

Today: BIAS - VARIANCE TRADE-OFF
 SPLIT INTO TRAIN AND TEST SET (+ VALIDATION SET)
 CROSS-VALIDATION
 ... BOOTSTRAPPING

"Plurality non est ponenda sine necessitate" (William of Ockham)

$n=30$ (y_i, x_i) $i=1, \dots, 30$ from $y = f(x) + \varepsilon$
 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

GOAL: estimate $f(x)$

$$y = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \varepsilon$$

Increasing the complexity of the model (linear Gaussian regression $\rightarrow$ complexity = p)
 increasing p $\rightarrow$ smaller RSS(β)
 $\searrow$ larger R^2

Our goal is NOT fit perfectly the data, but try to fit $f(x)$, the systematic part $\rightarrow$ we want that our model can predict correctly the response on new data

When we evaluate our model $\hat{f}(x)$ on new data
 $\rightarrow$ residual error that initially decreases, then increases again
 $\rightarrow R^2$ the other way around

Related to the model complexity:

- $\rightarrow$ a model not complex enough cannot catch the systematic part ($f(x)$)
- $\rightarrow$ a model too complex not only fits the systematic part, but also noise

Mathematically:

$$E[(\hat{f}(x_0) - f(x_0))^2] = \underbrace{(E[\hat{f}(x_0)] - f(x_0))^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{variance}}$$

$$\hat{y} = \hat{f}(x)$$

$$y \neq f(x)$$

$$E[(\hat{f}(x) - y_0)^2] = \underbrace{(E[\hat{f}(x_0)] - y_0)^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{variance}} + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible error}}$$

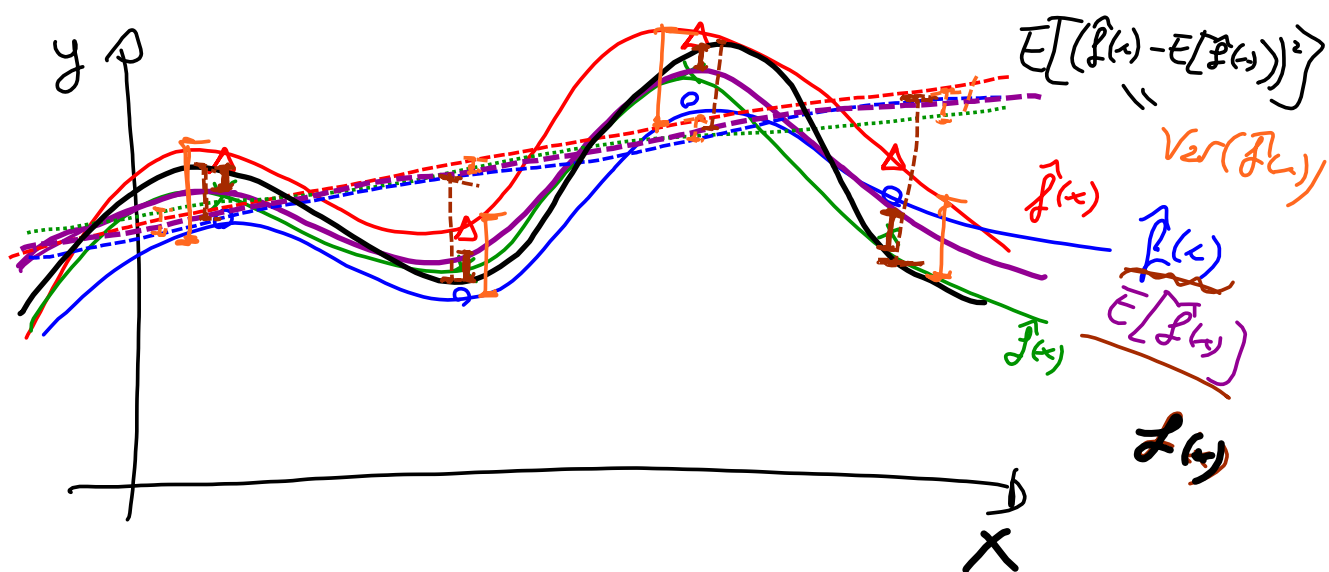
mean squared error

The error has two components, squared bias and variance

model with low complexity	→ high bias low variance	} the model is not able to capture the systematic part, but it is pretty stable (new data, approximately same model)
model with high complexity	→ low bias high variance	

} the model fits the data very well, but also captures the noisy part. = Unstable (fit on new data, or different model)

BIAS - VARIANCE TRADE-OFF: in order to minimize the ^{prediction} error, we need to find the right complexity, i.e. the model for which the sum of the squared bias and variance is minimum



Unfortunately we do not know $f(x)$ so the formula above cannot be used directly. Still, it is clear that we cannot find the right complexity by looking at the error computed on the data we used to train our model.

- The error is too small (optimism)
- We would always choose a model that is too complex ($\rightarrow$ overfitting)

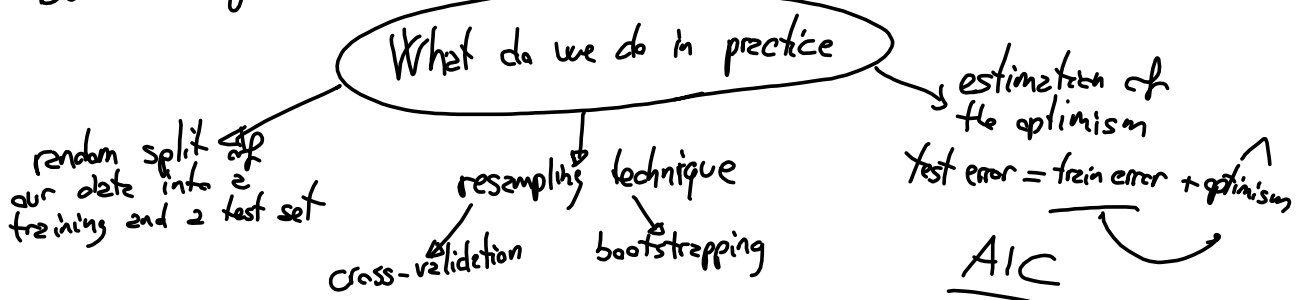
OVERFITTING: is that process for which we fit a model too specific to the data on hand
 $\rightarrow$ we are also (in addition to the systematic/structural part) modelling noise/characteristics of the data only present in the specific data set

$$y = f(x) + \epsilon$$

The key to avoid overfitting we need to compute the error (test/evaluate our model) on new data, i.e. independent of the data we used to train our model

This is not perfect, because it is influenced by the variability of the new data.

But it is good enough to avoid overfitting



Random split in a training and test set

The simplest way to evaluate a model on data that are not used to train it, is to split the available data in two sets

training set: data we use to train the model

test set: data we use to evaluate the model and choose the best one

we are both selecting and evaluating the model on the same data
still a bit of optimism

further split into ~~test~~ and validation set

training set	validation set	test set
50%	25%	25%
50%	0%	50%
66.6%	0%	33.3%
80%	0%	20%