Deadline: Fr. June 21 (online) or Mo. June 23, 10:00 AM (hand-in)
Drop your printed or legible handwritten submissions into the boxes at Samelsonplatz, or upload a `.pdf` file via LearnWeb.

1 Bias-Variance Dilemma (20 points)

Note: You can use Hastie's book as a reference. It is available as an ebook on his homepage https://web.stanford.edu/~hastie/ElemStatLearn/

Consider a regression problem $y = f(x) + \varepsilon$ with $\varepsilon \overset{\text{iid}}{\sim} \mathcal{N}$ and $x \overset{\text{iid}}{\sim} \mathcal{D}$ for some data distribution $\mathcal{D}$ and some noise distribution $\mathcal{N}$ with zero mean $\mathbb{E}[\varepsilon] = 0$ and finite variance $\mathrm{Var}[\varepsilon] = \sigma^2$. We are interested in finding the best model $\hat{y}$ from a class of models[1] $\mathcal{M}$ which minimizes the L^2 loss

$$\hat{y}^* = \operatorname*{argmin}_{\hat{y} \in \mathcal{M}} \mathbb{E}_{x \sim \mathcal{D}, \epsilon \sim \mathcal{N}}[(y(x) - \hat{y}(x))^2] \tag{1}$$

In practice, the data distribution $\mathcal{D}$ is unknown and we are only provided with a finite sample dataset $D = (x_i, y_i)_{i=1\ldots N}$. So instead of minimizing (1), we minimize the *empirical* L^2 loss

$$\hat{y}_D = \operatorname*{argmin}_{\hat{y} \in \mathcal{M}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}(x_i))^2 \tag{2}$$

Now, the best model depends on the specific sample D we have. If the experiment is repeated (i.e. a new dataset is gathered), then due to noise the optimal model we find will never be 100% the same.

A. [4p] Let $\bar{y} = \mathbb{E}_D[\hat{y}_D]$ and $\mathrm{Bias}(\hat{y}) = \bar{y} - f$. Show that

$$\mathbb{E}_{D,\varepsilon}[(y - \hat{y}_D)^2] = \mathrm{Bias}(\hat{y})^2 + \mathrm{Var}[\hat{y}] + \sigma^2 \tag{3}$$

This is known as the *Bias-Variance-Decomposition*

B. [4p] Explain what is meant by the *Bias-Variance-Tradeoff*. How does it relate to over- and underfitting?

C. [4p] Provide an example of a model where one can control the Bias-Variance-Tradeoff through the choice of a hyperparameter.

D. [4p] Explain why bagging reduces variance, but can lead to higher bias.

E. [4p] Explain why boosting reduces bias, but can lead to higher variance.

[1] usually a parametric family, e.g. in linear regression: $\mathcal{M} = \{g\colon X \to Y, x \mapsto \beta^T x \mid \beta \in \mathbb{R}^m\}$