# Validation

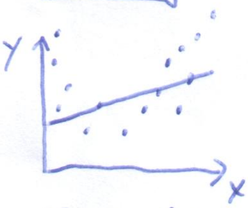
## a) What is validation?

Optimization of model parameters to training data does not imply good generalization performance (small $E^T \not\Rightarrow$ small $E^G$). Validation: check generalization of a model (training procedure) by estimating $\hat{E}^G \approx E^G = \int P(\underline{x}, y_T)\, e(y(\underline{x}), y_T)\, d\underline{x}\, dy_T$

## b) Over- vs. underfitting.

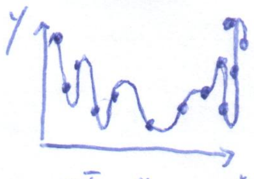
example regression:

underfitting



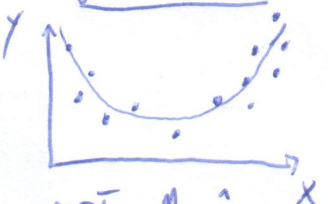
$\Rightarrow E^T$ high, $\hat{E}^G$ high

overfitting



$\Rightarrow E^T$ small, $\hat{E}^G$ high

good model



$\Rightarrow E^T$ small, $\hat{E}^G$ small

## c) Validation techniques (no hyperparameter optimization).

validation set: non-overlapping split of data:

data: | training | validation |

training: learn model via $\min_{\underline{w}} E^T(\underline{w})$

validation: estimate generalization performance $\hat{E}^G = \frac{1}{q}\sum_{\beta=1}^{q} e^{(\beta)}$

k-fold cross validation: k different nonoverlapping data splits

e.g. $k=4$: 1) $D_1$ valid., train; 2) $D_2$ V., train; 3) $D_3$; 4) $D_4$ $\Rightarrow$ 4× trained models, 4× $\hat{E}^G$

algorithm

for $j = 1, \dots, k$:

train model on $D \setminus D_j$ where $D = \{(\underline{x}^{(\alpha)}, y_T^{(\alpha)})\}_{\alpha=1}^{p}$ ($\Rightarrow \underline{w}_{(j)}$ model params)

test model on $D_j \Rightarrow \hat{E}^G_{(j)} = \frac{1}{|D_j|}\sum_{\beta \in D_j} e^{(\beta)}$

$$\hat{E}^G = \frac{1}{k}\sum_{j=1}^{k} \hat{E}^G_{(j)}$$

$$\mathrm{var}(\hat{E}_G) = \frac{k-1}{k}\sum_{j=1}^{k}\left(\hat{E}^G_{(j)} - \hat{E}^G\right)^2$$ ($\leftarrow$ informative measure!)

remarks:

- $\hat{E}^G \hat{=}$ bias, $\mathrm{var}(\hat{E}^G) \hat{=}$ variance (of model optimization)
- $k = p \Rightarrow$ leave one out cross validation (LOOCV) [e.g. for small datasets]
- k-fold CV yields k models $\rightarrow$ bias and variance are for training procedure
  $\Rightarrow$ TRAIN FINAL MODEL ON ALL DATA!

# Validation (cont'd)

## d) validation (with hyperparameters).

<u>test set method</u>: non-overlapping split of data


for fixed hyperparameters ~~learn~~ learn model via $\min_{\underline{w}} E^T(\underline{w})$

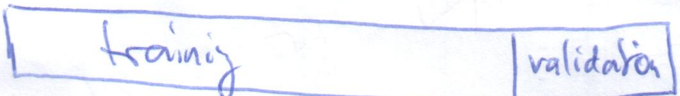
for that hyperparam. test that model $\hat{E}^G$

loop over hyperparam values and select those ~~hp~~ that min. $\hat{E}^G$!

~~test~~ the validate the model of the best hyperparams.
$\Rightarrow$ this $\hat{E}^G$ is estimate of generalization performance only for validation set!

<u>attention</u>: similarly as the model params $\underline{w}$ can overfit to the training data, the hyperparams can also overfit to training + test data!
$\Rightarrow$ <u>validation is always needed!</u>

<u>~~k-fold~~ cross-validation + validation set</u>: k non-overlapping splits of ~~training~~ training data




run cross validation for all hyperparam values, select hyperparams that min. $\hat{E}^G$!

validate the model with best hyperparams (trained on all training data)
$\Rightarrow$ this $\hat{E}_G$ estimation of generalization performance

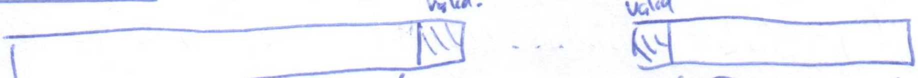
<u>attention</u>: (same as above)

---

NEVER USE VALIDATION DATA FOR OPTIMIZATION OF MODEL PARAMS OR HYPERPARAMS!

---

<u>nested cross validation</u> (optional)




inner CV $\rightarrow$ hyperparam opt.

inner CV: hyperparam opt

outer CV: estimation of generalization performance

→ as many opt. hyperparams as outer CV folds

$\Rightarrow$ final model: det. hypp on outer CV