Machine Learning

Dr. Mehran Safayani

safayani@iut.ac.ir

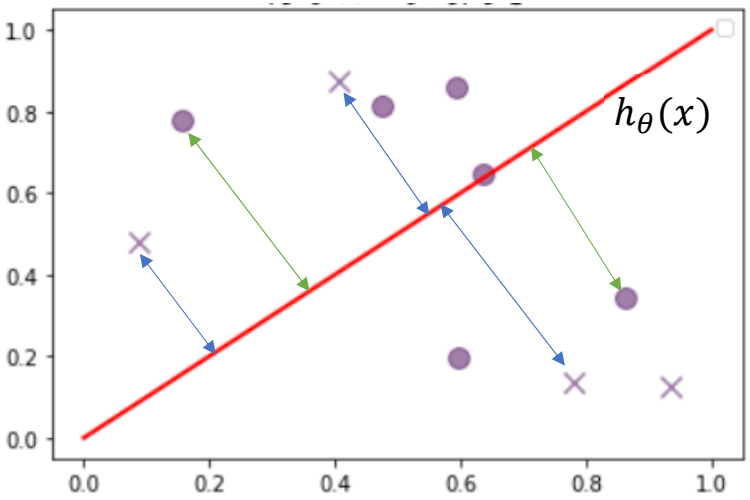
safayani.iut.ac.ir

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

Underfitting and Overfitting

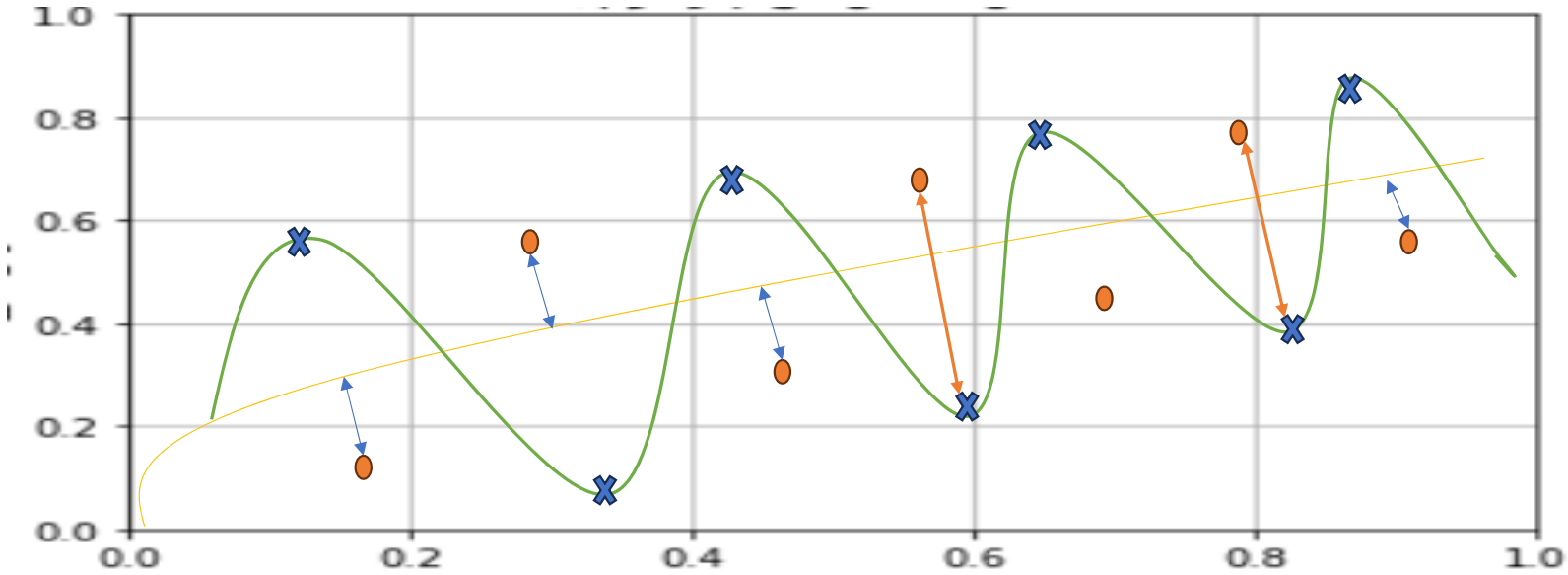


$h_\theta(x) = \theta_0 + \theta_1 x$

خطای زمان آموزش بسیار زیاد

خطای زمان آزمون بسیار زیاد

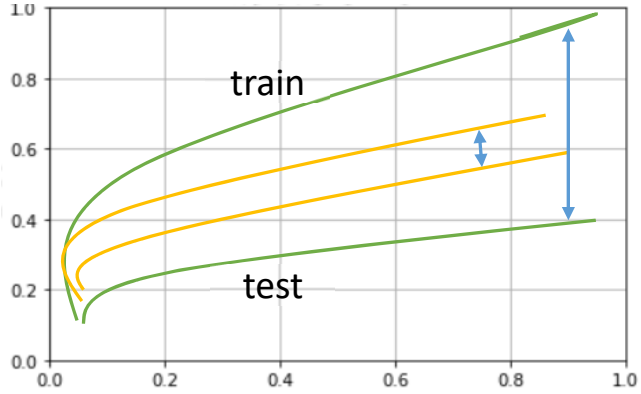
underfit



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

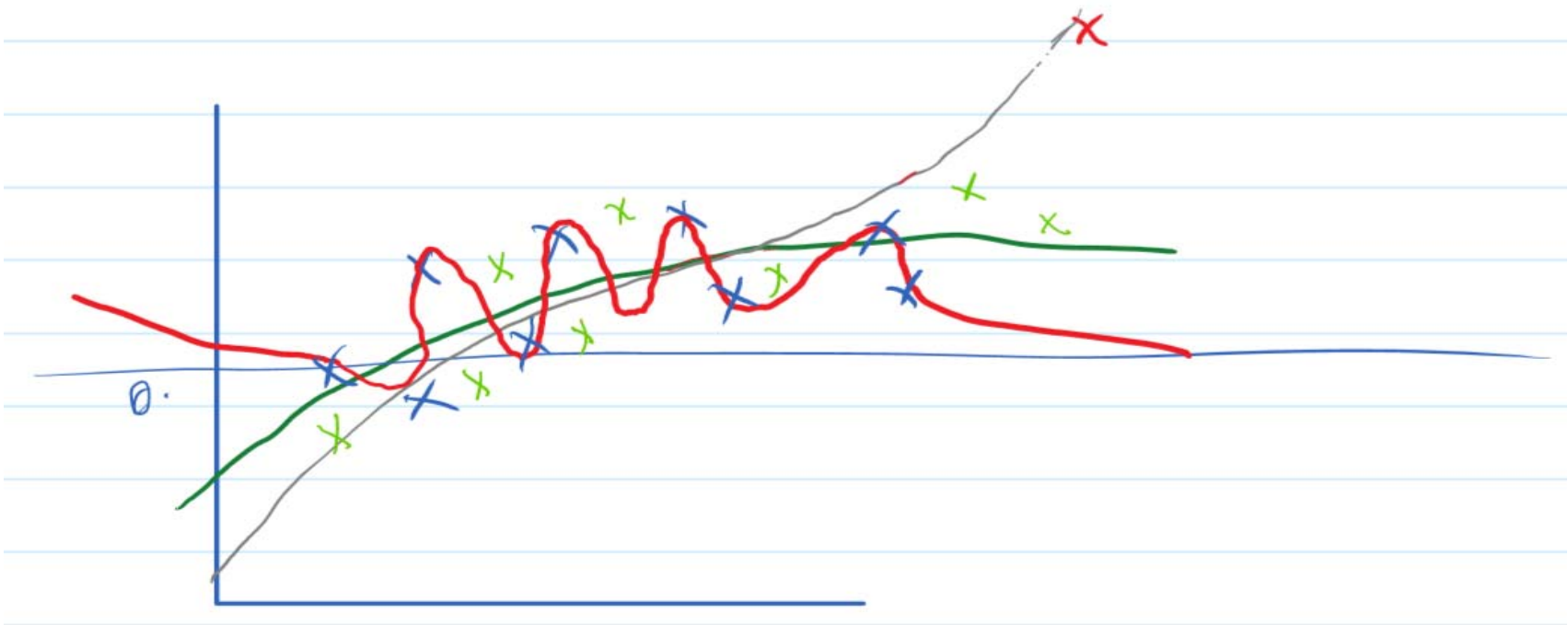
خطای آموزش بسیار کم
خطای آزمون بسیار زیاد

overfit



$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

Just right



Right : $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

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

Regularization

$$\underset{\theta}{\text{Minimize}} \ \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2 + 1000\,\theta_3^2 + 1000\,\theta_4^2$$

$\theta_3 = 0$, $\theta_4 = 0$

Features: $x_1, x_2, \dots, x_{100}$

Parameters: $\theta_0, \theta_1, \theta_2, \dots, \theta_{100}$

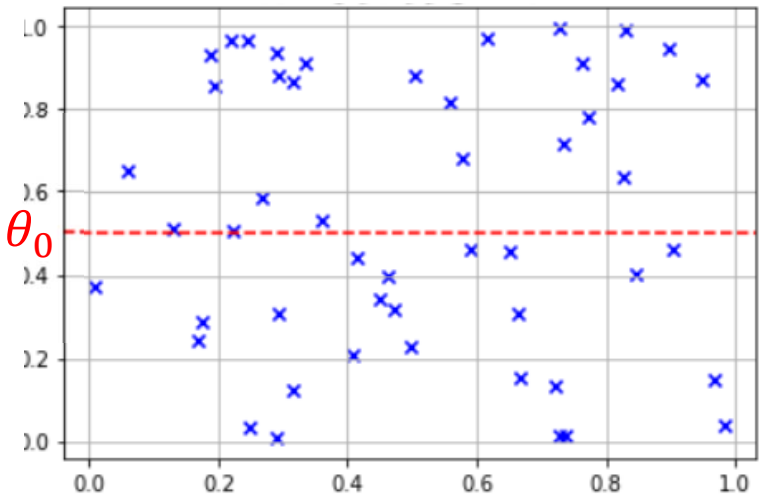
$$\underset{\theta}{\text{Min}}\ J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right]$$

Regularization ضريب

Regularization

If λ is very big ($\lambda = 10^{10}$) ⟹ underfit , $\theta_1, \theta_2, \dots, \theta_n = 0$, $h_\theta(x_i) = \theta_0$

If λ is very small ($\lambda = 0$) ⟹ overfit



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

}

Train set: 60% $(x^1, y^1), (x^2, y^2), \ldots, (x^M, y^M)$

Cross validation set: 20% $(x_{cv}^1, y_{cv}^1), (x_{cv}^2, y_{cv}^2), \ldots, (x_{cv}^M, y_{cv}^M)$

Test set: 20% $(x_{test}^1, y_{test}^1), (x_{test}^2, y_{test}^2), \ldots, (x_{test}^M, y_{test}^M)$

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

$$J_{cv}(\theta) = \frac{1}{2m} \sum_{i=1}^{M_{cv}} (h_\theta(x_{cv}^i) - y_i)^2$$

$$J_{test}(\theta) = \frac{1}{2m} \sum_{i=1}^{M_{test}} (h_\theta(x_{test}^i) - y_i)^2$$

Model Selection

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

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

.

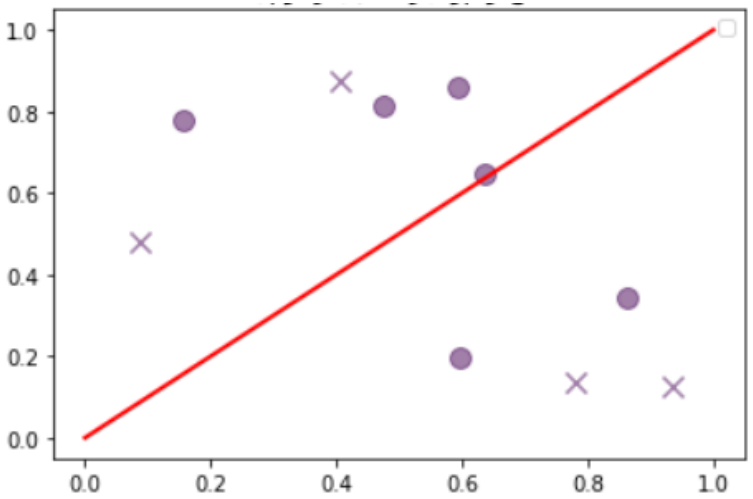
.

.

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

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

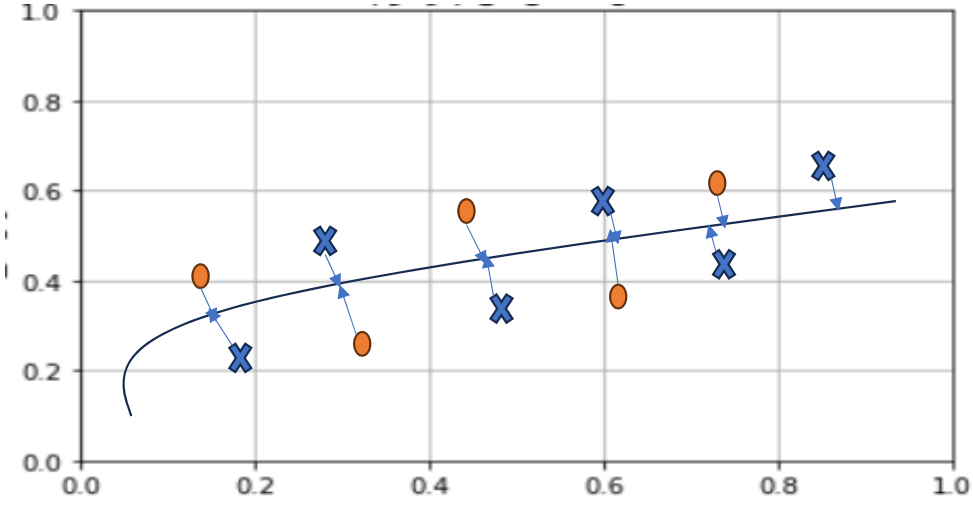
$J_{test}(\theta^{i^*})$



$\theta_0 + \theta_1 x$

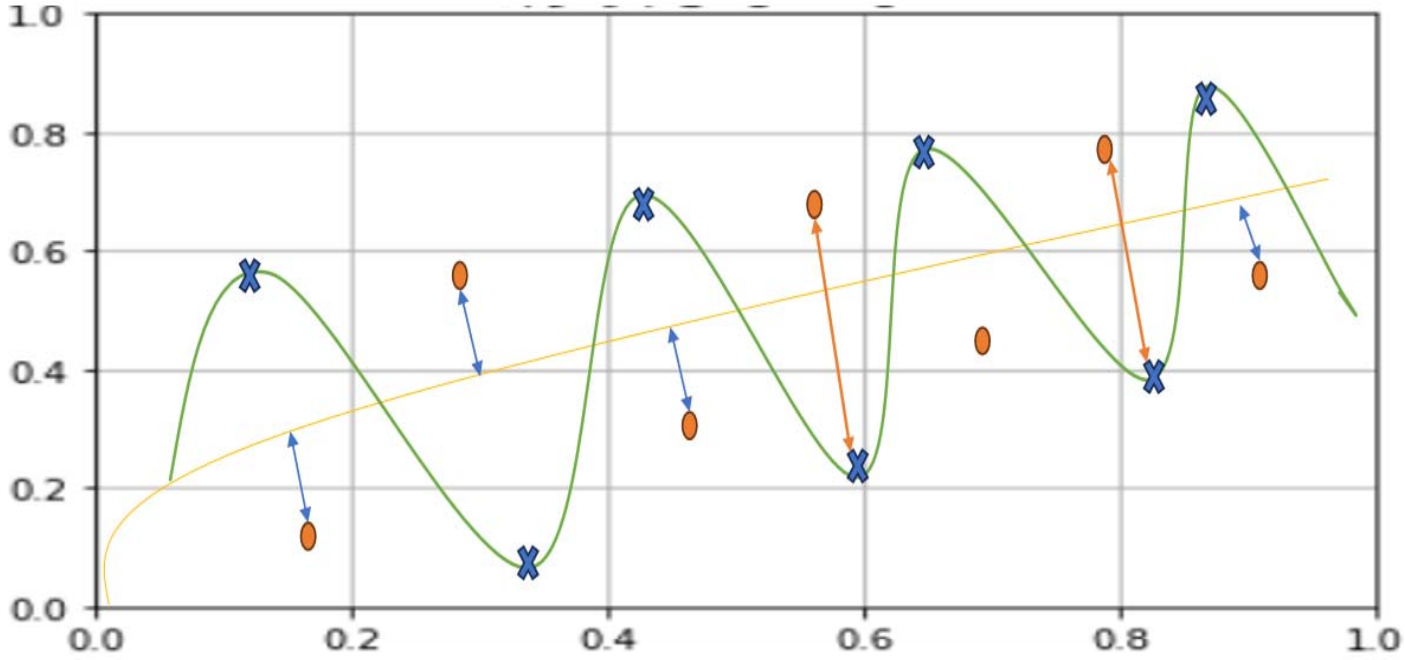
Underfit
High bias

خطای زیاد آموزش
خطای زیاد CV



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

خطای آموزش کم
خطای CV کم
نزدیک به یکدیگر



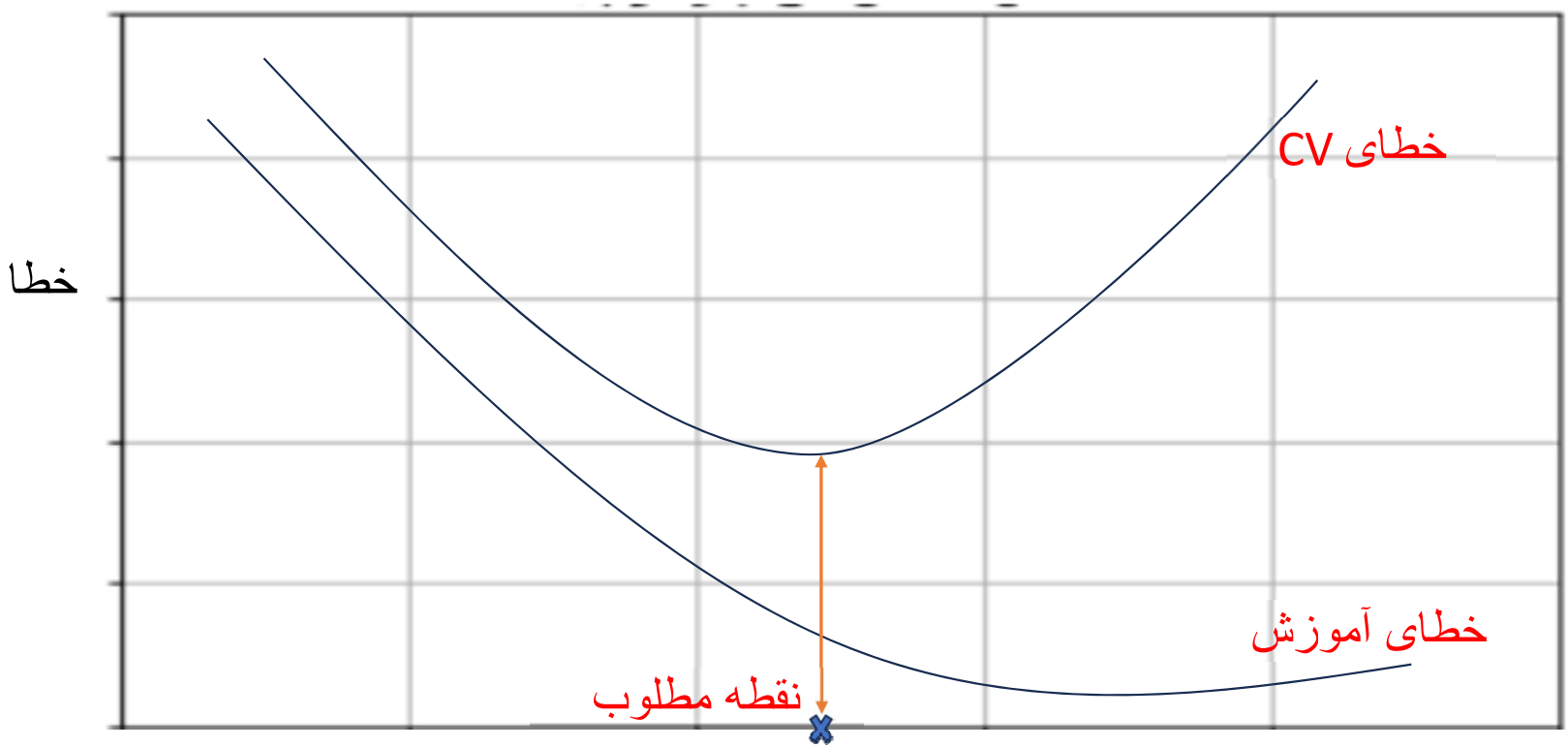
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

Overfit

High variance

خطای آموزشی خیلی کم

خطای CV زیاد


خطا
خطای CV
خطای آموزش
نقطه مطلوب
پارامترهای مدل

Regularization

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2 + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$$

If λ is too large ---> λ = 10000 ---> underfit ---> high bias
---> $\theta_1 = \theta_2 = \ldots = \theta_5 = 0$ ---> h(x) = θ_0

Appropriate λ ---> appropriate result

If λ is too small ---> λ = 0 ---> overfit ---> high variance

Regularization

λ = 0 ---> $J_{train}(m^1)$ ---> $J_{cv}(m^1)$

λ = 0.01 ---> $J_{train}(m^2)$ ---> $J_{cv}(m^2)$

λ = 0.02 ---> $J_{train}(m^3)$ ---> $J_{cv}(m^3)$

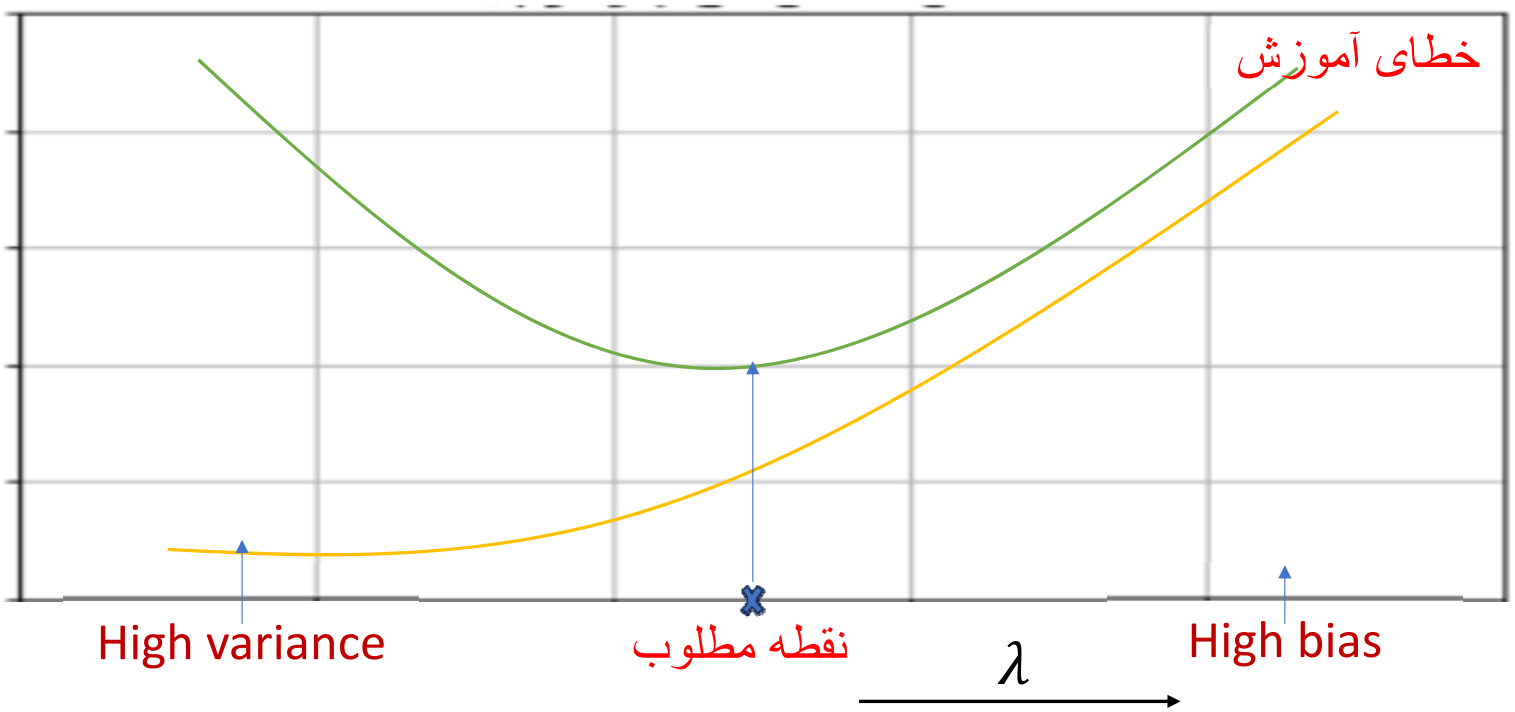
.

.

.

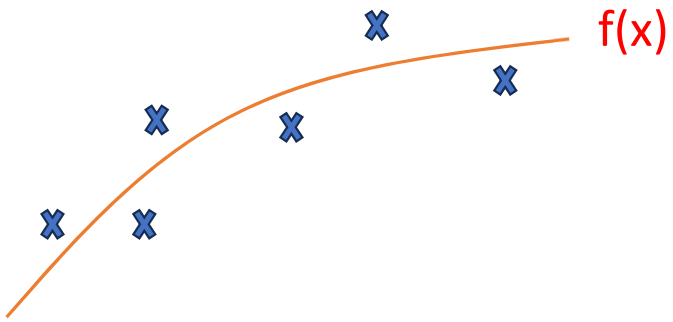
λ = 10 ---> $J_{train}(m^{10})$ ---> $J_{cv}(m^{10})$

i^* = $\underset{i}{\text{argmin}}\ J_{cv}(m^i)$


خطای آموزش
High variance
نقطه مطلوب
High bias
λ

تعریف تئوری باياس ـ واريانس

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

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

برای یک نقطه x_0 خطا به صورت زیر است:

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

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:

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

ادامه محاسبه رابطه خطا

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

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

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 $\mathbb{E}_{S'_{\text{train}}\sim\mathcal{D}}[f_{S'_{\text{train}}}(\mathbf{x}_0)]$ 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

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

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.

چند مثال

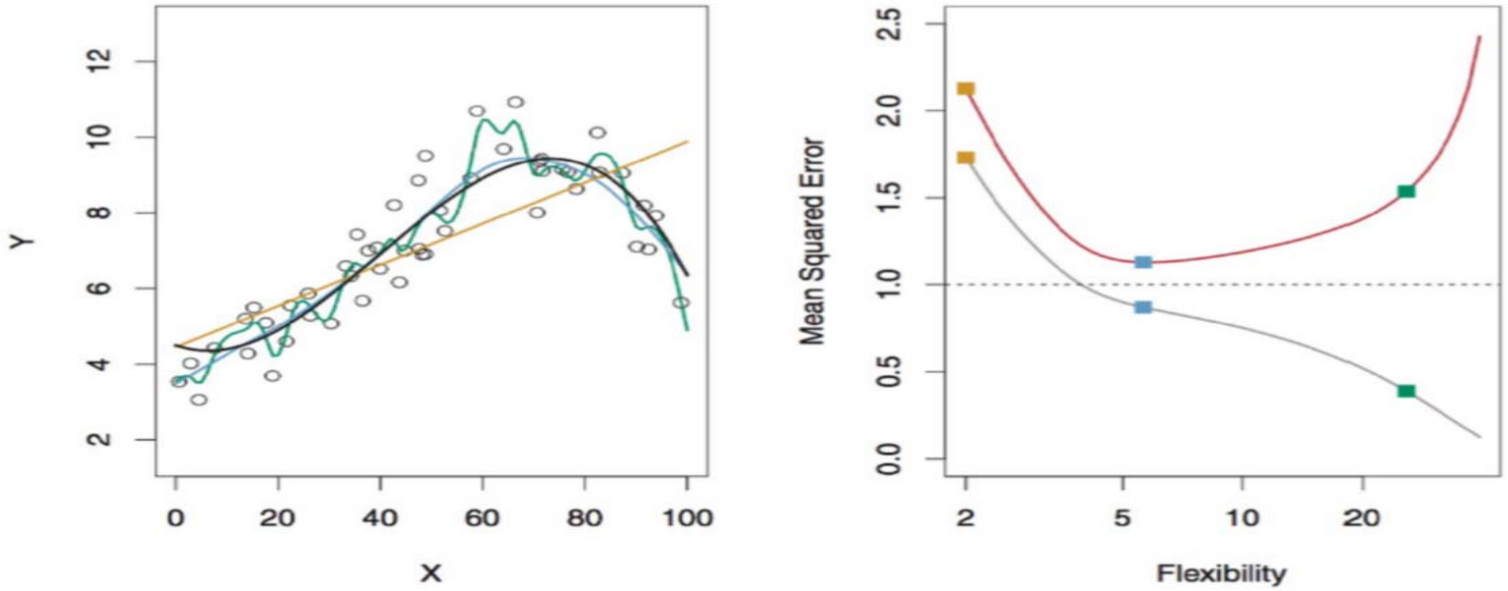


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed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

چند مثال

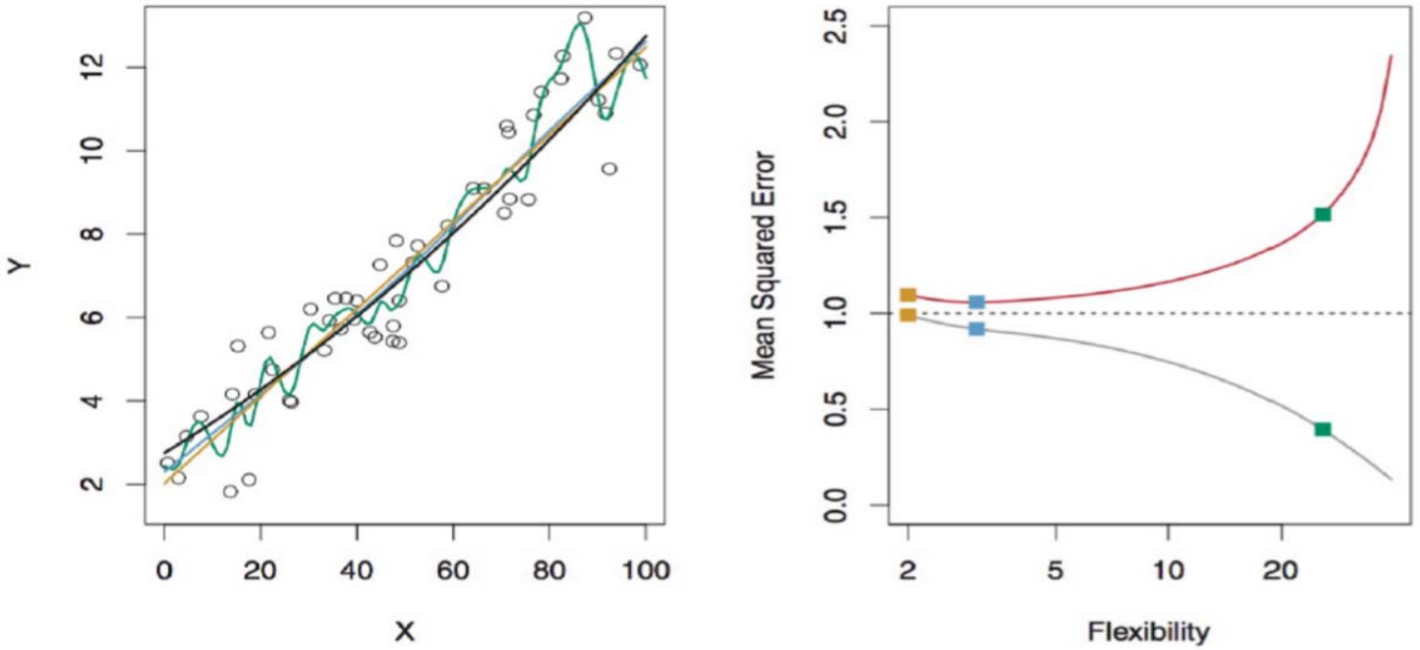


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

چند مثال

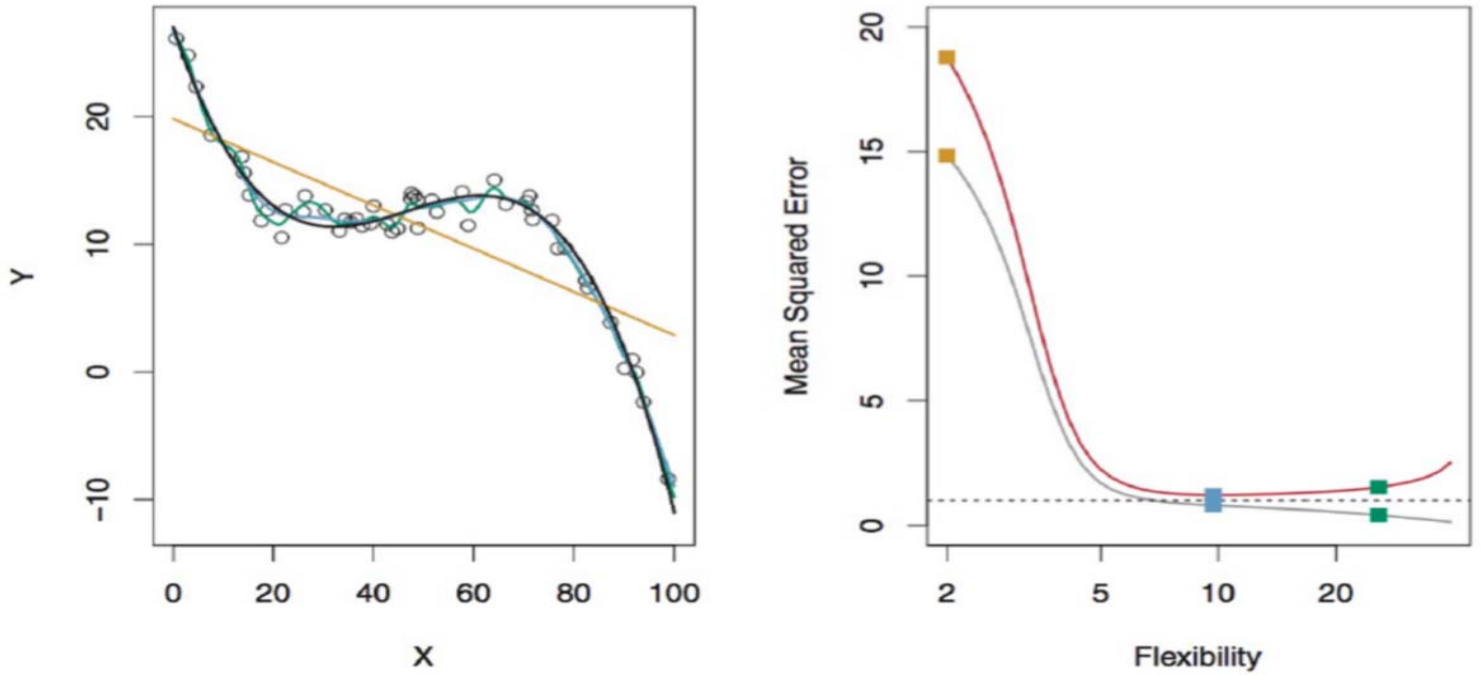


FIGURE 2.11. *Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

چند مثال

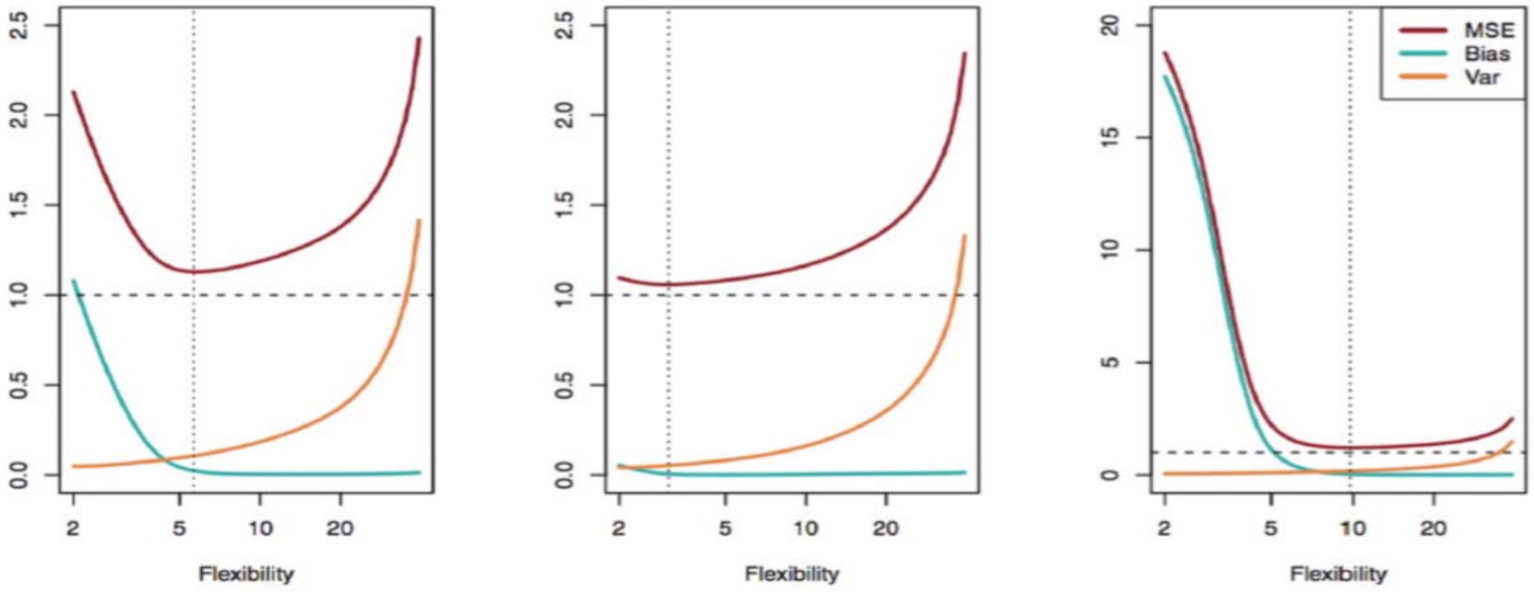


FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

K fold CV