

Zur statistischen Analyse überparametrisierter tiefer neuronaler Netze trainiert durch Gradientenabstieg

On the Statistical Analysis of Over-parametrized Deep Neural Networks Trained by Gradient Descent

Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)

Genehmigte Dissertation von Selina Katharina Drews

Tag der Einreichung: 23. Oktober 2024, Tag der Prüfung: 11. Dezember 2024

1. Gutachten: Prof. Dr. Michael Kohler

2. Gutachten: Prof. Dr. Frank Aurzada

Darmstadt, Technische Universität Darmstadt



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Mathematik
Arbeitsgruppe Stochastik

Zur statistischen Analyse überparametrisierter tiefer neuronaler Netze trainiert durch Gradientenabstieg
On the Statistical Analysis of Over-parametrized Deep Neural Networks Trained by Gradient Descent

Genehmigte Dissertation von Selina Katharina Drews

Tag der Einreichung: 23. Oktober 2024

Tag der Prüfung: 11. Dezember 2024

Darmstadt, Technische Universität Darmstadt

Bitte zitieren Sie dieses Dokument als:

URN: urn:nbn:de:tuda-tuprints-307035

URL: <https://tuprints.ulb.tu-darmstadt.de/id/eprint/30703>

Jahr der Veröffentlichung auf TUpriints: 2025

Dieses Dokument wird bereitgestellt von tuprints,

E-Publishing-Service der TU Darmstadt

<https://tuprints.ulb.tu-darmstadt.de>

tuprints@ulb.tu-darmstadt.de

Die Veröffentlichung steht unter folgender Creative Commons Lizenz:

Namensnennung 4.0 International

<https://creativecommons.org/licenses/by/4.0/>

Danksagung

An dieser Stelle möchte ich allen danken, die mich während der Promotion unterstützt haben und zum Gelingen dieser Arbeit beigetragen haben.

Ein besonderer Dank gilt Herrn Professor Michael Kohler für seine umfangreiche Unterstützung und fachliche Beratung während der letzten Jahre. Seine wertvollen Anregungen und umfassende Hilfe haben maßgeblich zum Gelingen dieser Arbeit beigetragen. Ebenso möchte ich Professor Frank Aurzada für die Zweitkorrektur dieser Arbeit danken. Darüber hinaus danke ich den Prüfern, Professor Yann Disser und Dr. Kersten Schmidt, für ihre Zeit und Mühe.

Mein Dank gilt außerdem der gesamten AG Stochastik für die gute gemeinsame Zeit. Besonders möchte ich mich bei Benjamin bedanken, dessen ermutigende Worte und emotionaler Beistand mir oft weitergeholfen haben. Ein weiterer großer Dank geht an Alisha und Max, mit denen die gemeinsamen Essenspausen und insbesondere die entspannten Kaffeepausen stets für gute Laune und wertvolle Ablenkung gesorgt haben. Zudem danke ich Alisha für ihre Unterstützung beim Korrekturlesen dieser Arbeit.

Ich möchte mich auch bei meinen Freunden, insbesondere bei Bilke, bedanken, die immer ein offenes Ohr für mich hatten und mir stets zur Seite standen.

Ein besonderer Dank gilt auch meinen Eltern, die mir in jeder Situation mit Rat und Tat zur Seite stehen und mich unermüdlich unterstützen. Darüber hinaus danke ich ihnen für ihre wertvolle Hilfe beim Korrekturlesen dieser Arbeit.

Abschließend möchte ich mich bei Laurin bedanken – nicht nur für das sorgfältige Korrekturlesen dieser Arbeit, sondern vor allem für die fachliche und persönliche Unterstützung, die er mir während der gesamten Zeit entgegengebracht hat. Seine klugen Kalendersprüche und bestärkenden Worte haben mir oft weitergeholfen und mich stets motiviert.

Zusammenfassung

Der Erfolg des Deep Learnings in den vergangenen Jahren ist unbestreitbar, insbesondere bei neuronalen Netzen mit einer großen Anzahl von Parametern. So verfügt beispielsweise ChatGPT-3 über 175 Milliarden Parameter und BERT-Large von Google über 340 Millionen Parameter. Diese Beispiele verdeutlichen den Trend zu immer größeren neuronalen Netzen mit wachsender Parameteranzahl, wodurch sie in ein überparametrisiertes Regime übergehen, in dem die Anzahl der Modellparameter die verfügbaren Trainingsdaten deutlich übersteigt.

Dieses Phänomen wirft grundlegende theoretische Fragen auf, da überparametrisierte neuronale Netze, die durch den Gradientenabstieg trainiert werden, trotz ihrer Komplexität oft gut generalisieren, was der klassischen Theorie widerspricht, die die Überparametrisierung als nachteilig ansieht. Sie geht davon aus, dass Netze mit einer sehr hohen Anzahl von Parametern dazu neigen, sich zu stark an die Trainingsdaten anzupassen, was zu einer Überanpassung führt und die Leistung auf neuen, unbekanntem Daten verschlechtert. Aufgrund dieses scheinbaren Widerspruchs werden wir uns in dieser Arbeit eingehender mit überparametrisierten neuronalen Netzen und ihrer Leistungsfähigkeit beschäftigen. Im Rahmen der nichtparametrischen Regression untersuchen wir dabei wichtige Eigenschaften wie die universelle Konsistenz und Konvergenzraten. Unsere Analyse führt zu der Schlussfolgerung, dass die Hypothese der Überanpassung in diesem Kontext nicht zutrifft.

Die zugrunde liegende Theorie verfolgt hierbei einen ganzheitlichen Ansatz, der die drei Kernbereiche des Deep Learnings vereint: Optimierung, Approximation und Generalisierung.

Bei der Optimierung trainieren wir unsere Schätzer, wie in der Praxis üblich, mittels Gradientenabstieg. Dabei werden die Parameter des Netzes so angepasst, dass die Verlustfunktion minimiert wird. Im Gegensatz zu herkömmlichen theoretischen Ansätzen verwenden wir jedoch keinen Regularisierungsterm, der normalerweise zur Vermeidung von Überanpassung eingesetzt wird.

Der Bereich der Approximation konzentriert sich auf die Fähigkeit neuronaler Netze, komplexe Funktionen möglichst genau zu rekonstruieren. Durch den Einsatz einer geeigneten Netzwerktopologie können wir zeigen, dass überparametrisierte neuronale Netze gute Approximationseigenschaften aufweisen.

Ein weiterer wichtiger Bestandteil ist die Generalisierung, also die Fähigkeit, auch auf neuen, unbekanntem Daten gute Ergebnisse zu erzielen. Unsere Untersuchungen zeigen, dass überparametrisierte neuronale Netze trotz der großen Anzahl von Parametern in der Lage sind, zuverlässige Vorhersagen zu treffen.

In der vorliegenden Arbeit werden wir drei zentrale Resultate präsentieren, die die statistische Leistungsfähigkeit dieser Netze verdeutlichen. So kann durch die Verwendung der sigmoidalen Aktivierungsfunktion $\sigma(z) = 1/(1 + \exp(-z))$ die universelle Konsistenz solcher Netze nachgewiesen werden. Darüber hinaus ist es möglich, dieses Ergebnis für Regressionsfunktionen mit geeigneten Glattheitsannahmen zu erweitern, indem eine gute Konvergenzrate hergeleitet wird, die nahezu optimal ist. Des Weiteren können wir für überparametrisierte Neuronale-Netze-Schätzer mit ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$ eine dimensionsunabhängige Konvergenzrate herleiten. Die Grundlage für diese Rate bildet eine kompositionelle Annahme an die Struktur der Regressionsfunktion, wodurch gezeigt werden kann, dass überparametrisierte neuronale Netze unter bestimmten Bedingungen in der Lage sind, den Fluch der Dimensionalität zu umgehen.

Abstract

The success of deep learning over the past few years has been undeniable, especially for neural networks with a large number of parameters. For example, ChatGPT-3 has 175 billion parameters, and BERT-Large has 340 million parameters. These examples illustrate the trend toward increasingly larger neural networks with a growing number of parameters, leading to an over-parametrized regime where the number of model parameters significantly exceeds the available training data.

This phenomenon raises fundamental theoretical questions, as over-parametrized neural networks trained by gradient descent often generalize well, contradicting the classical theory that over-parameterization should be avoided. It is believed that networks with a very large number of parameters tend to overfit the training data, resulting in reduced performance on new, unseen data. Based on this apparent contradiction, we will take a closer look at over-parametrized neural networks and evaluate their performance. In the context of nonparametric regression, we examine important properties such as consistency and rates of convergence. Our analysis leads to the conclusion that the hypothesis of overfitting does not hold in this context and thus provides an important contribution to the understanding of over-parametrized neural networks.

Our theory takes a comprehensive approach by integrating the three fundamental aspects of deep learning: optimization, approximation, and generalization.

Regarding optimization, our estimates are trained using gradient descent, as is commonly done in practice. The parameters of the network are adjusted to minimize the loss function. Unlike traditional theoretical approaches, we do not use a regularization term, which is commonly included to prevent overfitting.

The aspect of approximation focuses on the ability of neural networks to reconstruct complex functions as accurately as possible. By using a suitable network topology, we show that over-parametrized neural networks exhibit good approximation properties.

Another crucial aspect is generalization, which refers to the ability of a model to perform well on new, unseen data. Our analysis shows that over-parametrized neural networks are able to make reliable predictions despite their large number of parameters.

In this thesis, we will present three main results that demonstrate the statistical performance of over-parametrized neural networks.

By using the sigmoidal activation function $\sigma(z) = 1/(1 + \exp(-z))$, the good generalization capability of such networks can be confirmed, as we prove their universal consistency.

Furthermore, it is possible to extend this result for regression functions under suitable smoothness assumptions by deriving a good rate of convergence that is close to optimal.

Finally, we can derive a dimension-independent rate of convergence for over-parametrized neural network estimates with ReLU activation function $\sigma(z) = \max\{z, 0\}$. The foundation of this theorem is a compositional assumption about the structure of the regression function. This result shows that over-parametrized neural networks are able to avoid the curse of dimensionality under appropriate conditions.

Inhaltsverzeichnis

1	Einführung	1
1.1	Motivation	1
1.2	Nichtparametrische Regression und Fluch der Dimensionalität	3
1.3	Neuronale Netze	9
1.4	Gradientenabstieg	14
1.5	Konvergenzverhalten überparametrisierter tiefer Neuronale-Netze-Schätzer gelernt durch Gradientenabstieg	18
1.6	Dimensionsreduktion durch überparametrisierte tiefe Neuronale-Netze-Schätzer	21
1.7	Notation	23
2	Zur universellen Konsistenz von überparametrisierten tiefen Neuronale-Netze-Schätzern	25
2.1	Einführung des Neuronale-Netze-Schätzers	25
2.2	Universelle Konsistenz des Schätzers	27
2.2.1	Analyse des Gradientenabstiegs	28
2.2.2	Komplexität des Funktionsraums überparametrisierter tiefer neuronaler Netze	39
2.2.3	Approximationseigenschaft überparametrisierter tiefer neuronaler Netze	46
2.2.4	Beweis des Resultats zur universellen Konsistenz	56
3	Zur Konvergenzgeschwindigkeit von überparametrisierten tiefen Neuronale-Netze-Schätzern	71
3.1	Fehlerschranke des Schätzers	72
3.2	Konvergenzgeschwindigkeit des Neuronale-Netze-Schätzers	87
3.3	Konvergenzgeschwindigkeit des Schätzers im Interaktionsmodell	89
4	Dimensionsreduktion überparametrisierter tiefer Neuronale-Netze-Schätzer trainiert durch Gradientenabstieg	97
4.1	Einführung des Schätzers	97
4.2	Ein allgemeines Resultat zur Fehlerschranke des Schätzers	100
4.3	Approximation hierarchischer Kompositionsmodelle durch neuronale Netze	112
4.4	Konvergenzgeschwindigkeit des Schätzers	121
5	Fazit	133
A	Ergänzende Resultate und Beweise	135
A.1	Beweis von Lemma 16	135
A.1.1	Approximation einer (p, C) -glatten Funktion durch Taylorpolynome	135
A.1.2	Idee des Beweises von Lemma 16	136
A.1.3	Schritt 1: Eine rekursive Definition von $T_{f,q,(C_{\mathcal{P}_2(\mathbf{x}))_{\text{left}}}}(\mathbf{x})$	137
A.1.4	Schritt 2: Approximation von $\phi_{1,3}$ durch neuronale Netze	139
A.1.5	Schritt 3: Approximation von $w_{\mathcal{P}_2(\mathbf{x})} \cdot f(\mathbf{x})$ durch neuronale Netze	168



A.1.6	Schritt 4: Anwendung von \hat{f} auf leicht verschobene Partitionen	186
A.2	Hilfsresultate	188
Abbildungsverzeichnis		191
Literaturverzeichnis		193

1 Einführung

1.1 Motivation

Künstliche Intelligenz (KI) ist kein neuartiges Phänomen. Bereits in den 1940er Jahren beschäftigten sich Forscher mit der Idee, Maschinen zu entwickeln, die in der Lage sind, menschliches Denken nachzuahmen (siehe Goodfellow et al. (2016)). Das Problem besteht darin, dass menschliches Denken oft intuitiv erfolgt. Somit ist es sowohl schwer zu formalisieren als auch zu implementieren. Um diese Herausforderung zu meistern, müssen KI-Systeme in der Lage sein, sich selbst Wissen anzueignen, indem sie aus gegebenen Daten die notwendigen Informationen extrahieren. Dieser Prozess wird mit dem Begriff *Machine Learning* bezeichnet.

Im Jahr 2006 wurde das sogenannte *Deep Learning* von Hinton et al. (2006) eingeführt. Es basiert auf künstlichen neuronalen Netzen (KNN) mit mehreren Schichten und ermöglicht es, Muster in großen Datensätzen zu erkennen. Durch die zunehmende Verfügbarkeit großer Datensätze und die Verbesserung der Rechenleistung konnte das Deep Learning in praktischen Anwendungen beachtliche Erfolge erzielen. Bekannte Beispiele für den Einsatz von Deep Learning sind die Bereiche der Bilderkennung (Rawat und Wang (2017)), der Spracherkennung (Hinton et al. (2012)) oder der Medizin (Litjens et al. (2017)). Insbesondere im Bereich der medizinischen Bildanalyse ermöglichen Deep-Learning-Algorithmen die Früherkennung von Krankheiten wie zum Beispiel Hautkrebs, indem sie präzise Diagnosen aus radiologischen Bildern liefern (Lai et al. (2023)).

Das Deep Learning wird heutzutage als Kerntechnologie der vierten industriellen Revolution, die auch als Industrie 4.0 bezeichnet wird, betrachtet (Sarker (2021)). Industrie 4.0 steht für die intelligente Vernetzung von Maschinen und Abläufen. Diese soll mithilfe von Informations- und Kommunikationstechnologien wie zum Beispiel dem Internet der Dinge, Big Data und künstlicher Intelligenz erreicht werden. In diesem Zusammenhang spielt das Deep Learning eine entscheidende Rolle, da es die Grundlage für intelligente Automatisierungssysteme bildet. Tiefe neuronale Netze können so zum Beispiel in der Produktion oder für die Produktqualitätsprüfung eingesetzt werden (Hernavts et al. (2018)).

Trotz der praktischen Erfolge des Deep Learnings fehlen oft die theoretischen Grundlagen, um diese vollumfänglich zu erklären. Die Suche nach geeigneten Netzwerkarchitekturen ist daher sehr zeitaufwendig und basiert häufig auf dem Prinzip von Versuch und Irrtum. Die Forschung auf diesem Gebiet hat in den letzten Jahren stark zugenommen und zielt darauf ab, diese Lücken zu schließen.

Die mathematische Analyse neuronaler Netze ist sehr komplex, da verschiedene Aspekte berücksichtigt werden müssen. Beispielsweise spielen die Wahl der richtigen Netzwerkarchitektur, der geeigneten Aktivierungsfunktion und dem optimalen Lernalgorithmus eine große Rolle. Im Deep Learning haben sich im Laufe der Zeit drei zentrale Forschungsbereiche herauskristallisiert: *Generalisierung*, *Approximation* und *Optimierung* (siehe Kutyniok (2022)). Diese drei Bereiche sind in Abbildung 1.1 dargestellt.

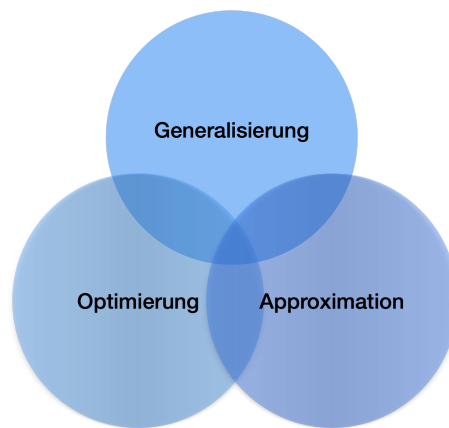


Abbildung 1.1: Fundamentale Forschungsbereiche des Deep Learnings

Im Bereich der Generalisierung liegt der Fokus auf der Analyse, wie gut neuronale Netze die gewünschte Funktion auf neuen, unbekanntem Daten nachbilden. Eine interessante Beobachtung ist, dass Netze, die mehr Parameter als verfügbare Daten haben, neue, unbekannte Daten sehr gut wiedergeben und nicht, wie lange angenommen, zur Überanpassung neigen (Neyshabur et al. (2019)).

Im Bereich der Optimierung wird der Trainingsalgorithmus, mit dem das neuronale Netz effizient trainiert werden kann, untersucht. Hierbei hat sich insbesondere der (stochastische) Gradientenabstieg durchgesetzt. Mit seiner Hilfe können oft geeignete lokale Minima erreicht werden, was wesentlich zur erfolgreichen Anwendung neuronaler Netze beiträgt (Ruder (2016)).

Der Bereich der Approximation beschäftigt sich mit der Frage, wie gut eine gegebene Funktionsklasse durch eine Klasse von tiefen neuronalen Netzen beschrieben werden kann und wie groß der damit einhergehende Fehler ist. Bereits Hornik et al. (1989) konnten die universelle Approximationsfähigkeit für neuronale Netze mit einer einzigen verdeckten Schicht nachweisen, welche besagt, dass ein solches Netzwerk mit genügend Parametern jede stetige Funktion beliebig genau approximieren kann. Diese Aussage lässt sich auf tiefe neuronale Netzwerke erweitern (siehe zum Beispiel Yarotsky (2017)).

Viele bereits existierende Resultate beschäftigen sich lediglich mit zwei der drei Forschungsbereiche und lassen einen der Bereiche außer Acht. Um neuronale Netze jedoch ganzheitlich verstehen zu können, sind Resultate notwendig, die alle drei Bereiche berücksichtigen. Im Rahmen der vorliegenden Arbeit betrachten wir neuronale Netze als Schätzer im Kontext der nichtparametrischen Regression und können dabei Resultate herleiten, welche in der Schnittmenge der drei Forschungsbereiche liegen. Damit leisten wir einen wichtigen Beitrag zum theoretischen Verständnis des Deep Learnings und zeigen, wie Generalisierung, Approximation und Optimierung zusammenwirken, um die Leistungsfähigkeit neuronaler Netze zu erklären.

Diese Arbeit ist folgendermaßen gegliedert:

In Kapitel 1 geben wir eine Einführung in die nichtparametrische Regression sowie die Grundlagen neuronaler Netze. Am Ende dieses Kapitels präsentieren wir die erzielten Resultate im Kontext bereits bestehender Forschungsergebnisse.

Im zweiten Kapitel führen wir einen überparametrisierten Neuronale-Netze-Schätzer ein, der durch den Gradientenabstieg trainiert wird, und weisen für diesen die Eigenschaft der universellen Konsistenz nach. Hierfür leiten wir Resultate für die drei fundamentalen Forschungsbereiche des Deep Learnings her.

In Kapitel 3 stellen wir für diesen überparametrisierten Neuronale-Netze-Schätzer eine Fehlerschranke vor und analysieren auf Basis dieser Schranke die Konvergenzgeschwindigkeit des Schätzers. Dabei greifen wir auf einige der im zweiten Kapitel aufgeführten Ergebnisse zurück.

Im vierten Kapitel werden wir einen weiteren durch den Gradientenabstieg gelernten Neuronale-Netze-Schätzer einführen, für den wir, unter geeigneten Voraussetzungen an die Regressionsfunktion, eine dimensionsunabhängige Konvergenzgeschwindigkeit nachweisen. Hierfür verwenden wir unter anderem die Approximationsfähigkeit vollständig verbundener neuronaler Netze und begrenzen den Generalisierungsfehler durch die Rademacher-Komplexität.

Das letzte Kapitel schließt mit einem Fazit ab, das die wesentlichen Ergebnisse der Arbeit zusammenfasst und bietet darüber hinaus einen Ausblick auf zukünftige Forschungsmöglichkeiten.

1.2 Nichtparametrische Regression und Fluch der Dimensionalität

Die Regressionsanalyse ist ein wichtiger Bereich der Statistik, der darauf abzielt, die Beziehung zwischen einer abhängigen Variablen und einer oder mehreren unabhängigen Variablen zu modellieren. Hierfür betrachten wir einen Zufallsvektor (\mathbf{X}, Y) , wobei $\mathbf{X} \in \mathbb{R}^d$ der sogenannte *Prädiktorvektor* und $Y \in \mathbb{R}$ die sogenannte *abhängige Variable* mit $\mathbf{E}\{Y^2\} < \infty$ ist. Wie in Györfi et al. (2002) beschrieben, interessieren wir uns für den funktionalen Zusammenhang zwischen der abhängigen Zufallsvariable Y und dem Zufallsvektor \mathbf{X} . Das Ziel der Regression ist es, eine Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}$ zu bestimmen, so dass die abhängige Zufallsvariable Y möglichst gut durch $f(\mathbf{X})$ approximiert wird. Als Maß für die Güte dieser Approximation verwenden wir den mittleren quadratischen Fehler, auch bekannt als das L_2 -Risiko

$$\mathbf{E} \{ |f(\mathbf{X}) - Y|^2 \},$$

welches minimiert werden soll. Das bedeutet, wir möchten eine Funktion $m^* : \mathbb{R}^d \rightarrow \mathbb{R}$ finden, die

$$\mathbf{E} \{ |m^*(\mathbf{X}) - Y|^2 \} = \min_{f: \mathbb{R}^d \rightarrow \mathbb{R}} \mathbf{E} \{ |f(\mathbf{X}) - Y|^2 \} \quad (1.1)$$

erfüllt.

Die sogenannte *Regressionsfunktion*

$$m(\mathbf{x}) = \mathbf{E} \{ Y | \mathbf{X} = \mathbf{x} \}$$

ist eine Lösung des Minimierungsproblems in (1.1). Dies ergibt sich aus der Tatsache, dass

$$\begin{aligned} \mathbf{E} \{ |f(\mathbf{X}) - Y|^2 \} &= \mathbf{E} \{ |f(\mathbf{X}) - m(\mathbf{X})|^2 \} + \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \\ &= \int_{\mathbb{R}^d} |f(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) + \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \end{aligned} \quad (1.2)$$

gilt (für eine detaillierte Herleitung siehe Abschnitt 1.1 in Györfi et al. (2002)). Somit kann das L_2 -Risiko einer Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}$ durch die Summe des sogenannten L_2 -Fehlers

$$\int_{\mathbb{R}^d} |f(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})$$

und des L_2 -Risikos der Regressionsfunktion dargestellt werden. Da der L_2 -Fehler für $f(\mathbf{x}) = m(\mathbf{x})$ gleich 0 ist, erfüllt die Regressionsfunktion das Minimierungsproblem in (1.1).

In praktischen Anwendungen sind sowohl die Verteilung von (\mathbf{X}, Y) als auch die Regressionsfunktion m im Allgemeinen unbekannt. Es ist jedoch oft möglich, Beobachtungen von (\mathbf{X}, Y) zu erhalten. Wir bezeichnen mit

$$\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$$

die Menge aller Beobachtungen, wobei $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängig und identisch verteilte Zufallsvariablen sind. Mithilfe dieser Daten wird ein Regressionssschätzer

$$m_n : \mathbb{R}^d \rightarrow \mathbb{R} \quad \text{mit} \quad m_n(\mathbf{x}) = m_n(\mathbf{x}, \mathcal{D}_n)$$

konstruiert, um die Regressionsfunktion m möglichst gut zu approximieren. Dies bedeutet, dass der L_2 -Fehler des Schätzers

$$\int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \tag{1.3}$$

möglichst klein sein soll.

Die Wahl des L_2 -Fehlers als Maß für die Approximationsgüte wird durch die folgende Überlegung motiviert: Das L_2 -Risiko des Regressionssschätzers

$$\mathbf{E} \left\{ |m_n(\mathbf{X}) - Y|^2 \middle| \mathcal{D}_n \right\}$$

lässt sich analog zu Gleichung (1.2) durch

$$\mathbf{E} \left\{ |m_n(\mathbf{X}) - Y|^2 \middle| \mathcal{D}_n \right\} = \int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) + \mathbf{E} \left\{ |m(\mathbf{X}) - Y|^2 \right\}$$

darstellen. Damit ist das L_2 -Risiko des Schätzers m_n nur nahe am optimalen Wert, wenn der L_2 -Fehler in (1.3) sehr klein wird. Da der Schätzer m_n von der Datenmenge \mathcal{D}_n abhängt und sein L_2 -Fehler damit selbst eine Zufallsvariable ist, verwenden wir in dieser Arbeit den erwarteten L_2 -Fehler als Grundlage für die Beurteilung der Güte eines Schätzers.

Grundlagen der parametrischen und nichtparametrischen Regression

Im Allgemeinen wird zwischen der parametrischen und der nichtparametrischen Regression unterschieden. Parametrische Verfahren setzen die Bauart der zu schätzenden Funktion als bekannt voraus, wohingegen in der nichtparametrischen Regression keine Annahmen über den funktionalen Zusammenhang getroffen werden. Ein klassisches Beispiel für die parametrische Regression ist die *lineare Regression*. Hierbei wird ein linearer Zusammenhang, das heißt

$$m(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b \quad \text{für } \mathbf{a} \in \mathbb{R}^d, b \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^d$$

unterstellt. Die Aufgabe besteht nun darin, die Parameter \mathbf{a} und b so zu bestimmen, dass das Modell die Daten möglichst gut beschreibt. Der Nachteil parametrischer Schätzverfahren besteht darin, dass eine falsche Annahme über die Bauart zu einem großen Fehler führen kann. Ein linearer Regressionssschätzer wird beispielsweise bei Datensätzen, deren zugrundeliegende Regressionsfunktion nicht linear ist und nicht gut durch lineare Funktionen approximiert werden kann, große Fehler verursachen. Dieses Problem wird bei der nichtparametrischen Regression umgangen.

Nichtparametrische Verfahren werden zum Beispiel in der Finanz- und Wirtschaftsanalyse bei der Schätzung von Renditen, der Volatilität von Finanzinstrumenten sowie zur Bewertung von Anleihekursen und Aktienoptionen eingesetzt (Härdle und Simar (2019)). In diesem Zusammenhang gibt es in der Literatur eine Vielzahl nichtparametrischer Schätzer, darunter der *lokale Mittelungsschätzer*, der *lokale Modellierungsschätzer*, der *Kleinste-Quadrate-Schätzer* und der *penalisierte Kleinste-Quadrate-Schätzer*. Bei den lokalen Mittelungsschätzverfahren wird der Schätzer der Regressionsfunktion m als gewichtete Summe der Y_i gebildet. Beispiele für solche Schätzer sind der *Partitionierungsschätzer*, welcher zuerst unter dem Namen Regressogramm in Tukey (1947) und Tukey (1961) eingeführt wurde, oder die sogenannten *Kernschätzer*, bei denen als Gewichte zum Beispiel der Kern von Nadaraya-Watson (Nadaraya (1964), Nadaraya (1970), Watson (1964)) oder der naive Kern (Zambom und Dias (2013)) verwendet werden. Ein weiterer Schätzer dieser Klasse ist der *Nächste-Nachbar-Schätzer* (Cover (1968), Stone (1977)). Eine Verallgemeinerung der lokalen Mittelungsschätzer führt zu dem lokalen Modellierungsschätzer. Hierbei wird anstelle einer konstanten Gewichtung eine allgemeinere Funktion verwendet, um die Daten zu approximieren. Diese Funktion hängt dabei von mehreren Parametern ab. Der bekannteste Schätzer dieser Kategorie ist der *lokale Polynomschätzer* (Härdle und Tsybakov (1997), Fan (1996)). Ein weiterer Ansatz ist der Kleinste-Quadrate-Schätzer, der für eine Menge von Funktion \mathcal{F}_n definiert ist durch

$$m_n(\mathbf{x}) = \arg \min_{f \in \mathcal{F}_n} \left\{ \frac{1}{n} \sum_{s=1}^n |f(\mathbf{X}_s) - Y_s|^2 \right\}.$$

Dieser minimiert das empirische L_2 -Risiko

$$\frac{1}{n} \sum_{s=1}^n |f(\mathbf{X}_s) - Y_s|^2$$

über eine gegebene Funktionsklasse \mathcal{F}_n . Da der beste Schätzer derjenige wäre, der die Daten interpoliert, ist es sinnvoll, die Funktionsklasse \mathcal{F}_n einzuschränken. Eine Alternative zur Einschränkung der Funktionsklasse \mathcal{F}_n bietet die Verwendung von penalisierten Kleinste-Quadrate-Schätzern. Dabei wird der zu minimierenden Funktion ein Strafterm hinzugefügt, welcher die „Rauheit“ der Funktion bestraft. Ein bekanntes Beispiel für einen penalisierten Kleinste-Quadrate-Schätzer ist der sogenannte *Smoothing-Spline-Schätzer* (Wahba (1990)). Hierbei wird der Strafterm als das Integral der quadrierten zweiten Ableitung der Funktion definiert. Für einen detaillierten Einblick in nichtparametrische Regressionsschätzer verweisen wir auf Györfi et al. (2002).

Konvergenzeigenschaften von Schätzern

Es gibt verschiedene Konvergenzeigenschaften, die ein Schätzer erfüllen sollte. Eine grundlegende Eigenschaft besteht darin, dass der Schätzer für einen wachsenden Stichprobenumfang in allen möglichen Situationen gegen die zu schätzende Funktion konvergiert. Anders ausgedrückt bedeutet dies, dass der erwartete L_2 -Fehler beziehungsweise der L_2 -Fehler des Schätzers gegen 0 konvergiert, wenn der Stichprobenumfang gegen unendlich geht. Dies führt zu folgender Definition:

Definition 1. Wir bezeichnen eine Folge von Schätzern m_n als schwach universell konsistent, wenn

$$\lim_{n \rightarrow \infty} \mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) = 0$$

für jede Verteilung von (\mathbf{X}, Y) mit $\mathbf{E}\{Y^2\} < \infty$ gilt. Im Folgenden werden wir diese Eigenschaft der Einfachheit halber als universelle Konsistenz bezeichnen.

Eine Folge von Schätzern wird stark universell konsistent genannt, wenn

$$\lim_{n \rightarrow \infty} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) = 0 \quad \text{fast sicher}$$

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

Hierbei wollen wir noch einmal betonen, dass die Verteilung von (\mathbf{X}, Y) im Allgemeinen unbekannt ist. Aus diesem Grund stellt die universelle Konsistenz eine essentielle Eigenschaft für einen Schätzer dar.

In Stone (1977) wurde zum ersten Mal nachgewiesen, dass schwache universelle Schätzer existieren. Inzwischen konnte für eine Reihe von Schätzern nachgewiesen werden, dass diese schwach und stark universell konsistent sind (siehe Györfi et al. (2002) für einen Überblick über universell konsistente Schätzer).

Allerdings sagt die Eigenschaft der universellen Konsistenz nichts darüber aus, wie schnell der Schätzer gegen die zu schätzende Funktion konvergiert. Daher betrachten wir zusätzlich die Konvergenzgeschwindigkeit des Schätzers, die mithilfe des erwarteten L_2 -Fehlers gemessen wird, wobei wir insbesondere an einer Schranke interessiert sind, die vom Stichprobenumfang n abhängt. Diese liefert schließlich die Konvergenzgeschwindigkeit.

Es ist nicht möglich, eine allgemeingültige Konvergenzgeschwindigkeit herzuleiten. In Kapitel 3 von Györfi et al. (2002) wird gezeigt, dass es immer eine Verteilung von (\mathbf{X}, Y) gibt, bei der der erwartete L_2 -Fehler beliebig langsam gegen 0 konvergiert. Für den Nachweis einer Konvergenzgeschwindigkeit ist es daher erforderlich, die Klasse der Verteilungen einzuschränken. Dies kann zum Beispiel erreicht werden, indem wir die folgende spezifische Glattheitsbedingung an die Regressionsfunktion stellen.

Definition 2. Sei $p = q + s$ für ein $q \in \mathbb{N}_0$ und ein $0 < s \leq 1$. Sei $C > 0$. Eine Funktion $m : \mathbb{R}^d \rightarrow \mathbb{R}$ wird als (p, C) -glatt bezeichnet, wenn für jedes $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ mit $\sum_{j=1}^d \alpha_j = q$ die partielle Ableitung $\frac{\partial^q m}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ existiert und

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

für alle $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$ erfüllt ist.

Die (p, C) -Glattheit stellt somit eine Erweiterung der Hölder-Stetigkeit dar, indem sie auch Ableitungen höherer Ordnung berücksichtigt.

Der Fluch der Dimensionalität

Für Regressionsfunktionen in der Klasse der (p, C) -glatten Funktionen wurde von Stone (1982) gezeigt, dass die optimale Konvergenzgeschwindigkeit durch

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

gegeben ist. Damit hängt die Konvergenzgeschwindigkeit von der Dimension d unserer Eingabevariablen ab. Bleibt nun die Glattheitsanforderung an die Regressionsfunktion mit steigender Dimension gleich, so konvergiert der Schätzer immer langsamer gegen die zu schätzende Funktion. Wir verdeutlichen dies durch ein Beispiel: Sei $p = 1$ und seien $d_1 = 10$, $d_2 = 50$ und $d_3 = 500$, so erhalten wir Konvergenzraten von $n^{-\frac{1}{6}}$, $n^{-\frac{1}{26}}$ und $n^{-\frac{1}{251}}$. Es ist also deutlich erkennbar, dass die Konvergenzgeschwindigkeit abnimmt. Der Grund hierfür ist, dass mit steigender Dimension der Eingaberaum exponentiell wächst, wodurch die vorhandenen Datenpunkte den Raum nicht mehr ausreichend abdecken. Dadurch liegen die Datenpunkte sehr isoliert, was die Schätzung der Funktion erschwert. Dieses Phänomen wird als *Fluch der Dimensionalität* bezeichnet (Bellman (1957)). Durch zusätzliche Annahmen an die Regressionsfunktion ist es jedoch möglich, den Fluch der Dimensionalität zu umgehen.

In Stone (1985) wurde gezeigt, dass eine Regressionsfunktion m mit *additiver Struktur*, das heißt

$$m(\mathbf{x}) = \sum_{i=1}^d m_i(x^{(i)}) \quad (\mathbf{x} = (x^{(1)}, \dots, x^{(d)})^\top),$$

für univariate (p, C) -glatte Funktionen $m_1, \dots, m_d : \mathbb{R} \rightarrow \mathbb{R}$, die optimale Konvergenzgeschwindigkeit von $n^{-2p/(2p+1)}$ erzielt. Dadurch konnte der Fluch der Dimensionalität umgangen werden.

Stone (1994) weitete dieses Modell auf sogenannte *Interaktionsmodelle* aus. Die Regressionsfunktion ist hier für $1 \leq d^* \leq d$ gegeben durch

$$m(\mathbf{x}) = \sum_{I \subseteq \{1, \dots, d\} : |I|=d^*} m_I(\mathbf{x}_I) \quad (\mathbf{x} = (x^{(1)}, \dots, x^{(d)})^\top) \quad (1.4)$$

mit (p, C) -glatte Funktionen $m_I : \mathbb{R}^{|I|} \rightarrow \mathbb{R}$, wobei $I \subseteq \{1, \dots, d\}$, $|I| = d^*$ und $\mathbf{x}_I = (x^{(i_1)}, \dots, x^{(i_{d^*})})^\top$. Die optimale Konvergenzgeschwindigkeit für Interaktionsmodelle beträgt $n^{-2p/(2p+d^*)}$ und ist somit unabhängig von der Eingabedimension d .

Ein weiterer Ansatz aus der Literatur, um eine Dimensionsreduktion zu erzielen, ist das sogenannte *Single-Index-Modell*. Hier wird angenommen, dass die Regressionsfunktion durch

$$m(\mathbf{x}) = g(\mathbf{a}^\top \mathbf{x}) \quad (\mathbf{x} \in \mathbb{R}^d)$$

gegeben ist, wobei $\mathbf{a} \in \mathbb{R}^d$ und $g : \mathbb{R} \rightarrow \mathbb{R}$ eine univariate Funktion ist (Härdle et al. (1993), Härdle und Stoker (1989), Yu und Ruppert (2002), Kong und Xia (2007)).

Dieses Modell lässt sich zum sogenannten *Projection Pursuit* erweitern. Dabei kann die Regressionsfunktion als Summe von Single-Index-Modellen dargestellt werden. Das heißt

$$m(\mathbf{x}) = \sum_{k=1}^K g_k(\mathbf{a}_k^\top \mathbf{x}) \quad (\mathbf{x} \in \mathbb{R}^d)$$

für $K \in \mathbb{N}$, $g_k : \mathbb{R} \rightarrow \mathbb{R}$ und $\mathbf{a}_k \in \mathbb{R}^d$ (Friedman und Stuetzle (1981)). Sind in diesen beiden Modellen die univariaten Funktionen (p, C) -glatte, so erhalten wir, bis auf einen logarithmischen Faktor, die gleiche Konvergenzgeschwindigkeit wie bei den Modellen mit additiver Struktur (siehe Theorem 22.2 in Györfi et al. (2002)).

In Horowitz und Mammen (2007) wurde eine spezielle Klasse von Regressionsfunktionen untersucht, welche durch

$$m(\mathbf{x}) = g \left(\sum_{l_1=1}^{L_1} g_{l_1} \left(\sum_{l_2=1}^{L_2} g_{l_1, l_2} \left(\dots \sum_{l_r=1}^{L_r} g_{l_1, \dots, l_r} (x^{l_1, \dots, l_r}) \right) \right) \right)$$

definiert ist, wobei $g, g_{l_1}, \dots, g_{l_1, \dots, l_r} : \mathbb{R} \rightarrow \mathbb{R}$ univariate (p, C) -glatte Funktionen sind und x^{l_1, \dots, l_r} die einzelnen Komponenten von $\mathbf{x} \in \mathbb{R}^d$ repräsentieren. Die Komponenten müssen dabei für unterschiedliche Indizes nicht notwendigerweise verschieden sein. In diesem Zusammenhang wurde gezeigt, dass ein penalisierter Kleinste-Quadrate-Schätzer eine Konvergenzrate von $n^{-2p/(2p+1)}$ erzielen kann.

Es ist also durchaus möglich, unter geeigneten Voraussetzungen dem Fluch der Dimensionalität bei der Schätzung der Regressionsfunktion zu entgehen.

Hierarchische Modelle

Die oben beschriebenen Modelle können durch das in Kohler und Krzyzak (2017) eingeführte *verallgemeinerte hierarchische Interaktionsmodell* erweitert werden.

Definition 3. Sei $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ und sei $m : \mathbb{R}^d \rightarrow \mathbb{R}$ eine Funktion.

- a) Die Funktion m genügt einem verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level 0, wenn Vektoren $\mathbf{a}_1, \dots, \mathbf{a}_{d^*} \in \mathbb{R}^d$ sowie eine Funktion $f : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ existieren, so dass

$$m(\mathbf{x}) = f(\mathbf{a}_1^\top \mathbf{x}, \dots, \mathbf{a}_{d^*}^\top \mathbf{x}) \quad \text{für alle } \mathbf{x} \in \mathbb{R}^d$$

gilt.

- b) Die Funktion m genügt einem verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level $\ell + 1$, falls ein $K \in \mathbb{N}$, Funktionen $g_k : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ ($k \in \{1, \dots, K\}$) sowie Funktionen $f_{1,k}, \dots, f_{d^*,k} : \mathbb{R}^d \rightarrow \mathbb{R}$ ($k \in \{1, \dots, K\}$) existieren, so dass

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

gilt und die Funktionen $f_{1,k}, \dots, f_{d^*,k}$ ($k \in \{1, \dots, K\}$) einem verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level ℓ genügen.

- c) Das verallgemeinerte hierarchische Interaktionsmodell ist (p, C) -glatt, wenn alle in der Definition vorkommenden Funktionen f und g_k ebenfalls (p, C) -glatt sind.

Sowohl die additiven Modelle als auch die Projection Pursuit Modelle gehören zu den verallgemeinerten hierarchischen Interaktionsmodellen der Ordnung 1 und Level 1. Das Interaktionsmodell von Stone in (1.4) entspricht einem verallgemeinerten hierarchischen Interaktionsmodell der Ordnung d^* und Level 1. Auch die Single-Index-Modelle lassen sich durch ein verallgemeinertes hierarchisches Interaktionsmodell der Ordnung 1 und Level 0 darstellen. Das Modell in Horowitz und Mammen (2007) hingegen ist ein verallgemeinertes hierarchisches Interaktionsmodell der Ordnung 1 und Level $r + 1$.

In Schmidt-Hieber (2020) wurde das verallgemeinerte hierarchische Interaktionsmodell erweitert, indem in den verschiedenen Leveln unterschiedliche Glattheiten p_ℓ sowie unterschiedliche Dimensionen d_ℓ zugelassen wurden. Darüber hinaus wurde der additive Zusammenhang durch einen allgemeineren funktionalen Zusammenhang ersetzt.

Dieses Modell wurde in Kohler und Langer (2021) dahingehend weiterentwickelt, dass auch innerhalb desselben Levels einer Komposition unterschiedliche Glattheiten und Dimensionen zugelassen werden. Dies führte zum sogenannten *hierarchischen Kompositionsmodell*:

Definition 4. Sei $d \in \mathbb{N}$ und $m : \mathbb{R}^d \rightarrow \mathbb{R}$. Sei weiter \mathcal{P} eine Teilmenge von $(0, \infty) \times \mathbb{N}$.

- a) Die Funktion m genügt einem hierarchischen Kompositionsmodell von Level 0 mit Ordnungs- und Glattheitsbedingung \mathcal{P} , wenn ein $K \in \{1, \dots, d\}$ existiert, so dass

$$m(\mathbf{x}) = x^{(K)} \quad \text{für alle } \mathbf{x} = (x^{(1)}, \dots, x^{(d)})^\top \in \mathbb{R}^d$$

gilt.

- b) Die Funktion m genügt einem hierarchischen Kompositionsmodell von Level $\ell + 1$ mit Ordnungs- und Glattheitsbedingung \mathcal{P} , wenn $(p, K) \in \mathcal{P}$, $C > 0$, $g : \mathbb{R}^K \rightarrow \mathbb{R}$ und $f_1, \dots, f_K : \mathbb{R}^d \rightarrow \mathbb{R}$ existieren, so dass g eine (p, C) -glatte Funktion ist sowie

$$m(\mathbf{x}) = g(f_1(\mathbf{x}), \dots, f_K(\mathbf{x})) \quad \text{für alle } \mathbf{x} \in \mathbb{R}^d$$

gilt, wobei f_1, \dots, f_K einem hierarchischen Kompositionsmodell von Level ℓ mit Ordnungs- und Glattheitsbedingung \mathcal{P} genügen.

In Kohler und Krzyżak (2017) sowie in Kohler und Langer (2021) werden hierarchische Kompositionsmodelle durch die Anwendung in komplexen technischen Systemen, die in modularer Form aufgebaut sind, motiviert. Nach diesem Vorbild scheint es realistisch zu sein, dass die Eingabe-Ausgabe-Beziehung eines solchen Systems durch eine Regressionsfunktion beschrieben werden kann, die aus der Klasse der hierarchischen Kompositionsmodelle stammt. Für diese Klasse an Regressionsfunktionen konnte in Kohler und Langer (2021) unter Verwendung eines Kleinste-Quadrate-Schätzers die, bis auf einen logarithmischen Term, optimale Konvergenzrate von

$$\max_{(p,C) \in \mathcal{P}} n^{-\frac{2p}{2p+K}}$$

mit Ordnungs- und Glattheitsbedingung $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$ nachgewiesen werden.

In diesem Abschnitt haben wir uns eingehend mit der Struktur von Regressionsfunktionen befasst, die zur Erzielung guter Konvergenzgeschwindigkeiten erforderlich sind. Dabei spielt natürlich auch die Wahl des Schätzers eine entscheidende Rolle. Als besonders erfolgreich haben sich sogenannte *Neuronale-Netze-Schätzer* erwiesen (siehe zum Beispiel Györfi et al. (2002), Devroye et al. (1996), Anthony und Bartlett (1999) und Ripley (1996)). Diese wurden unter anderem auch in den Resultaten von Bauer und Kohler (2019), Schmidt-Hieber (2020) sowie Kohler und Langer (2021) verwendet. Im folgenden Abschnitt werden wir daher neuronale Netze und ihre theoretischen Grundlagen näher betrachten.

1.3 Neuronale Netze

Ursprünglich diente das menschliche Gehirn als Vorbild für die Entwicklung künstlicher neuronaler Netze. Heutzutage steht jedoch nicht mehr die Nachbildung des menschlichen Gehirns im Vordergrund, sondern vielmehr die Entwicklung leistungsfähiger und spezialisierter neuronaler Netze für verschiedene Anwendungsbereiche. Ein neuronales Netz besteht aus verschiedenen Komponenten. Eine der Hauptkomponenten neuronaler Netze ist das *Neuron*. Die künstlichen Neuronen wurden nach dem Vorbild menschlicher Nervenzellen entwickelt. Sie nehmen Informationen von außen oder von anderen Neuronen auf und leiten diese an andere Neuronen oder an die Umwelt in modifizierter Form weiter. Der Effekt, den Neuronen aufeinander haben, wird durch die sogenannten *Gewichte* ausgedrückt. Je größer der absolute Betrag eines Gewichts ist, desto größer ist der Einfluss, den ein Neuron auf ein anderes Neuron hat.

Die Eingabe eines Neurons ergibt sich aus der gewichteten Summe der Komponenten des Eingabevektors, welche um den sogenannten *Bias*-Term verschoben wird. Im Folgenden fassen wir sowohl die Gewichte des Netzes als auch die Bias-Terme unter dem Begriff *Gewichte* zusammen.

Bevor das Neuron die empfangene Eingabe an die nachfolgenden Neuronen weiterleitet, wird die sogenannte *Aktivierungsfunktion* auf die Eingabe angewendet. Diese Funktion bestimmt, wie die Informationen an die folgenden Neuronen weitergegeben werden. In der Literatur wird eine Vielzahl unterschiedlicher Aktivierungsfunktionen beschrieben. Dabei stellen sigmoidale Aktivierungsfunktionen eine besonders bekannte Klasse dar. Diese sind monoton wachsend mit

$$\lim_{z \rightarrow -\infty} \sigma(z) = 0 \quad \text{und} \quad \lim_{z \rightarrow \infty} \sigma(z) = 1.$$

Die am weitesten verbreitete Aktivierungsfunktion dieser Klasse ist die *logistische Sigmoidfunktion*

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \quad \text{für } z \in \mathbb{R}.$$

Diese ist in Abbildung 1.2a dargestellt. Die logistische Sigmoidfunktion war lange Zeit die bevorzugte Wahl in neuronalen Netzen.

Ein weiteres Beispiel für eine Aktivierungsfunktion ist der *Tangens Hyperbolicus* $\tanh(z)$ mit

$$\lim_{z \rightarrow -\infty} \sigma(z) = -1 \quad \text{und} \quad \lim_{z \rightarrow \infty} \sigma(z) = 1.$$

Eine grafische Darstellung des Tangens Hyperbolicus befindet sich in Abbildung 1.2b.

Die logistische Sigmoidfunktion und der Tangens Hyperbolicus wurden vor allem in den späten 1980er und 1990er Jahren häufig verwendet (Goodfellow et al. (2016)). Sie spielten eine entscheidende Rolle bei der Entwicklung neuronaler Netze und sind bekannt für ihre Fähigkeit, nichtlineare Zusammenhänge zu modellieren.

Eine der bekanntesten Aktivierungsfunktionen, welche heutzutage häufig in praktischen Anwendungen zum Einsatz kommt, ist die *Rectifier Linear Unit (ReLU)*

$$\sigma(z) = \max\{z, 0\} \quad \text{für } z \in \mathbb{R}.$$

Hierbei handelt es sich um eine lineare Aktivierungsfunktion mit einer Schwelle. Das bedeutet, dass sie erst „aktiv“ wird, wenn die Schwelle überwunden wird. Im Fall der ReLU-Aktivierungsfunktion wird also einfach die Eingabe ausgegeben, sobald der Wert 0 überschritten wird. In Abbildung 1.2c befindet sich die Darstellung der ReLU-Aktivierungsfunktion.

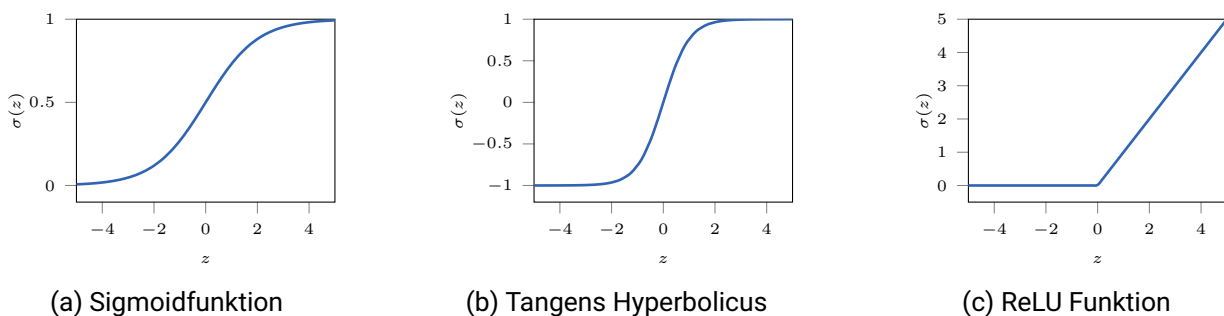


Abbildung 1.2: Aktivierungsfunktionen

Nachdem wir die Funktionsweise der Aktivierungsfunktion erläutert haben, möchten wir nun die Struktur eines einzelnen Neurons veranschaulichen. In Abbildung 1.3 ist ein Neuron mit drei Eingaben dargestellt. Das Neuron wird hier als Knoten abgebildet, während die Eingaben durch Kanten mit dem Neuron verbunden sind. Diese Kanten beschreiben sowohl die Verbindung zwischen den Eingaben und dem Neuron als auch den Informationsfluss im Netzwerk. Aus Gründen der Übersichtlichkeit haben wir auf die Beschriftung der Kanten sowie die Darstellung der Bias-Terme verzichtet.

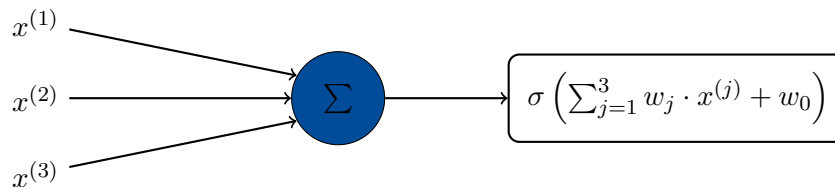


Abbildung 1.3: Graphische Darstellung eines Neurons

Formal ausgedrückt handelt es sich bei einem Neuron um eine Funktion $g : \mathbb{R}^d \rightarrow \mathbb{R}$ mit

$$g(\mathbf{x}) = \sigma \left(\sum_{j=1}^d w_j \cdot x^{(j)} + w_0 \right),$$

wobei $\mathbf{x} \in \mathbb{R}^d$ der Inputvektor, w_0, \dots, w_d die Gewichte und $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die Aktivierungsfunktion sind. Das Gewicht w_0 wird dabei, wie oben bereits erwähnt, als Bias bezeichnet. Ist der Bias w_0 positiv, so erhöht sich der Eingabewert und es wird sichergestellt, dass das Neuron aktiviert wird, auch wenn die Eingabe nur gering positiv war. Ist der Bias w_0 negativ, so sorgt er dafür, dass ein Neuron weiterhin inaktiv bleibt oder in einen inaktiven Zustand übergeht. Dies geschieht natürlich immer in Abhängigkeit von der gewählten Aktivierungsfunktion.

Kombinieren und verbinden wir mehrere Neuronen miteinander, so erhalten wir ein neuronales Netz. Bildlich gesprochen bilden übereinander angeordnete Knoten eine sogenannte *Schicht* (englisch: *layer*). Die Schichten, die sich zwischen der Eingabe und der Ausgabe befinden, nennt man *verdeckte Schichten* (englisch: *hidden layers*). In Abbildung 1.4 ist ein neuronales Netz mit zwei Eingaben, je drei Neuronen pro verdeckter Schicht sowie einer Ausgabe dargestellt. Besteht ein neuronales Netz aus mehreren Schichten, in denen der Informationsfluss vorwärtsgerichtet ist, spricht man von einem *mehrschichtigen vorwärtsgerichteten neuronalen Netz* (englisch: *multilayer feedforward neural network*).

Ist die Anzahl der Gewichte in einem Netzwerk durch eine Zahl $s \in \mathbb{N}$ beschränkt, so sind je nach Größe von s nicht alle Neuronen einer Schicht mit denen der nächsten Schicht verbunden. Das bedeutet, dass die Gesamtanzahl der Verbindungen in dem Netz höchstens s beträgt. Ein solches Netzwerk wird als *unvollständig verbundenes neuronales Netz* (englisch: *sparse neural network*) bezeichnet. Existiert keine solche Beschränkung und alle Neuronen einer Schicht sind mit allen Neuronen in der folgenden Schicht verbunden, sprechen wir von einem *vollständig verbundenen neuronalen Netz* (englisch: *fully connected neural network*).

Formal lässt sich diese Art von neuronalen Netzen durch eine Funktion $f_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}$ definieren, die durch

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{k_L} w_{1,j}^{(L)} \cdot f_{\mathbf{w},j}^{(L)}(\mathbf{x}) + w_{1,0}^{(L)} \quad (1.5)$$

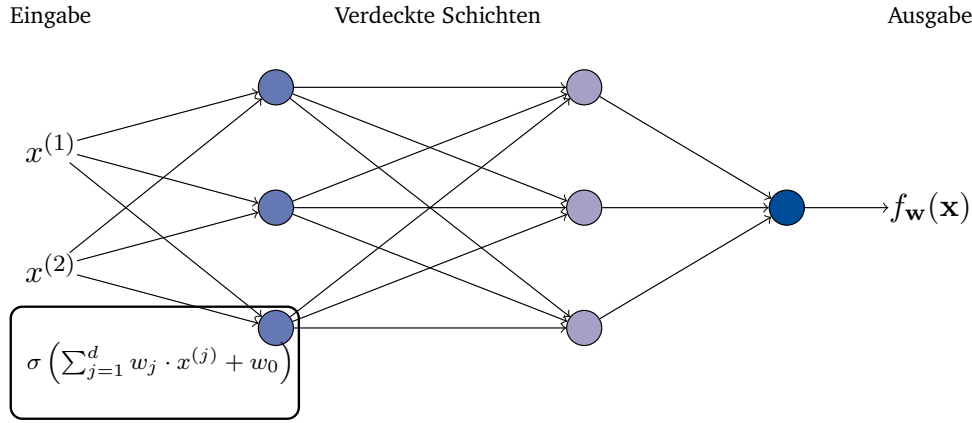


Abbildung 1.4: Vollständig verbundenes vorwärtsgerichtetes neuronales Netz

gegeben ist, wobei $w_{1,0}^{(L)}, \dots, w_{1,k_L}^{(L)} \in \mathbb{R}$ die Gewichte sind. Die Funktion $f_{\mathbf{w},i}^{(L)} : \mathbb{R}^d \rightarrow \mathbb{R}$ ist hierbei rekursiv definiert durch

$$f_{\mathbf{w},i}^{(l)}(\mathbf{x}) = \sigma \left(\sum_{j=1}^{k_{l-1}} w_{i,j}^{(l-1)} \cdot f_{\mathbf{w},j}^{(l-1)}(\mathbf{x}) + w_{i,0}^{(l-1)} \right) \quad (1.6)$$

für Gewichte $w_{i,0}^{(l-1)}, \dots, w_{i,k_{l-1}}^{(l-1)} \in \mathbb{R}$ für $l \in \{2, \dots, L\}$ und

$$f_{\mathbf{w},i}^{(1)}(\mathbf{x}) = \sigma \left(\sum_{j=1}^d w_{i,j}^{(0)} \cdot x^{(j)} + w_{i,0}^{(0)} \right) \quad (1.7)$$

für Gewichte $w_{i,0}^{(0)}, \dots, w_{i,d}^{(0)} \in \mathbb{R}$.

Dabei bezeichnet $f_i^{(l)}(\mathbf{x})$ die Ausgabe des i -ten Neurons in Schicht l und $w_{i,j}^{(l)}$ das Gewicht zwischen dem Neuron j in der $(l-1)$ -ten Schicht und dem Neuron i in Schicht l .

Wir definieren die Klasse der vollständig verbundenen neuronalen Netze durch

$$\mathcal{F}(L, r) := \{f : f \text{ wird durch (1.5) beschrieben, wobei } k_1 = \dots = k_L = r\}.$$

Mit anderen Worten enthält die Klasse $\mathcal{F}(L, r)$ vollständig verbundene neuronale Netze mit L verdeckten Schichten, wobei jede dieser Schichten die gleiche Anzahl von r Neuronen besitzt.

Neben diesen Netzwerkstrukturen gibt es noch viele weitere. Zum Beispiel die Klasse der *rekurrenten neuronalen Netze* (englisch: *recurrent networks*). Bei dieser Netzwerkart können Neuronen Verbindungen zu sich selbst oder zu Neuronen aus vorherigen Schichten haben. Eine weitere sehr bekannte Netzwerkstruktur bilden die sogenannten *faltenden neuronalen Netze* (englisch: *convolutional neural networks*), welche insbesondere im Bereich der Bilderkennung maßgebliche Erfolge erzielen (Krizhevsky et al. (2012)). Sie basieren auf der Faltungsoperation in mindestens einer Schicht des Netzes und ermöglichen so eine effiziente Verarbeitung von Bilddaten. Faltende neuronale Netze werden vor allem in den Bereichen des autonomen Fahrens (Grigorescu et al. (2019)), im medizinischen Bereich bei bildgebenden Verfahren (Yamashita et al. (2018)) oder bei der Gesichtserkennung (Coşkun et al. (2017)) eingesetzt.

Hat ein neuronales Netz „viele“ Schichten, so bezeichnen wir es als *tiefes neuronales Netzwerk* (englisch: *deep neural network*). Im letzten Jahrzehnt gewannen tiefe neuronale Netze, wie oben bereits erwähnt, in verschiedenen Bereichen an großer Bedeutung. Dieser Erfolg ist auf die Verfügbarkeit großer Datensätze sowie den Anstieg der Rechenleistung und Leistungsfähigkeit moderner Grafikkarten zurückzuführen (siehe Goodfellow et al. (2016)). Dadurch kann ein tiefes neuronales Netz mit einem Lernalgorithmus so trainiert werden, dass es bei bestimmten Aufgaben nahezu menschliche Leistungen erbringt.

Training neuronaler Netze

Das Training eines neuronalen Netzes bedeutet, dass die Gewichte des Netzes mithilfe sogenannter *Trainingsdaten* modifiziert werden. Das neuronale Netz erhält eine Eingabe, für die es eine Ausgabe bestimmt. Durch eine vorgegebene Lernregel sollen die Gewichte so optimiert werden, dass die Ausgabe des Netzes sich dem korrekten Output der Trainingsdaten annähert. Die Fehlergröße für die Trainingsdatenmenge wird als *Trainingsfehler* bezeichnet. Um herauszufinden, wie gut ein trainiertes Netz auf unbekanntem Daten abschneidet, werden sogenannte *Testdaten*, die sich von den Trainingsdaten unterscheiden, verwendet. Als Maß für die Leistung des Netzes auf diesen Daten wird der sogenannte *Testfehler* verwendet. Die genaue Vorgehensweise beim Training eines künstlichen neuronalen Netzes wird im Abschnitt 1.4 erläutert.

Überparametrisierte neuronale Netze

In der klassischen statistischen Theorie herrschte lange die Überzeugung, dass ein Modell, welches im Vergleich zu den Trainingsdaten zu viele Gewichte enthält, keine gute Generalisierung auf unbekanntem Daten erzielen kann. Man ging davon aus, dass es eine optimale Kapazität des Modells gibt, für die der Testfehler minimal wird. Diese Theorie wird in Abbildung 1.5 veranschaulicht. In der Abbildung wird deutlich, dass sowohl der Trainings- als auch der Testfehler hoch sind, wenn das Modell zu einfach ist. Dies ist darin begründet, dass das Modell bei einer geringen Kapazität zu simpel ist, um die Struktur der Daten zu erkennen. Ist die Kapazität des Modells jedoch zu hoch, so passt es sich zu stark an die Trainingsdaten an, was zu einem geringen Trainingsfehler führt. Da es jedoch die zugrunde liegende Struktur der Daten nicht gelernt hat und stattdessen etwaiges Rauschen in den Daten mitlernt, erzielt es eine schlechte Leistung auf den Testdaten (Györfi et al. (2002)). Dieses Problem wird als *Überanpassung* (englisch: *overfitting*) bezeichnet. Um dies zu vermeiden, verwenden viele Ansätze eine sogenannte *Regularisierung* (Goodfellow et al. (2016)). Für einen Überblick über verschiedene Regularisierungsmethoden siehe Kukačka et al. (2017).

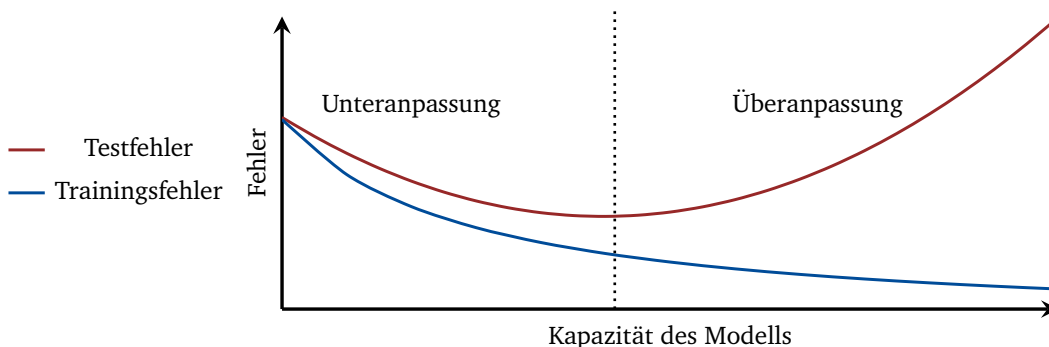


Abbildung 1.5: Klassische U-förmige Risikokurve

In den letzten Jahren zeigte sich allerdings in der Praxis, dass neuronale Netze mit einer sehr großen Anzahl von Gewichten herausragende Ergebnisse erzielen können. So hat zum Beispiel eines der leistungsfähigsten Sprachprogramme, ChatGPT-3, eine Anzahl von 175 Milliarden Gewichten (Kalyan (2024)). Auch BERT-Large von Google, ein ebenfalls sehr leistungsfähiges Sprachmodell, verfügt über 340 Millionen Gewichte (Devlin et al. (2019)). Zudem verfügt das Text-zu-Bild-Modell DALL-E-2 mit 3,5 Milliarden Gewichten über eine beachtliche Komplexität (Hunt (2023)).

Enthält ein neuronales Netz eine große Anzahl an Gewichten, das heißt, ist die Anzahl der Gewichte in dem Modell wesentlich größer als die Anzahl der Trainingsdaten, so bezeichnen wir dieses Netz als *überparametrisiert* (englisch: *over-parametrized*).

Belkin et al. (2019) konnten die klassische Theorie mit den neuen Erkenntnissen aus praktischen Anwendungen vereinen. Sie identifizierten ein Muster, welches die Abhängigkeit der Modelleistung von der Modellkapazität darstellt. Dieses Muster wird durch die sogenannte *Double Descent Curve* in Abbildung 1.6 veranschaulicht. Sie beschreibt das Phänomen, bei dem der Testfehler nach einem anfänglichen Anstieg mit zunehmender Modellkomplexität wieder abnimmt, wenn das Modell in den überparametrisierten Bereich eintritt. Das bedeutet, dass das Modell, trotz der hohen Kapazität und dem damit sehr kleinen Trainingsfehler, eine präzise Vorhersage neuer Daten liefern kann. Aus theoretischer Sicht konnte allerdings noch nicht gezeigt werden, warum dies geschieht.

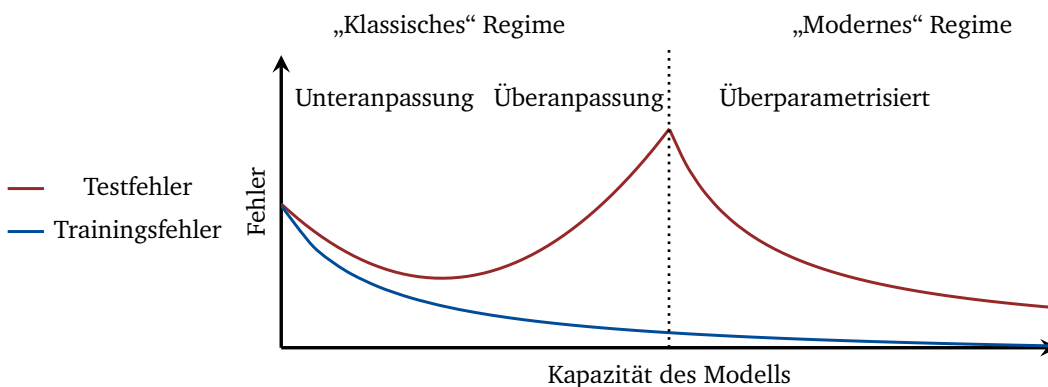


Abbildung 1.6: Double Descent Curve

Dieses Phänomen veranlasst uns, überparametrisierte neuronale Netze genauer zu untersuchen. Dafür werden wir diese in dem zuvor vorgestellten statistischen Kontext der nichtparametrischen Regression betrachten. Um für überparametrisierte neuronale Netze ein Schätzverfahren zu definieren, werden wir im folgenden Abschnitt den Gradientenabstieg als zentrale genauer Lernmethode betrachten.

1.4 Gradientenabstieg

In diesem Abschnitt werden wir uns näher damit befassen, wie neuronale Netze lernen, beziehungsweise wie sie trainiert werden. Das Training eines neuronalen Netzes besteht im Wesentlichen in der Modifizierung seiner Gewichte. Wie bereits in Abschnitt 1.3 erwähnt, bestimmen diese Gewichte, wie die Eingaben im Netzwerk verarbeitet werden und welche Ausgabe erzeugt wird.

Machine-Learning-Algorithmen können grundsätzlich in drei Kategorien eingeteilt werden: *Überwachtes Lernen* (englisch: *supervised learning*), *Unüberwachtes Lernen* (englisch: *unsupervised learning*) und *Verstärkendes Lernen* (englisch: *reinforcement learning*).

Regressionsprobleme gehören zum Bereich des überwachten Lernens. Aus diesem Grund werden wir uns im Folgenden eingehender mit diesem Themenfeld beschäftigen. Für einen tieferen Einblick in das unüberwachte und verstärkende Lernen verweisen wir auf Goodfellow et al. (2016).

Beim überwachten Lernen wird ein Datensatz verwendet, in dem die beobachteten Werte der abhängigen Variable bereits in den Trainingsdaten enthalten sind. Diese Daten dienen zur Anpassung der Gewichte des neuronalen Netzes. Das Ziel des Lernprozesses besteht darin, die Gewichte des Netzes so zu optimieren, dass die Vorhersage des Modells möglichst mit den tatsächlichen Werten der abhängigen Variable übereinstimmt. Um dies zu erreichen, wird der Datensatz in zwei Teile aufgeteilt: einen Trainings- und einen Testdatensatz. Der Trainingsdatensatz dient dazu, das neuronale Netz zu trainieren und die optimalen Gewichte zu ermitteln. Mit dem Testdatensatz wird anschließend die Generalisierungsfähigkeit des Netzes überprüft.

Einer der bekanntesten und in praktischen Anwendungen weitverbreiteten Algorithmen im Bereich des überwachten Lernens ist der sogenannte *Gradientenabstieg*. Das Ziel dieses Algorithmus ist es, die sogenannte *Verlustfunktion* für einen gegebenen Datensatz $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ zu minimieren. Die Verlustfunktion wird für eine Klasse von Funktionen $f_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}$ definiert, welche unsere neuronalen Netze darstellen, sowie eine Gütefunktion $c : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty]$, welche den Output der Funktion $f_{\mathbf{w}}$ mit den Werten der abhängigen Variablen Y_s bezüglich der Eingabe \mathbf{X}_s vergleicht. Die Verlustfunktion ist gegeben durch

$$\mathcal{L}(\mathbf{w}) = \sum_{s=1}^n c(f_{\mathbf{w}}(\mathbf{X}_s), Y_s).$$

Im Kontext der nichtparametrischen Regression ist die gängigste Wahl für die Verlustfunktion das empirische L_2 -Risiko

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{s=1}^n |f_{\mathbf{w}}(\mathbf{X}_s) - Y_s|^2.$$

Gesucht sind nun die Gewichte, die die Verlustfunktion minimieren.

Um das Minimierungsproblem

$$\min_{\mathbf{w}} F_n(\mathbf{w})$$

algorithmisch zu lösen, wird in der Praxis überwiegend das *Gradientenabstiegsverfahren*, auch *Gradientenabstieg* genannt, verwendet. Die Grundidee hierbei ist, dass man sich von einer Startposition ausgehend immer in die Richtung des steilsten Abstiegs bewegt. Dabei ist der steilste Abstieg gerade der negative Gradient der Verlustfunktion. Auf diese Weise werden die neuen Gewichte des neuronalen Netzes bestimmt. Formal ausgedrückt, ergeben sich die neuen Gewichte in Iteration $t + 1$ durch

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda \cdot \nabla_{\mathbf{w}} F_n(\mathbf{w}^{(t)}).$$

Die sogenannte *Schrittweite* λ gibt die Größe der Schritte im Verfahren an. Die Gradientenschritte werden so lange durchgeführt, bis ein Abbruchkriterium erreicht ist. In der Praxis wird häufig eine vorher festgelegte Anzahl an Gradientenschritten verwendet oder das Verfahren wird abgebrochen, falls die Änderung der Verlustfunktion zwischen zwei Iterationen unter einen bestimmten Schwellenwert fällt.

Die Berechnung des Gradienten

Um die Gewichte eines neuronalen Netzes anzupassen, benötigen wir die partiellen Ableitung der Verlustfunktion bezüglich jedes einzelnen Gewichts. Dieser Gradient kann mithilfe des *Backpropagation-Verfahrens*, das auf den Artikel von Rumelhart et al. (1986) zurückgeht, analytisch berechnet werden.

Für die Berechnung des Gradienten der Verlustfunktion nehmen wir im Folgenden an, dass die Aktivierungsfunktion differenzierbar oder wenigstens subdifferenzierbar ist. Ziel des Backpropagation-Verfahrens ist es, die partiellen Ableitungen der Verlustfunktion F_n bezüglich aller Gewichte zu berechnen. Erinnern wir uns hierfür daran, dass die Ausgabe des neuronalen Netzes durch

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{k_L} w_{1,i}^{(L)} \cdot f_{\mathbf{w},i}^{(L)}(\mathbf{x}) + w_{1,0}^{(L)}$$

gegeben ist und die Ausgabe von Neuron i in der l -ten verdeckten Schicht eines neuronalen Netzes für $l = 1, \dots, L$ durch

$$f_{\mathbf{w},i}^{(l)}(\mathbf{x}) = \sigma \left(\sum_{j=1}^{k_{l-1}} w_{i,j}^{(l-1)} \cdot f_{\mathbf{w},j}^{(l-1)}(\mathbf{x}) + w_{i,0}^{(l-1)} \right).$$

Die Ausgabe eines Neurons vor Anwendung der Aktivierungsfunktion σ bezeichnen wir mit

$$z_i^{(l)}(\mathbf{x}) = \sum_{j=1}^{k_{l-1}} w_{i,j}^{(l-1)} \cdot f_{\mathbf{w},j}^{(l-1)}(\mathbf{x}) + w_{i,0}^{(l-1)},$$

wobei $f_{\mathbf{w},j}^{(0)}(\mathbf{x}) = x^{(j)}$ ist. Wir fassen die Ausgaben der einzelnen Neuronen und die Ausgaben vor Anwendung der Aktivierungsfunktion jeweils in einem Vektor $f_{\mathbf{w}}^{(l)} = (f_{\mathbf{w},1}^{(l)}(\mathbf{x}), \dots, f_{\mathbf{w},k_l}^{(l)}(\mathbf{x}))$ beziehungsweise $z^{(l)} = (z_1^{(l)}(\mathbf{x}), \dots, z_{k_l}^{(l)}(\mathbf{x}))$ zusammen. Die Verlustfunktion definieren wir durch

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{s=1}^n r_s(\mathbf{w})$$

mit $r_s(\mathbf{w}) = |f_{\mathbf{w}}(\mathbf{x}_s) - y_s|^2$.

Aufgrund der Linearität der Ableitung genügt es, die partielle Ableitung von $r_s(\mathbf{w})$ für $s = 1, \dots, n$ bezüglich des Gewichtsvektors \mathbf{w} zu bestimmen. Hierfür benötigen wir $\delta_k^{(l)} = \partial r_s(\mathbf{w}) / \partial z_k^{(l)}$ für $l = 1, \dots, L + 1$. Da die Berechnung von $\delta_k^{(l)}$ auf den Werten von $z_i^{(l)}$ sowie $f_{\mathbf{w},i}^{(l)}$ der vorherigen Schichten basiert, müssen die Netze zuerst *vorwärts* (englisch: *forwards*) durchlaufen werden, um $z_i^{(l)}$ für $l = 1, \dots, L + 1$ sowie $f_{\mathbf{w}}$ beziehungsweise $f_{\mathbf{w},i}^{(l)}$ für $l = 1, \dots, L$ zu berechnen. In einem zweiten Schritt müssen die Netze *rückwärts* (englisch: *backwards*) durchlaufen werden, um $\delta^{(l)}$ für $l = 1, \dots, L + 1$ zu bestimmen.

Um nun $\delta_k^{(l)}$ für $l = 1, \dots, L + 1$ zu bestimmen, starten wir mit der Ausgabeschicht des Netzes. Wir bestimmen $\delta_1^{(L+1)}$, indem wir die korrekte Ausgabe y_s mit der Ausgabe des Netzes $f_{\mathbf{w}}(\mathbf{x}_s)$ vergleichen. Damit ergibt sich

$$\delta_1^{(L+1)} = \frac{\partial r_s(\mathbf{w})}{\partial z_1^{(L+1)}} = \frac{\partial r_s(\mathbf{w})}{\partial f_{\mathbf{w}}} = 2 \cdot (f_{\mathbf{w}}(\mathbf{x}_s) - y_s).$$

Zudem ist

$$\delta_k^{(L)} = \frac{\partial r_s(\mathbf{w})}{\partial z_k^{(L)}} = \frac{\partial r_s(\mathbf{w})}{\partial z_1^{(L+1)}} \cdot \frac{\partial z_1^{(L+1)}}{\partial z_k^{(L)}} = \delta_1^{(L+1)} \cdot w_{1,k}^{(L)} \cdot \sigma' \left(z_k^{(L)} \right).$$

Nun können wir die Terme $\delta_k^{(l)}$ für $l = 1, \dots, L - 1$ durch $\delta_k^{(l+1)}$ ausdrücken. Daher erhalten wir

$$\delta_k^{(l)} = \frac{\partial r_s(\mathbf{w})}{\partial z_k^{(l)}} = \sum_{j=1}^{k_{l+1}} \frac{\partial r_s(\mathbf{w})}{\partial z_j^{(l+1)}} \cdot \frac{\partial z_j^{(l+1)}}{\partial z_k^{(l)}} = \sum_{j=1}^{k_{l+1}} \delta_j^{(l+1)} \cdot w_{j,k}^{(l)} \cdot \sigma' \left(z_k^{(l)} \right).$$

Aus diesen Termen ergeben sich die partiellen Ableitungen der Gewichte in der Ausgangsschicht wie folgt

$$\frac{\partial r_s(\mathbf{w})}{\partial w_{1,k}^{(L)}} = \frac{\partial r_s(\mathbf{w})}{\partial f_{\mathbf{w}}} \cdot f_{\mathbf{w},k}^{(L)}(\mathbf{x}) = \delta_1^{(L+1)} \cdot f_{\mathbf{w},k}^{(L)}(\mathbf{x}) \quad \text{und} \quad \frac{\partial r_s(\mathbf{w})}{\partial w_{1,0}^{(L)}} = \frac{\partial r_s(\mathbf{w})}{\partial f_{\mathbf{w}}} = \delta_1^{(L+1)}.$$

Die partiellen Ableitungen in den Schichten $l = 1, \dots, L - 1$ ergeben sich aufgrund der Tatsache, dass

$$\frac{\partial z_j^{(l+1)}}{\partial w_{i,k}^{(l)}} = 0 \quad \text{für } j \neq i$$

ist durch

$$\frac{\partial r_s(\mathbf{w})}{\partial w_{i,k}^{(l)}} = \sum_{j=1}^{k_{l+1}} \frac{\partial r_s(\mathbf{w})}{\partial z_j^{(l+1)}} \cdot \frac{\partial z_j^{(l+1)}}{\partial w_{i,k}^{(l)}} = \delta_i^{(l+1)} \cdot f_{\mathbf{w},k}^{(l)}(\mathbf{x}) \quad \text{und} \quad \frac{\partial r_s(\mathbf{w})}{\partial w_{i,0}^{(l)}} = \sum_{j=1}^{k_{l+1}} \frac{\partial r_s(\mathbf{w})}{\partial z_j^{(l+1)}} \cdot \frac{\partial z_j^{(l+1)}}{\partial w_{i,0}^{(l)}} = \delta_i^{(l+1)}.$$

Auf diese Weise liefert das Backpropagation-Verfahren die partielle Ableitung der Verlustfunktion für jedes einzelne Gewicht.

Bei einem großen Stichprobenumfang n und einer großen Anzahl von Gewichten kann die Berechnung des Gradienten so zeitaufwendig werden, dass ein Computer nicht mehr in der Lage ist, diese in einem angemessenen Zeitraum durchzuführen. Aus diesem Grund wird in der Praxis häufig auf den sogenannten *stochastischen Gradientenabstieg* zurückgegriffen. Bei diesem basiert die Berechnung des Gradienten auf einer einzelnen Stichprobe (\mathbf{X}_i, Y_i) oder auf einer zufälligen Stichprobe $(\mathbf{X}_{i(1)}, Y_{i(1)}), \dots, (\mathbf{X}_{i(s)}, Y_{i(s)})$. Somit reduziert sich der Rechenaufwand bei dieser Variante des Gradientenabstiegs deutlich.

Der Gradientenabstieg findet nicht notwendigerweise ein globales Minimum. In den meisten Fällen konvergiert lediglich eine Teilfolge der Gewichte $\mathbf{w}^{(t)}$ gegen einen stationären Punkt der Zielfunktion, der häufig nur ein lokales Minimum ist. Die Konvergenz des Gradientenabstiegs zu einem globalen Minimum kann nur garantiert werden, wenn die Zielfunktion konvex ist (siehe Ulbrich und Ulbrich (2012)). In Choromanska et al. (2015) wurde jedoch mithilfe der Random-Matrix-Theory empirisch belegt, dass das empirische L_2 -Risiko lokaler Minima oft nicht wesentlich größer ist als das globaler Minima. Dies konnte für neuronale Netze mit speziellen Aktivierungsfunktionen in Arora et al. (2019), Kawaguchi (2016) und Du und Lee (2018) bestätigt werden. Diese Arbeiten untersuchten den Gradientenabstieg für neuronale Netze mit linearen sowie quadratischen Aktivierungsfunktionen und zeigten, dass diese Netzwerke trotz zahlreicher lokaler Minima oft Lösungen finden, deren L_2 -Risiko nahe am globalen Minimum liegt.

Da überparametrisierte neuronale Netze, die durch den Gradientenabstieg trainiert werden, zunehmend an Bedeutung gewinnen, wollen wir in dieser Arbeit die Kombination aus Überparametrisierung und Gradientenabstieg aus statistischer Perspektive näher untersuchen.

In den folgenden beiden Abschnitten werden wir zunächst die bestehende Literatur zur statistischen Theorie tiefer neuronaler Netze vorstellen und anschließend auf unsere Forschungsergebnisse eingehen, die einen wichtigen Beitrag zum Verständnis des Erfolgs von überparametrisierten Netzen, die durch den Gradientenabstieg trainiert werden, leisten. Die im Rahmen dieser Arbeit erzielten Resultate lassen sich hierbei in zwei Bereiche aufteilen, die jeweils in Abschnitt 1.5 und Abschnitt 1.6 vorgestellt werden.

1.5 Konvergenzverhalten überparametrisierter tiefer Neuronale-Netze-Schätzer trainiert durch Gradientenabstieg

In der Literatur gibt es bereits eine Reihe wichtiger Forschungsergebnisse zu Neuronale-Netze-Schätzern. So konnte Barron (1994) zeigen, dass Kleinste-Quadrate-Neuronale-Netze-Schätzer eine Konvergenzgeschwindigkeit von $n^{-1/2}$ erreichen können, vorausgesetzt, die Fourier-Transformierte der Regressionsfunktion hat ein endliches erstes Moment. Dieses Resultat wurde von Kohler und Krzyżak (2005) erweitert, indem sie nachwiesen, dass Neuronale-Netze-Schätzer mit zwei verdeckten Schichten für (p, C) -glatte Regressionsfunktionen mit $p \leq 1$ die optimale Konvergenzgeschwindigkeit von $n^{-2p/(2p+d)}$ (bis auf einen logarithmischen Faktor) erzielen können.

Darüber hinaus konnten Kohler und Krzyżak (2017) zeigen, dass unter der Annahme eines verallgemeinerten hierarchischen Interaktionsmodells (siehe Definition 3) die Konvergenzgeschwindigkeit eines geeigneten Kleinste-Quadrate-Schätzers, der auf einem mehrschichtigen neuronalen Netz basiert, eine dimensionsunabhängige Konvergenzgeschwindigkeit von $n^{-2p/(2p+d^*)}$ (bis auf einen logarithmischen Term) für $p \leq 1$ erreicht. Dieses Resultat wurde in Bauer und Kohler (2019) auf den Fall $p \geq 1$ verallgemeinert.

In Schmidt-Hieber (2020) konnte die gleiche Konvergenzgeschwindigkeit für einen Schätzer mit ReLU-Aktivierungsfunktion, der auf einem unvollständig verbundenen neuronalen Netz basiert, hergeleitet werden, sofern die Regressionsfunktion geeignete Kompositionsannahmen (siehe Abschnitt 1.2) erfüllt. In diesem Fall ist es sogar möglich, eine dimensionsunabhängige Konvergenzgeschwindigkeit für einen Kleinste-Quadrate-Schätzer zu zeigen, der auf vollständig verbundenen neuronalen Netzen basiert (siehe in Abschnitt 1.2 und Kohler und Langer (2021)). In Suzuki (2019) und Suzuki und Nitanda (2021) konnte gezeigt werden, dass auch unter schwächeren Annahmen an die Glattheit der Regressionsfunktion eine Dimensionsreduktion erreicht werden kann. Zu diesem Zweck untersuchten sie tiefe neuronale Netze mit ReLU-Aktivierungsfunktion als Schätzer für Funktionen in Besov-Räumen.

Diese Resultate vernachlässigen jedoch zwei wichtige Eigenschaften, die für praktische Anwendungen essentiell sind. Zum einen handelt es sich hierbei um die bereits erwähnte Überparametrisierung eines neuronalen Netzes. Zum anderen werden die Schätzer in der Praxis mithilfe des Gradientenabstiegs optimiert und nicht, wie in den oben genannten Resultaten, mit der Methode der kleinsten Quadrate. Da Gradientenverfahren jedoch nicht zwangsläufig ein globales Minimum finden, ist es in praktischen Anwendungen selten möglich, das globale Minimum des empirischen L_2 -Risikos für eine Klasse neuronaler Netze zu bestimmen. Stattdessen wird in der Regel nur ein lokales Minimum erreicht.

Ein für diese Arbeit zentrales Resultat konnte in Braun et al. (2024) erzielt werden. In diesem Artikel entwickelten die Autoren das zuvor erwähnte Resultat von Barron (1994) weiter. Sie konnten unter der Annahme, dass die Fourier-Transformierte der Regressionsfunktion hinreichend schnell abfällt sowie einer geeigneten Initialisierung der Startgewichte, zeigen, dass ein Neuronale-Netze-Schätzer, der durch den Gradientenabstieg gelernt wurde, ebenfalls eine Konvergenzgeschwindigkeit von $n^{-1/2}$ (bis auf einen logarithmischen Faktor) erzielt. Durch die geeignete Wahl der inneren Gewichte und die Anpassung der

äußeren Gewichte konnte die Methode der kleinsten Quadrate genutzt werden, um die äußeren Gewichte zu optimieren.

In Kohler und Krzyżak (2022a) konnte weiter gezeigt werden, dass unter ähnlichen Voraussetzungen an eine univariate Regressionsfunktion sowie der Verwendung des Gradientenabstiegs eine Verbesserung der Konvergenzgeschwindigkeit auf $n^{-2/3}$ für einen überparametrisierten Neuronale-Netze-Schätzer erreicht werden kann. Zusätzlich konnte nachgewiesen werden, dass der (stochastische) Gradientenabstieg bei überparametrisierten neuronalen Netzen sogar in der Lage ist, ein globales Minimum des empirischen L_2 -Risikos zu finden (Du et al. (2018), Allen-Zhu et al. (2019), Kawaguchi und Huang (2019) und Zou et al. (2020)).

Trotz dieser bemerkenswerten Eigenschaften gibt es Hinweise darauf, dass überparametrisierte neuronale Netze nicht immer gut generalisieren. So präsentieren Kohler und Krzyżak (2021) ein Gegenbeispiel, welches zeigt, dass überparametrisierte neuronale Netze, die im Wesentlichen die Trainingsdaten interpolieren, auf neuen Daten nicht zwangsläufig gut abschneiden. Sie konnten nachweisen, dass es für solche Schätzer eine Verteilung von (\mathbf{X}, Y) gibt, für die der L_2 -Fehler einer (p, C) -glatten Regressionsfunktion von unten beschränkt ist. Dies impliziert, dass überparametrisierte neuronale Netze auf neuen Daten nicht gut abschneiden, da Netzwerke, die das empirische L_2 -Risiko minimieren, nicht die optimale Minimax-Konvergenzrate bei der Schätzung glatter Regressionsfunktionen erreichen.

In Drews und Kohler (2024) konnte jedoch gezeigt werden, dass es aus theoretischer Sicht möglich ist, gute Schätzungen mit überparametrisierten neuronalen Netzen zu erzielen, wenn das Netzwerk nicht so lange trainiert wird, bis das empirische L_2 -Risiko minimal wird. Diese Erkenntnis führte zu einer detaillierten Untersuchung des Konvergenzverhaltens von überparametrisierten Neuronale-Netze-Schätzern, die bezüglich des empirischen L_2 -Risikos mit Regularisierungsterm durch den Gradientenabstieg trainiert wurden. Dabei konnte nachgewiesen werden, dass ein überparametrisierter Neuronale-Netze-Schätzer, der durch den Gradientenabstieg gelernt wurde, die Eigenschaft der universellen Konsistenz besitzt und somit auf neuen unabhängigen Daten gut generalisieren kann.

In dem Resultat von Drews und Kohler (2024) wurde der verwendete Regularisierungsterm allerdings nicht zur Beschränkung der Komplexität des Schätzers benötigt, sondern für die Analyse des Gradientenabstiegs.

Im Rahmen dieser Arbeit werden wir das Resultat erweitern, indem wir nachweisen, dass der Regularisierungsterm nicht nötig ist, um die universelle Konsistenz des Schätzers zu erhalten. Dies bedeutet, dass unser Schätzer trotz seiner hohen Komplexität in der Lage ist, das zugrunde liegende Muster in den Daten korrekt zu erfassen und auf neue Daten zu verallgemeinern.

Universelle Konsistenz überparametrisierter Neuronale-Netze-Schätzer trainiert durch Gradientenabstieg

Der Schätzer m_n , den wir in den ersten beiden Resultaten dieser Arbeit verwenden, basiert auf einem überparametrisierten tiefen neuronalen Netz. Dieses ergibt sich, indem wir mehrere tiefe vollständig verbundene neuronale Netze parallel berechnen und ihre Ausgabeneuronen miteinander verbinden, um die Ausgabe des finalen Netzes zu erhalten. Die Überparametrisierung erreichen wir, indem wir eine polynomielle Anzahl vollständig verbundener neuronaler Netze in Abhängigkeit von der Stichprobengröße n verwenden. Die Werte der Gewichte im neuronalen Netz werden durch eine festgelegte Anzahl von Gradientenschritten bestimmt, die ebenfalls von der Stichprobengröße n abhängt. Als Schätzer wählen wir dann das daraus resultierende neuronale Netz. Bevor der Gradientenabstieg angewendet werden kann, um den Neuronale-Netze-Schätzer zu bestimmen, wird eine Startinitialisierung der Gewichte des

Netzes benötigt. Hierfür setzen wir die äußeren Gewichte, also die Gewichte der Ausgabeschicht, auf 0. Die übrigen Gewichte werden unabhängig voneinander gemäß einer Gleichverteilung auf geeigneten Intervallen gewählt. Eine detaillierte Einführung des Schätzers m_n erfolgt in Abschnitt 2.1. Mit diesem Schätzer erhalten wir das folgende Resultat:

Resultat I (Universelle Konsistenz eines überparametrisierten tiefen Neuronale-Netze-Schätzers trainiert durch Gradientenabstieg). Sei $\sigma(z) = 1/(1 + \exp(-z))$ die logistische Sigmoidfunktion. Wir betrachten den überparametrisierten tiefen Neuronale-Netze-Schätzer m_n für ein Netzwerk mit mindestens zwei verdeckten Schichten und $2d$ Neuronen in jeder Schicht. Für eine geeignete Anzahl von Gradientenschritten, welche von der Größe des beobachteten Datensatzes n abhängt (siehe Theorem 1), und einer Schrittweite, die dem Kehrwert der Anzahl der Gradientenschritte entspricht, gilt

$$\lim_{n \rightarrow \infty} \mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) = 0$$

für jede Verteilung von (\mathbf{X}, Y) , wobei $\text{supp}(\mathbf{X})$ beschränkt ist sowie $\mathbf{E} \{Y^2\} < \infty$.

Die universelle Konsistenz eines Schätzers ist jedoch nur ein Aspekt seiner Leistungsfähigkeit. Wie bereits in Abschnitt 1.2 erwähnt, ist die Konvergenzgeschwindigkeit ein weiterer wichtiger Maßstab zur Beurteilung der Qualität eines Schätzers.

Konvergenzgeschwindigkeit überparametrisierter Neuronale-Netze-Schätzer trainiert durch Gradientenabstieg

Um die Konvergenzgeschwindigkeit eines überparametrisierten tiefen Neuronale-Netze-Schätzers bestimmen zu können, müssen wir die Klasse der Regressionsfunktionen einschränken. Dabei konzentrieren wir uns im Folgenden auf die Klasse der (p, C) -glatte Funktionen mit $p \in [1/2, 1]$. Für diese Klasse von Regressionsfunktionen konnte in Kohler und Krzyżak (2022b) für einen überparametrisierten Neuronale-Netze-Schätzer, der durch den Gradientenabstieg bezüglich des empirischen L_2 -Risikos mit einem Regularisierungsterm gelernt wurde, eine Konvergenzgeschwindigkeit von $n^{-1/(1+d)+\varepsilon}$ hergeleitet werden.

In dieser Arbeit wird dieses Resultat erweitert, indem wir zeigen, dass diese Konvergenzgeschwindigkeit auch ohne Regularisierungsterm erreicht werden kann.

Resultat II (Konvergenzgeschwindigkeit eines überparametrisierten tiefen Neuronale-Netze-Schätzers trainiert durch Gradientenabstieg). Sei $\text{supp}(\mathbf{X}) \subseteq [0, 1]^d$ und $\mathbf{E} \{\exp(c_1 \cdot Y^2)\} < \infty$ für eine Konstante $c_1 > 0$. Weiterhin sei die Regressionsfunktion $m(\mathbf{x})$ eine (p, C) -glatte Funktion für $p \in [1/2, 1]$ und $C > 0$. Für eine geeignete Anzahl der Gradientenschritte, welche von der Größe n des beobachteten Datensatzes abhängt (siehe Theorem 3), und einer Schrittweite, die durch den Kehrwert der Anzahl der Gradientenschritte gegeben ist, erhalten wir für einen ausreichend überparametrisierten Neuronale-Netze-Schätzer m_n , dass für alle $\varepsilon > 0$ die Ungleichung

$$\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_2 \cdot n^{-\frac{1}{1+d}+\varepsilon}$$

für eine Konstante $c_2 = c_2(\varepsilon) > 0$, die von ε abhängt, erfüllt ist.

Für $p = 1/2$ liegt der Exponent dieser Konvergenzgeschwindigkeit beliebig nahe an dem der optimalen Konvergenzrate von Stone (1982).

Genügt die Regressionsfunktion m zudem einem Interaktionsmodell (1.4) mit $p \in [1/2, 1]$, so können wir für diesen überparametrisierten Neuronale-Netze-Schätzer m_n nachweisen, dass er eine Konvergenzgeschwindigkeit von $n^{-\frac{1}{1+d^*} + \varepsilon}$ besitzt (siehe Theorem 4). Auf diese Weise kann der Fluch der Dimensionalität umgangen werden.

Diese Ergebnisse sind insofern bedeutend, da sie zeigen, dass der Regularisierungsterm nicht notwendig ist, damit ein überparametrisierter tiefer Neuronale-Netze-Schätzer, der durch den Gradientenabstieg trainiert wurde, gute Ergebnisse auf neuen Daten liefern kann. Dies bietet aus theoretischer Perspektive neue Einblicke in die Fähigkeiten überparametrisierter tiefer neuronaler Netze.

Für die Beweise der Resultate wird der L_2 -Fehler aufgespalten, so dass Approximations-, Generalisierungs- und Optimierungsfehler analysiert werden können. Zur Untersuchung dieser Fehlerkomponenten verwenden wir verschiedene Ansätze.

Für die Analyse des Optimierungsfehlers kombinieren wir Methoden aus Braun et al. (2024) und Yehudai und Shamir (2019). Während Braun et al. (2024) die Lipschitz-Stetigkeit des empirischen L_2 -Risikos zur Fehleranalyse verwenden, nutzen Yehudai und Shamir (2019) die Konvexität des empirischen L_2 -Risikos. Durch die Kombination dieser beiden Techniken gelingt es uns, den Optimierungsfehler des empirischen L_2 -Risikos ohne Regularisierungsterm zu begrenzen (siehe Lemma 1).

Um den Generalisierungsfehler zu beschränken, verwenden wir wie in Kohler und Krzyżak (2022b) einen Ansatz, der in Li und Ding (2021) vorgestellt wurde. Dieser Ansatz ermöglicht es uns, die Komplexität überparametrisierter tiefer neuronaler Netze mithilfe einer metrischen Entropie-Schranke zu kontrollieren (siehe Lemma 5).

Zur Analyse des Approximationsfehlers werden wir Resultate herleiten, die die guten Approximationseigenschaften überparametrisierter tiefer neuronaler Netze ausnutzen und zeigen, dass diese in der Lage sind, (p, C) -glatte Funktionen gut zu approximieren.

1.6 Dimensionsreduktion durch überparametrisierte tiefe Neuronale-Netze-Schätzer

Aufbauend auf dem Resultat von Kohler und Langer (2021) werden wir im Rahmen dieser Arbeit eine Konvergenzgeschwindigkeit für überparametrisierte Neuronale-Netze-Schätzer herleiten, die durch den Gradientenabstieg gelernt werden. Genügt die Regressionsfunktion hierbei einem hierarchischen Kompositionsmodell (siehe Definition 4), so ist diese Konvergenzgeschwindigkeit ebenfalls unabhängig von der Dimension der Eingabedaten d .

Für dieses Resultat betrachten wir einen Neuronale-Netze-Schätzer, der aus einer Linearkombination von vollständig verbundenen neuronalen Netzen besteht, deren Anzahl exponentiell in n wächst. Auch dieser Schätzer wird durch den Gradientenabstieg trainiert. Hierbei projizieren wir sowohl die äußeren als auch die inneren Gewichte auf geeignete Intervalle und führen eine geeignete Anzahl von Gradientenschritten durch, die von der Größe des Trainingsdatensatzes n abhängt. Als Schätzer wählen wir dann das neuronale Netz mit den Gewichten, die im Verlauf der Gradientenschritte das kleinste empirische L_2 -Risiko erreicht haben. In Abschnitt 4.1 befindet sich eine ausführlichere Beschreibung des Schätzers.

Resultat III (Dimensionsreduktion überparametrisierter tiefer Neuronale-Netze-Schätzer trainiert durch Gradientenabstieg). Sei $\sigma(z) = \max\{z, 0\}$ die ReLU-Aktivierungsfunktion sowie $\text{supp}(\mathbf{X})$ beschränkt und $\mathbf{E} \{ \exp(c_3 \cdot Y^2) \} < \infty$ für $c_3 > 0$. Die Regressionsfunktion genüge einem hierarchischen Kompositionsmodell mit geeigneten Ordnungs- und Glattheitsbedingungen $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$ (siehe Theorem 6). Unter der Voraussetzung, dass die Anzahl der Gradientenschritte von der Anzahl der parallel berechneten Netze abhängt, die Schrittweite als Kehrwert der Anzahl der Gradientenschritte gewählt wird und das Netzwerk aus L_n Schichten mit jeweils r_n Neuronen besteht, existiert eine Konstante $c_4 > 0$, so dass

$$\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_4 \cdot (\log n)^5 \cdot \max_{(p,K) \in \mathcal{P}} n^{-\frac{p}{2p+K}}$$

gilt.

Auch in diesem Beweis spielen der Optimierungs-, Approximations- und Generalisierungsfehler eine zentrale Rolle.

Den Approximationsfehler können wir mithilfe des Approximationsresultats aus Kohler und Langer (2021) abschätzen, wodurch eine Fehlerschranke für die Approximation einer (p, C) -glatte Funktion durch vollständig verbundene neuronale Netze hergeleitet werden kann. Um dieses Resultat anwenden zu können, muss sichergestellt sein, dass das bestapproximierende Netzwerk auch in dem benötigten Bereich liegt. Dies erreichen wir durch die geeignete Wahl der Startgewichte.

Das Approximationsresultat aus Kohler und Langer (2021) ist eine Erweiterung der folgenden Resultate: Bauer und Kohler (2019) zeigten, dass für ein neuronales Netz mit zwei verdeckten Schichten, der logistischen Sigmoid-Aktivierungsfunktion und einer Anzahl von W Gewichten, ein Approximationsfehler der Ordnung $W^{-p/d}$ hergeleitet werden kann. Dieselbe Fehlerordnung konnte von Schmidt-Hieber (2020) für Netzwerke mit ReLU-Aktivierungsfunktion nachgewiesen werden. Yarotsky (2017) zeigte für die Klasse der Hölder-stetigen Funktionen, dass diese durch sehr tiefe vollständig verbundene neuronale Netze mit ReLU-Aktivierungsfunktion mit einem Fehler von $W^{-2p/d}$ approximiert werden können. Aufbauend auf diesen Arbeiten wurde in Kohler und Langer (2021) ein Approximationsfehler der gleichen Größenordnung für alle (p, C) -glatten Funktionen mit $p \geq 1$ hergeleitet.

Für die Betrachtung des Optimierungsfehlers verwenden wir ähnliche Techniken wie in Resultat I und Resultat II.

Die Komplexität eines neuronalen Netzes kann für die Analyse des Generalisierungsfehlers entweder im Rahmen der klassischen Vapnik-Chervonenkis-Theorie (siehe Bartlett et al. (2019)) oder, im Fall sehr großer überparametrisierter tiefer neuronaler Netze, mithilfe von Schranken für die Rademacher-Komplexität (siehe z. B. Liang et al. (2015), Golowich et al. (2018), Lin und Zhang (2019) und Wang und Ma (2023)) untersucht werden. Da wir exponentiell überparametrisierte tiefe neuronale Netze betrachten, verwenden wir hier die Rademacher-Komplexität, um eine Schranke für den Generalisierungsfehler herzuleiten. Hierbei spielt die Projektion der Gewichte im Gradientenabstieg eine wichtige Rolle und ermöglicht es uns den Generalisierungsfehler kontrollieren zu können. Wie in Kohler und Langer (2021) betrachten wir ebenfalls vollständig verbundene neuronale Netze. Für unseren Schätzer schalten wir jedoch mehrere dieser Netze parallel, wodurch wir ein überparametrisiertes neuronales Netz erhalten. Dies führt zu einer deutlichen Vergrößerung der Komplexität des Netzes. Über den Ansatz der Rademacher-Komplexität können wir dennoch eine sehr gute Konvergenzgeschwindigkeit von

$$\max_{(p,C) \in \mathcal{P}} n^{-\frac{p}{2p+K}}$$

herleiten, welche nicht von der Eingabedimension d abhängt.

1.7 Notation

In diesem Abschnitt wird eine detaillierte Übersicht über die Symbole und Bezeichnungen gegeben, die in der vorliegenden Arbeit verwendet werden. Die verwendeten Symbole werden vorgestellt und kurz erläutert, um einen Überblick über die Notationen zu geben. In Tabelle 1.1 sind alle wichtigen Abkürzungen mit ihren zugehörigen Bedeutungen aufgelistet.

Symbol	Bedeutung
\mathbb{N}	Menge der natürlichen Zahlen
\mathbb{N}_0	Menge der natürlichen Zahlen inklusive 0
\mathbb{R}	Menge der reellen Zahlen
\mathbb{R}^+	Menge der positiven reellen Zahlen
\mathbb{R}_0^+	Menge der positiven reellen Zahlen inklusive 0
\mathbf{P}_X	Wahrscheinlichkeitsverteilung einer Zufallsvariablen X
\mathbf{E}	Erwartungswert
$\arg \min_{\mathbf{x} \in D} f(\mathbf{x})$	Bezeichnet das Argument \mathbf{x} , für das die Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ihr Minimum auf der Menge D annimmt (sofern das Minimum $\min_{\mathbf{x} \in D} f(\mathbf{x})$ existiert)
ξ	Multiindex $\xi = (\xi_1, \xi_2, \dots, \xi_d)$, wobei $\xi_i \in \mathbb{N}_0$ für $i = 1, \dots, d$
$\xi!$	Fakultät des Multiindex: $\xi! = \xi_1! \cdot \xi_2! \cdot \dots \cdot \xi_d!$
\mathbf{x}^ξ	Potenzprodukt: $\mathbf{x}^\xi = (x^{(1)})^{\xi_1} \cdot (x^{(2)})^{\xi_2} \cdot \dots \cdot (x^{(d)})^{\xi_d}$ für $\mathbf{x} \in \mathbb{R}^d$
$\partial^\xi f$	Partielle Ableitung der Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}$: $\partial^\xi f(\mathbf{x}) = \frac{\partial^{\xi_1 + \dots + \xi_d}}{\partial^{\xi_1} x^{(1)} \dots \partial^{\xi_d} x^{(d)}} f(\mathbf{x})$
$\ \mathbf{x}\ $	Euklidische Norm des Vektors $\mathbf{x} \in \mathbb{R}^d$
$\ \mathbf{x}\ _1$	1-Norm des Vektors $\mathbf{x} \in \mathbb{R}^d$
$\ \mathbf{x}\ _\infty$	Supremumsnorm des Vektors $\mathbf{x} \in \mathbb{R}^d$
$\ f\ _\infty$	Supremumsnorm $\sup_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) $ einer Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}$
$\ f\ _{A, \infty}$	Supremumsnorm $\sup_{\mathbf{x} \in A} f(\mathbf{x}) $ einer Funktion $f : A \rightarrow \mathbb{R}$ auf $A \subseteq \mathbb{R}^d$
$\ f\ _{C^q(A)}$	C^q -Norm einer Funktion $f : A \rightarrow \mathbb{R}$, wobei $A \subseteq \mathbb{R}^d$: $\max \{ \ \partial^\xi f\ _{\infty, A} : \ \xi\ _1 \leq q, \xi \in \mathbb{N}^d \}$
$f \circ g$	Komposition zweier Funktionen f und g
$\log(\cdot)$	Logarithmus zur Basis e
$\lceil z \rceil$	Aufrundungsfunktion: $\lceil z \rceil = \min\{n \in \mathbb{Z} : n \geq z\}$
$\lfloor z \rfloor$	Abrundungsfunktion: $\lfloor z \rfloor = \max\{n \in \mathbb{Z} : n \leq z\}$
$\mathbb{1}_A(\cdot)$	Indikatorfunktion einer Menge $A \subseteq \mathbb{R}^d$: $\mathbb{1}_A(\mathbf{x}) = \begin{cases} 1, & \text{falls } \mathbf{x} \in A \\ 0, & \text{sonst} \end{cases}$
$T_\beta z$	Stützungsfunktion für $z \in \mathbb{R}$ auf Höhe $\beta > 0$: $T_\beta z = \begin{cases} z, & \text{falls } -\beta \leq z \leq \beta \\ \beta \cdot \text{sgn}(z), & \text{falls } z > \beta \end{cases}$
\mathbf{x}_1^n	Sequenz von n Vektoren $\mathbf{x}_1, \dots, \mathbf{x}_n$ mit $\mathbf{x}_1^n = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, wobei $\mathbf{x}_i \in \mathbb{R}^d$ für alle $i = 1, \dots, n$
c_i	Positive Konstante c_i mit $i \in \mathbb{N}$ zur Unterscheidung

Tabelle 1.1: Notationsverzeichnis

2 Zur universellen Konsistenz von überparametrisierten tiefen Neuronale-Netze-Schätzern

In diesem Kapitel werden wir die universelle Konsistenz eines überparametrisierten tiefen Neuronale-Netze-Schätzers, der durch den Gradientenabstieg gelernt wurde, nachweisen.

Zu Beginn des Kapitels führen wir den Neuronale-Netze-Schätzer ein, dessen spezielle Topologie für die spätere Analyse von Bedeutung ist. Im zweiten Abschnitt präsentieren wir das Resultat zur universellen Konsistenz des überparametrisierten Neuronale-Netze-Schätzers. Für den Nachweis der Konsistenz benötigen wir Ergebnisse aus den Bereichen Optimierung, Generalisierung und Approximation.

In Unterabschnitt 2.2.1 werden wir die Resultate für den Optimierungsfehler präsentieren. Dabei leiten wir eine Abschätzung für das empirische L_2 -Risiko her und zeigen, dass sowohl die inneren als auch die äußeren Gewichte, die durch den Gradientenabstieg gelernt werden, nicht zu weit von den Startgewichten abweichen. Zusätzlich geben wir eine Schranke für den Gradienten des empirischen L_2 -Risikos an und weisen die Lipschitz-Stetigkeit des Gradienten des empirischen L_2 -Risikos nach.

Für die Analyse des Generalisierungsfehlers schätzen wir die Komplexität des Funktionsraums der überparametrisierten neuronalen Netze mithilfe der sogenannten *Überdeckungsanzahl* ab. Diese wird durch eine in Unterabschnitt 2.2.2 hergeleitete obere Schranke begrenzt.

In Unterabschnitt 2.2.3 zeigen wir, dass überparametrisierte neuronale Netze in der Lage sind, beschränkte Lipschitz-stetige Funktionen gut anzunähern. Dadurch lässt sich der Approximationsfehler kontrollieren.

Zum Abschluss des Kapitels kombinieren wir die verschiedenen Resultate, um zu beweisen, dass der betrachtete überparametrisierte tiefe Neuronale-Netze-Schätzer für jede Verteilung von (\mathbf{X}, Y) mit beschränktem $\text{supp}(\mathbf{X})$ bei wachsender Stichprobengröße n gegen die Regressionsfunktion konvergiert.

2.1 Einführung des Neuronale-Netze-Schätzers

Für die Definition des Schätzers, benötigen wir eine Aktivierungsfunktion. Bei diesem Schätzer fällt unsere Wahl auf die logistische Sigmoidfunktion $\sigma(z) = 1/(1 + \exp(-z))$. Für die Topologie des neuronalen Netzes verwenden wir mehrere der in Abschnitt 1.3 beschriebenen, vollständig verbundenen neuronalen Netze, führen deren Berechnungen parallel aus und bilden abschließend eine Linearkombination ihrer Ausgaben. Eine beispielhafte Darstellung eines solchen Netzes mit zwei verdeckten Schichten und drei Neuronen pro Schicht ist in Abbildung 2.1 zu finden.

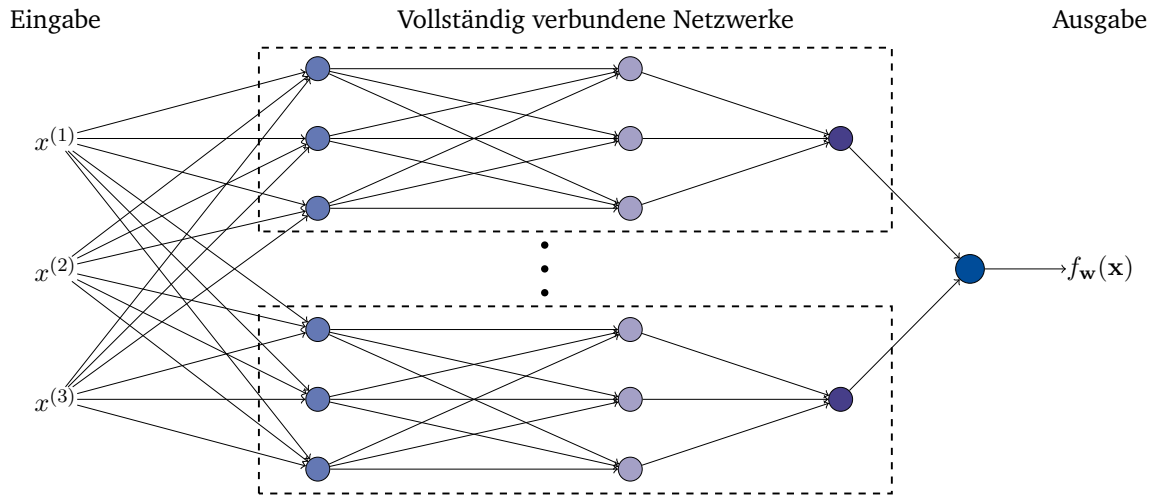


Abbildung 2.1: Neuronales Netz, bestehend aus parallel berechneten, vollständig verbundenen Netzen

Formal lässt sich die Ausgabe des Netzes durch

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^{\widehat{K}_n} w_{1,1,k}^{(L)} \cdot f_{\mathbf{w},k,1}^{(L)}(\mathbf{x}) \quad (2.1)$$

darstellen, wobei die Gewichte $w_{1,1,1}^{(L)}, \dots, w_{1,1,\widehat{K}_n}^{(L)} \in \mathbb{R}$ sind. Die vollständig verbundenen neuronalen Netze $f_{\mathbf{w},k,1}^{(L)}$ sind rekursiv durch

$$f_{\mathbf{w},k,i}^{(l)}(\mathbf{x}) = \sigma \left(\sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot f_{\mathbf{w},k,j}^{(l-1)}(\mathbf{x}) + w_{k,i,0}^{(l-1)} \right) \quad (2.2)$$

mit Gewichten $w_{k,i,0}^{(l-1)}, \dots, w_{k,i,r}^{(l-1)} \in \mathbb{R}$ für $l = 2, \dots, L$ und

$$f_{\mathbf{w},k,i}^{(1)}(\mathbf{x}) = \sigma \left(\sum_{j=1}^d w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)} \right) \quad (2.3)$$

mit Gewichten $w_{k,i,0}^{(0)}, \dots, w_{k,i,d}^{(0)} \in \mathbb{R}$ definiert.

Mit $f_{\mathbf{w},k,i}^{(l)}(\mathbf{x})$ wird das i -te Neuron aus der l -ten verdeckten Schicht bezeichnet. Der Index k gibt an, in welchem der \widehat{K}_n vollständig verbundenen neuronalen Netze wir uns befinden. Mit $w_{k,i,j}^{(l)}$ wird das Gewicht im k -ten vollständig verbundenen neuronalen Netz zwischen Neuron j in der $(l-1)$ -ten und Neuron i in der l -ten verdeckten Schicht bezeichnet.

Die Überparametrisierung des Netzwerks wird durch die geeignete Wahl von \widehat{K}_n , also der Anzahl von parallel berechneten vollständig verbundenen neuronalen Netzen, in Theorem 1 erzeugt.

Den Schätzer erhalten wir, indem wir die Gewichte mithilfe des Gradientenabstiegs trainieren. Hierfür wird der Gewichtsvektor $\mathbf{w}^{(0)} = ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l}$ wie folgt initialisiert: Die Gewichte der Ausgabeschicht werden auf 0 gesetzt, das bedeutet

$$(\mathbf{w}^{(0)})_{1,1,k}^{(L)} = 0 \quad \text{für } k = 1, \dots, \widehat{K}_n. \quad (2.4)$$

Die anderen Komponenten des Vektors $\mathbf{w}^{(0)}$ werden derart gewählt, dass diese unabhängig sind. Die Gewichte $(\mathbf{w}^{(0)})_{k,i,j}^{(l)}$ für $l \in \{1, \dots, L - 1\}$ werden zudem gleichverteilt aus dem Intervall $[-20d \cdot (\log n)^2, 20d \cdot (\log n)^2]$ gewählt, während die Gewichte der ersten Schicht $(\mathbf{w}^{(0)})_{k,i,j}^{(0)}$ gleichverteilt aus dem Intervall $[-8d \cdot (\log n)^2 \cdot n^\tau, 8d \cdot (\log n)^2 \cdot n^\tau]$ für ein festes $\tau > 0$ stammen.

Durch den Gradientenabstieg wird eine Folge von Gewichten $\mathbf{w}^{(t)}$ für $t = 1, \dots, t_n$ berechnet. Die Anzahl der Schritte des Gradientenabstiegs t_n hängt von der Größe n des Stichprobenumfangs \mathcal{D}_n ab und wird in Theorem 1 gewählt. Wie im Abschnitt 1.4 bereits erläutert, erfolgt die Aktualisierung der Gewichtsvektoren durch

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)}).$$

Die Verlustfunktion ist hierbei durch das empirische L_2 -Risiko bezüglich des Netzes $f_{\mathbf{w}}$ auf den Trainingsdaten gegeben, das heißt

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{s=1}^n |f_{\mathbf{w}}(\mathbf{X}_s) - Y_s|^2.$$

Für die Schrittweite verwenden wir den Kehrwert der Anzahl der Gradientenschritte, also

$$\lambda_n = \frac{1}{t_n}.$$

Als Schätzer wählen wir dann das neuronale Netz mit den Gewichten, die sich nach t_n Gradientenschritten ergeben, wobei wir die Ausgabe des Netzes auf das Intervall $[-\beta_n, \beta_n]$ beschränken. Der Schätzer ist für $\beta_n = c_5 \cdot \log n$ gegeben durch

$$m_n(\mathbf{x}) = T_{\beta_n} f_{\mathbf{w}^{(t_n)}}(\mathbf{x}). \quad (2.5)$$

2.2 Universelle Konsistenz des Schätzers

Das folgende Theorem zeigt die universelle Konsistenz eines überparametrisierten tiefen Neuronale-Netze-Schätzers, der mittels Gradientenabstieg trainiert wird. Es weist nach, dass der Schätzer unter geeigneten Voraussetzungen und bei wachsender Stichprobengröße n für jede Verteilung von (\mathbf{X}, Y) gegen die Regressionsfunktion konvergiert, sofern $\text{supp}(\mathbf{X})$ beschränkt ist.

Theorem 1. Sei $\sigma(z) = 1/(1 + \exp(-z))$ die logistische Sigmoidfunktion. Seien weiter $\widehat{K}_n, L, r, t_n \in \mathbb{N}$ mit $L \geq 2$ und $r \geq 2d$ sowie $\tau \in \mathbb{R}^+$ mit $\tau = 1/(d + 1)$. Zudem gelte

$$\frac{\widehat{K}_n}{n^\kappa} \rightarrow 0 \quad (n \rightarrow \infty) \quad (2.6)$$

für ein $\kappa > 0$ und

$$\frac{\widehat{K}_n}{n^{r+2}} \rightarrow \infty \quad (n \rightarrow \infty). \quad (2.7)$$

Der Neuronale-Netze-Schätzer $m_n(\mathbf{x})$ sei definiert wie in (2.5) mit

$$\beta_n = c_5 \cdot \log n$$

sowie

$$\lambda_n = \frac{1}{t_n} \quad \text{und} \quad t_n = \lceil c_6 \cdot C_n \rceil \quad (2.8)$$

für eine Konstante $c_6 \geq 2$, wobei $C_n > 0$ gegeben ist durch

$$C_n \geq \widehat{K}_n^{3/2} \cdot (\log n)^{6L+5}.$$

Dann konvergiert

$$\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \rightarrow 0 \quad (n \rightarrow \infty)$$

für jede Verteilung von (\mathbf{X}, Y) , deren $\text{supp}(\mathbf{X})$ beschränkt ist und für die $\mathbf{E}Y^2 < \infty$ gilt.

Das Theorem zeigt, dass die Verwendung des empirischen L_2 -Risikos mit einer geeigneten Anzahl von Gradientenschritten und einer geeigneten Schrittweite eine gute Generalisierung auf neuen, unabhängigen Daten liefert. Die Voraussetzung (2.7) stellt hierbei sicher, dass das neuronale Netz überparametrisiert ist.

Der Beweis von Theorem 1 (siehe Unterabschnitt 2.2.4) basiert darauf, dass sich die inneren Gewichte während des Gradientenabstiegs nur geringfügig ändern und der Gradientenabstieg in der Lage ist, geeignete Werte für die äußeren Gewichte des Netzes zu bestimmen. Die inneren Gewichte werden mit hoher Wahrscheinlichkeit so gewählt, dass geeignete Werte für die äußeren Gewichte existieren, für die das entsprechende neuronale Netz ein geringes empirisches L_2 -Risiko erzielt. Der Beweis besteht dabei aus drei wichtigen Aspekten:

Zunächst widmen wir uns der Analyse des Optimierungsfehlers. Dafür verwenden wir eine Methode, die sowohl die Lipschitz-Stetigkeit des Gradienten des empirischen L_2 -Risikos als auch die Konvexität des empirischen L_2 -Risikos ausnutzt (siehe Lemma 1).

Um den Generalisierungsfehler des Schätzers kontrollieren zu können, benötigen wir eine Aussage über die Komplexität des Funktionsraums der überparametrisierten neuronalen Netze, welche wir in Lemma 5 herleiten werden.

Für die Analyse des Approximationsfehlers haben wir die gute Approximationseigenschaft eines tiefen, vollständig verbundenen neuronalen Netzes für die Indikatorfunktion ausgenutzt. Darüber hinaus konnten wir unter geeigneten Bedingungen für die Gewichte des Netzes die Beschränktheit des überparametrisierten neuronalen Netzes nachweisen. Die Kombination dieser beiden Eigenschaften führt zu einer insgesamt guten Approximationseigenschaft des überparametrisierten tiefen neuronalen Netzes (siehe Lemma 9).

2.2.1 Analyse des Gradientenabstiegs

Das erste Lemma ermöglicht es uns, den Gradientenabstieg zu analysieren und zeigt zudem, dass die Gewichte während des Trainingsprozesses nahe an ihren Startwerten bleiben.

Lemma 1. Seien $d_1, d_2 \in \mathbb{N}$. Sei $(\mathbf{u}_0, \mathbf{v}_0) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ und sei $F : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}^+$ eine stetig differenzierbare Funktion. Wir definieren im Folgenden die Menge A durch

$$A := \left\{ (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} : \|(\mathbf{u}, \mathbf{v}) - (\mathbf{u}_0, \mathbf{v}_0)\| \leq 2 \cdot \sqrt{F(\mathbf{u}_0, \mathbf{v}_0) + 1} \right\}.$$

Des Weiteren sei

$$\mathbf{u} \mapsto F(\mathbf{u}, \mathbf{v})$$

für alle $\mathbf{v} \in \mathbb{R}^{d_2}$ konvex. Für $C_n, D_n > 0$ gelten die folgenden Bedingungen

$$\|(\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}, \mathbf{v})\| \leq C_n \quad (2.9)$$

für alle $(\mathbf{u}, \mathbf{v}) \in A$,

$$\|(\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_1, \mathbf{v}_1) - (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_2, \mathbf{v}_2)\| \leq C_n \cdot \|(\mathbf{u}_1, \mathbf{v}_1) - (\mathbf{u}_2, \mathbf{v}_2)\| \quad (2.10)$$

für alle $(\mathbf{u}_1, \mathbf{v}_1), (\mathbf{u}_2, \mathbf{v}_2) \in A$ sowie

$$|F(\mathbf{u}^*, \mathbf{v}) - F(\mathbf{u}^*, \mathbf{v}_0)| \leq D_n \cdot \|\mathbf{u}^*\| \cdot \|\mathbf{v} - \mathbf{v}_0\| \quad (2.11)$$

für ein $\mathbf{u}^* \in \mathbb{R}^{d_1}$ und alle $\mathbf{v} \in \{\tilde{\mathbf{v}} : \|\tilde{\mathbf{v}} - \mathbf{v}_0\| \leq \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)}\}$.

Zudem sei $t_n \in \mathbb{N}$ sowie

$$t_n \geq C_n \quad \text{und} \quad \lambda_n = \frac{1}{t_n}.$$

Für $t = 0, 1, \dots, t_n - 1$ setzen wir

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \lambda_n \cdot (\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t),$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \lambda_n \cdot (\nabla_{\mathbf{v}} F)(\mathbf{u}_t, \mathbf{v}_t).$$

Dann gilt für \mathbf{u}^* aus (2.11) die folgende Ungleichung

$$F(\mathbf{u}_{t_n}, \mathbf{v}_{t_n}) \leq F(\mathbf{u}^*, \mathbf{v}_0) + D_n \cdot \|\mathbf{u}^*\| \cdot \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)} + \frac{\|\mathbf{u}^* - \mathbf{u}_0\|^2}{2} + \frac{F(\mathbf{u}_0, \mathbf{v}_0)}{t_n}.$$

Unabhängig davon, ob die Bedingung (2.11) erfüllt ist, erhalten wir

$$\|\mathbf{u}_t - \mathbf{u}_0\| \leq \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)} \quad \text{und} \quad \|\mathbf{v}_t - \mathbf{v}_0\| \leq \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)}$$

für alle $t = 0, 1, \dots, t_n$.

Beweis. Im ersten Schritt des Beweises zeigen wir, dass

$$\frac{1}{t_n} \sum_{t=0}^{t_n-1} F(\mathbf{u}_t, \mathbf{v}_t) \leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(\mathbf{u}^*, \mathbf{v}_t) + \frac{\|\mathbf{u}^* - \mathbf{u}_0\|^2}{2} + \frac{1}{2 \cdot t_n} \sum_{t=0}^{t_n-1} \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \quad (2.12)$$

gilt. Aus der Konvexität der Funktion $\mathbf{u} \mapsto F(\mathbf{u}, \mathbf{v}_t)$ ergibt sich

$$\begin{aligned} & F(\mathbf{u}_t, \mathbf{v}_t) - F(\mathbf{u}^*, \mathbf{v}_t) \\ & \leq \langle (\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t), \mathbf{u}_t - \mathbf{u}^* \rangle \\ & = \frac{1}{2 \cdot \lambda_n} \cdot 2 \cdot \langle \lambda_n \cdot (\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t), \mathbf{u}_t - \mathbf{u}^* \rangle \\ & = \frac{1}{2 \cdot \lambda_n} \cdot (-\|\mathbf{u}_t - \mathbf{u}^* - \lambda_n \cdot (\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 + \|\mathbf{u}_t - \mathbf{u}^*\|^2 + \|\lambda_n \cdot (\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2) \end{aligned}$$

$$= \frac{1}{2 \cdot \lambda_n} \cdot (-\|\mathbf{u}_{t+1} - \mathbf{u}^*\|^2 + \|\mathbf{u}_t - \mathbf{u}^*\|^2 + \lambda_n^2 \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2).$$

Zusammen mit der Wahl der Schrittweite $\lambda_n = \frac{1}{t_n}$ folgt hieraus

$$\begin{aligned} & \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(\mathbf{u}_t, \mathbf{v}_t) - \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(\mathbf{u}^*, \mathbf{v}_t) \\ &= \frac{1}{t_n} \sum_{t=0}^{t_n-1} (F(\mathbf{u}_t, \mathbf{v}_t) - F(\mathbf{u}^*, \mathbf{v}_t)) \\ &\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} \frac{1}{2 \cdot \lambda_n} \cdot (-\|\mathbf{u}_{t+1} - \mathbf{u}^*\|^2 + \|\mathbf{u}_t - \mathbf{u}^*\|^2) + \frac{1}{t_n} \sum_{t=0}^{t_n-1} \frac{\lambda_n}{2} \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \\ &= \frac{1}{2} \cdot \sum_{t=0}^{t_n-1} (\|\mathbf{u}_t - \mathbf{u}^*\|^2 - \|\mathbf{u}_{t+1} - \mathbf{u}^*\|^2) + \frac{1}{2 \cdot t_n} \sum_{t=0}^{t_n-1} \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \\ &\leq \frac{\|\mathbf{u}_0 - \mathbf{u}^*\|^2}{2} + \frac{1}{2 \cdot t_n} \sum_{t=0}^{t_n-1} \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2. \end{aligned}$$

Damit erhalten wir Ungleichung (2.12).

Im zweiten Schritt des Beweises zeigen wir, dass aus

$$\begin{aligned} & \|(\nabla_{(\mathbf{u}, \mathbf{v})} F)((\mathbf{u}_t, \mathbf{v}_t) + \tau \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t))) - (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t)\| \\ & \leq C_n \cdot \tau \cdot \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)\| \end{aligned} \quad (2.13)$$

für alle $\tau \in [0, 1]$ folgt, dass

$$F(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - F(\mathbf{u}_t, \mathbf{v}_t) \leq -\frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 - \frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2$$

gilt.

Hierfür definieren wir eine stetig differenzierbare Funktion $H : [0, 1] \rightarrow \mathbb{R}$ mit

$$H(\tau) = F((\mathbf{u}_t, \mathbf{v}_t) + \tau \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t))).$$

Aufgrund von Annahme (2.13) erhalten wir durch den Hauptsatz der Differential- und Integralrechnung

$$\begin{aligned} F(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - F(\mathbf{u}_t, \mathbf{v}_t) &= H(1) - H(0) = \int_0^1 H'(\tau) d\tau \\ &= \int_0^1 (\nabla_{(\mathbf{u}, \mathbf{v})} F)((\mathbf{u}_t, \mathbf{v}_t) + \tau \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t))) \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)) d\tau \\ &= \int_0^1 \left((\nabla_{(\mathbf{u}, \mathbf{v})} F)((\mathbf{u}_t, \mathbf{v}_t) + \tau \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t))) - (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t) \right) \\ & \quad \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)) d\tau \\ & \quad + \int_0^1 (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t) \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)) d\tau \\ &\leq \int_0^1 \|(\nabla_{(\mathbf{u}, \mathbf{v})} F)((\mathbf{u}_t, \mathbf{v}_t) + \tau \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t))) - (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t)\| \end{aligned}$$

$$\begin{aligned}
& \cdot \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)\| d\tau \\
& + (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t) \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)) \\
\leq & \int_0^1 C_n \cdot \tau \cdot \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)\|^2 d\tau + (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t) \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)) \\
= & \frac{1}{2} \cdot C_n \cdot \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)\|^2 + (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t) \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)).
\end{aligned}$$

Die Vorschrift des Gradientenverfahrens zusammen mit der Tatsache, dass die Schrittweite

$$\lambda_n = \frac{1}{t_n} \leq \frac{1}{C_n}$$

ist, liefert

$$\begin{aligned}
& \frac{1}{2} \cdot C_n \cdot \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)\|^2 + (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_t, \mathbf{v}_t) \cdot ((\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_t, \mathbf{v}_t)) \\
= & \frac{1}{2} \cdot C_n \cdot (\lambda_n^2 \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 + \lambda_n^2 \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2) - \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \\
& - \lambda_n \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \\
= & \left(\frac{1}{2} \cdot C_n \cdot \lambda_n - 1\right) \cdot \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 + \left(\frac{1}{2} \cdot C_n \cdot \lambda_n - 1\right) \cdot \lambda_n \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \\
\leq & -\frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 - \frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2.
\end{aligned}$$

Im dritten Schritt des Beweises zeigen wir, dass die Ungleichung

$$F(\mathbf{u}_1, \mathbf{v}_1) - F(\mathbf{u}_0, \mathbf{v}_0) \leq -\frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_0, \mathbf{v}_0)\|^2 - \frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_0, \mathbf{v}_0)\|^2$$

gilt. Um diese Ungleichung zu belegen, werden wir nachweisen, dass Ungleichung (2.13) aus dem zweiten Beweisschritt für $t = 0$ erfüllt ist.

Da $(\mathbf{u}_0, \mathbf{v}_0) \in A$ folgt aus Voraussetzung (2.9), dass

$$\|(\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_0, \mathbf{v}_0)\| \leq C_n$$

erfüllt ist. Dies impliziert für jedes $\tau \in [0, 1]$ zusammen mit der Vorschrift des Gradientenverfahrens und $\lambda_n \leq \frac{1}{C_n}$, dass

$$\|(\mathbf{u}_0, \mathbf{v}_0) + \tau \cdot ((\mathbf{u}_1, \mathbf{v}_1) - (\mathbf{u}_0, \mathbf{v}_0)) - (\mathbf{u}_0, \mathbf{v}_0)\| \leq \tau \cdot \|\lambda_n \cdot (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_0, \mathbf{v}_0)\| \leq \lambda_n \cdot C_n \leq 1$$

gilt, woraus

$$(\mathbf{u}_0, \mathbf{v}_0) + \tau \cdot ((\mathbf{u}_1, \mathbf{v}_1) - (\mathbf{u}_0, \mathbf{v}_0)) \in A$$

folgt. Somit können wir aus Voraussetzung (2.10) folgern, dass Ungleichung (2.13) aus dem zweiten Beweisschritt für $t = 0$ erfüllt ist, denn

$$\|(\nabla_{(\mathbf{u}, \mathbf{v})} F)((\mathbf{u}_0, \mathbf{v}_0) + \tau \cdot ((\mathbf{u}_1, \mathbf{v}_1) - (\mathbf{u}_0, \mathbf{v}_0))) - (\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_0, \mathbf{v}_0)\| \leq C_n \cdot \tau \cdot \|(\mathbf{u}_1, \mathbf{v}_1) - (\mathbf{u}_0, \mathbf{v}_0)\|.$$

Wenden wir darauf nun die Behauptung aus Beweisschritt zwei für $t = 0$ an, so erhalten wir die Aussage des dritten Beweisschrittes.

Im vierten Schritt des Beweises zeigen wir mittels Induktion über t und \tilde{t} , dass

$$F(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - F(\mathbf{u}_t, \mathbf{v}_t) \leq -\frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 - \frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \quad (2.14)$$

für alle $t \in \{0, 1, \dots, t_n - 1\}$ gilt, und dass

$$\|\mathbf{u}_{\tilde{t}} - \mathbf{u}_0\| \leq \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)} \quad \text{und} \quad \|\mathbf{v}_{\tilde{t}} - \mathbf{v}_0\| \leq \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)} \quad (2.15)$$

für alle $\tilde{t} \in \{0, 1, \dots, t_n\}$ erfüllt ist.

Für $t = 0$ wurde die Aussage von (2.14) bereits im dritten Beweisschritt gezeigt, und die Ungleichungen in (2.15) sind für $\tilde{t} = 0$ direkt erfüllt. Wir nehmen nun an, dass Behauptung (2.14) für ein $t \in \{0, 1, \dots, t_n - 2\}$ und Behauptung (2.15) für ein $\tilde{t} \in \{0, 1, \dots, t_n - 1\}$ erfüllt sind. Dann ist $(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) \in A$, woraus aus Voraussetzung (2.9) die Ungleichung

$$\|(\nabla_{(\mathbf{u}, \mathbf{v})} F)(\mathbf{u}_{t+1}, \mathbf{v}_{t+1})\| \leq C_n$$

folgt. Diese impliziert für alle $\tau \in [0, 1]$ und $t \in \{0, \dots, t_n - 2\}$ die folgende Abschätzung

$$\begin{aligned} & \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) + \tau \cdot ((\mathbf{u}_{t+2}, \mathbf{v}_{t+2}) - (\mathbf{u}_{t+1}, \mathbf{v}_{t+1})) - (\mathbf{u}_0, \mathbf{v}_0)\| \\ & \leq \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_0, \mathbf{v}_0)\| + \|\lambda_n \cdot \nabla_{(\mathbf{u}, \mathbf{v})} F(\mathbf{u}_{t+1}, \mathbf{v}_{t+1})\| \\ & \leq \|(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) - (\mathbf{u}_0, \mathbf{v}_0)\| + \lambda_n \cdot C_n \\ & \leq \sqrt{\|\mathbf{u}_{t+1} - \mathbf{u}_0\|^2 + \|\mathbf{v}_{t+1} - \mathbf{v}_0\|^2} + \lambda_n \cdot C_n \\ & \leq 2 \cdot \sqrt{F(\mathbf{u}_0, \mathbf{v}_0)} + \lambda_n \cdot C_n \\ & \leq 2 \cdot \sqrt{F(\mathbf{u}_0, \mathbf{v}_0)} + 1. \end{aligned}$$

Damit ist auch $(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) + \tau \cdot ((\mathbf{u}_{t+2}, \mathbf{v}_{t+2}) - (\mathbf{u}_{t+1}, \mathbf{v}_{t+1})) \in A$. Wie im vorherigen Beweisschritt können wir aus Voraussetzung (2.10) folgern, dass Ungleichung (2.13) erfüllt ist. Wegen Beweisschritt zwei folgt hieraus

$$F(\mathbf{u}_{t+2}, \mathbf{v}_{t+2}) - F(\mathbf{u}_{t+1}, \mathbf{v}_{t+1}) \leq -\frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_{t+1}, \mathbf{v}_{t+1})\|^2 - \frac{1}{2} \cdot \lambda_n \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_{t+1}, \mathbf{v}_{t+1})\|^2.$$

Daher erhalten wir mit der Ungleichung von Cauchy-Schwarz, der Vorschrift des Gradientenverfahrens sowie $\tilde{t} \in \{0, 1, \dots, t_n - 1\}$, dass

$$\begin{aligned} \|\mathbf{u}_{\tilde{t}+1} - \mathbf{u}_0\| & \leq \sum_{s=0}^{\tilde{t}} \|\mathbf{u}_{s+1} - \mathbf{u}_s\| \\ & \leq \sqrt{(\tilde{t} + 1) \cdot \sum_{s=0}^{\tilde{t}} \|\mathbf{u}_{s+1} - \mathbf{u}_s\|^2} \\ & = \sqrt{(\tilde{t} + 1) \cdot \sum_{s=0}^{\tilde{t}} \lambda_n^2 \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_s, \mathbf{v}_s)\|^2} \\ & \leq \sqrt{2 \cdot (\tilde{t} + 1) \cdot \lambda_n \cdot \sum_{s=0}^{\tilde{t}} (F(\mathbf{u}_s, \mathbf{v}_s) - F(\mathbf{u}_{s+1}, \mathbf{v}_{s+1}))} \end{aligned}$$

$$\begin{aligned}
&\leq \sqrt{2 \cdot (\tilde{t} + 1) \cdot \frac{1}{t_n} \cdot F(\mathbf{u}_0, \mathbf{v}_0)} \\
&\leq \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)}
\end{aligned}$$

und

$$\begin{aligned}
\|\mathbf{v}_{\tilde{t}+1} - \mathbf{v}_0\| &\leq \sum_{s=0}^{\tilde{t}} \|\mathbf{v}_{s+1} - \mathbf{v}_s\| \\
&\leq \sqrt{(\tilde{t} + 1) \cdot \sum_{s=0}^{\tilde{t}} \|\mathbf{v}_{s+1} - \mathbf{v}_s\|^2} \\
&= \sqrt{(\tilde{t} + 1) \cdot \sum_{s=0}^{\tilde{t}} \lambda_n^2 \cdot \|(\nabla_{\mathbf{v}} F)(\mathbf{u}_s, \mathbf{v}_s)\|^2} \\
&\leq \sqrt{2 \cdot (\tilde{t} + 1) \cdot \lambda_n \cdot \sum_{s=0}^{\tilde{t}} (F(\mathbf{u}_s, \mathbf{v}_s) - F(\mathbf{u}_{s+1}, \mathbf{v}_{s+1}))} \\
&\leq \sqrt{2 \cdot (\tilde{t} + 1) \cdot \frac{1}{t_n} \cdot F(\mathbf{u}_0, \mathbf{v}_0)} \\
&\leq \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)}
\end{aligned}$$

gelten.

Im *fünften Schritt des Beweises* zeigen wir nun die Behauptung des Lemmas. Aus dem vierten Beweisschritt folgt, dass die Funktion F monoton fallend ist. Daher gilt

$$F(\mathbf{u}_{t_n}, \mathbf{v}_{t_n}) \leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(\mathbf{u}_t, \mathbf{v}_t).$$

Aus Ungleichung (2.12) aus dem ersten Beweisschritt können wir damit folgern, dass

$$\begin{aligned}
F(\mathbf{u}_{t_n}, \mathbf{v}_{t_n}) &\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(\mathbf{u}_t, \mathbf{v}_t) \\
&\leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} F(\mathbf{u}^*, \mathbf{v}_t) + \frac{\|\mathbf{u}^* - \mathbf{u}_0\|^2}{2} + \frac{1}{2 \cdot t_n} \sum_{t=0}^{t_n-1} \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \\
&\leq F(\mathbf{u}^*, \mathbf{v}_0) + \frac{1}{t_n} \sum_{t=0}^{t_n-1} |F(\mathbf{u}^*, \mathbf{v}_t) - F(\mathbf{u}^*, \mathbf{v}_0)| \\
&\quad + \frac{\|\mathbf{u}^* - \mathbf{u}_0\|^2}{2} + \frac{1}{2 \cdot t_n} \sum_{t=0}^{t_n-1} \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 \tag{2.16}
\end{aligned}$$

gilt. Mit Voraussetzung (2.11) und den Ungleichungen in (2.15) aus dem vierten Schritt des Beweises erhalten wir dann

$$\frac{1}{t_n} \sum_{t=0}^{t_n-1} |F(\mathbf{u}^*, \mathbf{v}_t) - F(\mathbf{u}^*, \mathbf{v}_0)| \leq \frac{1}{t_n} \sum_{t=0}^{t_n-1} D_n \cdot \|\mathbf{u}^*\| \cdot \|\mathbf{v}_t - \mathbf{v}_0\|$$

$$\leq D_n \cdot \|\mathbf{u}^*\| \cdot \sqrt{2 \cdot F(\mathbf{u}_0, \mathbf{v}_0)}. \quad (2.17)$$

Zudem ergibt sich mit Ungleichung (2.14) aus dem vierten Beweisschritt

$$\begin{aligned} \sum_{t=0}^{t_n-1} \lambda_n \cdot \|(\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)\|^2 &\leq 2 \cdot \sum_{t=0}^{t_n-1} (F(\mathbf{u}_t, \mathbf{v}_t) - F(\mathbf{u}_{t+1}, \mathbf{v}_{t+1})) \\ &\leq 2 \cdot F(\mathbf{u}_0, \mathbf{v}_0). \end{aligned} \quad (2.18)$$

Setzen wir nun (2.17) und (2.18) in Ungleichung (2.16) ein, so erhalten wir die Aussage von Lemma 1. \square

Um Lemma 1 im Beweis von Theorem 1 anwenden zu können, benötigen wir die beiden folgenden Lemmata. Mit ihrer Hilfe können wir zeigen, dass die Voraussetzungen von Lemma 1 im Beweis von Theorem 1 erfüllt sind.

Lemma 2. *Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ eine beschränkte und differenzierbare Funktion, deren Ableitung ebenfalls beschränkt ist. Des Weiteren seien $\alpha_n \geq 1$, $t_n \geq C_n$, $\gamma_n^* \geq 1$, $B_n \geq 1$ und $d, r \in \mathbb{N}$ mit $r \geq 2d$. Ist zudem*

$$|w_{1,1,k}^{(L)}| \leq \gamma_n^* \quad \text{für } k \in \{1, \dots, \widehat{K}_n\}, \quad (2.19)$$

$$|w_{k,i,j}^{(l)}| \leq B_n \quad \text{für } l \in \{1, \dots, L-1\}, k \in \{1, \dots, \widehat{K}_n\}, i, j \in \{1, \dots, r\} \quad (2.20)$$

und

$$\|\mathbf{w} - \mathbf{v}\|_\infty^2 \leq \frac{8 \cdot t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\} \quad (2.21)$$

erfüllt, so gilt für $\mathbf{X}_1, \dots, \mathbf{X}_n \in [-\alpha_n, \alpha_n]^d$ die Ungleichung

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq c_7 \cdot \widehat{K}_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2 \cdot \sqrt{\frac{t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\}}.$$

Beweis. Durch die Anwendung der Kettenregel sowie der Ungleichung von Cauchy-Schwarz ergibt sich

$$\begin{aligned} \|\nabla_{\mathbf{w}} F_n(\mathbf{w})\|^2 &= \sum_{k,i,j,l} \left(\frac{2}{n} \sum_{s=1}^n (f_{\mathbf{w}}(\mathbf{X}_s) - Y_s) \cdot \frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \\ &\leq 4 \cdot \sum_{k,i,j,l} \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}}(\mathbf{X}_s) - Y_s)^2 \cdot \frac{1}{n} \sum_{s=1}^n \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \\ &\leq 4 \cdot \widehat{K}_n \cdot L \cdot r^2 \cdot d \cdot \max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \cdot \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}}(\mathbf{X}_s) - Y_s)^2. \end{aligned} \quad (2.22)$$

Um die Aussage des Lemmas zu zeigen, werden wir im Folgenden die beiden letzten Faktoren beschränken. Die partielle Ableitung des neuronalen Netzes $f_{\mathbf{w}}$ bezüglich des Gewichts $w_{k,i,j}^{(l)}$ ist gegeben durch

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(\mathbf{x}) = \sum_{s_{l+2}=1}^r \cdots \sum_{s_{L-1}=1}^r f_{\mathbf{w},k,j}^{(l)}(\mathbf{x}) \cdot \sigma' \left(\sum_{t=1}^r w_{k,i,t}^{(l)} \cdot f_{\mathbf{w},k,t}^{(l)}(\mathbf{x}) + w_{k,i,0}^{(l)} \right) \cdot w_{k,s_{l+2},i}^{(l+1)}$$

$$\begin{aligned}
& \cdot \sigma' \left(\sum_{t=1}^r w_{k,s_{l+2},t}^{(l+1)} \cdot f_{\mathbf{w},k,t}^{(l+1)}(\mathbf{x}) + w_{k,s_{l+2},0}^{(l+1)} \right) \cdot w_{k,s_{l+3},s_{l+2}}^{(l+2)} \\
& \cdot \sigma' \left(\sum_{t=1}^r w_{k,s_{l+3},t}^{(l+2)} \cdot f_{\mathbf{w},k,t}^{(l+2)}(\mathbf{x}) + w_{k,s_{l+3},0}^{(l+2)} \right) \cdots w_{k,s_{L-1},s_{L-2}}^{(L-2)} \\
& \cdot \sigma' \left(\sum_{t=1}^r w_{k,s_{L-1},t}^{(L-2)} \cdot f_{\mathbf{w},k,t}^{(L-2)}(\mathbf{x}) + w_{k,s_{L-1},0}^{(L-2)} \right) \cdot w_{k,1,s_{L-1}}^{(L-1)} \\
& \cdot \sigma' \left(\sum_{t=1}^r w_{k,1,t}^{(L-1)} \cdot f_{\mathbf{w},k,t}^{(L-1)}(\mathbf{x}) + w_{k,1,0}^{(L-1)} \right) \cdot w_{1,1,k}^{(L)}, \tag{2.23}
\end{aligned}$$

wobei

$$f_{\mathbf{w},k,j}^{(0)}(\mathbf{x}) = \begin{cases} x^{(j)}, & \text{falls } j \in \{1, \dots, d\} \\ 1, & \text{falls } j = 0 \end{cases}$$

und

$$f_{\mathbf{w},k,0}^{(l)}(\mathbf{x}) = 1 \quad \text{für } l = 1, \dots, L-1.$$

Zusammen mit den Voraussetzungen (2.19) und (2.20) ergibt sich damit

$$\max_{k,i,j,l,s} \left(\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \leq c_8 \cdot r^{2L} \cdot \max\{\|\sigma\|_{\infty}^2, 1\} \cdot \max\{\|\sigma'\|_{\infty}^{2L}, 1\} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2. \tag{2.24}$$

Für die Abschätzung des zweiten Faktors bezeichnen wir mit $f_{\mathbf{v}}$ das neuronale Netz, welches durch die Gewichte $(v_{k,i,j}^{(l)})_{k,i,j,l}$ gemäß (2.1)–(2.3) definiert ist. Durch eine geeignete Nulladdition folgt daraus die Ungleichung

$$\frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}}(\mathbf{X}_s) - Y_s)^2 \leq 2 \cdot F_n(\mathbf{v}) + \frac{2}{n} \sum_{s=1}^n (f_{\mathbf{v}}(\mathbf{X}_s) - f_{\mathbf{w}}(\mathbf{X}_s))^2.$$

Es bleibt noch zu zeigen, dass

$$|f_{\mathbf{w}}(\mathbf{x}) - f_{\mathbf{v}}(\mathbf{x})| \leq c_9 \cdot \widehat{K}_n \cdot \max\{\|\sigma'\|_{\infty}^L, 1\} \cdot \gamma_n^* \cdot (2r+1)^L \cdot B_n^L \cdot \alpha_n \cdot \|\mathbf{w} - \mathbf{v}\|_{\infty} \cdot \max\{\|\sigma\|_{\infty}, 1\}$$

gilt.

Hierfür zeigen wir im Folgenden per Induktion, dass

$$\begin{aligned}
& \left| f_{\mathbf{w},k,i}^{(l)}(\mathbf{x}) - f_{\mathbf{v},k,i}^{(l)}(\mathbf{x}) \right| \\
& \leq l \cdot c_{10} \cdot \max\{\|\sigma'\|_{\infty}^l, 1\} \cdot (2r+1)^l \cdot B_n^l \cdot \alpha_n \cdot \max_{i,j,s:s < L} |w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)}| \cdot \max\{\|\sigma\|_{\infty}, 1\} \tag{2.25}
\end{aligned}$$

für $l \in \{1, \dots, L\}$, $k \in \{1, \dots, \widehat{K}_n\}$ und $\mathbf{x} \in [-\alpha_n, \alpha_n]^d$ erfüllt ist.

Sei hierfür $k \in \{1, \dots, \widehat{K}_n\}$. Die Funktion σ ist nach Voraussetzung differenzierbar und besitzt eine beschränkte Ableitung. Damit ist σ Lipschitz-stetig mit einer Lipschitz-Konstanten $\|\sigma'\|_{\infty}$. Wegen $r \geq 2d$ folgt daher

$$\left| f_{\mathbf{w},k,i}^{(1)}(\mathbf{x}) - f_{\mathbf{v},k,i}^{(1)}(\mathbf{x}) \right| \leq \|\sigma'\|_{\infty} \cdot \left(\sum_{j=1}^d |w_{k,i,j}^{(0)} - v_{k,i,j}^{(0)}| \cdot |x^{(j)}| + |w_{k,i,0}^{(0)} - v_{k,i,0}^{(0)}| \right)$$

$$\leq \|\sigma'\|_\infty \cdot (2r + 1) \cdot \alpha_n \cdot \max_{i,j,s:s < L} |w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)}|$$

für $\mathbf{x} \in [-\alpha_n, \alpha_n]^d$.

Angenommen, die Induktionsannahme (2.25) ist für ein $l - 1$ mit $l \in \{2, \dots, L\}$ erfüllt, dann ist

$$\begin{aligned} & \left| f_{\mathbf{w},k,i}^{(l)}(\mathbf{x}) - f_{\mathbf{v},k,i}^{(l)}(\mathbf{x}) \right| \\ & \leq \|\sigma'\|_\infty \cdot \left(\sum_{j=1}^r |w_{k,i,j}^{(l-1)}| \cdot \left| f_{\mathbf{w},k,j}^{(l-1)}(\mathbf{x}) - f_{\mathbf{v},k,j}^{(l-1)}(\mathbf{x}) \right| \right. \\ & \quad \left. + \sum_{j=1}^r \left| w_{k,i,j}^{(l-1)} - v_{k,i,j}^{(l-1)} \right| \cdot \left| f_{\mathbf{v},k,j}^{(l-1)}(\mathbf{x}) \right| + \left| w_{k,i,0}^{(l-1)} - v_{k,i,0}^{(l-1)} \right| \right) \\ & \leq \|\sigma'\|_\infty \cdot \left(r \cdot B_n \cdot \max_{j=1,\dots,r} \left| f_{\mathbf{w},k,j}^{(l-1)}(\mathbf{x}) - f_{\mathbf{v},k,j}^{(l-1)}(\mathbf{x}) \right| \right. \\ & \quad \left. + (r + 1) \cdot \max_{i,j,s:s < L} \left| w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)} \right| \cdot \max\{\|\sigma\|_\infty, 1\} \right) \\ & \leq l \cdot c_{10} \cdot \max\{\|\sigma'\|_\infty^l, 1\} \cdot (2r + 1)^l \cdot B_n^l \cdot \alpha_n \cdot \max_{i,j,s:s < L} \left| w_{k,i,j}^{(s)} - v_{k,i,j}^{(s)} \right| \cdot \max\{\|\sigma\|_\infty, 1\} \end{aligned}$$

für $\mathbf{x} \in [-\alpha_n, \alpha_n]$.

Für die Differenz der Ausgaben der beiden neuronalen Netze erhalten wir damit

$$\begin{aligned} & |f_{\mathbf{w}}(\mathbf{x}) - f_{\mathbf{v}}(\mathbf{x})| \\ & = \left| \sum_{k=1}^{\widehat{K}_n} w_{1,1,k}^{(L)} \cdot f_{\mathbf{w},k,1}^{(L)}(\mathbf{x}) - \sum_{k=1}^{\widehat{K}_n} v_{1,1,k}^{(L)} \cdot f_{\mathbf{v},k,1}^{(L)}(\mathbf{x}) \right| \\ & \leq \left| \sum_{k=1}^{\widehat{K}_n} w_{1,1,k}^{(L)} \cdot \left(f_{\mathbf{w},k,1}^{(L)}(\mathbf{x}) - f_{\mathbf{v},k,1}^{(L)}(\mathbf{x}) \right) \right| + \left| \sum_{k=1}^{\widehat{K}_n} \left(w_{1,1,k}^{(L)} - v_{1,1,k}^{(L)} \right) \cdot f_{\mathbf{v},k,1}^{(L)}(\mathbf{x}) \right| \\ & \leq \widehat{K}_n \cdot \max_{k=1,\dots,\widehat{K}_n} |w_{1,1,k}^{(L)}| \cdot \max_{k=1,\dots,\widehat{K}_n} \left| f_{\mathbf{w},k,1}^{(L)}(\mathbf{x}) - f_{\mathbf{v},k,1}^{(L)}(\mathbf{x}) \right| + \widehat{K}_n \cdot \max_{k=1,\dots,\widehat{K}_n} \left| w_{1,1,k}^{(L)} - v_{1,1,k}^{(L)} \right| \cdot \max\{\|\sigma\|_\infty, 1\} \\ & \leq c_9 \cdot \widehat{K}_n \cdot \gamma_n^* \cdot \max\{\|\sigma'\|_\infty^L, 1\} \cdot (2r + 1)^L \cdot B_n^L \cdot \alpha_n \cdot \|\mathbf{w} - \mathbf{v}\|_\infty \cdot \max\{\|\sigma\|_\infty, 1\}. \end{aligned} \quad (2.26)$$

Mit Voraussetzung (2.21) folgt hieraus

$$\begin{aligned} & \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}}(\mathbf{X}_s) - Y_s)^2 \\ & \leq 2 \cdot F_n(\mathbf{v}) + \frac{2}{n} \sum_{s=1}^n (f_{\mathbf{v}}(\mathbf{X}_s) - f_{\mathbf{w}}(\mathbf{X}_s))^2 \\ & \leq 2 \cdot F_n(\mathbf{v}) + 2 \cdot c_9^2 \cdot \widehat{K}_n^2 \cdot (\gamma_n^*)^2 \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot (2r + 1)^{2L} \cdot B_n^{2L} \cdot \alpha_n^2 \\ & \quad \cdot \max\{\|\sigma\|_\infty, 1\}^2 \cdot \frac{8 \cdot t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\}. \end{aligned} \quad (2.27)$$

Setzen wir nun (2.24) und (2.27) in Ungleichung (2.22) vom Beginn des Beweises ein, so erhalten wir die Aussage des Lemmas. \square

Lemma 3. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ eine beschränkte und differenzierbare Funktion, deren Ableitung Lipschitz-stetig mit Lipschitz-Konstante σ'_{Lip} und beschränkt ist. Seien weiter $t_n \in \mathbb{N}$ und $C_n > 0$ mit $t_n \geq C_n$ sowie $\gamma_n^* \geq 1$, $B_n \geq 1$ und $d, r \in \mathbb{N}$ mit $r \geq 2d$. Zudem sei $\alpha_n \geq 1$ und die Zufallsvariablen $\mathbf{X}_1, \dots, \mathbf{X}_n \in [-\alpha_n, \alpha_n]^d$. Angenommen, es gelte

$$\left| \max \left\{ (\mathbf{w}_1)_{1,1,k}^{(L)}, (\mathbf{w}_2)_{1,1,k}^{(L)} \right\} \right| \leq \gamma_n^* \quad \text{für } k \in \{1, \dots, \widehat{K}_n\}, \quad (2.28)$$

$$\left| \max \left\{ (\mathbf{w}_1)_{k,i,j}^{(l)}, (\mathbf{w}_2)_{k,i,j}^{(l)} \right\} \right| \leq B_n \quad \text{für } l \in \{1, \dots, L-1\}, k \in \{1, \dots, \widehat{K}_n\}, i, j \in \{1, \dots, r\} \quad (2.29)$$

und

$$\|\mathbf{w}_2 - \mathbf{v}\|^2 \leq \frac{8 \cdot t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\}, \quad (2.30)$$

dann ist die Ungleichung

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}_1) - (\nabla_{\mathbf{w}} F_n)(\mathbf{w}_2)\| \leq c_{11} \cdot \widehat{K}_n^{3/2} \cdot B_n^{3L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^3 \cdot \sqrt{\frac{t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\}} \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|$$

erfüllt.

Beweis. Wir beginnen damit, die Norm der Differenz der Gradienten der Verlustfunktion nach oben abzuschätzen. Durch Anwenden der Ungleichung $(a - b)^2 \leq 2 \cdot a^2 + 2 \cdot b^2$ für $a, b \in \mathbb{R}$ und der Ungleichung von Cauchy-Schwarz erhalten wir

$$\begin{aligned} & \| \nabla_{\mathbf{w}} F_n(\mathbf{w}_1) - \nabla_{\mathbf{w}} F_n(\mathbf{w}_2) \|^2 \\ &= \sum_{k,i,j,l} \left(\frac{2}{n} \sum_{s=1}^n (f_{\mathbf{w}_1}(\mathbf{X}_s) - Y_s) \cdot \frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) - \left(\frac{2}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(\mathbf{X}_s) - Y_s) \cdot \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right) \right)^2 \\ &\leq 8 \cdot \sum_{k,i,j,l} \left(\frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(\mathbf{X}_s) - f_{\mathbf{w}_1}(\mathbf{X}_s)) \cdot \frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \\ &\quad + 8 \cdot \sum_{k,i,j,l} \left(\frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(\mathbf{X}_s) - Y_s) \cdot \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right) \right)^2 \\ &\leq 8 \cdot \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(\mathbf{X}_s) - f_{\mathbf{w}_1}(\mathbf{X}_s))^2 \cdot \sum_{k,i,j,l} \max_{s=1, \dots, n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \\ &\quad + 8 \cdot \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(\mathbf{X}_s) - Y_s)^2 \cdot \sum_{k,i,j,l} \max_{s=1, \dots, n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2. \end{aligned} \quad (2.31)$$

Aus Ungleichung (2.24) im Beweis von Lemma 2 können wir

$$\begin{aligned} & \sum_{k,i,j,l} \max_{s=1, \dots, n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \\ & \leq \widehat{K}_n \cdot L \cdot r^2 \cdot d \cdot c_8 \cdot r^{2L} \cdot \max\{\|\sigma\|_{\infty}^2, 1\} \cdot \max\{\|\sigma'\|_{\infty}^{2L}, 1\} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^2 \end{aligned} \quad (2.32)$$

folgern. Zudem erhalten wir aus den Ungleichungen (2.26) und (2.27), dass

$$\begin{aligned} & \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(\mathbf{X}_s) - f_{\mathbf{w}_1}(\mathbf{X}_s))^2 \\ & \leq c_9^2 \cdot \widehat{K}_n^2 \cdot (\gamma_n^*)^2 \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \cdot (2r+1)^{2L} \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \end{aligned} \quad (2.33)$$

und

$$\begin{aligned} \frac{1}{n} \sum_{s=1}^n (f_{\mathbf{w}_2}(\mathbf{X}_s) - Y_s)^2 & \leq 2 \cdot F_n(\mathbf{v}) + 2 \cdot c_9^2 \cdot \widehat{K}_n^2 \cdot (\gamma_n^*)^2 \cdot \max\{\|\sigma'\|_\infty^{2L}, 1\} \\ & \quad \cdot (2r+1)^{2L} \cdot B_n^{2L} \cdot \alpha_n^2 \cdot \max\{\|\sigma\|_\infty, 1\}^2 \cdot \frac{8 \cdot t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\} \end{aligned} \quad (2.34)$$

gilt. Daher genügt es im Folgenden,

$$\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2$$

zu beschränken.

Aus Gleichung (2.23) im Beweis von Lemma 2 folgt, dass die partielle Ableitung

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(\mathbf{x})$$

für feste $\mathbf{x} \in [-\alpha_n, \alpha_n]^d$ eine Summe von höchstens r^{L-2} Produkten ist. Jedes dieser Produkte besteht aus höchstens $2L+1$ Faktoren. Von diesen Faktoren sind höchstens $L-1$ Faktoren betragsmäßig durch B_n beschränkt. Der letzte Faktor ist im Betrag durch γ_n^* beschränkt. Betrachten wir diese Faktoren als eine Funktion in \mathbf{w} , so sind diese Lipschitz-stetig mit einer Lipschitz-Konstanten, welche durch 1 beschränkt ist.

Aufgrund des Beweises von Lemma 2 ist bekannt, dass $f_{\mathbf{w},k,j}^{(l)}(\mathbf{x})$ für $k \in \{1, \dots, \widehat{K}_n\}$, $j \in \{1, \dots, r\}$ und $l \in \{1, \dots, L\}$, bezüglich des Gewichtsvektors \mathbf{w} , Lipschitz-stetig mit einer Lipschitz-Konstanten kleiner gleich $l \cdot c_{10} \cdot \max\{\|\sigma'\|_\infty^l, 1\} \cdot \max\{\|\sigma\|_\infty, 1\} \cdot (2r+1)^l \cdot B_n^l \cdot \alpha_n$ ist. Zudem ist $f_{\mathbf{w},k,j}^{(l)}(\mathbf{x})$ aufgrund seiner Definition entweder durch $\|\sigma\|_\infty$ oder α_n beschränkt.

Die verbleibenden L Faktoren, also die Ableitungen der Funktion σ , sind durch $\max\{\|\sigma'\|_\infty, 1\}$ beschränkt. Aus der Lipschitz-Stetigkeit von σ' und dem Beweis von Ungleichung (2.25) im Beweis von Lemma 2 folgt, dass die Lipschitz-Konstante dieser L Faktoren kleiner gleich $\sigma'_{\text{Lip}} \cdot L \cdot c_{10} \cdot (2r+1)^L \cdot B_n^L \cdot \alpha_n \cdot \max\{\|\sigma\|_\infty, 1\}$ ist.

Mit Hilfslemma 3, welches im Anhang enthalten ist, folgt, dass die Lipschitz-Konstante von

$$\frac{\partial f_{\mathbf{w}}}{\partial w_{k,i,j}^{(l)}}(\mathbf{x})$$

durch

$$r^{L-2} \cdot (2L+1) \cdot c_{12} \cdot (2r+1)^L \cdot B_n^L \cdot \alpha_n \cdot \max\{\|\sigma\|_\infty, 1\} \cdot B_n^L \cdot (\gamma_n^*) \cdot \max\{\|\sigma\|_\infty, \alpha_n\}$$

beschränkt ist, wobei $c_{12} = \max\{l \cdot c_{10} \cdot \|\sigma'\|_\infty, \sigma'_{\text{Lip}} \cdot L \cdot c_{10}, 1\}$ ist. Damit ergibt sich mit der Beschränktheit von σ sowie der Beschränktheit von σ' , dass

$$\sum_{k,i,j,l} \max_{s=1,\dots,n} \left(\frac{\partial f_{\mathbf{w}_1}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) - \frac{\partial f_{\mathbf{w}_2}}{\partial w_{k,i,j}^{(l)}}(\mathbf{X}_s) \right)^2 \leq c_{13} \cdot \widehat{K}_n \cdot B_n^{4L} \cdot \alpha_n^4 \cdot (\gamma_n^*)^2 \cdot \|\mathbf{w}_1 - \mathbf{w}_2\|^2 \quad (2.35)$$

gilt. Setzen wir nun die Resultate aus (2.32)–(2.35) in Ungleichung (2.31) ein, erhalten wir die Behauptung von Lemma 3. \square

2.2.2 Komplexität des Funktionsraums überparametrisierter tiefer neuronaler Netze

In diesem Abschnitt leiten wir eine obere Schranke für die Überdeckungszahl her. Damit können wir zeigen, dass die Komplexität des Funktionsraums überparametrisierter tiefer neuronaler Netze beschränkt ist. Bevor wir diese Schranke herleiten, definieren wir zunächst, was unter einer Überdeckungszahl zu verstehen ist.

Definition 5. Sei $\varepsilon > 0$. Sei zudem \mathcal{G} eine Menge von Funktionen $g : \mathbb{R}^d \rightarrow \mathbb{R}$, $\mathbf{x}_1^n = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in (\mathbb{R}^d)^n$ sowie $1 \leq p < \infty$.

- a) Eine endliche Menge von Funktionen $g_1, \dots, g_N : \mathbb{R}^d \rightarrow \mathbb{R}$ wird als **L_p - ε -Überdeckung von \mathcal{G} auf \mathbf{x}_1^n** bezeichnet, falls für jedes $g \in \mathcal{G}$ ein $j = j(g) \in \{1, \dots, N\}$ existiert, so dass

$$\left(\frac{1}{n} \sum_{i=1}^n |g(\mathbf{x}_i) - g_j(\mathbf{x}_i)|^p \right)^{\frac{1}{p}} < \varepsilon$$

gilt.

- b) Wir bezeichnen mit

$$\mathcal{N}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n)$$

die **L_p - ε -Überdeckungszahl von \mathcal{G} auf \mathbf{x}_1^n** , welche durch die Anzahl N der kleinsten L_p - ε -Überdeckung von \mathcal{G} auf \mathbf{x}_1^n definiert ist. Wir setzen $\mathcal{N}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n) = \infty$, falls keine endliche L_p - ε -Überdeckung von \mathcal{G} auf \mathbf{x}_1^n existiert.

Das folgende Lemma liefert eine Schranke für die Wahrscheinlichkeit, dass der empirische Mittelwert einer beschränkten Funktion vom Erwartungswert abweicht. Diese Aussage dient im Beweis von Theorem 1 als Grundlage, um weitere Aussagen über die Komplexität des überparametrisierten neuronalen Netzes treffen zu können.

Lemma 4. Seien $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$ unabhängig und identisch verteilte Zufallsvariablen mit Werten in \mathbb{R}^d . Sei zudem $B > 0$ und \mathcal{G} eine Menge messbarer Funktionen $g : \mathbb{R}^d \rightarrow [0, B]$. Dann gilt für alle $n \in \mathbb{N}$ und alle $\varepsilon > 0$ die Ungleichung

$$\mathbf{P} \left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) - \mathbf{E} \{g(\mathbf{X})\} \right| > \varepsilon \right\} \leq 8 \cdot \mathbf{E} \left\{ \mathcal{N}_1 \left(\frac{\varepsilon}{8}, \mathcal{G}, \mathbf{X}_1^n \right) \right\} \cdot \exp \left(-\frac{n \cdot \varepsilon^2}{128 \cdot B^2} \right).$$

Bemerkung 1. Da die Menge, über die wir das Supremum in der obigen Wahrscheinlichkeit bilden, möglicherweise nicht abzählbar ist, können Messbarkeitsprobleme auftreten. Diese Probleme können, wie in Van der Vaart und Wellner (1996) beschrieben, durch die Verwendung der äußeren Wahrscheinlichkeit umgangen werden.

Beweis. Siehe Theorem 9.1 in Györfi et al. (2002). □

Das folgende Lemma präsentiert eine obere Schranke für die ε -Überdeckungsanzahl, die für die Kontrolle der Komplexität der Funktionsklasse der überparametrisierten tiefen neuronalen Netze notwendig ist.

Lemma 5. Seien $\alpha \geq 1$, $\beta > 0$ und $A, B, C \geq 1$. Sei weiter $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ eine ℓ -mal differenzierbare Funktion derart, dass alle ihre Ableitungen bis Ordnung ℓ auf \mathbb{R} beschränkt sind. Wir bezeichnen mit \mathcal{F} die Menge aller Funktionen $f_{\mathbf{w}}$, welche durch (2.1)–(2.3) definiert sind und deren Gewichtsvektor \mathbf{w} die Eigenschaften

$$\sum_{k=1}^{\widehat{K}_n} |w_{1,1,k}^{(L)}| \leq C, \quad (2.36)$$

$$|w_{k,i,j}^{(l)}| \leq B \quad \text{für } k \in \{1, \dots, \widehat{K}_n\}, i, j \in \{1, \dots, r\}, l \in \{1, \dots, L-1\} \quad (2.37)$$

sowie

$$|w_{k,i,j}^{(0)}| \leq A \quad \text{für } k \in \{1, \dots, \widehat{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\} \quad (2.38)$$

besitzt. Dann existieren Konstanten $c_{14}, c_{15}, c_{16} > 0$, so dass für alle $p \in [1, \infty)$, $\varepsilon \in (0, \beta)$ und $\mathbf{x}_1^n \in [-\alpha, \alpha]^{d \cdot n}$ die Ungleichung

$$\mathcal{N}_p(\varepsilon, \{T_\beta f : f \in \mathcal{F}\}, \mathbf{x}_1^n) \leq \left(c_{14} \cdot \frac{\beta^p}{\varepsilon^p} \right)^{c_{15} \cdot \alpha^d \cdot B^{(L-1) \cdot d} \cdot A^d \cdot \left(\frac{C}{\varepsilon}\right)^{d/\ell} + c_{16}}$$

gilt.

Für den Beweis dieses Lemmas benötigen wir einige Hilfsresultate, die wir im Folgenden vorstellen.

Um die Überdeckungsanzahl einer Klasse von Funktionen abzuschätzen, kann diese in Beziehung zur so genannten *Packzahl* gesetzt werden.

Definition 6. Sei $\varepsilon > 0$ und sei \mathcal{G} die Menge aller Funktionen $g : \mathbb{R}^d \rightarrow \mathbb{R}$. Sei zudem $\mathbf{x}_1^n \in (\mathbb{R}^d)^n$ sowie $1 \leq p < \infty$.

a) Eine endliche Menge von Funktionen $g_1, \dots, g_N \in \mathcal{G}$, die

$$\left(\frac{1}{n} \sum_{i=1}^n |g_j(\mathbf{x}_i) - g_k(\mathbf{x}_i)|^p \right)^{\frac{1}{p}} \geq \varepsilon$$

für alle $1 \leq j < k \leq N$ erfüllt, bezeichnen wir als **L_p - ε -Packung von \mathcal{G} auf \mathbf{x}_1^n** .

b) Die maximale Anzahl $N \in \mathbb{N}$, für die eine solche L_p - ε -Packung von \mathcal{G} auf \mathbf{x}_1^n existiert, wird mit $\mathcal{M}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n)$ bezeichnet. Existiert für jedes $N \in \mathbb{N}$ eine ε -Packung von \mathcal{G} bezüglich der L_p -Norm, so setzen wir $\mathcal{M}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n) = \infty$.

Wir nennen $\mathcal{M}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n)$ die **ε -Packzahl von \mathcal{G} auf \mathbf{x}_1^n bezüglich der L_p -Norm**.

Das folgende Hilfslemma zeigt, dass die ε -Überdeckungszahl durch die ε -Packzahl nach oben beschränkt werden kann.

Hilfslemma 1. Sei \mathcal{G} eine Klasse von Funktionen auf \mathbb{R}^d , $p \geq 1$ und $\varepsilon > 0$. Dann gilt

$$\mathcal{M}_p(2\varepsilon, \mathcal{G}, \mathbf{x}_1^n) \leq \mathcal{N}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n) \leq \mathcal{M}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n)$$

für alle $\mathbf{x}_1^n \in (\mathbb{R}^d)^n$.

Beweis. Siehe Lemma 9.2 in Györfi et al. (2002). □

Eine in der Literatur gängige Methode zur Abschätzung der Überdeckungszahl basiert auf der sogenannten *Vapnik-Chervonenkis-Dimension*. Sie spielt eine entscheidende Rolle bei der Analyse der Komplexität des Funktionsraums eines neuronalen Netzes.

Definition 7. Sei \mathcal{A} eine Klasse von Teilmengen des \mathbb{R}^d mit $\mathcal{A} \neq \emptyset$ und $n \in \mathbb{N}$.

a) Für $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ definieren wir durch

$$s(\mathcal{A}, n) = \max_{\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subseteq \mathbb{R}^d} |\{A \cap \{\mathbf{x}_1, \dots, \mathbf{x}_n\} : A \in \mathcal{A}\}|$$

den *n-ten Zerlegungskoeffizienten* von \mathcal{A} .

b) Die *Vapnik-Chervonenkis-Dimension* (VC-Dimension) $\mathcal{V}_{\mathcal{A}}$ ist definiert durch

$$\mathcal{V}_{\mathcal{A}} = \sup\{n \in \mathbb{N} : s(\mathcal{A}, n) = 2^n\}.$$

In Worten ausgedrückt, bedeutet dies, dass die VC-Dimension $\mathcal{V}_{\mathcal{A}}$ die größte natürliche Zahl n ist, so dass es eine Menge von n Punkten in \mathbb{R}^d gibt, die durch \mathcal{A} vollständig zerlegt werden kann.

Das folgende Hilfslemma zeigt, dass die VC-Dimension einer Menge von Subgraphen, die durch reellwertige Funktionen aus dem Funktionsraum \mathcal{G} definiert sind, durch die Dimension von \mathcal{G} beschränkt ist.

Hilfslemma 2. Sei \mathcal{G} ein r -dimensionaler Vektorraum von reellen Funktionen auf \mathbb{R}^d und sei

$$\mathcal{G}^+ = \left\{ \left\{ (\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R} : g(\mathbf{x}) \geq t \right\} : g \in \mathcal{G} \right\}.$$

Dann gilt

$$\mathcal{V}_{\mathcal{G}^+} \leq r + 1.$$

Beweis. Folgt aus Theorem 9.5 in Györfi et al. (2002). □

Mithilfe des folgenden Lemmas erhalten wir unter Verwendung der VC-Dimension eine obere Schranke für die ε -Packzahl bezüglich der L_p -Norm.

Lemma 6. Sei \mathcal{G} eine Klasse von Funktionen $g : \mathbb{R}^d \rightarrow [-B, B]$ mit $\mathcal{V}_{\mathcal{G}^+} \geq 2$, wobei

$$\mathcal{G}^+ := \left\{ \left\{ (\mathbf{x}, t) \in \mathbb{R}^d \times \mathbb{R} : g(\mathbf{x}) \geq t \right\} : g \in \mathcal{G} \right\}.$$

Seien zudem $p \geq 1$ und $0 < \varepsilon < \frac{B}{2}$. Dann gilt

$$\mathcal{M}_p(\varepsilon, \mathcal{G}, \mathbf{x}_1^n) \leq 3 \left(\frac{2e(2B)^p}{\varepsilon^p} \log \left(\frac{3e(2B)^p}{\varepsilon^p} \right) \right)^{\mathcal{V}_{\mathcal{G}^+}}.$$

Beweis. Siehe Theorem 9.4 in Györfi et al. (2002). □

Auf Basis dieser Resultate können wir nun Lemma 5 beweisen.

Beweis von Lemma 5. Da es sich bei σ um eine ℓ -mal differenzierbare Funktion handelt, deren Ableitungen bis Ordnung ℓ auf \mathbb{R} beschränkt sind, können wir annehmen, dass $|\sigma^{(s)}| \leq c_{17}$ für $s \in \{1, \dots, \ell\}$ ist.

Im ersten Schritt des Beweises zeigen wir, dass

$$\left| \partial^{\xi} f_{\mathbf{w}}(\mathbf{x}) \right| \leq c_{18} \cdot C \cdot B^{(L-1) \cdot \ell} \cdot A^{\ell} =: c \quad (2.39)$$

für alle Funktionen $f_{\mathbf{w}} \in \mathcal{F}$, alle $\mathbf{x} \in \mathbb{R}^d$ sowie $\|\xi\|_1 = \ell$ gilt. Die Definition von $f_{\mathbf{w}}$ impliziert

$$\partial^{\xi} f_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^{\hat{K}_n} w_{1,1,k}^{(L)} \cdot \partial^{\xi} f_{\mathbf{w},k,1}^{(L)}(\mathbf{x}).$$

Wir beginnen damit zu zeigen, dass

$$\left| \partial^{\xi} f_{\mathbf{w},k,1}^{(L)}(\mathbf{x}) \right| \leq c_{18} \cdot B^{(L-1) \cdot \ell} \cdot A^{\ell} \quad (2.40)$$

für alle $\mathbf{x} \in \mathbb{R}^d$ ist.

Die partielle Ableitung von $f_{\mathbf{w},k,i}^{(l)}$ nach $x^{(\rho)}$ mit $\rho \in \{1, \dots, d\}$ ist gegeben durch

$$\begin{aligned} \frac{\partial f_{\mathbf{w},k,i}^{(l)}}{\partial x^{(\rho)}}(\mathbf{x}) &= \sigma' \left(\sum_{t=1}^r w_{k,i,t}^{(l-1)} \cdot f_{\mathbf{w},k,t}^{(l-1)}(\mathbf{x}) + w_{k,i,0}^{(l-1)} \right) \cdot \sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot \frac{\partial f_{\mathbf{w},k,j}^{(l-1)}}{\partial x^{(\rho)}}(\mathbf{x}) \\ &= \sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot \sigma' \left(\sum_{t=1}^r w_{k,i,t}^{(l-1)} \cdot f_{\mathbf{w},k,t}^{(l-1)}(\mathbf{x}) + w_{k,i,0}^{(l-1)} \right) \cdot \frac{\partial f_{\mathbf{w},k,j}^{(l-1)}}{\partial x^{(\rho)}}(\mathbf{x}). \end{aligned}$$

Für $l = 1$ gilt dementsprechend

$$\frac{\partial f_{\mathbf{w},k,i}^{(1)}}{\partial x^{(\rho)}}(\mathbf{x}) = \sigma' \left(\sum_{j=1}^d w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)} \right) \cdot w_{k,i,\rho}^{(0)}.$$

Die Anwendung der Produktregel der Differentialrechnung liefert für $l > 1$, dass die partielle Ableitung

$$\partial^{\xi} f_{\mathbf{w},k,i}^{(l)}(\mathbf{x}) \quad (2.41)$$

eine Summe von höchstens $r \cdot (r + \ell)^{\ell-1}$ vielen Termen der Form

$$w \cdot \sigma^{(s)} \left(\sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot f_{\mathbf{w},k,j}^{(l-1)}(\mathbf{x}) + w_{k,i,0}^{(l-1)} \right) \cdot \partial^{\mathbf{t}_1} f_{\mathbf{w},k,j_1}^{(l-1)}(\mathbf{x}) \cdots \partial^{\mathbf{t}_s} f_{\mathbf{w},k,j_s}^{(l-1)}(\mathbf{x}) \quad (2.42)$$

für $s \in \{1, \dots, \ell\}$, $j_1, \dots, j_s \in \{1, \dots, r\}$, $|w| \leq B^s$ und $\|\mathbf{t}_1\|_1 + \dots + \|\mathbf{t}_s\|_1 = \ell$ ist. Zudem ist

$$\partial^{\xi} f_{\mathbf{w},k,i}^{(1)}(\mathbf{x}) = \prod_{j=1}^{\ell} w_{k,i,s_j}^{(0)} \cdot \sigma^{(\ell)} \left(\sum_{t=1}^d w_{k,i,t}^{(0)} \cdot x^{(t)} + w_{k,i,0}^{(0)} \right)$$

für $\|\xi\|_1 = \ell$.

Da die Ableitungen der Funktion σ bis Ordnung ℓ beschränkt sind, erhalten wir zusammen mit Voraussetzung (2.38), dass

$$\left| \partial^{\xi} f_{\mathbf{w},k,i}^{(1)}(\mathbf{x}) \right| \leq c_{17} \cdot A^{\ell}$$

für alle $\ell \in \mathbb{N}$ und $\|\xi\|_1 = \ell$ gilt.

Mithilfe einer Induktion wollen wir nun nachweisen, dass

$$\left| \partial^{\xi} f_{\mathbf{w},k,i}^{(l)}(\mathbf{x}) \right| \leq c_{17}^l \cdot B^{(l-1) \cdot \ell} \cdot A^{\ell} \quad (2.43)$$

für $\|\xi\|_1 = \ell$ und $l \in \{1, \dots, L\}$ erfüllt ist.

Für $l = 1$ folgt die Aussage direkt aus der obigen Ungleichung mit

$$\left| \partial^{\xi} f_{\mathbf{w},k,i}^{(1)}(\mathbf{x}) \right| \leq c_{17} \cdot A^{\ell} = c_{17}^1 \cdot B^{(1-1) \cdot \ell} \cdot A^{\ell}.$$

Im Folgenden nehmen wir an, dass Ungleichung (2.43) für ein $l \in \{1, \dots, L-1\}$ gilt.

Wir wollen nun zeigen, dass

$$\left| \partial^{\xi} f_{\mathbf{w},k,i}^{(l+1)}(\mathbf{x}) \right| \leq c_{17}^{(l+1)} \cdot B^{l \cdot \ell} \cdot A^{\ell}$$

für $l \in \{1, \dots, L-1\}$ ebenfalls erfüllt ist. Wie oben bereits gezeigt, ist diese partielle Ableitung eine Summe von höchstens $r \cdot (r + \ell)^{\ell-1}$ vielen Termen der Form (2.42). Wegen der Induktionsannahme sind die hinteren partiellen Ableitungen durch

$$c_{17}^l \cdot B^{(l-1) \cdot \ell} \cdot A^{\ell}$$

begrenzt. Das erste Produkt ist aufgrund der Voraussetzung, dass σ bis Ordnung ℓ auf \mathbb{R} beschränkt ist sowie der Voraussetzung (2.37) kleiner gleich

$$c_{17} \cdot B.$$

Daher ist

$$\left| \partial^{\xi} f_{\mathbf{w},k,i}^{(l+1)}(\mathbf{x}) \right| \leq c_{17} \cdot B \cdot c_{17}^l \cdot B^{(l-1) \cdot \ell} \cdot A^{\ell} = c_{17}^{(l+1)} \cdot B^{l \cdot \ell} \cdot A^{\ell}.$$

Setzen wir $l + 1 = L$ und $c_{17}^{(l+1)} = c_{18}$, so erhalten wir Ungleichung (2.40).

Um die Behauptung (2.39) des ersten Beweisschrittes zu zeigen, betrachten wir nun erneut die partielle Ableitung des neuronalen Netzes $f_{\mathbf{w}}$, welche gegeben ist durch

$$\partial^{\xi} f_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^{\widehat{K}_n} w_{1,1,k}^{(L)} \cdot \partial^{\xi} f_{\mathbf{w},k,1}^{(L)}(\mathbf{x}).$$

Die Anwendung von Voraussetzung (2.36) zusammen mit Ungleichung (2.40) liefert die Behauptung des ersten Schrittes.

Im Folgenden sei \mathcal{G} die Menge aller Polynome vom Grad kleiner gleich $\ell - 1$. Weiter sei Π eine Partition von $[-\alpha, \alpha]^d$ in K Wurfel mit Seitenlange

$$\left(c_{19} \cdot \frac{\varepsilon}{c} \right)^{1/\ell},$$

wobei $0 < c_{19} = c_{19}(d, \ell) \leq \frac{1}{2} \cdot \frac{\ell!}{d^\ell}$ gilt.

Im *zweiten Schritt des Beweises* zeigen wir, dass fur jedes $p \in [1, \infty)$, $\varepsilon \in (0, \beta)$ sowie $\mathbf{x}_1^n \in [-\alpha, \alpha]^{d \cdot n}$ die folgende Ungleichung gilt

$$\mathcal{N}_p(\varepsilon, \{T_{\beta} f : f \in \mathcal{F}\}, \mathbf{x}_1^n) \leq \mathcal{N}_p\left(\frac{\varepsilon}{2}, \{T_{\beta} g : g \in \mathcal{G} \circ \Pi\}, \mathbf{x}_1^n\right). \quad (2.44)$$

Hierfur werden wir zuerst zeigen, dass

$$|f_{\mathbf{w}}(\mathbf{x}) - g(\mathbf{x})| \leq \frac{\varepsilon}{2}$$

fur $f_{\mathbf{w}} \in \mathcal{F}$, $g \in \mathcal{G} \circ \Pi$ und alle $\mathbf{x} \in [-\alpha, \alpha]^d$ erfullt ist.

Da $\mathcal{G} \circ \Pi$ gema der obigen Definition die Menge aller stuckweisen Polynome bezuglich der Partition Π ist, konnen wir $f_{\mathbf{w}}$ nach dem Satz von Taylor auf jedem Teilbereich der Partition durch eine Funktion $g \in \mathcal{G} \circ \Pi$ zuzuglich dem mehrdimensionalen Integralrestglied der Ordnung $\ell - 1$ darstellen. Das heit, fur jeden Teilbereich P der Partition gibt es ein $\mathbf{x}_0 \in P$ und eine Funktion $g \in \mathcal{G} \circ \Pi$ derart, dass

$$f_{\mathbf{w}}(\mathbf{x}) = g(\mathbf{x}) + R_{P, \mathbf{x}_0, \ell-1} f(\mathbf{h}),$$

fur alle $\mathbf{x} \in P$ und $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$, wobei

$$R_{P, \mathbf{x}_0, \ell-1} f(\mathbf{h}) = \ell \sum_{\|\xi\|_1 = \ell} \int_0^1 \frac{(1-t)^{\ell-1} \cdot \mathbf{h}^{\xi}}{\xi!} \cdot \partial^{\xi} f(\mathbf{x}_0 + t\mathbf{h}) dt.$$

Aus dem ersten Beweisschritt wissen wir, dass die partiellen Ableitungen von f beschrankt sind. Durch Anwendung des Multinomialtheorems (siehe z. B. Arens et al. (2022)) erhalten wir die Identitat

$$(h_1 + \dots + h_d)^{\ell} = \sum_{\|\xi\|_1 = \ell} \frac{\ell!}{\xi!} \cdot \mathbf{h}^{\xi},$$

woraus sich ergibt, dass

$$|R_{P, \mathbf{x}_0, \ell-1} f(\mathbf{h})| \leq \sum_{\|\xi\|_1 = \ell} \frac{|\mathbf{h}^{\xi}|}{\xi!} \cdot c = \frac{1}{\ell!} \cdot (h_1 + \dots + h_d)^{\ell} \cdot c$$

für alle $\mathbf{x} \in P$ und $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$ erfüllt ist.

Ersetzen wir nun $c = c_{18} \cdot C \cdot B^{(L-1) \cdot \ell} \cdot A^\ell$ und $h_i = (c_{19} \cdot \frac{\varepsilon}{c})^{1/\ell}$ für $i \in \{1, \dots, d\}$, so erhalten wir für jedes Restglied

$$\begin{aligned} |R_{P, \mathbf{x}_0, \ell-1} f(\mathbf{h})| &\leq c_{18} \cdot C \cdot B^{(L-1) \cdot \ell} \cdot A^\ell \cdot \frac{1}{\ell!} \cdot d^\ell \cdot c_{19} \cdot \frac{\varepsilon}{c_{18} \cdot C \cdot B^{(L-1) \cdot \ell} \cdot A^\ell} \\ &\leq c_{19} \cdot \frac{d^\ell}{\ell!} \cdot \varepsilon \leq \frac{1}{2} \cdot \frac{\ell!}{d^\ell} \cdot \frac{d^\ell}{\ell!} \cdot \varepsilon \leq \frac{\varepsilon}{2} \end{aligned}$$

für alle $\mathbf{x} \in P$ und $\mathbf{h} = \mathbf{x} - \mathbf{x}_0$. Daher gilt diese Abschätzung insbesondere für alle $\mathbf{x} \in [-\alpha, \alpha]^d$, womit

$$|f_{\mathbf{w}}(\mathbf{x}) - g(\mathbf{x})| \leq \frac{\varepsilon}{2}$$

für alle $\mathbf{x} \in [-\alpha, \alpha]^d$ folgt.

Sei nun $\{g_1, \dots, g_N\}$ eine L_p - $\frac{\varepsilon}{2}$ -Überdeckung von $T_\beta(\mathcal{G} \circ \Pi)$ auf \mathbf{x}_1^n , dann existiert ein $j = j(g) \in \{1, \dots, N\}$, so dass

$$\left\{ \frac{1}{n} \sum_{i=1}^n |g(\mathbf{x}_i) - g_j(\mathbf{x}_i)|^p \right\}^{\frac{1}{p}} \leq \frac{\varepsilon}{2}$$

für $g \in T_\beta(\mathcal{G} \circ \Pi)$ gilt. Damit erhalten wir für $f \in T_\beta \mathcal{F}$ die folgende Ungleichung

$$\left\{ \frac{1}{n} \sum_{i=1}^n |f(\mathbf{x}_i) - g_j(\mathbf{x}_i)|^p \right\}^{\frac{1}{p}} \leq \left\{ \frac{1}{n} \sum_{i=1}^n |f(\mathbf{x}_i) - g(\mathbf{x}_i)|^p \right\}^{\frac{1}{p}} + \left\{ \frac{1}{n} \sum_{i=1}^n |g(\mathbf{x}_i) - g_j(\mathbf{x}_i)|^p \right\}^{\frac{1}{p}} \leq \varepsilon.$$

Daher ist jede L_p - $\frac{\varepsilon}{2}$ -Überdeckung von $T_\beta(\mathcal{G} \circ \Pi)$ auch eine L_p - ε -Überdeckung von $T_\beta \mathcal{F}$, woraus Ungleichung (2.44) folgt.

Im dritten Schritt des Beweises zeigen wir die Aussage des Lemmas.

Aus Hilfslemma 1 folgt

$$\mathcal{N}_p \left(\frac{\varepsilon}{2}, \{T_\beta g : g \in \mathcal{G} \circ \Pi\}, \mathbf{x}_1^n \right) \leq \mathcal{M}_p \left(\frac{\varepsilon}{2}, \{T_\beta g : g \in \mathcal{G} \circ \Pi\}, \mathbf{x}_1^n \right).$$

Die Anwendung von Lemma 6 liefert dann die Abschätzung

$$\mathcal{M}_p \left(\frac{\varepsilon}{2}, \{T_\beta g : g \in \mathcal{G} \circ \Pi\}, \mathbf{x}_1^n \right) \leq 3 \left(\frac{2e(2\beta)^p}{(\varepsilon/2)^p} \log \left(\frac{3e(2\beta)^p}{(\varepsilon/2)^p} \right) \right)^{V_{(\mathcal{G} \circ \Pi)^+}} \leq \left(c_{14} \cdot \frac{\beta^p}{\varepsilon^p} \right)^{V_{(\mathcal{G} \circ \Pi)^+}}.$$

Da $\mathcal{G} \circ \Pi$ ein linearer Vektorraum mit einer Dimension kleiner als

$$c_{20} \cdot \alpha^d \cdot \left(\frac{c}{\varepsilon} \right)^{d/\ell}$$

ist, ergibt sich aus Hilfslemma 2 die folgende obere Schranke für die VC-Dimension:

$$V_{(\mathcal{G} \circ \Pi)^+} \leq c_{20} \cdot \alpha^d \cdot \left(\frac{c}{\varepsilon} \right)^{d/\ell} + 1.$$

Somit erhalten wir

$$\mathcal{N}_p \left(\frac{\varepsilon}{2}, \{T_\beta g : g \in \mathcal{G} \circ \Pi\}, \mathbf{x}_1^n \right) \leq \left(c_{14} \cdot \frac{\beta^p}{\varepsilon^p} \right)^{c_{15} \cdot \alpha^d \cdot B^{d \cdot (L-1)} \cdot A^d \cdot \left(\frac{c}{\varepsilon} \right)^{d/\ell} + c_{16}}$$

für Konstanten $c_{14}, c_{15}, c_{16} > 0$.

Mit Ungleichung (2.44) folgt hieraus die Aussage des Lemmas. \square

2.2.3 Approximationseigenschaft überparametrisierter tiefer neuronaler Netze

In diesem Abschnitt wollen wir zeigen, dass eine Lipschitz-stetige Funktion hinreichend gut durch die Linearkombination tiefer neuronaler Netze approximiert werden kann. Für den Beweis dieser Aussage benötigen wir die zwei folgenden Hilfsresultate.

Das folgende Lemma zeigt, dass ein tiefes, vollständig verbundenes neuronales Netz unter geeigneten Voraussetzungen eine Indikatorfunktion, die im Inneren eines Würfels den Wert 1 und außerhalb den Wert 0 annimmt, bis auf an den Rändern des Würfels beliebig genau approximieren kann.

Lemma 7. Sei $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}$ die logistische Sigmoidfunktion. Seien weiter $0 < \delta \leq 1$ und $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ mit

$$v^{(s)} - u^{(s)} \geq 2\delta \quad \text{für } s \in \{1, \dots, d\}$$

sowie $1 \leq \alpha_n \leq \log n$ und $\mathbf{x} \in [-\alpha_n, \alpha_n]^d$. Zudem seien $L, n, r \in \mathbb{N}$ mit $L \geq 2$, $r \geq 2d$, $n \geq 8d$ und $n \geq \exp(r+1)$. Die Ausgabe des vollständig verbundenen neuronalen Netzes $f_{\bar{\mathbf{w}},1,1}^{(L)}(\mathbf{x})$ sei definiert wie in (2.2) und (2.3).

Angenommen, für die Gewichte gelten die Bedingungen

$$w_{1,j,j}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \quad \text{und} \quad w_{1,j,0}^{(0)} = -\frac{4d \cdot (\log n)^2}{\delta} \cdot u^{(j)} \quad \text{für } j \in \{1, \dots, d\}, \quad (2.45)$$

$$w_{1,j+d,j}^{(0)} = -\frac{4d \cdot (\log n)^2}{\delta} \quad \text{und} \quad w_{1,j+d,0}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \cdot v^{(j)} \quad \text{für } j \in \{1, \dots, d\}, \quad (2.46)$$

$$w_{1,s,t}^{(0)} = 0 \quad \text{falls } s \leq 2d, s \neq t, s \neq t+d \text{ und } t > 0, \quad (2.47)$$

$$w_{1,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \text{für } t \in \{1, \dots, 2d\}, \quad (2.48)$$

$$w_{1,1,0}^{(1)} = -8 \cdot (\log n)^2 \cdot \left(2d - \frac{1}{2}\right), \quad (2.49)$$

$$w_{1,1,t}^{(1)} = 0 \quad \text{für } t > 2d, \quad (2.50)$$

$$w_{1,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \text{für } l \in \{2, \dots, L-1\}, \quad (2.51)$$

$$w_{1,1,0}^{(l)} = -3 \cdot (\log n)^2 \quad \text{für } l \in \{2, \dots, L-1\} \quad (2.52)$$

sowie

$$w_{1,1,t}^{(l)} = 0 \quad \text{für } t > 1 \text{ und } l \in \{2, \dots, L-1\}. \quad (2.53)$$

Zudem sei der Gewichtsvektor $\bar{\mathbf{w}}$ so gewählt, dass

$$|\bar{w}_{1,i,j}^{(l)} - w_{1,i,j}^{(l)}| \leq \log n \quad \text{für alle } l \in \{0, \dots, L-1\} \quad (2.54)$$

erfüllt ist. Dann gilt sowohl

$$f_{\bar{\mathbf{w}},1,1}^{(L)}(\mathbf{x}) \geq 1 - \frac{1}{n}, \quad \text{falls } \mathbf{x} \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \dots \times [u^{(d)} + \delta, v^{(d)} - \delta]$$

als auch

$$f_{\bar{\mathbf{w}},1,1}^{(L)}(\mathbf{x}) \leq \frac{1}{n}, \quad \text{falls } x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta] \text{ für ein } i \in \{1, \dots, d\}.$$

Beweis. Im ersten Schritt des Beweises zeigen wir, dass

$$f_{\bar{\mathbf{w}},1,1}^{(L)}(\mathbf{x}) \geq 1 - \frac{1}{n}$$

für alle $\mathbf{x} \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \dots \times [u^{(d)} + \delta, v^{(d)} - \delta]$ erfüllt ist.

Sei hierfür $\mathbf{x} \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \dots \times [u^{(d)} + \delta, v^{(d)} - \delta]$. Durch die Bedingungen (2.45), (2.47), (2.54) und $1 \leq \alpha_n \leq \log n$ erhalten wir für alle $i \in \{1, \dots, d\}$, dass

$$\begin{aligned} & \sum_{j=1}^d \bar{w}_{1,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i,0}^{(0)} \\ &= \sum_{j=1}^d (\bar{w}_{1,i,j}^{(0)} - w_{1,i,j}^{(0)}) \cdot x^{(j)} + (\bar{w}_{1,i,0}^{(0)} - w_{1,i,0}^{(0)}) + \sum_{j=1}^d w_{1,i,j}^{(0)} \cdot x^{(j)} + w_{1,i,0}^{(0)} \\ &\geq -d \cdot \log n \cdot \alpha_n - \log n + \frac{4d \cdot (\log n)^2}{\delta} \cdot (u^{(i)} + \delta) - \frac{4d \cdot (\log n)^2}{\delta} \cdot u^{(i)} \\ &\geq 3d \cdot (\log n)^2 - \log n \\ &\geq \log n \end{aligned}$$

gilt. Mit den Bedingungen (2.46), (2.47) und (2.54) ergibt sich für alle $i \in \{1, \dots, d\}$ die Abschätzung

$$\begin{aligned} & \sum_{j=1}^d \bar{w}_{1,i+d,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i+d,0}^{(0)} \\ &= \sum_{j=1}^d (\bar{w}_{1,i+d,j}^{(0)} - w_{1,i+d,j}^{(0)}) \cdot x^{(j)} + (\bar{w}_{1,i+d,0}^{(0)} - w_{1,i+d,0}^{(0)}) + \sum_{j=1}^d w_{1,i+d,j}^{(0)} \cdot x^{(j)} + w_{1,i+d,0}^{(0)} \\ &\geq -d \cdot \log n \cdot \alpha_n - \log n - \frac{4d \cdot (\log n)^2}{\delta} (v^{(i)} - \delta) + \frac{4d \cdot (\log n)^2}{\delta} v^{(i)} \\ &\geq 3d \cdot (\log n)^2 - \log n \\ &\geq \log n. \end{aligned}$$

Da die logarithmische Sigmoidfunktion $\sigma(z) = \frac{1}{1+\exp(-z)}$ monoton wachsend ist, impliziert dies

$$f_{\bar{\mathbf{w}},1,i}^{(1)}(\mathbf{x}) \geq \sigma(\log n) = \frac{1}{1 + \frac{1}{n}} = \frac{n}{n+1} = \frac{n+1-1}{n+1} = 1 - \frac{1}{n+1} \geq 1 - \frac{1}{n}$$

für alle $i \in \{1, \dots, 2d\}$.

Analog erhalten wir durch Anwendung der zuvor gezeigten Ungleichungen sowie den Bedingungen (2.48)–(2.50), (2.54) und unter Berücksichtigung von $|\sigma(z)| \leq 1$ für alle $z \in \mathbb{R}$, dass

$$\begin{aligned} & \sum_{j=1}^r \bar{w}_{1,1,j}^{(1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(1)}(\mathbf{x}) + \bar{w}_{1,1,0}^{(1)} \\ &\geq -(r+1) \cdot \log n + \sum_{j=1}^r w_{1,1,j}^{(1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(1)}(\mathbf{x}) + w_{1,1,0}^{(1)} \end{aligned}$$

$$\begin{aligned}
&= -(r+1) \cdot \log n + \sum_{j=1}^{2d} w_{1,1,j}^{(1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(1)}(\mathbf{x}) + \sum_{j=2d+1}^r w_{1,1,j}^{(1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(1)}(\mathbf{x}) + w_{1,1,0}^{(1)} \\
&\geq -(r+1) \cdot \log n + 2d \cdot 8 \cdot (\log n)^2 \left(1 - \frac{1}{n}\right) - 8 \cdot (\log n)^2 \left(2d - \frac{1}{2}\right) \\
&= -(r+1) \cdot \log n + 8 \cdot (\log n)^2 \left(\frac{1}{2} - \frac{1}{n}\right) \\
&\geq \log n
\end{aligned}$$

gilt. Die letzte Aussage folgt hierbei aus $n \geq \exp(r+1)$, was äquivalent zu $\log n \geq r+1$ ist. Somit ergibt sich analog zu oben

$$f_{\bar{\mathbf{w}},1,1}^{(2)}(\mathbf{x}) \geq 1 - \frac{1}{n}.$$

Durch wiederholtes Anwenden des gleichen Arguments können wir, unter Berücksichtigung der Bedingungen (2.51)–(2.54) sowie der Voraussetzung $n \geq \exp(r+1)$ schließen, dass

$$\begin{aligned}
&\sum_{j=1}^r \bar{w}_{1,1,j}^{(l-1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(l-1)}(\mathbf{x}) + \bar{w}_{1,1,0}^{(l-1)} \\
&\geq -(r+1) \cdot \log n + \sum_{j=1}^r w_{1,1,j}^{(l-1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(l-1)}(\mathbf{x}) + w_{1,1,0}^{(l-1)} \\
&\geq -(r+1) \cdot \log n + 6 \cdot (\log n)^2 \left(1 - \frac{1}{n}\right) - 3 \cdot (\log n)^2 \\
&= -(r+1) \cdot \log n + 3 \cdot (\log n)^2 - \frac{6}{n} \cdot (\log n)^2 \\
&\geq 2 \cdot (\log n)^2 - \frac{6}{n} \cdot (\log n)^2 \\
&\geq \log n
\end{aligned}$$

für $l = 3, \dots, L$ ist. Daraus ergibt sich

$$f_{\bar{\mathbf{w}},1,1}^{(l)}(\mathbf{x}) \geq 1 - \frac{1}{n}$$

für $l = 3, \dots, L$.

Somit gilt

$$f_{\bar{\mathbf{w}},1,1}^{(L)}(\mathbf{x}) \geq 1 - \frac{1}{n},$$

falls $\mathbf{x} \in [u^{(1)} + \delta, v^{(1)} - \delta] \times \dots \times [u^{(d)} + \delta, v^{(d)} - \delta]$ ist, was die Aussage des ersten Schrittes beweist.

Im zweiten Schritt des Beweises wollen wir zeigen, dass

$$f_{\bar{\mathbf{w}},1,1}^{(L)}(\mathbf{x}) \leq \frac{1}{n}$$

ist, falls $x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta]$ für ein $i \in \{1, \dots, d\}$.

Sei $i \in \{1, \dots, d\}$. Wir nehmen im Folgenden an, dass $x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta]$. Das Vorgehen ist ähnlich wie im ersten Beweisschritt. Für den Fall, dass $x^{(i)} < u^{(i)} - \delta$ für ein $i \in \{1, \dots, d\}$ ist, erhalten wir zusammen

mit den Bedingungen (2.45), (2.47), (2.54) und $1 \leq \alpha_n \leq \log n$, dass

$$\begin{aligned}
& \sum_{j=1}^d \bar{w}_{1,i,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i,0}^{(0)} \\
& \leq d \cdot \log n \cdot \alpha_n + \log n + \frac{4d \cdot (\log n)^2}{\delta} \cdot (u^{(i)} - \delta) - \frac{4d \cdot (\log n)^2}{\delta} \cdot u^{(i)} \\
& \leq -3d \cdot (\log n)^2 + \log n \\
& \leq -\log n
\end{aligned}$$

gilt. Im Fall, dass $x^{(i)} > v^{(i)} + \delta$ für ein $i \in \{1, \dots, d\}$ ist, folgt aus den Bedingungen (2.46), (2.47) und (2.54), dass

$$\begin{aligned}
& \sum_{j=1}^d \bar{w}_{1,i+d,j}^{(0)} \cdot x^{(j)} + \bar{w}_{1,i+d,0}^{(0)} \\
& \leq d \cdot \log n \cdot \alpha_n + \log n - \frac{4d \cdot (\log n)^2}{\delta} \cdot (v^{(i)} + \delta) + \frac{4d \cdot (\log n)^2}{\delta} \cdot v^{(i)} \\
& \leq -3d \cdot (\log n)^2 + \log n \\
& \leq -\log n
\end{aligned}$$

erfüllt ist. Wenden wir hierauf die logistische Sigmoidfunktion an, ergibt sich

$$f_{\bar{\mathbf{w}},1,i}^{(1)}(\mathbf{x}) \leq \sigma(-\log n) = \frac{1}{1+n} \leq \frac{1}{n}$$

für ein $i \in \{1, \dots, 2d\}$.

Mit dem gleichen Argument wie im ersten Schritt, unter Berücksichtigung der Bedingungen (2.48)–(2.50), (2.54), sowie der Tatsache, dass $n \geq \exp(r+1)$ ist, und dem eben Gezeigten, erhalten wir die folgende Abschätzung

$$\begin{aligned}
& \sum_{j=1}^r \bar{w}_{1,1,j}^{(1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(1)}(\mathbf{x}) + \bar{w}_{1,1,0}^{(1)} \\
& \leq (r+1) \cdot \log n + \sum_{j=1}^r w_{1,1,j}^{(1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(1)}(\mathbf{x}) + w_{1,1,0}^{(1)} \\
& \leq (r+1) \cdot \log n + w_{1,1,i}^{(1)} \cdot f_{\bar{\mathbf{w}},1,i}^{(1)}(\mathbf{x}) + \sum_{j=1, j \neq i}^r w_{1,1,j}^{(1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(1)}(\mathbf{x}) + w_{1,1,0}^{(1)} \\
& \leq (r+1) \cdot \log n + 8 \cdot (\log n)^2 \cdot \frac{1}{n} + (2d-1) \cdot 8 \cdot (\log n)^2 - 8 \cdot (\log n)^2 \cdot \left(2d - \frac{1}{2}\right) \\
& = (r+1) \cdot \log n + 8 \cdot (\log n)^2 \cdot \left(\frac{1}{n} - \frac{1}{2}\right) \\
& \leq (\log n)^2 + 8 \cdot (\log n)^2 \cdot \frac{1}{n} - 8 \cdot (\log n)^2 \cdot \frac{1}{2} \\
& \leq -\log n.
\end{aligned}$$

Damit gilt

$$f_{\bar{\mathbf{w}},1,1}^{(2)}(\mathbf{x}) \leq \frac{1}{n}.$$

Durch die gleiche Argumentation wie zuvor, in Verbindung mit den Bedingungen (2.51)–(2.54) und der Annahme $n \geq \exp(r+1)$, können wir folgern, dass

$$\begin{aligned}
& \sum_{j=1}^r \bar{w}_{1,1,j}^{(l-1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(l-1)}(\mathbf{x}) + \bar{w}_{1,1,0}^{(l-1)} \\
& \leq (r+1) \cdot \log n + \sum_{j=1}^r w_{1,1,j}^{(l-1)} \cdot f_{\bar{\mathbf{w}},1,j}^{(l-1)}(\mathbf{x}) + w_{1,1,0}^{(l-1)} \\
& \leq (r+1) \cdot \log n + 6 \cdot (\log n)^2 \cdot \frac{1}{n} - 3 \cdot (\log n)^2 \\
& \leq (\log n)^2 + 6 \cdot (\log n)^2 \cdot \frac{1}{n} - 3 \cdot (\log n)^2 \\
& = -2 \cdot (\log n)^2 + 6 \cdot (\log n)^2 \cdot \frac{1}{n} \\
& \leq -\log n
\end{aligned}$$

für $l = 3, \dots, L$ gilt, wobei die letzte Ungleichung aus $n \geq 8d \geq 8$ folgt.

Somit erhalten wir $f_{\bar{\mathbf{w}},1,1}^{(l)}(\mathbf{x}) \leq \frac{1}{n}$ für $l = 3, \dots, L$.

Folglich ist

$$f_{\bar{\mathbf{w}},1,1}^{(L)}(\mathbf{x}) \leq \frac{1}{n}, \quad \text{falls } x^{(i)} \notin [u^{(i)} - \delta, v^{(i)} + \delta] \text{ für ein } i \in \{1, \dots, d\},$$

woraus sich die Aussage des Lemmas ergibt. \square

Durch das folgende Lemma erhalten wir eine obere Schranke für den Abstand zwischen der Ausgabe eines tiefen neuronalen Netzes und einer Lipschitz-stetigen Funktion im Inneren der Würfel einer Partition. Darüber hinaus stellt es sicher, dass die Ausgabe des neuronalen Netzes auf ganz \mathbb{R} beschränkt ist.

Lemma 8. Sei $0 < \delta \leq 1$ sowie $1 \leq \alpha_n \leq \log n$. Sei σ die logistische Sigmoidfunktion und sei $m : \mathbb{R}^d \rightarrow \mathbb{R}$ eine Lipschitz-stetige Funktion mit Lipschitz-Konstante $C_{\text{Lip}} > 0$. Zudem seien $L, n, r \in \mathbb{N}$ mit $L \geq 2, r \geq 2d, n \geq 8d, n \geq \exp(r+1)$ und $K, \hat{K}_n \in \mathbb{N}$ mit $K^d \leq \hat{K}_n$. Des Weiteren seien $a^{(1)}, \dots, a^{(d)}, b^{(1)}, \dots, b^{(d)} \in [-\alpha_n, \alpha_n]$ mit $b^{(i)} - a^{(i)} = \Delta$ für alle $i \in \{1, \dots, d\}$ und $\Delta \in \mathbb{R}_+$. Wir teilen den Würfel $[a^{(i)}, b^{(i)}] \times \dots \times [a^{(d)}, b^{(d)}]$ in K^d disjunkte gleichgroße Würfel der Seitenlänge $\frac{\Delta}{K}$, welche gegeben sich durch

$$[\mathbf{u}_s, \mathbf{v}_s] = [u_s^{(1)}, v_s^{(1)}] \times \dots \times [u_s^{(d)}, v_s^{(d)}]$$

für $s \in \{1, \dots, K^d\}$.

Das neuronale Netz $f_{\bar{\mathbf{w}}}$ sei definiert wie in (2.1)–(2.3) und der Gewichtsvektor \mathbf{w} erfülle in Abhängigkeit von $\mathbf{u}_s = (u_s^{(1)}, \dots, u_s^{(d)})$ und $\mathbf{v}_s = (v_s^{(1)}, \dots, v_s^{(d)})$ die Bedingungen

$$w_{s,j,j}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \quad \text{und} \quad w_{s,j,0}^{(0)} = \frac{-4d \cdot (\log n)^2}{\delta} \cdot u_s^{(j)} \quad \text{für } j \in \{1, \dots, d\}, \quad (2.55)$$

$$w_{s,j+d,j}^{(0)} = \frac{-4d \cdot (\log n)^2}{\delta} \quad \text{und} \quad w_{s,j+d,0}^{(0)} = \frac{4d \cdot (\log n)^2}{\delta} \cdot v_s^{(j)} \quad \text{für } j \in \{1, \dots, d\}, \quad (2.56)$$

$$w_{s,i,t}^{(0)} = 0 \quad \text{falls } i \leq 2d, i \neq t, i \neq t+d \text{ und } t > 0, \quad (2.57)$$

$$w_{s,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \text{für } t \in \{1, \dots, 2d\}, \quad (2.58)$$

$$w_{s,1,0}^{(1)} = -8 \cdot (\log n)^2 \cdot \left(2d - \frac{1}{2}\right) \quad (2.59)$$

$$w_{s,1,t}^{(1)} = 0 \quad \text{für } t > 2d, \quad (2.60)$$

$$w_{s,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \text{für } l \in \{2, \dots, L-1\}, \quad (2.61)$$

$$w_{s,1,0}^{(l)} = -3 \cdot (\log n)^2 \quad \text{für } l \in \{2, \dots, L-1\} \quad (2.62)$$

und

$$w_{s,1,t}^{(l)} = 0 \quad \text{für } t > 1 \text{ und } l \in \{2, \dots, L-1\} \quad (2.63)$$

für alle $s \in \{1, \dots, K^d\}$. Des Weiteren sei

$$w_{s,i,t}^{(l)} = 0 \quad \text{für } s \notin \{1, \dots, K^d\}.$$

Dann existieren

$$\bar{\alpha}_1, \dots, \bar{\alpha}_{K^d} \in [-\|m\|_\infty, \|m\|_\infty],$$

so dass für alle paarweise verschiedenen $k_1, \dots, k_{K^d} \in \{1, \dots, \hat{K}_n\}$ und für alle Gewichtsvektoren $\bar{\mathbf{w}}$ mit

$$\bar{w}_{1,1,k_s}^{(L)} = \bar{\alpha}_s \quad (s \in \{1, \dots, K^d\}), \quad \bar{w}_{1,1,s}^{(L)} = 0 \quad (s \notin \{k_1, \dots, k_{K^d}\}) \quad (2.64)$$

und

$$|w_{s,i,j}^{(l)} - \bar{w}_{k_s,i,j}^{(l)}| \leq \log n \quad \text{für alle } l \in \{0, \dots, L-1\}, s \in \{1, \dots, K^d\}$$

die Ungleichung

$$|f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})| \leq c_{21} \cdot \left(C_{\text{Lip}} \cdot \frac{\Delta}{K} + K^d \cdot \frac{1}{n} \right)$$

für alle $\mathbf{x} \in [a^{(1)}, b^{(1)}] \times \dots \times [a^{(d)}, b^{(d)}]$ gilt, welche nicht in der Menge

$$\bigcup_{j \in \{0,1,\dots,K\}} \bigcup_{i \in \{1,\dots,d\}} \left\{ \mathbf{x} \in \mathbb{R}^d : \left| x^{(i)} - \left(a^{(i)} + j \cdot \frac{b^{(i)} - a^{(i)}}{K} \right) \right| < \delta \right\} \quad (2.65)$$

enthalten sind.

Für $\delta \leq \frac{\Delta}{K}$ und $\mathbf{x} \in \mathbb{R}^d$ erhalten wir zusätzlich

$$|f_{\bar{\mathbf{w}}}(\mathbf{x})| \leq \|m\|_\infty \cdot \left(3^d + \frac{K^d}{n} \right). \quad (2.66)$$

Beweis. Der Einfachheit halber bezeichnen wir die Würfel $(\mathbf{u}_s, \mathbf{v}_s)$ mit C_s für $s \in \{1, \dots, K^d\}$.

Da die Funktion m Lipschitz-stetig mit Lipschitz-Konstante C_{Lip} ist, können wir sie durch eine Funktion $S : \mathbb{R}^d \rightarrow \mathbb{R}$, die stückweise konstant auf jedem der Teilwürfel ist, approximieren. Die Funktion S hat dann die Form

$$S(\mathbf{x}) = \sum_{s=1}^{K^d} \bar{\alpha}_s \cdot \mathbb{1}_{C_s}(\mathbf{x})$$

mit $\bar{\alpha}_s = m(\mathbf{z}_s)$ für $s \in \{1, \dots, K^d\}$, wobei \mathbf{z}_s den Mittelpunkt des Würfels C_s beschreibt.

Für $\mathbf{x} \in C_s$ nimmt der Funktionswert $S(\mathbf{x})$ den Wert $m(\mathbf{z}_s)$ an. Damit folgt aus der Lipschitz-Stetigkeit der Funktion m , dass

$$|S(\mathbf{x}) - m(\mathbf{x})| = |m(\mathbf{z}_s) - m(\mathbf{x})| \leq C_{\text{Lip}} \cdot \|\mathbf{z}_s - \mathbf{x}\|$$

gilt, falls $\mathbf{x} \in C_s$ für ein $s \in \{1, \dots, K^d\}$. Da jeder Würfel eine Seitenlänge von $\frac{\Delta}{K}$ hat, erhalten wir

$$|S(\mathbf{x}) - m(\mathbf{x})| \leq C_{\text{Lip}} \cdot \sqrt{d} \cdot \frac{\Delta}{K} \quad \text{für } \mathbf{x} \in [a^{(1)}, b^{(1)}] \times \dots \times [a^{(d)}, b^{(d)}].$$

Mit Bedingung (2.64), der Definition der Funktion S sowie der Definition des neuronalen Netzes $f_{\bar{\mathbf{w}}}$ ergibt sich

$$S(\mathbf{x}) - f_{\bar{\mathbf{w}}}(\mathbf{x}) = \sum_{s=1}^{K^d} \bar{\alpha}_s \left(\mathbb{1}_{C_s}(\mathbf{x}) - f_{\bar{\mathbf{w}}, k_s, 1}^{(L)}(\mathbf{x}) \right).$$

Für alle $\mathbf{x} \in [a^{(1)}, b^{(1)}] \times \dots \times [a^{(d)}, b^{(d)}]$, welche nicht in der Menge (2.65) enthalten sind, liefert die Anwendung von Lemma 7 die Ungleichung

$$|S(\mathbf{x}) - f_{\bar{\mathbf{w}}}(\mathbf{x})| \leq K^d \cdot \|m\|_{\infty} \cdot \frac{1}{n}.$$

Die letzte Ungleichung folgt hierbei aus $\bar{\alpha}_1, \dots, \bar{\alpha}_{K^d} \in [-\|m\|_{\infty}, \|m\|_{\infty}]$. Somit ergibt sich

$$|f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})| = |f_{\bar{\mathbf{w}}}(\mathbf{x}) - S(\mathbf{x})| + |S(\mathbf{x}) - m(\mathbf{x})| \leq c_{21} \cdot \left(K^d \cdot \frac{1}{n} + C_{\text{Lip}} \cdot \frac{\Delta}{K} \right)$$

für alle $\mathbf{x} \in [a^{(1)}, b^{(1)}] \times \dots \times [a^{(d)}, b^{(d)}]$, die nicht in der Menge (2.65) enthalten sind.

Um Ungleichung (2.66) nachzuweisen, nehmen wir an, dass $\mathbf{x} \in \mathbb{R}^d$ ist. Dann gilt für $\delta \leq \frac{\Delta}{K}$ und jedes feste $\mathbf{x} \in \mathbb{R}^d$ die folgende Gleichheit

$$\begin{aligned} |f_{\bar{\mathbf{w}}}(\mathbf{x})| &= \sum_{s=1}^{K^d} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}}, k_s, 1}^{(L)}(\mathbf{x}) \\ &= \sum_{s \in \{1, \dots, K^d\} : \mathbf{x} \in [u_s^{(1)} - \delta, v_s^{(1)} + \delta] \times \dots \times [u_s^{(d)} - \delta, v_s^{(d)} + \delta]} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}}, k_s, 1}^{(L)}(\mathbf{x}) \\ &\quad + \sum_{s \in \{1, \dots, K^d\} : \mathbf{x} \notin [u_s^{(1)} - \delta, v_s^{(1)} + \delta] \times \dots \times [u_s^{(d)} - \delta, v_s^{(d)} + \delta]} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}}, k_s, 1}^{(L)}(\mathbf{x}). \end{aligned}$$

Da höchstens 3^d Würfel der Form $[u_i^{(1)} - \delta, v_i^{(1)} + \delta] \times \dots \times [u_i^{(d)} - \delta, v_i^{(d)} + \delta]$ den Vektor \mathbf{x} enthalten, ergibt sich aufgrund der Definition der Gewichte $\bar{w}_{1,1,k_s}^{(L)}$ für $s \in \{1, \dots, K^d\}$ sowie der Schranke $|\sigma(z)| \leq 1$ für alle $z \in \mathbb{R}$ die Ungleichung

$$\sum_{s \in \{1, \dots, K^d\} : \mathbf{x} \in [u_s^{(1)} - \delta, v_s^{(1)} + \delta] \times \dots \times [u_s^{(d)} - \delta, v_s^{(d)} + \delta]} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}}, k_s, 1}^{(L)}(\mathbf{x}) \leq 3^d \cdot \|m\|_{\infty}.$$

Aus Lemma 7 ist bekannt, dass für die Ausgabe eines vollständig verbundenen neuronalen Netzes

$$f_{\bar{\mathbf{w}}, k_s, 1}^{(L)}(\mathbf{x}) \leq \frac{1}{n} \quad \text{für } s \in \{1, \dots, K^d\}$$

gilt, falls $\mathbf{x} \notin [u_s^{(1)} - \delta, v_s^{(1)} + \delta] \times \cdots \times [u_s^{(d)} - \delta, v_s^{(d)} + \delta]$. Daher erhalten wir aus der Definition der Gewichte $\bar{w}_{1,1,k_s}^{(L)}$ für $s \in \{1, \dots, K^d\}$ die Ungleichung

$$\sum_{s \in \{1, \dots, K^d\} : \mathbf{x} \notin [u_s^{(1)} - \delta, v_s^{(1)} + \delta] \times \cdots \times [u_s^{(d)} - \delta, v_s^{(d)} + \delta]} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}},k_s,1}^{(L)}(\mathbf{x}) \leq K^d \cdot \|m\|_\infty \cdot \frac{1}{n}.$$

Die Kombination dieser beiden Abschätzungen führt zu

$$\begin{aligned} |f_{\bar{\mathbf{w}}}(\mathbf{x})| &= \sum_{s=1}^{K^d} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}},k_s,1}^{(L)}(\mathbf{x}) \\ &= \sum_{s \in \{1, \dots, K^d\} : \mathbf{x} \in [u_s^{(1)} - \delta, v_s^{(1)} + \delta] \times \cdots \times [u_s^{(d)} - \delta, v_s^{(d)} + \delta]} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}},k_s,1}^{(L)}(\mathbf{x}) \\ &\quad + \sum_{s \in \{1, \dots, K^d\} : \mathbf{x} \notin [u_s^{(1)} - \delta, v_s^{(1)} + \delta] \times \cdots \times [u_s^{(d)} - \delta, v_s^{(d)} + \delta]} \bar{w}_{1,1,k_s}^{(L)} \cdot f_{\bar{\mathbf{w}},k_s,1}^{(L)}(\mathbf{x}) \\ &\leq 3^d \cdot \|m\|_\infty + K^d \cdot \|m\|_\infty \cdot \frac{1}{n} \end{aligned}$$

für $\mathbf{x} \in \mathbb{R}^d$, woraus die Behauptung von Lemma 8 folgt. \square

Im nächsten Lemma leiten wir eine Schranke für den Approximationsfehler her, der bei der Approximation einer beschränkten, Lipschitz-stetigen Funktion durch ein tiefes neuronales Netz entsteht.

Lemma 9. Sei $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}$ die logistische Sigmoidfunktion und $1 \leq \alpha_n \leq \log n$. Die Funktion $m : \mathbb{R}^d \rightarrow \mathbb{R}$ sei sowohl beschränkt als auch Lipschitz-stetig. Zudem gelte $L, n, r \in \mathbb{N}$ mit $L \geq 2$, $r \geq 2d$, $n \geq 8d$ und $n \geq \exp(r+1)$ sowie $K, \hat{K}_n, N_n \in \mathbb{N}$ mit $2 \leq K \leq \alpha_n - 1$ und $N_n \cdot (K^2 + 1)^d \leq \hat{K}_n$. Des Weiteren unterteilen wir den Würfel $[-K - \frac{2}{K}, K]^d$ in ein Gitter von $(K^2 + 1)^d$ äquidistanten Würfeln der Seitenlänge $\frac{2}{K}$ und bezeichnen diese mit

$$C_s := [u_s^{(1)}, v_s^{(1)}] \times \cdots \times [u_s^{(d)}, v_s^{(d)}]$$

für $s \in \{1, \dots, (K^2 + 1)^d\}$.

Der Gewichtsvektor \mathbf{w} erfülle die Bedingungen

$$w_{s,j,j}^{(0)} = 4d \cdot K^2 \cdot (\log n)^2 \quad \text{und} \quad w_{s,j,0}^{(0)} = -4d \cdot K^2 \cdot (\log n)^2 \cdot u_s^{(j)} \quad \text{für } j \in \{1, \dots, d\}, \quad (2.67)$$

$$w_{s,j+d,j}^{(0)} = -4d \cdot K^2 \cdot (\log n)^2 \quad \text{und} \quad w_{s,j+d,0}^{(0)} = 4d \cdot K^2 \cdot (\log n)^2 \cdot v_s^{(j)} \quad \text{für } j \in \{1, \dots, d\}, \quad (2.68)$$

$$w_{s,i,t}^{(0)} = 0 \quad \text{falls } i \leq 2d, i \neq t, i \neq t + d \text{ und } t > 0, \quad (2.69)$$

$$w_{s,1,t}^{(1)} = 8 \cdot (\log n)^2 \quad \text{für } t \in \{1, \dots, 2d\}, \quad (2.70)$$

$$w_{s,1,0}^{(1)} = -8 \cdot (\log n)^2 \left(2d - \frac{1}{n}\right), \quad (2.71)$$

$$w_{s,1,t}^{(1)} = 0 \quad \text{für } t > 2d, \quad (2.72)$$

$$w_{s,1,1}^{(l)} = 6 \cdot (\log n)^2 \quad \text{für } l \in \{2, \dots, L-1\}, \quad (2.73)$$

$$w_{s,1,0}^{(l)} = -3 \cdot (\log n)^2 \quad \text{für } l \in \{2, \dots, L-1\} \quad (2.74)$$

und

$$w_{s,1,t}^{(l)} = 0 \quad \text{für } t > 1 \text{ und } l \in \{2, \dots, L-1\} \quad (2.75)$$

für alle $s \in \{1, \dots, N_n \cdot (K^2 + 1)^d\}$.

Zudem sei

$$w_{s,i,t}^{(l)} = 0 \quad \text{für } s \notin \{1, \dots, N_n \cdot (K^2 + 1)^d\}. \quad (2.76)$$

Dann existieren

$$\bar{\alpha}_1, \dots, \bar{\alpha}_{N_n(K^2+1)^d} \in \left[-\frac{\|m\|_\infty}{N_n}, \frac{\|m\|_\infty}{N_n} \right], \quad (2.77)$$

so dass für alle paarweise verschiedenen $k_1, \dots, k_{N_n(K^2+1)^d} \in \{1, \dots, \widehat{K}_n\}$ und für alle Gewichtsvektoren $\bar{\mathbf{w}}$ mit

$$\bar{w}_{1,1,k_s}^{(L)} = \bar{\alpha}_s \quad \left(s \in \{1, \dots, N_n \cdot (K^2 + 1)^d\} \right), \quad \bar{w}_{1,1,s}^{(L)} = 0 \quad \left(s \notin \{k_1, \dots, k_{N_n(K^2+1)^d}\} \right) \quad (2.78)$$

sowie

$$|w_{s,i,j}^{(l)} - \bar{w}_{k_s,i,j}^{(l)}| \leq \log n \quad \text{für alle } l \in \{0, \dots, L-1\}, s \in \{1, \dots, N_n \cdot (K^2 + 1)^d\} \quad (2.79)$$

die Ungleichung

$$\int |f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_{22} \cdot \left(\frac{1}{K} + \frac{N_n^2 \cdot K^{4d}}{n^2} + \left(\frac{K^{2d}}{n} + 1 \right)^2 \cdot \mathbf{P}_{\mathbf{X}}(\mathbb{R}^d \setminus [-K, K]^d) \right) \quad (2.80)$$

gilt. Zusätzlich erhalten wir durch

$$\|f_{\bar{\mathbf{w}}}\|_\infty \leq c_{23} \cdot \left(3^d + \frac{(K^2 + 1)^d}{n} \right) \quad (2.81)$$

eine obere Schranke für die Ausgabe des neuronalen Netzes, wobei die Konstante $c_{23} > 0$ von $\|m\|_\infty$ abhängt.

Beweis. Sei $C_{\text{Lip}} > 0$ die Lipschitz-Konstante der Funktion m . Für den Beweis von Lemma 9 werden wir Lemma 8 auf m/N_n anwenden. Hierbei ersetzen wir K in Lemma 8 durch $K^2 + 1$ und setzen $a^{(i)} = -K - \frac{2}{K}$, $b^{(i)} = K$. Dies führt zu

$$\left| f_{\bar{\mathbf{w}}}(\mathbf{x}) - \frac{1}{N_n} \cdot m(\mathbf{x}) \right| \leq c_{21} \cdot \left(\frac{C_{\text{Lip}}}{N_n} \cdot \frac{2}{K} + (K^2 + 1)^d \cdot \frac{1}{n} \right)$$

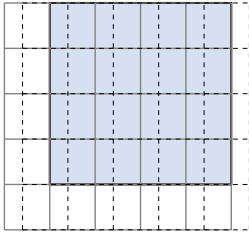
für alle $\mathbf{x} \in [-K - \frac{2}{K}, K]^d$, welche nicht in der Menge

$$A := \bigcup_{j \in \{0, 1, \dots, K^2+1\}} \bigcup_{i \in \{1, \dots, d\}} \left\{ \mathbf{x} \in \mathbb{R}^d : \left| x^{(i)} - \left(-K - \frac{2}{K} + j \cdot \frac{2}{K} \right) \right| < \delta \right\} \quad (2.82)$$

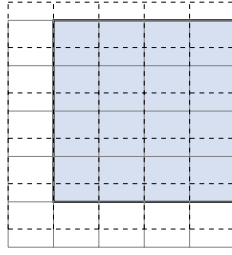
enthalten sind. Durch N_n -maliges Wiederholen dieser Konstruktion erhalten wir eine Approximation $f_{\bar{\mathbf{w}}}$ von $N_n \cdot (1/N_n) \cdot m$, welche die Ungleichung

$$|f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})| \leq c_{24} \cdot \left(\frac{1}{K} + N_n \cdot (K^2 + 1)^d \cdot \frac{1}{n} \right) \quad (2.83)$$

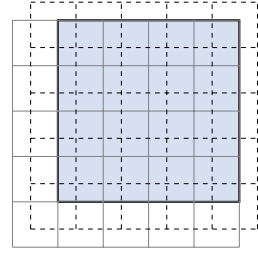
für $\mathbf{x} \in [-K - \frac{2}{K}, K]^d \setminus A$ erfüllt.



(a) Verschiebung in $x^{(1)}$ -Richtung



(b) Verschiebung in $x^{(2)}$ -Richtung



(c) Verschiebung in $x^{(1)}$ - und $x^{(2)}$ -Richtung

Abbildung 2.2: Verschiedene Verschiebungen des Gitters, um den Würfel $[-K, K]^d$ abzudecken

Im nächsten Schritt verschieben wir das Gitter entlang der i -ten Komponente, so dass der Würfel $[-K, K]^d$ immer überdeckt ist. Die zugrundeliegende Idee ist in Abbildung 2.2 dargestellt. Die blaue Fläche kennzeichnet das Quadrat $[-K, K]^2$, welches von $[-K - \frac{2}{K}, K]$ überdeckt wird. In jedem der drei Bilder wird das Gitter so verschoben, dass das gesamte Quadrat weiterhin vollständig abgedeckt bleibt.

Wir addieren also für festes $i \in \{1, \dots, d\}$ auf alle $u_j^{(i)}, v_j^{(i)}$ denselben Summanden, welcher aus der Menge

$$\left\{ k \cdot \frac{2}{K^2} \quad : \quad k = 0, 1, \dots, K-1 \right\}$$

stammt. Damit erhalten wir K verschiedene Versionen der Approximation $f_{\bar{w}}$, welche die Ungleichung (2.83) für alle $\mathbf{x} \in [-K, K]^d$, die nicht in den entsprechenden Versionen der Menge A enthalten sind, erfüllen.

Durch Verschieben des Gitters erhalten wir für ein festes $i \in \{1, \dots, d\}$ eine Anzahl von K disjunkten Versionen der Menge

$$\bigcup_{j \in \{0, 1, \dots, K^2+1\}} \left\{ \mathbf{x} \in \mathbb{R}^d : \left| x^{(i)} - \left(-K - \frac{2}{K} + j \cdot \frac{2}{K} \right) \right| < \delta \right\}.$$

Die Summe der Wahrscheinlichkeitsmaße $\mathbf{P}_{\mathbf{x}}$ dieser K disjunkten Mengen ist kleiner gleich eins. Aus diesem Grund muss mindestens eines dieser Maße kleiner gleich $\frac{1}{K}$ sein. Daher können wir \mathbf{u}_j und \mathbf{v}_j so verschieben, dass

$$\mathbf{P}_{\mathbf{x}}(A) \leq \sum_{i=1}^d \frac{1}{K} = \frac{d}{K}$$

gilt.

Damit haben wir gezeigt, dass eine Version dieses Gitters existiert, so dass die Menge A ein Maß kleiner gleich $\frac{d}{K}$ hat.

Des Weiteren folgt aufgrund der oberen Schranke des neuronalen Netzes in (2.66) aus Lemma 8 sowie der Beschränktheit von m , dass

$$|f_{\bar{w}}(\mathbf{x})| \leq c_{25} \cdot \left(3^d + (K^2 + 1)^d \cdot \frac{1}{n} \right)$$

für $\mathbf{x} \in \mathbb{R}^d$ erfüllt ist.

Dies zusammen mit der Tatsache, dass

$$|f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})| \leq c_{26} \cdot \left(\frac{1}{K} + \frac{N_n \cdot K^{2d}}{n} \right)$$

für $\mathbf{x} \in [-K, K]^d \setminus A$ gilt, was sich aus Ungleichung (2.83) ergibt, liefert

$$\begin{aligned} & \int |f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &= \int_{[-K, K]^d \setminus A} |f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) + \int_A |f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &\quad + \int_{\mathbb{R}^d \setminus [-K, K]^d} |f_{\bar{\mathbf{w}}}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &\leq c_{27} \cdot \left(\frac{1}{K} + \frac{N_n \cdot K^{2d}}{n} \right)^2 + c_{28} \cdot \left(3^d + \frac{K^{2d}}{n} \right)^2 \cdot \frac{d}{K} \\ &\quad + c_{29} \cdot \left(3^d + \frac{K^{2d}}{n} \right)^2 \cdot \mathbf{P}_{\mathbf{X}}(\mathbb{R}^d \setminus [-K, K]^d) \end{aligned}$$

für Konstanten $c_{27}, c_{28}, c_{29} > 0$. Durch das Vereinfachen der Terme ergibt sich die Aussage des Lemmas. \square

2.2.4 Beweis des Resultats zur universellen Konsistenz

Die Idee des Beweises zur universellen Konsistenz des überparametrisierten Neuronale-Netze-Schätzers m_n ist es, ein Ereignis A_n zu betrachten, bei dem die inneren Startgewichte höchstens um einen logarithmischen Faktor von den Werten der inneren Gewichte des Netzes mit guten Approximationseigenschaften abweichen, wobei die äußeren Gewichte geeignet gewählt sind. Zusätzlich soll das arithmetische Mittel der Zufallsvariablen Y_i^2 durch einen kubischen logarithmischen Faktor beschränkt sein. Dann kann gezeigt werden, dass die Wahrscheinlichkeit, sich nicht in diesem Ereignis zu befinden, für eine wachsende Stichprobengröße n gegen 0 geht. Durch Aufspalten des L_2 -Fehlers in mehrere Terme können wir mithilfe der vorherigen Lemmata die universelle Konsistenz des Schätzers nachweisen.

Beweis von Theorem 1. Sei $\varepsilon > 0$, $K \in \mathbb{N}$ mit $K \geq 2$ beliebig und $N_n = \lceil c_{30} \cdot (\log n)^{18L} \rceil$. Aufgrund der Ungleichung von Jensen für bedingte Erwartungswerte gilt

$$\mathbf{E} \{ |m(\mathbf{X})|^2 \} = \mathbf{E} \{ |\mathbf{E}\{Y|\mathbf{X}\}|^2 \} \leq \mathbf{E} \{ |\mathbf{E}\{Y^2|\mathbf{X}\}| \} = \mathbf{E} \{ Y^2 \} < \infty.$$

Daher existiert eine beschränkte sowie Lipschitz-stetige Funktion $\bar{m} : \mathbb{R}^d \rightarrow \mathbb{R}$ mit

$$\int |\bar{m}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq \varepsilon.$$

Wir bezeichnen mit A_n das Ereignis, dass der Gewichtsvektor $\mathbf{w}^{(0)}$ für paarweise verschiedene Indizes $k_1, \dots, k_{N_n(K^2+1)^d}$ die Bedingung

$$\left| (\mathbf{w}^{(0)})_{k_s, i, j}^{(l)} - w_{s, i, j}^{(l)} \right| \leq \log n \quad \text{für alle } l \in \{0, \dots, L-1\}, s \in \{1, \dots, N_n \cdot (K^2+1)^d\} \quad (2.84)$$

erfüllt, wobei \mathbf{w} ein Gewichtsvektor ist, der die Voraussetzungen (2.67)–(2.76) in Lemma 9 für \bar{m} erfüllt. Zusätzlich soll auf A_n die Ungleichung

$$\frac{1}{n} \sum_{i=1}^n Y_i^2 \leq \beta_n^3$$

erfüllt sein.

Tritt das Ereignis A_n ein, so wählen wir $\bar{\alpha}_s$ gemäß Lemma 9 für \bar{m} . Andernfalls setzen wir $\bar{\alpha}_1 = \dots = \bar{\alpha}_{N_n(K^2+1)^d} = 0$. Den Gewichtsvektor \mathbf{w}^* definieren wir für einen gegebenen Gewichtsvektor \mathbf{w} durch

$$\begin{aligned} (\mathbf{w}^*)_{k,i,j}^{(l)} &= w_{k,i,j}^{(l)} \quad \text{für alle } l \in \{0, \dots, L-1\}, \\ (\mathbf{w}^*)_{1,1,k_s}^{(L)} &= \bar{\alpha}_s \quad \text{für alle } s \in \{1, \dots, N_n \cdot (K^2+1)^d\}, \\ (\mathbf{w}^*)_{1,1,s}^{(L)} &= 0 \quad \text{für alle } s \notin \{k_1, \dots, k_{N_n(K^2+1)^d}\}. \end{aligned}$$

Insbesondere ist dann der Gewichtsvektor $(\mathbf{w}^{(0)})^*$ gegeben durch

$$\begin{aligned} ((\mathbf{w}^{(0)})^*)_{k,i,j}^{(l)} &= (\mathbf{w}^{(0)})_{k,i,j}^{(l)} \quad \text{für alle } l \in \{0, \dots, L-1\}, \\ ((\mathbf{w}^{(0)})^*)_{1,1,k_s}^{(L)} &= \bar{\alpha}_s \quad \text{für alle } s \in \{1, \dots, N_n \cdot (K^2+1)^d\}, \\ ((\mathbf{w}^{(0)})^*)_{1,1,s}^{(L)} &= 0 \quad \text{für alle } s \notin \{k_1, \dots, k_{N_n(K^2+1)^d}\}. \end{aligned}$$

Im *ersten Schritt des Beweises* beginnen wir damit, den L_2 -Fehler in eine Summe mehrerer Terme zu zerlegen. Damit erhalten wir

$$\begin{aligned} & \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &= (\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}) \cdot \mathbf{1}_{A_n} \\ & \quad + \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{A_n^c} \\ &= (\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - (1 + \varepsilon) \cdot \mathbf{E}\{|m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\}) \cdot \mathbf{1}_{A_n} \\ & \quad + \left((1 + \varepsilon) \cdot \mathbf{E}\{|m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} - (1 + \varepsilon) \cdot \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \cdot \mathbf{1}_{A_n} \\ & \quad + \left((1 + \varepsilon) \cdot \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - (1 + \varepsilon) \cdot \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t_n)}}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \cdot \mathbf{1}_{A_n} \\ & \quad + \left((1 + \varepsilon) \cdot \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t_n)}}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - (1 + \varepsilon)^2 \cdot \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t_n)}}(\mathbf{X}_i) - Y_i|^2 \right) \cdot \mathbf{1}_{A_n} \\ & \quad + \left((1 + \varepsilon)^2 \cdot \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t_n)}}(\mathbf{X}_i) - Y_i|^2 - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \right) \cdot \mathbf{1}_{A_n} \\ & \quad + \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{A_n^c} \\ &= \sum_{j=1}^6 T_{j,n}. \end{aligned}$$

Im zweiten Schritt des Beweises zeigen wir, dass

$$\limsup_{n \rightarrow \infty} \mathbf{E}\{T_{j,n}\} \leq 0 \quad \text{für } j \in \{1, 4\}$$

gilt.

Die Aussage ergibt sich mithilfe der Ungleichung

$$(a + b)^2 \leq (1 + \varepsilon) \cdot a^2 + \left(1 + \frac{1}{\varepsilon}\right) \cdot b^2 \quad \text{für } a, b \in \mathbb{R}.$$

Aus den Definitionen der Terme $T_{1,n}$ und $T_{4,n}$ erhalten wir damit

$$\begin{aligned} \mathbf{E}\{T_{1,n}\} &= \mathbf{E}\left\{\left(\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - (1 + \varepsilon) \cdot \mathbf{E}\{|m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\}\right) \cdot \mathbf{1}_{A_n}\right\} \\ &\leq \left(1 + \frac{1}{\varepsilon}\right) \cdot \mathbf{E}\{|T_{\beta_n} Y - Y|^2\} \end{aligned}$$

und

$$\begin{aligned} \mathbf{E}\{T_{4,n}\} &= (1 + \varepsilon) \cdot \mathbf{E}\left\{\left(\frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}(t_n)}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - (1 + \varepsilon) \cdot \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}(t_n)}(\mathbf{X}_i) - Y_i|^2\right) \cdot \mathbf{1}_{A_n}\right\} \\ &\leq (1 + \varepsilon) \cdot \left(1 + \frac{1}{\varepsilon}\right) \cdot \mathbf{E}\{|T_{\beta_n} Y - Y|^2\}. \end{aligned}$$

Betrachten wir die beiden Ausdrücke für $n \rightarrow \infty$, womit auch $\beta_n \rightarrow \infty$ gilt, folgt wegen der Voraussetzung $\mathbf{E}\{Y^2\} < \infty$ die Behauptung des zweiten Beweisschrittes.

Im dritten Schritt des Beweises zeigen wir, dass

$$\limsup_{n \rightarrow \infty} \mathbf{E}\{T_{3,n}\} \leq 0$$

gilt.

Angenommen es ist $|y| \leq \beta_n$, dann gilt für alle $z \in \mathbb{R}$ die Ungleichung

$$|T_{\beta_n} z - y| \leq |z - y|.$$

Aufgrund der Definition des Schätzers m_n erhalten wir durch diese Ungleichung

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 &= \frac{1}{n} \sum_{i=1}^n |T_{\beta_n} f_{\mathbf{w}(t_n)}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}(t_n)}(\mathbf{X}_i) - Y_i|^2. \end{aligned}$$

Damit ergibt sich

$$\frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}(t_n)}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \leq 0,$$

woraus die Behauptung des dritten Beweisschrittes folgt.

Im vierten Schritt des Beweises weisen wir nach, dass die Voraussetzungen (2.9)–(2.11) aus Lemma 1 erfüllt sind, sofern das Ereignis A_n eintritt. Sei hierfür die Menge S definiert durch

$$S := \left\{ \mathbf{v} : \|\mathbf{v} - \mathbf{w}^{(0)}\| \leq 2 \cdot \sqrt{F(\mathbf{w}^{(0)}) + 1} \right\}.$$

Des Weiteren definieren wir die Menge \tilde{S} durch

$$\tilde{S} := \left\{ (\tilde{w}_{k,i,j}^{(l)})_{k,i,j,l:l < L} : \|(\tilde{w}_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\| \leq \sqrt{2 \cdot F(\mathbf{w}^{(0)})} \right\}.$$

Im Folgenden werden wir zeigen, dass die Ungleichungen

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w})\| \leq C_n \quad (2.85)$$

für alle $\mathbf{w} \in S$,

$$\|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}) - (\nabla_{\mathbf{w}} F_n)(\bar{\mathbf{w}})\| \leq C_n \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \quad (2.86)$$

für alle $\mathbf{w}, \bar{\mathbf{w}} \in S$ und

$$\left| F_n(\mathbf{w}^*) - F_n((\mathbf{w}^{(0)})^*) \right| \leq D_n \cdot \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\| \cdot \|(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\| \quad (2.87)$$

für alle \mathbf{w} mit $(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} \in \tilde{S}$ erfüllt sind, wenn das Ereignis A_n eintritt.

Da die äußeren Startgewichte $(\mathbf{w}^{(0)})_{1,1,k}^{(L)} = 0$ für $k = 1, \dots, \hat{K}_n$ sind, gilt auf dem Ereignis A_n , dass

$$F_n(\mathbf{w}^{(0)}) = \frac{1}{n} \sum_{i=1}^n Y_i^2 \leq \beta_n^3$$

ist.

Sei nun $\mathbf{w} \in S$. Aufgrund der Startinitialisierung der Gewichte können wir folgern, dass die Abschätzungen

$$\begin{aligned} \|(w_{k,i,j}^{(l)})_{k,i,j,l:l \leq l < L}\|_{\infty} &\leq \|(w_{k,i,j}^{(l)})_{k,i,j,l:l \leq l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{i,j,k,l:l \leq l < L}\|_{\infty} + \|((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{i,j,k,l:l \leq l < L}\|_{\infty} \\ &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\|_{\infty} + \|((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{i,j,k,l:l \leq l < L}\|_{\infty} \\ &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l \leq l < L}\|_{\infty} \\ &\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)}) + 1} + 20d \cdot (\log n)^2 \\ &\leq c_{31} \cdot (\log n)^2 \end{aligned}$$

und

$$\begin{aligned} \|(w_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\|_{\infty} &\leq \|(w_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n} - ((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\|_{\infty} + \|((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\|_{\infty} \\ &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\|_{\infty} + \|((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\|_{\infty} \\ &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\|_{\infty} \\ &\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)}) + 1} \\ &\leq c_{32} \cdot (\log n)^{3/2} \end{aligned}$$

gelten. Für $\bar{\mathbf{w}} \in S$ folgt dies analog. Damit sind die Voraussetzungen (2.19) und (2.20) von Lemma 2 sowie die Voraussetzungen (2.28) und (2.29) von Lemma 3 für $B_n = c_{31} \cdot (\log n)^2$ und $\gamma_n^* = c_{32} \cdot (\log n)^{3/2}$ erfüllt. Zusätzlich ist

$$\begin{aligned} \|\mathbf{w} - \mathbf{w}^{(0)}\|_\infty^2 &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\|^2 \\ &\leq \left(2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1\right)^2 \\ &\leq 2 \cdot \left(2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})}\right)^2 + 2 \\ &\leq 8 \cdot F_n(\mathbf{w}^{(0)}) + 2. \end{aligned}$$

Wegen $t_n = \lceil c_6 \cdot C_n \rceil$ mit $c_6 \geq 2$ sind daher die Voraussetzungen (2.21) und (2.30) ebenfalls erfüllt. Zudem können wir wegen des beschränkten Trägers von \mathbf{X} annehmen, dass $\alpha_n = c_{33}$ ist. Unter der Voraussetzung, dass $C_n \geq \widehat{K}_n^{3/2} \cdot (\log n)^{6L+5}$ gilt, erhalten wir aus Lemma 2 die Ungleichung

$$\begin{aligned} \|\nabla_{\mathbf{w}} F_n(\mathbf{w})\| &\leq c_7 \cdot \widehat{K}_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n \cdot \sqrt{\frac{t_n}{C_n} \cdot \max\{F_n(\mathbf{w}^{(0)}), 1\}} \\ &\leq c_{34} \cdot \widehat{K}_n^{3/2} \cdot (\log n)^{4L} \cdot (\log n)^3 \cdot \sqrt{(\log n)^3} \\ &\leq c_{34} \cdot \widehat{K}_n^{3/2} \cdot (\log n)^{4L} \cdot (\log n)^{3+3/2} \\ &\leq C_n \end{aligned}$$

und aus Lemma 3 die Ungleichung

$$\begin{aligned} \|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}) - (\nabla_{\mathbf{w}} F_n)(\bar{\mathbf{w}})\| &\leq c_{11} \cdot \widehat{K}_n^{3/2} \cdot B_n^{3L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^3 \cdot \sqrt{\frac{t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\}} \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \\ &\leq c_{35} \cdot \widehat{K}_n^{3/2} \cdot (\log n)^{6L} \cdot (\log n)^3 \cdot \sqrt{(\log n)^3} \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \\ &\leq c_{35} \cdot \widehat{K}_n^{3/2} \cdot (\log n)^{6L} \cdot (\log n)^{3+3/2} \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \\ &\leq C_n \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \end{aligned}$$

für hinreichend großes n . Damit folgt die Gültigkeit der Bedingungen (2.85) und (2.86).

Es bleibt noch zu zeigen, dass Ungleichung (2.87) gilt. Sei hierfür der Gewichtsvektor \mathbf{w} , so dass $(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} \in \widetilde{S}$ ist. Dann erhalten wir durch Anwenden der dritten binomischen Formel, der Ungleichung von Cauchy-Schwarz sowie der Ungleichung $(a+b)^2 \leq 2a^2 + 2b^2$ für $a, b \in \mathbb{R}$, dass

$$\begin{aligned} &\left| F_n(\mathbf{w}^*) - F_n((\mathbf{w}^{(0)})^*) \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^*}(\mathbf{X}_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i|^2 \right| \\ &= \frac{1}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) - Y_i + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i \right) \left(f_{\mathbf{w}^*}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right) \\ &\leq \left(\frac{1}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) - Y_i + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i \right)^2 \right)^{1/2} \cdot \left(\frac{1}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 \right)^{1/2} \\ &\leq \left(\frac{2}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \end{aligned}$$

$$\begin{aligned} & \cdot \left(\frac{1}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 \right)^{1/2} \\ & = T_{7,n} \cdot T_{8,n} \end{aligned}$$

gilt.

Im Folgenden werden wir diese beiden Terme abschätzen. Gemäß der Definition des Gewichtsvektors \mathbf{w}^* ist $(\mathbf{w}^*)_{1,1,k}^{(L)} = ((\mathbf{w}^{(0)})^*)_{1,1,k}^{(L)}$ für alle $k \in \{1, \dots, \widehat{K}_n\}$. Für den Term $T_{7,n}$ ergibt sich damit

$$\begin{aligned} T_{7,n} & = \left(\frac{2}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ & = \left(\frac{2}{n} \sum_{i=1}^n \left(f_{(\mathbf{w}^*)^*}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) + 2 \cdot f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ & \leq \left(\frac{2}{n} \sum_{i=1}^n 2 \cdot \left(f_{\mathbf{w}^*}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 + \frac{2}{n} \sum_{i=1}^n 8 \cdot f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ & \leq \left(4 \cdot \frac{1}{n} \sum_{i=1}^n \left(\sum_{k=1}^{\widehat{K}_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{\widehat{K}_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \right) \right. \\ & \quad \left. + 16 \cdot \frac{1}{n} \sum_{i=1}^n f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ & \leq \left(4 \cdot \sum_{k=1}^{\widehat{K}_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \max_{i=1, \dots, n} \sum_{k=1}^{\widehat{K}_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \right. \\ & \quad \left. + 16 \cdot \frac{1}{n} \sum_{i=1}^n f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2}. \end{aligned}$$

Durch Anwenden von Lemma 9 auf \bar{m} erhalten wir wegen der Beschränktheit von \bar{m} die Abschätzung

$$|\bar{\alpha}_s| \leq \frac{c_{36}}{N_n} \quad \text{für } s \in \{1, \dots, N_n \cdot (K^2 + 1)^d\} \quad (2.88)$$

sowie

$$\left\| f_{(\mathbf{w}^{(0)})^*} \right\|_{\infty} \leq c_{23} \cdot \left(3^d + \frac{(K^2 + 1)^d}{n} \right).$$

Wie bereits im Beweis von Lemma 2 gezeigt, ist

$$\begin{aligned} & \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{x}) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{x}) \right| \\ & \leq c_{37} \cdot \max\{\|\sigma'\|_{\infty}^L, 1\} \cdot (2r + 1)^L \cdot B_n^L \cdot \alpha_n \cdot \max_{i,j,\ell:\ell < L} \left| (\mathbf{w}^*)_{k,i,j}^{(\ell)} - ((\mathbf{w}^{(0)})^*)_{k,i,j}^{(\ell)} \right| \cdot \max\{\|\sigma\|_{\infty}, 1\} \end{aligned}$$

für $\mathbf{x} \in \mathbb{R}^d$. Mit der Wahl von $\alpha_n = c_{33}$ und $B_n = c_{31} \cdot (\log n)^2$ ergibt sich

$$\left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{x}) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{x}) \right| \leq c_{38} \cdot (\log n)^{2L} \cdot \max_{i,j,\ell:\ell < L} \left| (\mathbf{w}^*)_{k,i,j}^{(\ell)} - ((\mathbf{w}^{(0)})^*)_{k,i,j}^{(\ell)} \right|. \quad (2.89)$$

Daher erhalten wir für $(w_{i,j,k}^{(l)})_{i,j,k,l:l < L} \in \tilde{S}$, dass

$$\begin{aligned}
& \max_{i=1,\dots,n} \sum_{k=1}^{\hat{K}_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \\
& \leq c_{38}^2 \cdot (\log n)^{4L} \cdot \sum_{k=1}^{\hat{K}_n} \max_{i,j,l:l < L} \left| (\mathbf{w}^*)_{k,i,j}^{(l)} - ((\mathbf{w}^{(0)})^*)_{k,i,j}^{(l)} \right|^2 \\
& \leq c_{38}^2 \cdot (\log n)^{4L} \cdot \|(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\|^2 \\
& \leq c_{39} \cdot (\log n)^{4L+3}
\end{aligned}$$

gilt. Zusammen mit der Definition des Gewichtsvektors \mathbf{w}^* und Ungleichung (2.88) ergibt sich damit

$$\begin{aligned}
& T_{7,n} \\
& \leq \left(4 \cdot \sum_{k=1}^{\hat{K}_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \max_{i=1,\dots,n} \sum_{k=1}^{\hat{K}_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \right. \\
& \quad \left. + 16 \cdot \frac{1}{n} \sum_{i=1}^n f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\
& \leq \left(4 \cdot \sum_{s=1}^{N_n \cdot (K^2+1)^d} |\bar{\alpha}_s|^2 \cdot c_{39} \cdot (\log n)^{4L+3} + 16 \cdot \frac{1}{n} \sum_{i=1}^n f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\
& \leq \left(c_{40} \cdot \frac{N_n \cdot (K^2+1)^d}{N_n^2} \cdot (\log n)^{4L+3} + 16 \cdot \left(c_{23} \cdot \left(3^d + \frac{(K^2+1)^d}{n} \right) \right)^2 + 8 \cdot c_5^3 \cdot (\log n)^3 \right)^{1/2} \\
& \leq c_{41} \cdot \left(\frac{(K^2+1)^{2d}}{N_n} \cdot (\log n)^{4L+3} + (\log n)^3 \right)^{1/2}.
\end{aligned}$$

Für den Term $T_{8,n}$ erhalten wir aufgrund der Definition von \mathbf{w}^* und mithilfe von Ungleichung (2.89), dass

$$\begin{aligned}
& T_{8,n} = \left(\frac{1}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 \right)^{1/2} \\
& \leq \left(\frac{1}{n} \sum_{i=1}^n \left(\sum_{k=1}^{\hat{K}_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{\hat{K}_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \right) \right)^{1/2} \\
& \leq \left(\sum_{k=1}^{\hat{K}_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{\hat{K}_n} \max_{i=1,\dots,n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \right)^{1/2} \\
& \leq \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\| \cdot c_{38} \cdot (\log n)^{2L} \cdot \|(w_{i,j,k}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,kj}^{(l)})_{k,i,j,l:l < L}\|
\end{aligned}$$

gilt. Durch die Kombination dieser beiden Ergebnisse erhalten wir

$$\left| F_n(\mathbf{w}^*) - F_n((\mathbf{w}^{(0)})^*) \right|$$

$$\leq c_{42} \cdot \left(\frac{(K^2 + 1)^d}{N_n^{1/2}} \cdot (\log n)^{4L+3/2} + (\log n)^{2L+3/2} \right) \cdot \|((\mathbf{w}^*)_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)}\| \cdot \|(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\|.$$

Da

$$c_{42} \cdot \left(\frac{(K^2 + 1)^d}{N_n^{1/2}} \cdot (\log n)^{4L+3/2} + (\log n)^{2L+3/2} \right) \leq c_{43} \cdot (\log n)^{4L+3/2}$$

für hinreichend großes n gilt, ist Voraussetzung (2.87) für

$$D_n = c_{43} \cdot (\log n)^{4L+3/2}$$

erfüllt.

Im fünften Schritt des Beweises zeigen wir, dass

$$\mathbf{P}(A_n^c) \leq \frac{c_{44}}{(\log n)^3}$$

gilt. Wir beginnen damit, die Wahrscheinlichkeit, dass der Gewichtsvektor $\mathbf{w}^{(0)}$ Bedingung (2.84) des Ereignisses A_n nicht erfüllt, zu beschränken. Hierfür betrachten wir eine schrittweise Auswahl der Gewichte der \widehat{K}_n vollständig verbundenen neuronalen Netze.

Aus den Bedingungen in Lemma 9 folgt für $2 \leq K \leq \log n - 1$ und $v_s^{(j)} \in [-K - \frac{2}{K}, K]$ mit $s \in \{1, \dots, N \cdot (K^2 + 1)^d\}$, dass

$$|w_{k,i,j}^{(0)}| \leq 4d \cdot K^2 \cdot (\log n)^2 \cdot v_s^{(j)}$$

für $i \in \{1, \dots, r\}, j \in \{1, \dots, d\}, k \in \{1, \dots, \widehat{K}_n\}$ gilt. Da $v_s^{(j)} \leq K$ ist, ergibt sich

$$\begin{aligned} 4d \cdot K^2 \cdot (\log n)^2 \cdot v_s^{(j)} &\leq 4d \cdot K^2 \cdot (\log n)^2 \cdot K \\ &= 4d \cdot (\log n)^2 \cdot K^3. \end{aligned}$$

Für $K \leq \log n - 1 \leq \log n$ folgt für hinreichend große n die Abschätzung

$$\begin{aligned} 4d \cdot (\log n)^2 \cdot K^3 &\leq 4d \cdot (\log n)^2 \cdot (\log n)^3 \\ &\leq 8d \cdot (\log n)^2 \cdot n^{1/(1+d)}. \end{aligned}$$

Durch die Wahl von $\tau = \frac{1}{1+d}$ ergibt sich hieraus, dass die innersten Gewichte in dem gewählten Startintervall liegen. Weiterhin gilt

$$\begin{aligned} |w_{k,i,j}^{(l)}| &\leq 8 \cdot (\log n)^2 \cdot \left(2d - \frac{1}{n}\right) \\ &\leq 16d \cdot (\log n)^2 \\ &\leq 20d \cdot (\log n)^2 \end{aligned}$$

für $l \in \{1, \dots, L\}, i, j \in \{1, \dots, r\}, k \in \{1, \dots, \widehat{K}_n\}$, womit auch diese Gewichte in dem Intervall der Startinitialisierung liegen.

Jedes dieser \widehat{K}_n neuronalen Netze enthält $(r+1) + (L-2) \cdot r \cdot (r+1) + r \cdot (d+1)$ Gewichte. Deshalb ist die Wahrscheinlichkeit, dass alle diese Gewichte Bedingung (2.84) für $s = 1$ erfüllen, von unten beschränkt durch

$$\left(\frac{\log n}{40d \cdot (\log n)^2} \right)^{(r+1)+(L-2) \cdot r \cdot (r+1)} \cdot \left(\frac{\log n}{16d \cdot (\log n)^2 \cdot n^\tau} \right)^{r \cdot (d+1)}.$$

Folglich ist die Wahrscheinlichkeit, dass Bedingung (2.84) von den ersten $n^{r(d+1)\tau+1}$ vollständig verbundenen neuronalen Netzen für k_1 nicht erfüllt ist, für n hinreichend groß von oben beschränkt durch

$$\begin{aligned} & \left(1 - \left(\frac{1}{40d \cdot \log n} \right)^{(r+1)+(L-2) \cdot r \cdot (r+1)} \cdot \left(\frac{1}{16d \cdot \log n \cdot n^\tau} \right)^{r \cdot (d+1)} \right)^{n^{r(d+1)\tau+1}} \\ & \leq \left(1 - n^{-r(d+1)\tau-0,5} \right)^{n^{r(d+1)\tau+1}}. \end{aligned}$$

Aus Voraussetzung (2.7) folgt, dass für n hinreichend groß die Ungleichung $\widehat{K}_n \geq N_n \cdot (K^2 + 1)^d \cdot n^{r(d+1)\tau+1}$ gilt. Dies impliziert, dass für hinreichend große n die Bedingung (2.84) mit hoher Wahrscheinlichkeit erfüllt ist, da das Gegenereignis mit einer Wahrscheinlichkeit von höchstens

$$\begin{aligned} & N_n \cdot (K^2 + 1)^d \cdot \left(1 - n^{-r(d+1)\tau-0,5} \right)^{n^{r(d+1)\tau+1}} \\ & \leq N_n \cdot (K^2 + 1)^d \cdot \left(\exp \left(-n^{-r(d+1)\tau-0,5} \right) \right)^{n^{r(d+1)\tau+1}} \\ & \leq N_n \cdot (K^2 + 1)^d \cdot \exp \left(-n^{0,5} \right) \\ & \leq c_{45} \cdot (\log n)^{18L} \cdot \exp \left(-n^{0,5} \right) \\ & \leq \frac{c_{45}}{n} \end{aligned}$$

auftritt.

Für hinreichend große n erhalten wir dann mithilfe der Ungleichung von Markov, dass

$$\begin{aligned} \mathbf{P}(A_n^c) & \leq \frac{c_{45}}{n} + \mathbf{P} \left\{ \frac{1}{n} \sum_{i=1}^n Y_i^2 > \beta_n^3 \right\} \\ & \leq \frac{c_{45}}{n} + \frac{\mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n Y_i^2 \right\}}{\beta_n^3} \\ & \leq \frac{c_{45}}{n} + \frac{\mathbf{E} \{ Y^2 \}}{\beta_n^3} \\ & \leq \frac{c_{44}}{(\log n)^3} \end{aligned}$$

gilt, wobei die letzte Ungleichung aus $\mathbf{E}\{Y^2\} < \infty$ folgt.

Im *sechsten Schritt des Beweises* wollen wir nachweisen, dass

$$\limsup_{n \rightarrow \infty} \mathbf{E}\{T_{2,n}\} \leq 0$$

gilt.

Sei \mathcal{F} der Funktionsraum, der wie in Lemma 5 mit $C = c_{46} \cdot \widehat{K}_n \cdot (\log n)^{3/2}$, $B = c_{47} \cdot (\log n)^2$ und $A = c_{48} \cdot (\log n)^2 \cdot n^\tau$ definiert ist.

Es gilt

$$\frac{1}{1 + \varepsilon} \cdot \mathbf{E}\{T_{2,n}\}$$

$$\begin{aligned}
&= \frac{1}{1+\varepsilon} \cdot \mathbf{E} \{ \max\{T_{2,n}, 0\} \} \\
&= \int_0^{4\beta_n^2} \mathbf{P} \left\{ \left(\mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \mathbf{1}_{A_n} > t \right\} dt \\
&\leq n^{\frac{-1}{4(d+2)}} \\
&\quad + \int_{n^{\frac{-1}{4(d+2)}}}^{4\beta_n^2} \mathbf{P} \left\{ \left(\mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \mathbf{1}_{A_n} > t \right\} dt.
\end{aligned}$$

Da der Term $T_{2,n}$ nur ungleich 0 ist, wenn das Ereignis A_n eintritt, kann ohne Beschränkung der Allgemeinheit angenommen werden, dass das Ereignis A_n eintritt. Mit Lemma 1, das wir, wie in Beweisschritt 4 gezeigt, anwenden können, erhalten wir

$$\|((\mathbf{w}^{(t_n)})_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\| \leq c_{49} \cdot (\log n)^{3/2}$$

und

$$\|((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n} - ((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n}\| = \|((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n}\| \leq c_{46} \cdot (\log n)^{3/2}.$$

Aufgrund der Initialisierung der Startgewichte folgen hieraus die Ungleichungen

$$\begin{aligned}
&\|((\mathbf{w}^{(t_n)})_{k,i,j}^{(l)})_{k,i,j,l:1 \leq l < L}\|_{\infty} \\
&\leq \|((\mathbf{w}^{(t_n)})_{k,i,j}^{(l)})_{k,i,j,l:1 \leq l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:1 \leq l < L}\|_{\infty} + \|((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:1 \leq l < L}\|_{\infty} \\
&\leq c_{49} \cdot (\log n)^{3/2} + 20d \cdot (\log n)^2 \\
&\leq c_{47} \cdot (\log n)^2
\end{aligned}$$

und

$$\begin{aligned}
&\|((\mathbf{w}^{(t_n)})_{k,i,j}^{(0)})_{k,i,j}\|_{\infty} \\
&\leq \|((\mathbf{w}^{(t_n)})_{k,i,j}^{(0)})_{k,i,j} - ((\mathbf{w}^{(0)})_{k,i,j}^{(0)})_{k,i,j}\|_{\infty} + \|((\mathbf{w}^{(0)})_{k,i,j}^{(0)})_{k,i,j}\|_{\infty} \\
&\leq c_{49} \cdot (\log n)^{3/2} + 8d \cdot (\log n)^2 \cdot n^{\tau} \\
&\leq c_{48} \cdot (\log n)^2 \cdot n^{\tau}.
\end{aligned}$$

Für die äußeren Gewichte gilt zudem

$$\|((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n}\|_{\infty} \leq \|((\mathbf{w}^{(t_n)})_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n}\| \leq c_{46} \cdot (\log n)^{3/2}.$$

Dies impliziert, dass der Schätzer m_n in dem Funktionsraum

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

enthalten ist. Aus Lemma 4 und Lemma 5 folgt dann, da $|m_n(\mathbf{x})| \leq \beta_n$ für $\mathbf{x} \in \mathbb{R}^d$, dass

$$\mathbf{P} \left\{ \left(\mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \cdot \mathbf{1}_{A_n} > t \right\}$$

$$\begin{aligned} &\leq 8 \cdot \mathbf{E} \left\{ \mathcal{N}_1 \left(\frac{t}{8}, \{T_{\beta_n} f : f \in \mathcal{F}\}, \mathbf{X}_1^n \right) \right\} \cdot \exp \left(-\frac{n \cdot t^2}{128 \cdot \beta_n^4} \right) \\ &\leq 8 \cdot \left(c_{14} \cdot \frac{\beta_n}{t/8} \right)^{c_{50} \cdot (\log n)^{c_{51}} \cdot n^{\tau \cdot d} \cdot \left(\frac{c_{46} \cdot \hat{\kappa}_n \cdot (\log n)^{3/2}}{t/8} \right)^{d/\ell} + c_{16}} \cdot \exp \left(-\frac{n \cdot t^2}{128 \cdot \beta_n^4} \right) \end{aligned}$$

gilt. Mit Voraussetzung (2.6) und $\tau = \frac{1}{d+1}$ erhalten wir für $t > n^{-\frac{1}{4(d+2)}}$ und hinreichend große n die Abschätzung

$$\begin{aligned} &8 \cdot \left(c_{14} \cdot \frac{\beta_n}{t/8} \right)^{c_{50} \cdot (\log n)^{c_{51}} \cdot n^{\tau \cdot d} \cdot \left(\frac{c_{46} \cdot \hat{\kappa}_n \cdot (\log n)^{3/2}}{t/8} \right)^{d/\ell} + c_{16}} \cdot \exp \left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \cdot \exp \left(-\frac{n \cdot t^2}{256 \cdot \beta_n^4} \right) \\ &\leq \exp \left(c_{52} \cdot (\log n)^{c_{51}} \cdot n^{\tau \cdot d + (\kappa+1) \cdot \frac{d}{\ell}} \cdot \log \left(c_{14} \cdot \frac{\beta_n}{t/8} \right) \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \\ &\leq \exp \left(c_{53} \cdot \left((\log n)^{c_{51}} \cdot n^{\frac{d}{d+1} + (\kappa+1) \cdot \frac{d}{\ell}} - \frac{n^{\frac{2d+3}{2(d+2)}}}{(\log n)^4} \right) \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right). \end{aligned}$$

Für $\ell > (\kappa + 1) \cdot d \cdot (d + 1) \cdot (d + 2)$ mit $\kappa > 0$ ergibt sich

$$\begin{aligned} \frac{d}{d+1} + (\kappa + 1) \cdot \frac{d}{\ell} &< \frac{d}{d+1} + (\kappa + 1) \cdot \frac{d}{(\kappa + 1) \cdot d \cdot (d + 1) \cdot (d + 2)} \\ &= \frac{d}{d+1} + \frac{1}{(d+1) \cdot (d+2)} \\ &= \frac{(d+1)^2}{(d+1) \cdot (d+2)} \\ &= \frac{d+1}{d+2}. \end{aligned}$$

Somit ist

$$\begin{aligned} &\exp \left(c_{53} \cdot \left((\log n)^{c_{51}} \cdot n^{\frac{d}{d+1} + (\kappa+1) \cdot \frac{d}{\ell}} - \frac{n^{\frac{2d+3}{2(d+2)}}}{(\log n)^4} \right) \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \\ &\leq \exp \left(c_{53} \cdot \left((\log n)^{c_{51}} \cdot n^{\frac{d+1}{d+2}} - \frac{n^{\frac{2d+3}{2(d+2)}}}{(\log n)^4} \right) \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \\ &= \exp \left(c_{53} \cdot n^{\frac{d+1}{d+2}} \cdot \left((\log n)^{c_{51}} - \frac{1}{n^{\frac{2d+3}{2(d+2)}}} \right) \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \cdot \exp \left(-\frac{2d+3}{2(d+2)} \right) \\ &\leq c_{53} \cdot \exp \left(-\frac{n^{\frac{2d+3}{2(d+2)}}}{256 \cdot \beta_n^4} \right) \end{aligned}$$

für n hinreichend groß. Zusammenfassend resultiert dies in

$$\begin{aligned} \mathbf{E} \{T_{2,n}\} &\leq n^{-\frac{1}{4(d+2)}} \\ &\quad + \int_{n^{-\frac{1}{4(d+2)}}}^{4 \cdot \beta_n^2} \mathbf{P} \left\{ \left(\mathbf{E} \left\{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 \mid \mathcal{D}_n \right\} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \mathbf{1}_{A_n} > t \right\} dt \end{aligned}$$

$$\leq (1 + \varepsilon) \cdot \left(n^{\frac{-1}{4(d+2)}} + 4 \cdot \beta_n^4 \cdot c_{53} \cdot \exp\left(-\frac{n^{\frac{2d+3}{2(d+2)}}}{256 \cdot \beta_n^4}\right) \right).$$

Für $n \rightarrow \infty$ erhalten wir damit die Aussage des sechsten Beweisschrittes.

Im *siebten Schritt des Beweises* zeigen wir

$$\limsup_{n \rightarrow \infty} \mathbf{E}\{T_{6,n}\} \leq 0.$$

Aufgrund der Integrierbarkeit der Regressionsfunktion m und der Aussage des fünften Beweisschrittes, dass

$$\mathbf{P}(A_n^c) \leq \frac{c_{44}}{(\log n)^3}$$

ist, ergibt sich

$$\begin{aligned} \mathbf{E}\{T_{6,n}\} &= \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbb{1}_{A_n^c} \\ &\leq \left(2 \cdot \int |m_n(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) + 2 \cdot \int |m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right) \cdot \mathbf{P}(A_n^c) \\ &\leq 2 \cdot (\beta_n^2 + c_{54}) \cdot \frac{c_{44}}{(\log n)^3} \\ &= c_{55} \cdot \frac{1}{\log n}, \end{aligned}$$

wobei die letzte Gleichheit aus $\beta_n = c_5 \cdot \log n$ folgt. Hieraus erhalten wir die Aussage des *siebten Beweisschrittes*.

Im *achten Schritt des Beweises* wollen wir

$$\mathbf{E}\{T_{5,n}\}$$

beschränken.

Sofern das Ereignis A_n eintritt, können wir, wie bereits im vierten Beweisschritt gezeigt, Lemma 1 mit $D_n = c_{43} \cdot (\log n)^{4L+3/2}$ anwenden. Dies führt zu

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}(t_n)}(\mathbf{X}_i) - Y_i|^2 &= F_n(\mathbf{w}(t_n)) \\ &\leq F_n((\mathbf{w}^{(0)})^*) + D_n \cdot \|((\mathbf{w}^*)_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)}\| \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} \\ &\quad + \frac{\|((\mathbf{w}^*)_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)} - ((\mathbf{w}^{(0)})_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)}\|^2}{2} + \frac{F_n(\mathbf{w}^{(0)})}{t_n}. \end{aligned} \quad (2.90)$$

Zusammen mit der Definition des Gewichtsvektors \mathbf{w}^* und der Tatsache, dass $|\bar{\alpha}_s| \leq \frac{c_{33}}{N_n}$ für $s \in \{1, \dots, N_n \cdot (K^2 + 1)^d\}$ ist, ergibt sich

$$\|((\mathbf{w}^*)_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)}\| \leq N_n^{1/2} \cdot (K^2 + 1)^{d/2} \cdot \frac{c_{33}}{N_n}$$

sowie

$$\frac{\|((\mathbf{w}^*)_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)} - ((\mathbf{w}^{(0)})_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)}\|^2}{2} = \frac{\|((\mathbf{w}^*)_{1,1,k})_{k=1,\dots,\widehat{K}_n}^{(L)}\|^2}{2} \leq \frac{c_{33}^2 \cdot N_n \cdot (K^2 + 1)^d}{2 \cdot N_n^2}.$$

Setzen wir dies in (2.90) ein, so ist

$$\begin{aligned}
& F_n((\mathbf{w}^{(0)})^*) + D_n \cdot \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n}\| \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} \\
& \quad + \frac{\|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n} - ((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\dots,\widehat{K}_n}\|^2}{2} + \frac{F_n(\mathbf{w}^{(0)})}{t_n} \\
& \leq F_n((\mathbf{w}^{(0)})^*) + c_{56} \cdot (\log n)^{4L+3/2} \cdot \frac{N_n^{1/2} \cdot (K^2 + 1)^{d/2}}{N_n} \cdot (\log n)^{3/2} \\
& \quad + \frac{c_{33}^2 \cdot N_n \cdot (K^2 + 1)^d}{2 \cdot N_n^2} + \frac{c_5^3 \cdot (\log n)^3}{t_n} \\
& = F_n((\mathbf{w}^{(0)})^*) + c_{56} \cdot (\log n)^{4L+3/2} \cdot \frac{(K^2 + 1)^{d/2}}{N_n^{1/2}} \cdot (\log n)^{3/2} \\
& \quad + \frac{c_{33}^2 \cdot (K^2 + 1)^d}{2 \cdot N_n} + \frac{c_5^3 \cdot (\log n)^3}{t_n} \\
& \leq F_n((\mathbf{w}^{(0)})^*) + c_{57} \cdot \frac{(K^2 + 1)^d}{N_n^{1/2}} \cdot (\log n)^{4L+3}.
\end{aligned}$$

Die letzte Ungleichung ergibt sich hierbei aus $t_n \geq c_6 \cdot \widehat{K}_n^{3/2} \cdot (\log n)^{6L+5}$.

Damit erhalten wir

$$\begin{aligned}
& \mathbf{E}\{T_{5,n}\} \\
& = (1 + \varepsilon)^2 \cdot \mathbf{E}\left\{\left(F_n(\mathbf{w}^{(t_n)}) - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}\right) \cdot \mathbf{1}_{A_n}\right\} \\
& \quad + ((1 + \varepsilon)^2 - 1) \cdot \mathbf{E}\left\{\mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \cdot \mathbf{1}_{A_n}\right\} \\
& \leq (1 + \varepsilon)^2 \cdot \left(\mathbf{E}\left\{F_n((\mathbf{w}^{(0)})^*) \cdot \mathbf{1}_{A_n} + c_{57} \cdot \left(\frac{(K^2 + 1)^d}{N_n^{1/2}} \cdot (\log n)^{4L+3}\right) \cdot \mathbf{1}_{A_n}\right\}\right. \\
& \quad \left. - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \cdot \mathbf{P}(A_n)\right) + ((1 + \varepsilon)^2 - 1) \cdot \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}. \tag{2.91}
\end{aligned}$$

Sei nun \widetilde{A}_n das Ereignis, bei dem der Gewichtsvektor $\mathbf{w}^{(0)}$ die Bedingung

$$\left|(\mathbf{w}^{(0)})_{k_s,i,j}^{(l)} - w_{s,i,j}^{(l)}\right| \leq \log n \quad \text{für alle } l \in \{0, \dots, L-1\}, s \in \{1, \dots, N_n \cdot (K^2 + 1)^d\}$$

erfüllt, wobei der Gewichtsvektor \mathbf{w} , den Voraussetzungen (2.67)–(2.76) in Lemma 9 für \bar{m} genügt. Dann erhalten wir, wie im fünften Beweisschritt, dass

$$\mathbf{P}(\widetilde{A}_n) - \mathbf{P}(A_n) \leq \mathbf{P}\left\{\frac{1}{n} \sum_{i=1}^n Y_i^2 > \beta_n^3\right\} \leq \frac{c_{58}}{(\log n)^3}$$

gilt.

Zusammen mit der Tatsache, dass die Zufallsvariablen $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängig vom Ereignis \widetilde{A}_n sind, führt dies zu

$$\mathbf{E}\left\{F_n((\mathbf{w}^{(0)})^*) \cdot \mathbf{1}_{A_n}\right\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \cdot \mathbf{P}(A_n)$$

$$\begin{aligned}
&\leq \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i|^2 \cdot \mathbf{1}_{\tilde{A}_n} \right\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \cdot \mathbf{P}(\tilde{A}_n) \\
&\quad + \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \cdot (\mathbf{P}(\tilde{A}_n) - \mathbf{P}(A_n)) \\
&\leq \mathbf{E} \left\{ \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i|^2 \cdot \mathbf{1}_{\tilde{A}_n} \middle| (\mathbf{w}^{(0)})^* \right\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \cdot \mathbf{1}_{\tilde{A}_n} \right\} \\
&\quad + (2 \cdot \mathbf{E}\{m(\mathbf{X})^2\} + 2 \cdot \mathbf{E}\{Y^2\}) \cdot \frac{c_{58}}{(\log n)^3} \\
&= \mathbf{E} \left\{ \left(\mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i|^2 \middle| (\mathbf{w}^{(0)})^* \right\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \right) \cdot \mathbf{1}_{\tilde{A}_n} \right\} + \frac{c_{59}}{(\log n)^3} \\
&= \mathbf{E} \left\{ \int |f_{(\mathbf{w}^{(0)})^*}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} \right\} + \frac{c_{59}}{(\log n)^3}. \tag{2.92}
\end{aligned}$$

Sei nun K hinreichend groß, so dass $\text{supp}(\mathbf{X}) \subseteq [-K, K]^d$ ist. Aufgrund der Wahl von \bar{m} und Lemma 9 erhalten wir

$$\begin{aligned}
&\mathbf{E} \left\{ \int |f_{(\mathbf{w}^{(0)})^*}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} \right\} \\
&\leq 2 \cdot \mathbf{E} \left\{ \int |f_{(\mathbf{w}^{(0)})^*}(\mathbf{x}) - \bar{m}(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} + 2 \int |\bar{m}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} \right\} \\
&\leq c_{60} \cdot \left(\frac{1}{K} + \frac{N_n^2 \cdot K^{4d}}{n^2} + \left(\frac{K^{2d}}{n} + 1 \right)^2 \cdot \mathbf{P}_{\mathbf{X}}(\mathbb{R}^d \setminus [-K, K]^d) \right) + 2\varepsilon \\
&\leq c_{61} \cdot \left(\frac{1}{K} + \frac{N_n^2 \cdot K^{4d}}{n^2} \right) + 2\varepsilon. \tag{2.93}
\end{aligned}$$

Des Weiteren folgt aus der Definition von N_n , dass

$$\frac{(K^2 + 1)^d}{N_n^{1/2}} \cdot (\log n)^{4L+3} \rightarrow 0 \quad (n \rightarrow \infty) \tag{2.94}$$

gilt. Setzen wir die Ergebnisse aus (2.92)–(2.94) nun in Ungleichung (2.91) ein, so führt dies zu

$$\limsup_{n \rightarrow \infty} \mathbf{E}\{T_{5,n}\} \leq c_{62} \cdot (1 + \varepsilon)^2 \cdot \left(\frac{1}{K} + 2\varepsilon \right) + ((1 + \varepsilon)^2 - 1) \cdot \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}.$$

Im neunten Schritt des Beweises schließen wir den Beweis ab, indem wir die Aussage von Theorem 1 zeigen. Durch Anwenden der Resultate aus den Beweisschritten 1,2,3,6,7 und 8 ergibt sich

$$\begin{aligned}
&\limsup_{n \rightarrow \infty} \mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\
&\leq c_{62} \cdot (1 + \varepsilon)^2 \cdot \left(\frac{1}{K} + 2\varepsilon \right) + ((1 + \varepsilon)^2 - 1) \cdot \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}.
\end{aligned}$$

Für $K \rightarrow \infty$ erhalten wir damit

$$\limsup_{n \rightarrow \infty} \mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_{63} \cdot ((1 + \varepsilon)^2 \cdot \varepsilon + ((1 + \varepsilon)^2 - 1) \cdot \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}),$$

woraus mit $\varepsilon \rightarrow 0$ die Aussage von Theorem 1 ergibt. \square

Das Theorem zeigt also, dass die Kombination aus Überparametrisierung und dem Gradientenabstieg bezüglich des empirischen L_2 -Risikos ohne Regularisierungsterm zu einem universell konsistenten Schätzer führt. Dieser Schätzer gewährleistet somit asymptotisch eine gute Generalisierung auf neue, unabhängige Daten.

Neben der universellen Konsistenz des Schätzers ist es jedoch auch von Interesse, mit welcher Geschwindigkeit er gegen die wahre Regressionsfunktion konvergiert und wie schnell der erwartete L_2 -Fehler gegen 0 geht. Daher werden wir uns im folgenden Kapitel mit der Konvergenzgeschwindigkeit von überparametrisierten tiefen Neuronale-Netze-Schätzern beschäftigen, die durch Gradientenabstieg trainiert werden.

3 Zur Konvergenzgeschwindigkeit von überparametrisierten tiefen Neuronale-Netze-Schätzern

Im vorherigen Kapitel haben wir die universelle Konsistenz überparametrisierter neuronaler Netze untersucht und diese Eigenschaft für einen überparametrisierten Neuronale-Netze-Schätzer nachgewiesen, der mittels Gradientenabstieg bezüglich des empirischen L_2 -Risikos ohne Regularisierungsterm trainiert wurde. Dabei haben wir gezeigt, dass sich dieser Schätzer asymptotisch der wahren Funktion annähert und somit für große Stichproben eine gute Generalisierungsfähigkeit auf neue, unabhängige Daten besitzt.

Ein weiterer wichtiger Aspekt bei der Analyse von Schätzern ist die Frage nach der Konvergenzrate, also der Geschwindigkeit, mit welcher der Schätzer gegen die wahre Regressionsfunktion konvergiert. Während die universelle Konsistenz eines Schätzers lediglich garantiert, dass er sich mit zunehmender Stichprobengröße asymptotisch dem wahren Wert annähert, liefert die Konvergenzrate eine Aussage darüber, wie schnell diese Annäherung erfolgt. Aussagen über die Konvergenzrate eines Schätzers sind insofern von Bedeutung, da sie es ermöglichen, die Leistungsfähigkeit verschiedener Schätzer zu vergleichen und zu beurteilen, wie gut ein Schätzer mit einer begrenzten Datenmenge arbeitet. Eine schnelle Konvergenzrate bedeutet, dass der Schätzer bereits bei kleineren Stichproben eine hohe Genauigkeit erreicht, während ein konsistenter Schätzer mit langsamer Konvergenzrate möglicherweise erst bei sehr großen Datenmengen eine akzeptable Genauigkeit erzielt.

Da, wie in Kapitel 1 bereits erwähnt, keine allgemeingültige Aussage über die Konvergenzrate eines Schätzers hergeleitet werden kann, müssen wir spezifische Annahmen über die zu schätzende Regressionsfunktion treffen. In diesem Kapitel nehmen wir daher an, dass die Regressionsfunktion (p, C) -glatt für $p \in [1/2, 1]$ ist. Diese Annahme ermöglicht es uns, eine Konvergenzrate für den überparametrisierten tiefen Neuronale-Netze-Schätzer aus dem vorherigen Kapitel herzuleiten.

In Abschnitt 3.1 werden wir eine Fehlerschranke für den erwarteten L_2 -Fehler nachweisen. Mit dieser Schranke und einem weiteren Resultat, welches es uns ermöglicht, den Approximationsfehler zu beschränken, können wir die Konvergenzgeschwindigkeit des Neuronale-Netze-Schätzers in Abschnitt 3.2 zeigen.

Unter der zusätzlichen Annahme, dass die Regressionsfunktion einem Interaktionsmodell mit $p \in [1/2, 1]$ genügt, können wir in Abschnitt 3.3 ergänzend zeigen, dass der überparametrisierte Neuronale-Netze-Schätzer eine Konvergenzgeschwindigkeit erzielen kann, die nicht von der Eingabedimension d abhängt.

3.1 Fehlerschranke des Schätzers

Das folgende Theorem liefert eine notwendige Fehlerschranke für den erwarteten L_2 -Fehler des Schätzers, die einen wichtigen Beitrag zur Herleitung der Konvergenzrate leisten wird.

Theorem 2. Sei $n \in \mathbb{N}$ mit $n \geq 2$ und $\beta_n = c_5 \cdot \log n$. Seien $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängig und identisch verteilte $\mathbb{R}^d \times \mathbb{R}$ -wertige Zufallsvariablen, wobei $\text{supp}(\mathbf{X})$ beschränkt ist sowie

$$\mathbf{E} \{ \exp(c_{64} \cdot Y^2) \} < \infty \quad (3.1)$$

für eine Konstante $c_{64} > 0$ mit $c_5 \cdot c_{64} \geq 2$ gilt. Die zugehörige Regressionsfunktion $m(\mathbf{x}) = \mathbf{E}\{Y|\mathbf{X} = \mathbf{x}\}$ sei beschränkt. Weiter sei $\sigma(z) = 1/(1 + \exp(-z))$ die logistische Sigmoidfunktion sowie $\widehat{K}_n, L, r, t_n \in \mathbb{N}$, $M_n \geq 1$ und $\lambda_n, \tau > 0$.

Zudem sei $\widetilde{K}_n \in \{1, \dots, \widehat{K}_n\}$ und es gelte

$$\begin{aligned} w_{k,i,j}^{(l)} &\in [-20d \cdot (\log n)^2, 20d \cdot (\log n)^2] \quad \text{für } l \in \{1, \dots, L\}, k \in \{1, \dots, \widetilde{K}_n\}, i, j \in \{1, \dots, r\}, \\ w_{k,i,j}^{(0)} &\in [-8d \cdot (\log n)^2 \cdot n^\tau, 8d \cdot (\log n)^2 \cdot n^\tau] \quad \text{für } k \in \{1, \dots, \widetilde{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}, \end{aligned}$$

$$\sqrt{\sum_{k=1}^{\widetilde{K}_n} |w_{1,1,k}^{(L)}|^2} \leq M_n \quad (3.2)$$

und

$$\left| \sum_{k=1}^{\widetilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(\mathbf{x}) \right| \leq \beta_n \quad (3.3)$$

für $\mathbf{x} \in \text{supp}(\mathbf{X})$ sowie für alle Gewichtsvektoren $\bar{\mathbf{w}}$ mit

$$\left| \bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)} \right| \leq \log n \quad \text{für } l = 0, \dots, L-1.$$

Des Weiteren sei

$$\frac{\widehat{K}_n}{n^\kappa} \rightarrow 0 \quad (n \rightarrow \infty) \quad (3.4)$$

für ein $\kappa > 0$ und

$$\frac{\widehat{K}_n}{\widetilde{K}_n \cdot n^{r \cdot (d+1) \cdot \tau + 1}} \rightarrow \infty \quad (n \rightarrow \infty). \quad (3.5)$$

Der Schätzer m_n sei definiert wie in Abschnitt 2.1, wobei

$$\lambda_n = \frac{1}{t_n} \quad \text{und} \quad t_n = \lceil c_{65} \cdot C_n \rceil$$

für eine Konstante $c_{65} \geq 2$ und für $C_n > 0$ mit

$$C_n \geq \widehat{K}_n^{3/2} \cdot (\log n)^{6L+2}.$$

Dann gilt für alle $\varepsilon > 0$ die Abschätzung

$$\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_{66} \cdot \left(\frac{n^{\tau \cdot d + \varepsilon}}{n} + M_n^2 \cdot (\log n)^{4L+3/2} \right) + \sup_{\substack{(\bar{w}_{k,i,j}^{(l)})_{k,i,j,l}: \\ |\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n \ (l=0, \dots, L-1)}} \int \left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})$$

für eine Konstante $c_{66} = c_{66}(\varepsilon) > 0$, die von ε abhängt.

Für den Beweis der Fehlerschranke in Theorem 2 benötigen wir das folgende Lemma, das eine notwendige Abschätzung zur Kontrolle der Komplexität der Funktionsklasse \mathcal{F} liefert.

Lemma 10. Sei $B \geq 1$ und sei \mathcal{F} eine Menge von Funktionen mit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ und $|f(\mathbf{x})| \leq B$ für alle $\mathbf{x} \in \mathbb{R}^d$. Angenommen, es ist $|Y| \leq B$ fast sicher, dann gilt für alle $n \geq 1$ die Ungleichung

$$\mathbf{P} \left\{ \exists f \in \mathcal{F} : \mathbf{E} \{|f(\mathbf{X}) - Y|^2\} - \mathbf{E} \{|m(\mathbf{X}) - Y|^2\} - \frac{1}{n} \sum_{i=1}^n (|f(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2) \geq \varepsilon \cdot (\alpha + \beta + \mathbf{E} \{|f(\mathbf{X}) - Y|^2\} - \mathbf{E} \{|m(\mathbf{X}) - Y|^2\}) \right\} \leq 14 \cdot \sup_{\mathbf{x}_1^n} \mathcal{N}_1 \left(\frac{\beta \varepsilon}{20B}, \mathcal{F}, \mathbf{x}_1^n \right) \exp \left(- \frac{\varepsilon^2 (1 - \varepsilon) \alpha n}{214(1 + \varepsilon) B^4} \right),$$

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

Beweis. Siehe Theorem 11.4 in Györfi et al. (2002). □

Bemerkung 2. Gemäß dem Beweis von Theorem 11.4 in Györfi et al. (2002) ist die gleiche Abschätzung auch für die Wahrscheinlichkeit

$$\mathbf{P} \left\{ \exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^n (|f(\mathbf{X}_i) - Y_i|^2 - |m(\mathbf{X}_i) - Y_i|^2) - (\mathbf{E} \{|f(\mathbf{X}) - Y|^2\} - \mathbf{E} \{|m(\mathbf{X}) - Y|^2\}) \geq \varepsilon \cdot (\alpha + \beta + \mathbf{E} \{|f(\mathbf{X}) - Y|^2\} - \mathbf{E} \{|m(\mathbf{X}) - Y|^2\}) \right\}$$

erfüllt.

Im Beweis von Theorem 2 leiten wir unter den gegebenen Voraussetzungen eine Fehlerschranke für den erwarteten L_2 -Fehler her. Der Beweis erfolgt in mehreren Schritten. Zunächst definieren wir ein Ereignis A_n , auf dem sichergestellt ist, dass die inneren Startgewichte höchstens um einen logarithmischen Faktor von den inneren Gewichten eines gegebenen Gewichtsvektors abweichen. Anschließend zerlegen wir den L_2 -Fehler in mehrere Komponenten, die wir jeweils separat abschätzen. Zur Herleitung der finalen Fehlerschranke greifen wir auf Resultate aus dem vorherigen Kapitel zurück.

Beweis von Theorem 2. Sei A_n das Ereignis, bei dem der Gewichtsvektor $\mathbf{w}^{(0)}$ die Bedingung

$$\left| (\mathbf{w}^{(0)})_{k_s, i, j}^{(l)} - w_{s, i, j}^{(l)} \right| \leq \log n \quad \text{für alle } l \in \{0, \dots, L-1\}, s \in \{1, \dots, \tilde{K}_n\}$$

für paarweise verschiedene Indizes $k_1, \dots, k_{\tilde{K}_n} \in \{1, \dots, \tilde{K}_n\}$ erfüllt und zusätzlich

$$\max_{i=1, \dots, n} |Y_i| \leq \sqrt{\beta_n}$$

gilt.

Wir definieren den Gewichtsvektor \mathbf{w}^* durch

$$\begin{aligned} (\mathbf{w}^*)_{k, i, j}^{(l)} &= w_{k, i, j}^{(l)} \quad \text{für alle } l \in \{0, \dots, L-1\}, \\ (\mathbf{w}^*)_{1, 1, k_s}^{(L)} &= w_{1, 1, s}^{(L)} \quad \text{für alle } s \in \{1, \dots, \tilde{K}_n\}, \\ (\mathbf{w}^*)_{1, 1, s}^{(L)} &= 0 \quad \text{für alle } s \notin \{k_1, \dots, k_{\tilde{K}_n}\}. \end{aligned}$$

Insbesondere ist der Gewichtsvektor $(\mathbf{w}^{(0)})^*$ dann gegeben durch

$$\begin{aligned} ((\mathbf{w}^{(0)})^*)_{k, i, j}^{(l)} &= (\mathbf{w}^{(0)})_{k, i, j}^{(l)} \quad \text{für alle } l = 0, \dots, L-1, \\ ((\mathbf{w}^{(0)})^*)_{1, 1, k_s}^{(L)} &= w_{1, 1, s}^{(L)} \quad \text{für alle } s \in \{1, \dots, \tilde{K}_n\}, \\ ((\mathbf{w}^{(0)})^*)_{1, 1, s}^{(L)} &= 0 \quad \text{für alle } s \notin \{k_1, \dots, k_{\tilde{K}_n}\}. \end{aligned}$$

Des Weiteren setzen wir

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

Da wir uns im Rahmen der Konvergenzgeschwindigkeit nur für große Stichprobenumfänge n interessieren, nehmen wir im Folgenden ohne Beschränkung der Allgemeinheit an, dass n hinreichend groß ist. Zusätzlich können wir aufgrund der Beschränktheit von m voraussetzen, dass $\|m\|_\infty \leq \beta_n$ gilt.

Im *ersten Schritt des Beweises* zerlegen wir den L_2 -Fehler des Schätzers m_n in eine Summe verschiedener Terme. Damit erhalten wir

$$\begin{aligned} & \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &= (\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\}) \cdot \mathbb{1}_{A_n} \\ & \quad + \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbb{1}_{A_n^c} \\ &= \left[\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \right. \\ & \quad \left. - (\mathbf{E}\{|m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m_{\beta_n}(\mathbf{X}) - T_{\beta_n} Y|^2\}) \right] \cdot \mathbb{1}_{A_n} \\ & \quad + \left[\mathbf{E}\{|m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m_{\beta_n}(\mathbf{X}) - T_{\beta_n} Y|^2\} \right] \end{aligned}$$

$$\begin{aligned}
& -2 \cdot \frac{1}{n} \sum_{i=1}^n (|m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2) \Big] \cdot \mathbb{1}_{A_n} \\
& + \left[2 \cdot \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m_{\beta_n}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right. \\
& \quad \left. - \left(2 \cdot \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right) \right] \cdot \mathbb{1}_{A_n} \\
& + \left[2 \cdot \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 - 2 \cdot \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbb{1}_{A_n} \\
& + \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbb{1}_{A_n^c} \\
& =: \sum_{j=1}^5 T_{j,n}.
\end{aligned}$$

Im zweiten Schritt des Beweises zeigen wir, dass

$$\mathbf{E}\{T_{1,n}\} \leq c_{67} \cdot \frac{\log n}{n} \quad \text{sowie} \quad \mathbf{E}\{T_{3,n}\} \leq c_{68} \cdot \frac{\log n}{n} \quad (3.6)$$

gilt.

Der Beweis hierzu orientiert sich an dem Beweis von Lemma 1 in Bauer und Kohler (2019). Der Vollständigkeit halber geben wir diesen im Folgenden an. Für den Term $T_{1,n}$ liefert die dritte binomische Formel

$$\begin{aligned}
T_{1,n} &= \left[\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m(\mathbf{X}) - Y|^2\} \right. \\
& \quad \left. - (\mathbf{E}\{|m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m_{\beta_n}(\mathbf{X}) - T_{\beta_n} Y|^2\}) \right] \cdot \mathbb{1}_{A_n} \\
&= \left[\mathbf{E}\{|m_n(\mathbf{X}) - Y|^2 - |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} \right. \\
& \quad \left. - \mathbf{E}\{|m(\mathbf{X}) - Y|^2 - |m_{\beta_n}(\mathbf{X}) - T_{\beta_n} Y|^2\} \right] \cdot \mathbb{1}_{A_n} \\
&= \mathbf{E}\{(T_{\beta_n} Y - Y) \cdot (2m_n(\mathbf{X}) - Y - T_{\beta_n} Y) | \mathcal{D}_n\} \cdot \mathbb{1}_{A_n} \\
& \quad - \mathbf{E}\{(m(\mathbf{X}) - Y - m_{\beta_n}(\mathbf{X}) + T_{\beta_n} Y) \cdot (m(\mathbf{X}) - Y + m_{\beta_n}(\mathbf{X}) - T_{\beta_n} Y)\} \cdot \mathbb{1}_{A_n} \\
&= T_{6,n} + T_{7,n}.
\end{aligned}$$

Wir starten damit zu zeigen, dass

$$\mathbf{E}\{T_{6,n}\} \leq c_{69} \cdot \frac{\log n}{n}$$

ist.

Da die Ungleichung $|Y| > \beta_n$ äquivalent zu $\exp((c_{64}/2) \cdot |Y|^2) > \exp((c_{64}/2) \cdot \beta_n^2)$ ist, gilt

$$\mathbb{1}_{\{|Y| > \beta_n\}} = \mathbb{1}_{\{\exp((c_{64}/2) \cdot |Y|^2) > \exp((c_{64}/2) \cdot \beta_n^2)\}}.$$

Damit folgt aus

$$\exp\left(\left(\frac{c_{64}}{2}\right) \cdot \beta_n^2\right) \cdot \mathbb{1}_{\{\exp\left(\left(\frac{c_{64}}{2}\right) \cdot |Y|^2\right) > \exp\left(\left(\frac{c_{64}}{2}\right) \cdot \beta_n^2\right)\}} \leq \exp\left(\left(\frac{c_{64}}{2}\right) \cdot |Y|^2\right)$$

die Gültigkeit von

$$\mathbb{1}_{\{|Y| > \beta_n\}} \leq \frac{\exp\left(\left(\frac{c_{64}}{2}\right) \cdot |Y|^2\right)}{\exp\left(\left(\frac{c_{64}}{2}\right) \cdot \beta_n^2\right)}.$$

Zusammen mit der Ungleichung von Cauchy-Schwarz erhalten wir daher

$$\begin{aligned} & \mathbf{E}\{(T_{\beta_n} Y - Y) \cdot (2m_n(\mathbf{X}) - Y - T_{\beta_n} Y) | \mathcal{D}_n\} \\ & \leq \sqrt{\mathbf{E}\{|T_{\beta_n} Y - Y|^2\}} \cdot \sqrt{\mathbf{E}\{|2m_n(\mathbf{X}) - Y - T_{\beta_n} Y|^2 | \mathcal{D}_n\}} \\ & \leq \sqrt{\mathbf{E}\{|Y|^2 \cdot \mathbb{1}_{\{|Y| > \beta_n\}}\}} \cdot \sqrt{\mathbf{E}\{2 \cdot |2m_n(\mathbf{X}) - T_{\beta_n} Y|^2 + 2 \cdot |Y|^2 | \mathcal{D}_n\}} \\ & \leq \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \frac{\exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right)}{\exp\left(\frac{c_{64}}{2} \cdot \beta_n^2\right)}\right\}} \cdot \sqrt{\mathbf{E}\{2 \cdot |2m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} + 2 \cdot \mathbf{E}\{|Y|^2\}} \\ & \leq \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right)\right\}} \cdot \exp\left(-\frac{c_{64}}{4} \cdot \beta_n^2\right) \cdot \sqrt{2 \cdot (3 \cdot \beta_n)^2 + 2 \cdot \mathbf{E}\{|Y|^2\}}, \end{aligned}$$

wobei wir in der letzten Ungleichung ausgenutzt haben, dass m_n und $T_{\beta_n} Y$ durch β_n beschränkt sind.

Wegen $y \leq \exp(y)$ für $y \in \mathbb{R}$ gilt

$$|Y|^2 \leq \frac{2}{c_{64}} \cdot \exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right).$$

Damit ist

$$\begin{aligned} \mathbf{E}\left\{|Y|^2 \cdot \exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right)\right\} & \leq \mathbf{E}\left\{\frac{2}{c_{64}} \cdot \exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right) \cdot \exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right)\right\} \\ & \leq \frac{2}{c_{64}} \cdot \mathbf{E}\{\exp(c_{64} \cdot |Y|^2)\}, \end{aligned}$$

woraus mit Voraussetzung (3.1) die Beschränktheit dieses Terms folgt.

Für den dritten Term erhalten wir aus

$$\mathbf{E}\{|Y|^2\} \leq \mathbf{E}\left\{\frac{1}{c_{64}} \cdot \exp(c_{64} \cdot |Y|^2)\right\} = c_{70} < \infty,$$

dass

$$\sqrt{2 \cdot (3 \cdot \beta_n)^2 + 2 \cdot \mathbf{E}\{|Y|^2\}} \leq \sqrt{18 \cdot \beta_n^2 + 2 \cdot c_{70}}$$

gilt.

Daher ergibt sich, mit $\beta_n = c_5 \cdot \log n$, dass

$$\mathbf{E}\{T_{6,n}\} \leq \sqrt{c_{71}} \cdot \exp\left(-\frac{c_{64}}{4} \cdot \beta_n^2\right) \cdot \sqrt{18 \cdot \beta_n^2 + 2 \cdot c_{70}} \cdot \mathbf{P}(A_n) \leq c_{69} \cdot \frac{\log n}{n}$$

erfüllt ist.

Als nächstes zeigen wir

$$\mathbf{E}\{T_{7,n}\} \leq c_{72} \cdot \frac{\log n}{n}.$$

Die Anwendung der Ungleichung von Cauchy-Schwarz führt zu

$$\begin{aligned} \mathbf{E}\{T_{7,n}\} &= -\mathbf{E}\{(m(\mathbf{X}) - Y - m_{\beta_n}(\mathbf{X}) + T_{\beta_n}Y) \cdot (m(\mathbf{X}) - Y + m_{\beta_n}(\mathbf{X}) - T_{\beta_n}Y)\} \cdot \mathbf{P}(A_n) \\ &\leq \sqrt{\mathbf{E}\{(m(\mathbf{X}) - m_{\beta_n}(\mathbf{X}) + T_{\beta_n}Y - Y)^2\}} \cdot \sqrt{\mathbf{E}\{(m(\mathbf{X}) - Y + m_{\beta_n}(\mathbf{X}) - T_{\beta_n}Y)^2\}} \\ &\leq \sqrt{2 \cdot \mathbf{E}\{|m(\mathbf{X}) - m_{\beta_n}(\mathbf{X})|^2\} + 2 \cdot \mathbf{E}\{|T_{\beta_n}Y - Y|^2\}} \\ &\quad \cdot \sqrt{\mathbf{E}\{|m(\mathbf{X}) - Y + m_{\beta_n}(\mathbf{X}) - T_{\beta_n}Y|^2\}}. \end{aligned} \quad (3.7)$$

Den zweiten Term können wir nun analog zu dem zweiten Faktor in $T_{6,n}$ abschätzen. Nach Voraussetzung des Theorems ist $\|m\|_\infty \leq \beta_n$. Zudem ist m_{β_n} per Definition ebenfalls durch β_n beschränkt. Damit gilt

$$\begin{aligned} \sqrt{\mathbf{E}\{|m(\mathbf{X}) - Y + m_{\beta_n}(\mathbf{X}) - T_{\beta_n}Y|^2\}} &\leq \sqrt{\mathbf{E}\{2 \cdot |m(\mathbf{X}) + m_{\beta_n}(\mathbf{X}) - T_{\beta_n}Y|^2\} + 2 \cdot \mathbf{E}\{|Y|^2\}} \\ &\leq \sqrt{\mathbf{E}\{4 \cdot |m(\mathbf{X})|^2 + 4 \cdot |m_{\beta_n}(\mathbf{X}) - T_{\beta_n}Y|^2\} + 2 \cdot \mathbf{E}\{|Y|^2\}} \\ &\leq \sqrt{\mathbf{E}\{20 \cdot \beta_n^2\} + 2 \cdot \mathbf{E}\{|Y|^2\}} \\ &\leq \sqrt{20 \cdot c_5^2 \cdot (\log n)^2 + 2 \cdot c_{70}} \\ &\leq c_{73} \cdot \log n. \end{aligned}$$

Für den ersten Term von $\mathbf{E}\{T_{7,n}\}$ ergibt sich durch die Ungleichung von Jensen, zusammen mit der Definition von m_n sowie der Definition von m_{β_n} die Abschätzung

$$\begin{aligned} &2 \cdot \mathbf{E}\{|m(\mathbf{X}) - m_{\beta_n}(\mathbf{X})|^2\} + 2 \cdot \mathbf{E}\{|T_{\beta_n}Y - Y|^2\} \\ &= 2 \cdot \mathbf{E}\left\{|\mathbf{E}\{Y - T_{\beta_n}Y | \mathbf{X}\}|^2\right\} + 2 \cdot \mathbf{E}\{|T_{\beta_n}Y - Y|^2\} \\ &\leq 2 \cdot \mathbf{E}\left\{|\mathbf{E}\{|Y - T_{\beta_n}Y|^2 | \mathbf{X}\}|\right\} + 2 \cdot \mathbf{E}\{|T_{\beta_n}Y - Y|^2\} \\ &= 4 \cdot \mathbf{E}\{|Y - T_{\beta_n}Y|^2\}. \end{aligned}$$

Einsetzen in Ungleichung (3.7) liefert

$$\mathbf{E}\{T_{7,n}\} \leq \sqrt{4 \cdot \mathbf{E}\{|Y - T_{\beta_n}Y|^2\}} \cdot c_{73} \cdot \log n.$$

Mit der gleichen Argumentation wie für den ersten Faktor in $T_{6,n}$ erhalten wir damit

$$\begin{aligned} \sqrt{4 \cdot \mathbf{E}\{|Y - T_{\beta_n}Y|^2\}} &\leq 2 \cdot \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \mathbf{1}_{\{|Y| > \beta_n\}}\right\}} \\ &\leq 2 \cdot \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \frac{\exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right)}{\exp\left(\frac{c_{64}}{2} \cdot \beta_n^2\right)}\right\}} \\ &\leq 2 \cdot \sqrt{\mathbf{E}\left\{|Y|^2 \cdot \exp\left(\frac{c_{64}}{2} \cdot |Y|^2\right)\right\}} \cdot \exp\left(-\frac{c_{64}}{4} \cdot \beta_n^2\right) \end{aligned}$$

$$\leq 2 \cdot \sqrt{c_{71}} \cdot \exp\left(-\frac{c_{64}}{4} \cdot \beta_n^2\right).$$

Damit ergibt sich

$$\mathbf{E}\{T_{7,n}\} \leq 2 \cdot \sqrt{c_{71}} \cdot \exp\left(-\frac{c_{64}}{4} \cdot \beta_n^2\right) \cdot c_{73} \cdot \log n \leq c_{72} \cdot \frac{\log n}{n},$$

woraus die Aussage des ersten Beweisschrittes folgt.

Der Beweis für die entsprechende Schranke von $\mathbf{E}\{T_{3,n}\}$ ergibt sich analog.

Im dritten Schritt des Beweises zeigen wir, dass

$$\mathbf{E}T_{5,n} \leq c_{74} \cdot \frac{(\log n)^2}{n}$$

ist.

Aus der Definition des Schätzers m_n und der Annahme, dass $\|m\|_\infty \leq \beta_n$ ist, folgt

$$\int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq 4 \cdot c_5^2 \cdot (\log n)^2.$$

Deshalb genügt es zu zeigen, dass

$$\mathbf{P}(A_n^c) \leq \frac{c_{75}}{n} \quad (3.8)$$

ist.

Um dies zu nachzuweisen, beschränken wir zunächst die Wahrscheinlichkeit, dass die Startgewichte im ersten der \widehat{K}_n vollständig verbundenen neuronalen Netze in jeder Komponente höchstens um einen Faktor $\log n$ von $(w_{1,i,j}^{(l)})_{i,j,l:l < L}$ abweichen. Da jedes der vollständig verbundenen neuronalen Netze $(r+1) + (L-2) \cdot r \cdot (r+1) + r \cdot (d+1)$ Gewichte enthält, ist diese Wahrscheinlichkeit für n hinreichend groß von unten beschränkt durch

$$\left(\frac{\log n}{40d \cdot (\log n)^2}\right)^{(r+1)+(L-2) \cdot r \cdot (r+1)} \cdot \left(\frac{\log n}{16d \cdot (\log n)^2 \cdot n^\tau}\right)^{r \cdot (d+1)} \geq n^{-r \cdot (d+1) \cdot \tau - 0,5}.$$

Damit ist die Wahrscheinlichkeit, dass keines der ersten $n^{r \cdot (d+1) \cdot \tau + 1}$ neuronalen Netze die Bedingung erfüllt, gegeben durch

$$\begin{aligned} (1 - n^{-r \cdot (d+1) \cdot \tau - 0,5})^{n^{r \cdot (d+1) \cdot \tau + 1}} &\leq \left(\exp\left(-n^{-r \cdot (d+1) \cdot \tau - 0,5}\right)\right)^{n^{r \cdot (d+1) \cdot \tau + 1}} \\ &= \exp(-n^{0,5}). \end{aligned}$$

Aus Voraussetzung (3.5) folgt, dass $\widehat{K}_n \geq \widetilde{K}_n \cdot n^{r \cdot (d+1) \cdot \tau + 1}$ für hinreichend große n gilt. Deshalb können wir diese Konstruktion schrittweise auf alle \widetilde{K}_n Gewichte $((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{i,j,l:l < L}$ anwenden. Die Wahrscheinlichkeit, dass ein $k \in \{1, \dots, \widetilde{K}_n\}$ existiert, so dass keiner der \widehat{K}_n Gewichtsvektoren des vollständig verbundenen neuronalen Netzes sich höchstens um $\log n$ von $(w_{k,i,j}^{(l)})_{i,j,l:l < L}$ unterscheidet, ist dann für hinreichend große n von oben beschränkt durch

$$\widetilde{K}_n \cdot \exp(-n^{0,5}) \leq n^\kappa \cdot \exp(-n^{0,5}) \leq \frac{c_{76}}{n}.$$

Mithilfe der Ungleichung von Markov erhalten wir

$$\begin{aligned} \mathbf{P}(A_n^c) &\leq \frac{c_{76}}{n} + \mathbf{P} \left\{ \max_{i=1,\dots,n} |Y_i| > \sqrt{\beta_n} \right\} \leq \frac{c_{76}}{n} + n \cdot \mathbf{P} \left\{ |Y| > \sqrt{\beta_n} \right\} \\ &\leq \frac{c_{76}}{n} + n \cdot \frac{\mathbf{E} \left\{ \exp(c_{64} \cdot Y^2) \right\}}{\exp(c_{64} \cdot \beta_n)} = \frac{c_{76}}{n} + n \cdot \frac{\mathbf{E} \left\{ \exp(c_{64} \cdot Y^2) \right\}}{\exp(c_{64} \cdot c_5 \cdot \log n)} \leq \frac{c_{74}}{n}, \end{aligned}$$

wobei die letzte Ungleichung aus Voraussetzung (3.1) und $c_5 \cdot c_{64} \geq 2$ ist, folgt.

Im vierten Schritt des Beweises wollen wir zeigen, dass die Voraussetzungen (2.9)–(2.11) von Lemma 1 erfüllt sind. Seien hierfür

$$S := \left\{ \mathbf{v} : \|\mathbf{v} - \mathbf{w}^{(0)}\| \leq 2 \cdot \sqrt{F(\mathbf{w}^{(0)})} + 1 \right\}$$

und

$$\tilde{S} := \left\{ (\bar{w}_{k,i,j}^{(l)})_{k,i,j,l:l < L} : \|(\bar{w}_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})^{(l)})_{k,i,j,l:l < L}\| \leq \sqrt{2 \cdot F(\mathbf{w}^{(0)})} \right\}.$$

Wir müssen nun nachweisen, dass bei Eintritt des Ereignisses A_n die Bedingungen

$$\|(\nabla_{\mathbf{w}} F)(\mathbf{w})\| \leq C_n \tag{3.9}$$

für alle $\mathbf{w} \in S$,

$$\|(\nabla_{\mathbf{w}} F)(\mathbf{w}) - (\nabla_{\mathbf{w}} F)(\bar{\mathbf{w}})\| \leq C_n \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \tag{3.10}$$

für alle $\mathbf{w}, \bar{\mathbf{w}} \in S$, sowie

$$\left| F(\mathbf{w}^*) - F((\mathbf{w}^{(0)})^*) \right| \leq D_n \cdot \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\| \cdot \|(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\|$$

für alle $(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} \in \tilde{S}$ gelten.

Sofern das Ereignis A_n eintritt, ist

$$F_n(\mathbf{w}^{(0)}) = \frac{1}{n} \sum_{i=1}^n Y_i^2 = \max_{i=1,\dots,n} Y_i^2 \leq \beta_n.$$

Sei nun $\mathbf{w} \in S$. Analog zu Beweisschritt vier in Theorem 1 gelten dann aufgrund der Initialisierung der Startgewichte die Ungleichungen

$$\begin{aligned} \|(w_{k,i,j}^{(l)})_{k,i,j,l:1 \leq l < L}\|_\infty &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:1 \leq l < L}\|_\infty \\ &\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1 + 20d \cdot (\log n)^2 \\ &\leq c_{77} \cdot (\log n)^2 \end{aligned}$$

und

$$\begin{aligned} \|(\mathbf{w}_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\|_\infty &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\| + \|((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\|_\infty \\ &\leq 2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1 \\ &\leq c_{78} \cdot (\log n)^{1/2}. \end{aligned}$$

Für $\bar{\mathbf{w}} \in S$ folgt dies analog. Somit sind die Voraussetzungen (2.19) und (2.20) von Lemma 2 sowie die Voraussetzungen (2.28) und (2.29) von Lemma 3 für $B_n = c_{77} \cdot (\log n)^2$ und $\gamma_n^* = c_{78} \cdot (\log n)^{1/2}$ erfüllt. Zusätzlich gilt

$$\begin{aligned} \|\mathbf{w} - \mathbf{w}^{(0)}\|_\infty^2 &\leq \|\mathbf{w} - \mathbf{w}^{(0)}\|^2 \\ &\leq \left(2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})} + 1\right)^2 \\ &\leq 2 \cdot \left(2 \cdot \sqrt{F_n(\mathbf{w}^{(0)})}\right)^2 + 2 \\ &\leq 8 \cdot F_n(\mathbf{w}^{(0)}) + 2. \end{aligned}$$

Wegen $t_n = \lceil c_{65} \cdot C_n \rceil$ mit $c_{65} \geq 2$ folgt hieraus, dass die Voraussetzungen (2.21) und (2.30) ebenfalls erfüllt sind.

Zudem können wir aufgrund des beschränkten Trägers von \mathbf{X} annehmen, dass $\alpha_n = c_{79}$ ist. Unter der Voraussetzung, dass $C_n \geq \hat{K}_n^{3/2} \cdot (\log n)^{6L+2}$ gilt, erhalten wir aus Lemma 2 die Ungleichung

$$\begin{aligned} \|\nabla_{\mathbf{w}} F_n(\mathbf{w})\| &\leq c_7 \cdot \hat{K}_n^{3/2} \cdot B_n^{2L} \cdot (\gamma_n^*)^2 \cdot \alpha_n \cdot \sqrt{\frac{t_n}{C_n} \cdot \max\{F_n(\mathbf{w}^{(0)}), 1\}} \\ &\leq c_{80} \cdot \hat{K}_n^{3/2} \cdot (\log n)^{4L} \cdot \log n \cdot \sqrt{\log n} \\ &\leq c_{80} \cdot \hat{K}_n^{3/2} \cdot (\log n)^{4L} \cdot (\log n)^{1+1/2} \\ &\leq C_n \end{aligned}$$

und aus Lemma 3 die Ungleichung

$$\begin{aligned} \|(\nabla_{\mathbf{w}} F_n)(\mathbf{w}) - (\nabla_{\mathbf{w}} F_n)(\bar{\mathbf{w}})\| &\leq c_{11} \cdot \hat{K}_n^{3/2} \cdot B_n^{3L} \cdot (\gamma_n^*)^2 \cdot \alpha_n^3 \cdot \sqrt{\frac{t_n}{C_n} \cdot \max\{F_n(\mathbf{v}), 1\}} \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \\ &\leq c_{81} \cdot \hat{K}_n^{3/2} \cdot (\log n)^{6L} \cdot \log n \cdot \sqrt{\log n} \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \\ &\leq c_{81} \cdot \hat{K}_n^{3/2} \cdot (\log n)^{6L} \cdot (\log n)^{1+1/2} \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \\ &\leq C_n \cdot \|\mathbf{w} - \bar{\mathbf{w}}\| \end{aligned}$$

für hinreichend großes n . Damit ist die Gültigkeit der Bedingungen (3.9) und (3.10) für $C_n \geq \hat{K}_n^{3/2} \cdot (\log n)^{6L+2}$ nachgewiesen.

Im Folgenden sei der Gewichtsvektor $\tilde{\mathbf{w}}$, so dass $(\tilde{w}_{k,i,j}^{(l)})_{k,i,j,l:l < L} \in \tilde{S}$ ist. Analog zum vierten Beweisschritt von Theorem 1 erhalten wir

$$\begin{aligned} &\left| F_n(\mathbf{w}^*) - F_n((\mathbf{w}^{(0)})^*) \right| \\ &\leq \left(\frac{2}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ &\quad \cdot \left(\frac{1}{n} \sum_{i=1}^n \left(\sum_{k=1}^{\hat{K}_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{\hat{K}_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \right) \right)^{1/2}. \end{aligned}$$

Aus dem Beweis von Lemma 2 ergibt sich mit $B_n = c_{77} \cdot (\log n)^2$ und aufgrund des beschränkten Trägers von \mathbf{X} die Ungleichung

$$\left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{x}) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{x}) \right|$$

$$\leq c_{82} \cdot (\log n)^{2L} \cdot \max_{i,j,l:l < L} \left| (\mathbf{w}^*)_{k,i,j}^{(l)} - ((\mathbf{w}^{(0)})^*)_{k,i,j}^{(l)} \right|. \quad (3.11)$$

Somit ist

$$\begin{aligned} & \left(\frac{1}{n} \sum_{i=1}^n \left(\sum_{k=1}^{\hat{K}_n} \left| (\mathbf{w}^*)_{1,1,k}^{(L)} \right|^2 \cdot \sum_{k=1}^{\hat{K}_n} \left| f_{\mathbf{w}^*,k,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}_i) \right|^2 \right) \right)^{1/2} \\ & \leq \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\| \cdot c_{83} \cdot (\log n)^{2L} \cdot \|(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\|. \end{aligned}$$

Für $(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} \in \tilde{S}$ ergibt sich, analog zu Beweisschritt vier in Theorem 1, durch die Voraussetzungen (3.2) und (3.3) sowie die Anwendung der Ungleichung (3.11), dass

$$\begin{aligned} & \left(\frac{2}{n} \sum_{i=1}^n \left(f_{\mathbf{w}^*}(\mathbf{X}_i) + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ & \leq \left(4 \cdot \sum_{s=1}^{\tilde{K}_n} \left| w_{1,1,s}^{(L)} \right|^2 \cdot \max_{i=1,\dots,n} \sum_{s=1}^{\tilde{K}_n} \left| f_{\mathbf{w}^*,k_s,1}^{(L)}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k_s,1}^{(L)}(\mathbf{X}_i) \right|^2 \right. \\ & \quad \left. + \frac{16}{n} \sum_{i=1}^n f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ & \leq \left(4 \cdot \sum_{s=1}^{\tilde{K}_n} \left| w_{1,1,s}^{(L)} \right|^2 \cdot c_{84} \cdot (\log n)^{4L} \cdot \|(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\|^2 \right. \\ & \quad \left. + \frac{16}{n} \cdot \sum_{i=1}^n \left(\sum_{s=1}^{\tilde{K}_n} w_{1,1,s}^{(L)} \cdot f_{(\mathbf{w}^{(0)})^*,k_s,1}^{(L)}(\mathbf{X}_i) \right)^2 + \frac{8}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \\ & \leq \left(c_{85} \cdot M_n^2 \cdot (\log n)^{4L+1} + 16 \cdot \beta_n^2 + 8 \cdot \beta_n \right)^{1/2} \\ & \leq c_{86} \cdot M_n \cdot (\log n)^{2L+1} \end{aligned}$$

gilt.

Zusammenfassend erhalten wir damit

$$\begin{aligned} & \left| F_n(\mathbf{w}^*) - F_n((\mathbf{w}^{(0)})^*) \right| \\ & \leq c_{87} \cdot M_n \cdot (\log n)^{4L+1} \cdot \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1,\dots,\hat{K}_n}\| \cdot \|(w_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L}\|. \end{aligned}$$

Daher ist Voraussetzung (2.11) in Lemma 1 für

$$D_n = c_{87} \cdot M_n \cdot (\log n)^{4L+1}$$

erfüllt.

Sei im Weiteren $\varepsilon > 0$ beliebig.

Im fünften Schritt des Beweises zeigen wir, dass

$$\mathbf{E}T_{2,n} \leq c_{88} \cdot \frac{n^{\tau \cdot d + \varepsilon}}{n}$$

gilt. Sei hierfür \mathcal{W}_n die Menge aller Gewichtsvektoren $(w_{k,i,j}^{(l)})_{k,i,j,l}$ mit den Eigenschaften

$$|w_{1,1,k}^{(L)}| \leq c_{89} \cdot (\log n)^2 \quad \text{für } k \in \{1, \dots, \widehat{K}_n\},$$

$$|w_{k,i,j}^{(l)}| \leq (20d+1) \cdot (\log n)^2 \quad \text{für } l \in \{1, \dots, L-1\}, k \in \{1, \dots, \widehat{K}_n\}, i, j \in \{1, \dots, r\}$$

sowie

$$|w_{k,i,j}^{(0)}| \leq (8d+1) \cdot (\log n)^2 \cdot n^\tau \quad \text{für } k \in \{1, \dots, \widehat{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}.$$

Wir definieren den Funktionsraum \mathcal{F}_n durch

$$\mathcal{F}_n = \{T_{\beta_n} f_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}_n\}.$$

Tritt das Ereignis A_n ein, so wissen wir aus Lemma 1, dass die Abschätzungen

$$\left\| ((\mathbf{w}^{(t)})_{1,1,k}^{(L)})_{k=1, \dots, \widehat{K}_n} - ((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1, \dots, \widehat{K}_n} \right\| \leq \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} \leq (\log n)^2$$

und

$$\left\| ((\mathbf{w}^{(t)})_{k,i,j}^{(l)})_{k,i,j,l:l < L} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l:l < L} \right\| \leq \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} \leq (\log n)^2$$

für alle $t \in \{0, \dots, t_n\}$ erfüllt sind. Somit erhalten wir aufgrund der Wahl des Startgewichtsvektors $\mathbf{w}^{(0)}$, dass $\mathbf{w}^{(t)}$ in der Menge \mathcal{W}_n für $t \in \{0, \dots, t_n\}$ enthalten ist, sofern das Ereignis A_n eintritt. Für alle $u > 0$ gilt daher

$$\begin{aligned} \mathbf{P}\{T_{2,n} > u\} &\leq \mathbf{P}\left\{ \mathbf{E}\{|m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n\} - \mathbf{E}\{|m_{\beta_n}(\mathbf{X}) - T_{\beta_n} Y|^2\} \right. \\ &\quad \left. - 2 \cdot \frac{1}{n} \sum_{i=1}^n (|m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - |m_{\beta_n}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2) > u \right\} \\ &\leq \mathbf{P}\left\{ \exists f \in \mathcal{F}_n : \mathbf{E}\left(\left| \frac{f(\mathbf{X})}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n} \right|^2 \right) - \mathbf{E}\left(\left| \frac{m_{\beta_n}(\mathbf{X})}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n} \right|^2 \right) \right. \\ &\quad \left. - \frac{1}{n} \sum_{i=1}^n \left(\left| \frac{f(\mathbf{X}_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n} \right|^2 - \left| \frac{m_{\beta_n}(\mathbf{X}_i)}{\beta_n} - \frac{T_{\beta_n} Y_i}{\beta_n} \right|^2 \right) \right. \\ &\quad \left. > \frac{1}{2} \cdot \left(\frac{u}{\beta_n^2} + \mathbf{E}\left(\left| \frac{f(\mathbf{X})}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n} \right|^2 \right) - \mathbf{E}\left(\left| \frac{m_{\beta_n}(\mathbf{X})}{\beta_n} - \frac{T_{\beta_n} Y}{\beta_n} \right|^2 \right) \right) \right\}. \end{aligned}$$

Durch die Verwendung von Lemma 10, wobei wir hier $B = 1$, $\varepsilon = 1/2$, $\alpha = \beta = \frac{u}{2 \cdot \beta_n^2}$ gesetzt haben, folgt, dass dieser Term durch

$$\begin{aligned} &14 \cdot \sup_{\mathbf{x}_1^n} \mathcal{N}_1 \left(\delta, \left\{ \frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n \right\}, \mathbf{x}_1^n \right) \cdot \exp \left(- \frac{\left(\frac{1}{2}\right)^2 \cdot \frac{1}{2} \cdot \frac{u}{2 \cdot \beta_n^2} \cdot n}{214 \cdot \left(1 + \frac{1}{2}\right)} \right) \\ &\leq 14 \cdot \sup_{\mathbf{x}_1^n} \mathcal{N}_1 \left(\delta, \left\{ \frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n \right\}, \mathbf{x}_1^n \right) \cdot \exp \left(- \frac{n \cdot u}{5136 \cdot \beta_n^2} \right) \end{aligned}$$

beschränkt ist. Wenden wir hierauf Lemma 5 mit $A = (8d + 1) \cdot (\log n)^2 \cdot n^\tau$, $B = (20d + 1) \cdot (\log n)^2$, $C = c_{89} \cdot \widehat{K}_n \cdot (\log n)^2$ sowie $\alpha = c_{90}$ an, so erhalten wir für $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in (\text{supp}(\mathbf{X}))^n$, dass

$$\begin{aligned} \mathcal{N}_1 \left(\delta, \left\{ \frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n \right\}, \mathbf{x}_1^n \right) &\leq \mathcal{N}_1(\delta \cdot \beta_n, \mathcal{F}_n, \mathbf{x}_1^n) \\ &\leq \left(\frac{c_{14} \cdot \beta_n}{\delta \cdot \beta_n} \right)^{c_{91} \cdot (\log n)^{2 \cdot (L-1) \cdot d} \cdot (\log n)^{2d} \cdot n^{\tau \cdot d} \cdot \left(\frac{\widehat{K}_n \cdot (\log n)^2}{\delta \cdot \beta_n} \right)^{d/\ell}} + c_{16} \end{aligned}$$

ist. Für $\delta > 1/n$ und $\ell > \frac{4d \cdot (\kappa+2)}{\varepsilon}$ gilt

$$\begin{aligned} \left(\frac{\widehat{K}_n \cdot (\log n)^2}{\delta \cdot \beta_n} \right)^{d/\ell} &\leq \left(\frac{n^{\kappa+1} \cdot (\log n)^2}{c_5 \cdot \log n} \right)^{d/\ell} \\ &\leq \left(\frac{1}{c_5} \cdot n^{\kappa+1} \cdot \log n \right)^{d/\ell} \\ &\leq \left(\frac{1}{c_5} \cdot n \right)^{(\kappa+2) \cdot \frac{d}{\varepsilon}} \\ &= c_{92}(\varepsilon) \cdot n^{\frac{\varepsilon}{4}}. \end{aligned}$$

Setzen wir dies in die obige Ungleichung ein, führt dies zu

$$\mathcal{N}_1 \left(\delta, \left\{ \frac{1}{\beta_n} \cdot f : f \in \mathcal{F}_n \right\}, \mathbf{x}_1^n \right) \leq c_{93}(\varepsilon) \cdot n^{c_{94} \cdot n^{\tau \cdot d + \varepsilon/2}}.$$

Zusammenfassen dieser Ergebnisse liefert für $u \geq 1/n$ die Ungleichung

$$\mathbf{P}\{T_{2,n} > u\} \leq 14 \cdot c_{93}(\varepsilon) \cdot n^{c_{94} \cdot n^{\tau \cdot d + \varepsilon/2}} \cdot \exp\left(-\frac{n \cdot u}{5136 \cdot \beta_n^2}\right).$$

Für $\varepsilon_n \geq 1/n$ ergibt sich dann

$$\begin{aligned} \mathbf{E}\{T_{2,n}\} &\leq \mathbf{E}\{\max\{T_{2,n}, 0\}\} \\ &\leq \varepsilon_n + \int_{\varepsilon_n}^{\infty} \mathbf{P}\{T_{2,n} > u\} du \\ &\leq \varepsilon_n + 14 \cdot c_{93}(\varepsilon) \cdot n^{c_{94} \cdot n^{\tau \cdot d + \varepsilon/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \varepsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n}. \end{aligned}$$

Setzen wir nun

$$\varepsilon_n = \frac{5136 \cdot \beta_n^2}{n} \cdot c_{94} \cdot n^{\tau \cdot d + \varepsilon/2} \cdot \log n,$$

so erhalten wir

$$\begin{aligned} &\varepsilon_n + 14 \cdot c_{93}(\varepsilon) \cdot n^{c_{94} \cdot n^{\tau \cdot d + \varepsilon/2}} \cdot \exp\left(-\frac{n}{5136 \cdot \beta_n^2} \cdot \varepsilon_n\right) \cdot \frac{5136 \cdot \beta_n^2}{n} \\ &\leq c_{95}(\varepsilon) \cdot \left(\frac{1}{n} \cdot n^{\tau \cdot d + \varepsilon} + n^{c_{94} \cdot n^{\tau \cdot d + \varepsilon/2}} \cdot \exp\left(-c_{94} \cdot n^{\tau \cdot d + \varepsilon/2} \cdot \log n\right) \cdot \frac{\beta_n^2}{n} \right) \\ &\leq c_{88}(\varepsilon) \cdot \frac{n^{\tau \cdot d + \varepsilon}}{n}, \end{aligned}$$

woraus die Behauptung des fünften Beweisschrittes folgt.

Im *sechsten Schritt des Beweises* zeigen wir, dass

$$\begin{aligned} & \mathbf{E}\{T_{4,n}\} \\ & \leq c_{96}(\varepsilon) \cdot \left(\sup_{\substack{(\bar{w}_{k,i,j}^{(l)})_{k,i,j,l}: \\ |\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n \ (l=0, \dots, L-1)}} \int \left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right. \\ & \quad \left. + \frac{n^{\tau \cdot d + \varepsilon}}{n} + M_n^2 \cdot (\log n)^{4L+3/2} \right) \end{aligned}$$

gilt.

Auf dem Ereignis A_n ist $|Y_i| \leq \sqrt{\beta_n} \leq \beta_n$ für alle $i = 1, \dots, n$ erfüllt. Aus der Ungleichung

$$|T_{\beta_n} z - y| \leq |z - y| \quad \text{für } |y| \leq \beta_n$$

folgt daher

$$\begin{aligned} T_{4,n}/2 &= \left[\frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbf{1}_{A_n} \\ &\leq \left[\frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t_n)}}(\mathbf{X}_i) - Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbf{1}_{A_n} \\ &= \left[F_n(\mathbf{w}^{(t_n)}) - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbf{1}_{A_n}. \end{aligned}$$

Die Anwendung von Lemma 1 führt dann zu

$$\begin{aligned} & \mathbf{E}\{T_{4,n}/2\} \\ & \leq \mathbf{E} \left\{ \left[F_n(\mathbf{w}^{(t_n)}) - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbf{1}_{A_n} \right\} \\ & \leq \mathbf{E} \left\{ \left[F_n((\mathbf{w}^{(0)})^*) + D_n \cdot \|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\| \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} \right. \right. \\ & \quad \left. \left. + \frac{\|((\mathbf{w}^*)_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n} - ((\mathbf{w}^{(0)})_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\|^2}{2} + \frac{F_n(\mathbf{w}^{(0)})}{t_n} \right. \right. \\ & \quad \left. \left. - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbf{1}_{A_n} \right\} \\ & = \mathbf{E} \left\{ \left[F_n((\mathbf{w}^{(0)})^*) + D_n \cdot \|(w_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\| \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} \right. \right. \\ & \quad \left. \left. + \frac{\|(w_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\|^2}{2} + \frac{F_n(\mathbf{w}^{(0)})}{t_n} - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbf{1}_{A_n} \right\}, \end{aligned}$$

wobei wir in der letzten Gleichheit ausgenutzt haben, dass $(\mathbf{w}^{(0)})_{1,1,k}^{(L)} = 0$ für alle $k \in \{1, \dots, \tilde{K}_n\}$ ist. Da wir aus dem vierten Beweisschritt bereits wissen, dass

$$D_n = c_{87} \cdot M_n \cdot (\log n)^{4L+1}$$

ist, ergibt sich zusammen mit Voraussetzung (3.2) die folgende Ungleichung

$$\begin{aligned} & D_n \cdot \|(w_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\| \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} + \frac{\|(w_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\|^2}{2} + \frac{F_n(\mathbf{w}^{(0)})}{t_n} \\ & \leq c_{87} \cdot M_n^2 \cdot (\log n)^{4L+3/2} + \frac{M_n^2}{2} + \frac{c_5 \cdot \log n}{t_n}. \end{aligned}$$

Für die Abschätzung des restlichen Terms zerlegen wir diesen zunächst in zwei Teile

$$\begin{aligned} & \mathbf{E} \left\{ \left[F_n((\mathbf{w}^{(0)})^*) - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right] \cdot \mathbb{1}_{A_n} \right\} \\ & = \mathbf{E} \left\{ \left[2 \cdot \left(\mathbf{E} \left\{ \left| \sum_{k=1}^{K_n} ((\mathbf{w}^{(0)})^*)_{1,1,k}^{(L)} \cdot f_{(\mathbf{w}^{(0)})^*, k, 1}^{(L)}(\mathbf{X}) - Y \right|^2 \middle| \mathcal{D}_n, \mathbf{w}^{(0)} \right\} - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \right. \right. \\ & \quad \left. \left. + \left(F_n((\mathbf{w}^{(0)})^*) - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right) \right. \right. \\ & \quad \left. \left. - 2 \cdot \left(\mathbf{E} \left\{ \left| \sum_{k=1}^{K_n} ((\mathbf{w}^{(0)})^*)_{1,1,k}^{(L)} \cdot f_{(\mathbf{w}^{(0)})^*, k, 1}^{(L)}(\mathbf{X}) - Y \right|^2 \middle| \mathcal{D}_n, \mathbf{w}^{(0)} \right\} - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \right] \cdot \mathbb{1}_{A_n} \right\} \\ & = \mathbf{E}\{T_{8,n}\} + \mathbf{E}\{T_{9,n}\}. \end{aligned}$$

Wir beginnen damit, $\mathbf{E}\{T_{8,n}\}$ abzuschätzen

$$\begin{aligned} \mathbf{E}\{T_{8,n}\} & = \mathbf{E} \left\{ 2 \cdot \left[\mathbf{E} \left\{ \left| \sum_{k=1}^{K_n} ((\mathbf{w}^{(0)})^*)_{1,1,k}^{(L)} \cdot f_{(\mathbf{w}^{(0)})^*, k, 1}^{(L)}(\mathbf{X}) - Y \right|^2 \middle| \mathcal{D}_n, \mathbf{w}^{(0)} \right\} - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right] \cdot \mathbb{1}_{A_n} \right\} \\ & = 2 \cdot \mathbf{E} \left\{ \int \left| \sum_{s=1}^{\tilde{K}_n} ((\mathbf{w}^{(0)})^*)_{1,1,k_s}^{(L)} \cdot f_{(\mathbf{w}^{(0)})^*, k_s, 1}^{(L)}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbb{1}_{A_n} \right\} \\ & \leq 2 \cdot \sup_{\substack{(\bar{w}_{k,i,j}^{(l)})_{k,i,j,l}: \\ |\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n \ (l=0, \dots, L-1)}} \int \left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}}, k, 1}^{(L)}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}). \end{aligned}$$

Um eine obere Schranke für $\mathbf{E}\{T_{9,n}\}$ zu bestimmen, nutzen wir die Tatsache, dass aus Voraussetzung (3.3) die Beschränktheit von $f_{(\mathbf{w}^{(0)})^*}$ folgt. Daher können wir ohne Beschränkung der Allgemeinheit einen Stützungsoperator einfügen. Gleiches gilt für die Regressionsfunktion m , so dass auch hier der Stützungsoperator verwendet werden kann. Zudem gilt $|Y_i| \leq \beta_n$ für $i \in \{1, \dots, n\}$, da wir uns auf dem Ereignis A_n befinden. Somit ist auch $|Y| \leq \beta_n$ fast sicher erfüllt und wir erhalten

$$\mathbf{E}\{T_{9,n}\}$$

$$\begin{aligned}
&= \mathbf{E} \left\{ \left[F_n((\mathbf{w}^{(0)})^*) - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right. \right. \\
&\quad \left. \left. - 2 \cdot \left(\mathbf{E} \left\{ \left| \sum_{k=1}^{K_n} ((\mathbf{w}^{(0)})^*)_{1,1,k}^{(L)} \cdot f_{(\mathbf{w}^{(0)})^*,k,1}^{(L)}(\mathbf{X}) - Y \right|^2 \middle| \mathcal{D}_n, \mathbf{w}^{(0)} \right\} - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \right] \cdot \mathbb{1}_{A_n} \right\} \\
&= \mathbf{E} \left\{ \left[\frac{1}{n} \sum_{i=1}^n |T_{\beta_n} f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - |T_{\beta_n} m(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right. \right. \\
&\quad \left. \left. - 2 \cdot \left(\mathbf{E} \left\{ |T_{\beta_n} f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}) - T_{\beta_n} Y|^2 \middle| \mathcal{D}_n, \mathbf{w}^{(0)} \right\} - \mathbf{E} \{ |T_{\beta_n} m(\mathbf{X}) - T_{\beta_n} Y|^2 \} \right) \right] \cdot \mathbb{1}_{A_n} \right\}.
\end{aligned}$$

Gemäß Bemerkung 2 können wir Lemma 10 anwenden und erhalten damit analog zu Beweisschritt fünf die Ungleichung

$$\begin{aligned}
&\mathbf{E} \left\{ \left[\frac{1}{n} \sum_{i=1}^n |T_{\beta_n} f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - |T_{\beta_n} m(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right. \right. \\
&\quad \left. \left. - 2 \cdot \left(\mathbf{E} \left\{ |T_{\beta_n} f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}) - T_{\beta_n} Y|^2 \middle| \mathcal{D}_n, \mathbf{w}^{(0)} \right\} - \mathbf{E} \{ |T_{\beta_n} m(\mathbf{X}) - T_{\beta_n} Y|^2 \} \right) \right] \cdot \mathbb{1}_{A_n} \right\} \\
&\leq c_{97}(\varepsilon) \cdot \frac{n^{\tau \cdot d + \varepsilon}}{n}.
\end{aligned}$$

Zusammenfassend folgt dann mit $\mathbf{P}(A_n^c) \leq \frac{c_{75}}{n}$, dass

$$\begin{aligned}
&\mathbf{E} \left\{ \left[F_n((\mathbf{w}^{(0)})^*) - \frac{1}{n} \sum_{i=1}^n |m(\mathbf{X}_i) - Y_i|^2 \right. \right. \\
&\quad \left. \left. + D_n \cdot \|(w_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\| \cdot \sqrt{2 \cdot F_n(\mathbf{w}^{(0)})} + \frac{\|(w_{1,1,k}^{(L)})_{k=1, \dots, \tilde{K}_n}\|^2}{2} + \frac{F_n(\mathbf{w}^{(0)})}{t_n} \right] \cdot \mathbb{1}_{A_n} \right\} \\
&\leq 2 \cdot \sup_{\substack{(\bar{w}_{k,i,j}^{(l)})_{k,i,j,l}: \\ |\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n \ (l=0, \dots, L-1)}} \int \left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\
&\quad + c_{97}(\varepsilon) \cdot \frac{n^{\tau \cdot d + \varepsilon}}{n} + c_{87} \cdot M_n^2 \cdot (\log n)^{4L+3/2} + \frac{M_n^2}{2} + \frac{c_5 \cdot \log n}{t_n}
\end{aligned}$$

gilt.

Dies impliziert

$$\begin{aligned}
&\mathbf{E}\{T_{4,n}\} \\
&\leq c_{96}(\varepsilon) \cdot \left(\sup_{\substack{(\bar{w}_{k,i,j}^{(l)})_{k,i,j,l}: \\ |\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n \ (l=0, \dots, L-1)}} \int \left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right. \\
&\quad \left. + \frac{n^{\tau \cdot d + \varepsilon}}{n} + M_n^2 \cdot (\log n)^{4L+3/2} \right).
\end{aligned}$$

Fassen wir alle Beweisschritte zusammen, erhalten wir die Aussage von Theorem 2. □

3.2 Konvergenzgeschwindigkeit des Neuronale-Netze-Schätzers

In diesem Abschnitt leiten wir die Konvergenzrate für den überparametrisierten tiefen Neuronale-Netze-Schätzer her. Grundlage dafür wird die zuvor hergeleitete Fehlerschranke aus Theorem 2 sein.

Theorem 3. Sei $n \in \mathbb{N}$ und seien $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängig identisch verteilte $\mathbb{R}^d \times \mathbb{R}$ -wertige Zufallsvariablen mit $\text{supp}(\mathbf{X}) \subseteq [0, 1]^d$ sowie

$$\mathbf{E} \left\{ \exp(c_{64} \cdot Y^2) \right\} < \infty \quad (3.12)$$

für eine Konstante $c_{64} > 0$. Wir nehmen an, dass die zugehörige Regressionsfunktion $m(\mathbf{x}) = \mathbf{E}\{Y|\mathbf{X} = \mathbf{x}\}$ für $p \in [1/2, 1]$ und $C > 0$ eine (p, C) -glatte Funktion ist.

Sei $\sigma(z) = 1/(1 + \exp(-z))$ die logistische Sigmoidfunktion. Seien zudem $L, r, t_n \in \mathbb{N}$ mit $L \geq 2$ und $r \geq 2d$. Wir setzen $\beta_n = c_5 \cdot \log n$ für eine Konstante $c_5 > 0$,

$$\widehat{K}_n = n^{6d+r+2} \quad \text{sowie} \quad \tau = \frac{1}{1+d}.$$

Der Schätzer m_n sei definiert wie in Abschnitt 2.1 mit

$$\lambda_n = \frac{1}{t_n} \quad \text{und} \quad t_n = \lceil c_{65} \cdot C_n \rceil \quad (3.13)$$

für eine Konstante $c_{65} \geq 2$ und $C_n > 0$ mit

$$C_n \geq \widehat{K}_n^{3/2} \cdot (\log n)^{6L+2}.$$

Des Weiteren sei

$$c_{64} \cdot c_5 \geq 2. \quad (3.14)$$

Dann gilt für jedes $\varepsilon > 0$, dass

$$\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_{98} \cdot n^{-\frac{1}{1+d} + \varepsilon}$$

für eine Konstante $c_{98} = c_{98}(\varepsilon) > 0$, die von ε abhängt, ist.

In Kapitel 1 haben wir gesehen, dass die optimale Konvergenzrate für (p, C) -glatte Funktionen durch

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

gegeben ist. Setzen wir $p = \frac{1}{2}$ in Theorem 3, so erhalten wir eine Konvergenzrate, deren Exponent annähernd optimal ist.

Für den Beweis von Theorem 3 benötigen wir neben der Fehlerschranke aus Theorem 2 ein weiteres Resultat, welches uns eine präzise Analyse des Approximationsfehler für (p, C) -glatte Funktionen liefert. Zudem stellt das folgende Lemma sicher, dass die äußeren Gewichte des neuronalen Netzes hinreichend klein sind.

Lemma 11. Sei $f : \mathbb{R}^d \rightarrow \mathbb{R}$ eine (p, C) -glatte Funktion mit $p \in [1/2, 1]$ und $C > 0$. Sei \mathbf{X} eine \mathbb{R}^d -wertige Zufallsvariable mit $\text{supp}(\mathbf{X}) \subseteq [0, 1]^d$. Des Weiteren seien $b \in \mathbb{N}$ und $0 < \delta < 1/2$ mit

$$c_{99} \cdot \delta \leq \frac{1}{2^b} \leq c_{100} \cdot \delta. \quad (3.15)$$

Ebenso seien $L, r, s \in \mathbb{N}$ mit $L \geq 2$ und $r \geq 2d$. Sei zudem $\tilde{K}_n \in \mathbb{N}$ mit

$$\tilde{K}_n \geq \left(b \cdot (2^b + 1)^{2d} + 1 \right)^3.$$

Dann existieren Gewichte

$$w_{k,i,j}^{(l)} \in [-20d \cdot (\log n)^2, 20d \cdot (\log n)^2] \quad \text{für } l \in \{1, \dots, L\}, k \in \{1, \dots, \tilde{K}_n\}, i, j \in \{1, \dots, r\}$$

und

$$w_{k,i,j}^{(0)} \in \left[-\frac{8 \cdot d \cdot (\log n)^2}{\delta}, \frac{8 \cdot d \cdot (\log n)^2}{\delta} \right] \quad \text{für } k \in \{1, \dots, \tilde{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}$$

derart, dass für alle Gewichtsvektoren $\bar{\mathbf{w}}$ mit $|\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n$ für $l = 0, \dots, L-1$ und hinreichend großes n die Ungleichungen

$$\int \left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(\mathbf{x}) - f(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_{101} \cdot \left(b^2 \cdot \delta + \delta^{2p} + \frac{b \cdot (2^b + 1)^{2d}}{n^s} \right), \quad (3.16)$$

$$\left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{\mathbf{w}},k,1}^{(L)}(\mathbf{x}) \right| \leq c_{102} \cdot \left(1 + \frac{(2^b + 1)^{2d}}{n^s} \right) \quad (\mathbf{x} \in [0, 1]^d) \quad (3.17)$$

und

$$\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \leq \frac{c_{103}}{2^{2 \cdot d \cdot b}} \quad (3.18)$$

gelten.

Beweis. Siehe Lemma 7 in Kohler und Krzyżak (2022b). □

Im Beweis von Theorem 3 werden wir die Aussagen von Theorem 2 und Lemma 11 kombinieren und damit die Konvergenzgeschwindigkeit für den überparametrisierten Neuronale-Netze-Schätzer nachweisen.

Beweis von Theorem 3. Sei $b = \lfloor \frac{1}{1+d} \log_2(n) \rfloor$. Dann gilt

$$\frac{1}{n^{\frac{1}{1+d}}} = \frac{1}{2^{\frac{1}{1+d} \log_2(n)}} \leq \frac{1}{2^b} \leq \frac{1}{2^{\frac{1}{1+d} \log_2(n) - 1}} = \frac{2}{n^{\frac{1}{1+d}}},$$

womit Voraussetzung (3.15) für $\delta = n^{-1/(1+d)}$ erfüllt ist. Wir setzen nun

$$\tilde{K}_n = \left(b \cdot (2^b + 1)^{2d} + 1 \right)^3 \approx c_{104} \cdot (\log n)^3 \cdot n^{\frac{6d}{1+d}}$$

und definieren den Gewichtsvektor w wie in Lemma 11. Somit ist

$$w_{k,i,j}^{(l)} \in [-20d \cdot (\log n)^2, 20d \cdot (\log n)^2] \quad \text{für } l \in \{1, \dots, L\}, k \in \{1, \dots, \tilde{K}_n\}, i, j \in \{1, \dots, r\},$$

sowie

$$w_{k,i,j}^{(0)} \in [-8d \cdot (\log n)^2 \cdot n^\tau, 8d \cdot (\log n)^2 \cdot n^\tau] \quad \text{für } k \in \{1, \dots, \tilde{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d\}$$

und $\tau = 1/(1+d)$.

Aus Ungleichung (3.18) in Lemma 11 ergibt sich, dass

$$\sum_{k=1}^{\tilde{K}_n} |w_{1,1,k}^{(L)}|^2 \leq c_{103} \cdot 2^{-2 \cdot d \cdot \frac{1}{1+d} \log_2(n)} \leq c_{103} \cdot n^{-\frac{2d}{1+d}}$$

gilt. Daraus folgt, dass Voraussetzung (3.2) für $M_n = \sqrt{c_{103}} \cdot n^{-d/(1+d)}$ erfüllt ist.

Da für eine spätere Abschätzung $s \geq \frac{3d}{1+d}$ gewählt wird, folgt Voraussetzung (3.3) direkt aus Bedingung (3.17) in Lemma 11, da

$$\left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{w},k,1}^{(L)}(\mathbf{x}) \right| \leq c_{102} \cdot \left(1 + \frac{(2^b + 1)^{2d}}{n^s} \right) \leq c_{105} \cdot \frac{n^{\frac{2d}{1+d}}}{n^s} \leq \beta_n$$

für alle Gewichtsvektoren \bar{w} mit $|\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n$ und hinreichend großes n gilt.

Durch die Anwendung von Theorem 2 mit $\tau = \frac{1}{1+d}$ sowie Ungleichung (3.16) aus Lemma 11 mit $s \geq \frac{3d}{1+d}$ erhalten wir für hinreichend großes n die Abschätzung

$$\begin{aligned} \mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) &\leq c_{66} \cdot \left(\frac{n^{\frac{1}{1+d} \cdot d + \varepsilon}}{n} + c_{103} \cdot \frac{(\log n)^{4L+3/2}}{n^{\frac{2d}{1+d}}} \right. \\ &\quad \left. + \sup_{\substack{(\bar{w}_{k,i,j}^{(l)})_{k,i,j,l} \\ |\bar{w}_{k,i,j}^{(l)} - w_{k,i,j}^{(l)}| \leq \log n \ (l=0, \dots, L-1)}} \int \left| \sum_{k=1}^{\tilde{K}_n} w_{1,1,k}^{(L)} \cdot f_{\bar{w},k,1}^{(L)}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right) \\ &\leq c_{106} \cdot \left(\frac{n^{\frac{1}{1+d} \cdot d + \varepsilon}}{n} + \frac{(\log n)^{4L+3/2}}{n^{\frac{2d}{1+d}}} + \frac{(\log n)^2}{n^{\frac{1}{1+d}}} + \frac{1}{n^{\frac{2p}{1+d}}} + \frac{\log n \cdot n^{\frac{2d}{1+d}}}{n^s} \right) \\ &\leq c_{98} \cdot n^{-\frac{1}{1+d} + \varepsilon} \end{aligned}$$

für eine Konstante $c_{98} = c_{98}(\varepsilon) > 0$, die von ε abhängt. \square

3.3 Konvergenzgeschwindigkeit des Schätzers im Interaktionsmodell

Das Ziel dieses Abschnitts ist es, den zu Beginn des Kapitels definierten Schätzer so zu modifizieren, dass wir eine Konvergenzgeschwindigkeit erzielen, die unabhängig von der Eingabedimension d ist. Hierfür

nehmen wir an, dass die Regressionsfunktion m einem Interaktionsmodell genügt. Wie bereits in Kapitel 1 erwähnt, lässt sich die Regressionsfunktion dann als eine Summe (p, C) -glatter Funktionen darstellen. Konkret bedeutet dies

$$m(\mathbf{x}) = \sum_{I \subseteq \{1, \dots, d\} : |I|=d^*} m_I(\mathbf{x}_I) \quad (\mathbf{x} = (x^{(1)}, \dots, x^{(d)})^\top),$$

wobei $1 \leq d^* < d$ gilt und die Funktionen $m_I : \mathbb{R}^{|I|} \rightarrow \mathbb{R}$ (p, C) -glatt sind.

Sei im Folgenden $f_{\mathbf{w}_I}$ für $I \subseteq \{1, \dots, d\}$ mit $|I| = d^*$ gemäß (2.1)–(2.3) definiert, wobei d durch d^* ersetzt wird. Dann ist das neuronale Netz, welches die Basis des Schätzers bildet, gegeben durch

$$\tilde{f}_{\mathbf{w}}(\mathbf{x}) = \sum_{I \subseteq \{1, \dots, d\} : |I|=d^*} f_{\mathbf{w}_I}(\mathbf{x}_I).$$

Der Gewichtsvektor \mathbf{w}_I ist definiert durch

$$\mathbf{w} = (\mathbf{w}_I)_{I \subseteq \{1, \dots, d\}, |I|=d^*}.$$

Für die Durchführung des Gradientenverfahrens benötigen wir eine Initialisierung der Startgewichte

$$\mathbf{w}^{(0)} = \left(\left(\left(\mathbf{w}_I^{(0)} \right)_{k,i,j}^{(l)} \right)_{k,i,j,l} \right)_{I \subseteq \{1, \dots, d\}, |I|=d^*}.$$

Im Folgenden sei $I \subseteq \{1, \dots, d\}$ mit $|I| = d^*$. Wir setzen die Gewichte der Ausgabeschicht

$$(\mathbf{w}_I^{(0)})_{1,1,k}^{(L)} = 0 \quad \text{für } k = 1, \dots, \hat{K}_n$$

und wählen alle anderen Gewichte $(\mathbf{w}_I^{(0)})_{k,i,j}^{(l)}$ unabhängig voneinander. Die Gewichte $(\mathbf{w}_I^{(0)})_{k,i,j}^{(l)}$ mit $l \in \{1, \dots, L-1\}$ werden gleichverteilt aus dem Intervall $[-20d^* \cdot (\log n)^2, 20d^* \cdot (\log n)^2]$ gewählt, während die Gewichte der ersten verdeckten Schicht $(\mathbf{w}_I^{(0)})_{k,i,j}^{(0)}$ gleichverteilt aus dem Intervall $[-8d^* \cdot (\log n)^2 \cdot n^\tau, 8d^* \cdot (\log n)^2 \cdot n^\tau]$ mit $\tau = \frac{1}{1+d^*}$ gewählt werden.

Wir setzen die Schrittweite

$$\lambda_n = \frac{1}{t_n},$$

wobei die Anzahl der Gradientenschritte t_n in Theorem 4 gewählt wird. Die Gewichte des Netzwerkes werden iterativ durch den Gradientenabstieg bestimmt. Das heißt,

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \lambda_n \cdot (\nabla_{\mathbf{w}} F_n)(\mathbf{w}^{(t)})$$

für $t = 0, \dots, t_n - 1$. Die Verlustfunktion entspricht hierbei dem L_2 -Risiko F_n des Netzwerkes $\tilde{f}_{\mathbf{w}}$ auf den Trainingsdaten und ist gegeben durch

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n |\tilde{f}_{\mathbf{w}}(\mathbf{X}_i) - Y_i|^2.$$

Der Schätzer ist dann definiert durch

$$\tilde{m}_n(\mathbf{x}) = T_{\beta_n} \tilde{f}_{\mathbf{w}^{(t_n)}}(\mathbf{x}),$$

mit $\beta_n = c_5 \cdot \log n$.

Das folgende Resultat zeigt, dass es unter geeigneten Annahmen möglich ist, durch einen überparametrisierten tiefen Neuronale-Netze-Schätzer eine gute sowie dimensionsunabhängige Konvergenzgeschwindigkeit zu erzielen, sofern die Regressionsfunktion einem Interaktionsmodell genügt.

Theorem 4. Seien $d \in \mathbb{N}$, $d^* \in \{1, \dots, d\}$ und $p \in [1/2, 1]$. Des Weiteren seien $C > 0$, $n \in \mathbb{N}$ und $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängige und identisch verteilte $\mathbb{R}^d \times \mathbb{R}$ -wertige Zufallsvariablen mit $\text{supp}(\mathbf{X}) \subseteq [0, 1]^d$ sowie

$$\mathbf{E}\{\exp(c_{107} \cdot Y^2)\} < \infty$$

mit einer Konstanten $c_{107} > 0$. Sei zudem $m(\mathbf{x}) = \mathbf{E}\{Y|\mathbf{X} = \mathbf{x}\}$ die zugehörige Regressionsfunktion mit

$$m(\mathbf{x}) = \sum_{I \subseteq \{1, \dots, d\}: |I|=d^*} m_I(\mathbf{x}_I) \quad \text{für } \mathbf{x} \in [0, 1]^d,$$

wobei jedes $m_I : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ ($I \subseteq \{1, \dots, d\}, |I| = d^*$) eine (p, C) -glatte Funktion ist.

Sei $\sigma(z) = 1/(1 + \exp(-z))$ die logistische Sigmoidfunktion und seien $L, r \in \mathbb{N}$ mit $L \geq 2$ und $r \geq 2d^*$. Setze $\beta_n = c_5 \cdot \log n$ für eine Konstante $c_5 > 0$,

$$\widehat{K}_n = n^{6d^*+r+2} \quad \text{sowie} \quad \tau = \frac{1}{1+d^*}.$$

Der Schätzer m_n sei definiert wie in Abschnitt 3.3 beschrieben mit

$$\lambda_n = \frac{1}{t_n} \quad \text{und} \quad t_n = \lceil c_{108} \cdot C_n \rceil$$

für eine Konstante $c_{108} \geq 2$ und $C_n > 0$ mit

$$C_n \geq \widehat{K}_n^{3/2} \cdot (\log n)^{6L+2}.$$

Zusätzlich gelte

$$c_{107} \cdot c_5 \geq 2. \tag{3.19}$$

Dann gilt für alle $\varepsilon > 0$, dass

$$\mathbf{E} \int |\tilde{m}_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \leq c_{109} \cdot n^{-\frac{1}{1+d^*} + \varepsilon}$$

für eine Konstante $c_{109} = c_{109}(\varepsilon) > 0$, die von ε abhängt, ist.

Die hergeleitete Konvergenzgeschwindigkeit ist unabhängig von der Eingabedimension d . Daher zeigt Theorem 4, dass unser Schätzer den Fluch der Dimensionalität umgehen kann.

Um Theorem 4 zu beweisen, benötigen wir das folgende Lemma, welches die Komplexität einer Menge überparametrisierter tiefer neuronaler Netze für Interaktionsmodelle kontrolliert.

Lemma 12. Seien $\alpha \geq 1$, $\beta > 0$ sowie $\tilde{A}, \tilde{B}, \tilde{C} \geq 1$. Weiter sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ eine ℓ -mal differenzierbare Funktion, so dass alle ihre Ableitungen bis zur Ordnung ℓ auf \mathbb{R} beschränkt sind. Sei \mathcal{F} die Menge aller Funktionen der Form

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{I \subseteq \{1, \dots, d\}: |I|=d^*} f_{\mathbf{w}_I}(\mathbf{x}_I),$$

mit $f_{\mathbf{w}_I}$, welches wie in (2.1)–(2.3) definiert ist, wobei d durch d^* ersetzt wurde und dem Gewichtsvektor \mathbf{w}_I ,

$$\mathbf{w} = (\mathbf{w}_I)_{I \subseteq \{1, \dots, d\}: |I|=d^*}.$$

Zusätzlich erfülle der Gewichtsvektor \mathbf{w}_I für alle $I \subseteq \{1, \dots, d\}$ mit $|I| = d^*$ die Bedingungen

$$\sum_{k=1}^{\hat{K}_n} |(\mathbf{w}_I)_{1,1,k}^{(L)}| \leq \tilde{C}, \quad (3.20)$$

$$|(\mathbf{w}_I)_{k,i,j}^{(l)}| \leq \tilde{B} \quad \text{für } k \in \{1, \dots, \hat{K}_n\}, i, j \in \{1, \dots, r\}, l \in \{1, \dots, L-1\} \quad (3.21)$$

und

$$|(\mathbf{w}_I)_{k,i,j}^{(0)}| \leq \tilde{A} \quad \text{für } k \in \{1, \dots, \hat{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d^*\}. \quad (3.22)$$

Dann gilt für alle $1 \leq p < \infty$, $0 < \varepsilon < \beta$ und $\mathbf{x}_1^n \in [-\alpha, \alpha]^{d \cdot n}$ die Ungleichung

$$\mathcal{N}_p(\varepsilon, \{T_\beta f : f \in \mathcal{F}\}, \mathbf{x}_1^n) \leq \left(c_{110} \cdot \frac{\beta^p}{\varepsilon^p} \right)^{c_{111} \cdot \alpha^{d^*} \cdot \tilde{A}^{d^*} \cdot \tilde{B}^{(L-1) \cdot d^*} \left(\frac{\tilde{C}}{\varepsilon} \right)^{d^*/\ell} + c_{112}}.$$

Beweis. Siehe Lemma 8 in Kohler und Krzyżak (2022b). □

Für den Nachweis der Konvergenzgeschwindigkeit im Interaktionsmodell benötigen wir die Aussage von Theorem 2 für Interaktionsmodelle, welche im folgenden Lemma formuliert wird.

Lemma 13. Sei $n \in \mathbb{N}$ mit $n \geq 2$ und $\beta_n = c_5 \cdot \log n$. Seien $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängig und identisch verteilte $\mathbb{R}^d \times \mathbb{R}$ -wertige Zufallsvariablen, wobei $\text{supp}(\mathbf{X})$ beschränkt ist und

$$\mathbf{E} \{ \exp(c_{107} \cdot Y^2) \} < \infty \quad (3.23)$$

für eine Konstante $c_{107} > 0$ mit $c_{107} \cdot c_5 \geq 2$ gilt. Die zugehörige Regressionsfunktion

$$m(\mathbf{x}) = \sum_{I \subseteq \{1, \dots, d\}: |I|=d^*} m_I(\mathbf{x}_I)$$

sei beschränkt, wobei $1 \leq d^* < d$ ist und die Funktionen $m_I : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ (p, C)-glatt sind. Weiter sei $\sigma(z) = 1/(1 + \exp(-z))$ die logistische Sigmoidfunktion sowie $\hat{K}_n, L, r, t_n \in \mathbb{N}$, $M_n \geq 1$ und $\lambda_n, \tau > 0$.

Des Weiteren sei $\tilde{K}_n \in \{1, \dots, \hat{K}_n\}$ und es gelte für $I \subseteq \{1, \dots, d\}$ mit $|I| = d^*$, dass

$$(\mathbf{w}_I)_{k,i,j}^{(l)} \in [-20d^* \cdot (\log n)^2, 20d^* \cdot (\log n)^2] \quad \text{für } l \in \{1, \dots, L\}, k \in \{1, \dots, \tilde{K}_n\}, i, j \in \{1, \dots, r\},$$

$$(\mathbf{w}_I)_{k,i,j}^{(0)} \in [-8d^* \cdot (\log n)^2 \cdot n^\tau, 8d^* \cdot (\log n)^2 \cdot n^\tau] \quad \text{für } k \in \{1, \dots, \tilde{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d^*\},$$

$$\sqrt{\sum_{k=1}^{\tilde{K}_n} |(\mathbf{w}_I)_{1,1,k}^{(L)}|^2} \leq M_n \quad (3.24)$$

und

$$\left| \sum_{k=1}^{\tilde{K}_n} (\mathbf{w}_I)_{1,1,k}^{(L)} \cdot \tilde{f}_{\bar{\mathbf{w}}_I,k,1}^{(L)}(\mathbf{x}_I) \right| \leq \beta_n \quad (3.25)$$

für $\mathbf{x} \in \text{supp}(\mathbf{X})$ sowie für alle Gewichtsvektoren $\bar{\mathbf{w}}$ mit

$$\left| (\bar{\mathbf{w}}_I)_{k,i,j}^{(l)} - (\mathbf{w}_I)_{k,i,j}^{(l)} \right| \leq \log n \quad \text{für } l = 0, \dots, L-1.$$

Zudem sei

$$\frac{\hat{K}_n}{n^\kappa} \rightarrow 0 \quad (n \rightarrow \infty) \quad (3.26)$$

für ein $\kappa > 0$ und

$$\frac{\hat{K}_n}{\tilde{K}_n \cdot n^{\tau \cdot (d^*+1) \cdot \tau + 1}} \rightarrow \infty \quad (n \rightarrow \infty). \quad (3.27)$$

Der Schätzer \tilde{m}_n sei definiert wie in Abschnitt 3.3, wobei

$$\lambda_n = \frac{1}{t_n} \quad \text{und} \quad t_n = \lceil c_{108} \cdot C_n \rceil \quad (3.28)$$

für eine Konstante $c_{108} \geq 2$ und für $C_n > 0$ mit

$$C_n \geq \hat{K}_n^{3/2} \cdot (\log n)^{6L+2}.$$

Dann gilt für alle $\varepsilon > 0$ die Abschätzung

$$\begin{aligned} & \mathbf{E} \int |\tilde{m}_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ & \leq c_{113} \cdot \left(\frac{n^{\tau \cdot d^* + \varepsilon}}{n} + M_n^2 \cdot (\log n)^{4L+3/2} \right. \\ & \quad \left. + \sup_{\substack{((\bar{\mathbf{w}}_I)_{k,i,j}^{(l)})_{k,i,j,l} : I \subseteq \{1, \dots, d\}, |I|=d^* \\ |(\bar{\mathbf{w}}_I)_{k,i,j}^{(l)} - (\mathbf{w}_I)_{k,i,j}^{(l)}| \leq \log n \\ (l=0, \dots, L-1, I \subseteq \{1, \dots, d\}, |I|=d^*)}} \int \left| \sum_{I \subseteq \{1, \dots, d^*\}: |I|=d^*} \sum_{k=1}^{\tilde{K}_n} (\mathbf{w}_I)_{1,1,k}^{(L)} \cdot \tilde{f}_{\bar{\mathbf{w}}_I,k,1}^{(L)}(\mathbf{x}_I) - m(\mathbf{x}_I) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right) \end{aligned}$$

für eine Konstante $c_{113} = c_{113}(\varepsilon) > 0$, die von ε abhängt.

Beweis. Unter Verwendung von Lemma 12 folgt der Beweis analog zu dem Beweis von Theorem 2. \square

Für den Nachweis der Konvergenzgeschwindigkeit im Interaktionsmodell können wir analog zu dem Beweis von Theorem 3 vorgehen.

Beweis von Theorem 4. Sei $b = \lfloor \frac{1}{1+d^*} \log_2(n) \rfloor$. Weiter sei

$$\tilde{K}_n = \binom{d}{d^*} \left(b \cdot (2^b + 1)^{2d^*} + 1 \right)^3 \approx c_{114} \cdot (\log n)^3 \cdot n^{\frac{6d^*}{1+d^*}}.$$

Dann ist wegen

$$\frac{1}{n^{\frac{1}{1+d^*}}} \leq \frac{1}{2^b} \leq \frac{2}{n^{\frac{1}{1+d^*}}}$$

Voraussetzung (3.15) für $\delta = n^{-1/(1+d^*)}$ erfüllt. Auch hier definieren wir den Gewichtsvektor \mathbf{w} , so dass die Komponenten gemäß Lemma 11 gewählt werden, um m_I zu approximieren.

Damit ist

$$(\mathbf{w}_I)_{k,i,j}^{(l)} \in [-20d^* \cdot (\log n)^2, 20d^* \cdot (\log n)^2] \quad \text{für } l \in \{1, \dots, L\}, k \in \{1, \dots, \tilde{K}_n\}, i, j \in \{1, \dots, r\}$$

sowie

$$(\mathbf{w}_I)_{k,i,j}^{(0)} \in [-8d^* \cdot (\log n)^2 \cdot n^\tau, 8d^* \cdot (\log n)^2 \cdot n^\tau] \quad \text{für } k \in \{1, \dots, \tilde{K}_n\}, i \in \{1, \dots, r\}, j \in \{1, \dots, d^*\}$$

und $\tau = 1/(1+d^*)$.

Die Voraussetzungen (3.24) und (3.25) in Lemma 13 sind für den Gewichtsvektor \mathbf{w} erfüllt, da

$$\sum_{I \subseteq \{1, \dots, d\}: |I|=d^*} \sum_{k=1}^{\tilde{K}_n} |(\mathbf{w}_I)_{1,1,k}^{(L)}|^2 \leq c_{115} \cdot n^{-\frac{2d^*}{1+d^*}}$$

beziehungsweise

$$\sum_{I \subseteq \{1, \dots, d\}: |I|=d^*} \left| \sum_{k=1}^{\tilde{K}_n} (\mathbf{w}_I)_{1,1,k}^{(L)} \cdot \tilde{f}_{(\mathbf{w}_I)_{k,1}}^{(L)}(\mathbf{x}) \right| \leq c_{116} \cdot \frac{n^{\frac{2d^*}{1+d^*}}}{n^s} \leq \beta_n$$

für $s \geq \frac{3d^*}{1+d^*}$ und alle Gewichtsvektoren $\bar{\mathbf{w}}$ mit $|(\bar{\mathbf{w}}_I)_{k,i,j}^{(l)} - (\mathbf{w}_I)_{k,i,j}^{(l)}| \leq \log n$ für $l = 0, \dots, L-1$ sowie n hinreichend groß gelten.

Mit Lemma 13 und Lemma 11 erhalten wir für $\tau = \frac{1}{1+d^*}$ und $s \geq \frac{3d^*}{1+d^*}$ die Ungleichung

$$\begin{aligned} \mathbf{E} \int |\tilde{m}_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) &\leq c_{113} \cdot \left(\frac{n^{\frac{1}{1+d^*} \cdot d^* + \varepsilon}}{n} + c_{115} \cdot \frac{(\log n)^{4L+3/2}}{n^{\frac{2d^*}{1+d^*}}} \right) \\ &+ \sup_{\substack{((\bar{\mathbf{w}}_I)_{k,i,j}^{(l)})_{k,i,j,l} : I \subseteq \{1, \dots, d\}, |I|=d^* : \\ |(\bar{\mathbf{w}}_I)_{k,i,j}^{(l)} - (\mathbf{w}_I)_{k,i,j}^{(l)}| \leq \log n \\ (l=0, \dots, L-1, I \subseteq \{1, \dots, d\}, |I|=d^*)}} \int \left| \sum_{I \subseteq \{1, \dots, d^*\}: |I|=d^*} \sum_{k=1}^{\tilde{K}_n} (\mathbf{w}_I)_{1,1,k}^{(L)} \cdot \tilde{f}_{\bar{\mathbf{w}}_I, k, 1}^{(L)}(\mathbf{x}_I) - m(\mathbf{x}_I) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &\leq c_{117} \cdot \left(\frac{n^{\frac{1}{1+d^*} \cdot d^* + \varepsilon}}{n} + \frac{(\log n)^{4L+3/2}}{n^{\frac{2d^*}{1+d^*}}} + \binom{d}{d^*} \cdot \left(\frac{(\log n)^2}{n^{\frac{1}{1+d^*}}} + \frac{1}{n^{\frac{2p}{1+d^*}}} + \frac{\log n \cdot n^{\frac{2d^*}{1+d^*}}}{n^s} \right) \right) \\ &\leq c_{118} \cdot n^{-\frac{1}{1+d^*} + \varepsilon} \end{aligned}$$

für eine Konstante $c_{118} = c_{118}(\varepsilon) > 0$, die von ε abhängt. □

In diesem Kapitel haben wir für einen überparametrisierten tiefen Neuronale-Netze-Schätzer, der durch Gradientenabstieg bezüglich des empirischen L_2 -Risikos ohne Regularisierungsterm trainiert wurde, eine Konvergenzgeschwindigkeit von $n^{-\frac{1}{1+d}+\varepsilon}$ für (p, C) -glatte Regressionsfunktionen mit $p \in [1/2, 1]$ hergeleitet. Für den Fall $p = \frac{1}{2}$ ist der Exponent der Konvergenzrate annähernd optimal.

Darüber hinaus konnten wir für den Fall, dass die Regressionsfunktion einem Interaktionsmodell für $p \in [1/2, 1]$ genügt, eine dimensionsunabhängige Konvergenzgeschwindigkeit erzielen. Dies belegt, dass der überparametrisierte Neuronale-Netze-Schätzer in der Lage ist, den Fluch der Dimensionalität zu umgehen.

Für den Nachweis der Konvergenzraten haben wir verschiedene Resultate aus den Bereichen der Optimierung, Generalisierung und Approximation verwendet, mit deren Hilfe die entsprechenden Fehler abgeschätzt werden konnten.

Aufbauend auf diesen Ergebnissen, die bezüglich weniger glatten Funktionen gelten, werden wir nun unsere Betrachtungen auf Regressionsfunktionen mit einer Glattheitsordnung von $p \geq 1$ erweitern.

Wir möchten untersuchen, wie die Einschränkung der Regressionsfunktion auf hierarchische Kompositionsmodelle die Konvergenzgeschwindigkeit des überparametrisierten Neuronale-Netze-Schätzer sowie dessen Abhängigkeit von der Eingabedimension d beeinflusst. Dafür werden wir im Folgenden die ReLU-Aktivierungsfunktion verwenden.

4 Dimensionsreduktion überparametrisierter tiefer Neuronale-Netze-Schätzer trainiert durch Gradientenabstieg

Im vorherigen Kapitel haben wir gezeigt, dass für eine (p, C) -glatte Regressionsfunktion eine gute Konvergenzrate eines überparametrisierten tiefen Neuronale-Netze-Schätzers erreicht werden kann, der durch Gradientenabstieg trainiert wurde. Diese Konvergenzrate kann allerdings nur für $p \in [1/2, 1]$ erzielt werden und somit ausschließlich für Funktionen mit geringer Glattheit.

Das Ziel dieses Kapitels ist es, eine Konvergenzgeschwindigkeit eines Neuronale-Netze-Schätzers für (p, C) -glatte Regressionsfunktionen mit $p \geq 1$ herzuleiten. Dabei nutzen wir die ReLU-Aktivierungsfunktion. Um den Fluch der Dimensionalität zu umgehen, werden wir zusätzlich annehmen, dass die Regressionsfunktion eine kompositionelle Struktur aufweist und somit eine Verallgemeinerung des Interaktionsmodells aus dem vorherigen Kapitel darstellt.

Eine zentrale Rolle im Beweis dieses Resultates spielt das Approximationsresultat aus Kohler und Langer (2021), welches die gute Approximationsfähigkeit neuronaler Netze für (p, C) -glatte Funktionen belegt.

Bevor wir jedoch die Konvergenzgeschwindigkeit im Detail betrachten, werden wir den Neuronale-Netze-Schätzer einführen.

4.1 Einführung des Schätzers

Für die Definition dieses Neuronale-Netze-Schätzers wählen wir die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Wir bezeichnen die Anzahl der Gewichte des neuronalen Netzes mit $W \in \mathbb{N}$ und definieren $\mathcal{W} \subseteq \mathbb{R}^W$ als eine abgeschlossene und konvexe Menge, welche die Gewichte des Netzes für die gegebene Topologie enthält.

Im Folgenden ist es unser Ziel, den Gewichtsvektor $\mathbf{w} \in \mathcal{W}$ mithilfe der Daten \mathcal{D}_n eines neuronalen Netzes

$$f_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}$$

so zu lernen, dass wir eine gute Vorhersage für die Regressionsfunktion m erhalten. Um dies zu erreichen, betrachten wir eine Linearkombination gestutzter vollständig verbundener neuronaler Netze

$$f_{\mathbf{w}}(\mathbf{x}) = \sum_{k=1}^{\hat{K}_n} w_k \cdot T_{\beta_n}(f_{\mathbf{w},k,1}(\mathbf{x})) \quad (4.1)$$

mit Gewichten $w_1, \dots, w_{\widehat{K}_n} \in \mathbb{R}$. Hierbei wird durch $f_{\mathbf{w},k,1}(\mathbf{x})$ die Ausgabe des k -ten vollständig verbundenen neuronalen Netzes bezeichnet. Diese ist gegeben durch

$$f_{\mathbf{w},k,1}(\mathbf{x}) = \sum_{j=1}^r w_{k,1,j}^{(L)} \cdot f_{\mathbf{w},k,j}^{(L)}(\mathbf{x}) \quad (4.2)$$

für Gewichte $w_{k,1,1}^{(L)}, \dots, w_{k,1,r}^{(L)} \in \mathbb{R}$, wobei $f_{\mathbf{w},k,i}^{(L)}(\mathbf{x})$ rekursiv definiert ist durch

$$f_{\mathbf{w},k,i}^{(l)}(\mathbf{x}) = \sigma \left(\sum_{j=1}^r w_{k,i,j}^{(l-1)} \cdot f_{\mathbf{w},k,j}^{(l-1)}(\mathbf{x}) + w_{k,i,0}^{(l-1)} \right) \quad (4.3)$$

für Gewichte $w_{k,i,0}^{(l-1)}, \dots, w_{k,i,r}^{(l-1)} \in \mathbb{R}$ und $l = 2, \dots, L$ sowie

$$f_{\mathbf{w},k,i}^{(1)}(\mathbf{x}) = \sigma \left(\sum_{j=1}^d w_{k,i,j}^{(0)} \cdot x^{(j)} + w_{k,i,0}^{(0)} \right) \quad (4.4)$$

für Gewichte $w_{k,i,0}^{(0)}, \dots, w_{k,i,d}^{(0)} \in \mathbb{R}$.

Für die Optimierung der Gewichte nutzen wir wieder den Gradientenabstieg bezüglich des empirischen L_2 -Risikos

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{\rho=1}^n |f_{\mathbf{w}}(\mathbf{X}_\rho) - Y_\rho|^2.$$

Zur Initialisierung wählen wir die inneren Gewichte unabhängig voneinander und gleichverteilt aus einer Menge $\mathcal{W}^0 \subseteq \mathcal{W}$. Anschließend setzen wir die äußeren Gewichte

$$(\mathbf{w}^{(0)})_k = 0 \quad \text{für } k = 1, \dots, \widehat{K}_n.$$

Darüber hinaus benötigen wir für die Analyse des Gradientenabstiegs je eine zusätzliche Bedingung für die Gewichte der Ausgabeschicht sowie die Gewichte der inneren Schichten. Sei A hierfür die Menge aller Gewichtsvektoren $(w_k)_{k=1, \dots, \widehat{K}_n}$, welche

$$\sum_{k=1}^{\widehat{K}_n} |w_k| \leq 1 \quad (4.5)$$

erfüllen. Die Menge B sei die Teilmenge aller Gewichtsvektoren, welche die Bedingung

$$\|(w_{k,i,j}^{(l)})_{k,i,j,l} - ((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{k,i,j,l}\| \leq c_{119} \quad (4.6)$$

für eine Konstante $c_{119} > 0$ erfüllen. Die inneren Gewichte sind also höchstens eine Konstante c_{119} von den Startgewichten entfernt. Durch die Anwendung des Gradientenabstiegs ergibt sich eine Folge von Gewichten. Diese wird für die Gewichte der Ausgabeschicht und die der anderen Schichten getrennt bestimmt. Daher erhalten wir für eine Schrittweite $\lambda_n > 0$ die folgende Vorschrift

$$\left((\mathbf{w}^{(t+1)})_k \right)_{k=1, \dots, \widehat{K}_n} = Proj_A \left(\left((\mathbf{w}^{(t)})_k - \lambda_n \cdot \frac{\partial F_n(\mathbf{w}^{(t)})}{\partial w_k} \right)_{k=1, \dots, \widehat{K}_n} \right)$$

und

$$\left((\mathbf{w}^{(t+1)})_{k,i,j}^{(l)} \right)_{k,i,j,l} = Proj_B \left(\left((\mathbf{w}^{(t)})_{k,i,j}^{(l)} - \lambda_n \cdot \frac{\partial F_n(\mathbf{w}^{(t)})}{\partial w_{k,i,j}^{(l)}} \right)_{k,i,j,l} \right).$$

Hierbei ist die partielle Ableitung der Verlustfunktion F_n bezüglich der äußeren Gewichte w_k definiert durch

$$\frac{\partial F_n(\mathbf{w}^{(t)})}{\partial w_k} = \frac{2}{n} \sum_{\rho=1}^n (f_{\mathbf{w}^{(t)}}(\mathbf{X}_\rho) - Y_\rho) \cdot T_{\beta_n}(f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho))$$

mit $k \in \{1, \dots, \widehat{K}_n\}$. Die partiellen Ableitungen der Verlustfunktion F_n bezüglich der Gewichte $w_{k,1,j}^{(L)}$ für $k \in \{1, \dots, \widehat{K}_n\}$, $j \in \{1, \dots, r\}$ sind gegeben durch

$$\frac{\partial F_n(\mathbf{w}^{(t)})}{\partial w_{k,1,j}^{(L)}} = \frac{2}{n} \sum_{\rho=1}^n (f_{\mathbf{w}^{(t)}}(\mathbf{X}_\rho) - Y_\rho) \cdot (\mathbf{w}^{(t)})_k \cdot \frac{\partial T_{\beta_n}(f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho))}{\partial w_{k,1,j}^{(L)}}$$

und die partiellen Ableitungen der Verlustfunktion F_n bezüglich der restlichen Gewichte $w_{k,i,j}^{(l)}$ für $k \in \{1, \dots, \widehat{K}_n\}$, $i, j \in \{1, \dots, r\}$, $l \in \{1, \dots, L-1\}$ sind definiert durch

$$\frac{\partial F_n(\mathbf{w}^{(t)})}{\partial w_{k,i,j}^{(l)}} = \frac{2}{n} \sum_{\rho=1}^n (f_{\mathbf{w}^{(t)}}(\mathbf{X}_\rho) - Y_\rho) \cdot (\mathbf{w}^{(t)})_k \cdot \frac{\partial T_{\beta_n}(f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho))}{\partial w_{k,i,j}^{(l)}}.$$

Hierbei ist zu beachten, dass für $l = 0$ lediglich $j \in \{1, \dots, d\}$ gilt.

Da

$$T_{\beta_n}(z) = \sigma(2 \cdot \beta_n - \sigma((-1) \cdot z + \beta_n)) - \beta_n$$

für $z \in \mathbb{R}$ gilt, impliziert die Kettenregel, dass

$$\begin{aligned} \frac{\partial T_{\beta_n}(f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho))}{\partial w_{k,i,j}^{(l)}} &= \frac{\partial \left[\sigma(2 \cdot \beta_n - \sigma((-1) \cdot f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho) + \beta_n)) - \beta_n \right]}{\partial w_{k,i,j}^{(l)}} \\ &= \sigma' \left(2 \cdot \beta_n - \sigma((-1) \cdot f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho) + \beta_n) \right) \\ &\quad \cdot (-1) \cdot \sigma' \left((-1) \cdot f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho) + \beta_n \right) \cdot (-1) \cdot \frac{\partial f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho)}{\partial w_{k,i,j}^{(l)}} \end{aligned}$$

gilt. Hierbei ist

$$\begin{aligned} &\frac{\partial f_{\mathbf{w}^{(t)},k,1}(\mathbf{X}_\rho)}{\partial w_{k,i,j}^{(l)}} \\ &= \sum_{s_{l+2}=1}^r \cdots \sum_{s_L=1}^r f_{\mathbf{w}^{(t)},k,j}^{(l)}(\mathbf{X}_\rho) \cdot \sigma' \left(\sum_{s=1}^r (\mathbf{w}^{(t)})_{k,i,s}^{(l)} \cdot f_{\mathbf{w}^{(t)},k,s}^{(l)}(\mathbf{X}_\rho) + (\mathbf{w}^{(t)})_{k,i,0}^{(l)} \right) \\ &\quad \cdot (\mathbf{w}^{(t)})_{k,s_{l+2},i}^{(l+1)} \cdot \sigma' \left(\sum_{s=1}^r (\mathbf{w}^{(t)})_{k,s_{l+2},s}^{(l+1)} \cdot f_{\mathbf{w}^{(t)},k,s}^{(l+1)}(\mathbf{X}_\rho) + (\mathbf{w}^{(t)})_{k,s_{l+2},0}^{(l+1)} \right) \cdot (\mathbf{w}^{(t)})_{k,s_{l+3},s_{l+2}}^{(l+2)} \end{aligned}$$

$$\begin{aligned}
& \cdot \sigma' \left(\sum_{s=1}^r (\mathbf{w}^{(t)})_{k,s_{l+3},s}^{(l+2)} \cdot f_{\mathbf{w}^{(t)},k,s}^{(l+2)}(\mathbf{X}_\rho) + (\mathbf{w}^{(t)})_{k,s_{l+3},0}^{(l+2)} \right) \cdots (\mathbf{w}^{(t)})_{k,s_{L-1},s_{L-2}}^{(L-2)} \\
& \cdot \sigma' \left(\sum_{s=1}^r (\mathbf{w}^{(t)})_{k,s_{L-2},s}^{(L-2)} \cdot f_{\mathbf{w}^{(t)},k,s}^{(L-2)}(\mathbf{X}_\rho) + (\mathbf{w}^{(t)})_{k,s_{L-1},0}^{(L-2)} \right) \cdot (\mathbf{w}^{(t)})_{k,s_L,s_{L-1}}^{(L-1)} \\
& \cdot \sigma' \left(\sum_{s=1}^r (\mathbf{w}^{(t)})_{k,s_L,s}^{(L-1)} \cdot f_{\mathbf{w}^{(t)},k,s}^{(L-1)}(\mathbf{X}_\rho) + (\mathbf{w}^{(t)})_{k,s_L,0}^{(L-1)} \right) \cdot (\mathbf{w}^{(t)})_{k,1,s_L}^{(L)},
\end{aligned}$$

wobei

$$f_{\mathbf{w}^{(t)},k,j}^{(0)}(\mathbf{x}) = \begin{cases} x^{(j)} & \text{für } j \in \{1, \dots, d\} \\ 1 & \text{für } j = 0 \end{cases}$$

sowie

$$f_{\mathbf{w}^{(t)},k,0}^{(l)}(\mathbf{x}) = 1 \quad \text{für } l \in \{1, \dots, L\}.$$

Die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$ ist an der Stelle $z = 0$ nicht differenzierbar. Daher ist der Gradientenabstieg nicht wohldefiniert, wenn wir die Ableitung bei $z = 0$ betrachten. Aus diesem Grund verwenden wir einen Subgradienten der konvexen Funktion σ an der Stelle 0 und setzen $\sigma'(0) = 0$.

Wir setzen $L = L_n$ sowie $r = r_n$ und wählen als Schätzer ein neuronales Netz in der Form von (4.1), dessen Gewichtsvektor innerhalb der t_n Gradientenschritte das geringste empirische L_2 -Risiko erzielt hat. Dieser Schätzer ist dann definiert durch

$$m_n(\mathbf{x}) = f_{\mathbf{w}^{(\hat{t})}}(\mathbf{x}), \quad (4.7)$$

wobei

$$\hat{t} = \arg \min_{t \in \{0, 1, \dots, t_n\}} \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t)}}(\mathbf{X}_i) - Y_i|^2.$$

In diesem Kapitel bezeichnen wir mit $(w_{k,i,j}^{(l)})_{i,j,l}$ den Gewichtsvektor, der alle Gewichte $w_{k,i,j}^{(l)}$ enthält, die dem k -ten vollständig verbundenen neuronalen Netzwerk entsprechen.

4.2 Ein allgemeines Resultat zur Fehlerschranke des Schätzers

Das folgende Theorem ist ein zentraler Bestandteil für den Nachweis einer Dimensionsreduktion von überparametrisierten tiefen Neuronale-Netze-Schätzern. Es liefert eine Fehlerschranke für den erwarteten L_2 -Fehler des Schätzers, die es uns ermöglicht, die Konvergenzrate für den Schätzer herzuleiten. Das Besondere hierbei ist, dass trotz Überparametrisierung die Gewichte eines einzelnen tiefen neuronalen Netzes durch den Trainingsprozess an die zugrunde liegende Datenstruktur angepasst werden und die Überparametrisierung kontrollierbar bleibt.

Theorem 5. Sei $n \in \mathbb{N}$ und seien $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängig und identisch verteilte Zufallsvariablen mit Werten in $\mathbb{R}^d \times \mathbb{R}$ und

$$\mathbf{E} \{ \exp(c_{120} \cdot Y^2) \} < \infty \quad (4.8)$$

für eine Konstante $c_{120} > 0$. Sei $m : \mathbb{R}^d \rightarrow \mathbb{R}$ die zugehörige Regressionsfunktion, welche beschränkt ist. Seien weiter $I_n, L_n, N_n, t_n \in \mathbb{N}$. Im Folgenden setzen wir

$$\lambda_n = \frac{1}{t_n} \quad \text{und} \quad \hat{K}_n = N_n \cdot I_n.$$

Des Weiteren seien $\mathcal{W}^* \subseteq \mathcal{W}^0$ und

$$\overline{\mathcal{W}} := \left\{ (w_{1,i,j}^{(l)})_{i,j,l} \in \mathcal{W} : \inf_{(v_{1,i,j}^{(l)})_{i,j,l} \in \mathcal{W}^0} \left\| (w_{1,i,j}^{(l)})_{i,j,l} - (v_{1,i,j}^{(l)})_{i,j,l} \right\| \leq c_{121} \right\}$$

für eine Konstante $c_{121} \geq c_{119} > 0$. Zudem sei der Schätzer m_n , wie in Abschnitt 4.1 eingeführt, mit $\beta_n = c_5 \cdot \log n$ definiert. Angenommen, für $C_n, D_n > 0$ gelte

$$\|(\nabla_{(w_k)_{k=1,\dots,\hat{K}_n}} F_n)(\mathbf{w})\| \leq C_n \quad (4.9)$$

sofern $Y_1, \dots, Y_n \in [-\beta_n, \beta_n]$,

$$\begin{aligned} \varepsilon_n &= \mathbf{P}\{((\mathbf{w}^{(0)})_{1,i,j}^{(l)})_{i,j,l} \in \mathcal{W}^*\}, \\ N_n \cdot (1 - \varepsilon_n)^{I_n} &\leq \frac{1}{n} \end{aligned} \quad (4.10)$$

sowie

$$|f_{\mathbf{w},1,1}(\mathbf{x}) - f_{\mathbf{v},1,1}(\mathbf{x})| \leq D_n \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (v_{1,i,j}^{(l)})_{i,j,l}\| \quad (4.11)$$

für alle $(w_{1,i,j}^{(l)})_{i,j,l}, (v_{1,i,j}^{(l)})_{i,j,l} \in \overline{\mathcal{W}}$ und $\mathbf{x} \in \mathbb{R}^d$. Dann ist

$$\begin{aligned} &\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &\leq c_{122} \cdot \left(\frac{(\log n)^2}{n} + \beta_n \cdot \mathbf{E} \left\{ \sup_{\mathbf{w} \in \overline{\mathcal{W}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right| \right\} \right. \\ &\quad \left. + \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*,1,1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} + \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{N_n} + \frac{C_n^2}{t_n} \right), \end{aligned}$$

wobei wir mit $\varepsilon_1, \dots, \varepsilon_n$ die sogenannten Rademacher Zufallsvariablen bezeichnen, welche unabhängig und gleichverteilt auf der Menge $\{-1, 1\}$ sind.

Da auch die Gewichte dieses Schätzers mithilfe des Gradientenabstiegs bestimmt werden, benötigen wir eine entsprechende Aussage zur Analyse des Gradientenabstiegs. Dies führt zu folgendem Lemma:

Lemma 14. Seien $d_1, d_2 \in \mathbb{N}$ und $C_n, \tilde{D}_n \geq 0$. Zudem seien $A \subset \mathbb{R}^{d_1}$ und $B \subseteq \mathbb{R}^{d_2}$ abgeschlossene und konvexe Mengen. Sei weiter $F : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}^+$ eine Funktion, so dass

$$\mathbf{u} \mapsto F(\mathbf{u}, \mathbf{v}) \quad \text{differenzierbar und konvex für alle } \mathbf{v} \in \mathbb{R}^{d_2}$$

ist und

$$\|\nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v})\| \leq C_n \quad (4.12)$$

für alle $(\mathbf{u}, \mathbf{v}) \in A \times B$ gilt. Des Weiteren seien $(\mathbf{u}_0, \mathbf{v}_0) \in A \times B$ und $\mathbf{v}_1, \dots, \mathbf{v}_{t_n} \in B$. Wir setzen

$$\mathbf{u}_{t+1} = \text{Proj}_A(\mathbf{u}_t - \lambda_n \cdot (\nabla_{\mathbf{u}} F)(\mathbf{u}_t, \mathbf{v}_t)) \quad \text{für } t = 1, \dots, t_n - 1$$

mit $\lambda_n = \frac{1}{t_n}$.

Angenommen, für $\mathbf{u}^* \in A$ ist die Ungleichung

$$|F(\mathbf{u}^*, \mathbf{v}_t) - F(\mathbf{u}^*, \mathbf{v}_0)| \leq \tilde{D}_n \cdot \|\mathbf{u}^*\| \cdot \|\mathbf{v}_t - \mathbf{v}_0\| \quad \text{für alle } t = 1, \dots, t_n \quad (4.13)$$

erfüllt, dann gilt

$$\min_{t=0, \dots, t_n} F(\mathbf{u}_t, \mathbf{v}_t) \leq F(\mathbf{u}^*, \mathbf{v}_0) + \tilde{D}_n \cdot \|\mathbf{u}^*\| \cdot \text{diam}(B) + \frac{\|\mathbf{u}^* - \mathbf{u}_0\|^2}{2} + \frac{C_n^2}{2 \cdot t_n}.$$

Beweis. Die Aussage des Lemmas lässt sich durch eine leichte Abwandlung des Beweises von Lemma 1 zeigen. Für einen detaillierten Beweis siehe Lemma 1 in Kohler und Krzyżak (2023). \square

Bevor wir den Beweis von Theorem 5 ausführen, möchten wir die zentralen Komponenten des Beweises darstellen. Zunächst definieren wir das Ereignis A_n , welches sicherstellt, dass bestimmte Bedingungen für die Gewichtsvektoren sowie die Zufallsvariablen Y_i erfüllt sind. Anschließend zerlegen wir den L_2 -Fehler in mehrere Terme, die wir separat analysieren. Dabei lässt sich zeigen, dass das Gegenereignis von A_n mit einer geringen Wahrscheinlichkeit eintritt.

Darüber hinaus kontrollieren wir den Generalisierungsfehler des Schätzers mithilfe der Rademacher-Komplexität. Der dritte Summand der Fehlerschranke misst den Approximationsfehler des vollständig verbundenen neuronalen Netzes, welches die Regressionsfunktion hinreichend genau approximiert. Der letzte Term ergibt durch die Analyse des Gradientenabstiegs und ermöglicht es uns, den Optimierungsfehler kontrollieren zu können.

Nachdem wir nun die Hauptidee skizziert haben, fahren wir mit dem eigentlichen Beweis des Theorems fort.

Beweis von Theorem 5. Sei A_n das Ereignis, bei dem erstens der Gewichtsvektor $\mathbf{w}^{(0)}$ die Bedingung

$$((\mathbf{w}^{(0)})^{(l)})_{k_s, i, j}^{(l)} \in \mathcal{W}^* \quad (4.14)$$

für paarweise verschiedene $k_s \in \{1, \dots, \hat{K}_n\}$ mit $s \in \{1, \dots, N_n\}$ erfüllt sowie die Bedingung

$$\max_{i=1, \dots, n} |Y_i| \leq \beta_n$$

gilt.

Auf dem Ereignis A_n definieren wir den Gewichtsvektor $(\mathbf{w}^{(t)})^*$ durch

$$((\mathbf{w}^{(t)})^*)_{k_s, i, j}^{(l)} = (\mathbf{w}^{(t)})_{k_s, i, j}^{(l)}$$

für alle Schichten $l \in \{0, \dots, L_n\}$. Des Weiteren definieren wir die Gewichte der Ausgabeschicht durch

$$((\mathbf{w}^{(t)})^*)_{k_s} = \frac{1}{N_n} \quad \text{für alle } s \in \{1, \dots, N_n\}$$

sowie

$$((\mathbf{w}^{(t)})^*)_k = 0 \quad \text{für } k \notin \{k_s : s = 1, \dots, N_n\}.$$

Wir beginnen den Beweis, indem wir den L_2 -Fehler des Schätzers m_n wie folgt zerlegen

$$\begin{aligned}
& \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\
&= (\mathbf{E} \{ |m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n \} - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \}) \cdot \mathbb{1}_{A_n} + \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbb{1}_{A_n^c} \\
&= (\mathbf{E} \{ |m_n(\mathbf{X}) - Y|^2 | \mathcal{D}_n \} - \mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \}) \cdot \mathbb{1}_{A_n} \\
&\quad + \left(\mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \cdot \mathbb{1}_{A_n} \\
&\quad + \left(\frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 \right) \cdot \mathbb{1}_{A_n} \\
&\quad + \left(\frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \cdot \mathbb{1}_{A_n} \\
&\quad + \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbb{1}_{A_n^c} \\
&=: \sum_{j=1}^5 T_{j,n}.
\end{aligned}$$

In den einzelnen Beweisschritten werden wir nun die Summanden $\mathbf{E}\{T_{j,n}\}$ für $j \in \{1, \dots, 5\}$ von oben beschränken.

Im *ersten Schritt des Beweises* weisen wir nach, dass

$$\mathbf{E}\{T_{1,n}\} \leq c_{123} \cdot \frac{\log n}{n}$$

ist. Diese Aussage ergibt sich analog zu der Abschätzung von $T_{6,n}$ im ersten Beweisschritt von Theorem 2.

Im *zweiten Schritt des Beweises* zeigen wir, dass

$$\mathbf{E}\{T_{3,n}\} = 0$$

gilt.

Da auf dem Ereignis A_n die Bedingung $\max_{i=1, \dots, n} |Y_i| \leq \beta_n$ erfüllt ist, folgt direkt

$$\begin{aligned}
\mathbf{E}\{T_{3,n}\} &= \left(\frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 \right) \cdot \mathbb{1}_{A_n} \\
&= \left(\frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \cdot \mathbb{1}_{A_n} \\
&= 0.
\end{aligned}$$

Im *dritten Schritt des Beweises* wollen wir zeigen, dass die Gegenwahrscheinlichkeit des Ereignisses A_n , welche wir mit $\mathbf{P}(A_n^c)$ bezeichnen, beschränkt ist.

Hierfür betrachten wir eine schrittweise Auswahl der Startgewichte $((\mathbf{w}^{(0)})_{k,i,j}^{(l)})_{i,j,l}$ für $k \in \{1, \dots, \hat{K}_n\}$. Die Wahrscheinlichkeit, dass sich keines der Gewichte $((\mathbf{w}^{(0)})_{1,i,j}^{(l)})_{i,j,l}, \dots, ((\mathbf{w}^{(0)})_{I_n,i,j}^{(l)})_{i,j,l}$ in der Menge \mathcal{W}^* befindet, ist durch

$$(1 - \varepsilon_n)^{I_n}$$

gegeben. Daher ist die Wahrscheinlichkeit, dass ein $s \in \{1, \dots, N_n\}$ existiert, so dass keines der Gewichte $((\mathbf{w}^{(0)})_{(s-1) \cdot I_n + 1, i, j}^{(l)})_{i,j,l}, \dots, ((\mathbf{w}^{(0)})_{s \cdot I_n, i, j}^{(l)})_{i,j,l}$ in der Menge \mathcal{W}^* enthalten ist, von oben beschränkt durch

$$N_n \cdot (1 - \varepsilon_n)^{I_n}.$$

Aus den Voraussetzungen (4.8) und (4.10) sowie der Ungleichung von Markov erhalten wir damit

$$\begin{aligned} P(A_n^c) &\leq N_n \cdot (1 - \varepsilon_n)^{I_n} + \mathbf{P} \left\{ \max_{i=1, \dots, n} |Y_i| > \beta_n \right\} \\ &\leq \frac{1}{n} + n \cdot \frac{\mathbf{E} \{ \exp(c_{120} \cdot Y^2) \}}{\exp(c_{120} \cdot \beta_n^2)} \\ &\leq \frac{c_{124}}{n} \end{aligned}$$

für n hinreichend groß.

In den folgenden Beweisschritten wollen wir $\mathbf{E}\{T_{2,n}\}$ beschränken. Sei hierfür \mathcal{F} die Menge aller Funktionen der Form (4.1), wobei der Gewichtsvektor \mathbf{w} in $\overline{\mathcal{W}}^{\hat{K}_n}$ liegt und die Bedingung

$$\sum_{k=1}^{\hat{K}_n} |w_k| \leq 1$$

erfüllt.

Im vierten Schritt des Beweises zeigen wir zunächst, dass die Ungleichung

$$\mathbf{E}\{T_{2,n}\} \leq \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\mathbf{E} \{ |f(\mathbf{X}) - T_{\beta_n} Y|^2 \} - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\} + 4 \cdot \beta_n^2 \cdot \frac{c_{124}}{n}$$

gilt.

Aufgrund von $\mathbf{P}(A_n^c) \leq \frac{c_{124}}{n}$ sowie der Beschränktheit von m_n und $T_{\beta_n} Y_i$ für $i \in \{1, \dots, n\}$ erhalten wir

$$\begin{aligned} \mathbf{E}\{T_{2,n}\} &= \mathbf{E} \left\{ \mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right\} \\ &\quad - \mathbf{E} \left\{ \left(\mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \cdot \mathbf{1}_{A_n^c} \right\} \\ &\leq \mathbf{E} \left\{ \mathbf{E} \{ |m_n(\mathbf{X}) - T_{\beta_n} Y|^2 | \mathcal{D}_n \} - \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right\} + 4 \cdot \beta_n^2 \cdot \frac{c_{124}}{n} \\ &\leq \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\mathbf{E} \{ |f(\mathbf{X}) - T_{\beta_n} Y|^2 \} - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\} + 4 \cdot \beta_n^2 \cdot \frac{c_{124}}{n}. \end{aligned}$$

Im fünften Schritt des Beweises zeigen wir, dass der Erwartungswert

$$\mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\mathbf{E} \{ |f(\mathbf{X}) - T_{\beta_n} Y|^2 \} - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\}$$

von oben beschränkt ist durch

$$8 \cdot \beta_n \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot f(\mathbf{X}_i) \right\},$$

wobei $\varepsilon_1, \dots, \varepsilon_n$ Zufallsvariablen mit $\mathbf{P}\{\varepsilon_i = 1\} = \mathbf{P}\{\varepsilon_i = -1\} = \frac{1}{2}$ für $i \in \{1, \dots, n\}$ sind.

Um dies zu zeigen, wählen wir die Zufallsvariablen $\mathbf{X}_1, \dots, \mathbf{X}_n, \varepsilon_1, \dots, \varepsilon_n$ unabhängig voneinander. Zudem wählen wir die Zufallsvariablen $(\mathbf{X}'_1, Y'_1), \dots, (\mathbf{X}'_n, Y'_n)$ derart, dass $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), \varepsilon_1, \dots, \varepsilon_n, (\mathbf{X}'_1, Y'_1), \dots, (\mathbf{X}'_n, Y'_n)$ unabhängig und $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), (\mathbf{X}'_1, Y'_1), \dots, (\mathbf{X}'_n, Y'_n)$ identisch verteilt sind.

Ausgehend von diesem Aufbau ergibt sich

$$\begin{aligned} & \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\mathbf{E} \{ |f(\mathbf{X}) - T_{\beta_n} Y|^2 \} - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\} \\ &= \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}'_i) - T_{\beta_n} Y'_i|^2 \middle| (\mathbf{X}, Y)_1^n \right\} - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\} \\ &= \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}'_i) - T_{\beta_n} Y'_i|^2 - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \middle| (\mathbf{X}, Y)_1^n \right\} \right) \right\} \\ &\leq \mathbf{E} \left\{ \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}'_i) - T_{\beta_n} Y'_i|^2 - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \middle| (\mathbf{X}, Y)_1^n \right\} \right\} \\ &= \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}'_i) - T_{\beta_n} Y'_i|^2 - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\}. \end{aligned}$$

Die gemeinsame Verteilung der Zufallsvariablen $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), (\mathbf{X}'_1, Y'_1), \dots, (\mathbf{X}'_n, Y'_n)$ bleibt unverändert, wenn wir die Paare (\mathbf{X}_i, Y_i) und (\mathbf{X}'_i, Y'_i) zufällig vertauschen. Daher erhalten wir

$$\begin{aligned} & \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}'_i) - T_{\beta_n} Y'_i|^2 - \frac{1}{n} \sum_{i=1}^n |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\} \\ &= \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (|f(\mathbf{X}'_i) - T_{\beta_n} Y'_i|^2 - |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2) \right) \right\} \\ &\leq \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot |f(\mathbf{X}'_i) - T_{\beta_n} Y'_i|^2 \right) \right\} + \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n (-\varepsilon_i) \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\} \\ &= 2 \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (|f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2) \right) \right\}. \end{aligned}$$

Aufgrund der Unabhängigkeit der Zufallsvariablen $\varepsilon_1, \dots, \varepsilon_n$ kann der Erwartungswert berechnet werden, indem zunächst der Erwartungswert bezüglich der Zufallsvariable ε_1 bestimmt und anschließend der

Erwartungswert bezüglich aller anderen Zufallsvariablen ε_i für $i \in \{2, \dots, n\}$ berechnet wird. Als Ergebnis erhalten wir

$$\begin{aligned}
& 2 \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (|f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2) \right) \right\} \\
&= 2 \cdot \mathbf{E} \left\{ \frac{1}{2} \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \cdot |f(\mathbf{X}_1) - T_{\beta_n} Y_1|^2 \right) \right. \\
&\quad \left. + \frac{1}{2} \sup_{g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - \frac{1}{n} \cdot |g(\mathbf{X}_1) - T_{\beta_n} Y_1|^2 \right) \right\} \\
&= \mathbf{E} \left\{ \sup_{f, g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right. \right. \\
&\quad \left. \left. + \frac{1}{n} \cdot |f(\mathbf{X}_1) - T_{\beta_n} Y_1|^2 - \frac{1}{n} \cdot |g(\mathbf{X}_1) - T_{\beta_n} Y_1|^2 \right) \right\}.
\end{aligned} \tag{4.15}$$

Durch die Anwendung der dritten binomischen Formel erhalten wir

$$\begin{aligned}
& \mathbf{E} \left\{ \sup_{f, g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right. \right. \\
&\quad \left. \left. + \frac{1}{n} \cdot |f(\mathbf{X}_1) - T_{\beta_n} Y_1|^2 - \frac{1}{n} \cdot |g(\mathbf{X}_1) - T_{\beta_n} Y_1|^2 \right) \right\} \\
&= \mathbf{E} \left\{ \sup_{f, g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right. \right. \\
&\quad \left. \left. + \frac{1}{n} \cdot (f(\mathbf{X}_1) - T_{\beta_n} Y_1 + (g(\mathbf{X}_1) - T_{\beta_n} Y_1)) \cdot (f(\mathbf{X}_1) - T_{\beta_n} Y_1 - (g(\mathbf{X}_1) - T_{\beta_n} Y_1)) \right) \right\} \\
&= \mathbf{E} \left\{ \sup_{f, g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right. \right. \\
&\quad \left. \left. + \frac{1}{n} \cdot (f(\mathbf{X}_1) - 2 \cdot T_{\beta_n} Y_1 + g(\mathbf{X}_1)) \cdot (f(\mathbf{X}_1) - g(\mathbf{X}_1)) \right) \right\} \\
&\leq \mathbf{E} \left\{ \sup_{f, g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot |f(\mathbf{X}_1) - g(\mathbf{X}_1)| \right) \right\},
\end{aligned}$$

wobei der letzte Schritt daraus folgt, dass die Funktionswerte von f und g aufgrund ihrer Definition betragsmäßig durch β_n nach oben beschränkt sind.

Der Term

$$\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot |f(\mathbf{X}_1) - g(\mathbf{X}_1)|$$

ist für feste $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), \varepsilon_2, \dots, \varepsilon_n$ in f und g symmetrisch. Deshalb können wir ohne Beschränkung der Allgemeinheit annehmen, dass $f(\mathbf{X}_1) \geq g(\mathbf{X}_1)$ ist. Dies führt zu

$$\sup_{f, g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot |f(\mathbf{X}_1) - g(\mathbf{X}_1)| \right)$$

$$= \sup_{f,g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot (f(\mathbf{X}_1) - g(\mathbf{X}_1)) \right).$$

Nehmen wir nun ohne Beschränkung der Allgemeinheit an, dass $f(\mathbf{X}_1) < g(\mathbf{X}_1)$ gilt, so erhalten wir

$$\begin{aligned} & \sup_{f,g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot |f(\mathbf{X}_1) - g(\mathbf{X}_1)| \right) \\ &= \sup_{f,g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - \frac{4 \cdot \beta_n}{n} \cdot (f(\mathbf{X}_1) - g(\mathbf{X}_1)) \right). \end{aligned}$$

Damit ergibt sich

$$\begin{aligned} & \mathbf{E} \left\{ \sup_{f,g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot |f(\mathbf{X}_1) - g(\mathbf{X}_1)| \right) \right\} \\ &= \mathbf{E} \left\{ \frac{1}{2} \sup_{f,g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot (f(\mathbf{X}_1) - g(\mathbf{X}_1)) \right) \right. \\ & \quad \left. + \frac{1}{2} \sup_{f,g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 - \frac{4 \cdot \beta_n}{n} \cdot (f(\mathbf{X}_1) - g(\mathbf{X}_1)) \right) \right\} \\ &= \mathbf{E} \left\{ \sup_{f,g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot \varepsilon_1 \cdot (f(\mathbf{X}_1) - g(\mathbf{X}_1)) \right) \right\} \\ &\leq \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot \varepsilon_1 \cdot f(\mathbf{X}_1) \right) \right. \\ & \quad \left. + \sup_{g \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |g(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot (-\varepsilon_1) \cdot g(\mathbf{X}_1) \right) \right\} \\ &= 2 \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot \varepsilon_1 \cdot f(\mathbf{X}_1) \right) \right\}. \end{aligned}$$

Die letzte Gleichheit folgt daraus, dass ε_1 und $-\varepsilon_1$ die gleiche Verteilung besitzen. Mit der gleichen Argumentation können wir dieses Vorgehen sukzessive für $\varepsilon_2, \dots, \varepsilon_n$ fortsetzen und erhalten damit die Abschätzung

$$\begin{aligned} & 2 \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 \right) \right\} \\ &\leq 2 \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=2}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot \varepsilon_1 \cdot f(\mathbf{X}_1) \right) \right\} \\ &\leq 2 \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=3}^n \varepsilon_i \cdot |f(\mathbf{X}_i) - T_{\beta_n} Y_i|^2 + \frac{4 \cdot \beta_n}{n} \cdot \varepsilon_1 \cdot f(\mathbf{X}_1) + \frac{4 \cdot \beta_n}{n} \cdot \varepsilon_2 \cdot f(\mathbf{X}_2) \right) \right\} \\ &\leq \dots \\ &\leq 2 \cdot \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \frac{4 \cdot \beta_n}{n} \sum_{i=1}^n \varepsilon_i \cdot f(\mathbf{X}_i) \right\}. \end{aligned}$$

Im sechsten Schritt des Beweises zeigen wir, dass

$$\mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot f(\mathbf{X}_i) \right\} \leq \mathbf{E} \left\{ \sup_{\mathbf{w} \in \overline{\mathcal{W}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right| \right\}$$

gilt. Sei hierfür A die Menge aller Gewichtsvektoren $(w_k)_{k=1, \dots, \widehat{K}_n}$, die die Bedingung $\sum_{k=1}^{\widehat{K}_n} |w_k| \leq 1$ erfüllen. Dann ist

$$\begin{aligned} \mathbf{E} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot f(\mathbf{X}_i) \right\} &\leq \mathbf{E} \left\{ \sup_{(w_k)_{k=1, \dots, \widehat{K}_n} \in A} \sup_{\mathbf{w} \in \overline{\mathcal{W}}^{\widehat{K}_n}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \sum_{k=1}^{\widehat{K}_n} w_k \cdot (T_{\beta_n} f_{\mathbf{w},k,1})(\mathbf{X}_i) \right\} \\ &= \mathbf{E} \left\{ \sup_{(w_k)_{k=1, \dots, \widehat{K}_n} \in A} \sup_{\mathbf{w} \in \overline{\mathcal{W}}^{\widehat{K}_n}} \sum_{k=1}^{\widehat{K}_n} w_k \cdot \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},k,1})(\mathbf{X}_i) \right\} \\ &\leq \mathbf{E} \left\{ \sup_{(w_k)_{k=1, \dots, \widehat{K}_n} \in A} \sup_{\mathbf{w} \in \overline{\mathcal{W}}^{\widehat{K}_n}} \sum_{k=1}^{\widehat{K}_n} |w_k| \cdot \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},k,1})(\mathbf{X}_i) \right| \right\} \\ &\leq \mathbf{E} \left\{ \sup_{(w_k)_{k=1, \dots, \widehat{K}_n} \in A} \sum_{k=1}^{\widehat{K}_n} |w_k| \cdot \sup_{\mathbf{w} \in \overline{\mathcal{W}}^{\widehat{K}_n}, k \in \{1, \dots, \widehat{K}_n\}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},k,1})(\mathbf{X}_i) \right| \right\} \\ &\leq \mathbf{E} \left\{ \sup_{\mathbf{w} \in \overline{\mathcal{W}}^{\widehat{K}_n}, k \in \{1, \dots, \widehat{K}_n\}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},k,1})(\mathbf{X}_i) \right| \right\} \\ &= \mathbf{E} \left\{ \sup_{\mathbf{w} \in \overline{\mathcal{W}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right| \right\}. \end{aligned}$$

Die letzte Gleichheit ergibt sich dabei aus der Tatsache, dass

$$\left\{ (T_{\beta_n} f_{\mathbf{w},k,1}) : \mathbf{w} \in \overline{\mathcal{W}}^{\widehat{K}_n}, k \in \{1, \dots, \widehat{K}_n\} \right\} = \left\{ (T_{\beta_n} f_{\mathbf{w},1,1}) : \mathbf{w} \in \overline{\mathcal{W}} \right\}$$

gilt.

Fassen wir die Beweisschritte vier bis sechs zusammen, so erhalten wir

$$\mathbf{E} \{T_{2,n}\} \leq 8 \cdot \beta_n \cdot \mathbf{E} \left\{ \sup_{\mathbf{w} \in \overline{\mathcal{W}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right| \right\} + 4 \cdot \beta_n^2 \cdot \frac{c_{124}}{n}.$$

Im siebten Schritt des Beweises wollen wir zeigen, dass die Voraussetzungen (4.12) und (4.13) in Lemma 14 erfüllt sind, sofern das Ereignis A_n eintritt.

Die Gültigkeit von Bedingung (4.12) in Lemma 14 folgt direkt aus Voraussetzung (4.9) in Theorem 5. Es bleibt also noch zu zeigen, dass

$$\begin{aligned} &\left| F_n((\mathbf{w}^{(t)})^*) - F_n((\mathbf{w}^{(0)})^*) \right| \\ &\leq \tilde{D}_n \cdot \left\| ((\mathbf{w}^{(t)})^*)_{k=1, \dots, \widehat{K}_n} \right\| \cdot \left\| (((\mathbf{w}^{(t)})^*)_{k,i,j}^{(l)})_{k,i,j,l} - (((\mathbf{w}^{(0)})^*)_{k,i,j}^{(l)})_{k,i,j,l} \right\| \end{aligned}$$

gilt, sofern das Ereignis A_n eintritt.

Durch die Anwendung der dritten binomischen Formel sowie der Ungleichung von Cauchy-Schwarz folgt, dass

$$\begin{aligned} & \left| F_n((\mathbf{w}^{(t)})^*) - F_n((\mathbf{w}^{(0)})^*) \right| \\ &= \frac{1}{n} \sum_{i=1}^n \left| f_{(\mathbf{w}^{(t)})^*}(\mathbf{X}_i) - Y_i \right|^2 - \frac{1}{n} \sum_{i=1}^n \left| f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i \right|^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left| \left(f_{(\mathbf{w}^{(t)})^*}(\mathbf{X}_i) - Y_i + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i \right) \left(f_{(\mathbf{w}^{(t)})^*}(\mathbf{X}_i) - Y_i - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) + Y_i \right) \right|. \end{aligned}$$

Aufgrund der Definition des neuronalen Netzes in (4.1) und der Projektion der äußeren Gewichte sind die Netzwerke $f_{(\mathbf{w}^{(t)})^*}$ und $f_{(\mathbf{w}^{(0)})^*}$ absolut durch β_n beschränkt. Zusätzlich gilt auf dem Ereignis A_n

$$\max_{i=1, \dots, n} |Y_i| \leq \beta_n,$$

wodurch wir

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left| \left(f_{(\mathbf{w}^{(t)})^*}(\mathbf{X}_i) - Y_i + f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) - Y_i \right) \left(f_{(\mathbf{w}^{(t)})^*}(\mathbf{X}_i) - Y_i - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) + Y_i \right) \right| \\ & \leq 4 \cdot \beta_n \cdot \frac{1}{n} \sum_{i=1}^n \left| f_{(\mathbf{w}^{(t)})^*}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*}(\mathbf{X}_i) \right| \\ & \leq 4 \cdot \beta_n \cdot \frac{1}{n} \sum_{i=1}^n \sqrt{\sum_{k=1}^{\widehat{K}_n} |((\mathbf{w}^{(t)})^*)_k|^2} \cdot \sqrt{\sum_{k=1}^{\widehat{K}_n} \left| f_{(\mathbf{w}^{(t)})^*,k,1}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}(\mathbf{X}_i) \right|^2} \end{aligned}$$

erhalten. Wegen Voraussetzung (4.11) folgt hieraus

$$\begin{aligned} & 4 \cdot \beta_n \cdot \frac{1}{n} \sum_{i=1}^n \sqrt{\sum_{k=1}^{\widehat{K}_n} |((\mathbf{w}^{(t)})^*)_k|^2} \cdot \sqrt{\sum_{k=1}^{\widehat{K}_n} \left| f_{(\mathbf{w}^{(t)})^*,k,1}(\mathbf{X}_i) - f_{(\mathbf{w}^{(0)})^*,k,1}(\mathbf{X}_i) \right|^2} \\ & \leq 4 \cdot \beta_n \cdot \left\| (((\mathbf{w}^{(t)})^*)_k)_{k=1, \dots, \widehat{K}_n} \right\| \cdot D_n \cdot \sqrt{\sum_{k=1}^{\widehat{K}_n} \left\| (((\mathbf{w}^{(t)})^*)_{k,i,j}^{(l)})_{i,j,l} - (((\mathbf{w}^{(0)})^*)_{k,i,j}^{(l)})_{i,j,l} \right\|^2} \\ & = 4 \cdot \beta_n \cdot D_n \cdot \left\| (((\mathbf{w}^{(t)})^*)_k)_{k=1, \dots, \widehat{K}_n} \right\| \cdot \left\| (((\mathbf{w}^{(t)})^*)_{k,i,j}^{(l)})_{k,i,j,l} - (((\mathbf{w}^{(0)})^*)_{k,i,j}^{(l)})_{k,i,j,l} \right\| \\ & = 4 \cdot \beta_n \cdot D_n \cdot \left\| (((\mathbf{w}^{(t)})^*)_k)_{k=1, \dots, \widehat{K}_n} \right\| \cdot \left\| (((\mathbf{w}^{(t)})^*)_{k,i,j}^{(l)})_{k,i,j,l} - (((\mathbf{w}^{(0)})^*)_{k,i,j}^{(l)})_{k,i,j,l} \right\|. \end{aligned}$$

Damit ist Voraussetzung (4.13) in Lemma 14 für $\widetilde{D}_n = 4 \cdot \beta_n \cdot D_n$ erfüllt.

Im achten Schritt des Beweises wollen wir mithilfe von Lemma 14 zeigen, dass

$$\begin{aligned} \mathbf{E}\{T_{4,n}\} & \leq \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*,1,1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \\ & \quad + c_{125} \cdot \left(\frac{1}{n} + \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{2 \cdot N_n} + \frac{C_n^2}{2 \cdot t_n} \right) \end{aligned}$$

für eine Konstante $c_{125} > 0$ gilt.

Aus der Definition des Schätzers m_n ergibt sich

$$\begin{aligned} \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 \cdot \mathbb{1}_{A_n} \right\} &= \mathbf{E} \left\{ \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t)}}(\mathbf{X}_i) - Y_i|^2 \cdot \mathbb{1}_{A_n} \right\} \\ &= \mathbf{E} \left\{ \min_{t=1, \dots, t_n} \frac{1}{n} \sum_{i=1}^n |f_{\mathbf{w}^{(t)}}(\mathbf{X}_i) - Y_i|^2 \cdot \mathbb{1}_{A_n} \right\} \\ &= \mathbf{E} \left\{ \min_{t=0, \dots, t_n} F_n(\mathbf{w}^{(t)}) \cdot \mathbb{1}_{A_n} \right\}. \end{aligned}$$

Die Anwendung von Lemma 14 zusammen mit der Projektion der inneren Gewichte sowie der Definition von $((\mathbf{w}^{(t)})^*)_k$ führt zu

$$\min_{t=0, \dots, t_n} F_n(\mathbf{w}^{(t)}) \leq F_n((\mathbf{w}^{(0)})^*) + c_{126} \cdot \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{2 \cdot N_n} + \frac{C_n^2}{2 \cdot t_n}.$$

Hieraus können wir schließen, dass

$$\begin{aligned} \mathbf{E}\{T_{4,n}\} &= \mathbf{E} \left\{ \left(\frac{1}{n} \sum_{i=1}^n |m_n(\mathbf{X}_i) - Y_i|^2 - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \cdot \mathbb{1}_{A_n} \right\} \\ &\leq \mathbf{E} \left\{ \left(F_n((\mathbf{w}^{(0)})^*) + c_{126} \cdot \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{2 \cdot N_n} + \frac{C_n^2}{2 \cdot t_n} - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \mathbb{1}_{A_n} \right\} \\ &= \mathbf{E} \left\{ \mathbf{E} \left\{ \left(F_n((\mathbf{w}^{(0)})^*) - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \cdot \mathbb{1}_{A_n} \mid \mathbf{w}^{(0)} \right\} \right\} \\ &\quad + \mathbf{E} \left\{ \left(c_{126} \cdot \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{2 \cdot N_n} + \frac{C_n^2}{2 \cdot t_n} \right) \cdot \mathbb{1}_{A_n} \right\} \end{aligned}$$

ist. Sei nun \tilde{A}_n das Ereignis, auf welchem der Gewichtsvektor $\mathbf{w}^{(0)}$ die Bedingung

$$((\mathbf{w}^{(0)})^{(l)})_{k_s, i, j} \in \mathcal{W}^* \tag{4.16}$$

für paarweise verschiedene $k_s \in \{1, \dots, \hat{K}_n\}$ mit $s \in \{1, \dots, N_n\}$ erfüllt. Dann erhalten wir aufgrund von

$$\mathbf{P}(\tilde{A}_n) - \mathbf{P}(A_n) \leq \mathbf{P}(A_n^c) \leq c_{124} \cdot \frac{1}{n},$$

die folgende Abschätzung

$$\begin{aligned} &\mathbf{E} \left\{ \left(F_n((\mathbf{w}^{(0)})^*) - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \cdot \mathbb{1}_{A_n} \mid \mathbf{w}^{(0)} \right\} \\ &\leq \mathbf{E} \left\{ \left(F_n((\mathbf{w}^{(0)})^*) - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \cdot \mathbb{1}_{\tilde{A}_n} \mid \mathbf{w}^{(0)} \right\} + \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \cdot (\mathbf{P}(\tilde{A}_n) - \mathbf{P}(A_n)) \\ &\leq (\mathbf{E} \{ F_n((\mathbf{w}^{(0)})^*) \mid \mathbf{w}^{(0)} \} - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \}) \cdot \mathbb{1}_{\tilde{A}_n} + \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \cdot \frac{c_{124}}{n} \\ &= \int |f_{(\mathbf{w}^{(0)})^*}(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbb{1}_{\tilde{A}_n} + \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \cdot \frac{c_{124}}{n}. \end{aligned}$$

Aus

$$\mathbf{E} \{ Y^2 \} \leq \frac{1}{c_{120}} \cdot \mathbf{E} \{ \exp(c_{120} \cdot Y^2) \} < \infty$$

folgt durch die Anwendung der Ungleichung von Jensen

$$\mathbf{E} \{ |m(\mathbf{X})|^2 \} \leq \mathbf{E} \{ \mathbf{E} \{ Y^2 | \mathbf{X} \} \} = \mathbf{E} \{ Y^2 \} < \infty.$$

Somit ist

$$\begin{aligned} & \int \left| f_{(\mathbf{w}^{(0)})^*}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} + \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \cdot \frac{c_{124}}{n} \\ & \leq \int \left| f_{(\mathbf{w}^{(0)})^*}(\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} + c_{127} \cdot \frac{1}{n} \\ & = \int \left| \sum_{k=1}^{\hat{K}_n} ((\mathbf{w}^{(0)})^*)_k \cdot \left(T_{\beta_n} f_{(\mathbf{w}^{(0)})^*, k, 1} \right) (\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} + c_{127} \cdot \frac{1}{n}. \end{aligned}$$

Mithilfe der Konvexität der quadratischen Funktion erhalten wir durch erneutes Anwenden der Ungleichung von Jensen, dass

$$\begin{aligned} & \int \left| \sum_{k=1}^{\hat{K}_n} ((\mathbf{w}^{(0)})^*)_k \cdot \left(T_{\beta_n} f_{(\mathbf{w}^{(0)})^*, k, 1} \right) (\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} \\ & = \int \left| \sum_{s=1}^{N_n} \frac{1}{N_n} \cdot \left(T_{\beta_n} f_{(\mathbf{w}^{(0)})^*, k_s, 1} \right) (\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} \\ & \leq \sum_{s=1}^{N_n} \frac{1}{N_n} \int \left| \left(T_{\beta_n} f_{(\mathbf{w}^{(0)})^*, k_s, 1} \right) (\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \cdot \mathbf{1}_{\tilde{A}_n} \\ & \leq \sup_{\mathbf{w}^* \in (\mathcal{W}^*)^{\hat{K}_n}, s \in \{1, \dots, N_n\}} \int \left| \left(T_{\beta_n} f_{\mathbf{w}^*, k_s, 1} \right) (\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ & = \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int \left| \left(T_{\beta_n} f_{\mathbf{w}^*, 1, 1} \right) (\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \end{aligned}$$

gilt. Die letzte Gleichheit folgt, wie auch schon im sechsten Beweisschritt, aus der Tatsache, dass

$$\left\{ \left(T_{\beta_n} f_{\mathbf{w}^*, k, 1} \right) : \mathbf{w}^* \in (\mathcal{W}^*)^{\hat{K}_n}, k \in \{1, \dots, \hat{K}_n\} \right\} = \left\{ \left(T_{\beta_n} f_{\mathbf{w}^*, 1, 1} \right) : \mathbf{w}^* \in \mathcal{W}^* \right\}$$

ist. Daher ergibt sich

$$\begin{aligned} \mathbf{E} \{ T_{4,n} \} & \leq \mathbf{E} \left\{ \mathbf{E} \left\{ \left(F_n((\mathbf{w}^{(0)})^*) - \mathbf{E} \{ |m(\mathbf{X}) - Y|^2 \} \right) \cdot \mathbf{1}_{A_n} \middle| \mathbf{w}^{(0)} \right\} \right\} \\ & \quad + \mathbf{E} \left\{ \left(c_{126} \cdot \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{2 \cdot N_n} + \frac{C_n^2}{2 \cdot t_n} \right) \cdot \mathbf{1}_{A_n} \right\} \\ & \leq \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int \left| \left(T_{\beta_n} f_{\mathbf{w}^*, 1, 1} \right) (\mathbf{x}) - m(\mathbf{x}) \right|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \\ & \quad + c_{125} \cdot \left(\frac{1}{n} + \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{2 \cdot N_n} + \frac{C_n^2}{2 \cdot t_n} \right). \end{aligned}$$

Im neunten Schritt des Beweises zeigen wir, dass

$$\mathbf{E} \{ T_{5,n} \} \leq 4 \cdot \beta_n^2 \cdot \frac{c_{124}}{n}$$

erfüllt ist. Aufgrund der Beschränktheit von m können wir ohne Beschränkung der Allgemeinheit annehmen, dass $\|m\|_\infty \leq \beta_n$ ist. Zudem folgt aus der Definition des neuronalen Netzes

$$\|m_n\|_\infty = \|f_{\mathbf{w}^{(i)}}\|_\infty \leq \beta_n.$$

Wegen $\mathbf{P}(A_n^c) \leq \frac{c_{124}}{n}$ ergibt sich daher die Ungleichung

$$\mathbf{E} \left\{ \int |m(\mathbf{x}) - m_n(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \cdot \mathbf{P}(A_n^c) \leq 4 \cdot \beta_n^2 \cdot \frac{c_{124}}{n}.$$

Durch Zusammenfassen der Beweisschritte erhalten wir

$$\begin{aligned} & \mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\ & \leq c_{122} \cdot \left(\frac{(\log n)^2}{n} + \beta_n \cdot \mathbf{E} \left\{ \sup_{\mathbf{w} \in \overline{\mathcal{W}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right| \right\} \right. \\ & \quad \left. + \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \overline{\mathcal{W}}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*,1,1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} + \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{N_n} + \frac{C_n^2}{t_n} \right) \end{aligned}$$

und damit die Aussage des Theorems. □

4.3 Approximation hierarchischer Kompositionsmodelle durch neuronale Netze

Im Rahmen der drei Forschungsbereiche des Deep Learnings benötigen wir für die Herleitung einer Konvergenzrate für den Neuronale-Netze-Schätzer ein Resultat, mit dessen Hilfe wir den Approximationsfehler kontrollieren und eine gute Annäherung des Schätzers an die Regressionsfunktion garantieren können.

In Kapitel 1 haben wir bereits erläutert, dass die Regressionsfunktion bestimmte strukturelle Voraussetzungen erfüllen muss, um eine Konvergenzgeschwindigkeit zu erreichen, die unabhängig von der Eingabedimension d ist. Deshalb nehmen wir an, dass die Regressionsfunktion einem hierarchischen Kompositionsmodell (siehe Definition 4) genügt. In diesem Abschnitt wird beschrieben, inwieweit neuronale Netze in der Lage sind, hierarchische Kompositionsmodelle unter Berücksichtigung von geeigneten Glattheits- und Ordnungsbeschränkungen zu approximieren.

Für $\ell = 1$ und eine Ordnungs- und Glattheitsbedingung $\mathcal{P} \subseteq (0, \infty) \times \mathbb{N}$ ist der Raum der hierarchischen Kompositionsmodelle gegeben durch

$$\mathcal{H}(1, \mathcal{P}) = \left\{ h : \mathbb{R}^d \rightarrow \mathbb{R} \mid h(\mathbf{x}) = g \left(x^{(\pi(1))}, \dots, x^{(\pi(K))} \right) \text{ für alle } \mathbf{x} \in \mathbb{R}^d, \text{ wobei} \right. \\ \left. \begin{aligned} & g : \mathbb{R}^K \rightarrow \mathbb{R} \text{ eine Funktion ist, die } (p, C)\text{-glatt ist für ein } (p, K) \in \mathcal{P}, C > 0, \text{ sowie} \\ & \pi : \{1, \dots, K\} \rightarrow \{1, \dots, d\} \end{aligned} \right\}.$$

Für $\ell > 1$ ergibt sich rekursiv

$$\mathcal{H}(\ell, \mathcal{P}) := \left\{ h : \mathbb{R}^d \rightarrow \mathbb{R} \mid h(\mathbf{x}) = g(f_1(\mathbf{x}), \dots, f_K(\mathbf{x})) \text{ für alle } \mathbf{x} \in \mathbb{R}^d, \text{ wobei} \right. \\ \left. \begin{aligned} &g : \mathbb{R}^K \rightarrow \mathbb{R} \text{ eine Funktion ist, die } (p, C)\text{-glatt ist für ein } (p, K) \in \mathcal{P}, C > 0 \text{ und} \\ &f_i \in \mathcal{H}(\ell - 1, \mathcal{P}) \text{ für } i \in \{1, \dots, K\} \end{aligned} \right\}.$$

Um eine Funktion $h_1^{(\ell)}$, welche in der Klasse $\mathcal{H}(\ell, \mathcal{P})$ liegt, berechnen zu können, benötigen wir verschiedene hierarchische Kompositionsmodelle auf den Leveln $i = 1, \dots, \ell - 1$. Die Anzahl der hierarchischen Kompositionsmodelle vom Level i , die zur Berechnung von $h_1^{(\ell)}$ erforderlich ist, bezeichnen wir mit \tilde{N}_i . Hierbei ist

$$h_j^{(i)} : \mathbb{R}^d \rightarrow \mathbb{R} \quad (4.17)$$

das j -te hierarchische Kompositionsmodell eines bestimmten Levels i ($j \in \{1, \dots, \tilde{N}_i\}, i \in \{1, \dots, \ell\}$), das eine $(p_j^{(i)}, C)$ -glatte Funktion $g_j^{(i)} : \mathbb{R}^{K_j^{(i)}} \rightarrow \mathbb{R}$ verwendet, wobei $p_j^{(i)} = q_j^{(i)} + s_j^{(i)}$ für $q_j^{(i)} \in \mathbb{N}_0$ sowie $s_j^{(i)} \in (0, 1]$ mit $(p_j^{(i)}, K_j^{(i)}) \in \mathcal{P}$ ist.

Die Berechnung von $h_1^{(\ell)}(\mathbf{x})$ erfolgt rekursiv über

$$h_j^{(i)}(\mathbf{x}) = g_j^{(i)} \left(h_{\sum_{t=1}^{j-1} K_t^{(i)} + 1}^{(i-1)}(\mathbf{x}), \dots, h_{\sum_{t=1}^j K_t^{(i)}}^{(i-1)}(\mathbf{x}) \right) \quad (4.18)$$

für $j \in \{1, \dots, \tilde{N}_i\}$ sowie $i \in \{2, \dots, \ell\}$ und

$$h_j^{(1)}(\mathbf{x}) = g_j^{(1)} \left(x^{(\pi(\sum_{t=1}^{j-1} K_t^{(1)} + 1))}, \dots, x^{(\pi(\sum_{t=1}^j K_t^{(1)}))} \right) \quad (4.19)$$

für eine Funktion $\pi : \{1, \dots, \tilde{N}_1\} \rightarrow \{1, \dots, d\}$.

Des Weiteren beschreibt die Rekursion

$$\tilde{N}_\ell = 1 \quad \text{und} \quad \tilde{N}_i = \sum_{j=1}^{\tilde{N}_{i+1}} K_j^{(i+1)} \quad (4.20)$$

für $i \in \{1, \dots, \ell - 1\}$ die Anzahl aller hierarchischen Kompositionsmodelle vom Level i .

In Abbildung 4.1 ist die beispielhafte Struktur eines hierarchischen Kompositionsmodells vom Level $\ell = 2$ mit Ordnungs- und Glattheitsbedingung \mathcal{P} dargestellt. Sie verdeutlicht, dass sich die Funktion $h_1^{(2)} \in \mathcal{H}(2, \mathcal{P})$ als Komposition von Funktionen vorheriger Level darstellen lässt. In diesem Beispiel ergibt sich

$$h_1^{(2)}(\mathbf{x}) = g_1^{(2)}(h_1^{(1)}(\mathbf{x}), h_2^{(1)}(\mathbf{x}))$$

mit

$$\begin{aligned} h_1^{(1)}(\mathbf{x}) &= g_1^{(1)}(x^{(\pi(1))}, x^{(\pi(2))}, x^{(\pi(3))}) \\ h_2^{(1)}(\mathbf{x}) &= g_2^{(1)}(x^{(\pi(4))}, x^{(\pi(5))}) \end{aligned}$$

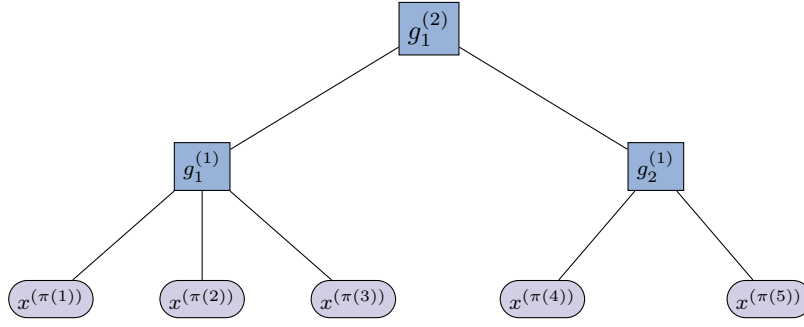


Abbildung 4.1: Hierarchisches Kompositionsmodell der Klasse $\mathcal{H}(2, \mathcal{P})$

sowie

$$\tilde{N}_2 = 1, \quad \tilde{N}_1 = 2, \quad K_1^{(1)} = 3, \quad K_2^{(1)} = 2 \quad \text{und} \quad K_1^{(2)} = 2.$$

Das folgende Resultat zeigt, dass tiefe vollständig verbundene neuronale Netze Funktionen aus der Klasse der hierarchischen Kompositionsmodelle approximieren können, sofern die Anzahl der verdeckten Schichten L und die Anzahl der Neuronen pro Schicht r geeigneten Voraussetzungen genügen.

Lemma 15. Sei \mathbf{X} eine \mathbb{R}^d -wertige Zufallsvariable und $m : \mathbb{R}^d \rightarrow \mathbb{R}$ eine Funktion, welche für $\ell \in \mathbb{N}$ und $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$ in der Klasse $\mathcal{H}(\ell, \mathcal{P})$ enthalten ist. Sei \tilde{N}_i definiert wie in (4.20). Des Weiteren bestehe jede Funktion $m = h_1^{(\ell)}$ aus verschiedenen Funktionen $h_j^{(i)}$ für $j \in \{1, \dots, \tilde{N}_i\}$ und $i \in \{1, \dots, \ell\}$, welche wie in (4.17), (4.18) und (4.19) definiert sind. Die zugehörigen Funktionen $g_j^{(i)}$ seien Lipschitz-stetig bezüglich der $\|\cdot\|_1$ -Norm mit einer Lipschitz-Konstante $C_{\text{Lip}} \geq 1$ und erfüllen zudem

$$\|g_j^{(i)}\|_{C^{q_j^{(i)}}(\mathbb{R}^d)} \leq c_{128}$$

für eine Konstante $c_{128} > 0$. Wir bezeichnen mit $K_{\max} = \max_{i,j} K_j^{(i)} < \infty$ die maximale Eingabedimension und mit $p_{\max} = \max_{i,j} p_j^{(i)} < \infty$ die maximale Glattheit der Funktionen $g_j^{(i)}$. Sei zusätzlich $a \geq 1$ und $M_{j,i} \in \mathbb{N}$ hinreichend groß, wobei jedes $M_{j,i}$ unabhängig von der Größe von a ist. Es sei jedoch

$$\min_{j,i} M_{j,i}^2 > \frac{c_{129} \cdot a^{4(p_{\max}+1)}}{(2^\ell \cdot K_{\max} \cdot C_{\text{Lip}})^\ell}$$

für eine hinreichend große Konstante $c_{129} \geq 1$ erfüllt.

Für $L, r \in \mathbb{N}$ mit

$$(i) \quad L \geq \ell \cdot \left(5 + \left\lceil \log_4 \left(\max_{j,i} M_{j,i}^{2p_j^{(i)}} \right) \right\rceil \cdot (\lceil \log_2(\max\{K_{\max}, p_{\max}\} + 1) \rceil + 1) \right)$$

$$(ii) \quad r \geq \max_{i \in \{1, \dots, \ell\}} \sum_{j=1}^{\tilde{N}_i} 2^{K_j^{(i)}} \cdot 64 \cdot \binom{K_j^{(i)} + q_j^{(i)}}{K_j^{(i)}} \cdot (K_j^{(i)})^2 \cdot (q_j^{(i)} + 1) \cdot M_{j,i}^{K_j^{(i)}}$$

existiert dann ein neuronales Netz t_1 der Netzwerkklassse $\mathcal{F}(L, r)$ mit Gewichten, die durch

$$\max_{j,i} M_{j,i}^{4p_j^{(i)}+4}$$

absolut beschränkt sind, so dass die Ungleichung

$$\|t_1 - m\|_{\infty, [-a, a]^d} \leq c_{130} \cdot a^{4(p_{\max}+1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \quad (4.21)$$

für eine Konstante $c_{130} > 0$ erfüllt ist.

Da sich gemäß Definition 3 eine Funktion, die einem hierarchischen Kompositionsmodell genügt, aus verschiedenen (p, C) -glatte Funktionen zusammensetzt, benötigen wir für den Beweis von Lemma 15 ein weiteres Resultat, mit dessen Hilfe wir zeigen können, dass vollständig verbundene neuronale Netze (p, C) -glatte Funktionen gut approximieren können.

Approximation (p, C) -glatte Funktionen durch neuronale Netze

Lemma 16. Sei $d \in \mathbb{N}$. Sei weiter $f : \mathbb{R}^d \rightarrow \mathbb{R}$ eine (p, C) -glatte Funktion mit $p = q + s$ für ein $q \in \mathbb{N}_0$ und ein $s \in (0, 1]$ sowie $C > 0$. Zudem sei $a \geq 1$ und $M \in \mathbb{N}$ hinreichend groß sowie unabhängig von der Größe von a , aber es soll

$$M \geq 2, \quad M^{2p} \geq 2 \cdot \exp(2ad) \cdot \max \left\{ \|f\|_{C^q([-a, a]^d)}, 1 \right\}$$

und

$$M^{2p} \geq c_{131} \cdot \left(\max \left\{ a, \|f\|_{C^q([-a, a]^d)} \right\} \right)^{4(q+1)}$$

für eine hinreichend große Konstante $c_{131} \geq 1$ gelten.

Des Weiteren seien $L, r \in \mathbb{N}$ mit

$$(i) \quad L \geq 5 + \lceil \log_4(M^{2p}) \rceil \cdot (\lceil \log_2(\max\{q, d\} + 1) \rceil + 1)$$

$$(ii) \quad r \geq 2^d \cdot 64 \cdot \binom{d+q}{d} \cdot d^2 \cdot (q+1) \cdot M^d.$$

Dann existiert ein neuronales Netz \hat{f}_{net} aus der Klasse $\mathcal{F}(L, r)$, dessen Gewichtsvektor $\mathbf{w}_{\hat{f}_{\text{net}}}$ die Eigenschaften

$$\|\mathbf{w}_{\hat{f}_{\text{net}}}\|_{\infty} \leq M^{4p+4}, \quad \left(\mathbf{w}_{\hat{f}_{\text{net}}} \right)_{1,0}^{(L)} = 0 \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{net}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1$$

besitzt und welches die Fehlerabschätzung

$$\|f - \hat{f}_{\text{net}}\|_{\infty, [-a, a]^d} \leq c_{132} \cdot \left(\max \left\{ a, \|f\|_{C^q([-a, a]^d)} \right\} \right)^{4 \cdot (q+1)} \cdot M^{-2p} \quad (4.22)$$

erfüllt.

Beweis. Der Beweis stellt eine Erweiterung des Beweises von Theorem 2 in Kohler und Langer (2021) dar. Eine vollständige Version ist im Anhang dieser Arbeit zu finden. An dieser Stelle wollen wir dennoch eine kurze Skizze des Beweises angeben.

Die zentrale Idee besteht in der Konstruktion neuronaler Netze, die ein stückweise definiertes Taylor-Polynom approximieren, welches auf einer Partition des Würfels $[-a, a]^d$ in M^{2d} gleichgroße Würfel basiert. Hierfür approximieren wir zunächst die Funktion auf einem groben Gitter aus M^d gleichgroßen Würfeln und bestimmen den Würfel C , der den Punkt \mathbf{x} enthält. Anschließend unterteilen wir diesen Würfel in M^d kleinere Würfel, um den Wert des Taylor-Polynoms auf dem feineren Gitter mit M^{2d} Würfeln zu berechnen.

In dem Beweis wird ausgenutzt, dass ein Netzwerk mit $c_{133} \cdot M^d$ Neuronen pro Schicht eine Anzahl von $c_{134} \cdot M^{2d}$ Verbindungen zwischen aufeinanderfolgenden Schichten hat. Jedes der $c_{134} \cdot M^{2d}$ Gewichte des Netzes wird einem der $c_{134} \cdot M^{2d}$ möglichen Werte der Ableitungen von f zugeordnet. Um die richtigen Werte der Ableitungen des Taylor-Polynoms zu finden, verfahren wir in zwei Schritten: In den ersten beiden Schichten des Netzes approximieren wir die Indikatorfunktion für jeden Würfel des groben Gitters. Die Ausgangsschicht dieser Netze wird dann jeweils mit den Ableitungen von f auf dem Würfel des feinen Gitters multipliziert. Diese Werte dienen als Eingabe für die nächsten $c_{135} \cdot M^d$ Neuronen in den folgenden zwei verdeckten Schichten, welche die Indikatorfunktion mit den jeweiligen Werten der Ableitung auf den M^d kleineren Würfeln der Unterteilung von C , welcher \mathbf{x} enthält, multiplizieren.

Durch diese zweistufige Approximation finden wir schließlich den Wert der Ableitungen auf den M^{2d} gleichgroßen Würfeln. In den folgenden Schichten wird dann das Taylor-Polynom bestimmt.

Die Schranken des Gewichtsvektors resultieren aus der Konstruktion des neuronalen Netzes. \square

Gewichtsschranken für zusammengesetzte neuronale Netzwerke

Für den Nachweis der Gewichtsschranken in Lemma 15 spielt das folgende Hilfsresultat eine zentrale Rolle. Das Netzwerk, welches für die Approximation in Lemma 16 benötigt wird, setzt sich aus kleineren Teilnetzwerken zusammen. Daher werden wir nun die Verknüpfung neuronaler Netze genauer betrachten. Hierfür folgen wir dem Ansatz von Kohler und Langer (2021), die das erste neuronale Netzwerk g als eine Eingabe für das zweite Netzwerk f verwenden. Die inneren Gewichte des Netzwerks f und die äußeren Gewichte des Netzes g werden daher „verschmolzen“, um die Gewichte des Netzwerks $f \circ g$ zu erhalten. Das folgende Beispiel sowie Abbildung 4.2 sollen diese Idee verdeutlichen.

Seien $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ und $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ zwei neuronale Netze mit

$$f(\mathbf{x}) = \beta_f \cdot \sigma(\alpha_f \cdot \mathbf{x}) \quad \text{mit } \alpha_f \in \mathbb{R}^{3 \times 2}, \beta_f \in \mathbb{R}^{1 \times 3}$$

und

$$g(\mathbf{x}) = \beta_g \cdot \sigma(\alpha_g \cdot \mathbf{x}) \quad \text{mit } \alpha_g \in \mathbb{R}^{3 \times 2}, \beta_g \in \mathbb{R}^{2 \times 3},$$

dann gilt für die Komposition der Netze

$$(f \circ g)(\mathbf{x}) = f(g(\mathbf{x})) = \beta_f \cdot \sigma(\alpha_f \cdot \beta_g \cdot \sigma(\alpha_g \cdot \mathbf{x})).$$

Für die Berechnung der Gewichtsschranken einer Komposition neuronaler Netze benötigen wir das folgende Lemma.

Lemma 17. Sei $g_0 : \mathbb{R}^k \rightarrow \mathbb{R}$ ein neuronales Netz der Klasse $\mathcal{F}(L, r)$ mit dem Gewichtsvektor \mathbf{w}_0 und seien $g_1, \dots, g_k : \mathbb{R}^d \rightarrow \mathbb{R}$ neuronale Netze der Klasse $\mathcal{F}(\bar{L}, \bar{r})$ mit Gewichtsvektoren $\mathbf{w}_1, \dots, \mathbf{w}_k$. Wir bezeichnen mit $\bar{\mathbf{w}}$ den Vektor, der alle Gewichtsvektoren $(\mathbf{w}_j)_{j \in \{1, \dots, k\}}$ enthält. Das Netzwerk $g = g_0(g_1, \dots, g_k)$ besitzt dann $L + \bar{L}$ Schichten und höchstens $\max\{k \cdot \bar{r}, r\}$ -viele Neuronen pro Schicht. Sei \mathbf{w} der Gewichtsvektor der Verknüpfung $g = g_0(g_1, \dots, g_k)$, dann gilt:

a) Der Gewichtsvektor \mathbf{w} ist beschränkt durch

$$\|\mathbf{w}\|_\infty \leq \max \left\{ \|\mathbf{w}_0\|_\infty, \|\bar{\mathbf{w}}\|_\infty, \left\| \mathbf{w}_0^{(0)} \right\|_\infty \cdot \left(k \cdot \left\| \bar{\mathbf{w}}^{(\bar{L})} \right\|_\infty + 1 \right) \right\}.$$

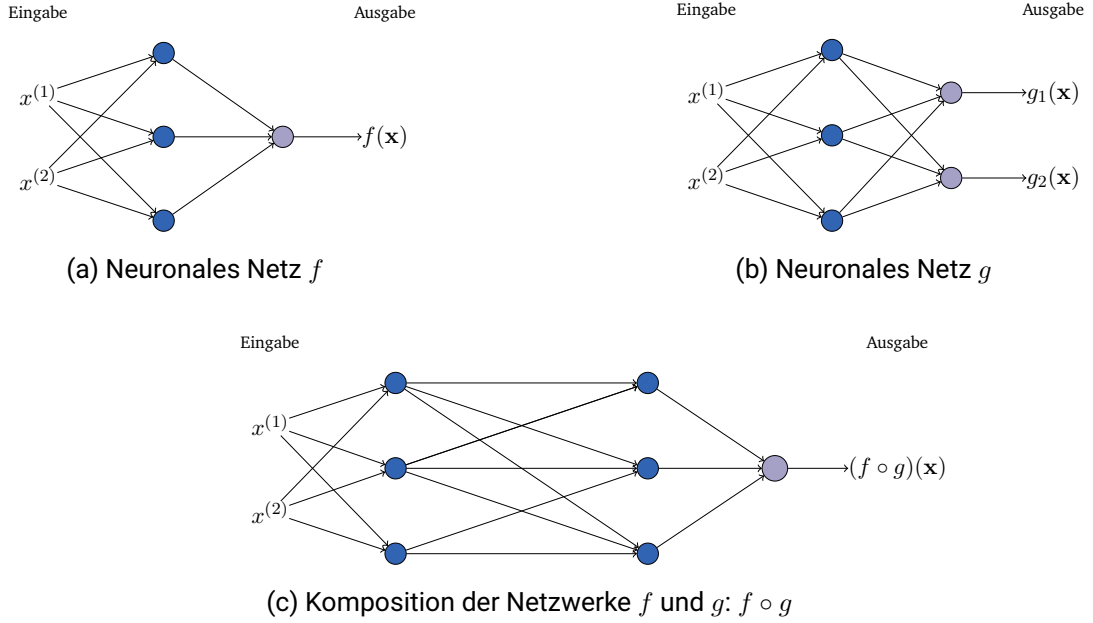


Abbildung 4.2: Darstellung der Komposition eines neuronalen Netzes $f \circ g$

b) Ist $(\mathbf{w}_j)_{1,0}^{(\bar{L})} = 0$ für alle $j \in \{1, \dots, k\}$, so ist der Gewichtsvektor \mathbf{w} beschränkt durch

$$\|\mathbf{w}\|_\infty \leq \max \left\{ \|\mathbf{w}_0\|_\infty, \|\bar{\mathbf{w}}\|_\infty, \left\| (\mathbf{w}_0)_{i,j>0}^{(0)} \right\|_\infty \cdot \left\| (\bar{\mathbf{w}})_{1,j>0}^{(\bar{L})} \right\|_\infty \right\}.$$

c) Gilt zusätzlich zu $(\mathbf{w}_j)_{1,0}^{(\bar{L})} = 0$ für alle $j \in \{1, \dots, k\}$, dass

$$\left\| (\mathbf{w}_0)_{i,j>0}^{(0)} \right\|_\infty \leq 1 \quad \text{oder} \quad \left\| (\bar{\mathbf{w}})_{i,j>0}^{(\bar{L})} \right\|_\infty \leq 1$$

erfüllt ist, so ist der Gewichtsvektor \mathbf{w} beschränkt durch

$$\|\mathbf{w}\|_\infty \leq \max \{ \|\mathbf{w}_0\|_\infty, \|\bar{\mathbf{w}}\|_\infty \}.$$

Beweis. Siehe Lemma 15 in Kohler et al. (2024). □

Nun können wir mithilfe von Lemma 16 und Lemma 17 die Aussage von Lemma 15 nachweisen.

Beweis von Lemma 15. Dieser Beweis orientiert sich in großen Teilen an dem Beweis von Theorem 3 in Kohler und Langer (2021). Der Vollständigkeit halber wird dieser dennoch im Folgenden angegeben.

Im *ersten Schritt des Beweises* wird die Berechnung der Funktion $m(\mathbf{x}) = h_1^{(\ell)}(\mathbf{x})$ beschrieben. Diese erfolgt gemäß den Gleichungen (4.18) und (4.19), weshalb die zentrale Idee darin besteht, ein zusammengesetztes Netzwerk zu definieren, das die Funktionen $h_1^{(1)}, \dots, h_{\tilde{N}_1}^{(1)}, h_1^{(2)}, \dots, h_{\tilde{N}_2}^{(2)}, \dots, h_1^{(\ell)}$ näherungsweise berechnet. Da die Funktionen $h_j^{(i)}$ durch die $(p_j^{(i)}, C)$ -glatte Funktionen $g_j^{(i)}$ beschrieben werden, konstruieren wir

zuerst eine Approximation von $g_j^{(i)}$. Für diese Approximation verwenden wir die in Lemma 16 eingeführten neuronalen Netze

$$\widehat{f}_{\text{net},g_j^{(i)}} \in \mathcal{F}(L_0, r_j^{(i)}),$$

wobei wir

$$L_0 = 5 + \left\lceil \log_4 \left(\max_{j,i} M_{j,i}^{2p_j^{(i)}} \right) \right\rceil \cdot (\lceil \log_2(\max\{K_{\max}, p_{\max}\} + 1) \rceil + 1)$$

und

$$r_j^{(i)} = 2^{K_j^{(i)}} \cdot 64 \cdot \binom{K_j^{(i)} + q_j^{(i)}}{K_j^{(i)}} \cdot (K_j^{(i)})^2 \cdot (q_j^{(i)} + 1) \cdot M_{j,i}^{K_j^{(i)}}$$

für $j \in \{1, \dots, \widetilde{N}_i\}$ und $i \in \{1, \dots, \ell\}$ wählen.

Für die Berechnung der Werte von $h_1^{(1)}, \dots, h_{\widetilde{N}_1}^{(1)}$ verwenden wir dann die neuronalen Netze

$$\begin{aligned} \widehat{h}_1^{(1)}(\mathbf{x}) &= \widehat{f}_{\text{net},g_1^{(1)}} \left(x^{(\pi(1))}, \dots, x^{(\pi(K_1^{(1)}))} \right) \\ &\vdots \\ \widehat{h}_{\widetilde{N}_1}^{(1)}(\mathbf{x}) &= \widehat{f}_{\text{net},g_{\widetilde{N}_1}^{(1)}} \left(x^{(\pi(\sum_{t=1}^{\widetilde{N}_1-1} K_t^{(1)}+1))}, \dots, x^{(\pi(\sum_{t=1}^{\widetilde{N}_1} K_t^{(1)}))} \right). \end{aligned}$$

Um die Werte von $h_1^{(i)}, \dots, h_{\widetilde{N}_i}^{(i)}$ für $i \in \{2, \dots, \ell\}$ zu berechnen, verwenden wir dementsprechend die Netzwerke

$$\widehat{h}_j^{(i)}(\mathbf{x}) = \widehat{f}_{\text{net},g_j^{(i)}} \left(\widehat{h}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}^{(i-1)}(\mathbf{x}), \dots, \widehat{h}_{\sum_{t=1}^j K_t^{(i)}}^{(i-1)}(\mathbf{x}) \right)$$

für $j \in \{1, \dots, \widetilde{N}_i\}$.

Schließlich setzen wir

$$t_1(\mathbf{x}) = \widehat{h}_1^{(\ell)}(\mathbf{x}).$$

In Abbildung 4.3 ist der Aufbau des Netzes t_1 grafisch dargestellt (vgl. Kohler und Langer (2021)).

Das Netzwerk t_1 bildet ein zusammengesetztes Netz, wobei die Netze $\widehat{h}_1^{(i)}, \dots, \widehat{h}_{\widetilde{N}_i}^{(i)}$ jeweils für jedes $i \in \{1, \dots, \ell\}$ parallel berechnet werden. Jedes Netzwerk $\widehat{h}_j^{(i)}$ für $j \in \{1, \dots, \widetilde{N}_i\}$ benötigt L_0 Schichten und $r_j^{(i)}$ Neuronen pro Schicht. Deshalb ist das finale Netzwerk t_1 in der Klasse

$$\mathcal{F} \left(\ell \cdot L_0, \max_{i \in \{1, \dots, \ell\}} \sum_{j=1}^{\widetilde{N}_i} r_j^{(i)} \right) \subseteq \mathcal{F}(L, r)$$

enthalten. Diese Beziehung ergibt sich aus den Annahmen für L und r .

Im zweiten Schritt des Beweises wollen wir die Gewichtsschranken des Netzes t_1 nachweisen.

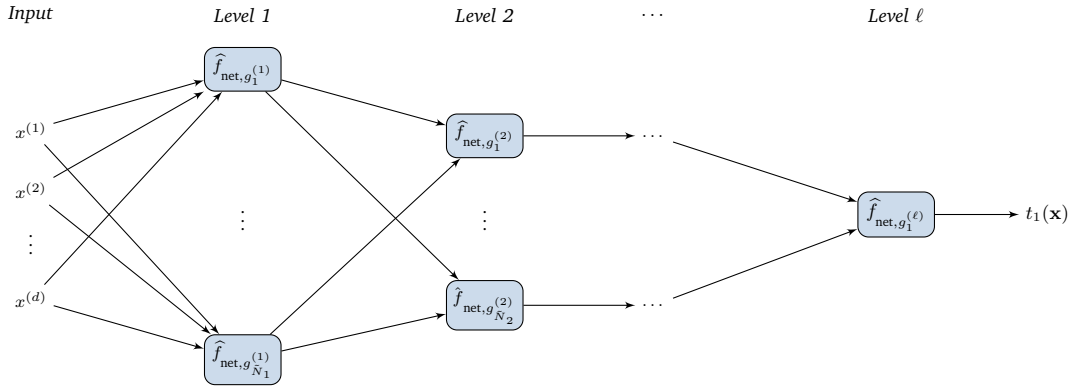


Abbildung 4.3: Darstellung des neuronalen Netzes t_1

Da das Netzwerk t_1 eine Komposition neuronaler Netze aus Lemma 16 ist, benötigen wir die Gewichtsschranken von $\hat{f}_{\text{net},g_j^{(i)}}$. Diese sind gemäß Lemma 16 durch

$$\|\mathbf{w}_{\hat{f}_{\text{net},g_j^{(i)}}}\|_\infty \leq M_{j,i}^{4p_j^{(i)}+4}, \quad \left(\mathbf{w}_{\hat{f}_{\text{net},g_j^{(i)}}}\right)_{1,0}^{(L)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{net},g_j^{(i)}}}\right)_{i,j>0}^{(0)} \right\|_\infty \leq 1 \quad (4.23)$$

gegeben.

Aus der Definition von $\hat{h}_j^{(1)}$ und den Gewichtsschranken des Netzes $\hat{f}_{\text{net},g_j^{(1)}}$ für $j \in \{1, \dots, \tilde{N}_1\}$ folgt direkt

$$\|\mathbf{w}_{\hat{h}_j^{(1)}}\|_\infty = \|\mathbf{w}_{\hat{f}_{\text{net},g_j^{(1)}}}\|_\infty \leq M_{j,1}^{4p_j^{(1)}+4}$$

für $j \in \{1, \dots, \tilde{N}_1\}$.

Aufgrund der Gewichtsschranken aus (4.23) führt das sukzessive Anwenden von Lemma 17c) auf die Abschätzung

$$\begin{aligned} \|\mathbf{w}_{\hat{h}_j^{(i)}}\|_\infty &\leq \max \left\{ \|\mathbf{w}_{\hat{f}_{\text{net},g_j^{(i)}}}\|_\infty, \|\mathbf{w}_{\hat{h}_j^{(i-1)}}\|_\infty \right\} \\ &\leq \max \left\{ M_{j,i}^{4p_j^{(i)}+4}, \max_{j,i} M_{j,i-1}^{4p_j^{(i-1)}+4} \right\} \\ &\leq \max_{j,i} M_{j,i}^{4p_j^{(i)}+4}. \end{aligned}$$

Damit ergibt sich für die Gewichte des neuronalen Netzes t_1 die folgende Schranke

$$\|\mathbf{w}_{t_1}\|_\infty = \|\mathbf{w}_{\hat{h}_1^{(\ell)}}\|_\infty \leq \max_{j,i} M_{j,i}^{4p_j^{(i)}+4}.$$

Im *dritten Beweisschritt* werden wir zeigen, dass das Netzwerk t_1 die Ungleichung

$$\|t_1 - m\|_{\infty, [-a, a]^d} \leq c_{130} \cdot a^{4(p_{\max}+1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \quad (4.24)$$

erfüllt. Sei hierfür

$$g_{\max} := \max \left\{ \max_{\substack{i \in \{1, \dots, \ell\} \\ j \in \{1, \dots, \tilde{N}_i\}}} \|g_j^{(i)}\|_{\infty}, 1 \right\}.$$

Da jedes $g_j^{(i)}$ die Annahmen aus Lemma 16 erfüllt, sowie

$$\|g_j^{(i)}\|_{C^{q_j^{(i)}}(\mathbb{R}^d)} \leq c_{128}$$

gilt, können wir

$$\begin{aligned} \left| g_j^{(i)}(\mathbf{x}) - \hat{f}_{\text{net}, g_j^{(i)}}(\mathbf{x}) \right| &\leq c_{132} \cdot (\max\{2 \cdot \max\{g_{\max}, a\}, c_{128}\})^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \\ &\leq c_{136} \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \end{aligned} \quad (4.25)$$

für $\mathbf{x} \in [-2 \cdot \max\{g_{\max}, a\}, 2 \cdot \max\{g_{\max}, a\}]^{K_j^{(i)}}$ folgern, wobei

$$c_{136} \geq c_{132} \cdot (2 \cdot g_{\max} \cdot \max\{c_{128}, 1\})^{4 \cdot (p_{\max} + 1)}.$$

Wir zeigen nun durch Induktion, dass

$$\left| \hat{h}_j^{(i)}(\mathbf{x}) - h_j^{(i)}(\mathbf{x}) \right| \leq c_{136} \cdot i \cdot (K_{\max} \cdot C_{\text{Lip}})^{i-1} \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \quad (4.26)$$

gilt, wobei C_{Lip} die Lipschitz-Konstante von $g_j^{(i)}$ ist. Aus Ungleichung (4.25) können wir für $\mathbf{x} \in [-2 \max\{g_{\max}, a\}, 2 \max\{g_{\max}, a\}]^{K_j^{(i)}}$ folgern, dass

$$\begin{aligned} &\left| \hat{h}_j^{(1)}(\mathbf{x}) - h_j^{(1)}(\mathbf{x}) \right| \\ &= \left| \hat{f}_{\text{net}, g_j^{(1)}} \left(x^{(\pi(\sum_{t=1}^{j-1} K_t^{(1)} + 1))}, \dots, x^{(\pi(\sum_{t=1}^j K_t^{(1)}))} \right) - g_j^{(1)} \left(x^{(\pi(\sum_{t=1}^{j-1} K_t^{(1)} + 1))}, \dots, x^{(\pi(\sum_{t=1}^j K_t^{(1)}))} \right) \right| \\ &\leq c_{136} \cdot 1 \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \\ &= c_{136} \cdot (K_{\max} \cdot C_{\text{Lip}})^{1-1} \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \end{aligned}$$

für $j \in \{1, \dots, \tilde{N}_1\}$ ist. Daher haben wir gezeigt, dass (4.26) für $i = 1$ gilt. Zusätzlich können wir damit für $\min_{j,i} M_{j,i}^{-2p_j^{(i)}} \geq c_{136} \cdot a^{4 \cdot (p_{\max} + 1)}$ den Wert des Netzes wie folgt abschätzen

$$\left| \hat{h}_j^{(1)}(\mathbf{x}) \right| \leq \left| \hat{h}_j^{(1)}(\mathbf{x}) - h_j^{(1)}(\mathbf{x}) \right| + g_{\max} \leq 1 + g_{\max} \leq 2 \cdot g_{\max}.$$

Damit ist die Ausgabe von $\hat{h}_j^{(1)}$ in dem Intervall enthalten, in dem die Ungleichung (4.25) gilt.

Wir nehmen im Folgenden an, dass Ungleichung (4.26) für ein $i - 1$ und alle $j \in \{1, \dots, \tilde{N}_{i-1}\}$ erfüllt ist. Dann ergibt sich mithilfe von (4.25) analog zu obigem

$$\left| \hat{h}_j^{(i-1)}(\mathbf{x}) \right| \leq \left| \hat{h}_j^{(i-1)}(\mathbf{x}) - h_j^{(i-1)}(\mathbf{x}) \right| + g_{\max} \leq 2 \cdot g_{\max}$$

für $\mathbf{x} \in [-2 \max\{g_{\max}, a\}, 2 \max\{g_{\max}, a\}]^{K_j^{(i-1)}}$. Daher liegt auch $\widehat{h}_j^{(i-1)}(\mathbf{x})$ in dem Intervall, für das Ungleichung (4.25) gilt. Die Anwendung von Ungleichung (4.25), die Induktionsannahme (4.26) sowie die Lipschitz-Stetigkeit von $g_j^{(i)}$ liefern

$$\begin{aligned}
& \left| \widehat{h}_j^{(i)}(\mathbf{x}) - h_j^{(i)}(\mathbf{x}) \right| \\
& \leq \left| \widehat{f}_{\text{net}, g_j^{(i)}} \left(\widehat{h}_{\sum_{t=1}^{j-1} K_t^{(i)+1}}^{(i-1)}(\mathbf{x}), \dots, \widehat{h}_{\sum_{t=1}^j K_t^{(i)}(\mathbf{x})}^{(i-1)}(\mathbf{x}) \right) - g_j^{(i)} \left(\widehat{h}_{\sum_{t=1}^{j-1} K_t^{(i)+1}}^{(i-1)}(\mathbf{x}), \dots, \widehat{h}_{\sum_{t=1}^j K_t^{(i)}(\mathbf{x})}^{(i-1)}(\mathbf{x}) \right) \right| \\
& + \left| g_j^{(i)} \left(\widehat{h}_{\sum_{t=1}^{j-1} K_t^{(i)+1}}^{(i-1)}(\mathbf{x}), \dots, \widehat{h}_{\sum_{t=1}^j K_t^{(i)}(\mathbf{x})}^{(i-1)}(\mathbf{x}) \right) - g_j^{(i)} \left(h_{\sum_{t=1}^{j-1} K_t^{(i)+1}}^{(i-1)}(\mathbf{x}), \dots, h_{\sum_{t=1}^j K_t^{(i)}(\mathbf{x})}^{(i-1)}(\mathbf{x}) \right) \right| \\
& \leq c_{136} \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \\
& \quad + K_j^{(i)} \cdot C_{\text{Lip}} \cdot c_{136} \cdot (i-1) \cdot (K_{\max} \cdot C_{\text{Lip}})^{i-2} \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}} \\
& \leq c_{136} \cdot i \cdot (K_{\max} \cdot C_{\text{Lip}})^{i-1} \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}}.
\end{aligned}$$

Somit haben wir gezeigt, dass ein Netzwerk $t_1(\mathbf{x})$ existiert, welches die Ungleichung

$$\|t_1 - m\|_{\infty, [-a, a]^d} \leq c_{130} \cdot a^{4 \cdot (p_{\max} + 1)} \cdot \max_{j,i} M_{j,i}^{-2p_j^{(i)}}$$

erfüllt, woraus die Behauptung des Lemmas folgt. \square

4.4 Konvergenzgeschwindigkeit des Schätzers

Aufgrund der strukturellen Annahmen an die Regressionsfunktion ist es möglich, eine Konvergenzgeschwindigkeit abzuleiten, die unabhängig von der Eingabedimension d ist. Mithilfe des folgenden Theorems können wir somit die tatsächliche Qualität des überparametrisierten Neuronale-Netze-Schätzers bewerten.

Theorem 6. *Seien $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ unabhängig und identisch verteilte Zufallsvariablen mit Werten in $\mathbb{R}^d \times \mathbb{R}$, wobei $\text{supp}(\mathbf{X})$ beschränkt ist und*

$$\mathbf{E} \left\{ \exp(c_{120} \cdot Y^2) \right\} < \infty$$

für eine Konstante $c_{120} > 0$ gilt. Sei m die zugehörige Regressionsfunktion, die für ein $\ell \in \mathbb{N}$ und $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$ in der Klasse $\mathcal{H}(\ell, \mathcal{P})$ enthalten ist. Zudem nehmen wir an, dass alle Funktionen $g : \mathbb{R}^K \rightarrow \mathbb{R}$ im hierarchischen Kompositionsmodell Lipschitz-stetig bezüglich der $\|\cdot\|_1$ -Norm mit Lipschitz-Konstante $C_{\text{Lip}} \geq 1$ sind und die Bedingung

$$\|g\|_{C^{q_g}(\mathbb{R}^K)} \leq c_{137}$$

für eine Konstante $c_{137} > 0$ erfüllen, wobei $p_g = q_g + s_g \geq 1$ mit $s_g \in (0, 1]$ und $q_g \in \mathbb{N}_0$ ist. Das Tupel $(p_g, K_g) \in \mathcal{P}$ gibt die entsprechende Glattheit und Ordnung von g im hierarchischen Kompositionsmodell an. Wir bezeichnen mit K_{\max} die maximale Eingabedimension und mit p_{\max} die maximale Glattheit einer der Funktionen g . Zudem seien $p_{\max}, K_{\max} < \infty$. Sei $\widehat{K}_n \in \mathbb{N}$, so dass

$$\frac{\widehat{K}_n}{\exp(2 \cdot n)} \rightarrow \infty \quad (n \rightarrow \infty).$$

Des Weiteren sei m_n der in (4.7) definierte Schätzer, wobei \mathcal{W}^0 die Menge aller Gewichtsvektoren ist, deren Komponenten im Intervall $[-c_{138} \cdot n^{4/3}, c_{138} \cdot n^{4/3}]$ für eine Konstante $c_{138} \geq 1$ liegen. Wir setzen

$$\beta_n = c_5 \cdot \log n, \quad t_n = n \cdot \widehat{K}_n, \quad \lambda_n = \frac{1}{t_n}$$

sowie

$$L_n = \lceil c_{139} \cdot \log n \rceil \quad \text{und} \quad r_n = \left\lceil c_{140} \cdot \max_{(p,K) \in \mathcal{P}} n^{\frac{K}{2(2p+K)}} \right\rceil.$$

Dann existiert eine Konstante $c_{141} > 0$, so dass

$$\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_X(d\mathbf{x}) \leq c_{141} \cdot (\log n)^5 \cdot \max_{(p,K) \in \mathcal{P}} n^{-\frac{p}{2p+K}}$$

für hinreichend große n gilt.

Für den Beweis von Theorem 6 möchten wir die in Theorem 5 hergeleitete Fehlerschranke verwenden. Um nachzuweisen, dass in Theorem 6 alle hierfür erforderlichen Voraussetzungen erfüllt sind, benötigen wir die folgenden zwei Hilfsresultate. Das erste Lemma liefert uns eine Schranke für die Norm des Gradienten der Verlustfunktion bezüglich der äußeren Gewichte.

Lemma 18. Seien $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ Zufallsvariablen mit Werten in $\mathbb{R}^d \times \mathbb{R}$. Sei weiter $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ eine beliebige Funktion und sei $\gamma_n \leq 1$. Das empirische L_2 -Risiko sei durch

$$F_n(\mathbf{w}) = \frac{1}{n} \sum_{s=1}^n |f_{\mathbf{w}}(\mathbf{X}_s) - Y_s|^2$$

gegeben. Angenommen, es gelte $Y_1, \dots, Y_n \in [-\beta_n, \beta_n]$ sowie

$$\sum_{k=1}^{\widehat{K}_n} |w_k| \leq \gamma_n, \tag{4.27}$$

so erhalten wir die folgende Schranke für den Gradienten des empirischen L_2 -Risikos bezüglich der äußeren Gewichte

$$\left\| \left(\nabla_{(w_k)_{k=1, \dots, \widehat{K}_n}} F_n \right) (\mathbf{w}) \right\|^2 \leq 8 \cdot \widehat{K}_n \cdot \beta_n^4 \cdot (\gamma_n^2 + 1). \tag{4.28}$$

Beweis. Aus der Definition des Gradienten des empirischen L_2 -Risikos bezüglich der Gewichte der Ausgabeschicht und unter Anwendung der Cauchy-Schwarz-Ungleichung ergibt sich

$$\begin{aligned} \left\| \left(\nabla_{(w_k)_{k=1, \dots, \widehat{K}_n}} F_n \right) (\mathbf{w}) \right\|^2 &= \sum_{k=1}^{\widehat{K}_n} \left| \frac{2}{n} \sum_{s=1}^n (f_{\mathbf{w}}(\mathbf{X}_s) - Y_s) \cdot T_{\beta_n}(f_{\mathbf{w},k,1}(\mathbf{X}_s)) \right|^2 \\ &\leq \sum_{k=1}^{\widehat{K}_n} \left(4 \cdot \frac{1}{n} \sum_{s=1}^n |f_{\mathbf{w}}(\mathbf{X}_s) - Y_s|^2 \cdot \frac{1}{n} \sum_{s=1}^n T_{\beta_n}(f_{\mathbf{w},k,1}(\mathbf{X}_s))^2 \right). \end{aligned}$$

Aufgrund der Voraussetzung (4.27), der Beschränktheit von Y_s für $s = 1, \dots, n$ sowie der Ungleichung $(a - b)^2 \leq 2 \cdot a^2 + 2 \cdot b^2$ für $a, b \in \mathbb{R}$ erhalten wir

$$\begin{aligned}
\frac{1}{n} \sum_{s=1}^n |f_{\mathbf{w}}(\mathbf{X}_s) - Y_s|^2 &\leq \frac{1}{n} \sum_{s=1}^n (2 \cdot f_{\mathbf{w}}(\mathbf{X}_s)^2 + 2 \cdot Y_s^2) \\
&= \frac{2}{n} \sum_{s=1}^n \left(\left| \sum_{k=1}^{\widehat{K}_n} w_k \cdot T_{\beta_n}(f_{\mathbf{w},k,1}(\mathbf{X}_s)) \right|^2 + Y_s^2 \right) \\
&\leq \frac{2}{n} \sum_{s=1}^n \left(\left(\sum_{k=1}^{\widehat{K}_n} |w_k| \cdot |T_{\beta_n}(f_{\mathbf{w},k,1}(\mathbf{X}_s))| \right)^2 + \beta_n^2 \right) \\
&\leq \frac{2}{n} \sum_{s=1}^n ((\gamma_n \cdot \beta_n)^2 + \beta_n^2) \\
&\leq 2 \cdot \beta_n^2 \cdot (\gamma_n^2 + 1).
\end{aligned}$$

Daher können wir folgern, dass

$$\begin{aligned}
\sum_{k=1}^{\widehat{K}_n} \left(4 \cdot \frac{1}{n} \sum_{s=1}^n |f_{\mathbf{w}}(\mathbf{X}_s) - Y_s|^2 \cdot \frac{1}{n} \sum_{s=1}^n T_{\beta_n}(f_{\mathbf{w},k,1}(\mathbf{X}_s))^2 \right) &\leq \sum_{k=1}^{\widehat{K}_n} (4 \cdot 2 \cdot \beta_n^2 \cdot (\gamma_n^2 + 1) \cdot \beta_n^2) \\
&\leq 8 \cdot \widehat{K}_n \cdot \beta_n^4 \cdot (\gamma_n^2 + 1)
\end{aligned}$$

gilt. □

Das zweite Hilfsresultat ermöglicht es, unter geeigneten Bedingungen an die Gewichte des neuronalen Netzes eine obere Schranke für die Differenz der Ausgaben zweier Netze herzuleiten.

Lemma 19. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion. Seien zudem $a, B, M \geq 1$ und $L, r \in \mathbb{N}$ mit $r \geq d$. Das vollständig verbundene neuronale Netz $f_{\mathbf{w},1,1}(\mathbf{x})$ sei definiert wie in (4.2)–(4.4). Angenommen, es gelte

$$\|(w_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \leq B \quad (4.29)$$

sowie

$$\|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \leq M, \quad (4.30)$$

dann ist für jeden Gewichtsvektor \mathbf{w} die Ungleichung

$$\begin{aligned}
&|f_{\mathbf{w},1,1}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,1}(\mathbf{x})| \\
&\leq (L + 1) \cdot (r + 1)^{L+1} \cdot a \cdot (B + M)^L \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty}
\end{aligned} \quad (4.31)$$

für $\mathbf{x} \in [-a, a]^d$ erfüllt.

Beweis. Sei $\mathbf{x} \in [-a, a]^d$. Zu Beginn des Beweises wollen wir zeigen, dass

$$0 \leq f_{\mathbf{w},1,i}^{(l)}(\mathbf{x}) \leq (r + 1)^l \cdot a \cdot B^l \quad (4.32)$$

für alle $l \in \{1, \dots, L\}$ und $i \in \{1, \dots, r\}$ gilt.

Aufgrund der Definition der ReLU-Aktivierungsfunktion ist $0 \leq \sigma(z) \leq |z|$ für alle $z \in \mathbb{R}$. Daher können wir mit Voraussetzung (4.29) folgern, dass

$$0 \leq f_{\mathbf{w},1,i}^{(1)}(\mathbf{x}) \leq \sum_{j=1}^d |w_{1,i,j}^{(0)}| \cdot |x^{(j)}| + |w_{1,i,0}^{(0)}| \leq (d+1) \cdot B \cdot a \leq (r+1) \cdot B \cdot a$$

für alle $i \in \{1, \dots, r\}$ gilt. Wiederholen wir dies sukzessive, so erhalten wir

$$0 \leq f_{\mathbf{w},1,i}^{(l)}(\mathbf{x}) \leq \sum_{j=1}^r |w_{1,i,j}^{(l-1)}| \cdot f_{\mathbf{w},1,j}^{(l-1)}(\mathbf{x}) + |w_{1,i,0}^{(l-1)}| \leq (r+1)^l \cdot a \cdot B^l$$

für $l \in \{2, \dots, L\}$ und $i \in \{1, \dots, r\}$.

Nun wollen wir mittels Induktion über l zeigen, dass die Ungleichung

$$|f_{\mathbf{w},1,i}^{(l)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,i}^{(l)}(\mathbf{x})| \leq l \cdot (r+1)^l \cdot a \cdot (B+M)^{l-1} \cdot \left\| (w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l} \right\|_{\infty} \quad (4.33)$$

für $l \in \{1, \dots, L\}$ und $i \in \{1, \dots, r\}$ erfüllt ist.

Die ReLU-Aktivierungsfunktion ist Lipschitz-stetig mit Lipschitz-Konstante 1, das heißt

$$|\sigma(y) - \sigma(z)| \leq |y - z|$$

für $y, z \in \mathbb{R}$. Daher gilt für $l = 1$ und $i \in \{1, \dots, r\}$, dass

$$\begin{aligned} |f_{\mathbf{w},1,i}^{(1)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,i}^{(1)}(\mathbf{x})| &\leq \sum_{j=1}^d |w_{1,i,j}^{(0)} - \bar{w}_{1,i,j}^{(0)}| \cdot |x^{(j)}| + |w_{1,i,0}^{(0)} - \bar{w}_{1,i,0}^{(0)}| \\ &\leq (d+1) \cdot a \cdot \|(w_{1,i,j}^{(0)})_{i,j,1} - (\bar{w}_{1,i,j}^{(0)})_{i,j,1}\|_{\infty} \\ &\leq 1 \cdot (r+1)^1 \cdot a \cdot (B+M)^0 \cdot \|(w_{1,i,j}^{(0)})_{i,j,1} - (\bar{w}_{1,i,j}^{(0)})_{i,j,1}\|_{\infty} \end{aligned}$$

ist.

Wir nehmen nun an, dass Ungleichung (4.33) für ein $l \in \{1, \dots, L-1\}$ gelte. Diese Annahme führt zusammen mit der Beschränktheit von $|f_{\mathbf{w},1,i}^{(l)}(\mathbf{x})|$ zu

$$\begin{aligned} &|f_{\mathbf{w},1,i}^{(l+1)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,i}^{(l+1)}(\mathbf{x})| \\ &\leq \sum_{j=1}^r |w_{1,i,j}^{(l)} - \bar{w}_{1,i,j}^{(l)}| \cdot |f_{\mathbf{w},1,j}^{(l)}(\mathbf{x})| + |w_{1,i,0}^{(l)} - \bar{w}_{1,i,0}^{(l)}| + \sum_{j=1}^r |\bar{w}_{1,i,j}^{(l)}| \cdot |f_{\mathbf{w},1,j}^{(l)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,j}^{(l)}(\mathbf{x})| \\ &\leq \sum_{j=1}^r |w_{1,i,j}^{(l)} - \bar{w}_{1,i,j}^{(l)}| \cdot |f_{\mathbf{w},1,j}^{(l)}(\mathbf{x})| + |w_{1,i,0}^{(l)} - \bar{w}_{1,i,0}^{(l)}| \\ &\quad + \sum_{j=1}^r \left(\|(\bar{w}_{1,i,j}^{(l)})_{i,j,l} - (w_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} + \|(w_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \right) \cdot |f_{\mathbf{w},1,j}^{(l)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,j}^{(l)}(\mathbf{x})|. \end{aligned}$$

Aus der Induktionsannahme und den Ungleichungen (4.29), (4.30) sowie (4.32) folgt dann

$$\sum_{j=1}^r |w_{1,i,j}^{(l)} - \bar{w}_{1,i,j}^{(l)}| \cdot |f_{\mathbf{w},1,j}^{(l)}(\mathbf{x})| + |w_{1,i,0}^{(l)} - \bar{w}_{1,i,0}^{(l)}|$$

$$\begin{aligned}
& + \sum_{j=1}^r \left(\|(\bar{w}_{1,i,j}^{(l)})_{i,j,l} - (w_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} + \|(w_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \right) \cdot |f_{\mathbf{w},1,j}^{(l)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,j}^{(l)}(\mathbf{x})| \\
& \leq (r+1) \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \cdot (r+1)^l \cdot a \cdot B^l \\
& \quad + r \cdot (M+B) \cdot l \cdot (r+1)^l \cdot a \cdot (B+M)^{l-1} \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \\
& \leq (r+1)^{l+1} \cdot a \cdot B^l \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \\
& \quad + l \cdot (r+1)^{l+1} \cdot a \cdot (B+M)^l \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \\
& \leq (l+1) \cdot (r+1)^{l+1} \cdot a \cdot (B+M)^l \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty}.
\end{aligned}$$

Wenden wir dies auf $|f_{\mathbf{w},1,1}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,1}(\mathbf{x})|$ an, so erhalten wir

$$\begin{aligned}
|f_{\mathbf{w},1,1}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,1}(\mathbf{x})| & \leq \sum_{j=1}^r |w_{1,1,j}^{(L)} - \bar{w}_{1,1,j}^{(L)}| \cdot |f_{\mathbf{w},1,j}^{(L)}(\mathbf{x})| + \sum_{j=1}^r |\bar{w}_{1,1,j}^{(L)}| \cdot |f_{\mathbf{w},1,j}^{(L)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,j}^{(L)}(\mathbf{x})| \\
& \leq \sum_{j=1}^r |w_{1,1,j}^{(L)} - \bar{w}_{1,1,j}^{(L)}| \cdot |f_{\mathbf{w},1,j}^{(L)}(\mathbf{x})| \\
& \quad + \sum_{j=1}^r \left(\|(\bar{w}_{1,i,j}^{(l)})_{i,j,l} - (w_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} + \|(w_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \right) \cdot |f_{\mathbf{w},1,j}^{(L)}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,j}^{(L)}(\mathbf{x})| \\
& \leq (r+1)^L \cdot a \cdot B^L \cdot r \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \\
& \quad + r \cdot (M+B) \cdot L \cdot (r+1)^L \cdot a \cdot (B+M)^{L-1} \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \\
& \leq (r+1)^{L+1} \cdot a \cdot B^L \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \\
& \quad + L \cdot (r+1)^{L+1} \cdot a \cdot (B+M)^L \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \\
& \leq (L+1) \cdot (r+1)^{L+1} \cdot a \cdot (B+M)^L \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty},
\end{aligned}$$

woraus die Behauptung des Lemmas folgt. \square

Mithilfe der beiden Hilfsresultate lässt sich die Gültigkeit der Voraussetzungen von Theorem 5 nachweisen, das einen wesentlichen Bestandteil des Beweises von Theorem 6 darstellt. Die Konvergenzrate lässt sich anschließend durch geeignete Abschätzungen der in Theorem 5 enthaltenen Terme herleiten.

Beweis von Theorem 6. Der Beweis von Theorem 6 besteht aus zwei Teilen. Im ersten Teil zeigen wir, dass die Voraussetzungen von Theorem 5 erfüllt sind. Im zweiten Teil werden wir die Konvergenzgeschwindigkeit des Schätzers durch die Anwendung der Fehlerschranke aus Theorem 5 beweisen.

Wir beginnen damit, die Gültigkeit der Voraussetzungen von Theorem 5 nachzuweisen. Erinnern wir uns daran, dass durch \mathcal{W}^0 die Menge aller Gewichtsvektoren definiert ist, deren Komponenten im Intervall $[-c_{138} \cdot n^{4/3}, c_{138} \cdot n^{4/3}]$ für $c_{138} \geq 1$ liegen. Sei $f_{\tilde{\mathbf{w}}}$ das neuronale Netz aus Lemma 15 mit dem zugehörigen Gewichtsvektor $\tilde{\mathbf{w}}$, wobei wir

$$M_{j,i} = \left[n^{\frac{1}{2(2p_j^{(i)} + K_j^{(i)})}} \right] \quad \text{sowie} \quad a_n = (\log n)^{\frac{1}{4(p_{\max} + 1)}}$$

setzen. Gemäß Lemma 15 sind diese Gewichte absolut durch $M_{j,i}^{4p_j^{(i)}+4}$ beschränkt. Aus der Wahl von $M_{j,i}$ folgt

$$M_{j,i}^{4p_j^{(i)}+4} \leq c_{138} \cdot n^{\frac{4p_j^{(i)}+4}{2(2p_j^{(i)}+K_j^{(i)})}} \leq c_{138} \cdot n^{\frac{4p_j^{(i)}+2K_j^{(i)}}{4p_j^{(i)}+2K_j^{(i)}}} \cdot n^{\frac{4-2K_j^{(i)}}{4p_j^{(i)}+2K_j^{(i)}}} \leq c_{138} \cdot n \cdot n^{1/3} \leq c_{138} \cdot n^{4/3},$$

womit die Gewichte des neuronalen Netzes aus Lemma 15 in \mathcal{W}^0 liegen. Zudem definieren wir die Menge \mathcal{W}^* durch

$$\mathcal{W}^* = \{((\mathbf{w}^*)_{1,i,j}^{(l)})_{i,j,l} \in \mathcal{W} : \|((\mathbf{w}^*)_{1,i,j}^{(l)})_{i,j,l} - (\tilde{w}_{1,i,j}^{(l)})_{i,j,l}\|_\infty \leq \bar{\delta}_n\}$$

mit

$$\bar{\delta}_n = \frac{1}{n(\log n)^2}.$$

Im *ersten Schritt des Beweises* wollen wir zeigen, dass die Voraussetzungen (4.9) und (4.11) in Theorem 5 für $C_n = 4 \cdot \sqrt{\widehat{K}_n} \cdot \beta_n^2$ und $D_n = c_{142} \cdot n^{c_{143} \cdot \log n}$ erfüllt sind. Hierfür nehmen wir im Folgenden an, dass n hinreichend groß ist.

Aus Lemma 18 erhalten wir für $Y_1, \dots, Y_n \in [-\beta_n, \beta_n]$ und $\gamma_n = 1$, dass

$$\left\| \left(\nabla_{(w_k)_{k=1, \dots, \widehat{K}_n}} F_n \right) (\mathbf{w}) \right\|^2 \leq 16 \cdot \widehat{K}_n \cdot \beta_n^4 \quad (4.34)$$

ist, woraus direkt $C_n = 4 \cdot \sqrt{\widehat{K}_n} \cdot \beta_n^2$ folgt.

Wir werden nun zeigen, dass Voraussetzung (4.11) in Theorem 5 für $D_n = c_{142} \cdot n^{c_{143} \cdot \log n}$ erfüllt ist. Seien hierfür $(w_{1,i,j}^{(l)})_{i,j,l}, (\bar{w}_{1,i,j}^{(l)})_{i,j,l} \in \overline{\mathcal{W}}$, wobei die Menge $\overline{\mathcal{W}}$ wie in Theorem 5 definiert ist.

Um Ungleichung (4.11) nachweisen zu können, benötigen wir die Aussage von Lemma 19. Daher werden wir nun prüfen, ob die Voraussetzungen (4.29) und (4.30) in Lemma 19 erfüllt sind. Durch die Wahl der Startgewichte ergibt sich

$$\begin{aligned} \|(w_{1,i,j}^{(l)})_{i,j,l}\|_\infty &\leq \|(w_{1,i,j}^{(l)})_{i,j,l} - ((\mathbf{w}^{(0)})_{1,i,j}^{(l)})_{i,j,l}\|_\infty + \|((\mathbf{w}^{(0)})_{1,i,j}^{(l)})_{i,j,l}\|_\infty \\ &\leq c_{121} + c_{138} \cdot n^{4/3}, \end{aligned}$$

woraus die Gültigkeit von Ungleichung (4.30) für $B = c_{144} \cdot n^{4/3}$ in Lemma 19 folgt.

Zudem gilt

$$\begin{aligned} &\|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_\infty \\ &\leq \|(w_{1,i,j}^{(l)})_{i,j,l} - ((\mathbf{w}^{(0)})_{1,i,j}^{(l)})_{i,j,l}\|_\infty + \|((\mathbf{w}^{(0)})_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|_\infty \\ &\leq 2 \cdot c_{121} \end{aligned}$$

für $(w_{1,i,j}^{(l)})_{i,j,l}, (\bar{w}_{1,i,j}^{(l)})_{i,j,l} \in \overline{\mathcal{W}}$. Daher ist auch Voraussetzung (4.30) für $M = 2 \cdot c_{121}$ gegeben. Somit können wir aus Lemma 19 sowie der Wahl von L_n und r_n folgern, dass

$$|f_{\mathbf{w},1,1}(\mathbf{x}) - f_{\bar{\mathbf{w}},1,1}(\mathbf{x})| \leq c_{142} \cdot n^{c_{143} \cdot \log n} \cdot \|(w_{1,i,j}^{(l)})_{i,j,l} - (\bar{w}_{1,i,j}^{(l)})_{i,j,l}\|$$

für $k \in \{1, \dots, \widehat{K}_n\}$ ist, woraus die Gültigkeit von Bedingung (4.11) für $D_n = c_{142} \cdot n^{c_{143} \cdot \log n}$ resultiert.

Im zweiten Schritt des Beweises wollen wir zeigen, dass Voraussetzung (4.10) ebenfalls erfüllt ist. Wir bezeichnen hierfür mit $(\bar{p}, \bar{K}) \in \mathcal{P}$ das Tupel, für welches

$$(\bar{p}, \bar{K}) = \arg \min_{(p, K) \in \mathcal{P}} \frac{p}{K}$$

gilt.

Für eines der vollständig verbundenen neuronalen Netze $f_{\mathbf{w}, k, 1}$ ($k \in \{1, \dots, \widehat{K}_n\}$) ist die Wahrscheinlichkeit, dass jedes der zugehörigen $r_n + (L_n - 1) \cdot r_n \cdot (r_n + 1) + r_n \cdot (d + 1)$ Gewichte in der Menge \mathcal{W}^* liegt, für $r_n \leq c_{145} \cdot n^{\frac{\bar{K}}{2(2\bar{p} + \bar{K})}}$ nach unten beschränkt durch

$$\left(\frac{1}{2 \cdot c_{138} \cdot n^{4/3} \cdot n^{(\log n)^2}} \right)^{r_n + (L_n - 1) \cdot r_n \cdot (r_n + 1) + r_n \cdot (d + 1)}.$$

Dies ergibt sich daraus, dass jedes Startgewicht unabhängig und gleichverteilt aus einem Intervall der Länge $2 \cdot c_{138} \cdot n^{4/3}$ gewählt wird und zugleich in das kleinere Intervall der Breite $\frac{1}{n^{(\log n)^2}}$ fallen muss.

Durch geeignetes Abschätzen nach oben erhalten wir dann

$$\begin{aligned} & \left(\frac{1}{2 \cdot c_{138} \cdot n^{4/3} \cdot n^{(\log n)^2}} \right)^{r_n + (L_n - 1) \cdot r_n \cdot (r_n + 1) + r_n \cdot (d + 1)} \geq \left(\frac{1}{2 \cdot c_{138} \cdot n^{c_{146} \cdot (\log n)^2}} \right)^{c_{147} \cdot \log n \cdot n^{\frac{\bar{K}}{2\bar{p} + \bar{K}}}} \\ & = \exp \left(c_{147} \cdot \log n \cdot n^{\frac{\bar{K}}{2\bar{p} + \bar{K}}} \cdot \left(-\log \left(2 \cdot c_{138} \cdot n^{c_{146} \cdot (\log n)^2} \right) \right) \right) \\ & = \exp \left(-c_{148} \cdot n^{\frac{\bar{K}}{2\bar{p} + \bar{K}}} \cdot (\log n)^4 \right) \\ & \geq \exp(-n). \end{aligned}$$

Die letzte Ungleichung folgt dabei aus der Tatsache, dass $\frac{\bar{K}}{2\bar{p} + \bar{K}} < 1$ gilt.

Die Wahrscheinlichkeit, dass alle $r_n + (L_n - 1) \cdot r_n \cdot (r_n + 1) + r_n \cdot (d + 1)$ Gewichte in der Menge \mathcal{W}^* liegen, ist somit von unten durch $\exp(-n)$ beschränkt. Nach Theorem 5 ist dann $\varepsilon_n \geq \exp(-n)$. Für $N_n = \exp((\log n)^3)$ und $I_n = (\log n)^4 \cdot \exp(n)$ ergibt sich hieraus die Abschätzung

$$\begin{aligned} N_n \cdot (1 - \varepsilon_n)^{I_n} & \leq \exp((\log n)^3) \cdot \left(1 - \frac{1}{\exp(n)} \right)^{(\log n)^4 \cdot \exp(n)} \\ & \leq \exp((\log n)^3) \cdot \exp \left(-\frac{1}{\exp(n)} \right)^{(\log n)^4 \cdot \exp(n)} \\ & \leq \exp((\log n)^3) \cdot \exp(-(\log n)^4) \\ & \leq \frac{1}{n}, \end{aligned}$$

womit Voraussetzung (4.10) aus Theorem 5 nachgewiesen ist.

Im zweiten Teil des Beweises zeigen wir die Aussage von Theorem 6. Hierfür wenden wir Theorem 5 an und erhalten die Ungleichung

$$\mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x})$$

$$\leq c_{122} \cdot \left(\frac{(\log n)^2}{n} + \beta_n \cdot \mathbf{E} \left\{ \sup_{\mathbf{w} \in \mathcal{W}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right| \right\} \right. \\ \left. + \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*,1,1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} + \beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{N_n} + \frac{C_n^2}{t_n} \right).$$

Im dritten Schritt des Beweises zeigen wir, dass

$$\beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{N_n} + \frac{C_n^2}{t_n} \leq \frac{c_{149}}{\sqrt{n}}$$

gilt. Die Definition von t_n in Theorem 6 sowie die Wahl von C_n, D_n und N_n im ersten und zweiten Beweisschritt implizieren

$$\beta_n \cdot D_n \cdot \frac{1}{\sqrt{N_n}} + \frac{1}{N_n} + \frac{C_n^2}{t_n} \leq \frac{c_5 \cdot \log n \cdot c_{142} \cdot n^{c_{143} \cdot \log n}}{\exp(0.5 \cdot (\log n)^3)} + \frac{1}{\exp((\log n)^3)} + \frac{16 \cdot \widehat{K}_n \cdot c_5^4 \cdot (\log n)^4}{n \cdot \widehat{K}_n} \\ \leq c_{150} \cdot \left(\frac{\exp(c_{143} \cdot (\log n)^2)}{\exp(0.5 \cdot (\log n)^3)} + \frac{(\log n)^4}{n} \right) \\ \leq \frac{c_{149}}{\sqrt{n}}.$$

Im vierten Schritt des Beweises zeigen wir

$$\mathbf{E} \left\{ \sup_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right\} \leq c_{151} \cdot (\log n)^5 \cdot \max_{(p,K) \in \mathcal{P}} n^{\frac{-p}{2p+K}}.$$

Sei im Folgenden $\mathcal{G} = \{f_{\mathbf{w},1,1} : \mathbf{w} \in \overline{\mathcal{W}}\}$. Für $\eta_n > 0$ ergibt sich damit

$$\mathbf{E} \left\{ \sup_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right\} = \mathbf{E} \left\{ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} g)(\mathbf{X}_i) \right\} \\ \leq \int_0^\infty \mathbf{P} \left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} g)(\mathbf{X}_i) \right| > t \right\} dt \\ \leq \eta_n + \int_{\eta_n}^\infty \mathbf{P} \left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} g)(\mathbf{X}_i) \right| > t \right\} dt.$$

Gemäß dem Beweis von Theorem 9.1 in Györfi et al. (2002) erhalten wir für $t \geq \eta_n$, dass die Ungleichung

$$\mathbf{P} \left\{ \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} g)(\mathbf{X}_i) \right| > t \right\} \\ \leq \sup_{\mathbf{x}_1^n \in [0,1]^{d \cdot n}} \mathcal{M}_1 \left(\frac{\eta_n}{2}, \{(T_{\beta_n} g) : g \in \mathcal{G}\}, \mathbf{x}_1^n \right) \cdot \sup_{g \in \mathcal{G}} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} g)(\mathbf{X}_i) \right| > \frac{t}{2} \right\}$$

gilt. Aus Lemma 6 folgt dann

$$\sup_{\mathbf{x}_1^n \in [0,1]^{d \cdot n}} \mathcal{M}_1 \left(\frac{\eta_n}{2}, \{(T_{\beta_n} g) : g \in \mathcal{G}\}, \mathbf{x}_1^n \right) \leq c_{152} \cdot \left(\frac{c_{153} \cdot \beta_n}{\eta_n} \right)^{V_{\mathcal{G}^+}}.$$

Die Netzwerke in der Menge \mathcal{G} haben höchstens $c_{154} \cdot r_n^2 \cdot L_n$ Gewichte und eine Tiefe von $L_n + 1$, weshalb die VC-Dimension $V_{\mathcal{G}^+}$ gemäß Lemma 31 aus dem Anhang von oben durch

$$\begin{aligned} V_{\mathcal{G}^+} &\leq c_{179} \cdot (L_n + 1) \cdot r_n^2 \cdot L_n \cdot \log_2(L_n \cdot r_n^2) \\ &\leq c_{155} \cdot r_n^2 \cdot L_n^2 \cdot \log(L_n \cdot r_n^2) \end{aligned}$$

beschränkt ist.

Auf den zweiten Term wenden wir die Ungleichung von Hoeffding (vgl. Lemma A.3 in Györfi et al. (2002)) an und nutzen dabei die Tatsache, dass $|(T_{\beta_n} g)(\mathbf{x})| \leq \beta_n$ für alle $\mathbf{x} \in \mathbb{R}^d$ ist. Daraus folgt

$$\mathbf{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} g)(\mathbf{X}_i) \right| > t/2 \right\} \leq 2 \cdot \exp \left(-\frac{2 \cdot n \cdot t^2}{16 \cdot \beta_n^2} \right).$$

Durch Zusammenfassen dieser Abschätzungen, erhalten wir dann

$$\begin{aligned} &\mathbf{E} \left\{ \sup_{\mathbf{w} \in \overline{\mathcal{W}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \cdot (T_{\beta_n} f_{\mathbf{w},1,1})(\mathbf{X}_i) \right| \right\} \\ &\leq \eta_n + \int_{\eta_n}^{\infty} c_{152} \cdot \left(\frac{c_{153} \cdot \beta_n}{\eta_n} \right)^{c_{155} \cdot r_n^2 \cdot L_n^2 \cdot \log(L_n \cdot r_n^2)} \cdot 2 \cdot \exp \left(-\frac{2 \cdot n \cdot t^2}{16 \cdot \beta_n^2} \right) dt \\ &\leq \eta_n + \int_{\eta_n}^{\infty} c_{152} \cdot \left(\frac{c_{153} \cdot \beta_n}{\eta_n} \right)^{c_{155} \cdot r_n^2 \cdot L_n^2 \cdot \log(L_n \cdot r_n^2)} \cdot 2 \cdot \exp \left(-\frac{2 \cdot n \cdot \eta_n \cdot t}{16 \cdot \beta_n^2} \right) dt \\ &\leq \eta_n + c_{152} \cdot \left(\frac{c_{153} \cdot \beta_n}{\eta_n} \right)^{c_{155} \cdot r_n^2 \cdot L_n^2 \cdot \log(L_n \cdot r_n^2)} \cdot \frac{16 \cdot \beta_n^2}{n \cdot \eta_n} \cdot \exp \left(-\frac{n \cdot \eta_n^2}{8 \cdot \beta_n^2} \right). \end{aligned}$$

Im Weiteren wählen wir

$$\eta_n = L_n \cdot \sqrt{\log(L_n \cdot r_n^2)} \cdot r_n \cdot (\log n)^2 \cdot \sqrt{\frac{8 \cdot \beta_n^2}{n}},$$

womit sich

$$\begin{aligned} &c_{152} \cdot \left(\frac{c_{153} \cdot \beta_n}{\eta_n} \right)^{c_{155} \cdot r_n^2 \cdot L_n^2 \cdot \log(L_n \cdot r_n^2)} \cdot \frac{16 \cdot \beta_n^2}{n \cdot \eta_n} \cdot \exp \left(-\frac{n \cdot \eta_n^2}{8 \cdot \beta_n^2} \right) \\ &\leq c_{156} \cdot \left(\frac{c_{157} \cdot \sqrt{n}}{r_n} \right)^{c_{155} \cdot r_n^2 \cdot L_n^2 \cdot \log(L_n \cdot r_n^2)} \cdot \frac{1}{\sqrt{n} \cdot r_n} \cdot \exp \left(-L_n^2 \cdot \log(L_n \cdot r_n^2) \cdot r_n^2 \cdot (\log n)^4 \right) \\ &\leq \exp \left(c_{158} \cdot r_n^2 \cdot (\log n)^3 \cdot \log(\sqrt{n}) \right) \cdot \frac{1}{\sqrt{n} \cdot r_n} \cdot \exp \left(-c_{139}^2 \cdot (\log n)^6 \cdot r_n^2 \right) \\ &\leq \exp \left(c_{158} \cdot r_n^2 \cdot (\log n)^4 \right) \cdot \frac{1}{\sqrt{n} \cdot r_n} \cdot \exp \left(-c_{139}^2 \cdot (\log n)^6 \cdot r_n^2 \right) \\ &\leq \exp \left(c_{159} \cdot r_n^2 \cdot (\log n)^4 \cdot (1 - (\log n)^2) \right) \cdot \frac{1}{\sqrt{n} \cdot r_n} \\ &\leq \frac{1}{\sqrt{n}} \end{aligned}$$

für n hinreichend groß ergibt.

Wegen $r_n \leq c_{145} \cdot n^{\frac{\bar{K}}{2(\bar{p}+K)}}$ erhalten wir zudem

$$\begin{aligned}\eta_n &= L_n \cdot \sqrt{\log(L_n \cdot r_n^2)} \cdot r_n \cdot (\log n)^2 \cdot \sqrt{\frac{8 \cdot \beta_n^2}{n}} \\ &\leq c_{160} \cdot (\log n)^5 \cdot \frac{n^{\frac{\bar{K}}{2(2\bar{p}+K)}}}{\sqrt{n}} \\ &\leq c_{160} \cdot (\log n)^5 \cdot n^{\frac{-\bar{p}}{2\bar{p}+K}}.\end{aligned}$$

Zusammenfassend ist damit die Ungleichung

$$\begin{aligned}\eta_n + c_{152} \cdot \left(\frac{c_{153} \cdot \beta_n}{\eta_n} \right)^{c_{155} \cdot L_n^2 \cdot r_n^2 \cdot \log(L_n \cdot r_n^2)} \cdot \frac{16 \cdot \beta_n^2}{n \cdot \eta_n} \cdot \exp\left(-\frac{n \cdot \eta_n^2}{8 \cdot \beta_n^2}\right) \\ \leq c_{160} \cdot (\log n)^5 \cdot n^{\frac{-\bar{p}}{2\bar{p}+K}} + \frac{1}{\sqrt{n}} \\ \leq c_{151} \cdot (\log n)^5 \cdot n^{\frac{-\bar{p}}{2\bar{p}+K}}\end{aligned}$$

für n hinreichend groß erfüllt, woraus die Aussage des vierten Beweisschrittes folgt.

Im *fünften Schritt des Beweises* wollen wir die Ungleichung

$$\mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*, 1, 1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \leq c_{161} \cdot (\log n)^2 \cdot \max_{(p, K) \in \mathcal{P}} n^{-\frac{2p}{2\bar{p}+K}}$$

nachweisen. Hierfür erinnern wir daran, dass $f_{\tilde{\mathbf{w}}}$ das in Lemma 15 eingeführte neuronale Netz bezeichnet.

Aus der Definition von \tilde{N}_i , der Beschränktheit von K_{\max} und p_{\max} sowie der elementaren Ungleichung

$$\lceil z \rceil^s \leq (2 \cdot z)^s \quad \text{für } z \geq 1 \text{ und } s > 0$$

ergeben sich die Abschätzungen

$$\begin{aligned}\ell \cdot \left(5 + \left\lceil \log_4 \left(\max_{j, i} M_{j, i}^{2p_j^{(i)}} \right) \right\rceil \cdot (\lceil \log_2(\max\{K_{\max}, p_{\max}\} + 1) \rceil + 1) \right) \\ \leq c_{162} \cdot \log \left(\max_{(p, K) \in \mathcal{P}} n^{\frac{p}{2\bar{p}+K}} \right) \\ \leq c_{139} \cdot \log n\end{aligned}$$

und

$$\begin{aligned}\max_{i \in \{1, \dots, \ell\}} \sum_{j=1}^{\tilde{N}_i} 2^{K_j^{(i)}} \cdot 64 \cdot \binom{K_j^{(i)} + q_j^{(i)}}{K_j^{(i)}} \cdot (K_j^{(i)})^2 \cdot (q_j^{(i)} + 1) \cdot M_{j, i}^{K_j^{(i)}} \\ \leq \max_{i \in \{1, \dots, \ell\}} \tilde{N}_i \cdot 2^{K_{\max}} \cdot 64 \cdot \binom{K_{\max} + p_{\max}}{K_{\max}} \cdot (K_{\max})^2 \cdot (p_{\max} + 1) \cdot \left[\max_{(p, K) \in \mathcal{P}} n^{\frac{K}{2(2\bar{p}+K)}} \right] \\ \leq c_{140} \cdot \max_{(p, K) \in \mathcal{P}} n^{\frac{K}{2(2\bar{p}+K)}}.\end{aligned}$$

Somit erfüllen L_n und r_n die Voraussetzungen in Lemma 15.

Des Weiteren gilt

$$\begin{aligned} & \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*,1,1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \\ & \leq 2 \cdot \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*,1,1})(\mathbf{x}) - (T_{\beta_n} f_{\tilde{\mathbf{w}},1,1})(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \\ & \quad + 2 \cdot \mathbf{E} \left\{ \int |(T_{\beta_n} f_{\tilde{\mathbf{w}},1,1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\}. \end{aligned}$$

Da der Träger von \mathbf{X} nach Voraussetzung beschränkt ist, können wir ohne Beschränkung der Allgemeinheit annehmen, dass $\text{supp}(\mathbf{X}) \subseteq [-a_n, a_n]^d$ ist. Durch Anwendung von Lemma 15 mit

$$M_{j,i} = \left\lceil n^{\frac{1}{2(2p_j^{(i)} + K_j^{(i)})}} \right\rceil \quad \text{sowie} \quad a_n = (\log n)^{\frac{1}{4(p_{\max}+1)}},$$

ergibt sich dann

$$\begin{aligned} \mathbf{E} \left\{ \int |(T_{\beta_n} f_{\tilde{\mathbf{w}},1,1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} & \leq c_{130}^2 \cdot a_n^{8(p_{\max}+1)} \cdot \max_{j,i} M_{j,i}^{-4p_j^{(i)}} \\ & \leq c_{163} \cdot (\log n)^2 \cdot \max_{j,i} n^{\frac{-2p_j^{(i)}}{2p_j^{(i)} + K_j^{(i)}}} \\ & = c_{163} \cdot (\log n)^2 \cdot \max_{(p,K) \in \mathcal{P}} n^{\frac{-2p}{2p+K}}. \end{aligned}$$

Wie bereits zu Beginn des Beweises gezeigt wurde, ist

$$\|(\tilde{w}_{1,i,j}^{(l)})_{i,j,l}\|_{\infty} \leq c_{138} \cdot n^{4/3}.$$

Zudem folgt aus der Definition der Menge \mathcal{W}^* , dass

$$\|((\mathbf{w}^*)_{1,i,j}^{(l)})_{i,j,l} - ((\tilde{w}_{1,i,j}^{(l)})_{i,j,k})_{i,j,k}\|_{\infty} \leq \bar{\delta}_n = n^{-(\log n)^2}$$

für $((\mathbf{w}^*)_{1,i,j}^{(l)})_{i,j,l} \in \mathcal{W}^*$ gilt. Mit Lemma 19 und der Wahl $B = c_{138} \cdot n^{4/3}$ sowie $M = \bar{\delta}_n$ ergibt sich daher die Abschätzung

$$\begin{aligned} |(T_{\beta_n} f_{\mathbf{w}^*,1,1})(\mathbf{x}) - (T_{\beta_n} f_{\tilde{\mathbf{w}},1,1})(\mathbf{x})|^2 & \leq |f_{\mathbf{w}^*,1,1}(\mathbf{x}) - f_{\tilde{\mathbf{w}},1,1}(\mathbf{x})|^2 \\ & \leq c_{164} \cdot (\log n)^2 \cdot n^{c_{165} \cdot \log n} \cdot (\log n)^{\frac{1}{2(p_{\max}+1)}} \cdot (n^{4/3} + \bar{\delta}_n)^{c_{166} \cdot \log n} \cdot \bar{\delta}_n^2 \\ & \leq c_{167} \cdot n^{c_{168} \cdot \log n} \cdot \bar{\delta}_n^2 \\ & \leq c_{167} \cdot n^{c_{168} \cdot \log n} \cdot \frac{1}{n^{2 \cdot (\log n)^2}} \\ & \leq \frac{c_{167}}{n} \end{aligned}$$

für hinreichend große n .

Insgesamt gilt damit

$$\begin{aligned}
& \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*, 1, 1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \\
& \leq 2 \cdot \mathbf{E} \left\{ \sup_{\mathbf{w}^* \in \mathcal{W}^*} \int |(T_{\beta_n} f_{\mathbf{w}^*, 1, 1})(\mathbf{x}) - (T_{\beta_n} f_{\tilde{\mathbf{w}}, 1, 1})(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \\
& \quad + 2 \cdot \mathbf{E} \left\{ \int |(T_{\beta_n} f_{\tilde{\mathbf{w}}, 1, 1})(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \right\} \\
& \leq c_{161} \cdot (\log n)^2 \cdot \max_{(p, K) \in \mathcal{P}} n^{\frac{-2p}{2p+K}},
\end{aligned}$$

womit der fünfte Schritt des Beweises gezeigt ist.

Im letzten Schritt des Beweises kombinieren wir alle bisherigen Resultate und erhalten

$$\begin{aligned}
& \mathbf{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mathbf{P}_{\mathbf{X}}(d\mathbf{x}) \\
& \leq c_{141} \left(\frac{(\log n)^2}{n} + (\log n)^5 \cdot \max_{(p, K) \in \mathcal{P}} n^{\frac{-p}{2p+K}} + (\log n)^2 \cdot \max_{(p, K) \in \mathcal{P}} n^{\frac{-2p}{2p+K}} + \frac{1}{\sqrt{n}} \right) \\
& \leq c_{141} \cdot (\log n)^5 \cdot \max_{(p, K) \in \mathcal{P}} n^{-\frac{p}{2p+K}}.
\end{aligned}$$

Damit ist die Aussage von Theorem 6 bewiesen. □

5 Fazit

Im Rahmen dieser Arbeit haben wir überparametrisierte neuronale Netze, die durch den Gradientenabstieg bezüglich des empirischen L_2 -Risikos ohne Regularisierungsterm trainiert wurden, in einem statistischen Kontext betrachtet. Dabei konnten wir insbesondere drei theoretische Hauptresultate formulieren, die die Leistungsfähigkeit überparametrisierter Netze beleuchten.

In den ersten beiden Resultaten haben wir als Grundlage für unseren Schätzer ein überparametrisiertes tiefes neuronales Netz mit der logistischen Sigmoidfunktion als Aktivierungsfunktion gewählt. Dieses Netzwerk besteht aus einer polynomiellen Anzahl tiefer vollständig verbundener neuronaler Netze, die parallel berechnet werden. Abschließend haben wir das neuronale Netz als Schätzer verwendet, dessen Gewichte nach einer geeigneten Anzahl von Gradientenschritten bestimmt wurden.

Im ersten Resultat konnten wir nachweisen, dass dieser Schätzer universell konsistent ist und somit in der Lage, auch auf neuen, unbekanntem Daten gut zu generalisieren.

Unter der Annahme, dass die Regressionsfunktion (p, C) -glatt für $p \in [1/2, 1]$ ist, konnten wir im zweiten Resultat nachweisen, dass der überparametrisierte Neuronale-Netze-Schätzer für alle $\varepsilon > 0$ mit einer Konvergenzgeschwindigkeit von

$$n^{-\frac{1}{1+d}+\varepsilon}$$

gegen die Regressionsfunktion konvergiert. Im Fall $p = 1/2$ entspricht diese Konvergenzrate annähernd der Minimax-Rate, wie sie in Stone (1982) gezeigt wurde, und ist somit nahezu optimal.

Wenn wir zusätzlich annehmen, dass die Regressionsfunktion einem Interaktionsmodell genügt, also als eine Summe Hölder-stetiger Funktionen für $p \in [1/2, 1]$ mit d^* Komponenten dargestellt werden kann, ist es möglich, für $\varepsilon > 0$ eine dimensionsunabhängige Konvergenzrate von

$$n^{-\frac{1}{1+d^*}+\varepsilon}$$

nachzuweisen. Dies zeigt, dass überparametrisierte Neuronale-Netze-Schätzer, die durch den Gradientenabstieg trainiert wurden, den Fluch der Dimensionalität umgehen können.

Im dritten Resultat dieser Arbeit haben wir eine (p, C) -glatte Regressionsfunktion für $p \geq 1$ betrachtet und zeigten unter der Annahme, dass die Regressionsfunktion zusätzlich einem hierarchischen Kompositionsmodell genügt, wie die Gewichte eines tiefen neuronalen Netzes durch die Überparametrisierung gelernt werden können. Dabei haben wir eine Konvergenzgeschwindigkeit von

$$n^{-\frac{p}{2p+K}}$$

hergeleitet und somit eine Dimensionsreduktion erzielt. Um diese Rate zu erreichen, haben wir in der Definition des Schätzers die ReLU-Aktivierungsfunktion verwendet und durch den Einsatz eines Projektionsoperators sichergestellt, dass die Gewichte während des Trainings nicht zu stark voneinander abweichen.

Für die Beweise der Hauptresultate haben wir die drei zentralen Forschungsbereiche des Deep Learnings berücksichtigt. Durch den Einsatz verschiedener Ansätze in diesen Bereichen konnten wir für beide Schätzer den Generalisierungs-, Optimierungs- und Approximationsfehler analysieren und damit sowohl die universelle Konsistenz des überparametrisierten tiefen Neuronale-Netze-Schätzers als auch die Konvergenzraten für unterschiedliche Aktivierungsfunktionen nachweisen.

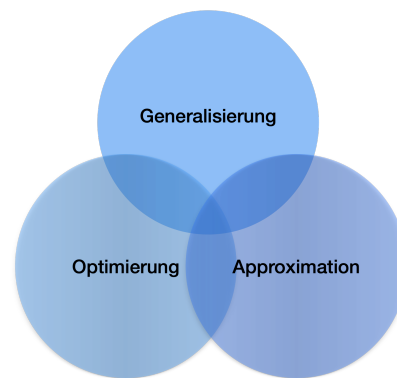


Abbildung 5.1: Fundamentale Forschungsbereiche des Deep Learnings

Ein zentraler Aspekt in den Beweisen der Resultate ist die Kontrolle des Generalisierungsfehlers, da die Überparametrisierung zu einer hohen Modellkapazität führt. In den ersten beiden Resultaten konnte durch die Verwendung der glatten logistischen Sigmoidfunktion die Komplexität des Netzes mittels einer metrischen Entropie-Schranke begrenzt werden. Diese Methode basiert auf der Glattheit der Aktivierungsfunktion und ist daher für die ReLU-Aktivierungsfunktion, die nicht glatt ist, nicht anwendbar. Im dritten Resultat wurde der Generalisierungsfehler daher durch die Rademacher-Komplexität abgeschätzt. Diese Abschätzung war maßgeblich dafür verantwortlich, dass die optimale Konvergenzrate nicht erreicht werden konnte. Daher wäre es für zukünftige Arbeiten von großem Interesse, eine präzisere Schranke zur Kontrolle dieses Generalisierungsfehlers zu entwickeln. Dies könnte wesentlich zur Verbesserung der Konvergenzrate beitragen und somit die Effizienz der Schätzverfahren steigern.

Um den Generalisierungsfehler mithilfe der Rademacher-Komplexität beschränken zu können, haben wir einen Projektionsoperator eingesetzt, der sicherstellt, dass sich die Gewichte nicht zu weit voneinander entfernen. Eine genauere Untersuchung der Projektionsoperatoren bei der Verwendung der ReLU-Aktivierungsfunktion wäre daher ein weiterer interessanter Aspekt. Hierbei stellt sich insbesondere die Frage, ob es möglich ist, ähnliche oder sogar bessere Ergebnisse zu erzielen, ohne die Projektion der Gewichte im Gradientenabstieg zu verwenden. Da diese Projektion jedoch ein notwendiger Bestandteil unseres dritten Ergebnisses war, um den Approximations- und Generalisierungsfehler zu kontrollieren, wird wahrscheinlich ein alternativer Ansatz erforderlich sein, um diese Fehler begrenzen zu können.

Ein weiterer wesentlicher Aspekt in der Betrachtung überparametrisierter neuronaler Netze ist die Wahl des Optimierungsalgorithmus. In praktischen Anwendungen kommen häufig der stochastische Gradientenabstieg oder Varianten mit Batches anstelle des klassischen Gradientenabstiegs zum Einsatz. Daher wäre es von Interesse zu untersuchen, ob sich die erzielten Ergebnisse auf den stochastischen Gradientenabstieg oder den Gradientenabstieg mit Batches übertragen lassen und gegebenenfalls sogar verbessert werden können.

Zusammenfassend haben wir in dieser Arbeit wertvolle Einblicke in die Eigenschaften überparametrisierter neuronaler Netze gegeben und deren Leistungsfähigkeit aufgezeigt. Durch die gemeinsame Betrachtung der drei Forschungsbereiche des Deep Learnings trägt diese Analyse zu einem umfassenderen Verständnis überparametrisierter tiefer neuronaler Netze bei. Dieser ganzheitliche Ansatz ist nicht nur für ein tieferes Verständnis der theoretischen Grundlagen von Bedeutung, sondern eröffnet auch neue Perspektiven für die praktische Anwendung und Weiterentwicklung dieser leistungsstarken Modelle.

A Ergänzende Resultate und Beweise

A.1 Beweis von Lemma 16

In diesem Abschnitt präsentieren wir den vollständigen Beweis von Lemma 16. Dabei orientieren wir uns am Beweis von Theorem 2 in Kohler und Langer (2021) und erweitern diesen um die Gewichtsschranken der jeweiligen Netzwerke. Die Gewichtsschranken in den Lemmata 23–27 gehen auf Kohler et al. (2024) zurück, während die entsprechenden Schranken in den Lemmata 28–30 erstmals in diesem Kontext hergeleitet werden. Aus Gründen der Vollständigkeit wird der gesamte Beweis im Folgenden ausführlich dargestellt.

Die Hauptidee des Beweises besteht darin, eine (p, C) -glatte Funktion f durch stückweise Taylor-Polynome auf einer Partition von Würfeln in $[-a, a]^d$ zu konstruieren. Hierfür benötigen wir die folgenden Konventionen: Ist C ein Würfel, so bezeichnen wir die linke untere Ecke von C durch C_{left} . Jeder halboffene Würfel C mit Seitenlänge μ kann als Polytop dargestellt werden, das durch die folgenden Ungleichungen definiert ist

$$-x^{(j)} + C_{\text{left}}^{(j)} \leq 0 \quad \text{und} \quad x^{(j)} - C_{\text{left}}^{(j)} - \mu < 0 \quad (j \in \{1, \dots, d\}).$$

Zudem bezeichnen wir durch $C_\delta^0 \subset C$ den Würfel, der alle $\mathbf{x} \in C$ enthält, die einen Abstand von mindestens δ zu den Grenzen von C haben. Dieses Polytop ist durch die Ungleichungen

$$-x^{(j)} + C_{\text{left}}^{(j)} \leq -\delta \quad \text{und} \quad x^{(j)} - C_{\text{left}}^{(j)} - \mu < -\delta \quad (j \in \{1, \dots, d\})$$

definiert.

Für eine Partition \mathcal{P} von Würfeln in $[-a, a]^d$ und $\mathbf{x} \in [-a, a]^d$ bezeichnen wir mit $C_{\mathcal{P}}(\mathbf{x})$ denjenigen Würfel $C \in \mathcal{P}$, der \mathbf{x} enthält.

A.1.1 Approximation einer (p, C) -glatten Funktion durch Taylorpolynome

Ein zentraler Bestandteil des Beweises von Lemma 16 ist das folgende Hilfsresultat. Es zeigt, dass jede (p, C) -glatte Funktion durch ein Taylorpolynom approximiert werden kann.

Lemma 20. Sei $p = q + s$ für ein $q \in \mathbb{N}_0$ und ein $s \in (0, 1]$ sowie $C > 0$. Weiter sei $f : \mathbb{R}^d \rightarrow \mathbb{R}$ eine (p, C) -glatte Funktion, $\mathbf{x}_0 \in \mathbb{R}^d$ und T_{f,q,\mathbf{x}_0} ein Taylorpolynom vom Grad q um \mathbf{x}_0 , welches durch

$$T_{f,q,\mathbf{x}_0}(\mathbf{x}) = \sum_{\xi \in \mathbb{N}_0^d: \|\xi\|_1 \leq q} (\partial^\xi f)(\mathbf{x}_0) \cdot \frac{(\mathbf{x} - \mathbf{x}_0)^\xi}{\xi!}$$

definiert ist. Dann gilt für jedes $\mathbf{x} \in \mathbb{R}^d$ die Ungleichung

$$|f(\mathbf{x}) - T_{f,q,\mathbf{x}_0}(\mathbf{x})| \leq c_{169} \cdot C \cdot \|\mathbf{x} - \mathbf{x}_0\|^p$$

für eine Konstante $c_{169} > 0$, die nur von q und d abhängt.

Beweis. Siehe Lemma 1 in Kohler (2014). □

A.1.2 Idee des Beweises von Lemma 16

Für den Beweis von Lemma 16 werden wir die (p, C) -glatte Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}$ in Lemma 20 durch ein stückweises Taylor-Polynom approximieren. Um das stückweise Taylor-Polynom zu definieren, unterteilen wir $[-a, a]^d$ in jeweils M^d und M^{2d} halboffene und gleichgroße Würfel der Form

$$[\boldsymbol{\alpha}, \boldsymbol{\beta}] = [\alpha^{(1)}, \beta^{(1)}] \times \cdots \times [\alpha^{(d)}, \beta^{(d)}] \quad \text{mit } \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d.$$

Seien des Weiteren

$$\mathcal{P}_1 = \{C_{k,1}\}_{k \in \{1, \dots, M^d\}} \quad \text{und} \quad \mathcal{P}_2 = \{C_{j,2}\}_{j \in \{1, \dots, M^{2d}\}} \quad (\text{A.1})$$

die zugehörigen Partitionen.

Wir bezeichnen die Würfel der Partition \mathcal{P}_2 , die in einem Würfel $C_{i,1}$ mit $i \in \{1, \dots, M^d\}$ enthalten sind, mit $\tilde{C}_{1,i}, \dots, \tilde{C}_{M^d,i}$. Im Folgenden sei $(C_{i,1})_{\text{left}}$ die linke untere Ecke des Würfels $C_{i,1}$. Zudem ordnen wir die Würfel derart an, dass

$$(\tilde{C}_{k,i})_{\text{left}} = (C_{i,1})_{\text{left}} + \tilde{\mathbf{v}}_k \quad (\text{A.2})$$

für alle $k \in \{1, \dots, M^d\}, i \in \{1, \dots, M^d\}$ und für einen Vektor $\tilde{\mathbf{v}}_k$ mit Einträgen aus $\{0, 2a/M^2, \dots, (M-1) \cdot 2a/M^2\}$ gilt. Der Vektor $\tilde{\mathbf{v}}_k$ beschreibt hierbei die Position von $(\tilde{C}_{k,i})_{\text{left}}$ im Vergleich zu $(C_{i,1})_{\text{left}}$. Wir ordnen die Würfel $\tilde{C}_{k,i}$ derart an, dass ihre Position unabhängig von i ist. Dies bedeutet nichts anderes, als dass der Vektor $\tilde{\mathbf{v}}_k$ für alle $i \in \{1, \dots, M^d\}$ gleich ist. Damit ist leicht erkennbar, dass die Partition \mathcal{P}_2 durch die Würfel $\tilde{C}_{k,i}$ dargestellt werden kann. Insbesondere ist

$$\mathcal{P}_2 = \{\tilde{C}_{k,i}\}_{k \in \{1, \dots, M^d\}, i \in \{1, \dots, M^d\}}.$$

Die Taylorentwicklung in Lemma 20 kann mithilfe des stückweise auf \mathcal{P}_2 definierten Taylor-Polynoms berechnet werden. Dies ergibt

$$T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x}) = \sum_{k \in \{1, \dots, M^d\}, i \in \{1, \dots, M^d\}} T_{f,q,(\tilde{C}_{k,i})_{\text{left}}}(\mathbf{x}) \cdot \mathbb{1}_{\tilde{C}_{k,i}}(\mathbf{x}).$$

Gemäß Lemma 20 erfüllt dieses stückweise Taylor-Polynom die Ungleichung

$$\sup_{\mathbf{x} \in [-a, a]^d} \left| f(\mathbf{x}) - T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x}) \right| \leq c_{169} \cdot (2 \cdot a \cdot d)^p \cdot C \cdot \frac{1}{M^{2p}}$$

für eine Konstante $c_{169} > 0$. Damit bleibt noch zu zeigen, dass das stückweise Taylor-Polynom bezüglich der Partition von $[-a, a]^d$ durch ein neuronales Netz approximiert werden kann. Dieser Beweis gliedert sich in vier wesentliche Schritte:

1. Berechnung von $T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x})$ durch rekursiv definierte Funktionen.
2. Approximation der rekursiven Funktionen durch neuronale Netze. Das resultierende Netzwerk ist eine gute Approximation für $f(\mathbf{x})$, falls

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{1/M^{2p+2}}^0.$$

3. Konstruktion eines neuronalen Netzes zur Approximation von $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$, wobei

$$w_{\mathcal{P}_2}(\mathbf{x}) = \prod_{j=1}^d \left(1 - \frac{M^2}{a} \cdot \left| (\mathbf{C}_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2} - x^{(j)} \right| \right)_+$$

ein linearer Tensorprodukt-B-Spline ist. Dieser nimmt seinen Maximalwert im Zentrum von $C_{\mathcal{P}_2}(\mathbf{x})$ an, ist ungleich 0 im Inneren von $C_{\mathcal{P}_2}(\mathbf{x})$ und verschwindet außerhalb von $C_{\mathcal{P}_2}(\mathbf{x})$.

4. Anwendung dieses Netzwerks auf 2^d leicht verschobene Partitionen von \mathcal{P}_2 , um $f(\mathbf{x})$ in der Supremumsnorm zu approximieren.

A.1.3 Schritt 1: Eine rekursive Definition von $T_{f,q,(\mathbf{C}_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x})$

Die folgende rekursive Definition des stückweisen Taylor-Polynoms wird uns später dabei helfen ein neuronales Netz zu definieren, welches die Funktion f approximiert.

Sei im Folgenden $i \in \{1, \dots, M^d\}$ und $C_{\mathcal{P}_1}(\mathbf{x}) = C_{i,1}$. Die Rekursion besteht aus zwei Schritten: Im ersten Schritt werden wir die Werte von $(\mathbf{C}_{\mathcal{P}_1}(\mathbf{x}))_{\text{left}} = (\mathbf{C}_{i,1})_{\text{left}}$ und die Werte von $(\partial^{\xi} f)((\tilde{\mathbf{C}}_{j,i})_{\text{left}})$ für $j \in \{1, \dots, M^d\}$ und $\xi \in \mathbb{N}_0^d$ mit $\|\xi\|_1 \leq q$ berechnen. Hierfür multiplizieren wir die Indikatorfunktion $\mathbb{1}_{C_{i,1}}$ mit $(\mathbf{C}_{i,1})_{\text{left}}$ oder mit $(\partial^{\xi} f)((\tilde{\mathbf{C}}_{j,i})_{\text{left}})$ für jedes $i \in \{1, \dots, M^d\}$. Zudem benötigen wir in der weiteren Rekursion den Eingabewert \mathbf{x} , weshalb wir diesen unverändert über die Identitätsfunktion in die nächste Schicht weitergeben. Wir setzen nun

$$\phi_{1,1} = (\phi_{1,1}^{(1)}, \dots, \phi_{1,1}^{(d)}) = \mathbf{x},$$

$$\phi_{2,1} = (\phi_{2,1}^{(1)}, \dots, \phi_{2,1}^{(d)}) = \sum_{i=1}^{M^d} (\mathbf{C}_{i,1})_{\text{left}} \cdot \mathbb{1}_{C_{i,1}}(\mathbf{x})$$

und

$$\phi_{3,1}^{(\xi,j)} = \sum_{i=1}^{M^d} (\partial^{\xi} f)((\tilde{\mathbf{C}}_{j,i})_{\text{left}}) \cdot \mathbb{1}_{C_{i,1}}(\mathbf{x})$$

für $j \in \{1, \dots, M^d\}$ und $\xi \in \mathbb{N}_0^d$ mit $\|\xi\|_1 \leq q$.

Seien im Weiteren $j, i \in \{1, \dots, M^d\}$ und $(\mathbf{C}_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}} = (\tilde{\mathbf{C}}_{j,i})_{\text{left}}$. Im zweiten Schritt der Rekursion berechnen wir den Wert von $(\mathbf{C}_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}} = (\tilde{\mathbf{C}}_{j,i})_{\text{left}}$ sowie die Werte von $(\partial^{\xi} f)((\mathbf{C}_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}})$ für $\xi \in \mathbb{N}_0^d$ mit $\|\xi\|_1 \leq q$. Aufgrund der Gleichheit in (A.2) folgt, dass jeder Würfel $\tilde{C}_{j,i}$ durch

$$\mathcal{A}^{(j)} = \left\{ \mathbf{x} \in \mathbb{R}^d : -x^{(k)} + \phi_{2,1}^{(k)} + \tilde{v}_j^{(k)} \leq 0 \right. \\ \left. \text{und } x^{(k)} - \phi_{2,1}^{(k)} - \tilde{v}_j^{(k)} - \frac{2a}{M^2} < 0 \text{ für alle } k \in \{1, \dots, d\} \right\} \quad (\text{A.3})$$

dargestellt werden kann. Daher berechnen wir in dieser Rekursion für jedes $j \in \{1, \dots, M^d\}$ die Indikatorfunktion $\mathbb{1}_{\mathcal{A}^{(j)}}$ und multiplizieren diese entweder mit $\phi_{2,1} + \tilde{\mathbf{v}}_j$ oder mit $\phi_{3,1}^{(\xi,j)}$ für $\xi \in \mathbb{N}_0^d$, $\|\xi\|_1 \leq q$. Auch hier möchten wir den Wert von \mathbf{x} unverändert durch die Anwendung der Identitätsfunktion weitergeben.

Wir setzen

$$\phi_{1,2} = (\phi_{1,2}^{(1)}, \dots, \phi_{1,2}^{(d)})^T = \phi_{1,1},$$

$$\phi_{2,2} = (\phi_{2,2}^{(1)}, \dots, \phi_{2,2}^{(d)}) = \sum_{j=1}^{M^d} (\phi_{2,1} + \mathbf{v}_j) \cdot \mathbb{1}_{\mathcal{A}^{(j)}}(\phi_{1,1})$$

und

$$\phi_{3,2}^{(\boldsymbol{\xi})} = \sum_{j=1}^{M^d} \phi_{3,1}^{(\boldsymbol{\xi},j)} \cdot \mathbb{1}_{\mathcal{A}^{(j)}}(\phi_{1,1})$$

für $\boldsymbol{\xi} \in \mathbb{N}_0^d$ mit $\|\boldsymbol{\xi}\|_1 \leq q$. Im letzten Schritt bestimmen wir das Taylor-Polynom durch

$$\phi_{1,3} = \sum_{\boldsymbol{\xi} \in \mathbb{N}_0^d: \|\boldsymbol{\xi}\|_1 \leq q} \frac{\phi_{3,2}^{(\boldsymbol{\xi})}}{\boldsymbol{\xi}!} \cdot (\phi_{1,2} - \phi_{2,2})^{\boldsymbol{\xi}}. \quad (\text{A.4})$$

Das folgende Lemma zeigt, dass wir durch diese Rekursion das stückweise Taylor-Polynom erhalten.

Lemma 21. Sei $p = q + s$ für ein $q \in \mathbb{N}_0$ und ein $s \in (0, 1]$, sei $C > 0$ und $\mathbf{x} \in [-a, a]^d$. Sei zudem $f : \mathbb{R}^d \rightarrow \mathbb{R}$ eine (p, C) -glatte Funktion und $T_{f,q,(\mathbf{C}_{\mathcal{P}_2(\mathbf{x}))_{\text{left}}}}$ das Taylor-Polynom vom Totalgrad q um den Punkt $(\mathbf{C}_{\mathcal{P}_2(\mathbf{x}))_{\text{left}}}$. Weiter sei $\phi_{1,3}$ rekursiv definiert, wie es in der Gleichung (A.4) beschrieben ist. Dann gilt

$$\phi_{1,3} = T_{f,q,(\mathbf{C}_{\mathcal{P}_2(\mathbf{x}))_{\text{left}}}}(\mathbf{x}).$$

Beweis. Seien $j, i \in \{1, \dots, M^d\}$ und $\mathbf{x} \in \tilde{\mathbf{C}}_{j,i}$. Daraus folgt, dass $\mathbf{C}_{\mathcal{P}_2(\mathbf{x})} = \tilde{\mathbf{C}}_{j,i}$ und $\mathbf{x} \in \mathbf{C}_{i,1}$ ist. Da \mathbf{x} nur in einem der Würfel $\mathbf{C}_{i,1}$ liegen kann, nimmt die Indikatorfunktion $\mathbb{1}_{\mathbf{C}_{i,1}}$ nur für einen Würfel den Wert 1 an. Damit folgt

$$\phi_{2,1} = (\mathbf{C}_{i,1})_{\text{left}}$$

sowie

$$\phi_{3,1}^{(\boldsymbol{\xi},j)} = (\partial^{\boldsymbol{\xi}} f)((\tilde{\mathbf{C}}_{j,i})_{\text{left}}) \quad \text{für } j \in \{1, \dots, M^d\}.$$

Nach Gleichung (A.2) ist daher

$$(\tilde{\mathbf{C}}_{j,i})_{\text{left}} = \phi_{2,1} + \tilde{\mathbf{v}}_j$$

und die Menge $\mathcal{A}^{(j)}$ beschreibt den Würfel $\tilde{\mathbf{C}}_{j,i}$. Zusammen mit $\phi_{1,1} = \mathbf{x}$ folgt aus den jeweiligen Definitionen, dass

$$\phi_{2,2} = (\tilde{\mathbf{C}}_{j,i})_{\text{left}}$$

und

$$\phi_{3,2}^{(\boldsymbol{\xi})} = (\partial^{\boldsymbol{\xi}} f)((\tilde{\mathbf{C}}_{j,i})_{\text{left}})$$

gilt. Setzen wir dies in Gleichung (A.4) ein, so erhalten wir

$$\phi_{1,3} = \sum_{\boldsymbol{\xi} \in \mathbb{N}_0^d: \|\boldsymbol{\xi}\|_1 \leq q} \frac{(\partial^{\boldsymbol{\xi}} f)((\tilde{\mathbf{C}}_{j,i})_{\text{left}})}{\boldsymbol{\xi}!} \cdot (\mathbf{x} - (\tilde{\mathbf{C}}_{j,i})_{\text{left}})^{\boldsymbol{\xi}} = T_{f,q,(\tilde{\mathbf{C}}_{j,i})_{\text{left}}}(\mathbf{x}) = T_{f,q,(\mathbf{C}_{\mathcal{P}_2(\mathbf{x}))_{\text{left}}}}(\mathbf{x}).$$

□

A.1.4 Schritt 2: Approximation von $\phi_{1,3}$ durch neuronale Netze

Die Grundidee des Beweises ist, ein zusammengesetztes neuronales Netz zu definieren, welches die Funktionen in den Definitionen von $\phi_{1,1}$, $\phi_{2,1}$, $\phi_{3,1}^{(\xi,j)}$, $\phi_{1,2}$, $\phi_{2,2}$, $\phi_{3,2}^{(\xi)}$, $\phi_{1,3}$ für $j \in \{1, \dots, M^d\}$ und $\xi \in \mathbb{N}_0^d$ mit $\|\xi\|_1 \leq q$ näherungsweise bestimmt. Wir werden zeigen, dass dieses neuronale Netz eine gute Approximation für $f(\mathbf{x})$ ist, sofern \mathbf{x} nicht in der Nähe des Randes einer der Würfel aus \mathcal{P}_2 liegt, das bedeutet, sofern gilt

$$\mathbf{x} \in \bigcup_{j \in \{1, \dots, M^{2d}\}} (C_{j,2})_{1/M^{2p+2}}^0.$$

Lemma 22. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Sei \mathcal{P}_2 definiert wie in (A.1). Sei zudem $p = q + s$ für ein $q \in \mathbb{N}_0$ und ein $s \in (0, 1]$ sowie $C > 0$. Die Funktion $f : \mathbb{R}^d \rightarrow \mathbb{R}$ sei eine (p, C) -glatte Funktion. Des Weiteren sei $1 \leq a < \infty$. Dann existiert für hinreichend großes $M \in \mathbb{N}$, das unabhängig von der Größe von a ist, aber

$$M^{2p} \geq c_{169} \cdot \max \left\{ \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4 \cdot (q+1)}, (2 \cdot a \cdot d)^p \cdot C \right\} \quad (\text{A.5})$$

erfüllt, ein neuronales Netzwerk $\widehat{f}_{\mathcal{P}_2} \in \mathcal{F}(L, r)$ mit

$$(i) \quad L = 4 + \lceil \log_4(M^{2p}) \rceil \cdot \lceil \log_2(\max\{q + 1, 2\}) \rceil$$

$$(ii) \quad r = \max \left\{ \left(\binom{d+q}{d} + d \right) \cdot M^d \cdot 2 \cdot (2 + 2d) + 2d, 18 \cdot (q + 1) \cdot \binom{d+q}{d} \right\},$$

so dass

$$\left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - f(\mathbf{x}) \right| \leq c_{170} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4 \cdot (q+1)} \cdot \frac{1}{M^{2p}}$$

für alle $\mathbf{x} \in \bigcup_{j \in \{1, \dots, M^{2d}\}} (C_{j,2})_{1/M^{2p+2}}^0$ gilt. Die Ausgabe des Netzes ist für alle $\mathbf{x} \in [-a, a]^d$ beschränkt durch

$$\left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| \leq 2 \cdot \exp(2ad) \cdot \max \left\{ \|f\|_{C^q([-a,a]^d)}, 1 \right\}.$$

Zusätzlich erfüllen die Gewichte des Netzes die folgenden Bedingungen

$$\|\mathbf{w}_{\widehat{f}_{\mathcal{P}_2}}\|_\infty \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_\infty \leq 1 \quad \text{und} \quad \left(\mathbf{w}_{\widehat{f}_{\mathcal{P}_2}} \right)_{1,0}^{(L)} = 0.$$

Lemma 22 zeigt, dass neuronale Netze mit einer Tiefe von $c_{171} \cdot \log_4(M)$ und einer Breite von $c_{172} \cdot M^d$ in der Lage sind, eine (p, C) -glatte Funktion f mit einer Genauigkeit der Größenordnung $1/M^{2p}$ zu approximieren, sofern $\mathbf{x} \in \bigcup_{j \in \{1, \dots, M^{2d}\}} (C_{j,2})_{1/M^{2p+2}}^0$. Für den Beweis dieses Resultats benötigen wir zusätzliche Aussagen über die Approximationsfähigkeiten neuronaler Netze.

Identitätsnetzwerk zur Schichtweitergabe und Tiefenanpassung

Um einen Eingabewert an die nächste Schicht weiterzugeben oder die Anzahl der verdeckten Schichten zweier Netzwerke anzugleichen, verwenden wir das Identitätsnetzwerk $\widehat{f}_{\text{id}} : \mathbb{R} \rightarrow \mathbb{R}$ mit

$$\widehat{f}_{\text{id}}(z) = \sigma(z) - \sigma(-z) = z \quad \text{für } z \in \mathbb{R}.$$

Für den Vektor $\mathbf{x} \in \mathbb{R}^d$ setzen wir entsprechend

$$\widehat{f}_{\text{id}}(\mathbf{x}) = \left(\widehat{f}_{\text{id}}(x^{(1)}), \dots, \widehat{f}_{\text{id}}(x^{(d)}) \right) = (x^{(1)}, \dots, x^{(d)}).$$

Die Gewichte des Netzes erfüllen hierbei die Bedingungen

$$\|\mathbf{w}_{\widehat{f}_{\text{id}}}\|_{\infty} = 1, \quad \left(\mathbf{w}_{\widehat{f}_{\text{id}}} \right)_{1,0}^{(1)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{id}}} \right)_{1,j>0}^{(1)} \right\|_{\infty} = 1.$$

Im Folgenden werden wir die Abkürzungen

$$\widehat{f}_{\text{id}}^0(\mathbf{x}) = \mathbf{x} \quad \text{für } \mathbf{x} \in \mathbb{R}^d$$

sowie

$$\widehat{f}_{\text{id}}^{t+1}(\mathbf{x}) = \widehat{f}_{\text{id}}\left(\widehat{f}_{\text{id}}^t(\mathbf{x})\right) = \mathbf{x} \quad \text{für } t \in \mathbb{N}_0 \text{ und } \mathbf{x} \in \mathbb{R}^d$$

verwenden.

Approximation der Quadratfunktion

Lemma 23. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Dann existiert für jedes $R \in \mathbb{N}$ und jedes $a \geq 1$ ein neuronales Netz

$$\widehat{f}_{\text{sq}} \in \mathcal{F}(R, 9)$$

mit den Gewichtsschranken

$$\left\| \mathbf{w}_{\widehat{f}_{\text{sq}}} \right\|_{\infty} \leq 4a^2, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{sq}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1 \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{sq}}} \right)_{1,0}^{(R)} \right\|_{\infty} = a^2,$$

so dass die Ungleichung

$$\left| \widehat{f}_{\text{sq}}(x) - x^2 \right| \leq a^2 \cdot 4^{-R}$$

für $x \in [-a, a]$ erfüllt ist.

Beweis. Dieser Beweis ergibt sich durch eine leichte Modifikation des Beweises von Proposition 2 in Yarotsky (2017). Der Vollständigkeit halber geben wir diesen Beweis dennoch im Folgenden an.

Sei $g : [0, 1] \rightarrow [0, 1]$ mit

$$g(x) = \begin{cases} 2x, & \text{falls } x \leq \frac{1}{2} \\ 2 \cdot (1 - x), & \text{falls } x > \frac{1}{2} \end{cases}$$

die sogenannte *Zackenfunktion* und

$$g_s(x) = \underbrace{g \circ g \circ \dots \circ g}_s(x)$$

die sogenannte *Sägezahnfunktion*. Im *ersten Schritt des Beweises* zeigen wir mithilfe einer Induktion über s , dass

$$g_s(x) = \begin{cases} 2^s \left(x - \frac{2k}{2^s}\right), & x \in \left[\frac{2k}{2^s}, \frac{2k+1}{2^s}\right], k \in \{0, 1, \dots, 2^{s-1} - 1\} \\ 2^s \left(\frac{2k}{2^s} - x\right), & x \in \left[\frac{2k-1}{2^s}, \frac{2k}{2^s}\right], k \in \{1, 2, \dots, 2^{s-1}\} \end{cases}$$

gilt. Für $s = 1$ folgt die Aussage direkt aus der Definition von g , da $g_1(x) = g(x)$ für $x \in [0, 1]$ ist. Im Folgenden nehmen wir an, dass die Aussage für ein $s \in \mathbb{N}$ erfüllt ist. Um die Aussage für $s + 1$ zu zeigen, werden wir die Eigenschaften der Sägezahnfunktion ausnutzen. Aufgrund ihrer Definition ist

$$(g_s \circ g)(x) = g_s(g(x)) = g_s(2x) \quad \text{für } x \in \left[0, \frac{1}{2}\right]$$

für beliebiges $s \in \mathbb{N}$. Des Weiteren gilt

$$g(x) = g(1 - x) \quad \text{für } x \in [0, 1]$$

und aufgrund der Induktionsannahme sowie der Symmetrie von g_s ist zusätzlich

$$g_s(x) = g_s(1 - x) \quad \text{für } x \in [0, 1]$$

erfüllt. Kombinieren wir diese Eigenschaften, so erhalten wir

$$\begin{aligned} g_{s+1}(x) &= g_s(g(x)) = g_s(2x) = g_s(1 - 2x) = g_s\left(2 \cdot \left(\frac{1}{2} - x\right)\right) \\ &= g_s\left(g\left(\frac{1}{2} - x\right)\right) = g_s\left(g\left(x + \frac{1}{2}\right)\right) = g_{s+1}\left(x + \frac{1}{2}\right) \end{aligned}$$

für jedes $x \in [0, \frac{1}{2}]$. Daher genügt es im Folgenden, die Aussage für $x \in [0, \frac{1}{2}]$ zu zeigen. Zusammen mit der Induktionsannahme erhalten wir dann

$$\begin{aligned} (g_s \circ g)(x) = g_s(2x) &= \begin{cases} 2^s \cdot \left(2x - \frac{2k}{2^s}\right), & 2x \in \left[\frac{2k}{2^s}, \frac{2k+1}{2^s}\right], k \in \{0, 1, \dots, 2^{s-1} - 1\} \\ 2^s \cdot \left(\frac{2k}{2^s} - 2x\right), & 2x \in \left[\frac{2k-1}{2^s}, \frac{2k}{2^s}\right], k \in \{1, 2, \dots, 2^{s-1}\} \end{cases} \\ &= \begin{cases} 2^{s+1} \cdot \left(x - \frac{2k}{2^{s+1}}\right), & x \in \left[\frac{2k}{2^{s+1}}, \frac{2k+1}{2^{s+1}}\right], k \in \{0, 1, \dots, 2^s - 1\} \\ 2^{s+1} \cdot \left(\frac{2k}{2^{s+1}} - x\right), & x \in \left[\frac{2k-1}{2^{s+1}}, \frac{2k}{2^{s+1}}\right], k \in \{1, 2, \dots, 2^s\}. \end{cases} \end{aligned}$$

Damit folgt die Aussage des ersten Schrittes.

Im *zweiten Schritt des Beweises* zeigen wir, dass die Funktion $f(x) = x^2$ für $x \in [0, 1]$ durch eine Linearkombination von Funktionen g_s approximiert werden kann. Sei hierfür S_R eine stückweise lineare Interpolation der Funktion f mit $2^R + 1$ gleichmäßig verteilten Stützstellen an den Punkten $\frac{k}{2^R}$ für $k = 0, \dots, 2^R$, wobei

$$S_R\left(\frac{k}{2^R}\right) = \left(\frac{k}{2^R}\right)^2.$$

Um den Fehler der stückweisen linearen Interpolation bestimmen zu können, definieren wir die Funktion

$$F(z) := f(z) - S_R(z) + \frac{S_R(x) - f(x)}{\left(x - \frac{k}{2^R}\right) \cdot \left(x - \frac{k+1}{2^R}\right)} \cdot \left(z - \frac{k}{2^R}\right) \cdot \left(z - \frac{k+1}{2^R}\right)$$

für $x \in [\frac{k}{2^R}, \frac{k+1}{2^R}]$ und $k = 0, \dots, 2^R - 1$.
Damit ergibt sich

$$F\left(\frac{k}{2^R}\right) = 0, \quad F\left(\frac{k+1}{2^R}\right) = 0 \quad \text{sowie} \quad F(x) = 0.$$

Nach dem Satz von Rolle (siehe beispielsweise Forster (2023)) existieren dann zwei Punkte z_1 und z_2 mit $\frac{k}{2^R} < z_1 < x$ und $F'(z_1) = 0$ sowie $x < z_2 < \frac{k+1}{2^R}$ und $F'(z_2) = 0$. Die erneute Anwendung des Satzes von Rolle liefert einen Punkt η , für den $z_1 < \eta < z_2$ mit $F''(\eta) = 0$ gilt. Mit

$$F''(z) = f''(z) + 2 \cdot \frac{S_R(x) - f(x)}{\left(x - \frac{k}{2^R}\right) \cdot \left(x - \frac{k+1}{2^R}\right)}$$

erhalten wir für $x \in [\frac{k}{2^R}, \frac{k+1}{2^R}]$, dass

$$\begin{aligned} |S_R(x) - f(x)| &= \left| -\frac{f''(\eta)}{2} \cdot \left(x - \frac{k}{2^R}\right) \cdot \left(x - \frac{k+1}{2^R}\right) \right| \\ &\leq \left| \left(x - \frac{k}{2^R}\right) \cdot \left(x - \frac{k+1}{2^R}\right) \right| \\ &= \left(x - \frac{k}{2^R}\right) \cdot \left(\frac{k+1}{2^R} - x\right) \end{aligned}$$

gilt. Das Maximum der Funktion

$$h(x) := \left(x - \frac{k}{2^R}\right) \cdot \left(\frac{k+1}{2^R} - x\right)$$

wird an der Stelle $x = \frac{k}{2^R} + \frac{1}{2} \cdot \frac{1}{2^R}$ angenommen. Daher gilt

$$\begin{aligned} |S_R(x) - f(x)| &\leq \left| \left(x - \frac{k}{2^R}\right) \cdot \left(x - \frac{k+1}{2^R}\right) \right| \\ &\leq \left| \left(\frac{k}{2^R} + \frac{1}{2} \cdot \frac{1}{2^R} - \frac{k}{2^R}\right) \cdot \left(\frac{k}{2^R} + \frac{1}{2} \cdot \frac{1}{2^R} - \frac{k+1}{2^R}\right) \right| \\ &\leq 2^{-2R-2} \end{aligned}$$

für $x \in [\frac{k}{2^R}, \frac{k+1}{2^R}]$. Die Verfeinerung der Interpolation von S_{R-1} zu S_R entspricht der Anpassung durch eine Funktion, die proportional zu einer Sägezahnfunktion ist, das heißt

$$S_{R-1}(x) - S_R(x) = \frac{g_R(x)}{2^{2R}}.$$

Dieser Zusammenhang ergibt sich für ein $x \in [\frac{k}{2^{R-1}}, \frac{k+1}{2^{R-1}}]$ mit $k \in \{0, \dots, 2^{R-1} - 1\}$ aus

$$\begin{aligned} S_{R-1}(x) &= S_{R-1}\left(\frac{k}{2^{R-1}}\right) + \frac{S_{R-1}\left(\frac{k+1}{2^{R-1}}\right) - S_{R-1}\left(\frac{k}{2^{R-1}}\right)}{\frac{1}{2^{R-1}}} \cdot \left(x - \frac{k}{2^{R-1}}\right) \\ &= \left(\frac{k}{2^{R-1}}\right)^2 + 2^{R-1} \cdot \left(\left(\frac{k+1}{2^{R-1}}\right)^2 - \left(\frac{k}{2^{R-1}}\right)^2 \right) \cdot \left(x - \frac{k}{2^{R-1}}\right) \\ &= \left(\frac{k}{2^{R-1}}\right)^2 + \left(\frac{2k+1}{2^{R-1}}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) \end{aligned}$$

und

$$\begin{aligned}
 S_R(x) &= \begin{cases} S_R\left(\frac{k}{2^{R-1}}\right) + \frac{S_R\left(\frac{k}{2^{R-1}} + \frac{1}{2^R}\right) - S_R\left(\frac{k}{2^{R-1}}\right)}{\frac{1}{2^R}} \cdot \left(x - \frac{k}{2^{R-1}}\right), \\ \text{falls } x \in \left[\frac{k}{2^{R-1}}, \frac{k}{2^{R-1}} + \frac{1}{2^R}\right] \\ S_R\left(\frac{k}{2^{R-1}} + \frac{1}{2^R}\right) + \frac{S_R\left(\frac{k+1}{2^{R-1}}\right) - S_R\left(\frac{k}{2^{R-1}} + \frac{1}{2^R}\right)}{\frac{1}{2^R}} \cdot \left(x - \frac{k}{2^{R-1}} - \frac{1}{2^R}\right), \\ \text{falls } x \in \left[\frac{k}{2^{R-1}} + \frac{1}{2^R}, \frac{k+1}{2^{R-1}}\right] \end{cases} \\
 &= \begin{cases} \left(\frac{k}{2^{R-1}}\right)^2 + \left(\frac{2k}{2^{R-1}} + \frac{1}{2^R}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right), & \text{falls } x \in \left[\frac{k}{2^{R-1}}, \frac{k}{2^{R-1}} + \frac{1}{2^R}\right] \\ \left(\frac{k}{2^{R-1}}\right)^2 - \frac{2}{2^{2R}} + \left(\frac{4k+3}{2^R}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right), & \text{falls } x \in \left[\frac{k}{2^{R-1}} + \frac{1}{2^R}, \frac{k+1}{2^{R-1}}\right]. \end{cases}
 \end{aligned}$$

Für $x \in \left[\frac{k}{2^{R-1}}, \frac{k}{2^{R-1}} + \frac{1}{2^R}\right] = \left[\frac{2k}{2^R}, \frac{2k+1}{2^R}\right]$ ist gemäß des ersten Beweisschrittes

$$g_R(x) = 2^R \cdot \left(x - \frac{2k}{2^R}\right),$$

woraus wir

$$\begin{aligned}
 S_{R-1}(x) - S_R(x) &= \left(\frac{k}{2^{R-1}}\right)^2 + \left(\frac{2k+1}{2^{R-1}}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) - \left(\frac{k}{2^{R-1}}\right)^2 - \left(\frac{2k}{2^{R-1}} + \frac{1}{2^R}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) \\
 &= \left(\frac{2k+1}{2^{R-1}}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) + \left(\frac{2k}{2^{R-1}} + \frac{1}{2^R}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) \\
 &= \left(\left(\frac{2k+1}{2^{R-1}}\right) - \left(\frac{2k}{2^{R-1}} + \frac{1}{2^R}\right)\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) \\
 &= \frac{1}{2^R} \cdot \left(x - \frac{k}{2^{R-1}}\right) = \frac{2^R}{2^{2R}} \cdot \left(x - \frac{2k}{2^R}\right) = \frac{g_R(x)}{2^{2R}}
 \end{aligned}$$

erhalten. Zudem ergibt sich für $x \in \left[\frac{k}{2^{R-1}} + \frac{1}{2^R}, \frac{k+1}{2^{R-1}}\right] = \left[\frac{2k+1}{2^R}, \frac{2k+2}{2^R}\right]$ aus dem ersten Schritt des Beweises

$$g_R(x) = 2^R \cdot \left(\frac{2k+2}{2^R} - x\right).$$

Somit ist

$$\begin{aligned}
 &S_{R-1}(x) - S_R(x) \\
 &= \left(\frac{k}{2^{R-1}}\right)^2 + \left(\frac{2k+1}{2^{R-1}}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) - \left(\frac{k}{2^{R-1}}\right)^2 + \frac{2}{2^{2R}} - \left(\frac{4k+3}{2^R}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) \\
 &= \left(\frac{2k+1}{2^{R-1}}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) + \frac{2}{2^{2R}} - \left(\frac{4k+3}{2^R}\right) \cdot \left(x - \frac{k}{2^{R-1}}\right) \\
 &= \left(\left(\frac{4k+3}{2^R}\right) - \left(\frac{2k+1}{2^{R-1}}\right)\right) \cdot \left(\frac{k}{2^{R-1}} - x\right) + \frac{2}{2^{2R}} \\
 &= \frac{1}{2^R} \cdot (4k+3 - 2(2k+1)) \cdot \left(\frac{k}{2^{R-1}} - x\right) + \frac{2}{2^{2R}} = \frac{1}{2^R} \cdot \left(\frac{2k+2}{2^R} - x\right) = \frac{g_R(x)}{2^{2R}},
 \end{aligned}$$

womit wir den Zusammenhang

$$S_{R-1}(x) - S_R(x) = \frac{g_R(x)}{2^{2R}}$$

gezeigt haben.

Da $S_0(x) = x$ ist, können wir rekursiv folgern, dass

$$S_R(x) = x - \sum_{s=1}^R \frac{g_s(x)}{2^{2s}}$$

mit

$$|S_R(x) - x^2| \leq 2^{-2R-2}$$

für $x \in [0, 1]$ gilt. Das heißt, jede Funktion $f(x) = x^2$ kann für $x \in [0, 1]$ durch eine Linearkombination von g_s approximiert werden.

Im *dritten Schritt des Beweises* zeigen wir, dass ein vorwärtsgerichtetes neuronales Netz existiert, das $S_R(x)$ für $x \in [0, 1]$ berechnet. Die Funktion $g(x)$ kann durch das Netzwerk

$$\widehat{f}_g(x) = 2 \cdot \sigma(x) - 4 \cdot \sigma\left(x - \frac{1}{2}\right) + 2 \cdot \sigma(x - 1)$$

mit den Gewichtsschranken

$$\left\| \mathbf{w}_{\widehat{f}_g}^{(1)} \right\|_{\infty} \leq 4, \quad \left(\mathbf{w}_{\widehat{f}_g} \right)_{1,0}^{(1)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_g} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1$$

dargestellt werden.

Die Funktion $g_s(x)$ kann dann durch ein Netzwerk

$$\widehat{f}_{g_s} \in \mathcal{F}(s, 3)$$

mit

$$\widehat{f}_{g_s}(x) = \underbrace{\widehat{f}_g(\widehat{f}_g(\dots(\widehat{f}_g(x))))}_s$$

bestimmt werden. Durch die Anwendung von Lemma 17c) erhalten wir für dieses Netzwerk die folgenden Gewichtsschranken

$$\left\| \mathbf{w}_{\widehat{f}_{g_s}}^{(s)} \right\|_{\infty} \leq 4, \quad \left(\mathbf{w}_{\widehat{f}_{g_s}} \right)_{1,0}^{(s)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{g_s}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1.$$

Mithilfe des Identitätsnetzwerks können wir die Funktion $S_R(x)$ durch das Netzwerk

$$\widehat{f}_{\text{sq}_{[0,1]}} \in \mathcal{F}(R, 7)$$

darstellen, das folgendermaßen definiert ist: Zunächst setzen wir $\widehat{f}_{1,0}(x) = \widehat{f}_{2,0}(x) = x$ und $\widehat{f}_{3,0}(x) = 0$. Dann definieren wir rekursiv für $i \in \{0, 1, \dots, R-2\}$

$$\widehat{f}_{1,i+1}(x) = \widehat{f}_{\text{id}}(\widehat{f}_{1,i}(x)),$$

$$\widehat{f}_{2,i+1}(x) = \widehat{f}_g(\widehat{f}_{2,i}(x))$$

und

$$\widehat{f}_{3,i+1}(x) = \widehat{f}_{\text{id}}(\widehat{f}_{3,i}(x)) - \frac{\widehat{f}_g(\widehat{f}_{2,i}(x))}{2^{2(i+1)}}.$$

Schließlich setzen wir

$$\widehat{f}_{\text{sq}_{[0,1]}}(x) = \widehat{f}_{\text{id}}(\widehat{f}_{1,R-1}(x)) - \frac{\widehat{f}_g(\widehat{f}_{2,R-1}(x))}{2^{2R}} + \widehat{f}_{\text{id}}(\widehat{f}_{3,R-1}(x)). \quad (\text{A.6})$$

Daraus folgt durch die Verwendung der positiven Homogenität der ReLU-Aktivierungsfunktion, dass

$$\begin{aligned} \widehat{f}_{\text{sq}_{[0,1]}}(x) &= \widehat{f}_{\text{id}}^R(x) - \frac{1}{2^{2R}} \cdot \widehat{f}_{gR}(x) - \widehat{f}_{\text{id}} \left(\frac{1}{2^{2(R-1)}} \cdot \widehat{f}_{g_{R-1}}(x) \right. \\ &\quad \left. - \widehat{f}_{\text{id}} \left(\frac{1}{2^{2(R-2)}} \cdot \widehat{f}_{g_{R-2}}(x) - \dots - \widehat{f}_{\text{id}} \left(\frac{1}{2^2} \cdot \widehat{f}_{g_1}(x) \right) \right) \right) \\ &= S_R(x) \end{aligned}$$

gilt. Also erfüllt $\widehat{f}_{\text{sq}_{[0,1]}}(x)$ die Ungleichung

$$|\widehat{f}_{\text{sq}_{[0,1]}}(x) - x^2| \leq 2^{-2R-2} \quad (\text{A.7})$$

für $x \in [0, 1]$.

Wegen der Definition des Netzwerks $\widehat{f}_{\text{sq}_{[0,1]}}$ in (A.6) lassen sich die Gewichtsschranken des Netzes durch die Gewichtsschranken des Netzwerks

$$\frac{\widehat{f}_g(\widehat{f}_{2,R-1}(x))}{2^{2R}} = \frac{\widehat{f}_{2,R}(x)}{2^{2R}} =: \widetilde{f}_{2,R}(x)$$

bestimmen.

Wir zeigen nun mittels Induktion, dass die Gewichte von $\widetilde{f}_{2,i}(x)$ für $i \in \{1, \dots, R\}$ durch 1 beschränkt sind.

Sei hierfür $i = 1$. Dann gilt

$$\widetilde{f}_{2,1}(x) = \frac{\widehat{f}_{2,1}(x)}{2^{2 \cdot 1}} = \frac{\widehat{f}_g(x)}{4} = \frac{1}{2} \cdot \sigma(x) - \sigma\left(x - \frac{1}{2}\right) + \frac{1}{2} \cdot \sigma(x - 1),$$

womit $\|\mathbf{w}_{\widetilde{f}_{2,1}(x)}\|_\infty \leq 1$ ist.

Im Folgenden nehmen wir an, dass die Gewichte von $\widetilde{f}_{2,i}$ für ein $i \in \{1, \dots, R\}$ durch 1 beschränkt sind.

Durch die Definition von \widehat{f}_g erhalten wir

$$\begin{aligned} \widetilde{f}_{2,i+1}(x) &= \frac{\widehat{f}_{2,i+1}(x)}{2^{2(i+1)}} = \frac{1}{2^2} \cdot \frac{1}{2^{2i}} \cdot \widehat{f}_g(\widehat{f}_{2,i}(x)) \\ &= \frac{1}{2^{2i}} \cdot \left(\frac{1}{2^2} \cdot 2 \cdot \sigma\left(\widehat{f}_{2,i}(x)\right) - \frac{1}{2^2} \cdot 4 \cdot \sigma\left(\widehat{f}_{2,i}(x) - \frac{1}{2}\right) + \frac{1}{2^2} \cdot 2 \cdot \sigma\left(\widehat{f}_{2,i}(x) - 1\right) \right) \\ &= \frac{1}{2^{2i}} \cdot \left(\frac{1}{2} \cdot \sigma\left(\widehat{f}_{2,i}(x)\right) - \sigma\left(\widehat{f}_{2,i}(x) - \frac{1}{2}\right) + \frac{1}{2} \cdot \sigma\left(\widehat{f}_{2,i}(x) - 1\right) \right). \end{aligned}$$

Wenden wir hierauf die positive Homogenität der ReLU-Aktivierungsfunktion an, so führt dies zu

$$\begin{aligned}\tilde{f}_{2,i+1}(x) &= \frac{1}{2^{2i}} \cdot \left(\frac{1}{2} \cdot \sigma \left(\hat{f}_{2,i}(x) \right) - \sigma \left(\hat{f}_{2,i}(x) - \frac{1}{2} \right) + \frac{1}{2} \cdot \sigma \left(\hat{f}_{2,i}(x) - 1 \right) \right) \\ &= \frac{1}{2} \cdot \sigma \left(\frac{1}{2^{2i}} \cdot \hat{f}_{2,i}(x) \right) - \sigma \left(\frac{1}{2^{2i}} \cdot \hat{f}_{2,i}(x) - \frac{1}{2^{2i}} \cdot \frac{1}{2} \right) + \frac{1}{2} \cdot \sigma \left(\frac{1}{2^{2i}} \cdot \hat{f}_{2,i}(x) - \frac{1}{2^{2i}} \right) \\ &= \frac{1}{2} \cdot \sigma \left(\tilde{f}_{2,i}(x) \right) - \sigma \left(\tilde{f}_{2,i}(x) - \frac{1}{2^{2i}} \cdot \frac{1}{2} \right) + \frac{1}{2} \cdot \sigma \left(\tilde{f}_{2,i}(x) - \frac{1}{2^{2i}} \right).\end{aligned}$$

Daraus folgt direkt, dass $\|\mathbf{w}_{\tilde{f}_{2,R}(x)}\|_\infty \leq 1$ gilt. Zudem ergibt sich hieraus, dass $\left(\mathbf{w}_{\tilde{f}_{2,R}}\right)_{1,0}^{(R)} = 0$ ist. Wegen der Induktionsannahme gilt außerdem $\left\| \left(\mathbf{w}_{\tilde{f}_{2,1}(x)}\right)_{i,j>0}^{(0)} \right\|_\infty \leq 1$.

Unter Berücksichtigung der Gewichtsschranken von \hat{f}_{id} erhalten wir damit

$$\left\| \mathbf{w}_{\hat{f}_{\text{sq}[0,1]}} \right\|_\infty \leq \frac{1}{4} \cdot \left\| \mathbf{w}_{\hat{f}_g} \right\|_\infty = 1, \quad \left(\mathbf{w}_{\hat{f}_{\text{sq}[0,1]}}\right)_{1,0}^{(R)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{sq}[0,1]}}\right)_{i,j>0}^{(0)} \right\|_\infty \leq 1.$$

Im letzten Schritt des Beweises zeigen wir, dass die Funktion $f(x) = x^2$ durch ein neuronales Netzwerk approximiert werden kann, wenn $x \in [-a, a]$ ist. Sei hierfür $f_{\text{tran}} : [-a, a] \rightarrow [0, 1]$ mit

$$f_{\text{tran}}(z) = \frac{z}{2a} + \frac{1}{2}$$

die Funktion, die den Wert von $x \in [-a, a]$ in das Intervall $[0, 1]$ überführt, in dem Ungleichung (A.7) gültig ist. Hierfür setzen wir

$$\hat{f}_{\text{sq}}(x) = 4a^2 \cdot \hat{f}_{\text{sq}[0,1]}(f_{\text{tran}}(x)) - 2a \cdot \hat{f}_{\text{id}}^R(x) - a^2.$$

Aufgrund der Tatsache, dass

$$x^2 = 4a^2 \cdot \left(\frac{x}{2a} + \frac{1}{2} \right)^2 - 2 \cdot ax - a^2$$

gilt, erhalten wir

$$\begin{aligned}\left| \hat{f}_{\text{sq}}(x) - x^2 \right| &= \left| 4a^2 \cdot \hat{f}_{\text{sq}[0,1]}(f_{\text{tran}}(x)) - 2a \cdot \hat{f}_{\text{id}}^R(x) - 4a^2 \cdot \left(\frac{x}{2a} + \frac{1}{2} \right)^2 + 2a \cdot x \right| \\ &\leq 4a^2 \cdot \left| \hat{f}_{\text{sq}[0,1]}(f_{\text{tran}}(x)) - (f_{\text{tran}}(x))^2 \right| + 2a \cdot \left| \hat{f}_{\text{id}}^R(x) - x \right| \\ &\leq 4a^2 \cdot 2^{-2R-2} = a^2 \cdot 4^{-R}.\end{aligned}$$

Die Schranken der Gewichte des Netzes \hat{f}_{sq} ergeben sich direkt aus der Definition des Netzwerks sowie den Gewichtsschranken des Netzes $\hat{f}_{\text{sq}[0,1]}$ und sind gegeben durch

$$\left\| \mathbf{w}_{\hat{f}_{\text{sq}}} \right\|_\infty \leq 4a^2, \quad \left| \left(\mathbf{w}_{\hat{f}_{\text{sq}}}\right)_{1,0}^{(R)} \right| = a^2 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{sq}}}\right)_{i,j>0}^{(0)} \right\|_\infty \leq 1.$$

□

Approximation einer Multiplikation

Das nächste Lemma stellt ein Netzwerk vor, das für gegebene Eingaben x und y eine Approximation von $x \cdot y$ liefert.

Lemma 24. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Dann existiert für alle $R \in \mathbb{N}$ und alle $a \geq 1$ ein neuronales Netz

$$\widehat{f}_{\text{mult}} \in \mathcal{F}(R, 18)$$

mit den Gewichtsschranken

$$\|\mathbf{w}_{\widehat{f}_{\text{mult}}}\|_{\infty} \leq 4a^2, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{mult}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1 \quad \text{und} \quad \left(\mathbf{w}_{\widehat{f}_{\text{mult}}} \right)_{1,0}^{(R)} = 0,$$

so dass

$$\left| \widehat{f}_{\text{mult}}(x, y) - x \cdot y \right| \leq 2 \cdot a^2 \cdot 4^{-R}$$

für alle $x, y \in [-a, a]$ gilt.

Beweis. Sei

$$\widehat{f}_{\text{sq}} \in \mathcal{F}(R, 9)$$

das neuronale Netz aus Lemma 23, welches

$$\left| \widehat{f}_{\text{sq}}(x) - x^2 \right| \leq 4 \cdot a^2 \cdot 4^{-R}$$

für $x \in [-2a, 2a]$ erfüllt. Wir setzen

$$\widehat{f}_{\text{mult}}(x, y) = \frac{1}{4} \cdot \left(\widehat{f}_{\text{sq}}(x + y) - \widehat{f}_{\text{sq}}(x - y) \right).$$

Da sich die Multiplikation durch

$$x \cdot y = \frac{1}{4} \left((x + y)^2 - (x - y)^2 \right)$$

darstellen lässt, gilt mithilfe des Netzwerks aus Lemma 23

$$\begin{aligned} \left| \widehat{f}_{\text{mult}}(x, y) - x \cdot y \right| &= \frac{1}{4} \cdot \left| \widehat{f}_{\text{sq}}(x + y) - \widehat{f}_{\text{sq}}(x - y) - ((x + y)^2 - (x - y)^2) \right| \\ &\leq \frac{1}{4} \cdot \left| \widehat{f}_{\text{sq}}(x + y) - (x + y)^2 \right| + \frac{1}{4} \cdot \left| (x - y)^2 - \widehat{f}_{\text{sq}}(x - y) \right| \\ &\leq \frac{1}{4} \cdot 4 \cdot a^2 \cdot 4^{-R} + \frac{1}{4} \cdot 4 \cdot a^2 \cdot 4^{-R} \\ &= 2 \cdot a^2 \cdot 4^{-R} \end{aligned}$$

für $x, y \in [-a, a]$.

Um die Gewichtsschranken zu erhalten, müssen wir das Netz $\widehat{f}_{\text{mult}}$ umschreiben zu

$$\begin{aligned} \widehat{f}_{\text{mult}}(x, y) &= \frac{1}{4} \cdot \left(\widehat{f}_{\text{sq}}(x + y) - \widehat{f}_{\text{sq}}(x - y) \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{4} \cdot \left(4 \cdot (2a)^2 \cdot \widehat{f}_{\text{sq}[0,1]}(\widehat{f}_{\text{tran}}(x+y)) - 2 \cdot (2a) \cdot \widehat{f}_{\text{id}}^R(x+y) - (2a)^2 \right. \\
&\quad \left. - \left(4 \cdot (2a)^2 \cdot \widehat{f}_{\text{sq}[0,1]}(\widehat{f}_{\text{tran}}(x-y)) - 2 \cdot (2a) \cdot \widehat{f}_{\text{id}}^R(x-y) - (2a)^2 \right) \right) \\
&= \frac{1}{4} \cdot \left(4 \cdot (2a)^2 \cdot \widehat{f}_{\text{sq}[0,1]} \left(\frac{x+y}{2 \cdot 2a} + \frac{1}{2} \right) - 2 \cdot (2a) \cdot \widehat{f}_{\text{id}}^R(x+y) - (2a)^2 \right. \\
&\quad \left. - \left(4 \cdot (2a)^2 \cdot \widehat{f}_{\text{sq}[0,1]} \left(\frac{x-y}{2 \cdot 2a} + \frac{1}{2} \right) - 2 \cdot (2a) \cdot \widehat{f}_{\text{id}}^R(x-y) - (2a)^2 \right) \right) \\
&= 4a^2 \cdot \widehat{f}_{\text{sq}[0,1]} \left(\frac{x+y}{2 \cdot 2a} + \frac{1}{2} \right) - a \cdot \widehat{f}_{\text{id}}^R(x+y) - \left(4a^2 \cdot \widehat{f}_{\text{sq}[0,1]} \left(\frac{x-y}{2 \cdot 2a} + \frac{1}{2} \right) - a \cdot \widehat{f}_{\text{id}}^R(x-y) \right).
\end{aligned}$$

Da sich die Bias-Terme gegenseitig aufgehoben haben, erhalten wir

$$\left(\mathbf{w}_{\widehat{f}_{\text{mult}}} \right)_{1,0}^{(R)} = 0.$$

Aus dieser Darstellung sowie dem Beweis von Lemma 23, in dem wir gezeigt haben, dass $\|\mathbf{w}_{\widehat{f}_{\text{sq}[0,1]}}\|_{\infty} \leq 1$ ist, folgt daher

$$\left\| \mathbf{w}_{\widehat{f}_{\text{mult}}} \right\|_{\infty} \leq 4a^2 \cdot \left\| \mathbf{w}_{\widehat{f}_{\text{sq}[0,1]}} \right\|_{\infty} \leq 4a^2 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{mult}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1.$$

□

Approximation eines Produkts mit d Komponenten

Im folgenden Lemma beschreiben wir ein neuronales Netzwerk, welches das Produkt von d Eingabekomponenten näherungsweise berechnet. Dadurch wird es möglich sein, Multiplikationen zu approximieren.

Lemma 25. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Dann existiert für jedes $R \in \mathbb{N}$ und jedes $a \geq 1$ ein neuronales Netzwerk

$$\widehat{f}_{\text{mult},d} \in \mathcal{F}(R \cdot \lceil \log_2(d) \rceil, 18 \cdot d)$$

mit den Gewichtsschranken

$$\left\| \mathbf{w}_{\widehat{f}_{\text{mult},d}} \right\|_{\infty} \leq 4 \cdot 4^{2d} \cdot a^{2d}, \quad \left(\mathbf{w}_{\widehat{f}_{\text{mult},d}} \right)_{1,0}^{(R \cdot \lceil \log_2(d) \rceil)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{mult},d}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1,$$

so dass

$$\left| \widehat{f}_{\text{mult},d}(\mathbf{x}) - \prod_{i=1}^d x^{(i)} \right| \leq 4 \cdot 4^{4d} \cdot a^{4d} \cdot d \cdot 4^{-R}$$

für alle $\mathbf{x} \in [-a, a]^d$ gilt.

Beweis. In diesem Beweis setzen wir $q = \lceil \log_2(d) \rceil$ und konstruieren im Folgenden das vorwärtsgerichtete neuronale Netz $\widehat{f}_{\text{mult},d}$ mit $L = R \cdot q$ verdeckten Schichten und $r = 18 \cdot d$ Neuronen in jeder Schicht. Hierfür setzen wir

$$(z_1, \dots, z_{2^q}) = \left(x^{(1)}, x^{(2)}, \dots, x^{(d)}, \underbrace{1, \dots, 1}_{2^q - d} \right). \quad (\text{A.8})$$

Für die Konstruktion des Netzes werden wir das Netzwerk $\widehat{f}_{\text{mult}}$ aus Lemma 24 verwenden, welches

$$\left| \widehat{f}_{\text{mult}}(x, y) - x \cdot y \right| \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R} \quad (\text{A.9})$$

für $x, y \in [-4^d \cdot a^d, 4^d \cdot a^d]$ erfüllt. In den ersten R Schichten werden wir

$$\widehat{f}_{\text{mult}}(z_1, z_2), \widehat{f}_{\text{mult}}(z_3, z_4), \dots, \widehat{f}_{\text{mult}}(z_{2^q-1}, z_{2^q})$$

berechnen. Hierbei handelt es sich um 2^{q-1} Netze, die parallel berechnet werden. Jedes Netz $\widehat{f}_{\text{mult}}$ besitzt R Schichten und 18 Neuronen pro Schicht. Daraus folgt, dass die parallel berechneten Netze insgesamt R Schichten und maximal $18 \cdot 2^{q-1} \leq 18 \cdot d$ Neuronen benötigen.

In diesen R verdeckten Schichten werden also benachbarte Werte (z_i, z_{i+1}) für $i = 1, \dots, 2^d$ miteinander multipliziert. Im Fall, dass $z_l = x^{(d)}$ und $z_{l+1} = 1$ ist, gilt beispielsweise

$$\widehat{f}_{\text{mult}}(z_l, z_{l+1}) = \widehat{f}_{\text{mult}}(x^{(d)}, 1).$$

Als Ausgabe der ersten R Schichten erhalten wir einen Vektor, welcher eine Länge von 2^{q-1} hat. Als nächstes kombinieren wir diese Ausgaben und wenden erneut $\widehat{f}_{\text{mult}}$ an. Dieses Verfahren wird sukzessive fortgesetzt, bis nur noch eine Ausgabe übrig bleibt. Aus diesem Grund benötigen wir $L = R \cdot q$ verdeckte Schichten und höchstens $18 \cdot d$ Neuronen in jeder Schicht.

Ist nun $R \geq \log_4(2 \cdot 4^{2 \cdot d} \cdot a^{2 \cdot d})$, so ergibt sich aus Ungleichung (A.9) für alle $l \in \{1, \dots, d\}$ und alle $z_1, z_2 \in [-(4^l - 1) \cdot a^l, (4^l - 1) \cdot a^l] \subseteq [-4^l \cdot a^l, 4^l \cdot a^l]$ die Abschätzung

$$\left| \widehat{f}_{\text{mult}}(z_1, z_2) \right| \leq |z_1 \cdot z_2| + \left| \widehat{f}_{\text{mult}}(z_1, z_2) - z_1 \cdot z_2 \right| \leq (4^l - 1)^2 a^{2l} + 1 \leq (4^{2l} - 1) \cdot a^{2l}.$$

Damit erhalten wir schrittweise, dass die Ausgaben jedes Netzes $\widehat{f}_{\text{mult}}$ des Levels $l \in \{1, \dots, q-1\}$ in dem Intervall $[-4^{2^l} \cdot a^{2^l}, 4^{2^l} \cdot a^{2^l}]$ liegen. Wegen

$$4^{2^l} \cdot a^{2^l} \leq 4^d \cdot a^d$$

für $l \in \{1, \dots, q-1\}$ sind die Ausgaben insbesondere in dem Intervall $[-4^d \cdot a^d, 4^d \cdot a^d]$ enthalten, in dem Ungleichung (A.9) erfüllt ist.

Bei $\widehat{f}_{\text{mult},d}$ handelt es sich um eine Komposition der neuronalen Netze $\widehat{f}_{\text{mult}}$. Aus Lemma 24 wissen wir, dass die Gewichte von $\widehat{f}_{\text{mult}}$ die Bedingungen

$$\left(\mathbf{w}_{\widehat{f}_{\text{mult}}} \right)_{1,0}^{(R)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{mult}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1$$

erfüllen. Daher können wir Lemma 17c) anwenden, woraus für $a = 4^d \cdot a^d$ in Lemma 24 folgt, dass

$$\left\| \mathbf{w}_{\widehat{f}_{\text{mult},d}} \right\|_{\infty} \leq 4 \cdot 4^{2d} \cdot a^{2d}$$

gilt. Die restlichen Gewichtsschranken folgen direkt aus Lemma 24 und sind gegeben durch

$$\left(\mathbf{w}_{\widehat{f}_{\text{mult},d}}\right)_{1,0}^{(R \cdot \lceil \log_2(d) \rceil)} = 0 \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{mult},d}}\right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1.$$

Für $z_1, \dots, z_{2^q} \in [-a, a]$ setzen wir

$$\begin{aligned} \widehat{f}_2(z_1, z_2) &= \widehat{f}_{\text{mult}}(z_1, z_2) \\ \widehat{f}_4(z_1, z_2, z_3, z_4) &= \widehat{f}_{\text{mult}}(\widehat{f}_2(z_1, z_2), \widehat{f}_2(z_3, z_4)) \\ &\vdots \\ \widehat{f}_{2^q}(z_1, \dots, z_{2^q}) &= \widehat{f}_{\text{mult}}(\widehat{f}_{2^{q-1}}(z_1, \dots, z_{2^{q-1}}), \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q})). \end{aligned}$$

Zudem setzen wir

$$\Delta_l = \sup_{z_1, \dots, z_{2^l} \in [-a, a]} \left| \widehat{f}_{2^l}(z_1, \dots, z_{2^l}) - \prod_{i=1}^{2^l} z_i \right|.$$

Dann ist

$$\left| \widehat{f}_{\text{mult},d}(\mathbf{x}) - \prod_{i=1}^d x^{(i)} \right| \leq \Delta_q.$$

Aus Ungleichung (A.9) folgt

$$\Delta_1 \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R},$$

woraus sich mithilfe der Dreiecksungleichung die folgende Abschätzung ergibt:

$$\begin{aligned} \Delta_q &\leq \sup_{z_1, \dots, z_{2^q} \in [-a, a]} \left| \widehat{f}_{\text{mult}}(\widehat{f}_{2^{q-1}}(z_1, \dots, z_{2^{q-1}}), \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q})) \right. \\ &\quad \left. - \widehat{f}_{2^{q-1}}(z_1, \dots, z_{2^{q-1}}) \cdot \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) \right| \\ &+ \sup_{z_1, \dots, z_{2^q} \in [-a, a]} \left| \widehat{f}_{2^{q-1}}(z_1, \dots, z_{2^{q-1}}) \cdot \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) \right. \\ &\quad \left. - \left(\prod_{i=1}^{2^{q-1}} z_i \right) \cdot \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) \right| \\ &+ \sup_{z_1, \dots, z_{2^q} \in [-a, a]} \left| \left(\prod_{i=1}^{2^{q-1}} z_i \right) \cdot \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) \right. \\ &\quad \left. - \left(\prod_{i=1}^{2^{q-1}} z_i \right) \cdot \prod_{i=2^{q-1}+1}^{2^q} z_i \right| \end{aligned}$$

Da alle Ausgaben des Levels $l \in \{1, \dots, q-1\}$ im Intervall $[-4^{2^l} \cdot a^{2^l}, 4^{2^l} \cdot a^{2^l}]$ enthalten sind, folgt für den ersten Summanden aus Ungleichung (A.9), dass dieser kleiner gleich

$$2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R}$$

ist.

Für den zweiten Summanden gilt, aufgrund der Tatsache, dass das Netz $\widehat{f}_{2^{q-1}}$ im Intervall $[-4^{2^{q-1}} \cdot a^{2^{q-1}}, 4^{2^{q-1}} \cdot a^{2^{q-1}}]$ liegt

$$\begin{aligned} & \sup_{z_1, \dots, z_{2^q} \in [-a, a]} \left| \widehat{f}_{2^{q-1}}(z_1, \dots, z_{2^{q-1}}) \cdot \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) - \left(\prod_{i=1}^{2^{q-1}} z_i \right) \cdot \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) \right| \\ & \leq \sup_{z_1, \dots, z_{2^q} \in [-a, a]} \left| \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) \right| \cdot \sup_{z_1, \dots, z_{2^q} \in [-a, a]} \left| \widehat{f}_{2^{q-1}}(z_1, \dots, z_{2^{q-1}}) - \left(\prod_{i=1}^{2^{q-1}} z_i \right) \right| \\ & \leq 4^{2^{q-1}} \cdot a^{2^{q-1}} \cdot \Delta_{q-1}. \end{aligned}$$

Der dritte Summand ist von oben beschränkt durch

$$\sup_{z_1, \dots, z_{2^q} \in [-a, a]} \left| \prod_{i=1}^{2^{q-1}} z_i \right| \cdot \left| \widehat{f}_{2^{q-1}}(z_{2^{q-1}+1}, \dots, z_{2^q}) - \prod_{i=2^{q-1}+1}^{2^q} z_i \right| \leq a^{2^{q-1}} \cdot \Delta_{q-1}.$$

Damit ist

$$\begin{aligned} & \left| \widehat{f}_{\text{mult},d}(\mathbf{x}) - \prod_{i=1}^d x^{(i)} \right| \\ & \leq \Delta_q \\ & \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R} + 2 \cdot 4^{2^{q-1}} \cdot a^{2^{q-1}} \cdot \Delta_{q-1} \\ & \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R} + 2 \cdot 4^{2^{q-1}} \cdot a^{2^{q-1}} \cdot (2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R} + 2 \cdot 4^{2^{q-2}} \cdot a^{2^{q-2}} \cdot \Delta_{q-2}) \\ & \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R} \cdot 4^{2+\dots+2^{q-1}} \cdot a^{2+\dots+2^{q-1}} \cdot (1 + 2 + \dots + 2^{q-1}) \\ & \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R} \cdot 4^{\sum_{k=0}^{q-1} 2^k - 1} \cdot a^{\sum_{k=0}^{q-1} 2^k - 1} \cdot \sum_{k=0}^{q-1} 2^k. \end{aligned}$$

Bei der Summe $\sum_{k=0}^{q-1} 2^k$ handelt es sich um eine geometrische Summe, für die

$$\sum_{k=0}^{q-1} 2^k = \frac{1 - 2^q}{1 - 2} = 2^q - 1 \leq 2^{\lceil \log_2(d) \rceil} \leq 2^{\log_2(d)+1} = 2 \cdot d$$

gilt. Damit ergibt sich

$$\begin{aligned} \left| \widehat{f}_{\text{mult},d}(\mathbf{x}) - \prod_{i=1}^d x^{(i)} \right| & \leq 2 \cdot (4^d \cdot a^d)^2 \cdot 4^{-R} \cdot 4^{2d} \cdot a^{2d} \cdot 2 \cdot d \\ & = 4 \cdot 4^{4d} \cdot a^{4d} \cdot d \cdot 4^{-R}, \end{aligned}$$

woraus die Behauptung des Lemmas folgt. □

Approximation multivariater Funktionen

Sei \mathcal{P}_N die lineare Hülle aller Monome der Form

$$\prod_{k=1}^d (x^{(k)})^{r_k}$$

für $r_1, \dots, r_d \in \mathbb{N}_0$, wobei $r_1 + \dots + r_d \leq N$. Dann ist \mathcal{P}_N ein linearer Vektorraum von Funktionen der Dimension

$$\dim(\mathcal{P}_N) = \left| \left\{ (r_0, \dots, r_d) \in \mathbb{N}_0^{d+1} : r_0 + \dots + r_d = N \right\} \right| = \binom{d+N}{d}.$$

Im nächsten Lemma konstruieren wir ein neuronales Netzwerk, das Funktionen der Klasse \mathcal{P}_N , die mit einem zusätzlichen Faktor multipliziert sind, annähert. Diese modifizierte Form von Polynomen wird später bei der Konstruktion unseres Netzwerks für das Resultat benötigt.

Lemma 26. Wir bezeichnen mit $m_1, \dots, m_{\binom{d+N}{d}}$ alle Monome der Klasse \mathcal{P}_N für ein $N \in \mathbb{N}$. Seien $r_1, \dots, r_{\binom{d+N}{d}} \in \mathbb{R}$. Wir setzen

$$p(\mathbf{x}, y_1, \dots, y_{\binom{d+N}{d}}) = \sum_{i=1}^{\binom{d+N}{d}} r_i \cdot y_i \cdot m_i(\mathbf{x}) \quad \text{für } \mathbf{x} \in [-a, a]^d \text{ und } y_i \in [-a, a]$$

sowie $\bar{r}(p) = \max_{i \in \{1, \dots, \binom{d+N}{d}\}} |r_i|$. Sei zudem $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Dann existiert für jedes $a \geq 1$ und

$$R \geq \log_4 \left(2 \cdot 4^{2 \cdot (N+1)} \cdot a^{2 \cdot (N+1)} \right) \tag{A.10}$$

ein neuronales Netz

$$\hat{f}_p \in \mathcal{F}(L, r)$$

mit

$$L = R \cdot \lceil \log_2(N+1) \rceil$$

und

$$r = 18 \cdot (N+1) \cdot \binom{d+N}{d},$$

welches

$$\left| \hat{f}_p(\mathbf{x}, y_1, \dots, y_{\binom{d+N}{d}}) - p(\mathbf{x}, y_1, \dots, y_{\binom{d+N}{d}}) \right| \leq c_{173} \cdot \bar{r}(p) \cdot a^{4 \cdot (N+1)} \cdot 4^{-R}$$

für alle $\mathbf{x} \in [-a, a]^d$, $y_1, \dots, y_{\binom{d+N}{d}} \in [-a, a]$ erfüllt. Dabei hängt die Konstante $c_{173} > 0$ von d und N ab. Zudem hat das neuronale Netz die folgenden Gewichtsbeschränkungen

$$\|\mathbf{w}_{\hat{f}_p}\|_\infty \leq \bar{r}(p) \cdot 4 \cdot 4^{2 \cdot (N+1)} \cdot a^{2 \cdot (N+1)}, \quad (\mathbf{w}_{\hat{f}_p})_{1,0}^{(L)} = 0 \quad \text{und} \quad \left\| (\mathbf{w}_{\hat{f}_p})_{i,j>0}^{(0)} \right\|_\infty \leq 1.$$

Beweis. Sei m ein beliebiges Monom der Klasse \mathcal{P}_N und $r_1, \dots, r_d \in \mathbb{N}_0$ mit $r_1 + \dots + r_d \leq N$.

Im *ersten Schritt des Beweises* werden wir ein neuronales Netz \widehat{f}_m konstruieren, das

$$y \cdot m(\mathbf{x}) = y \cdot \prod_{k=1}^d \left(x^{(k)}\right)^{r_k} \quad \text{für } \mathbf{x} \in [-a, a]^d \text{ und } y \in [-a, a]$$

approximiert. Durch die einmalige Verwendung von y und, falls erforderlich, die mehrfache Verwendung einiger $x^{(i)}$ kann Lemma 25 auf Monome erweitert werden. Hierfür substituieren wir d durch $N + 1$ in Lemma 25. Damit erhalten wir schließlich, dass ein Netzwerk

$$\widehat{f}_m \in \mathcal{F}(R \cdot \lceil \log_2(N + 1) \rceil, 18 \cdot (N + 1))$$

mit den Gewichtsschranken

$$\left\| \mathbf{w}_{\widehat{f}_m} \right\|_{\infty} \leq 4 \cdot 4^{2 \cdot (N+1)} \cdot a^{2 \cdot (N+1)}, \quad \left(\mathbf{w}_{\widehat{f}_m} \right)_{1,0}^{(R \cdot \lceil \log_2(N+1) \rceil)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_m} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1$$

existiert, das einen Approximationsfehler von

$$\left| \widehat{f}_m(\mathbf{x}, y) - y \cdot m(\mathbf{x}) \right| = 4 \cdot 4^{4 \cdot (N+1)} \cdot a^{4 \cdot (N+1)} \cdot (N + 1) \cdot 4^{-R} \quad (\text{A.11})$$

erzielt.

Im *zweiten Schritt des Beweises* zeigen wir die Aussage des Lemmas. Sei hierfür

$$p(\mathbf{x}, y_1, \dots, y_{\binom{d+N}{d}}) = \sum_{i=1}^{\binom{d+N}{d}} r_i \cdot y_i \cdot m_i(\mathbf{x}). \quad (\text{A.12})$$

Dieses Polynom kann durch ein neuronales Netz $\widehat{f}_p \in \mathcal{F}(L, r)$ mit $L = R \cdot \lceil \log_2(N + 1) \rceil$ sowie $r = 18 \cdot (N + 1) \cdot \binom{d+N}{d}$ der Form

$$\widehat{f}_p(\mathbf{x}, y_1, \dots, y_{\binom{d+N}{d}}) = \sum_{i=1}^{\binom{d+N}{d}} r_i \cdot \widehat{f}_{m_i}(\mathbf{x}, y_i) \quad (\text{A.13})$$

approximiert werden, wobei \widehat{f}_{m_i} das neuronale Netzwerk aus Beweisschritt 1 ist.

Mit den Gewichtsschranken des Netzes \widehat{f}_{m_i} aus dem ersten Beweisschritt erhalten wir dann

$$\left\| \mathbf{w}_{\widehat{f}_p} \right\|_{\infty} \leq \bar{r}(p) \cdot 4 \cdot 4^{2 \cdot (N+1)} \cdot a^{2 \cdot (N+1)}, \quad \left(\mathbf{w}_{\widehat{f}_p} \right)_{1,0}^{(L)} = 0 \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\widehat{f}_p} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1.$$

Aufgrund der Gleichungen (A.11), (A.12) und (A.13) können wir folgern, dass

$$\begin{aligned} & \left| p(\mathbf{x}, y_1, \dots, y_{\binom{d+N}{d}}) - \widehat{f}_p(\mathbf{x}, y_1, \dots, y_{\binom{d+N}{d}}) \right| \\ &= \left| \sum_{i=1}^{\binom{d+N}{d}} r_i \cdot y_i \cdot m_i(\mathbf{x}) - \sum_{i=1}^{\binom{d+N}{d}} r_i \cdot \widehat{f}_{m_i}(\mathbf{x}, y_i) \right| \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{i=1}^{\binom{d+N}{d}} |r_i| \cdot \left| y_i \cdot m_i(\mathbf{x}) - \widehat{f}_{m_i}(\mathbf{x}, y_i) \right| \\
&\leq \binom{d+N}{d} \cdot \bar{r}(p) \cdot 4 \cdot 4^{4(N+1)} \cdot a^{4(N+1)} \cdot (N+1) \cdot 4^{-R}
\end{aligned}$$

erfüllt ist, woraus die Aussage des Lemmas folgt. \square

Approximation der mehrdimensionalen Indikatorfunktion

In dem folgenden Lemma werden zwei neuronale Netze konstruiert, die zum einen eine mehrdimensionale Indikatorfunktion und zum anderen deren Produkt mit einem zusätzlichen Faktor approximieren. In diesen Netzwerken nutzen wir die Eigenschaft der ReLU-Aktivierungsfunktion aus, die bei negativen Eingabewerten den Wert 0 annimmt. Insbesondere verwenden wir, dass

$$\sigma(1 - R \cdot \sigma(x)) = \begin{cases} 0 & \text{für } x \geq \frac{1}{R} \\ 1 & \text{für } x \leq 0 \end{cases}$$

für $R \in \mathbb{N}$ gilt, sowie

$$\sigma\left(\widehat{f}_{\text{id}}(s) - R^2 \cdot \sigma(x)\right) + \sigma\left(-\widehat{f}_{\text{id}}(s) - R^2 \cdot \sigma(x)\right) = \begin{cases} 0 & \text{für } x \geq \frac{1}{R} \\ s & \text{für } x \leq 0 \end{cases}$$

für alle $|s| \leq R$.

Lemma 27. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$ und $R \in \mathbb{N}$. Seien zudem $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ mit

$$b^{(i)} - a^{(i)} \geq \frac{2}{R} \quad \text{für alle } i \in \{1, \dots, d\}$$

und

$$K_{1/R} = \left\{ \mathbf{x} \in \mathbb{R}^d : x^{(i)} \notin [a^{(i)}, a^{(i)} + 1/R) \cup (b^{(i)} - 1/R, b^{(i)}) \text{ für alle } i \in \{1, \dots, d\} \right\}.$$

a) Dann erfüllt das Netzwerk

$$\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x}) = \sigma\left(1 - R \cdot \sum_{i=1}^d \left(\sigma\left(a^{(i)} + \frac{1}{R} - x^{(i)}\right) + \sigma\left(x^{(i)} - b^{(i)} + \frac{1}{R}\right)\right)\right)$$

der Klasse $\mathcal{F}(2, 2d)$ die Eigenschaften

$$\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x}) = \mathbf{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}) \quad \text{für } \mathbf{x} \in K_{1/R}$$

sowie

$$\left| \widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x}) - \mathbf{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}) \right| \leq 1 \quad \text{für } \mathbf{x} \in \mathbb{R}^d.$$

Zusätzlich gelten für die Gewichte des Netzwerks die folgenden Schranken

$$\|\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}}\|_{\infty} \leq \max\left\{R, \|\mathbf{a}\|_{\infty} + \frac{1}{R}, \|\mathbf{b}\|_{\infty} + \frac{1}{R}\right\},$$

$$\left\| \left(\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}} \right)_{1,j>0}^{(2)} \right\|_{\infty} = 1, \quad \left(\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}} \right)_{1,0}^{(2)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = 1.$$

b) Für $|t| \leq R$ hat das Netzwerk

$$\begin{aligned} \widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) &= \sigma \left(\widehat{f}_{\text{id}}(t) - R^2 \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \right) \\ &\quad - \sigma \left(-\widehat{f}_{\text{id}}(t) - R^2 \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \right) \end{aligned}$$

der Klasse $\mathcal{F}(2, 2 \cdot (2d + 2))$ die Eigenschaften

$$\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) = t \cdot \mathbb{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}) \quad \text{für } \mathbf{x} \in K_{1/R}$$

und

$$\left| \widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) - t \cdot \mathbb{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}) \right| \leq |t| \quad \text{für } \mathbf{x} \in \mathbb{R}^d.$$

Zudem sind die Gewichtsschranken des neuronalen Netzes $\widehat{f}_{\text{test}}$ gegeben durch

$$\|\mathbf{w}_{\widehat{f}_{\text{test}}}\|_{\infty} \leq R^2, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{i,0}^{(0)} \right\|_{\infty} = \frac{1}{R},$$

$$\left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{1,j>0}^{(2)} \right\|_{\infty} = 1, \quad \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{1,0}^{(2)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = 1.$$

Beweis. a) Sei $\mathbf{x} \in [\mathbf{a} + 1/R \cdot \mathbf{1}, \mathbf{b} - 1/R \cdot \mathbf{1}]$. Dann ist

$$a^{(i)} + \frac{1}{R} - x^{(i)} \leq 0 \quad \text{und} \quad x^{(i)} - b^{(i)} + \frac{1}{R} \leq 0 \quad \text{für alle } i \in \{1, \dots, d\},$$

woraus

$$\sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) = 0$$

und damit

$$\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x}) = \sigma(1 - R \cdot 0) = 1 = \mathbb{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$$

folgt.

Sei $\mathbf{x} \notin [\mathbf{a}, \mathbf{b}]$. Dann existiert mindestens ein $j \in \{1, \dots, d\}$, so dass

$$x^{(j)} < a^{(j)} \quad \text{oder} \quad x^{(j)} \geq b^{(j)}$$

gilt. Sei ohne Beschränkung der Allgemeinheit $x^{(j)} < a^{(j)} < a^{(j)} + 1/R$ für ein $j \in \{1, \dots, d\}$, so ergibt sich

$$1 - R \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right)$$

$$\begin{aligned} &\leq 1 - R \cdot \left(a^{(j)} + \frac{1}{R} - x^{(j)} \right) \\ &\leq 0. \end{aligned}$$

Dies führt zu

$$\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x}) = 0 = \mathbb{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}).$$

Für $\mathbf{x} \in \mathbb{R}^d$ erhalten wir aufgrund der Tatsache, dass die ReLU-Aktivierungsfunktion immer größer oder gleich 0 ist, sowie der Ungleichung

$$1 - R \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \leq 1$$

die folgende Abschätzung

$$0 \leq \widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x}) \leq 1.$$

Zudem ist

$$\mathbb{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}) \in \{0, 1\}$$

für $\mathbf{x} \in \mathbb{R}^d$ erfüllt. Dies impliziert, dass

$$\left| \widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x}) - \mathbb{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}) \right| \leq 1$$

gilt.

Die Schranken für die Gewichte ergeben sich direkt aus der Definition des neuronalen Netzes $\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}(\mathbf{x})$ und sind gegeben durch

$$\|\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}}\|_{\infty} \leq \max \left\{ R, \|\mathbf{a}\|_{\infty} + \frac{1}{R}, \|\mathbf{b}\|_{\infty} + \frac{1}{R} \right\}$$

sowie

$$\left\| \left(\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = 1, \quad \left(\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}} \right)_{1,0}^{(2)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{ind},[\mathbf{a},\mathbf{b}]}} \right)_{1,j>0}^{(2)} \right\|_{\infty} = 1.$$

b) Sei $\mathbf{x} \in [\mathbf{a} + 1/R \cdot \mathbf{1}, \mathbf{b} - 1/R \cdot \mathbf{1}]$. Analog zu Teil a) erhalten wir

$$a^{(i)} + \frac{1}{R} - x^{(i)} \leq 0 \quad \text{und} \quad x^{(i)} - b^{(i)} + \frac{1}{R} \leq 0 \quad \text{für alle } i \in \{1, \dots, d\},$$

woraus

$$\sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) = 0$$

folgt. Damit ergibt sich

$$\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) = \sigma(\widehat{f}_{\text{id}}(t)) - \sigma(-\widehat{f}_{\text{id}}(t)) = t = t \cdot \mathbb{1}_{[\mathbf{a},\mathbf{b}]}(\mathbf{x}).$$

Für $\mathbf{x} \notin [\mathbf{a}, \mathbf{b}]$ existiert mindestens ein $j \in \{1, \dots, d\}$, welches

$$x^{(j)} < a^{(j)} \quad \text{oder} \quad x^{(j)} \geq b^{(j)}$$

erfüllt.

Sei $0 \leq t \leq R$. Zudem gelte ohne Beschränkung der Allgemeinheit $x^{(j)} < a^{(j)}$ für ein $j \in \{1, \dots, d\}$. Wegen $a^{(j)} + 1/R - x^{(j)} > 1/R \geq 0$ folgt dann die Ungleichung

$$\begin{aligned} \widehat{f}_{\text{id}}(t) - R^2 \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \\ \leq \widehat{f}_{\text{id}}(t) - R^2 \cdot \left(a^{(j)} + \frac{1}{R} - x^{(j)} \right) \\ \leq \widehat{f}_{\text{id}}(t) - R \\ \leq 0. \end{aligned}$$

Aus diesem Grund ist

$$\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) = 0 = t \cdot \mathbb{1}_{[\mathbf{a}, \mathbf{b}]}(\mathbf{x}).$$

Im Fall $-R \leq t < 0$ folgt die Behauptung analog.

Sei nun $\mathbf{x} \in \mathbb{R}^d$ und $t \geq 0$. Aufgrund der Tatsache, dass die ReLU-Aktivierungsfunktion nichtnegativ ist sowie der Definition des Netzes $\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t)$ ergibt sich

$$\begin{aligned} \widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) &= \sigma \left(\widehat{f}_{\text{id}}(t) - R^2 \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \right) \\ &\quad - \sigma \left(-\widehat{f}_{\text{id}}(t) - R^2 \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \right) \\ &\leq \sigma \left(\widehat{f}_{\text{id}}(t) - R^2 \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \right) \\ &\leq \sigma \left(\widehat{f}_{\text{id}}(t) \right) \\ &= t. \end{aligned}$$

Zusätzlich gilt

$$\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) \geq -\sigma \left(-\widehat{f}_{\text{id}}(t) - R^2 \cdot \sum_{i=1}^d \left(\sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) + \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \right) \right).$$

Da die ReLU-Aktivierungsfunktion nur Werte größer oder gleich 0 annehmen kann, sind

$$R^2 \cdot \sum_{i=1}^d \sigma \left(a^{(i)} + \frac{1}{R} - x^{(i)} \right) \geq 0 \quad \text{und} \quad R^2 \cdot \sum_{i=1}^d \sigma \left(x^{(i)} - b^{(i)} + \frac{1}{R} \right) \geq 0.$$

Außerdem folgt für $t \geq 0$ die Gleichheit $-\widehat{f}_{\text{id}}(t) = -t$, woraus sich

$$\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) \geq 0$$

ergibt.

Somit lässt sich schließen, dass

$$\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) \in [0, t]$$

für alle $\mathbf{x} \in \mathbb{R}^d$ und $t \geq 0$ ist. Des Weiteren ist

$$t \cdot \mathbb{1}_{[\mathbf{a}, \mathbf{b})}(\mathbf{x}) \in \{0, t\},$$

für alle $\mathbf{x} \in \mathbb{R}^d$. Dies liefert uns die Ungleichung

$$|\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t) - t \cdot \mathbb{1}_{[\mathbf{a}, \mathbf{b})}(\mathbf{x})| \leq |t|.$$

Die Behauptung für den Fall $\mathbf{x} \in \mathbb{R}^d$ und $t < 0$ folgt analog.

Die Schranken für die Gewichte des Netzes folgen direkt aus der Definition des Netzes $\widehat{f}_{\text{test}}$ sowie den Beschränkungen der Gewichte des Netzes \widehat{f}_{id} und sind gegeben durch

$$\begin{aligned} \|\mathbf{w}_{\widehat{f}_{\text{test}}}\|_{\infty} &\leq R^2, & \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{i,0}^{(0)} \right\|_{\infty} &= \frac{1}{R}, \\ \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} &= 1, & \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{1,0}^{(2)} &= 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{1,j>0}^{(2)} \right\|_{\infty} &= 1. \end{aligned}$$

□

Die Verwendung der Netzwerke aus Lemma 26 und Lemma 27 wird es uns nun ermöglichen, die rekursiv definierten Funktionen in der Definition von $\phi_{1,3}$ zu konstruieren.

Beweis von Lemma 22. Im ersten Schritt des Beweises werden wir beschreiben, wie die rekursiv definierte Funktion $\phi_{1,3}$ aus Lemma 21 mithilfe eines neuronalen Netzes approximiert werden kann.

Für die Konstruktion von $\widehat{\phi}_{1,3}$ werden wir die Netzwerke

$$\widehat{f}_{\text{ind},[\mathbf{a}, \mathbf{b})} \in \mathcal{F}(2, 2d) \quad \text{und} \quad \widehat{f}_{\text{test}} \in \mathcal{F}(2, 2 \cdot (2d + 2))$$

aus Lemma 27 mit $R = B_M = M^{2p+2}$ verwenden. Diese approximieren die Indikatorfunktion $\mathbb{1}_{[\mathbf{a}, \mathbf{b})}(\mathbf{x})$ beziehungsweise $t \cdot \mathbb{1}_{[\mathbf{a}, \mathbf{b})}(\mathbf{x})$ für ein $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ und $B_M \in \mathbb{N}$ mit

$$b^{(i)} - a^{(i)} \geq \frac{2}{B_M} \quad \text{für alle } i \in \{1, \dots, d\}.$$

Unter Beachtung von $B_M \in \mathbb{N}$, $|t| \leq B_M$ sowie

$$x^{(i)} \notin \left[a^{(i)}, a^{(i)} + \frac{1}{B_M} \right) \cup \left(b^{(i)} - \frac{1}{B_M}, b^{(i)} \right) \quad \text{für alle } i \in \{1, \dots, d\}$$

erhalten wir aus Lemma 27, dass

$$\widehat{f}_{\text{ind},[\mathbf{a}, \mathbf{b})}(\mathbf{x}) = \mathbb{1}_{[\mathbf{a}, \mathbf{b})}(\mathbf{x})$$

und

$$\widehat{f}_{\text{test}}(\mathbf{x}, \mathbf{a}, \mathbf{b}, t)(\mathbf{x}) = t \cdot \mathbb{1}_{[\mathbf{a}, \mathbf{b}]}(\mathbf{x})$$

gilt.

Zusätzlich liefert Lemma 27 für $R = B_M = M^{2p+2}$ die folgenden Beschränkungen für die Gewichte der beiden Netzwerke

$$\|\mathbf{w}_{\widehat{f}_{\text{ind}, [\mathbf{a}, \mathbf{b}]}}\|_{\infty} \leq \max \left\{ M^{2p+2}, \|\mathbf{a}\|_{\infty} + \frac{1}{M^{2p+2}}, \|\mathbf{b}\|_{\infty} + \frac{1}{M^{2p+2}} \right\},$$

$$\left\| \left(\mathbf{w}_{\widehat{f}_{\text{ind}, [\mathbf{a}, \mathbf{b}]}} \right)_{i, \widehat{j} > 0}^{(0)} \right\|_{\infty} = 1, \quad \left(\mathbf{w}_{\widehat{f}_{\text{ind}, [\mathbf{a}, \mathbf{b}]}} \right)_{1,0}^{(2)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{ind}, [\mathbf{a}, \mathbf{b}]}} \right)_{1, \widehat{j} > 0}^{(2)} \right\|_{\infty} = 1$$

sowie

$$\|\mathbf{w}_{\widehat{f}_{\text{test}}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{i,0}^{(0)} \right\|_{\infty} = \frac{1}{M^{2p+2}},$$

$$\left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{i, \widehat{j} > 0}^{(0)} \right\|_{\infty} = 1, \quad \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{1,0}^{(2)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)_{1, \widehat{j} > 0}^{(2)} \right\|_{\infty} = 1.$$

Für einen Vektor $\bar{\mathbf{v}} \in \mathbb{R}^d$ gilt

$$\bar{\mathbf{v}} \cdot \widehat{f}_{\text{ind}, [\mathbf{a}, \mathbf{b}]}(\mathbf{x}) = \left(\bar{v}^{(1)} \cdot \widehat{f}_{\text{ind}, [\mathbf{a}, \mathbf{b}]}(\mathbf{x}), \dots, \bar{v}^{(d)} \cdot \widehat{f}_{\text{ind}, [\mathbf{a}, \mathbf{b}]}(\mathbf{x}) \right).$$

Um das finale Taylor-Polynom $\phi_{1,3}$ zu bestimmen, verwenden wir das Netzwerk

$$\widehat{f}_p \in \mathcal{F}(L, r)$$

mit

$$L = B_{M,p} \cdot \lceil \log_2(\max\{q+1, 2\}) \rceil$$

und

$$r = 18 \cdot (q+1) \cdot \binom{d+q}{d}$$

aus Lemma 26. Dieses Netz erfüllt für $R = B_{M,p}$ und $q = N$ die Ungleichung

$$\left| \widehat{f}_p(\mathbf{z}, y_1, \dots, y_{\binom{d+q}{q}}) - p(\mathbf{z}, y_1, \dots, y_{\binom{d+q}{q}}) \right|$$

$$\leq c_{173} \cdot \bar{r}(p) \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a, a]^d)} \right\} \right)^{4 \cdot (q+1)} \cdot 4^{-B_{M,p}} \quad (\text{A.14})$$

für alle $z^{(1)}, \dots, z^{(d)}, y_1, \dots, y_{\binom{d+q}{d}}$, die in dem Intervall

$$\left[-\max \left\{ 2a, \|f\|_{C^q([-a, a]^d)} \right\}, \max \left\{ 2a, \|f\|_{C^q([-a, a]^d)} \right\} \right]$$

enthalten sind, sowie $B_{M,p} \in \mathbb{N}$ mit

$$B_{M,p} \geq \log_4 \left(2 \cdot 4^{2 \cdot (q+1)} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a, a]^d)} \right\} \right)^{2 \cdot (q+1)} \right)$$

(vgl. Ungleichung (A.10)).

Für $a = \max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\}$ erfüllen die Gewichte des Netzes \hat{f}_p gemäß Lemma 26 die Bedingungen

$$\|\mathbf{w}_{\hat{f}_p}\|_\infty \leq \bar{r}(p) \cdot 4 \cdot 4^{2 \cdot (q+1)} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{2 \cdot (q+1)}$$

und

$$\left(\mathbf{w}_{\hat{f}_p} \right)_{1,0}^{(L)} = 0 \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\hat{f}_p} \right)_{i,j>0}^{(0)} \right\|_\infty \leq 1.$$

Falls $q = 0$ ist, verwenden wir ein Polynom ersten Grades, bei dem die Koeffizienten r_i aller Terme höheren Grades auf 0 gesetzt werden. Um in diesem Fall zu vermeiden, dass das Netzwerk keine verdeckten Schichten hat, haben wir in Lemma 26 die Definition der verdeckten Schichten L von $\log_2(q+1)$ auf $\log_2(\max\{q+1, 2\})$ geändert.

Um ein neuronales Netz zu erhalten, welches $\phi_{1,3}$ approximiert, werden wir die einzelnen Elemente der Rekursion von $\phi_{1,3}$ durch neuronale Netze berechnen. Für die Berechnung der Werte von $\phi_{1,1}$, $\phi_{2,1}$ und $\phi_{3,1}^{(\xi,j)}$ verwenden wir für $j \in \{1, \dots, M^d\}$, $\xi \in \mathbb{N}_0^d$ mit $\|\xi\|_1 \leq q$ und $i \in \{1, \dots, d\}$ die Netzwerke

$$\hat{\phi}_{1,1} = \left(\hat{\phi}_{1,1}^{(1)}, \dots, \hat{\phi}_{1,1}^{(d)} \right) = \hat{f}_{\text{id}}^2(\mathbf{x}),$$

$$\hat{\phi}_{2,1} = \left(\hat{\phi}_{2,1}^{(1)}, \dots, \hat{\phi}_{2,1}^{(d)} \right) = \sum_{i=1}^{M^d} (\mathbf{C}_{i,1})_{\text{left}} \cdot \hat{f}_{\text{ind}, C_{i,1}}(\mathbf{x})$$

und

$$\hat{\phi}_{3,1}^{(\xi,j)} = \sum_{i=1}^{M^d} (\partial^{\xi} f) \left((\tilde{\mathbf{C}}_{j,i})_{\text{left}} \right) \cdot \hat{f}_{\text{ind}, C_{i,1}}(\mathbf{x}).$$

Die wiederholte Anwendung des Netzes \hat{f}_{id} verändert die Werte der Gewichte des Netzes nicht. Daher erhalten wir

$$\|\mathbf{w}_{\hat{\phi}_{1,1}}\|_\infty \leq 1.$$

Zusätzlich wissen wir aufgrund der Definition der Würfel $(C_{i,1})$, dass $\|(\mathbf{C}_{i,1})_{\text{left}}\|_\infty \leq a$ für alle $i \in \{1, \dots, M^d\}$ gilt. Nach Voraussetzung ist zudem $M^{2p} \geq a$ sowie $M^{2p} \geq \|f\|_{C^q([-a,a]^d)}$. Somit ergeben sich für das Netzwerk $\hat{\phi}_{2,1}$ die Gewichtsschranken

$$\begin{aligned} \|\mathbf{w}_{\hat{\phi}_{2,1}}\|_\infty &\leq \max \left\{ \left\| \mathbf{w}_{\hat{f}_{\text{ind}, C_{i,1}}} \right\|_\infty, \|(\mathbf{C}_{i,1})_{\text{left}}\|_\infty \cdot \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind}, C_{i,1}}} \right)_{1,j>0}^{(2)} \right\|_\infty \right\} \\ &\leq \max \left\{ M^{2p+2}, a + \frac{1}{M^{2p+2}}, a \cdot 1 \right\} \\ &= M^{2p+2}, \end{aligned}$$

$$\left\| \left(\mathbf{w}_{\hat{\phi}_{2,1}} \right)_{1,j>0}^{(2)} \right\|_\infty \leq \|(\mathbf{C}_{i,1})_{\text{left}}\|_\infty \cdot \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind}, C_{i,1}}} \right)_{1,j>0}^{(2)} \right\|_\infty \leq a$$

sowie

$$\left\| \left(\mathbf{w}_{\hat{\phi}_{2,1}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind},C_{i,1}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = 1.$$

Für das Netzwerk $\hat{\phi}_{3,1}^{(\xi,j)}$ ergeben sich auf ähnliche Weise die Gewichtsschranken

$$\begin{aligned} \left\| \mathbf{w}_{\hat{\phi}_{3,1}}^{(\xi,j)} \right\|_{\infty} &\leq \max \left\{ \left\| \mathbf{w}_{\hat{f}_{\text{ind},C_{i,1}}} \right\|_{\infty}, \|f\|_{C^q([-a,a]^d)} \cdot \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind},C_{i,1}}} \right)_{1,j>0}^{(2)} \right\|_{\infty} \right\} \\ &\leq \max \left\{ M^{2p+2}, a + \frac{1}{M^{2p+2}}, \|f\|_{C^q([-a,a]^d)} \cdot 1 \right\} \\ &= M^{2p+2}, \end{aligned}$$

$$\left\| \left(\mathbf{w}_{\hat{\phi}_{3,1}}^{(\xi,j)} \right)^{(2)} \right\|_{\infty} \leq \|f\|_{C^q([-a,a]^d)} \cdot \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind},C_{i,1}}} \right)_{1,j>0}^{(2)} \right\|_{\infty} \leq \|f\|_{C^q([-a,a]^d)}$$

sowie

$$\left\| \left(\mathbf{w}_{\hat{\phi}_{3,1}}^{(\xi,j)} \right)_{i,j>0}^{(0)} \right\|_{\infty} = \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind},C_{i,1}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = 1.$$

Um $\phi_{1,2}$, $\phi_{2,2}$ und $\phi_{3,2}^{(\xi)}$ zu berechnen, verwenden wir die Netze

$$\begin{aligned} \hat{\phi}_{1,2} &= \left(\hat{\phi}_{1,2}^{(1)}, \dots, \hat{\phi}_{1,2}^{(d)} \right) = \hat{f}_{\text{id}}^2(\hat{\phi}_{1,1}), \\ \hat{\phi}_{2,2} &= \left(\hat{\phi}_{2,2}^{(1)}, \dots, \hat{\phi}_{2,2}^{(d)} \right) \end{aligned}$$

mit

$$\hat{\phi}_{2,2}^{(i)} = \sum_{j=1}^{M^d} \hat{f}_{\text{test}} \left(\hat{\phi}_{1,1}, \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j, \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{2a}{M^2} \cdot \mathbf{1}, \hat{\phi}_{2,1}^{(i)} + \tilde{v}_j^{(i)} \right)$$

für $i \in \{1, \dots, d\}$ und

$$\hat{\phi}_{3,2}^{(\xi)} = \sum_{j=1}^{M^d} \hat{f}_{\text{test}} \left(\hat{\phi}_{1,1}, \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j, \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{2a}{M^2} \cdot \mathbf{1}, \hat{\phi}_{3,1}^{(\xi,j)} \right). \quad (\text{A.15})$$

Die Gewichtsschranken für das neuronale Netz $\hat{\phi}_{1,2}$ folgen direkt aus den Gewichtsschranken von \hat{f}_{id} und sind gegeben durch

$$\left\| \mathbf{w}_{\hat{\phi}_{1,2}} \right\|_{\infty} \leq 1, \quad \left\| \left(\mathbf{w}_{\hat{\phi}_{1,2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1, \quad \left\| \left(\mathbf{w}_{\hat{\phi}_{1,2}} \right)_{1,j>0}^{(4)} \right\|_{\infty} \leq 1 \quad \text{und} \quad \left(\mathbf{w}_{\hat{\phi}_{1,2}} \right)_{1,0}^{(4)} = 0.$$

Im Folgenden werden wir die Gewichtsschranken für die Netzwerke $\hat{\phi}_{2,2}$ und $\hat{\phi}_{3,2}^{(\xi)}$ herleiten. Hierfür wenden wir Lemma 17a) an, wofür die Gewichtsschranken der einzelnen Teilnetzwerke benötigt werden.

Das Netzwerk

$$\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j = \sum_{i=1}^{M^d} (\mathbf{C}_{i,1})_{\text{left}} \cdot \widehat{f}_{\text{ind},C_{i,1}}(\mathbf{x}) + \widetilde{\mathbf{v}}_j$$

lässt sich als Komposition mehrerer neuronaler Teilnetze auffassen. Die entsprechenden Gewichtsschranken lassen sich daher mithilfe von Lemma 17 bestimmen.

Die Einträge des Vektors $\widetilde{\mathbf{v}}_j$ für $j \in \{1, \dots, M^d\}$ nehmen nur Werte aus der Menge $\{0, \frac{2a}{M^2}, \dots, (M-1) \cdot \frac{2a}{M^2}\}$ an. Die Schranken der äußeren Gewichte lassen sich dann direkt aus der Konstruktion des Netzes ablesen und sind gegeben durch

$$\left\| \left(\mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j} \right)_{1,j>0}^{(2)} \right\|_{\infty} = \|(\mathbf{C}_{i,1})_{\text{left}}\|_{\infty} \leq a \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j} \right)_{1,0}^{(2)} \right\|_{\infty} = \|\widetilde{\mathbf{v}}_j\|_{\infty} \leq \frac{2a}{M}.$$

Da zusätzlich

$$\left(\mathbf{w}_{\widehat{f}_{\text{ind},C_{i,1}}} \right)_{1,0}^{(2)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{ind},C_{i,1}}} \right)_{1,j>0}^{(2)} \right\|_{\infty} = 1$$

gilt, können wir Lemma 17c) anwenden und erhalten

$$\left\| \mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j} \right\|_{\infty} \leq \max \left\{ a, \frac{2a}{M}, \left\| \mathbf{w}_{\widehat{f}_{\text{ind},C_{i,1}}} \right\|_{\infty} \right\} = \max \left\{ a, \frac{2a}{M}, M^{2p+2} \right\} = M^{2p+2},$$

wobei die letzte Gleichheit aus $M^{2p+2} \geq a$ folgt. Analog ergeben sich wegen $\widetilde{v}_j^{(k)} + 2a/M^2 \leq 2a/M$ für $k \in \{1, \dots, d\}$ die Gewichtsschranken

$$\left\| \left(\mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j + 2a/M^2 \cdot \mathbf{1}} \right)_{1,j>0}^{(2)} \right\|_{\infty} = \|(\mathbf{C}_{i,1})_{\text{left}}\|_{\infty} \leq a, \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j + 2a/M^2 \cdot \mathbf{1}} \right)_{1,0}^{(2)} \right\|_{\infty} = \|\widetilde{\mathbf{v}}_j\|_{\infty} \leq \frac{2a}{M}$$

und

$$\left\| \mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j + 2a/M^2 \cdot \mathbf{1}} \right\|_{\infty} \leq M^{2p+2}.$$

Die Anwendung von Lemma 17a) liefert dann

$$\begin{aligned} & \left\| \mathbf{w}_{\widehat{\phi}_{2,2}} \right\|_{\infty} \\ & \leq \max \left\{ \left\| \mathbf{w}_{\widehat{f}_{\text{test}}} \right\|_{\infty}, \left\| \mathbf{w}_{\phi_{1,1}} \right\|_{\infty}, \left\| \mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j} \right\|_{\infty}, \left\| \mathbf{w}_{\widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j + 2a/M^2 \cdot \mathbf{1}} \right\|_{\infty}, \left\| \mathbf{w}_{\phi_{2,1}^{(i)} + \widetilde{v}_j^{(i)}} \right\|_{\infty}, \left\| \left(\mathbf{w}_{\widehat{f}_{\text{test}}} \right)^{(0)} \right\|_{\infty} \right. \\ & \cdot \left(4 \cdot \max \left\{ \left\| \left(\mathbf{w}_{\phi_{1,1}} \right)^{(2)} \right\|_{\infty}, \left\| \left(\mathbf{v}_{\phi_{2,1} + \widetilde{\mathbf{v}}_j} \right)^{(2)} \right\|_{\infty}, \left\| \left(\mathbf{v}_{\phi_{2,1} + \widetilde{\mathbf{v}}_j + 2a/M^2 \cdot \mathbf{1}} \right)^{(2)} \right\|_{\infty}, \left\| \left(\mathbf{w}_{\phi_{2,1}^{(i)} + \widetilde{v}_j^{(i)}} \right)^{(2)} \right\|_{\infty} \right\} + 1 \right) \\ & \leq M^{4p+4}. \end{aligned}$$

Aufgrund der bekannten Gewichtsschranken der Netzwerke $\widehat{\phi}_{1,1}$ und $\widehat{\phi}_{2,1}$ gilt zudem

$$\left\| \left(\mathbf{w}_{\widehat{\phi}_{2,2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1.$$

Zusätzlich folgt aus Lemma 27

$$\left(\mathbf{w}_{\widehat{\phi}_{2,2}}\right)_{1,0}^{(4)} = 0 \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{2,2}}\right)_{1,j>0}^{(4)} \right\|_{\infty} \leq 1.$$

Analog ergibt sich für das Netzwerk $\widehat{\phi}_{3,2}^{(\xi)}$, unter Verwendung der Gewichtsschranken für $\widehat{\phi}_{3,1}^{(\xi,j)}$ mit

$$\left\| \mathbf{w}_{\widehat{\phi}_{3,1}}^{(\xi,j)} \right\|_{\infty} \leq M^{2p+2} \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{3,1}}\right)^{(2)} \right\|_{\infty} \leq \|f\|_{C^q([-a,a]^d)},$$

dass

$$\left\| \mathbf{w}_{\widehat{\phi}_{3,2}}^{(\xi)} \right\|_{\infty} \leq M^{4p+4}$$

ist. Darüber hinaus folgt aus Lemma 27 für die äußersten Gewichte

$$\left(\mathbf{w}_{\widehat{\phi}_{3,2}}^{(\xi)}\right)_{1,0}^{(4)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{3,2}}^{(\xi)}\right)_{1,j>0}^{(4)} \right\|_{\infty} \leq 1.$$

Zudem implizieren die bekannten Schranken der innersten Gewichte der Netze $\widehat{\phi}_{1,1}$, $\widehat{\phi}_{2,1}$ und $\widehat{\phi}_{3,1}^{(\xi,j)}$, dass

$$\left\| \left(\mathbf{w}_{\widehat{\phi}_{3,2}}^{(\xi)}\right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1$$

ist.

Wir wählen nun $\xi_1, \dots, \xi_{\binom{d+q}{d}}$, so dass

$$\left\{ \xi_1, \dots, \xi_{\binom{d+q}{d}} \right\} = \left\{ (r_1, \dots, r_d) \in \mathbb{N}_0^d : r_1 + \dots + r_d \leq q \right\} \quad (\text{A.16})$$

gilt. Der Wert von $\phi_{1,3}$ kann durch

$$\widehat{\phi}_{1,3} = \widehat{f}_p \left(\mathbf{z}, y_1, \dots, y_{\binom{d+q}{d}} \right) \quad (\text{A.17})$$

bestimmt werden, wobei

$$\mathbf{z} = \widehat{\phi}_{1,2} - \widehat{\phi}_{2,2}$$

und

$$y_{\nu} = \widehat{\phi}_{3,2}^{(\xi_{\nu})}$$

für $\nu \in \left\{ 1, \dots, \binom{d+q}{d} \right\}$ ist. Wir wählen die Koeffizienten $r_1, \dots, r_{\binom{d+q}{d}}$ in Lemma 26 durch

$$r_i = \frac{1}{\xi_i!}, \quad i \in \left\{ 1, \dots, \binom{d+q}{d} \right\}.$$

Das Netzwerk $\widehat{\phi}_{1,3}$ bildet ein zusammengesetztes Netzwerk, wobei die Netzwerke $\widehat{\phi}_{1,1}$, $\widehat{\phi}_{2,1}$, $\widehat{\phi}_{3,1}^{(\xi_{\nu,1})}$, \dots , $\widehat{\phi}_{3,1}^{(\xi_{\nu,M^d})}$ und die Netzwerke $\widehat{\phi}_{1,2}$, $\widehat{\phi}_{2,2}$, $\widehat{\phi}_{3,2}^{(\xi_{\nu})}$ für $\nu \in \{1, \dots, \binom{d+q}{d}\}$ jeweils parallel berechnet werden. Da das Netzwerk $\widehat{f}_{\text{ind},C_{i,1}}$ zwei verdeckte Schichten hat, können wir schließen, dass

$$\left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1}, \widehat{\phi}_{3,1}^{(\xi_{\nu,1})}, \dots, \widehat{\phi}_{3,1}^{(\xi_{\nu,M^d})}\right)$$

$L_1 = 2$ verdeckte Schichten hat.

Das Netzwerk $\widehat{\phi}_{1,1}$ benötigt aufgrund der mehrdimensionalen Eingabe $2d$ Neuronen pro Schicht. Das Netzwerk $\widehat{f}_{\text{ind},C_{i,1}}$ besitzt ebenfalls eine Anzahl von $2d$ Neuronen pro Schicht. Da dieses Netzwerk M^d -mal aufsummiert wird und sich aus d Komponenten zusammensetzt, ergibt sich eine Anzahl von $d \cdot M^d \cdot 2d$ Neuronen pro Schicht. Des Weiteren erfüllen gemäß Gleichung (A.16) eine Anzahl von $\binom{d+q}{d}$ Vektoren $\xi = (\xi_1, \dots, \xi_d)$ die Ungleichung $\xi_1 + \dots + \xi_d \leq N$. Daher werden $\binom{d+q}{d}$ Netze $\widehat{\phi}_{3,2}^{(\xi_{\nu,i})}$ für $i \in \{1, \dots, M^d\}$ parallel berechnet. Aus diesem Grund benötigen wir $\binom{d+q}{d} \cdot 2d$ Neuronen pro Schicht für $\widehat{\phi}_{3,2}^{(\xi_{\nu,i})}$ mit $i \in \{1, \dots, M^d\}$. Da all diese Netze parallel berechnet werden, ergibt sich somit insgesamt eine Anzahl von

$$r_1 = 2d + d \cdot M^d \cdot 2d + M^d \cdot \binom{d+q}{d} \cdot 2d$$

Neuronen, die pro Schicht benötigt werden.

Abbildung A.1 veranschaulicht die beschriebene Berechnung in einem Graphen, basierend auf der Darstellung in Kohler und Langer (2021). Daraus wird ersichtlich, warum wir nur $c_{174} \cdot M^d$ statt $c_{174} \cdot M^{2d}$ Neuronen pro Schicht benötigen, um dieses Netzwerk zu realisieren. Neuronale Netze mit einer Breite von M^d besitzen M^{2d} Verbindungen zwischen zwei benachbarten Schichten. Dadurch ist es möglich, jede Ableitung von f für jeden Würfel aus \mathcal{P}_2 zu berechnen, indem die Ableitungen als Gewichte im Netzwerk verwendet werden.

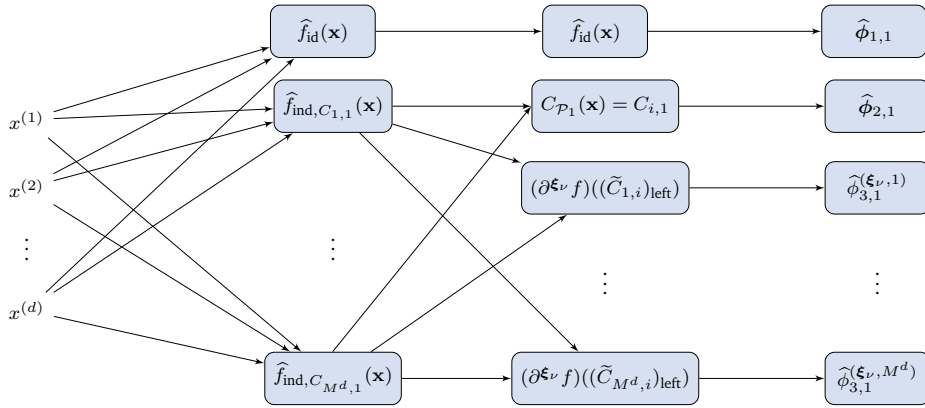


Abbildung A.1: Berechnung von $(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1}, \widehat{\phi}_{3,1}^{(\xi_{\nu,1})}, \dots, \widehat{\phi}_{3,1}^{(\xi_{\nu,M^d})})$

Zudem besitzt das Netzwerk

$$\left(\widehat{\phi}_{1,2}, \widehat{\phi}_{2,2}, \widehat{\phi}_{3,2}^{(\xi_{\nu})}\right)$$

insgesamt $L_2 = L_1 + 2 = 4$ verdeckte Schichten und

$$r_2 = \max\{r_1, 2d + d \cdot M^d \cdot 2 \cdot (2d + 2) + \binom{d+q}{d} \cdot M^d \cdot 2 \cdot (2d + 2)\}$$

$$= 2d + \left(d + \binom{d+q}{d} \right) \cdot M^d \cdot 2 \cdot (2d+2)$$

Neuronen pro Schicht. Die Anzahl der Neuronen ergibt sich hierbei analog zu r_1 , wobei wir für \hat{f}_{test} , gemäß Lemma 27 eine Anzahl von $2 \cdot (2d+2)$ Neuronen pro Schicht benötigen.

Aufgrund von Lemma 26 erhalten wir schließlich, dass $\hat{\phi}_{1,3}$ in der Klasse

$$\mathcal{F}(4 + B_{M,p} \cdot \lceil \log_2(\max\{q+1, 2\}) \rceil, r)$$

mit

$$B_{M,p} = \lceil \log_4(M^{2p}) \rceil$$

sowie

$$r = \max \left\{ r_2, 18 \cdot (q+1) \cdot \binom{d+q}{d} \right\}$$

liegt. Hierbei haben wir verwendet, dass

$$\mathcal{F}(L, r') \subseteq \mathcal{F}(L, r)$$

für $r' \leq r$ gilt. Wir setzen nun

$$\hat{f}_{\mathcal{P}_2}(\mathbf{x}) = \hat{\phi}_{1,3}.$$

Wegen der oben gezeigten Gewichtsschranken

$$\begin{aligned} \|\mathbf{w}_{\hat{\phi}_{1,2}}\|_\infty \leq 1, \quad \left(\mathbf{w}_{\hat{\phi}_{1,2}} \right)_{1,0}^{(4)} = 0, \quad \|\mathbf{w}_{\hat{\phi}_{2,2}}\|_\infty \leq M^{4p+4}, \quad \left(\mathbf{w}_{\hat{\phi}_{2,2}} \right)_{1,0}^{(4)} = 0, \\ \|\mathbf{w}_{\hat{\phi}_{3,2}(\xi_\nu)}\|_\infty \leq M^{4p+4} \quad \text{und} \quad \left(\mathbf{w}_{\hat{\phi}_{3,2}(\xi_\nu)} \right)_{1,0}^{(4)} = 0 \end{aligned}$$

sowie

$$\|\mathbf{w}_{\hat{f}_p}\|_\infty \leq \bar{r}(p) \cdot 4 \cdot 4^{2 \cdot (q+1)} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{2 \cdot (q+1)} \quad \text{und} \quad \left\| \left(\mathbf{w}_{\hat{f}_p} \right)_{i,j>0}^{(0)} \right\|_\infty \leq 1$$

können wir Lemma 17c) anwenden. Daraus ergibt sich aufgrund der Mindestgröße von M sowie $|\bar{r}(p)| \leq 1$ die folgende Beschränkung

$$\begin{aligned} \|\mathbf{w}_{\hat{f}_{\mathcal{P}_2}}\|_\infty = \|\mathbf{w}_{\hat{\phi}_{1,3}}\|_\infty &\leq \max \{ 4 \cdot 4^{2 \cdot (q+1)} \cdot \max \{ 2a, \|f\|_{C^q([-a,a]^d)} \}^{2 \cdot (q+1)}, M^{4p+4} \} \\ &\leq M^{4p+4}. \end{aligned}$$

Darüber hinaus liefert Lemma 26 die Bedingung

$$\left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2}} \right)_{1,0}^{(L)} = 0$$

und aus der Definition des Netzes sowie den bekannten Gewichtsbeschränkungen der Teilnetzwerke folgt

$$\left\| \left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_\infty \leq 1,$$

womit wir die Gewichtsschranken des Netzes nachgewiesen haben.

Im *zweiten Schritt des Beweises* analysieren wir den Fehler, den das Netz $\widehat{f}_{\mathcal{P}_2}$ bei der Approximation einer (p, C) -glaten Funktion macht, wenn

$$B_M \geq M^{2p+2}$$

und

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{1/M^{2p+2}}^0$$

gilt.

Aus Lemma 27 können wir schließen, dass die Netzwerke $\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1}, \widehat{\phi}_{3,1}^{(\xi_{\nu,1})}, \dots, \widehat{\phi}_{3,1}^{(\xi_{\nu,M^d})}$ sowie die Netzwerke $\widehat{\phi}_{1,2}, \widehat{\phi}_{2,2}, \widehat{\phi}_{3,2}^{(\xi_{\nu})}$ für $\nu \in \{1, \dots, \binom{d+q}{d}\}$ die zugehörigen Funktionen $\phi_{1,1}, \phi_{2,1}, \phi_{3,1}^{(\xi_{\nu,1})}, \dots, \phi_{3,1}^{(\xi_{\nu,M^d})}$ und $\phi_{1,2}, \phi_{2,2}, \phi_{3,2}^{(\xi_{\nu})}$ für $\nu \in \{1, \dots, \binom{d+q}{d}\}$ ohne einen Fehler berechnen, sofern $\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{1/M^{2p+2}}^0$. Daher folgt, dass

$$\left| \widehat{\phi}_{1,2} - \widehat{\phi}_{2,2} \right| = |\mathbf{x} - \phi_{2,2}| \leq 2a$$

und

$$\left| \widehat{\phi}_{3,2}^{(\xi_{\nu})} \right| = \left| \phi_{3,2}^{(\xi_{\nu})} \right| \leq \|f\|_{C^q([-a,a]^d)}$$

gilt. Somit liegt die Eingabe von \widehat{f}_p in (A.17) in dem Intervall, in dem (A.14) gilt. Durch die Wahl von $B_{M,p}$ und $r_i = \frac{1}{\xi_i!}$ für $i \in \{1, \dots, \binom{d+q}{d}\}$ erhalten wir

$$\left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x}) \right| = \left| \widehat{\phi}_{1,3} - \phi_{1,3} \right| \leq c_{173} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)} \cdot \frac{1}{M^{2p}},$$

wobei wir ausgenutzt haben, dass $\bar{r}(p) \leq 1$ ist. Aus Lemma 20 folgt hieraus

$$\begin{aligned} & \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - f(\mathbf{x}) \right| \\ & \leq \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x}) \right| + \left| T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x}) - f(\mathbf{x}) \right| \\ & \leq c_{173} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)} \cdot \frac{1}{M^{2p}} + c_{169} \cdot (2 \cdot a \cdot d)^p \cdot C \cdot \frac{1}{M^{2p}} \\ & \leq c_{175} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)} \cdot \frac{1}{M^{2p}}. \end{aligned}$$

Die letzte Ungleichung ergibt sich hierbei aus Voraussetzung (A.5).

Der Wert des Netzes ist dann beschränkt durch

$$\begin{aligned} \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| & \leq \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x}) \right| + \left| T_{f,q,(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}}(\mathbf{x}) - f(\mathbf{x}) \right| + |f(\mathbf{x})| \\ & \leq c_{173} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)} \cdot \frac{1}{M^{2p}} + c_{169} \cdot (2 \cdot a \cdot d)^p \cdot C \cdot \frac{1}{M^{2p}} + |f(\mathbf{x})| \\ & \leq 2 \cdot \max \left\{ \sup_{\mathbf{x} \in [-a,a]^d} |f(\mathbf{x})|, 1 \right\}, \end{aligned}$$

wobei wir hier

$$M^{2p} \geq c_{173} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)} \quad (\text{A.18})$$

und

$$M^{2p} \geq c_{169} \cdot (2 \cdot a \cdot d)^p \cdot C$$

verwendet haben.

Im letzten Schritt des Beweises leiten wir eine Schranke für $\widehat{f}_{\mathcal{P}_2}(\mathbf{x})$ im Fall

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0$$

her. In diesem Randbereich sind die Netzwerke $\widehat{f}_{\text{test}}$ und $\widehat{f}_{\text{ind}, C_{i,1}}$ gemäß Lemma 27 nicht exakt. Für $\mathbf{x} \in C_{i,1}$ mit $i \in \{1, \dots, M^d\}$ impliziert dies

$$\left| \widehat{\phi}_{3,1}^{(\xi, j)} \right| \leq \left| (\partial^{\xi} f) \left((\widetilde{C}_{j,i})_{\text{left}} \right) \right| \quad \text{für } j \in \{1, \dots, M^d\}$$

und

$$\left| \widehat{\phi}_{2,1}^{(s)} \right| \leq a \quad \text{für } s \in \{1, \dots, d\}.$$

Da $\widehat{f}_{\text{test}}$ höchstens einen Summanden in (A.15) mit einem Wert ungleich 0 erzeugt, führt dies zu

$$\left| \widehat{\phi}_{3,2}^{(\xi)} \right| \leq \|f\|_{C^q([-a,a]^d)}$$

und

$$\left| \widehat{\phi}_{2,2}^{(s)} \right| \leq a, \quad \text{für } s \in \{1, \dots, d\}.$$

Somit liegt die Eingabe von \widehat{f}_p in (A.17) erneut in dem Intervall, in dem (A.14) gilt. Aufgrund der Wahl von

$$B_{M,p} = \lceil \log_4(M^{2p}) \rceil$$

und $r_i = \frac{1}{\xi_i!}$ für $i \in \{1, \dots, \binom{d+q}{d}\}$ können wir daher unter Verwendung der Voraussetzung (A.5) folgern, dass

$$\begin{aligned} \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| &\leq \left| \widehat{f}_p(\mathbf{z}, y_1, \dots, y_{\binom{d+q}{d}}) - p(\mathbf{z}, y_1, \dots, y_{\binom{d+q}{d}}) \right| + \left| p(\mathbf{z}, y_1, \dots, y_{\binom{d+q}{d}}) \right| \\ &\leq c_{173} \cdot a^{4(q+1)} \cdot 4^{-B_{M,p}} + \sum_{0 \leq \|\xi\|_1 \leq q} \frac{1}{\xi!} \cdot \|f\|_{C^q([-a,a]^d)} \cdot (2a)^{\|\xi\|_1} \\ &\leq 1 + \|f\|_{C^q([-a,a]^d)} \cdot \left(\sum_{\xi=0}^{\infty} \frac{(2a)^\xi}{\xi!} \right)^d \\ &= 1 + \exp(2ad) \cdot \|f\|_{C^q([-a,a]^d)} \end{aligned}$$

erfüllt ist. □

A.1.5 Schritt 3: Approximation von $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ durch neuronale Netze

Um $f(\mathbf{x})$ in Supremumsnorm zu approximieren, verwenden wir das neuronale Netz $\widehat{f}_{\mathcal{P}_2}$ aus Lemma 22. Mit dessen Hilfe kann ein Netzwerk konstruiert werden, welches

$$w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$$

approximiert, wobei

$$w_{\mathcal{P}_2}(\mathbf{x}) = \prod_{j=1}^d \left(1 - \frac{M^2}{a} \cdot \left| (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2} - x^{(j)} \right| \right)_+ \quad (\text{A.19})$$

ein linearer Tensorprodukt-B-Spline vom Grad 1 ist. Dies folgt aus

$$\begin{aligned} & \left(1 - \frac{M^2}{a} \cdot \left| (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2} - x^{(j)} \right| \right)_+ \\ &= \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} \right) \right)_+ \\ & \quad - 2 \cdot \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} - \frac{a}{M^2} \right) \right)_+ \\ & \quad + \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} - \frac{2 \cdot a}{M^2} \right) \right)_+ \\ &= \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} \right) \right) \cdot \mathbb{1}_{[(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)}, (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2}]} \left(x^{(j)} \right) \\ & \quad + \left(\frac{M^2}{a} \cdot \left((C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{2a}{M^2} - x^{(j)} \right) \right) \cdot \mathbb{1}_{[(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2}, (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{2a}{M^2}]} \left(x^{(j)} \right) \end{aligned}$$

für $j \in \{1, \dots, d\}$.

Dieser nimmt seinen Maximalwert im Zentrum von $C_{\mathcal{P}_2}(\mathbf{x})$ an, ist ungleich 0 im Inneren von $C_{\mathcal{P}_2}(\mathbf{x})$ und verschwindet außerhalb von $C_{\mathcal{P}_2}(\mathbf{x})$.

Ist \mathbf{x} in der Menge

$$\begin{aligned} & \bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0 \\ &= \bigcup_{j \in \{1, \dots, d\}} \bigcup_{k \in \{1, \dots, M^{2d}\}} \left\{ \mathbf{x} \in [-a, a]^d : \left| x^{(j)} - (C_{k,2})_{\text{left}}^{(j)} \right| < \frac{1}{M^{2p+2}} \right\} \end{aligned} \quad (\text{A.20})$$

enthalten, so ist $w_{\mathcal{P}_2}(\mathbf{x})$ kleiner oder gleich $1/M^{2p}$. Um dies zu zeigen, nehmen wir ohne Beschränkung der Allgemeinheit an, dass

$$\left| x^{(1)} - (C_{k,2})_{\text{left}}^{(1)} \right| < \frac{1}{M^{2p+2}}$$

gilt, woraus

$$w_{\mathcal{P}_2}(\mathbf{x}) = \left(1 - \frac{M^2}{a} \cdot \left| (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(1)} + \frac{a}{M^2} - x^{(1)} \right| \right)_+ \cdot \prod_{j=2}^d \left(1 - \frac{M^2}{a} \cdot \left| (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2} - x^{(j)} \right| \right)_+$$

$$\begin{aligned}
&\leq \left(1 - \frac{M^2}{a} \cdot \left(-\frac{1}{M^{2p+2}} + \frac{a}{M^2}\right)\right)_+ \cdot 1 \\
&\leq \frac{1}{a \cdot M^{2p}}
\end{aligned} \tag{A.21}$$

folgt.

Da $w_{\mathcal{P}_2}(\mathbf{x})$ in der Nähe des Randes $C_{\mathcal{P}_2}(\mathbf{x})$ nahe 0 ist, wird es mit Hilfe des Netzes aus Lemma 22 möglich sein, ein neuronales Netz zu konstruieren, welches $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ in der Supremumsnorm approximiert.

Lemma 28. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Sei $p = q + s$ für ein $q \in \mathbb{N}_0$ und ein $s \in (0, 1]$, sowie $C > 0$. Zudem sei $f : \mathbb{R}^d \rightarrow \mathbb{R}$ eine (p, C) -glatte Funktion. Des Weiteren seien $1 \leq a < \infty$ und $M \in \mathbb{N}_0$ hinreichend groß, wobei M unabhängig von der Größe von a ist, aber zumindest

$$\begin{aligned}
M^{2p} &\geq \max\{c_{170}, c_{173}\} \cdot \left(\max\left\{2a, \|f\|_{C^q([-a, a]^d)}\right\}\right)^{4 \cdot (q+1)}, \\
M^{2p} &\geq c_{169} \cdot (2 \cdot a \cdot d)^p \cdot C
\end{aligned}$$

und

$$M^{2p} \geq 2 \cdot \exp(2ad) \cdot \max\left\{\|f\|_{C^q([-a, a]^d)}, 1\right\}$$

erfüllt. Sei weiter $w_{\mathcal{P}_2}$ definiert wie in Gleichung (A.19). Dann existiert ein neuronales Netz

$$\hat{f} \in \mathcal{F}(L, r)$$

mit

$$L = 5 + \lceil \log_4(M^{2p}) \rceil \cdot (\lceil \log_2(\max\{q, d\} + 1) \rceil + 1)$$

und

$$r = 64 \cdot \binom{d+q}{d} \cdot d^2 \cdot (q+1) \cdot M^d$$

sowie den Gewichtsschranken

$$\|\mathbf{w}_{\hat{f}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\hat{f}}\right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1 \quad \text{und} \quad \left(\mathbf{w}_{\hat{f}}\right)_{1,0}^{(L)} = 0,$$

so dass

$$\left| \hat{f}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \leq c_{176} \cdot \left(\max\left\{2a, \|f\|_{C^q([-a, a]^d)}\right\}\right)^{4 \cdot (q+1)} \cdot \frac{1}{M^{2p}}$$

für $\mathbf{x} \in [-a, a]^d$ gilt.

Für den Beweis dieses Lemmas benötigen wir einige Hilfsresultate. In Lemma 29 werden wir zeigen, dass jedes Gewicht $w_{\mathcal{P}_2}(\mathbf{x})$ (siehe Gleichung (A.19)) durch ein neuronales Netz approximiert werden kann, sofern \mathbf{x} nicht nah am Rand eines Würfels der Partition \mathcal{P}_2 liegt. Die Werte von $(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)}$ können hierbei, wie für $\hat{\phi}_{2,2}$ im Beweis von Lemma 22 beschrieben, bestimmt werden. Zusätzlich verwenden wir, dass

$$\left(1 - \frac{M^2}{a} \cdot \left|(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2} - x^{(j)}\right|\right)_+ = \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)}\right)\right)_+$$

$$\begin{aligned}
& -2 \cdot \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} - \frac{a}{M^2} \right) \right)_+ \\
& + \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} - \frac{2 \cdot a}{M^2} \right) \right)_+
\end{aligned}$$

für $j \in \{1, \dots, d\}$ gilt, weshalb jeder Faktor durch die Anwendung der ReLU-Aktivierungsfunktion berechnet werden kann. Das finale Produkt wird mithilfe des folgenden Lemmas approximiert.

Lemma 29. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Sei $1 \leq a < \infty$ und $M^{2p} \geq 4^{4d+1} \cdot d$. Sei weiter \mathcal{P}_2 die Partition, welche in (A.1) definiert ist und seien $w_{\mathcal{P}_2}(\mathbf{x})$ die zugehörigen Gewichte, welche durch (A.19) definiert sind. Dann existiert ein neuronales Netz

$$\widehat{f}_{w_{\mathcal{P}_2}} \in \mathcal{F}(L, r)$$

mit

$$L = 5 + \lceil \log_4(M^{2p}) \rceil \cdot \lceil \log_2(d) \rceil$$

und

$$r = \max \left\{ 18d, 2d + d \cdot M^d \cdot 2 \cdot (2 + 2d) \right\}$$

sowie den Gewichtsschranken

$$\|\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2}}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1 \quad \text{und} \quad \left(\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2}}} \right)_{1,0}^{(L)} = 0,$$

so dass

$$\left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \right| \leq 4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}}$$

für $\mathbf{x} \in \bigcup_{i \in \{1, \dots, M^{2d}\}} (C_{i,2})_{1/M^{2p+2}}^0$ und

$$\left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \right| \leq 2$$

für $\mathbf{x} \in [-a, a]^d$ erfüllt ist.

Beweis. Im Folgenden werden wir den Aufbau des neuronalen Netzes $\widehat{f}_{w_{\mathcal{P}_2}}$ beschreiben.

Die ersten vier verdeckten Schichten von $\widehat{f}_{w_{\mathcal{P}_2}}$ berechnen den Wert von

$$(C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}$$

und verschieben anschließend den Wert von \mathbf{x} in die nächste verdeckte Schicht. Dies kann wie bei den neuronalen Netzen $\widehat{\phi}_{1,2}$ und $\widehat{\phi}_{2,2}$ im Beweis von Lemma 22 beschrieben, erfolgen.

Das Netz

$$\widehat{\phi}_{1,2} = \left(\widehat{\phi}_{1,2}^{(1)}, \dots, \widehat{\phi}_{1,2}^{(d)} \right) = \widehat{f}_{\text{id}}^2(\widehat{\phi}_{1,1})$$

hat $2d$ Neuronen pro Schicht.

Da

$$\widehat{\phi}_{2,2}^{(i)} = \sum_{j=1}^{M^d} \widehat{f}_{\text{test}} \left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1} + \mathbf{v}_j, \widehat{\phi}_{2,1} + \widetilde{\mathbf{v}}_j + \frac{2a}{M^2} \cdot \mathbf{1}, \widehat{\phi}_{2,1}^{(i)} + \widetilde{v}_j^{(i)} \right)$$

ist und das Netzwerk $\widehat{f}_{\text{test}}$ gemäß Lemma 27 eine Anzahl von $2 \cdot (2 + 2d)$ Neuronen hat, ergibt sich durch das M^d -malige Aufsummieren, dass dieses Netz $M^d \cdot 2 \cdot (2 + 2d)$ Neuronen benötigt. Für das Netz $\widehat{\phi}_{2,2}$, welches durch

$$\widehat{\phi}_{2,2} = \left(\widehat{\phi}_{2,2}^{(1)}, \dots, \widehat{\phi}_{2,2}^{(d)} \right),$$

gegeben ist, erhalten wir damit eine Anzahl von $d \cdot M^d \cdot 2 \cdot (2 + 2d)$ Neuronen pro Schicht. Insgesamt haben wir somit $2d + d \cdot M^d \cdot 2 \cdot (2 + 2d)$ Neuronen pro Schicht.

In der fünften verdeckten Schicht werden dann die Funktionen

$$\begin{aligned} \left(1 - \frac{M^2}{a} \cdot \left| (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2} - x^{(j)} \right| \right)_+ &= \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} \right) \right)_+ \\ &\quad - 2 \cdot \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} - \frac{a}{M^2} \right) \right)_+ \\ &\quad + \left(\frac{M^2}{a} \cdot \left(x^{(j)} - (C_{\mathcal{P}_2}(\mathbf{x}))_{\text{left}}^{(j)} - \frac{2 \cdot a}{M^2} \right) \right)_+ \end{aligned}$$

für $j \in \{1, \dots, d\}$ berechnet. Dies geschieht mithilfe der Netzwerke $\widehat{\phi}_{1,2}$ und $\widehat{\phi}_{2,2}$ gemäß

$$\begin{aligned} \widehat{f}_{w_{\mathcal{P}_2,j}}(\mathbf{x}) &= \sigma \left(\frac{M^2}{a} \cdot \left(\widehat{\phi}_{1,2}^{(j)} - \widehat{\phi}_{2,2}^{(j)} \right) \right) \\ &\quad - 2 \cdot \sigma \left(\frac{M^2}{a} \cdot \left(\widehat{\phi}_{1,2}^{(j)} - \widehat{\phi}_{2,2}^{(j)} - \frac{a}{M^2} \right) \right) \\ &\quad + \sigma \left(\frac{M^2}{a} \cdot \left(\widehat{\phi}_{1,2}^{(j)} - \widehat{\phi}_{2,2}^{(j)} - \frac{2 \cdot a}{M^2} \right) \right) \end{aligned}$$

mit $3d$ Neuronen.

Aus dem Beweis von Lemma 22 folgt, dass die Gewichtsschranken für die Netze $\widehat{\phi}_{1,2}$ und $\widehat{\phi}_{2,2}$ durch

$$\|\mathbf{w}_{\widehat{\phi}_{1,2}}\|_{\infty} \leq 1, \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{1,2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1, \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{1,2}} \right)_{1,j>0}^{(4)} \right\|_{\infty} \leq 1, \quad \left(\mathbf{w}_{\widehat{\phi}_{1,2}} \right)_{1,0}^{(4)} = 0$$

sowie

$$\|\mathbf{w}_{\widehat{\phi}_{2,2}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{2,2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1, \quad \left\| \left(\mathbf{w}_{\widehat{\phi}_{2,2}} \right)_{1,j>0}^{(4)} \right\|_{\infty} \leq 1 \quad \text{und} \quad \left(\mathbf{w}_{\widehat{\phi}_{2,2}} \right)_{1,0}^{(4)} = 0$$

gegeben sind. Damit erhalten wir durch die Anwendung von Lemma 17a) für $j \in \{1, \dots, d\}$ die Gewichtsschranke

$$\left\| \mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2,j}}} \right\|_{\infty} \leq M^{4p+4}.$$

Aus der Definition des Netzes $\widehat{f}_{w_{\mathcal{P}_2,j}}$ lässt sich ablesen, dass

$$\left(\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2,j}}} \right)_{1,0}^{(5)} = 0 \quad \text{sowie} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2,j}}} \right)_{1,j>0}^{(5)} \right\| \leq 2 \quad \text{für } j \in \{1, \dots, d\}$$

gelten.

Das Produkt in $w_{\mathcal{P}_2,j}(\mathbf{x})$ für $j \in \{1, \dots, d\}$ kann durch das Netzwerk $\widehat{f}_{\text{mult},d}$ aus Lemma 25 berechnet werden. Dabei haben wir $x^{(j)} = \widehat{f}_{w_{\mathcal{P}_2,j}}(\mathbf{x})$ und $R = \lceil \log_4(M^{2p}) \rceil$ gewählt. Schließlich setzen wir

$$\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) = \widehat{f}_{\text{mult},d} \left(\widehat{f}_{w_{\mathcal{P}_2,1}}(\mathbf{x}), \dots, \widehat{f}_{w_{\mathcal{P}_2,d}}(\mathbf{x}) \right).$$

Das Netzwerk liegt dann in der Klasse

$$\mathcal{F}(L, r)$$

mit

$$L = 4 + 1 + \lceil \log_4(M^{2p}) \rceil \cdot \lceil \log_2(d) \rceil$$

und

$$r = \max \left\{ 18d, 2d + d \cdot M^d \cdot 2 \cdot (2 + 2d), 3d \right\}.$$

Für die Bestimmung der Gewichtsschranken dieses Netzes benötigen wir eine Aussage darüber, in welchem Intervall sich die Netze $\widehat{f}_{w_{\mathcal{P}_2,j}}$ für $j \in \{1, \dots, d\}$ befinden. Bei dem Netzwerk $\widehat{f}_{w_{\mathcal{P}_2,j}}$ handelt es sich, unabhängig von der Eingabe, um eine Hutfunktion, die ihren Maximalwert bei 1 annimmt. Daher ergibt sich

$$\left| \widehat{f}_{w_{\mathcal{P}_2,j}}(\mathbf{x}) \right| \leq 1$$

für $j \in \{1, \dots, d\}$ und $\mathbf{x} \in [-a, a]^d$. Aus diesem Grund setzen wir $a = 1$ in Lemma 25 und erhalten die Gewichtsschranken

$$\left\| \left(\mathbf{w}_{\widehat{f}_{\text{mult},d}} \right) \right\|_{\infty} \leq 4^{2d+1}, \quad \left(\mathbf{w}_{\widehat{f}_{\text{mult},d}} \right)_{1,0}^{(R \cdot \lceil \log_2(d) \rceil)} = 0 \quad \text{und} \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{mult},d}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1.$$

Da zusätzlich $\left(\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2,j}}} \right)_{1,0}^{(5)} = 0$ ist, können wir Lemma 17c) anwenden und erhalten die Gewichtsschranke

$$\left\| \mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2}}} \right\|_{\infty} \leq \max \left\{ \left\| \mathbf{w}_{\widehat{f}_{\text{mult},d}} \right\|_{\infty}, \left\| \mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2,j}}} \right\|_{\infty} \right\} = \max \{ 4^{2d+1}, M^{4p+4} \} = M^{4p+4},$$

wobei die letzte Gleichheit aus $M^{2p} \geq 4^{4d+1}$ folgt. Des Weiteren ergibt sich

$$\left\| \left(\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = \max \left\{ \left\| \left(\mathbf{w}_{\widehat{\phi}_{1,2}} \right)_{i,j>0}^{(0)} \right\|_{\infty}, \left\| \left(\mathbf{w}_{\widehat{\phi}_{2,2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \right\} \leq 1$$

sowie aufgrund von Lemma 25

$$\left(\mathbf{w}_{\widehat{f}_{w_{\mathcal{P}_2}}} \right)_{1,0}^{(L)} = 0.$$

Gemäß Lemma 25, wobei wir $R = \lceil \log_4(M^{2p}) \rceil$ und $a = 1$ (da $|\widehat{f}_{w_{\mathcal{P}_2,j}}(\mathbf{x})| \leq 1$) setzen, approximiert das Netz $\widehat{f}_{w_{\mathcal{P}_2}}$ die Funktion $w_{\mathcal{P}_2}(\mathbf{x})$ mit einem Fehler der Größe

$$4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}},$$

sofern \mathbf{x} in $\bigcup_{i \in \{1, \dots, M^{2d}\}} (C_{i,2})_{1/M^{2p+2}}^0$ enthalten ist.

Aus der Tatsache, dass $|\widehat{f}_{w_{\mathcal{P}_2},j}(\mathbf{x})| \leq 1$ für $j \in \{1, \dots, d\}$ und $\mathbf{x} \in [-a, a]^d$ gilt, können wir den Wert des Netzes unter Verwendung der Dreiecksungleichung durch

$$\left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \right| \leq \left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) - \prod_{j=1}^d \widehat{f}_{w_{\mathcal{P}_2},j}(\mathbf{x}) \right| + \left| \prod_{j=1}^d \widehat{f}_{w_{\mathcal{P}_2},j}(\mathbf{x}) \right| \leq 4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}} + 1 \leq 2$$

für $\mathbf{x} \in [-a, a]^d$ abschätzen, wobei wir erneut

$$M^{2p} \geq 4^{4d+1} \cdot d$$

verwendet haben. Daraus folgt die Aussage des Lemmas. \square

Das Netzwerk $\widehat{f}_{\mathcal{P}_2}$ aus Lemma 22 und das Netzwerk $\widehat{f}_{w_{\mathcal{P}_2}}$ aus Lemma 29 sind jeweils nur gute Approximationen für $f(\mathbf{x})$ beziehungsweise $w_{\mathcal{P}_2}(\mathbf{x})$, sofern

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{1/M^{2p+2}}^0$$

ist. Aus diesem Grund konstruieren wir im Folgenden ein Netz, mit dem wir den Approximationsfehler im Fall, dass \mathbf{x} in

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0 \quad (\text{A.22})$$

liegt, kontrollieren können. Dieses Netzwerk wird den Wert 1 annehmen, wenn \mathbf{x} in (A.22) enthalten ist. Des Weiteren wird es den Wert 0 annehmen, wenn \mathbf{x} in der Menge

$$\bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{2/M^{2p+2}}^0$$

enthalten ist. Daher sagen wir, dass dieses Netzwerk die Position unserer Eingabe \mathbf{x} *überprüft* (englisch: *check*).

Eine Möglichkeit für die Approximation von

$$\mathbb{1}_{\bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0}(\mathbf{x}) = 1 - \sum_{k \in \{1, \dots, M^{2d}\}} \mathbb{1}_{(C_{k,2})_{1/M^{2p+2}}^0}(\mathbf{x})$$

wäre es, jede der M^{2d} Indikatorfunktionen durch neuronale Netze anzunähern. In diesem Fall würde allerdings die Gesamtzahl der Neuronen pro Schicht in der Größenordnung M^{2d} liegen.

Daher gehen wir für die Konstruktion des Netzes in zwei Schritten vor:

In einem ersten Schritt berechnen wir die Position von $(C_{\mathcal{P}_1}(\mathbf{x}))_{\text{left}}$, wie es für das Netz $\widehat{\phi}_{2,1}$ in Lemma 22 beschrieben wurde.

Sei nun $i \in \{1, \dots, M^d\}$, so dass $(C_{\mathcal{P}_1}(\mathbf{x})) = C_{i,1}$ gilt. In einem zweiten Schritt muss lediglich noch

$$\mathbb{1}_{\bigcup_{j \in \{1, \dots, M^d\}} \widetilde{C}_{j,i} \setminus (\widetilde{C}_{j,i})_{1/M^{2p+2}}^0}(\mathbf{x}) = 1 - \sum_{j \in \{1, \dots, M^d\}} \mathbb{1}_{(\widetilde{C}_{j,i})_{1/M^{2p+2}}^0}(\mathbf{x})$$

approximiert werden. Dies kann mit $c_{177} \cdot M^d$ Neuronen pro Schicht erreicht werden.

Das Netzwerk $\widehat{\phi}_{2,1}$ ist nur eine gute Approximation von $(C_{i,1})_{\text{left}}$, sofern

$$\mathbf{x} \notin \bigcup_{k \in \{1, \dots, M^d\}} C_{k,1} \setminus (C_{k,1})_{1/M^{2p+2}}^0$$

ist. Daher müssen wir überprüfen, ob \mathbf{x} am Rand der groben Partition \mathcal{P}_1 liegt, das heißt, ob

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^d\}} C_{k,1} \setminus (C_{k,1})_{1/M^{2p+2}}^0$$

gilt. Hierfür berechnen wir

$$\mathbb{1}_{\bigcup_{k \in \{1, \dots, M^d\}} C_{k,1} \setminus (C_{k,1})_{1/M^{2p+2}}^0} = 1 - \sum_{k \in \{1, \dots, M^d\}} \mathbb{1}_{(C_{k,1})_{1/M^{2p+2}}^0}(\mathbf{x}) \quad (\text{A.23})$$

mithilfe der neuronalen Netze aus Lemma 27a). Kombinieren wir dies, so soll unser finales Netzwerk

$$1 - \sigma(1 - \mathbb{1}_{\bigcup_{j \in \{1, \dots, M^d\}} \widetilde{C}_{j,i} / (\widetilde{C}_{j,i})_{1/M^{2p+2}}^0}(\mathbf{x}) - \mathbb{1}_{\bigcup_{k \in \{1, \dots, M^d\}} C_{k,1} / (C_{k,1})_{1/M^{2p+2}}^0}(\mathbf{x})) \quad (\text{A.24})$$

berechnen. Dabei haben wir ausgenutzt, dass die ReLU-Aktivierungsfunktion bei negativem Input 0 ist. Die zweite Indikatorfunktion in (A.24) wird durch die Netzwerke \widehat{f}_{ind} aus Lemma 27a) berechnet, während die erste Indikatorfunktion durch die Netzwerke $\widehat{f}_{\text{test}}$ in Lemma 27b) approximiert wird.

Lemma 30. Sei $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ die ReLU-Aktivierungsfunktion $\sigma(z) = \max\{z, 0\}$. Sei $1 \leq a < \infty$ und $M \in \mathbb{N}$ hinreichend groß sowie unabhängig von der Größe von a , mit einer Mindestgröße von $M^{2p} \geq a$. Zudem seien die Partitionen \mathcal{P}_1 und \mathcal{P}_2 wie in (A.1) definiert. Dann existiert ein neuronales Netzwerk

$$\widehat{f}_{\text{check}, \mathcal{P}_2} \in \mathcal{F} \left(5, 2d + (4d^2 + 4d) \cdot M^d \right)$$

mit den Gewichtsschranken

$$\|\mathbf{w}_{\widehat{f}_{\text{check}, \mathcal{P}_2}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{check}, \mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1, \quad \left\| \left(\mathbf{w}_{\widehat{f}_{\text{check}, \mathcal{P}_2}} \right)_{1,j>1}^{(5)} \right\|_{\infty} = 1 \quad \text{und} \quad \left(\mathbf{w}_{\widehat{f}_{\text{check}, \mathcal{P}_2}} \right)_{1,0}^{(5)} = 1,$$

welches

$$\widehat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) = \mathbb{1}_{\bigcup_{i \in \{1, \dots, M^{2d}\}} C_{i,2} \setminus (C_{i,2})_{1/M^{2p+2}}^0}(\mathbf{x})$$

für $\mathbf{x} \notin \bigcup_{i \in \{1, \dots, M^{2d}\}} (C_{i,2})_{1/M^{2p+2}}^0 \setminus (C_{i,2})_{2/M^{2p+2}}^0$ und

$$\widehat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) \in [0, 1]$$

für $\mathbf{x} \in [-a, a]^d$ erfüllt.

Beweis. Wir nehmen im gesamten Beweis an, dass die Bedingung $C_{\mathcal{P}_1}(\mathbf{x}) = C_{i,1}$ für ein $i \in \{1, \dots, M^d\}$ erfüllt ist. Die oben beschriebene Approximation verläuft wie folgt:

Im ersten Teil überprüfen wir, ob \mathbf{x} in der Menge

$$\bigcup_{k \in \{1, \dots, M^d\}} C_{k,1} \setminus (C_{k,1})_{1/M^{2p+2}}^0$$

enthalten ist. Hierfür wird in den ersten beiden Schichten des Netzwerks die Funktion

$$f_1(\mathbf{x}) = \mathbb{1}_{\bigcup_{i \in \{1, \dots, M^d\}} C_{i,1} \setminus (C_{i,1})_{1/M^{2p+2}}^0}(\mathbf{x}) = 1 - \sum_{i \in \{1, \dots, M^d\}} \mathbb{1}_{(C_{i,1})_{1/M^{2p+2}}^0}(\mathbf{x})$$

durch

$$\hat{f}_1(\mathbf{x}) = 1 - \sum_{k \in \{1, \dots, M^d\}} \hat{f}_{\text{ind}, (C_{k,1})_{1/M^{2p+2}}^0}(\mathbf{x})$$

approximiert, wobei $\hat{f}_{\text{ind}, (C_{k,1})_{1/M^{2p+2}}^0}$ die Netzwerke aus Lemma 27 mit $R = M^{2p+2}$ sind.

Aus den Gewichtsschranken in Lemma 27a) können wir für $R = M^{2p+2}$ folgern, dass

$$\|\mathbf{w}_{\hat{f}_1}\|_\infty \leq \max \left\{ 1, \|\mathbf{w}_{\hat{f}_{\text{ind}, (C_{k,1})_{1/M^{2p+2}}^0}}\|_\infty \right\} = M^{2p+2}$$

und

$$\left\| \left(\mathbf{w}_{\hat{f}_1} \right)_{i,j>0}^{(0)} \right\|_\infty = \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind}, (C_{k,1})_{1/M^{2p+2}}^0}} \right)_{i,j>0}^{(0)} \right\|_\infty \leq 1$$

gilt.

Um die Indikatorfunktion auf der Partition \mathcal{P}_2 nur für die Würfel $C_{k,2} \subset C_{\mathcal{P}_1}(\mathbf{x})$ zu approximieren, müssen wir die Position von $(C_{\mathcal{P}_1}(\mathbf{x}))_{\text{left}}$ bestimmen. Dies kann, wie im Beweis von Lemma 22 für das Netzwerk $\hat{\phi}_{2,1}$ beschrieben, mit $d \cdot M^d \cdot 2d$ Neuronen durchgeführt werden. Um den Wert von \mathbf{x} in die nächsten verdeckten Schichten zu verschieben, wenden wir außerdem das Netzwerk \hat{f}_{id}^2 an. Dieses Netz benötigt $2d$ Neuronen pro Schicht.

Analog zu Gleichung (A.3) können wir die Würfel $(\tilde{C}_{j,i})_{1/M^{2p+2}}^0$ für $j \in \{1, \dots, M^d\}$, die in $C_{i,1}$ enthalten sind, durch

$$\begin{aligned} (\mathcal{A}^{(j)})_{1/M^{2p+2}}^0 = & \left\{ \mathbf{x} \in \mathbb{R}^d : -x^{(k)} + \phi_{2,1}^{(k)} + \tilde{v}_j^{(k)} + \frac{1}{M^{2p+2}} \leq 0 \right. \\ & \left. \text{und } x^{(k)} - \phi_{2,1}^{(k)} - \tilde{v}_j^{(k)} - \frac{2a}{M^2} + \frac{1}{M^{2p+2}} < 0 \text{ für alle } k \in \{1, \dots, d\} \right\} \end{aligned}$$

beschreiben. Die Funktion

$$f_2(\mathbf{x}) = \mathbb{1}_{\bigcup_{j \in \{1, \dots, M^d\}} \tilde{C}_{j,i} \setminus (\tilde{C}_{j,i})_{1/M^{2p+2}}^0}(\mathbf{x}) = 1 - \sum_{j \in \{1, \dots, M^d\}} \mathbb{1}_{(\tilde{C}_{j,i})_{1/M^{2p+2}}^0}(\mathbf{x})$$

kann daher durch

$$\hat{f}_2(\mathbf{x}) = 1 - \sum_{j \in \{1, \dots, M^d\}} \hat{f}_{\text{test}} \left(\hat{f}_{\text{id}}^2(\mathbf{x}), \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{1}{M^{2p+2}} \cdot \mathbf{1}, \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{2a}{M^2} \cdot \mathbf{1} - \frac{1}{M^{2p+2}} \cdot \mathbf{1}, 1 \right)$$

approximiert werden, wobei \hat{f}_{test} das neuronale Netzwerk aus Lemma 27b) mit $R = M^{2p+2}$ ist.

Die Gewichtsschranken des Netzwerks $\hat{f}_2(\mathbf{x})$ erhalten wir wie im Beweis von Lemma 22. Da alle Einträge des Vektors $\tilde{\mathbf{v}}_j$ in der Menge $\{0, \frac{2a}{M^2}, \dots, (M-1) \cdot \frac{2a}{M^2}\}$ liegen, ist

$$\|\mathbf{w}_{\hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{1}{M^{2p+2}} \cdot \mathbf{1}}\|_{\infty} \leq \max \left\{ a, (M-1) \cdot \frac{2a}{M^2} + \frac{1}{M^{2p+2}}, \left\| \mathbf{w}_{\hat{f}_{\text{ind}, C_{i,1}}} \right\|_{\infty} \right\} = M^{2p+2}$$

und

$$\|\mathbf{w}_{\hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{2a}{M^2} \cdot \mathbf{1} - \frac{1}{M^{2p+2}} \cdot \mathbf{1}}\|_{\infty} \leq \max \left\{ a, \frac{2a}{M} - \frac{1}{M^{2p+2}}, \left\| \mathbf{w}_{\hat{f}_{\text{ind}, C_{i,1}}} \right\|_{\infty} \right\} = M^{2p+2}.$$

Daraus ergibt sich wie im Beweis von Lemma 22 aus Lemma 17a), dass

$$\|\mathbf{w}_{\hat{f}_2}\|_{\infty} \leq M^{4p+4}$$

gilt. Zudem folgt aus Lemma 27a) die Gewichtsschranke

$$\left\| \left(\mathbf{w}_{\hat{f}_2} \right)_{i,j>0}^{(0)} \right\|_{\infty} = \max \left\{ 1, \left\| \left(\mathbf{w}_{\phi_{2,1}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \right\} = \max \left\{ 1, \left\| \left(\mathbf{w}_{\hat{f}_{\text{ind}, C_{i,1}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \right\} \leq 1.$$

Durch die Kombination der Netzwerke \hat{f}_1 und \hat{f}_2 sowie der Tatsache, dass die ReLU-Aktivierungsfunktion bei negativer Eingabe 0 ist, können wir

$$\mathbb{1}_{\bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})^0_{1/M^{2p+2}}}(\mathbf{x})$$

durch

$$\begin{aligned} \hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) &= 1 - \sigma \left(1 - \hat{f}_2(\mathbf{x}) - \hat{f}_{\text{id}}^2 \left(\hat{f}_1(\mathbf{x}) \right) \right) \\ &= 1 - \sigma \left(\sum_{j \in \{1, \dots, M^d\}} \hat{f}_{\text{test}} \left(\hat{f}_{\text{id}}^2(\mathbf{x}), \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{1}{M^{2p+2}} \cdot \mathbf{1}, \hat{\phi}_{2,1} + \tilde{\mathbf{v}}_j + \frac{2a}{M^2} \cdot \mathbf{1} - \frac{1}{M^{2p+2}} \cdot \mathbf{1}, 1 \right) \right. \\ &\quad \left. - \hat{f}_{\text{id}}^2 \left(1 - \sum_{k \in \{1, \dots, M^d\}} \hat{f}_{\text{ind}, (C_{k,1})^0_{1/M^{2p+2}}}(\mathbf{x}) \right) \right) \end{aligned}$$

approximieren.

Aus dieser Darstellung erhalten wir für das Netzwerk $\hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x})$ die Gewichtsschranken

$$\|\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}}\|_{\infty} = \max \left\{ 1, \|\mathbf{w}_{\hat{f}_1}\|_{\infty}, \|\mathbf{w}_{\hat{f}_2}\|_{\infty} \right\} \leq \max \{ 1, M^{2p+2}, M^{4p+4} \} = M^{4p+4}$$

und

$$\left\| \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq \max \left\{ \left\| \left(\mathbf{w}_{\hat{f}_1} \right)_{i,j>0}^{(0)} \right\|_{\infty}, \left\| \left(\mathbf{w}_{\hat{f}_2} \right)_{i,j>0}^{(0)} \right\|_{\infty} \right\} \leq 1.$$

Die Schranken der äußersten Gewichte folgen aus der Definition von $\hat{f}_{\text{check}, \mathcal{P}_2}$ und sind gegeben durch

$$\left\| \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{1,j>0}^{(5)} \right\|_{\infty} = 1 \quad \text{und} \quad \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{1,0}^{(5)} = 1.$$

Das Netzwerk besteht aus fünf verdeckten Schichten. In den ersten drei Schichten besitzt das Netz eine Anzahl von $M^d \cdot 2d + 2d + d \cdot M^d \cdot 2d$ Neuronen pro Schicht. Dies ergibt sich, da das Netz \hat{f}_{id}^2 eine Anzahl von $2d$ Neuronen pro Schicht aufweist. Das Netz $\hat{\phi}_{2,1}$ hat aufgrund seiner Konstruktion $d \cdot M^d \cdot 2d$ Neuronen pro Schicht. Zusätzlich besteht das Netz \hat{f}_1 aus einer Summe von M^d Netzwerken der Form $\hat{f}_{\text{ind},(C_{k,1})^0_{1/M^{2p+2}}}$, weshalb es $M^d + 2d$ Neuronen pro Schicht hat.

Die Neuronenanzahl für die letzten beiden Schichten des Netzwerks ergibt sich wie folgt: Das Identitätsnetz hat 2 Neuronen pro Schicht. Das Netzwerk \hat{f}_{test} , das $2 \cdot (2d + 2)$ Neuronen pro Schicht benötigt, bestimmt die Neuronenanzahl für \hat{f}_2 , welche $M^d \cdot 2 \cdot (2d + 2)$ Neuronen pro Schicht beträgt. Folglich hat das Netzwerk $\hat{f}_{\text{check},\mathcal{P}_2}$ in den letzten beiden Schichten insgesamt $2 + M^d \cdot 2 \cdot (2d + 2)$ Neuronen.

Damit ist das finale neuronale Netzwerk in der Klasse

$$\mathcal{F}(5, r)$$

mit

$$\begin{aligned} r &= \max\{M^d \cdot 2d + 2d + d \cdot M^d \cdot 2d, 2 + M^d \cdot 2 \cdot (2d + 2)\} \\ &\leq 2d + (4d^2 + 4d) \cdot M^d \end{aligned}$$

enthalten.

Im Folgenden zeigen wir, dass

$$\hat{f}_{\text{check},\mathcal{P}_2}(\mathbf{x}) = \mathbb{1}_{\bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})^0_{1/M^{2p+2}}}(\mathbf{x})$$

für $\mathbf{x} \notin \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})^0_{1/M^{2p+2}} \setminus (C_{k,2})^0_{2/M^{2p+2}}$ gilt.

Wir unterscheiden hier zwischen drei Fällen. Im ersten Fall nehmen wir an, dass

$$\mathbf{x} \notin \bigcup_{k \in \{1, \dots, M^d\}} (C_{k,1})^0_{1/M^{2p+2}},$$

ist, woraus

$$\mathbf{x} \notin \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})^0_{1/M^{2p+2}}$$

folgt.

Für diese \mathbf{x} ergibt sich aufgrund von Lemma 27a), dass $\hat{f}_{\text{ind},(C_{k,1})^0_{1/M^{2p+2}}}(\mathbf{x}) = 0$ ist, woraus $\hat{f}_1(\mathbf{x}) = 1$ folgt. Daraus können wir schließen, dass

$$\begin{aligned} &1 - \hat{f}_2(\mathbf{x}) - \hat{f}_{\text{id}}^2(\hat{f}_1(\mathbf{x})) \\ &= \sum_{j \in \{1, \dots, M^d\}} \hat{f}_{\text{test}} \left(\hat{f}_{\text{id}}^2(\mathbf{x}), \hat{\phi}_{2,1} + \mathbf{v}_j + \frac{1}{M^{2p+2}} \cdot \mathbf{1}, \hat{\phi}_{2,1} + \mathbf{v}_j + \frac{2a}{M^2} \cdot \mathbf{1} - \frac{1}{M^{2p+2}} \cdot \mathbf{1}, 1 \right) - 1 \\ &\leq 0 \end{aligned}$$

gilt. Wir haben hier verwendet, dass jedes Netzwerk $\widehat{f}_{\text{test}}$ in dem Intervall $[0, 1]$ liegt und höchstens ein $\widehat{f}_{\text{test}}$ in der Summe größer als 0 ist. Daher erhalten wir

$$\widehat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) = 1 - 0 = 1 = \mathbb{1}_{\bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0}(\mathbf{x}).$$

Im zweiten Fall, nehmen wir an, dass

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^d\}} (C_{k,1})_{1/M^{2p+2}}^0$$

sowie

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{2/M^{2p+2}}^0$$

gilt. Dann ist $\widehat{\phi}_{2,1} = (\mathbf{C}_{\mathcal{P}_1}(\mathbf{x}))_{\text{left}} = (\mathbf{C}_{i,1})_{\text{left}}$. Zudem können wir folgern, dass

$$\begin{aligned} (\mathcal{A}^{(j)})_{1/M^{2p+2}}^0 &= \left\{ \mathbf{x} \in \mathbb{R}^d : -\widehat{\phi}_{1,1}^{(k)} + \widehat{\phi}_{2,1}^{(k)} + \widetilde{v}_j^{(k)} + \frac{1}{M^{2p+2}} \leq 0 \right. \\ &\quad \left. \text{und } \widehat{\phi}_{1,1}^{(k)} - \widehat{\phi}_{2,1}^{(k)} - \widetilde{v}_j^{(k)} - \frac{2a}{M^2} + \frac{1}{M^{2p+2}} < 0 \text{ für alle } k \in \{1, \dots, d\} \right\} \\ &= (\widetilde{C}_{j,i})_{1/M^{2p+2}}^0 \end{aligned}$$

für $j \in \{1, \dots, M^d\}$ gilt. Da die Aussage, aufgrund der Voraussetzung für die zu zeigende Ungleichung, nur für

$$\mathbf{x} \notin \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{1/M^{2p+2}}^0 \setminus (C_{k,2})_{2/M^{2p+2}}^0,$$

gezeigt werden muss, können wir aus Lemma 27b) folgern, dass

$$\widehat{f}_{\text{test}}\left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1} + \mathbf{v}_j + \frac{1}{M^{2p+2}} \cdot \mathbf{1}, \widehat{\phi}_{2,1} + \mathbf{v}_j + \frac{2a}{M^2} \cdot \mathbf{1} - \frac{1}{M^{2p+2}} \cdot \mathbf{1}, 1\right) = \mathbb{1}_{(\widetilde{C}_{j,i})_{1/M^{2p+2}}^0}(\mathbf{x})$$

für alle $j \in \{1, \dots, M^d\}$ gilt. Dies impliziert

$$\widehat{f}_2(\mathbf{x}) = f_2(\mathbf{x}).$$

Wegen

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{2/M^{2p+2}}^0$$

können wir weiter folgern, dass

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^d\}} (C_{k,1})_{2/M^{2p+2}}^0.$$

Daher ergibt sich aus Lemma 27a) die Gleichheit

$$\widehat{f}_1(\mathbf{x}) = f_1(\mathbf{x}) = 0.$$

Somit ist

$$1 - \widehat{f}_2(\mathbf{x}) - \widehat{f}_{\text{id}}^2(\widehat{f}_1(\mathbf{x})) = 1 - f_2(\mathbf{x}) = 1 - 0 = 1$$

und

$$\widehat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) = 1 - 1 = 0 = \mathbb{1}_{\bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0}(\mathbf{x}).$$

Im dritten Fall nehmen wir an, dass

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^d\}} (C_{k,1})_{1/M^{2p+2}}^0$$

und gleichzeitig

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2}) \setminus (C_{k,2})_{1/M^{2p+2}}^0$$

ist, woraus

$$\mathbf{x} \notin \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{1/M^{2p+2}}^0$$

folgt. In diesem Fall ist die Approximation $\widehat{f}_1(\mathbf{x})$ nicht exakt. Gemäß Lemma 27 sind dann alle Werte von $\widehat{f}_{\text{ind}, (C_{k,1})_{1/M^{2p+2}}^0}^*$ für $k \in \{1, \dots, M^d\}$ in der Definition von \widehat{f}_1 in $[0, 1]$ enthalten. Damit ist

$$\widehat{f}_1(\mathbf{x}) \in [0, 1].$$

Da

$$\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^d\}} (C_{k,1})_{1/M^{2p+2}}^0 \tag{A.25}$$

gilt, ergibt sich, wie bereits im zweiten Schritt gezeigt wurde, dass

$$\widehat{f}_2(\mathbf{x}) = f_2(\mathbf{x}).$$

Zusammenfassend erhalten wir

$$1 - f_2(\mathbf{x}) - \widehat{f}_{\text{id}}^2(\widehat{f}_1(\mathbf{x})) = \sum_{j \in \{1, \dots, M^d\}} \mathbb{1}_{(\widetilde{C}_{j,i})_{1/M^{2p+2}}^0}(\mathbf{x}) - \widehat{f}_{\text{id}}^2(\widehat{f}_1(\mathbf{x})) \leq 0 - 0 = 0.$$

Daher ist

$$\widehat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) = 1 - 0 = 1 = \mathbb{1}_{\bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0}(\mathbf{x}).$$

Wegen der Konstruktion des Netzes gilt zusätzlich

$$\widehat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) \in [0, 1]$$

für $\mathbf{x} \in [-a, a]^d$. □

Die Kombination der Netzwerke $\hat{f}_{\mathcal{P}_2}$ aus Lemma 22, $\hat{f}_{w_{\mathcal{P}_2}}$ aus Lemma 29 sowie \hat{f}_{check} aus Lemma 30 führt schließlich zu dem Netzwerk, welches $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ in der Supremumsnorm approximiert.

Die Hauptidee ist es, ein Netzwerk

$$\hat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) = \sigma \left(\hat{f}_{\mathcal{P}_2}(\mathbf{x}) - B_{\text{true}} \cdot \hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) \right) - \sigma \left(-\hat{f}_{\mathcal{P}_2}(\mathbf{x}) - B_{\text{true}} \cdot \hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) \right)$$

zu definieren, welches gleich $\hat{f}_{\mathcal{P}_2}(\mathbf{x})$ sein wird, so lange $\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} (C_{k,2})_{2/M^{2p+2}}^0$ und den Wert 0 annimmt, wenn $\mathbf{x} \in \bigcup_{k \in \{1, \dots, M^{2d}\}} C_{k,2} \setminus (C_{k,2})_{1/M^{2p+2}}^0$ ist. Hierfür werden wir das Netzwerk aus Lemma 30 verwenden. Als Wert für B_{true} werden wir die Schranke von $\hat{f}_{\mathcal{P}_2}(\mathbf{x})$ aus Lemma 22 heranziehen. Zudem machen wir uns die Eigenschaft der ReLU-Aktivierungsfunktion zunutze, die im Fall einer negativen Eingabe den Wert 0 annimmt. Im Fall, dass \mathbf{x} nah am Rand einer der Würfel der feineren Partition liegt, nimmt das Netzwerk $\hat{f}_{\text{check}, \mathcal{P}_2}$ den Wert 1 an. Daraus folgt $\hat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) = 0$. Ansonsten nimmt $\hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x})$ den Wert 0 an und es gilt

$$\hat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) = \hat{f}_{\mathcal{P}_2}(\mathbf{x}).$$

Abschließend multiplizieren wir dann das Netzwerk $\hat{f}_{\mathcal{P}_2, \text{true}}$ mit dem Netzwerk $\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x})$ aus Lemma 29.

Beweis von Lemma 28. Sei $\hat{f}_{\mathcal{P}_2}$ das Netzwerk aus Lemma 22 sowie $\hat{f}_{\text{check}, \mathcal{P}_2}$ das Netzwerk aus Lemma 30. Durch sukzessives Anwenden von \hat{f}_{id} auf die Ausgabe eines dieser Netzwerke können wir erreichen, dass beide Netze die gleiche Anzahl an verdeckten Schichten

$$L = 4 + \max \left\{ \lceil \log_4(M^{2p}) \rceil \cdot \lceil \log_2(\max\{q+1, 2\}) \rceil, 1 \right\}$$

haben. Wir definieren dann ein Netzwerk $\hat{f}_{\mathcal{P}_2, \text{true}}$ durch

$$\begin{aligned} \hat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) &= \sigma \left(\hat{f}_{\mathcal{P}_2}(\mathbf{x}) - B_{\text{true}} \cdot \hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) \right) \\ &\quad - \sigma \left(-\hat{f}_{\mathcal{P}_2}(\mathbf{x}) - B_{\text{true}} \cdot \hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) \right) \end{aligned}$$

mit

$$B_{\text{true}} = 2 \cdot \exp(2ad) \cdot \max \left\{ \|f\|_{C^q([-a, a]^d)}, 1 \right\}.$$

Gemäß Lemma 22 und Lemma 30 erhalten wir die Gewichtsschranken

$$\|\mathbf{w}_{\hat{f}_{\mathcal{P}_2}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1, \quad \left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2}} \right)_{1,0}^{(L)} = 0$$

sowie

$$\|\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1, \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{1,j>1}^{(5)} \right\|_{\infty} = 1 \text{ und } \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{1,0}^{(5)} = 1.$$

Aufgrund der Definition des Netzes $\hat{f}_{\mathcal{P}_2, \text{true}}$ resultiert daraus die Schranke

$$\|\mathbf{w}_{\hat{f}_{\mathcal{P}_2, \text{true}}}\|_{\infty} = \max \left\{ 1, \|\mathbf{w}_{\hat{f}_{\mathcal{P}_2}}\|_{\infty}, B_{\text{true}} \cdot \left\| \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{1,j}^{(5)} \right\|_{\infty}, \|\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}}\|_{\infty} \right\}$$

$$\leq M^{4p+4},$$

wobei wir verwendet haben, dass $M^{2p} \geq 2 \cdot \exp(2ad) \cdot \max \left\{ \|f\|_{C^q([-a,a]^d)}, 1 \right\}$ ist. Zudem gelten

$$\left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2, \text{true}}} \right)_{1,0}^{(L)} = 0, \quad \left\| \left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2, \text{true}}} \right)_{1,j>0}^{(L)} \right\|_{\infty} \leq 1$$

und

$$\left\| \left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2, \text{true}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} = \max \left\{ \left\| \left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_{\infty}, \left\| \left(\mathbf{w}_{\hat{f}_{\text{check}, \mathcal{P}_2}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \right\} \leq 1.$$

Da das Netzwerk $\hat{f}_{\mathcal{P}_2}$ zur Netzwerkklassse $\mathcal{F}(L_{\hat{f}_{\mathcal{P}_2}}, r_{\hat{f}_{\mathcal{P}_2}})$ mit

$$(i) \quad L_{\hat{f}_{\mathcal{P}_2}} = 4 + \lceil \log_4(M^{2p}) \rceil \cdot \lceil \log_2(\max\{q+1, 2\}) \rceil$$

$$(ii) \quad r_{\hat{f}_{\mathcal{P}_2}} = \max \left\{ \left(\binom{d+q}{d} + d \right) \cdot M^d \cdot 2 \cdot (2+2d) + 2d, 18 \cdot (q+1) \cdot \binom{d+q}{d} \right\}$$

gehört und zusätzlich das Netzwerk $\hat{f}_{\text{check}, \mathcal{P}_2}$ in der Klasse $\mathcal{F}(5, 2d + (4d^2 + 4d) \cdot M^d)$ enthalten ist, folgt daraus, dass auch das zusammengesetzte Netzwerk $\hat{f}_{\mathcal{P}_2, \text{true}}$ in der Klasse $\mathcal{F}(L_{\hat{f}_{\mathcal{P}_2, \text{true}}}, r_{\hat{f}_{\mathcal{P}_2, \text{true}}})$ mit

$$L_{\hat{f}_{\mathcal{P}_2, \text{true}}} = 5 + \lceil \log_4(M^{2p}) \rceil \cdot \lceil \log_2(\max\{q+1, 2\}) \rceil$$

und

$$r_{\hat{f}_{\mathcal{P}_2, \text{true}}} = \max \left\{ \left(\binom{d+q}{d} + d \right) \cdot M^d \cdot 2 \cdot (2+2d) + 2d, 18 \cdot (q+1) \cdot \binom{d+q}{d} \right\} \\ + 2d + (4d^2 + 4d) \cdot M^d$$

liegt. Dies ergibt sich daraus, dass wir durch Anwenden von \hat{f}_{id} die Netzwerke auf die gleiche Anzahl von Schichten bringen können.

Da der Wert von $|\hat{f}_{\mathcal{P}_2}|$ gemäß Lemma 22 durch B_{true} beschränkt ist und $\hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) = 1$ gilt, wenn \mathbf{x} in der Menge

$$\bigcup_{\{i \in \{1, \dots, M^{2d}\}\}} C_{i,2} \setminus (C_{i,2})_{1/M^{2p+2}}^0 \tag{A.26}$$

liegt ist, folgt aus der Definition der ReLU-Aktivierungsfunktion, dass $\hat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x})$ den Wert 0 annimmt, sofern \mathbf{x} in der Menge (A.26) enthalten ist.

Sei $\hat{f}_{w_{\mathcal{P}_2}}$ das Netzwerk aus Lemma 29. Um $\hat{f}_{\mathcal{P}_2, \text{true}}$ mit $\hat{f}_{w_{\mathcal{P}_2}}$ zu multiplizieren, verwenden wir das Netzwerk

$$\hat{f}_{\text{mult}} \in \mathcal{F}(\lceil \log_4(M^{2p}) \rceil, 18)$$

aus Lemma 24, wobei wir $R = \lceil \log_4(M^{2p}) \rceil$ wählen. Dieses Netz erfüllt nach Lemma 24 die Ungleichung

$$\left| \hat{f}_{\text{mult}}(x, y) - x \cdot y \right| \leq 8 \cdot \left(\max \left\{ \|f\|_{\infty, [-a,a]^d}, 1 \right\} \right)^2 \cdot \frac{1}{M^{2p}} \tag{A.27}$$

für alle x, y in

$$\left[-2 \cdot \max \left\{ \|f\|_{\infty, [-a, a]^d}, 1 \right\}, 2 \cdot \max \left\{ \|f\|_{\infty, [-a, a]^d}, 1 \right\} \right].$$

Durch die wiederholte Anwendung von \hat{f}_{id} auf die Ausgabe der Netze $\hat{f}_{w_{\mathcal{P}_2}}$ und $\hat{f}_{\mathcal{P}_2, \text{true}}$ können wir deren Tiefe angleichen, so dass beide Netzwerke

$$L = 5 + \lceil \log_4(M^{2p}) \rceil \cdot (\lceil \log_2(\max\{q, d\}) + 1 \rceil)$$

verdeckte Schichten haben.

Das finale Netzwerk ist dann gegeben durch

$$\hat{f}(\mathbf{x}) = \hat{f}_{\text{mult}} \left(\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \hat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right).$$

Aus Lemma 30 wissen wir, dass $\hat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) \in [0, 1]$ für $\mathbf{x} \in [-a, a]^d$ ist. Daher erhalten wir aus Lemma 22, dass

$$|\hat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x})| \leq |\hat{f}_{\mathcal{P}_2}(\mathbf{x})| \leq 2 \cdot \exp(2ad) \cdot \max \left\{ \|f\|_{C^q([-a, a]^d)}, 1 \right\}$$

gilt. Zusätzlich folgt aus Lemma 29 die Beschränktheit $|\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x})| \leq 2$. Daher setzen wir $a = 2 \cdot \exp(2ad) \cdot \max\{\|f\|_{C^q([-a, a]^d)}, 1\}$ in Lemma 24.

Wegen

$$\left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2, \text{true}}} \right)_{1,0}^{(L)} = 0, \quad \left(\mathbf{w}_{\hat{f}_{w_{\mathcal{P}_2}}} \right)_{1,0}^{(L)} = 0, \quad \text{und} \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{mult}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1$$

können wir Lemma 17c) anwenden und erhalten zusammen mit den Gewichtsschranken aus Lemma 29, dass

$$\begin{aligned} \|\mathbf{w}_{\hat{f}}\|_{\infty} &\leq \max \left\{ \|\mathbf{w}_{\hat{f}_{\text{mult}}}\|_{\infty}, \|\mathbf{w}_{\hat{f}_{w_{\mathcal{P}_2}}}\|_{\infty}, \|\mathbf{w}_{\hat{f}_{\mathcal{P}_2, \text{true}}}\|_{\infty} \right\} \\ &\leq \max \left\{ 4 \cdot (2 \exp(2ad) \cdot \max\{\|f\|_{C^q([-a, a]^d)}, 1\})^2, M^{4p+4}, M^{4p+4} \right\} \\ &\leq M^{4p+4}. \end{aligned}$$

Darüber hinaus ergibt sich

$$\left\| \left(\mathbf{w}_{\hat{f}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq \max \left\{ \left\| \left(\mathbf{w}_{\hat{f}_{w_{\mathcal{P}_2}}} \right)_{i,j>0}^{(0)} \right\|_{\infty}, \left\| \left(\mathbf{w}_{\hat{f}_{\mathcal{P}_2, \text{true}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \right\} \leq 1$$

sowie aus Lemma 24, dass

$$\left(\mathbf{w}_{\hat{f}} \right)_{1,0}^{(L)} = 0$$

gilt.

Wie bereits oben gezeigt, hat das Netz $\widehat{f}_{\mathcal{P}_2, \text{true}}$ eine Anzahl von

$$r_{\widehat{f}_{\mathcal{P}_2, \text{true}}} = \max \left\{ \left(\binom{d+q}{d} + d \right) \cdot M^d \cdot 2 \cdot (2+2d) + 2d, 18 \cdot (q+1) \cdot \binom{d+q}{d} \right\} \\ + 2d + (4d^2 + 4d) \cdot M^d$$

Neuronen. Das Netz $\widehat{f}_{w_{\mathcal{P}_2}}$ aus Lemma 29 hat eine Anzahl von

$$r_{\widehat{f}_{w_{\mathcal{P}_2}}} = \max \left\{ 18d, 2d + d \cdot M^d \cdot 2 \cdot (2+2d) \right\}$$

Neuronen. Daraus ergibt sich, dass das Netzwerk \widehat{f} eine Anzahl von

$$r = \max \left\{ \left(\binom{d+q}{d} + d \right) \cdot M^d \cdot 2 \cdot (2+2d) + 2d, 18 \cdot (q+1) \cdot \binom{d+q}{d} \right\} \\ + 2d + (4d^2 + 4d) \cdot M^d + \max \left\{ 18d, 2d + d \cdot M^d \cdot 2 \cdot (2+2d) \right\} \\ \leq \binom{d+q}{d} \cdot M^d \cdot 18 \cdot d^2 \cdot (q+1) + 16 \cdot M^d \cdot d^2 + 18 \cdot M^d \cdot d^2 \\ \leq 64 \cdot \binom{d+q}{d} \cdot d^2 \cdot (q+1) \cdot M^d$$

Neuronen hat.

Somit liegt das Netzwerk \widehat{f} in der Klasse $\mathcal{F}(L, r)$ mit

$$L = 5 + \lceil \log_4(M^{2p}) \rceil \cdot (\lceil \log_2(\max\{q, d\} + 1) \rceil + 1)$$

und

$$r \leq 64 \cdot \binom{d+q}{d} \cdot d^2 \cdot (q+1) \cdot M^d.$$

In dem Fall, dass

$$\mathbf{x} \in \bigcup_{i \in \{1, \dots, M^{2d}\}} (C_{i,2})_{2/M^{2p+2}}^0,$$

ist der Wert von \mathbf{x} weder in

$$\bigcup_{i \in \{1, \dots, M^{2d}\}} C_{i,2} \setminus (C_{i,2})_{1/M^{2p+2}}^0 \tag{A.28}$$

noch in

$$\bigcup_{i \in \{1, \dots, M^{2d}\}} (C_{i,2})_{1/M^{2p+2}}^0 \setminus (C_{i,2})_{2/M^{2p+2}}^0 \tag{A.29}$$

enthalten.

Gemäß Lemma 29 approximiert das Netz $\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x})$ für $\mathbf{x} \in \bigcup_{i \in \{1, \dots, M^{2d}\}} (C_{i,2})_{2/M^{2p+2}}^0$ den Wert $w_{\mathcal{P}_2}(\mathbf{x})$ mit einem Fehler der Größe

$$4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}}. \quad (\text{A.30})$$

Zudem gilt $w_{\mathcal{P}_2}(\mathbf{x}) \in [0, 1]$ für alle $\mathbf{x} \in \mathbb{R}^d$.

Aus Lemma 22 folgt, dass das Netz $\widehat{f}_{\mathcal{P}_2}$ die Funktion f mit einem Fehler der Größe

$$c_{170} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)} \cdot \frac{1}{M^{2p}} \quad (\text{A.31})$$

approximiert.

Ist das Netz $\widehat{f}_{\text{check}, \mathcal{P}_2}(\mathbf{x}) = 0$, so gilt

$$\widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) = \sigma(\widehat{f}_{\mathcal{P}_2}(\mathbf{x})) - \sigma(-\widehat{f}_{\mathcal{P}_2}(\mathbf{x})) = \widehat{f}_{\mathcal{P}_2}(\mathbf{x}).$$

Für $M^{2p} \geq 4^{4d+1} \cdot d$ können wir nun den Wert von $\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x})$ beschränken, indem wir die Dreiecksungleichung anwenden. Damit erhalten wir

$$\left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \right| \leq \left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \right| + |w_{\mathcal{P}_2}(\mathbf{x})| \leq 4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}} + 1 \leq 2.$$

Darüber hinaus folgt aus Lemma 22 die Ungleichung

$$\begin{aligned} \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| &\leq \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - f(\mathbf{x}) \right| + |f(\mathbf{x})| \\ &\leq c_{170} \cdot \max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\}^{4 \cdot (q+1)} \cdot \frac{1}{M^{2p}} + \|f\|_{\infty, [-a,a]^d} \\ &\leq 2 \cdot \max \left\{ \|f\|_{\infty, [-a,a]^d}, 1 \right\}, \end{aligned}$$

wobei wir verwendet haben, dass $M^{2p} \geq c_{170} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4 \cdot (q+1)}$ gilt. Somit sind beide Netzwerke in dem Intervall enthalten, in dem Ungleichung (A.27) erfüllt ist.

Aus (A.27) und der Verwendung der Dreiecksungleichung ergibt sich

$$\begin{aligned} &\left| \widehat{f}_{\text{mult}} \left(\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\ &\leq \left| \widehat{f}_{\text{mult}} \left(\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right) - \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| + \left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| \\ &\quad + \left| w_{\mathcal{P}_2}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\ &\leq \left| \widehat{f}_{\text{mult}} \left(\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right) - \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| + \left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \right| \cdot \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| \\ &\quad + \left| \widehat{f}_{\mathcal{P}_2}(\mathbf{x}) - f(\mathbf{x}) \right| \cdot |w_{\mathcal{P}_2}(\mathbf{x})| \\ &\leq 8 \cdot \left(\max \left\{ \|f\|_{\infty, [-a,a]^d}, 1 \right\} \right)^2 \cdot \frac{1}{M^{2p}} + 4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}} \cdot 2 \cdot \max \left\{ \|f\|_{\infty, [-a,a]^d}, 1 \right\} \\ &\quad + c_{170} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4 \cdot (q+1)} \cdot \frac{1}{M^{2p}} \\ &\leq c_{176} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4 \cdot (q+1)} \cdot \frac{1}{M^{2p}}. \end{aligned}$$

Sofern \mathbf{x} jedoch in der Menge (A.28) liegt, ist der Approximationsfehler von $\hat{f}_{\mathcal{P}_2}$ nicht in der Größenordnung von $c_{173} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)} \cdot \frac{1}{M^{2p}}$. In diesem Fall nimmt $\hat{f}_{\text{check},\mathcal{P}_2}(\mathbf{x})$ den Wert 1 an, womit $\hat{f}_{\mathcal{P}_2,\text{true}} = 0$ ist. Des Weiteren ist nach Lemma 29

$$\left| \hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \right| \leq 2$$

für alle $\mathbf{x} \in [-a, a]^d$. Somit sind $\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x})$ und $\hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x})$ in dem Intervall enthalten, in dem (A.27) gilt. Aus der Tatsache, dass

$$0 \leq w_{\mathcal{P}_2}(\mathbf{x}) \leq \frac{1}{a \cdot M^{2p}},$$

was aus Ungleichung (A.21) folgt, sowie der Dreiecksungleichung ergibt sich

$$\begin{aligned} & \left| \hat{f}_{\text{mult}} \left(\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\ & \leq \left| \hat{f}_{\text{mult}} \left(\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right) - \hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right| + \left| \hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right| \\ & \quad + \left| w_{\mathcal{P}_2}(\mathbf{x}) \cdot \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\ & \leq \left| \hat{f}_{\text{mult}} \left(\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right) - \hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right| + \left| \hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot 0 - w_{\mathcal{P}_2}(\mathbf{x}) \cdot 0 \right| \\ & \quad + \left| w_{\mathcal{P}_2}(\mathbf{x}) \cdot 0 - w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\ & \leq 8 \cdot \left(\max \left\{ \|f\|_{\infty,[-a,a]^d}, 1 \right\} \right)^2 \cdot \frac{1}{M^{2p}} + 0 + |w_{\mathcal{P}_2}(\mathbf{x})| \cdot \|f\|_{\infty,[-a,a]^d} \\ & \leq c_{176} \cdot \left(\max \left\{ \|f\|_{\infty,[-a,a]^d}, 1 \right\} \right)^2 \cdot \frac{1}{M^{2p}}. \end{aligned}$$

In dem Fall, dass \mathbf{x} in (A.29), aber nicht in (A.28) enthalten ist, approximiert das Netzwerk $\hat{f}_{\mathcal{P}_2}$ die Funktion f mit einem Fehler wie in (A.31). Zudem liegt $\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x})$ im Intervall $[-2, 2]$ und approximiert $w_{\mathcal{P}_2}(\mathbf{x})$ mit dem gleichen Fehler wie in (A.30).

Da der Wert von $\hat{f}_{\text{check},\mathcal{P}_2}(\mathbf{x})$ im Intervall $[0, 1]$ ist, erhalten wir analog zu oben

$$\begin{aligned} \left| \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right| & \leq \left| \hat{f}_{\mathcal{P}_2}(\mathbf{x}) \right| \\ & \leq \left| \hat{f}_{\mathcal{P}_2}(\mathbf{x}) - f(\mathbf{x}) \right| + |f(\mathbf{x})| \\ & \leq c_{170} \cdot \max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\}^{4(q+1)} \cdot \frac{1}{M^{2p}} + \|f\|_{\infty,[-a,a]^d} \\ & \leq 2 \cdot \max \left\{ \|f\|_{\infty,[-a,a]^d}, 1 \right\}, \end{aligned}$$

wobei wir wieder verwendet haben, dass $M^{2p} \geq c_{170} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4(q+1)}$. Also sind $\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x})$ und $\hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x})$ in dem Intervall enthalten, in dem (A.27) gilt. Zusammen mit

$$0 \leq w_{\mathcal{P}_2}(\mathbf{x}) \leq \frac{2}{a \cdot M^{2p}},$$

was ebenfalls aus Ungleichung (A.21) folgt, und der Dreiecksungleichung ergibt sich

$$\left| \hat{f}_{\text{mult}} \left(\hat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \hat{f}_{\mathcal{P}_2,\text{true}}(\mathbf{x}) \right) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right|$$

$$\begin{aligned}
&\leq \left| \widehat{f}_{\text{mult}} \left(\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right) - \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right| + \left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right| \\
&\quad + \left| w_{\mathcal{P}_2}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right| + \left| w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\
&\leq \left| \widehat{f}_{\text{mult}} \left(\widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}), \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right) - \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right| + \left| \widehat{f}_{w_{\mathcal{P}_2}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \right| \cdot \left| \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right| \\
&\quad + \left| w_{\mathcal{P}_2}(\mathbf{x}) \cdot \widehat{f}_{\mathcal{P}_2, \text{true}}(\mathbf{x}) \right| + \left| w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\
&\leq 8 \cdot \left(\max \left\{ \|f\|_{\infty, [-a, a]^d}, 1 \right\} \right)^2 \cdot \frac{1}{M^{2p}} + 4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}} \cdot 2 \cdot \max \left\{ \|f\|_{\infty, [-a, a]^d}, 1 \right\} \\
&\quad + \frac{2}{a \cdot M^{2p}} \cdot 2 \cdot \max \left\{ \|f\|_{\infty, [-a, a]^d}, 1 \right\} + \frac{2}{a \cdot M^{2p}} \cdot 2 \cdot \max \left\{ \|f\|_{\infty, [-a, a]^d}, 1 \right\} \\
&\leq c_{176} \cdot \left(\max \left\{ \|f\|_{\infty, [-a, a]^d}, 1 \right\} \right)^2 \cdot \frac{1}{M^{2p}}.
\end{aligned}$$

Damit haben wir die Aussage von Lemma 28 bewiesen. \square

A.1.6 Schritt 4: Anwendung von \widehat{f} auf leicht verschobene Partitionen

Im letzten Schritt des Beweises werden wir das Netzwerk aus Lemma 28 auf 2^d verschobene Partitionen von \mathcal{P}_2 anwenden. Diese Partitionen entstehen, indem mindestens eine Komponente des Gitters um a/M^2 verschoben wird. Das finale Netzwerk wird durch eine Linearkombination der 2^d Netze aus Lemma 28 konstruiert. Die Gewichte in der Linearkombination werden so gewählt, dass sie einen kleinen Wert annehmen, wenn sich \mathbf{x} am Rand des zugehörigen Würfels befindet.

Beweis von Lemma 16. Im Folgenden genügt es zu zeigen, dass ein Netz \widehat{f} existiert, das die Ungleichung

$$\sup_{\mathbf{x} \in [-a/2, a/2]^d} \left| f(\mathbf{x}) - \widehat{f}_{\text{net}}(\mathbf{x}) \right| \leq c_{178} \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a, a]^d)} \right\} \right)^{4 \cdot (q+1)} \cdot \frac{1}{M^{2p}}$$

erfüllt. Da a bei Bedarf einfach vergrößert werden kann, ist dies ausreichend.

Seien \mathcal{P}_1 und \mathcal{P}_2 die Partitionen, welche wie in (A.1) definiert sind. Wir setzen im Folgenden

$$\mathcal{P}_{1,1} = \mathcal{P}_1 \text{ sowie } \mathcal{P}_{2,1} = \mathcal{P}_2$$

und definieren für jedes $v \in \{2, \dots, 2^d\}$ Partitionen $\mathcal{P}_{1,v}$ und $\mathcal{P}_{2,v}$, die Modifikationen von $\mathcal{P}_{1,1}$ und $\mathcal{P}_{2,1}$ sind, wobei mindestens eine der Komponenten um a/M^2 verschoben ist.

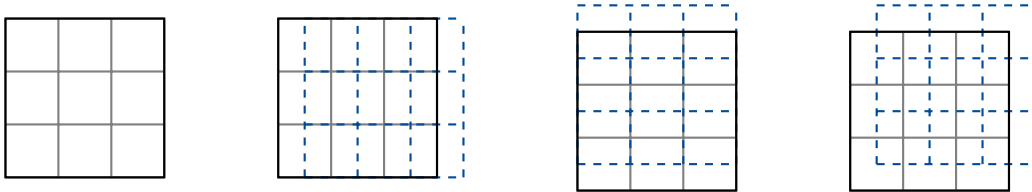


Abbildung A.2: 2^2 verschiedene Partitionen im Fall $d = 2$

Für den Fall $d = 2$ ist die Idee des Beweises in Abbildung A.2 dargestellt (vgl. Kohler und Langer (2021)). In der Abbildung kann man erkennen, dass eine Anzahl von $2^2 = 4$ verschiedenen Partitionen existiert,

sofern wir die Partition entlang mindestens einer Komponente um die gleiche Distanz verschieben. Wir bezeichnen mit $C_{k,2,v}$ für $k \in \{1, \dots, M^{2d}\}$ die zugehörigen Würfel der Partition $\mathcal{P}_{2,v}$.

Die Idee ist nun eine Linearkombination von Netzwerken $\hat{f}_{\mathcal{P}_{2,1}}, \dots, \hat{f}_{\mathcal{P}_{2,2^d}}$ aus Lemma 28 zu berechnen. Hierbei werden die Partitionen $\mathcal{P}_{2,v}$ wie in \mathcal{P}_2 in Lemma 28 behandelt.

Um zu vermeiden, dass der Approximationsfehler der Netzwerke nahe der Ränder eines Würfels der Partitionen zunimmt, multiplizieren wir jeden Wert von $\hat{f}_{\mathcal{P}_{2,v}}$ mit einem Gewicht

$$w_v(\mathbf{x}) = \prod_{j=1}^d \left(1 - \frac{M^2}{a} \cdot \left| (C_{\mathcal{P}_{2,v}}(\mathbf{x}))_{\text{left}}^{(j)} + \frac{a}{M^2} - x^{(j)} \right|_+ \right), \quad (\text{A.32})$$

welches, sobald sich \mathbf{x} am Rand befindet, einen kleinen Wert annimmt. Formal ausgedrückt handelt es sich bei $w_v(\mathbf{x})$ um einen linearen Tensorprodukt B-Spline. Dieser nimmt seinen Maximalwert im Zentrum von $C_{\mathcal{P}_{2,v}}(\mathbf{x})$ an, ist im Inneren von $C_{\mathcal{P}_{2,v}}(\mathbf{x})$ ungleich 0 und nimmt außerhalb von $C_{\mathcal{P}_{2,v}}(\mathbf{x})$ den Wert 0 an. Aus Lemma 15.2 in Györfi et al. (2002) folgt dann

$$w_1(\mathbf{x}) + \dots + w_{2^d}(\mathbf{x}) = 1$$

für $\mathbf{x} \in [-a/2, a/2]^d$, da $a/M^2 \leq a/2$ ist.

Seien $\hat{f}_1, \dots, \hat{f}_{2^d}$ die Netzwerke aus Lemma 28, die den Partitionen $\mathcal{P}_{1,v}$ und $\mathcal{P}_{2,v}$ für $v \in \{1, \dots, 2^d\}$ entsprechen. Da $[-a/2, a/2]^d \subset [-a + a/M^2, a]^d$, bildet jedes $\mathcal{P}_{1,v}$ und $\mathcal{P}_{2,v}$ eine Partition einer Menge, welche $[-a/2, a/2]^d$ enthält. Zudem gelten dann die Fehlerschranken von Lemma 28 für jedes Netzwerk \hat{f}_v in dem Intervall $[-a/2, a/2]^d$ sowie die entsprechenden Gewichtsschranken aus Lemma 28.

Wir setzen

$$\hat{f}_{\text{net}}(\mathbf{x}) = \sum_{v=1}^{2^d} \hat{f}_v(\mathbf{x}).$$

Unter Verwendung von Lemma 28 können wir erkennen, dass dieses Netzwerk in der Netzwerkklasse $\mathcal{F}(L, r)$ mit

$$L = 5 + \lceil \log_4(M^{2p}) \rceil \cdot (\lceil \log_2(\max\{q, d\} + 1) \rceil + 1)$$

und

$$r = 2^d \cdot 64 \cdot \binom{d+q}{d} \cdot d^2 \cdot (q+1) \cdot M^d$$

enthalten ist. Die Gewichtsschranken des Netzwerks ändern sich durch die Summation nicht und sind daher gemäß Lemma 28 gegeben durch

$$\|\mathbf{w}_{\hat{f}_{\text{net}}}\|_{\infty} \leq M^{4p+4}, \quad \left\| \left(\mathbf{w}_{\hat{f}_{\text{net}}} \right)_{i,j>0}^{(0)} \right\|_{\infty} \leq 1 \quad \text{und} \quad \left(\mathbf{w}_{\hat{f}_{\text{net}}} \right)_{1,0}^{(L)} = 0.$$

Da

$$f(\mathbf{x}) = \sum_{v=1}^{2^d} w_v(\mathbf{x}) \cdot \hat{f}_v(\mathbf{x})$$

ist, folgt unmittelbar aus Lemma 28, dass

$$\left| \hat{f}_{\text{net}}(\mathbf{x}) - f(\mathbf{x}) \right| = \left| \sum_{v=1}^{2^d} \hat{f}_{\text{mult}} \left(\hat{f}_{w_v}(\mathbf{x}), \hat{f}_{\mathcal{P}_{2,v}, \text{true}}(\mathbf{x}) \right) - \sum_{v=1}^{2^d} w_v(\mathbf{x}) \cdot \hat{f}_v(\mathbf{x}) \right|$$

$$\begin{aligned}
&\leq \sum_{v=1}^{2^d} \left| \widehat{f}_{\text{mult}} \left(\widehat{f}_{w_v}(\mathbf{x}), \widehat{f}_{\mathcal{P}_{2,v}, \text{true}}(\mathbf{x}) \right) - w_v(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\
&\leq c_{176} \cdot 2^d \cdot \left(\max \left\{ 2a, \|f\|_{C^q([-a,a]^d)} \right\} \right)^{4 \cdot (q+1)} \cdot \frac{1}{M^{2p}}
\end{aligned}$$

gilt. Damit ist die Aussage von Lemma 16 gezeigt. \square

A.2 Hilfsresultate

Hilfslemma 3. Sei $s \in \mathbb{N}$. Angenommen, $g_1, \dots, g_s : \mathbb{R} \rightarrow \mathbb{R}$ sind beschränkte und Lipschitz-stetige Funktionen mit den Lipschitz-Konstanten $C_{\text{Lip},g_1}, \dots, C_{\text{Lip},g_s}$. Dann sind sowohl die Summe $\sum_{j=1}^s g_j$ als auch das Produkt $\prod_{j=1}^s g_j$ Lipschitz-stetige Funktionen, deren Lipschitz-Konstanten durch

$$\sum_{j=1}^s C_{\text{Lip},g_j} \leq s \cdot \max_{j \in \{1, \dots, s\}} C_{\text{Lip},g_j}$$

beziehungsweise

$$\sum_{j=1}^s C_{\text{Lip},g_j} \cdot \prod_{k \in \{1, \dots, s\} \setminus \{j\}} \|g_k\|_{\infty} \leq s \cdot \max_{j \in \{1, \dots, s\}} C_{\text{Lip},g_j} \cdot \prod_{k \in \{1, \dots, s\} \setminus \{j\}} \|g_k\|_{\infty}$$

beschränkt sind.

Beweis. Wir beginnen damit nachzuweisen, dass die Lipschitz-Konstante einer Summe von Lipschitz-stetigen Funktionen durch die Summe der zugehörigen Lipschitz-Konstanten gegeben ist.

Es gilt

$$\begin{aligned}
\left| \sum_{j=1}^s g_j(z_1) - \sum_{j=1}^s g_j(z_2) \right| &\leq \sum_{j=1}^s |g_j(z_1) - g_j(z_2)| \\
&\leq \sum_{j=1}^s C_{\text{Lip},g_j} \cdot |z_1 - z_2|
\end{aligned}$$

für $z_1, z_2 \in \mathbb{R}$. Damit folgt, dass die Summe $\sum_{j=1}^s g_j$ Lipschitz-stetig ist mit einer Lipschitz-Konstanten, die durch

$$\sum_{j=1}^s C_{\text{Lip},g_j} \leq s \cdot \max_{j \in \{1, \dots, s\}} C_{\text{Lip},g_j}$$

beschränkt ist.

Im nächsten Schritt zeigen wir, dass das Produkt $\prod_{j=1}^s g_j$ ebenfalls Lipschitz-stetig ist.

Seien wieder $z_1, z_2 \in \mathbb{R}$, dann gilt für die Lipschitz-stetigen Funktionen g_1, \dots, g_s , dass

$$\left| \prod_{k=1}^s g_k(z_1) - \prod_{k=1}^s g_k(z_2) \right| \leq \sum_{j=1}^s \left| \prod_{k=1}^{j-1} g_k(z_1) \cdot (g_j(z_1) - g_j(z_2)) \cdot \prod_{k=j+1}^s g_k(z_2) \right|$$

$$\begin{aligned}
&\leq \sum_{j=1}^s \left(\prod_{k=1}^{j-1} \|g_k\|_\infty \cdot |g_j(z_1) - g_j(z_2)| \cdot \prod_{k=j+1}^s \|g_k\|_\infty \right) \\
&\leq \sum_{j=1}^s \left(\prod_{k=1}^{j-1} \|g_k\|_\infty \cdot C_{\text{Lip},g_j} \cdot |z_1 - z_2| \cdot \prod_{k=j+1}^s \|g_k\|_\infty \right) \\
&\leq \sum_{j=1}^s \left(\prod_{k \in \{1, \dots, s\} \setminus \{j\}} \|g_k\|_\infty \cdot C_{\text{Lip},g_j} \cdot |z_1 - z_2| \right) \\
&= \sum_{j=1}^s C_{\text{Lip},g_j} \cdot \prod_{k \in \{1, \dots, s\} \setminus \{j\}} \|g_k\|_\infty \cdot |z_1 - z_2| \\
&\leq s \cdot \max_{j \in \{1, \dots, s\}} C_{\text{Lip},g_j} \cdot \prod_{k \in \{1, \dots, s\} \setminus \{j\}} \|g_k\|_\infty \cdot |z_1 - z_2|.
\end{aligned}$$

Damit folgt die Aussage des Hilfslemmas. □

Lemma 31. Sei $\sigma(z) = \max\{z, 0\}$ die ReLU-Aktivierungsfunktion. Sei zudem \mathcal{F} die Menge aller vollständig verbundenen neuronalen Netze mit Tiefe $L \in \mathbb{N}$ und einer Anzahl von $W \in \mathbb{N}$ mit $W \geq 2$ Gewichten. Dann existiert eine Konstante $c_{179} > 0$, welche weder von L , W noch von der Anzahl der Neuronen des Netzes abhängt, so dass für die VC-Dimension

$$V_{\mathcal{F}_+} \leq c_{179} \cdot L \cdot W \cdot \log(W)$$

gilt.

Beweis. Siehe Theorem 7 in Bartlett et al. (2019). □

Abbildungsverzeichnis

1.1	Fundamentale Forschungsbereiche des Deep Learnings	2
1.2	Aktivierungsfunktionen	10
1.3	Graphische Darstellung eines Neurons	11
1.4	Vollständig verbundenes vorwärtsgerichtetes neuronales Netz	12
1.5	Klassische U-förmige Risikokurve	13
1.6	Double Descent Curve	14
2.1	Neuronales Netz, bestehend aus parallel berechneten, vollständig verbundenen Netzen . .	26
2.2	Verschiedene Verschiebungen des Gitters, um den Würfel $[-K, K]^d$ abzudecken	55
4.1	Hierarchisches Kompositionsmodell der Klasse $\mathcal{H}(2, \mathcal{P})$	114
4.2	Darstellung der Komposition eines neuronalen Netzes $f \circ g$	117
4.3	Darstellung des neuronalen Netzes t_1	119
5.1	Fundamentale Forschungsbereiche des Deep Learnings	134
A.1	Berechnung von $(\hat{\phi}_{1,1}, \hat{\phi}_{2,1}, \hat{\phi}_{3,1}^{(\xi_{\nu,1})}, \dots, \hat{\phi}_{3,1}^{(\xi_{\nu,M^d})})$	164
A.2	2^2 verschiedene Partitionen im Fall $d = 2$	186

Literaturverzeichnis

- [1] Allen-Zhu, Z., Li, Y., und Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, PMLR, Band 97, Seiten 242–252.
- [2] Anthony, M. und Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.
- [3] Arens, T., Hettlich, F., Karpfinger, C., Kockelkorn, U., Lichtenegger, K., und Stachel, H. (2022). *Mathematik*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- [4] Arora, S., Cohen, N., Hazan, E., und Hu, W. (2019). A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations (ICLR 2019)*. New Orleans, Louisiana.
- [5] Barron, A. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133.
- [6] Bartlett, P. L., Harvey, N., Liaw, C., und Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20:1–17.
- [7] Bauer, B. und Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47(4):2261–2285.
- [8] Belkin, M., Hsu, D., Ma, S., und Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116:15849–15854.
- [9] Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA.
- [10] Braun, A., Kohler, M., Langer, S., und Walk, H. (2024). Convergence rates for shallow neural networks learned by gradient descent. *Bernoulli*, 30(1):475–502.
- [11] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., und LeCun, Y. (2015). The loss surface of multilayer networks. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*, Band 38, Seiten 192–204. Proceedings of Machine Learning Research, San Diego, CA, USA.
- [12] Coşkun, M., Uçar, A., Yildirim, Ö., und Demir, Y. (2017). Face recognition based on convolutional neural network. In *Proceedings of the 2017 International Conference on Modern Electrical and Energy Systems (MEES)*. IEEE, Kremenchuk, Ukraine.
- [13] Cover, T. (1968). Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55.

-
- [14] Devlin, J., Chang, M.-W., Lee, K., und Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, Minneapolis, Minnesota.
- [15] Devroye, L., Györfi, L., und Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Band 31. New York, NY.
- [16] Drews, S. und Kohler, M. (2023). Analysis of the expected L_2 error of an over-parametrized deep neural network estimate learned by gradient descent without regularization. *Zur Veröffentlichung eingereicht*.
- [17] Drews, S. und Kohler, M. (2024). On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent. *Annals of the Institute of Statistical Mathematics*, 76:361–391.
- [18] Du, S. S. und Lee, J. D. (2018). On the power of over-parametrization in neural networks with quadratic activation. In *Proceedings of the 35th International Conference on Machine Learning (PMLR 2018)*, Band 80, Seiten 1329–1338. PMLR.
- [19] Du, S. S., Lee, J. D., Tian, Y., Póczos, B., und Singh, A. (2018). Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *Proceedings of the 35th International Conference on Machine Learning (ICML) 2018*, Band 3, Seiten 2142–2159. International Machine Learning Society.
- [20] Fan, J. (1996). *Local Polynomial Modelling and Its Applications*, Band 66 von *Monographs on Statistics and Applied Probability*. Routledge, New York.
- [21] Forster, O. (2023). *Analysis : 1. Differential- und Integralrechnung einer Veränderlichen*. Lehrbuch. Wiesbaden, 13. überarbeitete auflage Auflage.
- [22] Friedman, J. H. und Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823.
- [23] Golowich, N., Rakhlin, A., und Shamir, O. (2018). Size-independent sample complexity of neural networks. In S. Bubeck, V. Perchet, und P. Rigollet, Herausgeber, *Proceedings of the 31st Conference On Learning Theory*, Band 75 von *Proceedings of Machine Learning Research*, Seiten 297–299. PMLR.
- [24] Goodfellow, I., Bengio, Y., und Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA.
- [25] Grigorescu, S., Trasnea, B., Cocias, T., und Macesanu, G. (2019). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37.
- [26] Györfi, L., Kohler, M., Krzyzak, A., und Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.
- [27] Hernavs, J., Ficko, M., Klančnik, L., Rudolf, R., und Klančnik, S. (2018). Deep learning in industry 4.0 – brief overview. *Journal of Production Engineering*, 21(2):1–5.
- [28] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., und Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.
- [29] Hinton, G. E., Osindero, S., und Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

-
- [30] Hornik, K., Stinchcombe, M., und White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.
- [31] Horowitz, J. L. und Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Annals of Statistics*, 35(6):2589 – 2619.
- [32] Hunt, K. M. R. (2023). Could artificial intelligence win the next weather photographer of the year competition? *Weather*, 78(4):108–112.
- [33] Härdle, W., Hall, P., und Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, 21(1):157 – 178.
- [34] Härdle, W. und Simar, L. (2019). *Applied Multivariate Statistical Analysis*. Springer International Publishing, Cham.
- [35] Härdle, W. und Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995.
- [36] Härdle, W. und Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81(1):223–242.
- [37] Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6:100048.
- [38] Kawaguchi, K. (2016). Deep learning without poor local minima. *Advances in neural information processing systems*, 29.
- [39] Kawaguchi, K. und Huang, J. (2019). Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Seite 92–99. IEEE Press.
- [40] Kohler, M. (2014). Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *Journal of Multivariate Analysis*, 132:197–208.
- [41] Kohler, M. und Krzyżak, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *Journal of Nonparametric Statistics*, 17(8):891–913.
- [42] Kohler, M. und Krzyżak, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Transactions on Information Theory*, 63(3):1620–1630.
- [43] Kohler, M. und Krzyżak, A. (2021). Over-parametrized deep neural networks minimizing the empirical risk do not generalize well. *Bernoulli*, 27(4):2564–2597.
- [44] Kohler, M. und Krzyżak, A. (2022a). Over-parametrized neural networks learned by gradient descent can generalize especially well. Preprint.
- [45] Kohler, M. und Krzyżak, A. (2022b). Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent. ArXiv: 2210.01443, Preprint.
- [46] Kohler, M. und Krzyżak, A. (2023). On the rate of convergence of an over-parametrized deep neural network regression estimate with ReLU activation function learned by gradient descent. Preprint.
- [47] Kohler, M., Krzyżak, A., und Sänger, A. (2024). Learning of deep convolutional network image classifiers via stochastic gradient descent and over-parametrization. ArXiv: 2404.07128, Preprint.

-
- [48] Kohler, M. und Langer, S. (2021). On the rate of convergence of fully connected very deep neural network regression estimates. *Annals of Statistics*, 49:2231–2249.
- [49] Kong, E. und Xia, Y. (2007). Variable selection for the single-index model. *Biometrika*, 94(1):217–229.
- [50] Krizhevsky, A., Sutskever, I., und Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira et al. (Eds.), *Advances In Neural Information Processing Systems*, 25:1097–1105. Red Hook, NY: Curran.
- [51] Kukačka, J., Golkov, V., und Cremers, D. (2017). Regularization for deep learning: A taxonomy. ArXiv: 1710.10686, Preprint.
- [52] Kutyniok, G. (2022). The mathematics of artificial intelligence. ArXiv: 2203.08890, Preprint.
- [53] Lai, W., Kuang, M., Wang, X., Ghafariasl, P., Sabzalian, M. H., und Lee, S. (2023). Skin cancer diagnosis using artificial neural network and improved gray wolf optimization. *Scientific Reports*, 13(1):19377.
- [54] Li, G. und Ding, J. (2021). The rate of convergence of variation-constrained deep neural networks. ArXiv: 2106.12068, Preprint.
- [55] Liang, T., Rakhlin, A., und Sridharan, K. (2015). Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, Seiten 1260–1285. PMLR.
- [56] Lin, S. und Zhang, J. (2019). Generalization bounds for convolutional neural networks. ArXiv: 1910.01487, Preprint.
- [57] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., und Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.
- [58] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.
- [59] Nadaraya, É. A. (1970). Remarks on non-parametric estimates for density functions and regression curves. *Theory of Probability & Its Applications*, 15(1):134–137.
- [60] Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., und Srebro, N. (2019). The role of over-parametrization in generalization of neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.
- [61] Rawat, W. und Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449.
- [62] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.
- [63] Ruder, S. (2016). An overview of gradient descent optimization algorithms. ArXiv: 1609.04747, Preprint.
- [64] Rumelhart, D. E., Hinton, G. E., und Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.
- [65] Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):420.

-
- [66] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897.
- [67] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5(4):595–620.
- [68] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053.
- [69] Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705.
- [70] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics*, 22(1):118–171.
- [71] Suzuki, T. (2019). Adaptivity of Deep ReLU Network for Learning in Besov and Mixed Smooth Besov Spaces: Optimal Rate and Curse of Dimensionality. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA.
- [72] Suzuki, T. und Nitanda, A. (2021). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. ArXiv: 1910.12799, Preprint.
- [73] Tukey, J. W. (1947). Non-Parametric Estimation II. Statistically Equivalent Blocks and Tolerance Regions—The Continuous Case. *Annals of Mathematical Statistics*, 18(4):529 – 539.
- [74] Tukey, J. W. (1961). Curves as parameters, and touch estimation. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. University of California Press, Berkeley, CA, USA.
- [75] Ulbrich, M. und Ulbrich, S. (2012). *Nichtlineare Optimierung*. Springer Basel, Basel.
- [76] Van der Vaart, A. W. und Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, New York, NY.
- [77] Wahba, G. (1990). *Spline Models for Observational Data*, Band 59 von *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pennsylvania.
- [78] Wang, M. und Ma, C. (2023). Generalization error bounds for deep neural networks trained by SGD. ArXiv: 2206.03299, Preprint.
- [79] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372.
- [80] Yamashita, R., Nishio, M., Do, R. K. G., und Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629.
- [81] Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114.
- [82] Yehudai, G. und Shamir, O. (2019). On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, Band 32. Curran Associates, Inc.
- [83] Yu, Y. und Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054.

-
- [84] Zambom, A. Z. und Dias, R. (2013). A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1):20–42.
- [85] Zou, D., Cao, Y., Zhou, D., und Gu, Q. (2020). Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492.

Wissenschaftlicher Werdegang

- 10/2013 – 03/2019 **Technische Universität Darmstadt**
B.Sc. Mathematik mit Nebenfach Wirtschaft
Thesis: „Eine Einführung in die Lebensversicherungsmathematik“
- 04/2019 – 09/2021 **Technische Universität Darmstadt**
M.Sc. Mathematik mit Nebenfach Wirtschaft
Thesis: „Ein Beitrag zur statistischen Theorie der gefalteten neuronalen Netze“
- 10/2021 – 03/2025 **Technische Universität Darmstadt**
Promotion Mathematik
Wissenschaftliche Mitarbeiterin in der Arbeitsgruppe Stochastik