Advanced Learning Models

Jakob Verbeek
jakob.verbeek@inria.fr

December 10, 2015

Part I

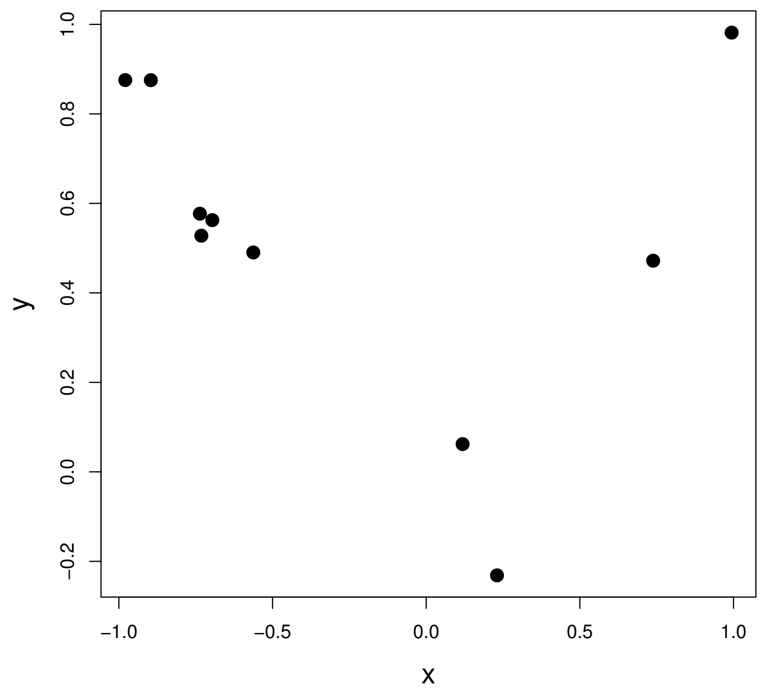
Overfitting, bias-variance tradeoff: what is the problem?

Thanks to Laurent Jacob for sharing slides!

Short term

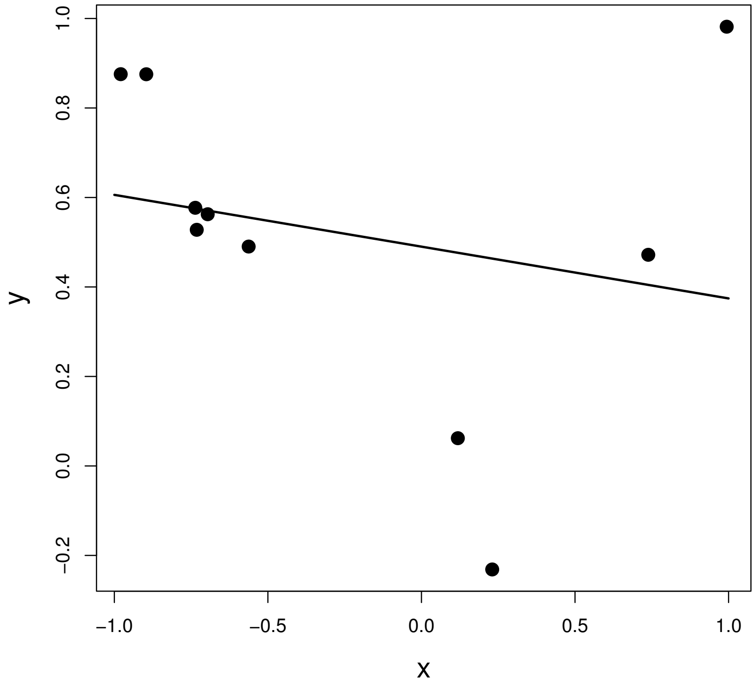
- We start with an informal example.
- We will formalize what we observe later.

Bias-variance tradeoff: intuition



- We observe 10 couples (x_i, y_i).
- We want to estimate y from x.
- **Our first strategy:** find f such that $f(x_i)$ is close to y_i.

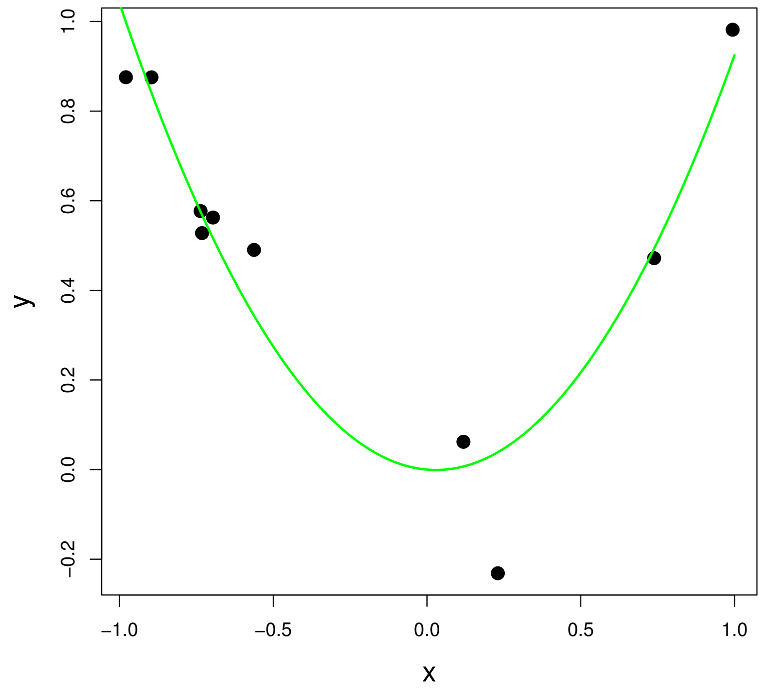
Bias-variance tradeoff: intuition



Find f as a line

$$\min_{f(x)=ax+b} \|Y - f(X)\|^2$$

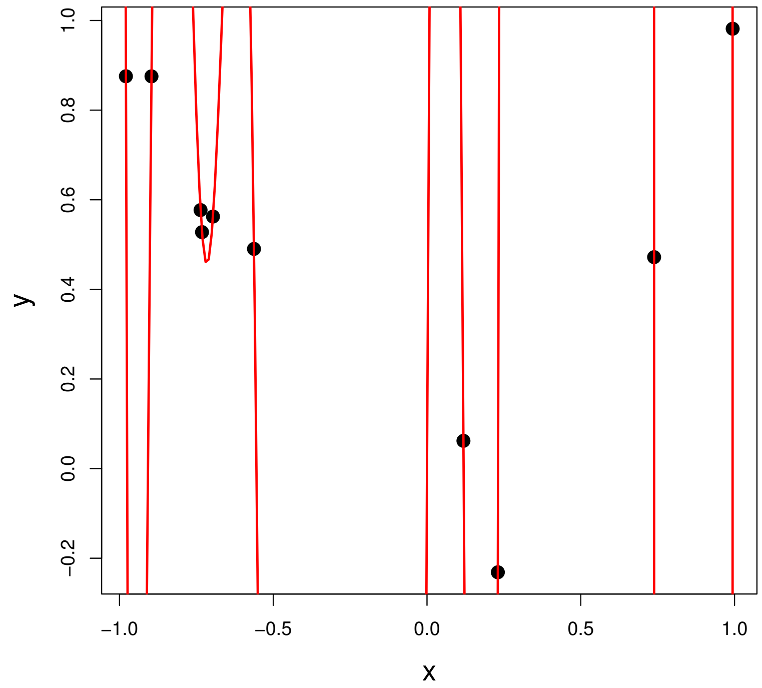
Bias-variance tradeoff: intuition



Find f as a quadratic function

$$\min_{f(x)=ax^2+bx+c} \|Y - f(X)\|^2$$

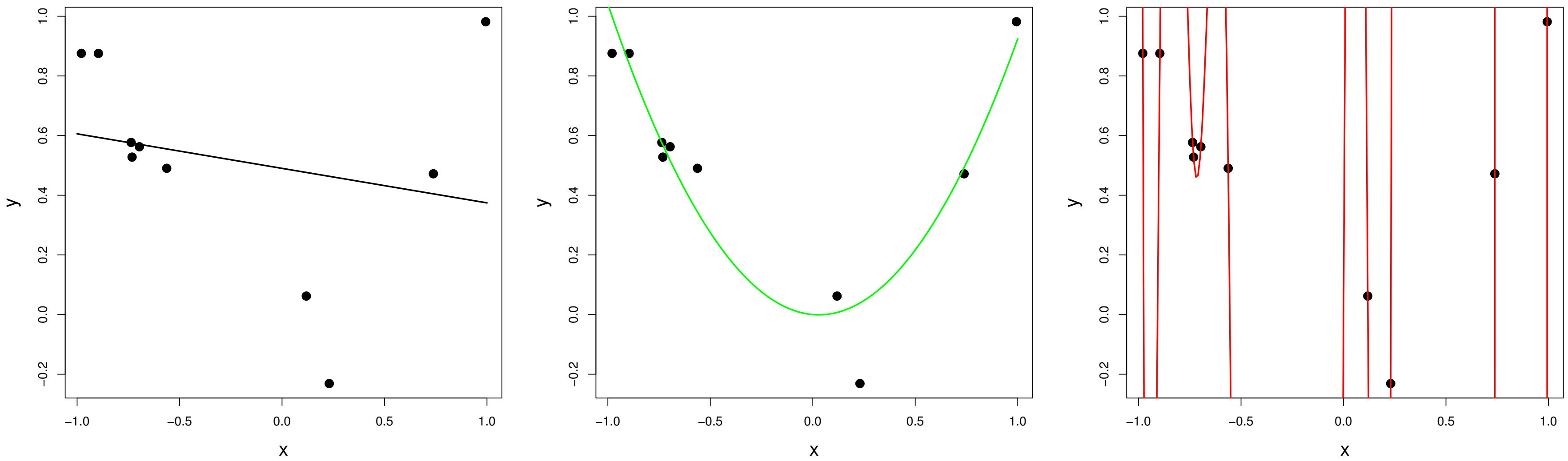
Bias-variance tradeoff: intuition



Find f as a polynomial of degree 10

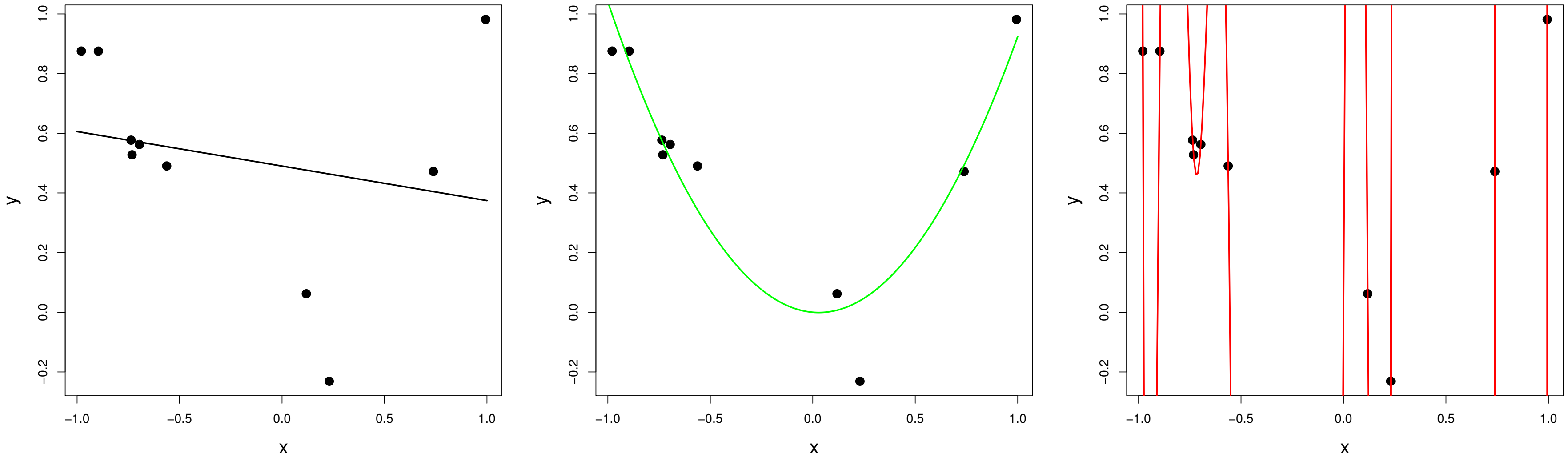
$$\min_{f(x)=\sum_{j=0}^{10} a_j x^j} \|Y - f(X)\|^2$$

Bias-variance tradeoff: intuition



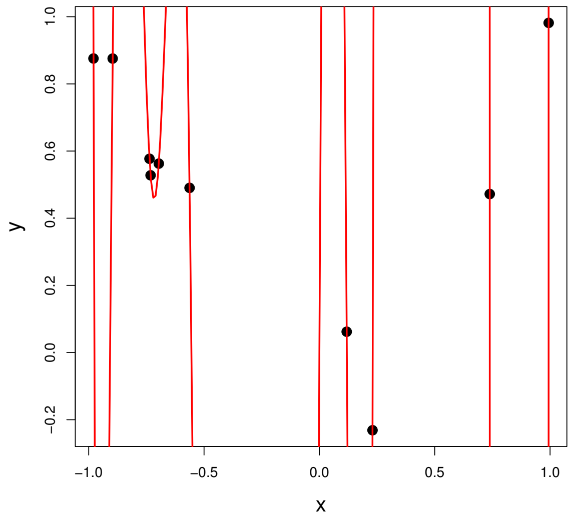
Which function would you trust to predict y corresponding to $x = 0.5$?

Bias-variance tradeoff: intuition



- Reminder: we aim at "finding f such that $f(x_i)$ is close to y_i".
- With the polynomial of degree 10, $f(x_i) - y_i = 0$ for all 10 points.
- There is something wrong with our objective.

Bias-variance tradeoff: intuition

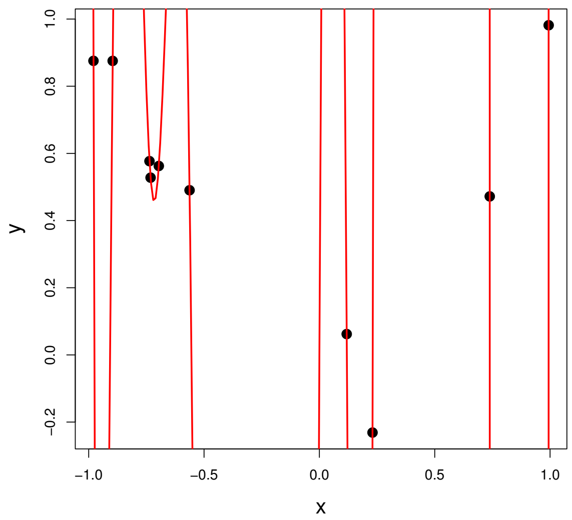


More precisely:

- If we allow **any** function f, we can find **a lot** of perfect solutions for the training data.
- Our actual goal is to estimate y for **new points** x from the same population :

$$\min_f \mathbb{E}_{(X,Y)} \|Y - f(X)\|^2$$

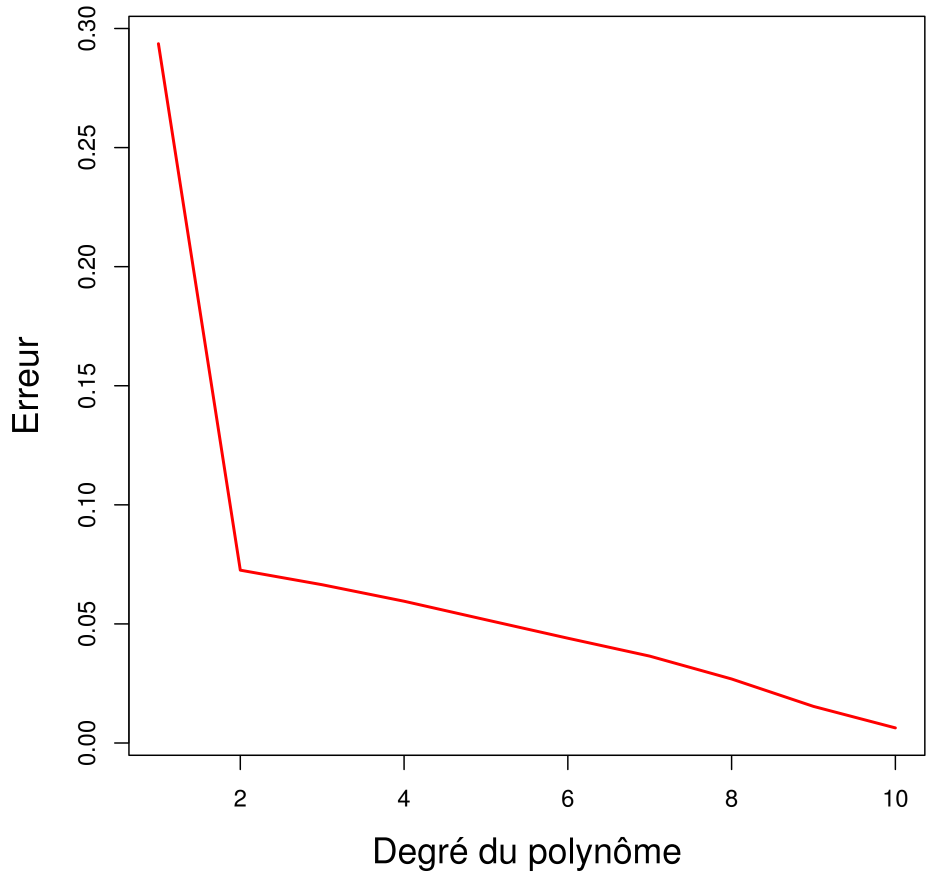
Biais-variance tradeoff: intuition



Even more precisely :

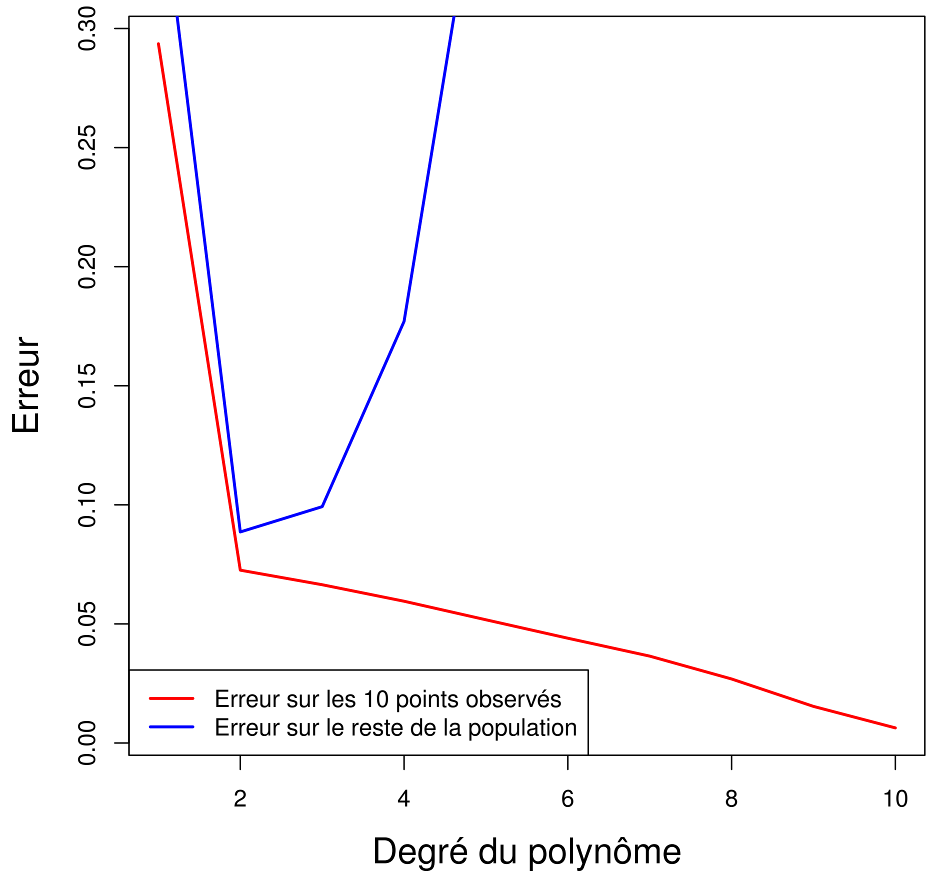
- We did not take into account the fact that our 10 points are a subsample from the population.
- If we sample 10 new points from the same population, the complex functions are likely to change more than the simple ones.
- Consequence: these fonctions will probably generalize less well to the rest of the population.

Overfitting



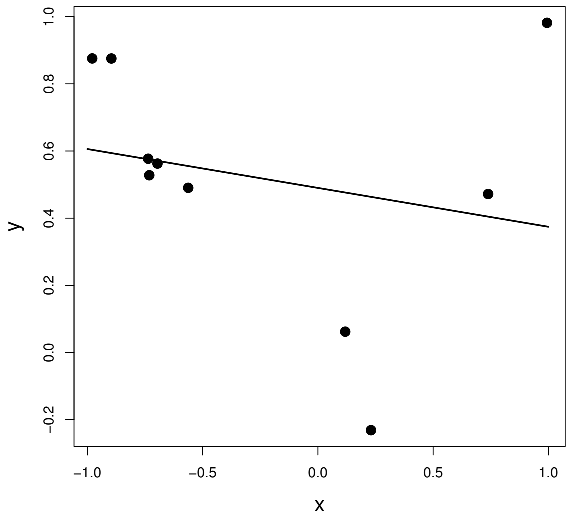
- When the degree increases, the error $\|y - f(x)\|^2$ over the 10 observations always decreases.
- Over the rest of the population, the error decreases, **then increases**.

Overfitting



- When the degree increases, the error $\|y - f(x)\|^2$ over the 10 observations always decreases.
- Over the rest of the population, the error decreases, **then increases**.

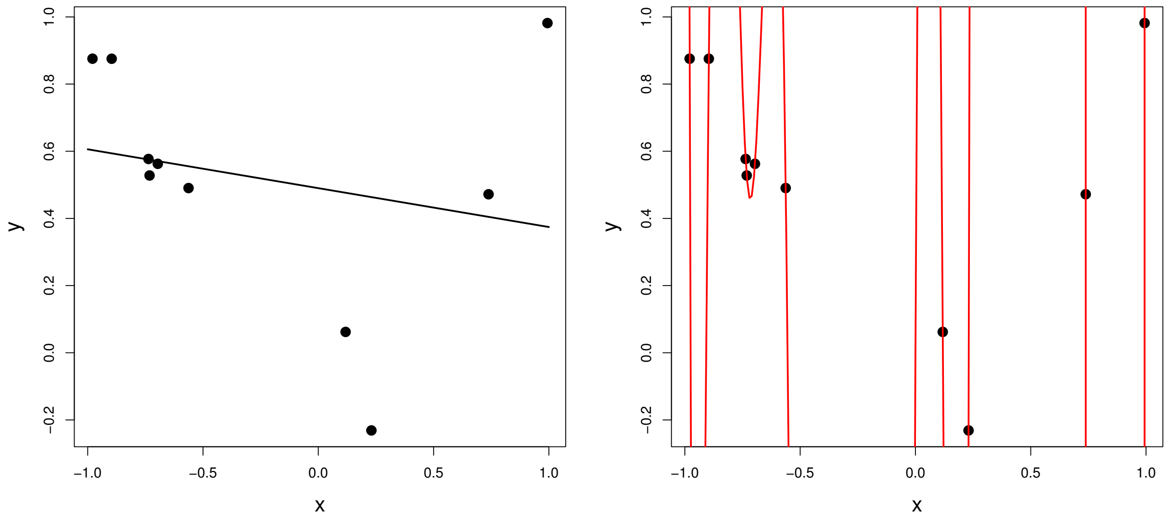
Overfitting



This suggests the existence of a **tradeoff** between two types of errors:

- Sets of functions which are too simple cannot contain functions which explain the data well enough.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample.

Overfitting



This suggests the existence of a **tradeoff** between two types of errors:

- Sets of functions which are too simple cannot contain functions which explain the data well enough.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample.

Parenthesis: complexity vs dimension (1/3)

- Our introductive examples had **a large number of descriptors**.
- This case involves increasingly **complex** functions of a single variable.

Parenthesis : complexity vs dimension (2/3)

- In fact, the two notions are related: here in particular, the three functions are linear in different representations.
- Reminder (linear regression): $\arg\min_{\theta \in \mathbb{R}^p} \|Y - X\theta\|^2 = (X^\top X)^{-1} X^\top Y$ (if $X^\top X$ is invertible).
- How can we use this fact to compute $\arg\min_{f(x)=\sum_{j=1}^{p} a_i x^j} \|Y - f(X)\|^2$?

Parenthesis : complexity vs dimension (3/3)

- We could have illustrated the same principle using linear functions involving more and more variables.
- Example : predicting a phenotype using the expression of an increasing number of genes.
- We sticked to polynomials, which allow for better visual representations.
- Along this class, the notion of complexity of a set of functions will become more and more precise.
- Complexity is what causes problems for inference, not just dimension.

Second parenthesis : models

- Until now, we did not need to introduce a **model** for the data, *i.e.*, a distribution over $\mathcal{X} \times \mathcal{Y}$:
  - Data could come from any population.
  - The functions we used to predict y can be derived from particular probabilistic models, but this is not necessary (they were in fact historically introduced without a model).
- The objective is not to criticize the use of models, but to show that the tradeoff problem we introduced goes beyond probabilistic models.
- We now show how using a model can give a better insight into the problem.

A little more formally: biais-variance decomposition

- We now assume that the data follow:

$$y = f(x) + \varepsilon, \tag{1}$$

  and $\mathbf{E}[\varepsilon] = 0$.
- Without loss of generality, we consider an estimator $\hat{f}$ of f, which is a function of training data $\mathcal{D} = (x_i, y_i)_{(i=1,\ldots,n)}$ sampled i.i.d. from (1)
- Note: $\hat{f}$ is a random function.
- We consider the mean **quadratic error** $\mathbf{E}[(y - \hat{f}(x))^2]$ incurred when using $\hat{f}$ to estimate for a given x the corresponding y sampled from (1) independently from $\mathcal{D}$.
- Expectation is taken over $\mathcal{D}$ used to estimate $\hat{f}$, and $\varepsilon = y - f(x)$.

A little more formally: biais-variance decomposition

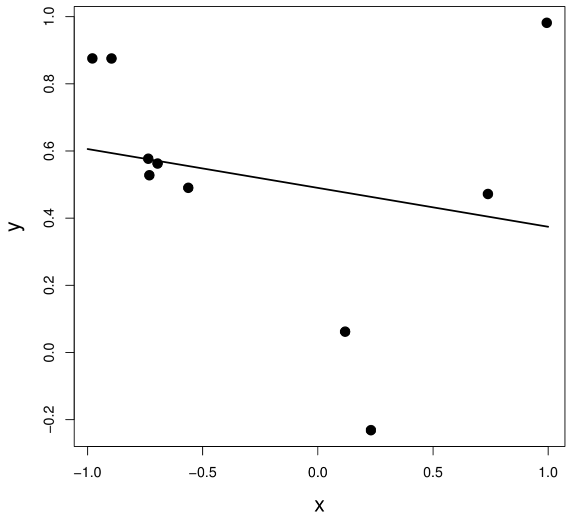
Proposition

Under the previous hypotheses,

$$\mathbf{E}[(y - \hat{f}(x))^2] = \left(\mathbf{E}[\hat{f}(x)] - f(x)\right)^2 + \mathbf{E}\left[\left(\mathbf{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right] + \mathbf{E}[(y - f(x))^2]$$

- The first term is the squared bias of $\hat{f}$: the difference between its mean (over the sample of $\mathcal{D}$) and the true f.
- The second term is the variance of $\hat{f}$: how much $\hat{f}$ varies around its average when the dataset $\mathcal{D}$ changes.
- The third term is the Bayes error, and does not depend on the estimator. The actual quantity of interest is the **excess of risk** $\mathbf{E}[(y - \hat{f}(x))^2] - \mathbf{E}[(y - f(x))^2]$.

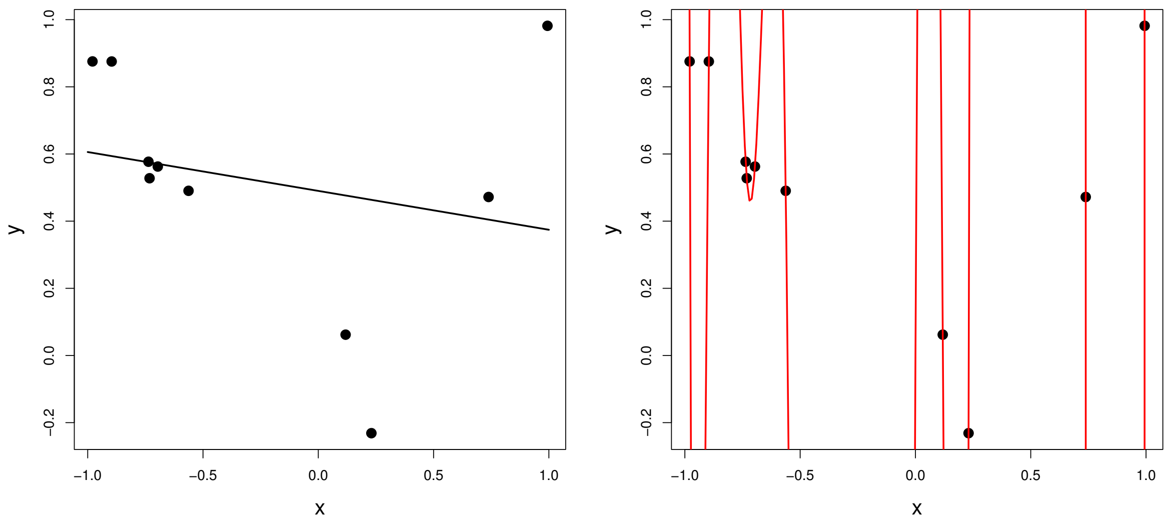
Back to our example



Tradeoff between two types of error:

- Sets of functions which are too simple cannot contain functions which explain the data well enough:
  these sets lead to estimators with a large bias.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample:
  these sets lead to estimators with a large variance.

Back to our example



Tradeoff between two types of error:

- Sets of functions which are too simple cannot contain functions which explain the data well enough:
  these sets lead to estimators with a large bias.
- Sets of functions which are too rich may contain functions which are too specific to the observed sample:
  these sets lead to estimators with a large variance.

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$\mathbf{E}[(y - \hat{f}(x))^2] = \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2]$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$\begin{aligned}\mathbf{E}[(y - \hat{f}(x))^2] =& \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2] \\ =& \mathbf{E}[y^2] - \mathbf{E}[2y\hat{f}(x)] + \mathbf{E}[\hat{f}(x)^2]\end{aligned}$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$
\begin{aligned}
\mathbf{E}[(y - \hat{f}(x))^2] =& \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2] \\
=& \mathbf{E}[y^2] - \mathbf{E}[2y\hat{f}(x)] + \mathbf{E}[\hat{f}(x)^2] \\
=& \mathbf{E}[y]^2 + \mathbf{E}[(y - \mathbf{E}[y])^2] \\
& - 2\mathbf{E}[y]\mathbf{E}[\hat{f}(x)] \\
& + \mathbf{E}[\hat{f}(x)]^2 + \mathbf{E}[(\hat{f}(x) - \mathbf{E}[\hat{f}(x)])^2]
\end{aligned}
$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$
\begin{aligned}
\mathbf{E}[(y - \hat{f}(x))^2] =& \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2] \\
=& \mathbf{E}[y^2] - \mathbf{E}[2y\hat{f}(x)] + \mathbf{E}[\hat{f}(x)^2] \\
=& f(x)^2 + \mathbf{E}[(y - f(x))^2] \\
& - 2f(x)\mathbf{E}[\hat{f}(x)] \\
& + \mathbf{E}[\hat{f}(x)]^2 + \mathbf{E}[(\hat{f}(x) - \mathbf{E}[\hat{f}(x)])^2]
\end{aligned}
$$

Biais-variance decomposition: proof

Reminder (König-Huygens)

For any real random variable Z, $\mathbf{E}\left[(Z - \mathbf{E}[Z])^2\right] = \mathbf{E}[Z^2] - \mathbf{E}[Z]^2$

$$\begin{aligned}
\mathbf{E}[(y - \hat{f}(x))^2] =& \mathbf{E}[y^2 - 2y\hat{f}(x) + \hat{f}(x)^2] \\
=& \mathbf{E}[y^2] - \mathbf{E}[2y\hat{f}(x)] + \mathbf{E}[\hat{f}(x)^2] \\
=& f(x)^2 + \mathbf{E}[(y - f(x))^2] \\
& - 2f(x)\mathbf{E}[\hat{f}(x)] \\
& + \mathbf{E}[\hat{f}(x)]^2 + \mathbf{E}[(\hat{f}(x) - \mathbf{E}[\hat{f}(x)])^2] \\
=& \mathbf{E}[(y - f(x))^2] + \mathbf{E}[(\hat{f}(x) - \mathbf{E}[\hat{f}(x)])^2] \\
& + \left(\mathbf{E}[\hat{f}(x)] - f(x)\right)^2
\end{aligned}$$

Biais-variance decomposition : perspective

$$\mathbf{E}[(y-\hat{f}(x))^2] = \left(\mathbf{E}[\hat{f}(x)] - f(x)\right)^2 + \mathbf{E}\left[\left(\mathbf{E}[\hat{f}(x)] - \hat{f}(x)\right)^2\right] + \mathbf{E}[(y-f(x))^2]$$

- Using a (rather general) model, we managed to start formalizing the tradeoff introduced with our example.
- Decomposition valid for any x, thus also in expectation over independent x.