STATG019 – Selected Topics in Statistics 2018

Lecture 2

Theory and Methodology for Estimating the Generalization Error

Dr Franz J. Király

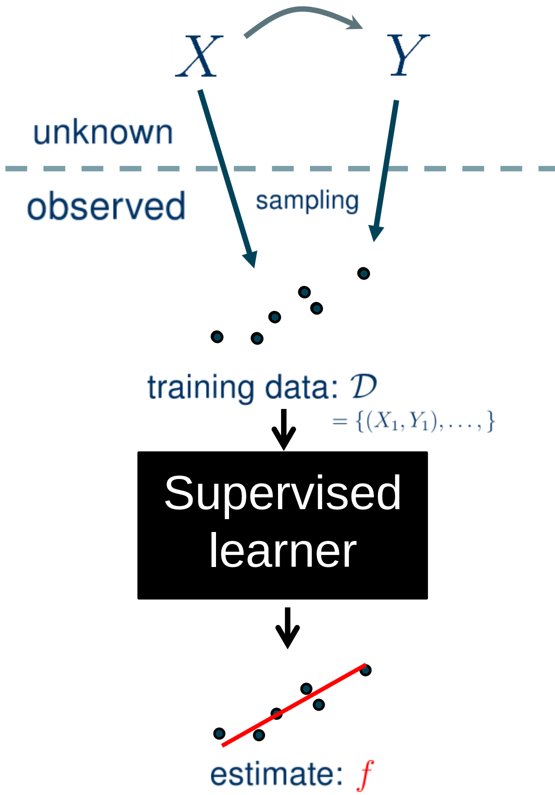
The statistical supervised learning setting

Given data from generative, unknown RV (X, Y) t.v.in $\mathcal{X} \times \mathcal{Y}$ (not independent of each other in general)

features labels domains

where we observe

$$(X_1, Y_1), \dots, (X_N, Y_N) \underset{\text{i.i.d.}}{\sim} (X, Y)$$



Estimate/learn

a prediction functional f t.v.in $[\mathcal{X} \to \mathcal{Y}]$

(via an estimator using only the data (X_i, Y_i))

Such that the expected generalization error

$\varepsilon(f) := \mathbb{E}\left[L\left(f(X), Y\right)\right]$ is small

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a chosen convex loss

i.e., $[\widehat{y} \mapsto L(\widehat{y}, y)]$ is convex for all $y \in \mathcal{Y}$

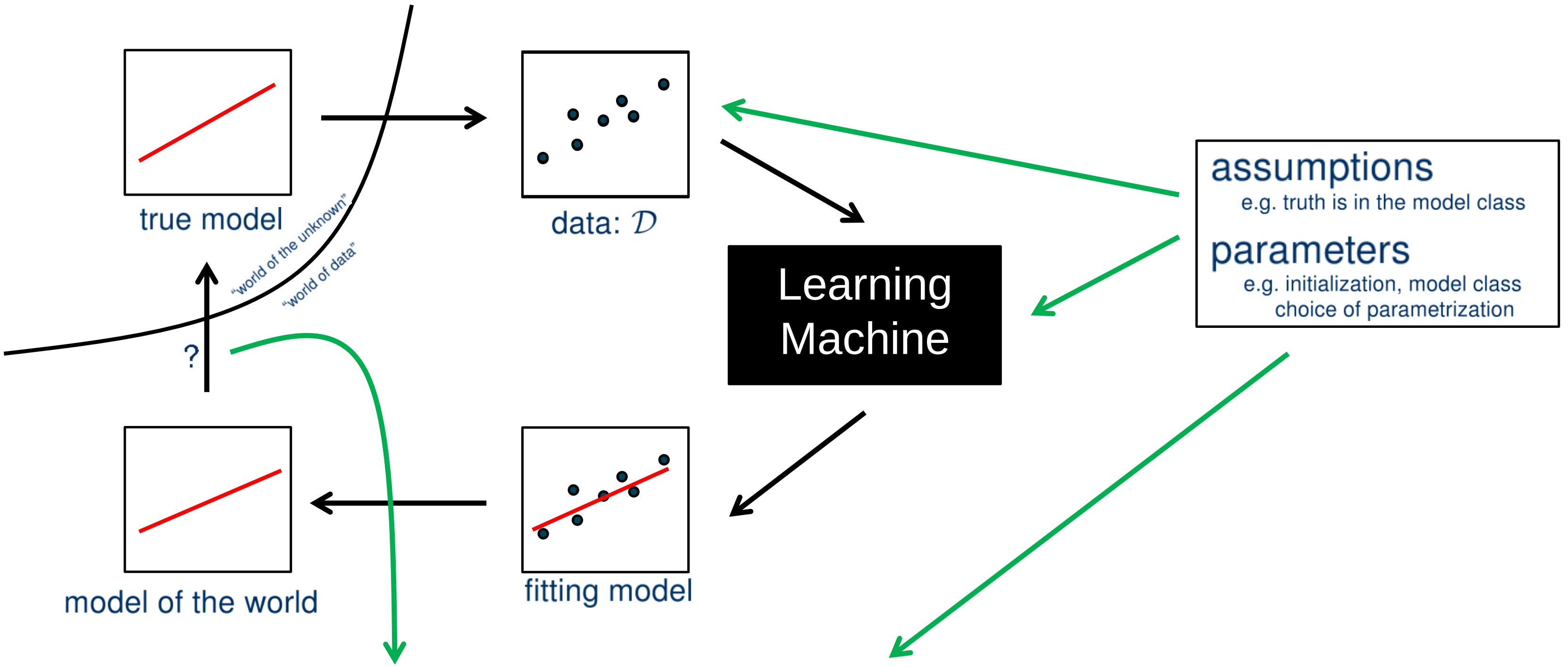
e.g., $L : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$

Validity: (i) under assumptions, prove that f has low error (model specific)

(ii) external estimate of f's error with guarantees (model agnostic)

(i) model-specific validity arguments

Classical statistics and statistical learning theory

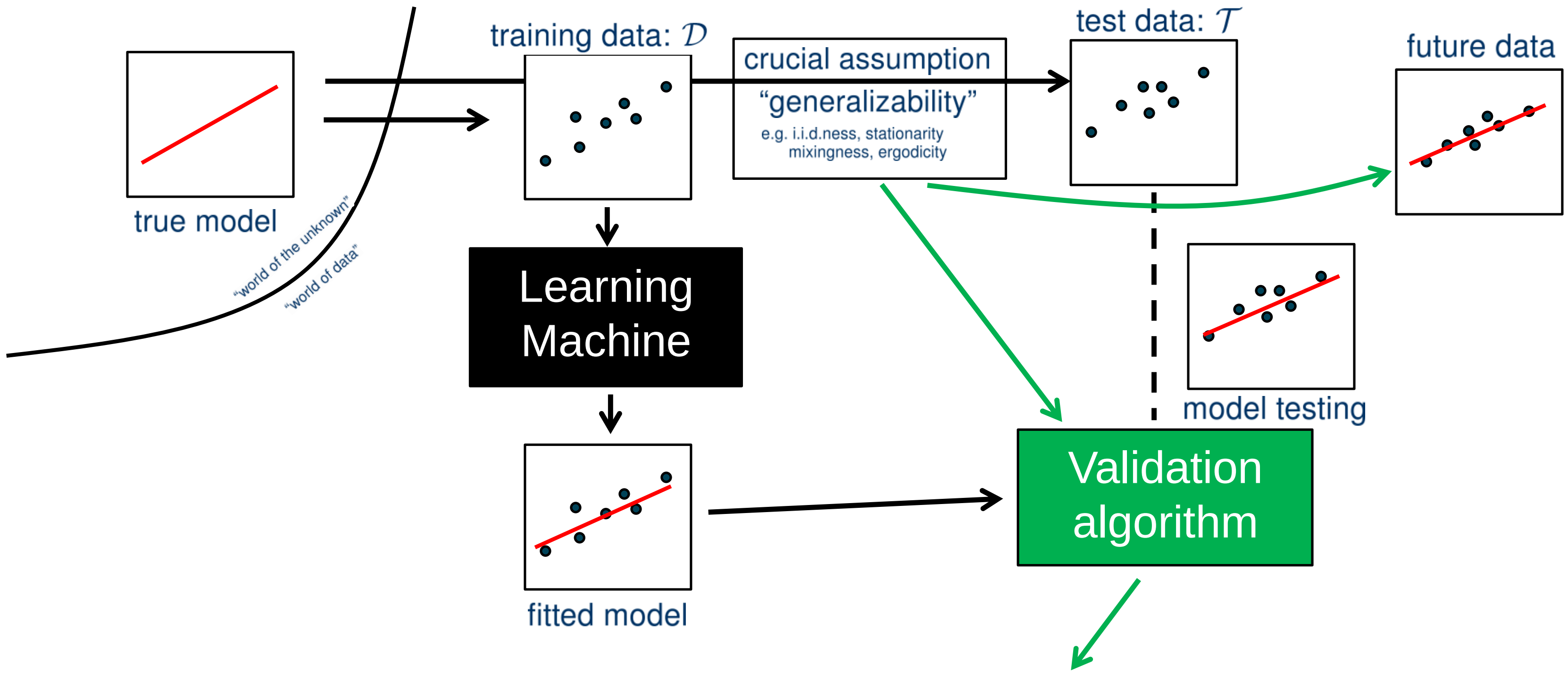


Guarantees implied by *assumptions on data/ and model properties*
in the form of mathematical/statistical theorems

Validity argument incomplete unless these are checked!

(ii) model-agnostic validity arguments

„model validation“, „model checking“, „model testing“



Guarantees implied by *empirical results and properties of the task*

in the form results plus theorems about the validation algorithm

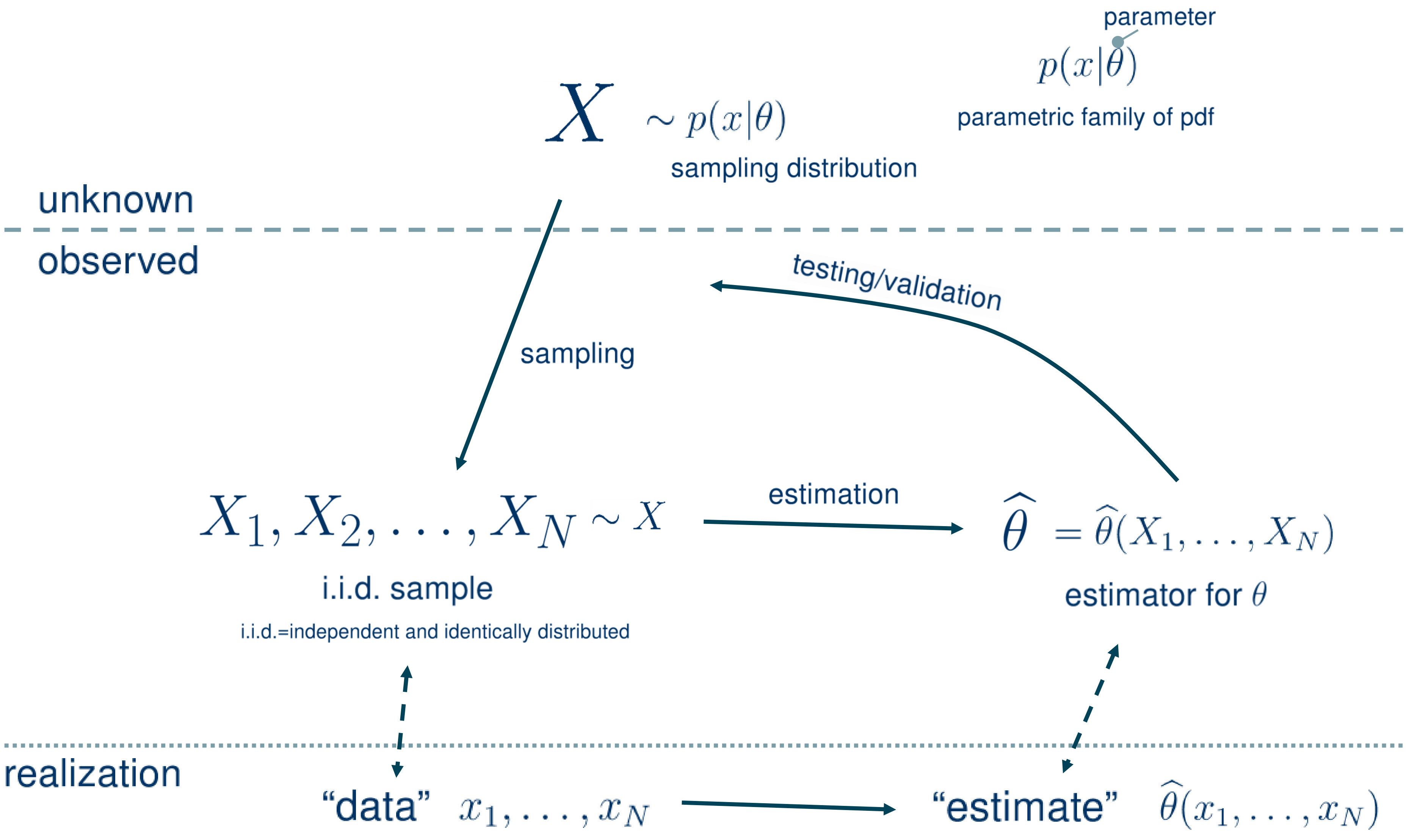
Guarantees always hold and allow comparison

but may be weaker than model-specific ones if theory available

Table of contents for lecture 2

Generalization guarantees

Reminder: Parametric Estimation



The parametric estimation setting

Data is generated as:

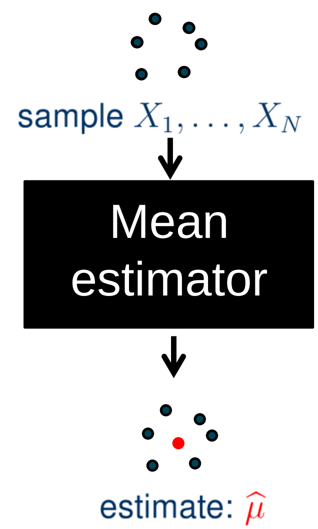
true parameter $\theta \in \mathbb{R}^m$ (e.g. $\theta = (\mu, \Sigma)$ for Gaussians)

$$X_1, \dots, X_N \underset{\text{i.i.d.}}{\sim} X \sim p(.|\theta)$$

(parametric) estimator $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_N)$

e.g. mean estimator $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$

covariance estimator $\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$



Accurate terminology: $\widehat{\theta}(X_1, \dots, X_N)$ is *„the estimate“* (a RV t.v.in $\mathbb{R}^m$)

$\widehat{\theta} : (x_1, \dots, x_N) \mapsto \widehat{\theta}(x_1, \dots, x_N)$ is *„the estimator“* (an algorithm)

function in $[\mathcal{X}^n \to \mathbb{R}^m]$

often confounded, but distinction will become important in later section

Guarantees: LLN, CLT and CI

hold in general, nonparametric setting $X_1, \dots, X_N \underset{\text{i.i.d.}}{\sim} X$ t.v.in $\mathbb{R}$ (assume finite moments)

$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ estimates $\mathbb{E}[X]$ $\qquad$ $\widehat{\Sigma} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \widehat{\mu})^2$ estimates $\mathrm{Var}[X]$

$\frac{\widehat{\Sigma}}{N}$ estimates $\mathrm{Var}[\widehat{\mu}]$ „variance of the sample mean“

Theorems: (i) $\mathbb{E}[\widehat{\mu}] = \mathbb{E}[X]$ $\qquad$ $\mathrm{Var}[\widehat{\mu}] = \frac{1}{N}\mathrm{Var}[X]$ (weak LLN implied by this & Chebyshev)

(statements you should have seen!)

(ii) $\sqrt{N}\left(\widehat{\mu} - \mathbb{E}[X]\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}(X)\right)$ as $N \to \infty$ (CLT as usually stated)

(iii) $\mathbb{E}[\widehat{\Sigma}] = \mathrm{Var}[X]$ $\qquad$ $\mathbb{E}[\widehat{\Sigma}/N] = \mathrm{Var}[\widehat{\mu}]$ (unbiasedness of estimates)

(statements you probably haven‘t seen but are simple computations)

(iv) $\mathrm{Var}[\widehat{\Sigma}] = \frac{1}{N}\cdot\left[M_4 - \frac{N-3}{N-1}\mathrm{Var}(X)^2\right]$ where $M_4 := \mathbb{E}\left[(X - \mathbb{E}(X))^4\right]$ (LLN for sample variance)

(v) $\sqrt{N}\left(\widehat{\Sigma} - \mathrm{Var}[X]\right) \xrightarrow{d} \mathcal{N}\left(0, \underbrace{M_4 - \mathrm{Var}(X)^2}_{= \mathrm{Var}\left[(X - \mathbb{E}[X])^2\right]}\right)$ as $N \to \infty$ (CLT for sample variance)

(stylized)
Consequence: $\left[\widehat{\mu} + \Phi^{-1}(\alpha/2)\cdot\sqrt{\widehat{\Sigma}/N}\,, \widehat{\mu} - \Phi^{-1}(\alpha/2)\cdot\sqrt{\widehat{\Sigma}/N}\right]$ is „good“ CI

i.e., α-CI for μ with good coverage probability as $N > 50$

Important exception: imbalanced & binary *X*

Estimation of the generalization error

Setting: **i.i.d. *test* data** $(X_1, Y_1), \dots, (X_M, Y_M) \underset{\text{i.i.d.}}{\sim} (X, Y)$ t.v.in $\mathcal{X} \times \mathcal{Y}$

prediction functional $f : \mathcal{X} \to \mathcal{Y}$ e.g., $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}$

loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ e.g., $L : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$

To estimate: $\varepsilon(f) = \mathbb{E}\left[L(f(X), Y)\right]$ *expected generalization loss*

Theorems suggest following estimators:

$$\widehat{\varepsilon}(f) := \frac{1}{M} \sum_{i=1}^{M} L(f(X_i), Y_i)$$

„empirical loss“

Observation: $L_i := L(f(X_i), Y_i)$ are i.i.d.
since pairs (X_i, Y_i) are i.i.d.

$$\widehat{v}(f) := \frac{1}{M-1} \sum_{i=1}^{M} (L_i - \widehat{\varepsilon})^2$$

„sample variance of the empirical losses“

Confidence interval: $\left[\widehat{\varepsilon}(f) + \Phi^{-1}(\alpha/2) \cdot \sqrt{\widehat{v}(f)/M}\ , \widehat{\varepsilon}(f) - \Phi^{-1}(\alpha/2) \cdot \sqrt{\widehat{v}(f)/M}\ \right]$

The end ... ?

Big problem: this is only valid if f is fixed (non-random), e.g., already trained/fitted!

Otherwise L_i are dependent through a random f. *(no guarantees for strategies!)*

But statements & guarantees about the trained prediction functionals are correct!

Bias and Variance in Parametric Estimation and Supervised Learning

Supervised learning as function estimation

(more restrictive formulation of the task due to stronger structural assumptions)

Given data from generative, unknown RV (X, Y) t.v.in $\mathcal{X} \times \mathcal{Y}$ (not independent of each other in general)

features labels domains

where we observe

$$(X_1, Y_1), \dots, (X_N, Y_N) \underset{\text{i.i.d.}}{\sim} (X, Y)$$

Parametric supervised assumption:

There is a „true“ labelling process $f = f_\theta$

(usually, one assumes an „additive error model“, that is:)

$$Y_i = f(X_i) + \varepsilon_i \qquad \varepsilon_i \text{ is error with } \mathbb{E}[\varepsilon_i] = 0$$

(errors assumed independent)

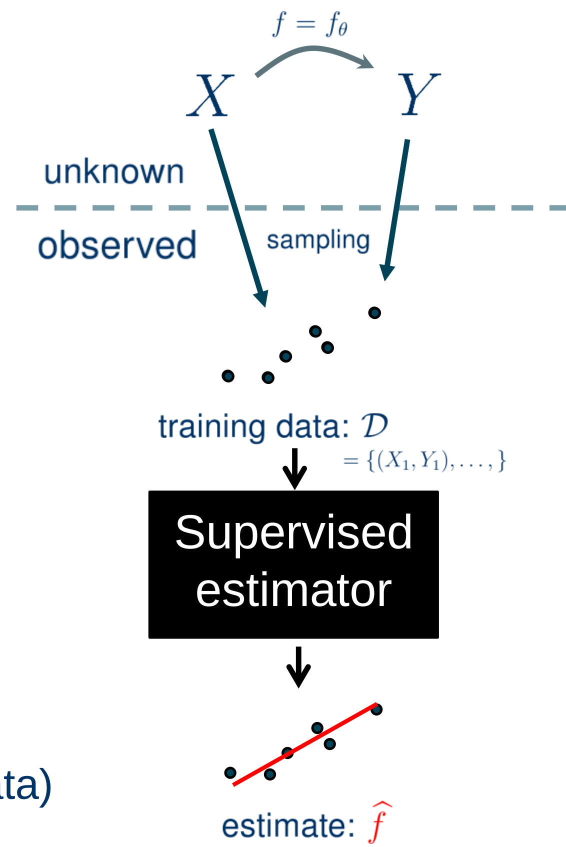
Example (Linear Regression):

$$f(x) = \beta^\top x + \alpha, \qquad \theta = (\alpha, \beta)$$

Supervised estimation task:

Learn a good approximation $\widehat{f} = f_{\widehat{\theta}}$ (using the training data)

such that $\varepsilon\left(\widehat{f}\right) = \mathbb{E}\left[L\left(\widehat{f}(X), Y\right)\right]$ is small



How good is an estimator?
Bias and variance

parametric estimation:

true parameter $\theta \in \mathbb{R}^m$ (e.g. $\theta = (\mu, \Sigma)$ for Gaussians)
estimator $\widehat{\theta} = \widehat{\theta}(X_1, \dots, X_N)$

bias of $\widehat{\theta}$: $\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$

measures expected deviation of the mean

variance of $\widehat{\theta}$: $\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$

measures scatter around estimator mean

MSE of $\widehat{\theta}$: $\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$

measures total estimation error

Convention: for $x \in \mathbb{R}^m$, denote $x^2 = x^\top x$

Cave: all these quantities may depend on θ

(yes, this includes the variance. Think why.)

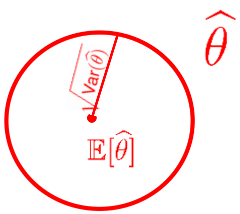
Bias and variance

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$$

measures expected deviation of the mean

$$\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$$

measures scatter around estimator mean



	low bias	high bias	
	θ 	 θ	low variance
	θ 	 θ	high variance

Bias-variance trade-off for estimators

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta]$$

onmeasures expected deviati

$$\text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right]$$

measures scatter around estimator mean

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$$

measures total estimation error

Proposition: $\text{MSE}(\widehat{\theta}) = \text{Bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta})$

proof:

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] = \mathbb{E}[\widehat{\theta^2}] - 2\theta\mathbb{E}[\widehat{\theta}] + \theta^2 \\
&= \mathbb{E}[\widehat{\theta^2}] - 2\left(\mathbb{E}[\widehat{\theta}]\right)^2 + \left(\mathbb{E}[\widehat{\theta}]\right)^2 + \left(\mathbb{E}[\widehat{\theta}]\right)^2 - 2\theta\mathbb{E}[\widehat{\theta}] + \theta^2 \\
&= \mathbb{E}[\widehat{\theta^2}] - 2\mathbb{E}\left[\widehat{\theta}\left(\mathbb{E}[\widehat{\theta}]\right)\right] + \mathbb{E}\left(\mathbb{E}[\widehat{\theta}]\right)^2 + \left(\mathbb{E}[\widehat{\theta}] - \theta\right)^2 \\
&= \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right] + \left(\mathbb{E}[\widehat{\theta}] - \theta\right)^2 \\
&= \text{Var}(\widehat{\theta}) + \text{Bias}(\widehat{\theta})^2
\end{aligned}$$

stays also valid for taking $xx^\top$ instead of $x^\top x$

Example: mean of Gaussian

$$\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta} - \theta] \qquad \text{Var}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2\right] \qquad \text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right]$$

$$\text{MSE}(\widehat{\theta}) = \text{Bias}(\widehat{\theta})^2 + \text{Var}(\widehat{\theta})$$

Example: estimation of mean

$X_1, \ldots, X_N$ i.i.d Gaussian $\sim \mathcal{N}(\mu, \sigma^2)$

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

"natural" estimator

$$\widehat{\mu}_s = 42$$

"universal" estimator

$$\text{Bias}(\widehat{\mu}) = 0 \qquad \text{MSE}(\widehat{\mu}) = \text{Var}(\widehat{\mu}) = \frac{\sigma^2}{N}$$

$$\text{Bias}(\widehat{\mu}_s) = 42 - \theta \qquad \text{Var}(\widehat{\mu}_s) = 0 \qquad \text{MSE}(\widehat{\mu}_s) = (42 - \theta)^2$$

Supervised learning as function estimation

Given data from generative, unknown RV (X, Y) t.v.in $\mathcal{X} \times \mathcal{Y}$ (not independent of each other in general)

features labels domains

where we observe

$$(X_1, Y_1), \ldots, (X_N, Y_N) \underset{\text{i.i.d.}}{\sim} (X, Y)$$

Parametric supervised assumption:

There is a „true" labelling process $f = f_\theta$

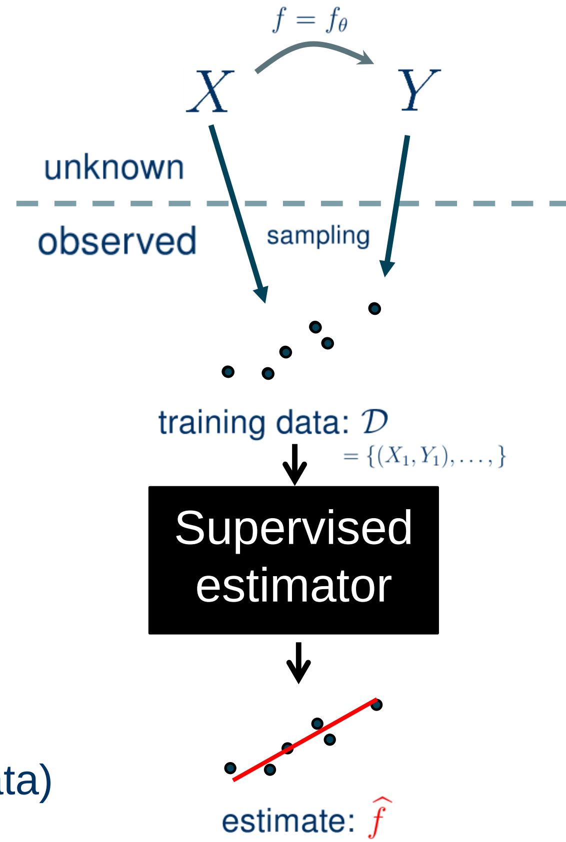
(usually, one assumes an „additive error model", that is:)

$$Y_i = f(X_i) + \epsilon_i \qquad \epsilon_i \text{ is error with } \mathbb{E}[\epsilon_i] = 0$$

(errors assumed independent)

Example (Linear Regression):

$$f(x) = \beta^\top x + \alpha, \qquad \theta = (\alpha, \beta)$$



Supervised estimation task:

Learn a good approximation $\widehat{f} = f_{\widehat{\theta}}$ (using the training data)

such that $\varepsilon\left(\widehat{f}\right) = \mathbb{E}\left[L\left(\widehat{f}(X), Y\right)\right]$ is small

Bias and variance in supervised regression

assume $\mathcal{Y} = \mathbb{R}$

test point X_*, test label Y_* where $(X_*, Y_*) \sim (X, Y)$

parameter θ "is" a generative function $f = f_\theta$:

$$Y_* = f(X_*) + \epsilon_*$$

$\widehat{f} = f_{\widehat{\theta}}$ learnt prediction rule (possibly depending on seen training data)

bias of $\widehat{f}$ at X_*: $\quad \text{Bias}(\widehat{f}|X_*) = \mathbb{E}_{Y|X_*}[\widehat{f}(X_*) - f(X_*)|X_*]$

$$= \mathbb{E}_{Y|X_*}[\widehat{f}(X_*) - Y_*|X_*]$$

variance of $\widehat{f}$ at X_*: $\quad \text{Var}(\widehat{f}|X_*) = \text{Var}_{Y|X_*}\left[\widehat{f}(X_*)|X_*\right] = \mathbb{E}_{Y|X_*}\left[(\widehat{f}(X_*)^2|X_*\right] - \mathbb{E}_{Y|X_*}\left[\widehat{f}(X_*)|X_*\right]^2$

MSE of $\widehat{f}$ at X_*: $\quad \text{MSE}(\widehat{f}|X_*) = \mathbb{E}_{Y|X_*}\left[(\widehat{f}(X_*) - Y_*)^2|X_*\right]$

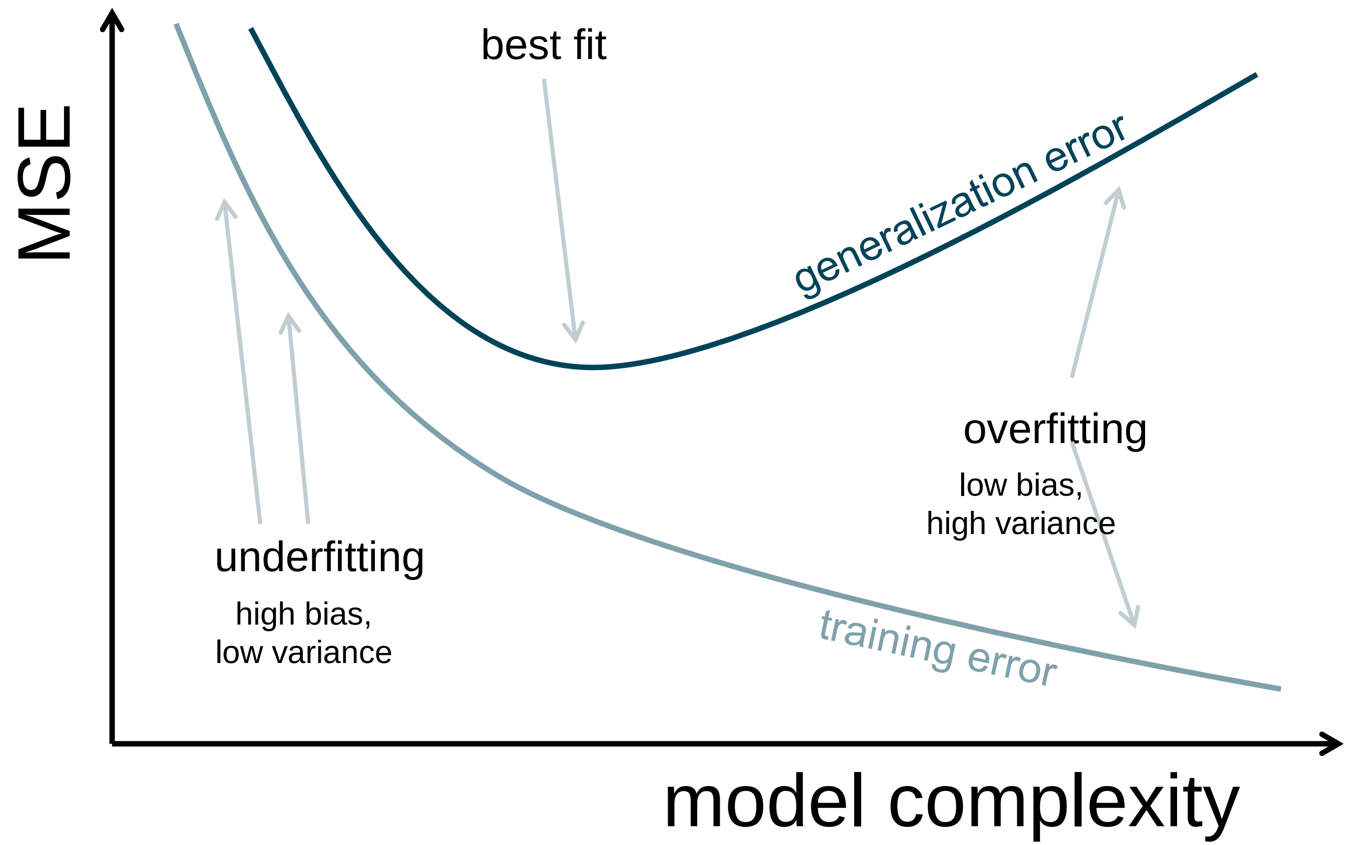
$$= \mathbb{E}_{Y|X_*}\left[L(\widehat{f}(X_*), Y_*)|X_*\right] \text{ for } L : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$$

Proposition: $\text{MSE}(\widehat{f}|X_*) = \text{Var}(\epsilon_*) + \text{Bias}(\widehat{f}|X_*)^2 + \text{Var}(\widehat{f}|X_*)$

expected out-of-sample- MSE „irreducible error“ from measurement noise Bias and variance of prediction

(proof in analogy to earlier bias-variance)

The Bias-variance-trade-off in prediction



Bias-variance-trade-off

An experiment from

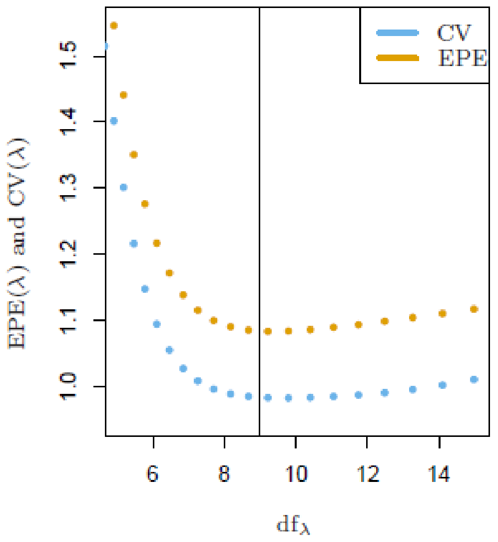
The Elements of Statistical Learning (Section 5.5)

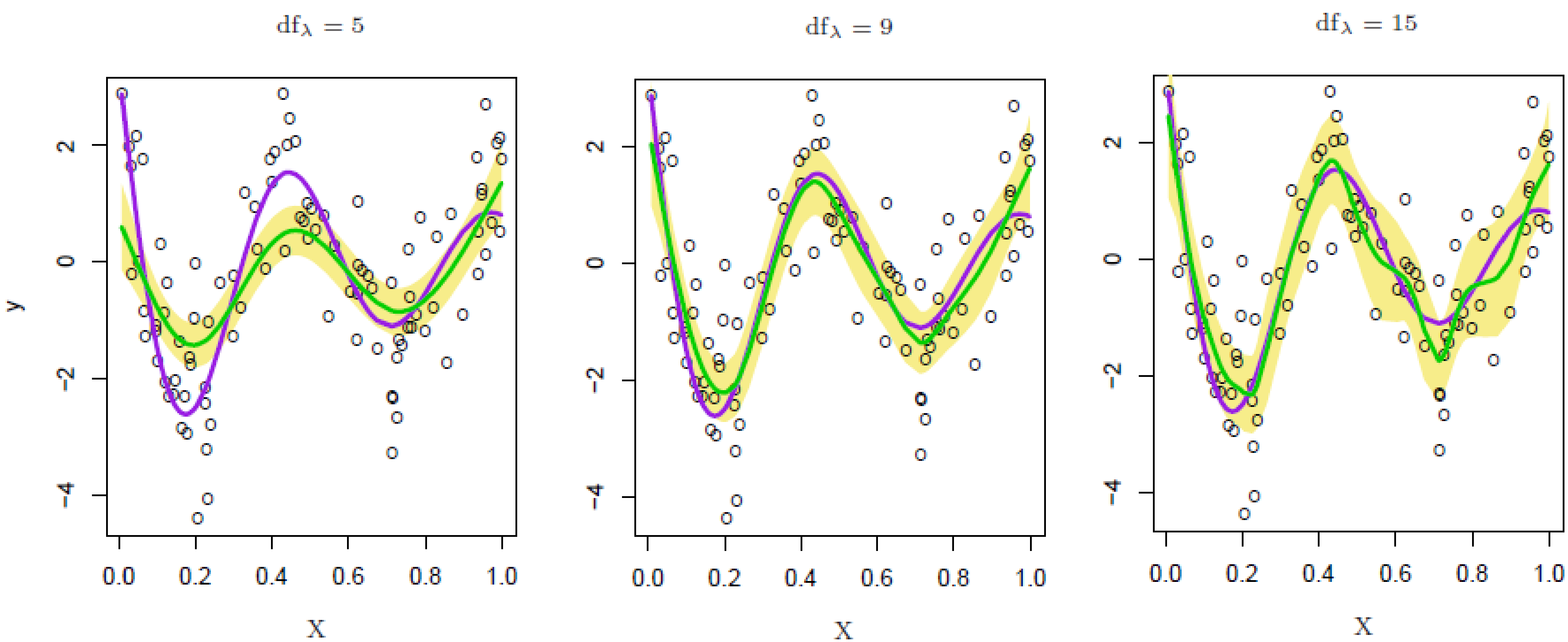
Interpolation using regularized splines

Strength of regularization set by parameter df_λ

Prediction variance grows with decreasing df_λ

Prediction bias decreases with increasing df_λ





Over-optimism of the training error

bias of $\widehat{f}$ at X_*: $\text{Bias}(\widehat{f}|X_*) = \mathbb{E}_{Y|X_*}[\widehat{f}(X_*) - f(X_*)|X_*]$

variance of $\widehat{f}$ at X_*: $\text{Var}(\widehat{f}|X_*) = \text{Var}_{Y|X_*}\left[\widehat{f}(X_*)|X_*\right]$

MSE of $\widehat{f}$ at X_*: $\text{MSE}(\widehat{f}|X_*) = \mathbb{E}_{Y|X_*}\left[(\widehat{f}(X_*) - Y_*)^2|X_*\right]$

Proposition: $\text{MSE}(\widehat{f}|X_*) = \text{Var}(\epsilon_*) + \text{Bias}(\widehat{f}|X_*)^2 + \text{Var}(\widehat{f}|X_*)$

is true only if ϵ_* is independent of $\widehat{f}$

i.e., if (X_*, Y_*) has not already been seen in the training set

(this was assumed on the earlier slide)

If it has been *seen*, i.e., we test on the training set, then:

$$\text{MSE}(\widehat{f}|X_*) = \text{Var}(\varepsilon_*) + \text{Bias}(\widehat{f}|X_*)^2 + \text{Var}(\widehat{f}|X_*)$$

(expected training MSE)

$$-2 \cdot \text{Cov}(\widehat{f}(X_*)|X_*, Y_*|X_*)$$

Sensible learning machines‘ prediction usually co-varies with observations

hence **expected training MSE < expected test MSE**

similar statements hold in more general settings in terms of noise and error statistic

Explicit form for Ordinary Least Squares

bias of $\widehat{f}$ at X_*: $\quad \text{Bias}(\widehat{f}|X_*) = \mathbb{E}_{Y|X_*}[\widehat{f}(X_*) - f(X_*)|X_*]$

variance of $\widehat{f}$ at X_*: $\quad \text{Var}(\widehat{f}|X_*) = \text{Var}_{Y|X_*}\left[\widehat{f}(X_*)|X_*\right]$

MSE of $\widehat{f}$ at X_*: $\quad \text{MSE}(\widehat{f}|X_*) = \mathbb{E}_{Y|X_*}\left[(\widehat{f}(X_*) - Y_*)^2|X_*\right]$

Linear Regression: $f(x) = \beta^\top x + \alpha,\ Y_i = f(X_i) + \varepsilon_i,\ \widehat{f}(x) = \widehat{\beta}^\top x + \widehat{\alpha}$

where objects with hats may (but don‘t need to) depend on training data

OLS: $\mathbb{E}[\widehat{\beta}] = \beta,\ \mathbb{E}[\widehat{\alpha}] = \alpha$ hence $\text{Bias}(\widehat{f}|X_*) = \langle \mathbb{E}[\widehat{\beta}] - \beta, X_*\rangle + \mathbb{E}[\widehat{\alpha}] - \alpha = 0$

for variance, use decomposition $\widehat{f}(x) = \overline{Y} + \widehat{\beta}^\top(x - \overline{X})$

so $\text{Var}(\widehat{f}|X_*) = \text{Var}(\overline{Y}|X_*) + \text{Var}(\widehat{\beta}^\top(X_* - \overline{X})|X_*)$

„error in estimating regression line“

$$= \sigma^2\left(\frac{1}{N} + (X_* - \overline{X})^\top C_{xx}^{-1}\right)(X_* - \overline{X}) = \frac{\sigma^2}{N}\left(1 + (X_* - \overline{X})^\top\Sigma^{-1}(X_* - \overline{X})\right)$$

Full MSE decomposition: $\text{MSE}(\widehat{f}|X_*) = \sigma^2 + 0 + \frac{\sigma^2}{N}\left(1 + (X_* - \overline{X})^\top\Sigma^{-1}(X_* - \overline{X})\right)$

Overfitting optimism for already seen test point X_* : $\quad (=$ coefficient of ε_* in $\widehat{f}(X_*))$

$$\text{Cov}(\widehat{f}(X_*)|X_*, Y_*|X_*) = \text{Cov}(\widehat{f}(X_*)|X_*, \varepsilon_*|X_*) = \frac{\sigma^2}{N}\left(1 + (X_* - \overline{X})\cdot\Sigma^{-1}\cdot(e_i - \frac{1}{N})\right)$$

Unconditional bias-variance trade-off

test point X_*, test label Y_* where $(X_*, Y_*) \sim (X, Y)$

unconditional quantities: taking expectations

total bias of $\widehat{f}$: $\text{Bias}^2(\widehat{f}) := \mathbb{E}\left[\text{Bias}^2(\widehat{f}|X_*)\right] \neq \mathbb{E}\left[\text{Bias}(\widehat{f})|X_*\right]^2$

total variance of $\widehat{f}$: $\text{Var}(\widehat{f}|X_*) := \mathbb{E}\left[(\widehat{f}(X_*) - \mathbb{E}[\widehat{f}(X_*)])^2\right]$

total MSE of $\widehat{f}$: $\text{MSE}(\widehat{f}) := \mathbb{E}\left[(\widehat{f}(X_*) - Y_*)^2\right] = \varepsilon(\widehat{f})$

Proposition: $\text{MSE}(\widehat{f}) = \text{Var}(\epsilon_*) + \text{Bias}^2(\widehat{f}) + \text{Var}(\widehat{f})$

Proof: take expectations in conditional trade-off

Note: the „obvious“ unconditional generalization

$$\text{MSE}(\widehat{f}) = \text{Var}(\epsilon_*) + \text{Bias}(\widehat{f})^2 + \text{Var}(\widehat{f})$$

where $\text{Bias}(\widehat{f}) := \mathbb{E}\left[\text{Bias}(\widehat{f}|X_*)\right]$ *is **wrong** in general!*

Prediction functionals vs prediction strategies

test point X_*, test label Y_* where $(X_*, Y_*) \sim (X, Y)$

unconditional quantities: taking expectations

Re-sampling strategies

Re-sampling

Given data vector $\mathcal{D} = (X_1, \dots, X_N) \quad X_1, \dots, X_N \sim X$ t.v.in $\mathcal{X}$

estimator $\widehat{\theta} : \mathcal{X}^* \to \mathbb{R}$ $\quad \mathcal{X}^* =$ vectors in $\mathcal{X}$ of arbitrary length

estimate $\widehat{\theta}(\mathcal{D})$

„Re-sampling estimator“ is constructed from re-samples

$\widehat{\theta}(\mathcal{D}[\pi_1]), \dots, \widehat{\theta}(\mathcal{D}[\pi_k])$

(random!)

where π_i t.v.in $\{1, \dots, N\}^*$ are re-sampling index vectors

Important cases:

π_i are i.i.d. random with/without replacement of fixed size m

π_i invariant under permutation e.g., missing block of size k non-overlapping

Important applications:

obtaining an improved version of $\widehat{\theta}$, e.g., variance-reduced

obtaining non-parametric estimates of expectation and variance

Estimation of the generalization error

Setting: **i.i.d. *test* data** $(X_1, Y_1), \dots, (X_M, Y_M) \underset{\text{i.i.d.}}{\sim} (X, Y)$ t.v.in $\mathcal{X} \times \mathcal{Y}$

prediction functional $f : \mathcal{X} \to \mathcal{Y}$ e.g., $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}$

loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ e.g., $L : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$

To estimate: $\varepsilon(f) = \mathbb{E}\left[L(\widehat{f}(X), Y)\right]$ *expected generalization loss*

Theorems suggest following estimators:

$$\widehat{\varepsilon}(f) := \frac{1}{M} \sum_{i=1}^{M} L(f(X_i), Y_i)$$

„empirical loss“

Observation: $L_i := L(f(X_i), Y_i)$ are i.i.d.
since pairs (X_i, Y_i) are i.i.d.

$$\widehat{v}(f) := \frac{1}{M(M-1)} \sum_{i=1}^{M} (L_i - \widehat{\varepsilon})^2$$

„standard error of the empirical loss“

Confidence interval: $\left[\widehat{\varepsilon}(f) + \Phi^{-1}(\alpha/2) \cdot \sqrt{\widehat{v}(f)}\ , \widehat{\varepsilon}(f) - \Phi^{-1}(\alpha/2) \cdot \sqrt{\widehat{v}(f)}\ \right]$

The end ... ?

Big problem: this is only valid if f is constant, e.g., already trained/fitted!
Otherwise L_i are dependent through a random f. *(no guarantees for strategies!)*

But statements & guarantees about the trained prediction functionals are correct!

Estimation of the generalization error

Setting: **i.i.d. data** $(X_1, Y_1), \dots, (X_M, Y_M) \underset{\text{i.i.d.}}{\sim} (X, Y)$ t.v.in $\mathcal{X} \times \mathcal{Y}$

prediction strategy $f_\mathcal{T}$ t.v.in $[\mathcal{X} \to \mathcal{Y}]$ "trained" on subset $(X_i, Y_i), i \in \mathcal{T}$

loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ e.g., $L : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$

To estimate: $\varepsilon(f) = \mathbb{E}\left[L(f_\mathcal{T}(X), Y)\right]$ *expected generalization loss*

One-split estimation: training indices $\mathcal{T} \subseteq \{1, \dots, N\}$ with $\#\mathcal{T} = M$

test indices $\mathcal{V} = \{1, \dots, N\} \setminus \mathcal{T}$

$$\widehat{\varepsilon}(f|\mathcal{T}) := \frac{1}{\#\mathcal{V}} \sum_{j \in \mathcal{V}} L(f_\mathcal{T}(X_j), Y_j)$$

„out-of-sample estimated“ since $f_\mathcal{T}$ indep. of $\{(X_i, Y_i) \ : \ i \in \mathcal{T}\}$

„empirical loss“

Corollary of CLT: $\mathbb{E}\left[\widehat{\varepsilon}(f|\mathcal{T})\right] = \varepsilon(f)$ i.e., $\widehat{\varepsilon}(f|\mathcal{T})$ unbiasedly estimates $\varepsilon(f)$

$$\widehat{\varepsilon}(f|\mathcal{T}) \xrightarrow{p} \varepsilon(f|\mathcal{T}) = \mathbb{E}[L(f_\mathcal{T}(X), Y)|f]$$ when $N \to \infty$ (with $\mathcal{T}$ fixed)

i.e., $\widehat{\varepsilon}(f|\mathcal{T})$ consistently estimates $\varepsilon(f|\mathcal{T})$

Problem: one-split estimation depends on single training set and single run

so variance of the one-split estimator may be high if algorithm is unstable

Solution: Re-sample averaging, $\widehat{\varepsilon}_{CV} = \frac{1}{K} \sum_{\kappa=1}^{K} \widehat{\varepsilon}(f|\mathcal{T}_\kappa)$

Potential issue:

$\widehat{\varepsilon}(f|\mathcal{T}_\kappa)$ are correlated

Variance reduction by averaging

Variance reduction lemma: for any number of correlated random variables

$Z_1, \ldots, Z_M$ with $\text{Var}(Z_i) = \sigma^2$ and $\text{Corr}(Z_i, Z_j) = \rho$ for $i \neq j$

one has $$\text{Var}\left(\frac{1}{M}\sum_{i=1}^{M} Z_i\right) = \rho \cdot \sigma^2 + \frac{1-\rho}{M} \cdot \sigma^2 \quad \lneq \sigma^2$$ if $\rho > 0$ (unless Z_i are equal)

So if Z_i are correlated re-samples with the same bias

averaging reduces the variance while not changing the bias

Hence by bias-variance trade-off, expected mean-squared-error is reduced

For cross-validation estimator $$\widehat{\varepsilon}_{CV} = \frac{1}{K}\sum_{\kappa=1}^{K} \widehat{\varepsilon}(f|\mathcal{T}_\kappa)$$

(any type of CV or re-sampling where folds are exchangeably sampled!)

$$\text{Var}(\widehat{\varepsilon}_{CV}) = \left(\rho + \frac{1-\rho}{K}\right) \cdot \text{Var}\left(\widehat{\varepsilon}[\mathcal{T}]\right)$$ where $$\rho = \text{Corr}\left(\widehat{\varepsilon}(f|\mathcal{T}_i), \widehat{\varepsilon}(f|\mathcal{T}_j)\right)$$ for $i \neq j$

„variance of CV“ „1 or smaller“ „variance of one-split“ „correlation of fold-wise estimates“

Since one-split estimates are unbiased, CV-estimator also is.

Hence by bias-variance trade-off, resample averaging reduces expected error.

Estimating the CV performance estimates‘ variance

Why? $\widehat{\varepsilon}_{CV}$ is an estimate for the expected prediction error of algorithm f

Variance is needed for *confidence intervals*; also useful in *comparing* strategies

Naive approach: $\widehat{v}[\mathcal{T}] := \frac{1}{\#\mathcal{V}\cdot(\#\mathcal{V}-1)}\sum_{j\in\mathcal{V}}\left(L(f_{\mathcal{T}}(x_j),Y_j)-\widehat{\varepsilon}(f|\mathcal{T})\right)^2$

is an (unbiased, consistent) estimator for $v\left[\mathcal{T}\right] = \text{Var}\left[L(f(X|\mathcal{T}),Y)|\mathcal{T}\right]$

Problem (?): conditional on training set, so variance from that is not included

may be fine if algorithm is „stable“ w.r.t. training set choice and repetitions

Bad but frequently seen approach: sample variance of $\widehat{\varepsilon}(f|\mathcal{T}_i),\ i=1\dots K$

LLN/CLT does not apply: K is usually too small, and sample is correlated!

Averaging approach: $\widehat{v}_{CV} := \frac{1}{K}\sum_{\kappa=1}^{K}\widehat{v}[\mathcal{T}_i]$ reduces variance from training set

(by variance reduction lemma)

Problem: this is not the variance of $\widehat{\varepsilon}_{CV}$ which it usually overestimates!

(but being conservative is fine for avoiding type I errors)

Theorem: There is no unbiased estimator for the variance of $\widehat{\varepsilon}_{CV}$ (of certain form)

(Bengio, Grandvalet 2004) Key realization: non-identifiability of inter-fold co-variance

There could be biased estimators with low MSE… but no substantial (?) progress since.

Black-box estimates of mean and variance

Given data vector $\mathcal{D} = (X_1, \dots, X_N)$ $\quad X_1, \dots, X_N \underset{\text{i.i.d.}}{\sim} X$ t.v.in $\mathcal{X}$

estimator $\widehat{\theta} : \mathcal{X}^* \to \mathbb{R}$ $\qquad$ estimate $\widehat{\theta}(\mathcal{D})$

Jackknife estimation of $\mathbb{E}\left[\widehat{\theta}(\mathcal{D})\right]$ and Var $\left[\widehat{\theta}(\mathcal{D})\right]$

$\widehat{\theta}_{/i} := N \cdot \widehat{\theta}(\mathcal{D}) - (N-1) \cdot \widehat{\theta}(\mathcal{D}_{/i})$ where $\mathcal{D}_{/i} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_N)$

„ordinary jackknife pseudo-samples“ $\qquad$ „dataset minus i-th data point“

Bootstrap estimation of Var $\left[\widehat{\theta}(\mathcal{D})\right]$

$B_i := \mathcal{D}[\pi_i]$ where $\pi_1, \dots, \pi_B$ are i.i.d. t.v.in $\mathbb{N}^N$

$\pi_i = (n_1, \dots, n_N),\ n_i \underset{\text{i.i.d.}}{\sim}$ Unif$\{1, \dots, N\}$

Sample mean and variance are „good“ estimates of $\mathbb{E}\left[\widehat{\theta}(\mathcal{D})\right]$ and Var $\left[\widehat{\theta}(\mathcal{D})\right]$

under certain regularity assumptions on $\widehat{\theta}$ $\qquad$ (these are often unclear in literature!)

Applicability: Re-sample estimates of full training/test variance

Confidence intervals for loss statistics which are not means

Cave: Regularity conditions may be strong! Inappropriate for medians/quantiles

Testing & Comparison

Principles of model comparison testing

First interesting and simpler case: prediction *functionals*

Setting: **i.i.d. *test* data** $(X_1, Y_1), \dots, (X_M, Y_M) \underset{\text{i.i.d.}}{\sim} (X, Y)$ t.v.in $\mathcal{X} \times \mathcal{Y}$

prediction functionals $f_1, \dots, f_S : \mathcal{X} \to \mathcal{Y}$ e.g., $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}$

loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ e.g., $L : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$

Testing approach: obtain loss residuals

$$L_{ij} := L(f_j(X_i), Y_i),\ i = 1 \dots M,\ j = 1 \dots S$$

the S samples $\mathcal{S}_j := (L_{1j}, \dots, L_{Mj})$ are tupled/paired

any paired and tupled portmanteau test for location comparison is applicable

Recommendation: non-parametric Wilcoxon (rank-sign) test

„comparing confidence intervals“ is also fine (unpaired z-test or t-test, conservative)

Prediction strategies? *Unclear what to do.*

Nadeau, Bengio (2004) discuss a few options based on the t-test

Re-sampling compatibility? *2x median p-value is conservative aggregation.*

Generally for exchangeable re-samples: $\gamma^{-1} p_{(\gamma k)}$ is conservative aggregation of $(p_1, \dots, p_k)$

WORQ - widely open research questions!

(aka MSc/PhD topics for the ambitious and theoretically inclined)

Theory of variance estimators for black-box functionals

- Bootstrap & re-sampling seem to be the only semi-solid strategies
- Also, just semi-solid, without source that has exact assumptions (?)
- How best to separate test set variance and training set variance?

Theory of variance estimators for re-sampled statistics

- Bengio/Grandvalet: no unbiased estimator *of a special form*
- This does not preclude a good estimate of another, simple form
- ... such as re-sampling the re-sample statistic (??)

Do all this for the „complicated tasks“

- Time series, on-line learning, anomaly detection, reinforcement learning
- Structured and heterogeneous prediction tasks
- Probabilistic and Bayesian modelling (credibility intervals?)

Hypothesis testing & portmanteau comparison

- Best way for strategies unclear – how to incorporate training set variance?

Misunderstandings and Statistical Learning Theory

Frequently mis-interpreted results

The No Free Lunch Theorems (Wolpert and Macready, 1997 onwards)

… for all data there is a model, for all models there is data…

Frequent mis-interpretation:

All of statistics and machine learning is arbitrary anyway.

More correct interpretation:

No meaningful definition of „learning“ is possible without assumptions on how training and test data relate. For example, they should be similar (e.g. distributionally).

Shao – Linear Model Selection by Cross-Validation (1993)

… some types of CV fail to identify the „correct“ model…

Frequent mis-interpretation:

Cross-validation should not be done. Or: Bayesian statistics is the only way.

More correct interpretation:

Model identification is more difficult than accurate prediction, sometimes considerably so.

Bengio, Grandvalet – *No Unbiased Estimator for the Variance of K-Fold Cross-Validation* (2004)

Frequent mis-interpretation:

Confidence intervals for error metrics cannot be computed, quantitative comparisons between different methods are futile.

More correct interpretation:

Predictions on different folds are correlated, one needs to be careful in aggregating them.

Recall: all models are wrong (George Box), but some are useful.
This is similarly true for meta-methods and model checking.

Overview: Model-specific Learning Theory

multiple „flavours“ of model-specific guarantees based on „model class complexity“

bound generalization loss $R(f) = \varepsilon(f)$ by empirical *training* loss $R_N(f)$ plus “complexity term”

approach/field	***Scope and assumptions***	***some notable statements***
Statistical Learning Theory **Vapnik, Chervonenkis**	Training and test data follow the *same* distribution Learning machine is in a fixed set of functions $\mathcal{F}$ Questions: asymptotic behaviour of machine Relation between training and test error	$R(f) \leq R_N(f) + \frac{3}{\sqrt{N}} \cdot \sqrt{\log S_{\mathcal{F}}(2n) + \log \frac{2}{\varepsilon}}$ for any learner f with probability $\geq 1-\varepsilon$ $R(f), R_N(f)$ expected and empirical loss/risk $S_{\mathcal{F}}(n)$ number of classification rules on n points general extensions via Rademacher/covering theory
Bayesian/ Parametric	Training and test data follow the *same* distribution Learning machine in parametric function class $\mathcal{F}$ Predictions are distributional „posterior“	AIC $= -2\mathcal{L}(f) + 2d$ BIC $= -2\mathcal{L}(f) + d\log N$ „Akaike information criterion“ „Bayesian IC“ for many simple model classes: Selection by AIC is asymptotically equivalent to LOOCV Selection by BIC is as. eq. to certain leave-out-CV
PAC-Bayesian	Training and test data follow the *same* distribution Learning machine in stochastic function class $\mathcal{F}$ Inspired by SLT and Bayesian paradigm	$R(f) \leq R_N(f) + \frac{3}{\sqrt{N}} \cdot \sqrt{\mathrm{KL}(f\|\pi) + \log \frac{2}{\varepsilon}}$ for any learner f with probability $\geq 1-\varepsilon$ KL$(f\|\pi)$ Kullback-Leibler divergence π prior belief π in reference class $\mathcal{F}$
Minimum Description Length	Training and test data follow the *same* distribution Parametric function class $\mathcal{F}$ similar to Bayesian Based on information theoretical argumentation maximizing Bayes posterior is posited best	similar to PAC-Bayesian (quantities are interpreted information theoretically)