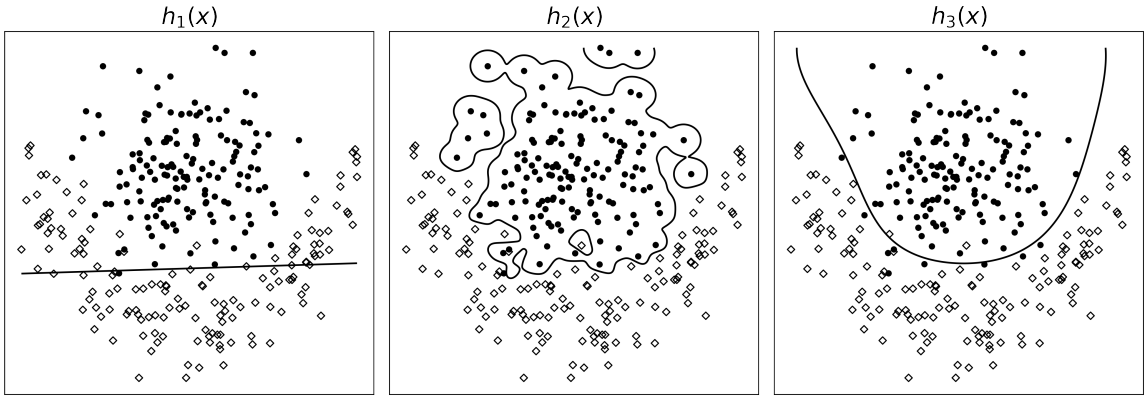
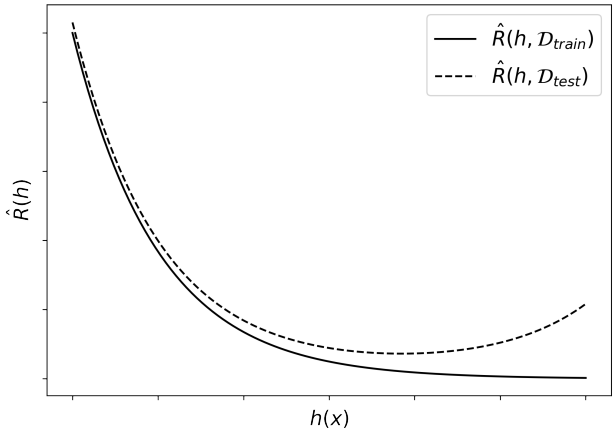
TUM EI 70360: Machine Learning and Optimization
Fall 2023
Lecturer: Reinhard Heckel
Teaching Assistant: Tobit Klug

Problem Set 7
Issued: Tuesday, Nov. 28, 2023
Due: Thursday, Dec. 7, 2023

Problem 1 (Training and test errors of binary classifiers)**.** Consider binary classification based on a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ with binary labels y_i and features $\mathbf{x}_i \in \mathbb{R}^2$ for all $i = 1, \ldots, n$. We want to compare three different classifiers $h_j(\mathbf{x}),\ j = 1, 2, 3$ that are obtained by running three different classification algorithms on the same training set. The first three plots below show the training set as well as the decision boundaries obtained by the three classifiers. The fourth plot shows the empirical training and test error as a function of classifiers estimated by different algorithms.

Copy the fourth plot to your solution sheet and draw vertical lines at positions, where the performances of the classifiers $h_j(\mathbf{x}),\ j = 1, 2, 3$ are approximately allocated. For each, provide one sentence of justification.





Problem 2 (Estimation, approximation and excess error)**.** Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a set of n data points drawn from some data distribution, and let $\hat{h} = \min_{h\in\mathcal{H}}(1/n)\sum_{i=1}^n \ell(h(\mathbf{x}_i), y_i)$ be the empirical risk minimizer with respect to a loss function ℓ over a hypothesis class $\mathcal{H}$. Let $h^* = \arg\min_h R(h)$ be the optimal estimator out of all possible estimators, and assume that h^* is not in the hypothesis class $\mathcal{H}$. Here, $R(h) = \mathbb{E}_{\mathbf{x},y}\left[\ell(h(\mathbf{x}), y)\right]$ is the risk of h. Draw curves that shows each of the following errors as a function of n:

(a) The estimation error $R(\hat{h}) - \inf_{h\in\mathcal{H}} R(h)$.

(b) The approximation error $\inf_{h\in\mathcal{H}} R(h) - R(h^*)$.

(c) The excess error $R(\hat{h}) - R(h^*)$.

Problem 3 (Generalization bound for binary weight neural networks)**.** Consider a neural network $f_{\boldsymbol{\theta}}(\mathbf{x})$ with weights $\boldsymbol{\theta} \in \mathbb{R}^p$. Assume n training examples are generated from an unknown joint distribution over $(\mathbf{x}, y)$, and assume that you train the model by minimizing the empirical risk defined as

$$\hat{R}(f_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{f_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq y_i\},$$

over model parameters $\boldsymbol{\theta}$ that are such that each entry of $\boldsymbol{\theta}$ is either 1 or -1. In the definition of the empirical risk, $\mathbb{1}\{A\}$ is the indicator function of the event A, which is one if the event is true and zero otherwise. Give the best possible bound on the risk $R(f_{\boldsymbol{\theta}}) = \mathbb{E}\left[\mathbb{1}\{f_{\boldsymbol{\theta}}(\mathbf{x}) \neq y\}\right]$ as a function of the problem parameters n, p and $\hat{R}(\boldsymbol{\theta})$, that holds with high probability.