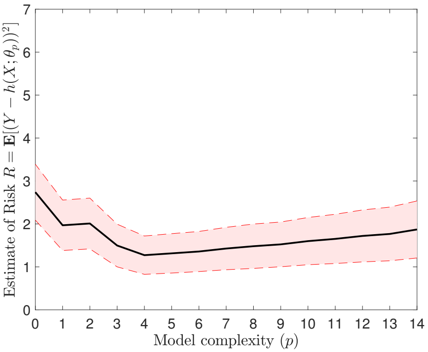
Statistical Learning

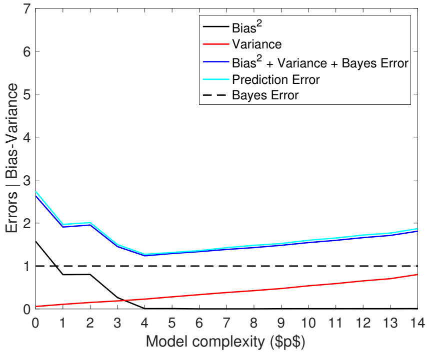
Master Spécialisé Intelligence Artificielle de Confiance (IAC)
@ Centrale Supélec en partenariat avec l'IRT SystemX
2024/2025.

Faïcel Chamroukhi

chamroukhi.com

- Risk Decomposition (Continued)
- Bias-Variance Decomposition





Bias-Variance Decomposition

Setting : Prediction under the squared loss

- Prediction function

$$h\colon \mathbb{R}^p \to \mathbb{R}^d$$
$$x \mapsto h(x)$$

- Squared (ℓ_2)-loss function :

$$\ell\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$
$$(h(x), y) \mapsto \ell(y, h(x)) = (y - h(x))^2$$

Expected Risk

- Consider the Risk :
  $R_x(h) = \mathbb{E}_P[\ell(Y, h(X))|X = x] = \mathbb{E}_{Y|X=x}((Y - h(X))^2|X = x)$
- Best prediction function (Bayes predictor) : $h^*(x) = \mathbb{E}(Y|X = x)$.
- Bayes Risk : $R(h^*)$
- Excess Risk : $R(h) - R(h^*)$

Bias-Variance Decomposition

Setting : Prediction under the squared loss

- Prediction function

$$\begin{aligned} h\colon \mathbb{R}^p &\to \mathbb{R}^d \\ x &\mapsto h(x) \end{aligned}$$

- Squared (ℓ_2)-loss function :

$$\begin{aligned} \ell\colon \mathbb{R}^d \times \mathbb{R}^d &\to \mathbb{R} \\ (h(x), y) &\mapsto \ell(y, h(x)) = (y - h(x))^2 \end{aligned}$$

Expected Risk

- Consider the Risk :
  $R_x(h) = \mathbb{E}_P[\ell(Y, h(X))|X = x] = \mathbb{E}_{Y|X=x}((Y - h(X))^2|X = x)$
- Best prediction function (Bayes predictor) : $h^*(x) = \mathbb{E}(Y|X = x)$.
- Bayes Risk : $R(h^*)$
- Excess Risk : $R(h) - R(h^*)$

Bias-Variance Decomposition

$$\begin{aligned}
\mathbb{E}[(h(X) - h^*(X))^2] &= \mathbb{E}[(h(X) - \mathbb{E}[h(X)] + \mathbb{E}[h(X)] - h^*(X))^2] \\
&= \mathbb{E}[(h(X) - \mathbb{E}[h(X)])^2] + \mathbb{E}[(\mathbb{E}[h(X)] - h^*(X))^2] \\
&\quad +2\underbrace{\mathbb{E}[(h(X) - \mathbb{E}[h(X)])\,(\mathbb{E}[h(X)] - h^*(X))]}_{=0} \\
&= \underbrace{\mathbb{E}[(h(X) - \mathbb{E}[h(X)])^2]}_{\text{Variance}(h(X))} + \underbrace{\mathbb{E}[(\mathbb{E}[h(X)] - h^*(X))^2]}_{\text{Bias}^2(h(X),h^*(X))}
\end{aligned}$$

- **Bias** : Systematic deviation of the average prediction from the true value.
- **Variance** : Amount of variability in the predictions for different training sets.
- **Bayes Error** : Intrinsic randomness in the target variable that no model can eliminate.

Bias-Variance Decomposition

Bias-Variance Decomposition

$$
\begin{aligned}
\mathbb{E}[(h(X) - h^*(X))^2] &= \mathbb{E}[(h(X) - \mathbb{E}[h(X)] + \mathbb{E}[h(X)] - h^*(X))^2] \\
&= \mathbb{E}[(h(X) - \mathbb{E}[h(X)])^2] + \mathbb{E}[(\mathbb{E}[h(X)] - h^*(X))^2] \\
&\quad +2\,\underbrace{\mathbb{E}[(h(X) - \mathbb{E}[h(X)])\,(\mathbb{E}[h(X)] - h^*(X))]}_{=0} \\
&= \underbrace{\mathbb{E}[(h(X) - \mathbb{E}[h(X)])^2]}_{\text{Variance}(h(X))} + \underbrace{\mathbb{E}[(\mathbb{E}[h(X)] - h^*(X))^2]}_{\text{Bias}^2(h(X),h^*(X))}
\end{aligned}
$$

- **Bias** : Systematic deviation of the average prediction from the true value.
- **Variance** : Amount of variability in the predictions for different training sets.
- **Bayes Error** : Intrinsic randomness in the target variable that no model can eliminate.

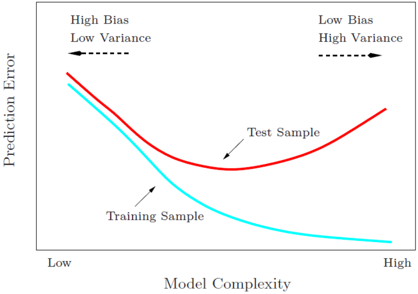
Bias-Variance Decomposition

The third term in the previous step vanishes because
by conditioning on X and using the law of total expectations we get :

$$\mathbb{E}\Big[(h(X)-\mathbb{E}[h(X)])(\mathbb{E}[h(X)]-h^*(X))\Big] = \mathbb{E}\Big[\mathbb{E}[(h(X)-\mathbb{E}[h(X)])|X]\cdot(\mathbb{E}[h(X)]-h^*(X))\Big].$$

and

$$\begin{aligned}\mathbb{E}[h(X) - \mathbb{E}[h(X)]|X] &= \mathbb{E}_X[\mathbb{E}[h(X) - \mathbb{E}[h(X)]|X]] \\ &= \mathbb{E}[\mathbb{E}[h(X)|X] - \mathbb{E}[\mathbb{E}[h(X)]|X]] \\ &= \mathbb{E}[h(X) - \mathbb{E}[h(X)]] \\ &= \mathbb{E}[h(X)] - \mathbb{E}[h(X)] \\ &= 0.\end{aligned}$$

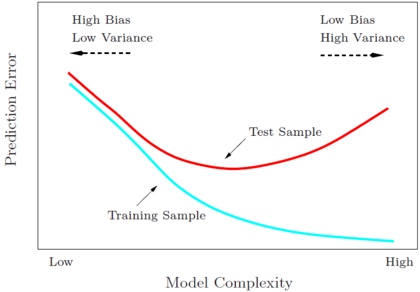


More complex models overfit while the simplest models underfit.

- If $\mathcal{H}$ has a large number of parameters, training a function $h \in \mathcal{H}$ can closely approximate h^*, thereby reducing bias. However, it becomes sensitive to variations in the training set, leading to increased variance.
- If $\mathcal{H}$ has a small number of parameters, any function $h \in \mathcal{H}$ deviates from h^*, increasing bias. However, it is less sensitive to fluctuations across different training sets, which results in lower variance.

↪ increasing model complexity reduces squared bias but increases variance. Conversely, decreasing model complexity raises bias while reducing variance.

↪ The goal is to find an optimal balance that **minimizes the generalization error**, which includes both bias and variance components.

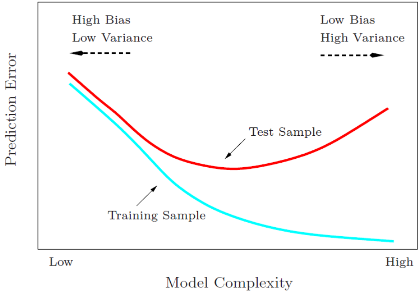


More complex models overfit while the simplest models underfit.

- If $\mathcal{H}$ has a large number of parameters, training a function $h \in \mathcal{H}$ can closely approximate h^*, thereby reducing bias. However, it becomes sensitive to variations in the training set, leading to increased variance.
- If $\mathcal{H}$ has a small number of parameters, any function $h \in \mathcal{H}$ deviates from h^*, increasing bias. However, it is less sensitive to fluctuations across different training sets, which results in lower variance.

↪ increasing model complexity reduces squared bias but increases variance. Conversely, decreasing model complexity raises bias while reducing variance.

↪ The goal is to find an optimal balance that **minimizes the generalization error**, which includes both bias and variance components.

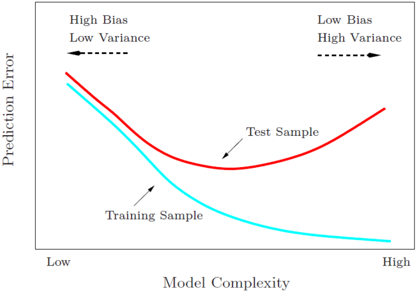


More complex models overfit while the simplest models underfit.

- If $\mathcal{H}$ has a large number of parameters, training a function $h \in \mathcal{H}$ can closely approximate h^*, thereby reducing bias. However, it becomes sensitive to variations in the training set, leading to increased variance.
- If $\mathcal{H}$ has a small number of parameters, any function $h \in \mathcal{H}$ deviates from h^*, increasing bias. However, it is less sensitive to fluctuations across different training sets, which results in lower variance.

↪ increasing model complexity reduces squared bias but increases variance. Conversely, decreasing model complexity raises bias while reducing variance.

↪ The goal is to find an optimal balance that **minimizes the generalization error**, which includes both bias and variance components.

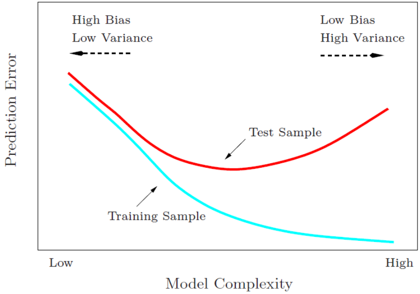


More complex models overfit while the simplest models underfit.

- If $\mathcal{H}$ has a large number of parameters, training a function $h \in \mathcal{H}$ can closely approximate h^*, thereby reducing bias. However, it becomes sensitive to variations in the training set, leading to increased variance.
- If $\mathcal{H}$ has a small number of parameters, any function $h \in \mathcal{H}$ deviates from h^*, increasing bias. However, it is less sensitive to fluctuations across different training sets, which results in lower variance.

$\hookrightarrow$ increasing model complexity reduces squared bias but increases variance. Conversely, decreasing model complexity raises bias while reducing variance.

$\hookrightarrow$ The goal is to find an optimal balance that **minimizes the generalization error**, which includes both bias and variance components.



More complex models overfit while the simplest models underfit.

- If $\mathcal{H}$ has a large number of parameters, training a function $h \in \mathcal{H}$ can closely approximate h^*, thereby reducing bias. However, it becomes sensitive to variations in the training set, leading to increased variance.
- If $\mathcal{H}$ has a small number of parameters, any function $h \in \mathcal{H}$ deviates from h^*, increasing bias. However, it is less sensitive to fluctuations across different training sets, which results in lower variance.

$\hookrightarrow$ increasing model complexity reduces squared bias but increases variance. Conversely, decreasing model complexity raises bias while reducing variance.

$\hookrightarrow$ The goal is to find an optimal balance that **minimizes the generalization error**, which includes both bias and variance components.

Risk decomposition for linear models

- Consider the statistical model $Y = f(X) + \varepsilon$, with f the true function
- ϵ_i's are independent with $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] = \sigma^2$
- Linear model : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in x of the form $\theta^T \widetilde{x}$ with $\widetilde{x} = (1, x^T)^T$, and $\theta = (\alpha, \beta^T)^T$. (denote $\widetilde{x}$ by x for simplicity)
- Bayes predictor h^* : for the squared loss : $h^*(x) = \mathbb{E}[Y|X = x] = f(x)$
- Let $\theta^* = (\alpha^*, \beta^{*T})^T$ be the optimal parameter. Then $h^*(x; \theta^*) = \theta^{*T} x = f(x)$

  Assume a fixed design, i.e. the x's are deterministic
- **Bayes Risk** $R^* = R(h^*) = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2|X = x] = \mathbb{E}[\epsilon^2|X = x] = \sigma^2$
- Risk for any **non-random** θ : $R(\theta) = \mathbb{E}[(Y - h(X;\theta))^2|X] = \sigma^2 + \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$:
- Excess risk of θ : $R(\theta) - R^* = \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$

  where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ and $\|u\|^2_A = u^T A u$.

  see proof in the next slide
- **ERM :** Solution : $\widehat{\theta}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$, whenever $\mathbf{X}^T\mathbf{X}$ has full rank. $(\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T)$

Risk decomposition for linear models

- Consider the statistical model $Y = f(X) + \varepsilon$, with f the true function
- ϵ_i's are independent with $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] = \sigma^2$
- Linear model : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in x of the form $\theta^T \widetilde{x}$ with $\widetilde{x} = (1, x^T)^T$, and $\theta = (\alpha, \beta^T)^T$. (denote $\widetilde{x}$ by x for simplicity)
- Bayes predictor h^* : for the squared loss : $h^*(x) = \mathbb{E}[Y|X = x] = f(x)$
- Let $\theta^* = (\alpha^*, \beta^{*T})^T$ be the optimal parameter. Then $h^*(x; \theta^*) = \theta^{*T} x = f(x)$

  Assume a fixed design, i.e. the x's are deterministic
- **Bayes Risk** $R^* = R(h^*) = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2 | X = x] = \mathbb{E}[\epsilon^2 | X = x] = \sigma^2$
- Risk for any **non-random** θ : $R(\theta) = \mathbb{E}[(Y - h(X;\theta))^2 | X] = \sigma^2 + \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$:
- Excess risk of θ : $R(\theta) - R^* = \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$

  where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ and $\|u\|^2_A = u^T A u$.

  see proof in the next slide
- **ERM :** Solution : $\widehat{\theta}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$, whenever $\mathbf{X}^T\mathbf{X}$ has full rank.
  $(\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T)$

Risk decomposition for linear models

- Consider the statistical model $Y = f(X) + \varepsilon$, with f the true function
- ϵ_i's are independent with $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] = \sigma^2$
- Linear model : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in x of the form $\theta^T \widetilde{x}$ with $\widetilde{x} = (1, x^T)^T$, and $\theta = (\alpha, \beta^T)^T$. (denote $\widetilde{x}$ by x for simplicity)
- Bayes predictor h^* : for the squared loss : $h^*(x) = \mathbb{E}[Y|X = x] = f(x)$
- Let $\theta^* = (\alpha^*, {\beta^*}^T)^T$ be the optimal parameter. Then $h^*(x; \theta^*) = {\theta^*}^T x = f(x)$

Assume a fixed design, i.e. the x's are deterministic

- **Bayes Risk** $R^* = R(h^*) = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2|X = x] = \mathbb{E}[\epsilon^2|X = x] = \sigma^2$
- Risk for any **non-random** θ : $R(\theta) = \mathbb{E}[(Y - h(X;\theta))^2|X] = \sigma^2 + \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$:
- Excess risk of θ : $R(\theta) - R^* = \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ and $\|u\|^2_A = u^T A u$.

see proof in the next slide

- **ERM :** Solution : $\widehat{\theta}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$, whenever $\mathbf{X}^T\mathbf{X}$ has full rank. $(\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T)$

Risk decomposition for linear models

- Consider the statistical model $Y = f(X) + \varepsilon$, with f the true function
- ϵ_i's are independent with $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] = \sigma^2$
- Linear model : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in x of the form $\theta^T \widetilde{x}$ with $\widetilde{x} = (1, x^T)^T$, and $\theta = (\alpha, \beta^T)^T$. (denote $\widetilde{x}$ by x for simplicity)
- Bayes predictor h^* : for the squared loss : $h^*(x) = \mathbb{E}[Y|X = x] = f(x)$
- Let $\theta^* = (\alpha^*, \beta^{*T})^T$ be the optimal parameter. Then $h^*(x; \theta^*) = \theta^{*T} x = f(x)$

  Assume a fixed design, i.e. the x's are deterministic
- **Bayes Risk** $R^* = R(h^*) = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2|X = x] = \mathbb{E}[\epsilon^2|X = x] = \sigma^2$
- Risk for any **non-random** θ : $R(\theta) = \mathbb{E}[(Y - h(X;\theta))^2|X] = \sigma^2 + \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$:
- Excess risk of θ : $R(\theta) - R^* = \|\theta - \theta^*\|^2_{\widehat{\Sigma}}$

  where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ and $\|u\|^2_A = u^T A u$.

  see proof in the next slide
- **ERM :** Solution : $\widehat{\theta}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$, whenever $\mathbf{X}^T\mathbf{X}$ has full rank. ($\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$)

Risk of any h (under the square loss) :

$$\begin{aligned}
r(h(x)|X=x) &= \mathbb{E}_{Y|X}[\ell(Y,h(X))|X=x] = \mathbb{E}_{Y|X}[(Y-h(X))^2|X=x] \\
&= \mathbb{E}[(f(X)+\epsilon-h(X))^2] \\
&= \mathbb{E}[(f(x)-h(x))^2] + 2\mathbb{E}[\epsilon(f(x)-h(x))] + \mathbb{E}[\epsilon^2] \\
&= \underbrace{\mathbb{E}[(f(x)-h(x))^2]}_{\text{Bias-Variance}} + 2\underbrace{\mathbb{E}[\varepsilon]}_{0}\mathbb{E}[(f(x)-h(x))] + \underbrace{\mathbb{E}[\epsilon^2]}_{\text{Irreducible Error:}\sigma^2} \\
&= \text{Excess Risk} + \text{Bayes Risk}
\end{aligned}$$

(cont.)

- Proof

$$\begin{aligned}
R(\theta) &= \mathbb{E}_Y \mathbb{E}_X[(Y - h(X))^2 | x_1, \ldots, x_n] = \mathbb{E}_Y \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - h_\theta(x_i))^2 \Big| x_1, \ldots, x_n \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\varepsilon[(x_i^T \theta + \varepsilon_i - x_i^T \theta^*)^2 | x_i] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \underbrace{\mathbb{E}_\varepsilon[\varepsilon_i^2 | x_i]}_{\sigma^2} + (x_i^T (\theta - \theta^*))^2 + 2 \underbrace{\mathbb{E}_\varepsilon[\varepsilon_i | x_i]}_{0} x_i^T (\theta - \theta^*) \right\} \\
&= \underbrace{\sigma^2}_{R^*} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} [x_i^T (\theta - \theta^*)]^2}_{\text{Excess Risk}} \\
&= R^* + \|\theta - \theta^*\|_{\widehat{\Sigma}}^2 \text{ where } \widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \text{ and } \|u\|_A^2 = u^T A u.
\end{aligned}$$

(cont.)

Random θ (and fixed design) : $R(\theta) = R^* + \text{Var}(\theta) + (\text{Bias}(\theta, \theta^*))^2$

$$
\begin{aligned}
R(\theta) &= \mathbb{E}_Y \mathbb{E}_X [(Y - h(X))^2 | x_1, \ldots, x_n] = \mathbb{E}_Y \left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - h_\theta(x_i))^2 \Big| x_1, \ldots, x_n \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \underbrace{\mathbb{E}_\varepsilon [\varepsilon_i^2 | x_i]}_{\sigma^2} + \mathbb{E}_Y [(x_i^T(\theta - \theta^*))^2] + 2 \underbrace{\mathbb{E}_\varepsilon [\varepsilon_i | x_i]}_{0} \mathbb{E}_Y [x_i^T(\theta - \theta^*)] \right\} \\
&= \underbrace{\sigma^2}_{R^*} + \underbrace{\mathbb{E}_Y [\frac{1}{n} \sum_{i=1}^{n} [x_i^T(\theta - \theta^*)]^2]}_{\text{Excess Risk}} \\
&= R^* + \mathbb{E}_Y \|\theta - \theta^*\|_{\widehat{\Sigma}}^2 \\
&= R^* + \mathbb{E}\|\theta - \mathbb{E}[\theta] + \mathbb{E}[\theta] - \theta^*\|_{\widehat{\Sigma}}^2 \\
&= R^* + \mathbb{E}\left[\|\theta - \mathbb{E}[\theta]\|_{\widehat{\Sigma}}^2\right] + 2\mathbb{E}\left[(\theta - \mathbb{E}[\theta])\widehat{\Sigma}(\mathbb{E}[\theta] - \theta^*)\right] + \mathbb{E}\left[\|\mathbb{E}[\theta] - \theta^*\|_{\widehat{\Sigma}}^2\right] \\
&= R^* + \text{Var}(\theta) + 0 + (\text{Bias}(\theta, \theta^*))^2
\end{aligned}
$$

MSE and Ordinary Least Squares (OLS)

- Empirical Risk : Under the squared loss the empirical risk $R_n(h)$ is

$$\begin{aligned} R_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\|y_i - h(x_i;\theta)\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|y_i - \theta^T x_i)\|_2^2 \\ &= \frac{1}{n}\|\boldsymbol{Y} - \mathbf{X}\theta\|_2^2 = \frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\theta)^T(\boldsymbol{Y} - \mathbf{X}\theta) \end{aligned}$$

with $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$

- ERM : $\widehat{\theta}_n R_n(\theta) = \arg\min_{\theta\in\Theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ (whenever $\mathbf{X}^T\mathbf{X}$ is positive definite) is the **Ordinary Least Squares Estimator** of θ
- Calculation detail :

$$\begin{aligned} \nabla R_n(\widehat{\boldsymbol{\theta}}) &= \mathbf{0} \text{ (FOC)} \\ -2\mathbf{X}^T\boldsymbol{Y} + 2\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\theta}} &= \mathbf{0} \\ \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\theta}} &= \mathbf{X}^T\boldsymbol{Y} \quad \text{Normal equations} \\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\theta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^T\boldsymbol{Y} \\ \widehat{\boldsymbol{\theta}} &= (\boldsymbol{X}^T\mathbf{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \end{aligned}$$

$\boxed{\widehat{\boldsymbol{\theta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}}$ whenever $\mathbf{X}^T\mathbf{X}$ is invertible.

MSE and Ordinary Least Squares (OLS)

- Empirical Risk : Under the squared loss the empirical risk $R_n(h)$ is

$$\begin{aligned} R_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\|y_i - h(x_i;\theta)\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|y_i - \theta^T x_i)\|_2^2 \\ &= \frac{1}{n}\|\boldsymbol{Y} - \mathbf{X}\theta\|_2^2 = \frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\theta)^T(\boldsymbol{Y} - \mathbf{X}\theta) \\ & \quad \text{with } \mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \text{ and } \boldsymbol{Y} = (Y_1, \ldots, Y_n)^T \end{aligned}$$

- ERM : $\widehat{\theta}_n R_n(\theta) = \arg\min_{\theta\in\Theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ (whenever $\mathbf{X}^T\mathbf{X}$ is positive definite) is the **Ordinary Least Squares Estimator** of θ
- Calculation detail :

$$\begin{aligned} \nabla R_n(\widehat{\boldsymbol{\theta}}) &= \mathbf{0} \text{ (FOC)} \\ -2\mathbf{X}^T\boldsymbol{Y} + 2\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\theta}} &= \mathbf{0} \\ \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\theta}} &= \mathbf{X}^T\boldsymbol{Y} \quad \text{Normal equations} \\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\theta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^T\boldsymbol{Y} \\ \widehat{\boldsymbol{\theta}} &= (\boldsymbol{X}^T\mathbf{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \end{aligned}$$

$$\boxed{\widehat{\boldsymbol{\theta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}} \text{ whenever } \mathbf{X}^T\mathbf{X} \text{ is invertible.}$$

MSE and Ordinary Least Squares (OLS)

- Empirical Risk : Under the squared loss the empirical risk $R_n(h)$ is

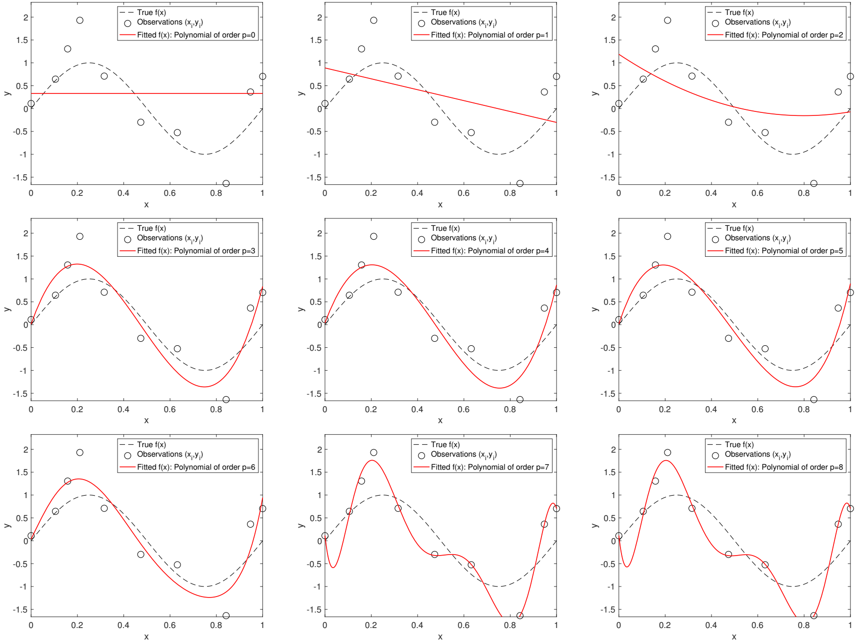
$$
\begin{aligned}
R_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\|y_i - h(x_i;\theta)\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|y_i - \theta^T x_i)\|_2^2 \\
&= \frac{1}{n}\|\boldsymbol{Y} - \mathbf{X}\theta\|_2^2 = \frac{1}{n}(\boldsymbol{Y} - \mathbf{X}\theta)^T(\boldsymbol{Y} - \mathbf{X}\theta) \\
&\quad \text{with } \mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \text{ and } \boldsymbol{Y} = (Y_1, \ldots, Y_n)^T
\end{aligned}
$$

- ERM : $\widehat{\theta}_n R_n(\theta) = \arg\min_{\theta\in\Theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}$ (whenever $\mathbf{X}^T\mathbf{X}$ is positive definite) is the **Ordinary Least Squares Estimator** of θ
- Calculation detail :

$$
\begin{aligned}
\nabla R_n(\widehat{\boldsymbol{\theta}}) &= \mathbf{0} \text{ (FOC)} \\
-2\mathbf{X}^T\boldsymbol{Y} + 2\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\theta}} &= \mathbf{0} \\
\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\theta}} &= \mathbf{X}^T\boldsymbol{Y} \quad \text{Normal equations} \\
(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\theta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^T\boldsymbol{Y} \\
\widehat{\boldsymbol{\theta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}
\end{aligned}
$$

$$
\boxed{\widehat{\boldsymbol{\theta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}} \text{ whenever } \mathbf{X}^T\mathbf{X} \text{ is invertible.}
$$

Figure on Bias-Variance Tradeoff/Underfitting and Overfitting



Setup to estimate the risk

Repeat :

- Fix an input x (or sample it from $P(X)$ in cas of random design)
- Sample the (true) target y from the conditional distribution $P(Y|x)$.
- Repeat :
  - Sample a training dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ i.i.d. from $P(x, Y)$.
  - Run the learning algorithm on $\mathcal{D}_n$ to obtain a predictor $\widehat{h}_n$.
  - Compute the prediction $\widehat{y} = \widehat{h}_n(x)$.
  - Compute the loss $\ell(\widehat{y}, y)$.
  - Average the losses.
- Average the losses.

  Notice : $\widehat{y}$ depends on $\mathcal{D}_n$, but y is sampled independently from $\mathcal{D}_n$.

Illustrations

Statistical learning of linear (polynomial) models

- True target function : $f(x) = 10 + 5x^2 \sin(2\pi x)$.
- The function is evaluated in the range $x \in [0, 1]$.
- Observations are generated as :

$$Y_i | x_i \sim f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

  - The dataset consists of $n = 20$ points.
  - The x_i values are either fixed or randomly sampled in $[0, 1]$.
  - The noise ε_i follows a Gaussian distribution :

$$\varepsilon_i \sim \mathcal{N}(\mu_e, \sigma_e^2), \quad \text{where } \mu_e = 0, \quad \sigma_e = 1.$$

- $N = 100$ replicates (samples) for averaging

Illustrations

Statistical learning of linear (polynomial) models

- True target function : $f(x) = 10 + 5x^2 \sin(2\pi x)$.
- The function is evaluated in the range $x \in [0, 1]$.
- Observations are generated as :

$$Y_i | x_i \sim f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

  - The dataset consists of $n = 20$ points.
  - The x_i values are either fixed or randomly sampled in $[0, 1]$.
  - The noise ε_i follows a Gaussian distribution :

$$\varepsilon_i \sim \mathcal{N}(\mu_e, \sigma_e^2), \quad \text{where } \mu_e = 0, \quad \sigma_e = 1.$$

- $N = 100$ replicates (samples) for averaging

Illustrations

Statistical learning of linear (polynomial) models

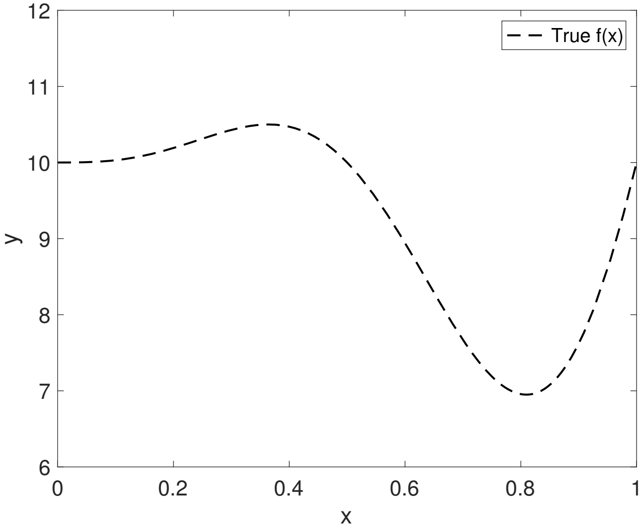
- True target function : $f(x) = 10 + 5x^2 \sin(2\pi x)$.
- The function is evaluated in the range $x \in [0, 1]$.
- Observations are generated as :

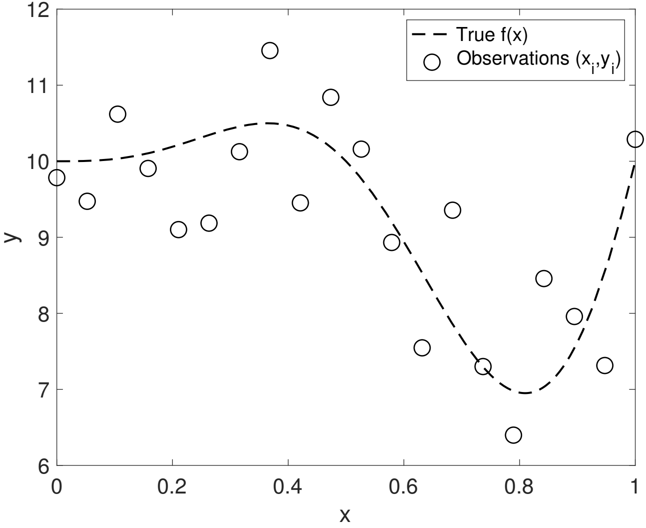
$$Y_i|x_i \sim f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

  - The dataset consists of $n = 20$ points.
  - The x_i values are either fixed or randomly sampled in $[0, 1]$.
  - The noise ε_i follows a Gaussian distribution :

$$\varepsilon_i \sim \mathcal{N}(\mu_e, \sigma_e^2), \quad \text{where } \mu_e = 0, \quad \sigma_e = 1.$$

- $N = 100$ replicates (samples) for averaging


True f(x)
12
11
10
9
8
7
6
y
0
0.2
0.4
0.6
0.8
1
x


12
11
10
9
8
7
6
y
0
0.2
0.4
0.6
0.8
1
x
True f(x)
Observations (xi,yi)

Polynomial regression

- Consider the class of polynomial models

$$\mathcal{H} = \{h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_p x^p\}$$

  the set of polynomials with p the polynomial degree

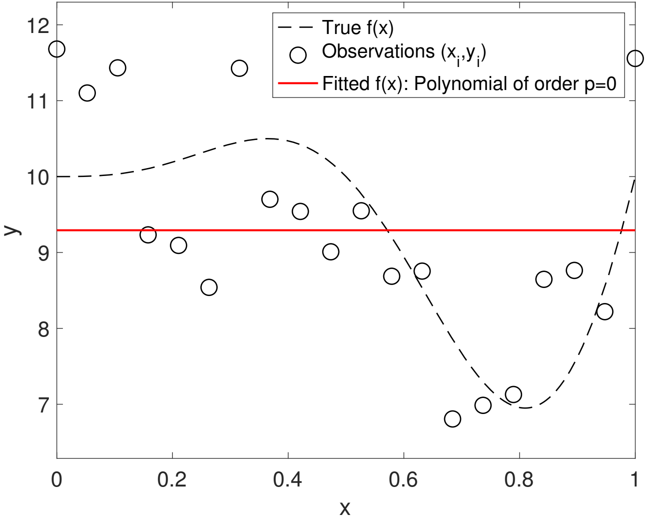
- p is ranging from 0 to 14
- ERM : $\widehat{\boldsymbol{\theta}}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$
  with
  - $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$,
  - $\boldsymbol{x}_i = (1, x_i, x_i^2, \ldots, x_i^p)^T$, and
  - $\boldsymbol{y} = (y_1, \ldots, y_n)^T$

Polynomial regression

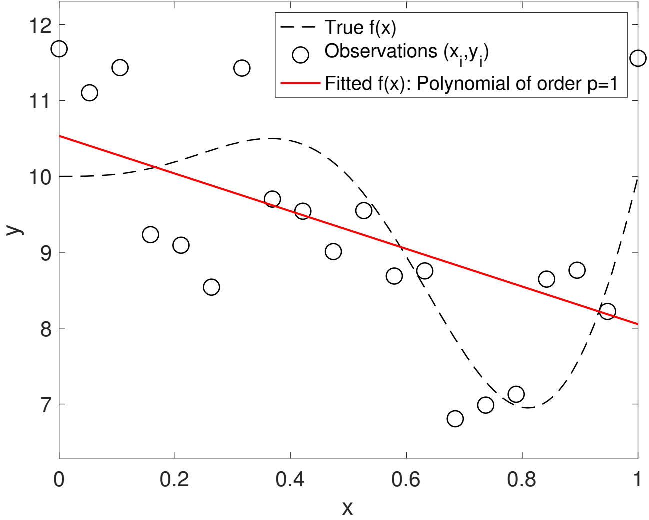
- Consider the class of polynomial models

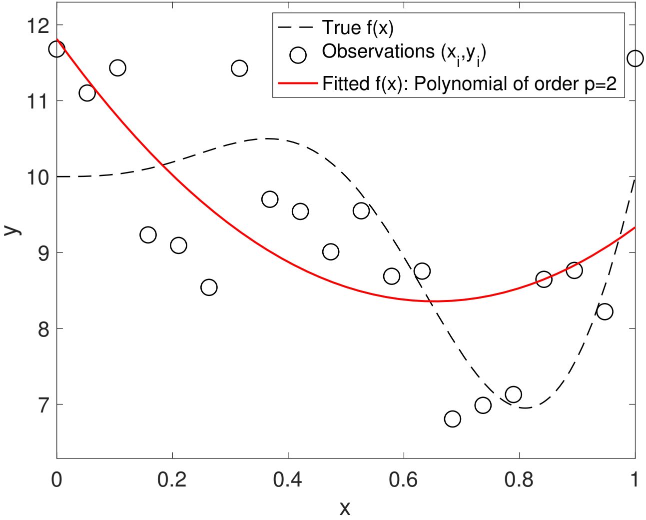
$$\mathcal{H} = \{h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_p x^p\}$$

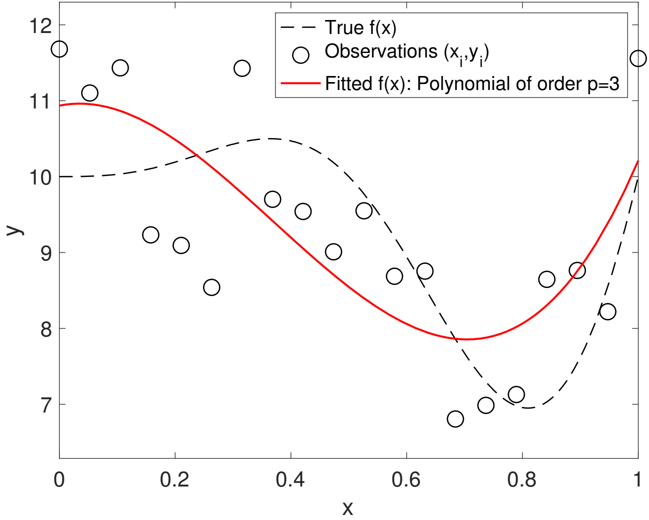
  the set of polynomials with p the polynomial degree
- p is ranging from 0 to 14
- ERM : $\widehat{\boldsymbol{\theta}}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$
  with
  - $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$,
  - $\boldsymbol{x}_i = (1, x_i, x_i^2, \ldots, x_i^p)^T$, and
  - $\boldsymbol{y} = (y_1, \ldots, y_n)^T$

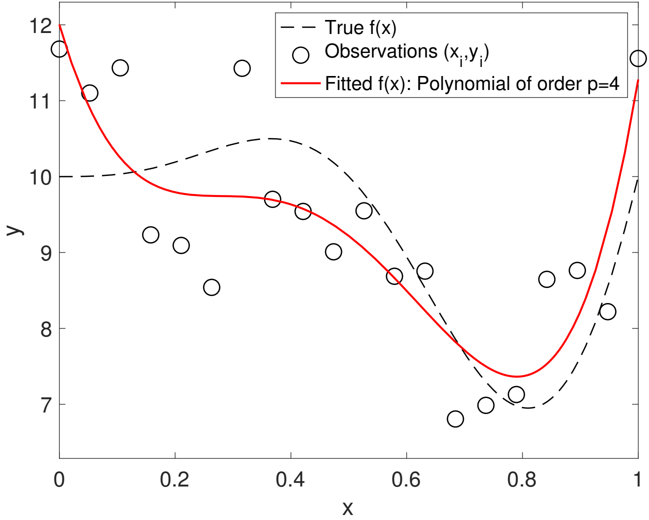


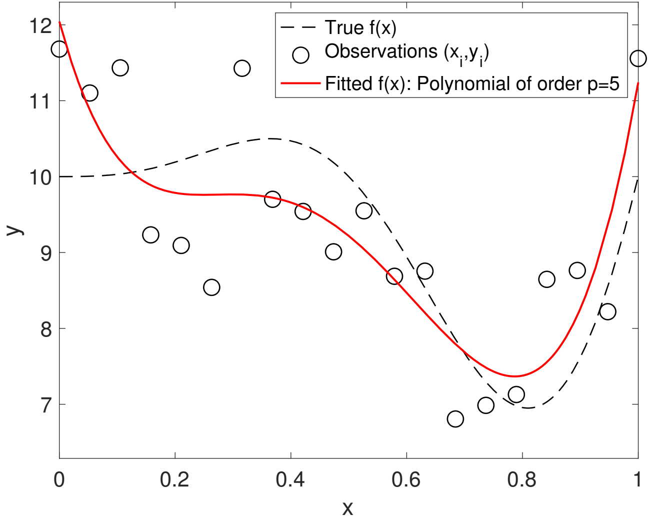
True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=0
y
x
12
11
10
9
8
7
0
0.2
0.4
0.6
0.8
1

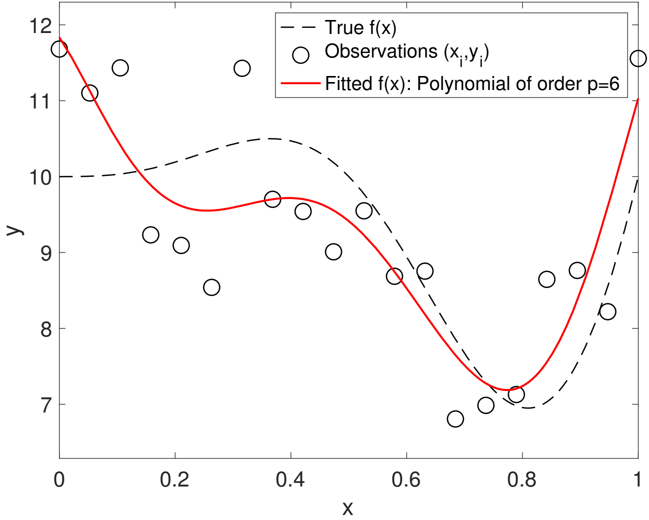

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=1
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

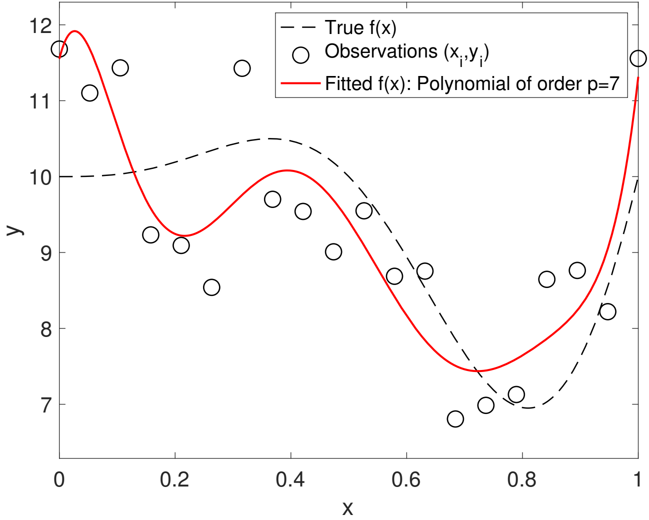

True f(x)
Observations (x_i, y_i)
Fitted f(x): Polynomial of order p=2
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

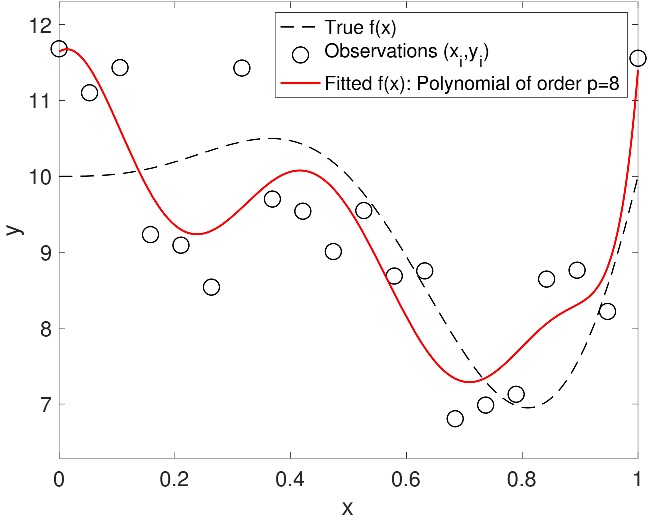

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=3
y
x
12
11
10
9
8
7
0
0.2
0.4
0.6
0.8
1

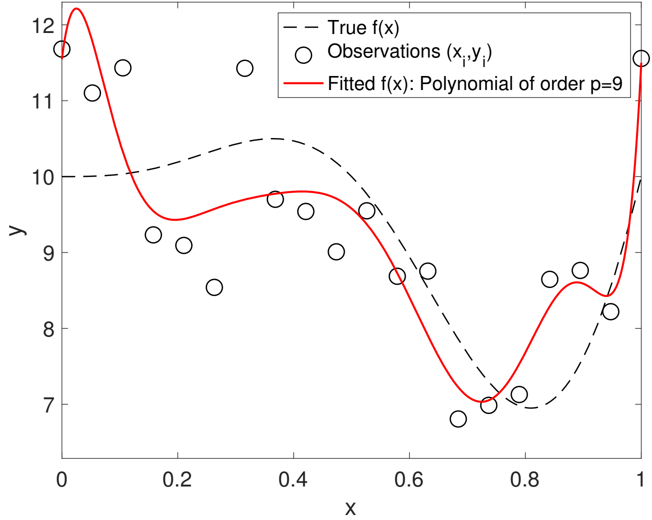

True f(x)
Observations (x_i, y_i)
Fitted f(x): Polynomial of order p=4
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

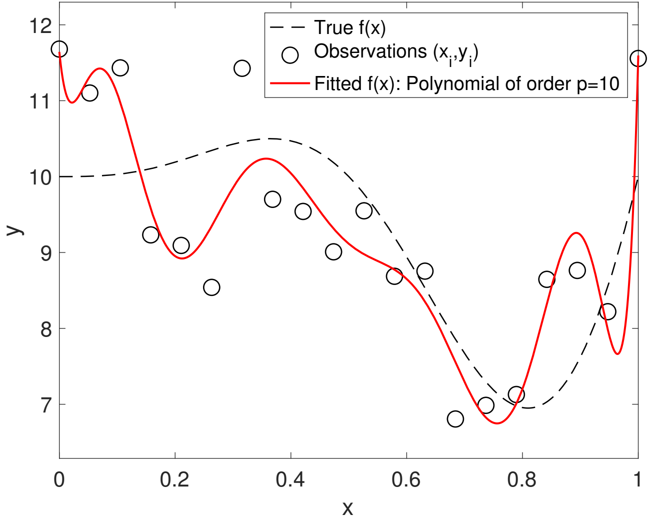

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=5
y
x
12
11
10
9
8
7
0
0.2
0.4
0.6
0.8
1

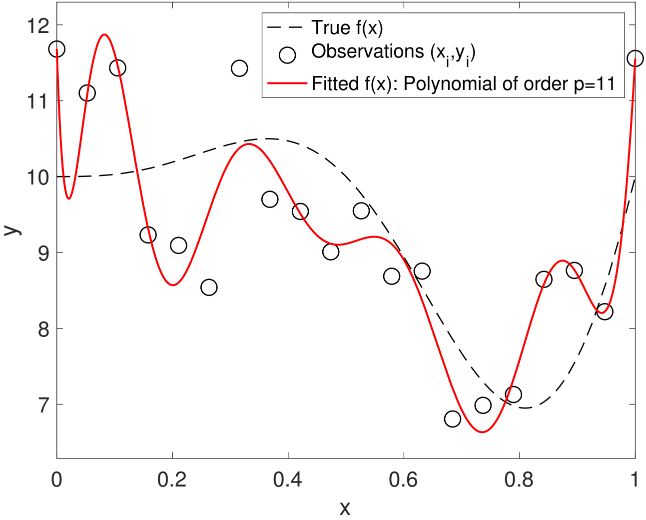

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=6
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

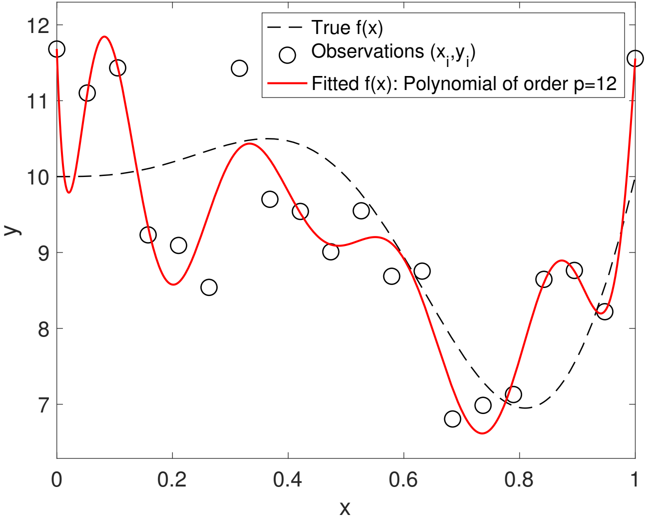

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=7
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

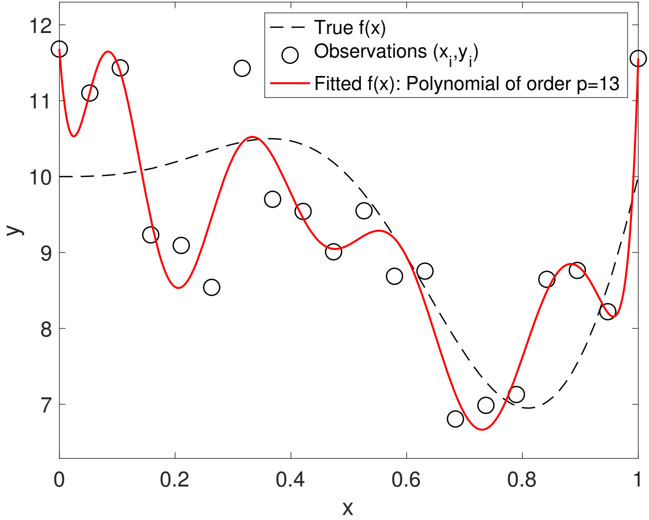

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=8
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

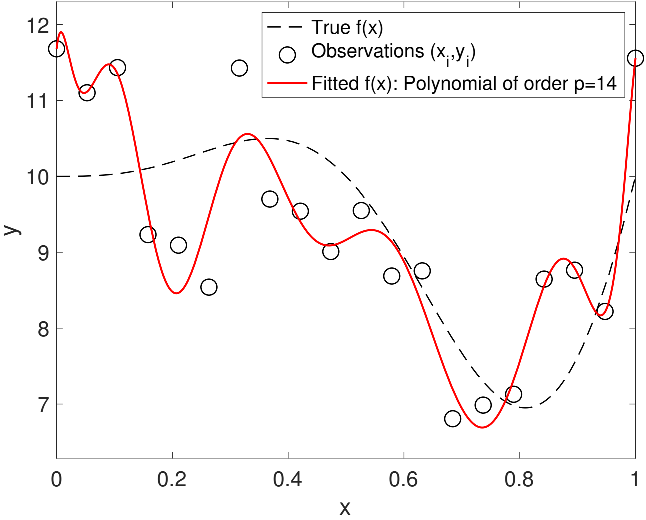

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=9
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

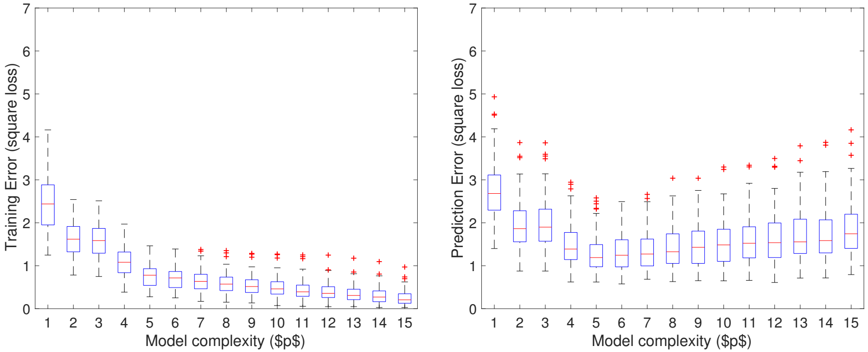

True f(x)
Observations (x_i, y_i)
Fitted f(x): Polynomial of order p=10
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

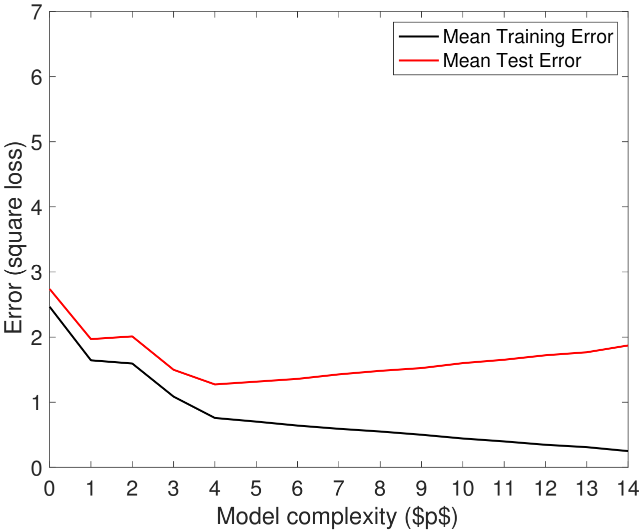

12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x
True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=11

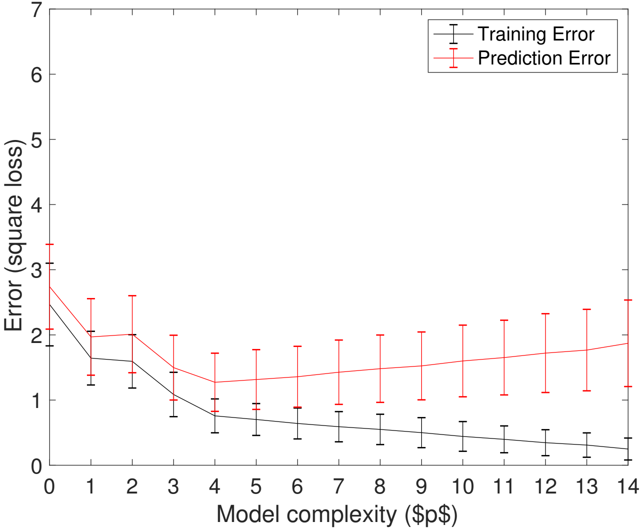

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=12
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

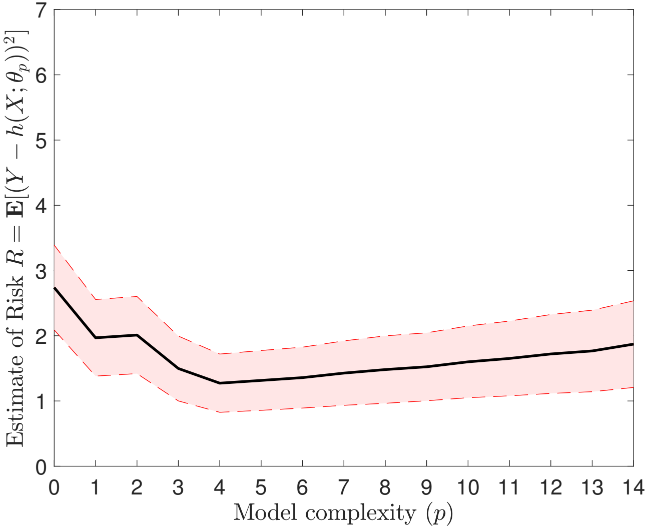

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=13
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

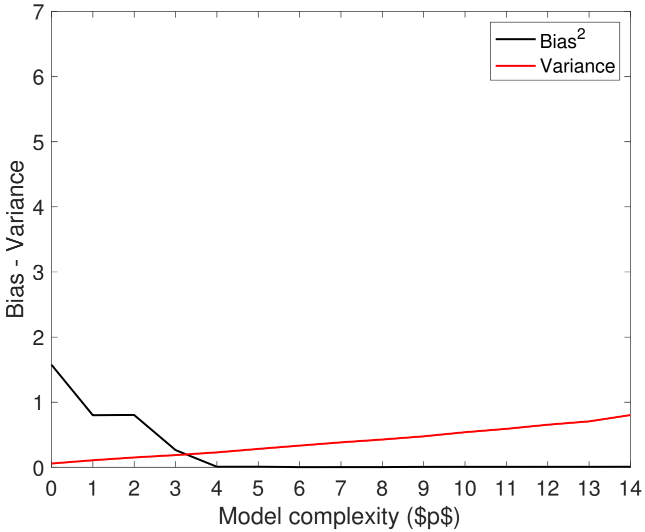

True f(x)
Observations (x_i,y_i)
Fitted f(x): Polynomial of order p=14
12
11
10
9
8
7
y
0
0.2
0.4
0.6
0.8
1
x

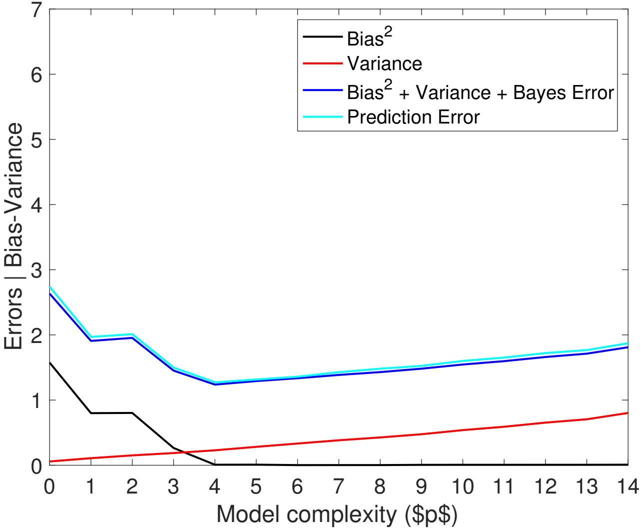

Training Error (square loss)
Model complexity (p)
Prediction Error (square loss)
Model complexity (p)
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

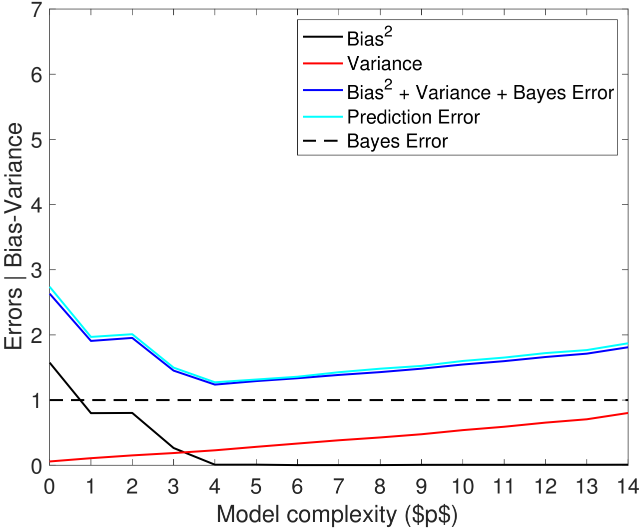

Mean Training Error
Mean Test Error
Error (square loss)
Model complexity (p)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

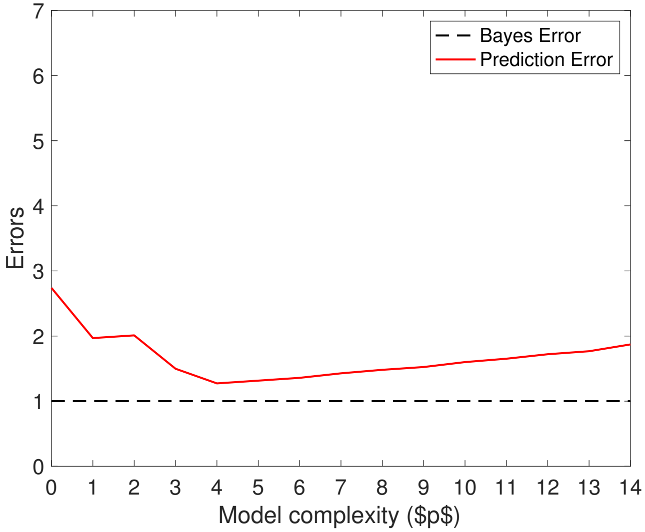

Training Error
Prediction Error
Error (square loss)
Model complexity (p)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14


Estimate of Risk $R = \mathbf{E}[(Y - h(X;\theta_p))^2]$
0
1
2
3
4
5
6
7
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Model complexity (p)


Bias^2
Variance
Bias - Variance
Model complexity (p)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14


Bias^2
Variance
Bias^2 + Variance + Bayes Error
Prediction Error
Errors | Bias-Variance
Model complexity (p)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14


Bias^2
Variance
Bias^2 + Variance + Bayes Error
Prediction Error
Bayes Error
Errors | Bias-Variance
Model complexity (p)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14


Bayes Error
Prediction Error
Errors
Model complexity (p)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14