Overfitting and Bias-Variance Trade-Off

Jérémie Cabessa

Laboratoire DAVID, UVSQ

Formulation du problème: apprentissage supervisé

- Soient $X_1, \ldots, X_p$ et Y des variables aléatoires.
- $X_1, \ldots, X_p$ sont appelées **variables d'inputs**, **variables indépendantes**, **variables explicatives**, **prédicteurs (features)**.
- Y est appelée **variable d'output**, **variable dépendante**, **réponse (response, target)**.

Formulation du problème: apprentissage supervisé

- Soient $X_1, \dots, X_p$ et Y des variables aléatoires.
- $X_1, \dots, X_p$ sont appelées **variables d'inputs**, **variables indépendantes**, **variables explicatives**, **prédicteurs** (**features**).
- Y est appelée **variable d'output**, **variable dépendante**, **réponse** (**response**, **target**).

Formulation du problème: apprentissage supervisé

- Soient $X_1, \ldots, X_p$ et Y des variables aléatoires.
- $X_1, \ldots, X_p$ sont appelées **variables d'inputs**, **variables indépendantes**, **variables explicatives**, **prédicteurs** (**features**).
- Y est appelée **variable d'output**, **variable dépendante**, **réponse** (**response, target**).

Formulation du problème: apprentissage supervisé

- On suppose qu'il existe une (vraie) **relation** f entre $X_1, \dots, X_p$ et Y de la forme

$$Y = f(X_1, \dots, X_p) + \epsilon$$

où f est une fonction inconnue et ϵ est une variable aléatoire indépendante de $X_1, \dots, X_p$ et de moyenne 0, le bruit.

- On aimerait apprendre une (bonne) **estimation** $\hat{f}$ de f. On aura alors

$$\hat{Y} = \hat{f}(X_1, \dots, X_p)$$

où $\hat{f}$ est l'**estimation** de f et $\hat{Y}$ est la **prediction** de Y.

Formulation du problème: apprentissage supervisé

- On suppose qu'il existe une (vraie) **relation** f entre $X_1, \ldots, X_p$ et Y de la forme

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

où f est une fonction inconnue et ϵ est une variable aléatoire indépendante de $X_1, \ldots, X_p$ et de moyenne 0, le bruit.

- On aimerait apprendre une (bonne) **estimation** $\hat{f}$ de f. On aura alors

$$\hat{Y} = \hat{f}(X_1, \ldots, X_p)$$

où $\hat{f}$ est l'**estimation** de f et $\hat{Y}$ est la **prediction** de Y.

Formulation du problème: apprentissage supervisé

- Pour apprendre l'estimation $\hat{f}$ de f, on dispose de données (data)

$$S_{\text{train}} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\}$$

où $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{i_p})$ pour tout $i = 1, \ldots, n$.

Formulation du problème: apprentissage supervisé

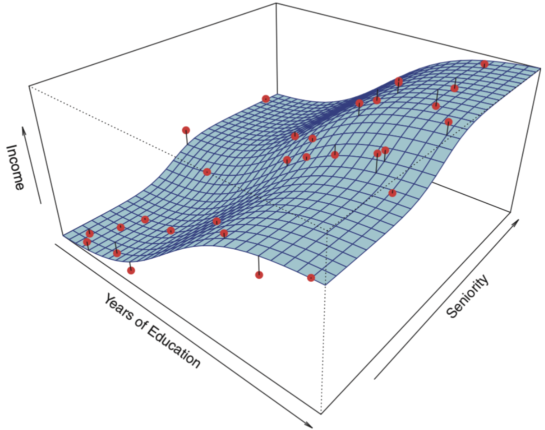


Figure taken from [James et al., 2013]

Dataset

- L'ensemble de toutes les données dont on dispose au départ s'appelle un **dataset**
- Les données utilisées pour entraîner le modèle s'appellent le **train set** (e.g., 80% du dataset).
- Les données, distinctes du train set, utilisées pour évaluer le modèle s'appellent le **test set** (e.g., 20% du dataset).

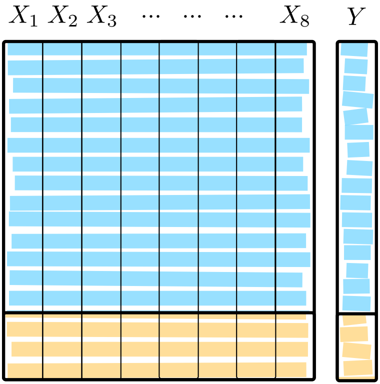
Dataset

- L'ensemble de toutes les données dont on dispose au départ s'appelle un **dataset**
- Les données utilisées pour entraîner le modèle s'appellent le **train set** (e.g., 80% du dataset).
- Les données, distinctes du train set, utilisées pour évaluer le modèle s'appellent le **test set** (e.g., 20% du dataset).

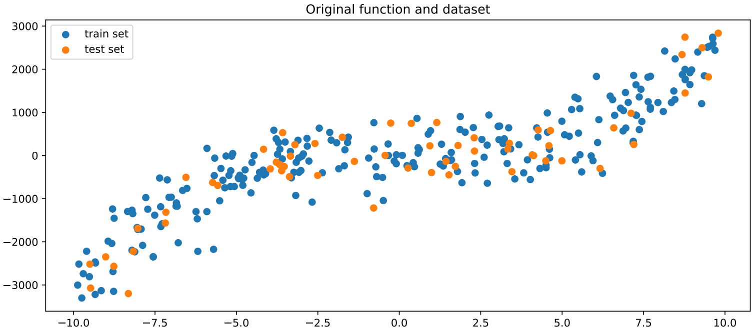
Dataset

- L'ensemble de toutes les données dont on dispose au départ s'appelle un **dataset**
- Les données utilisées pour entraîner le modèle s'appellent le **train set** (e.g., 80% du dataset).
- Les données, distinctes du train set, utilisées pour évaluer le modèle s'appellent le **test set** (e.g., 20% du dataset).

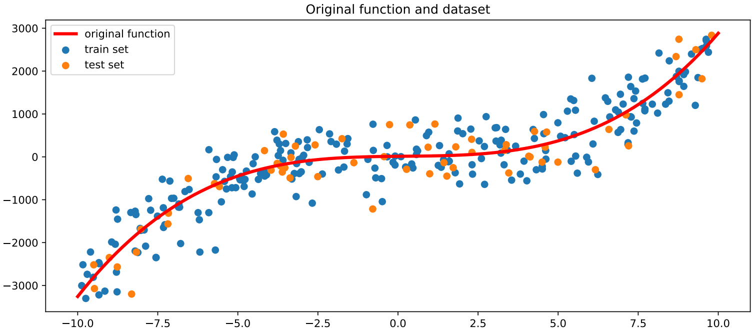
Dataset



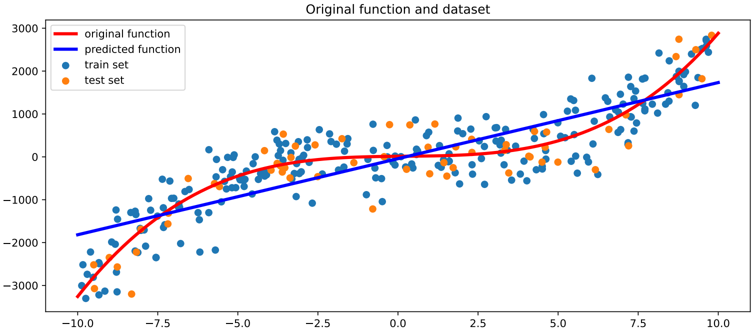
Ensembles de train et de test (train/test sets)



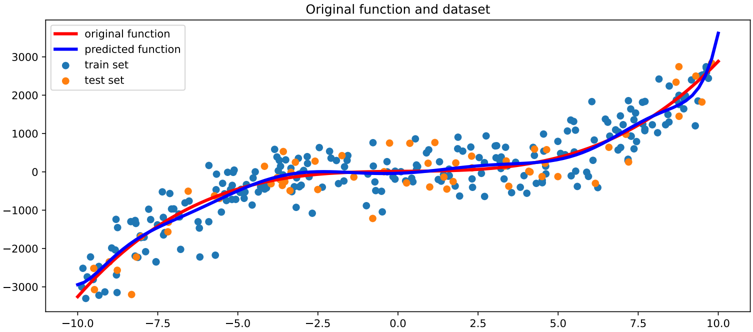
Ensembles de train et de test (train/test sets)



Ensembles de train et de test (train/test sets)



Ensembles de train et de test (train/test sets)



Erreur réductible et irréductible

- On a donc

$$\begin{aligned} Y &= f(X_1, \ldots, X_p) + \epsilon && \text{vraie relation} \\ \hat{Y} &= \hat{f}(X_1, \ldots, X_p) && \text{estimation} \end{aligned}$$

- En que supposant les X_i et $\hat{f}$ sont fixes, et en utilisant $\mathbb{E}(\epsilon) = 0$, on a:

$$\begin{aligned} \mathbb{E}\left[(Y - \hat{Y})^2\right] &= \mathbb{E}\left[\left(f(X_1, \ldots, X_p) + \epsilon - \hat{f}(X_1, \ldots, X_p)\right)^2\right] \\ &= \mathbb{E}\left[\left(f(\boldsymbol{X}) + \epsilon - \hat{f}(\boldsymbol{X})\right)^2\right] \\ &= \underbrace{\left(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\right)^2}_{\text{erreur réductible}} + \underbrace{\mathrm{Var}[\epsilon]}_{\text{erreur irréductible}} \end{aligned}$$

où $\boldsymbol{X} = (X_1, \ldots, X_p)$.

Erreur réductible et irréductible

- On a donc

$$
\begin{array}{rcll}
Y & = & f(X_1, \dots, X_p) + \epsilon & \text{vraie relation} \\
\hat{Y} & = & \hat{f}(X_1, \dots, X_p) & \text{estimation}
\end{array}
$$

- En que supposant les X_i et $\hat{f}$ sont fixes, et en utilisant $\mathbb{E}(\epsilon) = 0$, on a:

$$
\begin{aligned}
\mathbb{E}\left[(Y - \hat{Y})^2\right] &= \mathbb{E}\left[(f(X_1, \dots, X_p) + \epsilon - \hat{f}(X_1, \dots, X_p))^2\right] \\
&= \mathbb{E}\left[(f(\boldsymbol{X}) + \epsilon - \hat{f}(\boldsymbol{X}))^2\right] \\
&= \underbrace{\left(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\right)^2}_{\text{erreur réductible}} + \underbrace{\mathrm{Var}[\epsilon]}_{\text{erreur irréductible}}
\end{aligned}
$$

où $\boldsymbol{X} = (X_1, \dots, X_p)$.

Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de f, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" intrinsèque au modèle issu de:
  - aléatoire dans la vraie relation entre X et Y
    $\Rightarrow$ intrinsic randomness
  - variables explicatives manquantes pour expliquer Y
    $\Rightarrow$ epistemic randomness

Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de f, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" intrinsèque au modèle issu de:
  - aléatoire dans la vraie relation entre X et Y

    $\Rightarrow$ intrinsic randomness
  - variables explicatives manquantes pour expliquer Y

    $\Rightarrow$ epistemic randomness

Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de f, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" intrinsèque au modèle issu de:
  - aléatoire dans la vraie relation entre $\boldsymbol{X}$ et Y

    $\Rightarrow$ intrinsic randomness
  - variables explicatives manquantes pour expliquer Y

    $\Rightarrow$ epistemic randomness

Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de f, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" intrinsèque au modèle issu de:
  - aléatoire dans la vraie relation entre $\boldsymbol{X}$ et Y

    $\Rightarrow$ intrinsic randomness
  - variables explicatives manquantes pour expliquer Y

    $\Rightarrow$ epistemic randomness

Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de f, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" intrinsèque au modèle issu de:
  - aléatoire dans la vraie relation entre $\boldsymbol{X}$ et Y

    $\Rightarrow$ intrinsic randomness
  - variables explicatives manquantes pour expliquer Y

    $\Rightarrow$ epistemic randomness

Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de f, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" intrinsèque au modèle issu de:
  - aléatoire dans la vraie relation entre $\boldsymbol{X}$ et Y

    $\Rightarrow$ intrinsic randomness
  - variables explicatives manquantes pour expliquer Y

    $\Rightarrow$ epistemic randomness

Erreur réductible et irréductible

- Exemple de deux estimations $\hat{f}$. La deuxième estimation est meilleure car elle est associée à une erreur réductible plus faible.

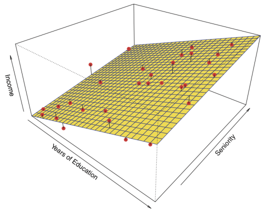


FIGURE 2.4. *A linear model fit by least squares to the* `Income` *data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.*

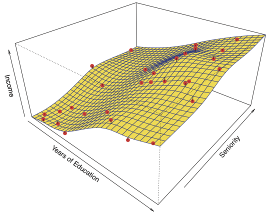


FIGURE 2.5. *A smooth thin-plate spline fit to the* `Income` *data from Figure 2.3 is shown in yellow; the observations are displayed in red. Splines are discussed in Chapter 7.*

Figure taken from [James et al., 2013]

Fonction de coût (cost function)

- Comment mesurer la qualité d'un modèle $\hat{f}$?
- On utilise une **fonction de coût (cost or loss function)**.
- La plus célèbre est **l'erreur des moindre carrés (mean squared error) MSE**. Étant donné un training set

$$S_{\text{train}} = \left\{ (\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n) \right\}$$

on définit

$$\text{MSE}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(\boldsymbol{x_i}) \right)^2$$

Fonction de coût (cost function)

- Comment mesurer la qualité d'un modèle $\hat{f}$?
- On utilise une **fonction de coût (cost or loss function)**.
- La plus célèbre est **l'erreur des moindre carrés (mean squared error) MSE**. Étant donné un training set

$$S_{\text{train}} = \left\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\right\}$$

on définit

$$\text{MSE}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\boldsymbol{x_i})\right)^2$$

Fonction de coût (cost function)

- Comment mesurer la qualité d'un modèle $\hat{f}$?
- On utilise une **fonction de coût (cost or loss function)**.
- La plus célèbre est **l'erreur des moindre carrés (mean squared error) MSE**. Étant donné un training set

$$S_{\text{train}} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\}$$

on définit

$$\text{MSE}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(\boldsymbol{x_i})\right)^2$$

Overfitting

- **Overfitting:** Le modèle est très performant sur le **training set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ basse) alors qu'il est bien moins performant sur sur le **test set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ élevée).
- Le modèle a donc "sur-appris" (overfit) les données d'apprentissage (train set), et ses performances ne se généralisent pas bien sur des données inconnues (test set).
- En fait, le modèle a appris le bruit des données d'apprentissage, au lieu de l'ignorer.

Overfitting

- **Overfitting:** Le modèle est très performant sur le **training set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ basse) alors qu'il est bien moins performant sur sur le **test set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ élevée).
- Le modèle a donc "sur-appris" (overfit) les données d'apprentissage (train set), et ses performances ne se généralisent pas bien sur des données inconnues (test set).
- En fait, le modèle a appris le bruit des données d'apprentissage, au lieu de l'ignorer.

Overfitting

- **Overfitting:** Le modèle est très performant sur le **training set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ basse) alors qu'il est bien moins performant sur sur le **test set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ élevée).
- Le modèle a donc "sur-appris" (overfit) les données d'apprentissage (train set), et ses performances ne se généralisent pas bien sur des données inconnues (test set).
- En fait, le modèle a appris le bruit des données d'apprentissage, au lieu de l'ignorer.

Erreur réductible et irréductible

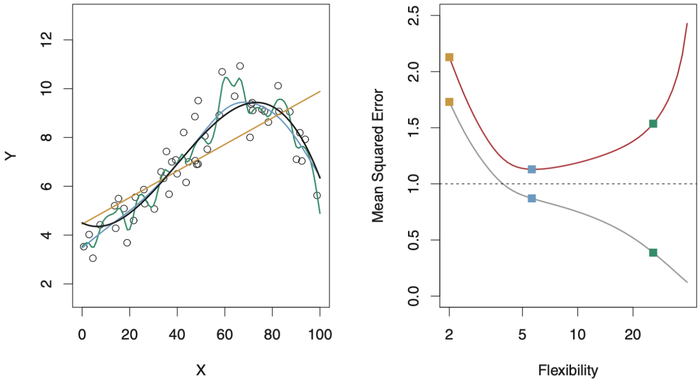


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed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

Figure taken from [James et al., 2013]

Erreur réductible et irréductible

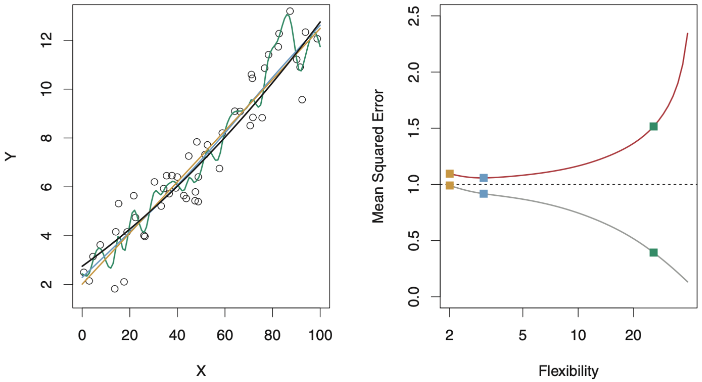


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

Figure taken from [James et al., 2013]

Bias-variance trade-off

- Soit une relation fonctionnelle

$$Y = f(\boldsymbol{X}) + \epsilon$$

  entre des variables d'input $\boldsymbol{X} = (X_1, \ldots, X_p)$ et d'output Y. Le bruit ϵ satisfait $\mathbb{E}[\epsilon] = 0$.

- On cherche à obtenir un modèle

$$\hat{Y} = \hat{f}(\boldsymbol{X})$$

  qui soit le plus performant possible *sur le test set* !

Bias-variance trade-off

- Soit une relation fonctionnelle

$$Y = f(\boldsymbol{X}) + \epsilon$$

  entre des variables d'input $\boldsymbol{X} = (X_1, \ldots, X_p)$ et d'output Y. Le bruit ϵ satisfait $\mathbb{E}[\epsilon] = 0$.

- On cherche à obtenir un modèle

$$\hat{Y} = \hat{f}(\boldsymbol{X})$$

  qui soit le plus performant possible *sur le test set* !

Bias-variance trade-off

- Soit un dataset

$$S = \left\{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \ldots, N\right\}$$

et soit $(\boldsymbol{x}, y) \in S$ un point du dataset.

- On note $\mathbb{E}_{S,\epsilon}[\ldots] := \mathbb{E}[\ldots]$.
- Par hypothèse, on a: $\mathbb{E}[\epsilon] = 0$. Et puisque la "vraie" relation fonctionnelle f est déterministe, on a: $\mathbb{E}[f] = f$.
- On rappelle que

$$\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Bias-variance trade-off

- Soit un dataset

$$S = \left\{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \ldots, N\right\}$$

et soit $(\boldsymbol{x}, y) \in S$ un point du dataset.

- On note $\mathbb{E}_{S,\epsilon}\left[\ldots\right] := \mathbb{E}\left[\ldots\right]$.
- Par hypothèse, on a: $\mathbb{E}\left[\epsilon\right] = 0$. Et puisque la “vraie” relation fonctionnelle f est déterministe, on a: $\mathbb{E}\left[f\right] = f$.
- On rappelle que

$$\mathrm{Var}\left[X\right] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2.$$

Bias-variance trade-off

- Soit un dataset

$$S = \left\{(\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \dots, N\right\}$$

et soit $(\boldsymbol{x}, y) \in S$ un point du dataset.

- On note $\mathbb{E}_{S,\epsilon}[\dots] := \mathbb{E}[\dots]$.
- Par hypothèse, on a: $\mathbb{E}[\epsilon] = 0$. Et puisque la “vraie” relation fonctionnelle f est déterministe, on a: $\mathbb{E}[f] = f$.
- On rappelle que

$$\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.$$

Bias-variance trade-off

- Soit un dataset

$$S = \left\{ (\boldsymbol{x_i}, y_i) \in \mathbb{R}^M \times \mathbb{R} : i = 1, \ldots, N \right\}$$

et soit $(\boldsymbol{x}, y) \in S$ un point du dataset.

- On note $\mathbb{E}_{S,\epsilon}\left[\ldots\right] := \mathbb{E}\left[\ldots\right]$.
- Par hypothèse, on a: $\mathbb{E}\left[\epsilon\right] = 0$. Et puisque la "vraie" relation fonctionnelle f est déterministe, on a: $\mathbb{E}\left[f\right] = f$.
- On rappelle que

$$\mathrm{Var}\left[X\right] := \mathbb{E}\left[(X - \mathbb{E}[X])^2 \right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2.$$

Bias-variance trade-off

$$
\begin{aligned}
\mathbb{E}\left[\left(y - \hat{f}(\boldsymbol{x})\right)^2\right] &= \mathbb{E}\left[\left(f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})}_{=C}\right)^2\right] \\
&= \mathbb{E}\left[A^2 + B^2 + C^2 + 2AB + 2BC + 2CA\right] \\
(\mathbb{E}[B] = \mathbb{E}[C] = 0) \quad &= \mathbb{E}\left[A^2\right] + \mathbb{E}\left[B^2\right] + \mathbb{E}\left[C^2\right] \\
(\mathbb{E}[\epsilon] = 0) \quad &= \mathbb{E}\left[\left(f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] + \mathbb{E}\left[\left(\epsilon - \mathbb{E}[\epsilon]\right)^2\right] + \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] \\
(\mathbb{E}[f(\boldsymbol{x})] = f(\boldsymbol{x})) \quad &= \left(\mathbb{E}\left[\hat{f}(\boldsymbol{x})\right] - f(\boldsymbol{x})\right)^2 + \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] + \mathrm{Var}[\epsilon] \\
&= \mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right]^2 + \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] + \mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\hat{f}(\boldsymbol{x})\right] - f(\boldsymbol{x}) \quad \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]
$$

Bias-variance trade-off

$$
\begin{aligned}
\mathbb{E}\left[\left(y - \hat{f}(\boldsymbol{x})\right)^2\right] &= \mathbb{E}\left[\left(f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}\Bigg[\Big(\underbrace{f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})}_{=C}\Big)^2\Bigg] \\
&= \mathbb{E}\left[A^2 + B^2 + C^2 + 2AB + 2BC + 2CA\right] \\
(\mathbb{E}[B] = \mathbb{E}[C] = 0) \quad &= \mathbb{E}\left[A^2\right] + \mathbb{E}\left[B^2\right] + \mathbb{E}\left[C^2\right] \\
(\mathbb{E}[\epsilon] = 0) \quad &= \mathbb{E}\left[\left(f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] + \mathbb{E}\left[\left(\epsilon - \mathbb{E}[\epsilon]\right)^2\right] + \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] \\
(\mathbb{E}[f(\boldsymbol{x})] = f(\boldsymbol{x})) \quad &= \left(\mathbb{E}\left[\hat{f}(\boldsymbol{x})\right] - f(\boldsymbol{x})\right)^2 + \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] + \mathrm{Var}[\epsilon] \\
&= \mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right]^2 + \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] + \mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\hat{f}(\boldsymbol{x})\right] - f(\boldsymbol{x}) \qquad \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]
$$

Bias-variance trade-off

$$
\begin{aligned}
\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right] &= \mathbb{E}\left[\left(f(\boldsymbol{x})+\epsilon-\hat{f}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{f(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A}+\underbrace{\epsilon}_{=B}+\underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})]-\hat{f}(\boldsymbol{x})}_{=C}\right)^2\right] \\
&= \mathbb{E}\left[A^2+B^2+C^2+2AB+2BC+2CA\right] \\
(\mathbb{E}[B]=\mathbb{E}[C]=0) \quad &= \mathbb{E}\left[A^2\right]+\mathbb{E}\left[B^2\right]+\mathbb{E}\left[C^2\right] \\
(\mathbb{E}[\epsilon]=0) \quad &= \mathbb{E}\left[\left(f(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]+\mathbb{E}\left[\left(\epsilon-\mathbb{E}[\epsilon]\right)^2\right]+\mathbb{E}\left[\left(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] \\
(\mathbb{E}[f(\boldsymbol{x})]=f(\boldsymbol{x})) \quad &= \left(\mathbb{E}\left[\hat{f}(\boldsymbol{x})\right]-f(\boldsymbol{x})\right)^2+\mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right]+\mathrm{Var}[\epsilon] \\
&= \mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right]^2+\mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right]+\mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\hat{f}(\boldsymbol{x})\right]-f(\boldsymbol{x}) \quad \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]
$$

Bias-variance trade-off

$$
\begin{aligned}
& \mathbb{E}\left[\left(y - \hat{f}(\boldsymbol{x})\right)^2\right] && = \mathbb{E}\left[\left(f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\right)^2\right] \\
& && = \mathbb{E}\left[\left(\underbrace{f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})}_{=C}\right)^2\right] \\
& && = \mathbb{E}\left[A^2 + B^2 + C^2 + 2AB + 2BC + 2CA\right] \\
& (\mathbb{E}[B] = \mathbb{E}[C] = 0) && = \mathbb{E}\left[A^2\right] + \mathbb{E}\left[B^2\right] + \mathbb{E}\left[C^2\right] \\
& (\mathbb{E}[\epsilon] = 0) && = \mathbb{E}\left[\left(f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] + \mathbb{E}\left[\left(\epsilon - \mathbb{E}[\epsilon]\right)^2\right] + \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] \\
& (\mathbb{E}[f(\boldsymbol{x})] = f(\boldsymbol{x})) && = \left(\mathbb{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x})\right)^2 + \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] + \mathrm{Var}[\epsilon] \\
& && = \mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right]^2 + \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] + \mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\hat{f}(\boldsymbol{x})\right] - f(\boldsymbol{x}) \quad \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]
$$

Bias-variance trade-off

$$
\begin{aligned}
\mathbb{E}\Big[\big(y-\hat{f}(\boldsymbol{x})\big)^2\Big] &= \mathbb{E}\Big[\big(f(\boldsymbol{x})+\epsilon-\hat{f}(\boldsymbol{x})\big)^2\Big] \\
&= \mathbb{E}\Big[\big(\underbrace{f(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A}+\underbrace{\epsilon}_{=B}+\underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})]-\hat{f}(\boldsymbol{x})}_{=C}\big)^2\Big] \\
&= \mathbb{E}\left[A^2+B^2+C^2+2AB+2BC+2CA\right] \\
(\mathbb{E}[B]=\mathbb{E}[C]=0) \quad &= \mathbb{E}\left[A^2\right]+\mathbb{E}\left[B^2\right]+\mathbb{E}\left[C^2\right] \\
(\mathbb{E}[\epsilon]=0) \quad &= \mathbb{E}\Big[\big(f(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big]+\mathbb{E}\Big[\big(\epsilon-\mathbb{E}[\epsilon]\big)^2\Big]+\mathbb{E}\Big[\big(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big] \\
(\mathbb{E}[f(\boldsymbol{x})]=f(\boldsymbol{x})) \quad &= \Big(\mathbb{E}[\hat{f}(\boldsymbol{x})]-f(\boldsymbol{x})\Big)^2+\mathrm{Var}\big[\hat{f}(\boldsymbol{x})\big]+\mathrm{Var}[\epsilon] \\
&= \mathrm{Biais}\big[\hat{f}(\boldsymbol{x})\big]^2+\mathrm{Var}\big[\hat{f}(\boldsymbol{x})\big]+\mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\big[\hat{f}(\boldsymbol{x})\big] := \mathbb{E}\big[\hat{f}(\boldsymbol{x})\big]-f(\boldsymbol{x}) \qquad \mathrm{Var}\big[\hat{f}(\boldsymbol{x})\big] := \mathbb{E}\Big[\big(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big]
$$

Bias-variance trade-off

$$
\begin{aligned}
\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right] &= \mathbb{E}\left[\left(f(\boldsymbol{x})+\epsilon-\hat{f}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{f(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A}+\underbrace{\epsilon}_{=B}+\underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})]-\hat{f}(\boldsymbol{x})}_{=C}\right)^2\right] \\
&= \mathbb{E}\left[A^2+B^2+C^2+2AB+2BC+2CA\right] \\
(\mathbb{E}[B]=\mathbb{E}[C]=0) \quad &= \mathbb{E}\left[A^2\right]+\mathbb{E}\left[B^2\right]+\mathbb{E}\left[C^2\right] \\
(\mathbb{E}[\epsilon]=0) \quad &= \mathbb{E}\left[\left(f(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]+\mathbb{E}\left[\left(\epsilon-\mathbb{E}[\epsilon]\right)^2\right]+\mathbb{E}\left[\left(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] \\
(\mathbb{E}[f(\boldsymbol{x})]=f(\boldsymbol{x})) \quad &= \left(\mathbb{E}[\hat{f}(\boldsymbol{x})]-f(\boldsymbol{x})\right)^2+\mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right]+\mathrm{Var}[\epsilon] \\
&= \mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right]^2+\mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right]+\mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\hat{f}(\boldsymbol{x})\right]-f(\boldsymbol{x}) \qquad \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]
$$

Bias-variance trade-off

$$
\begin{aligned}
\mathbb{E}\Big[\big(y - \hat{f}(\boldsymbol{x})\big)^2\Big] &= \mathbb{E}\Big[\big(f(\boldsymbol{x}) + \epsilon - \hat{f}(\boldsymbol{x})\big)^2\Big] \\
&= \mathbb{E}\Big[\big(\underbrace{f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})}_{=C}\big)^2\Big] \\
&= \mathbb{E}\left[A^2 + B^2 + C^2 + 2AB + 2BC + 2CA\right] \\
(\mathbb{E}[B] = \mathbb{E}[C] = 0) \quad &= \mathbb{E}\left[A^2\right] + \mathbb{E}\left[B^2\right] + \mathbb{E}\left[C^2\right] \\
(\mathbb{E}[\epsilon] = 0) \quad &= \mathbb{E}\Big[\big(f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big] + \mathbb{E}\Big[\big(\epsilon - \mathbb{E}[\epsilon]\big)^2\Big] + \mathbb{E}\Big[\big(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big] \\
(\mathbb{E}[f(\boldsymbol{x})] = f(\boldsymbol{x})) \quad &= \Big(\mathbb{E}[\hat{f}(\boldsymbol{x})] - f(\boldsymbol{x})\Big)^2 + \mathrm{Var}\Big[\hat{f}(\boldsymbol{x})\Big] + \mathrm{Var}[\epsilon] \\
&= \mathrm{Biais}\big[\hat{f}(\boldsymbol{x})\big]^2 + \mathrm{Var}\big[\hat{f}(\boldsymbol{x})\big] + \mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\big[\hat{f}(\boldsymbol{x})\big] := \mathbb{E}\big[\hat{f}(\boldsymbol{x})\big] - f(\boldsymbol{x}) \quad \mathrm{Var}\big[\hat{f}(\boldsymbol{x})\big] := \mathbb{E}\Big[\big(\hat{f}(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]\big)^2\Big]
$$

Bias-variance trade-off

$$
\begin{aligned}
\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right] &= \mathbb{E}\left[\left(f(\boldsymbol{x})+\epsilon-\hat{f}(\boldsymbol{x})\right)^2\right] \\
&= \mathbb{E}\left[\left(\underbrace{f(\boldsymbol{x}) - \mathbb{E}[\hat{f}(\boldsymbol{x})]}_{=A} + \underbrace{\epsilon}_{=B} + \underbrace{\mathbb{E}[\hat{f}(\boldsymbol{x})] - \hat{f}(\boldsymbol{x})}_{=C}\right)^2\right] \\
&= \mathbb{E}\left[A^2+B^2+C^2+2AB+2BC+2CA\right] \\
(\mathbb{E}[B]=\mathbb{E}[C]=0) \quad &= \mathbb{E}\left[A^2\right]+\mathbb{E}\left[B^2\right]+\mathbb{E}\left[C^2\right] \\
(\mathbb{E}[\epsilon]=0) \quad &= \mathbb{E}\left[\left(f(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]+\mathbb{E}\left[\left(\epsilon-\mathbb{E}[\epsilon]\right)^2\right]+\mathbb{E}\left[\left(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right] \\
(\mathbb{E}[f(\boldsymbol{x})]=f(\boldsymbol{x})) \quad &= \left(\mathbb{E}[\hat{f}(\boldsymbol{x})]-f(\boldsymbol{x})\right)^2+\mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right]+\mathrm{Var}[\epsilon] \\
&= \mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right]^2+\mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right]+\mathrm{Var}[\epsilon]
\end{aligned}
$$

où

$$
\mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\hat{f}(\boldsymbol{x})\right]-f(\boldsymbol{x}) \quad \mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right] := \mathbb{E}\left[\left(\hat{f}(\boldsymbol{x})-\mathbb{E}[\hat{f}(\boldsymbol{x})]\right)^2\right]
$$

Bias-variance trade-off

- Le biais $\mathrm{Biais}\big[\hat{f}(\boldsymbol{x})\big]$ représente **l'erreur due à la complexité du modèle** $\hat{f}$.
- La variance $\mathrm{Var}\big[\hat{f}(\boldsymbol{x})\big]$ représente **la sensibilité du modèle** $\hat{f}$ **par rapport à son training set**, i.e., la variation moyenne de $\hat{f}(\boldsymbol{x})$ autour de sa moyenne $\mathbb{E}[\hat{f}(\boldsymbol{x})]$ si le modèle $\hat{f}$ était estimé à partir de différents training sets.
- $\mathrm{Var}[\epsilon]$ représente **l'erreur irréductible** liée au bruit inhérent à la vraie relation fonctionnelle f.
- **Dilemme biais-variance (bias-variance trade-off):** plus le modèle $\hat{f}$ est complexe, plus le biais sera faible, mais plus la variance sera élevée.

Bias-variance trade-off

- Le biais $\mathrm{Biais}\left[\hat{f}(\boldsymbol{x})\right]$ représente **l'erreur due à la complexité du modèle $\hat{f}$.**
- La variance $\mathrm{Var}\left[\hat{f}(\boldsymbol{x})\right]$ représente **la sensibilité du modèle $\hat{f}$ par rapport à son training set**, i.e., la variation moyenne de $\hat{f}(\boldsymbol{x})$ autour de sa moyenne $\mathbb{E}[\hat{f}(\boldsymbol{x})]$ si le modèle $\hat{f}$ était estimé à partir de différents training sets.
- $\mathrm{Var}[\epsilon]$ représente **l'erreur irréductible** liée au bruit inhérent à la vraie relation fonctionnelle f.
- **Dilemme biais-variance (bias-variance trade-off):** plus le modèle $\hat{f}$ est complexe, plus le biais sera faible, mais plus la variance sera élevée.

Bias-variance trade-off

- Le biais $\mathrm{Biais}\big[\hat{f}(\boldsymbol{x})\big]$ représente **l'erreur due à la complexité du modèle** $\hat{f}$.
- La variance $\mathrm{Var}\big[\hat{f}(\boldsymbol{x})\big]$ représente **la sensibilité du modèle** $\hat{f}$ **par rapport à son training set**, i.e., la variation moyenne de $\hat{f}(\boldsymbol{x})$ autour de sa moyenne $\mathbb{E}[\hat{f}(\boldsymbol{x})]$ si le modèle $\hat{f}$ était estimé à partir de différents training sets.
- $\mathrm{Var}[\epsilon]$ représente **l'erreur irréductible** liée au bruit inhérent à la vraie relation fonctionnelle f.
- **Dilemme biais-variance (bias-variance trade-off):** plus le modèle $\hat{f}$ est complexe, plus le biais sera faible, mais plus la variance sera élevée.

Bias-variance trade-off

- Le biais $\text{Biais}\big[\hat{f}(\boldsymbol{x})\big]$ représente **l'erreur due à la complexité du modèle** $\hat{f}$.
- La variance $\text{Var}\big[\hat{f}(\boldsymbol{x})\big]$ représente **la sensibilité du modèle** $\hat{f}$ **par rapport à son training set**, i.e., la variation moyenne de $\hat{f}(\boldsymbol{x})$ autour de sa moyenne $\mathbb{E}[\hat{f}(\boldsymbol{x})]$ si le modèle $\hat{f}$ était estimé à partir de différents training sets.
- $\text{Var}[\epsilon]$ représente **l'erreur irréductible** liée au bruit inhérent à la vraie relation fonctionnelle f.
- **Dilemme biais-variance (bias-variance trade-off):** plus le modèle $\hat{f}$ est complexe, plus le biais sera faible, mais plus la variance sera élevée.

Bias-variance trade-off

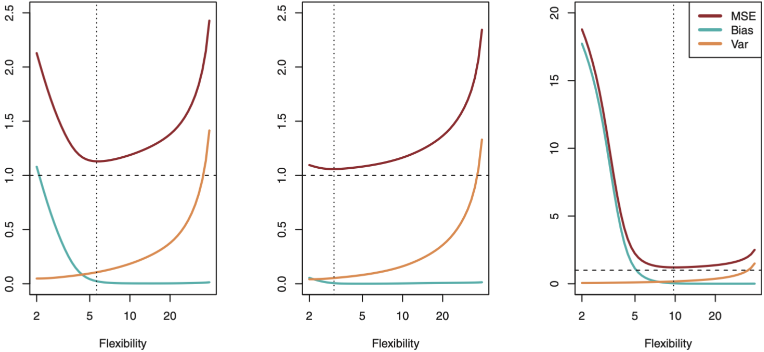


FIGURE 2.12. *Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

Figure taken from [James et al., 2013]

Bibliographie

Fleuret, F. (2022).
Deep Learning Course.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
An Introduction to Statistical Learning: with Applications in R, volume 103 of *Springer Texts in Statistics*.
Springer, New York.

Wikipedia contributors (2022).
Bias–variance tradeoff — Wikipedia, the free encyclopedia.