Statistical analysis of machine learning algorithms

Dr. Stefan Richter

Winter term 2020/2021

Contents

1 Statistical decision theory **4**
- 1.1 Statistical decision theory . . . 4
- 1.2 Evaluation of algorithms . . . 6
- 1.3 Standard approaches to derive algorithms . . . 7

2 Linear algorithms for regression problems **10**
- 2.1 Ridge Regression . . . 14
- 2.2 LASSO estimator . . . 16
- 2.3 Restricted Eigenvalue property . . . 17
- 2.4 Exercises . . . 21

3 Basics of classification problems; linear models in classification problems **25**
- 3.1 Decision regions, decision boundaries, discriminant functions . . . 25
- 3.2 Logistic regression . . . 27
  - 3.2.1 Theoretical statements . . . 29
- 3.3 Calibration condition and risk transfer formula . . . 39
- 3.4 Generalization to nonlinear models . . . 46
- 3.5 Exercises . . . 47

4 Support Vector Machines **52**
- 4.1 Separating hyperplanes . . . 52
- 4.2 Support vector machines (SVM) . . . 54
- 4.3 Dual formulation of the SVM . . . 56
- 4.4 Generalized SVM . . . 58
- 4.5 The SVM algorithm as a minimizer of an empirical risk . . . 61
- 4.6 Theoretical statements . . . 64
- 4.7 Simplified problem . . . 69
  - 4.7.1 Step (a) . . . 71
  - 4.7.2 Step (b) . . . 73
  - 4.7.3 Step (c) . . . 73
  - 4.7.4 Selection of r . . . 75
  - 4.7.5 Discussion of the complex problem . . . 76
- 4.8 Appendix . . . 80
- 4.9 Exercises . . . 81

5 A short excursus: kernel nonparametric statistics **88**
- 5.1 The kernel regression estimator . . . 88
- 5.2 Classification . . . 91
- 5.3 The naive Bayes classifier . . . 94
- 5.4 Exercises . . . 98

6 Regression and classification trees; Boosting **101**
6.1 Binary trees . . . 101
6.2 Dyadic trees . . . 102
6.3 Proof of Theorem 6.7 . . . 106
6.4 Boosting . . . 112
6.4.1 The exact boosting algorithm . . . 114
6.4.2 The approximated boosting algorithm . . . 115
6.5 Boosting: theoretical results . . . 118
6.6 Excursus: empirical process theory . . . 122
6.7 Proof of the boosting oracle inequality . . . 128
6.8 Exercises . . . 132

7 Neural networks **139**
7.1 The neural network algorithm . . . 141
7.2 Approximation theory for neural networks . . . 143
7.3 Theoretical results . . . 144
7.4 Convergence rate of the neural network algorithm . . . 149
7.5 Exercises . . . 154

8 Solutions of the exercises **160**
8.1 Solutions of Chapter 2 . . . 160
8.2 Solutions of Chapter 3 . . . 166
8.3 Solutions of Chapter 4 . . . 173
8.4 Solutions of Chapter 5 . . . 185
8.5 Solutions of Chapter 6 . . . 188
8.6 Solutions of Chapter 7 . . . 197

1 Statistical decision theory

Basic difference between machine learning and statistical learning:

- Machine learning invents models and algorithms which can 'learn' from training data and are available to generalize these findings to predict *new* outcomes
- Statistical learning is a discipline of mathematical statistics which formalizes the models from machine learning and quantifies their (statistical) uncertainty. Furthermore, the theoretical findings can be used to invent new or at least *improve* existing machine learning algorithms by proposing meaningful rules for tuning parameters.

Im this lecture, we consider the subdiscipline of *supervised learning*. In this setting, we observe pairs of data (X, Y), where X is called *input* and Y is called *output*. The goal is to predict Y from X.

1.1 Statistical decision theory

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

Definition 1.1 (Supervised learning setting)**.** Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, ..., K\}$ or $\mathcal{Y} = \mathbb{R}$. Let $\mathbb{P}^{(X,Y)}$ be some distribution on $\mathcal{X} \times \mathcal{Y}$. Let $(X_i, Y_i), i = 1, ..., n$ be i.i.d. random variables with distribution $\mathbb{P}^{(X,Y)}$, the so-called *training samples* or *training data*.

- If $\mathcal{Y} = \{1, ..., K\}$, then we say that the training data stems from a *classification problem*,
- If $\mathcal{Y} = \mathbb{R}$, then we say that the training data stems from a *regression problem*.

Throughout the lecture, let $(X, Y) \sim \mathbb{P}^{(X,Y)}$ be a realization independent of (X_i, Y_i), $i = 1, ..., n$.

General goal of machine learning: Find a mapping $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ (computed from the training samples (X_i, Y_i), $i = 1, ..., n$), such that the expression $\hat{f}_n(X)$ is near to Y. This is formalized as follows. We equip $\mathcal{X}$, $\mathcal{Y}$ with the Borel-σ-algebras $\mathcal{B}(\mathcal{X})$, $\mathcal{B}(\mathcal{Y})$ such that measurability is well-defined.

Definition 1.2.
- A decision rule is a measurable mapping $f : \mathcal{X} \to \mathcal{Y}$.

- A loss function is a measurable mapping $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Application of a decision rule f to some X produces the *loss* $L(Y, f(X))$.
- The risk of a decision rule f is defined by $R(f) := \mathbb{E}L(Y, f(X))$.

Example 1.3 (Examples of loss functions).
- For regression problems $\mathcal{Y} = \mathbb{R}$: Quadratic / squared loss $L(y, s) = (y - s)^2$ or absolute loss $L(y, s) = |y - s|$.
- For classification problems $\mathcal{Y} = \{1, ..., K\}$: 0-1-loss $L(y, s) = \mathbb{1}_{\{y \neq s\}}$.

Given a loss function L and a distribution $\mathbb{P}^{(X,Y)}$, one can define the *optimal* decision function associated to the corresponding supervised learning problem.

Definition 1.4. A measurable mapping $f^* : \mathcal{X} \to \mathcal{Y}$ which satisfies $R(f^*) = \min_{f:\mathcal{X}\to\mathcal{Y} \text{ meas.}} R(f)$ is called *Bayes rule* or *Bayes classifier* (for classification problems). The corresponding risk $R(f^*)$ is called *Bayes risk*.

For specific loss functions, one can provide more explicit expressions for f^*.

Remark 1.5 (Optimal decision functions when $\mathbb{P}^{(X,Y)}$ is known).
- Regression problem, squared loss:

$$R(f) = \mathbb{E}L(Y, f(X)) = \mathbb{E}[(Y - f(X))^2] = \int \int (y - f(x))^2 \, \mathrm{d}\mathbb{P}^{Y|X=x}(y) \, \mathrm{d}\mathbb{P}^X(x)$$

  is minimal for $f^*(x) = \mathbb{E}[Y|X = x]$ (conditional expectation of Y given $X = x$).
- Classification problem, 0-1-loss:

$$R(f) = \mathbb{E}L(Y, f(X)) = \mathbb{P}(Y \neq f(X)) = \int \mathbb{P}(Y \neq f(X)|X = x) \, \mathrm{d}\mathbb{P}^X(x)$$

  is minimal for

$$f^*(x) = \operatorname*{arg\,min}_{k\in\{1,...,K\}} \mathbb{P}(Y \neq k|X = x) = \operatorname*{arg\,max}_{k\in\{1,...,K\}} \mathbb{P}(Y = k|X = x).$$

  $f^*(x)$ selects the class which is most probable for the given observation x (f^* is also called *MAP-classifier*, MAP = maximum a posteriori).

1.2 Evaluation of algorithms

Given training data (X_i, Y_i), $i = 1, ..., n$, our goal is to define a decision rule $\hat{f}_n$ which is as near as possible to f^*. The function $\hat{f}_n$ which maps the training samples to a decision rule is called algorithm.

Definition 1.6 (Algorithm)**.** A measurable mapping $\hat{f}_n : \Omega \times \mathcal{X} \to \mathcal{Y}$ is called *algorithm*, if

(i) For each $x \in \mathcal{X}$, the mapping $\hat{f}_n(x) : \Omega \to \mathcal{Y}$ is measurable with respect to $T_n := ((X_i, Y_i))_{i=1,...,n}$,

(ii) For each $\omega \in \Omega$, $\hat{f}_n(\omega) : \mathcal{X} \to \mathcal{Y}$ is a decision rule.

With some abuse of this terminology, we also may call $\hat{f}_n$ a decision rule (in the sense of (ii)). The evaluation of an algorithm takes place in two steps: For given training samples (X_i, Y_i), $i = 1, ..., n$, the expression $\hat{f}_n$ is a decision rule and therefore can be evaluated by the risk $R(\hat{f}_n)$. Note that this expression *still* is a random variable dependent on (X_i, Y_i), $i = 1, ..., n$. By performing another expectation $\mathbb{E}R(\hat{f}_n)$, the influence of the training samples on the decision rule $\hat{f}_n$ is averaged. $\mathbb{E}R(\hat{f}_n)$ is a real number which therefore expresses the 'average risk' of $\hat{f}_n$. There are specific names for these quantities.

Definition 1.7. Let $\hat{f}_n$ be an algorithm.

- $R(\hat{f}_n)$ or $\mathbb{E}R(\hat{f}_n)$ is called *generalization error*,
- $R(\hat{f}_n) - R(f^*)$ or $\mathbb{E}R(\hat{f}_n) - R(f^*)$ is called *excess Bayes risk*.

An algorithm learns with convergence rate $\psi(n)$, if

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq \psi(n).$$

In the following chapters, we will also use a different formulation for convergence rate results of an algorithm, namely

$$\forall t \geq 0 : \quad \mathbb{P}\big(R(\hat{f}_n) - R(f^*) \geq \psi(n) + t\big) \leq p(t)$$

with some function $p : [0, \infty) \to [0, 1]$ which is decreasing in t and satisfies $\lim_{t \downarrow 0} p(t) = 0$. Such statements can be interpreted in the sense that $R(\hat{f}_n) - R(f^*)$ is at most $\psi(n)$ with large probability.

Based on model assumptions on $\mathbb{P}^{(X,Y)}$ and the specific approach which is used to derive an algorithm, the decision rules produced by $\hat{f}_n$ lie in some function class

$$\mathcal{F} = \mathcal{F}_n \subset \{f : \mathcal{X} \to \mathcal{Y} \text{ measurable}\}.$$

In general, one cannot hope that the model is 'correct' in the sense that $f^* \in \mathcal{F}$. Therefore, one typically has the following decomposition.

$$R(\hat{f}_n) - R(f^*) = \underbrace{\left[R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{F}} R(f) - R(f^*)\right]}_{\text{approximation error}}.$$

A larger class $\mathcal{F}$ leads to a smaller approximation error but to a larger estimation error. With the next example, we provide an intuitive explanation of the approximation error.

Example 1.8. We consider a regression problem with $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ and squared loss. We assume that the distribution $\mathbb{P}^{(X,Y)}$ satisfies the following relation:

$$Y = f_0(X) + \varepsilon$$

with some *non-continuous* function $f_0 : \mathbb{R} \to \mathbb{R}$ and ε independent of X with $\mathbb{E}\varepsilon = 0$. Note that this does *not* describe the whole distribution $\mathbb{P}^{(X,Y)}$, but it clearly puts some conditions on the relation between X and Y. These conditions are enough to derive a meaningful expression of f^* and the approximation error. It holds that

$$f^*(x) = \mathbb{E}[Y|X = x] = f_0(x) \quad \Rightarrow \quad R(f^*) = \mathbb{E}L(Y, f^*(X)) = \mathbb{E}[\varepsilon^2].$$

If the algorithm $\hat{f}_n$ is obtained from a procedure which forces $\hat{f}_n \in \mathcal{F} \subset \{f : \mathbb{R} \to \mathbb{R}, f \text{ continuous}\}$ then it holds for $f \in \mathcal{F}$ that

$$R(f) = \mathbb{E}[(\underbrace{Y}_{=f_0(X)+\varepsilon} - f(X))^2] = \mathbb{E}[(f_0(X) - f(X))^2] + \mathbb{E}[\varepsilon^2].$$

Then, the approximation reads $\inf_{f \in \mathcal{F}} R(f) - R(f^*) = \inf_{f \in \mathcal{F}} \mathbb{E}[(f_0(X) - f(X))^2]$ and is clearly nonzero if $\mathcal{F}$ can not approximate non-continuous functions well.

1.3 Standard approaches to derive algorithms

Let $\mathcal{F} \subset \{f : \mathcal{X} \to \mathcal{Y} \text{ meas.}\}$ be some function class and L a loss function.

Definition 1.9 (Standard algorithm)**.** Given some loss function L, the minimizer of the so-called *empirical risk* $\hat{R}_n(f)$

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \hat{R}_n(f), \qquad R_n(f) := \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$$

is called *standard algorithm*.

If $\mathcal{F}$ is large, then the above optimization problem has no unique solution. There may even exist $f \in \mathcal{F}$ such that $\hat{R}_n(f) = 0$ ('overfitting'). In this case, one typically introduces a *penalty term* $J_\lambda(f)$ which penalizes solutions with wrong properties. Here, $\lambda \in \Lambda$ is called *dimensionality parameter* or *tuning parameter*.

One example could be that one only wants to derive mapping with small second derivative. Then one could choose $J_\lambda(f) = \lambda \cdot \int_{\mathcal{X}} f''(x) \, \mathrm{d}x$.

The corresponding optimization problem then reads as follows.

Definition 1.10 (Standard algorithm with penalization)**.** Let L be a loss function and $J : \mathcal{F} \to \mathbb{R}$ a mapping. The minimizer of the empirical risk

$$\hat{f}_{n,\lambda} \in \arg\min_{f \in \mathcal{F}} \left\{ \hat{R}_n(f) + J_\lambda(f) \right\}$$

is called *standard algorithm with penalization* J.

Instead of L, one sometimes uses different loss functions $\tilde{L}$ in the empirical risk. Then $\hat{R}_n(f)$ is replaced by $\tilde{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \tilde{L}(Y_i, f(X_i))$. This will be often the case for classification problems (see the later chapters).
For standard algorithms from Definition 1.9 and Definition 1.10, the following very simple bound for the excess Bayes risk holds. We will never use this elementary bound, but it gives an idea how the typical proof strategy looks like.

Proposition 1.11 (Basic inequality)**.** Let $\tilde{f} \in \arg\min_{f \in \mathcal{F}} R(f)$. Then for $\hat{f}_n$ from Definition 1.9 it holds that

$$R(\hat{f}_n) - R(\tilde{f}) \leq 2 \sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right|.$$

Proof. It holds that

$$R(\hat{f}_n) - R(\tilde{f}) = \hat{R}_n(\hat{f}_n) - \hat{R}_n(\tilde{f}) + \big[R(\hat{f}_n) - R(\tilde{f})\big] - \big[\hat{R}_n(\hat{f}_n) - \hat{R}_n(\tilde{f}))\big].$$

Definition $\hat{f}_n \Rightarrow \hat{R}_n(\hat{f}_n) - \hat{R}_n(\tilde{f}) \leq 0 \Rightarrow$

$$R(\hat{f}_n) - R(\tilde{f}) \leq 2 \sup_{f \in \mathcal{F}} \big|\hat{R}_n(f) - R(f)\big|.$$

□

Remarks:

- Many machine learning algorithms (or at least parts of them) can be written in the above standard algorithm form. The representation as an optimization problems also allows for many approximative solution techniques (for instance, gradient descent methods or iterative solvers).
- If we boil down all machine learning algorithms to the above standard algorithm formulations, different algorithms *only* differ in the choice of the function class $\mathcal{F}$ and the corresponding penalty $J_\lambda(f)$. Theory is then derived under different model assumptions on $\mathbb{P}^{(X,Y)}$. One goal of this lecture is to introduce several machine learning algorithms and to work out the corresponding values of $\mathcal{F}$ and J.
- The basic inequality given in Proposition 1.11 is too weak to derive sharp results for the convergence rate of the excess Bayes risk. However, it yields an important information which quantity has to be analyzed theoretically: The supremum over function classes $\sup_{f \in \mathcal{F}} \big\{\hat{R}_n(f) - R(f)\big\}$.

2 Linear algorithms for regression problems

In this chapter, we consider regression problems with $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and loss function $L(y, s) = (y - s)^2$.

Definition 2.1 (Model assumption: Linear regression)**.** With parameters $\beta^* = (\beta_1^*, ..., \beta_d^*) \in \mathbb{R}^d$ it holds that

$$Y = (\beta^*)^T X + \varepsilon = \sum_{j=1}^{d} \beta_j^* X_j,$$

where ε is independent of X with $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[\varepsilon^2] = \sigma^2$.

If this model assumption holds true, we have

$$f^*(x) = \mathbb{E}[Y|X = x] = \sum_{j=1}^{d} \beta_j^* x_j.$$

This means that the model assumption implies some special (linear) structure of the Bayes rule. Thus it is reasonable to search for corresponding decision rules also in the space of linear functions,

$$\mathcal{F} := \{f : \mathbb{R}^d \to \mathbb{R} : f(x) = \sum_{j=1}^{d} \beta_j x_j \text{ with } \beta = (\beta_1, ..., \beta_d)^T \in \mathbb{R}^d\}.$$

In the following, we abbreviate $R(f) = R(\beta)$ for $f(x) = \sum_{j=1}^{d} \beta_j x_j \in \mathcal{F}$.

Lemma 2.2 (Risk and Bayes risk)**.** Let $\Sigma := \mathbb{E}[XX^T]$. Then it holds that

$$R(\beta) - R(\beta^*) = \|\Sigma^{1/2}(\beta^* - \beta)\|_2^2$$

and $R(\beta^*) = \mathbb{E}[\varepsilon^2]$.

Proof. It holds that

$$\begin{aligned} R(\beta) &= \mathbb{E}[(Y - \beta^T X)^2] = \mathbb{E}[((Y - (\beta^*)^T X) + (\beta^* - \beta)^T X)^2] \\ &= \mathbb{E}[\varepsilon^2] + \mathbb{E}[\varepsilon \cdot (\beta^* - \beta)^T X] + \mathbb{E}[(\beta^* - \beta)^T X)^2] \\ &= \mathbb{E}[\varepsilon^2] + \mathbb{E}[\|\Sigma^{1/2}(\beta^* - \beta)\|_2^2], \end{aligned}$$

where we have used $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[(a^T X)^2] = a^T \mathbb{E}[XX^T]a = a^T \Sigma a = \|\Sigma^{1/2} a\|_2^2$. The statement $R(\beta^*) = \mathbb{E}[\varepsilon^2]$ can be obtained from the above formula by plugging in $\beta = \beta^*$. □

We now apply the standard approach to derive an algorithm.

Definition 2.3 (LS estimator)**.** Let (X_i, Y_i) be training samples and

$$\hat{R}_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \sum_{j=1}^{d} \beta_j X_{ij}\big)^2.$$

The estimator

$$\hat{\beta} = \hat{\beta}^{KQ} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \hat{R}_n(\beta)$$

is called *least squares estimator* (LS estimator) of β. The corresponding algorithm reads

$$\hat{f}_n(x) = \hat{f}_n^{KQ}(x) = \sum_{j=1}^{d} \hat{\beta}_j x_j.$$

Lemma 2.4. Define the design matrix $\mathbb{X}$, the regression vector $\mathbb{Y}$ and the noise vector $\mathbb{e}$ via

$$\mathbb{X} = \begin{pmatrix} X_{11} & \dots & X_{1d} \\ \vdots & \ddots & \vdots \\ X_{n1} & \dots & X_{nd} \end{pmatrix}, \qquad \mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbb{e} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then the training samples fulfill $\mathbb{Y} = \mathbb{X}\beta^* + \mathbb{e}$ and the empirical risk has the form $\hat{R}_n(\beta) = \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2$.
Moreover, it holds that $(\mathbb{X}^T\mathbb{X})\hat{\beta} = \mathbb{X}^T\mathbb{Y}$ (the so-called normal equation). If $\mathbb{X}$ has full rank, then it holds that

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}.$$

Proof. These are basic calculations. The representation of $\hat{\beta}$ can be obtained by setting the derivativ equal to zero,

$$0 = \partial_\beta \hat{R}_n(\beta) = 2\mathbb{X}^T\mathbb{Y} - 2\mathbb{X}^T\mathbb{X}\beta.$$

□

We now try to calculate the excess Bayes risk of the corresponding algorithm. Define $\hat{\Sigma} := \frac{1}{n}\mathbb{X}^T\mathbb{X}$. Then it holds that

$$\hat{\beta}^{KQ} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta^* + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{e} = \beta^* + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{e},$$

thus

$$\begin{aligned} R(\hat{\beta}^{KQ}) - R(\beta^*) &= \big\|\Sigma^{1/2}(\hat{\beta}^{KQ} - \beta^*)\big\|_2^2 = \big\|\Sigma^{1/2}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{e}\|_2^2 = \frac{1}{n}\big\|\Sigma^{1/2}\hat{\Sigma}^{-1}\frac{\mathbb{X}^T\mathbb{e}}{\sqrt{n}}\big\|_2^2 \\ &= \frac{1}{n}\|A\mathbb{e}\|_2^2, \end{aligned}$$

where $A := \Sigma^{1/2}\hat{\Sigma}^{-1}\frac{\mathbb{X}^T}{\sqrt{n}}$.

We analyze this quantity by first calculating the excess Bayes risk under for given $\mathbb{X}$. Since $\mathbb{e}$ is independent of $\mathbb{X}$, $\mathrm{tr}(\cdot)$ is linear and due to the rule $\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$ for matrices A, B, C, we have

$$\begin{aligned} \mathbb{E}[R(\hat{\beta}^{KQ})|\mathbb{X}] - R(\beta^*) &= \frac{1}{n}\mathbb{E}[\|A\mathbb{e}\|_2^2|\mathbb{X}] \\ &= \frac{1}{n}\mathbb{E}\Big[\mathrm{tr}(A\mathbb{e}\mathbb{e}^TA^T)\Big|\mathbb{X}\Big] = \frac{1}{n}\mathrm{tr}(A\underbrace{\mathbb{E}[\mathbb{e}\mathbb{e}^T]}_{=\sigma^2 I_{d\times d}}A^T) = \frac{\sigma^2}{n}\|A\|_F^2 \\ &= \frac{\sigma^2}{n}\cdot\mathrm{tr}(\Sigma^{1/2}\hat{\Sigma}^{-1}\hat{\Sigma}\hat{\Sigma}^{-1}\Sigma^{1/2}) \\ &= \frac{\sigma^2}{n}\cdot\mathrm{tr}(\Sigma\hat{\Sigma}^{-1}). \end{aligned}$$

We conclude that if $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_iX_i^T \approx \mathbb{E}[XX^T] = \Sigma$, then it holds that $\mathbb{E}R(\hat{f}^{KQ}) - R(\beta^*) \approx \frac{\sigma^2 d}{n}$.
This approximation does not take into account the variation of X. Due to the inverse, it is not easily possible to derive an upper bound for the expectation of $\hat{\Sigma}^{-1}$. Therefore, we only obtain a result for the excess Bayes risk which holds with high probability and under additional assumptions (which could be relaxed but would lead to much more complicated proofs). Throughout the lecture, we call a constant c *universal* if it is a number not depending on any variable defined before (so in principle, it is something like '32' but we are too lazy to provide the explicit value).

Theorem 2.5. Let $\varepsilon \sim N(0,\sigma^2)$, $X \sim N(0,\Sigma)$. Then there exist universal constants $c_1, c_2 > 0$ such that the condition

$$\frac{\lambda_{min}(\Sigma)}{2} \geq c_1\|\Sigma\|\Big(\frac{\max\{\log(n), d\}}{n}\Big)^{1/2}$$

implies for all $t \geq 1$:

$$\mathbb{P}(R(\hat{\beta}^{KQ}) - R(\beta^*) > 4c_2^2 t \cdot \frac{\sigma^2 d}{n}) \leq e^{-t} + n^{-1}$$

Proof. Let $\|A\|$ denote the spectral norm, $\|A\|_F^2$ the Frobenius norm and $\lambda_{min}(A)$ the smallest eigenvalue of the matrix $A \in \mathbb{R}^{d\times d}$.

The basic idea is to define an event on which $\hat{\Sigma}$ is near to Σ. Put

$$E_n := \{\|\hat{\Sigma} - \Sigma\| \leq \frac{\lambda_{min}(\Sigma)}{2}\}.$$

On E_n it holds that $\lambda_{min}(\hat{\Sigma}) \geq \frac{\lambda_{min}(\Sigma)}{2}$ and thus $\|\hat{\Sigma}^{-1}\| \leq \frac{2}{\lambda_{min}(\Sigma)}$. On E_n it holds that

$$\begin{aligned}
\|A\|_F^2 &= \operatorname{tr}(AA^T) = \operatorname{tr}(\Sigma \underbrace{\hat{\Sigma}^{-1}}_{=\Sigma^{-1}+(\hat{\Sigma}^{-1}-\Sigma^{-1})}) = \operatorname{tr}(I_{d\times d}) + \operatorname{tr}(\hat{\Sigma}^{-1}(\Sigma - \hat{\Sigma})) \\
&\leq d + \|\hat{\Sigma}^{-1}\|_F \|\Sigma - \hat{\Sigma}\|_F \\
&\leq d + d\|\hat{\Sigma}^{-1}\| \cdot \|\Sigma - \hat{\Sigma}\| \leq 2d. \qquad (1)
\end{aligned}$$

Now we have

$$\begin{aligned}
\mathbb{P}(R(\hat{\beta}^{KQ}) - R(\beta^*) > \gamma_n) &\leq \mathbb{P}(R(\hat{\beta}^{KQ}) - R(\beta^*) > \gamma_n, E_n) + \mathbb{P}(E_n^c) \\
&\leq \mathbb{E}[\mathbb{P}(R(\hat{\beta}^{KQ}) - R(\beta^*) > \gamma_n|\mathbb{X})\mathbb{1}_{E_n}] + \mathbb{P}(E_n^c) \\
&\leq \mathbb{E}[\mathbb{P}(\|A\mathbb{e}\|_2^2 > n\gamma_n|\mathbb{X})\mathbb{1}_{E_n}] + \mathbb{P}(E_n^c),
\end{aligned}$$

To discuss these quantities, we use two lemmas from [7, Theorem 1] and [8, Corollary 2.2] without proof (so-called isoperimetric inequalities for normally distributed random vectors): There are universal constants $c_1, c_2 > 0$, such that for all $s, t > 0$,

$$\mathbb{P}(\|\hat{\Sigma} - \Sigma\| > c_1\|\Sigma\| \cdot \max\{\frac{d}{n}, \sqrt{\frac{d}{n}}, \frac{s}{n}, \sqrt{\frac{s}{n}}\}) \leq e^{-s}, \qquad (2)$$

$$\mathbb{P}(\|A\mathbb{e}\|_2 \geq c_2\sigma\|A\|_F \max\{1, \sqrt{t}\}) \leq e^{-t}. \qquad (3)$$

Let $t \geq 1$. Based on these inequalities, on E_n it holds that

$$\mathbb{P}(\|A\mathbb{e}\|_2^2 \geq 4c_2^2 t \cdot \frac{\sigma^2 d}{n}|\mathbb{X}) \overset{(1)}{\leq} \mathbb{P}(\|A\mathbb{e}\|_2 \geq c_2\sqrt{t}\sigma\|A\|_F|\mathbb{X}) \overset{(3)}{\leq} e^{-t}.$$

With $n \geq d$ and $s = \log(n)$ it holds that $\max\{\frac{d}{n}, \sqrt{\frac{d}{n}}, \frac{s}{n}, \sqrt{\frac{s}{n}}\} = (\frac{\max\{\log(n), d\}}{n})^{1/2}$. Then for $\frac{\lambda_{min}(\Sigma)}{2} \geq c_1\|\Sigma\|(\frac{\max\{\log(n),d\}}{n})^{1/2}$,

$$\mathbb{P}(E_n^c) \leq \mathbb{P}(\|\hat{\Sigma}-\Sigma\| \geq \frac{\lambda_{min}(\Sigma)}{2}) = \mathbb{P}(\|\hat{\Sigma}-\Sigma\| > c_1\|\Sigma\|\cdot\max\{\frac{d}{n}, \sqrt{\frac{d}{n}}, \frac{s}{n}, \sqrt{\frac{s}{n}}\}) \overset{(2)}{\leq} e^{-s} = n^{-1}.$$

□

Problem: In many applications, d is large, in particular it may occur that $d > n$. Then $\mathbb{X}^T\mathbb{X}$ is no longer invertible and the LS-estimator is not unique.

2.1 Ridge Regression

In the following, we use the following **basic assumption:** Only few components of X are relevant for Y, that is, many components of β are zero or at least near to zero.

Thus we introduce a penalization for the size of the entries of β.

We use the following standard approach with penalization $J_\lambda(\beta) = \lambda \cdot \|\beta\|_2^2$.

Definition 2.6 (Ridge estimator)**.** Let $\lambda \geq 0$ and

$$J_\lambda(\beta) = \lambda\|\beta\|_2^2 = \lambda \sum_{j=1}^{d} \beta_j^2.$$

The Ridge estimator is defined as

$$\begin{aligned} \hat{\beta}_\lambda = \hat{\beta}_\lambda^{ridge} &= \arg\min_{\beta\in\mathbb{R}^d}\{\hat{R}_n(\beta) + J_\lambda(\beta)\} \\ &= \arg\min_{\beta\in\mathbb{R}^d}\left\{\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\right\}. \end{aligned}$$

The corresponding algorithm is

$$\hat{f}_{n,\lambda}(x) = \hat{f}_{n,\lambda}^{ridge}(x) = \sum_{j=1}^{d} \hat{\beta}_{\lambda,j} x_j.$$

The ridge estimator has an explicit representation.

Lemma 2.7. Let $\lambda > 0$. Then

$$\hat{\beta}_\lambda = (\mathbb{X}^T\mathbb{X} + \lambda n I_{d\times d})^{-1}\mathbb{X}^T\mathbb{Y}.$$

Proof. Building derivatives equal to zero for $\beta \mapsto \hat{R}_n(\beta) + \lambda \cdot J(\beta)$. □

Remarks:

- In ridge regression, the matrix $\mathbb{X}^T\mathbb{X}$ is 'made invertible' by adding a positive multiple of the identity matrix. Therefore, the ridge estimator also can be used in the case $d > n$.
- The name 'ridge' stems from the fact that the optimization problem is equivalent to

$$\min_{\beta\in\mathbb{R}^d} \hat{R}_n(\beta) \qquad \text{s.t.} \qquad \|\beta\|_2 \leq t$$

for some suitable $t > 0$. Here, some explicit limit is provided for t.

We now start to analyze the excess Bayes risk of the ridge estimator, again given $\mathbb{X}$. It holds that

$$\begin{aligned}\hat{\beta}_\lambda - \beta^* &= -\lambda n(\mathbb{X}^T\mathbb{X} + \lambda n I_{d\times d})^{-1}\beta^* + (\mathbb{X}^T\mathbb{X} + \lambda n I_{d\times d})^{-1}\mathbb{X}^T\mathbb{e} \\ &= -\lambda(\hat{\Sigma} + \lambda I_{d\times d})^{-1}\beta^* + \frac{1}{n}(\hat{\Sigma} + \lambda I_{d\times d})^{-1}\mathbb{X}^T\mathbb{e},\end{aligned}$$

thus

$$R(\hat{\beta}_\lambda) - R(\beta^*) = \|B + \frac{1}{\sqrt{n}}A\mathbb{e}\|_2^2 = \|B\|_2^2 + \frac{2}{\sqrt{n}}\langle B, A\mathbb{e}\rangle + \frac{1}{n}\|A\mathbb{e}\|_2^2$$

where $A = \Sigma^{1/2}(\hat{\Sigma}+\lambda I_{d\times d})^{-1}\frac{\mathbb{X}^T}{\sqrt{n}}$ and $B := \lambda\Sigma^{1/2}(\hat{\Sigma}+\lambda I_{d\times d})^{-1}\beta^*$. Since $\mathbb{E}\mathbb{e} = 0$, we have

$$\begin{aligned}\mathbb{E}[R(\hat{\beta}_\lambda)|\mathbb{X}] - R(\beta^*) &= \frac{\sigma^2}{n}\|A\|_F^2 + \|B\|_2^2 \\ &= \frac{\sigma^2}{n}\cdot\mathrm{tr}\big(\Sigma(\hat{\Sigma} + \lambda I_{d\times d})^{-1}\hat{\Sigma}(\hat{\Sigma} + \lambda I_{d\times d})^{-1}\big) + \lambda^2\|\Sigma^{1/2}(\hat{\Sigma} + \lambda I_{d\times d})^{-1}\beta^*\|_2^2.\end{aligned}$$

To discuss the approximation $\hat{\Sigma} \approx \Sigma$ in more detail, one has to define a similar event as for the LS estimator, but this would be much more complicated here. Therefore, we restrict ourselves to a 'heuristic' analysis of the ridge estimator by simply replacing $\hat{\Sigma} \approx \Sigma$. Let $\Sigma = UDU^T$ be the spectral decomposition of Σ with orthogonal matrix U and diagonal matrix $D = \mathrm{diag}(s_1, ..., s_d)$ (entries are the eigenvalues of Σ). Then it holds that

$$\begin{aligned}\mathbb{E}[R(\hat{\beta}_\lambda)|\mathbb{X}] - R(\beta^*) &\approx \frac{\sigma^2}{n}\cdot\mathrm{tr}\big(\Sigma(\Sigma + \lambda I_{d\times d})^{-1}\Sigma(\Sigma + \lambda I_{d\times d})^{-1}\big) + \lambda^2\|\Sigma^{1/2}(\Sigma + \lambda I_{d\times d})^{-1}\beta^*\|_2^2 \\ &= \frac{\sigma^2}{n}\cdot\mathrm{tr}\big(D(D + \lambda I_{d\times d})^{-1}D(D + \lambda I_{d\times d})^{-1}\big) + \lambda^2\|D^{1/2}(D + \lambda I_{d\times d})^{-1}U^T\beta^*\|_2^2 \\ &= \frac{\sigma^2}{n}\sum_{j=1}^d \frac{s_j^2}{(s_j+\lambda)^2} + \lambda^2\cdot\sum_{j=1}^d \frac{s_j(U^T\beta^*)_j^2}{(s_j+\lambda)^2}.\end{aligned}$$

If all eigenvalues are equal, that is, $s_j = s$ and if additionally $(U^T\beta^*)_j = b$ $(j = 1, ..., d)$, then the above expression simplifies to

$$\frac{\sigma^2 d}{n} \cdot \frac{s^2}{(s+\lambda)^2} + \lambda^2 \frac{sb^2 d}{(s+\lambda)^2} \underset{\lambda = \frac{\sigma^2}{b^2}}{\overset{\min}{\to}} \frac{\sigma^2 d}{n} \cdot \frac{b^2 s}{\frac{\sigma^2}{n} + b^2 s} \leq \frac{\sigma^2 d}{n}.$$

We see that for a suitable choice of the penalization parameter λ, the excess Bayes risk of the ridge estimator can be smaller than the corresponding upper bound of the LS estimator.

2.2 LASSO estimator

Approach: The most obvious choice to penalize β would be of the form $\|\beta\|_0 = \#\{j = 1, ..., d : \beta_j \neq 0\}$. Then, one would simply penalize the number of non-zero entries of β. However, this leads to NP-hard optimization problems whose solutions are not accessible in practice. One therefore uses a different norm which has similar properties but leads to convex optimization problems.

Definition 2.8 (Lasso - Least absolute shrinkage and selection operator)**.** Let $\lambda \geq 0$ and

$$J_\lambda(\beta) = \lambda \cdot \|\beta\|_1 = \lambda \sum_{j=1}^{d} |\beta_j|.$$

The LASSO estimator ('least absolute shrinkage and selection operator') is given by

$$\begin{aligned} \hat{\beta}_\lambda = \hat{\beta}_\lambda^{lasso} &\in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \hat{R}_n(\beta) + J_\lambda(\beta) \right\} \\ &= \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \lambda \cdot \|\beta\|_1 \right\}. \end{aligned}$$

The corresponding algorithm reads

$$\hat{f}_{n,\lambda}(x) = \hat{f}_{n,\lambda}^{lasso}(x) = \sum_{j=1}^{d} \hat{\beta}_{\lambda,j} x_j.$$

There exists no easy closed-form solution for $\hat{\beta}_\lambda^{lasso}$ besides some special cases.
Remark: 'selection operator' means that $\hat{\beta}_\lambda$ really 'selects' components j of β^*. The quantity $\hat{\beta}_\lambda$ has some entries which are exactly zero.

2.3 Restricted Eigenvalue property

We now investigate the theoretical properties of the LASO estimator. In the following we work with the following assumption: Only few entries of X are relevant for Y, that is, many entries of β^* are zero.

Definition 2.9 (Notation)**.** For $\beta \in \mathbb{R}^d$, define

$$S(\beta) := \{j \in \{1, ..., d\} : \beta_j \neq 0\}.$$

For $S \subset \{1, ..., d\}$ and $v \in \mathbb{R}^d$, put $v_S := (v_j \mathbb{1}_{\{j \in S\}})_{j=1,...,d}$.

If it would hold that $d \ll n$, then $\hat{\Sigma}$ would be invertible and the smallest eigenvalue would satisfy (we use the Rayleigh quotient formulation of the smallest eigenvalue)

$$\lambda_{min}(\hat{\Sigma}) := \inf_{v \in \mathbb{R}^d} \frac{v^T \hat{\Sigma} v}{\|v\|_2^2} > 0.$$

Then $\hat{\Sigma}$ would be one-to-one (injective) and the linear equation system $\hat{\Sigma}\beta = \frac{1}{n}\mathbb{X}^T\mathbb{Y}$ would lead to the (unique) least squares estimator.

In the case $d \gg n$, one has $\lambda_{min}(\hat{\Sigma}) = 0$. The important difference is that we only search for estimators $\hat{\beta}$ which non-zero entries at the components $S(\beta^*)$. This means that in principle we only need injectivity of $\hat{\Sigma}$ on the set

$$\tilde{C} = \{\beta \in \mathbb{R}^d : S(\beta) = S(\beta^*)\} = \{\beta \in \mathbb{R}^d : \|\beta_{S(\beta^*)^c}\|_1 = 0\},$$

or translated to the smallest eigenvalue,

$$\inf_{v \in \tilde{C}} \frac{v^T \hat{\Sigma} v}{\|v\|_2^2} = \inf_{v \in \tilde{C}} \frac{v^T \hat{\Sigma} v}{\|v_{S(\beta^*)}\|_2^2} > 0. \tag{4}$$

However, a theoretical statement cannot pose assumptions on the *random* matrix $\hat{\Sigma}$. Because of the noise in the model, we also cannot transfer condition (4) directly to Σ. Instead, we ask that on the set

$$C := \{\beta \in \mathbb{R}^d : \|\beta_{S(\beta^*)^c}\|_1 \leq 3\|\beta_{S(\beta^*)}\|_1\}$$

it holds that

$$\Lambda_{min}(\Sigma) := \inf_{v \in C} \frac{v^T \Sigma v}{\|v_{S(\beta^*)}\|_2^2} > 0,$$

the so-called *restricted eigenvalue property (REP)*. Using $\Lambda_{min}(\Sigma)$, we can state the following theorem for the LASSO estimator.

Theorem 2.10. Let $\varepsilon \sim N(0, \sigma^2)$, $X \sim N(0, \Sigma)$ and $\Sigma_{jj} = 1$ $(j = 1, ..., d)$. Define $s := \#S(\beta^*)$. Then there exist universal constants $c_1, c_2 > 0$ such that the condition

$$n \geq c_1 \frac{\|\Sigma\|}{\Lambda_{min}(\Sigma)^2} s \log(ed/s)$$

implies: For each $t \geq 0$ and

$$\lambda \geq \frac{6\sqrt{2}\sigma}{\sqrt{n}} \sqrt{\log(d) + t},$$

it holds that

$$\mathbb{P}(R(\hat{\beta}_\lambda) - R(\beta^*) > 16\lambda^2 \frac{s}{\Lambda_{min}(\Sigma)}) \leq e^{-t} + 2de^{-c_2 n}.$$

Proof. We abbreviate $S = S(\beta^*)$. The basic idea is to define two events on which we can replace $\hat{\Sigma} \approx \Sigma$. These events read

$$B_1 := \{\forall j \in \{1, ..., d\} : \hat{\Sigma}_{jj} \leq \frac{3}{2}\Sigma_{jj}\}$$

and

$$B_2 := \{\forall v \in C : \frac{1}{\sqrt{2}} \|\Sigma^{1/2} v\|_2 \leq \|\hat{\Sigma}^{1/2} v\|_2\}.$$

Then it holds that

$$\begin{aligned}
& \mathbb{P}\big(R(\hat{\beta}) - R(\beta^*) > 16\lambda^2 \frac{s}{\Lambda_{min}(\Sigma)}\big) \\
\leq\ & \mathbb{P}\big(\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 > 16\lambda^2 \frac{s}{\Lambda_{min}(\Sigma)}, B_1 \cap B_2\big) + \mathbb{P}(B_1^c) + \mathbb{P}(B_2^c) \\
\leq\ & \mathbb{E}\Big[\mathbb{P}\Big(\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 > 16\lambda^2 \frac{s}{\Lambda_{min}(\Sigma)} \Big| \mathbb{X}\Big) \mathbb{1}_{B_1 \cap B_2}\Big] + \mathbb{P}(B_1^c) + \mathbb{P}(B_2^c).
\end{aligned}$$

In the first probability, we can consider all terms depending on $\mathbb{X}$ as deterministic, that is, one only has to analyze the variation from $\mathbb{e}$.

1. Basic inequality: $\hat{R}_n(\hat{\beta}) + J_\lambda(\hat{\beta}) \leq \hat{R}_n(\beta^*) + J_\lambda(\beta)$ implies

$$\begin{aligned}
\frac{1}{n}\|\mathbb{X}(\hat{\beta} - \beta^*)\|_2^2 + \lambda\|\hat{\beta}\|_1 &\leq \frac{2}{n}\mathbb{e}^T\mathbb{X}(\hat{\beta} - \beta) + \lambda\|\beta^*\|_1 \\
&\leq \frac{2}{n}\|\mathbb{e}^T\mathbb{X}\|_\infty \cdot \|\hat{\beta} - \beta^*\|_1 + \lambda\|\beta^*\|_1.
\end{aligned}$$

2. Define $A := \{\frac{2}{n}\|\mathrm{e}^T\mathbb{X}\|_\infty \leq \frac{\lambda}{2}\}$. On A it holds that

$$\begin{aligned}\frac{2}{n}\|\mathbb{X}(\hat\beta-\beta^*)\|_2^2 &\overset{2\cdot(1)}{\leq} \lambda\cdot\Big\{\underbrace{\|\hat\beta-\beta^*\|_1}_{=\|\hat\beta_S-\beta_S^*\|_1+\|\hat\beta_{S^c}\|_1} + 2\underbrace{\|\beta^*\|_1}_{=\|\beta_S^*\|_1} - 2\underbrace{\|\hat\beta\|_1}_{=\|\hat\beta_S\|_1+\|\hat\beta_{S^c}\|_1}\Big\}\\ &\leq \lambda\cdot\Big\{\|\hat\beta_S-\beta_S^*\|_1 - \|\hat\beta_{S^c}\|_1 + 2(\|\beta_S^*\|_1-\|\hat\beta_S\|_1)\Big\}\\ &\leq \lambda\cdot\big\{3\|\hat\beta_S-\beta_S^*\|_1 - \|\hat\beta_{S^c}\|_1\big\}.\end{aligned}$$

3. Because of (2), on A it holds that: $\hat\beta-\beta^*\in C$.

4. On B_1 it holds that

$$\begin{aligned}\mathbb{P}(A^c|\mathbb{X}) &= \mathbb{P}\Big(\max_{j=1,\dots,d}\Big|\frac{1}{n}\sum_{i=1}^n X_{ij}\varepsilon_i\Big| > \frac{\lambda}{2}\Big|\mathbb{X}\Big)\\ &\leq d\max_{j=1,\dots,d}\mathbb{P}\Big(\Big|\underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\varepsilon_i}_{\sim N(0,\frac{\sigma^2}{n}\sum_{i=1}^n X_{ij}^2)=N(0,\sigma^2\hat\Sigma_{jj})} > \frac{\lambda\sqrt{n}}{4}\Big|\mathbb{X}\Big)\\ &\leq d\cdot 2(1-\Phi(\frac{\lambda\sqrt{n}}{4\sigma\hat\Sigma_{jj}^{1/2}})) \overset{\text{auf } D}{\leq} 2d\cdot\exp\Big(-\frac{1}{2}\big(\frac{\lambda\sqrt{n}}{6\sigma}\big)^2\Big) \overset{\text{Condition on }\lambda}{\leq} e^{-t}.\end{aligned}$$

   because of $1-\Phi(x)\leq e^{-\frac{x^2}{2}}$ (Φ the distribution function of $N(0,1)$).

5. On B_2 it holds that $\Lambda_{min}(\hat\Sigma)\geq\frac{\Lambda_{min}(\Sigma)}{2}$. On $A\cap B_2$ it holds that

$$\|\hat\beta_S-\beta_S^*\|_1^2 \leq s\cdot\|\hat\beta_S-\beta_S^*\|_2^2 \overset{\text{Def. }\Lambda_{min}(\hat\Sigma),\ (3)}{\leq} \underbrace{\|\hat\Sigma^{1/2}(\hat\beta-\beta^*)\|_2^2}_{=\frac{1}{n}\|\mathbb{X}(\hat\beta-\beta^*)\|_2^2}\cdot\frac{2s}{\Lambda_{min}(\Sigma)}.$$

6. On $A\cap B_2$ it holds that

$$\begin{aligned}\frac{2}{n}\|\mathbb{X}(\hat\beta-\beta^*)\|_2^2 + \lambda\cdot\underbrace{\|\hat\beta-\beta^*\|_1}_{=\|\hat\beta_S-\beta_S^*\|_1+\|\hat\beta_S\|_1} &\overset{(2)}{\leq} 4\lambda\|\hat\beta_S-\beta_S^*\|_1\\ &\overset{(5)}{\leq} \frac{1}{\sqrt{n}}\|\mathbb{X}(\hat\beta-\beta^*)\|_2\cdot 4\lambda\Big(\frac{2s}{\Lambda_{min}(\Sigma)}\Big)^{1/2}\\ &\overset{4ab\leq a^2+4b^2}{\leq} \frac{1}{n}\|\mathbb{X}(\hat\beta-\beta^*)\|_2^2 + 8\lambda^2\frac{s}{\Lambda_{min}(\Sigma)}.\end{aligned}$$

   Subtracting $\frac{1}{n}\|\mathbb{X}(\hat\beta-\beta^*)\|_2^2$ from both sides yields

$$\|\hat\Sigma^{1/2}(\hat\beta-\beta^*)\|_2^2 = \frac{1}{n}\|\mathbb{X}(\hat\beta-\beta^*)\|_2^2 \leq 8\lambda^2\frac{s}{\Lambda_{min}(\Sigma)}. \tag{5}$$

7. Replacement of $\hat{\Sigma}$ by Σ: On $A \cap B_2$ it holds that

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \leq 16\lambda^2 \frac{s}{\Lambda_{min}(\Sigma)}.$$

8. Bounding the probability ofB_1, B_2: [20, Theorem 1.6] (cf. also [13]) implies that there exists constants $c_1, c_2 > 0$, such that for $n \geq c_1 \frac{\|\Sigma\|}{\Lambda_{min}(\Sigma)^2} s \log(ed/s)$ it holds that

$$\mathbb{P}(B_2^c) \leq 2\exp(-c_2 n).$$

Thus,

$$\begin{aligned} \mathbb{P}(B_1^c) &\leq d \max_{j=1,\ldots,d} \mathbb{P}(\hat{\Sigma}_{jj} \geq \frac{3\Sigma_{jj}}{2}) \leq d \max_{j=1,\ldots,d} \mathbb{P}(\underbrace{\sum_{i=1}^{n} X_{ij}^2}_{\sim\chi^2(n)} > \frac{3n}{2}\Big) \\ &\leq d(\frac{3}{2}e^{1-\frac{3}{2}})^{n/2}, \end{aligned}$$

since $1 - F_{\chi^2(n)}(zn) \leq (ze^{1-z})^{n/2}$ (here, $F_{\chi^2(n)}$ is the distribution function of the χ^2-distribution, cf. [6, Lemma 2.2]).

□

Remarks:

- The upper bound for the convergence rate of the excess Bayes risk of the LASSO estimator is minimized for $\lambda = 6\sqrt{2} \cdot \frac{\sigma}{\sqrt{n}} \sqrt{\log(d)}$. Then, it reads

$$16\lambda^2 \frac{s}{\Lambda_{min}(\Sigma)} = \frac{\text{Zahl}}{\Lambda_{min}(\Sigma)} \cdot \frac{\sigma^2 s}{n} \cdot \log(d).$$

  This rate can be interpreted as follows: $\hat{\beta}_\lambda$ behaves like the LQ estimator in a model with s instead of d dimensions / components instead of d. The LASSO estimator $\hat{\beta}_\lambda$ has to 'pay' with a factor $\log(d)$ for the missing insight which components are non-zero. This is a rather small price to pay even if d is large.

- One can prove similar theoretical statements without the conditions $\varepsilon \sim N(0, \sigma^2)$ and $X \sim N(0, \Sigma)$ and still can preserve the small $\log(d)$ term.

- Regarding the REP: Under the normalization condition $\Sigma_{jj} = 1$, the eigenvalues of $\hat{\Sigma}$ and Σ are small if many components of X are strongly correlated (for instance, Σ is not invertible if the first components coincide, $X_1 = X_2$). Σ would also be non-invertible if one component of X has no variance, but this is excluded by the assumption that $\Sigma_{jj} = 1$. Therefore, the smallest eigenvalue $\lambda_{min}(\Sigma)$ measures

how strongly the components of X are correlated. Note that a strong correlation of X is a problem for estimation of β^*, but not for the excess Bayes risk itself: In the extreme case $X_1 = X_2$, it is clear that $\hat{\beta}$ cannot distinguish the values of β_1^* and β_2^*, but it can still provide good predictions through $X\hat{\beta}$. Unfortunately, our proof technique transfers the estimation quality of β^* via (5) to an upper bound of the excess Bayes risk, therefore this fact is not adequately represented in the result.
Roughly speaking, the REP reduces the condition on the injectivity on Σ to a condition on the injectivity on the components with indices $S(\beta^*)$.

- The assumption $\Sigma_{jj} = 1$ is only to provide an easier result. In practice, this normalization can be obtained by standardizing $X_1, ..., X_n$ before computing the LASSO estimator (that is, center X_i and divide by the empirical standard deviation).

2.4 Exercises

Task 1 (Discussion: Proof of Theorem 2.5)**.** Let $A \in \mathbb{R}^{d\times d}$ and $\mathbb{e} \sim N(0, \sigma I_{d\times d})$.

1. Show that
$$\mathbb{P}(\|A\mathbb{e}\|_2 \geq \sigma\|A\|_F\sqrt{t}) \leq \frac{1}{t}.$$
2. Discuss in which sense this result is weaker than the result given in the lecture,
$$\mathbb{P}(\|A\mathbb{e}\|_2 \geq c_2\sigma\|A\|_F \max\{1, \sqrt{t}\}) \leq e^{-t}.$$

Now let $X_1, ..., X_n : \Omega \to \mathbb{R}^d$ be i.i.d. $N(0, \Sigma)$ distributed. Let $E_n := \{\|\hat{\Sigma} - \Sigma\| \leq \frac{\lambda_{min}(\Sigma)}{2}\}$. Denote by $\|\Sigma\|$ the spectral norm and $\|\Sigma\|_F^2 := \sum_{j,k=1}^d \Sigma_{jk}^2$ the Frobenius norm.

(c) Show that
$$\mathbb{P}(\|\hat{\Sigma} - \Sigma\| \geq x) \leq \frac{1}{n}\frac{\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}}{x^2}.$$
Hint: It holds that $\|\Sigma\| \leq \|\Sigma\|_F$ *and* $\mathbb{E}[X_{1j}^2 X_{1k}^2] = \Sigma_{jj}\Sigma_{kk} + 2\Sigma_{jk}^2$.

(d) Conclude that
$$\mathbb{P}\Big(\|\hat{\Sigma} - \Sigma\| \geq \frac{1}{\sqrt{n}}\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}^{1/2}x\Big) \leq \frac{1}{x^2}.$$

(e) Discuss in which sense the above inequality is weaker than the lemma from the lecture,
$$\mathbb{P}\Big(\|\hat{\Sigma} - \Sigma\| \geq c_1\|\Sigma\| \cdot \max\{\frac{d}{n}, \sqrt{\frac{d}{n}}, \frac{x}{n}, \sqrt{\frac{x}{n}}\}\Big) \leq e^{-x}.$$

Task 2 (Ridge estimator)**.** In practice it is common to consider the linear model $Y = X^T\beta + \varepsilon$ with $X_1 = 1$, that is, the first component of X is constantly 1 and not random. In this case, the ridge estimator is replaced by

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^d} \big\{\hat{R}_n(\beta) + \lambda \cdot \sum_{j=2}^{d} \beta_j^2\big\},$$

that is, the first component is not penalized in its size.

1. Why this is meaningful?
2. Derive an explicit expression of $\hat{\beta}$ in terms of $\mathbb{X}, \mathbb{Y}$.

Task 3 (Large deviations)**.**

1. Let $X \sim N(0,1)$. Show that for $t > 0$, $\mathbb{P}(X \geq t) \leq \exp(-\frac{t^2}{2})$.
   Hint: You may use Markov's inequality with $g(t) = e^{ct}$ and appropriately chosen $c \geq 0$.
2. Let $X \sim N(0,\sigma^2)$. Show that

$$\mathbb{P}(|X| > \sigma t) \leq 2\exp(-\frac{t^2}{2}), \qquad \mathbb{P}(|X| > \sigma\sqrt{2t}) \leq 2e^{-t}.$$

Let $X_1, .., X_n$ be i.i.d. with $\mathbb{E}[X_1^2] = \sigma^2$ and $|X_1| \leq M$, where $\sigma^2, M > 0$ are constants. Then, the so-called Bernstein inequality holds: For all $x > 0$,

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq x\Big) \leq \exp\Big(-\frac{1}{2}\frac{x^2}{\sigma^2 + \frac{Mx}{\sqrt{n}}}\Big).$$

(c) Discuss the relation between Bernstein's inequality and the statement in (a).

(d) Show that for any $t > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq \sqrt{n}\sigma\sqrt{t} + Mt\Big) \leq e^{-t/4}.$$

(e) Show that with probability $\geq 1 - \delta$ ($\delta \in (0,1)$), it holds that

$$\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \leq 2\sqrt{n}\sigma\sqrt{\log(\frac{1}{\delta})} + 4M\log(\frac{1}{\delta}).$$

Task 4 (Discussion: Theorem 2.10)**.** The convergence rate of the LASSO estimator from Theorem 2.10 reads

$$\frac{1}{\Lambda_{min}(\Sigma)} \cdot \frac{\sigma^2 s}{n} \cdot \log(d).$$

One of the assumptions in the theorem is that $n \geq c_1 \frac{\|\Sigma\|}{\Lambda_{min}(\Sigma)^2} s \cdot \log(ed/s)$.

1. Compare this assumption on n with the assumption $n \geq c_1' \cdot \frac{\|\Sigma\|^2}{\lambda_{min}(\Sigma)^2} \cdot d$ stated in the theorem of the LS estimator.
2. Recall: Why the additional factor $\frac{1}{\Lambda_{min}(\Sigma)}$ occurs in the rate of the LASSO estimator which is not present for the LS estimator?

Regarding the Restricted Eigenvalue Property: Recall that

$$\Lambda_{min}(\Sigma) = \inf_{v \in C} \frac{v^T \Sigma v}{\|v_S\|_2^2}, \qquad C := \{v \in \mathbb{R}^d : \|v_{S^c}\|_1 \leq 3\|v_S\|_1\}.$$

(c) Show the following statement: If $\Sigma = I_{d\times d}$, then $\Lambda_{min}(\Sigma) \geq 1$.

(d) Discuss why the scenario in (c) in not realistic, in particular for large d.

Consider for $\rho \in (0,1)$

$$\Sigma = \begin{pmatrix} 1 & 0 & \dots & 0 & \rho \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \rho \\ 0 & \dots & 0 & 1 & \rho \\ \rho & \dots & \rho & \rho & 1 \end{pmatrix}$$

One can show that the eigenvalues of Σ are given by $\{1, ..., 1, 1-(d-1)^{1/2}\rho, 1+(d-1)^{1/2}\rho\}$.

(e) Determine a condition for ρ such that $\lambda_{min}(\Sigma) > 0$.

(f) Let $S = \{1, ..., s\}$. Derive a lower bound for $\Lambda_{min}(\Sigma)$ which only depends on s but not on d. From this, derive a condition for ρ such that $\Lambda_{min}(\Sigma) > 0$.

Task 5 (Proof of Theorem 2.10: Analysis of B_1)**.** In Theorem 2.10 we defined the event

$$B_1 := \{\forall j \in \{1, ..., d\} : \hat{\Sigma}_{jj} \leq \frac{3}{2}\Sigma_{jj}\},$$

where $\Sigma = \mathbb{E}[XX^T]$, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T$ and $\Sigma_{jj} = 1$ ($j \in \{1, ..., d\}$).

1. Show that for $t < \frac{1}{2}$, one has

$$\mathbb{P}(B_1^c) \leq d \cdot ((1-2t)e^{3t})^{-n/2}.$$

Hint: For $Z \sim N(0,1)$ it holds that $\mathbb{E}e^{tZ^2} = (1-2t)^{-1/2}$.

2. Let $c := \frac{3}{4} + \log(\frac{1}{2}) > 0$. Show that for $n \geq \frac{2}{c}(\log(d) + x)$, we have $\mathbb{P}(B_1^c) \leq e^{-x}$.

Task 6 (Generalizations of Theorem 2.10)**.** In the situation of Theorem 2.10, we investigate the probability

$$\mathbb{P}\Big(\frac{2}{n}\|\mathrm{e}^T\mathbb{X}\|_\infty \geq \frac{\lambda}{2}\Big|\mathbb{X}\Big) = \mathbb{P}\Big(\max_{j=1,\dots,d}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big|\mathbb{X}\Big)$$

and the corresponding choice of λ. For simplicity, we assume that $\mathbb{X}$ is deterministic and satisfies $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$. Additionally, let ε_i, $i = 1, ..., n$ be i.i.d. with $\mathbb{E}[\varepsilon_1^2] = \sigma^2$.

1. Repeat the proof of the lecture: Show that if $\varepsilon \sim N(0, \sigma^2)$, we have

$$\mathbb{P}\big(\frac{2}{n}\|\mathrm{e}^T\mathbb{X}\|_\infty \geq \frac{\lambda}{2}\big) \leq 2d\exp\Big(-\frac{1}{2}\big(\frac{\lambda\sqrt{n}}{4\sigma}\big)^2\Big).$$

2. Prove that

$$\mathbb{P}\Big(\frac{2}{n}\|\mathrm{e}^T\mathbb{X}\|_\infty \geq \frac{\lambda}{2}\Big) \leq d\cdot\Big(\frac{4\sigma}{\sqrt{n}\lambda}\Big)^2.$$

3. Let $p \geq 2$. Show the following statement: If $\mathbb{E}[|\varepsilon|^p] = \mu_p^p$ with some $\mu_p \in \mathbb{R}$, then it holds that

$$\mathbb{P}\Big(\frac{2}{n}\|\mathrm{e}^T\mathbb{X}\|_\infty \geq \frac{\lambda}{2}\Big) \leq d\cdot\Big(\frac{4p^{1/2}\mu_p}{\lambda\sqrt{n}}\Big)^p.$$

   Hint: For independent random variables A_i, $i = 1, ..., n$ the inequality $\mathbb{E}[|\sum_{i=1}^n A_i|^p]^{1/p} \leq p^{1/2}(\sum_{i=1}^n \mathbb{E}[|A_i|^p]^{2/p})^{1/2}$ holds (cf. [?], Theorem 2.1).

Starting from now, suppose that $|X_{ij}| \leq C$ ($i = 1, ..., n$, $j = 1, ..., d$) with some constant $C > 0$.

(d) The following inequality is called Nemirovski's inequality (cf. Lemma 14.24 in [5]): For independent random variables $A_i \in \mathbb{R}^d$, $i = 1, ..., n$, we have

$$\mathbb{E}\big[\max_{j=1,\dots,d}\big|\sum_{i=1}^n (A_{ij} - \mathbb{E}A_{ij})\big|\big] \quad\leq\quad (8\log(2d))^{1/2}\cdot\mathbb{E}\big[\max_{j=1,\dots,d}\sum_{i=1}^n A_{ij}^2\big]^{1/2}.$$

Show that

$$\mathbb{P}\Big(\frac{2}{n}\|\mathrm{e}^T\mathbb{X}\|_\infty \geq \frac{\lambda}{2}\Big) \leq \frac{4(8\log(2d))^{1/2}C\sigma}{\lambda\sqrt{n}}.$$

3 Basics of classification problems; linear models in classification problems

In this chapter we assume that $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{1, ..., K\}$ with $K \in \mathbb{N}$ the number of classes, $L(y, s) = \mathbb{1}_{\{y \neq s\}}$ the 0-1-loss. We know that in this case, the Bayes rule reads

$$f^*(x) \in \underset{k \in \{1,...,K\}}{\arg\max} \mathbb{P}(Y = k | X = x).$$

Caution: For regression problems, f^* is $\mathbb{P}^X$-a.s. uniquely determined. In classification problems it can happen that f^* is *nowhere* unique. For instance, in the special case of $K = 2$ classes it could be that for all $x \in \mathcal{X}$, $\mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 2|X = x) = \frac{1}{2}$. In this chapter we make the high-level assumption that f^* is uniquely determined.

3.1 Decision regions, decision boundaries, discriminant functions

Each decision rule $f : \mathcal{X} \to \mathcal{Y}$ partitions the space $\mathcal{X}$ into decision regions which are separated by decision boundaries.

Definition 3.1. Let $f : \mathcal{X} \to \mathcal{Y}$ be a decision rule. Then

$$\Omega_k(f) := \{x \in \mathcal{X} : f(x) = k\}, \quad k \in \mathcal{Y}$$

is called the *decision region* associated to f and class k. The set

$$\partial\Omega_{kl}(f) := \partial\Omega_k(f) \cap \partial\Omega_l(f)$$

is called *decision boundary* between the classes $k, l \in \mathcal{Y}$.
Accordingly, $\Omega_k^* := \Omega_k(f^*)$ and $\partial\Omega_{kl}^* := \partial\Omega_{kl}(f^*)$ are called *optimal decision regions / decision boundaries*.

In many classification problems, it is hard to estimate f^* directly. One therefore tries to transfer the estimation problem to so-called discriminant functions.

Definition 3.2. Let $f : \mathcal{X} \to \mathcal{Y}$ be a decision rule. Measurable mappings $\delta_k = \delta_k(f) : \mathcal{X} \to \mathbb{R}$ are called discriminant functions with respect to f if

$$\Omega_k(f) = \{x \in \mathcal{X} : \delta_k(x) = \max_{l \in \mathcal{Y}} \delta_l(x)\}.$$

Accordingly, $\delta_k^* : \mathcal{X} \to \mathbb{R}$ are called *optimal discriminant functions* if

$$\Omega_k^* := \{x \in \mathcal{X} : \delta_k(x) = \max_{l \in \mathcal{Y}} \delta_l(x)\}.$$

There exist 'standard' discriminant functions.

Lemma 3.3. 1. The functions $\delta_k^*(x) = \mathbb{P}(Y = k|X = x)$ are optimal discriminant functions.

2. If $h : (0,1) \times \mathcal{X} \to \mathbb{R}$ is strictly increasing with respect to its first component, then

$$\delta_k^*(x) = h(\mathbb{P}(Y = k|X = x), x)$$

are optimal discriminant functions as well.

Proof. It holds that

$$\Omega_k^* := \{x \in \mathcal{X} : f^*(x) = k\} = \{x \in \mathcal{X} : \mathbb{P}(Y = k|X = x) = \max_{l \in \mathcal{Y}} \mathbb{P}(Y = l|X = x)\}.$$

The second statement follows from the equivalence

$$\begin{aligned} & \mathbb{P}(Y = k|X = x) = \max_{l \in \mathcal{Y}} \mathbb{P}(Y = l|X = x) \\ \iff \quad & h(\mathbb{P}(Y = k|X = x), x) = \max_{l \in \mathcal{Y}} h(\mathbb{P}(Y = l|X = x), x). \end{aligned}$$

□

One standard approach to estimate f^* is given by estimating δ_k^*, $k \in \mathcal{Y}$ under a specific model assumption. A very simple model is given by assuming that the decision boundaries have a linear structure.

Definition 3.4. A decision rule f has *linear decision boundaries* if there exist linear discrimininant functions for f. Formally: There exist $\beta^{(1)}, ..., \beta^{(K)} \in \mathbb{R}^d$ and discriminant functions $\delta_k(f)$ such that $\delta_k(f) = x^T\beta^{(k)}$, $k = 1, ..., K$.
A distribution $\mathbb{P}^{(X,Y)}$ has *linear optimal decision boundaries* if there exist linear optimal discriminant functions.

The motivation for this definition is as follows: The decision boundaries are sets where two discriminant functions coincide. That is,

$$\partial\Omega_{kl}(f) \subset \{x \in \mathcal{X} : x^T\beta^{(k)} = x^T\beta^{(l)}\},$$

are then subsets of a line.
Caution: If a distribution has linear decision boundaries, then one only knows that optimal linear discriminant function *exist*, but there are a lot of other nonlinear (optimal) discriminant functions. An example is given in the next subsection.

3.2 Logistic regression

Approach: We want to define a model with optimal *linear* decision boundaries, that is, we ask the model to allow for linear optimal discriminant functions. A possible assumption could look like

$$\mathbb{P}(Y = k|X = x) = \delta_k^*(x) \stackrel{!}{=} x^T \beta^{*(k)}$$

with some $\beta^{*(k)} \in \mathbb{R}^d$, $k = 1, ..., K$.

Problem: This is not a reasonable approach since the left hand side attains values in $\in [0, 1]$, but the right hand side can attain all values in $\mathbb{R}$.

Instead, we make use of Lemma 3.3(ii): We assume that

$$h(\mathbb{P}(Y = k|X = x), x) = \delta_k^*(x) \stackrel{!}{=} x^T \beta^{*(k)}, \quad k = 1, ..., K$$

with a suitable $h : [0, 1] \times \mathcal{X} \to \mathbb{R}$ which is strictly increasing with respect to its first argument.

Definition 3.5 (Model: Logistic regression)**.** For $k = 1, ..., K-1$ there exist $\beta^{*(k)} \in \mathbb{R}^d$ such that for all $x \in \mathcal{X}$,

$$\log\Big(\frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = K|X = x)}\Big) = x^T \beta^{*(k)}. \tag{6}$$

Lemma 3.6. The following conditions are equivalent to the model assumption from Definition (6)

$$\begin{aligned}
\mathbb{P}(Y = k|X = x) &= \frac{\exp(x^T \beta^{*(k)})}{1 + \sum_{k=1}^{K-1} \exp(x^T \beta^{*(k)})}, \quad k = 1, ..., K-1, \\
\mathbb{P}(Y = K|X = x) &= \frac{1}{1 + \sum_{k=1}^{K-1} \exp(x^T \beta^{*(k)})}.
\end{aligned}$$

Proof. Apply exp on both sides of (6) and make use of $\sum_{k=1}^{K} \mathbb{P}(Y = k|X = x) = 1$. □

Remark: The name 'logistic regression' is based on the representation of Lemma 3.6. $\mathbb{P}(Y = k|X = x)$ is modeled by a so-called logistic function $\frac{e^z}{1+e^z}$.

We motivate an algorithm by using a standard technique from statistics, the maximum likelihood approach for estimation of the parameters $\beta^{*(k)}$, $k = 1, ..., K-1$. We will see later that this approach is equivalent to minimizing an empirical risk.
Let $f(x_1, y_1, ..., x_n, y_n)$ denote the joint density of the random variables (X_i, Y_i), $i = 1, ..., n$, $f_{X,Y}$ the density of (X, Y) (with respect to the product measure of the Lebesgue and the counting measure), and f_X the density of X with respect to the Lebesgue measure. Then it holds that

$$\begin{aligned} \log f(x_1, y_1, ..., x_n, y_n) &= \sum_{i=1}^{n} \log \underbrace{f_{X,Y}(x_i, y_i)}_{= \frac{f_{X,Y}(x_i,y_i)}{f_X(x_i)} \cdot f_X(x_i)} \\ &= \sum_{i=1}^{n} \log \mathbb{P}(Y = y_i|X = x_i) + \underbrace{\sum_{i=1}^{n} \log f_X(x_i)}_{\text{unknown, we will omit it!}} . \end{aligned}$$

The maximum likelihood approach asks for maximizing this quantity with respect to $\beta^{(k)}$, $k = 1, ..., K-1$.
Lemma 3.6 $\Rightarrow$

$$\begin{aligned} \log \mathbb{P}(Y = y|X = x) &= \sum_{j=1}^{K-1} \mathbb{1}_{\{y=j\}} \Big\{ x^T \beta^{(j)} - \log \Big(1 + \sum_{k=1}^{K-1} \exp(x^T \beta^{(k)}) \Big) \Big\} \\ &\quad - \mathbb{1}_{\{y=K\}} \log \Big(1 + \sum_{k=1}^{K-1} \exp(x^T \beta^{(k)}) \Big) \\ &= \Big\{ \sum_{k=1}^{K-1} \mathbb{1}_{\{y=k\}} \cdot x^T \beta^{(k)} \Big\} - \log \Big(1 + \sum_{k=1}^{K-1} \exp(x^T \beta^{(k)}) \Big) \end{aligned}$$

Instead of $f(X_1, Y_1, ..., X_n, Y_n)$ we can therefore *minimize*

$$L_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \Big[\log \Big(1 + \sum_{k=1}^{K-1} \exp(X_i^T \beta^{(k)}) \Big) - \sum_{k=1}^{K-1} \mathbb{1}_{\{Y_i=k\}} \cdot X_i^T \beta^{(k)} \Big] \tag{7}$$

with respect to $\theta = ((\beta^{(1)})^T, ..., (\beta^{(K-1)})^T)^T \in \mathbb{R}^{(K-1)d}$.

Definition 3.7 (Classifier: Logistic regression)**.** Let

$$\hat{\theta} = ((\hat{\beta}^{(1)})^T, ..., (\hat{\beta}^{(K-1)})^T)^T \in \underset{\theta \in \mathbb{R}^{(K-1)d}}{\arg\min}\, L_n(\theta).$$

Let

$$\hat{\delta}_k^{LR}(x) = \begin{cases} x^T \hat{\beta}^{(k)}, & k = 1, ..., K-1, \\ 0, & k = K \end{cases}.$$

The logistic regression classifier is given by $\hat{f}_n^{LR}(x) = \arg\max_{k \in \{1,...,K\}} \hat{\delta}_k^{LR}(x)$.

3.2.1 Theoretical statements

The theoretical investigation of classification problems with more than two classes K is often laborious. To some extend, the results for $K = 2$ classes can be generalized to an arbitrary K classes. Therefore, we restrict ourselves to the investigation of problems with $K = 2$ classes. Moreover, we use a different notation for the classes:

$$\mathcal{Y} = \{+1, -1\} \quad \text{instead of} \quad \mathcal{Y} = \{1, 2\}.$$

This leads to more compact formulas for the classifiers (the main advantage is that we can express classifiers as sign functions of discriminant functions, see below). Put

$$\Delta = \{\delta_\beta(x) = x^T \beta : \beta \in \mathbb{R}^d\}.$$

Since $\mathbb{1}_{\{Y_i=1\}} = \frac{1}{2}(Y_i + 1)$, we have the following expression.

Definition 3.8 (Logistic regression for 2 classes, risk minimization formulation)**.**

$$\tilde{L}(y, s) := \log(1 + e^s) - \frac{1}{2}(y + 1)s.$$

Put

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^d}{\arg\min}\, \tilde{R}_n(\delta_\beta), \qquad \tilde{R}_n(\delta) := \frac{1}{n} \sum_{i=1}^n \tilde{L}(Y_i, \delta(X_i)).$$

Put $\hat{\delta}_n^{LR}(x) = x^T \hat{\beta}$. The logistic regression classifier reads

$$\hat{f}_n^{LR}(x) = \text{sign}(\hat{\delta}_n^{LR}(x)), \qquad \text{sign}(z) := \begin{cases} 1, & z > 0, \\ 0, & z = 0, \\ -1, & z < 0. \end{cases}$$

Contrary to the formulation before, we only have to define *one* discriminant function (instead of two). Moreover, we have a simple expression for the classifier in terms of the discriminant function.
Note that the loss function which is used in the above risk minimization problem is *not* the 0-1- loss but a more complicated function. Since the logistic regression classifier minimizes the risk with respect to $\tilde{L}$, theoretical statements can only proved 'directly' for the modified risk $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y, \delta(X))$. Therefore, our first goal is to show a theoretical result for the excess Bayes risk with respect to $\tilde{L}$,

$$\tilde{R}(\hat{\delta}_n^{LR}) - \tilde{R}(\delta^*). \tag{8}$$

Afterwards we will think about how an upper bound can be transferred to the excess Bayes risk

$$R(\hat{f}_n^{LR}) - R(f^*)$$

with respect to the 0-1-loss. For the next statements, we use the following important abbreviation.

Definition 3.9. $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

We now have to convince ourselves that under the model assumption of logistic regression (Definition 3.5)

$$\log(\frac{\eta(x)}{1-\eta(x)}) = x^T\beta^* \quad \text{für ein } \beta^* \in \mathbb{R}^d$$

the function $\delta_{\beta^*}(x) = x^T\beta^*$ is a Bayes rule with respect to the risk $\tilde{R}$ (even if we minimize over *all* possible measurable mappings $\delta : \mathcal{X} \to \mathbb{R}$!). Later in the proof of an upper bound of the excess Bayes risk, we will need also a more specific property of $\tilde{R}(\cdot)$, the so called *quadratic margin property*.

Lemma 3.10 (Quadratic margin property)**.** Suppose that there exists $c > 0$ such that $c \leq \eta(x) \leq 1 - c$ for all $x \in \mathcal{X}$. Then it holds that

$$\tilde{R}(\delta_{\beta^*}) = \min_{\delta:\mathcal{X}\to\mathbb{R}} \tilde{R}(\delta).$$

Suppose that $X \sim N(0, \Sigma)$. Then, for all $\beta \in \mathbb{R}^d$ with $\|\Sigma^{1/2}(\beta - \beta^*)\|_2 \leq c^2$ it holds that

$$\tilde{R}(\delta_\beta) - \tilde{R}(\delta_{\beta^*}) \geq \frac{c^2}{4} \cdot \|\Sigma^{1/2}(\beta - \beta^*)\|_2^2.$$

Remark: The assumption $X \sim N(0, \Sigma)$ can be weakened. Here we stick to this assumption since it simplifies the result and its proof significantly.

Proof. **Proof of the minimization property:** It holds that

$$\mathbb{E}[\tilde{L}(Y, \delta(X))|X = x] = -\eta(x)\delta(x) + \log(1 + \exp(\delta(x)) = g(\delta(x)). \tag{9}$$

with $g(s) = -\eta(x)s + \log(1 + e^s)$. It is enough to show that for all $x \in \mathcal{X}$ it holds that $g(\delta(x)) \geq g(\delta_{\beta^*}(x))$ (*). Then, by application of $\mathbb{E}[\cdot]$ to (9) we obtain: $R(\delta) \geq R(\delta_{\beta^*})$. It holds that

$$g'(s) = -\eta(x) + \frac{e^s}{1 + e^s}, \qquad g''(s) = \frac{e^s}{(1 + e^s)^2}, \qquad \|g'''\|_\infty \leq \frac{1}{6\sqrt{3}}.$$

By our model assumption, we have $\delta_{\beta^*}(x) = \log(\frac{\eta(x)}{1-\eta(x)})$, thus

$$g'(\delta_{\beta^*}(x)) = 0, \qquad g''(\delta_{\beta^*}(x)) = \eta(x) \cdot (1 - \eta(x)) \geq c^2.$$

Thus $\delta_{\beta^*}(x)$ is a (local) minimizer of $s \mapsto g(s)$. Since it is the only extremal point and g is continuous, we obtain (*).
Proof of the margin property: A Taylor expansion of g in $s = \delta_{\beta^*}(x)$ yields:

$$\begin{aligned} g(s) &= g(\delta_{\beta^*}(x)) + (s - \delta_{\beta^*}(x)) \cdot g'(\delta_{\beta^*}(x)) + \frac{1}{2}(s - \delta_{\beta^*}(x))^2 \cdot g''(\delta_{\beta^*}(x)) \\ &\quad + \frac{1}{2}(s - \delta_{\beta^*}(x))^2 \cdot [g''(\xi) - g''(\delta_{\beta^*}(x)], \end{aligned}$$

where $|\xi - \delta_{\beta^*}(x)| \leq |s - \delta_{\beta^*}(x)|$.

Thus,

$$\begin{aligned} \tilde{R}(\delta_\beta) - \tilde{R}(\delta_{\beta^*}) &\geq \frac{c^2}{2}\mathbb{E}[(\delta_\beta(X) - \delta^*(X))^2] - \frac{1}{12\sqrt{3}}\mathbb{E}[|\delta_\beta(X) - \delta_{\beta^*}(X)|^3] \\ &= \frac{c^2}{2}\|\Sigma^{1/2}(\beta - \beta^*)\|_2^2 - \frac{1}{12\sqrt{3}}\mathbb{E}[|X^T(\beta - \beta^*)|^3]. \end{aligned}$$

For $X \sim N(0, \Sigma)$ it holds that

$$\mathbb{E}[|X^T(\beta - \beta^*)|^3] \leq 3^{3/2}\mathbb{E}[|X^T(\beta - \beta^*)|^2]^{3/2} = 3^{3/2}\|\Sigma^{1/2}(\beta - \beta^*)\|_2^3.$$

The condition on β in the lemma implies

$$\tilde{R}(\delta_\beta) - \tilde{R}(\delta_{\beta^*}) \geq \frac{c^2}{2}\|\Sigma^{1/2}(\beta - \beta^*)\|_2^2 - \frac{1}{4}\|\Sigma^{1/2}(\beta - \beta^*)\|_2^3 \geq \frac{c^2}{4}\|\Sigma^{1/2}(\beta - \beta^*)\|_2^2.$$

□

To obtain an upper bound for the excess Bayes risk, we will need the following technical lemma (cf. [5], Theorem 14.3 and Theorem 14.4) which contains the so-called concentration inequalities of Talagrand and Ledoux.

Lemma 3.11. Suppose that $\tilde{L} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is such that $|\tilde{L}(y,s) - \tilde{L}(y,s')| \leq \ell \cdot |s - s'|$ for all $s, s' \in \mathbb{R}$ and $y \in \mathcal{Y}$. Suppose that $\Delta \subset \{\delta : \mathcal{X} \to \mathbb{R}\text{meas.}\}$ is a set of functions and $\delta^* \in \Delta$. Then it holds that

$$\begin{aligned} & \mathbb{E} \sup_{\delta \in \Delta} \Big| \frac{1}{n} \sum_{i=1}^{n} \{\tilde{L}(Y_i, \delta(X_i)) - \tilde{L}(Y_i, \delta^*(X_i)) - [\mathbb{E}\tilde{L}(Y_1, \delta(X_1)) - \mathbb{E}\tilde{L}(Y_1, \delta^*(X_1))]\} \Big| \\ \leq \; & 4\ell \cdot \mathbb{E} \sup_{\delta \in \Delta} \Big| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \big\{\delta(X_i) - \delta^*(X_i)\big\} \Big|, \end{aligned}$$

where $\varepsilon_i, i = 1, ..., n$ are i.i.d. and independent of (X_i, Y_i), $i = 1, ..., n$ with distribution $\mathbb{P}(\varepsilon_1 = 1) = \mathbb{P}(\varepsilon_1 = -1) = \frac{1}{2}$ (so-called Rademacher variables).

Caution: From now on, we abbreviate $\tilde{R}_n(\beta) = \tilde{R}_n(\delta_\beta)$, $\tilde{R}(\beta) = \tilde{R}(\delta_\beta)$.
We have the following result for the excess Bayes risk of logistic regression.

Theorem 3.12. Suppose that there exists $c \in (0,1)$ such that $c \leq \eta(x) \leq 1 - c$ for all $x \in \mathcal{X}$. Suppose that $X \sim N(0, \Sigma)$. Let $t \geq 1$. Suppose that $(\frac{d}{n})^{1/2} t \leq \frac{c^2}{128}$. Then it holds that

$$\mathbb{P}(\tilde{R}(\hat{\beta}) - \tilde{R}(\beta^*) \geq (\frac{32}{c})^2 t^2 \frac{d}{n}) \leq \frac{1}{t}.$$

Proof. General remark: In the proof, we introduce some variables $\gamma, a > 0$. They are first arbitrary and are chosen suitable at the end of the proof.

Derivation of a basic inequality: It holds that

$$\tilde{R}_n(\hat{\beta}) \leq \tilde{R}_n(\beta^*),$$

thus

$$\begin{aligned} \tilde{R}(\hat{\beta}) - \tilde{R}(\beta^*) &\leq \tilde{R}_n(\hat{\beta}) - \tilde{R}_n(\beta^*) - \{\tilde{R}_n(\beta) - \tilde{R}(\beta) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))\} \\ &\leq |\tilde{R}_n(\hat{\beta}) - \tilde{R}(\hat{\beta}) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))|. \end{aligned} \tag{10}$$

Now we use a 'trick': If $T \in [0,1]$ and $\tilde{\beta} = T\hat{\beta} + (1-T)\beta^*$, then convexity of $\tilde{L}(y,\cdot)$ implies

$$\tilde{R}_n(\tilde{\beta}) \leq T \underbrace{\tilde{R}_n(\hat{\beta})}_{\leq \tilde{R}_n(\beta^*)} + (1-T)\tilde{R}_n(\beta^*) \leq R_n(\beta^*).$$

As before, we have

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) \leq |\tilde{R}_n(\tilde{\beta}) - \tilde{R}(\tilde{\beta}) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))| \tag{11}$$

For $\gamma > 0$, define

$$T := \frac{\gamma}{\gamma + \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2}.$$

Then it holds that

$$\tilde{\beta} - \beta^* = T(\hat{\beta} - \beta^*) \quad \Rightarrow \quad \|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 = T\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq \gamma. \tag{12}$$

From (11) we obtain

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) \leq \sup_{\beta: \|\Sigma^{1/2}(\beta - \beta^*)\|_2 \leq \gamma} \left|\tilde{R}_n(\beta) - \tilde{R}(\beta) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))\right| =: Z_\gamma. \tag{13}$$

(The important point in this 'trick' is the replacement of $\hat{\beta}$ by $\tilde{\beta}$. Now we can bound the right hand side of the basic inequality with a supremum over a *bounded* set).

Definition of a 'nice' event: Let $A := \{Z_\gamma \leq \gamma \cdot a\}$.
If $\gamma \leq c^2$, then we have $\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 \leq \gamma \leq c^2$. Lemma 3.10 $\Rightarrow$

$$\frac{c^2}{4}\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2^2 \leq \tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) \overset{\text{auf } A}{\leq} \gamma \cdot a$$

We obtain (note that in the following calculation, '!' is a condition we have to meet at the end of the proof with our choices of a, γ):

$$\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 \leq \frac{2}{c}(\gamma a)^{1/2} \overset{!}{\leq} \frac{\gamma}{2}.$$

From (12) we obtain

$$\frac{\gamma\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2}{\gamma + \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2} = \|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 \leq \frac{\gamma}{2} \quad \overset{\text{Umstellen}}{\Rightarrow} \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq \gamma.$$

Now, we repeat the whole proof with $\hat{\beta}$ instead of $\tilde{\beta}$ (starting from (10) instead of (11)). Then on A, we have (cf. (13)):

$$\tilde{R}(\hat{\beta}) - \tilde{R}(\beta^*) \leq Z_\gamma \leq \gamma \cdot a.$$

Upper bound for the probability of A^c: It holds that

$$|\tilde{L}(y,s) - \tilde{L}(y,s')| \leq |\frac{1}{2}(y+1)| \cdot |s-s'| + 1 \cdot |s-s'| \leq 2|s-s'|,$$

thus Lemma 3.11 with $\ell = 2$ implies

$$\begin{aligned}
\mathbb{E}Z_\gamma &\leq 8 \cdot \mathbb{E} \sup_{\|\Sigma^{1/2}(\beta-\beta^*)\|_2 \leq \gamma} \Big| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i^T (\beta - \beta^*) \Big| \\
&\overset{\Sigma \text{ einfügen + CSU}}{\leq} 8\mathbb{E}\Big[\sup_{\|\Sigma^{1/2}(\beta-\beta^*)\|_2 \leq \gamma} \big\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (\Sigma^{-1/2} X_i) \big\|_2 \big\| \Sigma^{1/2}(\beta - \beta^*) \big\|_2 \Big] \\
&\overset{\text{Hölder } \mathbb{E}[|Z|] \leq \mathbb{E}[Z^2]^{1/2}}{\leq} \frac{8\gamma}{n} \mathbb{E}\Big[\underbrace{\Big\| \sum_{i=1}^n \varepsilon_i (\Sigma^{-1/2} X_i) \Big\|_2^2}_{= \sum_{j=1}^d (\sum_{i=1}^n \varepsilon_i (\Sigma^{-1/2} X_i)_j)^2} \Big]^{1/2} \\
&\overset{\text{unabh.}}{=} \frac{8\gamma}{n} \Big(\sum_{j=1}^d \sum_{i=1}^n \underbrace{\mathbb{E}[\varepsilon_i (\Sigma^{-1/2} X_i)_j^2]}_{=1} \Big)^{1/2} = 8\gamma \frac{\sqrt{d}}{\sqrt{n}}.
\end{aligned}$$

Choice of a, γ: Choose $a := 8(\frac{d}{n})^{1/2} \cdot t \Rightarrow \mathbb{P}(A^c) \leq \mathbb{P}(Z_\gamma > \gamma \cdot a) \overset{\text{Markov's ineq.}}{\leq} \frac{1}{t}$.
We have to meet the condition $\frac{2}{c}(\gamma a)^{1/2} \leq \frac{\gamma}{2} \Rightarrow$ Choose $\gamma := a \cdot (\frac{4}{c})^2$.
Condition from above: $\gamma \leq c^2 \iff (\frac{d}{n})^{1/2} \cdot t \leq \frac{c^4}{128}$.
Thus we obtain the rate

$$\gamma \cdot a = (\frac{4}{c})^2 \cdot a^2 = (\frac{32}{c})^2 t^2 \frac{d}{n}.$$

□

Logistic regression can also be performed with an additional penalization routine.

Definition 3.13 (Logistic regression with penalization for 2 classes)**.** Put

$$\hat{\beta}_\lambda \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \big\{ \tilde{R}_n(\beta) + \lambda \cdot \|\beta\|_1 \big\}.$$

Define $\hat{\delta}_{n,\lambda}^{LR}(x) = x^T \hat{\beta}_\lambda$. The *logistic regression classifier with penalization* is given by

$$\hat{f}_{n,\lambda}^{LR}(x) = \operatorname{sign}(\hat{\delta}_{n,\lambda}^{LR}(x)).$$

We are now able to provide the following theoretic result by combining techniques from Theorem 2.10 (Lasso) and Theorem 3.12 (Logistic regression without penalization). Put $S(\beta) := \{j \in \{1,...,d\} : \beta_j \neq 0\}$ and for $v \in \mathbb{R}^d$, $S \subset \{1,...,d\}$: $v_S := (v_j \mathbb{1}_{\{j \in S\}})_{j=1,...,d}$. Furthermore, set

$$C := \{v \in \mathbb{R}^d : \|v_{S(\beta^*)^c}\|_1 \leq 3\|v_{S(\beta^*)}\|_1\}, \qquad \Lambda_{min}(\Sigma) := \inf_{v \in C} \frac{v^T \Sigma v}{\|v_{S(\beta^*)}\|_2^2} \quad \text{(REC)}.$$

Theorem 3.14. Suppose that there exists $c \in (0,1)$ such that $c \leq \eta(x) \leq 1-c$ for all $x \in \mathcal{X}$. Suppose that $X \sim N(0,\Sigma)$. Put $s = \#S(\beta^*)$.
Let $t \geq 1$. Then for all

$$\lambda \geq 8(\frac{2\log(2d)}{n})^{1/2}\|\Sigma\|^{1/2} \cdot t$$

with the property that $(\frac{2\log(2d)}{n})^{1/2}\|\Sigma\|^{1/2}t \geq \frac{4\|\Sigma\|^{1/2}}{c^4}\frac{\lambda^2 s}{\Lambda_{min}(\Sigma)}$, it holds that

$$\mathbb{P}\Big(\tilde{R}(\hat{\beta}_{n,\lambda}) - \tilde{R}(\beta^*) \geq \frac{16}{c^2}\frac{\lambda^2 s}{\Lambda_{min}(\Sigma)}\Big) \leq \frac{1}{t}.$$

Proof. Abbreviate $S = S(\beta^*)$.
Deduction of the basic inequality: It holds that

$$\tilde{R}_n(\hat{\beta}) + \lambda\|\hat{\beta}\|_1 \leq \tilde{R}_n(\beta^*) + \lambda\|\beta^*\|_1,$$

thus

$$\begin{aligned}\tilde{R}(\hat{\beta}) - \tilde{R}(\beta^*) + \lambda\|\hat{\beta}\|_1 &\leq (\tilde{R}_n(\hat{\beta}) + \lambda\|\hat{\beta}\|_1) - (\tilde{R}_n(\beta^*) + \lambda\|\beta^*\|_1) \\ &\quad -\{\tilde{R}_n(\beta) - \tilde{R}(\beta) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))\} + \lambda\|\beta^*\|_1 \\ &\leq |\tilde{R}_n(\hat{\beta}) - \tilde{R}(\hat{\beta}) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))| + \lambda\|\beta^*\|_1. \qquad (14)\end{aligned}$$

Again we use the following 'trick': Let $T \in [0,1]$ and define $\tilde{\beta} = T\hat{\beta} + (1-T)\beta^*$. Then convexity of $\tilde{L}(y,\cdot)$ and $\|\cdot\|_1$ implies

$$\tilde{R}_n(\tilde{\beta}) + \lambda\|\tilde{\beta}\|_1 \leq T(\underbrace{\tilde{R}_n(\hat{\beta}) + \lambda\|\hat{\beta}\|_1}_{\leq \tilde{R}_n(\beta^*) + \lambda\|\beta^*\|_1}) + (1-T)(\tilde{R}_n(\beta^*) + \lambda\|\beta^*\|_1) \leq R_n(\beta^*) + \lambda\|\beta^*\|_1.$$

As before, we obtain

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) + \lambda\|\tilde{\beta}\|_1 \leq |\tilde{R}_n(\tilde{\beta}) - \tilde{R}(\tilde{\beta}) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))| + \lambda\|\beta^*\|_1 \qquad (15)$$

For $\gamma > 0$, define

$$T := \frac{\gamma}{\gamma + \|\hat{\beta} - \beta^*\|_1}.$$

It holds that

$$\tilde{\beta} - \beta^* = T(\hat{\beta} - \beta^*) \quad \Rightarrow \quad \|\tilde{\beta} - \beta^*\|_1 = T\|\hat{\beta} - \beta^*\|_1 \leq \gamma. \tag{16}$$

From (15) we get

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) + \lambda\|\tilde{\beta}\|_1 \leq \underbrace{\sup_{\beta:\|\beta-\beta^*\|_1\leq\gamma} \left|\tilde{R}_n(\beta) - \tilde{R}(\beta) - (\tilde{R}_n(\beta^*) - \tilde{R}(\beta^*))\right|}_{=:Z_\gamma} + \lambda\|\beta^*\|_1. \tag{17}$$

Definition of a 'nice' event: Put $A := \{Z_\gamma \leq \gamma \cdot a\}$.
On A it holds that

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) + \lambda\|\tilde{\beta}_{S^c}\|_1 \quad \leq \gamma \cdot a + \lambda(\|\beta^*_S\|_1 - \|\tilde{\beta}_S\|_1) \leq \quad \gamma \cdot a + \lambda\|(\beta^* - \tilde{\beta})_S\|_1. \tag{18}$$

Case 1: $2\lambda\|(\tilde{\beta} - \beta^*)_S\|_1 \geq \gamma \cdot a$. Then we have

$$\lambda\|(\tilde{\beta} - \beta^*)_{S^c}\|_1 \leq \underbrace{\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*)}_{0\leq} + \lambda\|\tilde{\beta}_{S^c}\|_1 \leq \gamma \cdot a + \lambda\|(\beta^* - \tilde{\beta})_S\|_1 \leq 3\lambda\|(\beta^* - \tilde{\beta})_S\|_1.$$

$\Rightarrow \tilde{\beta} - \beta^* \in C$
If $\|\Sigma\|^{1/2}\gamma \leq c^2$ (then $\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 \leq \|\Sigma\|^{1/2} \cdot \|\tilde{\beta} - \beta^*\|_2 \leq \|\Sigma\|^{1/2} \cdot \|\tilde{\beta} - \beta^*\|_1 \leq \|\Sigma\|^{1/2}\gamma \leq c^2$, that is, the margin property from Lemma 3.10 can be applied):

$$\begin{aligned}
& \tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) + \lambda \quad \underbrace{\|\tilde{\beta} - \beta^*\|_1}_{=\|\tilde{\beta}_{S^c}\|_1 + \|(\tilde{\beta}-\beta^*)_S\|_1} \\
\overset{(18)}{\leq} \quad & \gamma \cdot a + 2\lambda\|(\beta^* - \tilde{\beta})_S\|_1 \\
\overset{\text{REP}}{\leq} \quad & \gamma \cdot a + 2\lambda\Big(\frac{s}{\Lambda_{min}(\Sigma)}\Big)^{1/2}\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 \\
\overset{\text{Lemma 3.10}}{\leq} \quad & \gamma \cdot a + 2\lambda\Big(\frac{s}{\Lambda_{min}(\Sigma)}\Big)^{1/2} \cdot \frac{2}{c}\big(\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*)\big)^{1/2} \\
\overset{\substack{4xy\leq x^2+4y^2 \\ x=\frac{1}{\sqrt{2}}(\tilde{R}(\tilde{\beta})-\tilde{R}(\beta^*))}}{\leq} \quad & \gamma \cdot a + 4\frac{\lambda^2 s}{c^2\Lambda_{min}(\Sigma)} + \frac{1}{2}\big(\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*)\big).
\end{aligned}$$

Rearranging terms yields

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) + 2\lambda\|\tilde{\beta} - \beta^*\|_1 \leq 2\gamma \cdot a + 8\frac{\lambda^2 s}{c^2\Lambda_{min}(\Sigma)}.$$

Case 2: $2\lambda\|(\tilde{\beta} - \beta^*)_S\|_1 \leq \gamma \cdot a$. Dann ist mit (18):

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) + \lambda \cdot \|\tilde{\beta} - \beta^*\|_1 \leq \gamma \cdot a + 2\lambda\|(\tilde{\beta} - \beta^*)_S\|_1 \leq 2\gamma \cdot a.$$

End of case distinction.

We conclude that if we choose λ such that

$$4\frac{\lambda^2 s}{c^2\Lambda_{min}(\Sigma)} \leq \gamma \cdot a,$$

then on A it holds that

$$\tilde{R}(\tilde{\beta}) - \tilde{R}(\beta^*) + \lambda \cdot \|\tilde{\beta} - \beta^*\|_1 \leq 4\gamma \cdot a. \tag{19}$$

If $4a \leq \frac{\lambda}{2}$, then it holds that

$$\|\tilde{\beta} - \beta^*\|_1 \leq \frac{\gamma}{2}.$$

From (16) we get

$$\frac{\gamma\|\hat{\beta} - \beta^*\|_1}{\gamma + \|\hat{\beta} - \beta^*\|_1} = \|\tilde{\beta} - \beta^*\|_1 \leq \frac{\gamma}{2} \quad \overset{\text{rearranging}}{\Rightarrow} \quad \|\hat{\beta} - \beta^*\|_1 \leq \gamma.$$

If we repeat the whole proof for $\hat{\beta}$ (now starting from (10) instead of (15)), then we obtain that on A it holds that (cf. (19)):

$$\tilde{R}(\hat{\beta}) - \tilde{R}(\beta^*) \leq 4\gamma \cdot a.$$

Upper bound for the probability of A^c: It holds that

$$|\tilde{L}(y,s) - \tilde{L}(y,s')| \leq |\frac{1}{2}(y+1)| \cdot |s - s'| + 1 \cdot |s - s'| \leq 2|s - s'|,$$

thus Lemma 3.11 with $\ell = 2$ implies

$$\begin{aligned}
\mathbb{E}Z_\gamma &\leq 8 \cdot \mathbb{E} \sup_{\|\beta-\beta^*\|_1 \leq \gamma} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i^T(\beta - \beta^*)\Big| \\
&\leq 8\mathbb{E}\big[\sup_{\|\beta-\beta^*\|_1 \leq \gamma} \big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\big\|_\infty \big\|\beta - \beta^*\big\|_1\big] \\
&\leq \frac{8\gamma}{n}\mathbb{E}\Big\|\sum_{i=1}^n \varepsilon_i X_i\Big\|_\infty.
\end{aligned}$$

Since $X_1, ..., X_n \sim N(0, \Sigma)$, it holds that (this is left as an exercise):

$$\mathbb{E}\Big[\max_{j=1,...,d}\Big|\sum_{i=1}^n \varepsilon_i X_{ij}\Big|\Big] \leq \sqrt{n}\sqrt{2\log(2d)} \cdot \max_{j=1,...,d} \Sigma_{jj}^{1/2}$$

We conclude that

$$\mathbb{E}Z_\gamma \leq \gamma \cdot 8\Big(\frac{2\log(2d)}{n}\Big)^{1/2}\|\Sigma\|^{1/2}.$$

Choice of a, γ, λ: Choose $a := 8(\frac{2\log(2d)}{n})^{1/2}\|\Sigma\|^{1/2}\cdot t \Rightarrow \mathbb{P}(A^c) \leq \mathbb{P}(Z_\gamma > \gamma\cdot a) \overset{\text{Markov's ineq.}}{\leq} \frac{1}{t}$.
It has to hold that $4a \leq \frac{\lambda}{2}$, that is, $\lambda \geq 8a$ (this leads to one condition in the theorem).
It has to hold that $4\frac{\lambda^2 s}{c^2\Lambda_{min}(\Sigma)} \leq \gamma a$, that is, we should choose $\gamma := 4\frac{\lambda^2 s}{ac^2\Lambda_{min}(\Sigma)}$.
Condition from above: $\|\Sigma\|^{1/2}\gamma \leq c^2 \iff 4\frac{\lambda^2 s}{ac^2\Lambda_{min}(\Sigma)} \leq c^2\|\Sigma\|^{-1/2}$ (this leads to one condition in the theorem).
The convergence rate then reads

$$4\gamma \cdot a = \frac{16}{c^2}\frac{\lambda^2 s}{\Lambda_{min}(\Sigma)}.$$

□

Remarks:

- The condition (*) is a bound for the dimensions s, d. For instance, if we choose the smallest possible λ,

  $$\lambda = 8(\frac{2\log(2d)}{n})^{1/2}\|\Sigma\|^{1/2}t,$$

  then (*) reads

  $$(\frac{2\log(2d)}{n})^{1/2}s \cdot t \leq \frac{c^4}{256}\frac{\Lambda_{min}(\Sigma)}{\|\Sigma\|}.$$

  and we obtain the convergence rate

  $$\frac{16}{c^2}\frac{\lambda^2 s}{\Lambda_{min}(\Sigma)} = \frac{\text{Zahl}}{c^2}\cdot\frac{\log(d)s}{n}\cdot\frac{\|\Sigma\|}{\Lambda_{min}(\Sigma)}\cdot t^2.$$

- Note that similar as in the LASSO case, the original dimension d of the space $\mathcal{X}$ only shows up in the convergence rate via the much smaller $\log(d)$.

- In both theorems, Theorem 3.12 and Theorem 3.14, *much* better upper bounds are available with respect to t. We only used the simple Markov inequality $\mathbb{P}(Z_\gamma \geq \gamma a) \leq \frac{\mathbb{E}Z_\gamma}{\gamma a}$ and obtained a result of the form

  $$\mathbb{P}(\tilde{R}(\hat{\beta}) - \tilde{R}(\beta^*) \geq \text{Rate}\cdot t^2) \leq \frac{1}{t}.$$

  Using more sophisticated concentration inequalities, one can obtain $\leq e^{-t}$. This is part of the exercises.

3.3 Calibration condition and risk transfer formula

Up to now, we have investigates classifiers for $K = 2$ classes which had the form

$$\hat{f}_n(x) = \operatorname{sign}(\hat{\delta}_n(x)), \tag{20}$$

Here, $\hat{\delta}_n$ was obtained via a minimization problem

$$\hat{\delta}_n :\in \operatorname*{arg\,min}_{\delta \in \Delta} \tilde{R}_n(\delta), \qquad \tilde{R}_n(\delta) := \frac{1}{n} \sum_{i=1}^{n} \tilde{L}(Y_i, \delta(X_i))$$

(or with additional penalization). Mathematically speaking, $\hat{\delta}_n(x)$ is a discriminant function for class 1 of the classifier $\hat{f}_n$ (and the discriminant function for class -1 is the constant zero function). Computing $\hat{\delta}_n$ therefore means to estimating an optimal discriminant function instead of the Bayes rule f^*.
We obtained upper bounds for the excess Bayes risk of $\hat{\delta}_n$ with respect to $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y, \delta(X))$, that is, upper bounds for

$$\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*), \qquad \delta^* :\in \operatorname*{arg\,min}_{\delta : \mathcal{X} \to \mathbb{R}} \tilde{R}(\delta).$$

Question: Can we derive upper bounds for the excess Bayes risk of $\hat{f}_n$ with respect to the 0-1 loss $L(y, s) = \mathbb{1}_{\{y \neq s\}}$ from the above upper bounds? That is, can we derive upper bounds for

$$R(\hat{f}_n) - R(f^*), \qquad f^* :\in \operatorname*{arg\,min}_{f : \mathcal{X} \to \{-1, +1\}} R(f) \quad ?$$

We are interested to transfer the results to the 0-1 loss of the classifier $\hat{f}_n$ since the 0-1 loss is the 'most natural' loss function for classification: From it, one can directly infer the expected number of false decisions.

Remark 3.15. The strategy to estimate optimal discriminant functions instead of the Bayes rule f^* and defining (20) afterwards is a common approach in several machine learning algorithms for classification. Many of these algorithms were originally motivated by estimation of discriminant functions. However, one could also think of defining a classifier directly via minimization over the 0-1 loss $L(y, s) = \mathbb{1}_{\{y \neq s\}}$,

$$\hat{f}_n := \operatorname*{arg\,min}_{f \in \mathcal{F}} \hat{R}_n(f), \qquad \hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) \tag{21}$$

(or with additional penalization). However, $L(y, s)$ is *non-convex* and *non-differentiable* in s, therefore the optimization problem is hard to solve in practice.

Due to (20) and $\hat{\delta}_n \to \delta^*$, we expect that

$$\hat{f}_n = \operatorname{sign}(\hat{\delta}_n) \to \operatorname{sign}(\delta^*).$$

On the other hand, we want that $\hat{f}_n$ converges (as a function) towards f^*. The approach (20) therefore can only lead to meaningful algorithms if the following condition is satisfied.

Definition 3.16 (Calibration condition)**.** A loss function $\tilde{L} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ satisfies the *calibration condition* if

$$f^* = \operatorname{sign}(\delta^*).$$

To transfer an upper bound of the excess Bayes risk of $\hat{\delta}_n$ to $\hat{f}_n$, we need the following abstract condition.

Definition 3.17 (Risk transfer formula)**.** A loss function $\tilde{L} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ produces a *risk transfer formula* if there exists a non-decreasing function $G : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for all measurable $\delta : \mathcal{X} \to \mathbb{R}$ it holds that

$$R(\operatorname{sign}(\delta)) - R(\operatorname{sign}(\delta^*)) \leq G\big(\tilde{R}(\delta) - \tilde{R}(\delta^*)\big).$$

Many loss functions $\tilde{L}$ can be written in the special form

$$\tilde{L}(y, s) = \phi(-ys), \tag{22}$$

where $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is non-decreasing. This can be motivated as follows: Starting from the optimization problem in Remark 3.15, it holds for $y, s \in \{+1, -1\}$ that

$$L(y, s) = \mathbb{1}_{\{y \neq s\}} = \mathbb{1}_{\{ys \leq 0\}} = \mathbb{1}_{\{-ys \geq 0\}} \approx \phi(-ys) =: \tilde{L}(y, s).$$

Note that starting from the term $\mathbb{1}_{\{-ys \geq 0\}}$ in the above chain of equations, also $s \in \mathbb{R}$ (instead of $s \in \{-1, +1\}$ would produce a meaningful 'loss' in the sense that the loss is 0 if s and y have the same sign. Thus, one can interpret the transition from estimating $f^* = \operatorname{sign}(\delta^*)$ to estimating δ^* as a *relaxation* of the original 'exact' optimization problem. Instead of searching for a decision rule $f(x) \in \{-1, +1\}$, we now allow for all real numbers $\delta(x) \in \mathbb{R}$ and use a different loss function $\tilde{L}(y, s)$.
We now convince ourselves that logistic regression indeed can be implanted into this framework.

Lemma 3.18. Let $\phi_{log}(x) = \log(1 + e^x)$. Then for $y \in \{-1, +1\}$, $s \in \mathbb{R}$ it holds that

$$\tilde{L}_{log}(y, s) := -\frac{1}{2}(y + 1)s + \log(1 + e^s) = \phi_{log}(-ys)$$

That is, the loss function $\tilde{L}_{log}$ of logistic regression has the form (22).

Proof. It holds that $\tilde{L}_{log}(1, s) = -s + \log(1 + e^s) = \log(1 + e^{-s}) = \phi_{log}(-s)$, $\tilde{L}_{log}(-1, s) = \log(1 + e^s) = \phi_{log}(s)$. □

We now show some general results for loss functions of the form (22). Recall that $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

Theorem 3.19. Let $\tilde{L}(y, s) = \phi(-ys)$ with measurable $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$. For $\eta \in [0, 1]$ let

$$\Phi_\eta(z) := \phi(-z)\eta + \phi(z)(1 - \eta).$$

Then $\delta^*(x) :\in \arg\min_{z \in \mathbb{R}} \Phi_{\eta(x)}(z)$ is a Bayes rule for the risk $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y, \delta(X))$.

Proof. It holds that $R(\delta) = \mathbb{E}[\mathbb{E}[\phi(-Y\delta(X))|X]]$, and

$$\begin{aligned}
\mathbb{E}[\phi(-Y\delta(X))|X = x] &= \phi(-\delta(x))\mathbb{P}(Y = 1|X = x) + \phi(\delta(x))\mathbb{P}(Y = -1|X = x) \\
&= \phi(-\delta(x))\eta(x) + \phi(\delta(x))(1 - \eta(x)) \\
&= \Phi_{\eta(x)}(\delta(x)) \\
&\geq \Phi_{\eta(x)}(\delta^*(x)) = \ldots = \mathbb{E}[\phi(-Y\delta^*(X))|X = x].
\end{aligned}$$

The tower property of conditional expectations yields the result. □

Based on this result, we can verify the calibration condition for logistic regression.

Example 3.20 (Logistic regression)**.** Here, we have $\phi_{log}(x) = \log(1 + e^x)$, and

$$\Phi_\eta(z) = \log(1 + e^{-z})\eta + \log(1 + e^z)(1 - \eta), \quad 0 = \Phi'_\eta(z) = -\frac{e^{-z}}{1 + e^{-z}}\eta + \frac{e^z}{1 + e^z}(1 - \eta)$$

$\Rightarrow -\eta + e^z(1 - \eta) = 0 \Rightarrow z = \log(\frac{\eta}{1-\eta}) \Rightarrow \delta^*(x) = \log(\frac{\eta(x)}{1-\eta(x)})$.
Note that we already have seen a similar calculation in the proof of Theorem 3.10, but

with an intermediate step using the linear model assumption. The above result was obtained without any model assumption.
It holds that

$$\delta^*(x) > 0 \quad \Longleftrightarrow \quad \eta(x) > \frac{1}{2}$$

Thus, $f^*(x) = \arg\max_{k\in\{-1,+1\}} \mathbb{P}(Y = k|X = x) = \operatorname{sign}(\delta^*(x))$. This shows that $\tilde{L}_{log}$ satisfies the calibration condition.

We now prove a general result how to obtain risk transfer formulas.

Theorem 3.21. Let $\tilde{L}(y, s) = \phi(-ys)$ with $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be convex and suppose that $\phi(0) = 1$. Suppose that there exist $C_H \geq 0$, $s \geq 1$, such that for all $\eta \in [0, 1]$ it holds that

$$\big|\eta - \frac{1}{2}\big|^s \leq C_H^s(1 - H(\eta)), \qquad H(\eta) := \min_{z\in\mathbb{R}} \Phi_\eta(z), \quad g(\eta) :\in \arg\min_{z\in\mathbb{R}} \Phi_\eta(z).$$

Suppose furthermore that g satisfies $g(\eta) > 0$ for $\eta > \frac{1}{2}$. Then $\tilde{L}$ satisfies the risk transfer formula with $G(r) = 2C_H r^{1/s}$.

Proof. Let $\delta : \mathcal{X} \to \mathbb{R}$ be measurable, $f = \operatorname{sign}(\delta) : \mathcal{X} \to \{-1, +1\}$. Then it holds that

$$\begin{aligned} R(f) &= \mathbb{E}[\mathbb{P}(Y \neq f(X)|X)] = \mathbb{E}[\eta(X)\mathbb{1}_{\{f(X)=-1\}} + (1 - \eta(X))\mathbb{1}_{\{f(X)=1\}}] \\ &= \mathbb{E}[\eta(X)\mathbb{1}_{\{\delta(X)<0\}} + (1 - \eta(X))\mathbb{1}_{\{\delta(X)>0\}}]. \end{aligned}$$

With $f^*(x) = \operatorname{sign}(2\eta(x) - 1)$, we obtain that

$$\begin{aligned} R(f) - R(f^*) &= \mathbb{E}[\eta(X)\mathbb{1}_{\{\delta(X)<0\}} + (1 - \eta(X))\mathbb{1}_{\{\delta(X)>0\}}] \\ &\quad -\mathbb{E}[\eta(X)\mathbb{1}_{\{\eta(X)<\frac{1}{2}\}} + (1 - \eta(X))\mathbb{1}_{\{\eta(X)>\frac{1}{2}\}}] \\ &\overset{\text{case dist.}}{=} \mathbb{E}\big[(2\eta(X) - 1)\mathbb{1}_{\{\eta(X)>\frac{1}{2},\delta(X)<0\}} + (1 - 2\eta(X))\mathbb{1}_{\{\eta(X)<\frac{1}{2},\delta(X)>0\}}\big] \\ &\leq \mathbb{E}\big[|2\eta(X) - 1|\mathbb{1}_{\{(2\eta(X)-1)\delta(X)<0\}}\big] \\ &\overset{\text{Hölder's ineq.}}{\leq} 2\big(\mathbb{E}\big[\big|\eta(X) - \frac{1}{2}\big|^s\mathbb{1}_{\{(2\eta(X)-1)\delta(X)<0\}}\big]\big)^{1/s} \\ &\overset{\text{Vorauss.}}{\leq} 2C_H\mathbb{E}\big[(1 - H(\eta(X)))\mathbb{1}_{\{(2\eta(X)-1)\delta(X)<0\}}\big]^{1/s}. \end{aligned}$$

We now show the following intermediate result: (*) If $(2\eta - 1)p < 0$, then $1 \leq \Phi_\eta(p)$. If this result is shown, we can conclude the result as follows:

$$R(f) - R(f^*) \leq 2C_H\mathbb{E}\big[\Phi_{\eta(X)}(\delta(X)) - \Phi_{\eta(X)}(\delta^*(X))\big]^{1/s} = 2C_H\big(\tilde{R}(\delta) - \tilde{R}(\delta^*)\big)^{1/s}$$

To prove (*), we use a case distinction. Note that convexity of ϕ implies that $z \mapsto \Phi_\eta(z)$ is convex.

- If $\eta > \frac{1}{2}$, then $g(\eta) > \frac{1}{2}$ and $p < 0$.
  $\Rightarrow 0 \in [p, g(\eta)]$. Convexity of $\Phi_\eta(\cdot) \Rightarrow 1 = \Phi_\eta(0) \leq \max\{\Phi_\eta(p), \Phi_\eta(g(\eta))\} = \Phi_\eta(p)$.
- If $\eta < \frac{1}{2}$, then $p > 0$. Based on the symmetry of $z \mapsto \Phi_\eta(\eta)$ we conclude that $g(\eta) = -g(1-\eta) < 0$. A similar argument as above yields $1 = \Phi_\eta(0) \leq \Phi_\eta(p)$.
- If $\eta = \frac{1}{2}$, then $g(\eta) = -g(1-\eta)$ implies $g(\eta) = 0$. Thus $1 = \Phi_\eta(0) = \Phi_\eta(g(\eta)) \leq \Phi_\eta(p)$ for all values of p.

□

Example 3.22 (Logistic regression)**.** Here, we have $\phi_{log}(x) = \log(1 + e^x)$. Here, we need a slight modification to make ϕ_{log} fit into the setting of Theorem 3.21: Define $\tilde{\phi}_{log}(x) = \frac{\log(1+e^x)}{\log(2)}$. Then we have $\tilde{\phi}_{log}(0) = 1$ (note that his does not change the optimization problem since the objective is just multiplied with a positive factor).

$$\Phi_\eta(z) = \frac{1}{\log(2)}(\log(1 + e^{-z})\eta + \log(1 + e^z)(1 - \eta)).$$

We have already seen that $g(\eta) = \log(\frac{\eta}{1-\eta})$,

$$H(\eta) = \Phi_\eta(g(\eta)) = -\frac{1}{\log(2)}\Big(\log(\eta)\eta + \log(1-\eta)(1-\eta)\Big).$$

The function $p(\eta) = 1 - H(\eta)$ satisfies

$$p'(\eta) = \frac{1}{\log(2)}(\log(\eta) - \log(1-\eta)), \qquad p''(\eta) = \frac{1}{\log(2)}(\frac{1}{\eta} + \frac{1}{1-\eta}) = \frac{1}{\log(2)}\frac{1}{\eta(1-\eta)} \geq \frac{1}{4\log(2)}.$$

A Taylor expansion of $p(\cdot)$ around $\eta = \frac{1}{2}$ yields with some intermediate value $\xi \in (\frac{1}{2}, \eta)$ that

$$p(\eta) = \frac{1}{2}(\eta - \frac{1}{2})^2 \cdot p''(\xi) \geq \frac{1}{8\log(2)}(\eta - \frac{1}{2})^2.$$

$\Rightarrow$ The condition of Theorem 3.21 is satisfied with $s = 2$, $C_H = 2\sqrt{2\log(2)}$.
$\Rightarrow$ The risk transfer formula of logistic regression reads

$$R(\hat{f}_n^{LR}) - R(f^*) \leq 4\sqrt{2\log(2)}\big(\tilde{R}(\hat{\delta}_n^{LR}) - \tilde{R}(\delta^*)\big)^{1/2}.$$

This result is a little bit unsatisfying: The (upper bound of the) convergence rate of $\tilde{R}(\hat{\delta}_n^{LR}) - \tilde{R}(\delta^*)$ is only transferred with an exponent $\frac{1}{2}$ to the rate of $R(\hat{f}_n^{LR}) - R(f^*)$. Instead of an upper bound $\frac{d}{n}$ we therefore only obtain an upper bound for $R(\hat{f}_n^{LR}) - R(f^*)$ of the form $(\frac{d}{n})^{1/2}$. Due to the exact calculation in the linear regression model, we are somehow 'sure' that the rate of $\tilde{R}(\hat{\delta}_n^{LR}) - \tilde{R}(\delta^*)$ cannot be better than $\frac{d}{n}$. However for $R(\hat{f}_n^{LR}) - R(f^*)$ it is not clear if the rate $(\frac{d}{n})^{1/2}$ could be improved. In [17] it was shown that even for the 0-1 loss one expects convergence rates of order n^{-1} under specific assumptions.
We now present a general result how specific assumptions on $\eta(x)$ can improve the risk transfer formula from Theorem 3.21.

Definition 3.23 (Noise condition)**.** There are $q \geq 0$, $C > 0$ such that

$$\forall t > 0 : \qquad \mathbb{P}\Big(\big|\eta(X) - \frac{1}{2}\big| \leq t\Big) \leq Ct^q.$$

The extreme case $q = \infty$ is allowed and should be interpreted as follows: There exists some $c > 0$ such that $|\eta(x) - \frac{1}{2}| \geq c$ for all $x \in \mathcal{X}$.

The variable q in the noise condition can be interpreted as a measure how 'often' $\eta(x)$ reaches $\frac{1}{2}$, that is, how often one has to make 'hard' decisions. Under this condition, the following improved risk transfer formula holds.

Theorem 3.24. Suppose that the conditions of Theorem 3.21 are met. Additionally, suppose that the noise condition holds. Then the risk transfer formula is satisfied with $G(r) = 4C_H^{\frac{s(q+1)}{q+s}} C^{-\frac{1}{q+s}} \cdot r^{\frac{q+1}{q+s}}$.
In the special case $q = \infty$ we have $G(r) = 2\frac{C_H^s}{c^{s-1}} r$.

Proof. Starting from the proof of Theorem 3.21, it holds for $f = \operatorname{sign}(\delta)$, $\gamma \geq 0$ and

$A := \{(2\eta(X)-1)\delta(X) < 0\}$ that

$$\begin{aligned}
R(f) - R(f^*) &\leq 2\mathbb{E}\big[\big|\eta(X) - \frac{1}{2}\big|\mathbb{1}_A\big] \\
&\leq 2\Big(\mathbb{E}\big[\big|\eta(X) - \frac{1}{2}\big|\mathbb{1}_A \mathbb{1}_{\{|\eta(X)-\frac{1}{2}|<\gamma\}}\big] + \mathbb{E}\big[\big|\eta(X) - \frac{1}{2}\big|\mathbb{1}_A \underbrace{\mathbb{1}_{\{|\eta(X)-\frac{1}{2}|\geq\gamma\}}}_{\leq \mathbb{1}_{\{\dots\}}\frac{|\eta(X)-\frac{1}{2}|^{s-1}}{\gamma^{s-1}}}\big]\Big) \\
&\leq 2\Big(\gamma \cdot \mathbb{P}(|\eta(X) - \frac{1}{2}| < \gamma) + \frac{1}{\gamma^{s-1}}\mathbb{E}\big[\big|\eta(X) - \frac{1}{2}\big|^s \mathbb{1}_A\big]\Big) \\
&\overset{\text{as in Thm. 3.21}}{\underset{\text{noise cond.}}{\leq}} 2\Big(C\gamma^{q+1} + \frac{1}{\gamma^{s-1}} \underbrace{C_H^s(\tilde{R}(\delta) - \tilde{R}(\delta^*))}_{=:B}\Big). \qquad (23)
\end{aligned}$$

Now we choose $\gamma > 0$ in such a way that both terms on the right hand side of (23) are nearly equal in size. This can be obtained by choosing

$$\gamma = \Big(\frac{B}{C}\Big)^{\frac{1}{q+s}}.$$

Putting this result into (23), we obtain

$$R(f) - R(f^*) \leq 4C^{-\frac{1}{q+s}} B^{\frac{q+1}{q+s}}.$$

In the special case $q = \infty$, the first summand in (23) does not appear for the choice $\gamma = c$ which then yields the result. □

Example 3.25 (Logistic regression)**.** To obtain the margin property in Theorem 3.10, we had to assume that $c \leq \eta(x) \leq 1 - c$ for all $x \in \mathcal{X}$. Now we additionally suppose that $|\eta(x) - \frac{1}{2}| \geq c$ for all $x \in \mathcal{X}$, that is we assume that the noise condition holds with $q = \infty$. We then obtain the following risk transfer formula for logistic regression:

$$R(\hat{f}_n^{LR}) - R(f^*) \leq \frac{16\log(2)}{c} \cdot \big(\tilde{R}(\hat{\delta}_n^{LR}) - \tilde{R}(\delta^*)\big).$$

In the situation of Theorem 3.12, we would obtain

$$\mathbb{P}\Big(R(\hat{f}_n^{LR}) - R(f^*) \geq (\frac{32}{c})^3 t^2 \frac{d}{n}\Big) \leq \frac{1}{t}.$$

This finishes the discussion of logistic regression.

3.4 Generalization to nonlinear models

Even though linear models seem of quite limited use in practice, they can be used to model *nonlinear* relations between Y and X.

'Trick': The algorithms are not applied to (X, Y) but to $(\tilde{X}, Y)$, where

$$\tilde{X} = h(X), \qquad h : \mathcal{X} \to \tilde{\mathcal{X}} \subset \mathbb{R}^{\tilde{d}}.$$

Often one then has $\tilde{d} \gg d$. Additionally, the model assumptions have to be satisfied for $(\tilde{X}, Y)$ instead of (X, Y). Accordingly, also the conditions of the theorems have to be satisfied for $(\tilde{X}, Y)$.

Example 3.26 (Linear regression)**.** The model

$$Y = \sum_{j=1}^{d} a_j^* X_j + \sum_{j=1}^{d} b_j^* X_j^2 + \varepsilon$$

with $a^*, b^* \in \mathbb{R}^d$ corresponds to the model assumption of linear regression with

$$Y = \tilde{X}^T \beta^* + \varepsilon, \qquad \beta^* = (a^*, b^*) \in \mathbb{R}^{2d}, \qquad \tilde{X} = h(X),$$

$h(x) = (x_1, ..., x_d, x_1^2, ..., x_d^2)$.

Example 3.27 (Logistic regression)**.** The model

$$\log(\frac{\eta(x)}{1 - \eta(x)}) = \sum_{j=1}^{d} a_j^* x_j + \sum_{j=1}^{d} b_j^* x_j^2$$

with $a^*, b^* \in \mathbb{R}^d$ corresponds to the model assumption of logistic regression with

$$\log(\frac{\tilde{\eta}(x)}{1 - \tilde{\eta}(x)}) = \tilde{x}^T \beta^*, \qquad \beta^* = (a^*, b^*) \in \mathbb{R}^{2d}, \qquad \tilde{x} = h(x),$$

$h(x) = (x_1, ..., x_d, x_1^2, ..., x_d^2)$ and $\tilde{\eta}(x) = \mathbb{P}(Y = 1 | h(X) = h(x)) = \eta(x)$.

In the classification example, the linear (optimal) decision boundaries then change to nonlinear decision boundaries as follows:

$$\Omega_1 = \{x \in \mathcal{X} : \tilde{x}^T \beta^* \geq 0\} = \{x \in \mathcal{X} : h(x)^T \beta^* \geq 0\}.$$

In Example 3.27, this leads to

$$\Omega_1 = \{x \in \mathcal{X} : \sum_{j=1}^{d} a_j^* x_j + \sum_{j=1}^{d} b_j^* x_j^2 \geq 0\},$$

that is, the decision boundaries are of quadratic form (ellipses).

Caution: There are several issues coming with these techniques:

- It is not clear which functions $h : \mathcal{X} \to \tilde{\mathcal{X}} \subset \mathbb{R}^{\tilde{d}}$ should be chosen if one wants to model a more complicated relationship in the data. A more complicated and diverse h often leads to $\tilde{d} \gg d$. A standard choice to overcome this selection problem is to choose the components of h as bases from Hilbert spaces $\mathcal{H} \subset \{f : \mathcal{X} \to \mathbb{R}\}$. This is discussed in more detail in the SVM chapter.
- Since one often has $\tilde{d} \gg d$, it is necessary to use algorithms with penalization of β.
- The covariance matrix in the nonlinear formulation reads $\tilde{\Sigma} = \mathbb{E}[\tilde{X}\tilde{X}^T] = \mathbb{E}[h(X)h(X)^T]$. In many results, the smallest eigenvalues of $\tilde{\Sigma}$ play an important role. This may lead to problems: If the functions h are not well-chosen, the eigenvalues of $\tilde{\Sigma}$ may be very small even if X itself had nearly independent components. As an example, consider $X \sim U[0,1]$ (1-dimensional uniform distribution on $[0,1]$) and

  $$h(x) = (x, x^2, ..., x^{\tilde{d}}).$$

  Then it holds that

  $$\tilde{\Sigma}_{jk} = \mathbb{E}[X^j X^k] = \mathbb{E}[X^{j+k}] = \frac{1}{j+k+1}, \quad j,k = 1, ..., \tilde{d}.$$

  Already for $\tilde{d} = 3$ it holds that $\lambda_{min}(\tilde{\Sigma}) \approx 0.0002$ since X, X^2, X^3 are strongly correlated.

Therefore the choice of h has to be made with care. Often the penalization term has to be changed and cannot be chosen simply as the 1- or 2-norm of β since then the results depend too heavily on the properties of $\tilde{\Sigma}$.

3.5 Exercises

Task 7 (Discussion: Application of the proof technique of Theorem 3.12)**.** In this task we aim to apply the proof technique of Theorem 3.12 on the LS estimator in the linear model with *deterministic* design, that is $X_1, ..., X_n$ are considered as deterministic and $\hat{\Sigma} = \frac{1}{n}\mathbb{X}^T\mathbb{X} = \Sigma$. Moreover, suppose that $\varepsilon_i \sim N(0, \sigma^2)$.
Recall that $\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \hat{R}_n(\beta)$ with $\hat{R}_n(\beta) = \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2$, $R(\beta) = \|\Sigma^{1/2}(\beta - \beta^*)\|_2^2 + \sigma^2$.

1. Show that
$$R(\hat{\beta}) - R(\beta^*) \leq \big|(\hat{R}_n(\hat{\beta}) - R(\hat{\beta})) - (\hat{R}_n(\beta^*) - R(\beta^*))\big|.$$

2. Let $\gamma > 0$, $T := \frac{\gamma}{\gamma + \|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2}$ and $\tilde{\beta} := T\hat{\beta} + (1-T)\beta^*$. Show that $\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 \leq \gamma$ and
$$R(\tilde{\beta}) - R(\beta^*) \leq \sup_{\beta:\|\Sigma^{1/2}(\beta-\beta^*)\|_2 \leq \gamma} \big|(\hat{R}_n(\beta) - R(\beta)) - (\hat{R}_n(\beta^*) - R(\beta^*))\big| =: Z_\gamma.$$

3. Let $a > 0$. Show the following statement: If $(a\gamma)^{1/2} \leq \frac{\gamma}{2}$, then we have on the event $A = \{Z_\gamma \leq a \cdot \gamma\}$ that
$$\|\Sigma^{1/2}(\hat{\beta} - \beta)\|_2 \leq \gamma.$$
Conclude that $R(\hat{\beta}) - R(\beta^*) \leq a \cdot \gamma$.

4. Conclude that
$$\mathbb{P}(R(\hat{\beta}) - R(\beta^*) > a \cdot \gamma) \leq \mathbb{P}(A^c).$$

Now, we investigate the choice of a, γ.

(e) Show that
$$(\hat{R}_n(\beta) - R(\beta)) - (\hat{R}_n(\beta^*) - R(\beta^*)) = -\frac{2}{n}\mathbb{e}^T\mathbb{X}(\beta - \beta^*)$$
and $|Z_\gamma| \leq \frac{2\gamma}{n}\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2$.
Hint: Add $\Sigma^{-1/2}\Sigma$ in the above expression and use that $|v^T w| \leq \|v\|_2\|w\|_2$.

(f) Show that $\mathbb{E}|Z_\gamma| \leq 2\gamma\sigma(\frac{d}{n})^{1/2}$.

(g) Choose a, γ appropriately to obtain the following statement:
$$\mathbb{P}(R(\hat{\beta}) - R(\beta^*) \geq 16\sigma^2\frac{d}{n} \cdot t^2) \leq \frac{1}{t}.$$

Now, we aim to find a better upper bound for $\mathbb{P}(Z_\gamma > a\gamma)$.

(h) Show that $\frac{1}{\sigma\sqrt{n}}\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2 \stackrel{d}{=} \|W\|_2$ with $W \sim N(0, I_{d\times d})$.

(i) Show that
$$\mathbb{P}(\|W\|_2^2 - 2\mathbb{E}\|W\|_2^2 \geq t) \leq e^{-t/4}.$$
Hint: Compute $\mathbb{E}\|W\|_2^2$, shift it to the right hand side of the above inequality and then apply Markov's inequality with $g(x) = e^{-x/4}$. For $S \sim N(0,1)$ it holds that $\mathbb{E}[\exp(\frac{S^2}{4})] = \sqrt{2}$, and one has $\sqrt{2}e^{-1/2} \leq 1$.

(j) Using the upper bound for Z_γ from (e), show that with $a = \frac{\sigma}{\sqrt{n}}\sqrt{2d+t}$ it holds that

$$\mathbb{P}(A^c) = \mathbb{P}(Z_\gamma > a\gamma) \leq e^{-t}.$$

(k) Conclude as in (g) that

$$\mathbb{P}\Big(R(\hat{\beta}) - R(\beta^*) \geq 4\frac{\sigma^2}{n} \cdot (2d+t)\Big) \leq e^{-t}.$$

Task 8 (Computation of the expected value in Theorem 3.14)**.** Let ε_i, $i = 1, ..., n$ be i.i.d. Rademacher distributed ($\mathbb{P}(\varepsilon_1 = 1) = \mathbb{P}(\varepsilon_1 = -1) = \frac{1}{2}$) and $X_i \sim N(0, \Sigma)$, $i = 1, ..., n$ be i.i.d (independent of ε_i). We now show that

$$\mathbb{E}\Big[\max_{j=1,...,d}\Big|\sum_{i=1}^n \varepsilon_i X_{ij}\Big|\Big] \leq 2\sqrt{n}\sqrt{\log(\sqrt{2}d)} \cdot \max_{j=1,...,d}\Sigma_{jj}^{1/2}.$$

1. Define $W_j := \frac{1}{\sqrt{n}\Sigma_{jj}^{1/2}}\sum_{i=1}^n \varepsilon_i X_{ij}$. Show that conditionally on ε_i, $i = 1, ..., n$, it holds that
$$W_j \sim N(0,1),$$
and $\mathbb{E}\Big[\max_{j=1,...,d}\Big|\sum_{i=1}^n \varepsilon_i X_{ij}\Big|\Big] \leq \mathbb{E}[\max_{j=1,...,d}|W_j|] \cdot \sqrt{n}\max_{j=1,...,d}\Sigma_{jj}^{1/2}$.

2. The function $\psi(x) = \exp(x^2/4)$ is convex. Show with Jensen's inequality that
$$\psi\big(\mathbb{E}[\max_{j=1,...,d}|W_j|]\big) \leq \sum_{j=1}^d \mathbb{E}\psi(|W_j|).$$

3. Conclude from (b) that $\mathbb{E}[\max_{j=1,...,d}|W_j|] \leq 2\sqrt{\log(\sqrt{2}d)}$.
Hint: It holds that $\mathbb{E}[\exp(W_j^2/4)] = \sqrt{2}$.

Task 9 (Discussion: The model assumption of logistic regression)**.** Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{+1, -1\}$ (classification problems with 2 classes). Let $(X, Y) \ll \lambda \times \mu_Z$, where λ denotes the Lebesgue measure on $\mathcal{X}$ and μ_Z the counting measure on $\mathcal{Y}$. Let g_k denote the conditional density of X given $Y = k$, and $\pi_1 := \mathbb{P}(Y = 1)$. Recall that $\eta(x) := \mathbb{P}(Y = 1|X = x)$.

1. Derive an expression for the density $g(x)$ of X in terms of π_1, g_k.
Hint: Use the law of total probability.

2. Show that
$$\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) = \log\Big(\frac{\pi_1}{1-\pi_1}\Big) + \log\Big(\frac{g_1(x)}{g_{-1}(x)}\Big).$$
Hint: Bayes' theorem.

Now, let for $k \in \mathcal{Y}$
$$g_k(x) = \frac{1}{((2\pi)^d \det(\Sigma_k))^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\Big)$$
be the density of a $N(\mu_k, \Sigma_k)$ distribution, where $\mu_k \in \mathbb{R}^d$, $\Sigma_k \in \mathbb{R}^{d\times d}$ is positive definite.

(c) Derive a simplified expression for $\log(\frac{g_1(x)}{g_{-1}(x)})$.

(d) Show the following statement: If $\Sigma_1 = \Sigma_2$, $\pi_1 = \frac{1}{2}$ and $\mu_1^T\Sigma^{-1}\mu_1 = \mu_{-1}^T\Sigma^{-1}\mu_{-1}$, then the model has linear optimal decision boundaries and (X, Y) satisfies the model assumption of logistic regression.

(e) Provide a simple condition for μ_1, μ_{-1} such that $\mu_1^T\Sigma^{-1}\mu_1 = \mu_{-1}^T\Sigma^{-1}\mu_{-1}$.

(f) Show the following statement: If $\Sigma_1 = \Sigma_2$, then the model has 'affine' linear optimal decision boundaries and $(\tilde{X}, Y)$ with $\tilde{X} = (1, X)$ satisfies the model assumption of logistic regression.

(g) Show that for general Σ_1, Σ_2, the model has quadratic optimal decision boundaries and $(\tilde{X}, Y)$ with $\tilde{X} = h(X)$ mit $h(x) = (1, (x_j)_{j=1,\dots,d}, (x_j x_l)_{1\le j\le l\le d})$ satisfies the model assumption of logistic regression.

(h) Which convergence rates do we expect for the logistic regression classifier from Theorem 3.12 applied to $(\tilde{X}, Y)$ in the cases (d),(f),(g)?

Task 10 (The Bayes risk in classification problems)**.** Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{+1, -1\}$ (classification problems with 2 classes). Suppose that the same assumptions as in task 2 are satisfied, in particular that X given $Y = k$ is $N(\mu_k, \Sigma)$ distributed with density
$$g_k(x) = \frac{1}{((2\pi)^d \det(\Sigma))^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\Big),$$
where $\mu_k \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ is positive definite. In this task we derive an expression for the Bayes risk $R(f^*)$.

1. Define $\delta^*(x) := \log(\frac{\eta(x)}{1-\eta(x)})$. Show that
$$f^* = \operatorname{sign}(\delta^*),$$
and derive $\delta^*(x)$.

2. Show that

$$R(f^*) = \pi_1 \mathbb{P}(\delta^*(X) < 0 | Y = 1) + (1 - \pi_1)\mathbb{P}(\delta^*(X) > 0 | Y = -1).$$

3. Show that conditionally on $Y = 1$ it holds that $\delta^*(X) \sim N(T + \frac{1}{2}\Delta, \Delta)$. Show that conditionally on $Y = -1$, it holds that $\delta^*(X) \sim N(T - \frac{1}{2}\Delta, \Delta)$, where

$$\Delta := (\mu_1 - \mu_{-1})^T \Sigma^{-1} (\mu_1 - \mu_{-1}) \qquad \text{so-called Mahalanobis distance,} \qquad T := \log\Big(\frac{\pi_1}{1 - \pi_1}\Big).$$

4. Show that

$$R(f^*) = \pi_1 \Phi\Big(\frac{-T - \frac{1}{2}\Delta}{\sqrt{\Delta}}\Big) + (1 - \pi_1)\Big(1 - \Phi\Big(\frac{-T + \frac{1}{2}\Delta}{\sqrt{\Delta}}\Big)\Big),$$

where Φ is the distribution function of the standard normal distribution.

5. Show that in the special case $\pi_1 = \frac{1}{2}$, $\Sigma = I_{d\times d}$, $\mu_{-1} = -\mu_1$, we have $R(f^*) = \Phi(-\|\mu_1\|_2)$.

4 Support Vector Machines

In this chapter, we consider classification problems with $\mathcal{X} \subset \mathbb{R}^d$ and $K = 2$ classes $\mathcal{Y} = \{+1, -1\}$. The measure the quality of the classifier with the 0-1 loss $L(y, s) = \mathbb{1}_{\{y \neq s\}}$.
Goal: Find new methods which yield linear decision boundaries but also allow for better generalizations to nonlinear decision boundaries.
We first present a very naive approach.

Example 4.1 (Classification with linear regression)**.** Let (X_i, Y_i), $i = 1, ..., n$ be training samples. Naive approach: Apply the LS-estimator to (X_i, Y_i) with squared loss $\tilde{L}(y, s) = (y - s)^2$.

$$\hat{\beta} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \tilde{R}_n(\beta), \qquad \tilde{R}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \tilde{L}(Y_i, X_i^T \beta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2.$$

Put $\hat{\delta}_n^{\text{naiv}}(x) = x^T \hat{\beta}$ and $\hat{f}_n^{\text{naiv}}(x) = \text{sign}(\hat{\delta}_n^{\text{naiv}}(x))$.

Problem: If the training samples X_i corresponding to different classes are linear separable (that is, there exists a line or a hyperplane which separates the two point clouds $\{X_i : Y_i = 1\}$,$\{X_i : Y_i = -1\}$), then a classifier should yield such a separation. The reason is that new training samples of class 1 are expected to be 'near' to observations which had class 1; therefore a clear separation of the sets $\{X_i : Y_i = 1\}$,$\{X_i : Y_i = -1\}$ yields a 'safety buffer' between the two classes. The above naive approach does *not* yield such a solution.

4.1 Separating hyperplanes

We first collect some elementary properties of hyperplanes.

Definition 4.2. An (affine) hyperplane with normal vector $\beta \in \mathbb{R}^d$ and base $\beta_0 \in \mathbb{R}$ is a set $\{x \in \mathbb{R}^d : x^T \beta + \beta_0 = 0\}$. Let

$$\Delta := \{\delta_{\beta, \beta_0}(x) = x^T \beta + \beta_0 \,|\, \beta \in \mathbb{R}^d, \beta \in \mathbb{R}\}$$

be the set of all hyperplane functions.

Lemma 4.3. The distance (with sign) of a point $x \in \mathbb{R}^d$ to an hyperplane δ_{β,β_0} is given by $\frac{1}{\|\beta\|_2}(x^T\beta + \beta_0)$.

Remark: The sets $\{x \in \mathbb{R}^d : \delta_{\beta,\beta_0}(x) < 0\}$ and $\{x \in \mathbb{R}^d : \delta_{\beta,\beta_0}(x) > 0\}$ correspond to the sets of points on both sides of the hyperplane δ_{β,β_0}.
The basic approach is to define a classifier as in the case of logistic regression via

$$\hat{f}_n(x) = \text{sign}(\hat{\delta}_n(x)),$$

where $\hat{\delta}_n(x) \in \Delta$. This means the problem of finding a classifier boils down to find a suitable hyperplane. To derive an estimator $\hat{\delta}_n$ we start with the following assumption:

There exists a hyperplane which separates $\{X_i : Y_i = -1\}$ and $\{X_i : Y_i = +1\}$.

Such a hyperplane is also called *separating hyperplane*. Formally, this assumption can be written as follows:

$$(A) \qquad \exists \delta_{\beta,\beta_0} \in \Delta : \quad Y_i = -1 \Rightarrow \delta_{\beta,\beta_0}(X_i) \leq 0 \quad \text{und} \quad Y_i = +1 \Rightarrow \delta_{\beta,\beta_0}(X_i) \geq 0$$

Approach (motivated graphically): Find a hyperplane which separates the point clouds corresponding to different classes such that on both sides of the hyperplane, there is still a large distance towards the elements of the point clouds. In the special case of $d = 2$, this can be visualized as follows: Find a street between the two point clouds which is as broad as possible. The middle of this street then is the hyperplane we are searching for.

Statistical justification for this approach: New points of class '-1' are located closely to the points which were already observed. The larger the distance of $\{X_i : Y_i = -1\}$ towards the hyperplane, the more probable it is that new points of class '-1' lie on the right side of the hyperplane and are therefore classified correctly through $\hat{f}_n$.

Observation: Each hyperplane δ_{β,β_0} from (A) satisfies the following: If Y_i is classified correctly by $f(x) = \text{sign}(\delta_{\beta,\beta_0}(x))$, then it holds that

$$Y_i \cdot \delta_{\beta,\beta_0}(x) > 0.$$

Derivation of *the* separating hyperplane:

- Step 1: For each $i \in \{1, ..., n\}$ let $m_i := Y_i \cdot \frac{1}{\|\beta\|_2}(X_i^T\beta + \beta_0)$ be the distance of X_i towards the hyperplane δ_{β,β_0}. Let

  $$M := \inf_{i=1,...,n} m_i$$

  be the minimal distance of all observations towards the hyperplane (note that $M < 0$ implies that δ_{β,β_0} misclassifies points)

- Step 2: The goal is to find δ_{β,β_0} such that M is as large as possible, that is, we aim to solve the following optimization problem

$$\max_{\beta,\beta_0} M \qquad \text{s.t.} \qquad \forall i = 1, ..., n : \qquad Y_i \cdot \frac{1}{\|\beta\|_2}(X_i^T\beta + \beta_0) = m_i \geq M.$$
$$\iff \quad Y_i(X_i^T\beta + \beta_0) \geq M\|\beta\|_2.$$

- Step 3: (β, β_0) is not uniquely determined! Each $(c \cdot \beta, c \cdot \beta_0)$ with $c > 0$ is also a solution if (β, β_0) is a solution. We therefore have to eliminate one degree of freedom. Let c be chosen such that $\|\beta\|_2 = 1/M$ (*). Then the above optimization problem is equivalent to

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 \qquad \text{s.t.} \qquad \forall i = 1, ..., n : \quad Y_i(X_i^T\beta + \beta_0) \geq 1.$$

Definition 4.4 (Optimal separating hyperplane)**.** Let $\hat{\beta}$, $\hat{\beta}_0$ be solutions of

$$\min_{\beta\in\mathbb{R}^d,\beta_0\in\mathbb{R}} \frac{1}{2}\|\beta\|_2^2 \qquad \text{s.t.} \qquad \forall i = 1, ..., n : \quad Y_i(X_i^T\beta + \beta_0) \geq 1, \tag{24}$$

then $\hat{\delta}_n^{OSH} := \delta_{\hat{\beta},\hat{\beta}_0}$ is called *optimal separating hyperplane* and $\hat{f}_n^{OSH}(x) := \text{sign}(\hat{\delta}_n^{OSH}(x))$ is the corresponding classifier.

Caution: Starting from the substitution (*) above (which assumes $M \geq 0$), the problem has no longer a solution if (A) is violated. Graphically, the set

$$\{x \in \mathbb{R}^d : \delta_{\hat{\beta},\hat{\beta}_0}(x) \in [-1, 1]\}$$

is the 'street' which separates the point clouds $\{X_i : Y_i = -1\}$ and $\{X_i : Y_i = +1\}$. Accordingly, $\{x \in \mathbb{R}^d : \delta_{\hat{\beta},\hat{\beta}_0}(x) = \pm 1\}$ is the roadside and $\{x \in \mathbb{R}^d : \delta_{\hat{\beta},\hat{\beta}_0}(x) = 0\}$ is the middle of the street.

4.2 Support vector machines (SVM)

We now solve the issue that Definition 4.4 has no solution of (A) is violated.

Idea: We allow for misclassified points by introducing a tolerance parameter ξ_i for each point (the so-called 'slack variables'). Additionally, we minimize the sum of these slack variables to obtain solutions which only make use of few misclassifications. To do so, we change the constraint to

$$Y_i(X_i^T\beta + \beta_0) \geq 1 - \xi_i$$

with $\xi_i \geq 0$, $\sum_{i=1}^n \xi_i \leq D$. Here, $D > 0$ is a limit specified by the user which should be proportional to the number and strength of the misclassifications allowed. Note that $\xi_i > 1$ corresponds to a misclassification.

New optimization problem:

$$\min_{\beta,\beta_0,\xi} \frac{1}{2}\|\beta\|_2^2 \qquad \text{s.t.} \qquad \forall i = 1,...,n: \qquad Y_i(X_i^T\beta + \beta_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0,$$
$$\sum_{i=1}^n \xi_i \leq D.$$

As we have seen in the chapter 2 about linear regression (ridge and LASSO estimator), we can find an equivalent formulation of the above optimization problem which replaces the constraint $\sum_{i=1}^n \xi_i \leq D$ by an additive term. This leads to the following definition.

Definition 4.5 (SVM classifier)**.** Let $C > 0$. Let $\hat{\beta}_C$, $\hat{\beta}_{0,C}, \hat{\xi}$ be solutions of

$$\min_{\beta\in\mathbb{R}^d,\beta_0\in\mathbb{R},\xi\in\mathbb{R}^n} \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \qquad \text{s.t.} \qquad \forall i = 1,...,n: \qquad Y_i(X_i^T\beta + \beta_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0.$$

Define $\hat{\delta}_{n,C}^{SVM}(x) = \delta_{\hat{\beta}_C,\hat{\beta}_{0,C}}(x)$. The algorithm $\hat{f}_{n,C}^{SVM}(x) := \text{sign}(\hat{\delta}_{n,C}^{SVM}(x))$ is called *SVM classifier*.

Remark 4.6. As it was the case for the optimal separating hyperplane, $\{x \in \mathbb{R}^d : \hat{\delta}_{n,C}^{SVM}(x) \in [-1,1]\}$ can be thought of the 'street' which is put between the point clouds by the SVM algorithm. According to the meaning of ξ_i, we have to distinct between three cases for each X_i:

- $\xi_i = 0$: X_i is classified correctly and lies outside of the street or on the roadside,
- $\xi_i \in (0,1]$: X_i is classified correctly but lies 'on' the street,
- $\xi_i > 1$: X_i is misclassified.

Remark: The condition $\sum_{i=1}^n \xi_i \leq D$ is a rather strong and unsatisfying constraint since it introduces a maximal limit for the number of misclassifications. If D is too small, there

may not exist a solution. instead, the additive formulation with penalization $C\sum_{i=1}^n \xi_i$ always has a solution.

4.3 Dual formulation of the SVM

In this chapter we reformulate the optimization problem from Definition 4.5. The new representation allows for a deeper insight, justifies the name SVM and also is the starting point for a generalization to nonlinear decision boundaries.
In the following theorem from convex analysis, inequalities which include vectors should be understood component-wise!

Theorem 4.7. Let $F : \mathbb{R}^r \to \mathbb{R}$, $G : \mathbb{R}^r \to \mathbb{R}^k$ be convex and continuously differentiable. Suppose that there exists $\theta \in \mathbb{R}^r$ such that $G(\theta) < 0$ (the so-called Slater condition). We consider the optimization problem

$$\min_{\theta\in\mathbb{R}^r} F(\theta) \quad \text{s.t.} \quad G(\theta) \leq 0. \tag{25}$$

The mapping

$$L(\theta, p) := F(\theta) + p^T G(\theta)$$

is called the corresponding Lagrange function. Then the following statement holds: $\hat\theta \in \mathbb{R}^r$ is a solution of (25) if there exists some $\hat p \in \mathbb{R}^k_{\geq 0}$ such that the *optimality conditions*

$$0 = \nabla_\theta L(\hat\theta, \hat p), \quad G(\hat\theta)^T \hat p = 0, \quad \hat p \geq 0, \quad \hat G(\hat\theta) \leq 0 \tag{26}$$

are satisfied. In this case, $(\hat\theta, \hat p)$ is also a solution of the so-called *Wolfe dual*

$$\sup_{p\in\mathbb{R}^k_{\geq 0},\theta\in\mathbb{R}^r} L(\theta, p) \quad \text{s.t.} \quad \nabla_\theta L(\theta, p) = 0. \tag{27}$$

Starting from the formulation in Definition 4.5, we define for $\theta = (\beta, \beta_0, \xi)$:

$$F(\theta) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i, \quad G(\theta) = \begin{pmatrix} \big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big)_{i=1,\dots,n} \\ -\xi \end{pmatrix}.$$

The corresponding Lagrange function (with $p = (\alpha, \gamma) \in \mathbb{R}^{2n}_{\geq 0}$) is given by

$$L(\theta, p) = F(\theta) + \sum_{i=1}^n \alpha_i\big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big) - \sum_{i=1}^n \gamma_i\xi_i.$$

One can show (this is left as an exercise), that the Wolfe dual (27) for the SVM optimization problem from Definition (4.5) is only an optimization problem with respect to

$\alpha \in \mathbb{R}^n$. With $\mathbb{Y} := (Y_1, ..., Y_n)^T$, $\mathbb{1} := (1, ..., 1)^T$ and $Q = (Q_{ij})_{i,j=1,...,n}$ defined via

$$Q_{ij} := Y_i Y_j X_i^T X_j,$$

the Wolfe dual reads

$$\sup_{\alpha \in \mathbb{R}^n} \{\frac{1}{2}\alpha^T Q \alpha - \mathbb{1}^T \alpha\} \quad \text{s.t.} \quad \mathbb{Y}^T \alpha = 0, \quad 0 \leq \alpha \leq C.$$

Based on the optimality conditions (26), we can infer a solution of the original solution from the Wolfe dual solution. From (26) we obtain

$$\hat{\beta}_C = \sum_{i=1}^n \hat{\alpha}_i Y_i X_i, \qquad \hat{\beta}_{0,C} = Y_i - X_i^T \hat{\beta}_C \quad \text{with some } i \text{ with } 0 < \hat{\alpha}_i < C.$$

This leads to the following theorem.

Theorem 4.8 (SVM classifier, dual formulation)**.** Let $C > 0$. Let $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_n)^T$ be a solution of

$$\min_{\alpha \in \mathbb{R}^n} \left\{\frac{1}{2}\alpha^T Q \alpha - \mathbb{1}^T \alpha\right\} \qquad \text{s.t.} \qquad \mathbb{Y}^T \alpha = 0, \quad 0 \leq \alpha \leq C.$$

Then the following holds: $\hat{\beta}_C = \sum_{i=1}^n \hat{\alpha}_i Y_i \cdot X_i$ and $\hat{\beta}_{0,C} = Y_i - X_i^T \hat{\beta}_C$, where $i \in \{1, ..., n\}$ is chosen such that $0 < \hat{\alpha}_i < C$.

Remark: We see that $\hat{\beta}_C$ is a linear combination of vectors X_i which correspond to $\hat{\alpha}_i \neq 0$. The optimality conditions (26) imply:

- $\hat{\alpha}_i = 0 \Rightarrow \hat{\xi}_i = 0$, $\delta_{\hat{\beta}_C, \hat{\beta}_{0,C}}(X_i) \geq 1$ (X_i is classified correctly and is not located on the street)
- $\hat{\alpha}_i \in (0, C) \Rightarrow \hat{\xi}_i = 0$, $\delta_{\hat{\beta}_C, \hat{\beta}_{0,C}}(X_i) = 1$ (X_i is classified correctly and is on the roadside of the street)
- $\hat{\alpha}_i = C \Rightarrow \hat{\xi}_i \neq 0$, $\delta_{\hat{\beta}_C, \hat{\beta}_{0,C}}(X_i) \geq 1 - \xi_i$ (X_i is either classified correctly or misclassified and is located either on the street or on the wrong side of the street).

Note the following important property: $\hat{\alpha}_i \neq 0$ does not holds for all $i \in \{1, ..., n\}$ but only for a small part of the training samples X_i. More precisely, $\hat{\alpha}_i \neq 0$ only holds for those points X_i which have an influence on the form and location of the street. The equation $\hat{\beta}_C = \sum_{i=1}^n \hat{\alpha}_i Y_i \cdot X_i$ can be interpreted in the way that the street itself is only determined by those points with $\hat{\alpha}_i \neq 0$ (which are those nearest to points of the other class). These points 'support' the street because of their location, thus they are also called *support vectors* and the whole algorithm support vector machine (SVM).

4.4 Generalized SVM

Problem: Up to now, the SVM only allows for linear decision boundaries since

$$\hat{\delta}_{n,C}^{SVM} \in \Delta = \{\delta_{\beta,\beta_0}(x) = \beta^T x + \beta_0 \,|\, \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}.$$

Solution: We apply the approach from Section 3.4: The original linear SVM algorithm is applied to transformed data points

$$\tilde{X} = h(X) \quad \text{with} \quad h : \mathcal{X} \subset \mathbb{R}^d \to \tilde{\mathcal{X}} \subset \mathbb{R}^{\tilde{d}}.$$

This leads to the following algorithm: Define $\tilde{Q} = (\tilde{Q}_{ij})_{i,j=1,...,n}$ via

$$\tilde{Q}_{ij} = y_i y_j h(X_i)^T h(X_j),$$

and put

Definition 4.9 (SVM classifier, dual nonlinear formulation)**.** Let $C > 0$. Let $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_n)^T$ be a solution of

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{2} \alpha^T \tilde{Q} \alpha - \mathbb{1}^T \alpha \right\} \qquad \text{s.t.} \qquad \mathbb{Y}^T \alpha = 0, \quad 0 \le \alpha \le C.$$

Put $\hat{\beta}_C^{nl} = \sum_{i=1}^n \hat{\alpha}_i Y_i \cdot h(X_i)$, and with some $i^* \in \{1, ..., n\}$ with $0 < \hat{\alpha}_{i^*} < C$: $\hat{\beta}_{0,C}^{nl} := Y_{i^*} - h(X_{i^*})^T \hat{\beta}_C$. Put

$$\hat{\delta}_{n,C}^{SVM,nl}(x) = \delta_{\hat{\beta}_C^{nl}, \hat{\beta}_{0,C}^{nl}}(h(x)), \qquad \hat{f}_{n,C}^{SVM,nl}(x) = \text{sign}(\hat{\delta}_{n,C}^{SVM,nl}(x)).$$

Important observation: Contrary to logistic regression, the transformed training samples $\tilde{X}_i = h(X_i)$ must *not* be calculated to apply the classifier. It is enough to know the values

$$K(x, x') := h(x)^T h(x')$$

for each $x, x' \in \mathcal{X}$. We now show this in more detail. Note that $\tilde{Q}_{ij} = y_i y_j K(X_i, X_j)$, and

$$\begin{aligned}
\hat{\delta}_{n,C}^{SVM,nl}(x) &= (\hat{\beta}_C^{nl})^T h(x) + \hat{\beta}_{0,C}^{nl} \\
&= \sum_{i=1}^n \hat{\alpha}_i Y_i \cdot h(X_i)^T h(x) + \{Y_{i^*} - \sum_{i=1}^n \hat{\alpha}_i Y_i h(X_i)^T h(X_{i^*})\} \\
&= \sum_{i=1}^n \hat{\alpha}_i Y_i \cdot K(X_i, x) + \big\{Y_{i^*} - \sum_{i=1}^n \hat{\alpha}_i Y_i K(X_i, X_{i^*})\big\}.
\end{aligned}$$

That is, all quantities which include $h(X_i)$ are removed. We obtain the following, so-called 'kernel-based' SVM algorithm (the terminology of a kernel is introduced below).

Theorem 4.10 (SVM classifier, dual formulation with kernel K)**.** Let $K(x, x') = h(x)^T h(x')$ with some $h : \mathcal{X} \to \mathbb{R}^{\tilde{d}}$. Let $C > 0$. Let $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_n)^T$ be a solution of

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{2} \alpha^T \tilde{Q} \alpha - \mathbb{1}^T \alpha \right\} \qquad \text{s.t.} \qquad \mathbb{Y}^T \alpha = 0, \quad 0 \leq \alpha \leq C.$$

Let $i^* \in \{1, ..., n\}$ be such that $0 < \hat{\alpha}_{i^*} < C$. Then it holds that

$$\hat{\delta}_{n,C}^{SVM,nl}(x) = \sum_{i=1}^{n} \hat{\alpha}_i Y_i \cdot K(X_i, x) + \left\{ Y_{i^*} - \sum_{i=1}^{n} \hat{\alpha}_i Y_i K(X_i, X_{i^*}) \right\}.$$

The fact that $h(X_i)$ does not occur in the above formulas frees us from the problem to compute the (possibly large) vectors $\tilde{X}_i = h(X_i)$ before applying the SVM algorithm. Moreover, the introduction of the function K allows us to formulate a more abstract version of the algorithm. Note that we can interprete the SVM algorithm with a function

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

in two ways:

- 1) The SVM algorithm is applied in the original space $\mathcal{X} \subset \mathbb{R}^d$ but with a *new scalar product* $K(x, x')$ instead of the standard scalar product $x^T x'$. The nonlinearity introduced by h can be interpreted as a change of the scalar product.
- 2) The observations are embedded in a high-dimensional space $\tilde{\mathcal{X}} \subset \mathbb{R}^{\tilde{d}}$ via the nonlinear h and the SVM is applied in this high-dimensional space with the standard scalar product.

K is called *kernel function* or *kernel*. In the following we will assume that K is a *Mercer kernel*. A Mercer kernel summarizes all properties which are needed to interprete K as a scalar product.

Definition 4.11 (Mercer kernel)**.** A mapping $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *Mercer kernel* on $\mathcal{X}$ if the following properties hold:

(i) K is continuous

(ii) K is symmetric, that is,

$$\forall x, x' \in \mathbb{R}^d : \quad K(x, x') = K(x', x)$$

(iii) K is positive semidefinite, that is,

$$\forall n \in \mathbb{N} : \quad \forall x^{(1)}, ..., x^{(n)} \in \mathbb{R}^d : \quad \left(K(x^{(i)}, x^{(j)})\right)_{i,j=1,...,n} \quad \text{is positive semidefinite.}$$

There exist several Mercer kernels which are often used in practice:

Example 4.12. (i) linear kernel: $K(x, x') = x^T x'$,

(ii) polynomial kernel of degree $p \in \mathbb{N}$:

$$K_p(x, x') = (1 + x^T x')^d$$

(iii) radial basis functions (Gaussian kernel) with descent $\gamma > 0$:

$$K_\gamma(x, x') = \exp(-\gamma \cdot \|x - x'\|_2^2)$$

By using kernels, we loose the graphical interpretation which nonlinear transformations h of the original observations were considered. However, this can be deduced from the kernels. For the polynomial kernel, we have the following result.

Example 4.13. For $d = p = 2$, it holds that

$$K_p(x, x') = h(x)^T h(x')$$

with $h : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$, $\tilde{d} = 6$ and

$$h(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)^T.$$

We now present a more general result which shows that every Mercer kernel allows for

such a representation. Define

$$\begin{aligned} L^2(\mathcal{X}) &:= \{f : \mathcal{X} \to \mathbb{R} \text{ measurable}, \|f\|_{L^2(\mathcal{X})} := \int_{\mathcal{X}} f^2(x)\, \mathrm{d}\lambda(x) < \infty\}, \\ \ell^2 &:= \{a = (a_k)_{k\in\mathbb{N}} : \|a\|_{\ell^2} := \sum_{k=1}^{\infty} a_k^2 < \infty\}. \end{aligned}$$

Then we have the following representation theorem.

Theorem 4.14 (Mercer's Theorem)**.** Let $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a Mercer kernel and $\mathcal{X} \subset \mathbb{R}^d$ be compact. Then there exists an orthonormal basis $\tilde{h}_k \in L^2(\mathcal{X})$ and $\lambda_k \geq 0$ $(k \in \mathbb{N})$ with

$$K(x, x') = \sum_{k=1}^{\infty} \lambda_k \tilde{h}_k(x)\tilde{h}_k(x') \qquad x, x' \in \mathcal{X}.$$

The function

$$h(\cdot) := (\sqrt{\lambda_k} \cdot \tilde{h}_k(\cdot))_{k\in\mathbb{N}} : \mathcal{X} \to \ell^2$$

is called *feature map*, is well-defined, continuous and satisfies

$$K(x, x') = h(x)^T h(x'), \qquad x, x' \in \mathcal{X}. \tag{28}$$

That is, each Mercer kernel allows (in principle) for a representation of the form (28).

4.5 The SVM algorithm as a minimizer of an empirical risk

From Theorem 4.10 we obtain the following result: A solution of the SVM optimization problem has the form

$$\hat{\delta}_{n,C}^{SVM,nl} = \sum_{i=1}^{n} a_i K(X_i, x) + b$$

with $a_i, b \in \mathbb{R}$ $(i = 1, ..., n)$. We want to represent $\hat{\delta}_{n,C}^{SVM,nl}$ as a minimizer of an empirical risk to prove statistical results. To do so, we have to define the space of functions in which $\hat{\delta}_{n,C}^{SVM,nl}$ lies.

Definition 4.15 (RKHS)**.** Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel. Let $K_x := K(x, \cdot)$ and

$$\begin{aligned} H_0 &:= \mathrm{Lin}(K_x : x \in \mathcal{X}) \\ &= \Big\{ g_{a,y}(x) = \sum_{i=1}^{m} a_i \cdot K(x, y^{(i)}) : y^{(1)}, ..., y^{(m)} \in \mathcal{X}, a_1, ..., a_m \in \mathbb{R}, m \in \mathbb{N} \Big\}. \end{aligned}$$

For $g_{a,y}, g_{b,z} \in H_0$, define the scalar product

$$\langle g_{a,y}, g_{b,z} \rangle_K := \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j K(y^{(i)}, z^{(j)}),$$

The closure of $(H_0, \langle \cdot, \cdot \rangle_K)$ is denoted by $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ and is called *Reproducing kernel Hilbert space* (RKHS) of K. The corresponding norm is denoted by $\|g\|_K := \sqrt{\langle g, g \rangle_K}$.

Definition 4.16. $\mathcal{H}_K^b := \{\delta + \beta_0 : \delta \in \mathcal{H}_K, \beta_0 \in \mathbb{R}\}$.

Note that we have always

$$\hat{\delta}_{n,C}^{SVM,nl} \in \mathcal{H}_K^b.$$

Theorem 4.17. $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K)$ has the so-called *reproducing property*:

$$\forall g \in \mathcal{H}_K : \quad g = \langle g, K_x \rangle_K.$$

For h being the function from Theorem 4.14, we can find for each $g \in \mathcal{H}_K$ a $\beta \in \ell^2$ such that

$$g = \beta^T h.$$

Then, the following isometry property holds:

$$\|g\|_K^2 = \|\beta\|_{\ell^2}^2$$

We now deduce that $\hat{\delta}_{n,C}^{SVM,nl}$ is indeed a minimizer of an empirical risk with respect to functions of $\mathcal{H}_K^b$. To do so, we start from the original definition 4.5 of the SVM but with nonlinear modified observations $h(X_i)$.

- Step 1: Let $h : \mathcal{X} \to \ell^2$ be the feature map from Theorem 4.14. Then the original definition of the SVM algorithm with modified observations $h(X_i)$ and $\beta \in \ell^2$ reads

$$\min_{\beta \in \ell^2, \beta_0 \in \mathbb{R}, \xi \in \mathbb{R}^n} \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad \forall i = 1, ..., n : \quad Y_i(\beta^T h(X_i) + \beta_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \tag{29}$$

and the classifier is given by

$$\hat{f}_{n,C}(x) = \operatorname{sign}(\hat{\delta}_{n,C}(x)), \qquad \hat{\delta}_{n,C}(x) = \hat{\beta}_C^T h(x) + \hat{\beta}_{0,C}$$

- Step 2: Due to $\mathcal{H}_K = \{g = \beta^T h : \beta \in \ell^2\}$ and $\|\beta\|_2^2 = \|g\|_K^2$, (29) is equivalent to

$$\min_{g\in\mathcal{H}_K,\beta_0\in\mathbb{R},\xi\in\mathbb{R}^n} \frac{1}{2}\|g\|_K^2 + C\sum_{i=1}^n \xi_i \quad \text{s.t.} \quad \forall i = 1,...,n: \quad Y_i(g(X_i)+\beta_0) \geq 1-\xi_i, \quad \xi_i \geq 0 \tag{30}$$

  and

$$\hat{\delta}_{n,C}(x) = \hat{g}(x) + \hat{\beta}_{0,C}.$$

- Step 3: We now simplify the minimization problem by using the constraints. For fixed $g \in \mathcal{H}_K, \beta_0 \in \mathbb{R}$, the optimization problem can be solved with respect to $\xi_1, ..., \xi_n$: The objective $\frac{1}{2}\|g\|_K^2 + C\sum_{i=1}^n \xi_i$ is minimal if

$$\hat{\xi}_i(g,\beta_0) = \begin{cases} 0, & Y_i(g(X_i)+\beta_0) \geq 1, \\ 1 - Y_i(g(X_i)+\beta_0), & Y_i(g(X_i)+\beta_0) < 1 \end{cases} = \big(1 - Y_i(g(X_i)+\beta_0)\big)^+,$$

  where $a^+ := \max\{a, 0\}$.

  Plugging in (30) and using the definition $\lambda := \frac{1}{2nC}$, we obtain that (30) is equivalent to

$$\min_{g\in\mathcal{H}_K,\beta_0\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \big(1 - Y_i(g(X_i)+\beta_0)\big)^+ + \lambda\|g\|_K^2,$$

  and

$$\hat{\delta}_{n,C}(x) = \hat{g}(x) + \hat{\beta}_{0,C}.$$

Thus, we obtain a third formulation of the SVM algorithm which is now accessible to statistical theory.

Definition 4.18 (SVM, risk minimization formulation)**.** Let $\lambda > 0$. Let $\tilde{L}(y,s) := (1 - y\cdot s)^+$ (the so-called *hinge loss*, $a_+ := \max\{a, 0\}$), and

$$\hat{\delta}_n :\in \operatorname*{arg\,min}_{\delta\in\mathcal{H}_K^b} \big\{\tilde{R}_n(\delta) + \lambda\cdot\|\delta\|_K^2\big\}, \qquad \tilde{R}_n(\delta) = \frac{1}{n}\sum_{i=1}^n \tilde{L}(Y_i, \delta(X_i)). \tag{31}$$

Put $\hat{f}_n(x) = \operatorname{sign}(\hat{\delta}_n(x))$.

We obtain the following theorem as a connection between the original formulation of the SVM and the risk minimization formulation.

Theorem 4.19. If $\lambda = \frac{1}{2nC}$, then $\hat{\delta}_n = \hat{\delta}_{n,C}^{SVM,nl}$, where $\hat{\delta}_{n,C}^{SVM,nl}$ is from Definition 4.9 or Theorem 4.10, respectively.

Proof. We have seen that $\hat{\delta}_n$ from Definition 4.18 is a solution of Definition 4.5 with observations $h(X_i)$ replaced by X_i. Theorem 4.8 applied to $h(X_i)$ instead of X_i yields that this is equivalent to the optimization problem from Definition 4.9 or Theorem 4.10, respectively. □

The SVM therefore tries to approximate δ^* with functions from $\mathcal{H}_K^b$.

4.6 Theoretical statements

Using Theorem 3.19 and Theorem 3.21, one can show (this is left as an exercise) that the hinge loss $\tilde{L}(y,s) := (1 - y \cdot s)^+$ satisfies the calibration condition and the risk transfer formula. To formulate the results, let

$$\delta^* \in \operatorname*{arg\,min}_{\delta:\mathcal{X}\to\mathbb{R}} \tilde{R}(\delta), \qquad \tilde{R}(\delta) = \mathbb{E}\tilde{L}(Y, \delta(X)),$$

and $f^* \in \arg\min_{f:\mathcal{X}\to\mathcal{Y}} R(f)$, where $R(f) = \mathbb{E}L(Y, f(X))$. Recall that $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

Lemma 4.20. On $\{x \in \mathcal{X} : \eta(x) \neq \frac{1}{2}\}$ we have $\mathbb{P}^X$-a.s. $\delta^* = f^*$. Then the risk transfer formula holds with $G(r) = r$, that is, for measurable $\delta : \mathcal{X} \to \mathbb{R}$ it holds that

$$R(\operatorname{sign}(\delta)) - R(f^*) \leq \tilde{R}(\delta) - \tilde{R}(\delta^*).$$

Caution: Contrary to the algorithms considered before we *can not* simply assume that $\delta^* \in \mathcal{H}_K^b$, that is, we can not set the approximation error to 0. The reason is that the elements of $\mathcal{H}_K^b$ are (in general) continuous while δ^* is not. A theoretical statement therefore also has to contain the approximation error.
We need another definition, the so-called integral operator associated to a kernel. Let

$$L^2(\mathbb{P}^X) := \{\delta : \mathcal{X} \to \mathbb{R}, \mathbb{E}[\delta(X)^2] = \int \delta(x)^2 d\mathbb{P}^X(x) < \infty\}.$$

The following theorem is stated without proof (functional analysis). The quantities γ_k occurring therein are used to formulate the theoretical result about the excess Bayes risk of the SVM.

Definition 4.21 (Integral operator: Definition and theorem)**.** Let K be Mercer kernel. The integral operator associated to K and the distribution $\mathbb{P}^X$ is defined via

$$T_K : L^2(\mathbb{P}^X) \to L^2(\mathbb{P}^X), \quad (T_K g)(x) := \mathbb{E}[K(X, x)g(X)].$$

T_K is symmetric positive semidefinite and thus diagonalizable. Let $\gamma_1 \geq \gamma_2 \geq \gamma_3 \geq ... \geq 0$ denote the ordered eigenvalues. It holds that $\sum_{k=1}^{\infty} \gamma_k < \infty$.

The operator T_K can be interpreted as the covariance matrix $\Sigma_h = \mathbb{E}[h(X)h(X)^T]$ of the transformed observations $h(X)$ (with h from Theorem 4.14, $K(x, x') = h(x)^T h(x')$), Heuristically, one can easily see that γ_k are the eigenvalues of Σ_h: If v is an eigenvector of Σ_h (and thus also from Σ_h^T) corresponding to eigenvalue λ, then $g(x) = v^T h(x)$ satisfies

$$\begin{aligned}(T_K g)(x) &= \mathbb{E}[K(X, x)g(X)] = \mathbb{E}[h(X)^T h(x)v^T h(X)] = v^T \mathbb{E}[h(X)h(X)^T]h(x) \\ &= v^T \Sigma_h h(x) = \lambda v^T h(x) = \lambda g(x),\end{aligned}$$

that is, T_K has eigenvalue λ. This interpretation is used below.
The following theoretical result is taken from [1], Theorem 3.1 therein. The statement is formulated for the empirical risk minimizer from (31) but in a little bit simpler case where the minimum is taken with respect to the space $\mathcal{H}_K$ instead of the space $\mathcal{H}_K^b$.

Theorem 4.22. Let $\mathcal{X} \subset \mathbb{R}^d$ be compact. Let K be a Mercer kernel with $\sup_{x \in \mathcal{X}} K(x, x) \leq 1$. Suppose that there exist $\eta_0, \eta_1 > 0$ such that

$$\forall x \in \mathcal{X} : \quad \big|\eta(x) - \frac{1}{2}\big| \geq \eta_0, \qquad \eta_1 \leq \eta(x) \leq 1 - \eta_1. \tag{32}$$

Let

$$\gamma(n) := \frac{1}{\sqrt{n}} \inf_{a \in \mathbb{N}} \Big\{ \frac{a}{\sqrt{n}} + \eta_1 \sqrt{\sum_{j>a} \gamma_j} \Big\}.$$

Then there exist universal constants $c_1, c_2 > 0$ such that the following holds: For each $t > 0$ and

$$\lambda \geq \frac{c_1}{\eta_1} \Big\{ \gamma(n) + \frac{\log(\log(n) + 1) + t}{n} \Big\} \tag{33}$$

it holds that

$$\mathbb{P}\Big(\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*) \geq 2 \inf_{\delta \in \mathcal{H}_K} \big\{ \tilde{R}(\delta) - \tilde{R}(\delta^*) + 5\lambda \|\delta\|_K^2 \big\} + c_2 \lambda (1 + \frac{\eta_1}{\eta_0}) \Big) \leq e^{-t}. \tag{34}$$

Remarks:

- The basic assertion of the theorem is that $\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*)$ behaves nearly (factor 2) as good as the optimal minimizer over $\mathcal{H}_K$

$$2 \inf_{\delta \in \mathcal{H}_K} \left\{ \tilde{R}(\delta) - \tilde{R}(\delta^*) + 5\lambda \|\delta\|_K^2 \right\}, \tag{35}$$

  but with some additional penalty term $5\lambda\|\delta\|_K^2$ included. The other terms $c_2\lambda(1 + \frac{\eta_1}{\eta_0})$ have not such a big influence on the rate since they stem mainly from the proof technique. A more detailed analysis of the approximation qualities of $\mathcal{H}_K$ can then be used to upper bound (35) to obtain a convergence rate which includes d, n. This was for instance done in [15], Theorem 2.7 under additional assumptions and using the Gaussian kernel. Here, we will not investigate these results in detail.

- Inequalities of the form (34) are called *oracle inequalities*: They connect the quality of $\hat{\delta}_n$ with the quality of the optimizer over the function set considered (here $\mathcal{H}_K$) who *knows* the distribution and the risk function. The underlying dimensionality of the problem and the number of observations enter the convergence rate via λ and its corresponding lower bound given in (33).

- The assumption on the behavior of $\eta(x)$ near $\frac{1}{2}$ in (32) corresponds to the noise condition with $q = \infty$. In principle, this assumption can be relaxed to the noise condition with some $q > 0$. However, then the proof becomes a little bit more technical.

We want to get a feeling which convergence rates can be achieved using the SVM classifier. We will do this by using the simple (but often wrong) assumption that $\delta^* \in \mathcal{H}_K$. Then, (34) would yield

$$\mathbb{P}\Big(\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*) \geq 10\lambda\|\delta^*\|_K^2 + c_2\lambda(1 + \frac{\eta_1}{\eta_0}) \Big) \leq e^{-t}.$$

Thus, the rate of $\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*)$ is mainly determined by $\lambda\|\delta^*\|_K^2$. Depending on the complexity of δ^*, $\|\delta^*\|_K^2$ can be quite large (we will see in the proofs that in principle, $\|\delta^*\|_K$ can be in the worst case of size n). The reason is that a complex δ^* needs a lot of non-zero coefficients β in the representation $\delta^* = \beta^T h$ to be well approximated (recall that $\|\delta^*\|_K^2 = \|\beta\|_{\ell^2}^2$). Thus, $\|\delta^*\|_K^2$ should not be viewed as independent of d, n, but here we will for simplicity assume exactly this and only investigate the rate through λ. The term λ itself is dominated by $\gamma(n)$.

Example 4.23. We now investigate some special cases of the sequence γ_j:

- If $\gamma_1 > 0$ but $\gamma_j = 0$ for all $j \geq 2$, then it holds that

$$\gamma(n) \overset{a=1}{=} \frac{1}{n},$$

that is, the condition $\lambda \geq \frac{c_1}{\eta_1}(\gamma(n) + \frac{\log(\log(n)+1)+t}{n}$ is independent of the input dimension d.
The assumption made for γ_j means that Σ_h only has one eigenvector $v \in \mathbb{R}^{\tilde{d}}$ with the property $\mathrm{Var}(v^T h(X)) > 0$ (all other eigenvector satisfy $\mathrm{Var}(v^T h(X)) = 0$), that is, all values of $h(X)$ are located on a line. In this case, the kernel was chosen *very suitable* for the problem and the underlying distribution $\mathbb{P}^X$. Of course one cannot expect such a good behavior in practice because one typically uses 'standard kernels' (like the Gaussian kernel) which are not adapted to the problem, and furthermore the distribution $\mathbb{P}^X$ is unknown. For the situation above, one would obtain

$$\lambda \approx C(\frac{1}{n} + \frac{\log\log(n)}{n})$$

with some constant $C > 0$.

- If instead $\gamma_j = 0$ for $j \geq k$ ($k \in \mathbb{N}$ fixed), then we have

$$\gamma(n) \overset{a=k}{=} \frac{k}{n}.$$

In this case, $h(X)$ lies in a k-dimensional subspace. For the situation above, one would obtain

$$\lambda \approx C(\frac{k}{n} + \frac{\log\log(n)}{n})$$

with some constant $C > 0$.

- In general, the variation of $h(X)$ is much more complex and it typically holds that $\gamma_j > 0$ for all $j \in \mathbb{N}$. Then, the variation of $h(X)$ can be measured via the decay rate of γ_j. For instance, if $\gamma_j = \gamma_1 \cdot \rho^j$ with some $\rho \in (0, 1)$, then we have with $a = -\frac{\log(n\gamma_1)}{\log(\rho)}$

$$\gamma(n) \leq \frac{1}{\sqrt{n}}\big(- \frac{\log(n\gamma_1)}{\sqrt{n}\log(\rho)} + \eta_1 \frac{1}{\sqrt{n}}\big) \leq c \cdot \frac{\log(n) + \log(\gamma_1)}{n}.$$

The rate of λ therefore is now given by $\frac{\log(n)}{n}$ (and additionally may depend on d through γ_1, ρ). Slower decay rates of $\gamma_j = \gamma_1 j^{-\alpha}$ ($\alpha > 1$) produce even slower convergence rates for λ (this is left as an exercise).

To prove Theorem 4.22, we need some preparation. We start with a margin property similar as in Lemma 3.10 (for logistic regression). The margin property relates $\tilde{R}(\delta) - \tilde{R}(\delta^*)$ to a more simple distance measure $D(\delta, \delta_0) := \mathbb{E}[(\delta(X) - \delta_0(X))^2]^{1/2}$. In the proofs, mainly this more simple distance measure is used.

Lemma 4.24 (Quadratic margin property of the SVM loss)**.** Let $\rho \geq 0$.

1. Suppose (32) (noise condition). Then for each measurable $\delta : \mathcal{X} \to \mathbb{R}$ with $\|\delta\|_\infty \leq \rho$, it holds that
$$D(\delta, \delta^*)^2 \leq c_\rho \big\{ \tilde{R}(\delta) - \tilde{R}(\delta^*) \big\},$$
where $c_\rho := 2(\frac{\rho}{\eta_1} + \frac{1}{\eta_0})$.
2. Suppose that $K(x,x) \leq 1$. Then for each $\delta \in \mathcal{H}_K$, it holds that $\|\delta\|_\infty \leq \|\delta\|_K$. In particular, we have
$$\|\delta\|_K \leq \rho \quad \Rightarrow \quad \|\delta\|_\infty \leq \rho.$$

Proof. 1. Fix $x \in \mathcal{X}$. Without loss of generality, suppose that $\eta(x) > \frac{1}{2}$ ($\Rightarrow \delta^*(x) = 1$). Then we have

$$\begin{aligned} A(x) &:= \frac{(\delta(x) - \delta^*(x))^2}{\mathbb{E}[\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta^*(X)) | X = x]} \\ &= \frac{(\delta - 1)^2}{\eta(1-\delta)_+ + (1-\eta)(1+\delta)_+ - 2(1-\eta)}, \end{aligned}$$

where in the last step, we introduced the abbreviations $\delta = \delta(x), \eta = \eta(x)$. It is now enough to show that $A(x) \leq c_\rho$ (then the assertion follows by applying another expectation with respect to X). For $\eta > \frac{1}{2} > \eta_1$, we have

$$A(x) = \begin{cases} \frac{(1-\delta)^2}{\eta(1-\delta)-2(1-\eta)}, & \delta \leq -1, \\ \frac{\delta - 1}{1-\eta}, & \delta \geq 1, \\ \frac{1-\delta}{2\eta - 1}, & \delta \in [-1,1] \end{cases} \overset{(*)}{\leq} \begin{cases} 4\|\delta\|_\infty + \frac{2}{\eta_0}, & \delta \leq -1, \\ \frac{\|\delta\|_\infty}{\eta_1}, & \delta \geq 1, \\ \frac{1}{\eta_0}, & \delta \in [-1,1] \end{cases} \leq 2\Big(\frac{\|\delta\|_\infty}{\eta_1} + \frac{1}{\eta_0}\Big).$$

The upper bound provided in (*) is due to an elementary calculation: Put $z := -\delta - 1 \in [0, \|\delta\|_\infty - 1]$. Then we have

$$\begin{aligned} \frac{(1-\delta)^2}{\eta(1-\delta) - 2(1-\eta)} &= \frac{(z+2)^2}{\eta z + 2(2\eta - 1)} \overset{(a+b)^2 \leq 2a^2 + 2b^2}{\leq} \frac{2z^2 + 8}{\eta z + 2(2\eta - 1)} \\ &\overset{\frac{a+b}{c+d} \leq \frac{a}{c} + \frac{b}{d}}{\underset{\text{für } a,b,c,d \geq 0}{\leq}} \frac{2z}{\eta} + \frac{2}{\eta - \frac{1}{2}} \overset{\eta > \frac{1}{2}}{\leq} 4\|\delta\|_\infty + \frac{2}{\eta_0}. \end{aligned}$$

2. Let $x \in \mathcal{X}$. Then it holds that

$$|\delta(x)| \overset{\text{Theorem 4.17}}{=} |\langle \delta, K_x \rangle_K| \overset{CSU}{\leq} \|\delta\|_K \cdot \|K_x\|_K \overset{\text{Def. } \|\cdot\|_K \text{ from Def. 4.15}}{=} \|\delta\|_K \sqrt{K(x,x)} \leq \|\delta\|_K.$$

□

Remark: We therefore can only use the margin property for δ if it is known that $\|\delta\|_K \leq \rho$ with some $\rho > 0$! This complicates the proof. Compared to the proofs for logistic regression, we therefore have *two additional difficulties*:

(S1) Discussion of the approximation error

(S2) The margin condition does not hold uniformly for all $\delta \in \mathcal{H}_K$ in the optimization problem (and additionally, it is also not possible to upper bound the expectation of a supremum uniformly over all $\delta \in \mathcal{H}_K$).

4.7 Simplified problem

Before considering Theorem 4.22, we will discuss a more simple estimation problem to understand the structure of the whole proof. Let

$$B(\rho) := \{\delta \in \mathcal{H}_K : \|\delta\|_K \leq \rho\}.$$

We consider

$$\hat{\delta} = \hat{\delta}_n^{fix} \in \underset{\delta \in B(\rho)}{\arg\min}\, \tilde{R}_n(\delta). \tag{36}$$

We have replaced the penalization $\lambda\|\delta\|_K$ by an optimization over a new function set $\delta \in B(\rho)$ with fixed maximal radius ρ.
For each $\delta_0 \in B(\rho)$ (later, we choose $\delta_0 \in \arg\min_{\delta \in B(\rho)}\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}$) it holds that

$$\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*) = \big\{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0)\big\} + \big\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\big\}, \tag{37}$$

and

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) = \underbrace{\tilde{R}_n(\hat{\delta}) - \tilde{R}_n(\delta_0)}_{\leq 0} + \big\{\tilde{R}(\hat{\delta}) - \tilde{R}_n(\hat{\delta}) - (\tilde{R}(\delta_0) - R_n(\delta_0))\big\}. \tag{38}$$

Remark 4.25. In the proof of Theorem 3.12 we introduced $\tilde{\delta} = T\hat{\delta} + (1-T)\delta_0$ with $T = \frac{r}{r + D(\hat{\delta}, \delta_0)}$ ($\Rightarrow D(\tilde{\delta}, \delta_0) \leq r$). We then derived the same inequality (38) for $\tilde{\delta}$ to obtain

$$\tilde{R}(\tilde{\delta}) - \tilde{R}(\delta_0) \leq \sup_{\delta \in B(\rho), D(\delta, \delta_0) \leq r} \big\{\tilde{R}(\delta) - \tilde{R}_n(\delta) - (\tilde{R}(\delta_0) - R_n(\delta_0))\big\} =: Z_{r,\rho}(\delta_0).$$

In principle, we could use this proof technique to obtain a result for $\hat{\delta}_n^{fix}$. It would even be possible to include the approximation error (this is left as an exercise). However, here we aim to use the intermediate results of this simpler proof technique also for the original estimator $\hat{\delta}_n^{SVM}$ (in particular, we want to deal with (S2)). To do so, the introduction of $\tilde{\delta}$ is obstructive. Therefore, we now present a different proof technique for $\hat{\delta}_n^{fix}$. Heuristically, an upper bound of $Z_{r,\rho}(\delta_0)$ for

$$\delta = \tilde{\delta} = \delta_0 + r \cdot \frac{\hat{\delta} - \delta_0}{r + D(\hat{\delta}, \delta_0)}$$

means that we have to investigate a rescaled supremum.

We will see that without introducing $\tilde{\delta}$, it is helpful to find an upper bound for the supremum

$$V_{r,\rho}(\delta_0) := \sup_{\delta \in B(\rho)} \frac{\left\{\tilde{R}(\delta) - \tilde{R}_n(\delta) - (\tilde{R}(\delta_0) - R_n(\delta_0))\right\}}{r^2 + D(\delta, \delta_0)^2}.$$

Let $C \geq 1$ be arbitrary, $r > 0$ and $A := \{V_{r,\rho}(\delta_0) \leq \frac{1}{C}\}$. From (38) we obtain:

$$\begin{aligned} \tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) &\leq \frac{\left\{\tilde{R}(\hat{\delta}) - \tilde{R}_n(\hat{\delta}) - (\tilde{R}(\delta_0) - R_n(\delta_0))\right\}}{D(\hat{\delta}, \delta_0)^2 + r^2} \cdot (D(\hat{\delta}, \delta_0)^2 + r^2) \\ &\leq C^{-1}(D(\hat{\delta}, \delta_0)^2 + r^2) \end{aligned}$$

With (37), we conclude that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq C^{-1}(D(\hat{\delta}, \delta_0)^2 + r^2) + \tilde{R}(\delta_0) - \tilde{R}(\delta^*).$$

Using the margin property from Lemma 4.24, we obtain

$$\begin{aligned} D(\hat{\delta}, \delta_0)^2 &\overset{(a+b)^2 \leq 2(a^2+b^2)}{\leq} 2D(\hat{\delta}, \delta^*)^2 + 2D(\delta^*, \delta_0)^2 \\ &\leq 2c_\rho(\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*)) + 2c_\rho(\tilde{R}(\delta_0) - \tilde{R}(\delta^*)). \end{aligned} \tag{39}$$

Putting the results together, this produces an implicit equation for $\tilde{R}(\hat{\delta}) - R(\delta^*)$:

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq 2c_\rho C^{-1}\{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*)\} + [1 + 2c_\rho C^{-1}]\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + C^{-1}r^2$$

Solving for $\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*)$ yields

$$\begin{aligned} \tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) &\leq \frac{1 + 2c_\rho C^{-1}}{1 - 2c_\rho C^{-1}}\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \frac{C^{-1}}{1 - 2c_\rho C^{-1}} r^2 \\ &\overset{\text{Def. } \rho_0,\, C = c_\rho N \text{ with } N \text{ large enough}}{=} 2 \inf_{\delta \in B(\rho)} \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + \frac{c_\rho^{-1} N^{-1}}{1 - N^{-1}} r^2 \end{aligned} \tag{40}$$

Caution: r can not be chosen arbitrarily! It has to have a specific size so that $\mathbb{P}(A^c)$ is small.
The main part of the proof is to derive an upper bound for $\mathbb{P}(A^c)$. This is done in three steps:

(a) Find an upper bound of $\mathbb{E}|Z_{r,\rho}(\delta_0)|$ (*only in this calculation* the specific space over which the supremum is taken plays a role)

(b) Find an upper bound of $\mathbb{E}V_{r,\rho}(\delta_0)$ by using the so-called 'peeling device'

(c) Derivation of a concentration inequality for $V_{r,\rho}(\delta_0)$ by using a Talagrand-type inequality.

4.7.1 Step (a)

We first need a suitable orthonormal basis (ONB) of $\mathcal{H}_K$ which allows to calculate variances of $g(X)$, $g \in \mathcal{H}_K$ in a convenient way. Such an ONB is introduced by the following lemma.

Lemma 4.26 (Calculation of the variance)**.** There exists an ONB $(\psi_j)_{j\in\mathbb{N}}$ of $\mathcal{H}_K$ with the following property: For $g \in \mathcal{H}_K$, it holds that $\mathbb{E}[g(X)^2] = \sum_{j=1}^{\infty} \gamma_j \langle g, \psi_j\rangle_K^2$.

Lemma 4.27. For all $r, \rho > 0$ and $\delta_0 \in B(\rho)$, it holds that

$$\mathbb{E}|Z_{r,\rho}(\delta_0)| \leq \frac{4}{\sqrt{n}} \inf_{a\in\mathbb{N}} \Big(\sqrt{a}r + 2\rho\sqrt{\sum_{j>a}\gamma_j}\Big) =: \phi_\rho(r^2). \tag{41}$$

Proof. $\tilde{L}(y,s) = (1-ys)_+$ is Lipschitz continuous with constant $\ell = 1$. Lemma 3.11 implies that

$$\mathbb{E}|Z_{r,\rho}(\delta_0)| \leq 4 \cdot \mathbb{E} \sup_{\delta\in B(\rho), D(\delta,\delta_0)\leq r} \Big|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \big\{\delta(X_i) - \delta_0(X_i)\big\}\Big|$$

with independent Rademacher variables ε_i, $i = 1, ..., n$.
$\delta, \delta_0 \in B(\rho) \Rightarrow \|\delta - \delta_0\|_K \leq 2\rho$. Defining $g := \delta - \delta_0$, it follows that

$$\mathbb{E}|Z_{r,\rho}(\delta_0)| \leq 4 \cdot \mathbb{E} \sup_{g\in B(2\rho), \mathbb{E}[g(X)^2]\leq r^2} \Big|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(X_i)\Big|.$$

We now make use of the ONB from Lemma 4.26. Each $g \in \mathcal{H}_K$ has a representation $g_\alpha = \sum_{j=1}^\infty \alpha_j \psi_j$ with

$$\|g\|_K^2 \overset{\psi_j \text{ ONB}}{=} \sum_{j=1}^\infty \alpha_j^2, \qquad \mathbb{E}[g(X)^2] = \sum_{j=1}^\infty \gamma_j \alpha_j^2.$$

Let

$$\Gamma(\rho, r) := \Big\{ \alpha \in \ell^2 : \sum_{j=1}^\infty \alpha_j^2 \le (2R)^2, \sum_{j=1}^\infty \gamma_j \alpha_j^2 \le r^2 \Big\}.$$

Then it holds that

$$\mathbb{E} Z_{r,\rho}(\delta_0) \le \frac{4}{n} \mathbb{E} \sup_{\alpha \in \Gamma(\rho,r)} \Big| \sum_{i=1}^n \varepsilon_i g_\alpha(X_i) \Big|. \tag{42}$$

Now the goal is to eliminate the supremum with respect to α by using pure analytical methods. In the following, a parameter $a \in \mathbb{N}$ is introduced over which we can optimize later on. *The introduction of a is not necessary but yields better upper bounds!* By the Cauchy-Schwarz inequality (CSI) $\sum_j a_j b_j \le (\sum_j a_j^2)^{1/2} (\sum_j b_j^2)^{1/2}$, we have

$$\begin{aligned}
\Big| \sum_{i=1}^n \varepsilon_i g_\alpha(X_i) \Big| \quad &\le \quad \Big| \sum_{j \le a} \alpha_j \sqrt{\gamma_j} \sum_{i=1}^n \varepsilon_i \frac{\psi_j(X_i)}{\sqrt{\gamma_j}} \Big| + \Big| \sum_{j>a} \alpha_j \sum_{i=1}^n \varepsilon_i \psi_j(X_i) \Big| \\
&\overset{CSI}{\le} \quad \Big(\sum_{j \le a} \alpha_j^2 \gamma_j \Big)^{1/2} \Big(\sum_{j \le a} \Big(\sum_{i=1}^n \varepsilon_i \frac{\psi_j(X_i)}{\sqrt{\gamma_j}} \Big)^2 \Big)^{1/2} \\
&\qquad\qquad + \Big(\sum_{j>a} \alpha_j^2 \Big)^{1/2} \Big(\sum_{j>a} \Big(\sum_{i=1}^n \varepsilon_i \psi_j(X_i) \Big)^2 \Big)^{1/2} \\
&\overset{\alpha \in \Gamma(\rho,r)}{\le} \quad \sqrt{r} \Big(\sum_{j \le a} \Big(\sum_{i=1}^n \varepsilon_i \frac{\psi_j(X_i)}{\sqrt{\gamma_j}} \Big)^2 \Big)^{1/2} + 2\rho \Big(\sum_{j>a} \Big(\sum_{i=1}^n \varepsilon_i \psi_j(X_i) \Big)^2 \Big)^{1/2}
\end{aligned} \tag{43}$$

The right hand side does no longer depend on α. Therefore, we can upper bound the expectation without taking care of the supremum with respect to α. Hölder's inequality $\mathbb{E}[Z^{1/2}] \le \mathbb{E}[Z]^{1/2}$ for random variables $Z \ge 0$ yields

$$\begin{aligned}
\mathbb{E} \Big(\sum_{j \le a} \Big(\sum_{i=1}^n \varepsilon_i \frac{\psi_j(X_i)}{\sqrt{\gamma_j}} \Big)^2 \Big)^{1/2} \quad &\le \quad \Big(\sum_{j \le a} \mathbb{E} \Big[\mathbb{E} \Big[\Big(\sum_{i=1}^n \varepsilon_i \frac{\psi_j(X_i)}{\sqrt{\gamma_j}} \Big)^2 \Big| X_1, ..., X_n \Big] \Big] \Big)^{1/2} \\
&= \quad \Big(\sum_{j \le a} \mathbb{E} \Big[\sum_{i=1}^n \frac{\psi_j(X_i)^2}{\gamma_j} \Big] \Big)^{1/2} \overset{\text{Lemma 4.26: } \mathbb{E}[\psi_j(X)^2] = \gamma_j}{=} \sqrt{na}.
\end{aligned}$$

We use the same technique for the second term in (43) (then only γ_j is left). Plugging in these upper bounds in (43) and afterwards in (42) yields the assertion,

$$\mathbb{E} Z_{r,\rho}(\delta_0) \le \frac{4}{n} \mathbb{E} \sup_{\alpha \in \Gamma(\rho,r)} \Big| \sum_{i=1}^n \varepsilon_i g_\alpha(X_i) \Big| \le \frac{4}{\sqrt{n}} \Big(\sqrt{a} r + 2\rho \Big(\sum_{j>a} \gamma_j \Big)^{1/2} \Big).$$

□

4.7.2 Step (b)

Lemma 4.28. For all $r, \rho > 0$ and $\delta_0 \in B(\rho)$, it holds that $\mathbb{E}V_{r,\rho}(\delta_0) \leq 5\frac{\phi_\rho(r^2)}{r^2}$.

Proof. Abbreviate $\Delta_n(\delta) := \tilde{R}(\delta) - \tilde{R}_n(\delta) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0))$. Then for each $x > 1$, it holds that

$$\begin{aligned} V_{r,\rho}(\delta_0) &= \sup_{\delta \in B(\rho)} \frac{\Delta_n(\delta)}{D(\delta,\delta_0)^2 + r^2} \\ &\leq \sup_{\delta \in B(\rho), D(\delta,\delta_0) \leq r} \underbrace{\frac{|\Delta_n(\delta)|}{D(\delta,\delta_0)^2 + r^2}}_{\leq \frac{1}{r^2}|\Delta_n(\delta)|} + \sum_{k \geq 0} \sup_{\delta \in B(\rho): rx^k \leq D(\delta,\delta_0) \leq rx^{k+1}} \underbrace{\frac{|\Delta_n(\delta)|}{D(\delta,\delta_0)^2 + r^2}}_{\leq \frac{|\Delta_n(\delta)|}{r^2 x^{2k} + r^2}} \\ &\leq \frac{1}{r^2}\Big(\sup_{\delta \in B(\rho), D(\delta,\delta_0) \leq r} |\Delta_n(\delta)| + \sum_{k \geq 0} \frac{1}{1 + x^{2k}} \sup_{\delta \in B(\rho): D(\delta,\delta_0) \leq rx^{k+1}} |\Delta_n(\delta)| \Big). \end{aligned}$$

Note that $\phi_\rho(rz) \leq \phi_\rho(r) z^{1/2}$ for $z \geq 1$. This implies

$$\begin{aligned} \mathbb{E}V_{r,\rho}(\delta_0) &\overset{\text{Lemma 4.27}}{\leq} \frac{1}{r^2}\Big(\phi_\rho(r^2) + \sum_{k \geq 0} \frac{\phi_\rho(r^2 x^{2(k+1)})}{1 + x^k} \Big) \\ &\leq \frac{\phi_\rho(r^2)}{r^2}\Big(1 + \underbrace{\sum_{k \geq 0} \frac{x^{k+1}}{1 + x^{2k}}}_{\leq x \cdot \sum_{k \geq 0} x^{-k} = \frac{x^2}{x-1}} \Big) \leq 5\frac{\phi_\rho(r^2)}{r^2}. \end{aligned}$$

The last step is obtained by choosing $x = 2$. □

4.7.3 Step (c)

To derive a concentration inequality for $V_{r,\rho}(\delta_0)$, we use a so-called Talagrand or Bousquet-type inequality (cf. [4], Theorem 2.3).

Theorem 4.29. Let $\mathcal{F} := \{f : \mathcal{W} \to \mathbb{R} \text{ measurable}\}$ be a countable set of functions on $\mathcal{W} \subset \mathbb{R}^d$. Let W_i, $i = 1, ..., n$ be i.i.d. random variables with values in $\mathcal{W}$. Suppose that there exist $\sigma^2, M \in (0, \infty)$ such that $\mathbb{E}f(W) = 0$, $\sup_{f \in \mathcal{F}} \mathrm{Var}(f(W)) \leq \sigma^2$ and

$\sup_{f\in\mathcal{F}}\|f\|_\infty \le M$. Let $Z := \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(W_i)$. Put $v := n\sigma^2 + 2M\mathbb{E}Z$. Then it holds for $t > 0$ that

$$\mathbb{P}(Z \ge \mathbb{E}Z + \sqrt{2tv} + \frac{tM}{3}) \le e^{-t}.$$

Let $\alpha > 0$. Due to $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ und $2\sqrt{ab} \le \alpha a + \frac{b}{\alpha}$, the right hand side can be upper bounded as follows (motivation for this new bound is to eliminate v)

$$\mathbb{E}Z + \sqrt{2tv} + \frac{tM}{3} \quad \le \quad (1+\alpha)\mathbb{E}Z + \sqrt{2tn}\cdot\sigma + (\frac{1}{\alpha} + \frac{1}{3})tM.$$

Lemma 4.30. Let $r, \rho > 0$ and $\delta_0 \in B(\rho)$. Then, for all $t \ge 0$ it holds that

$$\mathbb{P}\Big(V_{r,\rho}(\delta_0) \ge 6\frac{\phi_\rho(r^2)}{r^2} + \sqrt{\frac{2t}{nr^2}} + \frac{22(\rho+1)t}{nr^2}\Big) \le e^{-t}.$$

Proof. We apply Theorem 4.29 to the space $\mathcal{W} = \mathcal{X}\times\{-1,+1\}$, the random variables $W_i = (X_i, Y_i)$, $i = 1, ..., n$ and the class of functions

$$\mathcal{F} = \Big\{f_\delta(x,y) := \frac{1}{n}\frac{\tilde{R}(\delta) - \tilde{L}(y,\delta(x)) - (\tilde{R}(\delta_0) - \tilde{L}(y,\delta_0(x)))}{D(\delta,\delta_0)^2 + r^2} : \delta \in B(\rho)\Big\},$$

since

$$V_{r,\rho}(\delta_0) = \sup_{f\in\mathcal{F}}\sum_{i=1}^n f(X_i, Y_i).$$

Lemma 4.24(ii) implies that $|\tilde{L}(y,\delta(x))| \le 1 + |\delta(x)| \le 1 + \rho$. Thus, it holds that

$$\|f_\delta\|_\infty \le \frac{4(\rho+1)}{nr^2},$$

and (due to $|\tilde{L}(y,s) - \tilde{L}(y,s')| \le |s - s'|$)

$$\mathrm{Var}(f_\delta(X,Y)) = \frac{\mathbb{E}[(\tilde{L}(Y,\delta(X)) - \tilde{L}(Y,\delta(X)))^2]}{n^2[D(\delta,\delta_0)^2 + r^2]^2} \le \frac{D(\delta,\delta_0)^2}{n^2[D(\delta,\delta_0)^2 + r^2]^2} \overset{(a+b)^2\ge 4ab}{\le} \frac{1}{n^2r^2}.$$

Theorerm 4.29 and the upper bound mentioned directly afterwards yields

$$\mathbb{P}\Big(V_{r,\rho}(\delta_0) \ge (1+\alpha)\mathbb{E}V_{r,\rho}(\delta_0) + \sqrt{\frac{2t}{nr}} + 4(\frac{1}{\alpha} + \frac{1}{3})(\rho+1)\frac{t}{nr}\Big) \le e^{-t}.$$

Caution: Note that $\mathcal{F}$ is not countable here! We can still apply the theorem since $\mathcal{H}_K$ is separable (this is left as an exercise). The assertion now follows by choosing $\alpha = \frac{1}{5}$ and using the result of Lemma 4.28. □

4.7.4 Selection of r

The following lemma is pure analytical statement (see the appendix for a proof).

Lemma 4.31. The equation $\phi_\rho(z) = \frac{z}{c_\rho}$ has exactly one solution $z^*_{\rho,c_\rho} \in (0,\infty)$. This solution satisfies

$$z^*_{\rho,c_\rho} \leq 16c_\rho^2 \cdot \gamma(n),$$

where $\gamma(n)$ is from Theorem 4.22. Furthermore, it holds that

$$\frac{z}{c_\rho} \geq \phi_\rho(z) \quad \Longleftrightarrow \quad z \geq z^*_{\rho,c_\rho}. \tag{44}$$

Putting our preliminary results together, we obtain the following theorem.

Theorem 4.32. Suppose that the assumptions of Theorem 4.22 are satisfied. Let $\rho \geq 1$, and let $\hat{\delta}_n^{fix}$ be from (36). Then there exists a universal constant $c > 0$ such that for all $t \geq 0$, it holds that

$$\mathbb{P}\Big(\tilde{R}(\hat{\delta}_n^{fix}) - \tilde{R}(\delta^*) \geq 2 \inf_{\delta \in B(\rho)} \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + c \cdot \{c_\rho \gamma(n) + \rho \cdot \frac{t}{n}\}\Big) \leq e^{-t}.$$

Proof. From equation (40) we obtain that on $A = \{V_{r,\rho}(\delta_0) \leq \frac{1}{c_\rho N}\}$, it holds for N large enough (N had to be chosen large enough to obtain the '2' in front of the infimum) that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq 2 \inf_{\delta \in B(\rho)} \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + \frac{c_\rho^{-1} N^{-1}}{1 - N^{-1}} r^2. \tag{45}$$

It is left to choose r^2 to provide a meaningful bound of the probability of A^c. Lemma 4.30 implies that if

$$\frac{1}{N} \geq 6c_\rho \frac{\phi_\rho(r^2)}{r^2} + \sqrt{\frac{2c_\rho t}{nr^2}} + \frac{22c_\rho(\rho+1)t}{nr^2},$$

then it holds that $\mathbb{P}(A^c) \leq e^{-t}$. The above inequality is satisfied if

$$\begin{aligned} &\frac{1}{3N} \geq 6c_\rho \frac{\phi_\rho(r^2)}{r^2}, \quad \frac{1}{3N} \geq \sqrt{\frac{2c_\rho t}{nr^2}}, \quad \frac{1}{3N} \geq \frac{22c_\rho(\rho+1)t}{nr^2} \\ \Leftarrow \quad & r^2 \geq z^*_{\rho,18Nc_\rho}, \quad r^2 \geq (3N)^2 \cdot \frac{2c_\rho t}{n}, \quad r^2 \geq 66Nc_\rho(\rho+1)\frac{t}{n}, \end{aligned}$$

where we have used Lemma 4.31 for the first inequality. Therefore, we may choose any

$$r^2 \geq \max\Big\{z^*_{\rho,18Nc_\rho}, (3N)^2 \cdot \frac{2c_\rho t}{n}, 66Nc_\rho(\rho+1)\frac{t}{n}\Big\},$$

e.g. (cf. Lemma 4.31 for the first term):

$$r = c(N) \cdot c_\rho\big\{c_\rho\gamma(n) + \rho \cdot \frac{t}{n}\big\} \tag{46}$$

with suitable $c(N) > 0$ which is a universal constant only depending on N. Plugging this into (45) yields the assertion with the universal constant $c := \frac{c(N)N^{-1}}{1+N}$. □

4.7.5 Discussion of the complex problem

We now provide the proof of Theorem 4.22 which discusses the estimator

$$\hat{\delta} \in \operatorname*{arg\,min}_{\delta\in\mathcal{H}_K}\big\{\tilde{R}_n(\delta) + \lambda \cdot \|\delta\|_K^2\big\}.$$

The main difference to $\hat{\delta}_n^{fix}$ is that instead of optimizing over $\delta \in B(\rho)$, we now use a penalization term $\lambda\|\delta\|_K^2$ and $\hat{\delta}_n \in \mathcal{H}_K$ has not known apriori to have a bounded $\|\cdot\|_K$-norm.

Proof of Theorem 4.22. We start similar to the proof of the simplified problem. We choose

$$\delta_0 \in \operatorname*{arg\,min}_{\delta\in\mathcal{H}_K}\big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda\|\delta\|_K^2\big\}.$$

As before, it holds that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) = \big\{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0)\big\} + \big\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\big\}. \tag{47}$$

With a suitable chosen $\tilde{\rho} > 0$ (which is determined below), the first term can be upper bounded as follows:

$$\begin{aligned}\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) &= \underbrace{\{\tilde{R}_n(\hat{\delta}) - \tilde{R}_n(\delta_0)\}}_{\leq\lambda\|\delta_0\|_K^2 - \lambda\|\hat{\delta}\|_K^2} + \{\tilde{R}(\hat{\delta}) - \tilde{R}_n(\hat{\delta}) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0)\} \\ &\leq \{\lambda\|\delta_0\|_K^2 - \lambda\|\hat{\delta}\|_K^2\} + V_{r,\tilde{\rho}}(\delta_0) \cdot \{r^2 + D(\hat{\delta},\delta_0)^2\}. \end{aligned} \tag{48}$$

Caution: To ensure the last inequality, we need that $\tilde{\rho} \geq \|\hat{\delta}\|_K$ so that $\hat{\delta} \in B(\tilde{\rho})$ (cf. the Definition of $V_{r,\rho}(\delta_0)$). Moreover, we need that $\tilde{\rho} \geq \|\delta_0\|_K$ to ensure $\delta_0 \in B(\tilde{\rho})$. If these conditions are fulfilled, we can apply the concentration inequality from Lemma 4.30.

Definition of a 'nice' event A: In the proof of Theorem 4.32 the following was shown: For fixed $\rho > 0$, $\delta_0 \in B(\rho)$ and arbitrary $N \in \mathbb{N}$, the event

$$A_{\rho,t}(\rho_0) := \{V_{r^2_{\rho,t},\rho}(\delta_0) \leq \frac{1}{Nc_\rho}\} \tag{49}$$

with (cf. (46))

$$r^2_{\rho,t} = c(N) \cdot c_\rho \big\{ c_\rho \gamma(n) + \rho \cdot \frac{t}{n} \big\}$$

satisfies $\mathbb{P}(A_{\rho,t}(\rho_0)^c) \leq e^{-t}$.
Problem: We need to upper bound the probability of the event $A_{\tilde{\rho},t}(\delta_0)^c$, where $\tilde{\rho}$ is a *random variable* (since $\tilde{\rho}$ is determined by the size of $\|\hat{\delta}\|_K$).
We therefore have to ensure that the event $A_{\rho,t}(\delta_0)$ holds *uniformly* for all possible $\rho = \tilde{\rho}$. Thus, our nice event should have the form $A = \bigcap_{\rho>0} A_{\rho,t}(\delta_0)$. We can not use this event since we have no possibility to find a meaningful upper bound despite $\mathbb{P}(A^c) \leq \sum_{\rho>0} \mathbb{P}(A_{\rho,t}(\delta_0)^c)$ which is infinity.

In the following, we therefore use a discretization $\mathcal{R} \subset (0, \infty)$ of the space of possible norms $\|\hat{\delta}\|_K$ of $\hat{\delta}$, δ_0. Additionally, we search for upper bounds for $\|\hat{\delta}\|_K, \|\delta_0\|_K$. Then we define A as the intersection over all $A_{\rho,t}(\delta_0)$ with $\rho \in \mathcal{R}$. Of course, $\mathcal{R}$ should not be too coarse because then we obtain not so tight upper bounds.
We start by analyzing the maximal size of $\|\hat{\delta}\|_K$, $\|\delta_0\|_K$. Since $0 \in \mathcal{H}_K$, it holds that

$$\tilde{R}_n(\hat{\delta}) + \lambda \|\hat{\delta}\|^2_K \leq \tilde{R}_n(0) + \lambda \|0\|^2_K \leq 1. \tag{50}$$

$$\Rightarrow \lambda \|\hat{\delta}\|^2_K \leq 1 \Rightarrow \|\hat{\delta}\|_K \leq \lambda^{-1/2} \overset{\lambda \geq n^{-1}}{\leq} n^{1/2} \leq n. \ (*)$$

A similar approach for δ_0 implies that $\|\delta_0\|_K \leq n$. (*)

Therefore, define

$$\mathcal{R} = \{2^k : k \in \mathbb{N}, 0 \leq k \leq \lceil \log_2(n) \rceil\}$$

and put

$$\begin{aligned} \hat{\rho} &:= 2^{\hat{k}} \overset{(*)}{\in} \mathcal{R}, \qquad \hat{k} = \lceil \log_2(\|\hat{\delta}\|_K)_+ \rceil, \\ \rho_0 &:= 2^{k_0} \overset{(*)}{\in} \mathcal{R}, \qquad k_0 = \lceil \log_2(\|\delta_0\|_K)_+ \rceil, \end{aligned}$$

and

$$\tilde{\rho} := \max\{\hat{\rho}, \rho_0\}.$$

Then it holds that

$$\|\delta_0\|_K, \|\hat{\delta}\|_K \leq \hat{\rho},$$

and vice versa,

$$\hat{\rho} \leq \max\{2\|\hat{\delta}\|_K, 1\}, \qquad \rho_0 \leq \max\{2\|\delta_0\|_K, 1\}. \tag{51}$$

That is, $\hat{\rho}$ and ρ_0 are good approximations of $\|\hat{\delta}\|_K$, $\|\delta_0\|_K$ in the sense that we loose at most a factor 2.

If we would define $\tilde{A} := \bigcap_{\rho\in\mathcal{R},\rho\geq\rho_0} A_{\rho,t}(\delta_0)$, then we would obtain that

$$\mathbb{P}(\tilde{A}^c) \leq \sum_{\rho\in\mathcal{R},\rho\geq\rho_0} \mathbb{P}(A_{\rho,t}(\delta_0)^c) \leq \sum_{\rho\in\mathcal{R},\rho\geq\rho_0} e^{-t} \leq |\mathcal{R}|e^{-t},$$

that is, we have an additional factor $|\mathcal{R}|$ which we do not want to have. To obtain e^{-t} instead, we therefore put

$$A := \bigcap_{\rho\in\mathcal{R},\rho\geq\rho_0} A_{\rho,\tilde{t}}(\delta_0), \qquad \tilde{t} := t + \log|\mathcal{R}|.$$

Then we obtain

$$\mathbb{P}(A^c) \leq |\mathcal{R}|e^{-t-\log|\mathcal{R}|} = e^{-t}.$$

Derivation of the upper bound of the excess Bayes risk on A: On A, the results (50) and (49) imply

$$\begin{aligned}
&\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) \\
\leq\ & \{\lambda\|\delta_0\|_K^2 - \lambda\|\hat{\delta}\|_K^2\} + V_{r_{\tilde{\rho},\tilde{t}},\tilde{\rho}}(\delta_0)\cdot\{r_{\tilde{\rho},\tilde{t}}^2 + D(\hat{\delta},\delta_0)^2\} \\
\leq\ & \{\lambda\|\delta_0\|_K^2 - \lambda\|\hat{\delta}\|_K^2\} + \frac{1}{c_{\tilde{\rho}}N}\cdot\{ \underbrace{r_{\tilde{\rho},\tilde{t}}^2}_{=c(N)\cdot c_{\tilde{\rho}}(c_{\tilde{\rho}}\gamma(n)+\tilde{\rho}\cdot\frac{\tilde{t}}{n})} + \underbrace{D(\hat{\delta},\delta_0)^2}_{\overset{\text{wie in (39)}}{\leq} 2c_{\tilde{\rho}}\{\tilde{R}(\hat{\delta})-\tilde{R}(\delta^*)\}+2c_{\tilde{\rho}}\{\tilde{R}(\delta_0)-\tilde{R}(\delta^*)\}} \} \\
\leq\ & \{\lambda\|\delta_0\|_K^2 - \lambda\|\hat{\delta}\|_K^2\} + \frac{2c(N)}{N}\cdot\hat{\rho}\cdot(\gamma(n)+\frac{\tilde{t}}{n}) + \frac{2c(N)}{N}\cdot\rho_0\cdot(\gamma(n)+\frac{\tilde{t}}{n}) + \frac{4c(N)}{N}\frac{\gamma(n)}{\eta_0} \\
& + \frac{2}{N}\{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*)\} + \frac{2}{N}\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\}.
\end{aligned}$$

In the last step, we made use of the structure $c_\rho = 2(\frac{\rho}{\eta_1}+\frac{1}{\eta_0})$ and that $\tilde{\rho} = \max\{\hat{\rho},\rho_0\} \leq \hat{\rho}+\rho_0$. This was necessary to replace $\tilde{\rho}$ by terms involving $\hat{\rho},\rho_0$.
Using (47) and rearranging terms yields

$$\begin{aligned}
\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) \ \leq\ & \frac{1}{1-2N^{-1}}\Big[(1+\frac{2}{N})\{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) + \frac{4c(N)}{N}\frac{\gamma(n)}{\eta_0} \\
& + \{\lambda\|\delta_0\|_K^2 - \lambda\|\hat{\delta}\|_K^2\} + \frac{2c(N)}{N}\cdot\hat{\rho}\cdot(\frac{\gamma(n)}{\eta_1}+\frac{\tilde{t}}{n}) + \frac{2c(N)}{N}\cdot\rho_0\cdot(\frac{\gamma(n)}{\eta_1}+\frac{\tilde{t}}{n})\Big].
\end{aligned}$$

We now only have to eliminate the (non-deterministic) term involving $\hat{\rho}$ to obtain a deterministic convergence rate. Note that this term came into the upper bound due to

the statistical variation of $\{\tilde{R}(\hat{\delta}) - \tilde{R}_n(\hat{\delta}) - \tilde{R}(\delta_0) - \tilde{R}_n(\delta_0)\}$. The size of this term is controlled by the penalization term $\lambda\|\hat{\delta}\|_K^2$ in the minimization problem for $\hat{\delta}$. Now, we want to reflect this behavior in our proof (cp. the underlined terms above), that is, we want to use the term $\lambda\|\hat{\delta}\|_K^2$ to eliminate the term containing $\hat{\rho}$. To do so, we have to find a connection between $\hat{\rho}$ and $\|\hat{\delta}\|_K^2$.
From (51) and the elementary inequality $\max\{2a, 1\} \leq 2\max\{a, 1\} \leq 2\max\{a^2, 1\} \leq 2(a^2 + 1)$, we obtain

$$\hat{\rho} \leq 2(\|\hat{\delta}\|_K^2 + 1), \qquad \rho_0 \leq 2(\|\delta_0\|_K^2 + 1)$$

If

$$\lambda \geq \frac{4c(N)}{N}\left(\frac{\gamma(n)}{\eta_1} + \frac{\tilde{t}}{n}\right), \tag{52}$$

we therefore have

$$\begin{aligned}
\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) \;\leq\; & \frac{1}{1-2N^{-1}}\Big[(1+\frac{2}{N})\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \frac{4c(N)}{N}\frac{\gamma(n)}{\eta_0} \\
& -\lambda\|\hat{\delta}\|_K^2 + \frac{4c(N)}{N}\cdot(\|\hat{\delta}\|_K^2 + 1)\cdot\left(\frac{\gamma(n)}{\eta_1} + \frac{\tilde{t}}{n}\right) \\
& +\lambda\|\delta_0\|_K^2 + \frac{4c(N)}{N}\cdot(\|\delta_0\|_K^2 + 1)\cdot\left(\frac{\gamma(n)}{\eta_1} + \frac{\tilde{t}}{n}\right)\Big] \\
\leq\; & \frac{1}{1-2N^{-1}}\Big[(1+\frac{2}{N})\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + 2\lambda\|\delta_0\|_K^2 + 2\lambda + \underbrace{\frac{4c(N)}{N}\frac{\gamma(n)}{\eta_0}}_{\leq\lambda\cdot\frac{\eta_1}{\eta_0}}\Big].
\end{aligned}$$

Now, we choose N so large that $\frac{1+2N^{-1}}{1-2N^{-1}} \leq 2$ (that is, for instance, $N = 6$). Then it holds that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) \leq 2\underbrace{\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*) + 2\lambda\|\delta_0\|_K^2\}}_{\overset{\text{Def. } \delta_0}{=}\inf_{\delta\in\mathcal{H}_K}\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda\|\delta\|_K^2\}} + \underbrace{4}_{=:c_2}\lambda\left(\frac{\eta_1}{\eta_0} + 1\right).$$

Choice of λ: We simplify the right hand side of (52) as follows: It holds that $\log|\mathcal{R}| \leq \log(\lceil\log_2(n)\rceil + 1) \leq 2\log(\log_2(n) + 1)$, thus

$$\frac{4c(N)}{N}\left(\frac{\gamma(n)}{\eta_1} + \frac{\tilde{t}}{n}\right) \leq \underbrace{\frac{8c(N)}{N}}_{=:c_1}\left(\frac{\gamma(n)}{\eta_1} + \frac{\log(\log_2(n) + 1) + t}{n}\right).$$

The right hand side of the above inequality now coincides with the condition on λ in the theorem. This concludes the proof. □

4.8 Appendix

The proof of the following lemma is mainly based on functional analysis.

Proof of Lemma 4.26. Let

$$T : \mathcal{H}_K \to L^2(\mathbb{P}^X), \quad g \mapsto g$$

be the canonical injection and T^* its adjoint. Then it holds that

$$\begin{aligned}(T^*g)(x) &\overset{\text{Reproc. prop. Thm. 4.17}}{=} \langle K(x,\cdot), T^*g\rangle_{H_K} \overset{\text{Adj.}}{=} \langle TK(x,\cdot), g\rangle_{L^2(\mathbb{P}^X)} \\ &= \int K(x,x')g(x')d\mathbb{P}^X(x') = (T_K g)(x).\end{aligned}$$

It follows that $TT^* = T_K$. Now let $C := T^*T : \mathcal{H}_K \to \mathcal{H}_K$.
Functional analysis $\Rightarrow$ C has the same eigenvalues γ_i as $TT^* = T_K$.
Let $(\psi_j)_{j\in\mathbb{N}}$ be a basis consisting of eigenvectors of $C := T^*T : \mathcal{H}_K \to \mathcal{H}_K$ corresponding to these eigenvalues γ_j. For every $g \in \mathcal{H}_K$, it holds that

$$\begin{aligned}\mathbb{E}[g(X)^2] &\overset{T(g)=g}{=} \mathbb{E}[(Tg(X))^2] = \langle Tg, Tg\rangle_{L^2(\mathbb{P}^X)} \overset{\text{adjoint}}{=} \langle Cg, g\rangle_K \\ &\overset{\text{base representation } g=\sum_{j=1}^\infty \langle g,\psi_j\rangle\psi_j}{=} \sum_{j=1}^\infty \gamma_j \langle g, \psi_j\rangle_K^2\end{aligned}$$

□

Proof of Lemma 4.31. Let $a \in \mathbb{N}$ and

$$\phi_{\rho,a}(z) = \frac{4}{\sqrt{n}}\Big(\sqrt{az} + 2\rho\sqrt{\sum_{j>a}\lambda_j}\Big).$$

The equation $\phi_{\rho,a}(z) = \frac{z}{C_\rho}$ can be rearranged as follows:

$$z_\rho^*(a) = \frac{4c_\rho^2}{n}\Big(\sqrt{a} + \sqrt{a + 2\sqrt{n}\frac{\rho}{c_\rho}\sqrt{\sum_{j>a}\gamma_j}}\Big)^2 \overset{(x+y)^2\le 2x^2+2y^2}{\le} \frac{16c_\rho^2}{n}\Big(a + \sqrt{n}\frac{\rho}{c_\rho}\sqrt{\sum_{j>a}\gamma_j}\Big)^2.$$

Then it holds $z_\rho^* = z_\rho^*(a^*)$ for the minimizing value a^* in $\phi_\rho(z)$. For each $a \in \mathbb{N}$, we have

$$\frac{z_\rho^*}{C_\rho} = \phi_\rho(z_\rho^*) \le \phi_{\rho,a}(z_\rho^*),$$

that is, $z_\rho^* \le z_\rho^*(a)$. Thus,

$$z_\rho^* \le \inf_{a\in\mathbb{N}} z_\rho^*(a) \le \inf_{a\in\mathbb{N}} \frac{16c_\rho^2}{n}\Big(a + \sqrt{n}\frac{\rho}{c_\rho}\sqrt{\sum_{j>a}\gamma_j}\Big)^2 \overset{\frac{\rho}{c_\rho}\le\frac{\eta_1}{2}}{\le} 16c_\rho^2\gamma(n).$$

The remaining analytical properties of ϕ_ρ can be verified easily. □

4.9 Exercises

Task 11 (Discussion of Example 4.1: the naive classifier)**.** Suppose that we are given i.i.d. training samples (X_i, Y_i) of a classification problem with $\mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \{-1, +1\}$. Suppose that $\tilde{L}(y, s) = (y - s)^2$. We use the abbreviation

$$\Delta := \{\delta_\beta(x) := x^T\beta : \beta \in \mathbb{R}^d\}.$$

We propose the following algorithm for classification:

$$\hat{\beta} :\in \arg\min_{\beta\in\mathbb{R}^d} \tilde{R}_n(\delta_\beta), \quad \tilde{R}_n(\delta) := \frac{1}{n}\sum_{i=1}^n \tilde{L}(Y_i, \delta(X_i)),$$

and $\hat{f}_n(x) := \operatorname{sign}(\hat{\delta}_n(x))$, where $\hat{\delta}_n(x) := \delta_{\hat{\beta}}(x) = x^T\hat{\beta}$. We can interpret the above classifier as a simple linear regression applied to (X_i, Y_i).

1. Show that $\tilde{L}(y, s) = \phi(-ys)$ with suitable $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$.
2. Use Theorem 3.19 to derive a Bayes rule δ^* for the risk $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y, \delta(X))$.
3. Does the calibration condition hold?
4. Which model assumption should hold so that $\delta^* \in \Delta$? Which properties should $\mathcal{X}$ have such that this can be fulfilled?
5. Derive a risk transfer formula with Theorem 3.21.
6. Suppose that the noise condition with parameters $q \geq 0$, $C > 0$ hold. Derive the risk transfer formula from Theorem 3.24 in the above setting, in particular in the case $q = \infty$.
7. Let $X \sim N(\mu, I_{d\times d})$ and $\eta(x) = \mathbb{P}(Y = 1|X = x) = \frac{1}{2} + \frac{1}{2}x^T\beta^*$. Suppose that $\beta^* \in \mathbb{R}^d$ is chosen such that $\|\beta^*\|_2 = 2$ and $\mu^T\beta^* = 0$. Show that the noise condition is satisfied with $q = 1$.
   Hint: The distribution function $\Phi(t)$ of the standard normal distribution is concave for $t \geq 0$, thus it holds that $2\Phi(t) - 1 \leq 2\Phi'(0)t$.

Task 12 (Examples for calibration condition and risk transfer formulas)**.** Given is a classification problem with $\mathcal{X} \subset \mathbb{R}^d$, $Y_i \in \{-1, +1\}$. Let $\tilde{L}(y, s) = \phi(-ys)$, where we consider the following two functions $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$:

- $\phi(x) = \phi_{hinge}(x) = \max\{1 + x, 0\}$, the so-called hinge loss,
- $\phi(x) = \phi_{exp}(x) = e^x$, the so-called exponential loss.

1. Convince yourself that ϕ is non-decreasing, convex and satisfied $\phi(0) = 1$.
2. Use Theorem 3.19 to derive a Bayes rule δ^* for the risk $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y, \delta(X))$. *Hint for ϕ_{hinge}: Due to monotonicity reasons, $z \mapsto \Phi_\eta(z)$ has to attains its minimum for $z \in [-1, 1]$.*
3. Prove that the calibration condition is fulfilled in both cases.
4. Use Theorem 3.21 to derive a risk transfer formula in both cases.

Task 13 (Derivation of the Wolfe-Dual and the optimality conditions of the SVM)**.** Let $C > 0$. Let $\hat{\beta}_C$, $\hat{\beta}_{0,C}, \hat{\xi}$ be solutions of

$$\min_{\beta\in\mathbb{R}^k,\beta_0\in\mathbb{R},\xi\in\mathbb{R}^n} \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \qquad \text{s.t.} \qquad \forall i = 1, ..., n: \quad Y_i(X_i^T\beta + \beta_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0,$$

The optimization problem has the structure

$$\min_{\theta\in\mathbb{R}^r} F(\theta) \quad \text{s.t.} \quad G(\theta) \leq 0$$

with $\theta = (\beta, \beta_0, \xi)$ and

$$F(\theta) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i, \quad G(\theta) = \begin{pmatrix} \big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big)_{i=1,...,n} \\ -\xi \end{pmatrix}.$$

Now let $p = (\alpha, \gamma) \in \mathbb{R}^{2n}_{\geq 0}$, and define the Lagrange function $L(\theta, p) = F(\theta) + G(\theta)^T p$.

1. Show that $\nabla_\theta L(\theta, p) = 0$ is equivalent to

$$\beta = \sum_{i=1}^n \alpha_i Y_i X_i, \qquad \sum_{i=1}^n \alpha_i Y_i = 0, \qquad \forall i : C - \alpha_i - \gamma_i = 0.$$

2. Show that the Wolfe Dual $\sup_{\theta\in\mathbb{R}^r, p\in\mathbb{R}^k_{\geq 0}} L(\theta, p)$ under the constraint $\nabla_\theta L(\theta, p) = 0$ is equivalent to the optimization problem

$$\min_{\alpha\in\mathbb{R}^n} \left\{\frac{1}{2}\alpha^T Q\alpha - \mathbb{1}^T\alpha\right\} \qquad \text{s.t.} \qquad \mathbb{Y}^T\alpha = 0, \quad 0 \leq \alpha \leq C,$$

where $\mathbb{1} = (1, ..., 1)^T$, $\mathbb{Y} = (Y_1, ..., Y_n)^T$ and $Q := (Q_{ij})_{i,j=1,...,n}$ mit $Q_{ij} = Y_i Y_j X_i^T X_j$.

3. Conclude from the optimality conditions $\nabla_\theta L(\hat{\theta}, \hat{p}) = 0$, $G(\hat{\theta})^T \hat{p} = 0$, $\hat{p} \geq 0$ and $G(\hat{\theta}) \leq 0$ that the following connection is satisfied for the solution of the original optimization problem and the solutions $\hat{\alpha}$ of the Wolfe Dual:

$$\hat{\beta}_C = \sum_{i=1}^{n} \hat{\alpha}_i Y_i X_i, \qquad \hat{\beta}_{0,C} = Y_i - X_i^T \hat{\beta}_C \quad \text{with some } i \text{ with } 0 < \hat{\alpha}_i < C.$$

Task 14 (SVM: Kernels and their nonlinear transformations behind)**.** Let $K_p : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_p(x, x') := (1 + x^T x')^p$ be the polynomial kernel with degree $p \in \mathbb{N}$.

1. Let $d = 2$, $p = 2$. Find a mapping $h : \mathbb{R}^2 \to \mathbb{R}^6$ such that $K_p(x, x') = h(x)^T h(x')$.

2. Let p, d be arbitrary. Find a mapping $h : \mathbb{R}^d \to \mathbb{R}^m$ such that $K_p(x, x') = h(x)^T h(x')$. Provide an upper bound for $m \in \mathbb{N}$.
   Hint: Define $\tilde{x} := (1, x)$, $\tilde{x}' := (1, x')$ and write $K_p(x, x') = (\tilde{x}, \tilde{x}')^p$.

Let $K_\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, K_\gamma(x, x') = \exp(-\gamma \|x - x'\|_2^2)$ be the Gaussian kernel with descent γ.

(c) Let $d = 1$. Find a mapping $h : \mathbb{R} \to \ell^2$ such that $K_\gamma(x, x') = h(x)^T h(x')$.

(d) Let d be arbitrary. How does h look like?

Task 15 (Analysis of the convergence rates in Theorem 4.22)**.** Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel. In the oracle inequality in Theorem 4.22, the convergence rate of the estimation error is given by

$$\gamma(n) = \frac{1}{\sqrt{n}} \inf_{a \in \mathbb{N}} \Big\{ \frac{a}{\sqrt{n}} + \sqrt{\sum_{j > a} \gamma_j} \Big\}.$$

(here, we ignore η_1 for simplicity), where γ_j are the eigenvalues of the integral operator $T_K : L^2(\mathbb{P}^X) \to L^2(\mathbb{P}^X), (T_K g)(x) := \int K(x, x') g(x') \, \mathrm{d}\mathbb{P}^X(x')$. In this task we derive (non-optimal) upper bounds for different decay structures of γ_j.

1. Suppose that there exists $C > 0$ with $\gamma_j = 0$ for $j > C$. Show that $\gamma(n) \leq \frac{C}{n}$.

2. Suppose that there exists $C > 0$ with $\gamma_j \leq C j^{-\alpha}$ $(\alpha > 1)$. Show that $\gamma(n) \leq \tilde{c}_\alpha \cdot (\frac{1}{n} + C^{\frac{1}{\alpha+1}} n^{-\frac{\alpha}{\alpha+1}})$, where $\tilde{c}_\alpha$ is constant only depending on α.
   Hint: Use without proof that $\sum_{j>a} j^{-\alpha} \leq c_\alpha a^{-\alpha+1}$ with some $c_\alpha > 0$.

3. Suppose that there exists $C > 0$ with $\gamma_j \leq C \rho^j$ $(\rho \in (0, 1))$. Show that $\gamma(n) \leq c_\rho \cdot (\frac{1}{n} + 2\frac{\log(nC)}{nC^{1/2}})$, where c_ρ is a constant which only depends on ρ.

If $\mathcal{X} = [0,1]^d$ and $\mathbb{P}^X = U[0,1]^d$ is the uniform distribution on $[0,1]^d$, then we have $(T_K g)(x) = \int K(x,x')g(x')dx'$. Let $K(x,x') = h(x)^T h(x')$ with suitably chosen $h : \mathbb{R}^d \to \ell^2$.

(d) Show that if $(h_k)_{k\in\mathbb{N}}$ is orthogonal in $L^2(\lambda)$, then we have $\gamma_k = \int h_k(x)^2 dx$.

(e) Argue that the polynomial kernel $K_p(x,x')$ satisfies: $\gamma(n) \le (d+1)^p$.
Hint: Use (a),(d) and Task 14(b).

Task 16 (Margin property of the SVM for the general noise condition)**.** To prove the quadratric margin property of the SVM, we have assumed in Lemma 4.24 that there exists some $\eta_0 > 0$ such that $|\eta(x) - \frac{1}{2}| \ge \eta_0$ for all $s \in \mathcal{X}$, that is, the noise condition is fulfilled with $q = \infty$. In the following we suppose instead that the noise condition is only fulfilled with some $q \in (1,\infty)$, that is, there exists some $C > 0$ such that

$$\forall t > 0 : \quad \mathbb{P}(|\eta(X) - \frac{1}{2}| \le t) \le Ct^q.$$

Let $\delta^*(x)$ denote the Bayes rule with respect to $\tilde{L}(y,s) = (1-ys)_+$. Show that for any $\delta : \mathcal{X} \to \mathbb{R}$ with $\|\delta\|_\infty \le \rho$ it holds that

$$\mathbb{E}[(\delta(X) - \delta^*(X))^2] \le \tilde{c}_\rho \cdot \left\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\right\}^{\frac{q}{q+1}},$$

where $\tilde{c}_\rho > 0$ is a constant which only depends on ρ, η_0, η_1, C.
Hints:

- It was already shown in Lemma 4.24 that $A(x) = \frac{(\delta(x)-\delta^*(x))^2}{\mathbb{E}[\tilde{L}(Y,\delta(X))-\tilde{L}(Y,\delta^*(X))|X=x]} \le c_\rho(\eta_0) = 2(\frac{\rho}{\eta_1} + \frac{1}{\eta_0})$.
- Introduce $1 = \mathbb{1}_{\{|\eta(X)-\frac{1}{2}|>t\}} + \mathbb{1}_{\{|\eta(X)-\frac{1}{2}|\le t\}}$ in $\mathbb{E}[(\delta(X) - \delta^*(X))^2]$ and choose a suitable t (cf. the proof of the risk transfer formula, Theorem 3.24).
- Finally, make use of the inequality $\tilde{R}(\delta) - \tilde{R}(\delta^*) \le (\rho+1)^{\frac{1}{q}}(\tilde{R}(\delta) - \tilde{R}(\delta^*))^{\frac{q}{q+1}}$.

Task 17 (An alternative way to prove the convergence rate of the simplified SVM algorithm)**.** Let $\mathcal{H}_K$ be the RKHS corresponding to some Mercer kernel K. In this task we consider a fixed $\rho > 0$ and the algorithm

$$\hat{\delta} \in \operatorname*{arg\,min}_{\delta\in B(\rho)} \tilde{R}_n(\delta), \qquad B(\rho) := \{\delta \in \mathcal{H}_K : \|\delta\|_K \le \rho\},$$

where $\tilde{R}_n(\delta) = \frac{1}{n}\sum_{i=1}^n \tilde{L}(Y_i, \delta(X_i))$ and $\tilde{L}(y,s) = (1-ys)_+$ is the hinge loss. Let $D(\delta,\delta_0) := \mathbb{E}[(\delta(X) - \delta^*(X))^2]^{1/2}$. Let $\delta_0 \in \arg\min_{\delta\in B(\rho)} \tilde{R}(\delta)$, where $\tilde{R}(\delta) := \mathbb{E}\tilde{L}(Y,\delta(X))$.

1. Let $r > 0$ and $\tilde{\delta} := T\hat{\delta} + (1-T)\delta_0$ with $T := \frac{r}{r+D(\hat{\delta},\delta_0)}$. Show that

$$\tilde{\delta} \in B(\rho), \qquad D(\tilde{\delta}, \delta_0) \leq r.$$

2. Conclude that

$$\tilde{R}(\tilde{\delta}) - \tilde{R}(\delta^*) \leq \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + Z_r,$$

where $Z_r := \sup_{\delta \in B(\rho), D(\delta,\delta_0) \leq r} \{\tilde{R}(\delta) - \tilde{R}_n(\delta) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0))\}$.
Hint: Use that $\tilde{L}(y,s)$ is convex in s.

3. Define $A := \{Z_r \leq \frac{1}{8c_\rho}(\frac{r}{2})^2\}$. Show that on the event A and under the assumption $\tilde{R}(\delta_0) - \tilde{R}(\delta^*) \leq \frac{1}{8c_\rho}(\frac{r}{2})^2$, it holds that

$$D(\tilde{\delta}, \delta^*) \leq \frac{r}{4}.$$

Hint: You may use the quadratic margin property $D(\delta,\delta^*)^2 \leq c_\rho\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}$.

4. Suppose the conditions from (c). Show that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \frac{1}{8c_\rho}(\frac{r}{2})^2.$$

5. It was shown already in Lemma 4.27 that $\mathbb{E}|Z_r| \leq \phi_\rho(r^2)$. Use Talagrand's inequality (Theorem 4.29) and the conclusion below with $\alpha = 1$ to show that for each $t > 0$, it holds that

$$\mathbb{P}\Big(Z_r \geq 2\phi_\rho(r^2) + \sqrt{\frac{2t}{n}} \cdot r + \frac{16(\rho+1)}{3} \cdot \frac{t}{n}\Big) \leq e^{-t}.$$

Hint: For $\delta \in B(\rho)$, it holds that $\|\delta\|_\infty \leq \rho$.

6. Let

$$r = \max\Big\{4{\cdot}192c_\rho\gamma(n)^{1/2}, \quad 96c_\rho\sqrt{\frac{2t}{n}}, \quad 16\sqrt{c_\rho(\rho+1)}{\cdot}\sqrt{\frac{2t}{n}}, \quad 2\big(8c_\rho\{\tilde{R}(\delta_0){-}\tilde{R}(\delta^*)\}\big)^{1/2}\Big\}.$$

Show that $\mathbb{P}(A^c) \leq e^{-t}$ and $\tilde{R}(\delta_0) - \tilde{R}(\delta^*) \leq \frac{1}{2c_\rho}(\frac{r}{2})^2$.
Hints:

- *Upper bound all 3 terms in the probability in (e) individually by $\frac{1}{24c_\rho}(\frac{r}{2})^2$.*
- *The inequality $\frac{r^2}{192c_\rho} \geq \phi_\rho(r^2)$ is satisfied if $r^2 \geq 16(192c_\rho)^2\gamma(n)$, cf. Lemma 4.31.*

7. Conclude that with probability $\geq 1 - e^{-t}$, it holds that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq 2 \inf_{\delta \in B(\rho)} \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + c \cdot \big\{c_\rho \gamma(n) + (c_\rho + \rho + 1) \cdot \frac{t}{n}\big\},$$

where c is some universal constant.
Hint: For $a, b \geq 0$ it holds that $\max\{a, b\} \leq a + b$.

8. Discuss how one has to modify the proof so that the following assertion is obtained: With probability $\geq 1 - e^{-t}$, it holds that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq (1 + \varepsilon)\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + c(\varepsilon) \cdot \big\{c_\rho \gamma(n) + (c_\rho + \rho + 1) \cdot \frac{t}{n}\big\},$$

where $\varepsilon > 0$ is arbitrarily small and $c(\varepsilon) > 0$ is some constant only dependent on ε.

9. Discussion: The proof assumes (basically without any justification) that there exists some $\delta_0 \in B(\rho)$ such that $\tilde{R}(\delta_0) = \inf_{\delta \in B(\rho)} \tilde{R}(\delta)$. In general, one can only hope that there exists some sequence $(\delta_m)_{m \in \mathbb{N}} \subset B(\rho)$ with $\tilde{R}(\delta_m) \downarrow \inf_{\delta \in B(\rho)} \tilde{R}(\delta)$. Therefore, the proof has to be performed with each δ_m, $m \in \mathbb{N}$ (instead of a fixed δ_0). Summarize shortly where one has to modify the proof.

Task 18 (Discussion: Bernstein's inequality and Talagrand's inequality, separability)**.** In this task we discuss the relationship between Bernstein's and Talagrand's inequality. Let $X_i \in \mathcal{X} \subset \mathbb{R}^d$, $i = 1, ..., n$ be i.i.d. and $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$ be countable. Suppose that $\mathbb{E}f(X) = 0$, $\sup_{f \in \mathcal{F}} \mathrm{Var}(f(X)) \leq \sigma^2$ and $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq M$. Let $v := n\sigma^2 + 2M\mathbb{E}Z$. Then for $Z := \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)$ and $t > 0$, it holds that

$$\begin{aligned} \mathbb{P}(Z \geq \mathbb{E}Z + \sqrt{2tv} + \frac{tM}{3}) &\leq e^{-t} \quad \text{(Talagrand)}, \\ \mathbb{P}(\sum_{i=1}^n f(X_i) \geq \sqrt{2tn}\sigma + \frac{tM}{3}) &\leq e^{-t} \quad \text{(Bernstein)}. \end{aligned}$$

1. Show the statement after Theorem 4.29 in detail: For $\alpha > 0$, Talagrand's inequality implies

$$\mathbb{P}(Z \geq (1 + \alpha)\mathbb{E}Z + \sqrt{2tn}\sigma + (\frac{1}{\alpha} + \frac{1}{3})tM) \leq e^{-t}$$

2. Compare the statements from Talagrand's and Bernstein's inequality. Which term is needed additionally to explain the variation of $\sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)$ in opposite to $\sum_{i=1}^n f(X_i)$ for fixed $f \in \mathcal{F}$? Interprete this result.

3. Suppose that $\mathcal{F}$ is not countable, but (as a subset of a normed subspace $(\mathcal{E}, \|\cdot\|)$) separable, that is, there exists a dense subset $\mathcal{F}_{sep} \subset \mathcal{F}$. Moreover, suppose that there exists some constant $C > 0$ such that for all $x \in \mathcal{X}$ and $f, g \in \mathcal{F}$, it holds that $|f(x) - g(x)| \leq C\|f - g\|$. Show that it holds that

$$\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) = \sup_{g \in \mathcal{F}_{sep}} \sum_{i=1}^{n} g(X_i).$$

4. Application in the SVM proof (Lemma 4.30): As a Hilbert space with countable orthonormal basis, $\mathcal{H}_K$ is separable. Show with (c) that Talagrand's inequality can be applied to the set $\mathcal{F} \subset \{f_\delta(x, y) = \frac{(\mathbb{E}\tilde{L}(Y,\delta(X)) - \mathbb{E}\tilde{L}(Y,\delta_0(X))) - (\tilde{L}(y,\delta(x)) - \tilde{L}(y,\delta_0(x)))}{n(D(\delta,\delta_0)^2 + r^2)} : \delta \in B(\rho)\}$.

Task 19 (Expected excess Bayes risk instead of large deviation inequality)**.** In this task, we derive upper bounds for expectations based on large deviations inequalities. This is based on the equality

$$\mathbb{E}Z = \int_0^\infty \mathbb{P}(Z \geq x) dx$$

for random variables $Z \geq 0$.

1. Suppose that $\mathbb{P}(Z \geq A + B \cdot t) \leq g(t)$, where $\int_0^\infty g(t)dt < \infty$. Show that

$$\mathbb{E}Z \leq A + B \cdot \int g(t)dt.$$

2. Show that the estimator of Task 18 satisfies

$$\mathbb{E}\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq 2 \inf_{\delta \in B(\rho)} \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + c \cdot \{c_\rho \gamma(n) + (c_\rho + \rho + 1) \cdot \frac{1}{n}\}.$$

5 A short excursus: kernel nonparametric statistics

This chapter gives a (very) brief introduction to standard kernel-based estimators from nonparametric statistics and their convergence rates. It is important to have these results in mind so that the improvements obtained by machine learning algorithms in the next chapters can be appreciated. The results are (partly) discussed in the exercises.
Contrary to the chapters before, we do *not* assume any explicit parametric structure on f^* or on the optimal discriminant function δ^*. Instead, we ask for some structural assumptions. To do so, we define the class

$$\mathcal{F}(L) := \{g : \mathcal{X} \to \mathbb{R} \,|\, \forall x, x' \in \mathcal{X} : |g(x) - g(x')| \leq L \cdot \|x - x'\|_\infty\} \tag{53}$$

of Lipschitz continuous functions with Lipschitz constant L with respect to the maximum norm $\|\cdot\|_\infty$ on $\mathbb{R}^d$.

5.1 The kernel regression estimator

We first consider regression problems, that is, $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$ and $L(y, s) = (y - s)^2$ with Bayes rule $f^*(x) = \mathbb{E}[Y|X = x]$. We work under the following assumption.

Definition 5.1 (Model assumption: nonparametric regression)**.** Suppose that $f^* \in \mathcal{F}(L)$ with

$$Y = f^*(X) + \varepsilon,$$

where $\mathbb{E}[\varepsilon|X] = 0$, $\mathbb{E}[\varepsilon^2|X] = \sigma^2$.

Caution: In the above model we have $\mathbb{E}[Y|X = x] = f^*(x)$, that is, Definition 5.1 is a direct assumption on the structure of the Bayes rule f^*.
As in chapter 2, we obtain the following representation of the excess Bayes risk.

Lemma 5.2 (Excess Bayes risk)**.** For $f : \mathcal{X} \to \mathbb{R}$ measurable it holds that $R(f) - R(f^*) = \mathbb{E}[(f(X) - f^*(X))^2]$ and $R(f^*) = \sigma^2$.

Because of $Y \approx f^*(X)$ the standard approach in nonparametric statistics to estimate $f^*(x)$ at some $x \in \mathcal{X}$ as follows: Search for all observations X_i near to x and average over the corresponding Y_i. This leads to the following formula: For some parameter

$h > 0$ (the so-called 'bandwidth'), define

$$\begin{aligned}\hat{f}_{n,h}(x) &= \frac{\text{sum of observations } Y_i \text{ with } X_i \text{ near to } x}{\text{number of summands}} \\ &= \frac{\sum_{i\in\{1,\dots,n\}:|X_i-x|\le h/2} Y_i}{\sum_{i\in\{1,\dots,n\}:|X_i-x|\le h/2} 1} \\ &= \frac{\sum_{i=1}^n \mathbb{1}_{\{|\frac{X_i-x}{h}|\le\frac{1}{2}\}}}{\sum_{i=1}^n \mathbb{1}_{\{|\frac{X_i-x}{h}|\le\frac{1}{2}\}}} \\ &= \frac{\sum_{i=1}^n W\big(\frac{X_i-x}{h}\big)Y_i}{\sum_{i=1}^n W\big(\frac{X_i-x}{h}\big)}\end{aligned}$$

with $W(x) = \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(x)$. The bandwidth h should be understood as a limit which indicates which observations are still considered as 'near'. W weights the single observations Y_i according to their distance $|X_i - x|$ towards x. Note that however in the above formula, no real 'weighting' takes place since all observations are given the same factor 1 (if $|X_i - x| \le \frac{h}{2}$). We now allow for more complicated weightings.

Definition 5.3 (Kernel function)**.** A mapping $W : \mathbb{R}^d \to [0, \infty)$ with $\int_{\mathbb{R}^d} W(x)dx = 1$ is called *kernel function*.

Using such a general kernel function, we obtain the following algorithm.

Definition 5.4 (Kernel regression estimator)**.** Let $h > 0$ ('bandwidth') and W a kernel function. The algorithm

$$\hat{f}_{n,h}(x) := \frac{\sum_{i=1}^n W\big(\frac{X_i-x}{h}\big)Y_i}{\sum_{i=1}^n W\big(\frac{X_i-x}{h}\big)}$$

is called kernel regression estimator.

In the following let g denote the density of X with respect to the Lebesgue measure on $\mathcal{X}$.

Theorem 5.5 (Convergence rate of the kernel regression estimator)**.** Let $\mathcal{X} \subset \mathbb{R}^d$ be compact. Suppose that $g \in \mathcal{F}(L)$ and that there are constants $c_g > 0, C_g, C_W, C_\varepsilon \ge 1$

such that $c_g \leq g(x) \leq C_g$ for all $x \in \mathcal{X}$, $|\varepsilon_1| \leq C_\varepsilon$ and

$$\int W(u)^2 du, \quad \int W(u)\|u\|_\infty du, \quad \int W(u)^2\|u\|_\infty^2 du, \quad \|W\|_\infty \leq C_W.$$

Then the following statement holds: For all $h > 0$ with

$$\frac{c_g}{4LC_W} \geq h \geq \max\{\frac{128C_gC_W}{c_g^2}, \frac{8C_W}{3c_g}\}^{1/d} \cdot \big(\frac{\log(n)}{n}\big)^{1/d}, \tag{54}$$

we have

$$\mathbb{E}R(\hat{f}_{n,h}) - R(f^*) \leq \gamma(n,h) := \frac{8C_g^2C_W^2}{c_g^2} \cdot \big\{\frac{\sigma^2}{nh^d} + L^2 \cdot h^2 + L^2 \cdot \frac{h^2}{nh^d}\big\} + 8(\|f^*\|_\infty + C_\varepsilon)^2 n^{-1}.$$

Proof. The proof is left as an exercise. □

Remarks:

- Only the first two summands of $\gamma(n,h)$ are relevant for the convergence rate, the others are of smaller order. If we optimize the convergence rate with respect to h, we find that $h^* = n^{-\frac{1}{d+2}}$ yields

$$\gamma(n,h^*) \approx \textbf{const.} \cdot n^{-\frac{2}{d+2}}.$$

- We see that the upper bound $n^{-\frac{2}{2+d}}$ for the excess Bayes risk is very slow in n if d is large (even for the optimal h). d slows down the rate *exponentially* in n. This is called the 'curse of dimension'. The reason is that our model assumption is very general and does not produce much structural information on f^*. Since we only know that f^* is Lipschitz continuous, the observations X_i have to be located very dense in $\mathcal{X}$. As an example, consider $\mathcal{X} = [0,1]^d$. Then we need at least a^d training samples to guarantee that in each dimension, there lie a points. Vice versa, given n training samples, one can only use $a = n^{1/d}$ points along each dimension for estimation. This is a heuristic explanation of the exponent $1/d$ appearing in the convergence rate obtained above.

- $\mathcal{X}$ is compact is only needed to realize the assumption that g has a lower bound (that is, $g(x) \geq c_g$).

5.2 Classification

We now consider classification problems (for simplicity, again only with two classes), that is, $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{-1, +1\}$, $L(y, s) = \mathbb{1}_{\{y \neq s\}}$. The Bayes rule reads

$$f^*(x) = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(Y = k | X = x).$$

To motivate an estimator, we start by recalling Bayes' theorem.

Lemma 5.6 (Bayes' theorem)**.** Let λ denote the Lebesgue measure on $\mathbb{R}^d$. Let $\pi_k := \mathbb{P}(Y = k)$ and let g_k denote the conditional density of X given $Y = k$ with respect to λ, and g then density of X with respect to λ. Then it holds that

$$\mathbb{P}(Y = k | X = x) = \frac{\pi_k \cdot g_k(x)}{g(x)}, \quad k \in \mathcal{Y}.$$

Given a kernel function W, an estimator of the density $g(x)$ is given by

$$\hat{g}_{n,h}(x) := \frac{1}{nh^d} \sum_{i=1}^{n} W\Big(\frac{X_i - x}{h}\Big).$$

Here, for each $x \in \mathcal{X}$ the number of training samples X_i near to x is determined and then divided by the average number of samples in this 'volume' nh^d. Mathematically, it is also possible to derive the above estimator very naturally as derivative of the empirical distribution function.
Similarly, we estimate the conditional densities g_k only based on the observations X_i with $Y_i = k$. In summary, we obtain the following classifier.

Definition 5.7 (Kernel classification algorithm)**.** Let K be a kernel function and $h > 0$ ('bandwidth'). Put

$$\hat{\pi}_k = \frac{\#\{i : Y_i = k\}}{n}, \qquad \hat{g}_{k,n,h}(x) = \frac{1}{n\hat{\pi}_k h^d} \sum_{i:Y_i=k} W\Big(\frac{X_i - x}{h}\Big).$$

Let $\hat{\delta}_k(x) = \frac{\hat{\pi}_k \cdot \hat{g}_{k,n,h}(x)}{\hat{g}_{n,h}(x)}$. Then

$$\hat{f}_{n,h}(x) := \arg\max_{k \in \mathcal{Y}} \hat{\delta}_k(x)$$

is called the *kernel classification algorithm*.

If we restrict ourselves to two classes $\mathcal{Y} = \{-1, +1\}$ as mentioned at the beginning of this section, the above estimator has a more simple representation. In this case, it holds that $\hat{\delta}_1(x) + \hat{\delta}_{-1}(x) = 1$ and thus

$$\hat{f}_{n,h}(x) = \begin{cases} 1, & \hat{\delta}_1(x) > \hat{\delta}_{-1}(x), \\ -1, & \hat{\delta}_1(x) < \hat{\delta}_{-1}(x) \end{cases} = \begin{cases} 1, & 2\hat{\delta}_1(x) > 1, \\ -1, & 2\hat{\delta}_1(x) < 1 \end{cases} = \operatorname{sign}(2\hat{\delta}_1(x) - 1).$$

This leads to the following definition.

Definition 5.8 (Kernel classification algorithm for two classes)**.** Let K be a kernel function and $h > 0$ ('bandwidth'). Put

$$\hat{\delta}(x) = 2\hat{\delta}_1(x) - 1, \qquad \hat{\delta}_1(x) = \frac{\sum_{i=1}^n W(\frac{X_i - x}{h}) \mathbb{1}_{\{Y_i = 1\}}}{\sum_{i=1}^n W(\frac{X_i - x}{h})}.$$

Then

$$\hat{f}_{n,h}(x) := \operatorname{sign}(\hat{\delta}(x))$$

is called the *kernel classification algorithm*.

Note that $\hat{\delta}_1(x)$ has the same form as the kernel regression estimator from Definition 5.4 (but with $\mathbb{1}_{\{Y_i=1\}}$ instead of Y_i). This can be used to transfer properties from the regression to the classification case.
For regression estimators, we measured the risk based on the squared loss $\mathbb{E}[(\hat{g}_{n,h}(x) - g(x))^2]$. We now try to relate the excess Bayes risk $R(\hat{f}_{n,h}) - R(f^*)$ (with respect to the 0-1 loss) to such terms. To do so, we apply the risk transfer formula to the squared loss $\tilde{L}(y, s) = (y - s)^2$, $\tilde{R}(\delta) = \mathbb{E}\tilde{L}(Y, \delta(X))$.
Caution: In our setting, we can not use the noise condition with $q = \infty$ since later on we want to assume that $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ is Lipschitz continuous! Therefore, we do not impose any noise condition and only use the standard risk transfer formula.

Lemma 5.9. It holds that

$$R(\operatorname{sign}(\delta)) - R(f^*) \leq \left\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\right\}^{1/2},$$

where $\delta^*(x) = 2\eta(x) - 1$.

Proof. We apply Theorem 3.24 and Theorem 3.19 with $\tilde{L}(y, s) = (y - s)^2 = (1 - ys)^2 = \phi(-ys)$, $\phi(x) = (1 + x)^2$, $s = 2$ and $C_H = \frac{1}{2}$. The details are left as an exercise. □

Theorem 5.10 (Convergence rate of the kernel classification algorithm)**.** Suppose that $\mathcal{X} \subset \mathbb{R}^d$ is compact. Suppose that $g, \eta \in \mathcal{F}(L)$ and that there exist $c_g, C_g, C_W > 0$ such that $c_g \leq g(x) \leq C_g$ for all $x \in \mathcal{X}$ and

$$\int W(u)^2 du, \quad \int W(u)\|u\|_\infty du, \quad \int W(u)^2\|u\|_\infty^2 du, \quad \|W\|_\infty \leq C_W.$$

If the bandwidth satisfies (54), then it holds that

$$\mathbb{E}R(\hat{f}_{n,h}) - R(f^*) \leq \gamma_2(n,h) := 2\Big(\frac{8C_g^2 C_W^2}{c_g^2} \cdot \Big\{\frac{1}{nh^d} + L^2 \cdot h^2 + L^2 \cdot \frac{h^2}{nh^d}\Big\} + 32n^{-1}\Big)^{1/2}.$$

Proof. Lemma 5.2 $\Rightarrow$

$$\begin{aligned} \tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) &= \mathbb{E}[(\hat{\delta}(X) - \delta^*(X))^2] \\ &= 4\mathbb{E}[(\hat{\delta}_1(X) - \eta(X))^2] = 4\{\tilde{R}(\hat{\delta}_1) - \tilde{R}(\eta)\}. \end{aligned}$$

Lemma 5.9 $\Rightarrow$

$$R(\hat{f}_{n,h}) - R(f^*) \leq \{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*)\}^{1/2} \leq 2\{\tilde{R}(\hat{\delta}_1) - \tilde{R}(\eta)\}^{1/2}. \tag{55}$$

Note that $\hat{\delta}_1(x)$ is a kernel regression estimator in the model

$$\underbrace{\mathbb{1}_{\{Y=1\}}}_{=:\tilde{Y}} = \eta(X) + \underbrace{\{\mathbb{1}_{\{Y=1\}} - \eta(X)\}}_{=:\tilde{\varepsilon}}$$

It holds that $\mathbb{E}[\tilde{\varepsilon}|X] = \eta(X) - \eta(X) = 0$ and $\mathbb{E}[\tilde{\varepsilon}^2|X] = \mathrm{Var}(\mathbb{1}_{\{Y=1\}}) = \eta(X) - \eta(X)^2 \leq 1 =: \sigma^2$, as well as $|\eta(x)| \leq 1 =: C_{f^*}$, $|\tilde{\varepsilon}_1| \leq |\mathbb{1}_{\{Y=1\}} - \eta(X)| \leq 1$. This shows that the model assumption of the regression model from Definition 5.1 is satisfied. Theorem 5.5 implies

$$\mathbb{E}\tilde{R}(\hat{\delta}_1) - \tilde{R}(\eta) \leq \gamma(n,h).$$

With (55) and $\mathbb{E}[Z^{1/2}] \leq (\mathbb{E}Z)^{1/2}$, we obtain the assertion. □

Besides the additional exponent $\frac{1}{2}$, our bound has the same properties as in the regression case. In particular, we it suffers from the 'curse of dimension' (even for the best possible bandwidth h). Again, this is due to our general assumption which only asks for Lipschitz continuity of the density g and the function η.
One could ask oneself if the exponent $\frac{1}{2}$ can be improved. In fact, there is a complex relationship between the smoothness of $\eta(x)$ and the possible q for which the noise condition holds. We will not investigate this here, but refer to [11] or [17] for a detailed summary.

5.3 The naive Bayes classifier

The can overcome the curse of dimension by imposing more specific structural assumptions on either the distribution or the Bayes rule itself. Here, we will present one possibility of such a structural assumption. We will assume that the components $X_1, ..., X_d$ of X are independent conditional on each class $Y = k$. Note that this is a quite restrictive assumption.

Definition 5.11 (Model assumption: naive Bayes classifier)**.** For $k \in \mathcal{Y}$ there exist measurable functions $g_k^{(j)} : \mathbb{R} \to \mathbb{R}_{\geq 0}$ $(j = 1, ..., d)$, such that the conditional density of g_k of X given $Y = k$ satisfies

$$g_k(x) = \prod_{j=1}^{d} g_k^{(j)}(x_j), \quad x = (x_1, ..., x_d)^T \in \mathbb{R}^d.$$

Due to the given structure, we can estimate the conditional densities $g_k^{(j)}$ (which has a one-dimensional domain of definition) instead of estimating the whole function g_k (which has a d-dimensional domain of definition).

Definition 5.12 (Naive Bayes classifier)**.** Fix $x = (x_1, ..., x_d)^T \in \mathbb{R}^d$. Let $h > 0$, and let W be a kernel function. Define $\hat{\pi}_k := \frac{\#\{i \in \{1,...,n\} : Y_i = k\}}{n}$ and

$$\hat{g}_k^{(j)}(x_j) = \frac{1}{\hat{\pi}_k nh} \sum_{i:Y_i=k} W\Big(\frac{X_{ij} - x_j}{h}\Big).$$

Put $\hat{g}_k(x) = \prod_{j=1}^{d} \hat{g}_k^{(j)}(x_j)$. Let $\hat{\delta}_k(x) = \frac{\hat{\pi}_k \cdot \hat{g}_k(x)}{\sum_{l \in \mathcal{Y}} \hat{\pi}_l \hat{g}_l(x)}$. Then

$$\hat{f}_{n,h}(x) := \arg\min_{k \in \mathcal{Y}} \hat{\delta}_k(x)$$

is called the *naive Bayes classifier*.

Remark (about the name):

- The part 'Bayes' comes from the fact that we pose an assumption on the priori distribution $X|Y = k$ and the classifier is built based on Bayes' theorem (as in Definition 5.7).

- It is called 'naive' because the model assumption in Definition 5.11 is quite restrictive. In practice, one can not hope that the different features in a feature vector are independent; often they will be highly dependent.

An analysis of the naive Bayes classifier which yields good convergence rates both in d, n is out of the scope of this lecture. Here we restrict ourselves to a result which drastically improves the rate only with respect to n compared to Theorem 5.10.
As before, we restrict ourselves to two classes $\mathcal{Y} = \{-1, +1\}$ and note that $\hat{\delta}_1(x) + \hat{\delta}_{-1}(x) = 1$ as well as

$$\hat{f}_{n,h}(x) = \operatorname{sign}(\hat{\delta}(x)), \qquad \hat{\delta}(x) = 2\hat{\delta}_1(x) - 1.$$

As in the proof of Theorem 5.10, we therefore have (without posing a noise condition) that

$$R(\hat{f}_{n,h}) - R(f^*) \leq 2\{\tilde{R}(\hat{\delta}_1) - \tilde{R}(\eta)\}^{1/2} \tag{56}$$

with the quadratic loss $\tilde{L}(y, s) = (y - s)^2$. To analyze the right hand side, define

$$\hat{\delta}_1^{(j)}(x_j) := \frac{\hat{\pi}_1 \hat{g}_1^{(j)}(x_j)}{\hat{g}^{(j)}(x_j)},$$

where $\hat{g}^{(j)}(x_j) := \frac{1}{nh} \sum_{i=1}^n W\big(\frac{X_{ij} - x_j}{h}\big)$. Then we have

$$\begin{aligned}
\hat{\delta}_1(x) &= \frac{\hat{\pi}_1 \prod_{j=1}^d \hat{g}_1^{(j)}(x_j)}{\hat{\pi}_1 \prod_{j=1}^d \hat{g}_1^{(j)}(x_j) + \hat{\pi}_{-1} \prod_{j=1}^d \hat{g}_{-1}^{(j)}(x_j)} \\
&= \frac{\frac{1}{\hat{\pi}_1^{d-1}} \prod_{j=1}^d \frac{\hat{\pi}_1 \hat{g}_1^{(j)}(x_j)}{\hat{g}^{(j)}(x_j)}}{\frac{1}{\hat{\pi}_1^{d-1}} \prod_{j=1}^d \frac{\hat{\pi}_1 \hat{g}_1^{(j)}(x_j)}{\hat{g}^{(j)}(x_j)} + \frac{1}{\hat{\pi}_{-1}^{d-1}} \prod_{j=1}^d \frac{\hat{\pi}_{-1} \hat{g}_{-1}^{(j)}(x_j)}{\hat{g}^{(j)}(x_j)}} \\
&= \frac{\hat{\pi}_{-1}^{d-1} \prod_{j=1}^d \hat{\delta}_1^{(j)}(x_j)}{\hat{\pi}_{-1}^{d-1} \prod_{j=1}^d \hat{\delta}_1^{(j)}(x_j) + \hat{\pi}_1^{d-1} \prod_{j=1}^d (1 - \hat{\delta}_1^{(j)}(x_j))}.
\end{aligned} \tag{57}$$

Similarly, with $\eta^{(j)}(x_j) = \mathbb{P}(Y = 1 | X_j = x_j)$, we obtain

$$\eta(x) = \frac{\pi_{-1}^{d-1} \prod_{j=1}^d \eta^{(j)}(x_j)}{\pi_{-1}^{d-1} \prod_{j=1}^d \eta^{(j)}(x_j) + \pi_1^{d-1} \prod_{j=1}^d (1 - \eta^{(j)}(x_j))}. \tag{58}$$

Theorem 5.13. Suppose that for all $j \in \{1, ..., d\}$, $g^{(j)}, \eta^{(j)} \in \mathcal{F}(L)$ and that there exist $c_g, C_g, C_W, \eta_0 > 0$ such that $c_g \leq g^{(j)}(x) \leq C_g$, $\eta_0 \leq \eta^{(j)}(x) \leq 1 - \eta_0$ for all $x \in \mathcal{X}$ and

$$\int W(u)^2 du, \quad \int W(u) \|u\|_\infty du, \quad \int W(u)^2 \|u\|_\infty^2 du, \quad \|W\|_\infty \leq C_W.$$

Then the following statement holds: For all $h > 0$ with

$$\frac{c_g}{4LC_W} \geq h \geq \max\{\frac{128C_gC_W}{c_g^2}, \frac{8C_W}{3c_g}\} \cdot \frac{\log(n)}{n}, \tag{59}$$

we have with some universal constant $c > 0$ that

$$R(\hat{f}_{n,h}) - R(f^*) \leq \gamma^{naiv}(n,h) := \frac{cd}{\eta_0^d(\pi_{-1}^{d-1} + \pi_1^{d-1})} \cdot \Big\{ \frac{C_gC_W}{c_g} \cdot \Big\{ \frac{\sigma^2}{nh} + L^2 \cdot h^2 + L^2 \cdot \frac{h^2}{nh} \Big\}^{1/2} + n^{-1/2} \Big\}.$$

Proof. The structures in (57) and (58) have the form $\frac{a}{a+b}, \frac{a'}{a'+b'}$ with $a, a', b, b' \geq 0$. We now use

$$\Big| \frac{a}{a+b} - \frac{a'}{a'+b'} \Big| = \frac{|ab' - a'b|}{(a+b)(a'+b')} \leq \frac{a}{a+b} \cdot \Big(\frac{|b-b'|}{a'+b'} + \frac{|a-a'|}{a'+b'} \Big).$$

Suppose that there exists $\eta_0 > 0$ such that $1 - \eta_0 \geq \eta^{(j)}(x_j) \geq \eta_0$, $j = 1, ..., d$. Then we have $\frac{a}{a+b} \leq 1$, and $\frac{1}{a'+b'} \leq \frac{1}{\eta_0^d(\pi_{-1}^{d-1} + \pi_1^{d-1})}$, therefore

$$\begin{aligned} & |\hat{\delta}_1(x) - \eta(x)| \\ \leq \; & \frac{1}{\eta_0^d(\pi_{-1}^{d-1} + \pi_1^{d-1})} \cdot \Big(\Big| \hat{\pi}_{-1}^{d-1} \prod_{j=1}^d \hat{\delta}_1^{(j)}(x_j) - \pi_{-1}^{d-1} \prod_{j=1}^d \eta^{(j)}(x_j) \Big| \\ & \qquad + \Big| \hat{\pi}_1^{d-1} \prod_{j=1}^d (1 - \hat{\delta}_1^{(j)}(x_j)) - \pi_1^{d-1} \prod_{j=1}^d (1 - \eta^{(j)}(x_j)) \Big| \Big) \end{aligned} \tag{60}$$

By definition, $\hat{\pi}_k, \pi_k \in [0,1]$ and $\hat{\delta}_1^{(j)}, \eta^{(j)} \in [0,1]$. Therefore, we obtain

$$\begin{aligned} & \Big| \hat{\pi}_{-1}^{d-1} \prod_{j=1}^d \hat{\delta}_1^{(j)}(x_j) - \pi_{-1}^{d-1} \prod_{j=1}^d \eta^{(j)}(x_j) \Big| \\ \leq \; & \big| \hat{\pi}_{-1}^{d-1} - \pi_{-1}^{d-1} \big| \cdot \prod_{j=1}^d \hat{\delta}_1^{(j)}(x_j) + \pi_{-1}^{d-1} \cdot \Big| \prod_{j=1}^d \hat{\delta}_1^{(j)}(x_j) - \prod_{j=1}^d \eta^{(j)}(x_j) \Big| \\ \leq \; & (d-1) \cdot |\hat{\pi}_{-1} - \pi_{-1}| + \sum_{j=1}^d \big| \hat{\delta}_1^{(j)}(x_j) - \eta^{(j)}(x_j) \big|. \end{aligned}$$

Insertion into (60) yields

$$|\hat{\delta}_1(x) - \eta(x)| \leq \frac{d \cdot \big\{ |\hat{\pi}_{-1} - \pi_{-1}| + |\hat{\pi}_1 - \pi_1| \big\} + 2 \sum_{j=1}^d \big| \hat{\delta}_1^{(j)}(x_j) - \eta^{(j)}(x_j) \big|}{\eta_0^d(\pi_{-1}^{d-1} + \pi_1^{d-1})}.$$

We obtain that (note that $Z \mapsto \mathbb{E}[Z^2]^{1/2}$ is a norm which satisfies the triangle inequality):

$$\begin{aligned} & \mathbb{E}[(\hat{\delta}_1(X) - \eta(X))^2]^{1/2} \\ \leq \; & \frac{d \cdot \{\mathbb{E}[(\hat{\pi}_{-1} - \pi_{-1})^2]^{1/2} + \mathbb{E}[(\hat{\pi}_1 - \pi_1)^2]^{1/2}\} + 2\sum_{j=1}^d \mathbb{E}[\big(\hat{\delta}_1^{(j)}(X_j) - \eta^{(j)}(X_j)\big)^2]^{1/2}}{\eta_0^d(\pi_{-1}^{d-1} + \pi_1^{d-1})} \end{aligned}$$

We have

$$\mathbb{E}[(\hat{\pi}_1 - \pi_1)^2] = \frac{\pi_1(1-\pi_1)}{n} \leq \frac{1}{n},$$

and by the findings in the proof of Theorem 5.10 (with dimension $d = 1$),

$$\mathbb{E}[\big(\hat{\delta}_1^{(j)}(X_j) - \eta^{(j)}(X_j)\big)^2] \leq \frac{8C_g^2 C_W^2}{c_g^2} \cdot \{\frac{\sigma^2}{nh} + L^2 \cdot h^2 + L^2 \cdot \frac{h^2}{nh}\} + 32n^{-1}.$$

Summarizing the results into (56), we obtain that there exists a universal constant $c > 0$ such that

$$R(\hat{f}_{n,h}) - R(f^*) \leq \frac{cd}{\eta_0^d(\pi_{-1}^{d-1} + \pi_1^{d-1})} \cdot \Big\{\frac{C_g C_W}{c_g} \cdot \{\frac{\sigma^2}{nh} + L^2 \cdot h^2 + L^2 \cdot \frac{h^2}{nh}\}^{1/2} + n^{-1/2}\Big\}.$$

□

With $h = n^{-1/3}$, we obtain that

$$\gamma^{naiv}(n,h) \approx \frac{d}{\eta_0^d(\pi_{-1}^{d-1} + \pi_1^{d-1})} \cdot n^{-1/3}.$$

Contrary to the kernel classification algorithm from Definition 5.8 which has a rate $\approx n^{-\frac{1}{d+2}}$, we here obtain a convergence rate $n^{-1/3}$ which does *not* suffer from the curse of dimension. The main reason is that the model assumption allows us to reduce the d-dimensional estimation problem to d one-dimensional estimation problems. However, due to our rough upper bounds the rate suffers from a pre-factor which is exponential in d.

Remark 5.14 (Regarding model assumption and additive structure)**.** The model assumption of Definition 5.11 for two classes $\mathcal{Y} = \{-1, +1\}$ implies that there exist measurable functions $h^{(j)} : \mathbb{R} \to \mathbb{R}$, $j = 1, ..., d$ such that

$$\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) = \sum_{j=1}^d h^{(j)}(x_j). \tag{61}$$

Thus, the optimal decision regions have the form

$$\begin{aligned}\partial\Omega_1 &= \{x \in \mathcal{X} : \eta(x) > \frac{1}{2}\} = \{x \in \mathcal{X} : \log(\frac{\eta(x)}{1-\eta(x)}) > 0\} \\ &= \{x \in \mathcal{X} : \sum_{j=1}^{d} h^{(j)}(x_j) > 0\}.\end{aligned}$$

This means that the decision boundary is described by the level set of a function with an *additive structure*: The function is a sum of functions which only depend on one coordinate of x. In the following two chapters, we will investigate these structures in more detail.
Proof of (61): It holds that

$$\eta(x) = \frac{\pi_1 g_1(x)}{\pi_1 g_1(x) + \pi_{-1} g_{-1}(x)},$$

thus

$$\begin{aligned}\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) &= \log\Big(\frac{\pi_1 g_1(x)}{\pi_{-1} g_{-1}(x)}\Big) = \log\Big(\frac{\pi_1}{\pi_{-1}}\Big) + \log\Big(\frac{g_1(x)}{g_{-1}(x)}\Big) \\ &\overset{\text{Def. 5.11}}{=} \log\Big(\frac{\pi_1}{\pi_{-1}}\Big) + \sum_{j=1}^{d} \log\Big(\frac{g_1^{(j)}(x_j)}{g_{-1}^{(j)}(x_j)}\Big).\end{aligned}$$

Now we can define $h^{(j)}(x_j) := \log(\frac{g_1^{(j)}(x_j)}{g_{-1}^{(j)}(x_j)})$, $j = 2, ..., d$ and $h^{(1)}(x_1) := \log(\frac{\pi_1}{\pi_{-1}}) + \log(\frac{g_1^{(1)}(x_1)}{g_{-1}^{(1)}(x_1)})$.

5.4 Exercises

Task 20 (Analysis of the kernel regression algorithm)**.** In this task we investigate the excess Bayes risk (with respect to $L(y,s) = (y-s)^2$, $R(f) = \mathbb{E}L(Y, f(X))$) of the kernel regression algorithm

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} W\big(\frac{X_i - x}{h}\big) Y_i}{\sum_{i=1}^{n} W\big(\frac{X_i - x}{h}\big)}$$

in regression problems $Y = f^*(X) + \varepsilon$. We show that

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2] \leq \gamma(n,h) \quad (*)$$

with a deterministic rate $\gamma(n,h)$.

1. Show that (*) implies $\mathbb{E}R(\hat{f}) - R(f^*) \leq \gamma(n,h)$.

Suppose that $W : \mathbb{R}^d \to [0,\infty)$ is a kernel function and $f^* \in \mathcal{F}(L) = \{f : \mathcal{X} \to \mathbb{R}, \forall x,x' : |f(x)-f(x')| \leq L\|x-x'\|_\infty\}$. Let $g \in \mathcal{F}(L)$ be a density of X with respect to the Lebesgue measure on $\mathcal{X}$. Suppose further that there exist constants $c_g > 0, C_g \geq 1$ with $c_g \leq g(x) \leq C_g$. Let $A(x) := \{|\hat{g}(x) - g(x)| \leq \frac{c_g}{2}\}$, where

$$\hat{g}(x) = \frac{1}{n}\sum_{i=1}^n W_h(X_i - x), \qquad W_h(z) := \frac{1}{h^d}W(\frac{z}{h}).$$

In the following, we investigate the following two summands separately:

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2] = \mathbb{E}[(\hat{f}(x) - f^*(x))^2 \mathbb{1}_{A(x)}] + \mathbb{E}[(\hat{f}(x) - f^*(x))^2 \mathbb{1}_{A(x)^c}].$$

Analysis of the second summand:

(b) Show that if $|\varepsilon_1| \leq C_\varepsilon$ with some constant $C_\varepsilon > 0$, then we have

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2 \mathbb{1}_{A(x)^c}] \leq (2\|f^*\|_\infty + C_\varepsilon)^2 \mathbb{P}(A(x)^c).$$

(c) Suppose that $\|W\|_\infty, \int W(u)^2 du, \int W(u)\|u\|_\infty du \leq C_W$ with some constant $C_W \geq 1$. Show that

$$\mathbb{P}\big(|\hat{g}(x) - \mathbb{E}\hat{g}(x)| \geq \sqrt{\frac{2tC_gC_W}{nh^d}} + \frac{tC_W}{3nh^d}\big) \leq 2e^{-t}.$$

Hint: Apply Bernstein's inequality: For i.i.d. random variables Z_i with $|Z_i| \leq M$, $\mathbb{E}Z_i = 0$ and $\mathrm{Var}(Z_i) \leq V^2$ it holds that $\mathbb{P}(|\sum_{i=1}^n Z_i| \geq \sqrt{2tn}V + \frac{tM}{3}) \leq 2e^{-t}$.

(d) Suppose that $\int W(x)\|x\|_\infty dx \leq C_W$. Show that $\mathbb{E}\hat{g}(x) - g(x) \leq LC_W \cdot h$.

(e) Conclude from (c), (d): If

$$\frac{c_g}{4LC_W} \geq h \geq \max\{\frac{128C_gC_W}{c_g^2}, \frac{8C_W}{3c_g}\}^{1/d} \cdot \big(\frac{\log(n)}{n}\big)^{1/d}$$

then it holds that $\mathbb{P}(A(x)^c) \leq 2n^{-1}$.

(f) Conclude with (b):

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2 \mathbb{1}_{A(x)^c}] \leq 2(2\|f^*\|_\infty + C_\varepsilon)^2 n^{-1}.$$

Analysis of the first summand: Define

$$\hat{m}(x) := \frac{1}{n}\sum_{i=1}^n W_h(X_i - x)Y_i,$$

then one has $\hat{f} = \frac{\hat{m}}{\hat{g}}$.

(g) Show that

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2 \mathbb{1}_{A(x)}] \leq \frac{4}{c_g^2} \mathbb{E}[|\hat{m}(x) - f^*(x)\hat{g}(x)|^2].$$

(h) Show that

$$\begin{aligned} \mathbb{E}[|\hat{m}(x) - f^*(x)\hat{g}(x)|^2] \;\leq\; & \frac{2}{n} \operatorname{Var}\big(W_h(X_1 - x)\{f^*(X_1) - f^*(x)\}\big) \\ & + 2\mathbb{E}\big[W_h(X_1 - x)\{f^*(X_1) - f^*(x)\}\big]^2 \\ & + \frac{2}{n} \operatorname{Var}\big(W_h(X_1 - x)\varepsilon_1\big). \end{aligned}$$

Hint: First use the decomposition $Y_i - f(x) = \{f(X_i) - f(x)\} + \varepsilon_i$*, then apply* $(a+b)^2 \leq 2a^2 + 2b^2$*.*

(i) Show that

$$\frac{1}{n} \operatorname{Var}(W_h(X_1 - x)\varepsilon_1) \leq \frac{\sigma^2 C_g C_W}{nh^d}.$$

(j) Show that

$$\mathbb{E}\big[W_h(X_1 - x)\{f^*(X_1) - f^*(x)\}\big]^2 \leq (LC_g C_W)^2 h^2.$$

(k) It holds (without proof) that $\frac{1}{n} \operatorname{Var}\big(W_h(X_1 - x)\{f^*(X_1) - f^*(x)\}\big) \leq L^2 C_g C_W \frac{h^2}{nh^d}$. Combine the above results to show that

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2 \mathbb{1}_{A(x)}] \leq \frac{8 C_W^2 C_g^2}{c_g^2} \{\frac{\sigma^2}{nh^d} + L^2 h^2 + L^2 \frac{h^2}{nh^d}\}.$$

(l) Provide an upper bound for $\mathbb{E}[(\hat{f}(x) - f^*(x))^2]$.

6 Regression and classification trees; Boosting

In this chapter we investigate both regression and classification problems with a machine learning algorithm based on so-called *trees*. Trees are nonparametric estimators in the sense that their construction does not rely on a specific parametric model assumption on f^*.

6.1 Binary trees

Basic construction approach: The observation space $\mathcal{X}$ is split successively in *two* halves along a specific coordinate $j \in \{1, ..., d\}$ which is selected beforehand (*two* halves $\rightarrow$ *binary* trees). This construction produces a partition $\mathcal{X} = \bigcup A_k$ of $\mathcal{X}$. On this partition, we can define a decision rule which assigns to each element A_j a constant value y.

The successive partitioning of the space $\mathcal{X}$ along a coordinate can be formalizes with a tree structure. The first split corresponds to the root of the tree. There start two edges from the root which correspond to the two parts in which the space was split, and so on. We first recall the definition of a tree as a special case of a graph.

A directed graph $G = (V, E)$ is a tuple, where E is a set (the elements are called *vertices*) and $E \subset V \times V$ (these elements are called *edges*). For $(v, w) \in E$ we write $v \rightarrow w$ ('v points to w').

Definition 6.1 (Tree)**.** A tree $T = (V_T, E_T)$ is a directed graph with the following properties:

- T is acyclic, that is, there exist no sequence $v_1, ..., v_n \in V_T$ of vertices with $v_1 \rightarrow v_2 \rightarrow ... \rightarrow v_n \rightarrow v_1$ (there are no 'circles')
- From each vertex there starts either no edge (a so-called *leaf*) or exactly two edges (a so-called *inner vertex*).
- For each vertex besides one (the so-called *root*, denoted by $v_0 \in V_T$) there exists exactly one other vertex which points on it.

The set of all leafs is denoted with B_T, the number of leafs is abbreviated as $|T| := |B_T|$, where $|\cdot|$ denotes the cardinality of a set.

We now define regression and classification trees.

Definition 6.2 (Regression and classification trees)**.** Let T be a tree. Suppose that for each vertex $v \in V_T \backslash B_T$ (inner vertex) there exist

- $j(v) \in \{1, ..., d\}$ (the so-called *split index*) and
- $s(v) \in \mathbb{R}$ (the so-called *split point*).

Suppose that for each leaf $v \in B_T$ there exists $y(v) \in \mathcal{Y}$. Then T is called *CART* (classification and regression tree). T is called *regression tree* if $\mathcal{Y} = \mathbb{R}$ or *classification tree* if $\mathcal{Y} = \{1, ..., K\}$.

Given a CART T, one can define a decision rule based on T as follows:

Definition 6.3. Let T be a CART with root v_0. Define $A(v_0) := \mathcal{X}$. Recursively, we define the following quantities.

- If $A(v)$ is already defined for a vertex $v \in V_T \backslash B_T$, then let $v_1, v_2 \in V_T$ be the vertices with $v \to v_1$, $v \to v_2$. Define

$$A(v_1) := \{x \in A(v) : x_{j(v)} < s(v)\}, \qquad A(v_2) := \{x \in A(v) : x_{j(v)} \geq s(v)\}.$$

Then, $\mathcal{A}(T) := \{A(v) : v \in B_T\}$ is a partition of $\mathcal{X}$. The *decision rule corresponding to* T is defined by

$$f_T : \mathcal{X} \to \mathcal{Y}, \qquad f_T(x) := \sum_{v \in B_T} y(v) \cdot \mathbb{1}_{A(v)}(x). \tag{62}$$

Simply speaking, we define f_T in such a way that on $A(v)$, attains the value $y(v)$. Since $\bigcup_{v \in B_T} A(v) = \mathcal{X}$ is a partition of $\mathcal{X}$, (62) should not be understood as a sum but as a case distinction.

Note that the topological *closure* of the set of all CARTs is *equal* to the set $\{f : \mathcal{X} \to \mathcal{Y}$ messbar$\}$ since every measurable function can be approximated by piecewise constant functions. Thus, we have to reduce the size of the set of all CARTs. We do this by restricting the number of splits in each coordinate and the location of the split points.

6.2 Dyadic trees

For simplicity, we restrict ourselves to the cube $\mathcal{X} = [0, 1]^d$. Additionally, we will only allow the split points to be exactly in the middle of the cuboid which represents the

actual subspace.

Definition 6.4. A CART T is called *dyadic* if for each $v \in V_T \backslash B_T$ the following holds: If $A(v) = \prod_{j=1}^{d} [a_j(v), b_j(v)]$, then $s(v) := \frac{a_{j(v)}(v) + b_{j(v)}(v)}{2}$.

In the following, we consider the following subset of CARTs.

Definition 6.5. Let $S \in \mathbb{N}$, and

$$\begin{aligned} \mathcal{T}_S \;:=\; & \{T \text{ is a dyadic CART and along one path from} \\ & \text{root to leaf, each coordinate is split at most } S \text{ times}\}. \end{aligned}$$

Let (X, Y) be a regression or classification problem (Loss function $L(y, s) = (y - s)^2$ or $L(y, s) = \mathbb{1}_{\{y \neq s\}}$). To obtain a suitable tree and to avoid overfitting, we will additionally penalize the size of a tree via the number of leafs $|T|$.

Definition 6.6 (Dyadic CART algorithm)**.** Let $S \in \mathbb{N}$, $\lambda > 0$, and

$$\hat{T}_{n,\lambda} :\in \operatorname*{arg\,min}_{T \in \mathcal{T}_S} \left\{ \hat{R}_n(f_T) + \lambda \cdot |T| \right\}, \qquad \hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)).$$

Then $\hat{f}_{n,\lambda} := f_{\hat{T}_{n,\lambda}}$ is called *dyadic CART algorithm.*

Remark: As long as S is small enough, the global optimizer in Definition 6.6 is computable. As we will see below in Theorem 6.9, for classification we need $S \geq \lceil \frac{\log_2(n)}{d} \rceil$ to obtain a good upper bound for the excess Bayes risk. There exist procedures which can compute $\hat{T}_{n,\lambda}$ in $\leq c \cdot ndS^d \log(nS^d)$ steps (c a universal constant, cf. [3], Proposition 1).

We now provide a theoretical result for classification with $L(y, s) = \mathbb{1}_{\{y \neq s\}}$, $\mathcal{Y} = \{-1, +1\}$.

Theorem 6.7 (Oracle inequality for dyadic classification trees)**.** Let $\mathcal{X} = [0,1]^d$. Suppose that there exists $\eta_0 > 0$ such that $|\eta(x) - \frac{1}{2}| \geq \eta_0$ for all $x \in \mathcal{X}$. Then there exists a universal constant $c > 0$ such that for all $t \geq 0$ the following assertion holds: If

$$\lambda \geq c \frac{\log(d) + t}{\eta_0 n},$$

then

$$\mathbb{P}\Big(R(\hat{f}_{n,\lambda}) - R(f^*) \geq 2 \inf_{T \in \mathcal{T}_S} \big\{R(f_T) - R(f^*) + \lambda \cdot |T|\big\}\Big) \leq e^{-t}.$$

The proof is postponed to Subsection 6.3.
Remarks:

- The assertion in Theorem 6.7 is an 'oracle inequality'. It states that with high probability, $R(\hat{f}_{n,\lambda}) - R(f^*)$ behaves at most twice as bad as the best possible decision rule on $\mathcal{T}_S$ (which *knows* the true distribution) plus an additional penalization term.
- To derive a convergence rate for $R(\hat{f}_{n,\lambda}) - R(f^*)$, it is left to upper bound $R(f_T) - R(f^*)$ by introducing a model assumption.
- We will see that an oracle inequality is a very strong result on its own: The corresponding machine learning algorithm fulfills an optimality assumption 'uniformly' over all possible model assumptions. If we then pose a specific model assumption, we directly obtain a convergence rate. The important point is tht the machine learning algorithm *does not make use* of this model assumption but has this rate anyway. Note that the kernel estimators presented in Chapter 5 can *not* be written as empirical risk minimizers. They do not allow for such an oracle inequality.

Definition 6.8 (Model assumption: classification tree)**.** Let μ denote the Lebesgue measure on $\mathbb{R}^d$. Suppose that there exist constants $c_\mu, c_{box} > 0$ such that:

(a) For all $A \subset \mathcal{X}$, $\mathbb{P}(X \in A) \leq c_0 \mu(A)$.

(b) For all $m \in \mathbb{N}$, the optimal decision boundary $\partial\Omega_1^*$ intersects at most $c_{box} m^{d-1}$ of the m^d cubes in which the space $\mathcal{X} = [0,1]^d$ can be decomposed.

Remarks:

- Assumption (a) asks that there is not particular subset of $\mathcal{X}$ where much more training samples are realized. A tree with restricted depth (and thus only a restricted number of splits) could not represent such a distribution well.

- Assumption (b) provides an upper bound on the complexity of the optimal decision boundary. If for instance the optimal decision boundary $\partial\Omega_1^* = \{(x_1, ..., x_{d-1}, g(x_1, ..., x_{d-1})) : x_1, ..., x_{d-1} \in [0,1]\}$ is a surface with $g \in \mathcal{F}(L)$ (Lipschitz continuous, cf. (53)), then we have
$$c_{box} \leq L + 1.$$
Proof: If $\mathcal{X} = [0,1]^d$ is partitioned into m^d cubes with length $\frac{1}{m}$, then $(x_1, ..., x_{d-1}, 0)$ runs through m^{d-1} cubes. By running through one cube, all arguments of $g(x_1, ..., x_{d-1})$ change by at most $\frac{1}{m}$. Thus, g varies by at most $\frac{L}{m}$. Thus, the decision boundary can only run through at most $(L+1) \cdot m^{d-1}$ cubes in the direction of x_d.

Theorem 6.9 (Convergence rate of a dyadic classification tree)**.** Suppose that the model assumption in Definition 6.8 holds. Suppose that there exists $\eta_0 > 0$ with $|\eta(x) - \frac{1}{2}| \geq \eta_0$ for all $x \in \mathcal{X}$. Let $m = 2^S$. Then it holds that

$$\inf_{T \in \mathcal{T}_S} \left\{ R(f_T) - R(f^*) + \lambda \cdot |T| \right\} \leq c_\mu c_{box} m^{-1} + \lambda m^d.$$

If $S \in \mathbb{N}$, $S \geq \frac{\log_2(n)}{d+1}$ and $\lambda = c\frac{\log(2d)+t}{\eta_0 n}$, then we have

$$\inf_{T \in \mathcal{T}_S} \left\{ R(f_T) - R(f^*) + \lambda \cdot |T| \right\} \leq (2c_\mu c_{box} + \frac{c}{\eta_0}(\log(2d) + t)) \cdot n^{-\frac{1}{d+1}}.$$

Proof. This is left as an exercise. □

Remark: With a more detailed argumentation, one can achieve even λm^{d-1} instead of λm^d. The above result shows that trees are able to achieve the (bad) nonparametric rate $n^{-\frac{1}{d+1}}$ which suffers from the curse of dimension.
However, the oracle inequality can now be used to prove better convergence rates under stronger assumptions on the underlying distribution of X. We show this behavior on a very simple example that X is located in a s-dimensional subspace $[0,1]^s \times \{0\}^{d-s}$ of $\mathbb{R}^d$.

Definition 6.10 (Model assumption: classification tree, reduced dimension)**.** Let μ_s denote the Lebesgue measure on $\mathbb{R}^s$. Suppose that there exist constants $c_\mu, c_{box} > 0$ such that

(a) For all $A \subset [0,1]^s \times \{0\}^{d-s}$, $\mathbb{P}(X \in A) \leq c_0 \mu_s(A)$.

(b) For all $m \in \mathbb{N}$, the decision boundary $\partial\Omega_1^*$ intersects at most $c_{box} m^{s-1}$ of the m^d cubes in which the space $\mathcal{X}$ can be partitioned.

In (a) we ask that $X \in [0,1]^s \times \{0\}^{d-s}$ a.s.; accordingly in (b) we ask that the optimal decision boundary is only $(s-1)$-dimensional. Similar to Theorem 6.9, we obtain under the model assumption of Definition 6.10 that

$$\inf_{T \in \mathcal{T}_S} \left\{ R(f_T) - R(f^*) + \lambda \cdot |T| \right\} \leq (2 c_\mu c_{box} + \frac{c}{\eta_0}(\log(2d) + t)) \cdot n^{-\frac{1}{s+1}}.$$

First note that the convergence rate now is of order $n^{-\frac{1}{s+1}}$ which is much better than $n^{-\frac{1}{d+1}}$ if $s \ll d$. Second, note that the tree algorithm *does not need to know* during construction that this model assumption holds true; it is obtained in the same way as always. Of course, there is a price to pay to obtain this property: The cost for this adaptation is hidden in the term λ in Theorem 6.7. The original dimension d enters the convergence rate through the factor $\log(2d)$ (independent of what model assumption we use). This is still a remarkable result: It shows that the original dimension only enters the convergence rate logarithmically, so the 'price' of a tree to adapt to a more specific structure is very small compared to the original dimension of the space.

6.3 Proof of Theorem 6.7

In the following, we will abbreviate $\hat{f} = \hat{f}_{n,\lambda}$, $\hat{T}_{n,\lambda} = \hat{T}$.
Important observation: The minimization problem of a dyadic tree can be decomposed as follows: Let $\mathcal{A}(T) = \{A(v) : v \in V_T\}$ be the partition induced by a tree T, and

$$\mathbb{A} := \{\mathcal{A}(T) : T \in \mathcal{T}_S\}$$

the set of all partitions which can be induced by trees of $\mathcal{T}_S$. For one specific partition $\mathcal{A} \in \mathbb{A}$, let

$$\mathcal{F}_{\mathcal{A}} = \{f_{\mathcal{A},y}(x) := \sum_{A \in \mathcal{A}} y_A \cdot \mathbb{1}_A(x) \,|\, y_A \in \mathcal{Y} \text{ für } A \in \mathcal{A}\}.$$

be the set of all decision rules on this partition. Then we have

$$\begin{aligned} \min_{T \in \mathcal{T}_S} \{\hat{R}_n(f_T) + \lambda \cdot |T|\} &= \min_{\mathcal{A} \in \mathbb{A}} \min_{f \in \mathcal{F}_{\mathcal{A}}} \left\{ \hat{R}_n(f) + \lambda \cdot |\mathcal{A}| \right\} \\ &= \min_{\mathcal{A} \in \mathbb{A}} \Big\{ \min_{f \in \mathcal{F}_{\mathcal{A}}} \hat{R}_n(f) + \lambda \cdot |\mathcal{A}| \Big\}. \end{aligned}$$

In particular, $\hat{f}_{n,\lambda}$ has a representation of the form

$$\hat{f} = f_{\hat{\mathcal{A}},\hat{y}}, \qquad \hat{\mathcal{A}} = \mathcal{A}(\hat{T}), \qquad \hat{y} \in \mathcal{Y}^{|\mathcal{A}|}. \tag{63}$$

This shows that minimization over dyadic tree decision rules can be interpreted as a two-step minimization problem:

- First, we minimize over all possible partitions which can be generated by dyadic trees from $\mathcal{T}_S$,
- second, in parallel one minimizes over the values which the trees attain on these partitions.

Given a partition, the second minimization problem can be solved explicitly (this is left as an exercise). For the proof, we do not need this explicit representation but only the minimization property.

Step 1: Derivation of a basic inequality. Let

$$f_0 := f_{T_0}, \qquad T_0 :\in \operatorname*{arg\,min}_{T \in \mathcal{T}_S} \big\{ R(f_T) - R(f^*) + \lambda |T| \big\}.$$

Similar to (63), define $\mathcal{A}_0 = \mathcal{A}(T_0)$. We now decompose the excess Bayes risk into estimation and approximation error as follows:

$$R(\hat{f}) - R(f^*) \leq \{R(\hat{f}) - R(f_0)\} + \{R(f_0) - R(f^*)\} \tag{64}$$

In the following, we consider the estimation error

$$R(\hat{f}) - R(f_0) = \{\hat{R}_n(\hat{f}) - \hat{R}_n(f_0)\} + \big\{R(\hat{f}) - \hat{R}_n(\hat{f}) - (R(f_0) - \hat{R}_n(f_0))\big\}.$$

Recall the general proof technique from the SVM section. It holds that

$$R(\hat{f}) - R(f_0) \leq \lambda |\mathcal{A}_0| - \lambda |\hat{\mathcal{A}}| + V_{r,\hat{\mathcal{A}}}(f_0) \cdot (r^2 + D(\hat{f}, f_0)^2), \tag{65}$$

where $r > 0$, $\mathcal{A} \in \mathbb{A}$ and

$$V_{r,\mathcal{A}}(f_0) := \sup_{f \in \mathcal{F}_{\mathcal{A}}} \frac{\big\{R(f) - \hat{R}_n(f) - (R(f_0) - \hat{R}_n(f_0))\big\}}{r^2 + D(f, f_0)^2},$$

and $D(\cdot, \cdot)$ is a distance measure which we still have to define. To obtain upper bounds in (65), we therefore again have to derive two intermediate results:

- A concentration inequality for $V_{r,A}(f_0)$, and
- a margin property for $D(f, f^*)$ with respect to $R(f) - R(f^*)$.

In the following, we use the distance

$$D(f, f^*) := \mathbb{E}[(L(Y, f(X)) - L(Y, f^*(X)))^2]^{1/2}.$$

The overall reason is that V naturally includes $L(Y, f(X))$ as summands and therefore its variance can be controlled by bounding $D(f, f^*)$. We have the following margin property. Recall $\eta(x) = \mathbb{P}(Y = 1 | X = x)$.

Lemma 6.11. Suppose that there exists $\eta_0 > 0$ such that $|\eta(x) - \frac{1}{2}| \geq \eta_0$ for all $x \in \mathcal{X}$ (this corresponds to the noise condition with $q = \infty$). Then, for all $f : \mathcal{X} \to \mathbb{R}$ it holds that

$$D(f, f^*)^2 \leq \frac{1}{2\eta_0}\{R(f) - R(f^*)\}.$$

Proof. It holds that $f(X) = f^*(X) \Rightarrow L(Y, f(X)) = L(Y, f^*(X))$, and $|L(y, f(x)) - L(y, f^*(x))| \leq 1$. Thus we have

$$D(f, f^*)^2 \;=\; \mathbb{E}[(L(Y, f(X)) - L(Y, f^*(X)))^2] \leq \mathbb{E}[\mathbb{1}_{\{f(X) \neq f^*(X)\}}] = \mathbb{P}(f(X) \neq f^*(X)).$$

As in the proof of Theorem 3.21 (recall $\eta(x) > \frac{1}{2} \iff f^*(x) = 1$ and $\delta(x) < 0 \iff f(x) = -1$), it holds that

$$\begin{aligned} R(f) - R(f^*) &= \mathbb{E}\big[\underbrace{(2\eta(X) - 1)}_{\geq 2\eta_0}\mathbb{1}_{\{f^*(X)=1, f(X)=-1\}} + \underbrace{(1 - 2\eta(X))}_{\geq 2\eta_0}\mathbb{1}_{\{f^*(X)=-1, f(X)=1\}}\big] \\ &\geq 2\eta_0\mathbb{E}[\mathbb{1}_{\{f^*(X) \neq f(X)\}}] = 2\eta_0\mathbb{P}(f^*(X) \neq f(X)). \end{aligned}$$

Putting these inequalities together yields the result. □

Now we show a concentration inequality for $V_{r,\mathcal{A}}(f_0)$. As in the proof of the SVM algorithm, we start with upper bounds for the auxiliary term

$$Z_{r,\mathcal{A}}(f_0) := \sup_{f \in \mathcal{F}_{\mathcal{A}}, D(f, f_0) \leq r} \big|R(f) - \hat{R}_n(f) - (R(f_0) - \hat{R}_n(f_0))\big|.$$

Contrary to the SVM case, the set of functions $\mathcal{F}_{\mathcal{A}}$ in the supremum is *finite* which allows for different proof techniques. Here, we use a Bernstein inequality and a direct implication for expectations of finite suprema (cf. [18], Lemma 19.33, the derivation is left as an exercise).

Theorem 6.12. Let $Z_1, ..., Z_n$ be i.i.d. random variables with values in $\mathbb{R}$. Suppose that there exist $\sigma^2, M > 0$ such that $\mathbb{E}Z_1 = 0$, $\mathrm{Var}(Z_1) \leq \sigma^2$ and $|Z_1| \leq M$. Then it holds that

$$\mathbb{P}\Big(\sum_{i=1}^{n} Z_i \geq t\Big) \leq \exp\Big(-\frac{t^2}{2n\sigma^2 + 2Mt/3}\Big).$$

Theorem 6.13 (Maximal inequality for finite function classes)**.** Let $Z_1, ..., Z_n$ be i.i.d. random variables on $\mathcal{X} \subset \mathbb{R}^d$. Let $\mathcal{G} \subset \{g : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$ be a finite class of functions. Suppose that there exist $\sigma^2, M > 0$ such that for all $g \in \mathcal{G}$, it holds that $\mathbb{E}g(Z_1) = 0$, $\mathrm{Var}(g(Z_1)) \leq \sigma^2$ and $\|g\|_\infty \leq M$. Then

$$\mathbb{E} \sup_{g \in \mathcal{G}} \Big| \sum_{i=1}^{n} g(Z_i) \Big| \leq 4 \cdot \big\{ \sigma \sqrt{n} \sqrt{\log(|\mathcal{G}| + 1)} + M \cdot \log(|\mathcal{G}| + 1) \big\}.$$

Using these results, we obtain the following upper bound for $\mathbb{E}Z_{r,\mathcal{A}}(f_0)$.

Lemma 6.14. Let $|\mathcal{A}| \geq 1$. For $r \geq \frac{2\log(2)|\mathcal{A}|}{n}$, it holds that

$$\mathbb{E}Z_{r,\mathcal{A}}(f_0) \leq 8 \sqrt{\frac{|\mathcal{A}|}{n}} \cdot r.$$

Proof. We use that $Z_{r,\mathcal{A}}(f_0)$ has a representation

$$Z_{r,\mathcal{A}}(f_0) = \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} g(X_i, Y_i),$$

where

$$\mathcal{G} = \big\{ g(x,y) = \frac{1}{n} \big\{ \mathbb{E}L(Y, f(X)) - L(y, f(x)) - (\mathbb{E}L(Y, f_0(X)) - L(y, f_0(x))) \big\} : f \in \mathcal{F}_{\mathcal{A}} \big\}.$$

Moreover, it holds that $|L(y,s)| \leq 1$, thus $|g(x,y)| \leq \frac{4}{n}$ and

$$\mathrm{Var}(g(X,Y)) \leq \frac{\mathbb{E}[(L(Y, f(X)) - L(Y, f_0(X)))^2]}{n^2} = \frac{D(f, f_0)^2}{n^2} \leq \frac{r^2}{n^2}.$$

By Theorem 6.13, we conclude that

$$\mathbb{E}Z_{r,\mathcal{A}}(f_0) \leq 4 \cdot \Big\{ \sqrt{\frac{\log(|\mathcal{F}_{\mathcal{A}}| + 1) r^2}{n}} + \frac{\log(|\mathcal{F}_{\mathcal{A}}| + 1)}{n} \Big\}.$$

It holds that $|\mathcal{F}_{\mathcal{A}}| = 2^{|\mathcal{A}|}$ (for each $A \in \mathcal{A}$, one can select from two classes $y_A \in \{-1, +1\}$) $\Rightarrow$

$$\log(|\mathcal{F}_{\mathcal{A}}| + 1) = \log(2^{|\mathcal{A}|} + 1) \overset{|\mathcal{A}| \geq 1}{\leq} 2\log(2)|\mathcal{A}|.$$

Inserting this result into the upper bound above, we end up with

$$\mathbb{E}Z_{r,\mathcal{A}}(f_0) \le 4 \cdot \Big\{ \sqrt{\frac{2\log(2)|\mathcal{A}|r^2}{n}} + \frac{2\log(2)|\mathcal{A}|}{n} \Big\} \overset{r^2 \ge \frac{2\log(2)|\mathcal{A}|}{n}}{\le} 8\sqrt{\frac{|\mathcal{A}|}{n}} \cdot r.$$

Note that the condition $r^2 \ge \frac{2\log(2)|\mathcal{A}|}{n}$ is taken from our assumption and was used to guarantee that both terms are of the same size. □

As in the proof of the SVM (Lemma 4.28), we now use the peeling device and a Talagrand-type inequality to obtain the following statement for $V_{r,\mathcal{A}}(f_0)$ from $Z_{r,\mathcal{A}}(f_0)$. Note that here, $|L(y,s)| \le 1$.

Lemma 6.15. Let $|\mathcal{A}| \ge 1$. For $r \ge \frac{2\log(2)|\mathcal{A}|}{n}$, it holds that

$$\mathbb{E}V_{r,\mathcal{A}}(f_0) \le 40\sqrt{\frac{|\mathcal{A}|}{n}} \cdot \frac{1}{r}.$$

For $t \ge 0$, we have

$$\mathbb{P}\Big(V_{r,\mathcal{A}}(f_0) \ge 48\sqrt{\frac{|\mathcal{A}|}{n}} \cdot \frac{1}{r} + \sqrt{\frac{2t}{nr^2}} + \frac{22t}{nr^2} \Big) \le e^{-t}.$$

In particular, for every $N \in \mathbb{N}$ the following statement holds. With $r_{\mathcal{A},t} := c(N) \cdot \big\{ \frac{|\mathcal{A}|}{n\eta_0} + \frac{1}{\eta_0^2} \cdot \frac{t}{n} \big\}$, where $c(N)$ is some universal constant only depending on N, we have

$$\mathbb{P}\big(V_{r_{\mathcal{A},t},\mathcal{A}}(f_0) \ge \frac{\eta_0}{N} \big) \le e^{-t}.$$

Proof. The first two inequalities can be shown in the same way as in Lemma 4.28 and Lemma 4.30 (SVM). We only have to show that the given $r = r_{\mathcal{A},t}$ satisfies the inequality

$$\frac{\eta_0}{N} \ge 48\sqrt{\frac{|\mathcal{A}|}{n}} \cdot \frac{1}{r} + \sqrt{\frac{2t}{nr^2}} + \frac{22t}{nr^2}.$$

The inequality above is implied by

$$\frac{\eta_0}{3N} \ge 48\sqrt{\frac{|\mathcal{A}|}{n}} \cdot \frac{1}{r}, \quad \frac{\eta_0}{3N} \ge \sqrt{\frac{2t}{nr^2}}, \quad \frac{\eta_0}{3N} \ge \frac{22t}{nr^2},$$

and these inequalities are again implied by

$$r^2 \ge \max\Big\{ 144^2 N^2 \frac{|\mathcal{A}|}{n\eta_0}, \frac{18N^2 t}{n\eta_0^2}, \frac{66Nt}{n\eta_0} \Big\}.$$

This is implied by the choice $r = r_{\mathcal{A},t}$ if $c(N)$ is chosen large enough. □

Step 2: Definition of a 'nice' event E. At the moment, we have shown the basic inequality (65). To upper bound $V_{r,\hat{\mathcal{A}}}(f_0)$ we have to define an event which allows to upper bound $V_{r,\mathcal{A}}$ *uniformly* over all $\mathcal{A} \in \mathbb{A}$.

For $t > 0$ let $t(|\mathcal{A}|)$ denote a (still unknown) function dependent on $|\mathcal{A}|, t$. We will derive its concrete form in a second. Define

$$E := \bigcap_{\mathcal{A}\in\mathbb{A}} \{V_{r_{\mathcal{A},t(|\mathcal{A}|)},\mathcal{A}}(f_0) \leq \frac{\eta_0}{N}\}$$

with $r_{\mathcal{A},t} := c(N) \cdot \{\frac{|\mathcal{A}|}{n\eta_0} + \frac{1}{\eta_0^2} \cdot \frac{t}{n}\}$ as given in Lemma 6.15. Then we have

$$\begin{aligned}
\mathbb{P}(E^c) &\leq \mathbb{P}\Big(\bigcup_{\mathcal{A}\in\mathbb{A}} \{V_{r_{\mathcal{A},t(|\mathcal{A}|)},\mathcal{A}}(f_0) \geq \frac{\eta_0}{N}\}\Big) \leq \sum_{\mathcal{A}\in\mathbb{A}} \mathbb{P}\big(V_{r_{\mathcal{A},t(|\mathcal{A}|)},\mathcal{A}}(f_0) \geq \frac{\eta_0}{N}\big) \\
&\leq \sum_{\mathcal{A}\in\mathbb{A}} e^{-t(|\mathcal{A}|)} = \sum_{D=1}^{\infty} \sum_{\mathcal{A}\in\mathbb{A}:|\mathcal{A}|=D} e^{-t(D)} = \sum_{D=1}^{\infty} e^{-t(D)} \cdot |\{\mathcal{A} \in \mathbb{A} : |\mathcal{A}| = D\}| \overset{!}{\leq} e^{-t}.
\end{aligned}$$

It is our aim to satisfy the last inequality by choosing a suitable function $t(|\mathcal{A}|)$. To do so, we have to upper bound the second to last term in the above inequality. The number of binary trees, which produce a partition with $D+1$ elements corresponds to the $(D+1)$-th Catalan number $\frac{1}{D+1}\binom{2D}{D}$ (note that we count only the trees themselves, *not* any assignment with split coordinates $j(v)$). A tree with $(D+1)$ leafs has D internal vertices. For every internal vertex, we can choose from d dimensions. Then we have

$$|\{\mathcal{A} \in \mathbb{A} : |\mathcal{A}| = D\}| \leq d^D \cdot \frac{1}{D+1}\binom{2D}{D}.$$

Stirling's formula ($1 \leq \frac{k!}{\sqrt{2\pi k}(\frac{k}{e})^k} \leq 2$ for any $k \in \mathbb{N}$) yields the upper bound

$$\binom{2D}{D} = \frac{(2D)!}{(D!)^2} \leq \frac{2\sqrt{4\pi D}(\frac{2D}{e})^{2D}}{(\sqrt{2\pi D}(\frac{D}{e})^D)^2} = \frac{2}{\sqrt{\pi D}} \cdot 2^{2D}.$$

Therefore, we get

$$|\{\mathcal{A} \in \mathbb{A} : |\mathcal{A}| = D\}| \leq 2 \cdot (4d)^D.$$

We may therefore choose $t(D) = 2D\log(4d) + t$. Then we have

$$\begin{aligned}
\mathbb{P}(E^c) &\leq \sum_{D=1}^{\infty} e^{-t(D)} \cdot |\{\mathcal{A} \in \mathbb{A} : |\mathcal{A}| = D\}| = e^{-t} \sum_{D=1}^{\infty} (4d)^{-2D} \cdot 2 \cdot (4d)^D \\
&= 2e^{-t} \sum_{D=1}^{\infty} (4d)^{-D} \leq 2e^{-t} \frac{(4d)^{-1}}{1-(4d)^{-1}} \leq e^{-t}.
\end{aligned}$$

Step 3: Derivation of an upper bound on the excess Bayes risk. The margin property of Lemma 6.11 implies

$$D(\hat{f}, f_0)^2 \leq 2D(\hat{f}, f^*)^2 + 2D(f_0, f^*)^2 \leq \frac{1}{\eta_0}\Big[\{R(\hat{f}) - R(f^*)\} + \{R(f_0) - R(f^*)\}\Big].$$

Plugging this into (64) and (65) yields on the event E:

$$\begin{aligned} R(\hat{f}) - R(f^*) \;\; &\leq \;\; \lambda|\mathcal{A}_0| - \lambda|\hat{\mathcal{A}}| + \frac{\eta_0}{N} \cdot r^2_{\hat{\mathcal{A}},t(|\hat{\mathcal{A}}|)} + N^{-1}\{R(\hat{f}) - R(f^*)\} \\ &\quad +(1 + N^{-1})\{R(f_0) - R(f^*)\}. \end{aligned}$$

Rearranging terms yields

$$R(\hat{f}) - R(f^*) \leq \frac{1}{1 - N^{-1}}\Big[(1 + N^{-1})\{R(f_0) - R(f^*)\} + \lambda|\mathcal{A}_0| \underline{\underline{-\lambda|\hat{\mathcal{A}}| + \frac{\eta_0}{N} \cdot r^2_{\hat{\mathcal{A}},t(|\hat{\mathcal{A}}|)}}}\Big]$$

We have to eliminate the non-deterministic terms from the underlined expression above. It holds that

$$\frac{\eta_0}{N} r^2_{\hat{\mathcal{A}},t(|\hat{\mathcal{A}}|)} = \frac{C(N)}{N} \cdot \{\frac{|\mathcal{A}|}{n} + \frac{1}{\eta_0} \cdot \frac{2|\mathcal{A}|\log(4d) + t}{n}\} \leq \frac{4C(N)}{N} \cdot \frac{\log(d) + t}{\eta_0 n}|\mathcal{A}|,$$

therefore we should choose $\lambda \geq \frac{4C(N)}{N} \cdot \frac{\log(d)+t}{n\eta_0}$. Then for N large enough (choose c in λ accordingly):

$$R(\hat{f}) - R(f^*) \leq \frac{1}{1 - N^{-1}}\Big[(1+N^{-1})\{R(f_0) - R(f^*)\} + \lambda|\mathcal{A}_0|\Big] \leq 2 \inf_{T \in \mathcal{T}_S} \big\{R(f_T) - R(f^*) + \lambda|T|\big\}.$$

6.4 Boosting

The original idea of Boosting was introduced for classification problems, but it can be extended to regression problems. Here, we consider only classification problems with two classes, that is,

$$\mathcal{X} \subset [0,1]^d, \qquad \mathcal{Y} \in \{-1, +1\}, \qquad L(y, s) = \mathbb{1}_{\{y \neq s\}}.$$

Approach: The starting point of the procedure is to choose a (relatively small) class of *base decision rules* $\mathcal{C} \subset \{f : \mathcal{X} \to \mathcal{Y} \text{ measurable}\}$. A typical example could be $\mathcal{C} = \mathcal{C}_1$, where

$$\begin{aligned} \mathcal{C}_1 \;\; &= \;\; \{f_T : \; T \text{ CART with depth 1, that is, } |T| = 2\} \\ &= \;\; \{x \mapsto y_1 \cdot \mathbb{1}_{\{x_j < s\}} + y_2 \cdot \mathbb{1}_{\{x_j \geq s\}} : y_1, y_2 \in \mathcal{Y}, j \in \{1, ..., d\}, s \;\in \mathbb{R}\} \end{aligned} \tag{66}$$

the function class of so-called *tree stumps* or *decision stumps*.
Obviously, the decision rules f coming from $\mathcal{C}$ are not suitable to approximate more complex Bayes rules f^*. The idea is to extend $\mathcal{C}$ successively to obtain better decision rules. This is done as follows.

1. Find a classifier $\hat{\delta}^{(1)} \in \mathcal{C}$ which explains (X_i, Y_i), $i = 1, ..., n$ as good as possible.

2. Afterwards, find $\hat{g}^{(2)} \in \mathcal{C}$ which tries to explain *in particular those* data points (X_i, Y_i) which were not well explained by $\hat{\delta}^{(1)}$.

3. *Combine* $\hat{\delta}^{(2)} := \hat{\delta}^{(1)} + \hat{\beta}_2 \cdot \hat{g}^{(2)}$ with a suitably chosen scaling factor $\hat{\beta}_2 \geq 0$. From intuition it should hold that $\hat{\beta}_2 \in [0, 1]$ since the second classifier $\hat{g}^{(2)}$ shall only *correct* the first classifier $\hat{\delta}^{(1)}$, not replace its result.

4. By combining both classifiers, in general $\hat{\delta}^{(2)}$ is no longer a function with values in $\mathcal{Y}$. Instead, $\hat{\delta}^{(2)}$ is an indicator how sure we are to decide for one of the classes $y \in \mathcal{Y} = \{-1, +1\}$. An improved *classifier* which combines the 'knowledge' from both base classifiers $\hat{\delta}^{(1)}, \hat{g}^{(2)}$ is given by

$$\hat{f}^{(2)}(x) := \text{sign}(\hat{\delta}^{(2)}(x)).$$

5. Repeat the steps (2.)-(4.) for $m \in \mathbb{N}$ with $\hat{g}^{(m)}$ and $\hat{\delta}^{(m-1)}$, that is,

$$\hat{f}^{(m)}(x) := \text{sign}(\hat{\delta}^{(m)}(x)), \qquad \hat{\delta}^{(m)} := \hat{\delta}^{(m-1)} + \hat{\beta}_m \cdot \hat{g}^{(m)}.$$

   Then, $\hat{f}^{(m)}$ then combines the 'knowledge' of m classifiers.

In the literature, such methods which (successively) improve or combine base classifiers are called *ensemble learning methods*. The method presented here is called *Boosting*: A base classifier with weak quality is *boosted* in the way that many replications of it are summarized to a strong 'comitee'. The hope is that if the procedure is stopped at the right time, overfitting is avoided.

If we define $\hat{\beta}_1 = 1$, $\hat{g}^{(1)} := \hat{\delta}^{(1)}$, then the procedure produces classifiers of the form

$$\hat{f}^{(m)}(x) = \text{sign}(\hat{\delta}^{(m)}(x)), \qquad \hat{\delta}^{(m)}(x) := \sum_{j=1}^{m} \hat{\beta}_j \cdot \hat{g}^{(j)}(x), \quad \hat{g}^{(j)} \in \mathcal{C}, \hat{\beta}_j \geq 0.$$

Therefore, $\hat{\delta}_n$ is an element of the following function class.

Definition 6.16 (Function class of boosting decision rules)**.** Let $\mathcal{C} \subset \{f : \mathcal{X} \to \mathcal{Y} \text{ measurable}\}$ with $\mathcal{C} = -\mathcal{C}$. Define

$$\Delta := \{x \mapsto \delta_{\beta,g} := \sum_{j=1}^{N} \beta_j \cdot g_j(x) : N \in \mathbb{N}, \forall j \in \{1, ..., N\} : g_j \in \mathcal{C}, \beta_j \geq 0\}.$$

Since we assume that the weights β_j are nonnegative, we ask for $\mathcal{C} = -\mathcal{C} := \{-f : f \in \mathcal{C}\}$ so that the same decision rules are available in direction of both classes -1 and $+1$.

To include a suitable *stopping rule*, we introduce a quantity which measures the sum of weights which are needed to represent a specific $\delta \in \Delta$:

Definition 6.17.

$$\|\delta\|_1 := \inf\{\sum_{j=1}^{N} \beta_j : \delta_{\beta,g} \in \Delta \text{ with } \delta_{\beta,g} = \delta\}.$$

Since δ has several representations of the form $\delta = \delta_{\beta,g} \in \Delta$, we have to take the infimum so that $\|\cdot\|_1$ is well-defined and is a norm (in particular, $\|\cdot\|_1$ satisfied the triangle inequality).

6.4.1 The exact boosting algorithm

Up to now we have not described how the specific β_j, $\hat{g}_n^{(j)}$ are obtained in the above representation of $\hat{\delta}_n$. We now present a formal definition.

Definition 6.18 (The exact boosting algorithm)**.** Let $\tilde{L} : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function. For $\lambda > 0$ and some non-increasing $P : [0, \infty) \to [0, \infty)$, define

$$\hat{\delta}_n :\in \operatorname*{arg\,min}_{\delta \in \Delta} \left\{\tilde{R}_n(\delta) + \lambda \cdot P(\|\delta\|_1)\right\}, \qquad \tilde{R}_n(\delta) := \frac{1}{n}\sum_{i=1}^{n} \tilde{L}(Y_i, \delta(X_i)).$$

Then

$$\hat{f}_n(x) := \operatorname{sign}(\hat{\delta}_n(x))$$

is called *boosting algorithm* with respect to the base class $\mathcal{C}$.

We still have to define the function P. We will see later that for strong theoretical results, P has to depend on the loss function $\tilde{L}$.
In practice, $\hat{\delta}_n$ can not be computed exactly since Δ is too large. Instead, one uses approximative solutions. One possibility is shown in the next subsection. Note that this is the first time in this lecture that we will have a 'gap' between the theoretical result and the algorithm used in practice. For more complex algorithms, this is quite common

in recent articles in statistics for machine learning algorithms. It is also current research to close these gaps.
However, note that even the analysis of the exact boosting algorithm can give valuable hints how to choose P and which convergence rates might be obtained for different choices of the base class $\mathcal{C}$.

6.4.2 The approximated boosting algorithm

In practice, the optimization problem in Definition 6.18 is solved iteratively by the following procedure.

Definition 6.19. Let $\hat{\delta}^{(0)} := 0$. For $m \in \mathbb{N}$, compute

$$(\hat{\beta}_m, \hat{g}^{(m)}) :\in \operatorname*{arg\,min}_{\beta \geq 0, g \in \mathcal{C}} \tilde{R}_n(\hat{\delta}^{(m-1)} + \beta \cdot g), \tag{67}$$

and put $\hat{\delta}^{(m)} := \hat{\delta}^{(m-1)} + \hat{\beta}_m \cdot \hat{g}^{(m)}$. Then, define

$$\hat{\delta}_n^{\approx} :\in \operatorname*{arg\,min}_{m \in \mathbb{N}} \left\{ \tilde{R}_n(\hat{\delta}^{(m)}) + \lambda \cdot P(\|\hat{\delta}^{(m)}\|_1) \right\}.$$

The classifier $\hat{f}_n^{\approx}(x) = \operatorname{sign}(\hat{\delta}_n^{\approx}(x))$ is called the *approximated boosting classifier*.

In the above definition, one first derives a sequence of $\hat{\delta}_n^{(m)}$ of discriminant functions. Out of this sequence a final discriminant function is selected by solving an optimization problem which includes the penalization term $P(\|\hat{\delta}^{(m)}\|_1$.

For 'small' classes $\mathcal{C}$, the optimization problem in (6.19) can be solved exactly in practice. Note that the problem looks different for different loss functions $\tilde{L}$. In principle one may use each loss function; however there are several standard choices which lead to well interpretable (in view of the weights β_m) and also well computable algorithms. We now present two of these choices.
Let $\tilde{L}(y, s) = \phi(-ys)$ with some increasing function $\phi \in \{\phi_1, \phi_2\}$, where

$$\phi_1(x) = e^x, \qquad \phi_2(x) = \log(1 + e^x).$$

$\tilde{L}$ corresponding to ϕ_1 is called *exponential loss*, $\tilde{L}$ corresponding to ϕ_2 is called *logistic loss*.
We now investigate the optimization problem (67) for these loss functions.

Exponential loss $\phi = \phi_1$**:** For $\beta \geq 0$, $g \in \mathcal{C}$, we have

$$\begin{aligned}
& \tilde{R}_n(\hat{\delta}^{(m-1)} + \beta \cdot g) \\
= \quad & \frac{1}{n} \sum_{i=1}^{n} \tilde{L}(Y_i, \hat{\delta}^{(m-1)}(X_i) + \beta \cdot g(X_i)) \\
= \quad & \frac{1}{n} \sum_{i=1}^{n} \exp\big[- Y_i \cdot (\hat{\delta}^{(m-1)}(X_i) + \beta \cdot g(X_i))\big] \\
= \quad & \frac{1}{n} \sum_{i=1}^{n} w_i^{(m)} \cdot \exp(-\beta Y_i g(X_i)), \qquad w_i^{(m)} := \exp(-Y_i \hat{\delta}^{(m-1)}(X_i)), \\
\overset{g(X_i), Y_i \in \mathcal{Y} = \{-1,+1\}}{=} \quad & \frac{1}{n} \sum_{i=1}^{n} w_i^{(m)} \big[e^{\beta} \mathbb{1}_{\{Y_i \neq g(X_i)\}} + e^{-\beta} \mathbb{1}_{\{Y_i = g(X_i)\}} \big] \\
= \quad & (e^{\beta} - e^{-\beta}) \cdot \frac{1}{n} \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{Y_i \neq g(X_i)\}} + e^{-\beta} \cdot \frac{1}{n} \sum_{i=1}^{n} w_i^{(m)}.
\end{aligned}$$

For each $\beta > 0$ it holds that

$$\hat{g}^{(m)} := \operatorname*{arg\,min}_{g \in \mathcal{C}} \tilde{R}_n(\hat{\delta}^{(m-1)} + \beta \cdot g) = \operatorname*{arg\,min}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{Y_i \neq g(X_i)\}}. \tag{68}$$

that is, the optimization problem can be solved with respect to $g \in \mathcal{C}$ *independent* of the specific value of $\beta \geq 0$. If $\hat{g}^{(m)}$ is computed, one has

$$\hat{\beta}_m := \operatorname*{arg\,min}_{\beta \geq 0} \tilde{R}_n(\hat{\delta}^{(m-1)} + \beta \cdot \hat{g}^{(m)}) = \frac{1}{2} \cdot \log\Big(\frac{1 - E(m)}{E(m)}\Big), \qquad E(m) := \frac{\sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{Y_i \neq \hat{g}^{(m)}(X_i)\}}}{\sum_{i=1}^{n} w_i^{(m)}}.$$

In practice, the optimization problem (68) can be solved exactly for simple classes $\mathcal{C}$. Note that in (68), the optimization takes place with respect to a weighted 0-1 loss (no longer an exponential loss).

Example 6.20 (Solution of (68) for decision stumps)**.** Let $\mathcal{C} = \mathcal{C}_1$ from Definition (66). Then $\hat{g}^{(m)}$ can be obtained as follows: For fixed $j \in \{1, ..., d\}$, $s \in \mathbb{R}$ define

$$A_1(j, s) := \{x \in \mathbb{R}^d : x_j < s\}, \qquad A_2(j, s) := \{x \in \mathbb{R}^d : x_j \geq s\}.$$

Then

$$\begin{aligned}
\Phi(j, s, y_1, y_2) \quad &:= \quad \frac{1}{n} \sum_{i=1}^{n} w_i^{(m)} \mathbb{1}_{\{Y_i \neq g(X_i)\}} \\
&= \quad \frac{1}{n} \sum_{i \in \{1,...,n\} : X_i \in A_1(j,s)} w_i^{(m)} \mathbb{1}_{\{Y_i \neq y_1\}} + \frac{1}{n} \sum_{i \in \{1,...,n\} : X_i \in A_2(j,s)} w_i^{(m)} \mathbb{1}_{\{Y_i \neq y_2\}}.
\end{aligned}$$

We can minimize the sums in $\Phi(j, s, y_1, y_2)$ separately with respect to y_1, y_2:

$$\hat{y}_k(j, s) := \arg\max_{y \in \mathcal{Y}} \sum_{i \in \{1,...,n\}: X_i \in A_k(j,s), Y_i = y} w_i^{(m)} \quad k = 1, 2.$$

The solution can be interpreted as follows: 'Which class is (weighted with $w_i^{(m)}$) most common in $A_k(j, s)$?'. Therefore, it is left to solve

$$(\hat{j}, \hat{s}) := \arg\min_{j \in \{1,...,d\}, s \in \mathbb{R}} \Phi(j, s, \hat{y}_1(j, s), \hat{y}_2(j, s)).$$

The solution can be obtained by simply performing iterations over $j \in \{1, ..., d\}$ and $s \in \{X_{1:n}, ..., X_{n:n}\}$, where $X_{1:n} < ... < X_{n:n}$ denotes the ordered values of $X_{1j}, ..., X_{nj}$. Due to the discrete structure, $\Phi(j, s, \hat{y}_1(j, s), \hat{y}_2(j, s))$ attains all values on these pairs (j, s).

We therefore obtain the following procedure to calculate $\hat{\delta}^{(m)}$ in Definition 6.19.

Lemma 6.21 (Computation of the approximated boosting algorithm with exponential loss)**.** Let $\hat{\delta}^{(0)} = 0$, and $w^{(1)} = (w_1^{(1)}, ..., w_n^{(1)}) = (1, ..., 1)$. Then for $m = 1, 2, 3, ...$, it holds that

1. Let $\hat{g}^{(m)} \in \arg\min_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n w_i^{(m)} \mathbb{1}_{\{Y_i \neq g(X_i)\}}$,
2. $E(m) := \frac{\sum_{i=1}^n w_i^{(m)} \mathbb{1}_{\{Y_i \neq \hat{g}^{(m)}(X_i)\}}}{\sum_{i=1}^n w_i^{(m)}}$,
3. $\hat{\beta}_m := \frac{1}{2} \log(\frac{1-E(m)}{E(m)})$,
4. $w_i^{(m+1)} = w_i^{(m)} \cdot \exp\left[2\hat{\beta}_m \mathbb{1}_{\{Y_i \neq \hat{g}^{(m)}(X_i)\}} - \hat{\beta}_m\right]$, $i = 1, ..., n$.
5. It holds that $\hat{\delta}^{(m)} = \hat{\delta}^{(m-1)} + \hat{\beta}_m \hat{g}^{(m)}$.

Remarks: The evolution of the weights is obtained as follows. With $\hat{\delta}^{(m)} = \hat{\delta}^{(m-1)} + \hat{\beta}_m \hat{g}^{(m)}$, one has

$$\begin{aligned} w_i^{(m+1)} &= \exp(-Y_i \hat{\delta}^{(m)}(X_i)) = w_i^{(m)} \exp(-Y_i \hat{\beta}_m \hat{g}^{(m)}(X_i)) \\ &\overset{-Y_i \hat{g}^{(m)}(X_i) = 2 \cdot \mathbb{1}_{\{Y_i \neq \hat{g}^{(m)}(X_i)\}} - 1}{=} w_i^{(m)} \exp(2\hat{\beta}_m \mathbb{1}_{\{Y_i \neq \hat{g}^{(m)}(X_i)\}}) \exp(-\hat{\beta}_m). \end{aligned}$$

This can be interpreted graphically: The training samples X_i, Y_i which were misclassified before are weighted more heavy in the next steps.

Logistic loss $\phi = \phi_2$: In this case we can obtain a simple solution if we allow for another approximation. Similar as in the exponential loss case, $g(X_i), Y_i \in \mathcal{Y} = \{-1, +1\}$ implies

$$\begin{aligned}\tilde{R}_n(\hat{\delta}^{(m-1)} + \beta \cdot g) &= \frac{1}{n}\sum_{i=1}^n \log\big(1 + \exp\big[-Y_i \cdot (\hat{\delta}^{(m-1)}(X_i) + \beta \cdot g(X_i))\big]\big) \\ &= \frac{1}{n}\sum_{i=1}^n \big\{\log\big(1 + w_i^{(m)} e^{\beta}\big) - \log\big(1 + w_i^{(m)} e^{-\beta}\big)\big\} \cdot \mathbb{1}_{\{Y_i \neq g(X_i)\}} \\ &\qquad + \frac{1}{n}\sum_{i=1}^n \log(1 + w_i^{(m)} e^{-\beta}).\end{aligned}$$

For $a \approx 0$ and $c \geq 0$, a Taylor approximation yields

$$\log(1 + c + a) \approx \log(1 + c) + \frac{1}{1 + c} \cdot a.$$

With $c = w_i^{(m)}$, $a = w_i^{(m)}(e^{\pm\beta} - 1) \approx 0$ (for $|\beta| \ll 1$) we obtain that for $|\beta| \ll 1$,

$$\tilde{R}_n(\hat{\delta}^{(m-1)} + \beta \cdot g) \approx (e^{\beta} - e^{-\beta})\frac{1}{n}\sum_{i=1}^n \frac{w_i^{(m)}}{1 + w_i^{(m)}} \mathbb{1}_{\{Y_i \neq g(X_i)\}} + e^{-\beta} \cdot \frac{1}{n}\sum_{i=1}^n \frac{w_i^{(m)}}{1 + w_i^{(m)}} + \text{const.}$$

The only difference to Lemma 6.21 is therefore that in step (1.) and step (2.) one makes use of the rescaled versions $\frac{w_i^{(m)}}{1+w_i^{(m)}}$ of the weights instead of $w_i^{(m)}$ itself. All other statements stay similar (at least approximately). For the exponential loss, $w_i^{(m)}$ can grow arbitrarily large due to the successive multiplication in step (4.) and therefore misclassified training samples can be weighted with a really large $w_i^{(m)}$ which may lead to overfitting after only few steps. In contrary, the logistic loss only weights with $\frac{w_i^{(m)}}{1+w_i^{(m)}}$ which behaves much more regular. Roughly spoken, this enforces that boosting with logistic loss is more stable and also more robust against outliers and overfitting.

6.5 Boosting: theoretical results

We now investigate the excess Bayes risk $R(\hat{f}_n) - R(f^*)$ of the *exact* boosting algorithm of Definition 6.18. Here we discuss the results from [2] but with severe changes tue to didactical reasons.
In the exercises and examples we have already shown the calibration condition and the risk transfer formula for both exponential and logistic loss (for $\phi = \phi_1$, see the exercises; for $\phi = \phi_2$, see Example 3.20, 3.22). Recall that $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

Lemma 6.22. For

- $\phi = \phi_1$, put $C_s = \frac{1}{\sqrt{2}}$, $\delta^*(x) = \frac{1}{2}\log(\frac{\eta(x)}{1-\eta(x)})$.
- $\phi = \phi_2$, put $C_H = 2\sqrt{2\log(2)}$, $\delta^*(x) = \log(\frac{\eta(x)}{1-\eta(x)})$.

Then it holds that for all $\delta : \mathcal{X} \to \mathbb{R}$,

$$R(\text{sign}(\delta)) - R(f^*) \leq 2C_H\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}^{1/2}.$$

If additionally there exists some $\eta_0 > 0$ such that $|\eta(x) - \frac{1}{2}| \geq \eta_0$ for all $x \in \mathcal{X}$, then

$$R(\text{sign}(\delta)) - R(f^*) \leq \frac{2C_H^2}{\eta_0}\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}.$$

We conclude that in both cases it is enough to investigate the excess Bayes risk $\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*)$ with respect to the loss $\tilde{L}$. Due to Lemma 6.22, the convergence rates transfer to $R(\hat{f}_n) - R(f^*)$.
In the course of this section, we first discuss some results for general base function classes $\mathcal{C}$ and loss functions $\tilde{L}(y, s) = \phi(-ys)$. We first introduce a quantity which measures the size of function classes, the so-called *covering numbers*.

Definition 6.23 (Covering numbers)**.** Let $(E, \|\cdot\|)$ be a normed space and $W \subset E$ a subset. Then

- $\{v_1, ..., v_n\}$ is called an ε-*covering* of W if $W \subset \bigcup_{j=1}^N B_\varepsilon(v_j)$ (or equivalently, $\forall w \in W \exists j \in \{1, ..., N\} : \|w - v_j\| \leq \varepsilon$).
- $N(\varepsilon, W, \|\cdot\|) := \min\{N : \exists\ \varepsilon\text{-covering with } N \text{ elements}\}$ is called the *covering number* of W.

To prove a result for the boosting algorithm, we make the following assumptions.

Definition 6.24 (Boosting: general assumptions)**.** (A1) $\phi : \mathbb{R} \to [0, \infty)$ is convex, continuously differentiable, non-decreasing and satisfies $\phi'(0) > 0$.

(A2) There exist $c_{cov} \geq 1$, $V \geq 1$ such that the class $\mathcal{C}$ satisfies

$$N(\varepsilon,\ \mathcal{C}, \|\cdot\|_{2,n,X}) \leq (\frac{c_{cov}}{\varepsilon})^V,$$

where $\|g\|_{2,n,X} := \left(\frac{1}{n}\sum_{i=1}^n g(X_i)^2\right)^{1/2}$.

Remark: It is a relatively strong assumption to ask for a (deterministic) upper bound on $N(\varepsilon, \mathcal{C}, \|\cdot\|_{2,n,X})$ since this quantity is still a random variable. In general we can only ask that such a bound is independent of n if the functions in $\mathcal{C}$ are bounded. Here in the lecture, we only consider the following example of tree decision rules (part 1 is left as an exercise, for part 2 we refer to Corollary 10 in [9]):

Example 6.25 (Covering numbers for trees)**.**

- The class of decision stumps

$$\begin{aligned}\mathcal{C}_1 &:= \{f_T : \ T \text{ CART with depth } 2\}\\ &= \{x \mapsto y_1 \mathbb{1}_{\{x_j<s\}} + y_2 \mathbb{1}_{\{x_j \geq s\}} : j \in \{1,...,d\}, s \in \mathbb{R}, y_1, y_2 \in \mathcal{Y}\}\end{aligned}$$

  satisfies (A2) with $V = 2\lfloor 2\log_2(2d)\rfloor$, where $c > 0$ is a universal constant.

- Let $K \in \mathbb{N}$. The class of trees

$$\mathcal{C}_S = \{f_T : \ T \text{ CART with } K \text{ inner vertices}\}$$

  satisfies (A2) with $V = c_1 \cdot K\log_2(Kd)$ where $c, c_1 > 0$ are universal constants. If we consider the class of trees which are allowed to split in each dimension S times (cf. $\mathcal{T}_S$ in Definition 6.5), then we have $K = Sd$.

We now show the following result. Note that this result asks for a special form of the penalization term which then can be used in practice.

Theorem 6.26. Suppose that assumptions (A1),(A2) of Definition 6.24 hold. Define

$$P(\frac{\rho}{2}) := (\rho\phi'(\rho))^{\frac{V}{V+1}}\phi(\rho)^{\frac{1}{V+1}} + \phi(\rho).$$

Then there exists a constant $c > 0$ only dependent on $C_\phi, \phi'(0), \phi(0)$ and c_{cov} (for C_ϕ, see the Lemma 6.27 below). Then for all $t \geq 0$ the following statement holds: If

$$\lambda \geq c\Big\{\Big(\frac{(V+2)V^{1/2}}{\sqrt{n}}\Big)^{\frac{V+2}{V+1}} + \frac{t+\log(\log_2(n))}{n}\Big\}, \tag{69}$$

then

$$\mathbb{P}\Big(\tilde{R}(\hat{\delta}_n) - \tilde{R}(\delta^*) \geq 2\inf_{\delta\in\Delta}\big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} + 4P(1)\lambda\Big) \leq e^{-t}.$$

Remark: We shortly discuss the form of the penalization term (a more detailed discussion is presented below):

- For $\phi(\rho) = \log(1 + e^\rho)$ it holds that $\phi'(\rho) = \frac{e^\rho}{1+e^\rho} \leq 1$ and $\phi(\rho) \leq 1 + \rho$. Then it holds that
$$P(\frac{\rho}{2}) \leq c(\rho + 1).$$
- If $\phi(\rho) = e^\rho$, then we have $P(\frac{\rho}{2}) \leq c(\rho + 1)e^\rho$.

For logistic loss, one therefore only has penalize with $\|\delta\|_1$, where for exponential loss one has to penalize with $\approx \|\delta\|_1 e^{2\|\delta\|_1}$.

We now summarize all necessary tools for the proof. The main task is again to proof a concentration inequality for

$$V_{r,\rho}(\delta_0) := \sup_{\delta \in B(\rho)} \frac{\{\tilde{R}(\delta) - \tilde{R}_n(\delta) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0))\}}{r^2 + D(\delta, \delta_0)^2}.$$

As before, such a concentration inequality can be obtained in a straightforward way from an upper bound on the expectation of

$$Z_{r,\rho}(\delta_0) := \sup_{\delta \in B(\rho), D(\delta,\delta_0) \leq r} \{\tilde{R}(\delta) - \tilde{R}_n(\delta) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0))\}. \tag{70}$$

Here, $D(\cdot, \cdot)$ again is a distance which mimics the root of the variance of $\tilde{R}_n(\delta) - \tilde{R}_n(\delta_0)$, that is, we put

$$D(\delta, \delta^*)^2 := \mathbb{E}\big[(\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta^*(X)))^2\big].$$

The whole proof is now very similar to the SVM section. In particular, we can only prove a margin property of the distance $D(\delta, \delta^*)$ on a subspace

$$B(\rho) := \{\delta \in \Delta : \|\delta\|_1 \leq \rho\} \subset \Delta$$

with some fixed $\rho > 0$.

Lemma 6.27 (Margin property for boosting loss)**.** (i) For all $\delta \in \Delta$, it holds that $\|\delta\|_\infty \leq \|\delta\|_1$.

(ii) Put

- for $\phi = \phi_1$: $C_\phi := 0$,
- for $\phi = \phi_2$: $C_\phi := 2 - 2\log(2)$

Then it holds for all $\delta \in B(\rho)$ that

$$D(\delta, \delta^*)^2 \leq c_\rho \cdot \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\},$$

where $c_\rho := \phi(\rho) + \phi(-\rho) + C_\phi$.

Proof. 1. Let $\delta = \delta_{\beta,g}$. Then we have

$$\|\delta\|_\infty \leq \|\delta_{\beta,g}\|_\infty \leq \sum_{j=1}^{N} \beta_j \|g^{(j)}\|_\infty \leq \sum_{j=1}^{N} \beta_j.$$

Computing the infimum over all possible representations $\delta = \delta_{\beta,g}$ yields

$$\|\delta\|_\infty \leq \|\delta\|_1.$$

2. This is left as an exercise.

□

6.6 Excursus: empirical process theory

Our goal is to derive a (good) upper bound for $\mathbb{E}|Z_{r,\rho}(\delta_0)|$ in (70).
The following theorem provides upper bounds of expectations of suprema by using covering numbers. These results are taken from [10] (Theorem 4.12) and [19] (Corollary 2.2.8).
In the following, we write $\mathbb{E}_\varepsilon[\cdot] = \mathbb{E}[\cdot|Z_1, ..., Z_n]$ if the expectation is only taken with respect to $\varepsilon_1, ..., \varepsilon_n$. Thus, $\mathbb{E}_\varepsilon[\cdot]$ is still a random variable dependent on $X_1, ..., X_n$.

Lemma 6.28 (Symmetrization, contraction and entropy bounds)**.** Let Z_i, $i = 1, ..., n$ be i.i.d. random variables with values in $\mathcal{Z}$ and $\mathcal{F} \subset \{f : \mathcal{Z} \to \mathbb{R} \text{ measurable}\}$.

- *Symmetrization:* Let $\varepsilon_1, ..., \varepsilon_n$ be i.i.d. Rademacher variables (that is, $\mathbb{P}(\varepsilon_1 = 1) = \mathbb{P}(\varepsilon_1 = -1) = \frac{1}{2}$) independent of $Z_1, ..., Z_n$. Then it holds that

$$\mathbb{E} \sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{n} \{f(Z_i) - \mathbb{E}f(Z_i)\} \Big| \leq 2\mathbb{E} \sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{n} \varepsilon_i f(Z_i) \Big|.$$

- *Contraction inequality:* If $\ell_i : \mathbb{R} \to \mathbb{R}$, $i = 1, ..., n$ (here, ℓ_i may depend on $Z_1, ..., Z_n$!) is Lipschitz continuous with Lipschitz constant 1 and if $\ell_i(0) = 0$, then it holds that

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{n} \varepsilon_i \ell_i(f(Z_i)) \Big| \leq 2\ell \cdot \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{n} \varepsilon_i f(Z_i) \Big|$$

- *Entropy bound:* It holds that

$$\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \Big| \varepsilon_i f(Z_i) \Big| \leq 12 \cdot \sqrt{n} \cdot \int_0^a \sqrt{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,n})} d\varepsilon,$$

where $\|f\|_{2,n} := \left(\frac{1}{n}\sum_{i=1}^{n} f^2(Z_i)\right)^{1/2}$ and $a := \sup_{f\in\mathcal{F}} \|f\|_{2,n}$.

With these results, we can prove the following lemma (cf. [16], Corollary 3.4 and [12], Lemma 2.2).

Lemma 6.29. Let $Z_1, ..., Z_n$ be i.i.d. random variables with values in $\mathcal{Z}$ and $\mathcal{F} \subset \{f : \mathcal{Z} \to \mathbb{R} \text{ measurable}\}$. Suppose that for all $f \in \mathcal{F}$, it holds that $\mathbb{E}f(Z_1)^2 \leq \sigma^2$, $\|f\|_\infty \leq M$. Suppose that there exist $0 \leq p < 2$, $c_1, c_2 > 0$ such that

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,n}) \leq C' \cdot \varepsilon^{-p}.$$

Then there exists a universal constant $c > 0$ such that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \Big| \sum_{i=1}^n \{f(Z_i) - \mathbb{E}f(Z_i)\} \Big| \leq c \cdot \big[B n^{1/2} \sigma^{1-\frac{p}{2}} + B^{\frac{4}{2+p}} n^{\frac{p}{2+p}} M^{\frac{2-p}{2+p}} \big],$$

where $B := \frac{(C')^{1/2}}{2-p}$.

Proof. Step 1: Theorem 6.28(c) $\Rightarrow$

$$\begin{aligned} \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \Big| \sum_{i=1}^n \varepsilon_i f(Z_i) \Big| &\leq 12 \cdot \sqrt{n} \int_0^a \sqrt{\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,n})} d\varepsilon \\ &\leq 12 (C')^{1/2} \sqrt{n} \int_0^a \varepsilon^{-p/2} d\varepsilon \leq \underbrace{12 \frac{(C')^{1/2}}{1 - \frac{p}{2}}}_{=:\tilde{B}} \sqrt{n} \delta^{\frac{2-p}{2}}. \end{aligned} \tag{71}$$

Step 2: Upper bounding $\mathbb{E}a^2$: We have

$$a^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(Z_i)^2 \leq \sup_{f \in \mathcal{F}} \mathbb{E}f(Z_i)^2 + \sup_{f \in \mathcal{F}} \frac{1}{n} \Big| \sum_{i=1}^n \big\{ f(Z_i)^2 - \mathbb{E}f(Z_i)^2 \big\} \Big|$$

$\Rightarrow$

$$\begin{aligned} \mathbb{E}a^2 &\leq \sup_{f \in \mathcal{F}} \mathbb{E}f(Z_i)^2 + \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \Big| \sum_{i=1}^n \big\{ f(Z_i)^2 - \mathbb{E}f(Z_i)^2 \big\} \Big| \\ &\overset{\text{Theorem 6.28(a)}}{\leq} \sup_{f \in \mathcal{F}} \mathbb{E}f(Z_i)^2 + \frac{2}{n} \mathbb{E} \sup_{f \in \mathcal{F}} \Big| \sum_{i=1}^n \varepsilon_i f(Z_i)^2 \Big|. \end{aligned}$$

Theorem 6.28(b) implies (note that $\ell(z) = \frac{z^2}{2M}$ satisfies for $z \in [-M, M]$ that $|\ell(zz) - \ell(z')| \leq \frac{|z+z'|}{2M} \cdot |z - z'| \leq |z - z'|$)

$$\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \Big| \sum_{i=1}^n \varepsilon_i f(Z_i)^2 \Big| \leq 2 \cdot 2M \cdot \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \Big| \sum_{i=1}^n \varepsilon_i f(Z_i) \Big|.$$

We conclude that

$$\mathbb{E}a^2 \leq \sigma^2 + \frac{8M}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \varepsilon_i f(Z_i)\Big| = \sigma^2 + \frac{8M}{n}\cdot\mathbb{E}A, \tag{72}$$

where $A := \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\big|\sum_{i=1}^n \varepsilon_i f(Z_i)\big|$.
Step 3: Upper bounding $\mathbb{E}A$: Plugging in the above result into (71) yields

$$\begin{aligned}
\mathbb{E}A &= \mathbb{E}\sup_{f\in\mathcal{F}}\varepsilon_i f(Z_i) \leq \tilde{B}\sqrt{n}\mathbb{E}[(a^2)^{\frac{2-p}{4}}] \overset{\text{Jensen's ineq.}}{\leq} \tilde{B}\sqrt{n}\mathbb{E}[a^2]^{\frac{2-p}{4}} \\
&\overset{(72)}{\leq} \tilde{B}\sqrt{n}\Big(\sigma^2 + \frac{8M}{n}\cdot\mathbb{E}A\Big)^{\frac{2-p}{4}} \\
&\overset{(x+y)^q\leq x^q+y^q\,(q\leq 1)}{\leq} \tilde{B}\sqrt{n}\sigma^{\frac{2-p}{2}} + \underbrace{8^{\frac{2-p}{4}}\tilde{B}n^{\frac{p}{4}}\cdot M^{\frac{2-p}{4}}}_{=x}\cdot\underbrace{(\mathbb{E}A)^{\frac{2-p}{4}}}_{=y} \\
&\overset{\substack{\text{Young ineq. } r=\frac{4}{2-p} \\ xy\leq\frac{x^q}{q}+\frac{y^r}{r},\frac{1}{q}+\frac{1}{r}=1}}{\leq} \tilde{B}\sqrt{n}\sigma^{\frac{2-p}{2}} + \frac{2+p}{4}\big(8^{\frac{2-p}{4}}\tilde{B}n^{\frac{p}{4}}\cdot M^{\frac{2-p}{4}}\big)^{\frac{4}{2+p}} + \frac{2-p}{4}\cdot\mathbb{E}A
\end{aligned}$$

Rearranging terms and solving for $\mathbb{E}A$ yields

$$\mathbb{E}A \leq \underbrace{\frac{4}{2+p}}_{\leq 2}\cdot\Big[\tilde{B}\sqrt{n}\sigma^{\frac{2-p}{2}} + \frac{2+p}{4}8^{\frac{2-p}{2+p}}\tilde{B}^{\frac{4}{2+p}}n^{\frac{p}{2+p}}\cdot M^{\frac{2-p}{2+p}}\Big].$$

Step 4: Theorem 6.28(a) implies

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n\{f(Z_i)-\mathbb{E}f(Z_i)\}\Big| \leq 2\mathbb{E}A.$$

□

The following theorem allows to transfer an upper bound for the covering number of $\mathcal{C}$ to an upper bound of the convex hull $\mathrm{conv}(\mathcal{C})$ of $\mathcal{C}$ (cf. [19], Theorem 2.6.9):

Theorem 6.30. Suppose that (A2) of Definition 6.24 holds. Let

$$\mathrm{conv}(\mathcal{C}) := \{\sum_{j=1}^N \alpha_j g^{(j)} : N\in\mathbb{N}, \alpha_j\in[0,1], \sum_{j=1}^N \alpha_j \leq 1, g^{(j)}\in\mathcal{C}\}$$

be the convex hull of the function class $\mathcal{C}$. Then there exists a constant $c'_{cov} > 0$ only dependent on c_{cov} such that

$$\log N(\varepsilon, \mathrm{conv}(\mathcal{C}), \|\cdot\|_{2,n,X}) \leq V\cdot\Big(\frac{c'_{cov}}{\varepsilon}\Big)^{\frac{2V}{V+2}}.$$

Remark: Note that the upper bound for the covering number of $\text{conv}(\mathcal{C})$ is *much larger* than the covering number of $\mathcal{C}$ due to the additional log term on the left hand side.
Based on the above results, we can now derive an upper bound for $\mathbb{E}|Z_{r,\rho}(\delta_0)|$.

Lemma 6.31. Let $r > 0$, $\delta_0 \in B(\rho)$. Suppose that (A1),(A2) from Definition 6.24 hold. Then there exists a universal constant $c > 0$ such that

$$\mathbb{E}|Z_{r,\rho}(\delta_0)| \leq c \cdot \left[B_\rho n^{-1/2} r^{\frac{2}{V+2}} + \left(2B_\rho \phi(\rho)\right)^{\frac{1}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}}\right] =: \phi_\rho(r^2),$$

where $B_\rho := (V+2)V^{1/2}(c'_{cov}\rho\phi'(\rho))^{\frac{V}{V+2}}$.

Proof. Define

$$\mathcal{F} = \{f_\delta(x,y) := \tilde{L}(y,\delta(x)) - \tilde{L}(y,\delta_0(x)) : \delta \in B(\rho), D(\delta,\delta_0) \leq r\}.$$

Then we have

$$\mathbb{E}|Z_{r,\rho}(\delta_0)| = \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n \{f(X_i,Y_i) - \mathbb{E}f(X_i,Y_i)\}\Big|.$$

We want to apply Lemma 6.29. To do so, we have to upper bound $N(\varepsilon,\mathcal{F},\|\cdot\|_{2,n})$, where $\|f\|_{2,n} = \left(\frac{1}{n}\sum_{i=1}^n f(X_i,Y_i)^2\right)^{1/2}$. This is done by successively simplifying $\mathcal{F}$:

- $\mathcal{F}_1 = \{\tilde{L}(y,\delta(x)) = \phi(-y\delta(x)) : \delta \in B(\rho), D(\delta,\delta_0) \leq r\}$.
- $\mathcal{F}_2 = \{-y\delta(x) : \delta \in B(\rho), D(\delta,\delta_0) \leq r\}$
- $\mathcal{F}_3 = \{\delta(x) : \delta \in B(\rho), D(\delta,\delta_0) \leq r\}$ with new distance/norm $\|f\|_{2,n,X} := \left(\frac{1}{n}\sum_{i=1}^n f(X_i)^2\right)^{1/2}$
- $\mathcal{F}_4 = \{\frac{\delta(x)}{\rho} : \delta \in B(\rho), D(\delta,\delta_0) \leq r\} \subset \text{conv}(\mathcal{C})$ with $\|\cdot\|_{2,n,X}$.
  This inclusion holds since $\delta = \sum_{j=1}^N \beta_j g^{(j)} \in B(\rho)$ satisfies $\sum_{j=1}^N \beta_j \leq \rho$, that is, $\frac{\delta}{\rho} = \sum_{j=1}^N \frac{\beta_j}{\rho} g^{(j)}$ satisfies $\sum_{j=1}^N \frac{\beta_j}{\rho} \leq 1$.

Then we have

$$\begin{aligned}
& \log N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,n}) \\
\overset{(a)}{\leq} \quad & \log N(\varepsilon, \mathcal{F}_1, \|\cdot\|_{2,n}) \\
\overset{(b)}{\leq} \quad & \log N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_2, \|\cdot\|_{2,n}) \\
\overset{(c)}{\leq} \quad & \log N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_3, \|\cdot\|_{2,n,X}) \\
\overset{(d)}{\leq} \quad & \log N(\frac{\varepsilon}{\rho\phi'(\rho)}, \mathcal{F}_4, \|\cdot\|_{2,n,X}) \\
\overset{(e)}{\leq} \quad & \log N(\frac{\varepsilon}{\rho\phi'(\rho)}, \mathrm{conv}(\mathcal{C}), \|\cdot\|_{2,n,X}) \\
\overset{\text{Theorem 6.30}}{\leq} \quad & V \cdot \big(c'\rho\phi'(\rho)\big)^{\frac{2V}{V+2}} \cdot \varepsilon^{-\frac{2V}{V+2}}.
\end{aligned}$$

The justifications for (a)-(d) are left as an exercise.
For $f_\delta \in \mathcal{F}$, we have

$$\mathbb{E} f_\delta(X_i, Y_i)^2 \leq D(\delta, \delta_0)^2 \leq r^2 =: \sigma^2$$

and (since ϕ is non-decreasing)

$$\|f_\delta\|_\infty \leq 2\phi(\rho) =: M$$

Let $C' = V\big(c'\rho\phi'(\rho)\big)^{\frac{2V}{V+2}}$, $p = \frac{2V}{V+2}$ and

$$B = \frac{(C')^{1/2}}{2-p} = \frac{V+2}{4} V^{1/2} \big(c'_{cov}\rho\phi'(\rho)\big)^{\frac{V}{V+2}}.$$

Then Lemma 6.29 implies

$$\begin{aligned}
\mathbb{E}|Z_{r,\rho}(\delta_0)| \quad & \leq \quad \frac{c}{n} \cdot \big[Bn^{1/2} r^{1-\frac{p}{2}} + B^{\frac{4}{2+p}} n^{\frac{p}{2+p}} M^{\frac{2-p}{2+p}}\big] \\
& \overset{p=\frac{2V}{V+2}}{=} \quad c \cdot \big[Bn^{-1/2} r^{\frac{2}{V+2}} + \big(2B\phi(\rho)\big)^{\frac{1}{V+1}} \cdot n^{-\frac{1}{2}\frac{V+2}{V+1}}\big].
\end{aligned}$$

□

As in Section 4 (SVMs) we now apply the peeling device and a concentration inequality of Talagrand-type to obtain the following result (as before, we use $\|f_\delta\|_\infty \leq 2\phi(\rho)$).

Lemma 6.32. Let $r > 0$, $\delta_0 \in B(\rho)$. Suppose that (A1),(A2) from Definition 6.24 hold. Then there exists a universal constant $c > 0$ such that for all $t > 0$ it holds that

$$\mathbb{P}\Big(V_{r,\rho}(\delta_0) \geq c \cdot \big(\frac{\phi_\rho(r^2)}{r^2} + \sqrt{\frac{t}{nr^2}} + \frac{\phi(\rho)t}{nr^2}\big)\Big) \leq e^{-t}.$$

We already know from earlier proofs that we aim to choose r as small as possible such that still $V_{r,\rho}(\delta_0) \leq \frac{1}{Nc_\rho}$ holds with large probability. The following lemma provides the necessary lower bound on r.

Lemma 6.33. Let $\rho \geq 1$, $\delta_0 \in B(\rho)$. Suppose that (A1),(A2) from Definition 6.24 hold. Then for each $N \in \mathbb{N}$, there exists a constant $c(N) > 0$ only dependent on $C_\phi, \phi'(0), c'$ such that any

$$r^2 \geq r_{\rho,t}^2 := c(N) \cdot c_\rho \Big[((V+2)V^{1/2})^{\frac{V+2}{V+1}} \cdot (\rho\phi'(\rho))^{\frac{V}{V+1}} \phi(\rho)^{\frac{1}{V+1}} \cdot n^{-\frac{1}{2}\frac{V+2}{V+1}} + \phi(\rho) \cdot \frac{t}{n} \Big]$$

satisfies

$$\mathbb{P}\big(V_{r,\rho}(\delta_0) \geq \frac{1}{Nc_\rho}\big) \leq e^{-t}.$$

Proof. To follow the assertion from Lemma 6.32, we have to guarantee that

$$\frac{1}{Nc_\rho} \geq c \cdot \big(\frac{\phi_\rho(r^2)}{r^2} + \sqrt{\frac{t}{nr^2}} + \frac{\phi(\rho)t}{nr^2}\big).$$

This is implied by

$$\begin{aligned} & \frac{1}{3Nc_\rho} \geq c\frac{\phi_\rho(r^2)}{r^2}, \quad \frac{1}{3Nc_\rho} \geq \sqrt{\frac{t}{nr^2}}, \quad \frac{1}{3Nc_\rho} \geq \frac{\phi(\rho)t}{nr^2} \\ \Leftarrow \quad & \frac{1}{6Nc_\rho} \geq cB_\rho n^{-\frac{1}{2}} r^{-\frac{2(V+1)}{V+2}}, \quad \frac{1}{6Nc_\rho} \geq \frac{c}{r^2}\big(2B_\rho\phi(\rho)\big)^{\frac{1}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}}, \\ & \qquad\qquad \frac{1}{3Nc_\rho} \geq \sqrt{\frac{t}{nr^2}}, \quad \frac{1}{3Nc_\rho} \geq \frac{\phi(\rho)t}{nr^2}, \end{aligned}$$

which is furthermore implied by

$$\begin{aligned} r^2 \quad \geq \quad \max\Big\{ & (6cN)^{\frac{V+2}{V+1}} (c_\rho B_\rho)^{\frac{V+2}{V+1}} \cdot n^{-\frac{1}{2}\frac{V+2}{V+1}}, \quad 6cNc_\rho\big(2B_\rho\phi(\rho)\big)^{\frac{1}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}}, \\ & (3Nc_\rho)^2\frac{t}{n}, \quad (3N)\cdot c_\rho\phi(\rho)\frac{t}{n}\Big\}. \end{aligned}$$

Using $\phi(\rho) \leq c_\rho$ and choosing $\tilde{c}(N)$ (only dependent on N) large enough, we can summarize these conditions to

$$r^2 \geq \tilde{c}(N) \cdot c_\rho \cdot \Big[\big(B_\rho c_\rho\big)^{\frac{1}{V+1}} \cdot \big\{B_\rho + 1\big\} n^{-\frac{1}{2}\frac{V+2}{V+1}} + c_\rho\frac{t}{n}\Big]. \tag{73}$$

We now upper bound the right hand side even further to simplify the expression a little bit more. We do this by summarizing the constants $c'_{cov}, C_\phi, \phi'(0)$ into $\tilde{c}(N)$. It holds that

$$c_\rho \leq (2 + C_\phi) \cdot \phi(\rho),$$

and for $\rho \geq 1$, we have $\rho\phi'(\rho) \geq \phi'(1) \geq \phi'(0)$. Thus

$$B_\rho + 1 \leq \text{const.}(c', \phi'(0)) \cdot (V+2)V^{1/2}(\rho\phi'(\rho))^{\frac{V}{V+2}}.$$

Plugging in these estimates into (73) yields that (73) is implied by

$$r^2 \geq r_{\rho,t}^2 := c(N) \cdot c_\rho \Big[((V+2)V^{1/2})^{\frac{V+2}{V+1}} \cdot (\rho\phi'(\rho))^{\frac{V}{V+1}} \phi(\rho)^{\frac{1}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}} + \phi(\rho) \cdot \frac{t}{n} \Big].$$

□

6.7 Proof of the boosting oracle inequality

Proof. [Proof of Theorem 6.26] Let

$$\delta_0 \in \arg\min_{\delta \in \Delta} \big\{ \tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda \cdot P(\|\delta\|_1) \big\}.$$

Discretization: As in the proof of the SVM algorithm, discretize the radii via

$$\begin{aligned} \hat{\rho} &:= 2^{\hat{k}}, \qquad \hat{k} = \lceil \log(\|\hat{\delta}\|_1)_+ \rceil, \\ \rho_0 &:= 2^{k_0}, \qquad k_0 = \lceil \log(\|\delta_0\|_1)_+ \rceil. \end{aligned}$$

Derivation of the maximal radius: It holds that

$$\tilde{R}_n(\hat{\delta}) + \lambda \cdot P(\|\hat{\delta}\|_1) \leq \tilde{R}_n(0) + \lambda \cdot P(\|0\|_1) = \phi(0) + \lambda\phi(0),$$

thus $P(\|\hat{\delta}\|_1) \leq \frac{\phi(0)+2\lambda\phi(0)}{\lambda} \overset{\lambda \geq n^{-1}}{\leq} 2\phi(0)n$.
Convexity of ϕ implies that $\phi(x) \geq \phi'(0)x + \phi(0) \geq \phi'(0)x$, thus

$$\phi(0)n \geq P(\|\hat{\delta}\|_1) \geq \rho(2\|\hat{\delta}\|_1) \geq 2\phi'(0)\|\hat{\delta}\|_1 \quad \Rightarrow \quad \frac{\phi(0)}{\phi'(0)} n \geq \|\hat{\delta}\|_1.$$

Similarly, we have $P(\|\delta_0\|_1) \leq \frac{\phi(0)}{\phi'(0)} n$.
Thus

$$\hat{\rho}, \rho_0 \in \mathcal{R} := \{2^k : k \in \{0, ..., \lceil \log_2(\frac{\phi(0)}{\phi'(0)} n) \rceil\}\}.$$

By defining $\tilde{\rho} := \max\{\hat{\rho}, \rho_0\}$, we have

$$\|\hat{\delta}\|_1 \leq \hat{\rho}, \quad \|\delta_0\|_1 \leq \rho_0 \quad \Rightarrow \quad \hat{\delta}, \delta_0 \in B(\tilde{\rho}),$$

and vice versa,

$$\hat{\rho} \leq 2\max\{\|\hat{\delta}\|_1, 1\}, \qquad \rho_0 \leq 2\max\{\|\delta_0\|_1, 1\}. \tag{74}$$

Definition of a 'nice' event A: Put

$$A := \bigcap_{\rho\in\mathcal{R},\rho\geq\rho_0} \{V_{r_{\rho,\tilde{t}},\rho}(\delta_0) \leq \frac{1}{c_\rho N}\}.$$

Lemma 6.32 implies that

$$\mathbb{P}(A^c) \leq \sum_{\rho\in\mathcal{R},\rho\geq\rho_0} \mathbb{P}\Big(V_{r_{\rho,\tilde{t}},\rho}(\delta_0) \geq \frac{1}{c_\rho N}\Big) \leq |\mathcal{R}|\cdot e^{-\tilde{t}} \leq e^{-t}.$$

Derivation of the upper bound of excess Bayes risk on A: Standard techniques yield

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta) \leq \tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) + \tilde{R}(\delta_0) - \tilde{R}(\delta^*), \tag{75}$$

and (with $\tilde{t} := t + \log(|\mathcal{R}|)$)

$$\begin{aligned}\tilde{R}(\hat{\delta}) - \tilde{R}(\delta_0) &\leq \tilde{R}_n(\hat{\delta}) - \tilde{R}_n(\delta_0) + \big\{\tilde{R}(\hat{\delta}) - \tilde{R}_n(\hat{\delta}) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0))\big\} \\ &\overset{\hat{\delta},\delta_0\in B(\tilde{\rho})}{\leq} \big\{\lambda P(\|\delta_0\|_1) - \lambda P(\|\hat{\delta}\|_1)\big\} + V_{r_{\tilde{\rho},\tilde{t}},\tilde{\rho}}(\delta_0)\cdot\big(r^2_{\tilde{\rho},\tilde{t}} + D(\hat{\delta},\delta_0)^2\big) \end{aligned} \tag{76}$$

Lemma 6.27(ii) implies

$$\begin{aligned} D(\hat{\delta},\delta_0)^2 &\leq \big[D(\hat{\delta},\delta^*) + D(\delta_0,\delta^*)\big]^2 \leq 2D(\hat{\delta},\delta^*)^2 + 2D(\delta_0,\delta^*)^2 \\ &\overset{\hat{\delta},\delta_0\in B(\tilde{\rho})}{\leq} 2c_{\tilde{\rho}}\big\{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) + \tilde{R}(\delta_0) - \tilde{R}(\delta^*)\big\}. \end{aligned} \tag{77}$$

On A we conclude from (75), (76) and (77) that

$$\begin{aligned}\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \;\leq\; & (1+2N^{-1})\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + 2N^{-1}\{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*)\} \\ & + \big\{\lambda P(\|\delta_0\|_1) - \lambda P(\|\hat{\delta}\|_1)\big\} + \frac{1}{N}\frac{r^2_{\tilde{\rho},\tilde{t}}}{c_{\tilde{\rho}}}.\end{aligned}$$

Rearranging terms and solving for $\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*)$ yields

$$\begin{aligned}\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \;\leq\; & \frac{1}{1-2N^{-1}}\Big[(1+2N^{-1})\cdot\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} \\ & + \big\{\lambda P(\|\delta_0\|_1) - \lambda P(\|\hat{\delta}\|_1)\big\} + \frac{1}{N}\frac{r^2_{\tilde{\rho},\tilde{t}}}{c_{\tilde{\rho}}}\Big].\end{aligned} \tag{78}$$

We now discuss the second summand on the right hand side.
Lemma 6.33, Definition $P(\rho/2)$ amd λ, $\log(|\mathcal{R}|) \leq \log(\log_2(\frac{\phi(0)}{\phi'(0)}n)+1) \leq c(\phi(0),\phi'(0))\log(\log_2(n))$ yield

$$\frac{r^2_{\rho,t}}{Nc_\rho} \leq \frac{c(N)}{N}\big\{\big(\frac{(V+2)V^{1/2}}{\sqrt{n}}\big)^{\frac{V+2}{V+1}} + \frac{\tilde{t}}{n}\big\}\cdot\big\{(\rho\phi'(\rho))^{\frac{V}{V+1}}\phi(\rho)^{\frac{1}{V+1}} + \phi(\rho)\big\} \leq \lambda\cdot P(\frac{\rho}{2}).$$

Since $\rho \mapsto P(\rho)$ is non-decreasing, we conclude that

$$\begin{aligned}\frac{r_{\tilde{\rho},t}^2}{Nc_{\tilde{\rho}}} &\leq \lambda \cdot \max\{P(\frac{\hat{\rho}}{2}), P(\frac{\rho_0}{2})\} \leq \lambda \cdot \{P(\frac{\hat{\rho}}{2}) + P(\frac{\rho_0}{2})\} \\ &\overset{(74)}{\leq} \lambda \cdot \big\{ \max\{P(\|\hat{\delta}\|_1), P(1)\} + \max\{P(\|\delta_0\|_1), P(1)\} \\ &\leq \lambda P(\|\hat{\delta}\|_1) + \lambda P(\|\delta_0\|_1) + 2\lambda P(1).\end{aligned}$$

Plugging in this into (78) yields the final result,

$$\begin{aligned}\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) &\leq \frac{1}{1-2N^{-1}}\Big[(1+2N^{-1}) \cdot \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + 2\lambda P(\|\delta_0\|_1) + 2\lambda P(1)\Big] \\ &\overset{N \text{ large enough}}{\leq} 2\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta_0\|_1)\} + 4\lambda P(1) \\ &\overset{\text{Def. } \delta_0}{=} 2 \inf_{\delta \in \Delta} \big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} + 4\lambda P(1).\end{aligned}$$

□

We will now introduce a model assumption for the optimal decision boundary. This is done to derive an explicit convergence rate of the excess Bayes risk based on the oracle inequality from Theorem 6.26. The model assumption we introduce is dependent on the chosen base class $\mathcal{C}$. Here, we only consider decision stumps.

Definition 6.34 (Model assumption: Boosting with decision stumps)**.** Let $\mathcal{C} = \mathcal{C}_1$ and $\phi \in \{\phi_1, \phi_2\}$. Let $B > 0$ and suppose that there exist measurable functions $h_1, ..., h_d : [0,1] \to \mathbb{R}$ such that

$$\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) = \sum_{j=1}^{d} h_j(x_j) \tag{79}$$

and $|h_j(0)| + |h_j(1)| + |h_j|_{BV} \leq B$ (here, $|\cdot|_{BV}$ denotes the variation of a function).

Remarks:

- With decision stumps, we can approximate a one-dimensional function arbitrarily well by glueing decision stumps together. However, sums of decision stumps are not able introduce *interactions* between different coordinates. Therefore, we can only hope to approximate structures of the form (79) with boosting of decision stumps. This is also clear if we write down the explicit expression of a $\delta \in \Delta$:

  $$\delta(x) = \sum_{j=1}^{N} \beta_j g_j(x) = \sum_k a_{1,k}\big\{\mathbb{1}_{\{x_1 < s_{1,k}\}} - \mathbb{1}_{\{x_1 \geq s_k\}}\big\} + ... + \sum_k a_{d,k}\big\{\mathbb{1}_{\{x_d < s_k\}} - \mathbb{1}_{\{x_d \geq s_{d,k}\}}\big\}$$

  with suitably chosen $a_{j,k}, s_{j,k} \in \mathbb{R}$.

- Consider the more general class $\mathcal{C}$ of trees with 2 inner vertices. The elements of $\mathcal{C}$ have the form

$$\big(y_1 \mathbb{1}_{\{x_j<s\}} + y_2 \mathbb{1}_{\{x_j \geq s\}}\big) \cdot \big(\tilde{y}_1 \mathbb{1}_{\{x_{\tilde{j}}<\tilde{s}\}} + \tilde{y}_2 \mathbb{1}_{\{x_{\tilde{j}} \geq \tilde{s}\}}\big),$$

  in particular, there occur products $\mathbb{1}_{\{x_j<s\}} \cdot \mathbb{1}_{\{x_j<\tilde{s}\}}$ along two different coordinates. The corresponding model assumption reads

$$\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) = \sum_{j,k=1}^{d} h_{j,k}(x_j, x_k),$$

  that is, we can now approximate sums of functions which depend on 2 variables.

- Continuing this scheme shows that boosting of trees with d inner vertices can produce arbitrary functions starting from d dimensions.

Theorem 6.35. Suppose that the model assumption of Definition 6.34 holds. Suppose that $\phi \in \{\phi_1, \phi_2\}$. Then it holds that

$$\inf_{\delta \in \Delta} \big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} \leq 2\lambda P(\frac{Bd}{2}).$$

Let λ be chosen with equality in (69). Then with some universal constant $c > 0$, it holds that

$$\inf_{\delta \in \Delta} \big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} \leq c(1+t)\log(d+1)^3 \cdot n^{-\frac{1}{2}\frac{V+2}{V+1}} \cdot \begin{cases} (Bd+1)e^{Bd}, & \phi = \phi_1 \\ (Bd+1), & \phi = \phi_2, \end{cases}$$

where $V = 2\lfloor 2\log(2d)\rfloor$.

Proof. This is left as an exercise. □

Remarks:

- Even for high dimensions, the convergence rate is at most $n^{-\frac{1}{2}}$ in terms of n since $\frac{V+2}{V+1} \to 1$ $(d \to \infty)$.
- In the case of the exponential loss, one can not hope for a good behavior of the algorithm for large dimensions due to the factor e^{Bd}. For logistic loss, one only has a factor of size Bd and therefore can expect a much better behavior. Based on our theoretical results, we can therefore not hope for a good behavior of boosting

if $d \gg n$ and all functions $h_1, ..., h_d$ in eq. (79) enter $\log(\frac{\eta(x)}{1-\eta(x)})$ with the same strength. Note that the factor d is however reasonable since we have to estimate d different quantities.

- In practice, one observes that also boosting with exponential loss works quite well. One should always remember that we only prove *upper bounds* for the excess Bayes risk in this lecture under rather general assumptions. Under more specific assumptions, one may show much better results and of course there is also a chance that some of our inequalities are not tight with respect to the influence of the dimension d.

Finally, suppose that instead of Definition 6.34, we have that $\log(\frac{\eta(x)}{1-\eta(x)})$ only decomposes into $s \ll d$ summands, that is, only s summands in (79) are nonzero in the way that

$$\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) = \sum_{j \in S} h_j(x_j)$$

with $S \subset \{1, ..., d\}$, $|S| = s \ll d$. Then we obtain a similar result with

$$\inf_{\delta \in \Delta} \big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} \leq c(1+t)\log(d+1)^3 n^{-\frac{1}{2}\frac{V+2}{V+1}} \cdot \begin{cases} (Bs+1)e^{Bs}, & \phi = \phi_1 \\ (Bs+1), & \phi = \phi_2, \end{cases}.$$

This shows that the *original dimension* d only enters the rate with $\log(d)$ through λ, while the reduced dimension s produces much smaller factors $(Bs+1)e^{Bs}$ or $(Bs+1)$, respectively.

6.8 Exercises

Task 21 (A maximal inequality for finite function classes)**.** Let $Z_1, ..., Z_n$ be i.i.d. random variables on $\mathcal{X} \subset \mathbb{R}^d$. Let $\mathcal{G} \subset \{g : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$ be a finite function class. Suppose that there exist $\sigma^2, M > 0$ such that for all $g \in \mathcal{G}$, it holds that $\mathbb{E}g(Z_1) = 0$, $\mathrm{Var}(g(Z_1)) \leq \sigma^2$ and $\|g\|_\infty \leq M$.

1. Put $W(g) := \big|\sum_{i=1}^n g(Z_i)\big|$. Show with Bernstein's inequality: It holds that

$$\mathbb{P}(W(g) \geq t) \leq \begin{cases} 2\exp\Big(-\frac{t^2}{4n\sigma^2}\Big), & t < \frac{3\sigma^2 n}{M}, \\ 2\exp\Big(-\frac{3t}{4M}\Big), & t \geq \frac{3\sigma^2 n}{M}. \end{cases}$$

2. Define

$$A_1(g) := W(g)\mathbb{1}_{\{W(g) > \frac{3\sigma^2 n}{M}\}}, \qquad A_2(g) := W(g)\mathbb{1}_{\{W(g) \leq \frac{3\sigma^2 n}{M}\}},$$

and $\psi_p(x) = \exp(x^p) - 1$ für $p \in \{1, 2\}$. Show that

$$\mathbb{E}\psi_1\big(\frac{A_1(g)}{4M}\big) \leq 1.$$

Hint: Write $\psi_1(x) = \int_0^x e^t dt = \int_0^\infty \mathbb{1}_{\{x \geq t\}} e^t dt$, then apply the expected value. Note that $\mathbb{P}(A_1(g) \geq z) \leq \mathbb{P}(W(g) \geq \max\{z, \frac{3\sigma^2 n}{M}\})$, thus one can apply the first inequality from (a) and resubstituting $\max\{z, \frac{3\sigma^2 n}{M}\}$ by z afterwards.

3. Show that
$$\psi_1(\mathbb{E}\sup_{g \in \mathcal{G}} \frac{A_1(g)}{4M}\big) \leq |\mathcal{G}|.$$
Hint: Jensen's inequality and (b).

4. Conclude that
$$\mathbb{E}\sup_{g \in \mathcal{G}} \frac{A_1(g)}{4M} \leq \log(|\mathcal{G}| + 1).$$

5. Show that $\mathbb{E}\psi_2\big(\frac{A_2(g)}{2\sigma\sqrt{n}}\big) \leq 1$.
Hint: Write $\psi_2(x) = \int_0^{x^2} e^t dt = \int_0^\infty \mathbb{1}_{\{x \geq \sqrt{t}\}} e^t dt$ and $\mathbb{P}(A_2(g) \geq z) \leq \mathbb{P}(W(g) \geq \min\{z, \frac{3\sigma^2 n}{M}\})$.

6. Conclude that
$$\mathbb{E}\sup_{g \in \mathcal{G}} \frac{A_2(g)}{2\sigma\sqrt{n}} \leq \sqrt{\log(|\mathcal{G}| + 1)}.$$

7. Conclude from (d) and (f):
$$\mathbb{E}\sup_{g \in \mathcal{G}} W(g) \leq 4\big\{\sigma\sqrt{n}\sqrt{\log(|\mathcal{G}| + 1)} + M \cdot \log(|\mathcal{G}| + 1)\big\}.$$

Task 22 (Proof of Theorem 6.9)**.** Let $\mathcal{X} = [0, 1]^d$ and $S \in \mathbb{N}$. Denote by μ the Lebesgue measure on $\mathbb{R}^d$. Suppose that there are constants $c_\mu, c_{box} > 0$ such that

(A1) For all $A \subset \mathcal{X}$, it holds that $\mathbb{P}(X \in A) \leq c_0 \mu(A)$.

(A2) For $m = 2^S$, it holds that the optimal decision boundary $\partial\Omega_1^*$ only slices through at most $c_{box} m^{d-1}$ of the m^d cubes in which the space $\mathcal{X}$ decomposes.

Let $T_0 \in \mathcal{T}_S$ be a dyadic tree which is obtained as follows. Repeat S times: Split each vertex in dimension number 1,...., split each vertex in dimension number d.

1. Show that $R(f_{T_0}) - R(f^*) \leq \mathbb{P}(f_{T_0}(X) \neq f^*(X))$.

2. Suppose that (A1) and (A2). Let $\mathcal{C}$ denote the set of cubes which have a nonempty intersection with $\partial\Omega_1^*$. Show that

$$\mathbb{P}(f_{T_0}(X) \neq f^*(X)) \leq c_\mu c_{box} m^{-1}.$$

3. Show that $|T_0| \leq m^d$.

4. Conclude that

$$\inf_{T \in \mathcal{T}_S} \left\{ R(f_T) - R(f^*) + \lambda \cdot |T| \right\} \leq c_\mu c_{box} m^{-1} + \lambda m^d.$$

5. Suppose that $S \geq \frac{\log_2(n)}{d+1}$ and $\lambda = c \cdot \frac{\log(d)+t}{\eta_0 n}$. Show that

$$\inf_{T \in \mathcal{T}_S} \left\{ R(f_T) - R(f^*) + \lambda \cdot |T| \right\} \leq (2c_\mu c_{box} + \frac{c}{\eta_0}(\log(2d) + t)) \cdot n^{-\frac{1}{d+1}}.$$

Task 23 (Covering Numbers of decision stumps)**.** Let $\mathcal{X} \subset [0,1]^d$ and $\mathcal{Y} = \{-1, +1\}$. In this task we aim to show that the class of decision stumps,

$$\begin{aligned} \mathcal{C} &:= \{f_T : T \text{ CART with one inner vertex } \} \\ &= \{y_1 \mathbb{1}_{\{x_j < s\}} + y_2 \mathbb{1}_{\{x_j \geq s\}} : y_1, y_2 \in \mathcal{Y}, s \in \mathbb{R}, j \in \{1, ..., d\}\} \end{aligned}$$

satisfies

$$N(\varepsilon, \mathcal{C}, \|\cdot\|_{2,n,X}) \leq (\frac{c_{cov}}{\varepsilon})^V, \qquad \|g\|_{2,n,X} := \Big(\frac{1}{n}\sum_{i=1}^n g(X_i)^2\Big)^{1/2}, \qquad (*)$$

where $V = 2 \cdot \lfloor 2\log_2(2d) \rfloor$ and c_{cov} are universal constants. To do so, we use the following result: If $\mathcal{V}(\mathcal{C})$ is the VC-Dimension of $\mathcal{C}$, then it holds that

$$N(\varepsilon, \mathcal{C}, \|\cdot\|_{2,n,X}) \leq 13 \cdot \mathcal{V}(\mathcal{C}) \cdot (\frac{4e}{\varepsilon^2})^{\mathcal{V}(\mathcal{C})}.$$

Here, $\mathcal{V}(\mathcal{C})$ is defined as follows:

$$\mathcal{V}(\mathcal{C}) = \inf\{N \in \mathbb{N} : m_{\mathcal{C}}(N) < 2^N\},$$

where

$$m_{\mathcal{C}}(N) := \max_{x_1,...,x_N \in [0,1]^d} m_{\mathcal{C}}(x_1, ..., x_N), \qquad m_{\mathcal{C}}(x_1, ..., x_N) := |\{(f(x_1), ..., f(x_N)) : f \in \mathcal{C}\}|$$

denotes the maximal number of different labelings of N points in the space $[0,1]^d$ which can be obtained via $\mathcal{C}$.

1. Show that if $\mathcal{V}(\mathcal{C}) \leq \lfloor 2\log_2(2d) \rfloor$, then one has (*).

2. Let $d = 1$. Show that
$$m_{\mathcal{C}}(N) = 2N.$$

3. Let d be arbitrary. Show that
$$m_{\mathcal{C}}(N) \leq \min\{2Nd, 2^N\}.$$

4. Show that for $d \geq 1$, it holds that: $\mathcal{V}(\mathcal{C}) \leq \lfloor 2\log_2(2d) \rfloor$.
Hint: $d - 1 \geq \log_2(d)$ *for all* $d \geq 1$.

Task 24 (Proof of the margin property, Lemma 6.27(ii))**.** Let $\phi \in \{\phi_1(x) = e^x, \phi_2(x) = \log(1 + e^x)\}$ be non-decreasing, continuously differentiable and convex. Let $\tilde{L}(y, s) = \phi(-ys)$. We show that for all $\delta : \mathcal{X} \to \mathbb{R}$ with $\|\delta\|_\infty \leq \rho$ it holds that
$$\mathbb{E}[(\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta^*(X)))^2] \leq c_\rho \cdot \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\},$$
where $\delta^* \in \arg\min_{\delta:\mathcal{X}\to\mathbb{R}} \tilde{R}(\delta)$, $\tilde{R}(\delta) = \mathbb{E}\tilde{L}(Y, \delta(X))$ and $c_\rho = \phi(\rho) + \phi(-\rho) + C_\phi$, where
$$C_\phi = \begin{cases} 0, & \phi = \phi_1 \\ 2 - 2\log(2), & \phi = \phi_2 \end{cases}.$$
To do so, we proceed as follows.

1. Show that $\delta^*(x) = g(\eta(x))$, where $g(\eta)$ satisfies (cf. the risk transfer formula from Theorem 3.21):
$$0 = \phi'(-g(\eta)) \cdot \eta + \phi'(g(\eta)) \cdot (1 - \eta).$$

2. Show that
$$\begin{aligned} \mathbb{E}[(\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta^*(X)))^2 | X = x] &= A_2(\eta(x), \delta(x)), \\ \mathbb{E}[\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta^*(X)) | X = x] &= A_1(\eta(x), \delta(x)), \end{aligned}$$
and determine the functions A_1, A_2.
Hint: You should obtain $A_1(\eta, \delta) = \eta(\phi(-\delta) - \phi(-g(\eta))) + (1-\eta)(\phi(\delta) - \phi(g(\eta)))$.

3. Show with (a) that
$$\partial_\eta A_1(\eta, \delta) = \big[\phi(-\delta) - \phi(-g(\eta))\big] - \big[\phi(\delta) - \phi(g(\eta))\big].$$

4. *Caution: This task is tedious.* Show with (a) that
$$\partial_\eta A_2(\eta, \delta) = \partial_\eta A_1(\eta, \delta) \cdot (\phi(\delta) + \phi(-\delta) + B(\eta)),$$
where
$$B(\eta) := \big\{\big\{\eta\phi'(-g(\eta)) + (1 - \eta)\phi'(g(\eta))\big\}g'(\eta) - \phi(g(\eta)) - \phi(-g(\eta))\big\}.$$

5. Put $C_\phi := 0 \vee \max_{\eta\in[0,1]} B(\eta)$. Show that if $g(\eta) \geq \delta$, then one has

$$\partial_\eta A_1(\eta, \delta) \leq \{\phi(\rho) + \phi(-\rho) + C_\phi\} \cdot \partial_\eta A_2(\eta, \delta).$$

   Otherwise, the above inequality holds with '$<$' instead.
   Hint: It holds that $\phi(\delta) + \phi(-\delta) \leq \phi(\rho) + \phi(-\rho)$. Why?

6. Let g^{-1} denote the inverse function of g. Show that $A_j(g^{-1}(\delta), \delta) = 0$, $j = 1, 2$.

7. Show that $g(\eta) \geq \delta$ implies

$$A_2(\eta, \delta) \leq \{\phi(\rho) + \phi(-\rho) + C_\phi\} \cdot A_1(\eta, \delta).$$

   A similar result is obtained for $g(\eta) < \delta$. *Hint: Integration.*

8. Derive the values given for C_ϕ given above from (e).

Task 25 (Proof of Lemma 6.31 / transfer of Covering Numbers)**.** Let $\tilde{L}(y, s) = \phi(-ys)$ with some convex and non-decreasing, continuously differentiable ϕ. Let

$$\mathcal{F} = \{\tilde{L}(y, \delta(x)) - \tilde{L}(y, \delta_0(x)) : \delta \in B(\rho), D(\delta, \delta_0) \leq r\}$$

be the function class from Lemma 6.31 (in the following, we omit the additional condition $D(...) \leq r$). Let $\|f\|_{2,n} = \left(\frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)^2\right)^{1/2}$. Define

- $\mathcal{F}_1 = \{\tilde{L}(y, \delta(x)) = \phi(-y\delta(x)) : \delta \in B(\rho)\}$,
- $\mathcal{F}_2 = \{-y\delta(x) : \delta \in B(\rho)\}$,
- $\mathcal{F}_3 = \{\delta(x) : \delta \in B(\rho)\}$ with new distance measure $\|f\|_{2,n,X} := \left(\frac{1}{n}\sum_{i=1}^n f(X_i)^2\right)^{1/2}$,
- $\mathcal{F}_4 = \{\frac{\delta(x)}{\rho} : \delta \in B(\rho)\}$.

Show that

1. $N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,n}) \leq N(\varepsilon, \mathcal{F}_1, \|\cdot\|_{2,n})$,
2. $N(\varepsilon, \mathcal{F}_1, \|\cdot\|_{2,n}) \leq N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_2, \|\cdot\|_{2,n})$,
3. $N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_2, \|\cdot\|_{2,n}) \leq N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_3, \|\cdot\|_{2,n,X})$, where $\|f\|_{2,n,X} := (\frac{1}{n}\sum_{i=1}^n f(X_i)^2)^{1/2}$.
4. $N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_3, \|\cdot\|_{2,n,X}) \leq N(\frac{\varepsilon}{\rho\phi'(\rho)}, \mathcal{F}_4, \|\cdot\|_{2,n,X})$.

Task 26 (Approximations of functions with bounded variation)**.** Let $g : [0,1] \to \mathbb{R}$ be a function with bounded variation $|g|_{BV} < \infty$. Here, $|g|_{BV}$ is defined as follows:

$$|g|_{BV} = \sup_{M \in \mathbb{N}} \sup_{0 \le s_0 \le \ldots \le s_M \le 1} \sum_{i=1}^{M} \big|g(s_i) - g(s_{i-1})\big|.$$

From analysis it is know that there exist non-decreasing functions $u, v : [0,1] \to \mathbb{R}$ with $g = u - v$ and $|g|_{BV} = |u|_{BV} + |v|_{BV}$. Let $N \in \mathbb{N}$.

1. Define $C := u(1) - u(0)$. Show that $C = |u|_{BV}$.

2. Let $t_i := \sup\{z \in [0,1] : u(z) - u(0) \le C\frac{i}{N}\}$, $i = 0, ..., N$ be the 'quantiles' of u. Define
$$\tilde{u} : [0,1] \to \mathbb{R}, \qquad \tilde{u}(z) := u(0) + \sum_{i=1}^{N} \frac{C}{N} \mathbb{1}_{\{z \ge t_i\}}.$$
Show that
$$\|u - \tilde{u}\|_\infty \le \frac{|u|_{BV}}{N}.$$
Hint: Consider the expression for $z \in [t_j, t_{j+1})$, $j = 0, ..., N-1$.

3. Show that $\tilde{u}$ has the alternative representation
$$\tilde{u}(z) = \frac{u(0) + u(1)}{2}(\mathbb{1}_{\{z \ge 0\}} - \mathbb{1}_{\{z < 0\}}) + \sum_{i=1}^{N} \frac{C}{2N}(\mathbb{1}_{\{z \ge t_i\}} - \mathbb{1}_{\{z < t_i\}}).$$

4. Conclude that there exists a function $\tilde{g}$ of the form
$$\begin{aligned} \tilde{g}(z) \;=\; & \frac{g(0) + g(1)}{2}(\mathbb{1}_{\{z \ge 0\}} - \mathbb{1}_{\{z < 0\}}) \\ & + \sum_{i=1}^{N} \frac{|u|_{BV}}{2N}(\mathbb{1}_{\{z \ge t_i\}} - \mathbb{1}_{\{z < t_i\}}) - \sum_{i=1}^{N} \frac{|v|_{BV}}{2N}(\mathbb{1}_{\{z \ge q_i\}} - \mathbb{1}_{\{z < q_i\}}) \end{aligned}$$
with $0 \le q_0 \le .. \le q_N \le 1$ and the property $\|g - \tilde{g}\|_\infty \le \frac{|g|_{BV}}{N}$.

Task 27 (Convergence rate of boosting with decision stumps)**.** Let $\mathcal{X} \subset [0,1]^d$ and $\mathcal{Y} = \{-1, +1\}$. In this task we derive upper bounds on

$$\inf_{\delta \in \Delta} \big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\}, \qquad \Delta = \{\sum_{j=1}^{M} \beta_j g_j : M \in \mathbb{N}, \beta_j \ge 0, g_j \in \mathcal{C}\}$$

based on the class of decision stumps

$$\mathcal{C} := \{y_1 \mathbb{1}_{\{x_j<s\}} + y_2 \mathbb{1}_{\{x_j \geq s\}} : y_1, y_2 \in \mathcal{Y}, s \in \mathbb{R}, j \in \{1, ..., d\}\}$$

and the exponential or logistic loss $\tilde{L}(y, s) = \phi(-ys)$, $\phi \in \{\phi_1(x) = e^x, \phi_2(x) = \log(1 + e^x)\}$, respectively (cf. Theorem 6.35). To do so, we consider the model assumption from Definition 6.34 which states that there exists a decomposition

$$\delta^*(x) = \sum_{j=1}^{d} h_j(x_j)$$

with measurable functions $h_1, ..., h_d : [0, 1] \to \mathbb{R}$ with $|h_j(0)| + |h_j(1)| + |h_j|_{BV} \leq B$, $B > 0$.

1. Let $N \in \mathbb{N}$. Show with Task 26 that there exist $\tilde{h}_j : [0, 1] \to \mathbb{R}$, $j \in \{1, ..., d\}$ with

$$\|h_j - \tilde{h}_j\|_\infty \leq \frac{B}{N}, \qquad \|\tilde{h}_j\|_1 \leq \frac{B}{2}.$$

2. Conclude that $\tilde{h}_N(x) := \sum_{j=1}^{d} \tilde{h}_j(x_j) \in \Delta$ satisfies

$$\|\delta^* - \tilde{h}_N\|_\infty \leq \frac{Bd}{N}, \qquad \|\tilde{h}_N\|_1 \leq \frac{Bd}{2}.$$

3. Show that

$$\inf_{\delta \in \Delta} \left\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\right\} \leq 2\lambda P(\frac{Bd}{2}).$$

4. Let $\lambda = c \cdot (1 + t)((V + 2)V^{1/2})^{\frac{V+2}{V+1}} n^{-\frac{1}{2}\frac{V+2}{V+1}}$ be defined as in Theorem 6.26, where $V = 2\lfloor 2 \log_2(2d) \rfloor$. Show that with some universal constant $c' > 0$, it holds that

$$\inf_{\delta \in \Delta} \left\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\right\} \leq c'(1+t) \log(d+1)^3 \cdot n^{-\frac{1}{2}} \cdot \begin{cases} (Bd + 1)e^{Bd+1}, & \phi = \phi_1, \\ (Bd + 1), & \phi = \phi_2. \end{cases}$$

5. *Discussion:* What changes if instead, δ^* has the form $\delta^*(x) = \sum_{j=1}^{s} h_j(x_j)$ with some $1 \leq s \leq d$?

7 Neural networks

Similar to trees, neural networks are nonparametric algorithms which can be used both for regression and classification problems. Contrary to trees there exists a more explicit mathematical description of the corresponding decision rules.
Neural networks can be viewed as generalizations of the model classes for linear regression (in the case of regression problems) or logistic regression (in the case of classification problems). We shortly recall these methods.

- In linear regression, we considered the following decision rules for the Bayes rule f^* (with respect to squared loss):

$$f(x) = x^T\beta, \qquad \beta \in \mathbb{R}^d.$$

- In logistic regression, we considered the following decision rules for f^* (with respect to 0-1 loss):

$$f(x) = \underset{k\in\{1,...,K\}}{\arg\max}\, \delta_k(x), \qquad \delta_k(x) = \frac{\exp(x^T\beta^{(k)})}{\sum_{l=1}^K \exp(x^T\beta^{(l)})} = M(x^T\beta^{(1)}, ..., x^T\beta^{(K)})_k$$

  with $\beta^{(1)}, ..., \beta^{(K)} \in \mathbb{R}^d$, where

$$M(z) = \frac{1}{\sum_{l=1}^K e^{z_l}} \begin{pmatrix} e^{z_1} \\ \vdots \\ e^{z_K} \end{pmatrix}, \qquad z = (z_1, ..., z_K),$$

  cf. Lemma 3.6. Contrary to Lemma 3.6, we stick to the overparametrized model ($\beta^{(k)}$, $k = 1, ..., K$ instead of $k = 1, ..., K-1$).

In both cases, we now allow for more general (nonlinear) decision rules. More precisely, we replace the linear mappings $x \mapsto x^T\beta$ or $x \mapsto (x^T\beta^{(1)}, ..., x^T\beta^{(K)})$ by more complex mappings $f : \mathcal{X} \to \mathbb{R}$ or $g : \mathcal{X} \to \mathbb{R}^K$, the *neural networks*.
To define the function class of neural networks, we introduce an auxiliary function.

Definition 7.1 (Multivariate shift)**.** For some function $\sigma : \mathbb{R} \to \mathbb{R}$ and $v \in \mathbb{R}^r$, define

$$\sigma_v : \mathbb{R}^r \to \mathbb{R}^r, \quad \sigma_v(x) := \begin{pmatrix} \sigma(x_1 + v_1) \\ \vdots \\ \sigma(x_r + v_r) \end{pmatrix}, \qquad x = (x_1, ..., x_r)^T.$$

Definition 7.2 (Neural networks)**.** Let $L \in \mathbb{N}_0$, $p = (p_0, ..., p_{L+1})^T \in \mathbb{N}^{L+2}$. A *neural network* with *network architecture* (L, p) and *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$ is a function

$$\begin{aligned} &g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \\ &g(x) = W^{(L)} \cdot \sigma_{v^{(L)}}\Big(W^{(L-1)} \cdot \sigma_{v^{(L-1)}}\Big(\dots W^{(1)} \cdot \sigma_{v^{(1)}}\big(W^{(0)} \cdot x\big) \dots \Big)\Big), \end{aligned} \tag{80}$$

where $W^{(l)} \in \mathbb{R}^{p_l \times p_{l+1}}$, $l = 0, ..., L$ are *weight matrices* and $v^{(l)} \in \mathbb{R}^{p_l}$, $l = 1, ..., L$ are *bias vectors* associated to g. Here, L is called the number of *hidden layers* and p is called the *width vector*. Put

$$\mathcal{F}(L, p) := \{g \,|\, g \text{ neural network with architecture } (L, p)\}.$$

Remarks:

- A typical choice for the activation function is given by $\sigma(x) = \max\{x, 0\}$ (the so-called ReLU function) or $\sigma(x) = \frac{e^x}{1+e^x}$ (sigmoid function). It is very important that σ is *nonlinear* so that the composition of σ and linear mappings does lead to more complex mappings. Note that if σ would be linear, then g from (80) would only be a very complicated expression for a linear function which has no advantage over the basic linear regression model.

- There exists an important motivation for the above definition of neural networks from neuroscience: The original information (feature) x is successively processed and transformed. If we put
$$x^{(0)} := x$$
and
$$x^{(l+1)} := \sigma_{v^{(l+1)}}(W^{(l)} x^{(l)}), \quad l = 0, ..., L-1,$$
then $g(x) = W^{(L)} x^{(L)}$. The result of g therefore is obtained by applying a linear transformation L times followed by an 'activation' ('processing') by a nonlinear function σ. This shall simulate the process in the brain, where electric impulses are sent through a chain of neurons. This is also the reason why the elements of the l-th layer $x^{(l)}$ are also called 'neurons'.

- Clearly, σ has a big influence on the specific form of the function class $\mathcal{F}(L, p)$. Moreover, $\mathcal{F}(L, p)$ allows for more complex functions the larger L is and the larger the components of p are.

The decision rules for regression and classification are obtained as follows: In both cases, choose $p_0 = d$.

- For regression, choose $p_{L+1} = 1$ and $\mathcal{F} = \mathcal{F}(L, p)$.
- For classification, choose $p_{L+1} = K$ and $\Delta = \{M \circ g : g \in \mathcal{F}(L, p)\}$.

7.1 The neural network algorithm

Similar as in the section about boosting, we first define an exact neural network algorithm which is then analyzed theoretically. In practice however, an approximation of this exact algorithm is used.
For neural networks, it is necessary to understand the procedure in practice to define an appropriate 'exact' algorithm which shall be analyzed. We will therefore consider a network algorithm for regression used in practice. Let $\mathcal{Y} = \mathbb{R}$, $L(y, s) = (y - s)^2$.

Standard approach: With $p_0 = d$, $p_{L+1} = 1$, let

$$\hat{f}_n^{std} := \operatorname*{arg\,min}_{f \in \mathcal{F}(L,p)} \hat{R}_n(f), \qquad \hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)). \tag{81}$$

Clearly, $\hat{f}_n^{std}$ will overfit the training data if $\mathcal{F}(L, p)$ is too large (that is, if L and the entries of p are too large). Additionally, it is not possible to compute (81) in practice. Instead, one uses an approximation with a gradient descent algorithm. Write $\mathcal{F}(L, p) = \{f_\theta : \theta \in \Theta\}$, where

$$\theta = (v^{(1)}, ..., v^{(L)}, W^{(0)}, ..., W^{(L)}) \in \Theta = \mathbb{R}^T$$

is the vector which contains all the parameters of a network $f_\theta \in \mathcal{F}(L, p)$.

Definition 7.3 (Approximative network algorithm for regression)**.** Let $\lambda > 0$, $p_0 = d$, $p_{L+1} = 1$.

Let $\theta^{(0)} \sim U[-w, w]^D$ with $w > 0$ (*initialization of weights*). Let α_m, $m \in \mathbb{N}$ be a decreasing sequence of positive real numbers (the so-called *learning rate*). Put

$$J(\theta) := \sum_{l=0}^{L} \|W^{(l)}\|_2^2,$$

where $\|\cdot\|_2$ is the Frobenius norm of a matrix. For $m = 1, 2, ..., M$, define

$$\theta^{(m)} = \theta^{(m-1)} - \alpha_m \nabla_\theta \big\{\hat{R}_n(f_\theta) + \lambda \cdot J(\theta)\big\}\big|_{\theta=\theta^{(m-1)}}.$$

Put $\hat{f}_n^{\approx} := f_{\theta^{(M)}}$.

Remarks:

- Due to the random initialization and of the weights and the gradient descent algorithm, it happens in practice that only few of the weights really affect the final function value of $\hat{f}_n^{\approx}$.
- The penalization of the weights via $J(\theta)$ has the aim to prefer small weights.

The above approximative network algorithm *motivates* (not *justifies*) the definition of the following restricted class of neural networks which then serves as the basis for the exact network algorithm. For vectors $v \in \mathbb{R}^r$ let $\|v\|_0 := \#\{j \in \{1, ..., r\} : v_j \neq 0\}$ which counts the number of nonzero elements. A similar definition is made for matrices.

Definition 7.4. For $s \in \mathbb{N}$, $F > 0$ define

$$\mathcal{F}(L, p, s, F) := \Big\{ f \in \mathcal{F}(L, p) : \sum_{l=0}^{L} \|W^{(l)}\|_0 + \sum_{l=1}^{L} \|v^{(l)}\|_0 \leq s, \quad \max_{j \in \{1,\dots,p_{L+1}\}} \|f_j\|_\infty \leq F, \\ \forall l : \quad \|W^{(l)}\|_\infty \leq 1, \|v^{(l)}\|_\infty \leq 1 \Big\}.$$

This class measures the number of nonzero entries in the weight matrices and bias vectors with the parameter s. The additional parameter F is an upper bound on the maximum norm of the networks which is later needed to simplify the proofs. With this, we can formulate the following exact algorithm for regression.

Definition 7.5 (Exact network algorithm for regression)**.** Let $L(y, s) = (y - s)^2$, $p_0 = d$, $p_{L+1} = 1$. Then

$$\hat{f}_n \in \operatorname*{arg\,min}_{f \in \mathcal{F}(L,p,s,F)} \hat{R}_n(f), \qquad \hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)).$$

is called *exact network algorithm for regression.*

Note that $\hat{f}_n^{\approx}$ from Definition 7.3 and $\hat{f}_n$ from Definition 7.5 may differ quite a lot from each other:

- There is no explicit stopping rule for M in $\hat{f}_n^{\approx}$.

- In general, $\theta \mapsto \hat{R}_n(f_\theta)$ has a lot of local minima since it is not a convex function. $\hat{f}_n^{\approx}$ is obtained from a gradient descent algorithm which only converges towards local minima.
- The approximate procedure can not guarantee that only s parameters contribute to the final value of $\hat{f}_n^{\approx}$. Therefore, the assumption $\hat{f}_n^{\approx} \in \mathcal{F}(L, p, s, F)$ may not be satisfied.

Therefore, statements regarding $\hat{f}_n$ can only be transferred in a limited way to $\hat{f}_n^{\approx}$. However, they yield a reasonable first step to understand the performance of neural networks. A detailed analysis of approximative algorithms from Definition 7.3 is a topic of current research.
For completeness, we present also an exact network algorithm for classification.

Definition 7.6 (Exact network algorithm for classification)**.** Let $\tilde{L}(y, s) = -\log(s_y) = -\sum_{k=1}^{K} \mathbb{1}_{\{y=k\}} \log(s_k)$, $p_0 = d$ and $p_{L+1} = K$. Put

$$\hat{\delta}_n \in \operatorname*{arg\,min}_{\delta = M \circ g, g \in \mathcal{F}(L,p,s,F)} \hat{R}_n(\delta), \qquad \hat{R}_n(\delta) := \frac{1}{n} \sum_{i=1}^{n} \tilde{L}(Y_i, \delta(X_i)).$$

Then $\hat{f}_n(x) = \arg\max_{k \in \{1,\dots,K\}} \hat{\delta}_n(x)$ is called *exact network algorithm for classification.*

7.2 Approximation theory for neural networks

To get a better understanding for the set of decision functions $\mathcal{F}(L, p, s, F)$ and their approximation qualities, we first provide some approximation theory for the ReLU function $\sigma(x) = \max\{x, 0\}$.
The following result is taken from [14]. Here for $r \in \mathbb{N}$, $\alpha \in \mathbb{N}_0^r$ denotes a multi-index. We define $|\alpha| = \alpha_1 + \dots + \alpha_r$. For $f : \mathbb{R}^r \to \mathbb{R}$, we put $\partial^\alpha f := \partial_{x_1}^{\alpha_1} \dots \partial_{x_r}^{\alpha_r} f$.

Theorem 7.7. Let $\beta \in \mathbb{N}$. Let $f : [0,1]^r \to \mathbb{R}$ be β-times continuously differentiable with $\sum_{\alpha : 0 \le |\alpha| \le \beta} \|\partial^\alpha f\|_\infty \le K$. Then for all $m, N \in \mathbb{N}$ with $N \ge \max\{(\beta+1)^r, K+1\}$ there exists a network

$$\tilde{f} \in \mathcal{F}(L, p, s, \infty)$$

such that

$$\begin{aligned} p &= (r, 12rN, \dots, 12rN, 1), \\ L &= 8 + (m+5)(1 + \lceil \log_2(r) \rceil), \\ s &\le 94 r^2 (\beta+1)^{2r} N (m+6)(1 + \lceil \log_2(r) \rceil) \end{aligned}$$

and

$$\|f - \tilde{f}\|_\infty \leq (2K+1)3^{r+1}N2^{-m} + K2^\beta N^{-\beta/r}. \tag{82}$$

Proof. A proof sketch is left as an exercise. □

Based on a specific approximation quality (given through $m, N \in \mathbb{N}$), the theorem provides a network $\tilde{f}$ of a specific size which allows to approximate f accordingly. The result itself is not 'optimal' in any sense (in particular, the dimension r enters the approximation rate exponentially), but it gives an impression of the approximation quality of neural networks. In the exercises, we investigate a sketch of the proof.

7.3 Theoretical results

We now investigate a theoretical result for the exact network algorithm from Definition 7.5 for regression problems with quadratic loss $L(y,s) = (y-s)^2$ from [14]. We start with the following model assumption and afterwards derive an oracle inequality.

Definition 7.8 (Model assumption: regression)**.** It holds that $Y = f^*(X) + \varepsilon$, where ε, X are independent and $\varepsilon \sim N(0, v^2)$ with some $v > 0$.

Theorem 7.9 (Oracle inequality for neural networks)**.** Suppose that the model assumption from Definition 7.8 holds. Suppose that $F \geq K$. Let $\gamma \in (0,1]$. Define $H(\gamma) := \log N(\gamma, \mathcal{F}(L,p,s,F), \|\cdot\|_\infty)$. Then there exists a universal constant $c > 0$ such that

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq 2 \inf_{f \in \mathcal{F}(L,p,s,F)} \left\{R(f) - R(f^*)\right\} + c \cdot \left\{(F+v)^2 \cdot \frac{H(\gamma)}{n} + (F+v+1)\cdot\gamma\right\}.$$

Proof. **Step 1: Derivation of the basic inequality.** Note that

$$\begin{aligned}
\hat{R}_n(f) &= \frac{1}{n}\sum_{i=1}^n L(Y_i, f(X_i)) = \frac{1}{n}\sum_{i=1}^n (\varepsilon_i + f^*(X_i) - f(X_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i (f^*(X_i) - f(X_i)) + \underbrace{\frac{1}{n}\sum_{i=1}^n (f^*(X_i) - f(X_i))^2}_{=:\hat{D}_n(f)}.
\end{aligned} \tag{83}$$

Since $\mathbb{E}[\varepsilon|X] = 0$, we have

$$R(f) = \mathbb{E}L(Y, f(X)) = \mathbb{E}[(\varepsilon + f^*(X) - f(X))^2] = \mathbb{E}[\varepsilon^2] + \underbrace{\mathbb{E}[(f^*(X) - f(X))^2]}_{=:D(f)}. \quad (84)$$

Therefore, we have for any $f \in \mathcal{F}(L, p, s, F)$ that

$$R(\hat{f}_n) - R(f^*) \leq \{R(\hat{f}_n) - R(f)\} + \{R(f) - R(f^*)\} \quad (85)$$

and

$$\begin{aligned} R(\hat{f}_n) - R(f) &= \hat{R}_n(\hat{f}_n) - \hat{R}_n(f) + \{R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) - (R(f) - \hat{R}_n(f))\} \\ &\leq \{R(\hat{f}_n) - \hat{R}_n(\hat{f}_n) - (R(f) - \hat{R}_n(f))\} \\ &\overset{(83),(87)}{=} D(\hat{f}_n) - D(f) - (\hat{D}_n(\hat{f}_n) - \hat{D}_n(f)) \\ &\quad + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(\hat{f}_n(X_i) - f^*(X_i)) - \frac{2}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f^*(X_i)). \end{aligned}$$

With $\mathbb{E}\varepsilon = 0$ we conclude that

$$\mathbb{E}R(\hat{f}_n) - R(f) \leq \mathbb{E}[D(\hat{f}_n) - \hat{D}_n(\hat{f}_n)] + \mathbb{E}\Big[\frac{2}{n}\sum_{i=1}^n \varepsilon_i(\hat{f}_n(X_i) - f^*(X_i))\Big] =: A_1 + A_2 \quad (86)$$

Remark: The right hand side is not dependent on f which seems a little strange. This is due to the model assumption and since $\mathbb{E}\hat{D}_n(f) = D(f)$.

Step 2: An upper bound for the first summand in (86).
Let f_j, $j = 1, ..., N(\gamma) := N(\gamma, \mathcal{F}(L, p, s, F), \|\cdot\|_\infty)$ be a γ-covering of $\mathcal{F}(L, p, s, F)$ with oBdA. $\|f_j\|_\infty \leq F$. Let $\hat{j} \in \{1, ..., N\}$ be such that

$$\|\hat{f}_n - f_{\hat{j}}\|_\infty \leq \gamma.$$

By

$$\big|(f(x) - f^*(x))^2 - (\tilde{f}(x) - f^*(x))^2\big| \leq \big|f(x) - \tilde{f}(x)\big| \cdot \big|f(x) + \tilde{f}(x) - 2f^*(x)\big|$$

and $\|\hat{f}_n\|_\infty \leq F$, $\|f^*\|_\infty \leq K \leq F$, we have that

$$\big|D(\hat{f}_n) - D(f_{\hat{j}})\big| \leq 4F\gamma, \qquad \big|\hat{D}_n(\hat{f}_n) - \hat{D}_n(f_{\hat{j}})\big| \leq 4F\gamma.$$

We now aim to apply Lemma 7.10 on $\sup_{j=1,...,N(\gamma)} |D(f_j) - \hat{D}_n(f_j)|$. Choose $\mathcal{G} = \{g_j : j = 1, ..., N\}$, where

$$g_j(x) = (f_j(x) - f^*(x))^2.$$

Then it holds that

$$\mathbb{E}[g_j(X_1)^2] \leq (2F)^2 D(f_j), \qquad \|g_j\|_\infty \leq 4F^2 =: M.$$

Thus for $W := \sup_{j=1,\dots,N(\gamma)} \frac{n(D(f_j)-\hat{D}_n(f_j))}{2FD(f_j)^{1/2}+M\sqrt{\frac{H(\gamma)}{n}}}$, we have that

$$\mathbb{E}W \leq 8\sqrt{H(\gamma)n}, \qquad \mathbb{E}[W^2] \leq 33\sqrt{H(\gamma)n}. \tag{87}$$

We conclude that

$$\begin{aligned}
|D(\hat{f}_n) - \hat{D}_n(\hat{f}_n)| &\leq |D(f_{\hat{j}}) - \hat{D}_n(f_{\hat{j}})| + 8F\gamma \\
&\leq \frac{|D(f_{\hat{j}}) - \hat{D}_n(f_{\hat{j}})|}{2FD(f_{\hat{j}})^{1/2} + M\sqrt{\frac{H(\gamma)}{n}}} \cdot \{2FD(f_{\hat{j}})^{1/2} + M\sqrt{\frac{H(\gamma)}{n}}\} + 8F\gamma \\
&\leq \frac{|W|}{n} \cdot \Big\{2FD(f_{\hat{j}})^{1/2} + M\sqrt{\frac{H(\gamma)}{n}}\Big\} + 8F\gamma.
\end{aligned}$$

By the Cauchy Schwarz inequality, we have $\mathbb{E}[UV] \leq \mathbb{E}[U^2]^{1/2}\mathbb{E}[V^2]^{1/2}$ and $N \in \mathbb{N}$:

$$\begin{aligned}
|A_1| &\leq \mathbb{E}\big|D(\hat{f}_n) - \hat{D}_n(\hat{f}_n)\big| \\
&\leq \frac{1}{n}\mathbb{E}[W^2]^{1/2} \cdot 2F \underbrace{\mathbb{E}[D(f_{\hat{j}})]^{1/2}}_{\leq \mathbb{E}[D(\hat{f}_n)]^{1/2}+\gamma} + \frac{\mathbb{E}|W|}{n} \cdot M\sqrt{\frac{H(\gamma)}{n}} + 8F\gamma \\
&\overset{(87)}{\leq} 20\Big(\sqrt{\frac{H(\gamma)}{n}}F\mathbb{E}[D(\hat{f}_n)]^{1/2} + F^2\frac{H(\gamma)}{n} + \underbrace{\gamma \cdot F\sqrt{\frac{H(\gamma)}{n}}}_{\leq \frac{1}{2}\gamma^2 + \frac{1}{2}F^2\frac{H(\gamma)}{n}} + F\gamma\Big). \\
&\overset{\substack{2ab\leq a^2+b^2 \\ \leq \\ a=10\sqrt{N}F\sqrt{\frac{H(\gamma)}{n}}, b=\frac{1}{\sqrt{N}}\mathbb{E}[D(\hat{f}_n)]^{1/2}}}{\leq} \frac{1}{N}\mathbb{E}D(\hat{f}_n) + 100\big[(N+1)F^2\frac{H(\gamma)}{n} + (\gamma^2 + F\gamma)\big]
\end{aligned} \tag{88}$$

Step 3: Derivation of an upper bound for the second summand in (86). It holds that

$$\begin{aligned}
|A_2| &\leq \mathbb{E}\Big|\frac{2}{n}\sum_{i=1}^n \varepsilon_i(\hat{f}_n(X_i) - f^*(X_i))\Big| \\
&\leq \mathbb{E}\Big|\frac{2}{n}\sum_{i=1}^n \varepsilon_i(f_{\hat{j}}(X_i) - f^*(X_i))\Big| + \frac{2}{n}\sum_{i=1}^n \underbrace{\mathbb{E}|\varepsilon_i|}_{\leq \mathbb{E}[\varepsilon_1^2]^{1/2} \leq v} \cdot \underbrace{\|f_{\hat{j}} - \hat{f}_n\|_\infty}_{\leq \gamma}.
\end{aligned} \tag{89}$$

Define

$$W_j := \frac{\sum_{i=1}^n \varepsilon_i(f_j(X_i) - f^*(X_i))}{\big(\frac{1}{n}\sum_{i=1}^n (f_j(X_i) - f^*(X_i))^2\big)^{1/2}}$$

Then we have

$$\begin{aligned}
& \mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f_{\hat{j}}(X_i)-f^*(X_i))\Big| \\
\leq\quad & \frac{1}{n}\mathbb{E}\Big[W_{\hat{j}}\cdot\Big(\frac{1}{n}\sum_{i=1}^{n}(f_{\hat{j}}(X_i)-f^*(X_i))^2\Big)^{1/2}\Big] \\
\overset{\text{CSU}}{\leq}\quad & \frac{1}{n}\underbrace{\mathbb{E}[\max_{j=1,\dots,N(\gamma)}W_j^2]^{1/2}}_{=\mathbb{E}[\mathbb{E}[\max_{j=1,\dots,N(\gamma)}W_j^2|X_1,\dots,X_n]]}\cdot\underbrace{\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}(f_{\hat{j}}(X_i)-f^*(X_i))^2\Big)\Big]^{1/2}}_{\leq\mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(X_i)-f^*(X_i))^2)]^{1/2}+\gamma} \\
\overset{\text{Lemma 7.11}}{\leq}\quad & 2\Big\{\sqrt{\frac{H(\gamma)}{n}}\cdot v\Big\}\cdot\big\{\mathbb{E}[\hat{D}_n(\hat{f}_n)]^{1/2}+\gamma\big\} \\
=\quad & 2v\cdot\sqrt{\frac{H(\gamma)}{n}}\cdot\mathbb{E}[\hat{D}_n(\hat{f}_n)]^{1/2}+2v\cdot\sqrt{\frac{H(\gamma)}{n}}\cdot\gamma.
\end{aligned}$$

Let $N\in\mathbb{N}$ (N is chosen later large enough such that we can rearrange the terms in the final implicit equation for the excess Bayes risk). Plugging in this upper bound into (89), we obtain

$$\begin{aligned}
|A_2| \quad & \leq \quad \mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{f}_n(X_i)-f^*(X_i))\Big| \\
& \leq \quad 4v\cdot\sqrt{\frac{H(\gamma)}{n}}\cdot\mathbb{E}[\hat{D}_n(\hat{f}_n)]^{1/2}+\underbrace{4v\cdot\sqrt{\frac{H(\gamma)}{n}}\cdot\gamma}_{\leq 2\gamma^2+2v^2\frac{H(\gamma)}{n}}+2v\gamma \\
& \overset{4ab\leq 4a^2+b^2}{\underset{a=v\sqrt{N}\sqrt{\frac{H(\gamma)}{n}},b=\frac{1}{\sqrt{N}}\mathbb{E}[\hat{D}_n(\hat{f}_n)]^{1/2}}{\leq}} \quad \frac{1}{N}\underbrace{\mathbb{E}[\hat{D}_n(\hat{f}_n)]}_{\leq|\mathbb{E}\hat{D}_n(\hat{f}_n)-\mathbb{E}D(\hat{f}_n)|+\mathbb{E}D(\hat{f}_n)}+\big(2+N\big)v^2\frac{H(\gamma)}{n}+2(\gamma^2+v\gamma) \\
& \leq \quad \frac{1}{N}|A_1|+\frac{1}{N}\mathbb{E}D(\hat{f}_n)+\big(2+N\big)v^2\frac{H(\gamma)}{n}+2(\gamma^2+v\gamma).
\end{aligned} \tag{90}$$

Step 4: Summarizing the upper bounds. For $\gamma\leq 1$, we have $\gamma^2\leq\gamma$. Therefore,

$$\begin{aligned}
\mathbb{E}R(\hat{f}_n)-R(f) \quad & \overset{(90),(86)}{\leq} \quad (1+\frac{1}{N})|A_1|+\frac{1}{N}\mathbb{E}D(\hat{f}_n)+\big(2+N\big)v^2\frac{H(\gamma)}{n}+2(1+v)\gamma \\
& \overset{(88),D(f)=R(f)-R(f^*)}{\leq} \quad \Big(\frac{2}{N}+\frac{1}{N^2}\Big)\cdot\mathbb{E}D(\hat{f}_n) \\
& \qquad\qquad +c(N)\cdot\Big\{(F+v)^2\frac{H(\gamma)}{n}+(F+v+1)\cdot\gamma\Big\}
\end{aligned}$$

with suitable chosen $c(N) > 0$ only dependent on $N \in \mathbb{N}$. Plugging in this result into (85) and rearranging terms leads to

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \;\leq\; \Big(1 - \frac{2}{N} - \frac{1}{N^2}\Big)^{-1}\Big[\{R(f) - R(f^*)\} \\ + c(N) \cdot \big\{(F+v)^2 \frac{H(\gamma)}{n} + (F+v+1)\cdot\gamma\big\}\Big].$$

Choosing N large enough such that $(1 - \frac{2}{N} - \frac{1}{N^2})^{-1} \leq 2$ concludes the proof. □

Lemma 7.10. Let $Z_1, ..., Z_n$ be i.i.d. random variables with values in $\mathcal{Z}$. Let $\mathcal{G} \subset \{g : \mathcal{Z} \to \mathbb{R} \text{ measurable}\}$ such that for all $g \in \mathcal{G}$, it holds that $\|g\|_\infty \leq M$. Define $H := \log(|\mathcal{G}| + 1)$, and

$$W := \sup_{g\in\mathcal{G}} \frac{\big|\sum_{i=1}^n \{g(Z_i) - \mathbb{E}g(Z_i)\}\big|}{\mathbb{E}[g(Z_1)^2]^{1/2} + M\sqrt{\frac{H}{n}}}.$$

Then for all $|\mathcal{G}| \geq 2$, it holds that

$$\mathbb{E}W \leq 8\sqrt{nH}, \qquad \mathbb{E}[W^2] \leq 33nH.$$

Proof. Put $\tilde{g} := \frac{g}{\mathbb{E}[g(Z_1)^2]^{1/2} + M\sqrt{\frac{H}{n}}}$. Then we have

$$\mathbb{E}[\tilde{g}(Z_1)^2] \leq \frac{\mathbb{E}[g(Z_1)^2]}{\mathbb{E}[g(Z_1)^2]} \leq 1, \qquad \|\tilde{g}\|_\infty \leq \frac{\|g\|_\infty}{M\sqrt{\frac{H}{n}}} \leq \sqrt{\frac{n}{H}}.$$

Theorem 6.13 implies that

$$\mathbb{E}W = \mathbb{E}\sup_{g\in\mathcal{G}}\Big|\sum_{i=1}^n \{\tilde{g}(Z_i) - \mathbb{E}\tilde{g}(Z_1)\}\Big| \;\leq 4\cdot\Big\{1\cdot\sqrt{nH} + \sqrt{\frac{n}{H}}\cdot H\Big\} = 8\sqrt{nH}.$$

The second part regarding the upper bound for $\mathbb{E}[W^2]$ is a little bit more elaborated and is left as an exercise. □

Lemma 7.11 (Maximum of normal distributions)**.** Let $Z_1, ..., Z_n \sim N(0, v^2)$ be i.i.d. and $a_{ij} \in \mathbb{R}$, $i = 1, ..., n$, $j = 1, ..., N$. Let

$$W_j := \frac{1}{v\sqrt{n}} \frac{\sum_{i=1}^n a_{ij} Z_i}{\big(\frac{1}{n}\sum_{i=1}^n a_{ij}^2\big)^{1/2}}.$$

Then it holds that

$$\mathbb{E} \max_{j=1,\dots,N} |W_j| \leq 2\sqrt{\log(N+1)}, \qquad \mathbb{E} \max_{j=1,\dots,N} |W_j|^2 \leq 4\log(N+1).$$

Proof. This is left as an exercise. □

7.4 Convergence rate of the neural network algorithm

We use the oracle inequality to derive a convergence rate for the neural network algorithm. We first derive an upper bound for $H(\gamma) = \log N(\gamma, \mathcal{F}(L,p,s,F), \|\cdot\|_\infty)$ for a given $\gamma > 0$.

Lemma 7.12. (i) Define $V := \prod_{l=0}^{L+1}(p_l+1)$, Then it holds that

$$N(\gamma, \mathcal{F}(L,p,s,F), \|\cdot\|_\infty) \leq (2\gamma^{-1}V^2(L+1))^{s+1}.$$

(ii) Let $s \geq 2, L \geq 1$. Then there exists some universal constant $c > 0$ such that

$$H(\gamma) \leq c \cdot s \cdot \big\{L\log(s) + \log(\gamma^{-1}) + \log(p_0 p_{L+1})\big\}.$$

Proof. This is left as an exercise.

□

Definition 7.13 (Model assumption: Bayes rule is continuously differentiable ($\beta = 1$))**.** $f^* : [0,1]^d \to \mathbb{R}$ is continuously differentiable with $\|f^*\|_\infty + \sum_{j=1}^d \|\partial^j f^*\|_\infty \leq K$.

Then we have the following convergence rate.

Theorem 7.14 (Convergence rate of neural networks)**.** Suppose that the model assumptions of Definition 7.8 and Definition 7.13 hold. Define

$$N = \lceil \max\{2^d, K+1\} \cdot n^{\frac{d}{2+d}} \rceil.$$

Then there exist universal constants $c_1, c_2, c_3, c_4 > 0$, such that

- $F \geq K$,
- $L \geq c_1 \log(n) \log(d)$,
- $c_3 2^{3d} N \log(n) \geq s \geq c_2 2^{3d} N \log(n)$,
- $\forall l = 1, ..., L$: $p_l \geq c_4 dN$,

implies

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq C \cdot 95^d \cdot \log(n)^2 L \cdot n^{-\frac{2}{2+d}},$$

where C is a constant only depending on F, K, v.

Proof. We combine the statements from Theorem 7.7, Theorem 7.9 and Lemma 7.12. Theorem 7.9 implies that

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq 2 \inf_{f \in \mathcal{F}(L,p,s,F)} \big\{R(f) - R(f^*)\big\} + c \cdot \Big\{(F+v)^2 \cdot \frac{H(\gamma)}{n} + (F+v) \cdot \gamma\Big\}. \quad (91)$$

Let $m = \lceil \log_2(n) \rceil$.
Theorem 7.7 implies that there exists a network $\tilde{f} \in \mathcal{F}(L, p, s, \infty)$ with the given size (to meet the condition on s, we make use of the fact that $c 2^{3d} \geq 2^{2d} d^2 \log(d)$ holds uniformly in d for some universal constant c) such that

$$\|\tilde{f} - f^*\|_\infty \quad \leq \quad (2K+1) 3^{d+1} N \underbrace{2^{-m}}_{\leq n^{-1}} + 2K N^{-1/d}.$$

Caution: Up to now, $\tilde{f}$ may not be bounded by F. Define $\tilde{\tilde{f}} := \tilde{f} \cdot \min\{\frac{\|f^*\|_\infty}{\|\tilde{f}\|_\infty}, 1\}$. Then it holds that

$$\|\tilde{\tilde{f}}\|_\infty \leq \|f^*\|_\infty \ \leq K \leq F, \qquad \|\tilde{\tilde{f}} - f^*\|_\infty \leq \|\tilde{\tilde{f}} - \tilde{f}\|_\infty + \|\tilde{f} - f^*\|_\infty \leq 2\|\tilde{f} - f^*\|_\infty.$$

We therefore have $\tilde{\tilde{f}} \in \mathcal{F}(L, p, s, F)$ and $\tilde{\tilde{f}}$ still enjoys the approximation properties of $\tilde{f}$. We conclude that

$$\begin{aligned}
\inf_{f \in \mathcal{F}(L,p,s,F)} \big\{R(f) - R(f^*)\big\} \quad &= \quad R(\tilde{\tilde{f}}) - R(f^*) = \mathbb{E}[(\tilde{\tilde{f}}(X) - f^*(X))^2] \\
&\leq \quad \|\tilde{\tilde{f}} - f^*\|_\infty^2 \leq 4\big[(2K+1)3^{d+1}\frac{N}{n} + K 2^\beta N^{-\beta/d}\big]^2 \\
&\overset{(a+b)^2 \leq 2a^2+2b^2}{\leq} \quad 8(2K+1)^2 6^{d+1} \big(\frac{N}{n}\big)^2 + 32K^2 N^{-2/d}. \qquad (92)
\end{aligned}$$

Lemma 7.12 with $\gamma = n^{-1}$ implies that with some updated universal constants $c', c'' > 0$, it holds that

$$\begin{aligned} \frac{H(\gamma)}{n} &\leq c' \cdot \frac{s}{n} \cdot \big\{ L \log(s) + \log(n) + \log(d) \big\} \\ &\overset{L \geq c_1 \log(n)\log(d)}{\leq} c'' \cdot \frac{sL}{n} \log(s). \end{aligned} \tag{93}$$

Plugging in (92), (93) into (91) and summarizing the constants K, F, v into a new constant C yields

$$\begin{aligned} \mathbb{E}R(\hat{f}_n) - R(f^*) &\leq C \cdot \Big\{ 6^{d+1} \big(\frac{N}{n}\big)^2 + N^{-2/d} + \frac{sL}{n} \log(s) \Big\} \\ &\overset{N \leq 2^d n^d \Rightarrow \log(s) \leq C(\beta+1)^d \log(n)}{\leq} C \cdot \Big\{ 6^{d+1} \big(\frac{N}{n}\big)^2 + N^{-2/d} + \frac{NL}{n} 2^{4d} \log(n)^2 \Big\} \end{aligned}$$

The above terms now reflect a trade off between $\frac{N}{n}$ and $N^{-2/d}$ with respect to N. We can balance these terms with respect to n if we choose $N \approx n^{\frac{d}{2+d}}$. This is exactly the rate of N given in the assumptions. The additional factors given in the assumptions are due to the conditions in Theorem 7.7. Plugging in N into the above inequality yields

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq C \cdot 96^d \cdot \log(n)^2 L \cdot n^{-\frac{2}{2+d}}.$$

□

Remark: The number of layers should be chosen as small as possible to obtain a good upper bound for the convergence rate. The smallest possible value $L = c_1 \log(n) \log(d)$ yields

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq C \cdot \log(d) 96^d \cdot \log(n)^3 \cdot n^{-\frac{2}{2+d}}.$$

Obviously, this upper bound is quite bad. On the one hand, the factor $n^{-\frac{2}{2+d}}$ suffers from the curse of dimension; on the other hand, the factor 96^d grows exponentially in d. If we only know that f^* is continuously differentiable, we therefore should not expect any 'wonders' from neural network algorithms. The power of neural networks lies in the fact that they can *adapt* quite well to more specific structures (which is theoretically founded by the above oracle inequality). We show this in the special case of an reduced additive model.

Definition 7.15 (Model assumption: additive model)**.** Suppose that there exist measurable $h_j : [0,1] \to [0, \frac{1}{k}]$, $j \in \{1, ..., k\}$ with $\sup_{j=1,...,k} \{\|h_j\|_\infty + \|h_j'\|_\infty\} \leq K$ and

$$f^*(x) = \sum_{j=1}^{k} h_j(x_j), \quad x = (x_1, ..., x_d) \in \mathbb{R}^d.$$

Remark: We restrict ourselves to functions h_j with values in $[0, \frac{1}{k}]$ only for simplicity and to meet the assumptions of Theorem 7.7 in the following theorem.

Theorem 7.16 (Convergence rate of the neural network algorithm)**.** Suppose that the model assumptions from Definition 7.8 and Definition 7.15 hold. Define

$$N = \lceil \max\{(k+1)^k, K+1\} \cdot n^{1/3} \rceil.$$

Then there exist universal constants $c_1, c_2, c_3, c_4 > 0$ such that

- $F \geq 1$,
- $L \geq c_1 \log(n) \log(k)$,
- $c_3(3k)^{2k} N \log(n) \geq s \geq c_2(3k)^{2k} N \log(n)$,
- $\forall l = 1, ..., L\colon\ p_l \geq c_4 k N$,

implies

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq C \cdot (3k)^{3k} \cdot L \cdot \log(n)^2 \log(d) n^{-2/3},$$

where C is a constant only depending on F, K, v.

Proof. The most important change compared to the proof of Theorem 7.14 is a different application of the approximation result. The model assumption of Definition 7.15 can be written as follows:

$$f^* = g_1 \circ g_0,$$

where $g_0(x) := (h_1(x_1), ..., h_k(x_k))$, $g_1(y) = \sum_{j=1}^k y_j$.
We now apply Theorem 7.7 with $r = 1$ and k times (that is, for each h_j). The networks obtained are stacked up. This leads to a network where the width vector p and the number of non-zero entries s is multiplied with k. We therefore obtain $\tilde{g}_0 \in \mathcal{F}(L^{(0)}, p^{(0)}, s^{(0)}, \infty)$ such that

$$\sup_{j=1,...,k} \|\tilde{g}_{0j} - g_{0j}\|_\infty \leq (2K+1)3^2 \frac{N}{n} + 2KN^{-1}$$

and

$$\begin{aligned} p^{(0)} &= (d, 12kN, ..., 12kN, k), \\ L^{(0)} &= 8 + (\lceil \log_2(n) \rceil + 5)(1 + \lceil \log_2(1) \rceil), \\ s^{(0)} &= 94k \cdot 2^2 N (\lceil \log_2(n) \rceil + 6)(1 + \lceil \log_2(1) \rceil). \end{aligned}$$

Define $w(x_1, ..., x_k) := (1 - (1 - x_1)_+, ..., 1 - (1 - x_k)_+)$ (this corresponds to a network with 2 layers and $4 \cdot k$ non-zero parameters).
Theorem 7.7 applied with $r = k$, $K = 2k$, $\beta \geq 1$ arbitrarily chosen (g_1 is infinitely often differentiable with $\sum_{0 \leq |\alpha| \leq \beta} \|\partial^\alpha g_1\|_\infty = \|g_1\|_\infty + \sum_{j=1}^k \|\partial^j g_1\|_\infty \leq 2k$) implies that there exists $\tilde{g}_1 \in \mathcal{F}(L^{(1)}, p^{(1)}, s^{(1)}, \infty)$ such that

$$\|\tilde{g}_1 - g_1\|_\infty \leq (4k+1)3^{k+1}\frac{N}{n} + 4kN^{-\beta/k}$$

with

$$\begin{aligned} p^{(1)} &= (k, 12kN, ..., 12kN, 1), \\ L^{(1)} &= 8 + (\lceil \log_2(n) \rceil + 5)(1 + \lceil \log_2(k) \rceil), \\ s^{(1)} &= 94k^2 \cdot (\beta+1)^{2k} N(\lceil \log_2(n) \rceil + 6)(1 + \lceil \log_2(k) \rceil). \end{aligned}$$

Composition of the networks yields a new network

$$\tilde{f} := \tilde{g}_1 \circ (w \circ \tilde{g}_0) \in \mathcal{F}(\underbrace{L^{(0)} + L^{(1)} + 3}_{=:L}, \underbrace{(d, 12kN, ..., 12kN, 1)}_{=:p}, \underbrace{s^{(0)} + s^{(1)} + 4k}_{=:s}, \infty).$$

Note that $f^* = g_1 \circ w \circ g_0$. We therefore have (with $|\cdot|_\infty$ denoting the maximum norm in $\mathbb{R}^k$)

$$\begin{aligned} |f^*(x) - \tilde{f}(x)| &= |g_1(g_0(x)) - g_1(\tilde{g}_0(x))| + |g_1(w(\tilde{g}_0(x))) - \tilde{g}_1(w(\tilde{g}_0(x)))| \\ &\leq k \cdot |g_0(x) - \tilde{g}_0(x)|_\infty + |g_1(w(\tilde{g}_0(x()) - \tilde{g}_1(w(\tilde{g}_0(x)))| \\ &\leq k \cdot \big\{(2K+1)3^2\frac{N}{n} + 2KN^{-1}\big\} + (4k+1)3^{k+1}\frac{N}{n} + 4kN^{-\beta/k}. \end{aligned}$$

As in the proof of Theorem 7.14, in particular by truncating the network so that it meets the supremum norm F, we conclude with $\beta = k$ and some constant C only depending on K that

$$\begin{aligned} \inf_{f \in \mathcal{F}(L,p,s,F)} \big\{R(f) - R(f^*)\big\} &\leq 2\big[k \cdot \big\{(2K+1)3^2\frac{N}{n} + 2KN^{-1}\big\} + (4k+1)3^{k+1}\frac{N}{n} + 4kN^{-1}\big]^2 \\ &\leq C\big\{k^2 3^{2k}\big(\frac{N}{n}\big)^2 + k^2 N^{-2}\big\}. \end{aligned}$$

By Lemma 7.12 applied with $\gamma = n^{-1}$, we obtain with some updated universal constants $c', c'' > 0$ that

$$\begin{aligned} \frac{H(\gamma)}{n} &\leq c' \cdot \frac{s}{n} \cdot \big\{L\log(s) + \log(n) + \log(d)\big\} \\ &\overset{L \geq c_1 \log(n)\log(k)}{\leq} c'' \cdot \log(d)\frac{sL}{n}\log(s). \end{aligned}$$

Plugging in these results into (91) yields with some constant $C > 0$ only dependent on K, F, v that

$$\begin{aligned} & \mathbb{E}R(\hat{f}_n) - R(f^*) \\ \leq\ & C \cdot \big\{k^2 3^{2k} \big(\frac{N}{n}\big)^2 + k^2 N^{-2} + \log(d)\frac{sL}{n}\log(s)\big\} \\ \overset{s \leq c \cdot k^2 (k+1)^{2k} \log(k) N \log(n)}{\leq}\ & C \cdot \big\{k^2 3^{2k} \big(\frac{N}{n}\big)^2 + k^2 N^{-2} + \log(d)(2k)^{3k}\frac{LN}{n}\log(n)^2\big\}. \end{aligned}$$

Choose N as given in the theorem, then with some updated constant $C > 0$ only dependent on K, F, v, we obtain the final result

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq C \cdot (3k)^{3k} \cdot L \cdot \log(n)^2 \log(d) n^{-2/3}.$$

□

Remark: Choosing L minimal, we obtain that

$$\mathbb{E}R(\hat{f}_n) - R(f^*) \leq C \cdot (3k)^{4k} \cdot \log(n)^3 \log(d) n^{-2/3}.$$

Note that d only enters the rate via $\log(d)$ which comes from Lemma 7.12(ii) through the term $\log(p_0 p_{L+1})$. Now, only the reduced dimension k enters the rate exponentially via $(3k)^{4k}$. If k is small, this exponential factor is eliminated by $n^{-2/3}$. Therefore, we only have to pay with a factor $\log(d)$ that we do not know the structure of f^*.

7.5 Exercises

Task 28 (Approximation theory for neural networks)**.** Let $\sigma(x) = \max\{x, 0\}$ denote the ReLU activation function. The goal of this task is to give an idea of the proof of Theorem 7.7. For $k \in \mathbb{N}$, define

$$T_k : [0, 2^{2-2k}] \to [0, 2^{-2k}], \quad T_k(x) := \sigma(\frac{x}{2}) - \sigma(x - 2^{1-2k}),$$

and $R_k : [0,1] \to [0, 2^{-2k}]$, $R_k := T_k \circ T_{k-1} \circ ... \circ T_1$.

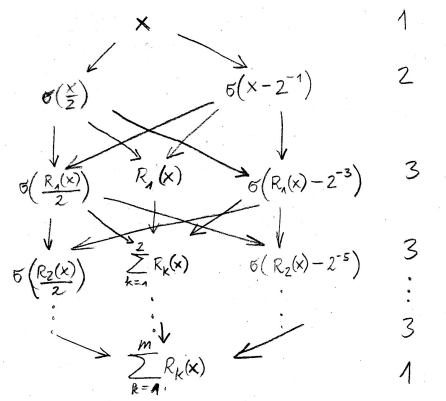
1. Plot a graph of R_1, $R_1 + R_2$ and $R_1 + R_2 + R_3$.
2. With induction one can show for $x \in [0,1]$ and $m \in \mathbb{N}$ that

$$\Big| x \cdot (1-x) - \sum_{k=1}^{m} R_k(x) \Big| \leq 2^{-m}.$$

Show that there exists a network $f_m \in \mathcal{F}(m, (1, 2, 3, ..., 3, 1))$ with

$$f_m(x) = \sum_{k=1}^{m} R_k(x).$$

Hint: Use the following sketch



3. Define $g(x) = x \cdot (1 - x)$. Show that

$$x \cdot y = g(\frac{x - y + 1}{2}) - g(\frac{x + y}{2}) + \frac{x + y}{2} - \frac{1}{4} = xy.$$

   How can this identity be used to approximate products $x \cdot y$ with networks?

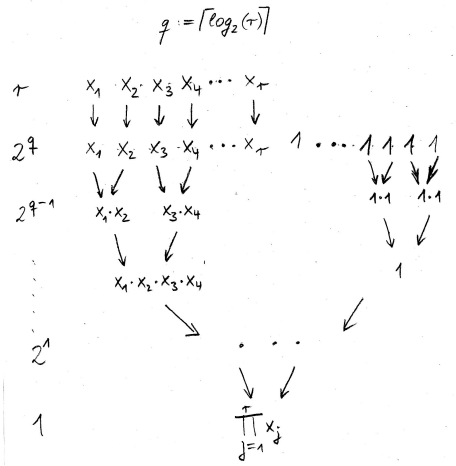
4. One can show that there exists a network $f_m \in \mathcal{F}(m+4, (2, 6, 6, ..., 6, 1))$ such that $|f_m(x, y) - xy| \leq 2^{-m}$ for all $x, y \in [0, 1]$. Show that there exists a network

$$f_{m,r} \in \mathcal{F}((m + 5)\lceil \log_2(r) \rceil, (r, 6r, 6r, ..., 6r, 1))$$

   with

$$\left| f_{m,r}(x) - \prod_{j=1}^{r} x_j \right| \leq r^2 2^{-m}, \qquad x = (x_1, ..., x_r) \in [0, 1]^r.$$

   Hint: Use the following sketch



5. For β-times differentiable $f : [0,1]^r \to \mathbb{R}$, there exists for each $a \in [0,1]^r$ a Taylor representation
$$T_a(x) := \sum_{\alpha \in \mathbb{N}_0^r : 0 \le |\alpha| \le \beta - 1} (\partial^\alpha f)(a) \cdot \frac{(x-a)^\alpha}{\alpha!},$$
where $x^\alpha := x_1^{\alpha_1} \cdot ... \cdot x_r^{\alpha_r}$ and $\alpha! := \alpha_1! \cdot ... \cdot \alpha_r!$. Show the following statement: If $\|\partial^\alpha f\|_\infty \le K$ for all $\alpha \in \mathbb{N}_0^r$ with $|\alpha| = \beta$, then one has
$$|f(x) - T_a(x)| \le K e^r \cdot |x-a|_\infty^\beta.$$

6. Let $M \in \mathbb{N}$ and $D(M) := \{a_\ell := (\frac{\ell_j}{M})_{j=1,...,r} : \ell = (\ell_1, ..., \ell_r) \in \{0, 1, ..., M\}^r\}$ be a grid. Let
$$T(x) := \sum_{a \in D(M)} T_a(x) \cdot \prod_{j=1}^r \big(1 - M \cdot |x_j - a_j|\big)_+$$
Show that
$$\|f - T\|_\infty \le K e^r M^{-\beta}.$$
Hint: Use that for all $x \in [0,1]^r$, it holds that $\sum_{a \in D(M)} \prod_{j=1}^r \big(1 - M \cdot |x_j - a_j|\big)_+ = 1$.

7. Discussion: How one can use the results of (d) and (f) to prove Theorem 7.7?

Task 29 (Maximal inequalities based on Bernstein's inequality, Lemma 7.10(ii))**.** Let $Z_1, ..., Z_n$ be i.i.d. random variables with values in some space $\mathcal{Z}$ and $\mathcal{G} \subset \{g : \mathcal{Z} \to \mathbb{R} \text{ measurable}\}$. Suppose that there exists $M \ge 0$ such that for all $g \in \mathcal{G}$, $\|g\|_\infty \le M$. Define $H := \log(|\mathcal{G}| + 1)$ and
$$W := \sup_{g \in \mathcal{G}} \frac{|\sum_{i=1}^n \{g(Z_i) - \mathbb{E}g(Z_i)\}}{\mathbb{E}[g(Z_1)^2]^{1/2} + M\sqrt{\frac{H}{n}}}.$$
We aim to show that $\mathbb{E}[W^2] \le 33nH$. Let $t_0 > 0$.

1. Show that
$$\mathbb{E}[W^2] \le t_0^2 + 2\int_{t_0}^\infty t\mathbb{P}(W \ge t)dt.$$
Hint: It holds that $\mathbb{E}[W^2] = \int \mathbb{P}(W^2 \ge u)du$.

2. Define $\tilde{g} := \frac{g}{\mathbb{E}[g(Z_1)^2]^{1/2} + M\sqrt{\frac{H}{n}}}$. Show with Bernstein's inequality ($\mathbb{P}\big(\big|\sum_{i=1}^n (X_i - \mathbb{E}X_1)\big| \ge x\big) \le 2\exp\big(-\frac{x^2}{2n\mathbb{E}[X_1^2] + \frac{2}{3}\|X_1\|_\infty \cdot x}\big)$ for i.i.d. $X_1, ..., X_n$) that for any $t \ge 3\sqrt{nH}$,
$$\mathbb{P}(W \ge t) \le 2|\mathcal{G}| \cdot \exp\Big(-\frac{3t}{4\sqrt{\frac{n}{H}}}\Big).$$

3. Conclude that
$$\mathbb{E}[W^2] \leq 33nH.$$
Hint: Define $t_0 = 3\sqrt{nH}$ *and* $a := \frac{3}{4}\sqrt{\frac{H}{n}}$. *With partial integration, one can show that* $\int te^{-at}dt = -(a^{-2} + a^{-1}t)e^{-at}$.

Task 30 (Maximal inequality for normal distributions, Lemma 7.11)**.** Let $Z_1, ..., Z_n \sim N(0, v^2)$ be i.i.d. and $a_{ij} \in \mathbb{R}$, $i = 1, ..., n$, $j = 1, ..., N$. Let
$$W_j := \frac{1}{v\sqrt{n}} \frac{\sum_{i=1}^n a_{ij} Z_i}{\left(\frac{1}{n}\sum_{i=1}^n a_{ij}^2\right)^{1/2}}.$$
Our aim it to show that
$$\mathbb{E} \max_{j=1,...,N} |W_j| \leq 2\sqrt{\log(N+1)}, \qquad \mathbb{E} \max_{j=1,...,N} |W_j|^2 \leq 4\log(N+1).$$

1. Show that $W_j \sim N(0, 1)$.
2. Let $\varphi_2(x) = \exp(x^2) - 1$. Show that $\varphi_2(\mathbb{E}\max_{j=1,...,N} \frac{|W_j|}{2}) \leq N$ and $\mathbb{E}\max_{j=1,...,N} \frac{|W_j|}{2} \leq 2\sqrt{\log(N+1)}$.
Hint: For $W_j \sim N(0, 1)$, *one has* $\mathbb{E}\exp(\frac{W_j^2}{4}) = \sqrt{2}$.
3. Let $\varphi_1(x) = \exp(x) - 1$. Show that $\varphi_1(\mathbb{E}\max_{j=1,...,N} \frac{|W_j|^2}{4}) \leq N$ and $\mathbb{E}\max_{j=1,...,N} |W_j|^2 \leq 4\log(N+1)$.

Task 31 (Covering Numbers of neural networks, Lemma 7.12)**.** In this task, we upper bound the log covering numbers $H(\gamma) = \log N(\gamma, \mathcal{F}(L, p, s, F), \|\cdot\|_\infty)$ of the class of neural networks
$$\mathcal{F}(L, p, s, \infty) := \Big\{ f \in \mathcal{F}(L, p) : \sum_{l=0}^{L} \|W^{(l)}\|_0 + \sum_{l=1}^{L} \|v^{(l)}\|_0 \leq s, \qquad \forall l : \|W^{(l)}\|_\infty \leq 1, \|v^{(l)}\|_\infty \leq 1 \Big\},$$
where
$$\mathcal{F}(L, p) = \Big\{ g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}} : g(x) = W^{(L)} \cdot \sigma_{v^{(L)}}\Big(W^{(L-1)} \cdot \sigma_{v^{(L-1)}}\Big(\dots W^{(1)} \cdot \sigma_{v^{(1)}}\big(W^{(0)} \cdot x\big) \dots \Big)\Big)$$
$$W^{(l)} \in \mathbb{R}^{p_l \times p_{l+1}} \quad (l = 0, ..., L), \qquad v^{(l)} \in \mathbb{R}^{p_l} \quad (l = 1, ..., L) \Big\}.$$

1. Show that each $\mathcal{F}(L, p)$ can be written as $\{f_\theta : \theta \in \Theta^{pre}\}$ with $\Theta^{pre} \subset [-1, 1]^T$, where
$$T = \sum_{l=0}^{L} p_l p_{l+1} + \sum_{l=1}^{L} p_l.$$

2. Show that $\mathcal{F}(L,p,s,\infty) = \{f_\theta : \theta \in \Theta\}$, where

$$\Theta \subset \bigcup_{S\subset\{1,\dots,T\}:|S|\leq s} \Theta_S, \qquad \Theta_S = \{\theta \in [-1,1]^T : \forall j \in S^c : \theta_j = 0\}.$$

3. Let $a > 0$ and $S \subset \{1,...,T\}$, $|S| \leq s$. Show that there exists $\tilde{\Theta}_S \subset [-1,1]^T$ with

$$\forall \theta \in \Theta_S \quad \exists \tilde{\theta} \in \tilde{\Theta}_S : \quad \|\theta - \tilde{\theta}\|_\infty \leq a, \qquad |\tilde{\Theta}_S| \leq \left\lfloor \frac{2}{a} \right\rfloor^s.$$

Define $\tilde{\Theta} := \bigcup_{S\subset\{1,\dots,T\}:|S|\leq s} \tilde{\Theta}_S$. Then the following property is fulfilled:

$$\forall \theta \in \Theta \quad \exists \tilde{\theta} \in \tilde{\Theta} : \quad \|\theta - \tilde{\theta}\|_\infty \leq a.$$

Show the following statements:

(d) $T \leq V := \prod_{l=0}^{L+1}(p_l + 1)$ and

$$|\tilde{\Theta}| \leq \big(\frac{2V}{a}\big)^{s+1}.$$

Hint: It holds that $\binom{T}{k} \leq T^k$.

For $f_\theta \in \mathcal{F}(L,p,s,F)$ choose $\tilde{\theta} \in \tilde{\Theta}$ with $\|\theta - \tilde{\theta}\|_\infty \leq a$. We aim to show that $\|f_\theta - f_{\tilde{\theta}}\|_\infty \leq a \cdot (L+1) \cdot V$. (*)

(e) Show that (*) and (d) imply:

$$N(\gamma, \mathcal{F}(L,p,s,F), \|\cdot\|_\infty) \leq (2\gamma^{-1}V^2(L+1))^{s+1}.$$

Define for $\theta = (v^1,...,v^{(L)}, W^{(0)},...,W^{(L)})$:

$$A_\theta^{k+}(x) = \sigma_{v^{(k)}} W^{(k-1)} \sigma_{v^{(k-1)}} ... W^{(1)} \sigma_{v^{(1)}} W^{(0)} x$$

and

$$A_\theta^{k-}(y) = W^{(L)} \sigma_{v^{(L)}} ... W^{(k)} \sigma_{v^{(k)}} W^{(k-1)} x.$$

(f) Show that $|A_\theta^{k-}(x) - A_\theta^{k-}(x')| \leq \big(\prod_{l=k-1}^{L} p_l\big) \cdot \|x - x'\|_\infty$.
Hint: The row-sum norm $\|A\|_Z$ *of a matrix* A *satisfies* $\|Ax\|_\infty \leq \|A\|_Z \|x\|_\infty$.

(g) Show that $|A_\theta^{k+}(x)|_\infty \leq \prod_{l=0}^{k-1}(p_l + 1)$.

(h) Conclude from (f),(g) that

$$|f_\theta(x) - f_{\tilde{\theta}}(x)| \leq a \cdot (L+1) \cdot V.$$

(i) Show that

$$N(\gamma, \mathcal{F}(L, p, s, \infty), \|\cdot\|_\infty) = N(\gamma, \mathcal{F}(L, (p_0, p_1 \wedge s, ..., p_L \wedge s, p_{L+1}), s, \infty), \|\cdot\|_\infty)$$

(j) Let $s \geq 2$, $L \geq 1$. Show that there exists a universal constant $c > 0$ such that $H(\gamma) = \log N(\gamma, \mathcal{F}(L, p, s, \infty), \|\cdot\|_\infty)$ satisfies

$$H(\gamma) \leq c \cdot s \cdot \big\{ L \log(s) + \log(\gamma^{-1}) + \log(p_0 p_{L+1}) \big\}.$$

8 Solutions of the exercises

8.1 Solutions of Chapter 2

Solution 1 (Solution of Task 1)**.** (a) Markov's inequality yields

$$\mathbb{P}(\|A\mathbb{e}\|_2 \geq \sigma\|A\|_F\sqrt{t}) \leq \frac{\mathbb{E}[\|A\mathbb{e}\|_2^2]}{(\sigma\|A\|_F\sqrt{t})^2}.$$

Here, we have

$$\mathbb{E}[\|A\mathbb{e}\|_2^2] = \mathbb{E}[\mathrm{tr}(A\mathbb{e}\mathbb{e}^T A^T)] = \mathrm{tr}(A\mathbb{E}[\mathbb{e}\mathbb{e}^T]A^T) = \mathrm{tr}(A\sigma^2 I_{d\times d}A^T) = \sigma^2\mathrm{tr}(AA^T) = \sigma^2\|A\|_F^2.$$

Plugging in this result into the first inequality yields

$$\frac{\mathbb{E}[\|A\mathbb{e}\|_2^2]}{(\sigma\|A\|_F\sqrt{t})^2} = \frac{1}{t}.$$

(b) Despite the constant c_2, the expressions in the probability are the same for $t \geq 1$. The lemma of the lecture yields a much better estimate for the probability of this event, since for large t, the term e^{-t} decays much faster than $\frac{1}{t}$. However, the lemma also needs stronger assumptions while we only needed the existence of second moments of $\mathbb{e}$ (no Gaussianity assumption) to obtain the result in (a).

(c) With $\|\hat{\Sigma} - \Sigma\| \leq \|\hat{\Sigma} - \Sigma\|_F$ and Markov's inequality, we have

$$\mathbb{P}(\|\hat{\Sigma} - \Sigma\| \geq x) \leq \frac{\mathbb{E}[\|\hat{\Sigma} - \Sigma\|_F^2]}{x^2}.$$

It holds that

$$\begin{aligned}
\mathbb{E}[\|\hat{\Sigma}-\Sigma\|_F^2] &= \sum_{j,k=1}^{d} \mathbb{E}\big[(\hat{\Sigma}_{jk}-\Sigma_{jk})^2\big] \\
&= \sum_{j,k=1}^{d} \mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik} - \mathbb{E}[X_{ij}X_{ik}]\Big)^2\Big] \\
&= \sum_{j,k=1}^{d} \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik}\Big) \\
&= \sum_{j,k=1}^{d} \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_{ij}X_{ik}) \\
&= \sum_{j,k=1}^{d} \frac{1}{n}\big\{\mathbb{E}[X_{1j}^2X_{1k}^2] - \mathbb{E}[X_{1j}X_{1k}]\big\} \\
&\overset{\text{Hint}}{=} \frac{1}{n}\sum_{j,k=1}^{d} \big\{\Sigma_{jj}\Sigma_{kk} + 2\Sigma_{jk}^2 - \Sigma_{jk}\big\} \\
&= \frac{1}{n}\Big\{\sum_{j,k=1}^{d}\Sigma_{jj}\Sigma_{kk} + \sum_{j,k=1}^{d}\Sigma_{jk}^2\Big\} \\
&= \frac{1}{n}\big\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\big\}.
\end{aligned}$$

This proves the claim.

(d) Substitute $x' := \frac{1}{\sqrt{n}}\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}^{1/2}x$. Then we have with (c):

$$\begin{aligned}
&\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq \frac{1}{\sqrt{n}}\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}^{1/2}x\Big) \\
\leq\ &\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq x'\Big) \leq \frac{1}{n}\frac{\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}}{(x')^2} = \frac{1}{x}.
\end{aligned}$$

(e) If $d \leq n$ and $n \geq x \geq d$ (these we simply assume here), then the statement of the lemma reads

$$\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq c_1\|\Sigma\| \cdot \sqrt{\frac{x}{n}}\Big) \leq e^{-x}.$$

The statement from (d) is

$$\mathbb{P}\Big(\|\hat{\Sigma}-\Sigma\| \geq \{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}^{1/2}\sqrt{\frac{x}{n}}\Big) \leq \frac{1}{x}.$$

We therefore have two disadvantages compared to the statement of the lemma: Again, $\frac{1}{x}$ is a much larger term than e^{-x} for large x. Moreover, the lemma compares the rate with $\|\Sigma\|$ which is much smaller than $\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}^{1/2}$ which appears in (d). For instance, if $\Sigma = I_{d\times d}$, then one has $\|\Sigma\| = 1$, but $\{\mathrm{tr}(\Sigma)^2 + \|\Sigma\|_F^2\}^{1/2} = \sqrt{2d}$.

Solution 2 (Solution of Task 2)**.** (a) Each component β_j of β corresponds to the influence of X_j towards Y. In the modified model, β_1 corresponds to the expectation of Y (and not to any influence of X, since X_1 is constantly 1).
The aim of penalizing β is to choose only the components of X which have an influence on Y. If Y has nonzero expectation, the first component $X_1 = 1$ is always necessary. Therefore, β_1 should not be penalized.

(b) For the objective function (here denoted with F), we have

$$F(\beta) = \hat{R}_n(\beta) + \lambda \cdot \sum_{j=2}^{d} \beta_j^2 = \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\beta\|_2^2 + \sum_{j=2}^{d} \beta_j^2.$$

Differentiating and setting it equal to zero yields (with $E = \mathrm{diag}(0, 1, ..., 1) \in \mathbb{R}^{d\times d}$):

$$0 = \nabla_\beta F(\beta) = -\frac{2}{n}\mathbb{X}^T(\mathbb{Y} - \mathbb{X}\beta) + 2\lambda E\beta = -\frac{2}{n}\mathbb{X}^T\mathbb{Y} + 2(\frac{1}{n}\mathbb{X}^T\mathbb{X} + \lambda E)\beta.$$

Thus $\hat{\beta} = (\mathbb{X}^T\mathbb{X} + n\lambda E)^{-1}\mathbb{X}^T\mathbb{Y}$.

Solution 3 (Solution of Task 3)**.** (a) Markov's inequality yields

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}e^{cX}}{e^{ct}} = e^{\frac{c^2}{2} - ct}.$$

Here we have used that for $X \sim N(0, 1)$, it holds that $\mathbb{E}e^{cX} = e^{\frac{c^2}{2}}$.
Plugging in $c = t$ yields

$$\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{2}}.$$

(b) Here we use that $\frac{X}{\sigma} \sim N(0, 1)$. With (a), we conclude that

$$\mathbb{P}(|X| > \sigma t) \leq \mathbb{P}(X > \sigma t) + \mathbb{P}(-X > \sigma t) = \mathbb{P}(\frac{X}{\sigma} > t) + \mathbb{P}(-\frac{X}{\sigma} > t) \leq 2e^{-\frac{t^2}{2}}.$$

Define $t' := \sqrt{2t}$. Then we have

$$\mathbb{P}(|X| > \sigma\sqrt{2t}) = \mathbb{P}(|X| > \sigma t') \leq 2e^{-\frac{(t')^2}{2}} = 2e^{-t}.$$

(c) If $X \sim N(0,\sigma^2)$, the statement in (a) reads $\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$.
The central limit theorem yields for $n \to \infty$ that $\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mathbb{E}X_i) \xrightarrow{d} N(0,\sigma^2)$, thus

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\Big) \to \mathbb{P}(X > t) \overset{s.o.}{\leq} e^{-\frac{t^2}{2\sigma^2}}.$$

Bernstein's inequality provides the corresponding non-asymptotic statement (without $n \to \infty$) via

$$\mathbb{P}\Big(\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\Big) \leq \exp\Big(-\frac{1}{2}\frac{t^2}{\sigma^2 + \frac{Mt}{\sqrt{n}}}\Big).$$

One can see that the above expression converges to $e^{-\frac{t^2}{2\sigma^2}}$ for $n \to \infty$. In this sense, Bernstein's inequality can be considered as 'optimal' since it converges to the non-avoidable upper bound given by the Gaussian limit. If the distribution of X_i is unknown, Bernstein's inequality therefore yields a very good upper bound, in particular for large n.

(d) Due to $\frac{a}{b+c} \geq \frac{1}{2}\min\{\frac{a}{b}, \frac{a}{c}\}$ for $a, b, c > 0$ we have

$$\exp\Big(-\frac{1}{2}\frac{x^2}{\sigma^2 + \frac{Mx}{\sqrt{n}}}\Big) \leq \exp\Big(-\frac{1}{4}\min\{\frac{x^2}{\sigma^2}, \frac{x}{M/\sqrt{n}}\}\Big).$$

We conclude that

$$\begin{aligned}
\mathbb{P}\Big(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq \sqrt{n}\sigma\sqrt{t} + Mt\Big) &= \mathbb{P}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq \underbrace{\sigma\sqrt{t} + \frac{M}{\sqrt{n}}t}_{=:x'}\Big) \\
&\leq \exp\Big(-\frac{1}{4}\min\{\frac{(x')^2}{\sigma^2}, \frac{x'}{M/\sqrt{n}}\}\Big) \\
&\leq e^{-\frac{t}{4}}.
\end{aligned}$$

The last inequality holds due to $(x')^2 \geq t\sigma^2$ (only use the first summand of x') and $x' \geq \frac{M}{\sqrt{n}}t$ (only use the second summand of x').

(e) Choose $t = 4\log(\frac{1}{\delta})$. Then we conclude with (d):

$$\mathbb{P}\Big(\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq 2\sqrt{n}\sigma\sqrt{\log(\frac{1}{\delta})} + 4M\log(\frac{1}{\delta})\Big) \leq e^{-t/4} = \delta.$$

We conclude that with probability $\geq 1 - \delta$,

$$\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq 2\sqrt{n}\sigma\sqrt{\log(\frac{1}{\delta})} + 4M\log(\frac{1}{\delta}).$$

Solution 4 (Solution of Task 4)**.** (a) The LS estimator only works well if d is much smaller than n. Instead, the LASSO estimator only needs that $s\log(d)$ is much smaller than n (s denotes the number of non-zero components of β^*). Therefore, the LASSO estimator behaves like a LS estimator if the non-zero components of β^* would be known, up to the additional factor $\log(d)$.
In opposite to the LS estimator, the convergence rate of the LASSO estimator contains $\Lambda_{min}(\Sigma)$ instead of $\lambda_{min}(\Sigma)$.

(b) The reason is the proof technique where a statement about $\|\hat{\Sigma}^{1/2}(\hat{\beta}-\beta^*)\|_2^2$ is derived from an upper bound of $\|(\hat{\beta}-\beta^*)_S\|_1$. Therefore, the convergence rate of the excess Bayes risks can only be as good as the convergence rate of $\|(\hat{\beta}-\beta^*)_S\|_1$. The proof technique therefore incorporates the estimation quality of $\hat{\beta}$ to derive a statement about the quality of the algorithm. The estimation quality of $\hat{\beta}$ however is always dependent of Σ.

(c) For $v \in C$, we have

$$v^T\Sigma v = v^T v = \|v\|_2^2 \geq \|v_S\|_2^2$$

$\Rightarrow \Lambda_{min}(\Sigma) \geq 1$.

(d) The assumption on Σ in (a) means that all components of X are independent. This is not realistic since for large d, the entries of X might be strongly correlated in practice. As an example, consider Y as the age of a person. X may contain 'measurable' quantities like the show size, the body height, the weight and so on. Obviously, weight and body height are correlated already. The more data is collected from a specific person, the more probable it is that these data contains correlations. It may even happen that some components of X are perfectly correlated in the sense that one component can be derived from other components. Then Σ would even not be invertible.
(the following statement holds: If there exists some $c \in \mathbb{R}^d$ with $c^T X = 0$, then Σ is not invertible).

(e) It holds that

$$\lambda_{min}(\Sigma) = 1 - (d-1)^{1/2}\rho > 0$$

if and only if $\rho < (d-1)^{-1/2}$. Thus, ρ (which can be interpreted as the strength of correlations) has to be very small to guarantee that Σ is invertible. As an example, if $\rho < \frac{1}{2}(d-1)^{-1/2}$, then we have $\lambda_{min}(\Sigma) \geq \frac{1}{2}$.

(f) For $v \in C$, we have $\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1 \leq 4\|v_S\|_1$, thus

$$\begin{aligned} v^T\Sigma v &= \sum_{j=1}^{d} v_j^2 + 2\rho v_d \sum_{j=1}^{d-1} v_j \geq \|v\|_2^2 - 2\rho\|v\|_1^2 \geq \|v\|_2^2 - 32\rho\|v_S\|_1^2 \\ &\geq \|v_S\|_2^2 - 32\rho s\|v_S\|_2^2 = (1-32\rho s)\|v_S\|_2^2. \end{aligned}$$

$\Rightarrow \Lambda_{min}(\Sigma) \geq 1 - 32\rho s$.
We therefore have $\Lambda_{min}(\Sigma) > 0$ as long as $\rho \leq (32s)^{-1}$. As an example, if $\rho \leq \frac{1}{64s}$, then $\Lambda_{min}(\Sigma) \geq \frac{1}{2}$.
Contrary to the smallest eigenvalue $\lambda_{min}(\Sigma)$, the Restricted Eigenvalue Property may only pose conditions on ρ dependent on s. For small s, the condition on ρ from (f) is much weaker than the condition on ρ from (e).

Solution 5 (Solution of Task 5)**.** (a) We have

$$\begin{aligned}\mathbb{P}(B_1^c) &\leq \mathbb{P}\Big(\exists j \in \{1,...,d\} \sum_{i=1}^n X_{ij}^2 > \frac{3n}{2}\Big) \leq d \cdot \max_{j\in\{1,...,d\}} \mathbb{P}\Big(\sum_{i=1}^n X_{ij}^2 > \frac{3n}{2}\Big) \\ &\leq d \cdot \frac{\mathbb{E}[e^{tX_{1j}^2}]^n}{e^{3tn/2}} \leq d \cdot ((1-2t)e^{3t})^{-n/2}.\end{aligned}$$

(b) With $t = \frac{1}{4}$, it holds that

$$\begin{aligned}\mathbb{P}(B_1^c) &\leq d \cdot (\frac{1}{2}e^{3/4})^{-n/2} = d\exp(-\frac{n}{2}(\frac{3}{4} + \log(\frac{1}{2}))) \\ &= d\exp(-\frac{nc}{2}) \overset{n\geq \frac{2}{c}(\log(d)+x)}{\leq} de^{-\log(d)}e^{-x} = e^{-x}.\end{aligned}$$

Solution 6 (Solution of Task 6)**.** First note that

$$\mathbb{P}\Big(\frac{2}{n}\|\mathbb{e}^T\mathbb{X}\|_\infty \geq \frac{\lambda}{2}\Big) = \mathbb{P}\Big(\max_{j=1,...,d} \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq d\cdot \max_{j=1,...,d} \mathbb{P}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big).$$

This inequality will be used in (a)-(c).

(a) It holds that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij} \sim N(0, \frac{\sigma^2}{n}\sum_{i=1}^n X_{ij}^2) = N(0,\sigma^2)$. Therefore, using $\mathbb{P}(N(0,1) > x) \leq e^{-x^2/2}$, we have

$$\mathbb{P}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq \mathbb{P}(|N(0,1)| > \frac{\lambda\sqrt{n}}{4\sigma}) \leq 2\exp\Big(-\frac{1}{2}(\frac{\lambda\sqrt{n}}{4\sigma})^2\Big).$$

(b) Markov's inequality applied with $g(x) = x^2$ yields

$$\mathbb{P}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq \frac{\mathbb{E}\Big[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big)^2\Big]}{\big(\frac{\lambda\sqrt{n}}{4}\big)^2},$$

and

$$\mathbb{E}\Big[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big)^2\Big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_i^2 X_{ij}^2] = \sigma^2 \cdot \frac{1}{n}\sum_{i=1}^n X_{ij}^2 = \sigma^2.$$

$\Rightarrow$

$$\mathbb{P}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq (\frac{4\sigma}{\lambda\sqrt{n}})^2$$

(c) Markov's inequality applied with $g(x) = x^p$ yields

$$\mathbb{P}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq \frac{\mathbb{E}\Big[\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big|^p\Big]}{\big(\frac{\lambda\sqrt{n}}{4}\big)^p},$$

and (using the hint)

$$\mathbb{E}\Big[\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big|^p\Big] \leq p^{p/2}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\varepsilon_i|^p |X_{ij}|^p]^{2/p}\Big)^{p/2} = p^{p/2}\mu_p^p\Big(\frac{1}{n}\sum_{i=1}^n X_{ij}^2\Big)^{p/2} = (p^{1/2}\mu_p)^p.$$

$\Rightarrow$

$$\mathbb{P}\Big(\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq \big(\frac{4p^{1/2}\mu_p}{\lambda\sqrt{n}}\big)^p.$$

(d) Markov's inequality yields

$$\mathbb{P}\Big(\max_{j=1,\dots,d}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq \frac{\mathbb{E}\big[\max_{j=1,\dots,d}\big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\big|\big]}{\big(\frac{\lambda\sqrt{n}}{4}\big)}.$$

Nemirovski's inequality yields

$$\begin{aligned}\mathbb{E}\big[\max_{j=1,\dots,d}\big|\sum_{i=1}^n \varepsilon_i X_{ij}\big|\big] &\leq (8\log(2d))^{1/2}\cdot\mathbb{E}\Big[\max_{j=1,\dots,d}\sum_{i=1}^n \varepsilon_i^2 X_{ij}^2\Big]^{1/2} \\ &\leq (8\log(2d))^{1/2}\cdot C\cdot\mathbb{E}\Big[\sum_{i=1}^n \varepsilon_i^2\Big]^{1/2} = (8\log(2d))^{1/2}\cdot C\sigma\sqrt{n}\end{aligned}$$

We conclude that

$$\mathbb{P}\Big(\max_{j=1,\dots,d}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big| > \frac{\lambda\sqrt{n}}{4}\Big) \leq \frac{4(8\log(2d))^{1/2}C\sigma}{\lambda\sqrt{n}}.$$

8.2 Solutions of Chapter 3

Solution 7 (Solution of Task 7)**.** (a) It holds that $\hat{R}_n(\hat{\beta}) \leq \hat{R}_n(\beta^*)$, thus

$$R(\hat{\beta}) - R(\beta^*) = \underbrace{\hat{R}_n(\hat{\beta}) - \hat{R}_n(\beta^*)}_{\leq 0} - \big\{\hat{R}_n(\hat{\beta}) - R(\hat{\beta}) - (\hat{R}_n(\beta^*) - R(\beta^*))\big\}.$$

(b) It holds that $\tilde{\beta} - \beta^* = T(\hat{\beta} - \beta^*)$, thus

$$\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 = T\cdot\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq \gamma.$$

By convexity of $\beta \mapsto \hat{R}_n(\beta)$, we have

$$\hat{R}_n(\tilde{\beta}) \leq T \underbrace{\hat{R}_n(\hat{\beta})}_{\leq \hat{R}_n(\beta^*)} + (1-T)\hat{R}_n(\beta^*) \leq \hat{R}_n(\beta^*),$$

that is, the statement from (a) also holds for $\tilde{\beta}$.
We obtain

$$\begin{aligned} R(\tilde{\beta}) - R(\beta^*) &\leq \left|\hat{R}_n(\tilde{\beta}) - R(\tilde{\beta}) - (\hat{R}_n(\beta^*) - R(\beta^*))\right| \\ &\leq \sup_{\beta: \|\Sigma^{1/2}(\beta-\beta^*)\|_2 \leq \gamma} \left|(\hat{R}_n(\beta) - R(\beta)) - (\hat{R}_n(\beta^*) - R(\beta^*))\right|. \end{aligned}$$

(c) On A it holds that

$$\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2^2 = R(\tilde{\beta}) - R(\beta^*) \leq Z_\gamma \leq a\gamma,$$

therefore

$$\|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2 \leq (a\gamma)^{1/2} \leq \frac{\gamma}{2}.$$

Using $T \cdot \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \overset{\text{see above}}{=} \|\Sigma^{1/2}(\tilde{\beta} - \beta^*)\|_2$, we obtain

$$\frac{\gamma \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2}{\gamma + \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2} \leq \frac{\gamma}{2}.$$

Rearranging terms yields $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq \gamma$.
Repeating the proof for $\hat{\beta}$ instead of $\tilde{\beta}$ yields that on the event A, it holds that

$$R(\hat{\beta}) - R(\beta^*) \leq a\gamma.$$

(d) We have seen that $A \subset \{R(\hat{\beta}) - R(\beta^*) \leq a\gamma\}$. We conclude that

$$\mathbb{P}(R(\hat{\beta}) - R(\beta^*) > a\gamma) \leq \mathbb{P}(A^c).$$

(e) An elementary calculation yields

$$\hat{R}_n(\beta) = \frac{1}{n}\| \underbrace{\mathbb{Y}}_{=\mathbb{e}+\mathbb{X}\beta^*} - \mathbb{X}\beta\|_2^2 = \frac{1}{n}\|\mathbb{e}\|_2^2 + \frac{2}{n}\mathbb{e}^T\mathbb{X}(\beta^* - \beta) + \underbrace{\frac{1}{n}\|\mathbb{X}(\beta^* - \beta)\|_2^2}_{X_i \overset{\text{determ.}}{=} \|\Sigma^{1/2}(\beta - \beta^*)\|_2^2}.$$

$\Rightarrow$

$$\hat{R}_n(\beta) - R(\beta) = \frac{1}{n}\|\mathbb{e}\|_2^2 + \frac{2}{n}\mathbb{e}^T\mathbb{X}(\beta^* - \beta).$$

$\Rightarrow$

$$(\hat{R}_n(\beta) - R(\beta)) - (\hat{R}_n(\beta^*) - R(\beta^*)) = \frac{2}{n}\mathbb{e}^T\mathbb{X}(\beta^* - \beta).$$

We therefore have

$$\begin{aligned} |Z_\gamma| &\leq \frac{2}{n}\sup|\mathbb{e}^T\mathbb{X}(\beta^*-\beta)| = \frac{2}{n}\sup|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\Sigma^{1/2}(\beta^*-\beta)| \\ &\leq \frac{2}{n}\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2 \sup\|\Sigma^{1/2}(\beta^*-\beta)\|_2 \\ &\leq \frac{2\gamma}{n}\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2. \end{aligned}$$

(f) It holds that

$$\mathbb{E}|Z_\gamma| = \frac{2\gamma}{n}\mathbb{E}\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2 \leq \frac{2\gamma}{n}\mathbb{E}[\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2^2]^{1/2}.$$

Here,

$$\mathbb{E}[\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2^2] = \mathrm{tr}(\Sigma^{-1/2}\mathbb{X}^T\underbrace{\mathbb{E}[\mathbb{e}\mathbb{e}^T]}_{=\sigma^2 I_{d\times d}}\mathbb{X}\Sigma^{-1/2}) = n\sigma^2\mathrm{tr}(\Sigma^{-1/2}\Sigma\Sigma^{-1/2}) = nd\sigma^2.$$

Thus

$$\mathbb{E}|Z_\gamma| \leq \frac{2\gamma}{n}\cdot\sqrt{nd}\sigma = 2\gamma\sigma(\frac{d}{n})^{1/2}.$$

(g) We want to guarantee that $\mathbb{P}(A^c) \leq \frac{1}{t}$. Markov's inequality yields

$$\mathbb{P}(A^c) \leq \frac{\mathbb{E}Z_\gamma}{a\gamma} \leq \frac{2\sigma(\frac{d}{n})^{1/2}}{a} = \frac{1}{t},$$

where the last equality holds if we choose $a = 2\sigma(\frac{d}{n})^{1/2}t$.
We have to satisfy $(a\gamma)^{1/2} \leq \frac{\gamma}{2}$, that is, $2a^{1/2} \leq \gamma^{1/2}$ or equivalently $4a \leq \gamma$.
We therefore choose $\gamma = 4a$.
We obtain the convergence rate

$$a\gamma = 4a^2 = 16\sigma^2\frac{d}{n}t^2.$$

(h) It holds that

$$\frac{1}{\sigma\sqrt{n}}\mathbb{e}^T\mathbb{X}\Sigma^{-1/2} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n \varepsilon_i\cdot(\Sigma^{-1/2}X_i) \sim N\Big(0, \underbrace{\frac{1}{n}\sum_{i=1}^n \Sigma^{-1/2}X_iX_i^T\Sigma^{-1/2}}_{=\Sigma^{-1/2}\underbrace{\frac{1}{n}\sum_{i=1}^n X_iX_i^T}_{=\Sigma}\Sigma^{-1/2}}\Big) = N(0, I_{d\times d}).$$

(i) It holds that $\mathbb{E}[\|W\|_2^2] = \sum_{j=1}^d \mathbb{E}[W_j^2] = d$. Markov's inequality yields

$$\mathbb{P}(\|W\|_2^2 - 2\mathbb{E}\|W\|_2^2 \geq t) = \mathbb{P}(\|W\|_2^2 \geq 2d + t) \leq \frac{\mathbb{E}[\exp(\frac{\|W\|_2^2}{4})]}{e^{-(2d+t)/4}}.$$

Since the components of W are independent, we conclude that

$$\mathbb{E}[\exp(\frac{\|W\|_2^2}{4})] = \mathbb{E}[\exp(W_1^2/4)]^d = \sqrt{2}^d.$$

Thus

$$\mathbb{P}(\|W\|_2^2 - 2\mathbb{E}\|W\|_2^2 \geq t) \leq (\sqrt{2}e^{-1/2})^d \cdot e^{-t/4} \leq e^{-t/4}.$$

(j) We have

$$\begin{aligned}\mathbb{P}(Z_\gamma > a\gamma) &\overset{(e)}{\leq} \mathbb{P}(\frac{2}{n}\|\mathbb{e}^T\mathbb{X}\Sigma^{-1/2}\|_2 > a) = \mathbb{P}(\frac{\sigma}{\sqrt{n}}\|W\|_2 > a) \\ &= \mathbb{P}(\|W\|_2 > \sqrt{2d+t}) = \mathbb{P}(\|W\|_2^2 > 2d + t) \leq e^{-t/4}.\end{aligned}$$

(k) a was already chosen. Additionally, we chose $\gamma = 4a$. We therefore obtain the convergence rate $\gamma a = 4a^2 = 4\frac{\sigma^2}{n}(2d + t)$.

Solution 8 (Solution of Task 8)**.** (a) We have

$$\begin{aligned}\mathbb{E}\Big[\max_{j=1,\dots,d}\Big|\sum_{i=1}^n \varepsilon_i X_{ij}\Big|\Big] &= \mathbb{E}\Big[\max_{j=1,\dots,d}\Big|\frac{1}{\sqrt{n}\Sigma_{jj}^{1/2}}\sum_{i=1}^n \varepsilon_i X_{ij}\Big|\Big] \cdot \sqrt{n}\max_{j=1,\dots,d}\Sigma_{jj}^{1/2} \\ &= \mathbb{E}[\max_{j=1,\dots,d}|W_j|] \cdot \sqrt{n}\max_{j=1,\dots,d}\Sigma_{jj}^{1/2}.\end{aligned}$$

Conditionally on ε_i, $i = 1, ..., n$, it holds that

$$W_j \sim N\Big(0, \frac{1}{n\Sigma_{jj}}\sum_{i=1}^n \varepsilon_i^2\Sigma_{jj}\Big) \overset{\varepsilon_i \in \{-1,1\}}{=} N(0,1).$$

(b) We have

$$\psi\big(\mathbb{E}[\max_{j=1,\dots,d}|W_j|]\big) \leq \mathbb{E}[\psi(\max_{j=1,\dots,d}|W_j|)] \overset{\psi \text{ nondecreas.}}{\leq} \mathbb{E}[\max_{j=1,\dots,d}\psi(|W_j|)] \leq \sum_{j=1}^d \mathbb{E}\psi(|W_j|).$$

(c) We have

$$\mathbb{E}\psi(|W_j|) = \mathbb{E}[\exp(W_j^2/4)] = \sqrt{2}.$$

$\Rightarrow \psi\big(\mathbb{E}[\max_{j=1,\dots,d}|W_j|]\big) \leq \sqrt{2}d$

$\Rightarrow \mathbb{E}[\max_{j=1,\dots,d}|W_j|] \leq \psi^{-1}(\sqrt{2}d) = 2\sqrt{\log(\sqrt{2}d)}.$

Solution 9 (Solution of Task 9)**.** (a) We have

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \le x) = \mathbb{P}(X \le x, Y = 1) + \mathbb{P}(X \le x, Y = -1) \\ &= \mathbb{P}(X \le x | Y = 1)\pi_1 + \mathbb{P}(X \le x | Y = -1)(1 - \pi_1) \\ &= \int_{-\infty}^{x} g_1(x)dx \cdot \pi_1 + \int_{-\infty}^{x} g_{-1}(x)dx \cdot (1 - \pi_1). \end{aligned}$$

$\Rightarrow g(x) = g_1(x)\pi_1 + g_{-1}(x)(1 - \pi_1)$.

(b) Bayes' theorem yields (here, $g_{X,Y}$ denotes the common density of X, Y, and $g_{X|Y=1} = g_1$ denotes the conditional density of X given $Y = 1$)

$$\eta(x) = \mathbb{P}(Y = 1 | X = x) = \frac{g_{X,Y}(x, 1)}{g(x)} = \frac{g_{X|Y=1}(x)\pi_1}{g(x)} = \frac{g_1(x)\pi_1}{g(x)}.$$

(a) $\Rightarrow 1 - \eta(x) = \frac{g_{-1}(x)(1-\pi_1)}{g(x)}$.
$\Rightarrow$

$$\log\Big(\frac{\eta(x)}{1 - \eta(x)}\Big) = \log\Big(\frac{g_1(x)\pi_1}{g_{-1}(x)(1 - \pi_1)}\Big) = \log\Big(\frac{\pi_1}{1 - \pi_1}\Big) + \log\Big(\frac{g_1(x)}{g_{-1}(x)}\Big).$$

(c) We have

$$\begin{aligned} \log\Big(\frac{g_1(x)}{g_{-1}(x)}\Big) &= \log\Big(\frac{\det(\Sigma_{-1})^{1/2} \exp\big(-\frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1)\big)}{\det(\Sigma_1)^{1/2} \exp\big(-\frac{1}{2}(x - \mu_{-1})^T \Sigma_2^{-1}(x - \mu_{-1})\big)}\Big) \\ &= \frac{1}{2}\log\Big(\frac{\det(\Sigma_{-1})}{\det(\Sigma_1)}\Big) - \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_{-1})^T \Sigma_{-1}^{-1}(x - \mu_{-1}). \end{aligned}$$

(d) (b),(c) $\Rightarrow$

$$\begin{aligned} \log\Big(\frac{\eta(x)}{1 - \eta(x)}\Big) &= \log\Big(\frac{g_1(x)}{g_{-1}(x)}\Big) = -\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_{-1})^T \Sigma^{-1}(x - \mu_{-1}) \\ &= \mu_1^T \Sigma^{-1} x - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 - \mu_{-1}^T \Sigma^{-1} x + \frac{1}{2}\mu_{-1}^T \Sigma^{-1}\mu_{-1} \\ &= \underbrace{(\mu_1 - \mu_{-1})^T \Sigma^{-1}}_{=:(\beta^*)^T} x. \qquad (*) \end{aligned}$$

The mappings $\delta_1(x) := \log\Big(\frac{\eta(x)}{1-\eta(x)}\Big)$ and $\delta_{-1}(x) := 0$ form optimal discriminant functions since

$$\delta_1(x) > \delta_{-1}(x) \quad \Longleftrightarrow \quad \eta(x) > \frac{1}{2} \quad \Longleftrightarrow \quad f^*(x) = 1.$$

Therefore, the model has optimal linear decision boundaries. Due to (*), the model assumption of logistic regression is fulfilled with $\beta^* := \Sigma^{-1}(\mu_1 - \mu_{-1})$.

(e) $\mu_{-1} = -\mu_1$.

(f) We rearrange terms as in (d) to obtain

$$\begin{aligned}
\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) &= \log\Big(\frac{\pi_1}{1-\pi_1}\Big) + \log\Big(\frac{g_1(x)}{g_{-1}(x)}\Big) = -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_{-1})^T\Sigma^{-1} \\
&= \log\Big(\frac{\pi_1}{1-\pi_1}\Big) + \mu_1^T\Sigma^{-1}x - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \mu_{-1}^T\Sigma^{-1}x + \frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} \\
&= \underbrace{\Big[\log\Big(\frac{\pi_1}{1-\pi_1}\Big) + \frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\Big]}_{=:\beta_0} + \underbrace{(\mu_1-\mu_{-1})^T\Sigma^{-1}}_{=:(\beta^*)^T}x. \qquad (*)
\end{aligned}$$

We conclude that the model has affine linear decision boundaries.
With $h(x) = (1, x)$, we have

$$\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) = (\beta_0, \beta^*)^T h(x),$$

that is, the model assumption of logistic regression holds for $(\tilde{X}, Y)$ with $\tilde{X} = h(X) = (1, X)$.

(g) It holds that

$$\begin{aligned}
\log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) &= \log\Big(\frac{\pi_1}{1-\pi_1}\Big) + \frac{1}{2}\log\Big(\frac{\det(\Sigma_{-1})}{\det(\Sigma_1)}\Big) \\
&\qquad -\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_{-1})^T\Sigma_{-1}^{-1}(x-\mu_{-1}) \\
&= \underbrace{\Big[\log\Big(\frac{\pi_1}{1-\pi_1}\Big) + \frac{1}{2}\log\Big(\frac{\det(\Sigma_{-1})}{\det(\Sigma_1)}\Big) + \frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\Big]}_{=:\beta_0} \\
&\qquad + \frac{1}{2}x^T\underbrace{(\Sigma_{-1}^{-1} - \Sigma_1^{-1})}_{=:2A}x + \underbrace{(\mu_1^T\Sigma_1^{-1} - \mu_{-1}\Sigma_{-1}^{-1})}_{=:a^T}x \\
&= \beta_0 + a^Tx + x^TAx \\
&= \beta_0 h(x)_1 + \sum_{j=1}^{d} a_j h(x)_{1+j} + \sum_{j,k=1}^{d} A_{jk} h(x)_{1+d+d(j-1)+k}
\end{aligned}$$

with $h(x) = (1, x_1, ..., x_d, x_1^2, ..., x_1x_d, x_2x_1, ..., x_2x_d, ..., x_dx_1, ..., x_d^2)$.

(h) We obtain the following rates: In (d), $\frac{d}{n}$, in (f), $\frac{d+1}{n}$, in (g): $\frac{1+d+d^2}{n}$.

Solution 10 (Solution of Task 10)**.** (a) This statement holds in general, even without the model assumption of logistic regression. We have

$$\delta^*(x) = \log\Big(\frac{\eta(x)}{1-\eta(x)}\Big) > 0 \quad \Longleftrightarrow \quad \eta(x) > \frac{1}{2} \quad \Longleftrightarrow \quad f^*(x) = 1.$$

Thus, $f^*(x) = \text{sign}(\delta^*(x))$.
Here, we have (cf. Task 9(f)(*)):

$$\delta^*(x) = \Big[\log\Big(\frac{\pi_1}{1-\pi_1}\Big) + \frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\Big] + (\mu_1-\mu_{-1})^T\Sigma^{-1}x.$$

(b) The law of total probability yields

$$\begin{aligned} R(f^*) &= \mathbb{P}(Y \neq f^*(X)) = \mathbb{P}(f^*(X) = -1|Y=1)\pi_1 + \mathbb{P}(f^*(X)=1|Y=-1)(1-\pi_1) \\ &= \mathbb{P}(\delta^*(X) < 0|Y=1)\pi_1 + \mathbb{P}(\delta^*(X) > 0|Y=-1)(1-\pi_1). \end{aligned}$$

(c) Conditionally on $Y=1$, we have $X \sim N(\mu_1, \Sigma)$, thus

$$\begin{aligned} \delta^*(X) &= T + \frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + (\mu_1-\mu_{-1})^T\Sigma^{-1}X \\ &\sim N\Big(T + \underbrace{\frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + (\mu_1-\mu_{-1})^T\Sigma^{-1}\mu_1}_{=\frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1}+\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1-\mu_{-1}^T\Sigma^{-1}\mu_1=\Delta/2}, (\mu_1-\mu_{-1})^T\Sigma^{-1}(\mu_1-\mu_{-1})\Big) \\ &= N(T+\frac{\Delta}{2}, \Delta). \end{aligned}$$

Similarly, conditionally on $Y=-1$ we have $X \sim N(\mu_{-1}, \Sigma)$, which yields

$$\begin{aligned} \delta^*(X) &\sim N\Big(T + \underbrace{\frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1} - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + (\mu_1-\mu_{-1})^T\Sigma^{-1}\mu_{-1}}_{=-\frac{1}{2}\mu_{-1}^T\Sigma^{-1}\mu_{-1}-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1+\mu_{-1}^T\Sigma^{-1}\mu_1=-\Delta/2}, (\mu_1-\mu_{-1})^T\Sigma^{-1}(\mu_1-\mu_{-1})\Big) \\ &= N(T-\frac{\Delta}{2}, \Delta). \end{aligned}$$

(d) For $Z \sim N(a,b)$ it holds that $\mathbb{P}(Z<0) = \mathbb{P}(\frac{Z-a}{\sqrt{b}} < -\frac{a}{\sqrt{b}}) = \Phi(-\frac{a}{\sqrt{b}})$ und $\mathbb{P}(Z > 0) = 1 - \mathbb{P}(Z \leq 0) = 1 - \Phi(-\frac{a}{\sqrt{b}})$.
(b),(c) $\Rightarrow$

$$\begin{aligned} R(f^*) &= \mathbb{P}(\delta^*(X) < 0|Y=1)\pi_1 + \mathbb{P}(\delta^*(X) > 0|Y=-1)(1-\pi_1) \\ &= \Phi\Big(\frac{-T-\frac{\Delta}{2}}{\sqrt{\Delta}}\Big)\pi_1 + \Big(1 - \Phi\Big(\frac{-T+\frac{\Delta}{2}}{\sqrt{\Delta}}\Big)\Big)(1-\pi_1). \end{aligned}$$

(e) $\pi_1 = \frac{1}{2} \Rightarrow T = 0$.
$\Sigma = I_{d\times d}$, $\mu_{-1} = -\mu_1 \Rightarrow \Delta = (\mu_1-\mu_{-1})^T\Sigma^{-1}(\mu_1-\mu_{-1}) = 2\mu_1^T(2\mu_1) = 4\|\mu_1\|_2^2$.
Plugging in the result from (d) and using $1-\Phi(x) = \Phi(-x)$, we have

$$\begin{aligned} R(f^*) &= \frac{1}{2}\Phi\Big(-\frac{2\|\mu_1\|_2^2}{2\|\mu_1\|_2}\Big) + \frac{1}{2}\Big(1 - \Phi\Big(\frac{2\|\mu_1\|_2^2}{2\|\mu_1\|_2}\Big)\Big) \\ &= \Phi(-\|\mu_1\|_2). \end{aligned}$$

Graphically this means that the more μ_1, $\mu_{-1} = -\mu_1$ are apart from each other, the smaller is $R(f^*)$. For $\mu_1 = 0$, the maximum is attained at $R(f^*) = \frac{1}{2}$. This corresponds to a random guess of the class since in this case, the distributions $X|Y = 1$ and $X|Y = -1$ are equal.

8.3 Solutions of Chapter 4

Solution 11 (Solution of Task 11)**.** (a) Due to $y \in \{+1, -1\}$, we have

$$\tilde{L}(y, s) = (y - s)^2 = y^2 - 2ys + s^2 = 1 - 2ys + (ys)^2 = (1 - ys)^2 = \phi(-ys)$$

with $\phi(x) = (1 + x)^2$.

(b) We have $\Phi_\eta(z) = \phi(-z)\eta + \phi(z)(1 - \eta) = (1 - z)^2\eta + (1 + z)^2(1 - \eta)$.
We now search for a minimizer of $z \mapsto \Phi_\eta(z)$ via

$$0 = \Phi'_\eta(z) = -2(1 - z)\eta + 2(1 + z)(1 - \eta) = 2 + 2z - 4\eta \quad \Rightarrow \quad z = 2\eta - 1.$$

Theorem 3.19 $\Rightarrow \delta^*(x) = 2\eta(x) - 1$ with $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

(c) It holds that $\delta^*(x) > 0 \iff \eta(x) > \frac{1}{2} \iff f^*(x) = \arg\max_{k\in\{+1,-1\}} \mathbb{P}(Y = k|X = x) = 1$.
$\Rightarrow f^*(x) = \text{sign}(\delta^*(x))$
$\Rightarrow$ The calibration condition is satisfied.

(d) $\delta^* \in \Delta$ is equivalent to $\forall x \in \mathcal{X} : 2\eta(x) - 1 = \delta^*(x) = x^T\beta^*$ for some $\beta^* \in \mathbb{R}^d$, that is, we have to satisfy

$$\eta(x) = \frac{1}{2} + \frac{1}{2}x^T\beta^*.$$

Note that the left hand side is in $\in [0, 1]$. Therefore this can only be fulfilled if $\mathcal{X}$ is bounded.

(e) We have already seen in (b) that $g(\eta) = \arg\min_{z\in\mathbb{R}} \Phi_\eta(z) = 2\eta - 1$.
Moreover,

$$H(\eta) = \Phi_\eta(g(\eta)) = (2 - 2\eta)^2\eta + (2\eta)^2(1 - \eta) = 4\eta(1 - \eta).$$

We conclude that

$$1 - H(\eta) = 1 - 4\eta + 4\eta^2 = 4(\eta - \frac{1}{2})^2 \quad \Rightarrow \quad (\frac{1}{2})^2(1 - H(\eta)) = (\eta - \frac{1}{2})^2.$$

Furthermore, $\phi(0) = 1$ and ϕ is convex. Theorem 3.21 $\Rightarrow$ With $s = 2$, $C_H = \frac{1}{2}$, the risk transfer formula is satisfied with

$$G(r) = 2C_H r^{1/s} = r^{1/2}.$$

(f) For $q \in [0, \infty)$, we have $G(r) = 4C_H^{\frac{s(q+1)}{q+s}} C^{-\frac{1}{q+s}} r^{\frac{q+1}{q+s}} = 4(\frac{1}{2})^{\frac{2(q+1)}{q+2}} C^{-\frac{1}{q+2}} r^{\frac{q+1}{q+2}}$.
For $q = \infty$, we have $G(r) = 2\frac{C_H^s}{c^{s-1}} r = \frac{1}{2c} \cdot r$.

(g) Due to $\frac{1}{2} X^T \beta^* \sim N(\frac{1}{2}\mu^T\beta^*, \frac{1}{4}\|\beta^*\|_2^2) = N(0,1)$, it holds that

$$\begin{aligned} \mathbb{P}(|\eta(X) - \frac{1}{2}| \leq t) &= \mathbb{P}(\frac{1}{2}|X^T\beta^*| \leq t) \overset{Z\sim N(0,1)}{=} \mathbb{P}(|Z| \leq t) = 2\Phi(t) - 1 \\ &\overset{\text{Hinweis}}{\leq} 2\Phi'(0)t = \sqrt{\frac{2}{\pi}} \cdot t. \end{aligned}$$

Here, $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-y^2/2) dy$ is the distribution function of the standard normal distribution and $\Phi'(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$.
Therefore, the noise condition

$$\mathbb{P}(|\eta(X) - \frac{1}{2}| \leq t) \leq Ct^q \quad \forall t > 0$$

is satisfied with $C = \sqrt{\frac{2}{\pi}}$, $q = 1$.

Solution 12 (Solution of Task 12)**.** Für $\phi(x) = e^x$:

(a) $\phi'(x) = e^x > 0 \Rightarrow \phi$ nondecreasing,
$\phi''(x) > 0 \Rightarrow \phi$ convex,
$\phi(0) = e^0 = 1$.

(b) We have $\Phi_\eta(z) = \phi(-z)\eta + \phi(z)(1-\eta) = e^{-z}\eta + e^z(1-\eta)$. A minimizer of $z \mapsto \Phi_\eta(z)$ is found via

$$0 = \Phi_\eta'(z) = -e^{-z}\eta + e^z(1-\eta) \quad \Rightarrow \quad z = \frac{1}{2}\log(\frac{\eta}{1-\eta}).$$

Theorem 3.19 $\Rightarrow \delta^*(x) = \frac{1}{2}\log(\frac{\eta(x)}{1-\eta(x)})$ with $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

(c) It holds that $\delta^*(x) = \frac{1}{2}\log(\frac{\eta(x)}{1-\eta(x)}) > 0 \iff \frac{\eta(x)}{1-\eta(x)} > 1 \iff \eta(x) > \frac{1}{2} \iff f^*(x) = \arg\max_{k\in\{+1,-1\}} \mathbb{P}(Y = k|X = x) = 1$.
$\Rightarrow f^*(x) = \text{sign}(\delta^*(x))$
$\Rightarrow$ The calibration condition is satisfied.

(d) We have already seen in (b) that $g(\eta) = \arg\min_{z\in\mathbb{R}} \Phi_\eta(z) = \frac{1}{2}\log(\frac{\eta}{1-\eta})$.
Moreover,

$$\begin{aligned} H(\eta) &= \Phi_\eta(g(\eta)) = \exp\Big(-\frac{1}{2}\log(\frac{\eta}{1-\eta})\Big)\eta + \exp\Big(\frac{1}{2}\log(\frac{\eta}{1-\eta})\Big)(1-\eta) \\ &= \Big(\frac{\eta}{1-\eta}\Big)^{-1/2}\eta + \Big(\frac{\eta}{1-\eta}\Big)^{1/2}(1-\eta) \\ &= 2(1-\eta)^{1/2}\eta^{1/2} = 2((1-\eta)\eta)^{1/2}. \end{aligned}$$

Thus,

$$p(\eta) := 1 - H(\eta) = 1 - 2((1-\eta)\eta)^{1/2}.$$

Here, it holds that $p'(\eta) = 2(\eta - \frac{1}{2}) \cdot (\eta(1-\eta))^{-1/2}$, $p''(\eta) = \frac{1}{2}(\eta(1-\eta))^{-3/2} \geq \frac{1}{2}(\frac{1}{4})^{-3/2} = 4$.
A Taylor expansion at $\eta = \frac{1}{2}$ yields

$$1-H(\eta) = p(\eta) = p(\frac{1}{2}) + (\eta - \frac{1}{2})p'(\frac{1}{2}) + \frac{1}{2}(\eta - \frac{1}{2})^2 p''(\xi) \geq \frac{1}{2}(\eta - \frac{1}{2})^2 \cdot 4 = 2(\eta - \frac{1}{2})^2.$$

Note that $\phi(0) = 1$ and ϕ is convex. Theorem 3.21 $\Rightarrow$ With $s = 2$, $C_H = \frac{1}{\sqrt{2}}$, the risk transfer formula is satisfied with

$$G(r) = 2C_H r^{1/s} = \sqrt{2} r^{1/2}.$$

For $\phi(x) = \max\{1 + x, 0\}$, we have:

(a) $1 + x, 0$ are increasing $\Rightarrow \phi(x) = \max\{1 + x, 0\}$ is increasing (maximum preserves monotonicity).
$1 + x, 0$ convex $\Rightarrow \phi(x) = \max\{1 + x, 0\}$ convex (maximum preserves convexity).
$\phi(0) = \max\{1, 0\} = 1$.

(b) We have $\Phi_\eta(z) = \phi(-z)\eta + \phi(z)(1-\eta) = \max\{1 - z, 0\}\eta + \max\{1 + z, 0\}(1-\eta)$.
We now search for a minimizer of $z \mapsto \Phi_\eta(z)$.
For $z < -1$, $\Phi_\eta(z) = (1-z)\eta$ is decreasing, for $z > 1$, $\Phi_\eta(z) = (1+z)(1-\eta)$ is increasing. $\Rightarrow$ the minimizer of Φ_η is located in $[-1, 1]$.
For $z \in [-1, 1]$, we have

$$\Phi_\eta(z) = (1-z)\eta + (1+z)(1-\eta) = 1 + 2z \cdot (\frac{1}{2} - \eta).$$

For $\eta > \frac{1}{2}$ this is minimal for $z = 1$, for $\eta < \frac{1}{2}$ this is minimal for $z = -1$. For $\eta = \frac{1}{2}$, there is no unique minimizer.
$\Rightarrow z = \text{sign}(\eta - \frac{1}{2})$ is a minimizer of Φ_η.
Theorem 3.19 implies that $\delta^*(x) = \text{sign}(\eta(x) - \frac{1}{2})$.

(c) It holds that $\delta^*(x) = \text{sign}(\eta(x) - \frac{1}{2}) = f^*(x)$, in particular we have $f^*(x) = \text{sign}(\delta^*(x))$
$\Rightarrow$ The calibration condition is satisfied.
Note that here, $\delta^*(x)$ itself also only takes discrete values $\{-1, +1\}$!

(d) We have already seen in (b) that $g(\eta) = \text{sign}(\eta - \frac{1}{2})$.
Moreover, we have

$$H(\eta) \;=\; \Phi_\eta(g(\eta)) = 1 + 2\text{sign}(\eta - \frac{1}{2}) \cdot (\frac{1}{2} - \eta) = 1 - 2|\eta - \frac{1}{2}|.$$

Thus

$$1 - H(\eta) = 2|\eta - \frac{1}{2}|.$$

Additionally, we have that $\phi(0) = 1$ and ϕ is convex. The conditions of Theorem 3.21 are fulfilled with $s = 1$, $C_H = \frac{1}{2}$, and we obtain the risk transfer formula with

$$G(r) = 2C_H r^{1/s} = r.$$

Solution 13 (Solution of Task 13)**.** (a) It holds that

$$L(\theta, p) \;=\; \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \quad + \sum_{i=1}^n \alpha_i\big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big) \quad - \sum_{i=1}^n \gamma_i\xi_i$$

We now investigate the three derivatives $\nabla_\theta = (\nabla_\beta, \nabla_{\beta_0}, \nabla_\xi)$:

$$\nabla_\beta L(\theta, p) = \beta - \sum_{i=1}^n \alpha_i Y_i X_i, \qquad \nabla_{\beta_0} L(\theta, p) = \sum_{i=1}^n \alpha_i Y_i, \qquad \nabla_{\xi_i} L(\theta, p) = C - \alpha_i - \gamma_i.$$

The optimality conditions $\nabla_\theta L(\theta, p) = 0$ now directly imply the assertion.

(b) We plug in the constraints $\nabla_\theta L(\theta, p) = 0$ in the objective function $L(\theta, p)$ and rearrange terms. During this procedure, we have to take care that the constraints regarding α are still preserved. It holds that

$$\begin{aligned}
L(\theta, p) \quad &= \quad \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \quad + \sum_{i=1}^n \alpha_i\big(1 - \xi_i - Y_i(X_i^T\beta + \beta_0)\big) \quad - \sum_{i=1}^n \gamma_i\xi_i \\
&= \quad \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^n \alpha_i + \sum_{i=1}^n \xi_i \cdot \underbrace{(C - \alpha_i - \gamma_i)}_{=0} - \beta_0 \underbrace{\sum_{i=1}^n \alpha_i Y_i}_{=0} - \Big(\underbrace{\sum_{i=1}^n \alpha_i Y_i X_i^T}_{=\beta^T}\Big)\beta \\
&\overset{\nabla_\theta L(\theta,p)=0}{=} \quad \frac{1}{2}\|\beta\|_2^2 + \sum_{i=1}^n \alpha_i - \|\beta\|_2^2 \\
&= \quad \sum_{i=1}^n \alpha_i - \frac{1}{2}\|\beta\|_2^2 \\
&= \quad \sum_{i=1}^n \alpha_i - \frac{1}{2}\Big\|\sum_{i=1}^n \alpha_i Y_i X_i\Big\|_2^2 \\
&= \quad \underbrace{\sum_{i=1}^n \alpha_i}_{=\mathbb{1}^T\alpha} - \frac{1}{2}\underbrace{\sum_{i,j=1}^n \alpha_i\alpha_j X_i^T X_j Y_i Y_j}_{=\alpha^T Q\alpha} \\
&= \quad \mathbb{1}^T\alpha - \frac{1}{2}\alpha^T Q\alpha.
\end{aligned}$$

A maximizer of $L(\theta, p)$ is a minimizer of $-L(\theta, p)$. This yields the optimizing function stated in the task.
The constraints posed on α read:

- $\nabla_{\beta_0} L(\theta, p) = 0$ is equivalent to $\mathbb{Y}^T \alpha = \sum_{i=1}^n \alpha_i Y_i = 0$.
- $\nabla_\beta L(\theta, p) = \beta - \sum_{i=1}^n \alpha_i Y_i X_i = 0$ is no longer relevant because β is no longer appearing in the optimization problem.
- $\nabla_{\xi_i} L(\theta, p) = C - \alpha_i - \gamma_i = 0$ yields $\alpha_i = C - \gamma_i \leq C$ (since $\gamma_i \geq 0$). This yields the constraint $\alpha_i \leq C$.

(c) The optimality condition $\nabla_\beta L(\hat{\theta}, \hat{p}) = 0$ yields $\hat{\beta} = \sum_{i=1}^n \alpha_i Y_i X_i$.
Due to $G(\hat{\theta}) \leq 0$, $\hat{p} \geq 0$ we have

$$0 = G(\hat{\theta})^T \hat{p} = \sum_j G(\hat{\theta})_j \hat{p}_j \quad \Longleftrightarrow \quad \forall j : G(\hat{\theta})_j \hat{p}_j = 0.$$

In particular, we have for all $i = 1, ..., n$ that

$$\hat{\gamma}_i \hat{\xi}_i = 0, \qquad \hat{\alpha}_i (1 - \hat{\xi}_i - Y_i (X_i^T \hat{\beta} + \hat{\beta}_0)) = 0. \qquad (*)$$

Let $i \in \{1, ..., n\}$ be some index with $\hat{\alpha}_i \in (0, C)$.
$\Rightarrow 0 = \nabla_{\beta_0} L(\hat{\theta}, \hat{p}) = C - \hat{\alpha}_i - \hat{\gamma}_i \Rightarrow \hat{\gamma}_i = C - \hat{\alpha}_i > 0$.
$(*) \Rightarrow \hat{\xi}_i = 0$
$(*)$, $\hat{\alpha}_i \neq 0 \Rightarrow 0 = 1 - \hat{\xi}_i - Y_i (X_i^T \hat{\beta} + \hat{\beta}_0) = 1 - Y_i (X_i^T \hat{\beta} + \hat{\beta}_0) \Rightarrow 0 = Y_i - Y_i^2 (X_i^T \hat{\beta} + \hat{\beta}_0) = Y_i - X_i^T \hat{\beta} - \hat{\beta}_0$
$\Rightarrow \hat{\beta}_0 = Y_i - X_i^T \hat{\beta}$.

Solution 14 (Solution of Task 14)**.** (a) It holds that

$$\begin{aligned} K_2(x, x') &= (1 + x^T x')^2 \\ &= (1 + x_1 x_1' + x_2 x_2')^2 = 1 + x_1^2 (x_1')^2 + x_2^2 (x_2')^2 + 2 x_1 x_1' + 2 x_2 x_2' + 2 x_1 x_2 x_1 x_2' \\ &= h(x)^T h(x') \end{aligned}$$

with

$$h(x) = (1, x_1^2, x_2^2, \sqrt{2} x_1, \sqrt{2} x_2, \sqrt{2} x_1 x_2).$$

(b) With $\tilde{x} := (1, x)$, $\tilde{x}' := (1, x')$, we have

$$\begin{aligned} K_p(x, x') &= (\tilde{x}, \tilde{x}')^p = \Big(\sum_{k=1}^{d+1} \tilde{x}_k \tilde{x}_k' \Big)^p \\ &= \sum_{i_1, ..., i_p = 1}^{d+1} (\tilde{x}_{i_1} \cdot ... \cdot x_{i_p}) \cdot (\tilde{x}_{i_1}' \cdot ... \cdot x_{i_p}') \\ &= h(x)^T h(x'), \end{aligned}$$

where $h(x) = (\tilde{x}_{i_1} \cdot ... \tilde{x}_{i_p})_{i_1,...,i_p=1,...,d+1}$.
Of course, this is not an 'optimal' representation, the above h even contains some components twice. However this representation is enough to get an idea which nonlinear transformations correspond to K.
Obviously, we have $m \leq |\{(i_1, ..., i_p) : i_1, ..., i_p \in \{1, ..., d+1\}\}| = (d+1)^p$.

(c) It holds that

$$\begin{aligned} K_\gamma(x, x') &= \exp(-\gamma(x - x')^2) = \exp(-\gamma x^2 + 2\gamma x x' - \gamma (x')^2) \\ &= e^{-\gamma(x^2 + (x')^2)} \sum_{k=0}^{\infty} \frac{(2\gamma x x')^k}{k!} \\ &= \sum_{k=0}^{\infty} \Big(\frac{(2\gamma)^{k/2}}{(k!)^{1/2}} x^k e^{-\gamma x^2} \Big) \cdot \Big(\frac{(2\gamma)^{k/2}}{(k!)^{1/2}} (x')^k e^{-\gamma (x')^2} \Big), \end{aligned}$$

that is, we can choose $h(x) = \big(\frac{(2\gamma)^{k/2}}{(k!)^{1/2}} x^k e^{-\gamma x^2} \big)_{k \in \mathbb{N}_0}$.
That is, h corresponds to polynomials x^k ($k \in \mathbb{N}_0$) whose oscillation for large x is damped with some exponential factor $e^{-\gamma x^2}$.

(d) Let H be the function h for the one-dimensional Gaussian kernel from (c). Then it holds that

$$\begin{aligned} K_\gamma(x, x') &= e^{-\gamma \|x - x'\|_2^2} = \prod_{j=1}^{d} e^{-\gamma (x_j - x'_j)^2} = \prod_{j=1}^{d} \sum_{k=0}^{\infty} H_k(x_k) H_k(x'_k) \\ &= \sum_{k_1,...,k_d=0}^{\infty} (H_{k_1}(x_{k_1}) \cdot ... \cdot H_{k_d}(x_{k_d})) \cdot (H_{k_1}(x'_{k_1}) \cdot ... \cdot H_{k_d}(x'_{k_d})), \end{aligned}$$

that is, we can choose

$$h(x) = (H_{k_1}(x_{k_1}) \cdot ... \cdot H_{k_d}(x_{k_d}))_{k_1,...,k_d \in \mathbb{N}_0} = \Big(\frac{(2\gamma)^{(k_1 + ... + k_d)/2}}{(k!)^{d/2}} x_1^{k_1} \cdot ... \cdot x_d^{k_d} e^{-\gamma \|x\|_2^2} \Big)_{k_1,...,k_d \in \mathbb{N}_0}.$$

Solution 15 (Solution of Task 15)**.** (a) Choose $a = C$ in the definition of $\gamma(n)$.
By assumption, we have $\sum_{k>a} \gamma_j = 0$ which implies $\gamma(n) \leq \frac{C}{n}$.

(b) Hint $\Rightarrow$ For all $a \in \mathbb{N}$, it holds that

$$\gamma(n) \leq \frac{1}{\sqrt{n}} \Big\{ \frac{a}{\sqrt{n}} + C^{1/2} c_\alpha^{1/2} a^{(1-\alpha)/2} \Big\}.$$

Derivation of the right hand side with respect to a yields: $\frac{1}{\sqrt{n}} - C^{1/2} c_\alpha^{1/2} \frac{\alpha - 1}{2} a^{-\frac{\alpha+1}{2}}$.
This motivates the choice

$$a = \lceil (nC)^{\frac{1}{\alpha+1}} \rceil.$$

We did *not* take into account the constants c_α, α for the choice of a since they do not have a direct influence on the rate with respect to n.
$\Rightarrow$

$$\gamma(n) \leq \frac{1}{\sqrt{n}}\Big\{\frac{1}{\sqrt{n}} + n^{\frac{1-\alpha}{2(\alpha+1)}} C^{\frac{1}{\alpha+1}} + c_\alpha^{1/2} n^{\frac{1-\alpha}{2(\alpha+1)}} C^{\frac{1}{\alpha+1}}\Big\} \leq \tilde{c}_\alpha \cdot \Big(\frac{1}{n} + C^{\frac{1}{\alpha+1}} n^{-\frac{\alpha}{\alpha+1}}\Big).$$

(c) We apply the same strategy as in (b): For all $a \in \mathbb{N}$, we have

$$\gamma(n) \leq \frac{1}{\sqrt{n}}\Big\{\frac{a}{\sqrt{n}} + C^{1/2}\frac{1}{(1-\rho)^{1/2}}\rho^{a/2}\Big\}.$$

Choose $a = \lceil 2\log_\rho((nC)^{-1})\rceil$. This implies

$$\gamma(n) \leq \frac{1}{\sqrt{n}}\Big\{\frac{1}{\sqrt{n}} + 2\log_\rho((nC)^{-1}) + C^{1/2}\frac{1}{(1-\rho)^{1/2}}(nC)^{-1}\Big\} \leq c_\rho \cdot \Big(\frac{1}{n} + 2\frac{\log(nC)}{nC^{1/2}}\Big).$$

(d) For h_k it holds that

$$(T_K h_k)(x) = \int K(x,x')h_k(x')dx' = \sum_{l=1}^{\infty} h_l(x)\int h_l(x')h_k(x')dx' \overset{\text{orthog.}}{=} h_k(x)\underbrace{\int h_k(x')^2 dx'}_{=:\gamma_k},$$

that is, h_k is an eigenfunction of T_K corresponding to the eigenvalue $\gamma_k = \int h_k(x')^2 dx'$.

(e) From (d) we conclude that T_K has at most as much nonzero eigenvalues as h has components.
A1(b) $\Rightarrow$ h has at most $(d+1)^p$ components.
(a) applied with $C = (d+1)^p \Rightarrow \gamma(n) \leq \frac{(d+1)^p}{n}$.

Solution 16 (Solution of Task 16)**.** For $t \in (0, \frac{1}{2})$, it holds that

$$\begin{aligned}
\mathbb{E}[(\delta(X) - \delta^*(X))^2] &= \mathbb{E}[\underbrace{(\delta(X) - \delta^*(X))^2}_{\leq A(x)\cdot\{\mathbb{E}[\tilde{L}(Y,\delta(X)) - \tilde{L}(Y,\delta^*(X))|X=x]\}} \underbrace{\mathbb{1}_{\{|\eta(X)-\frac{1}{2}|>t\}}}_{\leq 1}] \\
&\quad + \mathbb{E}[\underbrace{(\delta(X) - \delta^*(X))^2}_{\leq \|\delta-\delta^*\|_\infty^2 \leq (\rho+1)^2} \mathbb{1}_{\{|\eta(X)-\frac{1}{2}|\leq t\}}] \\
&\leq c_\rho(t)\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + (\rho+1)^2 \cdot \mathbb{P}(|\eta(X) - \frac{1}{2}| \leq t) \\
&\leq \frac{2\rho}{\eta_1}\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + \frac{2}{t}\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + C(\rho+1)^2 t^q.
\end{aligned}$$

Choose $t = \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}^{\frac{1}{q+1}}$. Then we have

$$\begin{aligned}
\mathbb{E}[(\delta(X) - \delta^*(X))^2] &\leq \frac{2\rho}{\eta_1}\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + \big(2 + C(\rho+1)^2\big)\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}^{\frac{q}{q+1}} \\
&\overset{\text{Hint 3}}{\leq} 2\{\rho\underbrace{(\rho+1)^{1/q}}_{\leq \rho+1} + 2 + C(\rho+1)^2\}\{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}^{\frac{q}{q+1}}.
\end{aligned}$$

This implies the assertion.

Solution 17 (Solution of Task 17)**.** (a) $\mathcal{H}_K$ is a vector space and $\delta_0, \hat{\delta} \in \mathcal{H}_K \Rightarrow \tilde{\delta} = T\hat{\delta} + (1-T)\delta_0 \in \mathcal{H}_K$. Moreover,

$$\|\tilde{\delta}\|_K \leq T\|\hat{\delta}\|_K + (1-T)\|\delta_0\|_K \leq T\rho + (1-T)\rho = \rho.$$

$\Rightarrow \tilde{\delta} \in B(\rho)$.
$\tilde{\delta} - \delta_0 = T(\hat{\delta} - \delta_0) \Rightarrow$

$$D(\tilde{\delta}, \delta_0) = TD(\hat{\delta}, \delta_0) = \frac{rD(\hat{\delta}, \delta_0)}{r + D(\hat{\delta}, \delta_0)} \leq r.$$

(b) It holds that $\tilde{R}_n(\hat{\delta}) \leq \tilde{R}_n(\delta_0)$, thus

$$\tilde{R}_n(\tilde{\delta}) = \tilde{R}_n(T\hat{\delta} + (1-T)\delta_0) \overset{s\mapsto\tilde{L}(y,s) \text{ convex}}{\leq} T \underbrace{\tilde{R}_n(\hat{\delta})}_{\leq \tilde{R}_n(\delta_0)} + (1-T)\tilde{R}_n(\delta_0) \leq \tilde{R}_n(\delta_0).$$

We conclude that

$$\tilde{R}(\tilde{\delta}) - \tilde{R}(\delta_0) = \underbrace{\{\tilde{R}_n(\tilde{\delta}) - \tilde{R}_n(\delta_0)\}}_{\leq 0} + \underbrace{\{\tilde{R}(\tilde{\delta}) - \tilde{R}_n(\tilde{\delta}) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0))\}}_{\overset{(a)}{\leq} Z_r}$$

$\Rightarrow$

$$\tilde{R}(\tilde{\delta}) - \tilde{R}(\delta^*) = \{\tilde{R}(\tilde{\delta}) - \tilde{R}(\delta_0)\} + \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} \leq \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + Z_r.$$

(c) Using the assumption, we have on A that

$$\tilde{R}(\tilde{\delta}) - \tilde{R}(\delta^*) \leq \frac{1}{8c_\rho}(\frac{r}{2})^2 + \frac{1}{8c_\rho}(\frac{r}{2})^2 = \frac{1}{4c_\rho}(\frac{r}{2})^2.$$

The quadratic margin property implies

$$\frac{1}{c_\rho} D(\tilde{\delta}, \delta^*)^2 \leq \tilde{R}(\tilde{\delta}) - \tilde{R}(\delta^*) \leq \frac{1}{4c_\rho}(\frac{r}{2})^2.$$

Rearranging terms implies

$$D(\tilde{\delta}, \delta^*) \leq \frac{r}{4}.$$

(d) By the quadratic margin property, we have

$$\frac{1}{c_\rho} D(\delta_0, \delta^*)^2 \leq \tilde{R}(\delta_0) - \tilde{R}(\delta^*) \leq \frac{1}{4c_\rho}(\frac{r}{2})^2.$$

$\Rightarrow D(\delta_0, \delta^*) \leq \frac{r}{4}$.
$\Rightarrow$

$$D(\tilde{\delta}, \delta_0) \leq D(\tilde{\delta}, \delta^*) + D(\delta_0, \delta^*) \leq \frac{r}{4} + \frac{r}{4} = \frac{r}{2}.$$

We have $\tilde{\delta} - \delta_0 = T(\hat{\delta} - \delta_0)$, thus

$$\frac{r}{2} \geq D(\tilde{\delta}, \delta^*) = TD(\hat{\delta}, \delta_0) = \frac{rD(\hat{\delta}, \delta_0)}{r + D(\hat{\delta}, \delta_0)}$$

Rearranging terms yields $D(\hat{\delta}, \delta_0) \leq r$.
Applying the technique from (b) to $\hat{\delta}$ yields on A that

$$\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \leq \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \frac{1}{8c_\rho}(\frac{r}{2})^2.$$

(e) We conclude that

$$\sup_{\delta \in B(\rho), D(\delta, \delta_0) \leq r} \{\tilde{R}(\delta) - \tilde{R}_n(\delta) - (\tilde{R}(\delta_0) - \tilde{R}_n(\delta_0))\} = \sup_{\delta \in \mathcal{F}} f_\delta(X_i, Y_i),$$

where $\mathcal{F} = \{\delta \in B(\rho) : D(\delta, \delta_0) \leq r\}$ and

$$f_\delta(x, y) = \frac{1}{n}\Big\{\mathbb{E}\tilde{L}(Y, \delta(X)) - \tilde{L}(y, \delta(x)) - (\mathbb{E}\tilde{L}(Y, \delta_0(X)) - \tilde{L}(y, \delta_0(x)))\Big\}.$$

For $\delta \in B(\rho)$, we have

$$|\tilde{L}(y, \delta(x))| \leq 1 + |\delta(x)| \leq 1 + \|\delta\|_\infty \leq 1 + \rho.$$

Thus

$$\|f_\delta\|_\infty \leq \frac{4(1+\rho)}{n},$$

and

$$\mathrm{Var}(f_\delta(X, Y)) \leq \frac{1}{n^2}\mathbb{E}[(\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta_0(X)))^2] \leq \frac{1}{n^2}\mathbb{E}[(\delta(X) - \delta_0(X))^2] = \frac{D(\delta, \delta_0)^2}{n^2} \leq \frac{r^2}{n^2}.$$

Talagrand's inequality and $\alpha > 0$ yields that with probability $\geq 1 - e^{-t}$, we have

$$\begin{aligned} Z_r &\leq (1+\alpha) \underbrace{\mathbb{E}Z_r}_{\leq \phi_\rho(r^2)} + \sqrt{2tn} \cdot \frac{r}{n} + 4(\frac{1}{\alpha} + \frac{1}{3})(\rho + 1) \cdot \frac{t}{n} \\ &\overset{\alpha=1}{\leq} 2\phi_\rho(r^2) + \sqrt{\frac{2t}{n}} \cdot r + \frac{16(\rho+1)}{3} \cdot \frac{t}{n}. \end{aligned}$$

(f) r has to satisfy

$$2\phi_\rho(r^2) + \sqrt{\frac{2t}{n}} \cdot r + \frac{16(\rho+1)}{3} \cdot \frac{t}{n} \leq \frac{1}{8c_\rho}(\frac{r}{2})^2,$$

then (e) yields $\mathbb{P}(A^c) \leq e^{-t}$. The condition on r is satisfied if

$$\frac{1}{24c_\rho}(\frac{r}{2})^2 \geq 2\phi_\rho(r^2), \qquad \frac{1}{24c_\rho}(\frac{r}{2})^2 \geq \sqrt{\frac{2t}{n}} \cdot r, \qquad \frac{1}{24c_\rho}(\frac{r}{2})^2 \geq \frac{16(\rho+1)}{3} \cdot \frac{t}{n},$$

or equivalently,

$$\frac{r^2}{192c_\rho} \geq \phi_\rho(r^2), \qquad r \geq 96c_\rho\sqrt{\frac{2t}{n}}, \qquad r \geq 16\sqrt{c_\rho(\rho+1)} \cdot \sqrt{\frac{2t}{n}}.$$

The first inequality is fulfilled if (cf. the hint): $r \geq 4 \cdot 192c_\rho\gamma(n)^{1/2}$.

Additionally, it has to hold that $\tilde{R}(\delta_0) - \tilde{R}(\delta^*) \leq \frac{1}{8c_\rho}(\frac{r}{2})^2$, that is,

$$r \geq 2\big(8c_\rho\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\}\big)^{1/2}.$$

Summarizing the inequalities yields the final statement.

(g) Plugging in r from (f) in the inequality from (d) yields (here c denotes some universal constant which we will not determine in detail).

$$\begin{aligned}
\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) &\leq \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \frac{1}{8c_\rho}(\frac{r}{2})^2 \\
&\leq 2\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \frac{1}{8c_\rho} \cdot \Big((4 \cdot 192)^2 c_\rho^2\gamma(n) + (96)^2 c_\rho^2 \cdot \frac{2t}{n} + 16^2 c_\rho(\rho+1) \cdot \frac{2t}{n}\Big) \\
&\leq 2\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + c \cdot \big\{c_\rho\gamma(n) + (c_\rho + \rho + 1) \cdot \frac{t}{n}\big\}.
\end{aligned}$$

The choice of δ_0 yields

$$\tilde{R}(\delta_0) - \tilde{R}(\delta^*) = \inf_{\delta \in B(\rho)} \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\}.$$

(h) We have to change the approach in (c). For $\varepsilon > 0$, define

$$A := \{Z_r \leq \frac{\varepsilon}{1+\varepsilon}\frac{1}{4c_\rho}(\frac{r}{2})^2\}$$

and assume that $\tilde{R}(\delta_0) - \tilde{R}(\delta^*) \leq \frac{1}{(1+\varepsilon)4c_\rho}(\frac{r}{2})^2$. Then in (d), we obtain $D(\tilde{\delta}, \delta_0) \leq \frac{r}{2}$ and thus $D(\hat{\delta}, \delta_0) \leq r$. In (f) we conclude that

$$r \geq \max\Big\{..., 2\big((1+\varepsilon)4c_\rho\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\}\big)^{1/2}\Big\},$$

thus in (g) we obtain

$$\begin{aligned}\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) &\leq \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \frac{\varepsilon}{(1+\varepsilon)4c_\rho}(\frac{r}{2})^2 \\ &\leq \{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + \varepsilon\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + ... \\ &= (1+\varepsilon)\cdot\{\tilde{R}(\delta_0) - \tilde{R}(\delta^*)\} + ...\end{aligned}$$

Of course, the residual terms in ... are larger now, because they include the larger factor $\frac{1}{\varepsilon}$.

(i) For each $m \in \mathbb{N}$, we have seen in (a)-(g) that

$$\mathbb{P}\Big(\underbrace{\tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*) \geq 2\{\tilde{R}(\delta_m) - \tilde{R}(\delta^*)\} + c\cdot\{c_\rho\gamma(n) + (c_\rho+\rho+1)\cdot\frac{t}{n}}_{=:B_m}\Big) \leq e^{-t}.$$

Due to $\tilde{R}(\delta_m) \downarrow \inf_{\delta\in B(\rho)} \tilde{R}(\delta)$, we have $B_m \uparrow$ and

$$\bigcup_{m\in\mathbb{N}} B_m = B := \Big\{ \inf_{\delta\in B(\rho)} \tilde{R}(\delta) - \tilde{R}(\delta^*) \geq 2\{\tilde{R}(\delta_m) - \tilde{R}(\delta^*)\} + c\cdot\{c_\rho\gamma(n) + (c_\rho+\rho+1)\cdot\frac{t}{n}\Big\}.$$

We conclude that

$$\mathbb{P}(B) = \lim_{m\to\infty} \mathbb{P}(B_m) \leq e^{-t}.$$

Solution 18 (Solution of Task 18)**.** (a) It holds that

$$\sqrt{2tv} \overset{\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}}{\leq} \sqrt{2tn}\sigma + 2\sqrt{tM\mathbb{E}Z} \overset{2\sqrt{ab}\leq\alpha a+\frac{b}{\alpha}}{\leq} \sqrt{2tn}\sigma + \alpha\mathbb{E}Z + \frac{1}{\alpha}tM.$$

Talagrand's inequality $\Rightarrow$ With probability $\geq 1-e^{-t}$, we have

$$Z \leq \mathbb{E}Z + \sqrt{2tv} + \frac{tM}{3} \leq (1+\alpha)\mathbb{E}Z + \sqrt{2tn}\sigma + (\frac{1}{3}+\frac{1}{\alpha})Mt.$$

(b) Compared to Bernstein's inequality, only $\mathbb{E}Z$ is needed additionally. The other terms $\sqrt{2tn}\sigma$ and tM appear in both inequalities.
For fixed f, the term $\sum_{i=1}^n f(X_i)$ varies around 0; this variation is explained only through $\sqrt{2tn}\sigma$ and tM.
$Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i)$ varies around $\mathbb{E}Z$ (of course, in general it holds that $\mathbb{E}Z > 0$); but apart from that, the variation behaves like $\sum_{i=1}^n f(X_i)$ for fixed f. The complexity of the function class $\mathcal{F}$ therefore enters the inequality through $\mathbb{E}Z$ (not through σ^2, M) and does not produce an additional *variation*, but a *shift* of Z away from zero. Therefore, the main ingredient for quantifying the variation of Z is to find a good upper bound for $\mathbb{E}Z$.

(c) $\mathcal{F}_{sep} \subset \mathcal{F} \Rightarrow \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i) \geq \sup_{g\in\mathcal{F}_{sep}} \sum_{i=1}^n g(X_i)$.
We now show '$\leq$': Let $\varepsilon > 0$. Let $f \in \mathcal{F}$ be arbitrary. Let $g \in \mathcal{F}_{sep}$ with $\|f - g\| \leq \varepsilon$. Then it holds that

$$\sum_{i=1}^n f(X_i) \leq \sum_{i=1}^n g(X_i) + \sum_{i=1}^n \underbrace{|f(X_i) - g(X_i)|}_{\leq C\|f-g\| \leq C\varepsilon} \leq \sum_{i=1}^n g(X_i) + nC\varepsilon.$$

$\Rightarrow$

$$\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i) \leq \sup_{g\in\mathcal{F}_{sep}} \sum_{i=1}^n g(X_i) + nC\varepsilon.$$

With $\varepsilon \to 0$, the assertion follows (note that n, C are fixed during this proof!)

(d) We show that for $\delta, \delta' \in B(\rho)$, it holds that $|f_\delta(x,y) - f_{\delta'}(x,y)| \leq \|\delta - \delta'\|_K$.
Note that $\|\delta\|_\infty \leq \|\delta\|_K \leq \rho$ (cf. Lemma 4.24(ii)). Then, we have $f_\delta = \frac{A(\delta)}{B(\delta)}$ with

$$A(\delta) := (\mathbb{E}\tilde{L}(Y,\delta(X)) - \mathbb{E}\tilde{L}(Y,\delta_0(X))) - (\tilde{L}(y,\delta(x)) - \tilde{L}(y,\delta_0(x))), \qquad B(\delta) := n(D(\delta,\delta_0)^2 + r^2).$$

Thus

$$|f_\delta(x,y) - f_{\delta'}(x,y)| \leq \frac{1}{B(\delta)} \cdot \big|A(\delta) - A(\delta')\big| + \frac{|A(\delta')|}{B(\delta)\cdot B(\delta')} \cdot \big|B(\delta) - B(\delta')\big|.$$

We have $\frac{1}{B(\delta)} \leq \frac{1}{nr^2}$, $|A(\delta)| \leq 4(\rho+1)$. Moreover,

$$|A(\delta) - A(\delta')| \leq \big|\mathbb{E}\tilde{L}(Y,\delta(X)) - \mathbb{E}\tilde{L}(Y,\delta'(X))\big| + \big|\tilde{L}(y,\delta(x)) - \tilde{L}(y,\delta'(x))\big| \leq 2\|\delta - \delta'\|_\infty.$$

Since $|D(\delta,\delta_0) - D(\delta',\delta_0)| \leq \|\delta - \delta'\|_\infty$, $D(\delta,\delta_0) \leq 2\rho$, we have

$$|B(\delta) - B(\delta')| \leq \big|D(\delta',\delta_0)^2 - D(\delta,\delta_0)^2\big| \leq \|\delta - \delta'\|_\infty \cdot 4\rho.$$

We conclude that

$$|f_\delta(x,y) - f_{\delta'}(x,y)| \leq \underbrace{\Big(\frac{2}{nr^2} + \frac{4(\rho+1)\cdot 4\rho}{(nr^2)^2}\Big)}_{=:C} \cdot \underbrace{\|\delta - \delta'\|_\infty}_{\leq \|\delta-\delta'\|_K}.$$

Solution 19 (Solution of Task 19)**.** (a) It holds that

$$\begin{aligned}
\mathbb{E}Z &= \int_0^\infty \mathbb{P}(Z \geq x)dx = \int_0^A \underbrace{\mathbb{P}(Z \geq x)}_{\leq 1} dx + \int_A^\infty \mathbb{P}(Z \geq x)dx \\
&\overset{\text{Subst. } x = y + A}{\leq} A + \int_0^\infty \mathbb{P}(Z \geq A + y)dy \\
&\overset{\text{Subst. } y = Bt}{=} A + B \cdot \int_0^\infty \underbrace{\mathbb{P}(Z \geq A + Bt)}_{\leq g(t)} dt \leq A + B \cdot \int_0^\infty g(t)dt.
\end{aligned}$$

(b) Choose $g(t) = e^{-t}$ (thus $\int_0^\infty g(t)dt = 1$),

$$\begin{aligned} Z &= \tilde{R}(\hat{\delta}) - \tilde{R}(\delta^*), \\ A &= 2 \inf_{\delta \in B(\rho)} \{\tilde{R}(\delta) - \tilde{R}(\delta^*)\} + c \cdot c_\rho \gamma(n), \\ B &= \frac{c}{n} \cdot (c_\rho + \rho + 1) \end{aligned}$$

The assertion now follows from (a) and Task 17(g).
(The proof technique is more complicated if there is an additional penalization term, since this penalization term also depends on t).

8.4 Solutions of Chapter 5

Solution 20 (Solution of Task 20)**.** (a) Lemma 5.2 implies that

$$\mathbb{E}R(\hat{f}_{n,h}) - R(f^*) = \mathbb{E}[(\hat{f}_{n,h}(X) - f^*(X))^2] = \mathbb{E}[\mathbb{E}[(\hat{f}_{n,h}(X) - f^*(X))|X = x]] \leq \mathbb{E}[\gamma(n,h)] = \gamma(n,h)$$

(b) It holds that

$$|\hat{f}_{n,h}(x)| \leq \|f^*\|_\infty + \sup_{i=1,\ldots,n} |\varepsilon_i| \quad \Rightarrow \quad |\hat{f}_{n,h}(x) - f^*(x)|^2 \leq (2\|f^*\|_\infty + C_\varepsilon)^2.$$

$\Rightarrow$

$$\mathbb{E}[(\hat{f}_{n,h}(x) - f^*(x))^2 \mathbb{1}_{A(x)^c}] \leq (2C_{f^*} + C_\varepsilon)^2 \mathbb{P}(A(x)^c).$$

(c) We have

$$\hat{g}(x) - \mathbb{E}\hat{g}(x) = \sum_{i=1}^n Z_i$$

with $Z_i = \frac{W(\frac{X_i - x}{h}) - \mathbb{E}W(\frac{X_i - x}{h})}{nh^d}$ i.i.d., $|Z_i| \leq \frac{\|W\|_\infty}{nh^d} =: M$ and

$$\begin{aligned} \mathrm{Var}(Z_1) &\leq \frac{1}{n^2 h^{2d}} \mathbb{E}[W(\frac{X_1 - x}{h})^2] = \frac{1}{n^2 h^{2d}} \int W(\frac{x_1 - x}{h}) g(x_1) dx_1 \\ &\overset{u = \frac{x_1 - x}{h}}{=} \frac{1}{n^2 h^d} \int W(u) g(x + uh) du \\ &\leq \frac{1}{n^2 h^d} \|g\|_\infty \int W(u)^2 du \leq \frac{C_W C_g}{n^2 h^d} =: V^2. \end{aligned}$$

We therefore have

$$\sqrt{2tn}V + \frac{tM}{3} = \sqrt{\frac{2tC_g C_W}{nh^d}} + \frac{tC_W}{3nh^d}.$$

Plugging in these results into Bernstein's inequality yields the assertion.

(d) Due to $g \in \mathcal{F}(L)$, we have

$$|\mathbb{E}\hat{g}(x) - g(x)| \overset{u=\frac{x_1-x}{h}}{\leq} \int W(u)|g(x+uh)-g(x)|du \leq L\int W(u)\|u\|_\infty du \cdot h \leq LC_W \cdot h.$$

(e) It holds that

$$\mathbb{P}(A(x)^c) \leq \mathbb{P}(|\hat{g}(x) - \mathbb{E}\hat{g}(x)| > \frac{c_g}{4}) + \mathbb{P}(|\mathbb{E}\hat{g}(x) - g(x)| > \frac{c_g}{4}). \qquad (*)$$

(d) $\Rightarrow$ If $LC_W \cdot h \leq \frac{c_g}{4} \iff h \leq \frac{c_g}{4LC_W}$, the last summand in (*) is 0.
(c) $\Rightarrow$ If

$$\sqrt{\frac{2tC_gC_W}{nh^d}} + \frac{tC_W}{3nh^d} \leq \frac{c_g}{4}, \qquad (**)$$

with $t = \log(n)$, the first summand in (*) is $\leq 2n^{-1}$.
The inequality in (**) is fulfilled if

$$(\frac{c_g}{8})^2 \geq \frac{2tC_gC_W}{nh^d}, \qquad \frac{c_g}{8} \geq \frac{tC_W}{3nh^d}.$$

Solving for h yields the condition on h presented in the task.

(f) The results from (b), (f) imply

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2 \mathbb{1}_{A(x)^c}] \leq (2\|f^*\|_\infty + C_\varepsilon)^2 \mathbb{P}(A(x)^c) \leq 2(2\|f^*\|_\infty + C_\varepsilon)^2 n^{-1}.$$

(g) On $A(x)$, we have $g(x) \geq c_g$. Thus $\hat{g}(x) \geq \frac{c_g}{2}$. With this, we conclude that

$$\hat{f}_{n,h} - f^* = \frac{\hat{m} - \hat{g} \cdot f^*}{\hat{g}} \quad \Rightarrow \quad \left|\hat{f}_{n,h} - f^*\right|^2 \mathbb{1}_{A(x)} \leq \frac{4}{c_g^2}\left|\hat{m} - \hat{g} \cdot f^*\right|^2.$$

Taking expectations yields the claim.

(h) It holds that

$$\begin{aligned}\hat{m}(x) - f^*(x)\hat{g}(x) &= \frac{1}{n}\sum_{i=1}^n W_h(X_i - x) \cdot \{Y_i - f^*(x)\} \\ &= \frac{1}{n}\sum_{i=1}^n W_h(X_i - x) \cdot \varepsilon_i + \frac{1}{n}\sum_{i=1}^n W_h(X_i - x) \cdot \{f^*(X_i) - f^*(x)\}.\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}&\mathbb{E}|\hat{m}(x) - f^*(x)\hat{g}(x)|^2] \\ \leq\ & \underbrace{2\,\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^n W_h(X_i - x) \cdot \varepsilon_i\Big)^2\Big]}_{=\frac{1}{n}\operatorname{Var}(W_h(X_1-x)\cdot\varepsilon_1)} + 2\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^n W_h(X_i - x) \cdot \{f^*(X_i) - f^*(x)\}\Big)^2\Big]\end{aligned}$$

For the last term, we use $\mathbb{E}[Z^2] = \mathbb{E}[Z]^2 + \mathrm{Var}(Z)$. This yields

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n} W_h(X_i - x)\cdot\{f^*(X_i) - f^*(x)\}\Big)^2\Big] \\
= \;& \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} W_h(X_i - x)\cdot\{f^*(X_i) - f^*(x)\}\Big]^2 + \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} W_h(X_i - x)\cdot\{f^*(X_i) - f^*(x)\}\Big) \\
= \;& \mathbb{E}[W_h(X_1 - x)\{f^*(X_1) - f^*(x)\}] + \frac{1}{n}\mathrm{Var}(W_h(X_1 - x)\cdot\{f^*(X_1) - f^*(x)\}).
\end{aligned}$$

(i) It holds that

$$\begin{aligned}
\mathbb{E}\big[\big(\frac{1}{n}\sum_{i=1}^{n} W_h(X_i - x)\varepsilon_i\big)^2\big] &= \frac{1}{n}\mathrm{Var}(W_h(X_1 - x)\varepsilon_1) \le \frac{1}{n}\mathbb{E}[W_h(X_1 - x)^2]\cdot\mathbb{E}[\varepsilon_1^2] \\
&= \frac{\sigma^2}{nh^{2d}}\int W(\frac{x_1 - x}{h})^2 g(x_1)dx_1 \\
&= \frac{\sigma^2}{nh^d}\int W(u)^2 g(x + uh)du \le \frac{\sigma^2\|g\|_\infty}{nh^d}\int W(u)^2 du \le \frac{\sigma^2 C_g C_W}{nh^d}
\end{aligned}$$

(j) It holds that

$$\begin{aligned}
\mathbb{E}[W_h(X_1 - x)\{f^*(X_1) - f^*(x)\}] &= \frac{1}{h^d}\int W(\frac{x_1 - x}{h})\{f^*(x_1) - f^*(x)\}g(x_1)dx_1 \\
&\overset{u=\frac{x_1-x}{h}}{=} \int W(u)\{f^*(x + uh) - f^*(x)\}g(x + uh)du.
\end{aligned}$$

Together with $f^* \in \mathcal{F}(L)$, we obtain

$$\big|\mathbb{E}[W_h(X_1 - x)\{f^*(X_1) - f^*(x)\}]\big| \le L\|g\|_\infty \int W(u)\|u\|_\infty du \cdot h \le LC_g C_W h.$$

Squaring both sides yields the claim.

(k) With the aforementioned results, we obtain

$$\begin{aligned}
\mathbb{E}[\big|\hat{f}_{n,h} - f^*\big|^2 \mathbb{1}_{A(x)}] &\le \frac{4}{c_g^2}\mathbb{E}[\big|\hat{m} - \hat{g}\cdot f^*\big|^2] \le \frac{8}{c_g^2}\Big\{\frac{\sigma^2 C_g C_W}{nh^d} + (LC_g C_W)^2 h^2 + L^2 C_g C_W \frac{h^2}{nh^d}\Big\} \\
&\le \frac{8C_g^2 C_W^2}{c_g^2}\Big\{\frac{\sigma^2}{nh^d} + L^2 h^2 + L^2\frac{h^2}{nh^d}\Big\}.
\end{aligned}$$

(l) With (f) and (k), we obtain

$$\mathbb{E}[(\hat{f}(x) - f^*(x))^2] \le \frac{8C_g^2 C_W^2}{c_g^2}\Big\{\frac{\sigma^2}{nh^d} + L^2 h^2 + L^2\frac{h^2}{nh^d}\Big\} + 2(2\|f^*\|_\infty + C_\varepsilon)n^{-1}.$$

8.5 Solutions of Chapter 6

Solution 21 (Solution of Task 21)**.** (a) Under the given assumptions, Bernstein's inequality yields

$$\mathbb{P}\Big(\sum_{i=1}^{n} Z_i \geq t\Big) \leq \exp\Big(-\frac{t^2}{2n\sigma^2 + 2Mt/3}\Big)$$

Application to $-\sum_{i=1}^{n} Z_i$ yields the same inequality, thus

$$\mathbb{P}\Big(W(g) \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{2n\sigma^2 + 2Mt/3}\Big).$$

If $t < \frac{3\sigma^2 n}{M}$, then we have $2n\sigma^2 \geq 2Mt/3$, thus $2n\sigma^2 + 2Mt/3 \leq 4n\sigma^2$. We conclude that

$$\mathbb{P}\big(W(g) \geq t\big) \leq 2\exp\Big(-\frac{t^2}{4n\sigma^2}\Big).$$

If $t > \frac{3\sigma^2 n}{M}$, then we have $2n\sigma^2 + 2Mt/3 \leq 4Mt/3$, thus

$$\mathbb{P}\big(W(g) \geq t\big) \leq 2\exp\Big(-\frac{t^2}{4Mt/3}\Big) = 2\exp\Big(-\frac{3t}{4M}\Big).$$

(b) It holds that

$$\mathbb{E}\psi_1\big(\frac{A_1(g)}{4M}\big) = \mathbb{E}\int_0^{A_1(g)/(4M)} e^t dt = \int_0^{\infty} \mathbb{P}(A_1(g) > t4M) e^t dt \overset{(a)}{\leq} \int_0^{\infty} 2e^{-3t} e^t dt = 1.$$

(c) It holds that

$$\psi_1\big(\mathbb{E}\sup_{g\in\mathcal{G}} \frac{A_1(g)}{4M}\big) \leq \mathbb{E}\psi_1\big(\sup_{g\in\mathcal{G}} \frac{A_1(g)}{4M}\big) \leq \sum_{g\in\mathcal{G}} \mathbb{E}\psi_1\big(\frac{A_1(g)}{4M}\big) \leq |\mathcal{G}|.$$

(d) The inverse function $\psi_1^{-1}(x) = \log(x+1)$ is increasing. Application on the result from (c) yields the result.

(e) It holds that

$$\mathbb{E}\psi_2\big(\frac{A_2(g)}{2\sigma\sqrt{3n}}\big) = \mathbb{E}\int_0^{(A_2(g)/(2\sigma\sqrt{3n}))^2} e^t dt = \int_0^{\infty} \mathbb{P}(A_2(g) > t\cdot 2\sigma\sqrt{3n}) e^t dt \leq \int_0^{\infty} 2e^{-3t} e^t dt = 1.$$

(f) As in (c),(d), we obtain that $\psi_2(x) = e^{x^2} - 1$ is convex. The inverse function is $\psi_2^{-1}(x) = \sqrt{\log(x+1)}$.

(g) It holds that

$$\begin{aligned}\mathbb{E}\sup_{g\in\mathcal{G}} W(g) &\leq \mathbb{E}\sup_{g\in\mathcal{G}} A_1(g) + \mathbb{E}\sup_{g\in\mathcal{G}} A_2(g)\\ &\leq 4M\cdot\mathbb{E}\sup_{g\in\mathcal{G}}\frac{A_1(g)}{4M} + 2\sqrt{3}\sigma\sqrt{n}\cdot\mathbb{E}\sup_{g\in\mathcal{G}}\frac{A_2(g)}{2\sqrt{3}\sigma\sqrt{n}}\\ &\leq 2\sqrt{3}\sigma\sqrt{n}\sqrt{\log(|\mathcal{G}|+1)} + 4M\cdot\log(|\mathcal{G}|+1).\end{aligned}$$

Solution 22 (Solution of Task 22)**.** (a) We have

$$\begin{aligned}R(f_{T_0}) - R(f^*) &= \mathbb{E}[\mathbb{1}_{\{Y\neq f_{T_0}(X)\}} - \mathbb{1}_{\{Y\neq f^*(X)\}}] \leq \mathbb{E}[\mathbb{1}_{\{f^*(X)\neq f_{T_0}(X)\}}]\\ &= \mathbb{P}(f_{T_0}(X)\neq f^*(X)).\end{aligned}$$

(b) It holds that

$$\begin{aligned}&\mathbb{P}(f_{T_0}(X)\neq f^*(X))\\ \leq\ &\mathbb{P}\Big(\bigcup_{A\in\mathcal{C}}\{X\in A\}\Big)\\ \leq\ &\sum_{A\in\mathcal{C}}\mathbb{P}(X\in A) \leq c_\mu\sum_{A\in\mathcal{C}}\mu(A)\\ =\ &c_\mu|\mathcal{C}|m^{-d}\\ =\ &c_\mu c_{box}m^{d-1}\cdot m^{-d} = c_\mu c_{box}m^{-1}.\end{aligned}$$

(c) Since each coordinate is splitted dS-times, T_0 satisfies $|T_0| = m^d$. Thus T_0 generates a partition with $2^{dS} = m^d$ cubes.

(d) Since $T_0\in\mathcal{T}_S$, we have

$$\inf_{T\in\mathcal{T}_S}\big\{R(f_T) - R(f^*) + \lambda\cdot|T|\big\} \leq R(f_{T_0}) - R(f^*) + \lambda\cdot|T_0| \leq c_\mu c_{box}m^{-1} + \lambda m^d.$$

(e) If $S\geq\frac{\log_2(n)}{d+1}$, then we can choose $m = 2^{\lfloor\log_2(n)/(d+1)\rfloor}$. Then we have $\frac{1}{2}n^{1/(d+1)} \leq m \leq n^{1/(d+1)}$. We conclude that

$$c_\mu c_{box}m^{-1} + \lambda m^{-d} \leq 2c_{mu}c_{box}n^{-\frac{1}{d+1}} + c\frac{\log(2d)+t}{\eta_0 n}\cdot n^{\frac{d}{d+1}} = (2c_\mu c_{box} + \frac{c}{\eta_0}(\log(2d)+t))\cdot n^{-\frac{1}{d+1}}.$$

Solution 23 (Solution of Task 23)**.** (a) By using the theorem stated in the task, we have

$$\begin{aligned}N(\varepsilon,\mathcal{C},\|\cdot\|_{2,n,X}) &\leq 13\cdot\mathcal{V}(\mathcal{C})\cdot(\frac{4e}{\varepsilon^2})^{\mathcal{V}(\mathcal{C})}\\ &= \Big(\frac{4e(13\mathcal{V}(\mathcal{C}))^{\frac{1}{2\mathcal{V}(\mathcal{C})}}}{\varepsilon}\Big)^{2\mathcal{V}(\mathcal{C})}\\ &\overset{\forall x>0: x^{1/2x}\leq 2}{\leq} \Big(\frac{13\cdot 2\cdot 4e}{\varepsilon}\Big)^{2\mathcal{V}(\mathcal{C})}\\ &\overset{2\mathcal{V}(\mathcal{C})\leq V}{\leq} \Big(\frac{13\cdot 2\cdot 4e}{\varepsilon}\Big)^{V}.\end{aligned}$$

(b) To calculate $m_{\mathcal{C}}(N)$ (at least a lower bound), we have to find out how many different labelings of N points in the space can be generated with decision stumps. Here, we can choose the location of the points as desired (due tot he max in $m_{\mathcal{C}}(N)$).
Let $x_1, ..., x_N \in [0,1]$ be arbitrary but with pairwise distinct locations. Then we have $m_{\mathcal{C}}(x_1, ..., x_N) = 2N$. Proof: Choose

$$f_{1j}(x) = \mathbb{1}_{\{x<x_j\}} - \mathbb{1}_{\{x\geq x_j\}}, \quad f_{2j}(x) = -\mathbb{1}_{\{x<x_j\}} + \mathbb{1}_{\{x\geq x_j\}}.$$

Then all of the $2N$ decision rules f_{1j}, f_{2j} above generate different labelings of $x_1, ..., x_N$. Graphically this can be understood as follows: If the points are located on a real line, we put a vertical line exactly at each point x_j; on the left of this vertical line, f_{1j} labels class all points with class 1, on the right hand side f_{1j} labels everything with class -1; vice versa for f_{2j}).
All other decision stumps do not provide different labelings.
If instead there exist indices $i, j \in \{1, ..., N\}$ with $x_i = x_j$, then we have $m_{\mathcal{C}}(x_1, ..., x_N) \leq 2N$.
$\Rightarrow m_{\mathcal{C}}(N) = 2N$.

(c) In d dimensions, the problem is more involved. For pairwise disjoint $x^{(1)}, ..., x^{(N)} \in [0,1]^d$ we choose

$$f_{1jk}(x) = \mathbb{1}_{\{x<x_j^{(k)}\}} - \mathbb{1}_{\{x\geq x_j^{(k)}\}}, \quad f_{2jk}(x) = -\mathbb{1}_{\{x<x_j^{(k)}\}} + \mathbb{1}_{\{x\geq x_j^{(k)}\}}, \quad j = 1, ..., d, k = 1, ..., N.$$

This yields $2Nd$ different labelings. All other decision stumps do not generate more labelings (one may use a picture to see this).
If instead there exist indices $i, j \in \{1, ..., N\}$ with $x^{(i)} = x^{(j)}$, then there exist only less labelings.
Note that there can exist at most 2^N labelings.
$\Rightarrow m_{\mathcal{C}}(x_1, ..., x_N) \leq \min\{2Nd, 2^N\}$.
The exact proof to show $m_{\mathcal{C}}(x_1, ..., x_N) = \min\{2Nd, 2^N\}$ is mathematically more involved and is omitted here.

(d) (c) implies

$$\begin{aligned} & m_{\mathcal{C}}(x_1, ..., x_N) \leq \min\{2Nd, 2^N\} \\ \Rightarrow \quad & \mathcal{V}(\mathcal{C}) \leq \inf\{N \in \mathbb{N} : \min\{2Nd, 2^N\} < 2^N\} = \inf\{N \in \mathbb{N} : 2Nd < 2^N\}. \end{aligned}$$

With $N = 2\log_2(2d)$, we obtain

$$2^N = (2d)^2 = 4d^2, \qquad 2Nd = 4d\log_2(2d).$$

Since $d > \log_2(2d)$ for $d \geq 2$, we obtain

$$2^N > 2Nd.$$

$\Rightarrow \mathcal{V}(\mathcal{C}) \leq N \overset{\mathcal{V}(\mathcal{C})\in\mathbb{N}}{\Rightarrow} \mathcal{V}(\mathcal{C}) \leq \lfloor 2\log_2(2d) \rfloor$.

Solution 24 (Solution of Task 24)**.** (a) It holds that $\mathbb{E}\tilde{L}(Y,\delta(X)) = \mathbb{E}[\mathbb{E}[\phi(-Y\delta(X))|X = x]]$ and

$$\mathbb{E}[\phi(-Y\delta(X))|X = x]] = \phi(-\delta(x))\eta(x) + \phi(\delta(x))(1-\eta(x)).$$

Pointwise minimization yields $\delta^*(x) \in \arg\min_{z\in\mathbb{R}} \Phi_{\eta(x)}(z)$ with $\Phi_\eta(z) := \phi(-z)\eta + \phi(z)(1-\eta)$.
ϕ is differentiable $\Rightarrow$ $z \mapsto \Phi_{\eta(x)}(z)$ is differentiable. Since δ^* is a minimizer, we have

$$0 = \Phi'_{\eta(x)}(\delta^*(x)).$$

We obtain that $\delta^*(x) = (\Phi'_{\eta(x)})^{-1}(0)$. Put $g(\eta) = (\Phi'_\eta)^{-1}(0)$. Moreover, it holds that

$$0 = \Phi'_\eta(g(\eta)) = -\phi'(-g(\eta))\eta + \phi'(\eta)(1-\eta) \qquad (*).$$

This yields the claim.

(b) By conditioning on X, we obtain

$$\begin{array}{ll} & \mathbb{E}[(\tilde{L}(Y,\delta(X)) - \tilde{L}(Y,\delta^*(X)))^2|X=x] \\ \overset{\delta^*(x)=g(\eta(x))}{=} & \eta(x)\big[\phi(-\delta(x)) - \phi(-g(\eta(x)))\big]^2 + (1-\eta(x))\big[\phi(\delta(x)) - \phi(g(\eta(x)))\big]^2 \\ =: & A_2(\eta(x),\delta(x)) \end{array}$$

with

$$A_2(\eta,\delta) := \eta\cdot\big[\phi(-\delta) - \phi(-g(\eta))\big]^2 + (1-\eta)\big[\phi(\delta) - \phi(g(\eta))\big]^2$$

Similarly,

$$\begin{array}{ll} & \mathbb{E}[\tilde{L}(Y,\delta(X)) - \tilde{L}(Y,\delta^*(X))|X=x] \\ = & \eta(x)\big[\phi(-\delta(x)) - \phi(-g(\eta((x)))\big] + (1-\eta(x))\big[\phi(\delta(x)) - \phi(g(\eta(x)))\big] =: A_1(\eta(x),\delta(x)) \end{array}$$

with

$$A_1(\eta,\delta) := \eta\cdot\big[\phi(-\delta) - \phi(-g(\eta))\big] + (1-\eta)\big[\phi(\delta) - \phi(g(\eta))\big].$$

(c) It holds that

$$\begin{array}{rcl} \partial_\eta A_1(\eta,\delta) & = & \big[\phi(-\delta) - \phi(-g(\eta))\big] - \big[\phi(\delta) - \phi(g(\eta))\big] \\ & & \quad + \big[\eta\phi'(-g(\eta)) - (1-\eta)\phi'(g(\eta))\big]\cdot g'(\eta) \\ & \overset{(*)}{=} & \big[\phi(-\delta) - \phi(-g(\eta))\big] - \big[\phi(\delta) - \phi(g(\eta))\big]. \end{array} \tag{94}$$

(d) We have

$$\begin{aligned}
&\partial_\eta A_2(\eta,\delta)\\
=\;& \big[\phi(-\delta)-\phi(-g(\eta))\big]^2-\big[\phi(\delta)-\phi(g(\eta))\big]^2\\
&\quad +2\big\{\eta\phi'(-g(\eta))\big[\phi(-\delta)-\phi(-g(\eta))\big]-(1-\eta)\phi'(g(\eta))\big[\phi(\delta)-\phi(g(\eta))\big]\big\}g'(\eta)\\
=\;& \underbrace{\big\{\phi(-\delta)-\phi(-g(\eta))-\phi(\delta)+\phi(g(\eta))\big\}}_{=\partial_\eta A_1(\eta,\delta)}\\
&\quad \times\big\{\phi(\delta)-\phi(g(\eta))+\phi(-\delta)-\phi(-g(\eta))+\big\{\eta\phi'(-g(\eta))+(1-\eta)\phi'(g(\eta))\big\}g'(\eta)\big\}.\\
=\;& \partial_\eta A_1(\eta,\delta)\cdot\{\phi(\delta)+\phi(-\delta)+B(\eta)\} \qquad (95)
\end{aligned}$$

For the third summand above, we have used $a^2-b^2=(a-b)(a+b)$ for real numbers $a,b\in\mathbb{R}$. For the second summand above, we used

$$\begin{aligned}
&2\eta\phi'(-g(\eta))\big[\phi(-\delta)-\phi(-g(\eta))\big]-2(1-\eta)\phi'(g(\eta))\big[\phi(\delta)-\phi(g(\eta))\big]\\
=\;& \underbrace{\eta\phi'(-g(\eta))}_{\overset{(*)}{=}(1-\eta)\phi'(g(\eta))}\big[\phi(-\delta)-\phi(-g(\eta))\big]-\underbrace{(1-\eta)\phi'(g(\eta))}_{\overset{(*)}{=}\eta\phi'(-g(\eta))}\big[\phi(\delta)-\phi(g(\eta))\big]\\
&\quad -(1-\eta)\phi'(g(\eta))\big[\phi(-\delta)-\phi(-g(\eta))\big]+\eta\phi'(-g(\eta))\big[\phi(\delta)-\phi(g(\eta))\big]\\
=\;& \big\{\eta\phi'(-g(\eta))+(1-\eta)\phi'(g(\eta))\big\}\cdot\big\{\phi(-\delta)-\phi(-g(\eta))-\phi(\delta)+\phi(g(\eta))\big\}.
\end{aligned}$$

(e) It holds that

$$\partial_\eta A_1(\eta,\delta)\begin{cases}\geq 0, & g(\eta)\geq\delta\\ \leq 0, & g(\eta)\leq\delta.\end{cases}$$

If $g(\eta)\geq\delta$, we conclude with (d) that

$$\begin{aligned}
\partial_\eta A_2(\eta,\delta) &\leq \{\underbrace{\phi(\delta)+\phi(-\delta)}_{\leq\phi(\rho)+\phi(-\rho)}+\underbrace{B(\eta)}_{\leq C_\phi}\}\cdot\underbrace{\partial_\eta A_1(\eta,\delta)}_{\geq 0}\\
&\leq \{\phi(\rho)+\phi(-\rho)+C_\phi\}\cdot\partial_\eta A_1(\eta,\delta).
\end{aligned}$$

If $g(\eta)\leq\delta$, we obtain the reverse inequality due to $\partial_\eta A_1(\eta,\delta)\leq 0$.
By convexity of ϕ (and thus ϕ' is increasing), we obtain

$$\frac{\phi(-\delta)-\phi(-\rho)}{\rho-\delta}\leq\phi'(-\delta)\leq\phi'(\delta)\leq\frac{\phi(\rho)-\phi(\delta)}{\rho-\delta}$$

which implies $\phi(\delta)+\phi(-\delta)\leq\phi(\rho)+\phi(-\rho)$.

(f) This follows directly from

$$\phi(-\delta)-\phi(-g(g^{-1}(\delta)))=0 \qquad\text{and}\qquad \phi(\delta)-\phi(g(g^{-1}(\delta)))=0.$$

(g) Integration yields for $g(\eta) \geq \delta$ that

$$\begin{aligned} A_2(\eta,\delta) &= \int_{g^{-1}(\delta)}^{\eta} \partial_\eta A_2(s,\delta)ds \overset{(e)}{\leq} \{\phi(\rho)+\phi(-\rho)+C_\phi\} \cdot \int_{g^{-1}(\delta)}^{\eta} \partial_\eta A_1(s,\delta)ds \\ &\leq \{\phi(\rho)+\phi(-\rho)+C_\phi\} A_1(\eta,\delta). \end{aligned}$$

For $g(\eta) \leq \delta$, we use $A_2(\eta,\delta) = \int_{g^{-1}(\delta)}^{\eta} \partial_\eta A_2(s,\delta)ds = -\int_{\eta}^{g^{-1}(\delta)} \partial_\eta A_2(s,\delta)ds$.

(h) Derivation of C_ϕ:

- Case $\phi(x) = e^x$: $g(\eta) = \frac{1}{2}\log(\frac{\eta}{1-\eta}) \Rightarrow g'(\eta) = \frac{1}{2}(\frac{1}{\eta} + \frac{1}{1-\eta}) = \frac{1}{2\eta(1-\eta)}$. Moreover,

$$\phi'(g(\eta)) = \eta^{1/2}(1-\eta)^{-1/2}, \qquad \phi'(-g(\eta)) = \eta^{-1/2}(1-\eta)^{1/2}.$$

Plugging in this result yields

$$\begin{aligned} C_\phi &= 0 \vee \max_\eta \Big\{ \frac{2(\eta(1-\eta))^{1/2}}{2\eta(1-\eta)} - \eta^{1/2}(1-\eta)^{-1/2} - \eta^{-1/2}(1-\eta)^{1/2} \Big\} \\ &= 0 \vee \max_\eta \big\{ \eta^{-1/2}(1-\eta)^{-1/2} - \eta^{1/2}(1-\eta)^{-1/2} - \eta^{-1/2}(1-\eta)^{1/2} \big\} = 0. \end{aligned}$$

- Case $\phi(x) = \log(1+e^x)$: $g(\eta) = \log(\frac{\eta}{1-\eta}) \Rightarrow g'(\eta) = \frac{1}{\eta(1-\eta)}$. Moreover,

$$\phi'(x) = \frac{e^x}{1+e^x}, \qquad \phi'(g(\eta)) = \eta, \qquad \phi'(-g(\eta)) = 1-\eta.$$

Plugging in this result yields

$$C_\phi = 0 \vee \max_\eta \Big\{ \frac{2\eta(1-\eta)}{\eta(1-\eta)} + \log(\eta(1-\eta)) \Big\} = \max_\eta \big\{ 2 + \log(\underbrace{\eta(1-\eta)}_{\leq 4^{-1}}) \big\} = 2 - 2\log(2).$$

Solution 25 (Solution of Task 25)**.** In the steps (a)-(d) we have to show that an $\tilde{\varepsilon}$-Covering of the right hand side yields an ε-Covering of the left hand side. In each of these steps we construct a covering of the left hand side from a covering of the right hand side.

(a) Let $(\tilde{f}_j)$ be a ε-Covering of $\mathcal{F}_1$. Then $f_j(x,y) := \tilde{f}_j(x,y) - \tilde{L}(y,\delta_0(x))$ is a ε-Covering of $\mathcal{F}$. Proof:
$f \in \mathcal{F}$ can be written as $f = \tilde{f} - \tilde{L}(y,\delta_0(x))$ with $\tilde{f} \in \mathcal{F}_1$. Let j be such that $\|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \varepsilon$. Then it holds that

$$\|f - f_j\|_{2,n} = \|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \varepsilon.$$

$\Rightarrow N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,n}) \leq N(\varepsilon, \mathcal{F}_1, \|\cdot\|_{2,n})$.

(b) Let $(\tilde{f}_j)$ be a ε-Covering of $\mathcal{F}_2$. Then $f_j := \phi \circ \tilde{f}_j$ is a $\varepsilon \cdot \phi'(\rho)$-Covering of $\mathcal{F}_1$. Proof:
For $f \in \mathcal{F}_1$ exists some $\tilde{f} \in \mathcal{F}_2$ with $f = \phi \circ \tilde{f}$. Let j be such that $\|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \varepsilon$. Then it holds that

$$\begin{aligned}\|f - f_j\|_{2,n} &= \big(\frac{1}{n}\sum_{i=1}^{n}\{\phi(\tilde{f}(X_i,Y_i)) - \phi(\tilde{f}_j(X_i,Y_i))\}^2\big)^{1/2} \\ &\leq \phi'(\rho)\cdot\big(\frac{1}{n}\sum_{i=1}^{n}\{\tilde{f}(X_i,Y_i) - \tilde{f}_j(X_i,Y_i)\}^2\big)^{1/2} = \phi'(\rho)\|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \phi'(\rho)\varepsilon,\end{aligned}$$

because ϕ is Lipschitz continuous on $[-\rho,\rho]$ with constant $\phi'(\rho)$.
Rearranging terms yields $N(\varepsilon, \mathcal{F}_1, \|\cdot\|_{2,n}) \leq N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_2, \|\cdot\|_{2,n})$.

(c) Let (δ_j) be a ε-Covering of $\mathcal{F}_3$ with respect to $\|\cdot\|_{2,n,X}$. Then $f_j(x,y) = y\delta_j(x)$ is a ε-Covering of $\mathcal{F}_2$ with respect to $\|\cdot\|_{2,n}$. Proof:
For $f \in \mathcal{F}_2$ exists some $\delta \in \mathcal{F}_3$ with $f(x,y) = y\delta(x)$. Let j be such that $\|\delta - \delta_j\|_{2,n,X} \leq \varepsilon$. Then it holds that

$$\|f - f_j\|_{2,n} = \big(\frac{1}{n}\sum_{i=1}^{n}\underbrace{\{Y_i\delta(X_i) - Y_i\delta_j(X_i)\}^2}_{=Y_i^2(\delta(X_i)-\delta_j(X_i))^2}\big)^{1/2} \overset{Y_i\in\{-1,1\}}{\leq} \|\delta - \delta_j\|_{2,n,X} \leq \varepsilon.$$

$\Rightarrow N(\varepsilon, \mathcal{F}_2, \|\cdot\|_{2,n}) \leq N(\varepsilon, \mathcal{F}_3, \|\cdot\|_{2,n,X})$.

(d) Let $(\tilde{f}_j)$ be a ε-Covering of $\mathcal{F}_4$. Then $f_j := \rho \cdot f_j$ is a $\rho\varepsilon$-Covering of $\mathcal{F}_3$. Proof: Similarly to (b) by using the specific function $\phi(x) = \rho \cdot x$.

Solution 26 (Solution of Task 26)**.** (a) u is increasing. We conclude that for all $M \in \mathbb{N}$, $0 \leq s_0 \leq ... \leq s_M \leq 1$,

$$\sum_{i=1}^{M}\big|u(s_i) - u(s_{i-1})\big| = \sum_{i=1}^{M}\big(u(s_i) - u(s_{i-1})\big) = u(s_M) - u(s_0) \leq u(1) - u(0),$$

with equality if $s_M = 1$, $s_0 = 0$. It follows that $|u|_{BV} = u(1) - u(0)$.

(b) For $z \in [t_j, t_{j+1})$ and due to $t_1 \leq t_2 \leq ... \leq t_N$, it holds that

$$\tilde{u}(z) = u(0) + \frac{C}{N}\cdot j.$$

By definition of t_j,

$$C\frac{j}{N} < u(z) - u(0) \leq C\frac{j+1}{N} \quad \Rightarrow \quad u(0) + C\frac{j+1}{N} < u(z) \leq u(0) + C\frac{j+1}{N}$$

(Note that both statements do not hold if $t_j = t_{j+1}$).
We conclude that

$$|u(z) - \tilde{u}(z)| \leq \frac{C}{N}.$$

For each $z \in [0,1)$ there exists $j \in \{0, ..., N-1\}$ with $z \in [t_j, t_{j+1})$. We conclude that

$$\|u - \tilde{u}\|_\infty \leq \frac{C}{N} = \frac{|u|_{BV}}{N}.$$

(c) Let $\tilde{\tilde{u}}$ be the alternative representation given in the task. For $z \in [t_j, t_{j+1})$, it holds that

$$\begin{aligned}
\tilde{\tilde{u}}(z) &= \frac{u(0)+u(1)}{2} + \sum_{i=1}^{j} \frac{C}{2N} - \sum_{i=j+1}^{N} \frac{C}{2N} \\
&= \frac{u(0)+u(1)}{2} + \frac{C}{2N}(j - (N-j)) \\
&= \frac{u(0)+u(1)}{2} - \frac{C}{2} + \frac{C}{N} j \\
&= u(0) + \frac{C}{N} j = \tilde{u}(z).
\end{aligned}$$

(d) We use a similar construction principle for v based on 'quantiles' $q_i := \sup\{z \in [0,1] : v(z) - v(0) \leq \frac{|v|_{BV}}{N} i\}$, $i = 0, ..., N$. We obtain that

$$\tilde{v}(z) = \frac{v(0)+v(1)}{2}(\mathbb{1}_{\{z \geq 0\}} - \mathbb{1}_{\{z<0\}}) + \sum_{i=1}^{N} \frac{|v|_{BV}}{2N}(\mathbb{1}_{\{z \geq q_i\}} - \mathbb{1}_{\{z<q_i\}}).$$

Define $\tilde{g} := \tilde{u} - \tilde{v}$. Then it holds that

$$\|\tilde{g} - g\|_\infty \leq \|\tilde{u} - u\|_\infty + \|\tilde{v} - v\|_\infty \leq \frac{|u|_{BV}}{N} + \frac{|v|_{BV}}{N} = \frac{|g|_{BV}}{N}$$

and

$$\begin{aligned}
\tilde{g}(z) &= \tilde{u}(z) - \tilde{v}(z) = \underbrace{\Big[\frac{u(0)+u(1)}{2} - \frac{v(0)+v(1)}{2}\Big]}_{= \frac{g(0)+g(1)}{2}} (\mathbb{1}_{\{z \geq 0\}} - \mathbb{1}_{\{z<0\}}) \\
&\quad + \sum_{i=1}^{N} \frac{|u|_{BV}}{2N}(\mathbb{1}_{\{z \geq t_i\}} - \mathbb{1}_{\{z<t_i\}}) - \sum_{i=1}^{N} \frac{|v|_{BV}}{2N}(\mathbb{1}_{\{z \geq q_i\}} - \mathbb{1}_{\{z<q_i\}}).
\end{aligned}$$

Solution 27 (Solution of Task 27)**.** (a) Task 26(e) implies that there exist increasing functions u_j, v_j with $h_j = u_j - v_j$ and $|h_j|_{BV} = |u_j|_{BV} + |v_j|_{BV}$. Moreover, there exist functions $\tilde{h}_j : [0,1] \to \mathbb{R}$ with

$$\|h_j - \tilde{h}_j\|_\infty \leq \frac{B}{N}$$

and

$$\|\tilde{h}_j\|_1 \leq \frac{|h_j(0)|+|h_j(1)|}{2} + \sum_{i=1}^{N} \big(\frac{|u|_{BV}}{2N} + \frac{|v|}{2N}\big) = \frac{|h_j(0)|+|h_j(1)|+|h_j|_{BV}}{2} \leq \frac{B}{2}.$$

(b) It holds that (be aware that we do not distinguish between the notations $\|\cdot\|_\infty$ and $\|\cdot\|_1$ for functions starting from d dimensions or 1 dimension)

$$\|\delta^* - \tilde{h}\|_\infty \leq \sum_{j=1}^{d} \|h_j - \tilde{h}_j\|_\infty \leq \frac{Bd}{N}$$

and

$$\|\tilde{h}\|_1 \leq \sum_{j=1}^{d} \|\tilde{h}_j\|_1 \leq \frac{Bd}{2}.$$

(c) For δ^* we choose $\tilde{h} \in \Delta$. Then it holds that

$$\inf_{\delta\in\Delta} \big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} \leq \tilde{R}(\tilde{h}_N) - \tilde{R}(\delta^*) + 2\lambda P(\|\tilde{h}_N\|_1).$$

Since $\|\delta^*\|_\infty \leq \sum_{j=1}^{d} \|h_j\|_\infty \leq \frac{Bd}{2}$ and $\|\tilde{h}_N\|_\infty \leq \|\tilde{h}_N - \delta^*\|_\infty + \|\delta^*\|_\infty \leq \frac{Bd}{N} + \frac{Bd}{2}$, we have

$$\phi(-y\tilde{h}_N(x)) \leq \phi(\|\tilde{h}_N\|_\infty) < \infty, \qquad \phi(-y\delta^*(x)) \leq \phi(\|\delta^*\|_\infty) < \infty.$$

By the dominated convergence theorem, Lipschitz continuity of ϕ with Lipschitz constant L and $\tilde{h}_N \to \delta^*$ uniformly, we have

$$\tilde{R}(\tilde{h}_N) - \tilde{R}(\delta^*) = \mathbb{E}\big[\underbrace{\phi(-Y\tilde{h}_N(X)) - \phi(-Y\delta^*(X))}_{\to 0 \quad (N\to\infty)}\big] \to 0.$$

It follows that

$$\begin{aligned} &\inf_{\delta\in\Delta} \big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} \\ \leq\ &\limsup_{N\to\infty} \big\{\tilde{R}(\tilde{h}_N) - \tilde{R}(\delta^*) + 2\lambda P(\|\tilde{h}_N\|_1)\big\} \leq 2\lambda P(\|\tilde{h}_N\|_1) \leq 2\lambda P(\frac{Bd}{2}). \end{aligned}$$

(d) We first consider λ. It holds that

$$V = 2\lfloor 2\log_2(2d)\rfloor \leq 4\log_2(2d) = \frac{4}{\log(2)}\log(2d) \leq \frac{4}{\log(2)}(\log(2)+\log(d+1)) \leq \frac{8}{\log(2)}\log(d+1).$$

Since $1 \leq \frac{V+2}{V+1} \leq 2$, we conclude that

$$((V+2)V^{1/2})^{\frac{V+2}{V+1}} \leq (3V^{3/2})^{\frac{V+2}{V+1}} \leq 9V^3 \leq 9\big(\frac{8}{\log(2)}\big)^3 \log(d+1)^3, \qquad n^{-\frac{1}{2}\frac{V+2}{V+1}} \leq n^{-\frac{1}{2}}.$$

Plugging in these findings into λ yields

$$\lambda \leq c \cdot 9\big(\frac{8}{\log(2)}\big)^3 \cdot (1+t)\log(d+1)^3 \cdot n^{-\frac{1}{2}}.$$

We now consider the whole penalization term: If $\phi(x) = e^x$, it holds that

$$P(\frac{\rho}{2}) = (\rho\phi'(\rho))^{\frac{V}{V+1}}\phi(\rho)^{\frac{1}{V+1}} + \phi(\rho) \leq 2(\rho+1)e^{\rho}.$$

This implies $P(\frac{Bd}{2}) \leq 2(Bd+1)e^{Bd}$.
If $\phi(x) = \log(1+e^x)$, we have $\phi'(x) = \frac{e^x}{1+e^x} \leq 1$ and $\phi(x) \leq 1+x$. Thus

$$P(\frac{\rho}{2}) = (\rho\phi'(\rho))^{\frac{V}{V+1}}\phi(\rho)^{\frac{1}{V+1}} + \phi(\rho) \leq 2(\rho+1).$$

This implies $P(\frac{Bd}{2}) \leq 2(Bd+1)$. Plugging in this results into (c) $2\lambda P(\frac{Bd}{2})$ yields the claim.

(e) In (b) we obtain

$$\|h - \tilde{h}\|_\infty \leq \frac{Bs}{N}, \qquad \|\tilde{h}\|_1 \leq \frac{Bs}{2}.$$

Thus, only the quantities in the case distinction (d) change, d is replaced by s. We obtain

$$\inf_{\delta\in\Delta}\big\{\tilde{R}(\delta) - \tilde{R}(\delta^*) + 2\lambda P(\|\delta\|_1)\big\} \leq c'(1+t)\log(d+1)^3 \cdot n^{-\frac{1}{2}} \cdot \begin{cases} (Bs+1)e^{Bs+1}, & \phi = \phi_1, \\ (Bs+1), & \phi = \phi_2. \end{cases}$$

We therefore 'pay' with the factor $\log(d+1)^3$ that the underlying dimension d is unknown. This term is present due to the oracle inequality. Instead, the 'real' dimension s of the underlying true function (that is, the number of non-zero summands) now enters the rate at the 'costly' locations $(Bs+1)e^{Bs+1}$ or $(Bs+1)$, respectively.

8.6 Solutions of Chapter 7

Solution 28 (Solution of Task 28)**.** (a) One may use a simple plot program.

(b) The network has $1 + (m-1) = m$ hidden layers with the layer widths given in the sketch.

(c) It holds that

$$\begin{aligned} & g(\frac{x-y+1}{2}) - g(\frac{x+y}{2}) + \frac{x+y}{2} - \frac{1}{4} \\ = \; & \frac{1+(x-y)}{2} \cdot \frac{1-(x-y)}{2} - \frac{x+y}{2} \cdot (1 - \frac{x+y}{2}) + \frac{x+y}{2} - \frac{1}{4} \\ = \; & \frac{1}{4}(1-(x-y)^2) + \frac{1}{4}(x+y)^2 - \frac{1}{4} \\ = \; & \frac{1}{4}((x+y)^2 - (x-y)^2) = \frac{1}{4} \cdot 4xy = xy. \end{aligned}$$

Therefore, we can approximate the function $x \cdot y$ by using networks which approximate g.

(d) The sketch contains $q = \lceil \log_2(r) \rceil$ steps ($2^q \to 2^{q-1}$, ..., $2^1 \to 2^0$). In each step, we have to insert multiplication networks $f_m \in \mathcal{F}((m+4), (2, 6, 6, ..., 6, 1))$. To connect the networks, we need one additional layer. In the first step, we need to summarize at most r networks (therefore, multiply the hidden layer size with r). Then we obtain a

$$\mathcal{F}(q(m+5), (r, 6r, 6r, ..., 6r, 1))$$

network.

Approximation: For simplicity, let $r = 2^q$ with some $q \in \mathbb{N}$. Then we have with $y_1 = (x_1, ..., x_{r/2}), y_2 = (x_{r/2+1}, ..., x_r)$ and since $0 \leq f_{m,r/2}(y_1) \leq 1$ that

$$\begin{aligned} \left|f_{m,r}(x) - \prod_{j=1}^{r} x_j\right| &= \left|f_m(f_{m,r/2}(y_1), f_{m,r/2}(y_2)) - \prod_{j=1}^{r/2} x_j \cdot \prod_{j=r/2+1}^{r} x_j\right| \\ &\leq \left|f_m(f_{m,r/2}(y_1), f_{m,r/2}(y_2)) - f_{m,r/2}(y_1) \cdot f_{m,r/2}(y_2)\right| \\ &\quad + \left|f_{m,r/2}(y_1) \cdot f_{m,r/2}(y_2) - \prod_{j=1}^{r/2} x_j \cdot \prod_{j=r/2+1}^{r} x_j\right| \\ &\leq 2^{-m} + \left|f_{m,r/2}(y_1) - \prod_{j=1}^{r/2} x_j\right| + \left|f_{m,r/2}(y_2) - \prod_{j=r/2+1}^{r} x_j\right| \end{aligned}$$

Because of $|f_{m,2}(x) - x_1 x_2| \leq 2^{-m}$, the two recursive terms above do no longer show up in step $r/2 = 1$.

By induction, we have $\left|f_{m,r}(x) - \prod_{j=1}^{r} x_j\right| \leq 3^q 2^{-m} \leq r^2 2^{-m}$.

(e) Using a Taylor expansion around a, we obtain that there exists $\xi_x \in [0, 1]$ with

$$f(x) = T_a(x) + \sum_{\alpha \in \mathbb{N}_0^r : |\alpha| = \beta} (\partial^\alpha f)(a + \xi(x-a)) \cdot \frac{(x-a)^\alpha}{\alpha!}.$$

$\Rightarrow$

$$|f(x)-T_a(x)| \leq \sum_{\alpha\in\mathbb{N}_0^r:|\alpha|=\beta} \underbrace{\frac{|(x-a)^\alpha|}{\alpha!}}_{\leq \frac{|x-a|_\infty^\beta}{\alpha!}} \cdot \underbrace{\left|(\partial^\alpha f)(a+\xi(x-a))\right|}_{\leq K} \leq K|x-a|_\infty^\beta \cdot \underbrace{\sum_{|\alpha|=\beta} \frac{1}{\alpha!}}_{\leq e^r}.$$

(f) It holds that

$$\sum_{a\in D(M)} \prod_{j=1}^r \left(1-M\cdot|x_j-a_j|\right)_+ = \prod_{j=1}^r \sum_{l=0}^M (1-M|x_j-\ell/M|)_+ = 1,$$

thus

$$|T(x)-f(x)| \leq \sum_{a\in D(M),\|a-x\|_\infty\leq 1/M} \underbrace{\left|T_a(x)-f(x)\right|}_{\overset{(e)}{\leq} Ke^rM^{-\beta}} \cdot \prod_{j=1}^r \left(1-M\cdot|x_j-a_j|\right)_+ \leq Ke^rM^{-\beta}.$$

(g) In statement (f), the approximation of arbitrary continuously differentiable functions is reduced to the approximation of products (or products of absolute values). The result of (d) shows how products can be approximated with neural networks.

Solution 29 (Solution of Task 29)**. Lösung:**

(a) For arbitrary $t_0 > 0$, it holds that

$$\begin{aligned}\mathbb{E}[W^2] &= \int_0^\infty \mathbb{P}(W^2\geq u)du = \int_0^\infty \mathbb{P}(W\geq\sqrt{u})du \overset{t=\sqrt{u}}{=} 2\int_0^\infty t\mathbb{P}(W\geq t)dt \\ &= \underbrace{2\int_0^{t_0} t\underbrace{\mathbb{P}(W\geq t)}_{\leq 1}dt}_{\leq t_0^2} + 2\int_{t_0}^\infty t\mathbb{P}(W\geq t)dt.\end{aligned}$$

(b) For each $g\in\mathcal{G}$, it holds that

$$\|\tilde{g}\|_\infty \leq \frac{\|g\|_\infty}{M\sqrt{\frac{H}{n}}} \leq \sqrt{\frac{n}{H}},$$

and

$$\mathbb{E}[g(Z_1)^2] = \frac{\mathbb{E}[g(Z_1)^2]}{\mathbb{E}[g(Z_1)^2]} \leq 1.$$

Using Bernstein's inequality, we have

$$\begin{aligned}\mathbb{P}\Big(\big|\sum_{i=1}^{n}\{\tilde{g}(Z_i)-\mathbb{E}\tilde{g}(Z_i)\}\big|\geq t\Big) &\leq 2\exp\Big(-\frac{t^2}{2n+\frac{2}{3}\sqrt{\frac{n}{H}}\cdot t}\Big)\\ &\overset{t\geq 3\sqrt{nH}}{\leq} 2\exp\Big(-\frac{t^2}{\frac{4}{3}\sqrt{\frac{n}{H}}\cdot t}\Big)=2\exp\big(-\frac{3t}{4\sqrt{\frac{n}{H}}}\big).\end{aligned}$$

We conclude that

$$\mathbb{P}(W\geq t)\leq|\mathcal{G}|\mathbb{P}\big(\big|\sum_{i=1}^{n}\{\tilde{g}(Z_i)-\mathbb{E}\tilde{g}(Z_i)\}\big|\geq t\big)\leq 2|\mathcal{G}|\exp\big(-\frac{3t}{4\sqrt{\frac{n}{H}}}\big).$$

(c) Choose $t_0=3\sqrt{nH}$ und $a=\frac{3}{4}\sqrt{\frac{H}{n}}$.
Partial integration yields that $\int te^{-at}dt=-(a^{-2}+a^{-1}t)e^{-at}\Rightarrow\int_{t_0}^{\infty}te^{-at}dt=(a^{-2}+a^{-1}t_0)e^{-at_0}$. We conclude that

$$\begin{aligned}2\cdot\int_{t_0}^{\infty}t\mathbb{P}(W\geq t)dt &\leq 4(a^{-2}+a^{-1}t_0)|\mathcal{G}|\underbrace{e^{-at_0}}_{=e^{-\frac{9}{4}H}\leq|\mathcal{G}|^{-1}}\\ &\leq 4\big\{(\frac{4}{3})^2\frac{n}{H}+4n\big\}\\ &\overset{|\mathcal{G}|\geq 2}{\leq} 24n.\end{aligned}$$

Together with $t_0^2=9nH$ the claim now follows from (a).

Solution 30 (Solution of Task 30)**.** (a) $Z_1,...,Z_n\sim N(0,v^2)$ i.i.d. $\Rightarrow$

$$W_j=\frac{1}{v\sqrt{n}}\frac{\sum_{i=1}^{n}a_{ij}Z_i}{\big(\frac{1}{n}\sum_{i=1}^{n}a_{ij}^2\big)^{1/2}}\sim N(0,1).$$

(b) For $W_j\sim N(0,1)$, it holds that $\mathbb{E}\exp(\frac{W_j^2}{4})=\sqrt{2}$. The function $\varphi_2(x)=\exp(x^2)-1$ is convex and increasing. It follows that

$$\varphi_2(\mathbb{E}\max_{j=1,...,N}\frac{|W_j|}{2})\leq\mathbb{E}\max_{j=1,...,N}\varphi_2(\frac{|W_j|}{2})\leq\sum_{j=1}^{N}\mathbb{E}\varphi_2(\frac{|W_j|}{2})\leq N\cdot(\sqrt{2}-1)\leq N.$$

$\Rightarrow$

$$\mathbb{E}\max_{j=1,...,N}|W_j|\leq 2\varphi_2^{-1}(N)=2\sqrt{\log(N+1)}.$$

(c) With $\varphi_1(x) = e^x - 1$, we have

$$\varphi_1(\mathbb{E}\max_{j=1,...,N}\frac{|W_j|^2}{4}) \leq \mathbb{E}\max_{j=1,...,N}\varphi_1(\frac{|W_j|^2}{4}) \leq \sum_{j=1}^{N}\mathbb{E}\varphi_1(\frac{|W_j|^2}{4}) \leq N\cdot(\sqrt{2}-1) \leq N.$$

$\Rightarrow$

$$\mathbb{E}\max_{j=1,...,N}|W_j|^2 \leq 4\varphi_1^{-1}(N) = 4\log(N+1).$$

Solution 31 (Solution of Task 31)**.** (a) Write $\theta = (v^1, ..., v^{(L)}, W^{(0)}, ..., W^{(L)})$ as a vector (the matrices are vectorized row by row). Then we have $\theta \in [-1,1]^T$ and

$$T = \sum_{l=1}^{L}\text{number of entries of } v^{(l)} + \sum_{l=0}^{L}\text{number of entries of } W^{(l)} = \sum_{l=1}^{L}p_l + \sum_{l=0}^{L}p_l p_{l+1}.$$

(b) Each $f_\theta \in \mathcal{F}(L,p,s,\infty)$ has at most s entries of θ being non-zero, that is, $\mathcal{F}(L,p,s,\infty) = \{f_\theta : \theta \in \Theta\}$ with

$$\begin{aligned}\Theta &\subset \{\theta \in [-1,1]^T : \leq s \text{ entries of } \theta \text{ non-zero}\}\\ &= \{\theta \in [-1,1]^T : \exists S \subset \{1,...,T\}, |S| \leq s : \forall j \in S^c : \theta_j = 0\}\\ &= \bigcup_{S\subset\{1,...,T\}:|S|\leq s}\underbrace{\{\theta \in [-1,1]^T : \forall j \in S^c : \theta_j = 0\}}_{=:\Theta_S}.\end{aligned}$$

(c) We choose a reasonable grid approximation of Θ_S. Without loss of generality, let $S = \{1,...,s\}$ (then we have $\Theta_S \subset [-1,1]^s \times \{0\}^{T-s}$). Put

$$\tilde{\Theta}_S = \Big\{-1 + j\cdot a : j = 1,...,m\Big\}^s \times \{0\}^{T-s}.$$

If $-1 + m\cdot a \geq 1 - a$, then $\tilde{\Theta}_S$ satisfies the approximation statement in the claim. This is fulfilled with $(m+1)\cdot a \geq 2$, that is, $m \geq \frac{2}{a} - 1$ or equivalently, $m \geq \lfloor\frac{2}{a}\rfloor$.

We conclude that

$$|\tilde{\Theta}_S| \leq m^s = \left\lfloor\frac{2}{a}\right\rfloor^s.$$

(d) It holds that

$$\begin{aligned}|\tilde{\Theta}| &\leq \sum_{S\subset\{1,...,T\}:|S|\leq s}|\tilde{\Theta}_S| \leq \sum_{k=0}^{s}\sum_{S\subset\{1,...,T\}:|S|=k}|\tilde{\Theta}_S|\\ &\overset{(c)}{=} \sum_{k=0}^{s}\#\{S\subset\{1,...,T\} : |S| = k\}\cdot\left\lfloor\frac{2}{a}\right\rfloor^s.\end{aligned}$$

We have

$$\#\{S \subset \{1,...,T\} : |S| = k\} = \binom{T}{k} \leq T^k \leq V^k$$

since

$$T = \sum_{l=0}^{L} p_l p_{l+1} + \sum_{l=1}^{L} p_l \leq \sum_{l=0}^{L} (p_l + 1) p_{l+1} \leq \prod_{l=0}^{L+1} (p_l + 1) = V.$$

Plugging in this result yields

$$|\tilde{\Theta}| \leq \sum_{k=0}^{s} V^k \cdot \big(\frac{2}{a}\big)^k \leq \big(\frac{2V}{a}\big)^{s+1}.$$

(e) For $\gamma > 0$, choose $a = \frac{\gamma}{(L+1)V}$. Then for each $f \in \mathcal{F}(L, p, s, F)$, there exists some $f_{\tilde{\theta}}$ with $\tilde{\theta} \in \tilde{\Theta}$ such that

$$\|f_\theta - f_{\tilde{\theta}}\|_\infty \leq a \cdot (L+1) \cdot V \leq \gamma.$$

$\Rightarrow$

$$N(\gamma, \mathcal{F}(L, p, s, F), \|\cdot\|_\infty) \leq \big(\frac{2V}{a}\big)^{s+1} = \big(2\gamma^{-1} V^2 (L+1)\big)^{s+1}.$$

(f) It holds that

$$\begin{aligned} |A_\theta^{k-}(x) - A_\theta^{k-}(x')| &\leq \|W^{(L)}\|_Z \cdot ... \cdot \|W^{(k-1)}\|_Z \|x - x'\|_\infty \\ &\overset{\|W^{(l)}\|_\infty \leq 1, W^{(l)} \in \mathbb{R}^{p_l \times p_{l+1}}}{\leq} \Big(\prod_{l=k-1}^{L} p_l\Big) \cdot \|x - x'\|_\infty. \end{aligned}$$

(g) It holds that

$$\begin{aligned} |A_\theta^{k+}(x)|_\infty &\leq \|W^{(k)}\|_Z \cdot \big\{...\|W^{(1)}\|_Z \big\{\|W^{(0)}\|_Z \|x\|_\infty + \|v^{(1)}\|_\infty) + \|v^{(2)}\|_\infty \big\}...\big\} + \|v^{(k)}\|_\infty \\ &\leq \prod_{l=0}^{k-1} (p_l + 1). \end{aligned}$$

(h) Using the convention $\sigma_{v^{(L+1)}}(x) := x$, we have

$$\begin{aligned}
|f(x) - f_{\tilde{\theta}}(x)| &\leq \sum_{k=1}^{L+1} \left| A_{\theta}^{(k+1)-} \circ \sigma_{v^{(k)}} W^{(k-1)} A_{\tilde{\theta}}^{(k-1)-}(x) - A_{\theta}^{(k+1)-} \circ \sigma_{\tilde{v}^{(k)}} \tilde{W}^{(k-1)} A_{\tilde{\theta}}^{(k-1)-}(x) \right| \\
&\overset{(f)}{\leq} \sum_{k=1}^{L+1} \Big(\prod_{l=k}^{L} p_l \Big) \left\| \sigma_{v^{(k)}} W^{(k-1)} A_{\tilde{\theta}}^{(k-1)-}(x) - \sigma_{\tilde{v}^{(k)}} \tilde{W}^{(k-1)} A_{\tilde{\theta}}^{(k-1)-}(x) \right\|_{\infty} \\
&\leq \sum_{k=1}^{L+1} \Big(\prod_{l=k}^{L} p_l \Big) \cdot \left[\|v^{(k)} - \tilde{v}^{(k)}\|_{\infty} + \|W^{(k-1)} - \tilde{W}^{(k-1)}\|_Z \cdot \|A_{\tilde{\theta}}^{(k-1)}(x)\|_{\infty} \right] \\
&\overset{(g)}{\leq} \sum_{k=1}^{L+1} \Big(\prod_{l=k}^{L} p_l \Big) \cdot \left[a + p_{k-1} \cdot a \cdot \prod_{l=0}^{k-2} (p_l + 1) \right] \\
&\leq a \cdot (L+1) \cdot \prod_{l=0}^{L} (p_l + 1) = a \cdot (L+1) \cdot V.
\end{aligned}$$

(i) If $f \in \mathcal{F}(L, p, s, F)$, then for each layer $l \in \{1, ..., L\}$ it holds that at most s columns of $W^{(l)} \in \mathbb{R}^{p_l \times p_{l+1}}$ are nonzero. If the j-th column of $W^{(l)}$ is a zero vector, we can eliminate the j-th row of $W^{(l-1)}$ and the j-th entry of $v^{(l)}$ from the model by still representing the same function. We then obtain that

$$f \in \mathcal{F}(L, (p_0, p_1, ..., p_{l-1}, p_l - 1, p_{l+1}, ..., p_L, p_{L+1}), s, F).$$

Repeating this argument, we obtain

$$f \in \mathcal{F}(L, (p_0, p_1 \wedge s, ..., p_L \wedge s, p_{L+1}), s, F).$$

(j) By $\tilde{V} = (p_0 + 1) \cdot (p_{L+1} + 1) \cdot \prod_{l=1}^{L} ((p_l \wedge s) + 1) \leq 2^{L+2} p_0 p_{L+1} s^L$, we have

$$\begin{aligned}
H(\gamma) &= \log N(\gamma, \mathcal{F}(L, p, s, F), \|\cdot\|_{\infty}) \\
&\overset{(i)}{=} \log N(\gamma, \mathcal{F}(L, (p_0, p_1 \wedge s, ..., p_L \wedge s, p_{L+1}), s, F), \|\cdot\|_{\infty}) \\
&\overset{(e)}{\leq} (s+1) \log(2\gamma^{-1} \tilde{V}^2 (L+1)) \\
&= (s+1) \log(2^{2L+5} \gamma^{-1} (L+1) p_0^2 p_{L+1}^2 s^{2L}) \\
&\leq 2s \cdot \big\{ (2L+5) \log(2) + \log(\gamma^{-1}) + \log(2L) + 2\log(p_0 p_{L+1}) + 2L \log(s) \big\} \\
&\leq c \cdot s \cdot \big\{ L \log(s) + \log(\gamma^{-1}) + \log(p_0 p_{L+1}) \big\}
\end{aligned}$$

where $c \geq 1$ is some large enough universal constant and $s \geq 2, L \geq 1$.

References

[1] Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36(2):489–531, 2008.

[2] Gilles Blanchard, Gábor Lugosi, and Nicolas Vayatis. On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.*, 4(5):861–894, 2004.

[3] Gilles Blanchard, C. Schäfer, Yves Rozenholc, and Klaus-Robert Müller. Optimal dyadic decision trees. *Machine Learning*, 66:209–241, 03 2007.

[4] Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[5] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[6] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[7] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators, 2014.

[8] Rafał Latała. On some inequalities for Gaussian measures. *arXiv Mathematics e-prints*, page math/0304343, April 2003.

[9] Jean-Samuel Leboeuf, Frédéric LeBlanc, and Mario Marchand. Decision trees as partitioning machines to characterize their generalization properties, 2020.

[10] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[11] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808 – 1829, 1999.

[12] S. Mendelson. Improving the sample complexity using global data. *IEEE Trans. Inf. Theory*, 48:1977–1991, 2002.

[13] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(78):2241–2259, 2010.

[14] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv e-prints*, page arXiv:1708.06633, Aug 2017.

[15] Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35(2):575–607, 2007.

[16] M. Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *The Annals of Probability*, 22(1):28 – 76, 1994.

[17] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135 – 166, 2004.

[18] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[19] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[20] Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.