Machine Learning

Bias-Variance Tradeoff

Alberto Maria Metelli and Francesco Trovò

Bias-Variance Dilemma

Known Process

To explicitly analyze the **variance** and the **bias** of a model we need to know the process generating the data:

$$t = \underbrace{f(x)}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{noise}} \qquad f(x) = 1 + \frac{1}{2}x + \frac{1}{10}x^2$$

- the input are x **uniformly** distributed in $[0, 5]$, i.e., $p(x) = \text{Uni}([0, 5])$
- the noise ε distribution $p(t|x)$ has $\mathbb{E}[\varepsilon|x] = 0$ and $\mathbb{V}\text{ar}[\varepsilon|x] = \sigma^2 = 0.7^2$

Known Process

To explicitly analyze the **variance** and the **bias** of a model we need to know the process generating the data:

$$t = \underbrace{f(x)}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{noise}} \qquad f(x) = 1 + \frac{1}{2}x + \frac{1}{10}x^2$$

- the input are x **uniformly** distributed in $[0, 5]$, i.e., $p(x) = \text{Uni}([0, 5])$
- the noise ε distribution $p(t|x)$ has $\mathbb{E}[\varepsilon|x] = 0$ and $\mathbb{V}\text{ar}[\varepsilon|x] = \sigma^2 = 0.7^2$

Known Process

To explicitly analyze the **variance** and the **bias** of a model we need to know the process generating the data:

$$t = \underbrace{f(x)}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{noise}} \qquad\qquad f(x) = 1 + \frac{1}{2}x + \frac{1}{10}x^2$$

- the input are x **uniformly** distributed in $[0, 5]$, i.e., $p(x) = \mathrm{Uni}([0, 5])$
- the noise ε distribution $p(t|x)$ has $\mathbb{E}[\varepsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\varepsilon|x] = \sigma^2 = 0.7^2$

Two-Model Dilemma

- Assume to approach the learning problem (we do not know the true model) using either one of the two following models:

$$\mathcal{H}_1: \quad y(x) = a + bx \quad \text{linear}$$
$$\mathcal{H}_2: \quad y(x) = a + bx + cx^2 \quad \text{quadratic}$$

- Hence, $\mathcal{H}_1 \subset \mathcal{H}_2$
- They can be both regarded as **linear models**: $y(x) = \mathbf{w}^\top \phi(x)$ with:

$$\mathcal{H}_1: \quad \phi(x) = (1, x)^\top \quad \text{and} \quad \mathbf{w} = (a, b)^\top$$
$$\mathcal{H}_2: \quad \phi(x) = (1, x, x^2)^\top \quad \text{and} \quad \mathbf{w} = (a, b, c)^\top$$

Two-Model Dilemma

- Assume to approach the learning problem (we do not know the true model) using either one of the two following models:

$$
\begin{aligned}
\mathcal{H}_1: &\quad y(x) = a + bx &\quad \text{linear}\\
\mathcal{H}_2: &\quad y(x) = a + bx + cx^2 &\quad \text{quadratic}
\end{aligned}
$$

- Hence, $\mathcal{H}_1 \subset \mathcal{H}_2$
- They can be both regarded as **linear models**: $y(x) = \mathbf{w}^\top \phi(x)$ with:

$$
\begin{aligned}
\mathcal{H}_1: &\quad \phi(x) = (1, x)^\top \quad \text{and} \quad \mathbf{w} = (a, b)^\top\\
\mathcal{H}_2: &\quad \phi(x) = (1, x, x^2)^\top \quad \text{and} \quad \mathbf{w} = (a, b, c)^\top
\end{aligned}
$$

Two-Model Dilemma

- Assume to approach the learning problem (we do not know the true model) using either one of the two following models:

$$
\begin{aligned}
\mathcal{H}_1 :& \quad y(x) = a + bx && \text{linear}\\
\mathcal{H}_2 :& \quad y(x) = a + bx + cx^2 && \text{quadratic}
\end{aligned}
$$

- Hence, $\mathcal{H}_1 \subset \mathcal{H}_2$
- They can be both regarded as **linear models**: $y(x) = \mathbf{w}^\top \phi(x)$ with:

$$
\begin{aligned}
\mathcal{H}_1 :& \quad \phi(x) = (1, x)^\top \quad \text{and} \quad \mathbf{w} = (a, b)^\top\\
\mathcal{H}_2 :& \quad \phi(x) = (1, x, x^2)^\top \quad \text{and} \quad \mathbf{w} = (a, b, c)^\top
\end{aligned}
$$

Population Risk Minimization (PRM)

Assumption: we know $p(x, t)$

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Population** risk minimization (PRM):

$$
y^* \in \arg\min_{y \in \mathcal{H}} \mathbb{E}_{t,x}[(t - y(x))^2] = \int p(x,t)(t - y(x))^2 \, \mathrm{d}x \, \mathrm{d}t
$$

$$
\overset{t=f(x)+\varepsilon}{=} \int p(x)(f(x) - y(x))^2 \, \mathrm{d}x
$$

We can solve this problem only if we know $p(x, t)$!

Population Risk Minimization (PRM)

Assumption: we know $p(x,t)$

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Population** risk minimization (PRM):

$$y^* \in \arg\min_{y \in \mathcal{H}} \mathbb{E}_{t,x}[(t - y(x))^2] = \int p(x,t)(t - y(x))^2 \, \mathrm{d}x \, \mathrm{d}t$$
$$\stackrel{t=f(x)+\varepsilon}{=} \int p(x)(f(x) - y(x))^2 \, \mathrm{d}x$$

We can solve this problem only if we know $p(x,t)$!

Population Risk Minimization (PRM)

Assumption: we know $p(x, t)$

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Population** risk minimization (PRM):

$$y^* \in \arg\min_{y \in \mathcal{H}} \mathbb{E}_{t,x}[(t - y(x))^2] = \int p(x,t)(t - y(x))^2 \,\mathrm{d}x \,\mathrm{d}t$$

$$\stackrel{t=f(x)+\varepsilon}{=} \int p(x)(f(x) - y(x))^2 \,\mathrm{d}x$$

We can solve this problem only if we know $p(x, t)$!

Population Risk Minimization (PRM)

Assumption: we know $p(x, t)$

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Population** risk minimization (PRM):

$$y^* \in \arg\min_{y \in \mathcal{H}} \mathbb{E}_{t,x}[(t - y(x))^2] = \int p(x,t)(t - y(x))^2 \, \mathrm{d}x \, \mathrm{d}t$$

$$\overset{t=f(x)+\varepsilon}{=} \int p(x)(f(x) - y(x))^2 \, \mathrm{d}x$$

We can solve this problem only if we know $p(x, t)$!

Population Risk Minimization (PRM)

Assumption: we know $p(x, t)$

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Population** risk minimization (PRM):

$$y^* \in \arg\min_{y \in \mathcal{H}} \mathbb{E}_{t,x}[(t - y(x))^2] = \int p(x,t)(t - y(x))^2 \, \mathrm{d}x \, \mathrm{d}t$$

$$\stackrel{t=f(x)+\varepsilon}{=} \int p(x)(f(x) - y(x))^2 \, \mathrm{d}x$$

We can solve this problem only if we know $p(x, t)$!

Population Risk Minimization

If the real model is known we can compute the optimal model for the two hypothesis space:

$$\mathcal{H}_1: \qquad \arg\min_{(a,b)\in\mathbb{R}^2} \int_0^5 \frac{1}{5}(f(x) - a - bx)^2 \, \mathrm{d}x = \left(\frac{7}{12},\ 1\right)^\top$$

$$\mathcal{H}_2: \qquad \arg\min_{(a,b,c)\in\mathbb{R}^3} \int_0^5 \frac{1}{5}(f(x) - a - bx - cx^2)^2 \, \mathrm{d}x = \left(1,\ \frac{1}{2},\ \frac{1}{10}\right)^\top$$

Empirical Risk Minimization (ERM)

Assumption: we **do not know** $p(x, t)$ but we have a **training dataset** $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^N$ i.i.d. from p

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Empirical** risk minimization (ERM):

$$\widehat{y} \in \arg\min_{y \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} (t_n - y(x_n))^2$$

$\widehat{y}$ is a **random variable** depending on the dataset $\mathcal{D}$!

Empirical Risk Minimization (ERM)

Assumption: we **do not know** $p(x,t)$ but we have a **training dataset** $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^{N}$ i.i.d. from p

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Empirical** risk minimization (ERM):

$$\widehat{y} \in \arg\min_{y \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} (t_n - y(x_n))^2$$

$\widehat{y}$ is a **random variable** depending on the dataset $\mathcal{D}$!

Empirical Risk Minimization (ERM)

Assumption: we **do not know** $p(x,t)$ but we have a **training dataset** $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^N$ i.i.d. from p

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Empirical** risk minimization (ERM):

$$\widehat{y} \in \arg\min_{y \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} (t_n - y(x_n))^2$$

$\widehat{y}$ is a **random variable** depending on the dataset $\mathcal{D}$!

Empirical Risk Minimization (ERM)

Assumption: we **do not know** $p(x,t)$ but we have a **training dataset** $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^N$ i.i.d. from p

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Empirical** risk minimization (ERM):

$$\widehat{y} \in \arg\min_{y \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} (t_n - y(x_n))^2$$

$\widehat{y}$ is a **random variable** depending on the dataset $\mathcal{D}$!

Empirical Risk Minimization (ERM)

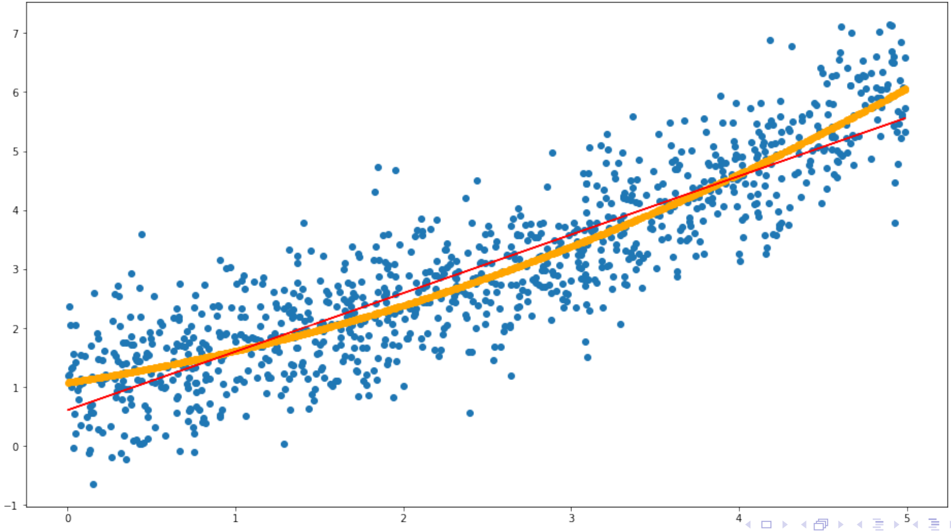
Assumption: we **do not know** $p(x, t)$ but we have a **training dataset** $\mathcal{D} = \{(x_n, t_n)\}_{n=1}^{N}$ i.i.d. from p

- Hypothesis space: $y(x) \in \mathcal{H}$
- Loss function: squared loss function $(t - y(x))^2$
- **Empirical** risk minimization (ERM):

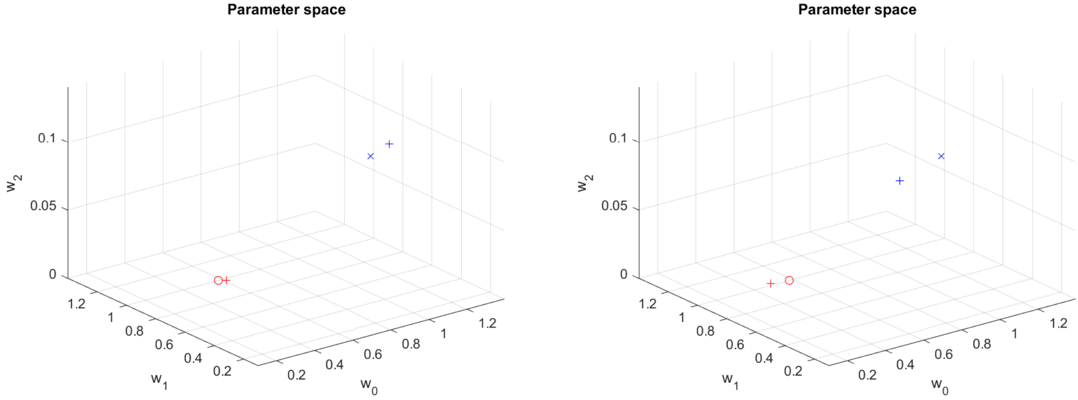
$$\widehat{y} \in \arg\min_{y \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} (t_n - y(x_n))^2$$

$\widehat{y}$ is a **random variable** depending on the dataset $\mathcal{D}$!

Visualizing the Fitting



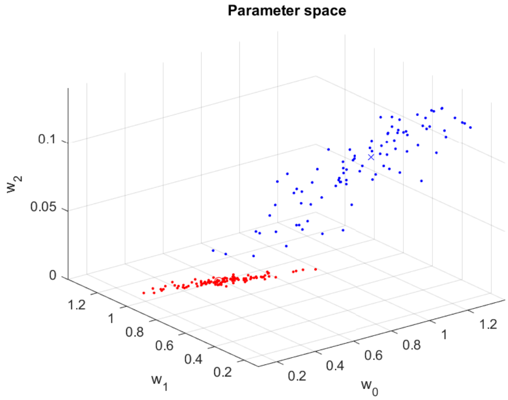
Optimal Parameters and Realized Parameters

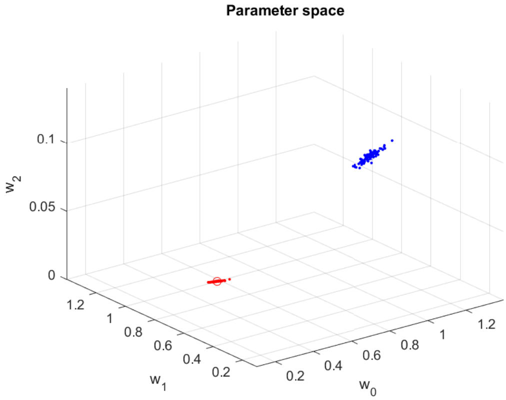


The blue $\times$ is the best model in $\mathcal{H}_2$ and the red $\circ$ is the best model in $\mathcal{H}_1 \rightarrow$ PRM
The $+$ are the optimal parameters for two realizations of the dataset $\mathcal{D}$ $(N = 1000) \rightarrow$ ERM

Visualization of Bias and Variance

If we repeat the ERM for multiple times (generation of 100 independent dataset) with different number of samples ($N = 100$ on the left and $N = 10000$ on the right)





Bias-Variance Decomposition

- $t = f(x) + \epsilon$ where $\mathbb{E}[\epsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\epsilon|x] = \sigma^2$
- $\widehat{y}(x)$ is the **empirical risk minimizer** over the training dataset $\mathcal{D}$
- x is a **fixed** (unseen point)

$$\underbrace{\mathbb{E}_{\mathcal{D},t}[(t - \widehat{y}(x))^2]}_{\text{error}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{V}\mathrm{ar}_{\mathcal{D}}[\widehat{y}(x)]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}}[f(x) - \widehat{y}(x)]^2}_{\text{bias}^2}$$

- **Error** in expectation taken w.r.t. the training dataset $\mathcal{D}$ and the target t
- **Irreducible error**
- **Variance** reduces with the number of samples $N = |\mathcal{D}|$
- **Bias** depends on the hypothesis space $\mathcal{H}$

Bias-Variance Decomposition

- $t = f(x) + \epsilon$ where $\mathbb{E}[\epsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\epsilon|x] = \sigma^2$
- $\widehat{y}(x)$ is the **empirical risk minimizer** over the training dataset $\mathcal{D}$
- x is a **fixed** (unseen point)

$$\underbrace{\mathbb{E}_{\mathcal{D},t}[(t-\widehat{y}(x))^2]}_{\text{error}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{V}\mathrm{ar}_{\mathcal{D}}[\widehat{y}(x)]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}}[f(x)-\widehat{y}(x)]^2}_{\text{bias}^2}$$

- **Error** in expectation taken w.r.t. the training dataset $\mathcal{D}$ and the target t
- **Irreducible error**
- **Variance** reduces with the number of samples $N = |\mathcal{D}|$
- **Bias** depends on the hypothesis space $\mathcal{H}$

Bias-Variance Decomposition

- $t = f(x) + \epsilon$ where $\mathbb{E}[\epsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\epsilon|x] = \sigma^2$
- $\widehat{y}(x)$ is the **empirical risk minimizer** over the training dataset $\mathcal{D}$
- x is a **fixed** (unseen point)

$$\underbrace{\mathbb{E}_{\mathcal{D},t}[(t - \widehat{y}(x))^2]}_{\text{error}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{V}\mathrm{ar}_{\mathcal{D}}[\widehat{y}(x)]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}}[f(x) - \widehat{y}(x)]^2}_{\text{bias}^2}$$

- **Error** in expectation taken w.r.t. the training dataset $\mathcal{D}$ and the target t
- **Irreducible error**
- **Variance** reduces with the number of samples $N = |\mathcal{D}|$
- **Bias** depends on the hypothesis space $\mathcal{H}$

Bias-Variance Decomposition

- $t = f(x) + \epsilon$ where $\mathbb{E}[\epsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\epsilon|x] = \sigma^2$
- $\widehat{y}(x)$ is the **empirical risk minimizer** over the training dataset $\mathcal{D}$
- x is a **fixed** (unseen point)

$$\underbrace{\mathbb{E}_{\mathcal{D},t}[(t - \widehat{y}(x))^2]}_{\text{error}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{V}\mathrm{ar}_{\mathcal{D}}[\widehat{y}(x)]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}}[f(x) - \widehat{y}(x)]^2}_{\text{bias}^2}$$

- **Error** in expectation taken w.r.t. the training dataset $\mathcal{D}$ and the target t
- **Irreducible error**
- **Variance** reduces with the number of samples $N = |\mathcal{D}|$
- **Bias** depends on the hypothesis space $\mathcal{H}$

Bias-Variance Decomposition

- $t = f(x) + \epsilon$ where $\mathbb{E}[\epsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\epsilon|x] = \sigma^2$
- $\widehat{y}(x)$ is the **empirical risk minimizer** over the training dataset $\mathcal{D}$
- x is a **fixed** (unseen point)

$$\underbrace{\mathbb{E}_{\mathcal{D},t}[(t - \widehat{y}(x))^2]}_{\text{error}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{V}\mathrm{ar}_{\mathcal{D}}[\widehat{y}(x)]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}}[f(x) - \widehat{y}(x)]^2}_{\text{bias}^2}$$

- **Error** in expectation taken w.r.t. the training dataset $\mathcal{D}$ and the target t
- **Irreducible error**
- **Variance** reduces with the number of samples $N = |\mathcal{D}|$
- **Bias** depends on the hypothesis space $\mathcal{H}$

Bias-Variance Decomposition

- $t = f(x) + \epsilon$ where $\mathbb{E}[\epsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\epsilon|x] = \sigma^2$
- $\widehat{y}(x)$ is the **empirical risk minimizer** over the training dataset $\mathcal{D}$
- x is a **fixed** (unseen point)

$$\underbrace{\mathbb{E}_{\mathcal{D},t}[(t - \widehat{y}(x))^2]}_{\text{error}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{V}\mathrm{ar}_{\mathcal{D}}[\widehat{y}(x)]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}}[f(x) - \widehat{y}(x)]^2}_{\text{bias}^2}$$

- **Error** in expectation taken w.r.t. the training dataset $\mathcal{D}$ and the target t
- **Irreducible error**
- **Variance** reduces with the number of samples $N = |\mathcal{D}|$
- **Bias** depends on the hypothesis space $\mathcal{H}$

Bias-Variance Decomposition

- $t = f(x) + \epsilon$ where $\mathbb{E}[\epsilon|x] = 0$ and $\mathbb{V}\mathrm{ar}[\epsilon|x] = \sigma^2$
- $\widehat{y}(x)$ is the **empirical risk minimizer** over the training dataset $\mathcal{D}$
- x is a **fixed** (unseen point)

$$\underbrace{\mathbb{E}_{\mathcal{D},t}[(t - \widehat{y}(x))^2]}_{\text{error}} = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\mathbb{V}\mathrm{ar}_{\mathcal{D}}[\widehat{y}(x)]}_{\text{variance}} + \underbrace{\mathbb{E}_{\mathcal{D}}[f(x) - \widehat{y}(x)]^2}_{\text{bias}^2}$$

- **Error** in expectation taken w.r.t. the training dataset $\mathcal{D}$ and the target t
- **Irreducible error**
- **Variance** reduces with the number of samples $N = |\mathcal{D}|$
- **Bias** depends on the hypothesis space $\mathcal{H}$

Computation of Bias and Variance

```
Linear error: 0.46867
Linear bias: 0.03613
Linear variance: 0.00011514
Linear sigma: 0.43242
Quadratic error: 0.42146
Quadratic bias: 1.412e-06
Quadratic variance: 0.00014674
Quadratic sigma: 0.42131
```

All the considerations holds on average, therefore there might be realizations for which the Bias and Variance of different models might not be coherent with what we saw.

Bias-Variance Tradeoff

Model Selection Problem

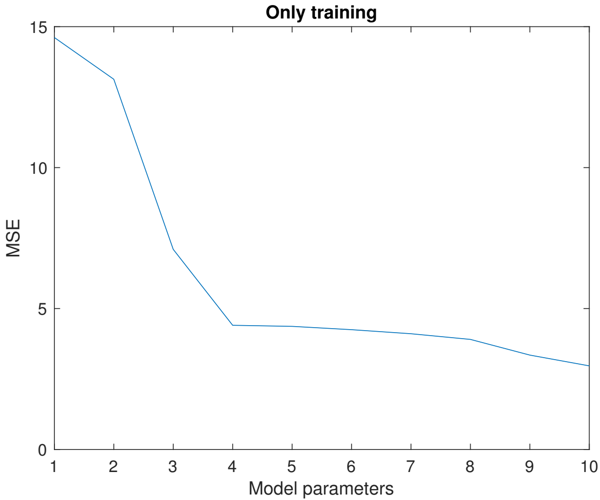
In real scenarios, we do not know the real model, so we should **select** the correct one among a set of models.

Consider the possible solutions for a regression problem:

- Hypothesis space: $y(x; \mathbf{w}) = f(x, \mathbf{w}) = \sum_{k=0}^{o} x^k w_k$
- Loss function: $\frac{1}{N} \sum_{(x,t) \in \mathcal{D}} (y(x_n; \mathbf{w}) - t_n)^2$ on a dataset $\mathcal{D}$
- Optimization method: Least Square (LS)

The order o and other parameters which should be chosen before performing the training phase are usually addressed as *hyperparameters*

Limits of Using the Training error



Why?

- The quality of a (fixed) model $\mathbf{w}$ is represented by the **expected MSE**:

$$\mathrm{MSE}(\mathbf{w}) := \mathbb{E}_{\mathbf{x},t}[(y(\mathbf{x};\mathbf{w}) - t)^2]$$

- We **train** on $\mathcal{D}_{\text{train}}$ with $N = |\mathcal{D}_{\text{train}}|$ by minimizing the **empirical MSE** (i.e., empirical risk minimization):

$$\widehat{\mathbf{w}} \in \arg \min_{\mathbf{w} \in \mathbb{R}^{o+1}} \widehat{\mathrm{MSE}}_{\text{train}}(\mathbf{w}) := \frac{1}{N} \sum_{(\mathbf{x},t) \in \mathcal{D}_{\text{train}}} (y(\mathbf{x};\mathbf{w}) - t)^2$$

  - $\widehat{\mathbf{w}}$ is **statistically dependent** on $\mathcal{D}_{\text{train}}$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ **is not a good estimate of** $\mathrm{MSE}(\widehat{\mathbf{w}})$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ cannot be used for **evaluating** the performance of $y(\cdot;\widehat{\mathbf{w}})$ nor for **selecting** among different models

Why?

- The quality of a (fixed) model $\mathbf{w}$ is represented by the **expected MSE**:

$$\mathrm{MSE}(\mathbf{w}) := \mathbb{E}_{\mathbf{x},t}[(y(\mathbf{x};\mathbf{w}) - t)^2]$$

- We **train** on $\mathcal{D}_{\text{train}}$ with $N = |\mathcal{D}_{\text{train}}|$ by minimizing the **empirical MSE** (i.e., empirical risk minimization):

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w} \in \mathbb{R}^{o+1}} \widehat{\mathrm{MSE}}_{\text{train}}(\mathbf{w}) := \frac{1}{N} \sum_{(\mathbf{x},t) \in \mathcal{D}_{\text{train}}} (y(\mathbf{x};\mathbf{w}) - t)^2$$

  - $\widehat{\mathbf{w}}$ is **statistically dependent** on $\mathcal{D}_{\text{train}}$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ **is not a good estimate of** $\mathrm{MSE}(\widehat{\mathbf{w}})$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ cannot be used for **evaluating** the performance of $y(\cdot;\widehat{\mathbf{w}})$ nor for **selecting** among different models

Why?

- The quality of a (fixed) model $\mathbf{w}$ is represented by the **expected MSE**:

$$\mathrm{MSE}(\mathbf{w}) := \mathbb{E}_{\mathbf{x},t}[(y(\mathbf{x};\mathbf{w}) - t)^2]$$

- We **train** on $\mathcal{D}_{\text{train}}$ with $N = |\mathcal{D}_{\text{train}}|$ by minimizing the **empirical MSE** (i.e., empirical risk minimization):

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathbb{R}^{o+1}} \widehat{\mathrm{MSE}}_{\text{train}}(\mathbf{w}) := \frac{1}{N}\sum_{(\mathbf{x},t)\in\mathcal{D}_{\text{train}}} (y(\mathbf{x};\mathbf{w}) - t)^2$$

  - $\widehat{\mathbf{w}}$ is **statistically dependent** on $\mathcal{D}_{\text{train}}$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ **is not a good estimate of** $\mathrm{MSE}(\widehat{\mathbf{w}})$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ cannot be used for **evaluating** the performance of $y(\cdot;\widehat{\mathbf{w}})$ nor for **selecting** among different models

Why?

- The quality of a (fixed) model $\mathbf{w}$ is represented by the **expected MSE**:

$$\mathrm{MSE}(\mathbf{w}) := \mathbb{E}_{\mathbf{x},t}[(y(\mathbf{x};\mathbf{w}) - t)^2]$$

- We **train** on $\mathcal{D}_{\text{train}}$ with $N = |\mathcal{D}_{\text{train}}|$ by minimizing the **empirical MSE** (i.e., empirical risk minimization):

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathbb{R}^{o+1}} \widehat{\mathrm{MSE}}_{\text{train}}(\mathbf{w}) := \frac{1}{N}\sum_{(\mathbf{x},t)\in\mathcal{D}_{\text{train}}} (y(\mathbf{x};\mathbf{w}) - t)^2$$

  - $\widehat{\mathbf{w}}$ is **statistically dependent** on $\mathcal{D}_{\text{train}}$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ **is not a good estimate of** $\mathrm{MSE}(\widehat{\mathbf{w}})$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ cannot be used for **evaluating** the performance of $y(\cdot;\widehat{\mathbf{w}})$ nor for **selecting** among different models

Why?

- The quality of a (fixed) model $\mathbf{w}$ is represented by the **expected MSE**:

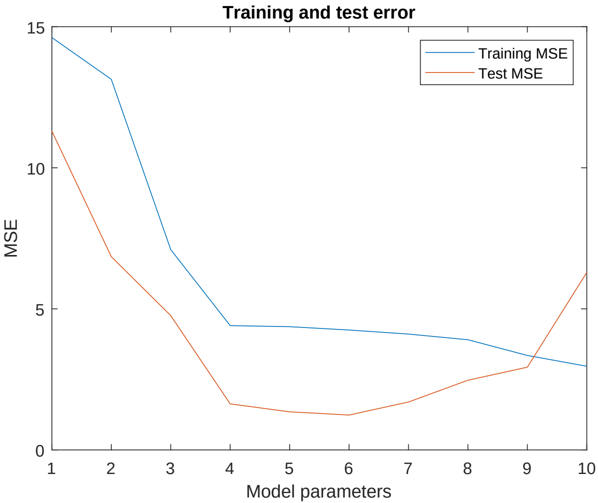
$$\mathrm{MSE}(\mathbf{w}) := \mathbb{E}_{\mathbf{x},t}[(y(\mathbf{x};\mathbf{w}) - t)^2]$$

- We **train** on $\mathcal{D}_{\text{train}}$ with $N = |\mathcal{D}_{\text{train}}|$ by minimizing the **empirical MSE** (i.e., empirical risk minimization):

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathbb{R}^{o+1}} \widehat{\mathrm{MSE}}_{\text{train}}(\mathbf{w}) := \frac{1}{N} \sum_{(\mathbf{x},t)\in\mathcal{D}_{\text{train}}} (y(\mathbf{x};\mathbf{w}) - t)^2$$

  - $\widehat{\mathbf{w}}$ is **statistically dependent** on $\mathcal{D}_{\text{train}}$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ **is not a good estimate of** $\mathrm{MSE}(\widehat{\mathbf{w}})$
  - $\widehat{\mathrm{MSE}}_{\text{train}}(\widehat{\mathbf{w}})$ cannot be used for **evaluating** the performance of $y(\cdot;\widehat{\mathbf{w}})$ nor for **selecting** among different models

Limits of Using the Training error



Validation

- Training set $\mathcal{D}_{\text{train}}$, i.e., the data we will use to **learn** the model parameters
- Validation set $\mathcal{D}_{\text{vali}}$, i.e., the data we will use to **select** the model
- Test set $\mathcal{D}_{\text{test}}$, i.e., the data we will use to **evaluate** the performance of our model

Usually, we use a split proportional to 50%-25%-25% for the three sets

Validation

- Training set $\mathcal{D}_{\text{train}}$, i.e., the data we will use to **learn** the model parameters
- Validation set $\mathcal{D}_{\text{vali}}$, i.e., the data we will use to **select** the model
- Test set $\mathcal{D}_{\text{test}}$, i.e., the data we will use to **evaluate** the performance of our model

Usually, we use a split proportional to 50%-25%-25% for the three sets

Validation

- Training set $\mathcal{D}_{\text{train}}$, i.e., the data we will use to **learn** the model parameters
- Validation set $\mathcal{D}_{\text{vali}}$, i.e., the data we will use to **select** the model
- Test set $\mathcal{D}_{\text{test}}$, i.e., the data we will use to **evaluate** the performance of our model

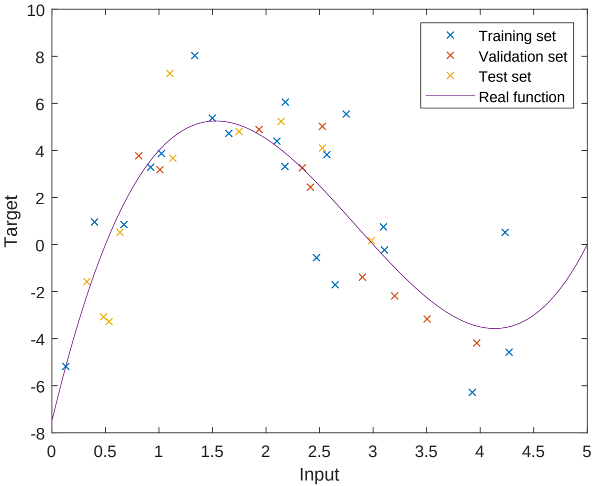
Usually, we use a split proportional to 50%-25%-25% for the three sets

Validation

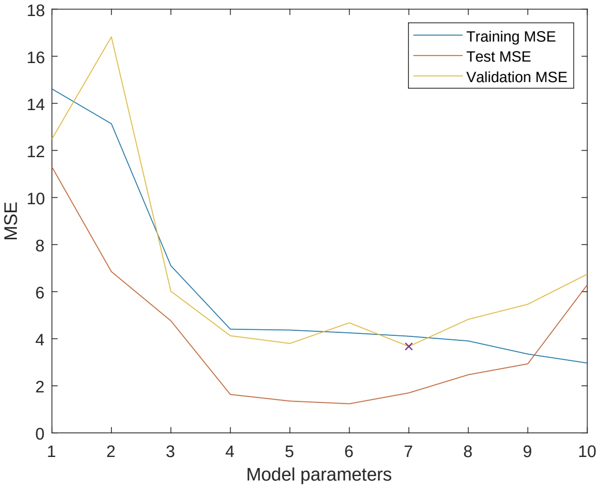
- Training set $\mathcal{D}_{\text{train}}$, i.e., the data we will use to **learn** the model parameters
- Validation set $\mathcal{D}_{\text{vali}}$, i.e., the data we will use to **select** the model
- Test set $\mathcal{D}_{\text{test}}$, i.e., the data we will use to **evaluate** the performance of our model

Usually, we use a split proportional to 50%-25%-25% for the three sets

Dataset Generated



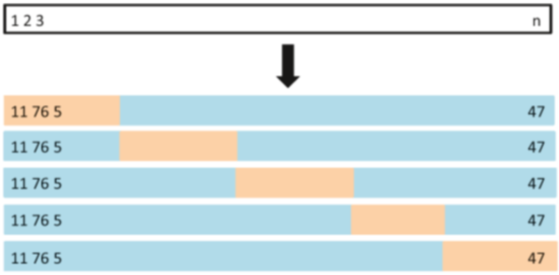
Validation Results



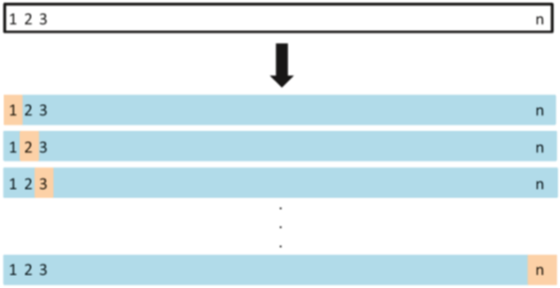
LOO and Crossvalidation

This way we reduce the amount of samples we could use for training of 33%, which could compromise the analysis since the training has been performed with a significantly smaller dataset

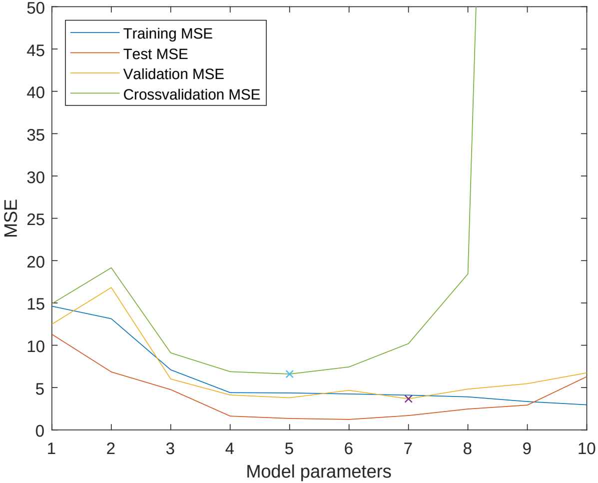
Crossvalidation



Leave One Out



Crossvalidation Results ($K = 5$)



Checking the Results

The data have been generated from the following model:

$$y = (0.5 - x)(5 - x)(x - 3) + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 1.5^2)$

The correct order is then $o = 3$ (4 in the graphs which considers also the constant term)

The procedure is correct on average, the realizations might return different orders than the correct one

Computational Times

Using different methods we have different time for the model selection:

```
Elapsed time is 0.016354 seconds. % Validation
Elapsed time is 0.431666 seconds. % Crossvalidation
Elapsed time is 4.308715 seconds. % LOO
```

Depending on the computational power available and the number of data we have we might choose different methods

Adjustment Techniques

- $C_p = \frac{1}{N}(RSS + 2d\tilde{\sigma})$
  where d is the total number of parameters, $\tilde{\sigma}$ is an estimate of the variance of the noise ϵ
- $AIC = -2\log L + 2d$
  where L is the maximized value of the likelihood function for the estimated model
- $BIC = \frac{1}{N}(RSS + \log(N)d\tilde{\sigma})$
  BIC replaces the $2d\tilde{\sigma}$ of C_p with $\log(N)d\tilde{\sigma}$ term. Since $\log N > 2$ for any $n > 7$, BIC selects smaller models
- Adjusted R^2 $R^2_{ad} = 1 - \frac{RSS/(N-d-1)}{TSS/(N-1)}$
  where TSS is the total sum of squares. Differently from the other criteria, here a **large value** indicates a model with a **small test error**

Test Error Fluctuations

Consider the following scenario:

- $N = 1000$ data
- $N_{\text{train}} = 800$ used to compute $\widehat{w}$
- $N_{\text{test}} = 200$

$\widehat{\text{MSE}}_{\text{test}}(\widehat{\mathbf{w}}) = \frac{1}{N} \sum_{(\mathbf{x},t) \in \mathcal{D}_{\text{test}}} (y(\mathbf{x}; \mathbf{w}) - t)^2$ is a *point estimate*, also subject to variance

Bootstrap Confidence Intervals

- Sample N data points from $\mathcal{D}_{\text{test}}$ *with replacement*
- Do this M times, forming M resamples $\mathcal{D}_1, \dots, \mathcal{D}_M$ of $\mathcal{D}_{\text{test}}$
- Compute MSE (or any other metric) on each resample: $\widehat{\text{MSE}}_1(\hat{\mathbf{w}}), \dots, \widehat{\text{MSE}}_M(\hat{\mathbf{w}})$
- Compute *percentile intervals* on $\widehat{\text{MSE}}_1(\hat{\mathbf{w}}), \dots, \widehat{\text{MSE}}_M(\hat{\mathbf{w}})$

Example: to build a 90% confidence interval, compute the 5-th and the 95-th percentiles.

You can say that the error test is within the two percentiles with 90% confidence.

N.B: A test dataset should be used *only once!*

If I reuse the same test data for K tests/confidence intervals, I must correct the significance level from α to α/K (Bonferroni correction)