Hochdimensionale nichtparametrische Regression mit tiefen neuronalen Netzen

Bachelorarbeit

am Fachbereich Mathematik und Informatik
der Philipps–Universität Marburg

zur Erlangung des akademischen Grades
Bachelor of Science

vorgelegt von
David Reifferscheidt

Oktober 2020
Betreuer: Prof. Dr. H. Holzmann

Inhaltsverzeichnis

Einleitung

Regression ist die statistische Methode, um den funktionalen Zusammenhang zwischen einer Kovariablen X und einer sogenannten Zielvariable Y zu schätzen bzw. modellieren. Sei dazu also (X, Y) eine $\mathbb{R}^d \times \mathbb{R}$–wertige Zufallsvariable mit $\mathbb{E}[|Y|] < \infty$, dann ist das Ziel der Regressionsanalyse eine messbare Funktion $f : \mathbb{R}^d \to \mathbb{R}$ zu finden, sodass $f(X)$ eine gute Approximation für Y liefert (siehe [Gyö+02]). Dazu gibt es verschiedene Konzepte. So wird bei der parametrischen Regression angenommen, dass die Regressionsfunktion zu einer bestimmten Klasse von Funktionen gehört, die durch endlich viele Parameter beschrieben werden kann. Oft wird dabei die Klasse der linearen Funktionen verwendet. Eine Schwäche der parametrischen Regression ist leicht erkennbar. Wählt man die falsche Funktionenklasse, hat dies schlechte Schätzungen der Regressionsfunktion zur Folge. Diese Spezifikationsfehler werden bei der *nichtparametrischen Regression* umgangen ([Bir07] Seite 5 ff.). Um eine nichtparametrische Regressionsfunktion zu schätzen, gibt es verschiedene Ansätze. [Här90] behandelt zum Beispiel Kernschätzer der Regressionsfunktion, [Dia06] untersucht Verfahren basierend auf Splines. Diese Arbeit widmet sich der Regressionsschätzung hochdimensionaler Funktionen durch tiefe neuronale Netze.
Der enorme Erfolg, den künstliche neuronale Netze in den vergangenen Jahren haben, ist zum einen auch auf die Vielfalt an Aufgaben zurückzuführen, die diese Algorithmen bearbeiten können [FMZ19]. So finden diese Verfahren des maschinellen Lernens auch in der nichtparametrischen Regression Anwendung (vgl. Kapitel 9 in [IM98]).
Kohler und Bauer nutzen in [BK19] tiefe neuronale Netze, um Regressionsschätzer zu konstruieren, die mit einer Konvergenzrate unabhängig von der Dimension d gegen die eigentliche Regressionsfunktion konvergieren. Um eine solche Konvergenzrate zu erhalten, müssen einige Anforderungen an die zugrundeliegende Verteilung gestellt werden. Kohler und Bauer fordern eine Glattheitsbedingung, sowie die Form eines hierarchischen Interaktionsmodells an die Regressionsfunktion und erhalten somit eine Konvergenzrate von $\log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}}$ mit $d^* \leq d$. Dieses Hauptresultat möchten wir im Umfang dieser Arbeit ausführen und die statistischen Zusammenhänge verdeutlichen.
Zu Beginn der Arbeit werden in Kapitel 1 einige grundlegende Konzepte, Aussagen und Definitionen eingeführt, die für die Erarbeitung der Materie dieser Untersuchung von Bedeutung sind. Dazu gehören Grundlagen zu nichtparametrischer Regression, sowie zu maschinellem Lernen. Die Einschränkung der Regressionsfunktion, sowie die Wahl eines geeigneten Schätzers motivieren wir in Kapitel 2, woraufhin wir die Funktionenklasse $\mathcal{H}^{(l)}$, der neuronalen Netze einführen. Um das Hauptresultat aus [BK19] zu beweisen, benutzen wir eine Orakel–Ungleichung, sowie einige vorbereitende Konzentrationsungleichungen, die wir in Kapitel 4 präzisieren und zeigen. Außerdem benötigen wir Approximationseigenschaften unserer Funktionenklasse, die wir in Kapitel 5 nennen. Zuletzt implementieren

wir die Klasse hierarchischer (dünnbesetzter) neuronaler Netze in `Python` und vergleichen dessen Schätzqualität mit derer herkömmlicher neuronaler feedforward–Netze.

Kapitel 1

Theoretische Grundlagen

Zunächst führen wir einige theoretischen Grundlagen zu maschinellem Lernen, tiefem Lernen und zu nichtparametrischer Regression ein, die wir im Verlauf dieser Arbeit benutzen werden.

1.1. Maschinelles Lernen

Maschinelles Lernen (engl. *Machine Learning* (ML)) behandelt Algorithmen, die aus (großen) Datensätzen lernen können. **Lernen** bedeutet in diesem Zusammenhang, dass sich die Performance des Computerprogramms durch die gegebenen Daten bezüglich einer Aufgabe verbessert [Mit97].
Übliche Machine–Learning Aufgaben sind z.B. Klassifizierung, Regression, maschinelle Übersetzung oder Dichteschätzung. Man unterscheidet außerdem zwischen überwachtem und unüberwachtem Lernen. Bei Algorithmen für **überwachtes Lernen** wird ein Datensatz genutzt, der zu jedem Merkmal auch ein **Label** oder **Zielwert** enthält. Bei **unüberwachtem Lernen** beinhalten die Daten keine solchen Zielwerte.

In dieser Ausarbeitung werden wir uns auf Methoden des überwachten Lernens fokussieren. Das heißt, gegeben sei eine Datenmenge

$$\mathcal{D}_n = \{ (x_1, y_1), \ldots, (x_n, y_n) \},$$

wobei $x_i \in \mathbb{R}^d$ die Eingabedaten und $y_i \in \mathbb{R}$ die Label sind. Die ML–Modelle sind meist durch Parameter aus einem Parameterraum $\Theta \subseteq \mathbb{R}^m$ definiert. Dann ist das Ziel des Machine–Learning–Algorithmus eine Funktion

$$f : \mathbb{R}^d \to \mathbb{R},\ f(x_i) = f(x_i; \theta) = \hat{y}_i, \quad \theta \in \Theta$$

einer bestimmten Funktionenklasse $\mathcal{F}$ zu finden, die eine gute Prädiktion auf einem Testdatensatz liefert. Der Begriff **Testdaten** beschreibt bereits die zentrale Herausforderung dieser Programme: Der Algorithmus muss mit bisher *unbeobachteten, neuen* Eingangsdaten zurechtkommen, nicht nur mit denen, auf die er trainiert wurde. Also spalten wir unseren Datensatz $\mathcal{D}_n$ in einen **Trainings**– und einen **Test**–Datensatz auf:

$$\mathcal{D}_n^{(\text{train})} = \{ (x_1, y_1), \ldots, (x_k, y_k) \}, \quad \mathcal{D}_n^{(\text{test})} = \{ (x_{k+1}, y_{k+1}), \ldots, (x_n, y_n) \}$$

Eine wichtige zugrundeliegende Annahme ist dabei, dass beide Datensätze dieselbe Verteilung besitzen. Die Spaltung in Trainings– und Testdaten hängt von mehreren Faktoren ab, oft wird allerdings $k = 0,8 \cdot n$ gewählt.

Das *Training* findet meistens statt, indem man eine sogenannte **Verlustfunktion** L bezüglich der Parameter $\theta \in \Theta$ minimiert. Diese vergleicht die Prädiktion $\hat{y}_i = f(x_i)$ mit dem (wahren) Label y. Zu lösen ist also außerdem das Optimierungsproblem

$$\min_{\theta \in \Theta} L(y, f(x; \theta)), \quad y = (y_1, \ldots, y_k).$$

Zur Messung der Modellleistung eines trainierten Algorithmus wird ein **Performance**–Maß benutzt. Ein trainiertes Programm kennt den Testdatensatz $\mathcal{D}_n^{(\text{test})}$ nicht, mit diesen Daten wird überprüft, wie gut es mit unbekannten Eingaben zurechtkommt. Diese Fähigkeit wird auch **Generalisierung** genannt.

Insgesamt wird beim Machine Learning also ein geringer **Generalisierungsfehler** sowie ein geringer **Trainingsfehler** angestrebt, wobei man letzteren berechnet, indem aus $\mathcal{D}_n^{(\text{train})}$ noch eine **Validierungsmenge** ausgewählt und die Trainingsleistung zunächst darauf getestet wird.

1.2. Deep Learning - Neuronale Netze

Deep Learning ist ein Unterbereich des maschinellen Lernens, dessen Popularität in den letzten Jahren enorm gewachsen ist. Machine Learning Algorithmen ermöglichen es Computern aus Erfahrung bzw. aus Datenmengen zu lernen, wodurch abstraktere Probleme für Maschinen zugänglicher werden. Die Methoden des Deep Learning nutzen dabei Zusammensetzungen mehrerer nichtlinearer Funktionen, um die Abhängigkeit von Eingabewerten und Zielwerten zu beschreiben [FMZ19]. Die Tiefe dieser Vorgehensweise wird genutzt, um komplexe Konzepte aus einfacheren Konzepten zu konstruieren [GBC18].
In dieser Ausarbeitung werden wir ausschließlich Deep Learning durch künstliche neuronale Feedforward–Netze betrachten. Trotz des Namens handelt es sich dabei nicht um Modelle von Gehirnfunktionen, sondern um Algorithmen zur Funktionsapproximation, die eine statistische Generalisierung zum Ziel haben. Die Netze sind aber entfernt von der Neurowissenschaft inspiriert [GBC18]. Sie bestehen zunächst aus einer Eingabe–Schicht, einer Ausgabe–Schicht und l verdeckten Schichten, $l \in \mathbb{N}$. Ein solches Netzwerk lässt sich also schreiben als Komposition mehrerer nichtlinearer Funktionen:

$$h^{(L)}(x) = g^{(L)} g^{(L-1)} \ldots g^{(1)}(x)$$

Die i–te verdeckte Schicht besteht wiederum aus $K_i \in \mathbb{N}$ Neuronen, welche die Parameter des Modells tragen und die *Breite* des Netzes definieren. Tiefe Neuronale Netze sind also reellwertige Funktionen, definiert auf $\mathbb{R}^d$, die aus der Verknüpfung mehrerer einfacher nichtlinearer Funktionen entstehen [BK19]. Ein mehrschichtiges neuronales Feedforward–Netz (oder auch Mehrschichtiges Perzeptron) besitzt eine spezifische Form der $g^{(l)}$:

$$h^{(l)} = g^{(l)}(h^{(l-1)}) = \sigma(W^{(l)} h^{(l-1)} + b^{(l)}),$$

wobei $W^{(l)}$ die Gewichte–Matrix und $b^{(l)}$ der Bias oder auch Intercept bezüglich der l–ten Schicht ist. $\sigma(\cdot)$ ist eine einfache (bekannte) nichtlineare Funktion, genannt **Aktivie-**

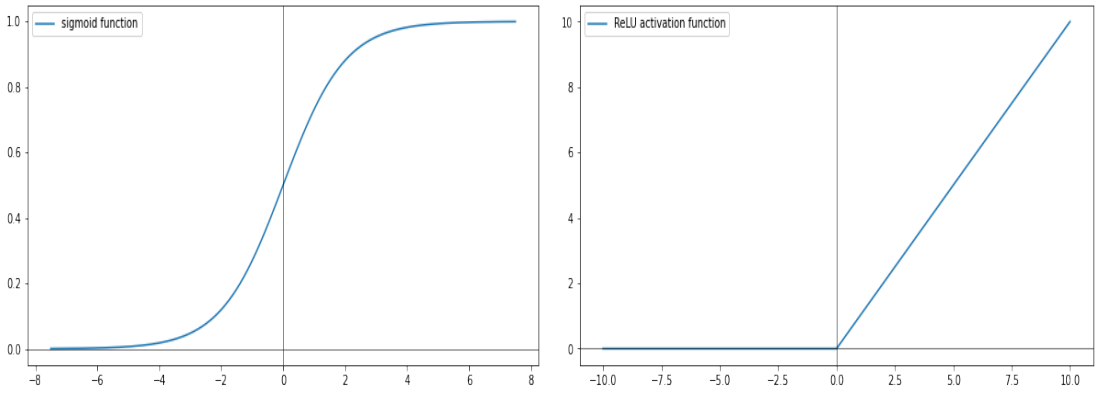

Abbildung 1.1: Sigmoid–Funktion (links) und ReLU–Funktion (rechts)

rungsfunktion. Gängige Aktivierungsfunktionen sind z.B. die *Sigmoid–Funktion*

$$\sigma : \mathbb{R} \to [0,1], \quad \sigma(x) = \frac{1}{1+\exp(-x)} \tag{1}$$

oder die *ReLU–Funktion* (engl. *Rectified Linear Unit*)

$$\sigma : \mathbb{R} \to [0,\infty), \quad \sigma(x) = \max(0,x), \tag{2}$$

dargestellt in Abbildung 1.1.
Das heißt, in jeder Schicht l durchläuft der Eingabevektor $h^{(l-1)}$ zunächst eine affine Transformation bevor die fixe, nichtlineare Funktion σ angewandt wird.

1.3. Nichtparametrische Regression

Wir betrachten in der nichtparametrischen Regression die Zufallsvariable (X, Y) mit Werten in $\mathbb{R}^d \times \mathbb{R}$ und $\mathbb{E}[|Y|] < \infty$. In der parametrischen Regression nehmen wir an, dass die Funktion, die die Abhängigkeit von X und Y modelliert, zu einer bestimmten parametrischen Funktionenklasse gehört. In nichtparametrischen Modellen dagegen fordert man zum einen meist nur die Glattheit der beschreibenden Funktion, zum anderen werden oft weitere Restriktionen wie z.B. Additivität vorausgesetzt.
Wir betrachten die Regressionsfunktion $m : \mathbb{R}^d \to \mathbb{R}$, definiert durch

$$m(x) = \mathbb{E}\big[Y|X = x\big] \quad (x \in \mathbb{R}^d).$$

Dies ist also die Borel–messbare Funktion m, mit

$$\forall B \in \mathcal{B}^d : \int_B m(x)\mathbf{P}_X(dx) = \int_{X^{-1}(B)} Y d\mathbf{P}.$$

Diese ist $\mathbf{P}_X$ fast sicher eindeutig und optimal in dem folgenden Sinne:

Lemma 1. *Sei (X, Y) eine $\mathbb{R}^d \times \mathbb{R}$–wertige Zufallsvariable mit $\mathbb{E}[Y^2] < \infty$, so minimiert die Regressionsfunktion $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbb{E}\big[Y|X = x\big]$ die mittleren quadratische Abweichung, d.h.*

$$\mathbb{E}\big[\,|m(x) - Y|^2\,\big] = \min_{f:\mathbb{R}^d \to \mathbb{R}\ messbar} \mathbb{E}\big[\,|f(X) - Y|^2\,\big].$$

Beweis. Wir zeigen zunächst, dass für jede messbare Funktion $f : \mathbb{R}^d \to \mathbb{R}$ gilt:

$$\mathbb{E}\big[\,|f(X) - Y|^2\,\big] = \mathbb{E}\big[\,|m(X) - Y|^2\,\big] + \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \, \mathbf{P}_X(dx). \tag{3}$$

Wegen

$$\int_{\mathbb{R}^d} |f(x) - m(x)|^2 \, \mathbf{P}_X(dx) \geq 0$$

folgt daraus die Behauptung.

Wir beachten, dass $\mathbb{E}[m(X)^2] < \infty$, denn die Jensensche Ungleichung impliziert

$$\mathbb{E}\big[\,|m(X)|^2\,\big] = \mathbb{E}\big[\,|\mathbb{E}[Y|X]|^2\,\big] \leq \mathbb{E}\big[\mathbb{E}[|Y|^2\,|X]\big] = \mathbb{E}[Y^2] < \infty.$$

Angenommen $\mathbb{E}\big[\,|f(X)|^2\,\big] = \infty$, dann ist

$$\mathbb{E}\big[\,|f(X) - Y|^2\,\big] = \infty = \int_{\mathbb{R}^d} |f(x) - m(x)|^2 \, \mathbf{P}_X(dx),$$

denn es gilt $\mathbb{E}\big[\,|f(X)|^2\,\big] \leq 2 \cdot \mathbb{E}\big[\,|f(X) - m(X)|^2\,\big] + 2 \cdot \mathbb{E}\big[\,|m(X)|^2\,\big]$, also (3).

Ist dagegen $\mathbb{E}\big[\,|f(X)|^2\,\big] < \infty$, so gilt

$$\begin{aligned}
\mathbb{E}\big[\,|f(X) - Y|^2\,\big] &= \mathbb{E}\big[\,|(f(X) - m(X)) - (m(X) - Y)|^2\,\big] \\
&= \mathbb{E}\big[\,|f(X) - m(X)|^2\,\big] + \mathbb{E}\big[\,|m(X) - Y|^2\,\big] \\
&\quad + 2 \cdot \mathbb{E}\big[(f(X) - m(X)) - (m(X) - Y)\big] \\
&= \mathbb{E}\big[\,|f(X) - m(X)|^2\,\big] + \mathbb{E}\big[\,|m(X) - Y|^2\,\big],
\end{aligned}$$

da

$$\begin{aligned}
\mathbb{E}\big[(f(X) - m(X)) \cdot (m(X) - Y)\big] &= \mathbb{E}\big[\mathbb{E}[f(X) - m(X)) \cdot (m(X) - Y)|X]\big] \\
&= \mathbb{E}\big[(f(X) - m(X)) \cdot (\mathbb{E}[Y|X] - m(X))\big] \\
&= 0.
\end{aligned}$$

Dabei nutzten wir, dass $(f(X) - m(X)) \cdot (m(X) - Y)$ integrierbar ist, denn nach Cauchy-Schwarz gilt

$$\mathbb{E}\big[\,|(f(X) - m(X)) \cdot (m(X) - Y)|\,\big] \leq \sqrt{\mathbb{E}\big[\,|f(X) - m(X)|^2\,\big]} \cdot \sqrt{\mathbb{E}\big[\,|m(X) - Y|^2\,\big]} < \infty.$$

□

Nach (3) lässt sich das L_2–*Risiko* durch zwei Summanden darstellen:

$$\mathbb{E}\big[\,|f(X)-Y|^2\,\big] = \mathbb{E}\big[\,|m(X)-Y|^2\,\big] + \int_{\mathbb{R}^d} |f(x)-m(x)|^2\,\mathbf{P}_X(dx)$$

Der erste Summand ist das L_2–Risiko der Regressionsfunktion m, welcher ein unvermeidbarer Fehler ist nach Lemma 1. Der zweite Summand ist der sogenannte L_2–*Fehler*

$$\int_{\mathbb{R}^d} |f(x)-m(x)|^2\,\mathbf{P}_X(dx). \tag{4}$$

1.3.1. Regressionsschätzung

In der praktischen Anwendung ist die Verteilung von (X,Y) und damit auch die Regressionsfunktion $m(x) = \mathbb{E}[Y|X=x]$ unbekannt. Das zugrundeliegende statistische Problem ist dann die Generierung einer Schätzung der Regressionsfunktion bzgl. einer Stichprobe $(X_i,Y_i)_{i=1,\dots,n}$ (vgl. Seite 1 [Sch17]), wobei $(X,Y),(X_1,Y_1),\dots,(X_n,Y_n)$ unabhängig identisch verteilte Zufallsvariablen sind.
Aus den obigen Bemerkungen wird klar, dass bei einer solchen Regressionsschätzung der L_2–Fehler (4) möglichst klein sein sollte, damit man eine Minimierung der mittleren quadratischen Abweichung erreicht.

Gegeben sei also die Stichprobe

$$\mathcal{D}_n = \{\,(X_1,Y_1),\dots,(X_n,Y_n)\,\},$$

dann möchten wir Schätzer m_n konstruieren, abhängig von der Datenmenge $\mathcal{D}_n$

$$m_n(\cdot) = m_n(\cdot,\mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R},$$

sodass der L_2–Fehler

$$\int_{\mathbb{R}^d} |m_n(x)-m(x)|^2\,\mathbf{P}_X(dx)$$

klein ist, denn es gilt äquivalent zu (3)

$$\mathbb{E}\big[\,|m_n(X)-Y|^2\,|\mathcal{D}_n\big] = \mathbb{E}\big[\,|m(X)-Y|^2\,\big] + \int_{\mathbb{R}^d} |m_n(x)-m(x)|^2\,\mathbf{P}_X(dx).$$

Um sinnvolle Schätzer zu konstruieren, fordern wir einige Eigenschaften der Regressionsschätzer.

1.3.2. Konsistenz und Konvergenzrate

Eine der wichtigsten Eigenschaften, die die Regressionsschätzer aufweisen sollten, ist die Konsistenz: Mit wachsendem Stichprobenumfang sollte der Fehler der Schätzung gegen

null konvergieren. Schätzer, die diese Eigenschaft erfüllen, nennt man *konsistent.* Wir messen den Fehler des Regressionsschätzers durch den L_2–Fehler

$$\int |m_n(x) - m(x)|^2 \, \mathbf{P}_X(dx).$$

Der Schätzer m_n hängt dabei von der Stichprobe $\mathcal{D}_n$ ab, sodass der L_2–Fehler eine Zufallsvariable ist. Das führt uns zu einer Definition, die sowohl die Konvergenz in Erwartung, als auch die fast sichere Konvergenz betrachtet:

Definition 1. Eine Folge von Schätzern m_n heißt **schwach universell konsistent**, falls

$$\mathbf{E}\left[\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)\right] \xrightarrow{n\to\infty} 0$$

für jede Verteilung von (X, Y) mit $\mathbf{E}Y^2 < \infty$. Die Folge heißt **stark universell konsistent**, falls

$$\int_{\mathbb{R}^d} |m_n(x) - m(x)|^2 \mathbf{P}_X(dx) \xrightarrow{f.s.} 0$$

für jede Verteilung von (X, Y) mit $\mathbf{E}Y^2 < \infty$.

Wir fordern hier die Konsistenz für beliebige Verteilungen (X, Y), da die Anwendung nichtparametrischer Regressionsschätzer oft aus mangelnden Informationen über die zugrundeliegende Verteilung folgt.
Diese Definition gibt allerdings keinerlei Auskunft darüber, *wie schnell* der L_2–Fehler gegen null konvergiert, was für uns von großem Interesse ist. Theorem 7.2 in [DGL96] zeigt, dass eine solche *universelle* Rate über alle Verteilungen (X, Y) nicht existiert. Um eine nichttriviale Konvergenzrate der Schätzer zu erhalten, muss man die Klasse der Verteilungen einschränken. Diese Einschränkung werden wir im Umfang dieser Arbeit erreichen, indem wir eine Forderung an die Glattheit der Regressionsfunktion $m(x) = \mathbb{E}[Y|X = x]$ stellen (siehe Kapitel 2).

Zunächst wollen wir jedoch definieren, was *optimale Minimax–Konvergenzraten* sind.

Definition 2. Für die Schätzung einer $(p, C)-$glatten Regressionsfunktion m über einer Klasse von Verteilungen $\mathcal{D}$ heißt eine Folge reeller, nicht-negativer Zahlen $(a_n)_{n\in\mathbb{N}}$ eine **untere Minimax-Konvergenzrate**, falls

$$\liminf_{n\to\infty} \inf_{m_n} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbb{E}\left[\int |m_n(x) - m(x)|^2 \, \mathbf{P}_X(dx)\right]}{a_n} = C_1 > 0.$$

Die Folge heißt **obere Minimax-Konvergenzrate**, falls ein Schätzer m_n existiert, sodass

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}} \frac{\mathbb{E}\left[\int |m_n(x) - m(x)|^2 \, \mathbf{P}_X(dx)\right]}{a_n} = C_2 < \infty.$$

Wir nennen die Folge $(a_n)_{n\in\mathbb{N}}$ **optimale Minimax-Konvergenzrate**, falls sie sowohl untere, als auch obere Minimax-Konvergenzrate ist.

Kapitel 2

Einschränkung und Schätzung der Regressionsfunktion

Wie bereits in Kapitel 1.3.2 erwähnt, schränken wir die Klasse der zugrundeliegenden Verteilungen ein, indem wir die folgende Glattheitsbedingung an die Regressionsfunktion m stellen:

Definition 3. Sei $p = q + s$ für ein $q \in \mathbb{N}_0$ und $0 < s \leq 1$. Eine Funktion $m : \mathbb{R}^d \to \mathbb{R}$ heißt $(\mathbf{p}, \mathbf{C}) - \mathbf{glatt}$, falls für jedes $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ mit $\sum_{j=1}^{d} \alpha_j = q$ die partielle Ableitung $\frac{\partial^q m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ existiert und wenn

$$\left| \frac{\partial^q m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(x) - \frac{\partial^q m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(z) \right| \leq C \cdot ||x - z||^s \tag{1}$$

für alle $x, z \in \mathbb{R}^d$, wobei $|| \cdot ||$ die euklidische Norm ist.

Mit dieser Einschränkung zeigte Stone in [Sto82], dass eine optimale Minimax–Konvergenzrate für die Schätzung der Regressionsfunktion von

$$n^{-\frac{2p}{2p+d}} \tag{2}$$

möglich ist. Diese Rate ist allerdings problematisch für hochdimensionale Funktionen.

2.1. Statistische Herausforderungen hochdimensionaler Probleme

Aktuelle technologische Innovationen erlauben es uns massive Mengen an Daten zu sammeln. Es entwickeln sich immer neue Methoden, um diese Daten zu verwerten und neue Erkenntnisse daraus zu gewinnen. Die Verfügbarkeit großer Datenmengen zusammen mit den auftretenden Problemen verändern Datenanalysen und statistisches Denken grundlegend [FL06].

Die Konvergenzrate $n^{-\frac{2p}{2p+d}}$ ist optimal, allerdings beschreibt sie ein Problem der Hochdimensionalität: den sogenannten Fluch der Dimensionen. Falls d im relativen Vergleich zu p groß ist, so kann diese Rate extrem langsam sein. Um diese Problematik zu umgehen und bessere Konvergenzraten zu erhalten, müssen wir weitere strukturelle Anforderungen an die Regressionsfunktion stellen.

Wir werden die Regressionsfunktion m einschränken auf Funktionen, die einem verallgemeinerten hierarchischen Interaktionsmodell genügen, was Kohler und Krzyżak (2016) wie folgt definieren:

Definition 4. Sei $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ und $m : \mathbb{R}^d \to \mathbb{R}$.

(a) Wir sagen, m genügt einem **verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level** 0, falls $a_1, \dots, a_{d^*} \in \mathbb{R}^d$ und $f : \mathbb{R}^{d^*} \to \mathbb{R}$ existieren, sodass

$$m(x) = f(a_1^\top x, \dots, a_{d^*}^\top x) \quad \text{für alle } x \in \mathbb{R}^d.$$

(b) Wir sagen, m genügt einem **verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level $l+1$**, falls für ein $K \in \mathbb{N}$ die Abbildungen

$$\begin{aligned} g_k &: \mathbb{R}^{d^*} \to \mathbb{R} \quad (k = 1, \dots, K) \\ f_{1,k}, \dots, f_{d^*,k} &: \mathbb{R}^d \to \mathbb{R} \quad (k = 1, \dots, K) \end{aligned}$$

existieren, sodass $f_{1,k}, \dots, f_{d^*,k}$ einem verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level l genügen und

$$m(x) = \sum_{k=1}^{K} g_k(f_{1,k}(x), \dots, f_{d^*,k}(x)) \quad \text{für alle } x \in \mathbb{R}^d$$

(c) Wir sagen, dass **verallgemeinerte hierarchische Interaktionsmodell** ist $(\mathbf{p}, \mathbf{C}) - \mathbf{glatt}$, falls alle in obiger Definition vorkommenden Funktionen $(p, C)-$glatt sind nach Definition 3.

Ein solches hierarchisches Modell umfasst weitreichende Funktionenklassen und mit dieser Struktur werden also komplexere Objekte iterativ aus einfacheren konstruiert [Sch17].

2.2. Motivation des trunkierten kleinsten Quadrate Schätzers

Aus Kapitel 1.3 ist uns bekannt, dass die Regressionsfunktion m die mittlere quadratische Abweichung minimiert (vgl. Lemma 1). Also ist die Berechnung von $\mathbb{E}\big[\,|m(X) - Y|^2\,\big]$ äquivalent zum Optimierungsproblem $\min_f \mathbb{E}\big[\,|f(X) - Y|^2\,\big]$.
Dieses ist allerdings nicht lösbar, da die Verteilung von (X, Y) unbekannt ist. Die Idee der kleinste Quadrate Methode ist nun das L_2–Risiko durch ein empirisches L_2–Risiko von unabhängigen Beobachtungen der Zufallsvariable (X, Y) zu schätzen:

$$\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 \tag{3}$$

Als Schätzung der Regressionsfunktion wählt man dann eine Funktion, die dieses empirische L_2–Risiko minimiert. Da dieser Schätzer im Allgemeinen nicht konsistent nach

Definition 1 ist, schränkt man die Klasse der Funktionen f auf einen Raum $\mathcal{F}_n = \mathcal{F}_n(\mathcal{D}_n)$, abhängig von der Stichprobe und dem Stichprobenumfang n, ein:

$$\tilde{m}_n(\cdot) = \operatorname*{argmin}_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 . \tag{4}$$

Wir bemerken, dass wir die Existenz dieses Minimierers voraussetzen, dieser aber nicht zwangsläufig eindeutig sein muss.
Um später den stochastischen Fehlerterm abschätzen zu können, trunkieren wir den Schätzer, nachdem wir ihn berechnet haben. Das heißt, wir nutzen im Folgenden (4), dieser erfüllt also

$$\tilde{m}_n \in \mathcal{F}_n \text{ und } \frac{1}{n} \sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 = \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2$$

und definieren den Schätzer m_n durch die Trunkierung von $\tilde{m}_n$,

$$m_n(x) = T_{\beta_n} \tilde{m}_n(x), \tag{5}$$

wobei β_n abhängig von dem Stichprobemumfang und T_L der folgende Trunkierungsoperator ist:

$$T_L u = \begin{cases} u & \text{falls } |u| \leq L, \\ L \cdot \operatorname{sign}(u) & \text{sonst.} \end{cases} \tag{6}$$

Dieser Schätzer verhält sich also gleich zu einem Schätzer definiert durch die empirische L_2–Risiko–Minimierung über einer Klasse von trunkierten Funktionen

$$T_{\beta_n} \mathcal{F}_n = \{ T_{\beta_n} f : f \in \mathcal{F}_n \}$$

und wird später im Hauptresultat von [BK19] benutzt.

Kapitel 3

Tiefe neuronale Netze in der Regressionsschätzung

Um eine Minimax–Konvergenzrate nach Definition 1 zu erhalten, die unabhängig von der Dimension d der zugrundeliegenden Funktion ist, definieren wir unsere Regressionsschätzer als trunkierte kleinste Quadrate Schätzer nach (5) basierend auf einer spezifischen Klasse neuronaler Netze.
Wir schränken also den Funktionenraum $\mathcal{F}_n$ auf neuronale Netze ein. Das universelle Approximationstheorem rechtfertigt diese Einschränkung. Dieses besagt, dass künstliche 2–schichtige Netze jede stetige Funktion auf einem Kompaktum beliebig gut approximieren können (vgl. [Lu+17]).
Das heißt, der Raum der neuronalen Netze liegt dicht in unserem Funktionenraum $\mathcal{F}_n$.

Die Klasse dieser mehrschichtigen neuronalen feedforward Netze definieren wir anhand der Voraussetzung an die Regressionsfunktion, also übereinstimmend mit den verallgemeinerten hierarchischen Interaktionsmodellen aus Definition 4.

3.1. Klassen hierarchischer neuronaler Netze

Definition 5. Seien $M \in \mathbb{N}$, $N \in \mathbb{N}_0$, $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ und $\alpha > 0$. Dann bezeichnen wir mit $\mathcal{F}_{M,N,d^*,d,\alpha}^{(neuronale\ Netze)}$ die Menge aller Funktionen $f : \mathbb{R}^d \to \mathbb{R}$ der Form

$$f(x) = \sum_{i=1}^{\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d^*} \lambda_{i,l} \cdot \sigma\left(\sum_{m=1}^{d} \theta_{i,l,m} \cdot x^{(m)} + \theta_{i,l,0}\right) + \lambda_{i,0}\right) + \mu_0 \quad (1)$$

mit Gewichten μ_i, $\lambda_{i,l}$, $\theta_{i,l,m} \in \mathbb{R}$, welche

$$|\mu_i| \le \alpha, \quad |\lambda_{i,l}| \le \alpha, \quad |\theta| \le \alpha$$

für alle $i \in \left\{0, 1, \dots, \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*}\right\}$, $l \in \{0, \dots, 4d^*\}$, $m \in \{0, \dots, d\}$ erfüllen.
Unsere **Räume hierarchischer neuronaler Netze** definieren wir für $l = 0$ durch

$$\mathcal{H}^{(0)} = \mathcal{F}_{M,N,d^*,d,\alpha}^{(neuronale\ Netze)} \quad (2)$$

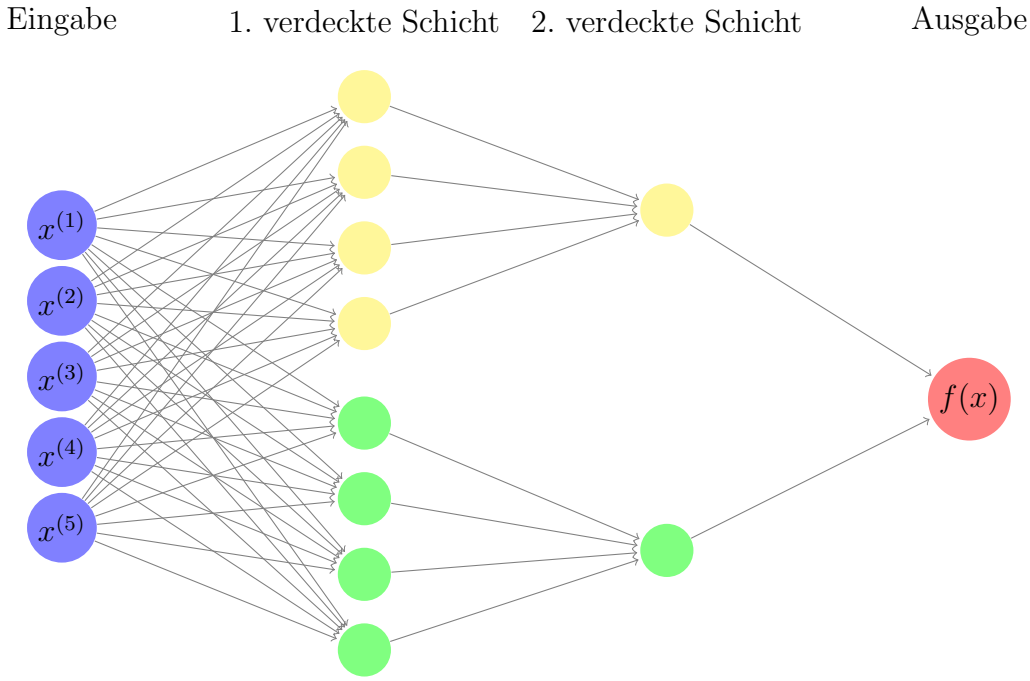

Abbildung 3.1: Ein neuronales Netz $f : \mathbb{R}^5 \to \mathbb{R}$ der Klasse $\mathcal{H}^{(0)} = \mathcal{F}_{1,0,1,5,\alpha}^{(neuronale\ Netze)}$ dargestellt als geordneter Graph, wobei $x = \left(x^{(1)}, \ldots, x^{(5)}\right)$.

und rekursiv für $l > 0$ durch

$$\mathcal{H}^{(l)} = \Bigg\{ h : \mathbb{R}^d \to \mathbb{R} \quad : \quad h(x) = \sum_{k=1}^{K} g_k(f_{1,k}(x), \ldots, f_{d^*,k}(x)) \quad (x \in \mathbb{R}^d) \\ \text{für } g_k \in \mathcal{F}_{M,N,d^*,d^*,\alpha}^{(neuronale\ Netze)} \text{ und } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\}. \tag{3}$$

Die Funktionenklasse $\mathcal{H}^{(0)}$ ist also eine Menge neuronaler Netze mit jeweils zwei verdeckten Schichten. Die Eingabeschicht (bestehend aus d Neuronen) ist mit der ersten verdeckten Schicht vollständig verbunden, also mit $4d^* \cdot \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*}$ Neuronen. Die zweite verdeckte Schicht enthält $\binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*}$ Neuronen, jedes davon ist mit $4d^*$ Neuronen aus der ersten verdeckten Schicht verbunden, während jedes Neuron der ersten mit genau einem Neuron aus der zweiten verdeckten Schicht verbunden ist (vgl. Abbildung 3.1).

Bemerkung 1 (Gewichte der Funktionenklasse)**.** Die Verbindungen zwischen den Neuronen der einzelnen Schichten ergeben, zusammen mit dem Bias, die Anzahl der Gewichte des neuronales Netzes. Das heißt, unsere Klasse hat

$$\begin{aligned}\mathcal{W}\left(\mathcal{F}_{M,N,d^*,d,\alpha}^{(neuronale\ Netze)}\right) &:= \binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}+1\\ &\quad+\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}\cdot(4d^*+1)\\ &\quad+\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}\cdot 4d^*\cdot(d+1)\\ &=\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}\cdot(4d^*\cdot(d+2)+2)+1 \qquad (4)\end{aligned}$$

Gewichte.

Die Funktionen aus dem Raum $\mathcal{H}^{(l)}$ sind für großes l sehr tiefe neuronale Netze, da sie $2\cdot l+2$ versteckte Schichten besitzen. Sei $N(\mathcal{H}^{(l)})$ die Anzahl der verbundenen zweischichtigen Netze aus $\mathcal{F}_{M,N,d^*,d,\alpha}^{(neuronale\ Netze)}$, aus denen sich die Funktionen aus $\mathcal{H}^{(l)}$ zusammensetzen. Dann gilt die Rekursion

$$\begin{aligned}N\left(\mathcal{H}^{(0)}\right)&=1\\ N\left(\mathcal{H}^{(l)}\right)&=K+Kd^*\cdot N\left(\mathcal{H}^{(l-1)}\right)\quad (l\in\mathbb{N})\end{aligned}$$

also

$$\begin{aligned}N\left(\mathcal{H}^{(l)}\right)&=K+Kd^*\cdot[K+Kd^*\cdot[\cdots[K+Kd^*\cdot 1]]]\\ &=K+K^2d^*+K^3d^{*2}+K^4d^{*3}+\ldots+K^ld^{*l-1}+(Kd^*)^l\\ &=\sum_{t=1}^{l}d^{*t-1}K^t+(d^*K)^l.\end{aligned}$$

Zusammenfassend hat also eine Funktion $h\in\mathcal{H}^{(l)}$ maximal

$$N(\mathcal{H}^{(l)})\cdot\mathcal{W}\left(\mathcal{F}_{M,N,d^*,d,\alpha}^{(neuronale\ Netze)}\right) \qquad (5)$$

variable Gewichte.

Im Folgenden werden wir auch die Notation $M^*=\binom{d^*+N}{d^*}\cdot(N+1)\cdot(M+1)^{d^*}$ verwenden.

3.2. Hauptresultat zur Konvergenzrate

Um nun das Hauptresultat zur Konvergenzrate formulieren zu können, benötigen wir noch einige Bedingungen an die *Aktivierungsfunktion*, die allerdings viele gängige Funktionen, wie z.B. die Sigmoid–Funktion (1) erfüllen. Diese fassen wir in der folgenden Definition zusammen:

Definition 6. Eine nicht-fallende und lipschitzstetige Funktion $\sigma:\mathbb{R}\to[0,1]$ heißt **N-zulässig**, wenn die folgenden drei Bedingungen erfüllt sind.

1. Die Funktion σ ist $N+1$ mal stetig differenzierbar mit beschränkten Ableitungen.
2. Es existiert ein Punkt $t_\sigma \in \mathbb{R}$, auf dem alle Ableitungen der Ordnung $\leq N$ von σ ungleich 0 sind.
3. Falls $y > 0$, so gilt $|\sigma(y) - 1| \leq \frac{1}{y}$. Falls $y \leq 0$, dann ist $|\sigma(y)| \leq \frac{1}{|y|}$.

Zusammenfassend schränken wir also die Verteilung von (X, Y) durch die Regressionsfunktion m ein und nutzen kleinste Quadrate Schätzer auf speziellen tiefen neuronalen Netzen, um eine Konvergenzrate unabhängig von der Dimension d zu erhalten. Diese optimale Rate ist

$$n^{-\frac{2p}{2p+d^*}},$$

was der folgende Satz formalisiert.

Satz 1 (Theorem 1 in [BK19]). *Seien $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ unabhängig und identisch verteilte Zuvallsvariablen in $\mathbb{R}^d \times \mathbb{R}$, sodass $\operatorname{supp}(X)$ beschränkt ist und*

$$\mathbb{E}[\exp(c_1 \cdot Y^2)] < \infty \tag{6}$$

für eine Konstante $c_1 > 0$. Sei $m(x) = \mathbb{E}[Y|X = x]$ die korrespondierende Regressionsfunktion, die einem (p, C)–glatten verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und endlichem Level l genügt mit $C > 0$, $p = q + s$ für $q \in \mathbb{N}_0$ und $s \in (0, 1]$. Sei weiterhin $N \in \mathbb{N}_0$ mit $N \geq q$. Alle partiellen Ableitungen der Funktionen $g_k, f_{j,k}$ in Definition 4 (b) der Ordnung $\leq q$ seien beschränkt, d.h. für jeder dieser Funktionen f gilt

$$\max_{\substack{j_1,\ldots,j_d \in \{0,1,\ldots,q\}, \\ j_1+\ldots+j_d \leq q}} \left\| \frac{\partial^{j_1+\ldots+j_d} f}{\partial^{j_1} x^{(1)} \ldots \partial^{j_d} x^{(d)}} \right\|_\infty \leq c_2. \tag{7}$$

Des Weiteren seien alle Funktionen g_k lipschitzstetig mit Lipschitz Konstante $L > 0$. Sei

$$\eta_n = \log(n)^{\frac{2\cdot(N+3)}{N+q+3}} \cdot n^{-\frac{2\cdot(N+1)\cdot)p+2d^*}{2p+d^*}} \tag{8}$$

und definiere den Raum hierarchischer neuronaler Netze $\mathcal{H}^{(l)}$ aus (3) durch die Parameter

$$M = M_n = \left\lceil n^{\frac{1}{2p+d^*}} \right\rceil, \tag{9}$$

$$\alpha = \log(n) \cdot \frac{M_n^{d^*+p\cdot(2N+3)+1}}{\eta_n}, \tag{10}$$

sowie durch K, d, d^* aus der Definition von m. Zudem sei $\sigma : \mathbb{R}^n \to [0, 1]$ N–zulässig wie in Definition 6 und $\tilde{m}_n$ der kleinste Quadrate Schätzer (16). Setze $\beta_n = c_5 \cdot \log(n)$ für $c_5 > 0$ und definiere $m_n = T_{\beta_n} \tilde{m}_n$ durch den Trunkierungsoperator (6).

Dann gilt

$$\mathbb{E}\left[\int |m_n(x) - m(x)|^2 \mathbf{P}_X(dx)\right] \leq c_4 \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}} \tag{11}$$

für hinreichend großes n.

Kapitel 4

Orakel–Ungleichung für den trunkierten kleinsten Quadrate Schätzer

4.1. Vorbereitungen: Konzentrationsungleichungen

Um Satz 1 beweisen zu können, nutzen wir eine Orakel Ungleichung für den L_2–Fehler, für die wir bekannte Ergebnisse aus der Stochastik verwenden, unter anderem Abschätzungen für Überdeckungszahlen, die wir wie folgt definieren.

Definition 7. Sei $\varepsilon > 0$, $A \subset \mathbb{R}^d$ und $\mathcal{G}$ ein Raum von Funktionen $\mathbb{R}^d \longrightarrow \mathbb{R}$.
$g_1, \ldots, g_n$ heißt **ε-Überdeckung** von $\mathcal{G}$ bezüglich $||\cdot||_{\infty,A}$, falls für jedes $g \in \mathcal{G}$ ein $i \in \{1, \ldots, n\}$ existiert, sodass

$$||g - g_i||_{\infty,A} = \sup_{x \in A} |g(x) - g_i(x)| < \varepsilon.$$

Sei $\mathcal{N}(\varepsilon, \mathcal{G}, ||\cdot||_{\infty,A})$ die Kardinalität der kleinsten Menge, die $\mathcal{G}$ ε-überdeckt und setze $\mathcal{N}(\varepsilon, \mathcal{G}, ||\cdot||_{\infty,A}) = \infty$, falls keine endliche Überdeckung existiert. Dann heißt $\mathcal{N}(\varepsilon, \mathcal{G}, ||\cdot||_{\infty,A})$ die **ε-Überdeckungszahl** von $\mathcal{G}$ bezüglich $||\cdot||_{\infty,A}$.

Sei $x_1, \ldots, x_n \in \mathbb{R}^d$ und setze $x_1^n = (x_1, \ldots, x_n)$. Sei $\mathcal{F}$ eine Klasse von Funktionen $f : \mathbb{R}^d \to \mathbb{R}$. Dann ist eine **$L_p$-$\varepsilon$-Überdeckung** von $\mathcal{F}$ über x_1^n eine endliche Menge von Funktionen $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ mit der Eigenschaft

$$\min_{1 \le j \le k} \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - f_j(x_i)|^p \right)^{1/p} < \varepsilon \quad \text{für alle } f \in \mathcal{F}.$$

Die **L_p-ε-Überdeckungszahl** $\mathcal{N}_p(\varepsilon, \mathcal{F}, x_1^n)$ von $\mathcal{F}$ über x_1^n ist die Größe der kleinsten existierenden L_p-ε-Überdeckung.

Für die Überdeckungszahl benötigen später wir folgendes Resultat:

Lemma 2. *Für einen beliebigen Funktionenraum $\mathcal{G}$ und $\delta > 0$ gilt*

$$\mathcal{N}\left(\delta, \left\{ \frac{1}{\beta_n} g : g \in \mathcal{G} \right\}, ||\cdot||_{\infty,\mathrm{supp}(X)}\right) = \mathcal{N}\left(\delta \cdot \beta_n, \mathcal{G}, ||\cdot||_{\infty,\mathrm{supp}(X)}\right).$$

Beweis. Für nicht endliche Überdeckungszahlen ist die Behauptung klar.
Sei also $g_1, \dots, g_N$ eine $\delta \cdot \beta_n-$Überdeckung von $\mathcal{G}$. D.h. für beliebiges $g \in \mathcal{G}$ existiert ein $i \in \{1, \dots, N\}$, sodass

$$\begin{aligned} &\sup_{x \in \operatorname{supp}(X)} |g(x) - g_i(x)| < \delta \cdot \beta_n \\ \iff &\sup_{x \in \operatorname{supp}(X)} \left| \frac{1}{\beta_n} g(x) - \frac{1}{\beta_n} g_i(x) \right| < \delta \end{aligned}$$

Also ist $\frac{1}{\beta_n} g_1, \dots, \frac{1}{\beta_n} g_N$ eine $\delta-$Überdeckung von $\left\{ \frac{1}{\beta_n} g : g \in \mathcal{G} \right\}$. Da dies für beliebige Überdeckungen gilt, folgt die Gleichheit der Überdeckungszahlen. □

Wir formulieren und zeigen zunächst den folgenden Satz aus [Gyö+02], der bewiesen wurde von Lee, Bartlett, und Williamson (1996).

Satz 2 (Theorem 11.4 in [Gyö+02]). *Sei $|Y| \leq B$ fast sicher und $B \geq 1$. Ist $\mathcal{F}$ eine Menge von Funktionen, $f : \mathbb{R}^d \to \mathbb{R}$ mit $|f(x)| \leq B$, dann gilt für jedes $n \geq 1$*

$$\begin{aligned} \mathbb{P}\Bigg(&\exists f \in \mathcal{F} : \mathbb{E}\big[\, |f(X) - Y|^2 \,\big] - \mathbb{E}\big[\, |m(X) - Y|^2 \,\big] \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} \left\{ |f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2 \right\} \\ &\quad \geq \varepsilon \cdot \left(\alpha + \beta + \mathbb{E}\big[\, |f(X) - Y|^2 \,\big] - \mathbb{E}\big[\, |m(X) - Y|^2 \,\big] \right) \Bigg) \\ \leq 14 \sup_{x_1^n} &\mathcal{N}_1 \left(\frac{\beta\varepsilon}{20B}, \mathcal{F}, x_1^n \right) \exp\left(-\frac{\varepsilon^2 (1-\varepsilon) \alpha n}{214(1+\varepsilon) B^4} \right), \end{aligned}$$

wobei $\alpha, \beta > 0$ und $0 < \varepsilon \leq 1/2$.

Für den Beweis von Satz 2 benötigen wir ein Hilfsresultat, sowie die Bernstein–Ungleichung, welche wir im Folgenden nennen wollen. Im Umfang dieser Bachelorarbeit werden wir diese Sätze nicht zeigen, die Beweise können aber in [Gyö+02] betrachtet werden.

Satz 3 (Theorem 11.6 in [Gyö+02]). *Sei $B \geq 1$ und $\mathcal{G}$ eine Klasse von Funktionen $g : \mathbb{R}^d \to [0, B]$. Weiterhin seien $Z, Z_1, \dots, Z_n$ unabhängig identisch verteilte $\mathbb{R}^d$-wertige Zufallsvariablen. Angenommen $\alpha > 0$, $0 < \varepsilon < 1$ und $n \geq 1$. Dann gilt*

$$\begin{aligned} &\mathbb{P}\left(\sup_{g \in \mathcal{G}} \frac{\frac{1}{n} \sum_{i=1}^{n} g(Z_i) - \mathbb{E}[g(Z)]}{\alpha + \frac{1}{n} \sum_{i=1}^{n} g(Z_i) + \mathbb{E}[g(Z)]} > \varepsilon \right) \\ &\leq 4 \cdot \mathbb{E}\left[\mathcal{N}_1\left(\frac{\alpha\varepsilon}{5}, \mathcal{G}, Z_1^n \right) \cdot \exp\left(-\frac{3\varepsilon^2 \alpha n}{40B} \right) \right]. \end{aligned}$$

Beweisskizze. Zunächst ersetzt man den auftretenden Erwartungswert der linken Seite durch das empirische Mittel einer Geisterstichprobe, führt dann zufällige Vorzeichen ein, um im nächsten Schritt Werte der Z_i zu fixieren und die Wahrscheinlichkeit mit Überdeckungen der Funktionenklasse $\mathcal{G}$ weiter abzuschätzen. Zuletzt wendet man die Hoeffding–Ungleichung an und schließt damit den Beweis. □

Satz 4 (Bernstein(1946)). *Seien $a, b \in \mathbb{R}$, $a < b$ und $X_1, \ldots, X_n$ seien unabhängige, reellwertige Zufallsvariablen mit $X_i \in [a, b]$ fast sicher, $i = 1, \ldots, n$. Außerdem gelte*

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}(X_i) > 0.$$

Dann gilt für alle $\varepsilon > 0$

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \right| > \varepsilon \right) \leq 2 \cdot \exp\left(- \frac{n\varepsilon^2}{2\sigma^2 + 2\varepsilon(b-a)/3} \right).$$

Beweis von Satz 2. Zur Vereinfachung setzen wir im Beweis zunächst

$$Z = (X, Y),\ Z_i = (X_i, Y_i),\ i = 1, \ldots, n,$$

und

$$g_f(x, y) = |f(x) - y|^2 - |m(x) - y|^2 .$$

Und wir bemerken, dass für $|f(x)| \leq B,\ |y| \leq B$, und $|m(x)| \leq B$ gilt

$$-4B^2 \leq g_f(x, y) \leq 4B^2. \tag{1}$$

Mit diesen Notationen können wir die Konzentrationsungleichung aus dem Lemma umschreiben als

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \mathbb{E}\big[g_f(Z)\big] - \frac{1}{n} \sum_{i=1}^{n} g_f(Z_i) \geq \varepsilon\big(\alpha + \beta + \mathbb{E}\big[g_f(Z)\big]\big) \right). \tag{2}$$

Nun gliedern wir den Beweis in mehrere Schritte. Zunächst nutzen wir die Tschebyscheff–Ungleichung und symmetrieren durch Einführung einer Geisterstichprobe. Danach ersetzen wir wiederum den Erwartungswert durch ein empirisches Mittel dieser Stichprobe. Im dritten und vierten Schritt führen wir zufällige Vorzeichen ein, um damit Überdeckungen der Funktionenklasse $\{ g_f : f \in \mathcal{F} \}$ über feste Werte $z_1, \ldots, z_n$ zu nutzen. Daraufhin können wir die Bernstein–Ungleichung anwenden und erhalten den exponentiellen Anteil der Ungleichung. Wir schließen den Beweis, indem wir die vorhandenen Überdeckungen auf Überdeckungszahlen der Funktionenenklasse $\mathcal{F}$ über unserer Stichprobe $X_1, \ldots, X_n$ zurückführen.

Schritt 1. Symmetrisierung durch Einführung einer Geisterstichprobe.
Wir ersetzen den Erwartungswert $\mathbb{E}[g_f(Z)]$ auf der linken Seite der Ungleichung in (2) durch das arithmetische Mittel $\frac{1}{n}\sum_{i=1}^{n} g_f(Z'_i)$ der u.i.v Geisterstichprobe $Z'^n_1 = Z'_1, \ldots, Z'_n$, wobei $Z'_1 \sim Z$ und Z'^n_1 unabbhängig von Z^n_1.

Wähle nun eine Funktion $f_n \in \mathcal{F}$ abhängig von der Stichprobe Z^n_1 so, dass

$$\mathbb{E}\big[g_{f_n}(Z) | Z^n_1\big] - \frac{1}{n} \sum_{i=1}^{n} g_{f_n}(Z_i) \geq \varepsilon(\alpha + \beta) + \varepsilon \mathbb{E}\big[g_{f_n}(Z) | Z^n_1\big],$$

falls eine solche Funktion existiert; andernfalls wähle eine beliebige Funktion in $\mathcal{F}$. Mit der Tschebyscheff–Ungleichung erhalten wir

$$\begin{aligned}\mathbb{P}\left(\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i') > \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\middle| Z_1^n\right)\\ \leq \frac{n\cdot\mathbb{V}\mathrm{ar}\big(g_{f_n}(Z)\big|Z_1^n\big)}{n^2\cdot\Big(\frac{\varepsilon}{2}\big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\big)\Big)^2}.\end{aligned} \tag{3}$$

Die Varianz von g_{f_n} ist beschränkt durch ihren Erwartungswert multipliziert mit einer Konstanten, denn

$$\begin{aligned} g_{f_n}(Z) &= \big(f_n(X)-Y+m(X)-Y\big)\big((f_n(X)-Y)-(m(X)-Y)\big)\\ &= \big(f_n(X)+m(X)-2Y\big)\big(f_n(X)-m(X)\big)\end{aligned}$$

und daher zusammen mit den Voraussetzungen des Satzes und (3)

$$\begin{aligned}\mathbb{V}\mathrm{ar}\big(g_{f_n}(Z)\big) &\leq \mathbb{E}\big[g_{f_n}(Z)^2\big] \leq 16B^2\cdot\mathbb{E}\left[|f_n(X)-m(X)|^2\right]\\ &= 16B^2\cdot\big(\mathbb{E}\big[\,|f_n(X)-Y|^2\,\big]-\mathbb{E}\big[\,|m(X)-Y|^2\,\big]\big)\\ &= 16B^2\cdot\mathbb{E}\big[g_{f_n}(Z)\big].\end{aligned}$$

Weiterhin gilt für alle $x\geq 0$ und $a>0$ die Ungleichung

$$f(x)=\frac{x}{(a+x)^2}\leq f(a)=\frac{1}{4a},$$

sodass wir (3) insgesamt abschätzen können durch

$$\frac{16B^2\cdot\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]}{n\cdot\Big(\frac{\varepsilon}{2}\big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\big)\Big)^2} \;\leq\; \frac{16B^2}{\varepsilon^2(\alpha+\beta)n}.$$

Durch diesen Zusammenhang eingesetzt in (3) haben wir also für jedes $n>\frac{128B^2}{\varepsilon^2(\alpha+\beta)}$

$$\mathbb{P}\left(\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]-\frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i')\leq\frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\middle|Z_1^n\right)\;\geq\;\frac{7}{8}. \tag{4}$$

Nun können wir die Wahrscheinlichkeit mit dem eingesetzten empirischen Mittel der Geisterstichprobe mit unserer ursprünglichen Konzentrationsungleichung in Verbindung bringen, denn es folgt

$$\begin{aligned}&\mathbb{P}\left(\exists f\in\mathcal{F}:\frac{1}{n}\sum_{i=1}^n g_f(Z_i')-\frac{1}{n}\sum_{i=1}^n g_f(Z_i)\geq\frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big)\right)\\ &\geq\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i')-\frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i)\geq\frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\right)\\ &\geq\mathbb{P}\Bigg(\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]-\frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i)\geq\varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big),\\ &\qquad\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]-\frac{1}{n}\sum_{i=1}^n g_{f_n}(Z_i')\leq\frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\Bigg).\end{aligned}$$

Die letzte Wahrscheinlichkeit kann man bestimmen, indem man zunächst die zugehörige Wahrscheinlichkeit bedingt auf Z_1^n ermittelt und in einem zweiten Schritt das Ergebnis bezüglich Z_1^n mittelt. Ob das erste Ereignis

$$\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i) \geq \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)$$

eintritt, ist dabei nur von der Stichprobe Z_1^n abhängig. Falls es gilt, ist die obige Wahrscheinlichkeit bedingt auf Z_1^n gleich

$$\mathbb{P}\left(\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i') \leq \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\middle| Z_1^n\right),$$

andernfalls ist sie 0. Damit können wir also die Indikatorfunktion nutzen und erhalten

$$\begin{aligned}
&\mathbb{P}\Bigg(\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i) \geq \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big),\\
&\qquad \mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i') \leq \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\Bigg)\\
&= \mathbb{E}\Bigg[\mathbb{1}\Big\{\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i) \geq \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\Big\}\\
&\qquad \cdot \mathbb{P}\left(\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i') \leq \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\middle| Z_1^n\right)\Bigg]\\
&\overset{(4)}{\geq} \mathbb{E}\left[\mathbb{1}\Big\{\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i) \geq \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\Big\}\cdot\frac{7}{8}\right]\\
&= \frac{7}{8}\cdot\mathbb{P}\left(\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big] - \frac{1}{n}\sum_{i=1}^{n} g_{f_n}(Z_i) \geq \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_{f_n}(Z)|Z_1^n\big]\Big)\right)\\
&= \frac{7}{8}\cdot\mathbb{P}\left(\exists f\in\mathcal{F}: \mathbb{E}\big[g_f(Z)\big] - \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i) \geq \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big)\right),
\end{aligned}$$

da wir f_n so wählten. Insgesamt haben wir nun also für jedes $n > \frac{128B^2}{\varepsilon^2(\alpha+\beta)}$

$$\begin{aligned}
&\mathbb{P}\left(\exists f\in\mathcal{F}: \mathbb{E}\big[g_f(Z)\big] - \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i) \geq \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big)\right)\\
&\leq \frac{8}{7}\cdot\mathbb{P}\left(\exists f\in\mathcal{F}: \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i') - \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i) \geq \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big)\right). \qquad (5)
\end{aligned}$$

Schritt 2. Ersetzen der Erwartung in (5) durch ein empirisches Mittel der Geisterstichprobe.

Im Folgenden möchten wir das Hilfsresultat Satz 3 anwenden. Dazu nutzen wir Subadditivität für Wahrscheinlichkeiten und die identische Verteilung von Z_1^n und $Z_1'^n$ und erhalten für unseren Term in (5)

$$\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \geq \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big)\right) \\
&\leq \mathbb{P}\Bigg(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n g_f(Z_i') - \frac{1}{n}\sum_{i=1}^n g_f(Z_i) \geq \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big), \\
&\qquad \frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \mathbb{E}\big[g_f^2(Z)\big] \leq \varepsilon\Big(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) + \mathbb{E}\big[g_f^2(Z)\big]\Big), \\
&\qquad \frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') - \mathbb{E}\big[g_f^2(Z)\big] \leq \varepsilon\Big(\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i') + \mathbb{E}\big[g_f^2(Z)\big]\Big)\Bigg) \\
&\quad + 2\cdot\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \mathbb{E}[g_f^2(Z)]}{\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) + \mathbb{E}[g_f^2(Z)]} > \varepsilon\right). \qquad (6)
\end{aligned}$$

Auf den letzten Summand können wir nun Satz 3 anwenden und es folgt

$$\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \mathbb{E}[g_f^2(Z)]}{\alpha+\beta+\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) + \mathbb{E}[g_f^2(Z)]} > \varepsilon\right) \\
&\leq 4\cdot\mathbb{E}\left[\mathcal{N}_1\left(\frac{\varepsilon(\alpha+\beta)}{5}, \{\, g_f : f \in \mathcal{F}\,\}, Z_1^n\right)\exp\left(-\frac{3\cdot\varepsilon^2(\alpha+\beta)n}{40\cdot(16B^2)}\right)\right].
\end{aligned}$$

Als nächstes betrachten wir die erste Wahrscheinlichkeit der rechten Seite in (6). Das zweite Ereignis innerhalb dieser Wahrscheinlichkeit impliziert

$$(1+\varepsilon)\cdot\mathbb{E}\big[g_f^2(Z)\big] \geq (1-\varepsilon)\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \varepsilon(\alpha+\beta),$$

was äquivalent ist zu

$$\frac{1}{32B^2}\cdot\mathbb{E}\big[g_f^2(Z)\big] \geq \frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^n g_f^2(Z_i) - \varepsilon\frac{\alpha+\beta}{32B^2(1+\varepsilon)}.$$

Weiterhin ist nach (3) $\mathbb{E}[g_f(Z)] \geq \frac{1}{16B^2}\mathbb{E}[g_f^2(Z)] = 2\frac{1}{32B^2}\mathbb{E}[g_f^2(Z)]$. Diese Argumente können wir gleichermaßen für das dritte Ereignis nutzen, sodass wir den ersten Term in (6) abschätzen können durch

$$\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i') - \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i) \\ \geq \varepsilon(\alpha+\beta)/2 + \frac{\varepsilon}{2}\Bigg(\frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n} g_f^2(Z_i) - \frac{\varepsilon(\alpha+\beta)}{32B^2(1+\varepsilon)} \\ + \frac{1-\varepsilon}{32B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n} g_f^2(Z_i') - \frac{\varepsilon(\alpha+\beta)}{32B^2(1+\varepsilon)}\Bigg)\Bigg).$$

Das zeigt zusammenfassend

$$\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i') - \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i) \geq \frac{\varepsilon}{2}\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big)\Bigg) \\ \leq \mathbb{P}\Bigg(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}\big(g_f(Z_i') - g_f(Z_i)\big) \\ \geq \varepsilon(\alpha+\beta)/2 - \frac{\varepsilon^2(\alpha+\beta)}{32B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(Z_i)+g_f^2(Z_i')\big)\Bigg) \\ + 8\cdot\mathbb{E}\Bigg[\mathcal{N}_1\Bigg(\frac{(\alpha+\beta)\varepsilon}{5}, \{\, g_f : f\in\mathcal{F}\,\}, Z_1^n\Bigg)\exp\Bigg(-\frac{3\varepsilon^2(\alpha+\beta)n}{640B^2}\Bigg)\Bigg]. \tag{7}$$

Schritt 3. Einführung zufälliger Vorzeichen.
Seien $U_1,\ldots,U_n$ gleichverteilte Zufallsvariablen über der Menge $\{-1,1\}$, das heißt

$$\mathbb{P}(U_i=1)=\mathbb{P}(U_i=-1)=\frac{1}{2}, \qquad i=1,\ldots,n$$

und es seien

$$Z_1,\ldots,Z_n,Z_1',\ldots,Z_n',U_1,\ldots,U_n \qquad \text{unabhängig.}$$

Dadurch, dass $Z_1,\ldots,Z_n,Z_1',\ldots,Z_n'$ unabhängig und identisch verteilt sind, ändert sich die gemeinsame Verteilung von $Z_1^n, Z_1'^n$ nicht, wenn man Komponenten von Z_1^n mit den entsprechenden Komponenten von $Z_1'^n$ zufällig vertauscht. Dadurch können wir die Wahrscheinlichkeit auf der rechten Seite der Ungleichung (7) ersetzen durch

$$\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(Z_i') - g_f(Z_i)\big) \\ \geq \frac{\varepsilon}{2}(\alpha+\beta) - \frac{\varepsilon^2(\alpha+\beta)}{32B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(Z_i)+g_f^2(Z_i')\big)\Bigg)$$

und das wiederum kann mit Subadditivität abgeschätzt werden durch

$$\begin{aligned}
&\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(Z_i')\big)\right| \\
&\qquad \geq \frac{1}{2}\Bigg(\varepsilon(\alpha+\beta)/2 - \frac{\varepsilon^2(\alpha+\beta)}{32B^2(1+\varepsilon)}\Bigg) + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(Z_i')\big)\Bigg) \\
&+ \mathbb{P}\Bigg(\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(Z_i)\big)\right| \\
&\qquad \geq \frac{1}{2}\Bigg(\varepsilon(\alpha+\beta)/2 - \frac{\varepsilon^2(\alpha+\beta)}{32B^2(1+\varepsilon)}\Bigg) + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(Z_i)\big)\Bigg) \\
&= 2\cdot\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(Z_i)\big)\right| \\
&\qquad \geq \varepsilon(\alpha+\beta)/4 - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(Z_i)\big)\Bigg). \qquad (8)
\end{aligned}$$

Schritt 4. Bedingen der Wahrscheinlichkeit und Verwenden von Überdeckungen.
Im nächsten Schritt bedingen wir zunächst die rechte Seite von (8) auf die Stichprobe Z_1^n, was gleichbedeutend damit ist, dass wir $z_1, \ldots, z_n$ fest wählen und damit den Term folgendermaßen umschreiben:

$$\begin{aligned}
&\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i)\big)\right| \\
&\qquad \geq \varepsilon(\alpha+\beta)/4 - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(z_i)\big)\Bigg).
\end{aligned}$$

Um die auftretenden Summen abschätzen zu können, nutzen wir nun Überdeckungen der Funktionenklasse. Sei dazu $\delta > 0$ und $\mathcal{G}_\delta$ bezeichne eine L_1–Überdeckung der Menge $\mathcal{G}_\mathcal{F} = \{\, g_f : f \in \mathcal{F} \,\}$ über $z_1, \ldots, z_n$. Weiterhin fixieren wir $f \in \mathcal{F}$. Dann existiert ein $g \in \mathcal{G}_\delta$, sodass

$$\frac{1}{n}\sum_{i=1}^{n} |g(z_i) - g_f(z_i)| < \delta.$$

Ohne Beschränkung der Allgemeinheit können wir annehmen, dass $-4B^2 \leq g(z) \leq 4B^2$.

Daraus folgt mit der Dreiecks–Ungleichung

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i)\big)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i)\big) + \frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i) - g(z_i)\big)\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i)\big)\right| + \frac{1}{n}\sum_{i=1}^{n} |g_f(z_i) - g(z_i)| \\ &< \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i)\big)\right| + \delta.\end{aligned}$$

Außerdem gilt

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} g_f^2(z_i) &= \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) + \frac{1}{n}\sum_{i=1}^{n} \big(g_f^2(z_i) - g^2(z_i)\big) \\ &= \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) + \frac{1}{n}\sum_{i=1}^{n} \big(g_f(z_i) + g(z_i)\big)\big(g_f(z_i) - g(z_i)\big) \\ &\geq \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) - 8B^2\frac{1}{n}\sum_{i=1}^{n} |g_f(z_i) - g(z_i)| \\ &\geq \frac{1}{n}\sum_{i=1}^{n} g^2(z_i) - 8B^2\delta.\end{aligned}$$

Damit erhalten wir

$$\begin{aligned}&\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i)\big)\right| \\ &\qquad \geq \varepsilon(\alpha+\beta)/4 - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(z_i)\big)\Bigg) \\ &\leq \mathbb{P}\Bigg(\exists g \in \mathcal{G}_\delta : \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g(z_i)\big)\right| + \delta \\ &\qquad \geq \varepsilon(\alpha+\beta)/4 - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\left(\frac{1}{n}\sum_{i=1}^{n} g^2(z_i) - 8B^2\delta\right)\Bigg) \\ &\leq |\mathcal{G}_\delta| \max_{g\in\mathcal{G}_\delta} \mathbb{P}\Bigg(\left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g(z_i)\big)\right| \\ &\qquad\qquad \geq \varepsilon(\alpha+\beta)/4 - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} - \delta - \delta\frac{\varepsilon(1-\varepsilon)}{8(1+\varepsilon)} \\ &\qquad\qquad\quad + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g^2(z_i)\big)\Bigg). \qquad (9)\end{aligned}$$

Nun setzen wir $\delta = \varepsilon\beta/5$, denn mit $B \geq 1$ und $0 < \varepsilon < \frac{1}{2}$ gilt

$$\begin{aligned}
&\frac{\varepsilon\beta}{4} - \frac{\varepsilon^2\beta}{64B^2(1+\varepsilon)} - \delta - \delta\frac{\varepsilon(1-\varepsilon)}{8(1+\varepsilon)} \\
&= \frac{\varepsilon\beta}{20} - \frac{\varepsilon^2\beta}{64B^2(1+\varepsilon)} - \frac{\varepsilon^2(1-\varepsilon)\beta}{40(1+\varepsilon)} \\
&\geq 0. \qquad (10)
\end{aligned}$$

Das heißt, mit der $\frac{\varepsilon\beta}{5}$–Überdeckung können wir (10) gleich 0 setzen, und es folgt zusammenfassend

$$\begin{aligned}
&\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g_f(z_i)\big)\right| \\
&\qquad \geq \varepsilon(\alpha+\beta)/4 - \frac{\varepsilon^2(\alpha+\beta)}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n} g_f^2(z_i)\Bigg) \\
&\leq \left|\mathcal{G}_{\frac{\varepsilon\beta}{5}}\right| \max_{g\in\mathcal{G}_{\frac{\varepsilon\beta}{5}}} \mathbb{P}\Bigg(\left|\frac{1}{n}\sum_{i=1}^{n} U_i\big(g(z_i)\big)\right| \\
&\qquad\qquad \geq \frac{\varepsilon\alpha}{4} - \frac{\varepsilon^2\alpha}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n} g^2(z_i)\Bigg). \qquad (11)
\end{aligned}$$

Schritt 5. Anwendung der Bernstein–Ungleichung.
In diesem Schritt wenden wir die Bernstein–Ungleichung an, um die Wahrscheinlichkeit in (11) weiter abzuschätzen, wobei $z_1, \ldots, z_n \in \mathbb{R}^d \times \mathbb{R}$ fest und für $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ gelte weiterhin $-4B^2 \leq g(z) \leq 4B^2$. Dazu stellen wir zunächst einen Zusammenhang zwischen $\frac{1}{n}\sum_{i=1}^{n} g^2(z_i)$ und der Varianz von $U_i(g(z_i))$ her, denn

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}\big(U_i g(z_i)\big) = \frac{1}{n}\sum_{i=1}^{n} g^2(z_i)\mathbb{V}\mathrm{ar}(U_i) = \frac{1}{n}\sum_{i=1}^{n} g^2(z_i).$$

Zur Vereinfachung setzen wir im Folgenden

$$V_i = U_i g(z_i),\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}\big(U_i g(z_i)\big),$$

und

$$A_1 = \frac{\varepsilon\alpha}{4} - \frac{\varepsilon^2\alpha}{64B^2(1+\varepsilon)},\ A_2 = \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)},$$

sodass sich die obige Wahrscheinlichkeit schreiben lässt als

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i\right| \geq A_1 + A_2\sigma^2\right).$$

Wir beachten, dass $V_1, \dots, V_n$ unabhängige Zufallsvariablen sind, die für jedes $i = 1, \dots, n$ $|V_i| \leq |g(z_i)| \leq 4B^2$ erfüllen und dass $A_1, A_2 \geq 0$. Daher können wir die Bernstein–Ungleichung (Satz 4) anwenden und erhalten

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i\right| \geq A_1 + A_2\sigma^2\right) &\leq 2 \cdot \exp\left(-\frac{n(A_1 + A_2\sigma^2)^2}{2\sigma^2 + 2(A_1 + A_2\sigma^2)\frac{8B^2}{3}}\right) \\
&= 2 \cdot \exp\left(-\frac{nA_2^2}{\frac{16}{3}B^2A_2} \cdot \frac{\left(\frac{A_1}{A_2} + \sigma^2\right)^2}{\frac{A_1}{A_2} + \left(1 + \frac{3}{8B^2A_2}\right)\sigma^2}\right) \\
&= 2 \cdot \exp\left(-\frac{3 \cdot nA_2}{16B^2} \cdot \frac{\left(\frac{A_1}{A_2} + \sigma^2\right)^2}{\frac{A_1}{A_2} + \left(1 + \frac{3}{8B^2A_2}\right)\sigma^2}\right). \qquad (12)
\end{aligned}$$

Für $a, b, u > 0$ hat die Funktion $f(u) = \frac{(a+u)^2}{a+b\cdot u}$ ein globales Minimum in $u = \frac{b-2}{b}a$, also gilt stets

$$\frac{(a+u)^2}{a+b\cdot u} \geq \frac{(a+\frac{b-2}{b}a)^2}{a+b\frac{b-2}{b}a} = 4a\frac{b-1}{b^2}. \qquad (13)$$

Das können wir für den Exponenten in (12) nutzen, denn mit $a = A_1/A_2$, $b = \left(1+\frac{3}{8B^2A_2)}\right)$, $u = \sigma^2$ folgt

$$\begin{aligned}
\frac{3 \cdot n \cdot A_2}{16B^2} \cdot \frac{\left(\frac{A_1}{A_2} + \sigma^2\right)^2}{\frac{A_1}{A_2} + \left(1 + \frac{3}{8B^2A_2}\right)\sigma^2} &\overset{(13)}{\geq} \frac{3 \cdot n \cdot A_2}{16B^2} \cdot 4 \cdot \frac{A_1}{A_2}\frac{\frac{3}{8B^2A_2}}{\left(1 + \frac{3}{8B^2A_2}\right)^2} \\
&= 18n \cdot \frac{A_1A_2}{(8B^2A_2 + 3)^2}.
\end{aligned}$$

Bevor wir die Definitionen von A_1 und A_2 wieder einsetzen, bemerken wir zunächst, dass

$$A_1 = \frac{\varepsilon\alpha}{4} - \frac{\varepsilon^2\alpha}{64B^2(1+\varepsilon)} = \frac{\varepsilon\alpha}{4} - \frac{\varepsilon^2\alpha}{64\underbrace{(B^2 + B^2\varepsilon)}_{\geq 1}} \geq \frac{\varepsilon\alpha}{4} - \frac{\varepsilon\alpha}{64} = \frac{15\varepsilon\alpha}{64}.$$

Außerdem hat die Funktion $f(x) = \frac{x(1-x)}{1+x}$ in dem Intervall $[0, 1]$ ein lokales Maximum bei $x^* = \sqrt{2} - 1$ und es gilt $\frac{1}{8}f(x^*) \leq \frac{1}{32}$, sodass wir weiter abschätzen können

$$
\begin{aligned}
18n \cdot \frac{A_1 A_2}{\left(8B^2 A_2 + 3\right)^2} &\geq 18n \cdot \frac{15\varepsilon\alpha}{64} \cdot \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)} \cdot \frac{1}{\left(\frac{\varepsilon(1-\varepsilon)}{8(1+\varepsilon)} + 3\right)^2} \\
&\geq 18n \cdot \frac{15 \cdot \varepsilon^2(1-\varepsilon) \cdot \alpha}{64^2 B^2(1+\varepsilon)} \cdot \frac{1}{\left(\frac{1}{32} + 3\right)^2} \\
&= \frac{9 \cdot 15}{2 \cdot 97 \cdot 97} \cdot \frac{\varepsilon^2(1-\varepsilon)}{1+\varepsilon} \cdot \frac{\alpha \cdot n}{B^2} \\
&\geq \frac{\varepsilon^2(1-\varepsilon) \cdot \alpha \cdot n}{140 \cdot B^2(1+\varepsilon)}.
\end{aligned}
$$

Diese untere Schranke für den Exponenten in (12) eingesetzt ergibt dann insgesamt

$$
\begin{aligned}
&\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} U_i\big(g(z_i)\big) \right| \geq \frac{\varepsilon\alpha}{4} - \frac{\varepsilon^2\alpha}{64B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)} \frac{1}{n} \sum_{i=1}^{n} g^2(z_i) \right) \\
&\qquad \leq 2 \cdot \exp\left(\frac{\varepsilon^2(1-\varepsilon) \cdot \alpha \cdot n}{140 \cdot B^2(1+\varepsilon)} \right). \qquad (14)
\end{aligned}
$$

Schritt 6. Abschätzen der Überdeckungszahl.
Im Folgenden konstruieren wir eine L_1–Überdeckung der Menge $\{ g_f : f \in \mathcal{F} \}$ über $z_1, \ldots, z_n$. Sei dazu zunächst $f_1, \ldots, f_l$, $l = \mathcal{N}_1\big(\frac{\varepsilon\beta}{20B}, \mathcal{F}, x_1^n\big)$ eine $\frac{\varepsilon\beta}{20B}$–Überdeckung von $\mathcal{F}$ über x_1^n. Ohne Einschränkung gelte $|f_j(x)| \leq B$ für alle $j \in \{ 1, \ldots, l \}$. Für ein beliebiges $f \in \mathcal{F}$ existiert dann ein f_j, sodass

$$
\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - f_j(x_i)| < \frac{\varepsilon\beta}{20B}.
$$

Daraus folgt

$$
\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} |g_f(z_i) - g_f(z_i)| \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \left| |f(x_i) - y_i|^2 - |m(x_i) - y_i|^2 - |f_j(x_i) - y_i|^2 + |m(x_i) - y_i|^2 \right| \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - y_i + f_j(x_i) - y_i| \, |f(x_i) - y_i - f_j(x_i) + y_i| \\
&\quad \leq 4B \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - f_j(x_i)| < \frac{\varepsilon\beta}{5}.
\end{aligned}
$$

Das heißt $g_{f_1}, \ldots, g_{f_l}$ ist eine $\frac{\varepsilon\beta}{5}$–Überdeckung von $\{ g_f : f \in \mathcal{F} \}$ über z_1^n der Größe $\mathcal{N}_1\big(\frac{\varepsilon\beta}{20B}, \mathcal{F}, x_1^n\big)$.

Daher folgt nun aus den Schritten 3 bis 6, dass

$$\begin{aligned}
&\mathbb{P}\Bigg(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}\big(g_f(Z_i') - g_f(Z_i)\big) \\
&\qquad \geq \varepsilon(\alpha+\beta)/2 - \frac{\varepsilon^2(\alpha+\beta)}{32B^2(1+\varepsilon)} + \frac{\varepsilon(1-\varepsilon)}{64B^2(1+\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\big(g_f^2(Z_i) + g_f^2(Z_i')\big)\Bigg) \\
&\leq 4 \sup_{x_1^n \in (\mathbb{R}^d)^n} \mathcal{N}_1\left(\frac{\varepsilon\beta}{20B}, \mathcal{F}, x_1^n\right) \cdot \exp\left(\frac{\varepsilon^2(1-\varepsilon)\alpha n}{140 \cdot B^2(1+\varepsilon)}\right). \qquad (15)
\end{aligned}$$

Schritt 7. Abschluss des Beweises.
Mit den Schritten 1, 2 und 6 können wir den Beweis vervollständigen, denn für $n > \frac{128B^2}{\varepsilon(\alpha+\beta)}$ folgt

$$\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{F} : \mathbb{E}\big[g_f(Z)\big] - \frac{1}{n}\sum_{i=1}^{n} g_f(Z_i) > \varepsilon\Big(\alpha+\beta+\mathbb{E}\big[g_f(Z)\big]\Big)\right) \\
&\leq \frac{32}{7}\sup_{x_1^n}\mathcal{N}_1\left(\frac{\varepsilon\beta}{20B}, \mathcal{F}, x_1^n\right) \cdot \exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{140B^2(1+\varepsilon)}\right) \\
&\quad + \frac{64}{7}\sup_{x_1^n}\mathcal{N}_1\left(\frac{\varepsilon(\alpha+\beta)}{20B}, \mathcal{F}, x_1^n\right) \cdot \exp\left(-\frac{3\varepsilon^2(\alpha+\beta)n}{640B^4}\right) \\
&\leq 14\sup_{x_1^n}\mathcal{N}_1\left(\frac{\varepsilon\beta}{20B}, \mathcal{F}, x_1^n\right) \cdot \exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)B^4}\right).
\end{aligned}$$

Für $n \leq \frac{128B^2}{\varepsilon^2(\alpha+\beta)}$ gilt für den Exponent

$$\exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)B^4}\right) \geq \exp\left(-\frac{128}{214}\right) \geq \frac{1}{14},$$

was die Behauptung zeigt. □

4.2. Orakel–Ungleichung und Abschätzen der Überdeckungszahl

Nun können wir eine Art Orakel–Ungleichung für den L_2–Fehler unseres trunkierten Schätzers formulieren und beweisen. Eine Orakel–Ungleichung im herkömmlichen Sinn vergleicht die Performance eines realen Schätzers mit der eines idealen Schätzers, welcher auf perfekter Information, gegeben durch ein Orakel, beruht [Can06].

Lemma 3 (Lemma 1 in [BK19]). *Sei $\beta_n = c_5 \cdot \log(n)$ mit $c_5 > 0$ konstant. Angenommen, die Verteilung von (X, Y) erfülle*

$$\mathbb{E}\left[e^{c_6 \cdot |Y|^2}\right] < \infty$$

für eine Konstante $c_6 > 0$ und, dass die Regressionsfunktion $m(x) = \mathbb{E}[Y|X = x]$ beschränkt sei. Weiterhin sei $\tilde{m}_n$ der kleinste Quadrate Schätzer

$$\tilde{m}_n = \underset{f \in \mathcal{F}_n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 \tag{16}$$

für einen Funktionenraum $\mathcal{F}_n$ und definiere $m_n = T_{\beta_n} \tilde{m}_n$ durch den Trunkierungsoperator (6). Dann gilt

$$\mathbb{E}\left[\int |m_n(x) - m(x)|^2 \, \mu(dx)\right] \leq \frac{c_7 \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, ||\cdot||_{\infty, \operatorname{supp}(X)}\right)\right)}{n} + 2 \cdot \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \, \mathbf{P}_X(dx)$$

für alle $n > 1$ und eine Konstante $c_7 > 0$, welche unabhängig von n, β_n und allen Parametern des Schätzers ist.

Beweis. Sei zunächst m_{β_n} die Regressionsfunktion der trunkierten Zufallsvariable $T_{\beta_n} Y$, das heißt

$$m_{\beta_n}(x) = \mathbb{E}[T_{\beta_n} Y | X = x].$$

Mit (3) zerlegen wir den L_2-Fehler wie folgt:

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \, \mu(dx) \\
&\quad = \Big\{ \mathbb{E}\Big[|m_n(X) - Y|^2 \, |D_n \Big] - \mathbb{E}\Big[|m(X) - Y|^2 \Big] \\
&\qquad - \mathbb{E}\Big[|m_n(X) - T_{\beta_n} Y|^2 \, |D_n \Big] + \mathbb{E}\Big[|m_{\beta_n}(X) - T_{\beta_n} Y|^2 \Big] \Big\} \\
&\qquad + \Big\{ \mathbb{E}\Big[|m_n(X) - T_{\beta_n} Y|^2 \, |D_n \Big] - \mathbb{E}\Big[|m_{\beta_n}(X) - T_{\beta_n} Y|^2 \Big] \\
&\qquad - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \Big(|m_n(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X) - T_{\beta_n} Y_i|^2 \Big) \Big\} \\
&\qquad + \Big\{ 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - T_{\beta_n} Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \\
&\qquad - \left(2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right) \Big\} \\
&\qquad + \left\{ 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} |m(X_i) - Y_i|^2 \right\} \\
&\quad = \sum_{i=1}^{4} \{S_{i,n}\} .
\end{aligned}$$

Im Folgenden schätzen wir die Summanden $S_{i,n}$ nacheinander ab, $i = 1, \dots, 4$. Dabei sind die Terme $S_{1,n}$ und $S_{3,n}$ Trunkierungsfehler, einmal in Erwartung und einmal im Mittel über der Stichprobe D_n. $S_{2,n}$ ist der empirische Prozessfehler oder auch stochastische Fehlerterm, der den Erwartungswert mit dem arithmetischen Mittel vergleicht und durch $S_{4,n}$ erhalten wir den zentralen Approximationsfehler dadurch, dass die Trunkierung vernachlässigbar wird.

Erster Summand $S_{1,n}$: Trunkierungsfehler.
Zunächst betrachten wir $S_{1,n}$. Mit $a^2 - b^2 = (a-b)(a+b)$ erhalten wir

$$\begin{aligned}
S_{1,n} &= \mathbb{E}\Big[\, |m_n(X) - Y|^2 - |m_n(X) - T_{\beta_n} Y|^2 \,|D_n\Big] \\
&\quad - \mathbb{E}\Big[\, |m(X) - Y|^2 + |m_{\beta_n}(X) - T_{\beta_n} Y|^2 \,\Big] \\
&= \mathbb{E}\Big[(T_{\beta_n} Y - Y)(2m_n(X) - Y - T_{\beta_n} Y)|D_n\Big] \\
&\quad - \mathbb{E}\Big[\Big((m(X) - m_{\beta_n}(X)) + (T_{\beta_n} Y - Y)\Big)\Big(m(X) + m_{\beta_n}(X) - Y - T_{\beta_n} Y\Big)\Big] \\
&= S_{5,n} + S_{6,n}.
\end{aligned}$$

Nun gilt $|T_{\beta_n} Y - Y|^2 \le |Y|^2 \cdot \mathbb{1}_{\{\,|Y| > \beta_n\,\}}$, denn für $|Y| \le \beta_n$ ist $|T_{\beta_n} Y - Y|^2 = 0$ und für $|Y| > \beta_n$ gilt $|T_{\beta_n} Y - Y|^2 = |\beta_n - Y|^2 \le |Y|^2$. Das nutzen wir für $S_{5,n}$ und mit $(a+b)^2 \le 2a^2 + 2b^2$, der Cauchy-Schwartz Ungleichung und

$$\mathbb{1}_{\{\,|Y| > \beta_n\,\}} \le \frac{\exp(c_8/2 \cdot |Y|^2)}{\exp(c_8/2 \cdot \beta_n^2)} \tag{17}$$

erhalten wir

$$\begin{aligned}
|S_{5,n}| &\le \sqrt{\mathbb{E}\big[\, |T_{\beta_n} Y - Y|^2 \,\big]} \cdot \sqrt{\mathbb{E}\big[\, |2m_n(X) - Y - T_{\beta_n} Y|^2 \,|D_n\big]} \\
&\le \sqrt{\mathbb{E}\big[\, |Y|^2 \cdot \mathbb{1}_{\{\,|Y| > \beta_n\,\}}\big]} \cdot \sqrt{\mathbb{E}\big[2 \cdot |2m_n(X) - T_{\beta_n} Y|^2 + 2 \cdot |Y|^2 \,|D_n\big]} \\
&\le \sqrt{\mathbb{E}\left[\, |Y|^2 \cdot \frac{\exp(c_8/2 \cdot |Y|^2)}{\exp(c_8/2 \cdot \beta_n^2)}\right]} \cdot \sqrt{\mathbb{E}\big[2 \cdot |2m_n(X) - T_{\beta_n} Y|^2 \,|D_n\big] + 2 \cdot \mathbb{E}\big[\, |Y|^2 \,\big]} \\
&\le \sqrt{\mathbb{E}\Big[\, |Y|^2 \cdot \exp(c_8/2 \cdot |Y|^2)\Big]} \cdot \exp\left(-\frac{c_8 \beta_n}{4}\right) \cdot \sqrt{2 \cdot (3\beta_n)^2 + 2\mathbb{E}\big[\, |Y|^2 \,\big]},
\end{aligned}$$

da $|m_n| \le \beta_n$ und $|T_{\beta_n} Y| \le \beta_n$.

Nun gilt $x \le \exp(x)$ für $x \in \mathbb{R}$, also

$$|Y|^2 \le \frac{2}{c_8} \exp\left(\frac{c_8}{2} \cdot |Y|^2\right),$$

sodass wir den ersten Faktor abschätzen können durch

$$\sqrt{\mathbb{E}\Big[\,|Y|^2\cdot\exp(c_1/2\cdot|Y|^2)\Big]}$$
$$\leq \mathbb{E}\left[\frac{2}{c_8}\exp\left(\frac{c_8}{2}\cdot|Y|^2\right)\cdot\exp\left(\frac{c_8}{2}\cdot|Y|^2\right)\right]\leq\mathbb{E}\left[\frac{2}{c_8}\exp(c_8\cdot|Y|^2)\right]\leq c_9<\infty$$

nach Voraussetzung aus dem Lemma. Mit $x\leq\exp(x)$ für $x\in\mathbb{R}$ und der Linearität des Erwartungswertes erhalten wir ebenso

$$\mathbb{E}\left[|Y|^2\right]\leq\mathbb{E}\left[1/c_8\cdot\exp(x_1\cdot|Y|^2)\right]\leq c_{10}<\infty, \tag{18}$$

sodass wir den dritten Term durch $\sqrt{18\beta_n^2+c_{10}}$ abschätzen können. Nun ist $\beta_n=c_5\cdot\log(n)$, also $\exp(-\frac{c_5\beta_n}{4})=n^{-c_{13}/4}$ und für Konstanten $c_{11},c_{12}>0$ folgt, dass

$$|S_{5,n}|\leq\sqrt{38}\cdot\exp\left(-c_{11}\cdot\log(n)^2\right)\cdot\sqrt{(18c_{14}\cdot\log(n)^2+c_{10}}\leq c_{12}\cdot\frac{\log(n)}{n}.$$

Für den Summand $S_{6,n}$ erhalten wir wieder mit der Cauchy-Schwartz Ungleichung und $(a+b)^2\leq 2a^2+2b^2$

$$|S_{6,n}|=\mathbb{E}\left[\Big((m(X)-m_{\beta_n}(X))+(T_{\beta_n}Y-Y)\Big)\Big(m(X)+m_{\beta_n}(X)-Y-T_{\beta_n}Y\Big)\right]$$

$$\leq\sqrt{2\cdot\mathbb{E}\Big[\,|m(X)-m_{\beta_n}(X)|^2\,\Big]+2\cdot\mathbb{E}\Big[\,|T_{\beta_n}Y-Y|^2\,\Big]} \tag{19}$$

$$\cdot\sqrt{\mathbb{E}\Big[\,|m(X)+m_{\beta_n}(X)-Y-T_{\beta_n}Y|^2\,\Big]}. \tag{20}$$

Für den ersten Term (19) benutzen wir die Linearität der bedingten Erwartung und es folgt

$$\begin{aligned}\mathbb{E}\Big[\,|m(X)-m_{\beta_n}(X)|^2\,\Big]&=\mathbb{E}\Big[\,|\mathbb{E}\left[Y|X\right]-\mathbb{E}\left[T_{\beta_n}Y|X\right]|^2\,\Big]\\&=\mathbb{E}\Big[\,|\mathbb{E}\left[Y-T_{\beta_n}Y|X\right]|^2\,\Big]\\&\overset{\text{Jensen}}{\leq}\mathbb{E}\Big[\,|Y-T_{\beta_n}Y|^2\,\Big],\end{aligned}$$

wobei wir in der letzten Zeile die Jensen-Ungleichung für bedingte Erwartungswerte, sowie die Tower Rule genutzt haben. Den zweiten Term (20) behandeln wir wie den zweiten Faktor in $S_{5,n}$:

$$\sqrt{\mathbb{E}\Big[\,|m(X)+m_{\beta_n}(X)-Y-T_{\beta_n}Y|^2\,\Big]}\leq\sqrt{\mathbb{E}\Big[2\cdot|m(X)+m_{\beta_n}(X)-T_{\beta_n}Y|^2\,\Big]+2\mathbb{E}\Big[\,|Y|^2\,\Big]}$$
$$\leq c_{15}\cdot\log(n)$$

für eine Konstante $c_{15} > 0$, da nach Voraussetzung die Regressionsfunktion m beschränkt ist und $|m_{\beta_n}| \leq \beta_n = c_5 \cdot \log(n)$. Daher gilt für $c_{15} > 0$

$$S_{6,n} \leq \sqrt{4 \cdot \mathbb{E}\left[\, |Y - T_{\beta_n} Y|^2 \,\right]} \cdot c_{15} \cdot \log(n) \leq c_{16} \cdot \log(n)/n$$

mit den selben Argumenten wie für $S_{5,n}$.
Insgesamt erhalten wir also

$$S_{1,n} \leq c_{17} \cdot \frac{\log(n)}{n}$$

für eine Konstante $c_{17} > 0$.

Dritter Summand $S_{3,n}$: Trunkierungsfehler.
Während $S_{1,n}$ ein Term aus Erwartungswerten ist, enthält $S_{3,n}$ die zugehörigen Mittelwerte über der Stichprobe $\mathcal{D}_n = \{\, (X_1, Y_1), \ldots, (X_n, Y_n) \,\}$. Daher sind für die Abschätzung alle obigen Argumente anwendbar, und wir erhalten analog zu $S_{1,n}$

$$\mathbb{E}[S_{3,n}] \leq c_{18} \cdot \frac{\log(n)}{n}$$

für eine hinreichend große Konstante $c_{18} > 0$.

Zweiter Summand $S_{2,n}$: Stochastischer Fehlerterm.
Nun schätzen wir $S_{2,n}$ ab. Dazu definieren wir zunächst die Menge $T_{\beta_n, \mathrm{supp}(X)} \mathcal{F}_n = \left\{\, T_{\beta_n} f \cdot \mathbb{1}_{\mathrm{supp}(X)} : f \in \mathcal{F}_n \,\right\}$. Unser trunkierter kleinste Quadrate Schätzer m_n liegt offensichtlich in dieser Menge und wir betrachten für $t > 0$ die Konzentrationsungleichung

$$\begin{aligned}
\mathbb{P}(S_{2,n} > t) = \mathbb{P}\Bigg(& \mathbb{E}\left[\, |m_n(X) - T_{\beta_n} Y|^2 \,|D_n\right] - \mathbb{E}\left[\, |m_{\beta_n}(X) - T_{\beta_n} Y|^2 \,\right] \\
& - 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \left(|m_n(X_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(X_i) - T_{\beta_n} Y_i|^2 \right) > t \Bigg) \\
\leq \mathbb{P}\Bigg(& \exists f \in T_{\beta_n, \mathrm{supp}(X)} \mathcal{F}_n : \mathbb{E}\left[\left| \frac{f(X) - T_{\beta_n} Y}{\beta_n} \right|^2 \right] - \mathbb{E}\left[\left| \frac{m_{\beta_n}(X) - T_{\beta_n} Y}{\beta_n} \right|^2 \right] \\
& - \frac{1}{n} \sum_{i=1}^{n} \left(\left| \frac{f(X_i) - T_{\beta_n} Y_i}{\beta_n} \right|^2 - \left| \frac{m_{\beta_n}(X_i) - T_{\beta_n} Y_i}{\beta_n} \right|^2 \right) \\
& > \frac{1}{2} \left(\frac{t}{\beta_n^2} + \mathbb{E}\left[\left| \frac{f(X) - T_{\beta_n} Y}{\beta_n} \right|^2 \right] - \mathbb{E}\left[\left| \frac{m_{\beta_n}(X) - T_{\beta_n} Y}{\beta_n} \right|^2 \right] \right) \Bigg)
\end{aligned}$$

Im nächsten Schritt können wir Satz 2, denn mit $B = 1$, $\varepsilon = \frac{1}{2}$, $\alpha = \beta = \frac{1}{2} \frac{t}{\beta_n^2}$ sind die Voraussetzungen $\left| \frac{f(X)}{\beta_n} \right| \leq B$, $\left| \frac{T_{\beta_n} Y}{\beta_n} \right| \leq B$ erfüllt, also

$$\mathbb{P}(S_{2,n} > t) \overset{\text{Th. 11.4}}{\leq} 14 \sup_{x_1^n} \mathcal{N}_1 \left(\frac{t}{80 \cdot \beta_n^2}, T_{\beta_n, \mathrm{supp}(X)} \mathcal{F}_n, x_1^n \right) \cdot \exp\left(- \frac{n}{5136 \cdot \beta_n^2} t \right).$$

Mit Lemma 2 erhalten wir weiterhin

$$\mathbb{P}(S_{2,n} > t) \leq 14 \cdot \mathcal{N}\left(\frac{t}{80 \cdot \beta_n}, \mathcal{F}_n, ||\cdot||_{\infty,\mathrm{supp}(X)}\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} t\right).$$

Da die Überdeckungszahl und der exponentielle Faktor fallend in t sind, folgt für $\varepsilon_n \geq \frac{80}{n}$

$$\begin{aligned}\mathbb{E}[S_{2,n}] &= \int_0^\infty \mathbb{P}(S_{2,n} > t)dt \leq \varepsilon_n + \int_{\varepsilon_n}^\infty \mathbb{P}(S_{2,n} > t)dt \\ &\leq \varepsilon_n + 14 \cdot \mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, ||\cdot||_{\infty,\mathrm{supp}(X)}\right) \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2}\varepsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}.\end{aligned}$$

Wir minimieren die rechte Seite, indem wir

$$\varepsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot \log\left(14 \cdot \mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, ||\cdot||_{\infty,\mathrm{supp}(X)}\right)\right)$$

setzen, welche die Bedingung $\varepsilon_n \geq \frac{80}{n}$ erfüllt, solange die Konstante c_5 in $\beta_n = c_5 \cdot \log(n)$ hinreichend groß ist. Daraus folgt direkt

$$\mathbb{E}[S_{2,n}] \leq \frac{c_{19} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, ||\cdot||_{\infty,\mathrm{supp}(X)}\right)\right)}{n}.$$

Sodass wir zusammen bereits

$$\mathbb{E}\left[\sum_{i=1}^3 S_{i,n}\right] \leq \frac{c_{20} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n \cdot \beta_n}, \mathcal{F}_n, ||\cdot||_{\infty,\mathrm{supp}(X)}\right)\right)}{n}$$

für eine ausreichend große Konstante $c_{20} > 0$ haben.

Vierter Summand $S_{4,n}$: Approximationsterm.
Damit fehlt nur noch der Approximationsfehler in unserer Orakel-Ungleichung. Dafür betrachten wir zuletzt $S_{4,n}$. Sei dafür zunächst A_n das Ereignis, dass ein $i \in \{1, \ldots, n\}$ existiert, sodass $|Y_i| > \beta_n$. $\mathbb{1}_{A_n}$ sei die Indikatorfunktion von A_n. Da die (X_i, Y_i), $i = 1, \ldots, n$ unabhängig identisch verteilt sind, ist $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |m_n(X_i) - Y_i|^2 \cdot \mathbb{1}_{A_n}\right] = \mathbb{E}\left[|m_n(X_1) - Y_1|^2 \cdot \mathbb{1}_{A_n}\right]$.
Somit zerlegen wir $\mathbb{E}[S_{4,n}]$ in zwei weitere Summanden:

$$\begin{aligned}
\mathbb{E}[S_{4,n}] &= 2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&= 2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot \mathbb{1}_{A_n}\right] \\
&\quad + 2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot \mathbb{1}_{A_n^c} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&= 2 \cdot \mathbb{E}\big[\, |m_n(X_1) - Y_1|^2 \cdot \mathbb{1}_{A_n}\big] \\
&\quad + 2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot \mathbb{1}_{A_n^c} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&= S_{7,n} + S_{8,n}.
\end{aligned}$$

Für den ersten Summand nutzen wir die Cauchy-Schwartz Ungleichung und es folgt

$$\begin{aligned}
\frac{1}{2} \cdot S_{7,n} &\leq \sqrt{\mathbb{E}\Big[\left(|m_n(X_1) - Y_1|^2\right)^2\Big]} \cdot \sqrt{\mathbb{P}(A_n)} \\
&\leq \sqrt{\mathbb{E}\Big[\left(2\,|m_n(X_1)|^2 + 2\,|Y_1|^2\right)^2\Big]} \cdot \sqrt{n \cdot \mathbb{P}(|Y_1| > \beta_n)} \\
&\leq \sqrt{\mathbb{E}\Big[8\,|m_n(X_1)|^4 + 8\,|Y_1|^4\Big]} \cdot \sqrt{n \cdot \frac{\mathbb{E}\big[\exp\left(c_1 \cdot |Y_1|^2\right)\big]}{\exp(c_1 \cdot \beta_n^2)}}.
\end{aligned}$$

Mit $x \leq \exp(x)$ für $x \in \mathbb{R}$ erhalten wir

$$\begin{aligned}
\mathbb{E}\big[\,|Y|^4\,\big] &= \mathbb{E}\big[\,|Y|^2 \cdot |Y|^2\,\big] \leq \mathbb{E}\left[\frac{2}{c_1} \cdot \exp(\frac{c_1}{2} \cdot |Y|^2) \cdot \frac{2}{c_1}\exp(\frac{c_1}{2} \cdot |Y|^2)\right] \\
&= \frac{4}{c_1^2} \cdot \mathbb{E}\big[\exp(c_1\,|Y|^2)\big] < \infty
\end{aligned}$$

nach Voraussetzung aus dem Lemma. Außerdem ist $||m_n||_\infty$ beschränkt durch β_n, sodass wir den ersten Faktor mit $c_{22} > 0$ durch

$$c_{21} \cdot \beta_n^2 = c_{22} \cdot \log(n)^2$$

abschätzen können. Mit $\mathbb{E}\big[\exp(c_1 \cdot |Y|^2)\big] \leq c_{23} < \infty$ für eine Konstante $c_{23} > 0$, erhalten wir für den zweiten Faktor

$$\sqrt{n \cdot \frac{\mathbb{E}\big[\exp\left(c_1 \cdot |Y_1|^2\right)\big]}{\exp(c_1 \cdot \beta_n^2)}} \leq \sqrt{n} \cdot \frac{\sqrt{c_{23}}}{\sqrt{\exp(c_1 \cdot \beta_n^2)}} \leq \frac{\sqrt{n} \cdot \sqrt{c_{23}}}{\exp((c_1 \cdot c_{14}^2 \cdot \log(n)^2)/2}.$$

Nun gilt $\exp(-c \cdot \log(n)^2) = \mathcal{O}(n^{-2})\ \forall c > 0$, sodass insgesamt die Abschätzung

$$S_{7,n} \leq c_{24} \cdot \frac{\log(n)^2\sqrt{n}}{n^2} \leq c_{25} \cdot \frac{\log(n)^2}{n}$$

folgt.

$|y| \leq \beta$ impliziert $|T_\beta z - y| \leq |z - y|$. Auf dem Ereignis $A_n^\complement$ ist $|Y_i| \leq \beta_n \ \forall i \in \{1, \ldots, n\}$, sodass wir auf $A_n^\complement$ unseren trunkierten Schätzer durch den Kleinste-Quadrate Schätzer $\tilde{m}_n$ ersetzen können und damit

$$\begin{aligned}
S_{8,n} &= 2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \cdot \mathbb{1}_{A_n^\complement} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&\leq 2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 \cdot \mathbb{1}_{A_n^\complement} - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&\leq 2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |\tilde{m}_n(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&\leq 2 \cdot \mathbb{E}\left[\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right],
\end{aligned}$$

wobei wir außerdem die Definition des kleinste Quadrate Schätzers $\tilde{m}_n$, sowie Eigenschaften der Indikatorfunktion genutzt haben.

Sei weiterhin $f^* \in \mathcal{F}_n$, sodass

$$\int |f^*(x) - m(x)|^2 \, \mathrm{P}_X(dx) \leq \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \, \mathbf{P}_X(dx) + \frac{1}{n}.$$

Damit ist

$$\begin{aligned}
&\mathbb{E}\left[\inf_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |f^*(X_i) - Y_i|^2 - \frac{1}{n}\sum_{i=1}^{n} |m(X_i) - Y_i|^2\right] \\
&= \mathbb{E}\left[|f^*(X) - Y|^2\right] - \mathbb{E}\left[|m(X) - Y|^2\right] \\
&= \mathbb{E}\left[|f^*(X) - m(X)|^2\right] + \mathbb{E}\left[|m(X) - Y|^2\right] - \mathbb{E}\left[|m(X) - Y|^2\right] \\
&\leq \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \, \mathbf{P}_X(dx) + \frac{1}{n}.
\end{aligned}$$

Dabei benutzten wir die Linearität des Erwartungswertes, sowie die Dreiecksungleichung mit der minimierenden Eigenschaft der Regressionsfunktion m (vgl. Lemma 1).

Die obigen Abschätzungen zusammengefasst ergeben

$$\begin{aligned}
\mathbb{E}\left[\int |m_n(x) - m(x)|^2 \mu(dx)\right] &= \mathbb{E}\left[\sum_{i=1}^{4}\{S_{i,n}\}\right] \\
&\leq \frac{c_{20} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}_n, ||\cdot||_{\infty,\mathrm{supp}(X)}\right)\right)}{n} \\
&\quad + c_{25} \cdot \frac{\log(n)^2}{n} + 2 \cdot \inf_{f\in\mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx) + \frac{1}{n} \\
&\leq \frac{c_{29} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n\cdot\beta_n}, \mathcal{F}_n, ||\cdot||_{\infty,\mathrm{supp}(X)}\right)\right)}{n} \\
&\quad + 2 \cdot \inf_{f\in\mathcal{F}_n} \int |f(x) - m(x)|^2 \mathbf{P}_X(dx),
\end{aligned}$$

für eine Konstante $c_{29} > 0$, was den Beweis vervollständigt. □

Um den ersten Summand in Lemma 3 abzuschätzen, nutzen wir das folgende Lemma über die Überdeckungszahl.

Lemma 4 (Lemma 2 in [BK19]). *Es gelten die Annahmen aus Satz 1. Sei* $\varepsilon_n \geq \frac{1}{n^{c_{26}}}$ *und* $\frac{M_n}{\eta_n} \leq n^{c_{27}}$ *für große* n. *Dann gilt*

$$\log\left(\mathcal{N}\left(\varepsilon_n, \mathcal{H}^{(l)}, ||\cdot||_{\infty,[-a_n,a_n]^d}\right)\right) \leq c_{28} \cdot \log(n) \cdot M_n^{d^*} \tag{21}$$

für n *hinreichend groß und eine Konstante* $c_{28} > 0$ *unabhängig von* n.

Für den Beweis von Lemma 4 benötigen wir folgendes Hilfsresultat:

Lemma 5 (Lemma 9 in [BK19]). *Sei* $l \in \mathbb{N}_0$ *und* $\sigma_r : \mathbb{R} \to \mathbb{R}$ *für* $r = 1, \ldots, l+1$ *lipschitzstetige Funktionen mit Lipschitz Konstante* $L \geq 1$, *für die gilt*

$$|\sigma_r(x)| \leq L \cdot \max\{|x|, 1\} \quad (x \in \mathbb{R}). \tag{22}$$

Sei ferner $K_0 = d$, $K_r \in \mathbb{N}$ *für* $r \in \{1, \ldots, l\}$, *sowie* $K_{l+1} = 1$. *Wir definieren rekursiv für* $r \in \{1, \ldots, l+1\}$ *und* $i \in \{1, \ldots, K_r\}$ *die Funktionen*

$$f_i^{(r)}(x) = \sigma_r\left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)}\right) \tag{23}$$

und

$$\overline{f}_i^{(r)}(x) = \sigma_r\left(\sum_{j=1}^{K_{r-1}} \overline{c}_{i,j}^{(r-1)} \cdot \overline{f}_j^{(r-1)}(x) + \overline{c}_{i,0}^{(r-1)}\right) \tag{24}$$

mit

$$f_j^{(0)} = \overline{f}_j^{(0)} = x^{(j)},$$

wobei $c_{i,0}^{(r-1)}, \overline{c}_{i,0}^{(r-1)}, \ldots, c_{i,K_{r-1}}^{(r-1)}, \overline{c}_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$. *Setzt man weiterhin*

$$\overline{C} = \max_{\substack{r=0,\ldots,l,\ i=1,\ldots,K_{r-1},\\ j=1,\ldots,K_r}} \max\left\{\left|c_{i,j}^{(r)}\right|, \left|\overline{c}_{i,j}^{(r)}\right|, 1\right\},$$

dann gilt

$$\begin{aligned}&\left|f_1^{(l+1)}(x) - \overline{f}_1^{(l+1)}(x)\right| \\ &\leq (l+1)\cdot L^{l+1}\cdot \prod_{r=0}^{l} \cdot \overline{C}^l \cdot \max\{||x||_\infty, 1\} \cdot \max_{\substack{r=0,\ldots,l,\ i=1,\ldots,K_{r+1},\\ j=0,\ldots,K_r}} \left|c_{i,j}^{(r)} - \overline{c}_{i,j}^{(r)}\right|\end{aligned}$$

für jedes $x \in \mathbb{R}^d$.

Beweis. Sei zunächst $\sigma_{id} : \mathbb{R} \to \mathbb{R}$ die Identität $\sigma_{id}(x) = x$ für alle $x \in \mathbb{R}$. Dann können wir den Raum hierarischer neuronaler Netze $\mathcal{H}^{(l)}$ für $l > 0$ beschreiben als

$$\mathcal{H}^{(l)} = \Bigg\{ h : \mathbb{R}^d \to \mathbb{R} \quad : \quad h(x) = \sum_{k=1}^{K} \sigma_{id}(g_k(\sigma_{id}(f_{1,k}(x), \ldots, f_{d^*,k}(x)))) \quad (x \in \mathbb{R}^d)$$
$$\text{für } g_k \in \mathcal{F}_{M_n,N,d^*,d^*,\alpha}^{(\textit{neuronale Netze})} \text{ und } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\}.$$

Wir setzen im Folgenden $M^* := \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M_n+1)^{d^*}$, so lässt sich jedes $g_k \in \mathcal{F}_{M_n,N,d^*,d^*,\alpha}^{(\textit{neuronale Netze})}$ schreiben als

$$g(x) = \sum_{i=1}^{M^*} \mu_i \cdot \sigma\left(\sum_{l=1}^{4d^*} \lambda_{i,l} \cdot \sigma\left(\sum_{v=1}^{d^*} \theta_{i,l,v} \cdot x^{(v)} + \theta_{i,l,0}\right) + \lambda_{i,0}\right) \tag{25}$$

$$= \sum_{i=1}^{M^*} \mu_i \cdot \sigma\left(\sum_{\substack{l=1,\ldots,4d^*\\ \bar{i}=1,\ldots,M^*}} \lambda_{i,\bar{i},l} \cdot \sigma\left(\sum_{v=1}^{d^*} \theta_{\bar{i},l,v} \cdot x^{(v)} + \theta_{\bar{i},l,0}\right) + \lambda_{i,\bar{i},0}\right), \tag{26}$$

wobei die neuen Koeffizienten definiert sind durch

$$\lambda_{i,\bar{i},l} := \begin{cases} \lambda_{i,l} & \text{falls } \bar{i} = i, \\ 0 & \text{sonst} \end{cases}$$

für alle $i, \bar{i} \in \{1, \ldots, M^*\}$ und $l \in \{0, \ldots, 4d^*\}$ (analog für $h \in \mathcal{H}^{(0)}$).

Im Folgenden wollen wir Lemma 5 anwenden, da die Funktionen $\sigma_{id}(h) = h$ für $h \in \mathcal{H}^{(l)}$ mit den obigen Darstellungen die Struktur der $f_1^{(l+1)}$ aus (23) erfüllen. Die Parameter in dem Lemma wählen wir wie folgt:

Als Lipschitz Konstante L wählen wir das Maximum der Lipschitz Konstanten von σ_{id} (was offensichtlich 1 ist) und der N-zulässigen Aktivierungsfunktion σ. Damit ist (22)

erfüllt, da $\|\sigma\|_\infty \leq 1,\ L \geq 1$ und $|\sigma_{id}(x)| = |x|$.
Bezüglich der Funktionenklasse $\mathcal{H}^{(l)}$ ist der Parameter l in dem Lemma $4l+2$ und für $r = 0, \ldots, l$ nehme K_r periodisch die Werte $\tilde{d},\ 4d^* \cdot M^*,\ M^*$ und K an, wobei $\tilde{d}$ für K_0 gleich d ist, sonst aber d^* ist.
Die Koeffizienten $c_{i,j}^{(r)}$ in unserer Funktionenklasse sind stets 0 oder μ_i, $\lambda_{i,l}$, $\theta_{i,l,v}$ aus der Definition von $\mathcal{F}_{M_n,N,d^*,d,\alpha}^{(neuronale\ Netze)}$ (1), wir können also $\overline{C} = \max\{\alpha, 1\}$ setzen.

Für Funktionen h und $\overline{h}$ aus $\mathcal{H}^{(l)}$ erhalten wir also mit Lemma 5

$$\begin{aligned}
&\left\|h - \overline{h}\right\|_{\infty,[-a_n,a_n]^d} \\
&\qquad \leq (4l+3) \cdot L^{4l+3} \cdot (4d^* \cdot M^* + 1)^{4l+3} \cdot \max\{\alpha, 1\}^{4l+2} \\
&\qquad\qquad \cdot \max\{a_n, 1\} \cdot \max_{\substack{r=0,\ldots,l,\ i=1,\ldots,K_{r+1},\\ j=0,\ldots,K_r}} \left|c_{i,j}^{(r)} - \overline{c}_{i,j}^{(r)}\right| \\
&\qquad \leq n^{c_{29}} \cdot \max_{\substack{r=0,\ldots,l,\ i=1,\ldots,K_{r+1},\\ j=0,\ldots,K_r}} \left|c_{i,j}^{(r)} - \overline{c}_{i,j}^{(r)}\right|
\end{aligned}$$

für n hinreichend groß und $c_{29} > 0$, da nach Voraussetzung gilt $a_n \leq M_n \leq \frac{M_n}{\eta_n} \leq n^{c_{27}}$.
Um also die Abschätzung $\left\|h - \overline{h}\right\|_{\infty,[-a_n,a_n]^d} \leq \varepsilon_n$ für ein ein beliebiges $h \in \mathcal{H}^{(l)}$ zu erhalten, müssen wir die Koeffizienten $\overline{c}_{i,j}^{(r)}$ einer Funktion $\overline{h} \in \mathcal{H}^{(l)}$ so wählen, dass

$$\left|c_{i,j}^{(r)} - \overline{c}_{i,j}^{(r)}\right| \leq \frac{\varepsilon}{n^{c_{29}}} \tag{27}$$

für alle möglichen Indizes i, j, r. Die $c_{i,j}^{(r)}$ müssen Werte in $[-\alpha, \alpha]$ annehmen und für hinreichend großes n gilt $\alpha = \log(n) \cdot \frac{M_n^{d^*+p\cdot(2N+3)+1}}{\eta_n} \leq n^{c_{53}}$. Das heißt, da $\varepsilon_n \geq \frac{1}{n^{c_{26}}}$, benötigt man

$$\left\lceil \frac{2 \cdot \alpha \cdot n^{c_{29}}}{2 \cdot \varepsilon_n} \right\rceil \leq n^{c_{30}}$$

unterschiedliche Koeffizienten $\overline{c}_{i,j}^{(r)}$, damit mindestens einer davon die Eigenschaft (27) für ein $c_{i,j}^{(r)}$ mit festen Indizes erfüllt. Ferner sind die $c_{i,j}^{(r)}$ für jedes $h \in \mathcal{H}^{(l)}$ verschieden, ergeben sich allerdings aus den Gewichten μ_i, $\lambda_{i,l}$, $\theta_{i,l,v}$ in der Definition von $\mathcal{F}_{M_n,N,d^*,d,\alpha}^{(neuronale\ Netze)}$.
Durch (4) und (5) können wir die Anzahl der Gewichte abschätzen mit

$$\begin{aligned}
&N\left(\mathcal{H}^{(l)}\right) \cdot \mathcal{W}\left(\mathcal{F}_{M,N,d^*,d^*,\alpha}^{(neuronale\ Netze)}\right) \\
&= \sum_{t=1}^{l} {d^*}^{t-1} \cdot K^t + (d^* \cdot K)^l \cdot \binom{d^*+N}{d^*} (N+1)(M_n+1)^{d^*} \cdot (4d^* \cdot (d+2)+2) + 1 \\
&\leq c_{31} \cdot M_n^{d^*},
\end{aligned}$$

für eine hinreichend große Konstante $c_{31} > 0$.
Insgesamt können wir nun also die ε_n-Überdeckungszahl $\mathcal{N}\left(\varepsilon_n, \mathcal{H}^{(l)}, \|\cdot\|_{\infty,[-a_n,a_n]^d}\right)$ abschätzen und erhalten für den Logarithmus

$$\log\left(\mathcal{N}\left(\varepsilon_n, \mathcal{H}^{(l)}, ||\cdot||_{\infty,[-a_n,a_n]^d}\right)\right) \leq \log\left((n^{c_{30}})^{c_{31}\cdot M_n^{d^*}}\right) \leq c_{28}\cdot\log(n)\cdot M_n^{d^*},$$

was die Aussage zeigt.

□

Kapitel 5

Approximationseigenschaft der Klasse $\mathcal{H}^{(l)}$ und Beweis von Satz 1

Als letzten Schritt für den Beweis von Satz 1 benötigen wir ein neues Resultat aus der Approximationstheorie mehrschichtiger neuronaler Netze.

Satz 5 (Theorem 3 in [BK19]). *Sei X eine $\mathbb{R}^d$-wertige Zuvallsvariable und $m : \mathbb{R}^d \to \mathbb{R}$ eine (p, C)-glatte Funkion, die einem verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level l genügt, wobei $p = q + s$ für ein $q \in \mathbb{N}_0$ und $s \in (0, 1]$. Sei $N \in \mathbb{N}_0$ mit $N \geq q$. Seien in Definition 4 (b) alle partiellen Ableitungen der Ordnung $\leq q$ der Funktionen $g_k, f_{j,k}$ beschränkt und jede solche Funktion f erfülle*

$$\max_{\substack{j_1,\dots,j_d \in \{0,1,\dots,q\},\\ j_1+\dots+j_d \leq q}} \left\| \frac{\partial^{j_1+\dots+j_d} f}{\partial^{j_1} x^{(1)} \dots \partial^{j_d} x^{(d)}} \right\|_{\infty,[-2a,2a]^d} \leq c_{32}. \tag{1}$$

Seien zudem alle Funktionen g_k lipschitzstetig mit Lipschitzkonstante $L > 0$. Sei nun $M_n \in \mathbb{N}$ und $(a_n)_{n\in\mathbb{N}}$ eine wachsende Folge reeller Zahlen mit $1 \leq a_n \leq M_n$, für die gilt

$$a_n^{N+q+3} \leq M_n^p \tag{2}$$

für hinreichend großes n. Sei $\mathcal{H}^{(l)}$ wie in (3) mit K, d, d^* aus der Definition von m und setze $M = M_n$, sowie

$$\alpha = \log(n) \cdot \frac{M_n^{d^*+p\cdot(2N+3)+1}}{\eta_n} \quad \text{mit } \eta_n \in (0, 1]$$

Sei $\sigma : \mathbb{R}^n \to [0, 1]$ N-zulässig gemäß Definition Definition 6. Dann existiert für beliebiges $c > 0$ und für alle $n > n_0(c) \in \mathbb{N}$ ein $t \in \mathcal{H}^{(l)}$, sodass außerhalb einer $\mathbf{P}_X$-messbaren Menge D_n mit $\mathbf{P}_X(D_n) \leq c \cdot \eta_n$ gilt

$$|t(x) - m(x)| \leq c_{33} \cdot a_n^{N+q+3} \cdot M_n^{-p} \tag{3}$$

für jedes $x \in [-a_n, a_n]^d$ und mit c_{33} unabhängig von allen anderen Faktoren auf der rechten Seite aber abhängig von festen Werten (c, d, d^*). Außerdem kann t so gewählt werden, dass

$$|t(x)| \leq c_{34} \cdot a_n^q \cdot M_n^{d^*+N\cdot p} \tag{4}$$

für alle $x \in \mathbb{R}^d$.

Beweis. siehe [BK19]. □

Nun haben wir alle erforderlichen Vorkenntnisse, um das Hauptresultat zu zeigen und eine Konvergenzrate von

$$\log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}}$$

nachzuweisen.

Beweis von Satz 1. Sei $a_n = \log(n)^{\frac{3}{2\cdot(N+q+3)}}$. Dann existiert nach Voraussetzung ein n_0, sodass $\text{supp}(X) \subseteq [-a_n, a_n]^d$ für alle $n > n_0$ gilt und damit auch $\mathcal{N}(\delta, \mathcal{G}, ||\cdot||_{\infty,\text{supp}(X)}) \leq \mathcal{N}\left(\delta, \mathcal{G}, ||\cdot||_{\infty,[-a_n,a_n]}\right)$ für $\delta > 0$ und einen beliebigen Funktionenraum $\mathcal{G}$ folgt. Wir wenden Lemma 3 an und erhalten

$$\mathbb{E}\left[\int |m_n(x) - m(x)|^2 \, \mathbf{P}_X(dx)\right] \leq \frac{c_{19} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n\cdot\beta_n}, \mathcal{H}^{(l)}, ||\cdot||_{\infty,\text{supp}(X)}\right)\right)}{n} \tag{5}$$

$$+ 2 \cdot \inf_{h\in\mathcal{H}^{(l)}} \int |h(x) - m(x)|^2 \, \mathbf{P}_X(dx). \tag{6}$$

Den ersten Term (5) mit der Überdeckungszahl schätzen wir durch Lemma 4 ab, denn es gilt $\frac{1}{n\cdot\beta_n} \geq \frac{1}{n^{c_{26}}}$ und $\frac{M_n}{\eta_n} \leq n^{c_{27}}$, sodass für die Überdeckungszahl folgt

$$\begin{aligned}
&\frac{c_{19} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n\cdot\beta_n}, \mathcal{H}^{(l)}, ||\cdot||_{\infty,\text{supp}(X)}\right)\right)}{n} \\
&\leq \frac{c_{19} \cdot \log(n)^2 \cdot \log\left(\mathcal{N}\left(\frac{1}{n\cdot\beta_n}, \mathcal{H}^{(l)}, ||\cdot||_{\infty,[-a_n,a_n]^d}\right)\right)}{n} \\
&\leq c_{19} \cdot \log(n)^2 \cdot \frac{c_{28} \cdot \log(n) \cdot M_n^{d^*}}{n} \\
&\leq c_{35} \cdot \log(n)^3 \cdot n^{\frac{d^*}{2p+d^*}+1} \\
&= c_{35} \cdot \log(n)^3 \cdot n^{\frac{d^*-(2p+d^*)}{2p+d^*}} \\
&= c_{35} \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}}.
\end{aligned}$$

Für den zweiten Term (6) nutzen wir Satz 5. Sei dazu $h^* \in \mathcal{H}^{(l)}$, welches die Approximationseigenschaft (3) für $c = 1$ erfüllt. Die Menge, auf der diese Eigenschaft nicht gilt, sei D_n, das heißt $\mathbf{P}_X(D_n) \leq \eta_n$. Nach (4) ist

$$|h^*(x)| \leq c_{34} \cdot a_n^q \cdot M_n^{d^*+N\cdot p} \quad \forall x \in \mathbb{R}^d$$

und für hinreichend großes n gilt auch für die Regressionsfunktion $|m(x)| \leq c_{34} \cdot a_n^q \cdot M_n^{d^*+N\cdot p}$ auf dem Träger $\text{supp}(X)$. Also können wir wie folgt abschätzen:

$$\begin{aligned}
&\inf_{h\in\mathcal{H}^{(l)}} \int |h(x) - m(x)|^2 \, \mathbf{P}_X(dx) \\
&\leq \int |h^*(x) - m(x)|^2 \cdot \mathbb{1}_{D_n^c} \mathbf{P}_X(dx) + \int |h^*(x) - m(x)|^2 \cdot \mathbb{1}_{D_n} \mathbf{P}_X(dx) \\
&\leq \left(c_{33} \cdot a_n^{N+q+3} \cdot M_n^{-p}\right)^2 + \left(2 \cdot c_{34} \cdot a_n^q \cdot M_n^{d^*+N\cdot p}\right)^2 \cdot \eta_n \\
&\leq c_{36} \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}} + c_{37} \cdot \log(n)^{\frac{3q}{N+q+3}} \cdot n^{\frac{2d^*+2N\cdot p}{2p+d^*}} \cdot \log(n)^{\frac{2\cdot(N+3)}{N+q+3}} \cdot n^{-\frac{2\cdot(N+1)\cdot p+2d^*}{2p+d^*}} \\
&\leq c_4 \cdot \log(n)^3 \cdot n^{-\frac{2p}{2p+d^*}}
\end{aligned}$$

Insgesamt folgt daraus die Aussage des Satzes.

□

Kapitel 6

Anwendungsorientierte Untersuchung der Klasse hierarchischer neuronaler Netze

In diesem Kapitel implementieren wir die in Kapitel 3.1 eingeführte Klasse der dünnbesetzen tiefen neuronalen Netze $\mathcal{H}^{(l)}$ mit Hilfe der Python–Bibliothek `Pytorch`. Dann vergleichen wir dessen Performance mit derer herkömmlicher vollständig verbundener neuronaler Netze. Wir schätzen mit den Netzwerken die Regressionsfunktion $m : \mathbb{R}^5 \to \mathbb{R}$, definiert durch

$$m(x) = x_1^2 \cdot x_3 + \big(\tanh(5x_2) + x_4\big)\big(\exp(x_4) + x_5\big) \tag{1}$$

Diese genügt einem hierarchischen Interaktionsmodell aus Definition 4 mit $K = 2$ und minimaler Ordnung $d^* = 2$, da sich die Summanden jeweils schreiben lassen als Verknüpfung mehrerer Funktionen mit einer Multiplikation.
Mit der (bekannten) Regressionsfunktion m generieren wir zunächst Trainings– und Testdaten durch

$$Y = m(X) + \sigma \cdot \varepsilon,$$

wobei $X \sim \mathcal{U}[0, 1]$ und ε standardnormalverteilt ist. Das Rauschen ε skalieren wir mit $\sigma = 0.05$.

Da die Klasse $\mathcal{H}^{(l)}$ sich rekursiv aus Funktionen der Klasse $\mathcal{H}^{(0)}$ bildet, implementieren wir zunächst diese. Funktionen der Klasse $\mathcal{H}^{(0)}$ wiederum bestehen aus kleineren, vollständig verbundenen Netzwerken, wie man (farblich abgetrennt) in Abbildung 3.1 sehen kann. Diese besitzen stets eine versteckte Schicht aus $4 \cdot d^*$ Neuronen, hier definiert als `smallDense`:

```
class smallDense(nn.Module):
  def __init__(self, d, d_star):
    super(smallDense, self).__init__()
    self.fc1 = nn.Linear(d,4*d_star)
    self.fc2 = nn.Linear(4*d_star,1)

  def forward(self, input):
    x = torch.sigmoid(self.fc1(input))
    x = torch.sigmoid(self.fc2(x))
    return x
```

Als Aktivierungsfunktion nutzen wir die Sigmoid–Funktion (1), welche die Voraussetzung

von Satz 1 erfüllt und N–zulässig ist. Eine Funktion der Klasse $\mathcal{H}^{(0)} = \mathcal{F}_{M,N,d^*,d,\alpha}^{(neuronale\ Netze)}$ enthält dann $M^* = \binom{d^*+N}{d^*} \cdot (N+1) \cdot (M+1)^{d^*}$ der kleineren `smallDense` Netzwerke, die wir folgendermaßen programmieren:

```
class H_0(nn.Module):
  def __init__(self, d, d_star, M_star):
    super(H_0, self).__init__()
    self.networks = nn.ModuleList(
        [smallDense(d, d_star) for i in range(M_star)])

    self.fc_out = nn.Linear(M_star, 1)

  def forward(self, input):
    outputs = []
    for i in range(len(self.networks)):
      x = self.networks[i](input)
      outputs.append(x)

    result = torch.cat(outputs, dim=1)

    x = self.fc_out(result)
    return x
```

Nun können wir diese Klasse benutzen, um unsere hierarchischen neuronalen Netze $\mathcal{H}^{(l)}$ (`H_network`) zu definieren. Diese Netzwerke bestehen aus l Schichten bestehend aus $\mathcal{H}^{(0)}$ Netzen und einer Additionsschicht, wobei die i–te Schicht $d^{*^{(l-i)}}$ Netzwerke enthält.

```
class H_network(nn.Module):
  def __init__(self, l, d, d_star, M_star, K):
    super(H_network, self).__init__()
    self.d_star = d_star
    self.first_networks = nn.ModuleList(
        H_0(d, d_star, M_star) for i in range(d_star**l))

    self.f_networks = nn.ModuleList(
        nn.ModuleList() for i in range(l-1))
    for i in range(l-1):
      self.f_networks[i] = nn.ModuleList(
          H_0(d_star, d_star, M_star) for j in range(d_star**(i+1)))

    self.f_networks = self.f_networks[::-1]

    self.g_networks = nn.ModuleList(
        [H_0(d_star, d_star, M_star) for i in range(K)])
```

```
  def forward(self, input):
    outputs = []
    for i in range(len(self.first_networks)):
      x = self.first_networks[i](input)
      outputs.append(x)

    for i in range(len(self.f_networks)):
      for j in range(len(self.f_networks[i])):
        x = self.f_networks[i][j](torch.cat(outputs[0:self.d_star], dim=1))
        del outputs[0:self.d_star]
        outputs.append(x)
    outputs_f = torch.cat(outputs, dim=1)

    outputs_g = []
    for i in range(len(self.g_networks)):
      x = self.g_networks[i](outputs_f)
      outputs_g.append(x)

    result = sum(outputs_g) #Zuletzt werden die Outputs nur addiert.

    return result
```

Um ein geeignetes Netzwerk für die Schätzung der Regressionsfunktion zu finden, trainieren wir die Netzwerke mit den verschiedenen Hyperparametern (l, d^*, M^*, K), durch die $\lfloor \frac{4}{5} \rfloor$ Trainingsdaten des generierten Datensatzes, wobei $l \in \{0, 1, 2\}$, $d^* \in \{1, \ldots, d\}$, $M^* \in \{1, \ldots, 6\}$ und $K \in \{1, 2\}$. Danach ermitteln wir die Hyperparameter anhand des minimalen empirischen L_2–Fehlers auf den Testdaten. Für die einzelnen Trainingsschleifen nutzen wir als Verlustfunktion die mittlere quadratische Abweichung und den Optimierungsalgorithmus Adam.

```
def train(network):
  EPOCHS = 20
  optimizer = torch.optim.Adam(network.parameters(), lr = 0.01)
  for epoch in range(EPOCHS):
    for data in x_train:
      output = network(x_train)
      loss = F.mse_loss(output, y_train)
      loss.backward()
      optimizer.step() #updates the weights
      optimizer.zero_grad()
```

Die datenabhängige Wahl der Hyperparameter führt zu einem Netzwerk der Klasse $\mathcal{H}^{(1)}$ mit $d^* = 2$, $M^* = 4$, $K = 2$. Im nächsten Schritt implementieren wir vollständig verbundene neuronale feedforward–Netze mit einem bzw. drei verdeckten

Schichten und die Anzahl der Neuronen in den Schichten ermitteln wir wie zuvor in Abhängigkeit der zugrundeliegenden Daten. Es folgt, dass wir die Performance unseres Netzwerkes m_1 = `H_network(1,d,2,4,2)` mit der eines einschichtigen Netzwerks m_2 = `OneLayer(d,33)` mit 33 Neuronen und der eines dreischichtigen Netzwerks m_3 = `ThreeLayer(d,8)` mit 8 Neuronen pro Schicht vergleichen.

Um die Qualität der Schätzer m_1, m_2 und m_3 zu bestimmen, betrachten wir ein empirischen L_2–Risiko, motiviert, durch die Eigenschaften des Schätzers in Kapitel 2.2.

$$r_k = \frac{1}{N} \sum_{i=1}^{N} \big(m(X_i) - m_k(X_i)\big)^2 \quad k = 1, 2, 3.$$

Wir bestimmen das L_2–Risiko über $N = 10^5$ unabhängigen Generierungen von X. Dies wiederholen wir 100 mal und bilden jeweils den Median über r_k Das führt uns zu den folgenden Ergebnissen:

m_1 = `H_network(1,d,2,4,2)`	m_2 = `OneLayer(d,33)`	m_3 = `ThreeLayer(d,8)`
0.0869	0.2995	0.1298

Das heißt, unser Schätzer durch die Funktionenklasse $\mathcal{H}^{(l)}$ beschreibt die (bekannte) Regressionsfunktion m etwas genauer als die vollständig verbundenen Netzwerke.

Literatur

[Bir07] Melanie Birke. "Schätz- und Testverfahren in der nichtparametrischen Regression unter qualitativen Annahmen". doctoralthesis. Ruhr-Universität Bochum, Universitätsbibliothek, 2007.

[BK19] Benedikt Bauer und Michael Kohler. "On deep learning as a remedy for the curse of dimensionality in nonparametric regression". In: *Ann. Statist.* 47.4 (Aug. 2019), S. 2261–2285. DOI: 10.1214/18-AOS1747. URL: https://doi.org/10.1214/18-AOS1747.

[Can06] Emmanuel Candès. "Modern statistical estimation via oracle inequalities". In: *Acta Numerica* 15 (Mai 2006), S. 257–325. DOI: 10.1017/S0962492906230010.

[DGL96] Luc Devroye, László Györfi und Gábor Lugosi. *A Probablistic Theory of Pattern Recognition*. Bd. 31. Jan. 1996. ISBN: 978-1-4612-6877-2. DOI: 10.1007/978-1-4612-0711-5.

[Dia06] Cheikh A.T. Diack. *A Consistent Nonparametric Test of the Convexity of Regression Based on Least Squares Splines*. Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, 2006. DOI: http://dx.doi.org/10.18452/3736.

[FL06] Jianqing Fan und Runze Li. "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery". In: *Proc. Madrid Int. Congress of Mathematicians* 3 (März 2006).

[FMZ19] Jianqing Fan, Cong Ma und Yiqiao Zhong. *A Selective Overview of Deep Learning*. 2019. arXiv: 1904.05526 [stat.ML].

[GBC18] Ian Goodfellow, Yoshua Bengio und Aaron Courville. *Deep Learning. Das umfassende Handbuch*. Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze. Frechen: MITP, 2018. ISBN: 978-3-95845-700-3.

[Gyö+02] László Györfi u. a. "A Distribution-Free Theory of Non-Parametric Regression". In: (Jan. 2002). DOI: 10.1007/b97848.

[Här90] Wolfgang Härdle. *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge University Press, 1990. DOI: 10.1017/CCOL0521382483.

[IM98] David Ríos Insua und Peter Müller. "Feedforward Neural Networks for Nonparametric Regression". In: *Practical Nonparametric and Semiparametric Bayesian Statistics*. Hrsg. von Dipak Dey, Peter Müller und Debajyoti Sinha. New York, NY: Springer New York, 1998, S. 181–193. ISBN: 978-1-4612-1732-9. DOI: 10.1007/978-1-4612-1732-9_9. URL: https://doi.org/10.1007/978-1-4612-1732-9_9.

[Lu+17] Zhou Lu u. a. *The Expressive Power of Neural Networks: A View from the Width*. 2017. arXiv: 1709.02540 [cs.LG].

[Mit97] Thomas M. Mitchell. *Machine Learning*. 1. Aufl. USA: McGraw-Hill, Inc., 1997. ISBN: 0070428077.

[Sch17] Johannes Schmidt-Hieber. *Nonparametric regression using deep neural networks with ReLU activation function*. 2017. arXiv: 1708.06633 [math.ST].

[Sto82] Charles J. Stone. "Optimal Global Rates of Convergence for Nonparametric Regression". In: *Ann. Statist.* 10.4 (Dez. 1982), S. 1040–1053. DOI: 10.1214/aos/1176345969. URL: https://doi.org/10.1214/aos/1176345969.

Selbstständigkeitserklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst und dabei keine anderen als die angegebenen Hilfsmittel benutzt habe. Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach Publikationen oder Vorträgen anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht. Die Arbeit wurde bisher weder gesamt noch in Teilen einer anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Marburg, den 28.10.2020