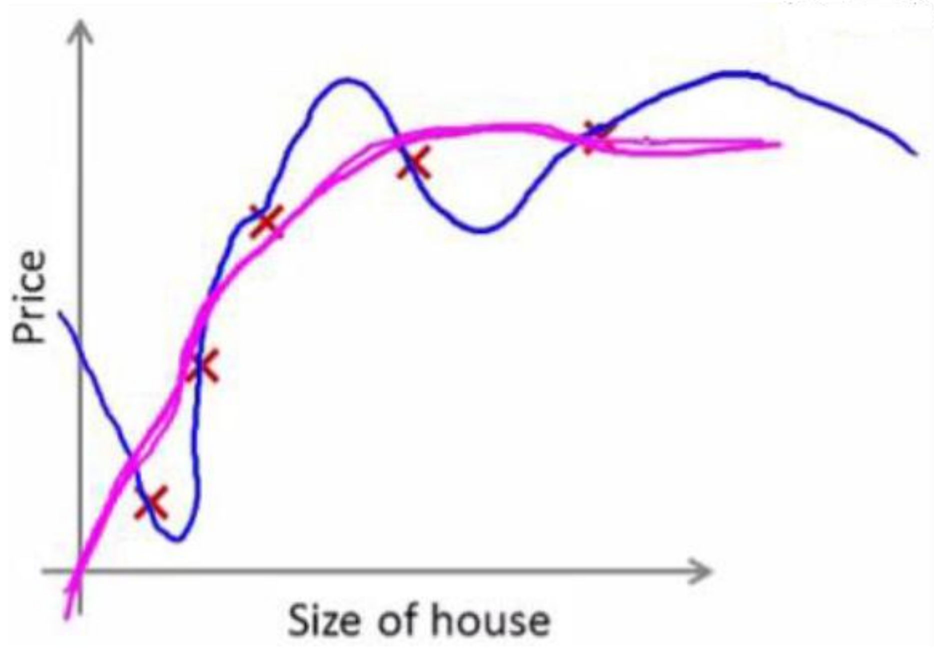
λ = regularization parameter

lambda lambd

Regularization

- $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$

$\min_{\theta} J(\theta)$

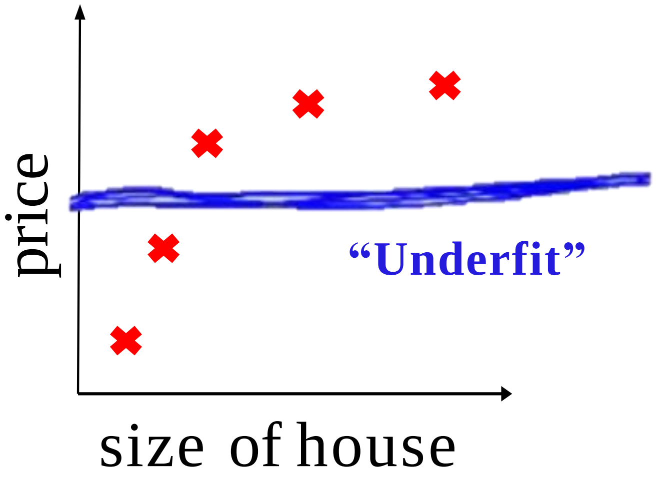


Regularization: Ridge Regression

- In regularized linear regression, we choose θ to minimize

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

- What if λ is set to an extremely large value (perhaps for too large for our problem, say / = 1010)?



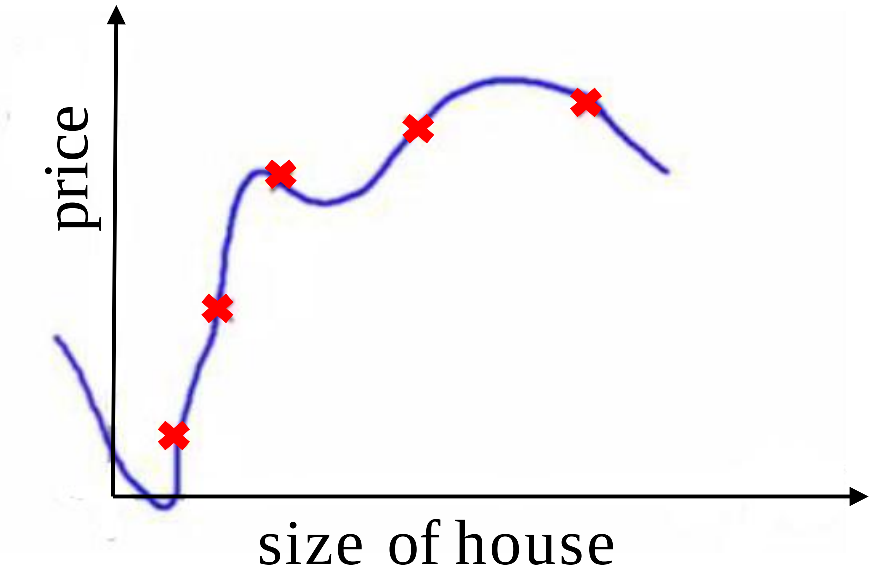
$$\theta_1, \theta_2, \theta_3, \theta_4$$

$$\theta_1 \approx 0, \theta_2 \approx 0$$
$$\theta_3 \approx 0, \theta_4 \approx 0$$
$$\boxed{h_\theta(x) = \theta_0}$$

$h_\theta(x)$

$$\theta_0 + \cancel{\theta_1} x + \cancel{\theta_2} x^2 + \cancel{\theta_3} x^3 + \cancel{\theta_4} x^4$$

Evaluating your hypothesis



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

- Fails to generalize to new examples not in training set.

  x_1 = size of house

  x_2 = no. of bedrooms

  x_3 = no. of floors

  x_4 = age of house

  x_5 = average income in neighborhood

  x_6 = kitchen size

  .

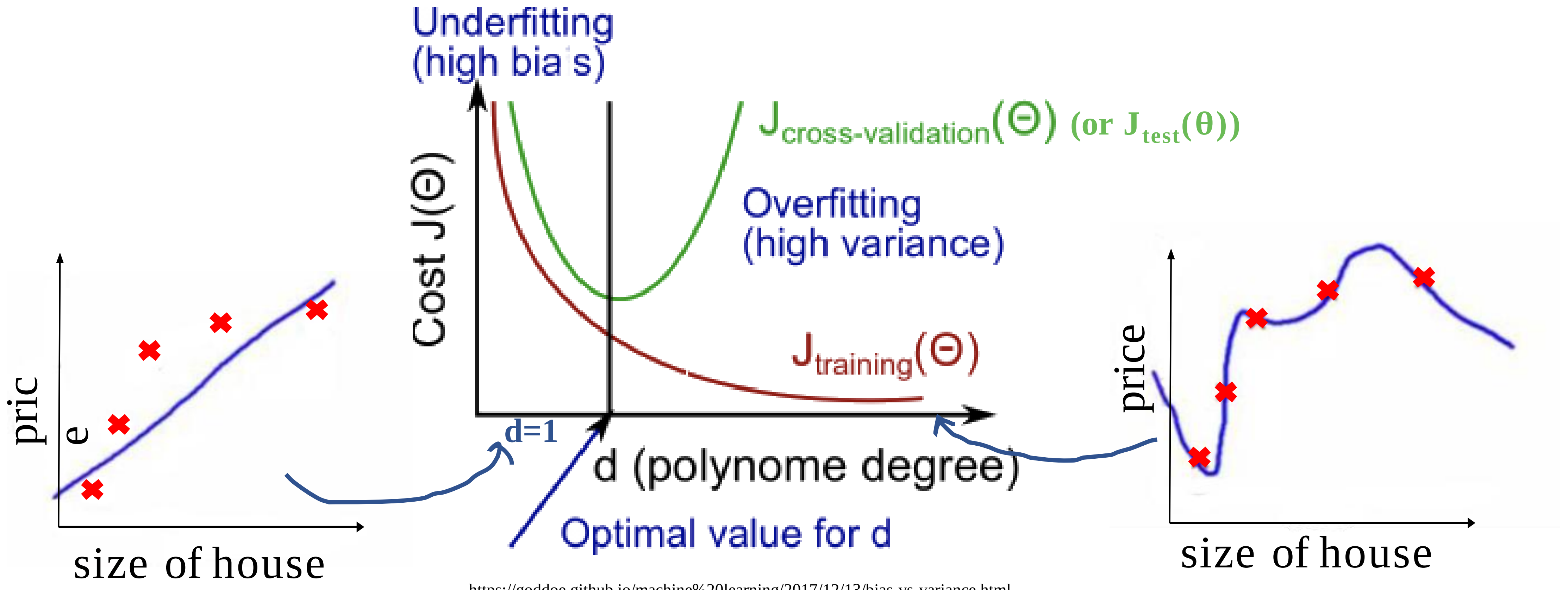
  .

  .

  x_{100}

Bias/variance

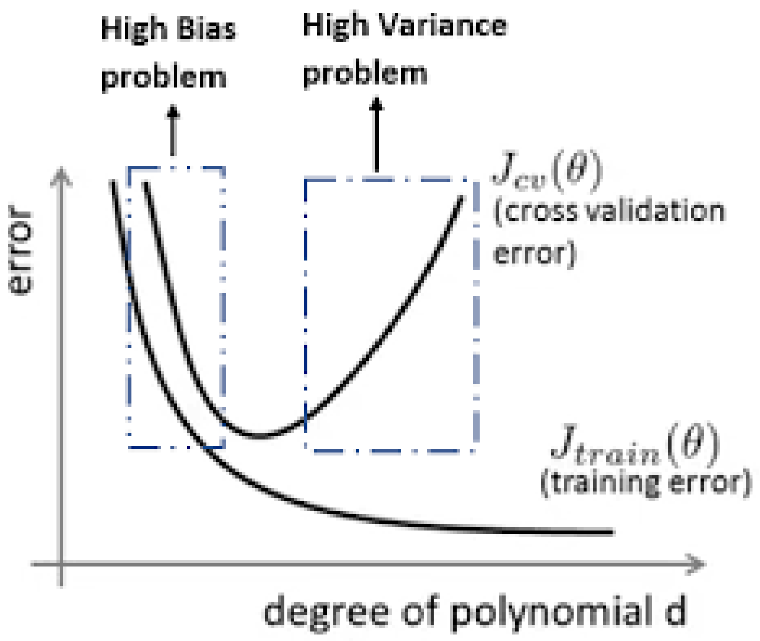
- Training error: $J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$
- Cross validation error: $J_{cv}(\theta) = \frac{1}{2mcv}\sum_{i=1}^{m_{cv}}\left(h_\theta\left(x_{cv}{}^{(i)}\right) - y_{cv}{}^{(i)}\right)^2$ **(or $J_{test}(\theta)$)**



https://goddoe.github.io/machine%20learning/2017/12/13/bias-vs-variance.html

Diagnosing bias vs. variance

- Suppose your learning algorithm is performing less well than you were hoping. ($J_{\text{cv}}(\theta)$ or $J_{\text{test}}(\theta)$) is high.) Is it a bias problem or a variance problem?



https://towardsdatascience.com/

Bias (underfit):

$J_{train}(\theta)$ *will be high*

$J_{cv}(\theta) \approx J_{train}(\theta)$

Variance (overfit):

$J_{train}(\theta)$ will be low

$J_{cv}(\theta) \gg J_{train}(\theta)$

Linear regression with regularization

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

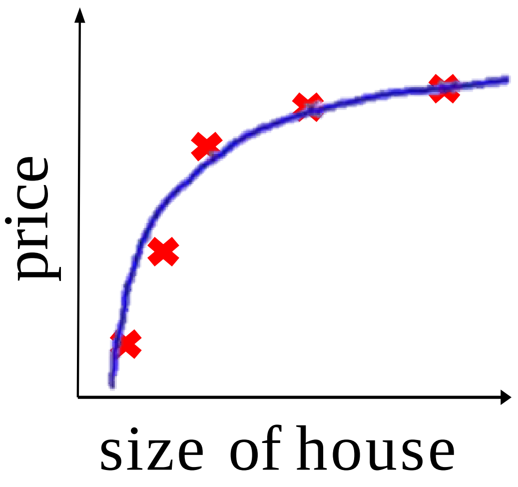
$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

Large λ

High bias(underfit)

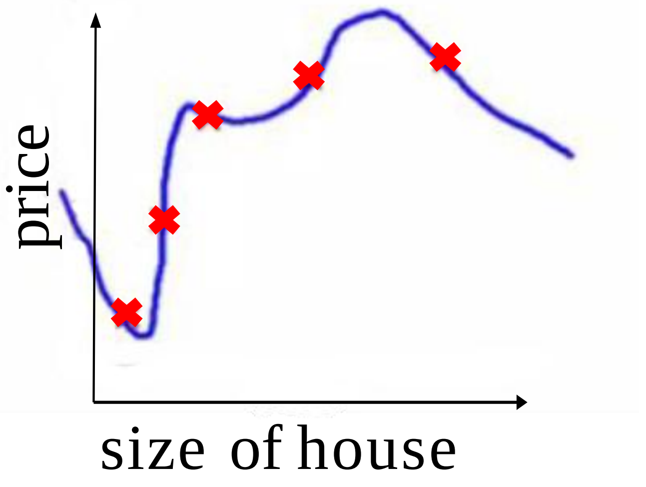
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

$h_\theta(x) \approx \theta_0$



Intermediate λ

"Just right"



Small λ

High variance (overfit)

$\lambda = 0$

Choosing the regularization parameter λ

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

1. Try λ = 0 $\rightarrow \min J(\theta) \rightarrow \theta^{(1)} \rightarrow J_{cv}(\theta^{(1)})$
2. Try λ = 0.01 $\rightarrow \min J(\theta) \rightarrow \theta^{(2)} \rightarrow J_{cv}(\theta^{(2)})$
3. Try λ = 0.02 $\rightarrow \min J(\theta) \rightarrow \theta^{(3)} \rightarrow J_{cv}(\theta^{(3)})$
4. Try λ = 0.04
5. Try λ = 0.08 $\rightarrow \min J(\theta) \rightarrow \theta^{(5)} \rightarrow J_{cv}(\theta^{(5)})$

.

.

.

12. Try λ = 10 $\rightarrow \min J(\theta) \rightarrow \theta^{(12)} \rightarrow J_{cv}(\theta^{(12)})$

10.24

Pick (say) $\theta^{(5)}$. **Test error:** $J_{test}(\theta^{(5)})$

Bias/variance as a function of the regularization parameter λ

- **Training error:**

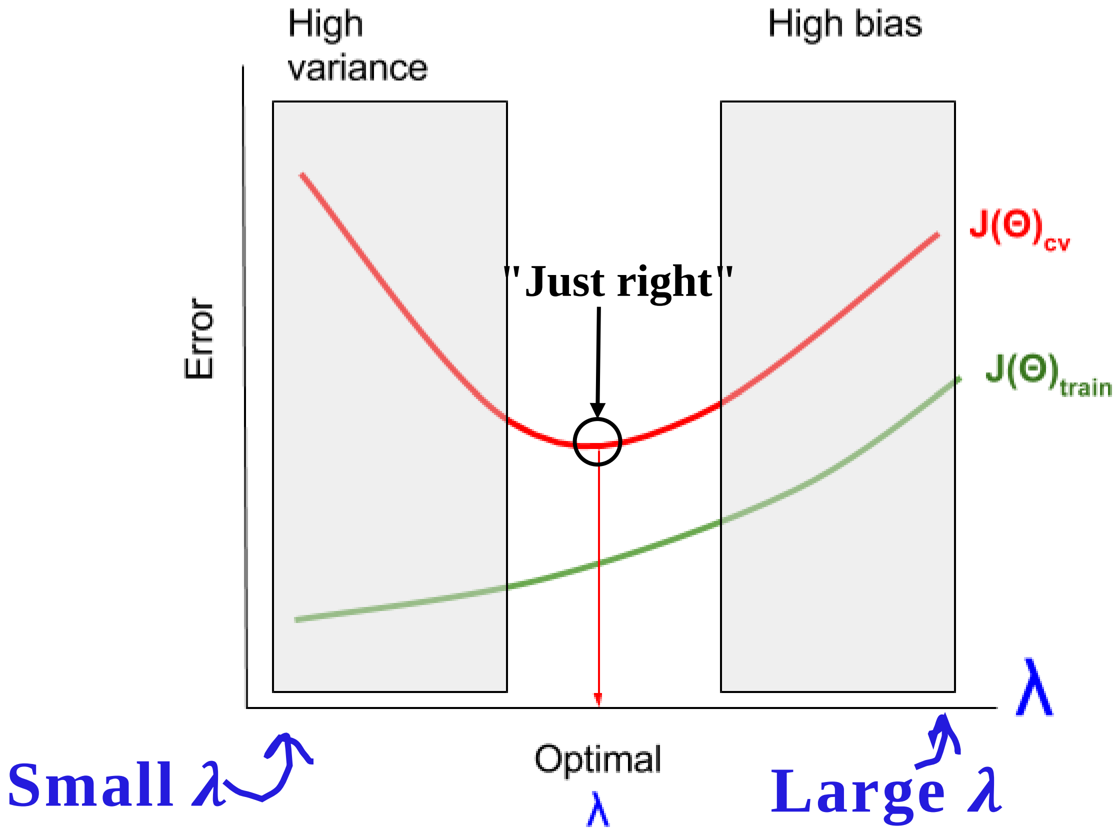
$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2$$

- **Cross Validation error:**

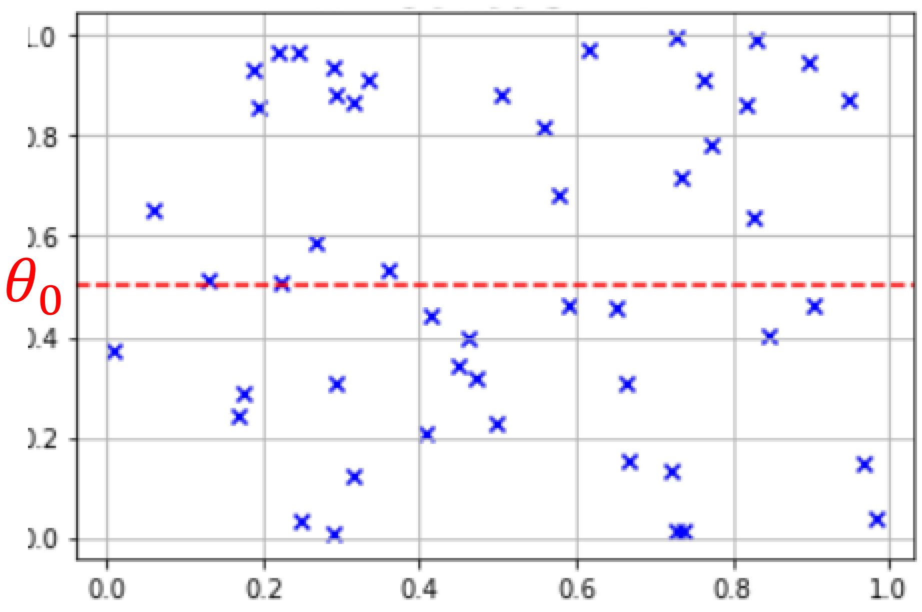
$$J_{cv}(\theta) = \frac{1}{2mcv}\sum_{i=1}^{m_{cv}}\left(h_\theta\left(x_{cv}{}^{(i)}\right) - y_{cv}{}^{(i)}\right)^2$$

- **Test error:**

$$J_{test}(\theta) = \frac{1}{2mtest}\sum_{i=1}^{m_{test}}\left(h_\theta\left(x_{test}{}^{(i)}\right) - y_{test}{}^{(i)}\right)^2$$



https://stats.stackexchange.com/questions/222493/why-do-we-use-the-unregularized-cost-to-plot-a-learning-curve



GD:

Repeat until convergence{

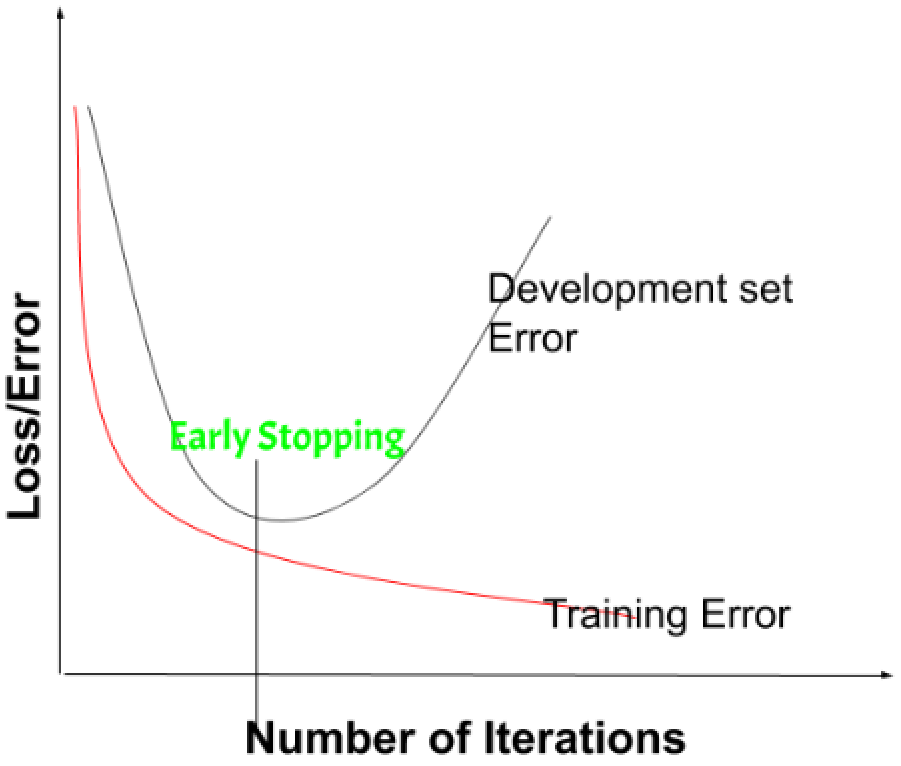
$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i) \qquad (x_0^i = 1)$$

$$\theta_j = \theta_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)\, x_j^i + \lambda \theta_j\right]$$

J = 1 , 2 , … , n

}

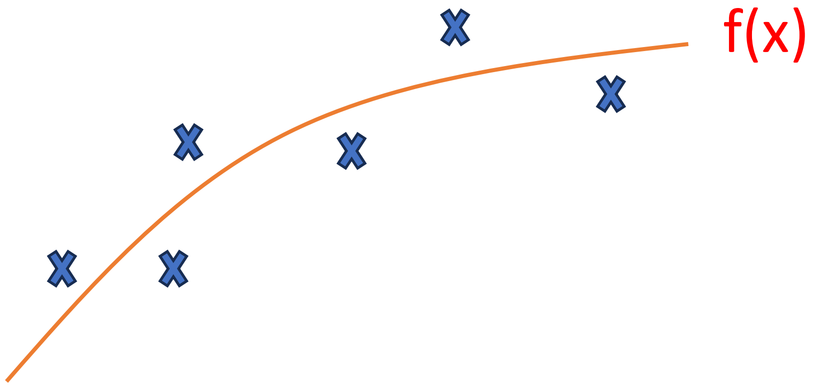
Early stopping



https://www.geeksforgeeks.org/regularization-by-early-stopping/

تعریف تئوری بایاس ـ واریانس

مدل مولد داده:



$Y = f(x) + \varepsilon$

ε: نویز با توزیع D_ε که مستقل از داده ها است.

S_{train}: داده های آموزشی

D: فضای داده ها

محاسبه رابطه خطا

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left(f(\mathbf{x})+\varepsilon-f_{S_{\text{train}}}(\mathbf{x})\right)^2\right]$$

برای یک نقطه x_0 خطا به صورت زیر است:

$$\left(f(\mathbf{x}_0)+\varepsilon-f_{S_{\text{train}}}(\mathbf{x}_0)\right)^2$$

فرض کنید که با داده های آموزشی مختلفی که از فضای داده D نمونه گیری شده اند آزمایش را تکرار میکنیم. در این حالت خطای داده x_0 به صورت زیر محاسبه می شود:

$$\mathbb{E}_{S_{\text{train}}\sim\mathcal{D},\varepsilon\sim\mathcal{D}_\varepsilon}\left[\left(f(\mathbf{x}_0)+\varepsilon-f_{S_{\text{train}}}(\mathbf{x}_0)\right)^2\right]$$

ادامه محاسبه رابطه خطا

می توانیم رابطه بالا را به صورت زیر به دست آوریم:

$$
\begin{aligned}
&\mathbb{E}_{S_{\text{train}} \sim \mathcal{D}, \varepsilon \sim \mathcal{D}_\varepsilon}\left[\left(f(\mathbf{x}_0) + \varepsilon - f_{S_{\text{train}}}(\mathbf{x}_0)\right)^2\right] \\
&\overset{(a)}{=} \mathbb{E}_{\varepsilon \sim \mathcal{D}_\varepsilon}[\varepsilon^2] + \mathbb{E}_{S_{\text{train}} \sim \mathcal{D}}\left[\left(f(\mathbf{x}_0) - f_{S_{\text{train}}}(\mathbf{x}_0)\right)^2\right] \\
&\overset{(b)}{=} \mathrm{Var}_{\varepsilon \sim \mathcal{D}_\varepsilon}[\varepsilon] + \mathbb{E}_{S_{\text{train}} \sim \mathcal{D}}\left[\left(f(\mathbf{x}_0) - f_{S_{\text{train}}}(\mathbf{x}_0)\right)^2\right] \\
&\overset{(c)}{=} \underbrace{\mathrm{Var}_{\varepsilon \sim \mathcal{D}_\varepsilon}[\varepsilon]}_{\text{noise variance}} \\
&+ \underbrace{\left(f(\mathbf{x}_0) - \mathbb{E}_{S'_{\text{train}} \sim \mathcal{D}}\left[f_{S'_{\text{train}}}(\mathbf{x}_0)\right]\right)^2}_{\text{bias}} \\
&+ \underbrace{\mathbb{E}_{S_{\text{train}} \sim \mathcal{D}}\left[\left(\mathbb{E}_{S'_{\text{train}} \sim \mathcal{D}}\left[f_{S'_{\text{train}}}(\mathbf{x}_0)\right] - f_{S_{\text{train}}}(\mathbf{x}_0)\right)^2\right]}_{\text{variance}}.
\end{aligned}
$$

ادامه محاسبه رابطه خطا

توجه کنید در بخش (a) عبارت زیر حذف شده است. چرا؟؟

$$\mathbb{E}_{S_{\text{train}} \sim \mathcal{D}, \varepsilon \sim \mathcal{D}_{\varepsilon}} \left[2\varepsilon \left(f(\mathbf{x}_0) - f_{S_{\text{train}}}(\mathbf{x}_0) \right) \right]$$

در بخش (b):

$$E_{\varepsilon \sim D_{\varepsilon}}[\varepsilon^2] = var_{\varepsilon \sim D_{\varepsilon}}[\varepsilon]$$

در بخش (c):

عبارت $\mathbb{E}_{S'_{\text{train}} \sim \mathcal{D}} \left[f_{S'_{\text{train}}}(\mathbf{x}_0) \right]$ که (S' یک مجموعه داده از D است) را به رابطه اضافه و کم می کنیم و سپس توان ۲ را اعمال می کنیم. در این رابطه یک ترم سوم هم وجود دارد که نشان می دهیم که به صورت زیر برابر با صفر است:

ادامه محاسبه رابطه خطا

$$
\begin{aligned}
& \mathbb{E}_{S\sim\mathcal{D}}\left[\left(f(\mathbf{x}_0) - \mathbb{E}_{S'\sim\mathcal{D}}[f_{S'}(\mathbf{x}_0)]\right)\cdot\Big(\mathbb{E}_{S'\sim\mathcal{D}}[f_{S'}(\mathbf{x}_0)] - f_S(\mathbf{x}_0)\Big)\right] \\
= \quad & \left(f(\mathbf{x}_0) - \mathbb{E}_{S'\sim\mathcal{D}}[f_{S'}(\mathbf{x}_0)]\right)\cdot \mathbb{E}_{S\sim\mathcal{D}}\left[\Big(\mathbb{E}_{S'\sim\mathcal{D}}[f_{S'}(\mathbf{x}_0)] - f_S(\mathbf{x}_0)\Big)\right] \\
= \quad & \left(f(\mathbf{x}_0) - \mathbb{E}_{S'\sim\mathcal{D}}[f_{S'}(\mathbf{x}_0)]\right)\cdot\left(\mathbb{E}_{S'\sim\mathcal{D}}[f_{S'}(\mathbf{x}_0)] - \mathbb{E}_{S\sim\mathcal{D}}[f_S(\mathbf{x}_0)]\right) \\
= \quad & 0
\end{aligned}
$$

ادامه محاسبه رابطه خطا

تعبیر رابطه (c):

از سه ترم مثبت تشکیل شده است. ترم اول ربطی به نحوه آموزش مدل ندارد و ناشی از عدم قطعیت ذاتی در داده ها است.

بایاس تفاضل مابین مقدار واقعی$f(x_0)$و متوسط مدل های مختلفی است که بر روی داده ها آموزش دیده اند. (مدل های ساده نمی توانند خوب بر روی داده ها تطبیق یابند. در نتیجه بایاس زیاد می شود.)

ترم واریانس در واقع واریانس مدل های مختلفی است که آموزش دیده اند. اگر مدل ما خیلی پیچیده باشد با تغییر اندکی در داده ها شکل مدل عوض می شود و پیش بینی بر روی x_0 به میزان زیادی متغیر می شود.

Examples

$$\underbrace{\left(f(\mathbf{x}_0) - \mathbb{E}_{S'_{\text{train}} \sim \mathcal{D}}\left[f_{S'_{\text{train}}}(\mathbf{x}_0)\right]\right)^2}_{\text{bias}}$$

$$\underbrace{\mathbb{E}_{S_{\text{train}} \sim \mathcal{D}}[(\mathbb{E}_{S'_{\text{train}} \sim \mathcal{D}}\left[f_{S'_{\text{train}}}(\mathbf{x}_0)\right] - f_{S_{\text{train}}}(\mathbf{x}_0))^2]}_{\text{variance}}.$$

- $f_S(x) = k$

Prove it

Bias is high

Variance is 0

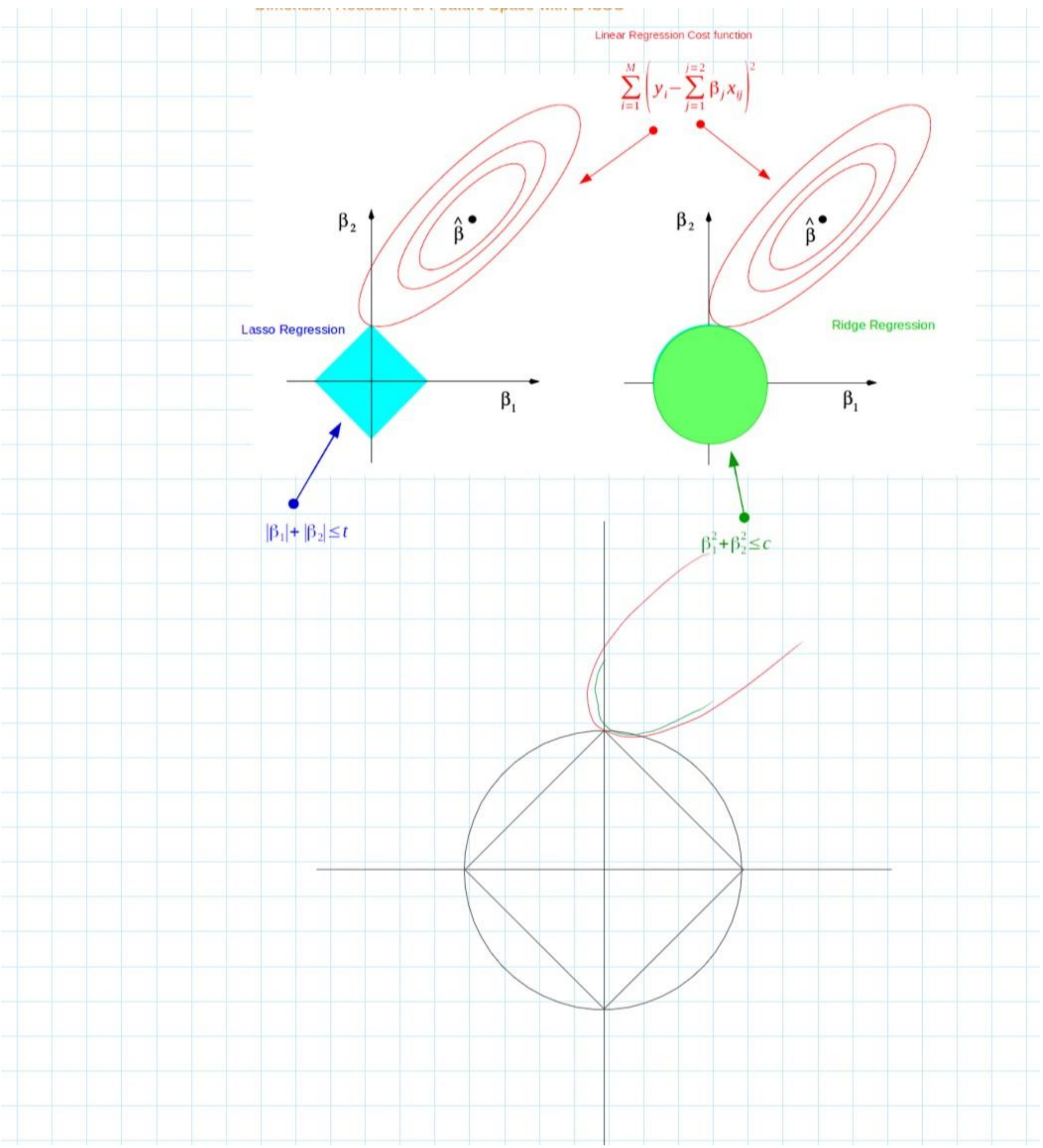
$$\boldsymbol{f_S(x) = f(x) + \varepsilon}$$

Bias is zero

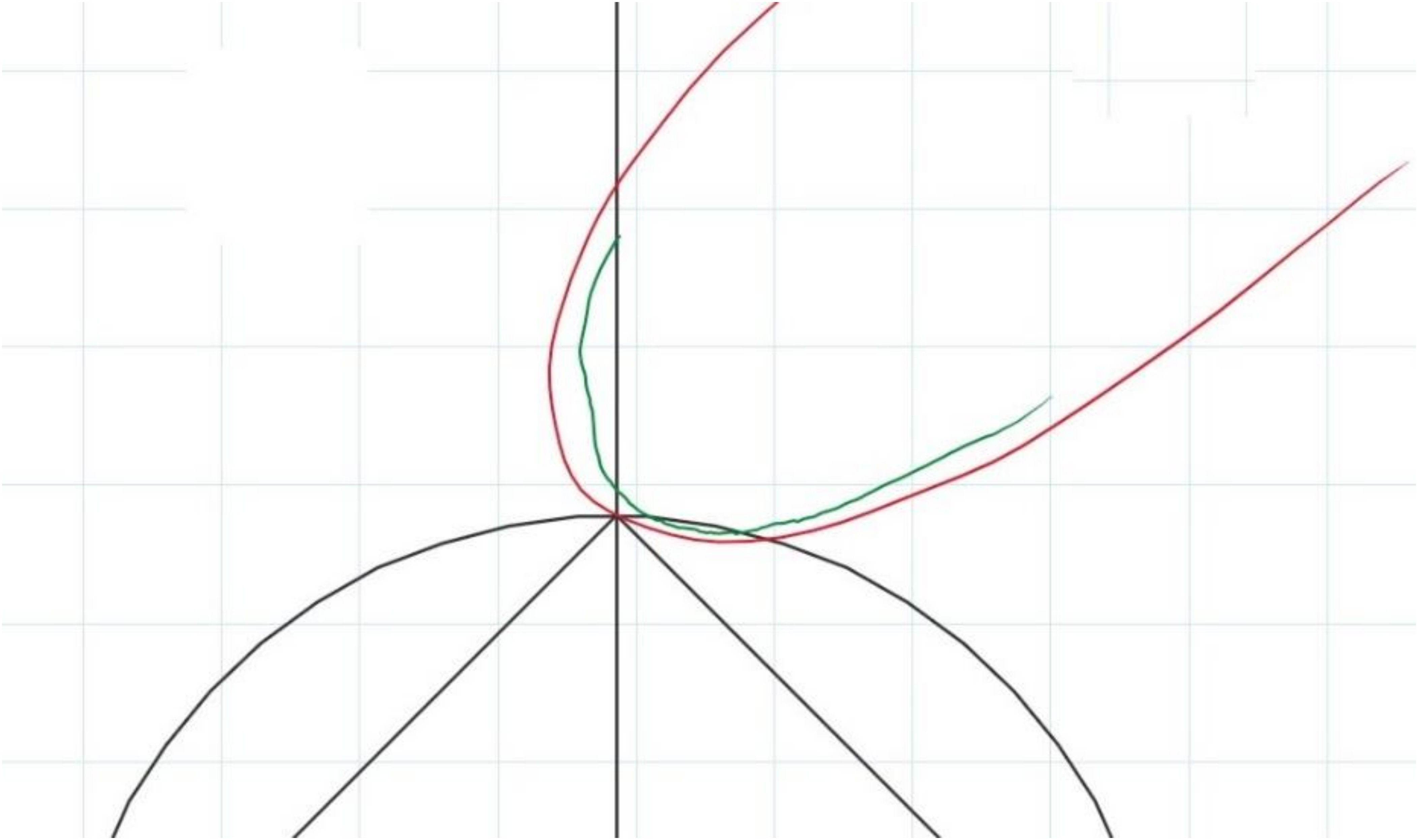
Variance is $var_{\varepsilon \sim D_\varepsilon}[\varepsilon]$

Lasso Regression

- $J(\theta) = \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda \sum_{j=1}^{n} \left|\theta_j\right|$



Linear Regression Cost function
$\sum_{i=1}^{M}\left(y_i-\sum_{j=1}^{j=2}\beta_j x_{ij}\right)^2$
β_2
$\hat{\beta}$
Lasso Regression
β_1
$|\beta_1|+|\beta_2|\leq t$
β_2
$\hat{\beta}$
Ridge Regression
β_1
$\beta_1^2+\beta_2^2\leq c$



چند مثال

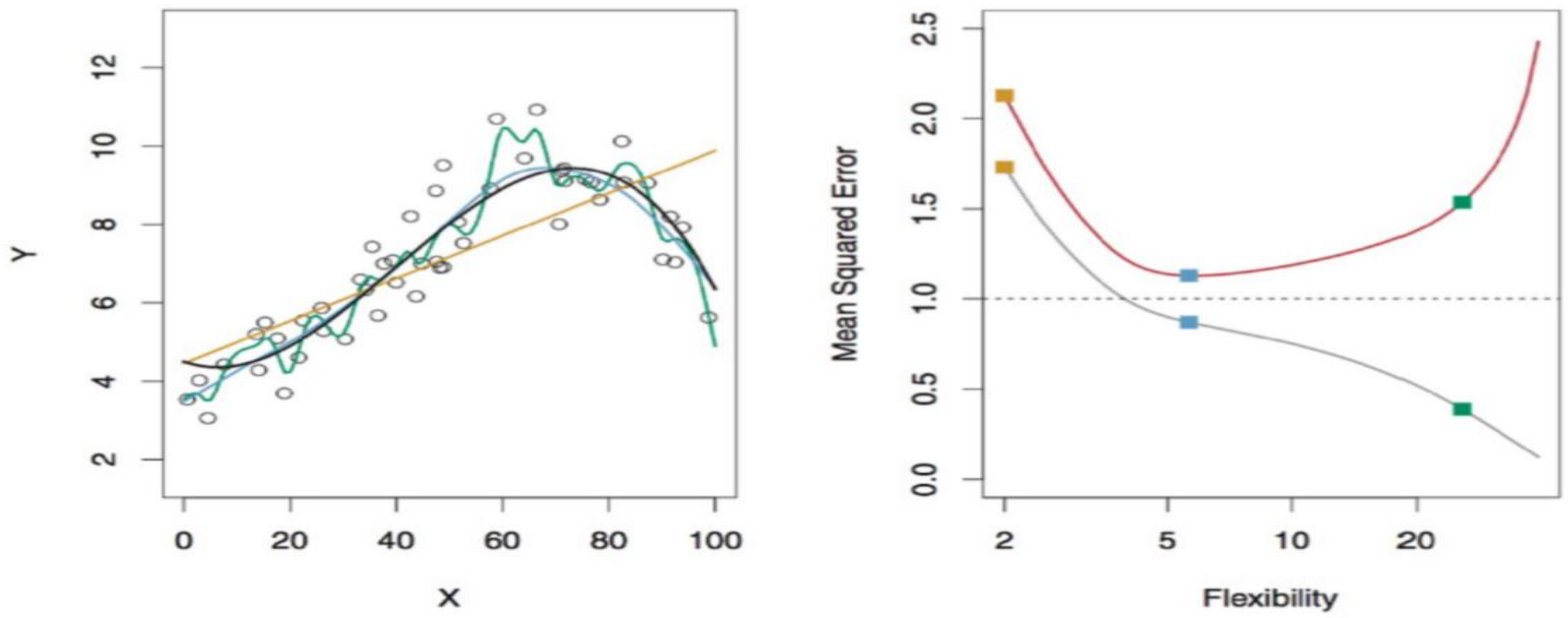


FIGURE 2.9. Left: *Data simulated from* f*, shown in black. Three estimates of* f *are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

چند مثال

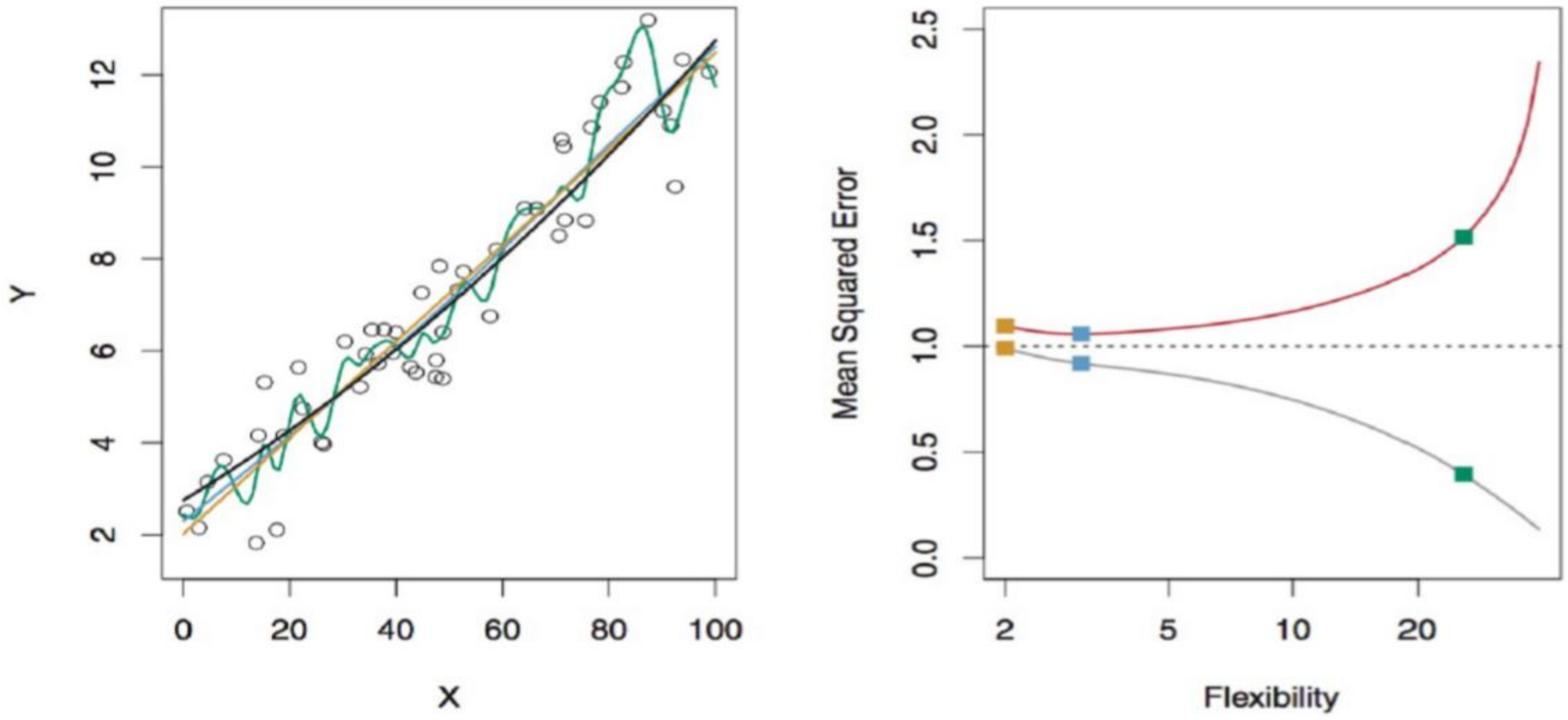


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

چند مثال

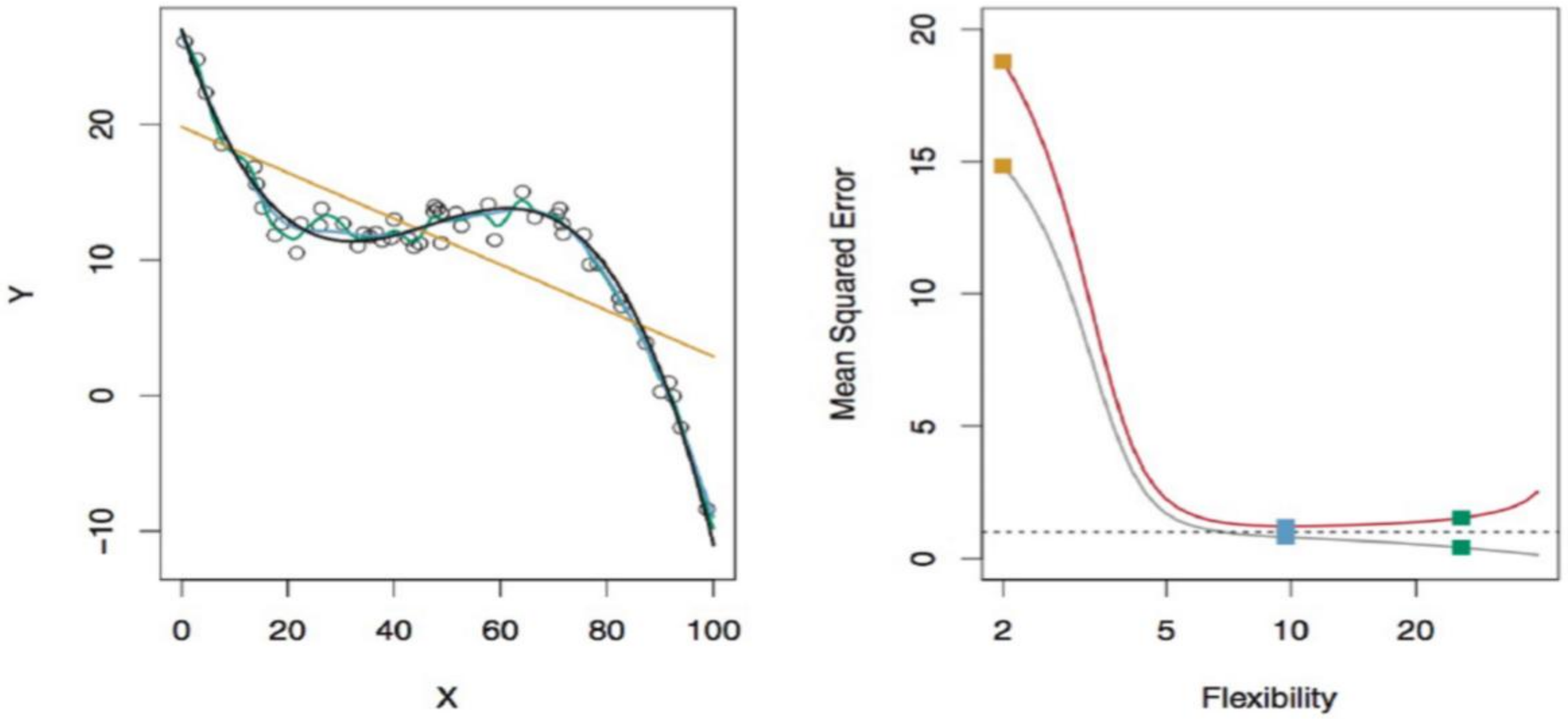


FIGURE 2.11. *Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

چند مثال

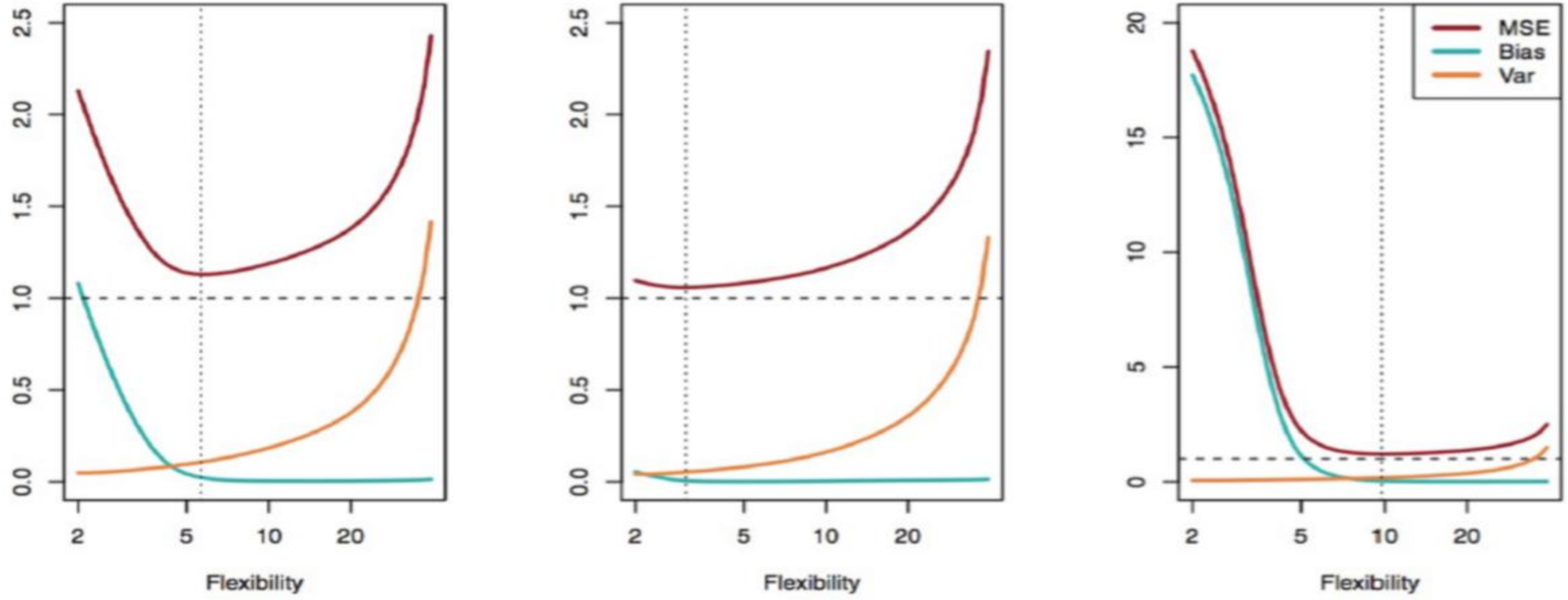


FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

Old way of splitting data

- Deep learning

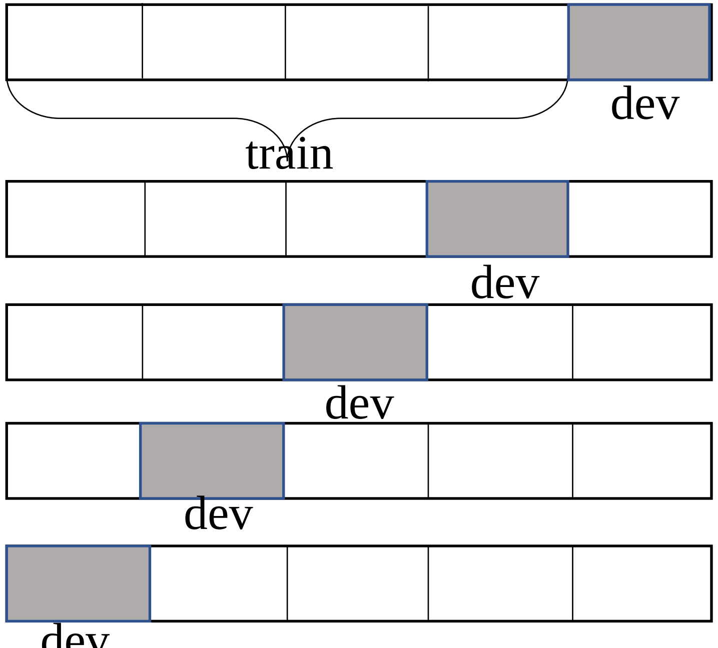
60% 20% 20%

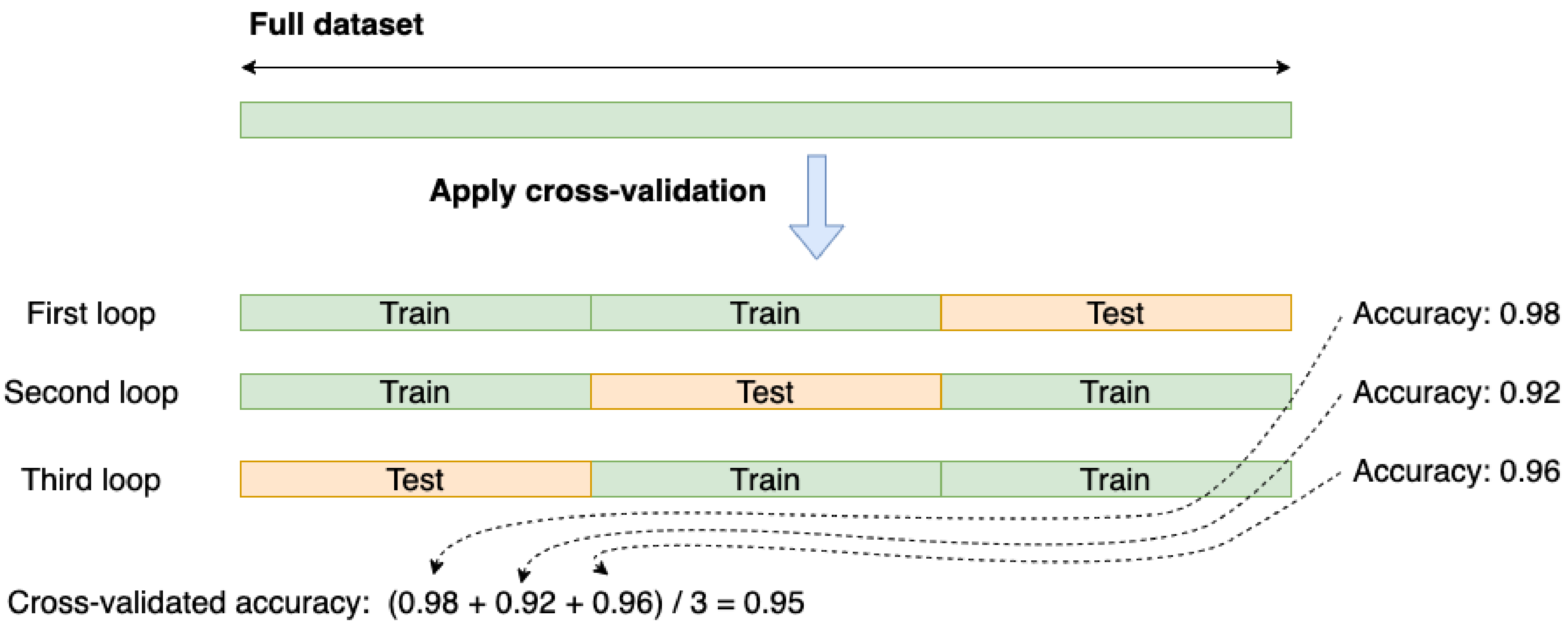
1000,000 data 1% 10,000

98% 1% 1%

99.5% 0.4% 0.1%

- K-fold cv




Full dataset
Apply cross-validation
First loop
Train
Train
Test
Accuracy: 0.98
Second loop
Train
Test
Train
Accuracy: 0.92
Third loop
Test
Train
Train
Accuracy: 0.96
Cross-validated accuracy: (0.98 + 0.92 + 0.96) / 3 = 0.95