Fachbereich Mathematik und Informatik

Bachelorarbeit

Eine theoretische Analyse von komplexen neuronalen Netzen und parametrischen partiellen Differentialgleichungen

A Theoretical Analysis of Deep Neural Networks and Parametric PDEs

Erstgutachter:	Prof. Dr. Benedikt Wirth
Zweitgutachter:	Dr. Stephan Rave
vorgelegt von:	Niek Maurits Jung
Matrikelnummer:	449890
E-Mail:	n_jung03@uni-muenster.de
Studiengang:	Ein-Fach-Bachelor Mathematik
eingereicht in:	Münster, 24. Januar 2022

Inhaltsverzeichnis

1 Einleitung

Der Artikel „A Theoretical Analysis of Deep Neural Networks and Parametric PDEs“ beschäftigt sich damit, ob und inwiefern sogenannte „Deep Neural Networks“ (DNNs) dabei behilflich sind, Lösungen für parametrische Probleme zu finden. Benutzt werden dafür parametrische Abbildungen. Diese bezeichnen Beziehungen zwischen den einzelnen Elementen, die innerhalb eines Raumes existieren. Dort sind alle Parameter zu finden, welche für die partielle Differentialgleichung relevant sind und es werden den einzelnen ihre jeweiligen Gewichtungen zugeordnet. Mit diesen Abbildungen wird der parametrische Raum auf den Lösungsraum abgebildet [1]. Parametrische Probleme können beispielsweise dann auftreten, wenn es um das Modellieren von stetigen Wärme- und Massentransfers, Akustiken, Strömungsmechaniken oder Elektromagnetik geht [11]. Dabei kommt es häufig vor, dass die Parameter eine physische oder geometrische Einschränkung zu erfüllen haben wie beispielsweise Randbedingungen. Diese Voraussetzungen werden meist durch den Input der gleich eingeführten NNs charakterisiert.
Beim lösen einer parametrischen PDE tritt häufig wenigstens eins von zwei Problemen auf. Zum einen kann es sein, dass der Aufwand zum Berechnen der PDEs für jeden individuellen Punkt viel zu hoch ist und somit nicht effizient. Dies passiert in der sogenannten Online-Phase. Sie berechnet für jeden Parameter die „reduced basis“ Approximation für den gewünschten Output. Dementsprechend kann auch die Zeit ein Problem darstellen, wenn diese begrenzt ist. Um dieses Problem zu lösen, wird die Vermutung aufgestellt, dass die Lösungsmannigfaltigkeit niedrigdimensional ist und somit wenig Parameter enthält, die als Lösung zulässig sind. Weniger Rechnungen müssen durchgeführt werden. Somit lässt sich die Rechenzeit beschränken. Zusätzlich wird eine Offline-Phase ausgenutzt. In der kann angenommen werden, dass unendlich viel Rechenkapazität zur Verfügung steht und auch die Zeit keine Rolle spielt. Genannt wird dies auch die Trainingsphase der PDE. In dieser Phase werden dann für Probleme mit endlich vielen Elementen die „reduced basis“ Mengen konstruiert. Diese ermöglichen es, einen geeigneten Approximationsraum zu konstruieren. Woraufhin die Online-Phase eingeleitet wird. Trotz dessen, dass in dieser die Kapazitäten und Zeiten begrenzt sind, denn es wurde in der Offline-Phase so viel Aufwand betrieben, dass der übriggebliebene Aufwand nun für Echtzeitanwendung angemessen ist [3]. Daraus schließt man, dass mit der Offline-Phase so viele Mengen wie notwendig konstruiert werden können, um diese dann für die Berechnungen zu benutzen und umgehen damit die zwei genannten Probleme. Übergeordnet bezeichnet das die „reduced base me-

thod“ (RBM), diese spielt für diese Thematik eine wichtige Rolle. Snapshots aus einer parametrisch induzierten Mannigfaltigkeit stellen eine Galerkinprojektion auf einem kleindimensionalen Approximationsraum dar, dieser wiederum gilt als RB Diskretisierung [2]. Mit Snapshots ist in unserem Fall eine ganze Lösung einer PDE gemeint und nicht wie ursprünglich ein Punkt auf einer Datenmannigfaltigkeit oder auch ein Datum. Hierbei handelt es sich um die Lösung für einen bestimmten, zufällig gewählten Parametersatz.
Genauer gesagt, wird sich in dem oben erwähnten Artikel damit beschäftigt, inwiefern uns die Niedrigdimensionalität der Lösungsmannigfaltigkeit dabei helfen kann, mit Hilfe von DNNs eine Approximation für parametrische Abbildungen zu berechnen. Damit ist gemeint, wie viel effektiver es ist, zuvor die reduzierte Basis zu berechnen und diese dann zu benutzen, um die Niedrigdimensionalität zu modellieren. Das Hauptresultat befindet sich in Kapitel 4, für das in Kapitel 2 und 3 Argumente verfasst und später dann auch bewiesen werden. Diese werden später verwendet, um schlussendlich Theorem 4.3 zu zeigen. Auf die Ergebnisse werden wir Schritt für Schritt hinarbeiten und in Abschnitt 1.3.2 werden die Schritte noch einmal vereinfacht erklärt.

1.1 Zielsetzung

Diese Arbeit wird in zwei Abschnitte aufgeteilt. Der erste Teil besteht darin zu vermitteln, wie NNs funktionieren, welche Eigenschaften sie besitzen und welche Möglichkeiten sie bieten. Der zweite Teil beschäftigt sich damit, relevante Beweise zu erläutern.
Meine Zielsetzung ist es, den Originalartikel so wiederzugeben, dass dieser leichter verständlich ist. Dazu gehört, dass ich an gewissen Stellen zusätzliche Informationen mit einbringe. Diese sind entweder dazu da, um Begrifflichkeiten zu erläutern oder dienen dazu, Vorgehensweisen zu erklären und ihre Funktion verständlicher zu gestalten. Um dies zu gewährleisten habe ich eigens ausgedachte Anwendungsbeispiele mit eingebracht.
Im zweiten Teil ist es meine Aufgabe, die bereits im Originalartikel vorhandenen Beweise zu vereinfachen. Dort wurden von mir Zwischenschritte eingefügt, benutzte Eigenschaften genauer erläutert. Dadurch wurde die Übersicht für den Beweis verbessert und es lässt sich diesen besser folgen. Außerdem habe ich zusätzliche Beweise mit Hilfe von anderen Arbeiten eingefügt. Dadurch wurden schwierig zu beweisende Aussagen bewiesen oder Eigenschaften, welche benutzt wurden, um die ursprünglichen Beweise durchführen zu können. Schlussendlich dient dies dazu, dass meine Ergänzungen in dem Artikel dazu führen, dass meine Arbeit als Lernmaterial genutzt werden kann.

1.2 Was ist ein NN und wie ist es aufgebaut?

Künstliche neuronale Netze (KNN) sind Rechenmodelle, welche an biologische neuronale Netze angelehnt sind. Diese KNNs sind so aufgebaut, dass immer mindestens eine Input- und eine Outputschicht vorhanden ist. Im einfachsten Fall gibt es nur diese beiden Schichten, allerdings können auch sogenannte „hidden layers“ bestehen. Diese liegen dann zwischen den oben genannten. Je mehr Schichten, desto komplexer ist das Netzwerk. Auf jeder Schicht sind Gruppen von Neuronen vorhanden, welche alle miteinander verbunden sind. Am Ende dieses Unterabschnitts folgt eine Abbildung die zeigt, wie ein solches KNN aussehen könnte.
Jedes Neuron in diesem Netzwerk bekommt eine bestimmte Gewichtung zugeordnet, abhängig davon, wie wichtig sie sind. Vergleichbar ist dies mit Hilfe eines Bildes von einem Tier. Wenn sich zum Beispiel ein Bild eines Hundes angesehen wird, werden zuerst essentielle Merkmale wie die Schnauze, die Ohren und die Rute wahrgenommen. Weiter gedacht wären dann Merkmale für die jeweilige Rasse relevant. Der Fakt, dass das Tier in dem Bild vier Beine hat oder behaart ist, fallen bei der Bilderkennung weniger ins Gewicht, so wie mögliche Hintergrundobjekte. Dementsprechend hat jedes Neuron einen Input (der seine Merkmale bestimmt) und einen Output (in unserem Fall die Wirkung der Merkmale). Um die Aktivität eines einzelnen Neurons zu bestimmen, wird eine Aktivierungsfunktion verwendet. Hier muss lediglich darauf geachtet werden, dass diese nicht linear ist, da ein NN nicht dafür gedacht ist, lineare Probleme zu lösen. Sehr geeignet dafür ist die ReLU-Funktion. Diese ist wie folgt darzustellen: $f(x) = max\{0, x\}$. Eine ReLU-Funktion ist eine Aktivierungsfunktion eines künstlichen Neurons, die als Positivteil seines Arguments definiert ist. x ist in dem Fall der Input des Neurons.
Der Output jedes Neurons wird von Schicht zu Schicht an das nächste Neuron weitergegeben. Praktisch zum Nutzen ist hier die Identitätsfunktion, denn sie gibt direkt die Aktivität des einzelnen Neurons an, $Id(x) = x$. Wenn die Grafik als Beispiel genutzt wird, kann wie folgt die Aktivität berechnet werden: Angenommen, es gibt 2 Inputvektoren $v1 = 0,5$ und $v2 = 0,8$, die Propagierungsfunktion berechnet sich dann wie folgt:

$$0,3 \cdot 0,5 + 0,8 \cdot 0,7 = 0,71$$

Nun kann die ReLU Funktion berechnet werden: $f(x) = max\{0, 0.71\} = 0.71$ und dementsprechend ergibt sich $Id(0.71) = 0.71$ als Ausgabefunktion. Bei diesem Vorgehen ist es tatsächlich der Fall, dass die Gewichtungen nach jedem Durchgang automa-

tisch angepasst werden, um dem gewünschten Ergebnis immer näher zu kommen und somit den Fehler zu minimieren [18].

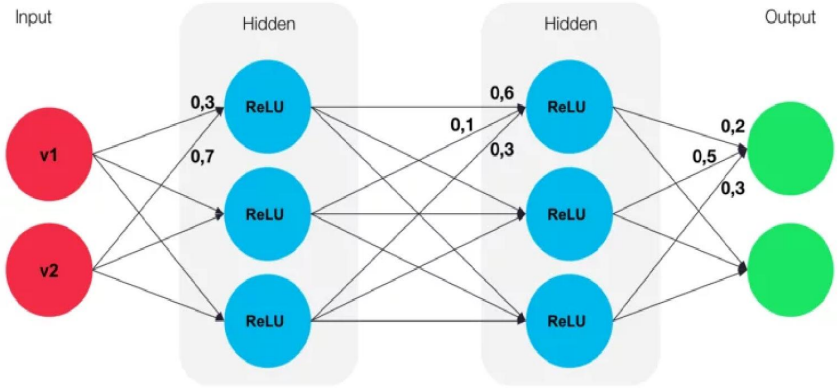

Abbildung 1: Beispielhafte Darstellung eines neuronalen Netzes mit gewichteten Verbindungen

1.3 Statistische Lernprobleme

Die Näherungsfähigkeit parametrischer Abbildungen durch DNNs zu studieren, ist durch die Ähnlichkeiten parametrischer Probleme mit statistischen Lernproblemen motiviert. Statistisches Lernen befasst sich unter anderem damit, Funktionen oder Funktionswerte basierend auf zuvor festgelegten Inputdaten zu finden. Es gibt verschiedene Arten des Lernens, relevant ist hier allerdings nur das beaufsichtigte Lernen (auch wenn das unbeaufsichtigte Lernen erwähnt und teilweise auch als Argument genutzt wird). Die entstehenden Probleme beim überwachten Lernen lassen sich aber noch in zwei Arten aufteilen, der Regression und der Klassifizierung [30]. Wir befassen uns nur mit der Klassifizierung (wie beim Beispiel des Bildes mit dem Hund). Wie folgend genauer beschrieben, gibt es immer eine Menge an Informationen als sogenannter Input zur Verfügung, durch den ein entsprechender Output entsteht. Nun soll eine Funktion gefunden werden, welche anhand des Inputs den Output mit bestimmter Genauigkeit vorhersagen kann. Nun zur mathematischen Beschreibung. Es existiert eine Definitionsmenge $X \subset \mathbb{R}^n, n \in \mathbb{N}$ und ein Label Set $Y \subset \mathbb{R}^k$, $k \in \mathbb{N}$. Das Label Set ist der Bildbereich einer Funktion, welcher mit dem neuronalen Netzwerk gelernt werden soll.

Weiter existiert eine Wahrscheinlichkeitsverteilung ρ auf $X\mathtt{x}\,Y$. Zusätzlich gibt es noch eine Verlustfunktion $L : Y\mathtt{x}Y \to R^+$. Die die Abweichung der exakten Funktion darstellt. Das Ziel ist es nun, eine Funktion zu finden, die in der Menge $H \subset \{h : X \to Y\}$ enthalten ist und dadurch den erwarteten Verlust $\mathbb{E}_{(x,y)\sim\rho}L(f(x), y)$ auf ein Minimum reduziert, sie dient als eine Art vorhersagende Funktion [24]. Angenommen, ein Arzt oder eine Ärztin möchte herausfinden, anhand welcher Symptome am ehesten gewisse Krankheiten diagnostiziert werden können, ohne einen genauen Test durchzuführen. Die Funktionen der Menge $H \subset \{h : X \to Y\}$ stellen in diesem Beispiel die Krankheiten dar. Für jedes einzeln betrachtete Symptom wird davon ausgegangen, dass der Patient eine bestimmte Krankheit hat. Diese Annahmen spiegeln die Funktionen wieder. Sie sagen voraus, wie das Endergebnis aussieht. Wenn diese Annahmen für jedes Symptom getroffen wurden, werden Tests durchgeführt, um die tatsächlichen Krankheiten der Patienten zu erfassen. Je nach Resultat werden die Symptome als Indiz für gewisse Krankheiten betrachtet. So lässt sich besser sagen, wie bei welchen Symptomen vorgegangen werden soll. Der zu trainierende Prozess nimmt sich Funktionen aus H und berechnet, bei welcher Funktion der erwartete Verlust minimiert wird. H besteht also aus allen zulässigen Funktionen. Da nun aber ρ unbekannt ist, müssen wir zunächst Paare von der Form $(x_i, y_i)_{i=1}^N N \in \mathbb{N}$ suchen, diese dienen uns als Trainingsdaten für den gewünschten Lerneffekt der Funktion und sind unabhängig von ρ gezogen worden. Diese Trainingsdaten sind in unserem Beispiel die Symptome und die resultierenden Krankheiten. Nun wird versucht, mit einem dieser Paare f zu finden, damit der empirische Verlust über H minimiert wird. Als Lernprozess verstehen wir die Optimierung des empirischen Verlusts.

$$\sum_{i=1}^{N} L(f(x_i), y_i). \tag{1.1}$$

Trotz dessen, dass die Konstruktion einer Funktion, die die relevanten partiellen Differentialgleichungen so genau wie möglich löst, mit hohem Rechenaufwand verbunden sein kann, sollte ihre Auswertung effizient sein. Die Intention ist es, mit so wenig Aufwand wie möglich, eine Funktion aus der Parametermenge auf einen Zustandsraum abzubilden.

Es wird angenommen, dass aufgrund dessen, dass Snapshots die Daten sind, die gebraucht werden, um den Prozess zu trainieren, und die Offline-Phase den Lernprozess imitiert, die Metrik auf dem Zustandsraum und die PDE, der Verlustfunktion und der Verteilung von ρ entsprechen. Das führt wiederum zur ursprünglichen Behauptung

zurück, dass Probleme parametrischer PDEs sehr denen statistischer Lernprobleme ähneln. Das Ergebnis ist, dass die parametrische Abbildung auf den Output des laufenden Prozesses für die zu dem Zeitpunkt genutzten Parameter verweist.
Es gibt mehrere Arten von Lernmethoden. Eine der effektivsten, mit der sich beschäftigt wird, ist das „deep learning“. Das „deep learning“ bezeichnet eine Art des Lernens, welches künstliche neuronale Netze verwendet, um Informationen zu verarbeiten. Diese sind wie in der Abbildung 1 aufgebaut und dienen dazu, auch komplexe Informationen verarbeiten zu können. Diese Methode beschreibt eine Reihe von Lernprozessen, welche zum Lösen statistischer Lernprobleme benutzt wird, in welchen H eine Menge von DNNs ist [21]. Diese Methode ist, vorausgesetzt die Aufgabe ist komplex genug, jeder anderen Methode überlegen, vor allem, wenn es um Spracherkennung und Bildklassifizierung geht. Zu beobachten ist, dass das Training eines KNNs rechnerisch sehr aufwendig ist. Im Vergleich dazu ist die Benutzung nach dem Training deutlich schneller als mit anderen Methoden, und direkt ist die erwähnte Unterscheidung der Offline-Online Prozesse sichtbar.
Gemessen an dem großen Erfolg dieser Techniken und den herrschenden Ähnlichkeiten statistischer Lernprobleme und parametrischer Probleme, ist es fast selbstverständlich, „deep learning“ Methoden auf statistische Lernprobleme anzuwenden, indem teilweise die parameterabhängigen Abbildungen durch DNNs ersetzt werden [22][25].

1.3.1 Approximation theoretischer Ergebnisse

Das Training der Netzwerke mit rektifizierenden Aktivierungsfunktionen ist wesentlich effizienter als mit zuvor bekannten Aktivierungsfunktionen. Eine Einheit, die diese Funktionen benutzt, nennt sich „rectified linear unit“. Diese Funktionen sind effizienter, da ihre Vorgänger wie der Sigmoid-Aktivierungsfunktion oder die hyperbolischen Tangentenfunktion, aufgrund der verschwindenden Gradienten in den einzelnen Schichten des NNs, nicht verwendet werden können. Durch die ReLU-Aktivierungsfunktion hingegen, werden diese Probleme überwunden und es ergibt sich ein Modell mit höherer Leistung und der Lernprozess schreitet schneller voran [26]. Nun wird versucht, eine Variation der parametrischen Abbildung

$$\mathcal{Y} \ni y \to u_y \in \mathcal{H},$$

zu lernen, mit $\mathcal{Y}$ als Parameterraum und $\mathcal{H}$ als Hilbertraum. Betrachtet wird ein kompakter Parameterraum, welcher ein Unterraum von $\mathbb{R}^p$ ist, und, dass $p \in \mathbb{N}$ endlich, aber vielleicht sehr groß ist. Daraus erschließt sich, dass hier eine endliche Anzahl von

Parametervektoren, die nicht auf die 0 abbilden, betrachtet wird. Wird davon ausgegangen, dass es eine Basis für eine Diskretisierung von hoher Genauigkeit von $\mathcal{H}$ gibt, welche sehr groß werden kann, dann sei $\mathbf{u}_y$ der Koeffizientenvektor von u_y bezüglich der High-Fidelity Diskretisierung (Diskretisierung hoher Genauigkeit. Mit Diskretisierung wird ein Vorgang bezeichnet, mit dem eine Teilmenge aus einer Menge herausgefiltert wird. In unserem Fall ist $\mathcal{H}$ die ursprüngliche Menge, aus der mit der Diskretisierung eine Teilmenge erstellt wird). Weiter wird angenommen, es gäbe eine RB die u_y hinreichend präzise approximiert $\forall y \in \mathcal{Y}$. Theorem 4.3 besagt, dass es mithilfe einiger technischer Annahmen ein DNN gibt, welches folgende diskrete Lösungsabbildung

$$\mathcal{Y} \ni y \to \mathbf{u}_y$$

bis zu einem einheitlichen Fehler $\epsilon > 0$, mit einer Größe polylogarithmisch in ϵ, kubisch in der Größe der reduzierten Basis und höchstens linear in der Größe der High-Fidelity-Basis approximiert [25]. Dadurch wird direkt sichtbar, dass DNNs tatsächlich hauptsächlich abhängig von der Größe der reduzierten Basis sind. Diese ist in der Regel immer geringer als die vorherigen, aus denen diese entstanden ist. Das Impliziert, dass unsere Abbildung schneller und mit weniger Rechenaufwand approximiert werden kann. Und somit ist klar, dass die DNNs direkt von reduzierten Basen profitieren können, wenn diese existieren. Dies wird im Laufe dieser Arbeit noch genauer erklärt und im Anschluss auch bewiesen. Zu erst werden in Unterabschnitt 1.3.2 die Schritte vorgestellt, bevor im Rest der Arbeit ins Detail gegangen wird.
Da es zu umfangreich wäre zu analysieren, wie der Prozess des Lernens abläuft, wird dieser hier nicht eingehend thematisiert.

1.3.2 Vereinfachte Präsentation der Argumente

„Jetzt versuchen wir, in vereinfachter Form die Argumente vorzustellen, die zu der in Unterabschnitt 1.3.1 beschriebenen Approximation führen. In diesem Aufbau stellen wir uns ein neuronales ReLU-Netz (ReLU NN) als eine Funktion

$$\mathbb{R}^n \to \mathbb{R}^k, \mathbf{x} \to T_L \varrho(T_{L-1} \varrho(...\varrho(T_1(x)))) \tag{1.2}$$

vor. Seien $L \in \mathbb{N}, T_1, ..., T_L$ affine Abbildungen und $\varrho : \mathbb{R} \to \mathbb{R}, \varrho(x) := max\{0, x\}$ ist die ReLU Aktivierungsfunktion, welche Koordinatenbasiert auf (1.2) angewandt wurde. L sei die Anzahl der Schichten innerhalb eines DNNs. Da T_l affin lineare Abbildungen sind, nehme ich an, dass für alle $\mathbf{x} \in dim T_l$ das $T_l(x) = A_l(x) + b_l$ für eine

Matrix A_l und einen Vektor b_l gilt. Die Größe des NNs definiert man als die Anzahl der Nicht-Null-Einträge aller A_l und b_l für $l \in \{1, ..., L\}$. Diese Definition wird später in Definition 3.1 genauer erklärt und erweitert werden. "[25, 1.2.2]

1. Im ersten Schritt wird die Konstruktion eines skalaren Multiplikationsoperators durch ReLU NNs aus [4] berücksichtigt. Für diese werden die beiden folgenden Beobachtungen benötigt:
Zuerst wird die Dreiecksfunktion $g : [0, 1] \to [0, 1], g(x) := min\{2x, 2 - 2x\}$ definiert. Um es genauer zu sagen, ergeben Kompositionen von g eine Sägezahnfunktion, also viele Dreiecksfunktionen mit unterschiedlichen Mittelpunkten. Für $s \in \mathbb{N}$ definiere $g_1 := g$ und $g_{s+1} := g \circ g_s$. In ([4], Proposition 2) wird folgendes gezeigt,

$$x^2 = lim_{n\to\infty} f_n(x) := lim_{n\to\infty} x - \sum_{s=1}^{n} \frac{g_s(x)}{2^{2s}}, \forall x \in [0, 1]. \tag{1.3}$$

Darüber hinaus kann g geschrieben werden als $g(x) = 2\varrho(x) - 4\varrho(x - \frac{1}{2}) + 2\varrho(x - 2)$. Das ermöglicht es g_s exakt durch ein ReLU NN darzustellen. Sei nun g_s durch die 1 beschränkt, zu sehen ist, dass f_n exponentiell gegen die Quadratfunktion konvergiert für $n \to \infty$. Außerdem kann f_n exakt als ein ReLU NN implementiert werden, wenn man die vorangegangenen Argumente nutzt [25]. Denn wie später noch gezeigt wird, lassen sich die f_n mit Hilfe von Linearkombinationen aus Teilen von g_s darstellen. Eine Approximation der quadratischen Funktion durch ein oder mehrere ReLU NNs entsteht durch eine approximative Realisierung der skalaren Multiplikation durch ReLu NNs. Ersichtlich wird dies durch die Identität des Parallelogramms $xz = \frac{1}{4}((x+z)^2-(x-z)^2)$ für $x, z \in \mathbb{R}$.
Intuitiv ist durch die exponentielle Konvergenz in (1.3) zu sehen, dass die Größe des NN, welches die skalare Multiplikation auf $[-1, 1]^2$ bis zu einem Fehler $\epsilon > 0$ approximiert, $\mathcal{O}(\log_2(\frac{1}{\epsilon}))$ ist. Beweis folgt in Kapitel 6.

2. Danach kann die Approximation der skalaren Multiplikation benutzt werden, um mithilfe von ReLU NNs eine Approximation für einen Multiplikationsoperator für Matrizen zu bekommen. Wie in [23] gezeigt wird, kann eine Matrixmultiplikation der Größenordnung $d\mathtt{x}d$ mit weniger als d^3 Operationen berechnet werden. Hier wird der Fall betrachtet, indem sich eine Multiplikation zweier Matrizen, dessen Einträge durch

die 1 beschränkt sind, durch NNs der Größe $\mathcal{O}(d^3 \log_2(\frac{1}{\epsilon}))$ mit einer Genauigkeit von $\epsilon > 0$ durchführen lässt. Genauer wird darauf in Proposition 3.7 eingegangen. Dort werden mit Hilfe von 3.6 und der später definierten Parallelisierung von KNNs Eigenschaften bezüglich der Schichten und der Gewichte eines NNs gezeigt. Auf die gleiche Weise kann gezeigt werden, wie ein NN konstruiert wird, das Matrix-Vektor-Multiplikationen imitiert.

3. Da ein NN ein Tupel aus vielen Matrix-Vektor-Multiplikationen bestehend ist, ist es möglich, durch die Verkettung dieser Matrixmultiplikationen, ganz einfach Matrixpolynome zu implementieren. Genauer gesagt, für $\mathbf{A} \in \mathbb{R}^{d \mathbf{x} d}$, so dass $\|\mathbf{A}\|_2 \leq 1 - \delta$ für manche $\delta \in (0, 1)$, kann die Abbildung $\mathbf{A} \to \sum_{s=0}^{m} \mathbf{A}^s$ annäherungsweise approximiert werden durch ein ReLU NN mit einer Genauigkeit von $\epsilon > 0$ und einer Größe von $\mathcal{O}(m \log_2^2(m) d^3 \cdot (\log(\frac{1}{\epsilon}) + \log_2(m))$, wobei der zusätzliche $\log_2$-Term in m dadurch entsteht, dass jede Approximation der Summe mit einer Genauigkeit von $\frac{\epsilon}{m}$ ausgeführt wird. Bekannt ist, dass die Neumann Reihe $\sum_{s=0}^{m} \mathbf{A}^s$ exponentiell gegen $(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A})^{-1}$ konvergiert für $m \to \infty$. Unter geeigneten Bedingungen ist es daher möglich, ein NN, ϕ_ϵ^{inv} für die Matrix $\mathbf{A}$ zu konstruieren, welche den Inverseoperator approximiert, also die Abbildung $\mathbf{A} \to \mathbf{A}^{-1}$, mit einer Genauigkeit von $\epsilon > 0$. Für eine Konstante $q > 0$ hat das NN eine Größe von $\mathcal{O}(d^3 \log_2^q(\frac{1}{\epsilon}))$ [25]. Dies zeigt Theorem 3.8 genauer, indem zu erst angenommen wird das die Norm jeder Matrix durch $\frac{1}{2}$ beschränkt ist und durch Anpassen der entsprechenden Konstante und der Behauptungen angenommen wird, dass die Normen der Matrizen durch ein beliebiges $Z > 0$ beschränkt werden können.

4. Wenn ein Lineares Gleichungssystem (LGS) existiert, kann durch die Existenz von ϕ_ϵ^{inv} und durch das Imitieren von approximativen Matrix-Vektor-Multiplikationen angenommen werden, dass ein NN existiert, dass eben genau dieses LGS approximativ löst. Als nächstes werden zwei Annahmen getroffen, die in vielen Anwendungen erfüllt sind:

- „Die Abbildung der Parameter auf die zugehörigen Steifigkeitsmatrizen der Galerkin - Diskretisierung der parametrischen PDE in Bezug auf eine reduzierte Basis kann durch NNs gut approximiert werden“ [25, 1.2.2 (4)].
- „Die Abbildung von den Parametern auf die rechte Seite der parametrischen

PDEs, die gemäß der reduzierten Basis diskretisiert sind, kann durch NNs gut approximiert werden“ [25, 1.2.2 (4)].

Mit diesen Annahmen, der Existenz von ϕ_ϵ^{inv} und einem ReLU NN, welches eine Matrix-Vektor-Multiplikation imitiert, ist es nicht schwer zu erkennen, dass es ein NN gibt, das näherungsweise die Abbildung von einem Parameter auf die zugehörige diskretisierte Lösung mit Bezug auf die reduzierte Basis implementiert. Angenommen, die reduzierte Basis hat die Größe d und die Implementierungen der Abbildung, die die Steifigkeitsmatrix und die rechte Seite ergeben, seien ausreichend effizient, dann hat mit der Konstruktion von ϕ_ϵ^{inv} das daraus entstehende NN die Größe $\mathcal{O}(d^3 \log_2^q(\frac{1}{\epsilon}))$. Definiere dieses NN mit ϕ_ϵ^{rb} [25].

5. Für den letzten Schritt ist das zuvor bereits konstruiertes ϕ_ϵ^{rb} wichtig, um das Ergebnis aus 1.3.1 abzuschließen. Begonnen wird mit der Größe der HFD. Wenn diese groß genug ist, angenommen D, kann approximativ jedes Element aus der reduzierten Basis mithilfe der HFD dargestellt werden. Impliziert wird dadurch, dass anstatt einer Approximation ein Basenwechsel vorgenommen werden kann, indem eine lineare Abbildung $\mathbf{V} \in \mathbb{R}^{D \mathtt{x} d}$ auf einen Vektor der zuvor erzeugten reduzierten Basis angewendet wird. Die erste Aussage von Unterabschnitt 1.3.1 folgt nun direkt aus der Betrachtung des NN $\mathbf{V} \circ \phi_\epsilon^{rb}$. Während des Prozesses wächst die Größe des NN auf $\mathcal{O}(d^3 \log_2^q(\frac{1}{\epsilon})) + dD)$. Der vollständige Beweis wird im Beweis von Theorem 4.3 vorgestellt.

1.4 Mögliche Auswirkungen und Erweiterungen

Die Verfasser hoffen, dass die Ergebnisse dieses Artikels das Potenzial besitzen, die Forschung über NNs und parametrische Probleme auf folgende Weise zu beeinflussen:

- *Theoretische Grundlage:* Das Endergebnis des Artikels ist so zu verstehen, dass, wenn es möglich ist, ein NN richtig zu trainieren, also die Approximation exakt genug werden zu lassen, diese genau so effizient wie RBMs bei der Lösung parametrischer PDEs sind, wenn sie annehmen, dass die Komplexität der NNs in Form ihrer freien Parameter gemessen wird. „Die Reduzierte-Basis-Methode ermöglicht eine effektive Modellreduktion bei der numerischen Lösung von parametrisierten partiellen Differentialgleichungen. Hierbei stellen die approximativen Lösungen der Differentialgleichung mit hoher Genauigkeit (z.B. die Lösung

einer hochdimensionalen Finite-Elemente-Methode) für eine kleine Anzahl an Parameterwerten eine reduzierte Basis des Lösungsraums auf. Mit Hilfe dieser reduzierten Basis können Näherungslösungen für weitere Parameter sehr kostengünstig berechnet werden. Entscheidener Bestandteil der reduzierten Basis Methode ist eine Abschätzung für den auftretenden Fehler, die ebenfalls mit geringem Rechenaufwand bestimmt werden kann“[28]. Durch die Anwendung des „Deep Learnings“ in Kombination mit der Approximationstheorie für parametrische PDE-Probleme könnte es möglich sein, noch komplexere Probleme anzugehen bzw. bestehende Lösungen noch exakter zu berechnen mit weniger hohem Aufwand, um diese wiederum in anderen Bereichen anzuwenden. Dies könnte dazu führen, dass neue Ergebnisse erzielt werden, dessen Berechnung zuvor zu aufwendig und dadurch nicht realisierbar waren.

- *Die Rolle der Umgebungsdimension verstehen:* Durch NNs ist es möglich, Approximationsraten zu erhalten, die sich im Vergleich zu anderen Methoden bei zunehmender Dimension weniger stark verschlechtern [5].
  Dafür wird als entscheidende Größe die Dimension der Lösungsmannigfaltigkeit betrachtet. Sie lässt direkt auf die Approximationsraten schließen, die NNs erreichen können, wenn es um parametrische Probleme geht. Denn umso weniger Parameter zur Verfügung stehen, desto genauer werden die Approximationen. Da für jede Rechenoperation ein Fehler entsteht, welche sich summieren. Diese Rate gilt es herauszufinden und nach Möglichkeit so klein wie möglich zu gestalten, damit die Methode effizient gestaltet werden kann. Der Zusammenhang zwischen den Approximationsraten den die NNs erreichen und der Umgebungsdimension wird detaillierter in Abschnitt 5 besprochen.

- *Indentifizierung geeigneter Architekturen:* Es lässt sich nicht genau sagen, wie das perfekte neuronale Netz für das jeweilige problem aussieht. Also die Anzahl der Schichten oder auch die Anzahl der Gewichte innerhalb des Netzes, können zuvor nicht genau bestimmt werden. Es wird nur klar, dass Netze, welche die entsprechenden Strukturen haben, sehr gute Approximationen ausgebe können, es wird aber nicht deutlich, wie groß diese Netze mindestens sein müssen.

Im Originalartikel ist es den Verfassern gelungen, einen Schritt auf dem Weg zu einer Theorie der auf „Deep Learning“ basierenden Lösungen für parametrische Probleme zu machen. Jedoch muss aufgrund der Komplexität dieses Feldes noch viel folgen, um es abzuschließen. Nachfolgend sind einige Fragen, die sich bei diesem Thema auftun:

- *Generelle parametrische Probleme:* Es werden lediglich koerzive, symmetrische und lineare parametrische Probleme angesprochen und es werden Fälle betrachtet, in denen endlich viele Parameter zur Verfügung stehen, alles andere wäre nicht dem Ziel dieser Arbeit dienlich. Wie sehen diese Anwendungen aber aus, wenn die Fälle nicht auf diese Eigenschaften beschränkt wären? Wenn das Interesse besteht, sich noch tiefer mit dieser Thematik zu beschäftigen, können in [7] und [8] noch weitere Fälle betrachtet werden.

- *Begrenzung der Anzahl von Snapshots:* Der bereits erwähnte empirische Verlust liegt sehr nahe am erwarteten Verlust. Das liegt daran, dass durch die Interpretation des parametrischen Problems als statistisches Lernproblem verschiedene Techniken angewendet werden können, wodurch die Anzahl der Stichproben N begrenzt wird. Es kann also angenommen werden, dass der Fehler beim Minimieren während des Lernprozesses so klein ist, dass unsere Funktion f, welche als Vorhersage dient, sehr gut funktioniert. Hier wird der Fehler in einer Norm gemessen, die durch die Verlustfunktion und die zugrunde liegende Wahrscheinlichkeitsverteilung induziert wird. Mit diesen Techniken ist es möglich, die Anzahl der für die Offline-Phase erforderlichen Snapshots zu begrenzen, um eine vermeintlich bessere Genauigkeit in der Online-Phase zu erreichen [25].

- *Notwendige Eigenschaften neuronaler Netzwerke:* Es ist schwierig zu sagen, wie die ideale Architektur eines NNs aussehen soll, also wie viele Schichten, wie viele Parameter sinnvoll oder auch wie viele Neuronen auf den einzelnen Schichten optimal sind. Man spricht hier wie bereits erwähnt lediglich über Approximationen und kann nicht exakt sagen, wie genau unsere Ergebnisse sind. Allerdings ist schon herausgefunden worden, dass NNs jede stetige Funktion approximieren können. Voraussetzung dafür ist, dass diese mindestens eine „hidden layer“ besitzen und auf einem beschränktem Raum operiert wird. Wenn diese dann auch noch genügend Neuronen besitzen, können diese Funktionen mit belibiger Genauigkeit approximiert werden. Trotzdem gilt auch hier die Aussage, dass dies keinen Aufschluss über die Größe des benötigten NNs gibt.

- *Allgemeine Matrixpolynome:* Oben wurde schon erwähnt, dass näherungsweise Implementierungen von Matrix-Polynomen für die Ergebnisse genutzt werden. Natürlich kann diese Konstruktion verwendet werden, um ein ReLU NN basierten Funktionskalkulus zu definieren und zu konstruieren. Mit anderen Worten, für jedes $d \in \mathbb{N}$ und jede stetige Funktion f, die durch Polynome gut approxi-

miert werden kann, können wir ein ReLU NN konstruieren, welches die Abbildung $\mathbf{A} \mapsto f(\mathbf{A})$ für jede entsprechend beschränkte Matrix $\mathbf{A}$ approximiert.
„Ein spezielles Beispiel für eine solche Funktion ist gegeben durch $f(\mathbf{A}) := e^{t\mathbf{A}}, t > 0$. Sie ist analytisch und spielt eine wichtige Rolle bei der Behandlung von Anfangswertproblemen“[25, Kapitel 1.3].

1.5 Vorgehen

„In Abschnitt 2 wird die Art der parametrischen PDEs, die wir in dieser Arbeit betrachten beschrieben, und es wird an die Theorie der RBs erinnert. In Abschnitt 3 wird ein NN-Kalkül eingeführt, das die Grundlage für alle Konstruktionen in dieser Arbeit bildet. Dort konstruiert man die NNs, die eine Matrix auf ihre ungefähre Inverse in Theorem 3.8 abbilden. In Abschnitt 4 werden NNs konstruiert, die parametrische Abbildungen approximieren. Zunächst approximiert man in Theorem 4.1 die parametrischen Abbildungen nach einer HFD. Diese Eigenschaften gelten in Anwendungsbeispielen oft als Voraussetzung.
In Abschnitt 5 schließen wir mit einer Diskussion unserer Ergebnisse im Hinblick auf die Abhängigkeit der zugrundeliegenden NN-Komplexität in Bezug auf die herrschenden Größen diese Arbeit ab. Alle Beweise werden erst in Kapitel 6 gezeigt, um den Lesefluss so gut wie möglich zu erhalten“[25, Kapitel 1.5].

1.6 Notation

Sei $\mathbb{N} = \{1, 2, ...\}$ die Menge der natürlichen Zahlen und definiere $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Für $a \in \mathbb{R}$ setzen wir $\lfloor a \rfloor := max\{b \in \mathbb{Z} : b \leq a\}$ und $\lceil a \rceil := min\{b \in \mathbb{Z} : b \geq a\}$. Sei $l, n \in \mathbb{N}$. Sei $\mathbf{Id}_{\mathbb{R}^n}$ die Identität und $\mathbf{0}_{\mathbb{R}^n}$ der Nullvektor auf $\mathbb{R}^n$. Für $\mathbf{A} \in \mathbb{R}^{n \times l}$ bezeichnen wir $\mathbf{A}^T$ als transponierte, $\sigma(\mathbf{A})$ das Spektrum von $\mathbf{A}$, mit $\|\mathbf{A}\|_2$ die Spektralnorm und mit $\|\mathbf{A}\|_0 := \sharp\{(i,j) : \mathbf{A}_{i,j} \neq 0\}$, wo $\sharp V$ die Kardinalität einer Menge V, die Anzahl der nicht Nulleinträge von $\mathbf{A}$, angibt. Außerdem für $\mathbf{v} \in \mathbb{R}^n$ beschreiben wir mit $|\mathbf{v}|$ die euklidische Norm. Sei V ein Vektorraum, dann ist $X \subset^s V$, falls X eine lineare Teilmenge von V ist. Darüber hinaus, falls $(V, \|\cdot\|_V)$ ein normierter Vektorraum ist, X ist eine Teilmenge von V und $v \in V$ dann beschreibt $dist(v, X) := inf\{\|x - v\|_V : x \in X\}$ den Abstand zwischen v, X und dem topologischen Dualraum, also die Menge aller skalarwertigen, stetigen, linearen Funktionen, ausgestattet mit der Operatornorm durch $(V^*, \|\cdot\|_{V^*})$. Für eine kompakte Menge $\Omega \subset \mathbb{R}^n$ beschreiben wir mit $C^r(\Omega), r \in \mathbb{N}_0 \cup \{\infty\}$, den Raum der r mal stetig differenzierbaren Funktionen, mit $L^p(\Omega, \mathbb{R}^n), p \in [1, \infty]$ den $\mathbb{R}^n$- wertigen Lebesgue Raum und setzen $L^p(\Omega, \mathbb{R})$ und

mit $H^1(\Omega) := W^{1,2}(\Omega)$ den Sobolev Raum erster Ordnung.

1.7 Mögliche Nutzung zukünftiger Resultate / Ausblick

In Abschnitt 1.3 wird darüber gesprochen, inwiefern das Deep Learning anderen Methoden gegenüber von Vorteil ist, wenn es unter anderem um Bildklassifizierung geht. Da die ganze Zeit davon gesprochen wird, dass NNs dazu in der Lage sind, komplexe Informationen zu verarbeiten, kann ich mir gut vorstellen, dass es in der Zukunft möglich wäre, diese Methoden in der Medizin anzuwenden. Gerade wenn es sich um bildgebende Diagnoseverfahren handelt wie die Magnetresonanztherapie (MRT), Röntgen oder auch Ultraschall. Vorstellen würde ich mir das so, dass die angesprochenen Parameter diese sind, die in den erstellten Bildern keine Abweichung von der zuvor definierten Norm beschreiben. Sofern diese nicht mehr erfüllt werden können, gerade weil zu große Abweichen bestehen, wird dies vom Algorithmus bemerkt. Ein einfaches Beispiel wäre ein Riss in einem Knochen bei einem Bruch oder, weiter gedacht, die Verfärbung einer Bandscheibe durch Überbelastung. In diesem Fall wäre es, sofern sich diese Methoden dahingehend implementieren lassen, von Vorteil, dass genau diese Bilder gesondert gekennzeichnet werden und somit auch kleinere unauffällige Verletzungen schneller erkannt und behandelt werden können. Im besten Fall werden die anormalen Stellen in den Bildern markiert.
Bezogen auf die erwähnte Spracherkennung in Abschnitt 1.3 könnte man dies gegebenenfalls beim Ultraschall anwenden, besonders für schwangere Frauen, da die Ärzte sich dort auf die Bildgebung, aber auch auf die Akustik verlässt, gerade beim Herzschlag. Wenn es anormale Geräusche gibt, welche im Augenblick nur schwer zu vernehmen sind, könnten sie durch diese Methoden auch diese wahrgenommen werden und so ggf. schlimmeres verhindert.

2 Parametrische PDEs und RBM

In diesem Abschnitt werden die Art parametrischer Probleme vorgestellt, die hier untersucht werden. Es wird davon ausgegangen, dass es ein *Parameterraum* $\mathcal{Y}$ und einen Lösungsraum $\mathcal{Z}$ gibt. Die Probleme basieren meist auf einer Abbildung $\mathcal{P} : \mathcal{Y} \to \mathcal{Z}$. Gesucht wird ein $\mathcal{P}(y) \in \mathcal{Z}$. Nun werden mit $\mathcal{Y}$ ausgewählte Parameter beschrieben und mit $\mathcal{Z}$ ein Funktionenraum oder ein Diskretisierungstheorem. Wenn nun angenommen wird, dass es sich um parametrische PDEs handelt und man die PDE mit dem Parameter y löst, gibt dies das gesuchte $\mathcal{P}(y) \in \mathcal{Z}$.
In Abschnitt 2.1 werden mehrere Annahmen zu den PDEs, die $\mathcal{P}$ zugrunde liegen, und zu den Parameterräumen $\mathcal{Y}$ getroffen. Anschließend wird in Abschnitt 2.2 ein abstrakter Überblick über Galerkin-Methoden gegeben, bevor in Abschnitt 2.3 einige grundlegende Fakten über RBs rekapituliert werden.

2.1 Parametrische partielle Differentialgleichung

„Im folgenden berücksichtigen wir parameterabhängige Gleichungen, welche in variierender Form gegeben sind durch:

$$b_y(u_y, v) = f_y(v), \quad \forall\, y \in \mathcal{Y},\ v \in \mathcal{H} \tag{2.1}$$

mit
(i) $\mathcal{Y}$ ist die Parametermenge, die klären wir direkt hinterher genauer,
(ii) $\mathcal{H}$ ist ein Hilbertraum,
(iii) $b_y : \mathcal{H}\mathrm{x}\mathcal{H} \to \mathbb{R}$ ist eine stetige Bilinearform welche die Bedingungen aus Annahme 2.1 erfüllt,
(iv) $f_y \in \mathcal{H}^*$ ist die parameterabhängige rechte Seite aus (2.1),
(v) $u_y \in \mathcal{H}$ ist die Lösung von (2.1). “[25, Kapitel 2.1]
Annahme 2.1. Folgende Annahmen werden im Rest der Arbeit für Gleichung (2.1) verwendet.

- **Parametermenge** $\mathcal{Y}$: Angenommen, $\mathcal{Y}$ sei eine kompakte Teilmenge von $\mathbb{R}^p$ mit $p \in \mathbb{N}$ fest, aber potentiell groß.
  Bemerkung. Weiter wird angenommen, dass $\mathcal{Y}$ kompakt und ein Unterraum eines Banachraumes V ist, dann ist wie in [12, Abschnitt 1.2] gezeigt, dass jeder Parameter aus $\mathcal{Y}$ durch eine Folge reeller Zahlen dargestellt werden kann.

Genauer gesagt, lässt sich ein $(v_i)_{i=0}^{\infty} \subset V$ finden, das, wenn eine Folge von Koeffizienten $\mathbf{c}_y$ gegeben ist, welche jeweils betragsmäßig durch die 1 beschränkt sind und zusätzlich $y \in \mathcal{Y}$ für alle y gilt, so dass angenommen werden kann, das $y = v_0 + \sum_{i=1}^{\infty} (\mathbf{c}_y)_i v_i$ gilt. Daraus ergibt sich, dass $\mathcal{Y}$ komplett durch Reihen von $\mathbf{c}_y$ darzustellen ist. Angenommen unser Raum V besitzt eine Schauderbasis. Diese wird in Form einer Folge $(b_n)_{n\in\mathbb{N}}$ repräsentiert, falls jeder Vektor ξ_n bezüglich dieser eine eindeutige Darstellung als konvergente Reihe $y = \sum_{n=1}^{\infty} \xi_n b_n$ besitzt. Daraus kann gefolgert werden, da $\mathcal{Y} \subset V$, dass jedes Element aus $\mathcal{Y}$ mit Hilfe einer Reihe reeller Zahlen angegeben werden kann [29]. Wenn nicht anders angegeben, gilt, dass diese Reihen $\mathbf{c}_y$ endlich mit festem, aber möglichst großem Support sind.

- **Symmetrie, einheitliche Stetigkeit und Koerzivität der Bilinearformen:** „Wir nehmen an, dass für alle $y \in \mathcal{Y}$ die Bilinearformen b_y symmetrisch sind, also:

$$b_y(u, v) = b_y(v, u), \quad \forall\ u, v \in \mathcal{H}.$$

Außerdem nehmen wir an, dass die Bilinearformen einheitlich stetig sind im Sinne, dass es eine Konstante $C_{cont} > 0$ gibt, so dass:

$$|b_y(u, v)| \leq C_{cont} \|u\|_{\mathcal{H}} \|v\|_{\mathcal{H}}, \quad \forall\ u \in \mathcal{H}, v \in \mathcal{H}, y \in \mathcal{Y}.$$

Schlussendlich nehmen wir an, dass die benutzten Bilinearformen koerziv in dem Sinne sind, dass eine Konstante $C_{coer} > 0$ existiert, so dass

$$\inf_{u\in\mathcal{H}\ ohne\ 0} \frac{b_y(u, u)}{\|u\|_{\mathcal{H}^2}} \geq C_{coer}, \quad \forall\ u \in \mathcal{H}, y \in \mathcal{Y}.$$

Somit sei mit dem Lax-Milgram Lemma (2.1) wohldefiniert, d.h. für jedes $y \in \mathcal{Y}$ und jedes $f_y \in \mathcal{H}^*$ existiert exakt ein $u_y \in \mathcal{H}$, das (2.1) löst und u_y ist stetig abhängig von f_y“[25, Annahme 2.1].

- **Beschränkt durch rechte Seite:** Angenommen, es existiert eine Konstante $C_{rhs} > 0$, so dass:

$$\|f_y\|_{\mathcal{H}^*} \leq C_{rhs}, \quad \forall\ y \in \mathcal{Y}.$$

- **Kompaktheit der Lösungsmannigfaltigkeit:** Weiter gilt, dass die Lösungs-

mannigfaltigkeit

$$S(\mathcal{Y}) := \{u_y : u_y \ \ ist \ \ die \ \ Loesung \ \ von \ \ (2.1), y \in \mathcal{Y}\}$$

kompakt in $\mathcal{H}$ ist.

Bemerkung. „Die Annahme $S(\mathcal{Y})$ sei kompakt, folgt direkt aus der Stetigkeit der Lösungsabbildung $y \mapsto u_y$. Die Bedingungen stimmen, falls $\forall \ \ u, v \in \mathcal{H}$ die Abbildungen $y \mapsto b_y(u, v)$ sowie $y \mapsto f_y(v)$ Lipschitzstetig sind [13, Proposition 5.1, Korollar 5.1] (Beweis in Kapitel 6)"[25, Kapitel 2.1].

2.2 High-Fidelity Approximation

Wie bereits beim Erklären der Offline oder auch Trainingsphase erwähnt, wird in dieser die Lösung nicht für jeden einzelnen Parameter ausgerechnet, sondern diese nur immer weiter approximiert. Nun kann angenommen werden, dass ein fester Parameter $y \in \mathcal{Y}$ genommen und dann für (2.1) die Galerkin-Methode verwendet wird. Im folgenden wird sich an [13, Kapitel 2.4] orientiert, um diese genauer zu erklären. Die Vorgehensweise ist es, nicht die exakte Lösung für (2.1) zu finden, sondern ein diskretes Schema der Form

$$b_y(u_y^{disc}, v) = f_y(v) \quad \forall \ v \in U^{disc}, \tag{2.2}$$

(jeder Punkt in U^{disc} hat eine Umgebung, in der kein anderer Punkt liegt) zu lösen, wo $U^{disc} \subset^s \mathcal{H}$ mit $dim(U^{disc}) < \infty$ und $u_y^{disc} \in U^{disc}$ die Lösung von (2.2) sei und auch Galerkin Approximation von u genannt wird. Für die Lösung von (2.2) haben wir

$$\left\|u_y^{disc}\right\|_{\mathcal{H}} \leq \frac{1}{C_{coer}} \left\|f_y\right\|_{\mathcal{H}^*}.$$

Nun sei u_y^{disc} die beste Annäherung an die Lösung u_y von (2.1) bis zu einer gewissen Konstante. Um es präziser auszudrücken, wird Cea's lemma [13, Lemma 2.2] verwendet,

$$\left\|u_y - u_y^{disc}\right\|_{\mathcal{H}} \leq \frac{C_{cont}}{C_{coer}} \inf_{\omega \in U^{disc}} \left\|u_y - \omega\right\|_{\mathcal{H}}. \tag{2.3}$$

U^{disc} sei nun gegeben. Falls nun $N := dim(U^{disc})$, sei $(\varphi_i)_{i=1}^N$ die Basis von U^{disc}. Dann existiert eine nicht singuläre (besitzt inverse) und positiv definite Matrix der

Form

$$\mathbf{B}_y := (b_y(\varphi_j, \varphi_i))_{i,j=1}^N.$$

Die Lösung von (2.2) erfüllt die Gleichung

$$u_y^{disc} = \sum_{i=1}^{N} (u_y)_i \varphi_i,$$

mit

$$\mathbf{u}_y := (\mathbf{B}_y)^{-1} \mathbf{f}_y \in \mathbb{R}^N$$

und $\mathbf{f}_y := (f_y(\varphi_i))_{i=1}^N \in \mathbb{R}^N$. „Üblicherweise beginnt man mit einer hochgradigen Diskretisierung des Raums $\mathcal{H}$, d. h. man wählt einen endlichen, aber potentiell hochdimensionalen Unterraum, für den die berechneten diskretisierten Lösungen ausreichend genau für jedes $y \in \mathcal{Y}$ sind“[25, Kapitel 2.2]. Um noch genauer zu sein wird folgendes angenommen:

Annahme 2.2. „Angenommen es existiert ein endlich dimensionaler Raum $U^h \subset^s \mathcal{H}$ mit endlicher Dimension D und Basis $(\varphi_i)_{i=1}^D$. Dieser Raum nennt sich HFD. Definiere mit $\mathbf{B}_y^h := (b_y(\varphi_j, \varphi_i))_{i,j=1}^D \in \mathbb{R}^{D \mathtt{x} D}$ die Steifigkeitsmatrix der HFD für $y \in \mathcal{Y}$, mit $\mathbf{f}_y^h := (f_y(\varphi_i))_{i=1}^D$ die diskretisierte rechte Seite und mit $\mathbf{u}_y := (\mathbf{B}_y)^{-1}\mathbf{f}_y \in \mathbb{R}^D$ den Koeffizientenvektor der Galerkinlösung mit Bezug zur HFD. Darüber hinaus beschreibt $u_y^h := \sum_{i=1}^D (\mathbf{u}_y^h)_i \varphi_i$ die Lösung für Galerkin. Angenommen für jedes $y \in \mathcal{Y}$ gilt $\sup_{y \in \mathcal{Y}} \left\| u_y - u_y^h \right\|_{\mathcal{H}} \leq \hat{\epsilon}$ für ein beliebig kleines, aber festes $\hat{\epsilon} > 0$. Im Folgenden wird nicht zwischen $\mathcal{H}$ und U^h unterschieden, es sei denn, die Unterscheidung ist notwendig“[25, Annahme 2.2].

Folgendes Problem kann bei diesem Ansatz auftreten. Und zwar das des zu hohen Rechenaufwands. Um $u_y^h \approx u_y$ zu berechnen, werden verschiedene Parameter gebraucht, um die beste Approximation zu erhalten. Das bedeutet, dass unser Raum U^h potentiell hohe Dimensionen annimmt, denn dieser enthält die relevanten Parameter. Nun ist die Aufgabe der RBM dafür zu sorgen, dass möglichst wenig Basisvektoren $(\varphi_i)_{i=1}^D$ bestehen bleiben. Denn für jeden dieser Vektoren muss ein LGS gelöst werden, um den gewünschten Koeffizientenvektor $\mathbf{u}_y^h$ zu erhalten. Dabei sind unter anderem die Kolmogorov N-Breiten hilfreich, welche im nächsten Abschnitt eingeführt werden.

Nachdem gleich noch zusätzliche Notation für den Rest der Arbeit eingeführt wurde, wird im nächsten Teil mehr auf die eben genannten Methoden eingegangen. Sei $\mathbf{G} := (\langle \varphi_i, \varphi_j \rangle_{\mathcal{H}})_{i,j=1}^{D} \in \mathbb{R}^{DxD}$, die symmetrische, positiv definite Gram Matrix der Basisvektoren $(\varphi_i)_{i=1}^{D}$. Diese beschreibt die Matrix von einer Menge von Vektoren eines inneren Produktraumes, welche die hermitische Matrix der inneren Produkte darstellt [27].
Mit einem Koeffizientenvektor $\mathbf{v}$ bezüglich der Basis $(\varphi_i)_{i=1}^{D}$ hat man für jedes $v \in U^h$: [13, Gleichung 2.41]

$$|\mathbf{v}|_{\mathbf{G}} := \left| \mathbf{G}^{\frac{1}{2}} \mathbf{v} \right| = \|v\|_{\mathcal{H}} . \tag{2.4}$$

2.3 Theorie der reduzierten Basen

In diesem Unterabschnitt wird, sofern nicht anders angegeben, [13, Kapitel 5] und den dortigen Verweisen gefolgt. Die Hauptmotivation für die Theorie der RBs liegt in der Tatsache, dass unter der Annahme 2.1 die Lösungsmannigfaltigkeit (Menge der Lösungen) $S(\mathcal{Y})$ eine kompakte Teilmenge von $\mathcal{H}$ ist. Diese Kompaktheiteigenschaft wirft die Frage auf, ob es entweder für jedes $\tilde{\epsilon} \neq \hat{\epsilon}$ möglich ist, ein endlich dimensionalen Unterraum $U_{\tilde{\epsilon}}^{rb}$ zu konstruieren, so dass $d(\tilde{\epsilon}) := dim(U_{\tilde{\epsilon}}^{rb}) \ll D$ und das

$$\sup_{y \in \mathcal{Y}} \inf_{\omega \in U_{\tilde{\epsilon}}^{rb}} \|u_y - \omega\|_{\mathcal{H}} \leq \tilde{\epsilon}, \tag{2.5}$$

gilt, oder äquivalent dazu, ob ein linear unabhängiger Vektor $(\psi_i)_{i=1}^{d(\tilde{\epsilon})}$ existiert mit der Eigenschaft, dass

$$\left\| \sum_{i=1}^{d(\tilde{\epsilon})} (\mathbf{c}_y)_i \psi_i - u_y \right\|_{\mathcal{H}} \leq \tilde{\epsilon} \;\; \forall \;\; y \in \mathcal{Y} \;\; und \;\; Koeffizientenvektoren \;\; \mathbf{c}_y \in \mathbb{R}^{d(\tilde{\epsilon})}.$$

Diese Eigenschaften werden in Theorem 2.4 nochmal aufgegriffen und später in Kapitel 6 bewiesen. Für diese Theorie werden nun die Kolmogorov N - Breiten eingeführt.

Definition 2.3 ([14]): „Für $N \in \mathbb{N}$ ist die Kolmogorov N-Breite einer begrenzten Teilmenge X eines normierten Raums V definiert durch

$$W_N(X) := \inf_{\substack{V_N \subset^s V \\ dim(V_N) \leq N}} \sup_{x \in X} \; dist(x, V_N)“$$

[25, Definition 2.3].
Hiermit lässt sich am besten der einheitliche Approximationsfehler von X beschreiben, wenn von einem höchstens N-dimensionalen linearen Unterraum V ausgegangen wird. In Kapitel 5 werden die oberen Grenzen für $W_N(S(\mathcal{Y}))$ besprechen. Das Ziel dieser RBMs ist es, einen Raum $U_{\bar{\epsilon}}^{rb}$ zu konstruieren, so dass die Menge des Supremums $\sup_{y\in\mathcal{Y}} dist(u_y, U_{\bar{\epsilon}}^{rb})$ sehr nah an $W_{d(\bar{\epsilon})}(S(\mathcal{Y}))$ liegt.
Die Identifizierung der Basisvektoren $(\psi_i)_{i=1}^{d(\bar{\epsilon})}$ von $U_{\bar{\epsilon}}^{rb}$ findet normalerweise in der Offline-Phase statt. Dort hat man für gewöhnlich große Rechenressourcen zur Verfügung, und dieser Prozess basiert in der Regel auf der Bestimmung von HFD von Stichproben der Parametermenge $\mathcal{Y}$. Die gebräuchlichsten Methoden beruhen auf einem (schwachen) gierigen Verfahren, dieses besteht in einer iterativen Stichprobe aus dem Parameterraum, die bei jedem Schritt ein geeignetes Optimalitätskriterium erfüllt, das sich auf die a posteriori-Fehlerschätzung stützt. Im Zusammenhang mit RB-Methoden ist ein gieriger Algorithmus ein Verfahren zur Konstruktion eines Unterraums durch iteratives Hinzufügen eines neuen Basisvektors bei jedem Schritt, anstatt über alle möglichen N-dimensionalen Unterräume zu optimieren. Mit anderen Worten, bei jedem Schritt ist der Punkt, an dem man ankommt, das Element der Lösungsmenge, das am schlechtesten durch die aktuelle RB-Approximation approximiert wird (siehe z. B. [13, Kapitel 7] und die dortigen Verweise). Oder auf echten orthogonalen Zerlegungen (POD). Das ist eine Technik zum Reduzieren der Dimension von Datenmengen oder auch von Systemen, indem diese auf einer Orthonormalbasis dargestellt werden, welche im Sinne der kleinsten Quadrate optimal ist. Dabei werden die ursprünglichen Variablen in eine neue Menge von unkorrelierten Variablen transformiert (POD-Modi oder Hauptkomponenten), wobei die ersten Komponenten idealerweise die meiste Energie der ursprünglichen Variablen behalten. Durch das Kürzen der neuen auf die erwähnten ersten Komponenten wird die neue Datenmenge mit geringerer Dimension als zuvor gewonnen(siehe z. B. [13, Kapitel 6] und die dortigen Verweise). Im letzten Schritt wird ein Orthogonalisierungsverfahren (z.B. Gram-Schmidt-Verfahren) durchgeführt, um eine orthonormale Menge der Basisvektoren $(\psi_i)_{i=1}^{d(\bar{\epsilon})}$ zu erhalten. In der anschließenden Online-Phase werden für gegebene Eingaben y die entsprechenden niedrigdimensionalen Steifigkeitsmatrizen und -vektoren gesammelt, mit welchen die Galerkin-Lösung durch das Lösen eines niedrigdimensionalen linearen Gleichungssystems berechnet wird.
Meist wird zwischen 3 verschiedenen Arten von RB unterschieden: die Lagrange RB, die Hermiten RB und der Taylor RB. Für das Verständnis dieser Arbeit sind die Einzel-

heiten dieser Methoden nicht wichtig. Wichtig ist nur, dass später angenommen wird, es existiere eine kleine RB $(\psi_i)_{i=1}^{d(\tilde{\epsilon})}$, welche durch beliebige lineare Kombinationen von High-Fidelity Elementen generiert wurde und dass alle 3 Typen der genannten RBM diese Annahme erfüllen.
Die nächste Aussage liefert eine (im Allgemeinen scharfe) obere Schranke, die die Möglichkeit der Konstruktion von kleinen Auszügen von RBs direkt mit der Kolmogorov-N-Breite in Zusammenhang bringt.

Theorem 2.4 [15, Theorem 4.1.]). „Sei $N \in \mathbb{N}$. Für eine kompakte Teilmenge X eines normierten Raums V definieren wir die innere N-Breite von X durch:

$$\bar{W}_N(X) := \inf_{V_N \in \mathcal{M}_N} \sup_{x \in X} \; dist(x, V_N),$$

wobei $\mathcal{M}_N := \{V_N \subset^s V : V_N = span(x_i)_{i=1}^N, x_1, ..., x_N \in X\}$. Dann ist

$$\bar{W}_N(X) \leq (N+1) W_N(X) \tag{2.6}$$

(Beweis folgt in Kapitel 6). Hier angewendet sagt Theorem 2.4, dass für jedes $N \in \mathbb{N}$ Lösungen $u^h(y^i) \approx u(y^i), i = 1, ..., N$ für (2.1) existieren, so dass:

$$\sup_{y \in \mathcal{Y}} \inf_{\omega \in span(u^h(y^i))_{i=1}^N} \|u_y - \omega\|_{\mathcal{H}} \leq (N+1) W_N(S(\mathcal{Y}))\text{“}$$

[25, Theorem 2.4].

Bemerkung 2.5. „Wir stellen fest, dass diese Schranke für allgemeine X, V scharf ist (gibt die höchste Laufzeit eines Algorithmus an.) Allerdings ist sie nicht unbedingt optimal für bestimmte Instanzen von $S(\mathcal{Y})$. Wenn beispielsweise $W_N(S(\mathcal{Y}))$ polynomial zerfällt, zerfällt $\bar{W}_N(S(\mathcal{Y}))$ genau so schnell [15, Theorem 3.1]. Falls nun $W_N(S(\mathcal{Y})) \leq Ce^{-cN^\beta}$ für manche $c, C, \beta > 0$ ist, bekommen wir $\bar{W}_N(S(\mathcal{Y})) \leq \tilde{C}e^{-\tilde{c}N^\beta}$ für einige $\tilde{c}, \tilde{C} > 0$“[25, Bemerkung 2.5].

Mit dem gerade Erwähnten als Rechtfertigung wird angenommen, dass für jedes $\tilde{\epsilon} \geq \hat{\epsilon}$ eine RB $U_{\tilde{\epsilon}}^{rb} = span(\psi_i)_{i=1}^{d(\tilde{\epsilon})}$ existiert, welche die Gleichung (2.5) erfüllt. Seien nun die linear unabhängigen Basisvektoren $(\psi_i)_{i=1}^{d(\tilde{\epsilon})}$ lineare Kombinationen der high-fidelity Basisvektoren $(\varphi_i)_{i=1}^D$ in dem Sinne, dass eine Matrix $\mathbf{V}_{\tilde{\epsilon}} \in \mathbb{R}^{D \times d(\tilde{\epsilon})}$ existiert, so dass

gilt

$$(\psi_i)_{i=1}^{d(\tilde{\epsilon})} = \left(\sum_{j=1}^{D} (\mathbf{V}_{\tilde{\epsilon}})_{j,i} \varphi_j \right)_{i=1}^{d(\tilde{\epsilon})}$$

mit $d(\tilde{\epsilon}) \ll D$ so klein wie möglich gewählt und trotzdem noch $(S(\mathcal{Y}), U_{\tilde{\epsilon}}^{rb}) \leq \bar{W}_{d(\tilde{\epsilon})}(S(\mathcal{Y}))$.

Zusätzlich gilt die Annahme, dass die Vektoren $(\psi_i)_{i=1}^{d(\tilde{\epsilon})}$ ein orthonormales System in $\mathcal{H}$ bilden, welches äquivalent dazu ist, dass die Spalten von $\mathbf{G}^{\frac{1}{2}} \mathbf{V}_{\tilde{\epsilon}}$ orthonormal sind [13, Bemerkung 4.1]. Dies impliziert

$$\left\| \mathbf{G}^{\frac{1}{2}} \mathbf{V}_{\tilde{\epsilon}} \right\|_2 = 1, \quad \forall\ \tilde{\epsilon} \geq \hat{\epsilon} \tag{2.7}$$

und auch

$$\left\| \sum_{i=1}^{d(\tilde{\epsilon})} \mathbf{c}_i \psi_i \right\|_{\mathcal{Y}} = |\mathbf{c}|, \quad \forall\ \mathbf{c} \in \mathbb{R}^{d(\tilde{\epsilon})}. \tag{2.8}$$

Für die Diskretisierungsmatrix kann folgendes gezeigt werden (siehe z. B. [13, Abschnitt 3.4.1])

$$\mathbf{B}_{y,\tilde{\epsilon}}^{rb} := (b_y(\psi_j, \psi_i))_{i,j}^{d(\tilde{\epsilon})} = \mathbf{V}_{\tilde{\epsilon}}^T \mathbf{B}_{y,\tilde{\epsilon}}^{h} \mathbf{V}_{\tilde{\epsilon}} \ \in \mathbb{R}^{d(\tilde{\epsilon}) \mathrm{x} d(\tilde{\epsilon})}, \quad \forall\ y \in \mathcal{Y}.$$

Nun können die zuvor erwähnten Eigenschaften der variierenden Bilinearform genutzt und die Orthonormalität der Basisvektoren $(\psi_i)_{i=1}^{d(\tilde{\epsilon})}$ hinzugezogen werden, dadurch lässt sich zeigen, dass [13, Bemerkung 3.5] gilt

$$C_{coer} \leq \left\| \mathbf{B}_{y,\tilde{\epsilon}}^{rb} \right\|_2 \leq C_{cont}, \ \textit{wie auch} \ \frac{1}{C_{cont}} \leq \left\| (\mathbf{B}_{y,\tilde{\epsilon}}^{rb})^{-1} \right\|_2 \leq \frac{1}{C_{coer}}, \ \forall\ y \in \mathcal{Y}. \tag{2.9}$$

Damit lässt sich sagen, dass die Steifigkeitsmatrix immer begrenzt und gleichzeitig unabhängig von y oder $d(\tilde{\epsilon})$ ist. Außerdem ist die diskretisierte rechte Seite bezüglich der RB gegeben durch

$$\mathbf{f}_{y,\tilde{\epsilon}}^{rb} := (f_y(\psi_i))_{i=1}^{d(\tilde{\epsilon})} = \mathbf{V}_{\tilde{\epsilon}}^T \mathbf{f}_{y,\tilde{\epsilon}}^{h} \in \mathbb{R}^{d(\tilde{\epsilon})}$$

und mit der Bessel-Ungleichheit folgt $|\mathbf{f}_{y,\tilde{\epsilon}}^{rb}| \leq \|f_y\|_{\mathcal{H}^*} \leq C_{rhs}$. Die besagt, dass ein Vektor eines Hilbertraums, die gleiche Länge besitzt, wie seine Orthogonalprojektion

auf einem beliebigen Untervektorraum. Darüber hinaus sei

$$\mathbf{u}^{rb}_{y,\tilde{\epsilon}} := (\mathbf{B}^{rb}_{y,\tilde{\epsilon}})^{-1}\mathbf{f}^{rb}_{y,\tilde{\epsilon}}$$

der Koeffizientenvektor des RB Raumes für die Galerkin-Lösung. Dann kann die Galerkinlösung $u^{rb}_{y,\tilde{\epsilon}}$ auch geschrieben werden als

$$u^{rb}_{y,\tilde{\epsilon}} = \sum_{i=1}^{d(\tilde{\epsilon})}(\mathbf{u}^{rb}_{y,\tilde{\epsilon}})_i\psi_i = \sum_{j=1}^{D}(\mathbf{V}_{\tilde{\epsilon}}\mathbf{u}^{rb}_{y,\tilde{\epsilon}})_j\varphi_j,$$

das heißt, wenn nun $\mathbf{u}^{rb}_{y,\tilde{\epsilon}} := (\mathbf{B}^{rb}_{y,\tilde{\epsilon}})^{-1}\mathbf{f}^{rb}_{y,\tilde{\epsilon}}$ benutzt und $\mathbf{f}^{rb}_{y,\tilde{\epsilon}}$ umgeschrieben wird wie zuvor, und letztendlich unser $(\mathbf{B}^{rb}_{y,\tilde{\epsilon}})^{-1}$ so geschrieben wird wie nach Gleichung (2.8), ergibt das

$$u^{rb}_{y,\epsilon} = \sum_{j=1}^{D}\left(\frac{\mathbf{V}_{\tilde{\epsilon}}\left(\mathbf{B}^{h}_{y,\tilde{\epsilon}}\right)^{-1}\mathbf{V}^{T}_{\tilde{\epsilon}}\mathbf{f}^{h}_{y,\tilde{\epsilon}}}{\mathbf{V}_{\tilde{\epsilon}}\mathbf{V}^{T}_{\tilde{\epsilon}}}\right)$$

und daraus resultiert, dass

$$\tilde{\mathbf{u}}^{h}_{y,\tilde{\epsilon}} := \mathbf{V}_{\tilde{\epsilon}}\mathbf{u}^{rb}_{y,\tilde{\epsilon}} \in \mathbb{R}^D$$

der Koeffizientenvektor der RB-Lösung ist, wenn er erweitert wird in Bezug auf die High-Fidelity-Basis $(\varphi_i)_{i=1}^{D}$. Abschließend, wie in Gleichung (2.3), erhalten wir

$$\sup_{y\in\mathcal{Y}}\left\|u_y - u^{rb}_{y,\tilde{\epsilon}}\right\|_{\mathcal{H}} \;\leq\; \sup_{y\in\mathcal{Y}}\frac{C_{cont}}{C_{coer}}\inf_{\omega\in U^{rb}_{\tilde{\epsilon}}}\|u_y-\omega\|_{\mathcal{H}} \;\leq\; \frac{C_{cont}}{C_{coer}}\tilde{\epsilon}.$$

Im folgenden wird versucht zu zeigen, dass es möglich ist, RBMs mit NNs zu imitieren. Dafür wird probiert, approximativ NNs zu konstruieren, mit denen die Abbildungen $\mathbf{u}^{rb}_{\cdot,\tilde{\epsilon}}, \mathbf{u}^{h}_{\cdot,\tilde{\epsilon}}$ approximiert werden können, wobei darauf geachtet werden muss, dass die Komplexität nur abhängig von der Größe der RB ist und gleichzeitig fast immer linear auf D ist. Erst werden kleine NNs konstruiert, um zu verstehen, wie diese funktionieren, und dann werden deren Grenzen Schritt für Schritt weiter hoch gesetzt, bis diese beliebig > 0 abgeschätzt werden können. Das gelingt durch das Implementieren einer approximativen Matrixinversion mithilfe der Richardson-Iteration. Danach wird sich darauf konzentriert, dass die Realisierungen der konstruierten NNs, also die Funktionen die daraus resultieren, die Abbildungen $\mathbf{u}^{rb}_{\cdot,\tilde{\epsilon}}, \mathbf{u}^{h}_{\cdot,\tilde{\epsilon}}$ approximieren.

3 Berechnungen mit neuronalen Netzen

„Das Ziel dieses Kapitels ist die Emulation der Matrixinversion durch NNs. In Abschnitt 3.1 werden grundlegende Begriffe im Zusammenhang mit NNs eingeführt, sowie einige grundlegende Operationen, die man mit diesen NNs durchführen kann. In Abschnitt 3.2 werden Ergebnisse überprüft, welche die Existenz von NNs zeigt, dessen ReLU Realisierungen eine Matrix $\mathbf{A} \in \mathbb{R}^{d\mathbf{x}d}, \|\mathbf{A}\|_2 < 1$ als deren Input nimmt und eine Approximation, basierend auf ihrer Neumannreihen Erweiterung, von $(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A})^{-1}$ berechnet. Die Beweise werde ich am Ende der Arbeit angeben, um den Lesefluss nicht zu stören“[25, Kapitel 3].

3.1 Grundlegende Definition und Operation

Nun folgen einige Definitionen, die später bei den Beweisen in Kapitel 6 verwendet werden. Anschließend können Operationen wie Parallelisierung oder Verkettung dazu genutzt werden, kleine, leicht verständliche NNs zu komplexeren zusammen zu setzen. In diesem Kapitel wird sich sehr stark an dem Originalartikel orientiert, da es sich hier hauptsächlich um Definitionen und Theoreme handelt. Im Originalartikel wird „NN“ als eine Familie von Gewichten bezeichnet. Allerdings ist hier das gesamte Konstrukt der neuronalen Netzes gemeint und alles was dazu gehört. Sprich der Input und Output, sowie die Architektur; Anzahl Schichten und Neuronen innerhalb des Netzes, sowie die Gewichte. Anders würden alle Verwendungen die bis jetzt in Kapitel 1 erwähnt wurden, nicht durchführbar sein, denn nur mit den Gewichten kann nichts berechnet werden, ohne das diese auf gegebenen Input und die Aktivierungsfunktion auf die Neuronen wirken kann. Die Gewichte selber sind im folgenden noch gesondert definiert. Dann gibt es noch die Realisierung. Das ist Die Funktion, die letztendlich von dem neuronalen Netz implementiert wird.

Definition 3.1. „Sei $n, L \in \mathbb{N}$. Ein NN Φ mit Inputdimension $dim_{in}(\Phi) := n$ und L Schichten ist eine Reihe von Matrix-Vektor Tupeln

$$\Phi = ((\mathbf{A}_1, \mathbf{b}_1), (\mathbf{A}_2, \mathbf{b}_2), ..., (\mathbf{A}_L, \mathbf{b}_L))$$

mit $N_0 = n$ und $N_1, ..., N_L \in \mathbb{N}$ und wo jedes $\mathbf{A}_l$ eine $N_l \mathbf{x} N_{l-1}$ Matrix ist und $\mathbf{b}_l \in \mathbb{R}^{N_l}$. Falls Φ ein NN ist wie oben beschrieben, sei $K \subset \mathbb{R}^n$ und falls $\varrho : \mathbb{R} \to \mathbb{R}$ beliebig ist, kann die zugehörige Realisierung von Φ mit Aktivierungsfunktion ϱ über K (ϱ-Realisierung von Φ über K) als die Abbildung $R_\varrho^K(\Phi) : K \to \mathbb{R}^{N_L}$ definiert werden,

so dass:

$$R_\varrho^K(\Phi)(x) = x_L,$$

wobei x_L aus dem folgenden Schema resultiert,

$$\begin{aligned} x_0 &:= x, \\ x_l &:= \varrho(\mathbf{A}_l x_l + b_l), \quad fuer\ l = 1, ..., L-1, \\ x_L &:= \mathbf{A}_L x_{L-1} + \mathbf{b}_L, \end{aligned}$$

wo ϱ komponentenweise agiert, also $\varrho(\mathbf{v}) = (\varrho(\mathbf{v}_1, ..., \varrho(\mathbf{v}_m))$ für jedes $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_m) \in \mathbb{R}^m$.
Sei $N(\Phi) := n + \sum_{j=1}^{L} N_j$ die Anzahl der Neuronen des NNs und $L = L(\Phi)$ die Anzahl der Schichten. Für $l \leq L$ sei $M_l(\Phi) := \|\mathbf{A}_l\|_0 + \|\mathbf{b}_l\|_0$ die Anzahl der Gewichte in der l-ten Schicht, und definiere $M(\Phi) := \sum_{l=1}^{L} M_l(\Phi)$ als die Anzahl der Gewichte von Φ. Außerdem sei $dim_{out}(\Phi) := N_L$ als Outputdimension von Φ.
Wichtig zu wissen ist, dass sich 2 NNs wie folgt verketten lassen"[25, Definition 3.1].

Definition 3.2. „Seien $L_1, L_2 \in \mathbb{N}$ und seien $\Phi^1 = ((\mathbf{A}_1^1, \mathbf{b}_1^1), ..., (\mathbf{A}_{L_1}^1, \mathbf{b}_{L_1}^1)), \Phi^2 = ((\mathbf{A}_1^2, \mathbf{b}_1^2), ..., (\mathbf{A}_{L_2}^2, \mathbf{b}_{L_2}^2))$ zwei NNs, so dass die Inputschicht von Φ^1 die selbe Dimension hat wie die Outputschicht von Φ^2. Dann beschreibt $\Phi^1 \bullet \Phi^2$ das folgende $L_1 + L_2 - 1$ schichtige NN:

$$\Phi^1 \bullet \Phi^2 := ((\mathbf{A}_1^2, \mathbf{b}_1^2), ..., (\mathbf{A}_{L_2-1}^2, \mathbf{b}_{L_2-1}^2), (\mathbf{A}_1^1 \mathbf{A}_{L_2}^2, \mathbf{A}_1^1 \mathbf{b}_{L_2}^2 + \mathbf{b}_1^1), (\mathbf{A}_2^1, \mathbf{b}_2^1), ..., (\mathbf{A}_{L_1}^1, \mathbf{b}_{L_1}^1))$$

Dies wird die Verkettung zweier NNs genannt.
Im Allgemeinen gibt es keine Begrenzung für $M(\Phi^1 \bullet \Phi^2)$, welches linear ist in $M(\Phi^1)$ und $M(\Phi^2)$.
Von nun an wird festgelegt, dass ϱ durch die ReLU Aktivierungsfunktion gegeben ist mit $\varrho(x) = max\{x, 0\}$ für $x \in \mathbb{R}$. Im folgenden wird klar, dass es möglich ist, Verkettungen einzuführen, welche dabei helfen, die Anzahl der Nicht-Null Gewichte zu kontrollieren. Das folgende Ergebnis zeigt, dass man NNs konstruieren kann, dessen ReLU Realisierungen die Identitätsfunktion von $\mathbb{R}^n$ darstellen"[25, Definition 3.2].

Lemma 3.3. „Für jedes $L \in \mathbb{N}$ existiert ein NN $\Phi_{n,L}^{\mathbf{Id}}$ mit Inputdimension n, Outputdimension n und höchstens $2nL$ Gewichte, die ungleich null sind und die Werte

$\{-1, 1\}$ annehmen, so dass

$$R_{\varrho}^{\mathbb{R}^n}\left(\Phi_{n,L}^{\mathbf{Id}}\right) = \mathbf{Id}_{\mathbb{R}^n}.$$

Es wird noch eine weitere Form der Verkettung eingeführt, die Sparse-Verkettung (also die Verkettung von dünnbesetzten Matrix-Vektor Tupeln, dementsprechend auch von dünnbesetzten NNs)"[Lemma 3.3].

Definition 3.4. „Seien Φ^1, Φ^2 zwei NNs, so dass die Outputdimension von Φ^2 und die Inputdimension von Φ^1 gleich $n \in \mathbb{N}$ sind. Dann ist die Sparse-Verkettung von Φ^1 und Φ^2 definiert als

$$\Phi^1 \odot \Phi^2 := \Phi^1 \bullet \Phi_{n,1}^{Id} \bullet \Phi^2.$$

Später in Lemma 3.6 werden die Eigenschaften einer Sparse-Verkettung näher gezeigt. Jetzt noch eine weitere Operation für NNs, die Parallelisierung"[25, Definition 3.4].

Definition 3.5 ([16]). „Seien $\Phi^1, ..., \Phi^k$ NNs mit gleicher Inputdimension, so dass gilt: $\Phi^i = ((\mathbf{A}_1^i, \mathbf{b}_1^i), ..., (\mathbf{A}_L^i, \mathbf{b}_L^i)$ für manche $L \in \mathbb{N}$. Dann wird die Parallelisierung von $\Phi^1, ..., \Phi^k$ wie folgt definiert

$$P(\Phi^1, ..., \Phi^k) = \left(\left(\begin{pmatrix} \mathbf{A}_1^1 & & \\ & \ddots & \\ & & \mathbf{A}_1^k \end{pmatrix}, \begin{pmatrix} \mathbf{b}_1^1 \\ \vdots \\ \mathbf{b}_1^k \end{pmatrix}\right), ..., \left(\begin{pmatrix} \mathbf{A}_L^1 & & \\ & \ddots & \\ & & \mathbf{A}_L^k \end{pmatrix}, \begin{pmatrix} \mathbf{b}_L^1 \\ \vdots \\ \mathbf{b}_L^k \end{pmatrix}\right)\right) \tag{3.1}$$

Sei nun Φ ein NN und $L \in \mathbb{N}$, so dass $L(\Phi) \leq L$. Dann definiere das NN mit

$$E_L(\Phi) = \begin{cases} \Phi, & \text{falls } L(\Phi) = L, \\ \Phi_{dim_{out}(\Phi), L-L(\Phi)}^{Id} \odot \Phi, & \text{falls } L(\Phi) < L. \end{cases} \tag{3.2}$$

Zuletzt seien $\tilde{\Phi}^1, ..., \tilde{\Phi}^k$ NNs mit derselben Inputdimension und sei

$$\tilde{L} := max\{L\left(\tilde{\Phi}^1\right), ..., L\left(\tilde{\Phi}^k\right)\}. \tag{3.3}$$

Dann definieren wir

$$P\left(\tilde{\Phi}^1, ..., \tilde{\Phi}^k\right) := P\left(E_{\tilde{L}}\left(\tilde{\Phi}^1\right), ..., E_{\tilde{L}}\left(\tilde{\Phi}^k\right)\right). \tag{3.4}$$

Wir nennen $P\left(\tilde{\Phi}^1, ..., \tilde{\Phi}^k\right)$ die Parallelisierung von $\tilde{\Phi}^1, ..., \tilde{\Phi}^k$.

Folgendes Lemma wurde in [17, Lemma 5.4] eingeführt und prüft die Eigenschaften der spärlichen Verkettung wie auch die der Parallelisierung der NNs“[25, Definition 3.5].

Lemma 3.6([17]). „Seien $\Phi^1, ..., \Phi^k$ NNs.

(a) Falls die Inputdimension von Φ^1, welche wir mit n_1 beschreiben, der Outputdimension von Φ^2 gleicht, und die Inputdimension sei n_2, dann gilt

$$R_\varrho^{\mathbb{R}^{n_1}}(\Phi^1) \circ R_\varrho^{\mathbb{R}^{n_2}}(\Phi^2) = R_\varrho^{\mathbb{R}^{n_2}}(\Phi^1 \odot \Phi^2)$$

und

$$\begin{aligned}
&(i) L(\Phi^1 \odot \Phi^2) \leq L(\Phi^1) + L(\Phi^2),\\
&(ii) M(\Phi^1 \odot \Phi^2) \leq M(\Phi^1) + M(\Phi^2) + M_1(\Phi^1) + M_{L(\Phi^2)}(\Phi^2)\\
&\qquad\qquad \leq 2M(\Phi^1) + 2M(\Phi^2),\\
&(iii) M_1(\Phi^1 \odot \Phi^2) = M_1(\Phi^2),\\
&(iv) M_{L(\Phi^1 \odot \Phi^2)}(\Phi^1 \odot \Phi^2) = M_{L(\Phi^1)}(\Phi^1).
\end{aligned}$$

(b) Falls die Inputdimension von Φ^i n der Inputdimension von Φ^j für alle i,j, gleicht, gilt für das NN $P(\Phi^1, ..., \Phi^k)$ und alle $x_1, ..., x_k \in \mathbb{R}^n$

$$R_\varrho^{\mathbb{R}^n}(P(\Phi^1, ..., \Phi^k))(x_1, ..., x_k) = \left(R_\varrho^{\mathbb{R}^n}(\Phi^1)(x_1), R_\varrho^{\mathbb{R}^n}(\Phi^2)(x_2), ..., R_\varrho^{\mathbb{R}^n}(\Phi^k)(x_k)\right)$$

so wie auch

$$\begin{aligned}
&(i) L(P(\Phi^1, ..., \Phi^k)) = max_{i=1,...,k} L(\Phi^i),\\
&(ii) M(P(\Phi^1, ..., \Phi^k)) \leq 2\left(\textstyle\sum_{i=1}^k M(\Phi^i)\right) + 4\left(\textstyle\sum_{i=1}^k dim_{out}(\Phi^i)\right) max_{i=1,...,k} L(\Phi^i),\\
&(iii) M(P(\Phi^1, ..., \Phi^k)) = \textstyle\sum_{i=1}^k M(\Phi^i), \text{ falls } L(\Phi^1) = ... = L(\Phi^k),
\end{aligned}$$

$(iv) M_1(P(\Phi^1, ..., \Phi^k)) = \sum_{i=1}^{k} M_1(\Phi^i)$,
$(v) M_{L(P(\Phi^1,...,\Phi^k))}(P(\Phi^1, ..., \Phi^k)) \leq \sum_{i=1}^{k} max\{2dim_{out}(\Phi^i), M_{L(\Phi^i)}(\Phi^i)\}$,
$(vi) M_{L(P(\Phi^1,...,\Phi^k))}(P(\Phi^1, ..., \Phi^k)) = \sum_{i=1}^{k} M_{L(\Phi^i)}(\Phi^i)$, falls $L(\Phi^1) = ... = L(\Phi^k)$. "[25, Lemma 3.6]

3.2 Ein Netzwerk-basierter Versuch der Matrixinversion

„Das Ziel dieses Abschnittes ist es, das Inverse der Quadratmatrizen mit Realisierungen der NNs zu imitieren, welche verhältnismäßig klein sind. Theorem 3.8 zeigt, dass für $d \in \mathbb{N}, \epsilon \in (0, \frac{1}{4})$ und $\delta \in (0, 1)$, NNs $\Phi_{inv;\epsilon}^{1-\delta,d}$ konstruiert werden können, dessen ReLU Realisierungen die folgende Abbildung

$$\{\mathbf{A} \in \mathbb{R}^{d\mathbf{x}d} : \|\mathbf{A}\|_2 \leq 1 - \delta\} \to \mathbb{R}^{d\mathbf{x}d}, \mathbf{A} \mapsto (\mathbf{Id}_{\mathbb{R}} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k$$

bis zu einem Fehler $\|\cdot\|_2$ von ϵ approximieren.
Um in der NN-Umgebung zu bleiben werden vektorisierte Matrizen genutzt. Dies sieht dann wie folgt aus. Sei $\mathbf{A} \in \mathbb{R}^{d\mathbf{x}l}$, dann

$$\mathbf{vec}(\mathbf{A}) := (\mathbf{A}_{1,1}, ..., \mathbf{A}_{d,1}, ..., \mathbf{A}_{1,l}, ..., \mathbf{A}_{d,l})^T \in \mathbb{R}^{dl}.$$

Dies geht aber auch in die andere Richtung, nämlich gibt ein Vektor $(\mathbf{v}_{1,1}, ..., \mathbf{v}_{d,1}, ..., \mathbf{v}_{1,d}, ..., \mathbf{v}_{d,l})^T \in \mathbb{R}^{dl}$ uns

$$\mathbf{matr}(\mathbf{v}) := (\mathbf{v}_{i,j})_{i=1,...d,j=1,...,l} \in \mathbb{R}^{d\mathbf{x}l}.$$

Zusätzlich wird für $d, n, l \in \mathbb{N}, Z > 0$ eine Menge festgesetzt, durch die später die Matrizen in ihrer Norm beschränkt werden können

$$K_{d,n,l}^{Z} := \{(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) : (\mathbf{A}, \mathbf{B}) \in \mathbb{R}^{d\mathbf{x}n}\mathbf{x}\mathbb{R}^{n\mathbf{x}l}, \|\mathbf{A}\|_2, \|\mathbf{B}\|_2 \leq Z\}$$

wie auch

$$K_{d}^{Z} := \{\mathbf{vec}(\mathbf{A}) : \mathbf{A} \in \mathbb{R}^{d\mathbf{x}d}, \|\mathbf{A}\|_2 \leq Z\}.$$

Der Grundbaustein für die Konstruktion von NNs, die eine Matrixinversion simulieren,

ist das folgende Ergebnis über NNs, die die Multiplikation von zwei Matrizen simulieren"[25, Kapitel 3.2].

Proposition 3.7. „Sei $d, n, l \in \mathbb{N}, \epsilon \in (0,1), Z > 0$: dann existiert ein NN der Form $\Phi_{mult;\epsilon}^{Z,d,n,l}$ mit $n \cdot (d+l)$ dimensionalem Input und dl dimensionalem Output, so dass für eine absolute Konstante $C_{mult} > 0$ die folgenden Eigenschaften gelten:

$(i) L\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right) \leq C_{mult} \cdot \left(\log_2(\frac{1}{\epsilon}) + \log_2\left(n\sqrt{dl}\right) + \log_2(max\{1, Z\})\right)$,
$(ii) M\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right) \leq C_{mult} \cdot \left(\log_2(\frac{1}{\epsilon}) + \log_2\left(n\sqrt{dl}\right) + \log_2(max\{1, Z\})\right) dnl$,
$(iii) M_1\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right) \leq C_{mult} dnl$, wie auch $M_{L\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right)}\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right) \leq C_{mult} dnl$,
(iv) $\sup_{(\mathbf{vec}(\mathbf{A}),\mathbf{vec}(\mathbf{B}))\in K_{d,n,l}^{Z}}$
$\left\|\mathbf{AB} - \mathbf{matr}\left(R_{\varrho}^{K_{d,n,l}^{Z}}\left(\Phi_{mult;\epsilon}^{Z,d,n,l}\right)(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B}))\right)\right\|_2 \leq \epsilon$,
(v) für alle $(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) \in K_{d,n,l}^{Z}$ bekommt man

$$\left\|\mathbf{matr}\left(R_{\varrho}^{K_{d,n,l}^{Z}}\left(\Phi_{mult;\epsilon}^{Z,d,n,l}\right)(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B}))\right)\right\|_2 \leq \epsilon + \|\mathbf{A}\|_2 \|\mathbf{B}\|_2 \leq \epsilon + Z^2 \leq 1 + Z^2.$$

Der Beweis ist in Kapitel 6 zu finden. Die dort konstruierten NNs werden dafür genutzt, um das folgende Resultat zu beweisen"[25, Proposition 3.7].

Theorem 3.8. „Für $\epsilon, \delta \in (0,1)$ definieren wir

$$m(\epsilon, \delta) := \left\lceil \frac{\log_2(0.5\epsilon\delta}{\log_2(1-\delta)} \right\rceil.$$

Dann gibt es eine Konstante $C_{inv} > 0$, so dass für jedes $d \in \mathbb{N}, \epsilon \in (0, \frac{1}{4})$ und jedes $\delta \in (0,1)$ ein NN $\Phi_{inv;\epsilon}^{1-\delta,d}$ existiert mit d^2 dimensionalem Output und den folgenden Eigenschaften:

$(i) L\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right) \leq C_{inv} \log_2(m(\epsilon,\delta)) \cdot (\log_2(\frac{1}{\epsilon}) + \log_2(m(\epsilon,\delta)) + \log_2(d))$,
$(ii) M\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right) \leq C_{inv} m(\epsilon,\delta) \log_2^2(m(\epsilon,\delta)) d^3 \cdot (\log_2(\frac{1}{\epsilon}) + \log_2(m(\epsilon,\delta)) + \log_2(d))$,
$(iii) sup_{\mathbf{vec}(\mathbf{A})\in K_d^{1-\delta}} \left\|(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A})^{-1} - \mathbf{matr}\left(R_{\varrho}^{1-\delta}\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \leq \epsilon$,

(*iv*) für alle $\mathbf{vec}(\mathbf{A}) \in K_d^{1-\delta}$ bekommen wir

$$\left\|\mathbf{matr}\left(R_\varrho^{1-\delta}\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \leq \epsilon + \left\|(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A})^{-1}\right\|_2 \leq \epsilon + \frac{1}{1 - \|\mathbf{A}\|_2} \leq \epsilon + \frac{1}{\delta}.$$

Im Beweis von Theorem 3.8 approximieren wir die Funktion, die eine Matrix auf ihre Inverse abbildet, durch die Neumann-Reihe und ersetzen dann diese Konstruktion durch NNs. Dies ließe sich auch mit Chebyshev Matrixpolynome durchführen. Allerdings wird der geringere Grad der Approximation dadurch nicht die Nachteile überspielen wie beispielsweise die Notwendigkeit größerer Netzwerke oder Koeffizienten, die exponentiell mit dem Grad des Polynoms wachsen“[25, Theorem 3.8].

4 Neuronale Netze und Lösungen von PDEs bei Nutzung der reduzierten Basen

In diesem Kapitel wird sich auf die Approximative Schätzung der Inversen Matrix aus Abschnitt 3.2 berufen, um die parameterabhängige Lösung der parametrischen PDEs durch NNs zu approximieren. Anders gesagt, für $\tilde{\epsilon} \geq \hat{\epsilon}$ werden NNs konstruiert, die die folgende Abbildung approximieren:

$$\mathcal{Y} \to \mathbb{R}^D : y \mapsto \tilde{\mathbf{u}}^h_{y,\tilde{\epsilon}}, \ und \ \mathcal{Y} \to \mathbb{R}^{d(\tilde{\epsilon})} : y \mapsto \mathbf{u}^{rb}_{y,\tilde{\epsilon}}.$$

Hier hängen die Größen der NNs im Wesentlichen nur von der Approximationsgenauigkeit $\tilde{\epsilon}$ und der Größe $d(\tilde{\epsilon})$ einer geeigneten RB ab, sind aber unabhängig oder höchsten linear in der Dimension der High-Fidelity-Diskretisierung D. In 4.1 wird zunächst unter einigen allgemeinen Annahmen zum parametrischen Problem ein NN konstruiert, das die Abbildungen $\tilde{\mathbf{u}}^h_{\cdot,\tilde{\epsilon}}$ und $\mathbf{u}^{rb}_{\cdot,\tilde{\epsilon}}$ imitiert.

4.1 Bestimmung der Koeffizienten der Lösung

Es wird jetzt eine Konstruktion eines NNs vorgestellt, dessen implizierte Funktionen dazu genutzt wird, die Abbildungen $\tilde{\mathbf{u}}^h_{\cdot,\tilde{\epsilon}}$ und $\mathbf{u}^{rb}_{\cdot,\tilde{\epsilon}}$ zu approximieren. Am Ende des Abschnitts wird deutlich, dass der Approximationsfehler mit der Norm $|\cdot|_{\mathbf{G}} - norm$ gemessen wird, wenn es um die NN Approximation für $\tilde{\mathbf{u}}^h_{\cdot,\tilde{\epsilon}}$ geht, da die Norm ein direkten Bezug zu der $\mathcal{H}$-Norm in (2.4) hat. Verglichen dazu wird der Approximationsfehler der zweiten Abbildung $\mathbf{u}^{rb}_{\cdot,\tilde{\epsilon}}$ anhand der $|\cdot| - norm$ gemessen, wenn man sich an Gleichung (2.8) orientiert.
Im folgenden werden viele Argumente bezüglich der Matrix $\mathbf{B}^{rb}_{y,\tilde{\epsilon}}$ aus Theorem 3.8 verwendet. Es lässt sich beweisen, dass die Matrix $\mathbf{B}^{rb}_{y,\tilde{\epsilon}}$ neu skalieren lässt. Dazu ist nur eine Konstante $\alpha := (C_{coer} + C_{cont})^{-1}$ nötig, welche zugleich unabhängig von y und $d(\tilde{\epsilon})$ ist. Wenn das erfüllt ist, kann folgende Annahme getroffen werden:

$$\left\| \mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon})}} - \alpha \mathbf{B}^{rb}_{y,\tilde{\epsilon}} \right\|_2 \leq 1 - \delta < 1.$$

„Wir setzen die Werte von α und δ für den Rest der Arbeit fest. Als nächstes nennen wir zwei abstrakte Annahmen über die Approximierbarkeit der Abbildung $\mathbf{B}^{rb}_{y,\tilde{\epsilon}}$, die wir später bei Betrachtung konkreter Beispiele in Unterabschnitt 4.2 benutzen“[25, Kapitel 4.1].

Annahme 4.1. „Wir nehmen an, dass für jedes $\tilde{\epsilon} \neq \hat{\epsilon}, \epsilon > 0$ und eine entsprechende RB $(\psi_i)_{i=1}^{d(\tilde{\epsilon})}$ ein NN $\Phi_{\tilde{\epsilon},\epsilon}^{B}$ existiert mit p-dimensionalem Input und $d(\tilde{\epsilon})^2$-dimensionalem Output, so dass

$$\sup_{y\in\mathcal{Y}} \left\| \alpha \mathbf{B}_{y,\tilde{\epsilon}}^{rb} - \mathbf{matr}\left(R_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{B}\right)(y)\right)\right\|_2 \leq \epsilon.$$

Wir setzen $B_M(\tilde{\epsilon},\epsilon) := M\left(\Phi_{\tilde{\epsilon},\epsilon}^{B}\right) \in \mathbb{N}$ und $B_L(\tilde{\epsilon},\epsilon) := L\left(\Phi_{\tilde{\epsilon},\epsilon}^{B}\right) \in \mathbb{N}$“[25, Annahme 4.1].

Annahme 4.2. „Wir nehmen an, dass für jedes $\tilde{\epsilon} \neq \hat{\epsilon}, \epsilon > 0$ und eine entsprechende RB $(\psi_i)_{i=1}^{d(\tilde{\epsilon}}$ ein NN $\Phi_{\tilde{\epsilon},\epsilon}^{f}$ existiert mit p-dimensionalem Input und $d(\tilde{\epsilon})$-dimensionalem Output, so dass

$$\sup_{y\in\mathcal{Y}} |\mathbf{f}_{y,\tilde{\epsilon}}^{rb} - R_{\varrho}^{\mathcal{Y}}(\Phi_{\tilde{\epsilon},\epsilon}^{f})(y)| \leq \epsilon.$$

Wir setzen $F_L(\tilde{\epsilon},\epsilon) := L(\Phi_{\tilde{\epsilon},\epsilon}^{f})$ *und* $F_M(\tilde{\epsilon},\epsilon) := M(\Phi_{\tilde{\epsilon},\epsilon}^{f})$.

Jetzt können wir NNs konsturieren deren ReLU-Realisierungen die Koeffizientenabbildungen $\tilde{\mathbf{u}}_{\cdot,\tilde{\epsilon}}^{h}, \mathbf{u}_{\cdot,\tilde{\epsilon}}^{rb}$ approximieren“[25, Annahme 4.2].

Theorem 4.3. „Sei $\tilde{\epsilon} \neq \hat{\epsilon}$ *und* $\epsilon \in (0, \frac{\alpha}{4} \cdot min\{1, C_{coer}\})$. Außerdem definieren $\epsilon^{'} := \frac{\epsilon}{max\{6, C_{rhs}\}}$, $\epsilon^{''} := \frac{\epsilon}{3} \cdot C_{coer}$, $\epsilon^{'''} := \frac{3}{8} \cdot \epsilon^{'} \alpha C_{coer}^2$ *und* $\mathcal{K} := 2max\left\{1, C_{rhs}, \frac{1}{C_{coer}}\right\}$.

Zusätzlich gehen wir davon aus, dass Annahmen 4.1 und 4.2 gelten. Dann existieren NNs $\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}$ *und* $\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}$, so dass die folgenden Eigenschaften gelten:

(i) $\sup_{y\in\mathcal{Y}} \left|\mathbf{u}_{y,\tilde{\epsilon}}^{rb} - R_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}\right)(y)\right| \leq \epsilon$, *und* $\sup_{y\in\mathcal{Y}} \left|\tilde{\mathbf{u}}_{y,\tilde{\epsilon}}^{h} - R_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}\right)(y)\right|_{\mathbf{G}} \leq \epsilon$,

(ii) es existiert eine Konstante $C_L^{\mathbf{u}} = C_L^{\mathbf{u}}(C_{coer}, C_{cont}, C_{rhs}) > 0$, so dass

$$\begin{aligned} L\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}\right) &\leq L\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}\right) \\ &\leq C_L^{\mathbf{u}} max\{log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right)\left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right) + log_2\left(d\left(\tilde{\epsilon}\right)\right)\right) \\ &\quad + B_L\left(\tilde{\epsilon}, \epsilon^{'''}\right), F_L\left(\tilde{\epsilon}, \epsilon^{''}\right)\}, \end{aligned}$$

(iii) es existiert eine Konstante $C_M^{\mathbf{u}} = C_M^{\mathbf{u}}(C_{coer}, C_{cont}, C_{rhs}) > 0$, so dass

$$\begin{aligned} M\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}\right) \leq\ & C_M^{\mathbf{u}} d\left(\tilde{\epsilon}\right)^2 \\ & \cdot \left(d\left(\tilde{\epsilon}\right) log_2\left(\frac{1}{\epsilon}\right) log_2^2\left(log_2\left(\frac{1}{\epsilon}\right)\right)\left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right)\right.\right. \\ & + log_2\left(d\left(\tilde{\epsilon}\right)\right) .. + B_L\left(\tilde{\epsilon}, \epsilon^{'''}\right) \\ & \left.\left. + F_L\left(\tilde{\epsilon},\ \epsilon^{''}\right)\right) + 2B_M\left(\tilde{\epsilon}, \epsilon^{'''}\right) + F_M\left(\tilde{\epsilon},\ \epsilon^{''}\right), \end{aligned}$$

(iv) $M\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}\right) \leq 2Dd(\tilde{\epsilon}) + 2M\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}\right)$,

(v) $\sup\limits_{y\in\mathcal{Y}}\left|R_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}\right)(y)\right| \leq \mathcal{K}^2 + \frac{\epsilon}{3}$, *und* $\sup\limits_{y\in\mathcal{Y}}\left|R_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}\right)(y)\right|_{\mathbf{G}} \leq \mathcal{K}^2 + \frac{\epsilon}{3}$“[25, Theorem 4.3].

Bemerkung 4.4. „Im Beweis zu Theorem 4.3 konstruieren wir ein NN $\Phi_{inv;\tilde{\epsilon},\epsilon}^{\mathbf{B}}$, dessen ϵ $(\mathbf{B}_y^{rb})^{-1}$ approximiert. Dann können die NNs aus Theorem 4.3 explizit wie folgt konstruiert werden:

$$\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb} := \Phi_{mult;\frac{\epsilon}{3}}^{\mathcal{K},d(\tilde{\epsilon}),d(\tilde{\epsilon}),1} \odot P\left(\Phi_{inv;\tilde{\epsilon},\epsilon'}^{\mathbf{B}}, \Phi_{inv;\tilde{\epsilon},\epsilon''}^{\mathbf{f}}\right) \bullet \left(\left(\begin{pmatrix}\mathbf{Id}_{\mathbb{R}^p}\\ \mathbf{Id}_{\mathbb{R}^p}\end{pmatrix}, \mathbf{0}_{\mathbb{R}^{2p}}\right)\right)$$
$$und\ \Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h} := \left(\left(\mathbf{V}_{\tilde{\epsilon}}, \mathbf{0}_{\mathbb{R}^D}\right)\right) \odot \Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}\text{“}$$

[25, Bemerkung 4.4].

Bemerkung 4.5. „Im genannten Beweis lässt sich, sehen dass besonders in den letzten beiden Abschätzungen die Konstanten $C_L^{\mathbf{u}}, C_M^{\mathbf{u}}$ wie folgt abhängig von den Konstanten $C_{coer}, C_{cont}, C_{rhs}$ sind (zur Erinnerung: $\frac{C_{coer}}{2C_{cont}} \leq \delta = \frac{C_{coer}}{C_{coer}+C_{cont}} \leq \frac{1}{2}$)

- $C_L^{\mathbf{u}}$ ist affin linear abhängig auf

$$\log_2^2\left(\frac{\log_2\left(\frac{\delta}{2}\right)}{\log_2(1-\delta)}\right),\ log_2\left(\frac{1}{C_{coer}+C_{cont}}\right),\ log_2\left(max\left\{1, C_{rhs}, \frac{1}{C_{coer}}\right\}\right)$$

- $C_M^{\mathbf{u}}$ ist affin linear abhängig auf

$$\log_2\left(\frac{1}{C_{coer}+C_{cont}}\right), \frac{\log_2\left(\frac{\delta}{2}\right)}{\log_2(1-\delta)} \cdot \log_2^3\left(\frac{\log_2\left(\frac{\delta}{2}\right)}{\log_2(1-\delta)}\right), \log_2\left(max\left\{, C_{rhs}, \frac{1}{C_{coer}}\right\}\right) \text{“}$$

[25, Bemerkung 4.5].

Bemerkung 4.6. Mit Hilfe von Theorem 4.3 die Annahme getätigt werden, dass es 2 NNs von angemessener Größe gibt, deren implizierten Funktionen eine Approximation für die folgenden diskretisierten Lösungsabbildungen darstellen:

$$\mathcal{Y} \to \mathbb{R}^D : y \mapsto \tilde{\mathbf{u}}^h_{y,\tilde{\epsilon}}, \ und \ \mathcal{Y} \to \mathbb{R}^{d(\tilde{\epsilon})} : y \mapsto \tilde{\mathbf{u}}^{rb}_{y,\tilde{\epsilon}}. \tag{4.1}$$

Nun kann auch die Approximation einer parametirisierten Lösung der PDE in Form einer Abbildung dargestellt werden. Das bedeutet, dass die Abbildung $\mathcal{Y}\mathbf{x}\Omega \to \mathbb{R} : (y,x) \mapsto u_y(x)$ auf genau den Raum abbildet, für den unsere PDE definiert ist. Jetzt ist bekannt, dass sich mit den durch die Realisierungen von NNs implizierten Funktionen sowohl Elemente der reduzierten Basis als auch die der High-Fidelity Basis approximieren lassen.

$$u_y(x) \approx \sum_{i=1}^{d(\tilde{\epsilon})} (\mathbf{u}^{rb}_{y,\tilde{\epsilon}})_i \psi_i(x) = \sum_{i=1}^{D} (\tilde{\mathbf{u}}^h_{y,\tilde{\epsilon}})_i \phi_i(x)$$

Tatsächlich kann für die Approximationen der beiden Gleichungen aus (4.1) in etwa der gleiche Aufwand betreiben werden wie für die Approximation der Abbildung $(y,x) \mapsto u_y(x)$. Mithilfe der implizierten Funktionen ist es möglich, viele der Basiselemente zu approximieren.

5 Diskussion: Abhängigkeit der Approximationsraten von den genannten Dimensionen

„In diesem Abschnitt werden wir unsere Ergebnisse im Hinblick auf die Abhängigkeit von den beteiligten Dimensionen diskutieren. Wir möchten betonen, dass sich die resultierenden Approximationsraten (die sich aus Theorem 4.3 ableiten lassen) deutlich von alternativen Ansätzen unterscheiden und oft wesentlich besser sind als diese. Wie in Abschnitt 2 beschrieben, gibt es drei zentrale Dimensionen, die die Härte des Problems beschreiben. Das sind die Dimensionen D des HF Diskretisierungsraums U^h, die Dimension $d(\tilde{\epsilon})$ des reduzierten Basisraums und die Dimension p vom Parameterraum $\mathcal{Y}$“[25, Kapitel 5].

Abhängigkeit von D: Grundlegend können Approximationsraten ermittelt werden, welche höchstens linear abhängig von D sind. Allerdings ist diese nicht an die Abhängigkeit von ϵ geknüpft.Es kann versucht werden, direkt die Lösung der linearen Systeme zu berechnen. Wenn lediglich die spärlichen Eigenschaften einer Matrix verwendet werden, erhält man eine Komplexität der Approximationsrate der Ordnung $\mathcal{O}(D^3)$ und hinzu käme noch der Aufwand zur Erstellung einer Steifigkeitsmatrix. Tatsächlich kann behauptet werden, dass die verwendeten Approximationsraten aus Theorem 4.3 deutlich besser sind, vorausgesetzt $D \gg d(\tilde{\epsilon})$.

Abhängigkeit von $d(\tilde{\epsilon})$: Angenommen, es will eine bereits gefundene reduzierte Basis mit dem Galerkin-Schema gelöst werden. In der Regel werden hierfür ein Rechenaufwand der Ordnung $\mathcal{O}\left(d(\tilde{\epsilon})^3\right)$ plus des Aufwands zum Finden der reduzierten Basis benötigt. Bekannt ist, dass es möglich ist, NNs zu konstruieren, die dieses Verfahren simulieren können und im Wesentlichen die gleiche Approximationsrate mit einem Faktor $C(\epsilon)$ haben, welcher polylogarithmysch auf der Approximationsgenauigkeit ϵ beruht.
Wichtig ist hier, dass die NNs vielseitiger einsetzbar sind als das Galerkin-Schema und je nach Komplexität der Rechnung immer vorteilhafter werden im Vergleich zu anderen Methoden. Außerdem können die NNs auch verwendet werden, wenn die relevante PDE gänzlich unbekannt ist, vorausgesetzt, es stehen genug Snapshots zur Verfügung.

Abhängigkeit von p: Wenn die Ergebnisse mit den naiven verglichen werden, sind

diese deutlich besser. Bei dem naiven Ansatz ist es notwendig, dass gewisse Glattheitseigenschaften für die Abbildung $y \mapsto \tilde{\mathbf{u}}^h_{\cdot,\tilde{\epsilon}}$ gelten, bei dem Ansatz, den wir gewählt haben, ist dies keine Voraussetzung, und es sind trotzdem die Annahmen 4.1 und 4.1 erfüllt. In diesem Fall ist es möglich eine Approximation bis zu einem Fehler ϵ für NNs mit beschränkter Anzahl von Gewichten zu bekommen. Wie auch zuvor ist die Abhängigkeit von D linear, dieses Mal ist sie aber gekoppelt mit einem potentiell schnell wachsendem Term $\epsilon^{\frac{-p}{n}}$.

Abschließend ist klar, dass diese Methode gegenüber allen anderen gegenüber mehr Vorteile mit sich bringen, da entweder genauere Approximationen ausgegeben werden, oder, wenn die Genauigkeiten gleich sind, der Aufwand geringer ist, da weniger Vorarbeit geleistet werden muss und vieles automatisch läuft.

6 Beweise

Bis auf die Beweise von Proposition 3.7, Theorem 3.8 und Theorem 4.3, welche ich durch zusätzliche Rechnungen und Erklärungen vereinfacht habe, sind alle Beweise von mir selbst aus anderen Artikeln hinzugefügt und falls notwendig auch mit zusätzlichen Zwischenschritten ausgestattet. Das soll das Verständnis für die Beweise erleichtern und dementsprechend auch die generellen Aussagen und das Endergebnis des ursprünglichen Artikels.

6.1 Beweise des ersten Abschnitts

Beweis Proposition 2 aus [4]:
Zuerst die Poposition.
Die Funktion $f(x) = x^2$ auf dem Intervall [0,1] kann durch ein ReLU-Netzwerk mit einer Tiefe und Anzahl von Gewichten und Recheneinheiten der Ordnung $\mathcal{O}\left(ln\left(\frac{1}{\epsilon}\right)\right)$ abgeschätzt werden und das für jeden Fehler $\epsilon > 0$.
Beweis Betrachte die Sägezahnfunktion (wie bereits in 1.3.2) $g : [0.1] \to [0, 1]$, für g gilt dann wie schon zuvor

$$g(x) = \begin{cases} 2x & , x < \frac{1}{2} \\ 2(1-x) & , x \geq \frac{1}{2} \end{cases}$$

und die iterierten Funktionen

$$g_s(x) = g \circ ... \circ g(x).$$

Mit einem weiteren Beweis von Telgarsky kann gezeigt werden, dass diese Funktion 2^{s-1} gleichmäßig verteilte Zähne hat. Jede neue Funktion hat doppelt so viele Zähne wie die vorherige:

$$g_s(x) = \begin{cases} 2^s\left(x - \frac{2k}{2^s}\right), & x \in \left[\frac{2k}{2^s}, \frac{2k+1}{2^s}\right], k = 0, ..., 2^{s-1} - 1, \\ 2^s\left(\frac{2k}{2^s} - x\right), & x \in \left[\frac{2k-1}{2^s}, \frac{2k}{2^s}\right], k = 1, ..., 2^{s-1}, \end{cases}$$

also hat g die Spitze seines „Zahns“ in [0,1] bei 0,5, g_1 hat 2 Zähne bei 0,25 und 0,75 usw. Die Zähne der Funktionen g_s lassen sich sehr gut mit zweidimensionalen Vektoren darstellen, da sie zwischen ihren Hoch- und Tiefpunkten gradlinig verlaufen. Das lässt vermuten, dass die Quadratfunktion $f(x) = x^2$ durch Linearkombinationen

der einzelnen Funktionen approximiert werden kann. Wird angenommen, dass f_m eine stückweise lineare Interpolation von f ist und dementsprechend $2^m + 1$ gleichmäßig verteilte Stützpunkte $\frac{k}{2^m}, K = 0, ..., 2^m$ hat, gilt:

$$f_m\left(\frac{k}{2^m}\right) = \left(\frac{k}{2^m}\right)^2, \quad k = 0, ..., 2^m.$$

Diese Funktionen approximieren f mit einem Fehler $\epsilon_m = 2^{-2m-2}$. Nun ist zu beachten, dass die Verfeinerung der Interpolation von f_m nach f_{m-1} einer Approximation durch eine Funktion proportional zu einer Sägezahnfunktion gleichkommt:

$$f_{m-1}(x) - f_m(x) = \frac{g_m(x)}{2^{2m}}.$$

Somit

$$f_m(x) = x - \sum_{s=1}^{m} \frac{g_s(x)}{2^{2s}}.$$

Bekannt ist bereits, dass Funktionen mit Hilfe von endlichen ReLU Netzwerken implementiert werden können. Anhand der Argumente in diesem Beweis ist zu sehen, dass für die Konstruktion von f_m nur $\mathcal{O}(m)$ lineare Operationen und Kompositionen von g benötigt werden. Das wiederum zeigt, dass f_m, welches aus g besteht, durch ein ReLU Netzwerk mit einer Tiefe, Anzahl von Gewichten und Recheneinheiten der Ordnung m implementiert werden kann. Und aus der Erkenntnis folgt dann die Behauptung von oben und der Beweis ist beendet. □

6.2 Beweise des zweiten Abschnitts

Beweis Gleichung (2.6): Sei V_N der optimale Kolmogorov-Raum für X mit Dimension n. Sei nun $\phi_1, ..., \phi_n$ eine Orthonormalbasis für V_N und sei P die Projektion auf V_N. $x_0, x_1, ...$ ist eine unendliche Reihe, wobei definiert ist, dass $x_m = 0$ für $m > N$. Darum gilt für jedes $x \in X$ mit dem Gram-Schmidt Orthogonalisierungsverfahren

$$Px = \sum_{j=1}^{n} \langle x, \phi_j \rangle \phi_j,$$

und $\|x - Px\| \leq W_N(X), x \in X$, gibt die kürzeste Strecke zwischen x und dem Kolmogorov-Raum an. Für jedes $\{x_1, ..., x_n\} \subset X$ beschreibt man die Determinan-

te mit $D(x_1, ..., x_n) := det(\langle x_i, \phi_j \rangle)$. Wähle $x_1, ..., x_n$ so, dass sie den absoluten Betrag der Determinante maximieren. Jetzt kann $Px = \sum_{i=1}^{n} \alpha_i P x_i$ für jedes $x \in X$ so geschrieben werden, da $\alpha_i = \langle x, \Phi_j \rangle$ und weil sich P als eine Linearkombination der Vektoren aus unserer ONB darstellen lässt. Das α kann dementsprechend wie folgt dargestellt werden

$$\alpha_i = \frac{D(Px_1, ..., Px_{i-1}, Px, Px_{i+1}, ..., Px_n)}{D(Px_1, ..., Px_n)} = \frac{D(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_n)}{D(x_1, ..., x_n)}.$$

Da P ein orthogonaler Projektor ist (P projiziert x in einem 90 Grad Winkel auf V_N, also gibt dies die minimale Distanz zwischen x und V_N an). Da α_i die Einträge einer unteren Dreiecksmatrix darstellt, gibt einem mit der Definition von $x_1, ..., x_n$, dass $|\alpha_i| \leq 1$. Da die Basisvektoren jeweils nur in der j-ten Zeile einen Eintrag eins haben und alle anderen Einträge 0 sind. Nun kann also $\|x - Px\| \leq W_N(X)$ umgeschrieben werden zu

$$x - \sum_{i=1}^{n} \alpha_i x_i = x - Px + \sum_{i=1}^{n} \alpha_i \left[Px_i - x_i \right].$$

Daraus folgt

$$\| \, x - \sum_{i=1}^{n} \alpha_i x_i \, \| \leq (n+1) W_N(X),$$

wie in der Behauptung [15, Theorem 4.1]. □

Beweis $S(\mathcal{Y})$ ist kompakt (Proposition 5.1 aus [13]): Sei $b(\cdot, \cdot; y)$ eine Bilinearform und $f(\cdot; y)$ eine Linearform und beide Lipschitzstetig bezüglich y. Dann ist auch die Lösung $u(y)$ von (2.1) Lipschitzstetig, es existiert also ein $L_u > 0$, so dass

$$\left\| u(y) - u(y') \right\| \;\leq\; L_u \left\| y - y' \right\| \quad \forall \; y, y' \in \mathcal{Y}$$

ist.

Beweis: Setze zu erst $u = u(y)$ und $u' = u(y')$, dann gilt

$$b(u, v; y) = f(v; y), \;\; b(u', v; y') = f(v; y') \;\; \forall \; v \in V$$

Ziehe nun diese beiden Terme voneinander ab und addiere 0, ergibt das zusammen mit

Definition 5.1 aus [13]

$$b(u,v;y) - b(u^{'},v;y) + b(u^{'},v;y) - b(u',v;y^{'}) = f(v;y) - f(v;y^{'})$$

nun gibt uns ([13], Def. 5.1)

$$\begin{aligned} |b(u,v;y) - b(u,v;y^{'})| \;&\leq\; L_b \left\|u\right\|_V \left\|v\right\|_V \left\|y - y^{'}\right\| \quad \forall\, y, y^{'} \in \mathcal{Y},\; u, v \in V \\ |f(v;y) - f(v;y^{'})| &\leq L_f \left\|v\right\|_V \left\|y - y^{'}\right\| \quad \forall\, y, y^{'} \in \mathcal{Y},\; v \in V \end{aligned}$$

womit sich folgende Abschätzung tätigen lässt

$$b(u - u^{'},v;y) \leq L_f \left\|v\right\|_V \left\|y - y^{'}\right\| + L_b \left\|u^{'}\right\|_V \left\|v\right\|_V \left\|y - y^{'}\right\|$$

wählt man jetzt $v = u - u^{'}$ bekommt man mit ([13], (2.5) $\rightarrow b(v,v) \;\neq\; \alpha \left\|v\right\|_V^2$)

$$\beta \left\|v\right\|_V^2 \leq L_f \left\|v\right\|_V \left\|y - y^{'}\right\| + L_b \left\|u^{'}\right\|_V \left\|v\right\|_V \left\|y - y^{'}\right\|$$

jetzt dividiere durch $\left\|v\right\|_V$:

$$\beta \left\|v\right\|_V \;\leq\; L_f \left\|y - y^{'}\right\| \;+\; L_b \left\|u^{'}\right\|_V \left\|y - y^{'}\right\|.$$

Nun wird die Stabilitätsabschätzung ([13], (3.10)) verwendet und es ergibt sich

$$\beta(y) \left\|u - u^{'}\right\|_V \leq L_f(y) \left\|y - y^{'}\right\| + L_b \frac{\left\|f(y^{'})\right\|_{V'}}{\beta_0(y^{'})} \left\|y - y^{'}\right\|.$$

Zum Schluss wird $L_u = \frac{1}{\beta_0}\left(L_f + L_b \frac{\bar{\gamma}_F}{\beta_0}\right)$ gesetzt und das gibt uns die Behauptung. □

6.3 Beweise des dritten Abschnitts

6.3.1 Beweis Lemma 3.6:

(a) Für die Behauptung definiere zu erst $i \in \{1,2\}, L_i \in \mathbb{N}, N_1^i, ..., N_{L_i}^i, \left(A_l^i, b_l^i\right) \in \mathbb{R}^{N_l^i \mathtt{x} N_{l-1}^i \mathtt{x} \mathbb{R}^{N_{L_i}}}, l \in \{1, ...L_i\}$, so dass $\Phi^i = \left(\left(A_1^i, b_1^i\right), ..., \left(A_{L_i}^i, b_{L_i}^i\right)\right)$. Außerdem seien $(A_l, b_l) \in \mathbb{R}^{N_l \mathtt{x} N_{l-1} \mathtt{x} \mathbb{R}^{N_L}}, l \in \{1, ..., L_1 + L_2\}$ die Matrix-Vektor Tupel, welche $\Phi_1 \odot \Phi_2 = \left(\left(A_1, b_1\right), ..., \left(A_{L_1+L_2}, b_{L_1+L_2}\right)\right)$ erfüllen und seien $r_l : \mathbb{R}^{N_0} \rightarrow \mathbb{R}^{N_l}, l \in \{1, ..., L_1 + L_2\}$

die Funktionen, für die für jedes $x \in \mathbb{R}^{N_0}$ gilt

$$r_l(x) = \begin{cases} \varrho^* \left(A_1 x + b_1\right) & : l = 1 \\ \varrho^* \left(A_l r_{l-1}(x) + b_l\right) & : 1 < l < L_1 + L_2 \\ \left(A_l r_{l-1}(x) + b_l\right) & : l = L_1 + L_2. \end{cases}$$

Zu beobachten ist, dass für alle $l \in \{1, ..., L_2 - 1\}, (A_l, b_l) = \left(A_l^2, b_l^2\right)$ gilt. Das Impliziert für jedes $x \in \mathbb{R}^{N_0}$

$$A_{L_2}^2 r_{L_2-1}(x) + b_{L_2}^2 = \left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x)$$

Da $1 < l < L_1 + L_2$, gibt das in Kombination mit (5.7) [17]

$$r_{L_2}(x) = \varrho^* \left(A_{L_2} r_{L_2-1}(x) + b_{L_2}\right) = \varrho^* \left(\begin{pmatrix} A_{L_2}^2 \\ -A_{L_2}^2 \end{pmatrix} r_{L_2-1}(x) + \begin{pmatrix} b_{L_2}^2 \\ -b_{L_2}^2 \end{pmatrix}\right)$$

und mit $l \in \{1, ..., L_2 - 1\}$, $(A_l, b_l) = \left(A_l^2, b_l^2\right)$ und der Gleichung zuvor gibt das

$$= \varrho^* \left(\begin{pmatrix} A_{L_2}^2 r_{L_2-1}(x) + b_{L_2}^2 \\ -A_{L_2}^2 r_{L_2-1}(x) + b_{L_2}^2 \end{pmatrix}\right) = \begin{pmatrix} \varrho^* \left(\left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x)\right) \\ -\varrho^* \left(\left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x)\right) \end{pmatrix}.$$

Zusätzlich gilt für jedes $d \in \mathbb{N}, y \in \{y_1, ..., y_d\} \in \mathbb{R}^d$

$$\varrho^*(y) - \varrho^*(-y) = (\varrho(y_1) - \varrho(-y_1), ..., \varrho(y_d) - \varrho(-y_d)) = y.$$

Das, (5.7) und (5.16) [17] ergibt dann, dass für $x \in \mathbb{R}^{N_0}$ gilt

$$\begin{aligned} r_{L_2+1}(x) &= A_{L_2+1} \begin{pmatrix} \varrho^* \left(\left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x)\right) \\ -\varrho^* \left(\left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x)\right) \end{pmatrix} + b_{L_2+1} \\ &= A_1^1 \varrho^* \left(\left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x)\right) - A_1^1 \varrho^* \left(-\left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x)\right) + b_{L_2+1} \\ &= A_1^1 \left[\mathcal{R}_\varrho \left(\Phi_2\right)\right](x) + b_1^1. \end{aligned}$$

Und das ist wiederum mit Definition 3.2 und da die Anzahl der Schichten $< L_1 + L_2 - 1$ (5.14)[17] ist, die Verkettung der beiden NNs und damit folgt die Behauptung.
Um (i) dieses Lemmas zu zeigen, benutzen man nun ([17], (5.7)). Sei also

$$\Phi^i = \left(\left(A_1^i, b_1^i\right), ..., \left(A_{L_i}^i, b_{L_i}^i\right)\right)$$

und daraus resultiert

$$\Phi^1 \odot \Phi^2 = \left(\left(A_1^1, b_1^1\right), ..., \left(A_{L_1}^i, b_{L_1}^i\right) \odot \left(A_1^2, b_1^2\right), ..., \left(A_{L_2}^2, b_{L_2}^2\right)\right)$$

nach ([17], (5.7)) ist dies das gleiche wie

$$\Phi^1 \odot \Phi^2 = \left(\begin{array}{c}\left(A_1^2, b_1^2\right), ..., \left(A_{L_2-1}^2, b_{L_2-1}^2\right), \left(\begin{pmatrix} A_{L_2}^2 \\ -A_{L_2}^2 \end{pmatrix}, \begin{pmatrix} b_{L_2}^2 \\ -b_{L_2}^2 \end{pmatrix}\right), \\ \left(\left(A_1^1 \ -A_1^1\right), b_1^1\right), ..., \left(A_{L_1}^1, b_{L_1}^1\right)\end{array}\right)$$

und damit folgt

$$\mathcal{L}\left(\Phi^1 \odot \Phi^2\right) \leq \mathcal{L}\left(\Phi^1\right) \ + \ \mathcal{L}\left(\Phi^2\right)$$

(ii), (iii) und (iv) werden mit der gleichen Formel bewiesen. Da für $M(.) \geq 0$ für alle NNs, gilt $M\left(\Phi^1 \odot \Phi^2\right) \leq M\left(\Phi^1\right) + M\left(\Phi^2\right)$. Wird nun [17] (5.7) verwendet, wird deutlich, dass $M_1(\Phi^1)$ und $M_{L(\Phi^2)}(\Phi^2) = 0$ sind. Damit wäre (ii) bewiesen. (iii) folgt direkt aus der Anordnung bei einer Verkettung von NNs nach [17] (5.7). (iv) folgt genau so aus ([17], (5.7)). □

(b) Zuerst beweisen wir die Behauptung

$$\mathcal{R}_\varrho\left(\mathcal{E}_{\mathbb{L}(\Phi)}\left(\phi^j\right)\right) = \mathcal{R}_\varrho\left(\phi^j\right).$$

Es lässt sich schreiben

$$\mathcal{R}_\varrho\left(\mathcal{E}_{\mathbb{L}(\Phi)}\left(\phi^j\right)\right) = \mathcal{R}_\varrho\left(P\left(\mathcal{E}_{\mathbb{L}(\Phi)}\left(\phi^1\right), ..., \mathcal{E}_{\mathbb{L}(\Phi)}\left(\phi^n\right)\right)\right)$$

und erhält damit, wenn (5.10) aus [17] und die Voraussetzung von Φ genutzt werden,

$$\mathcal{R}_\varrho\left(\mathcal{E}_{\mathbb{L}(\Phi)}\left(\phi^j\right)\right) = \mathcal{R}_\varrho\left(P\left(\mathcal{E}_{\mathbb{L}(\Phi)}\left(\phi^1\right), ..., \mathcal{E}_{\mathbb{L}(\Phi)}\left(\phi^n\right)\right)\right).$$

Für den nächsten Schritt wird (5.12) aus [17] verwendet und dann gilt mit 3.6(b)(iii) die Behauptung

$$= \mathcal{R}_\varrho\left(P\left(\phi^1, ..., \phi^n\right)\right) = \mathcal{R}_\varrho\left(\phi^j\right).$$

(i) gilt, da durch die Parallelisierung nicht die Anzahl der Schichten der einzelnen NNs beeinflusst wird. Dementsprechend entspricht die Anzahl der Schichten dieser Parallelisierung, der Anzahl der Schichten des NNs, welches die meisten Schichten hat. (iv) folgt direkt aus, dass keine Gewichte in der Parallelisierung beziehungsweise durch die Interaktion der einzelnen NNs verloren gehen. Darum hat die resultierende erste Schicht der Parallelisierung alle Gewichte der ersten Schichten der einzelnen NNs. Außerdem zeigt (3.1), dass für jedes $m \in \mathbb{N}, \psi_i \in \mathbb{R}, i \in \{1, ..., m\}$ mit $\forall \quad i, i^{'} \in \{1, ..., m\} : \ \mathcal{L}\left(\Phi^i\right) = \mathcal{L}\left(\Phi^{i^{'}}\right)$ gilt:

$$M\left(P\left(\Phi^1, ..., \Phi^k\right)\right) = \sum_{i=1}^{k} M\left(\Phi^i\right), \tag{6.1}$$

sofern für zwei verschiedene beliebige NNs gilt, dass sie gleich viele Schichten haben. Das beweist (iii) und (vi). Wenn nun (3.2) benutzt wird und $d, L \in \mathbb{N}$ so gewählt ist, so dass $M\left(\Phi_{d,L}^{Id}\right) \leq 2dL$. Das impliziert, dass für jedes $j \in \{1, ..., n\}$ gilt

$$\begin{aligned} M\left(\mathcal{E}_{L(\Phi)}\right)\left(\varphi^j\right) &\leq 2M\left(\Phi_{dim_{out}(\varphi^j), \mathcal{L}(\Phi) - \mathcal{L}(\varphi^j)}\right) + 2M\left(\varphi^j\right) \\ &\leq 4dim_{out}\left(\varphi^j\right) \mathcal{L}\left(\Phi + 2M\left(\varphi^j\right)\right). \end{aligned}$$

Wende darauf jetzt (6.2) an und beachte, dass $\Phi = P\left(\varphi^1, ..., \varphi^n\right)$, zusätzlich (i) und (3.2) genutzt werden, dann gilt (ii). Zusätzlich resultiert mit ([17], (5.8), (5.9)) und (3.2), dass für jedes $j \in \{1, ..., n\}$ gilt

$$M_{\mathcal{L}(\Phi)}\left(\mathcal{E}_{\mathcal{L}(\Phi)}\left(\varphi^j\right)\right) \ \leq \ max\{2dim_{out}\left(\varphi^j\right), M_{\mathcal{L}(\varphi^j)}\left(\varphi^j\right)\}.$$

Wird dies nun mit (3.1) kombiniert, beweist das (v). □

6.3.2 Beweis Proposition 3.7:

„Zu erst beweisen wir einen Spezialfall in welchem $M\left(\Phi^1 \bullet \Phi^2\right)$ durch $max\{M\left(\Phi^1\right), M\left(\Phi^2\right)\}$ abgeschätzt werden kann.

Lemma A.1. Sei Φ ein NN mit $m - dimensionalem$ Output und $d - dimnsionalem$

Input. Sei $\mathbf{a} \in \mathbb{R}^{1\mathtt{x}m}$, dann gilt für alle $l = 1, ..., L(\Phi)$,

$$M_l((\mathbf{a}, 0) \bullet \Phi) \leq M_l(\Phi).$$

Genauer gesagt gilt $M((\mathbf{a}, 0) \bullet \Phi) \leq M(\Phi)$, außerdem falls $\mathbf{D} \in \mathbb{R}^{d\mathtt{x}n}$, so dass für jedes $k \leq d$ es höchstens ein $l_k \leq n$ gibt mit $\mathbf{D}_{k,l_k} \neq 0$, dann gilt für alle $l = 1, ..., L(\Phi)$,

$$M_l(\Phi \bullet (\mathbf{D}, 0_{\mathbb{R}^d})) \leq M_l(\Phi).$$

Auch hier impliziert dies wieder $M_l(\Phi \bullet (\mathbf{D}, 0_{\mathbb{R}^d})) \leq M(\Phi)$.

Zum Beweis. Sei $\Phi = ((\mathbf{A}_1, \mathbf{b}_1), ..., (\mathbf{A}_L, \mathbf{b}_L))$, und $\mathbf{a}$, $\mathbf{D}$ wie im Lemma A.1. Dann bekommen wir die Behauptungen des Lemmas, falls gilt

$$\|\mathbf{a}\mathbf{A}_L\|_0 + \|\mathbf{a}\mathbf{b}\|_L \leq \|\mathbf{A}_L\|_0 + \|\mathbf{b}_L\|_0$$

und

$$\|\mathbf{A}_1\mathbf{D}\|_0 \leq \|\mathbf{A}_1\|_0 .$$

Klar ist, dass $\|\mathbf{a}\mathbf{A}_L\|_0$ kleiner ist als die Anzahl der Nicht-Null Spalten von $\mathbf{A}_L$, welche durch die Norm $\|\mathbf{A}_L\|_0$ beschränkt ist. Dasselbe Argument kann man für $\|\mathbf{a}\mathbf{b}\|_L \leq \|\mathbf{b}_L\|_0$ verwenden und erhält damit die erste Behauptung.
Jetzt gilt, dass für zwei Vektoren $\mathbf{p}, \mathbf{q} \in \mathbb{R}^k, k \in \mathbb{N}$ und für alle $\mu, \nu \in \mathbb{R}$:

$$\|\mu\mathbf{p} + \nu\mathbf{p}\|_0 \leq I(\mu) \|\mathbf{p}\|_0 + I(nu) \|\mathbf{q}\|_0 ,$$

wo $I(\gamma) = 0 \; falls \; \gamma = 0 \; und \; = 1$ sonst. Auch gilt natürlich

$$\|\mathbf{A}_1\mathbf{D}\|_0 = \left\|\mathbf{D}^T\mathbf{A}_1^T\right\|_0 = \sum_{l=1}^{n} \left\|\left(\mathbf{D}^T\mathbf{A}_1^T\right)_{l,-}\right\|_0 ,$$

wobei das $l, -$ die l-te Reihe bedeutet. Jetzt kann man für alle $l \leq n$ noch sagen, dass

$$\left(\mathbf{D}^T\mathbf{A}_1^T\right)_{l,-} = \sum_{k=1}^{d} \left(\mathbf{D}^T\right)_{l,k} \left(\mathbf{A}_1^T\right)_{k,-} = \sum_{k=1}^{d} \mathbf{D}_{k,l} \left(\mathbf{A}_1^T\right)_{k,-} .$$

Wenn wir nun die beiden letzten beiden Aussagen kombinieren und dann noch unsere Definition für I verwenden, können wir folgern

$$\begin{aligned}\|\mathbf{A}_1\mathbf{D}\|_0 &\leq \sum_{l=1}^{n}\left\|\sum_{k=1}^{d}\mathbf{D}_{k,l}\left(\mathbf{A}_1^T\right)_{k,-}\right\|_0 \leq \sum_{l=1}^{n}\sum_{k=1}^{d} I\left(\mathbf{D}_{k,l}\right)\left\|\left(\mathbf{A}_1^T\right)_{k,-}\right\|_0 \\ &= \sum_{k=1}^{d} I\left(\mathbf{D}_{k,l_k}\right)\left\|\left(\mathbf{A}_1^T\right)_{k,-}\right\|_0 \leq \|\mathbf{A}_1\|_0 \text{ “}\end{aligned}$$

[25, Lemma A.1]. □

Jetzt kann Proposition 3.7 bewiesen werden.

Ohne Verlust der Allgemeinheit wird angenommen, dass $Z \geq 1$. Wird [17, Lemma 6.2] genutzt, existiert ein NN $\mathtt{x}_\epsilon^Z$ mit Inputdimension 2 und Outputdimension 1, so dass für $\Phi_\epsilon := \mathtt{x}_\epsilon^Z$

$$L\left(\Phi_\epsilon\right) \leq 0.5 log_2\left(\frac{n\sqrt{dl}}{\epsilon}\right) + log_2\left(Z\right) + 6, \tag{6.2}$$

$$M\left(\Phi_\epsilon\right) \leq 90 \cdot \left(log_2\left(\frac{n\sqrt{dl}}{\epsilon}\right) + 2log_2\left(Z\right) + 6\right), \tag{6.3}$$

$$M_1\left(\Phi_\epsilon\right) \leq 16, \text{ wie auch } M_{L(\Phi_\epsilon)}\left(\Phi_\epsilon\right) \leq 2, \tag{6.4}$$

$$\sup_{|a|,|b|\leq Z}\left|ab - R_\varrho^{\mathbb{R}^2}\left(\Phi_\epsilon\right)(a,b)\right| \leq \frac{\epsilon}{n\sqrt{dl}}. \tag{6.5}$$

Beweis dieser vier Gleichungen: Dazu wird Lemma 6.2 [17] bewiesen. Für diesen Beweis wird die Umgebung von Annahme 5.2 [17] angenommen. Definiere eine Menge $\mathcal{N}^{N_0,\ldots,N_L} : \mathtt{x}_{l=1}^{L}\left(\mathcal{R}^{N_l}\mathtt{x}\mathcal{R}^{N_{l-1}}\mathtt{x}\mathcal{R}^{N_l}\right)$ und $R = \cup_{\substack{L\in\mathbb{N}\\ N_0,\ldots,N_L}} \mathcal{N}_L^{N_0,\ldots,N_L}, \Phi^i \in R$. Sei $\Theta \in \mathcal{N}_1^{1,1}$, das neurale Netz (0,0) und sei $\alpha \in \mathcal{N}_2^{2,6,3}$ das neurale Netz gegeben durch

$$\alpha_1 = \left(\left(\left(\begin{pmatrix}1&1\\-1&-1\\1&0\\-1&0\\0&1\\0&-1\end{pmatrix}, \begin{pmatrix}0\\0\\0\\0\\0\\0\end{pmatrix}\right), \left(\frac{1}{2B}\begin{pmatrix}1&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&1&1\end{pmatrix}, \begin{pmatrix}0\\0\\0\end{pmatrix}\right)\right),$$

und sei $\sum \in \mathcal{N}_1^{3,1}$ das neurale Netz gegeben durch $\sum = \left(\left(\begin{pmatrix}2B^2 & -2B^2 & -2B^2\end{pmatrix}, 0\right)\right)$. Sei mit Lemma 6.1 ein NN $(\sigma)_{\epsilon\in(0,\infty)} \subset R$, welches (i)-(iv) dieses Lemmas erfüllt. Sei

nun $(\mu)_{\epsilon\in(0,\infty)} \subset R$ das neurale Netz, für das gilt

$$\mu_\epsilon = \begin{cases} \sum \odot \mathcal{P}\left(\sigma_{\frac{\epsilon}{6B^2}}, \sigma_{\frac{\epsilon}{6B^2}}, \sigma_{\frac{\epsilon}{6B^2}}\right) \odot \alpha & : \epsilon < B^2 \\ \Theta & : \epsilon \geq B^2. \end{cases}$$

Jetzt lässt sich mit Lemma 6.1 und der Realisierung des neuralen Netzes $\sum$ (wenn wir das nach den jeweiligen Komponenten machen) sagen, dass für jedes $\epsilon \in (0,\infty)$ gilt

$$\begin{aligned} &\sup_{z\in[-2B,2B]} \left| \frac{1}{2}z^2 - 2B^2 \left[\left[\mathcal{R}_\varrho \left(\sigma_{\frac{\epsilon}{6B^2}} \right) \right] \left(\frac{|z|}{2B} \right) \right] \right| \\ &= \sup_{z\in[-2B,2B]} \left| 2B^2 \left[\frac{|z|}{2B} \right]^2 - 2B^2 \left[\left[\mathcal{R}_\varrho \left(\sigma_{\frac{\epsilon}{6B^2}} \right) \right] \left(\frac{|z|}{2B} \right) \right] \right| \\ &= 2B^2 \left[\sup_{t\in[0,1]} \left| t^2 - \left[\left[\mathcal{R}_\varrho \left(\sigma_{\frac{\epsilon}{6B^2}} \right) \right] (t) \right] \right| \right] \leq 2B^2 \left[\frac{\epsilon}{6B^2} \right] = \frac{\epsilon}{3}. \end{aligned}$$

Das und (A.46) [17] gibt für alle $\epsilon \in (0, B^2)$

$$\begin{aligned} &\sup_{x,y\in[-B,B]} |xy - [\mathcal{R}_\varrho(\mu_\epsilon)](x,y)| \\ &= \sup_{x,y\in[-B,B]} \left| \frac{1}{2}\left[(x+y)^2 - x^2 - y^2 \right] - [\mathcal{R}_\varrho(\mu_\epsilon)](x,y) \right| \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

Damit wäre (6.5) gezeigt.
Nun kann, um die erste Behauptung zu zeigen, anhand der Definitionen gesehen werden, dass $\mathcal{L}(\alpha) = 2, \mathcal{L}(\sum) = 1$ ist. Für die erste Ungleichung wird 3.6(a)(i) und für die zweite Lemma 6.1(i) [17] verwendet, danach werden nur noch die Rechenregeln für den Logarithmus verwendet.

$$\begin{aligned} \mathcal{L}(\mu_\epsilon) &= \mathcal{L}(\alpha) + \mathcal{L}(\sum) + \mathcal{L}\left(\frac{\sigma_\epsilon}{6B^2}\right) \\ &= 2 + 1 + \frac{1}{2}\left| log_2\left(\frac{\epsilon}{6B^2}\right)\right| + 1 = \frac{1}{2} log_2\left(\frac{6B^2}{\epsilon}\right) + 4 \\ &\leq \frac{1}{2}\left(log_2\left(\frac{1}{\epsilon}\right) + 2log_2(B) + 3 \right) + 4 \\ &\leq \frac{1}{2} log_2\left(\frac{1}{\epsilon}\right) + log_2(B) + 6. \end{aligned}$$

Für die zweite Behauptung lässt sich mit den Definitionen der neuronalen Netze sagen, dass $\mathcal{M}(\alpha) = 14, \mathcal{M}(\sum) = 3$. Nun gilt für alle $\epsilon \in (0, B)$ folgendes: Die erste Ungleichung verwendet 3.6(a)(ii) und die dritte das Lemma 6.1(ii) [17], danach werden wieder die Rechenregeln für den Logarithmus angewendet und ese wird zusammengefasst:

$$\begin{aligned}\mathcal{M}(\mu_\epsilon) &\leq 2\left(\mathcal{M}(\sum) + 3\mathcal{M}\left(\frac{\sigma_\epsilon}{6B^2}\right) + \mathcal{M}(\alpha)\right)\\ &\leq 2\left(3 + 3\cdot 15\left(\frac{1}{2}\left(log_2\left(\frac{6B^2}{\epsilon}\right) + 1\right) + 14\right)\right)\\ &\leq 45 log_2\left(\frac{1}{\epsilon}\right) + 90 log_2(B) + 259.\end{aligned}$$

Zuletzt zeigt sich, dass für $\epsilon \in \left(B^2, \infty\right)$ *gilt* $\mathcal{L}(\mu_\epsilon) = 1$ *und* $\mathcal{M}(\mu_\epsilon) = 0$. Mit Lemma 5.3 und 5.4 aus [17] und per Konstruktion von μ_ϵ gilt, dass für $\epsilon \in (0, \infty)$ gilt $\mathcal{M}_1(\mu_\epsilon) = \mathcal{M}_1(\alpha) = 8, \mathcal{M}_{\mathcal{L}(\mu_\epsilon)}(\mu_\epsilon) = \mathcal{M}(\sum) = 3$. Dies beendet den Beweis. □

Weiter mit dem ursprünglichen Beweis. Nach Definition $\|\mathbf{A}\|_2, \|\mathbf{B}\|_2 \leq Z$ ist für jedes $i = 1, ..., d, k = 1, ..., n, j = 1, ..., l$ bekannt, dass gilt $|\mathbf{A}_{i,k}|, |\mathbf{B}_{k,j}| \leq Z$. Für die Mengen $i \in \{1, ..., d\}, k \in \{1, ..., n\}, j \in \{1, ..., l\}$ wird die zuvor definierte Matrix definiert $\mathbf{D}_{i,k,j}$, so dass für alle $\mathbf{A} \in \mathbb{R}^{d\mathbf{x}n}, \mathbf{B} \in \mathbb{R}^{n\mathbf{x}l}$ gilt

$$\mathbf{D}_{i,k,j}\left(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})\right) = \left(\mathbf{A}_{i,k}, \mathbf{B}_{k,j}\right).$$

Außerdem sei

$$\Phi^Z_{i,k,j;\epsilon} := \mathfrak{x}^Z_\epsilon \bullet \left(\left(\mathbf{D}_{i,k,j}, \mathbf{0}_{\mathbb{R}^2}\right)\right).$$

Für alle $i \in \{1, ..., d\}, k \in \{1, ..., n\}, j \in \{1, ..., l\}$ ergibt sich, dass $L\left(\Phi^Z_{i,k,j;\epsilon}\right) = L\left(\mathfrak{x}^Z_\epsilon\right)$ zusammen mit Lemma A.1 die Gleichungen (6.2),(6.3),(6.4) mit $\Phi_\epsilon := \Phi^Z_{i,k,j;\epsilon}$ erfüllt. Mit (6.5) resultiert

$$\sup_{(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) \in K^Z_{d,n,l}} \left|\mathbf{A}_{i,k}, \mathbf{B}_{k,j} - R^{K^Z_{d,n,l}}_\varrho\left(\Phi^Z_{i,j,k;\epsilon}\right)\left(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})\right)\right| \leq \frac{\epsilon}{n\sqrt{dl}}. \tag{6.6}$$

Definiere als nächsten Schritt für $\mathbf{1}_{\mathbb{R}^n} \in \mathbb{R}^n$ einen Vektor mit nur Einsen als Einträgen

und zusätzlich noch

$$\Phi^Z_{i,j;\epsilon} := ((\mathbf{1}_{\mathbb{R}^n}, 0)) \bullet P\left(\Phi^Z_{i,1,j;\epsilon}, ..., \Phi^Z_{i,n,j;\epsilon}\right) \bullet \left(\begin{pmatrix}\begin{pmatrix}\mathbf{Id}_{\mathbb{R}^{n(d+l)}}\\ \vdots \\ \mathbf{Id}_{\mathbb{R}^{n(d+l)}}\end{pmatrix}, \mathbf{0}_{\mathbb{R}^{n^2(d+l)}}\end{pmatrix}\right),$$

das mit Lemma 3.6 ein NN ist mit $n\cdot(d+l)$- dimensionalem Input und 1-dimensionalem Output, so dass (6.3) gilt mit $\Phi_\epsilon := \Phi^Z_{i,k,j;\epsilon}$. Außerdem, wenn darauf M und A.1 doppelt angewendet wird, also $M\left((\mathbf{a}, 0) \bullet \Phi\right) \leq M\left(\Phi\right)$ und dann (6.3), gibt das

$$\begin{aligned} M\left(\Phi^Z_{i,j;\epsilon}\right) &\leq M\left(P\left(\Phi^Z_{i,1,j;\epsilon}, ..., \Phi^Z_{i,n,j;\epsilon}\right)\right) \\ &\leq 90n \cdot \left(log_2\left(\frac{n\sqrt{dl}}{\epsilon}\right) + 2log_2\left(Z\right) + 6\right). \end{aligned} \tag{6.7}$$

Jetzt wird genau so A.1 genutzt und außerdem (6.4) dann gilt

$$M_1\left(\Phi^Z_{i,j;\epsilon}\right) \leq M_1\left(P\left(\Phi^Z_{i,1,j;\epsilon}, ..., \Phi^Z_{i,n,j;\epsilon}\right)\right) \leq 16n$$

und

$$M_{L\left(\Phi^Z_{i,j;\epsilon}\right)}\left(\Phi^Z_{i,j;\epsilon}\right) = M_{L\left(\Phi^Z_{i,j;\epsilon}\right)}\left(P\left(\Phi^Z_{i,1,j;\epsilon}, ..., \Phi^Z_{i,n,j;\epsilon}\right)\right) \leq 2n. \tag{6.8}$$

Die nächste Gleichung gilt, da sich mit 3.6 b) (iii) sagen lässt

$$R_\varrho^{K^Z_{d,n,l}}\left(\Phi^Z_{i,j;\epsilon}\right) = R_\varrho^{K^Z_{d,n,l}}\left(P\left(\Phi^Z_{i,1,j;\epsilon}, ..., \Phi^Z_{i,n,j;\epsilon}\right)\right)$$

und

$$P\left(\Phi^Z_{i,1,j;\epsilon}, ..., \Phi^Z_{i,n,j;\epsilon}\right)$$

sich umschreiben lässt zu

$$\left(R_\varrho^{K^Z_{d,n,l}}\left(\Phi^Z_{i,1,j;\epsilon}\right), ..., R_\varrho^{K^Z_{d,n,l}}\left(\Phi^Z_{i,n,j;\epsilon}\right)\right)$$

und das lässt sich wiederum schreiben als

$$\sum_{k=1}^{n} R_\varrho^{K^Z_{d,n,l}}\left(\Phi^Z_{i,k,j;\epsilon}\right)$$

das gibt dann wiederum

$$R_{\varrho}^{K_{d,n,l}^{Z}}\left(\Phi_{i,j;\epsilon}^{Z}\right)\left(\mathbf{vec}\left(\mathbf{A}\right),\mathbf{vec}\left(\mathbf{B}\right)\right)=\sum_{k=1}^{n}R_{\varrho}^{K_{d,n,l}^{Z}}\left(\Phi_{i,k,j;\epsilon}^{Z}\right)\left(\mathbf{vec}\left(\mathbf{A}\right),\mathbf{vec}\left(\mathbf{B}\right)\right).$$

Somit gilt durch (6.6)

$$\sup_{(\mathbf{vec}(\mathbf{A}),\mathbf{vec}(\mathbf{B}))\in K_{d,n,l}^{Z}}\left|\sum_{k=1}^{n}\mathbf{A}_{i,k},\mathbf{B}_{k,j}-R_{\varrho}^{K_{d,n,l}^{Z}}\left(\Phi_{i,j;\epsilon}^{Z}\right)\left(\mathbf{vec}\left(\mathbf{A}\right),\mathbf{vec}\left(\mathbf{B}\right)\right)\right|\leq\frac{\epsilon}{\sqrt{dl}}.$$

Als letzten Schritt definiere

$$\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}:=P\left(\Phi_{1,1;\epsilon}^{Z},...,\Phi_{d,1;\epsilon}^{Z},...,\Phi_{1,l;\epsilon}^{Z},...,\Phi_{d,l;\epsilon}^{Z}\right)\bullet\left(\left(\begin{pmatrix}\mathbf{Id}_{\mathbb{R}^{n(d+l)}}\\ \vdots\\ \mathbf{Id}_{\mathbb{R}^{n(d+l)}}\end{pmatrix},\mathbf{0}_{\mathbb{R}^{dln(d+l)}}\right)\right)$$

Dann bekommt man mit Lemma 3.6, dass für $\Phi_{\epsilon}:=\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}$ die Gleichung (6.3) erfüllt ist und somit (i) bewiesen ist. Wenn jetzt Lemma 3.6, A.1 und (6.7) benutzt wird gibt uns die Definition von $\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}$

$$M\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right)\leq 90dln\cdot\left(log_{2}\left(\frac{n\sqrt{dl}}{\epsilon}\right)+2log_{2}\left(Z\right)+6\right),$$

womit (ii) gezeigt ist. Jetzt lässt sich mit Lemma 3.6 und (6.8) folgern, dass

$$M_{1}\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right)\leq 16dln\ und\ M_{\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}}\left(\Phi_{mult;\tilde{\epsilon}}^{Z,d,n,l}\right)\leq 2dln,$$

was den Beweis von (iii) komplettiert. Per Definition und dem Fakt, dass für jedes $\mathbf{N}\in\mathbb{R}^{dxl}$ gilt

$$\left\|\mathbf{N}\right\|_{2}\leq\sqrt{dl}\max_{i,j}\left|\mathbf{N}_{i,j}\right|,$$

erhält man

$$\begin{aligned}
&\sup_{(\mathbf{vec}(\mathbf{A}),\mathbf{vec}(\mathbf{B}))\in K^Z_{d,n,l}} \left\| \mathbf{AB} - \mathbf{matr}\left(R_\varrho^{K^Z_{d,n,l}} \left(\Phi^{Z,d,n,l}_{mult;\tilde{\epsilon}} \right) (\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) \right) \right\|_2 \\
&\qquad \leq \sqrt{dl} \sup_{(\mathbf{vec}(\mathbf{A}),\mathbf{vec}(\mathbf{B}))\in K^Z_{d,n,l}} \max_{i=1,\ldots,d,j=1,\ldots,l} \\
&\qquad \left| \sum_{k=1}^{n} \mathbf{A}_{i,k}, \mathbf{B}_{k,j} - R_\varrho^{K^Z_{d,n,l}} \left(\Phi^{Z}_{i,j;\epsilon} \right) (\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) \right| \\
&\qquad \leq \epsilon
\end{aligned} \tag{6.9}$$

und daraus resultiert (iv) der Behauptung. Zuletzt gilt für für $(\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) \in K^Z_{d,n,l}$ das

$$\left\| \mathbf{matr}\left(R_\varrho^{K^Z_{d,n,l}} \left(\Phi^{Z,d,n,l}_{mult;\tilde{\epsilon}} \right) (\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) \right) \right\|_2 \leq \epsilon,$$

nun wird $+0$ gerechnet, damit sich (iv) der Behauptung anwenden lässt und damit gilt

$$\left\| \mathbf{matr}\left(R_\varrho^{K^Z_{d,n,l}} \left(\Phi^{Z,d,n,l}_{mult;\tilde{\epsilon}} \right) (\mathbf{vec}(\mathbf{A}), \mathbf{vec}(\mathbf{B})) \right) - \mathbf{AB} \right\|_2 + \|\mathbf{AB}\|_2 \leq \epsilon$$

das wiederum kann, mit der Eigenschaft $\|\mathbf{A}\|_2, \|\mathbf{B}\|_2 \leq Z$, abgeschätzt werden und dann ergibt dies zusammen mit (iv), dass dies $\leq$ ist als

$$\epsilon + \|\mathbf{A}\|_2 \|\mathbf{B}\|_2 \leq 1 + Z^2 \leq 1 + Z^2.$$

Dies zeigt den fünften und letzten Part der Behauptung und beendet damit den Beweis. □

6.3.3 Beweis Theorem 3.8:

Bevor in diesem Abschnitt Theorem 3.8 bewiesen werden kann, werden NNs konstruiert, welche die Abbildung $\mathbf{A} \mapsto \mathbf{A}^k$ für $k \in \mathbb{N}$ und Quadratmatrizen $\mathbf{A}$ imitieren. Dafür wird unter anderem Proposition 3.7 verwendet. Aus Proposition 3.7 kann direkt die Größe der Abschätzung der Nachahmung der Multiplikation zweier Quadratmatrizen entnommen werden. Tatsächlich existiert eine Konstante $C_1 > 0$, so dass für $d \in \mathbb{N}, Z > 0, \epsilon \in (0,1)$ gilt

(i) $L\left(\Phi_{mult;\epsilon}^{Z,d,d,d}\right) \leq C_1 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(d\right) + log_2\left(max\{1,Z\}\right)\right)$,

(ii) $M\left(\Phi_{mult;\epsilon}^{Z,d,d,d}\right) \leq C_1 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(d\right) + log_2\left(max\{1,Z\}\right)\right) d^3$,

(iii) $M_1\left(\Phi_{mult;\epsilon}^{Z,d,d,d}\right) \leq C_1 d^3$, sowie $M_{L\left(\Phi_{mult;\epsilon}^{Z,d,d,d}\right)}\left(\Phi_{mult;\epsilon}^{Z,d,d,d}\right) \leq C_1 d^3$,

(iv) $\sup_{(\mathbf{vec}(\mathbf{A}),\mathbf{vec}(\mathbf{B}))\in K_{d,d,d}^{Z}} \left\|\mathbf{AB} - \mathbf{matr}\left(R_{\varrho}^{K_{d,d,d}^{Z}}\left(\Phi_{mult;\epsilon}^{Z,d,d,d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right),\mathbf{vec}\left(\mathbf{B}\right)\right)\right)\right\|_2 \leq \epsilon$,

(v) für alle $\left(\mathbf{vec}\left(\mathbf{A}\right),\mathbf{vec}\left(\mathbf{B}\right)\right) \in K_{d,d,d}^{Z}$ gilt

$$\left\|\mathbf{matr}\left(R_{\varrho}^{K_{d,d,d}^{Z}}\left(\Phi_{mult;\epsilon}^{Z,d,d,d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right),\mathbf{vec}\left(\mathbf{B}\right)\right)\right)\right\|_2 \leq \epsilon \left\|\mathbf{A}\right\|_2 \left\|\mathbf{B}\right\|_2 \leq \epsilon + Z^2 \leq 1 + Z^2.$$

Es lässt sich auch das Quadrat von Matrizen nachbilden. Genauer geschieht dies im folgenden.

Definition A.2. Für $d \in \mathbb{N}, Z > 0$, und $\epsilon \in (0,1)$ definiere das NN

$$\Phi_{2;\epsilon}^{Z,d} := \Phi_{mult;\epsilon}^{Z,d,d,d} \bullet \left(\left(\begin{pmatrix}\mathbf{Id}_{\mathbb{R}^{d^2}}\\ \mathbf{Id}_{\mathbb{R}^{d^2}}\end{pmatrix}, \mathbf{0}_{\mathbb{R}^{2d^2}}\right)\right),$$

welches eine d^2-In- und Outputdimension hat. Mit 3.6 ist leicht zu sehen, dass eine Konstante $C_{sq} > C_1$ existiert, so dass für alle $d \in \mathbb{N}, Z > 0, \epsilon \in (0,1)$

(i) $L\left(\Phi_{2;\epsilon}^{Z,d}\right) \leq C_{sq} \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(d\right) + log_2\left(max\{1,Z\}\right)\right)$,

(ii) $M\left(\Phi_{2;\epsilon}^{Z,d}\right) \leq C_{sq} d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(d\right) + log_2\left(max\{1,Z\}\right)\right)$,

(iii) $M_1\left(\Phi_{2;\epsilon}^{Z,d}\right) \leq C_{sq} d^3$, sowie $M_{L\left(\Phi_{2;\epsilon}^{Z,d}\right)}\left(\Phi_{2;\epsilon}^{Z,d}\right) \leq C_{sq} d^3$,

(iv) $\sup_{\mathbf{vec}(\mathbf{A})\in K_d^Z} \left\|\mathbf{A}^2 - \mathbf{matr}\left(R_{\varrho}^{K_d^Z}\left(\Phi_{2;\epsilon}^{Z,d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right)\right)\right)\right\|_2 \leq \epsilon$,

(v) für alle $\mathbf{vec}\left(\mathbf{A}\right) \in K_d^Z$ bekommen wir

$$\left\|\mathbf{matr}\left(R_{\varrho}^{K_d^Z}\left(\Phi_{2;\epsilon}^{Z,d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right)\right)\right)\right\|_2 \leq \epsilon \left\|\mathbf{A}\right\|^2 \leq \epsilon + Z^2 \leq 1 + Z^2.$$

Als nächstes will die Abbildung $\mathbf{A} \mapsto \mathbf{A}^k$ für ein beliebiges $k \in \mathbb{N}_0$ approximiert werden. Es wird angenommen, dass $k = 2$ ist und das die Norm der Matrizen durch $\frac{1}{2}$ beschränkt ist.

Proposition A.3. Sei $d \in \mathbb{N}, j \in \mathbb{N}$, sowie $\epsilon \in (0, \frac{1}{4})$. Dann existiert ein $NN\Phi_{2^j;\epsilon}^{\frac{1}{2},d}$ mit d^2-dimensionalem In- und Output mit den folgenden Eigenschaften:

(i) $L\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right) \leq C_{sq} j \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(d\right)\right) + 2C_{sq} \cdot (j-1)$,

(ii) $M\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right) \leq C_{sq} j d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(d\right)\right) + 4C_{sq} \cdot (j-1)\, d^3$,

(iii) $M_1\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right) \leq C_{sq} d^3$, sowie auch $M_{L\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right) \leq C_{sq} d^3$,

(iv) $\sup_{\mathbf{vec}(\mathbf{A}) \in K_d^{\frac{1}{2}}} \left\| \mathbf{A}^{2^j} - \mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right)\right)\right)\right\|_2 \leq \epsilon$,

(v) für jede $\mathbf{vec}\left(\mathbf{A}\right) \in K_d^{\frac{1}{2}}$ haben wir

$$\left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right)\right)\right)\right\|_2 \leq \epsilon + \left\|\mathbf{A}^{2^j}\right\|_2 \leq \epsilon + \left\|\mathbf{A}\right\|_2^{2^j} \leq \frac{1}{4} + \left(\frac{1}{2}\right)^{2^j} \leq \frac{1}{2}.$$

Jetzt zum Beweis dieser Eigenschaften. Die Aussage wird über Induktion über $j \in \mathbb{N}$ gezeigt, und für $j = 1$ gilt die Aussage, wenn $\Phi_{2;\epsilon}^{\frac{1}{2},d}$ wie in A.2 gewählt wird. Angenommen die Behauptung stimmt für ein beliebiges, aber festes $j \in \mathbb{N}$, dann existiert ein NN $\Phi_{2^j;\epsilon}^{\frac{1}{2},d}$, so dass

$$\begin{aligned} &\left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right)\right)\right) - \mathbf{A}^{2^j}\right\|_2 \leq \epsilon, \\ &\left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)\left(\mathbf{vec}\left(\mathbf{A}\right)\right)\right)\right\|_2 \leq \epsilon + \left(\frac{1}{2}\right)^{2^j} \end{aligned} \tag{6.10}$$

und $\Phi_{2^j;\epsilon}^{\frac{1}{2},d}$ erfüllt (i),(ii),(iii). Nun definiere

$$\Phi_{2^{j+1};\epsilon}^{\frac{1}{2},d} := \Phi_{2;\frac{\epsilon}{4}}^{1,d} \odot \Phi_{2^j;\epsilon}^{\frac{1}{2},d}.$$

Für $\mathbf{vec}(\mathbf{A}) \in K_d^{\frac{1}{2}}$ und indem zunächst 0 addiert wird, um dann die Dreiecksungleichung anzuwenden, ergibt sich

$$\begin{aligned}
&\left\| \mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^{2^{j+1}} \right\|_2 \\
&\leq \left\| \mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) \right\|_2 \\
&+ \left\| \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) - \left(\mathbf{A}^{2^j}\right)^2 \right\|_2
\end{aligned} \tag{6.11}$$

Da alle Matrizen durch $\frac{1}{2}$ beschränkt sind und Φ durch diese A's konstruiert wird, lässt sich die Matrix der Realisierung von $\Phi \cdot \mathbf{vec}(\mathbf{A})$ durch $\frac{1}{2}$ abschätzen. Durch die Konstruktion von $\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}$ gilt

$$\left\| \mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) - \left(\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right)\right)^2 \right\|_2 \leq \frac{\epsilon}{4}.$$

Dann benutze die Dreiecksungleichung und die Eigenschaft, dass $\|\cdot\|$ eine submultiplikative Operatornorm ist, dafür, den nächsten Term wie folgt abzuschätzen

$$\left\| \mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) \right\|_2$$

der erste Teil der Gleichung ist per Voraussetzung $\leq \frac{1}{2}$ ist und der zweite Teil $\leq \frac{1}{4}$. Darum kann dies abgeschätzt werden mit

$$\begin{aligned}
\leq \frac{\epsilon}{4} + \Bigg\| &\left(\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right)\right)^2 \\
&- \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)(\mathbf{vec}(\mathbf{A}))\right) \Bigg\|_2
\end{aligned}$$

nutze nun die Submultiplikativität

$$\leq \frac{\epsilon}{4} + \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) - \mathbf{A}^{2^j} \right\|_2 \cdot \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) \right\|_2 .$$

Wenn nun (6.11) verwendet wird, gilt

$$\leq \frac{\epsilon}{4} + \epsilon \cdot \left(\epsilon + \left(\frac{1}{2} \right)^{2^j} \right) \leq \frac{3}{4}\epsilon. \tag{6.12}$$

Also mit der Submultiplikatitvitätseigenschaft der Norm ergibt sich, dass wenn mit $\mathbf{A}^{2^j}$ multipliziert wird, gilt

$$\begin{aligned} &\left\| \mathbf{A}^{2^j} \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) - \left(\mathbf{A}^{2^j} \right)^2 \right\|_2 \\ &\leq \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) - \mathbf{A}^{2^j} \right\|_2 \left\| \mathbf{A}^{2^j} \right\|_2 \\ &\leq \frac{\epsilon}{4}. \end{aligned} \tag{6.13}$$

Nun kann (6.11) mit (6.12) wie folgt abgeschätzt werden

$$\begin{aligned} &\leq \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^{j+1};\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) - \mathbf{A}^{2^j} \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) \right\|_2 \\ &+ \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) - \mathbf{A}^{2^j} \right\|_2 \left\| \mathbf{A}^{2^j} \right\|_2 \\ &\leq \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^{j+1};\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) - \mathbf{A}^{2^j} \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d} \right)(\mathbf{vec}(\mathbf{A})) \right) \right\| + \frac{\epsilon}{4} \end{aligned}$$

das wiederum kann mit Hilfe von (6.12) abgeschätzt werden mit

$$\frac{3}{4}\epsilon + \frac{\epsilon}{4} = \epsilon \tag{6.14}$$

damit ist (iv) von A.3 bewiesen. Direkt daraus lässt sich (v) zeigen

$$\begin{aligned}\left\|\mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 &\leq \epsilon + \left\|\mathbf{A}^{2^{j+1}}\right\|_2 \leq \epsilon + \|\mathbf{A}\|_2^{2^{j+1}} \\ &\leq \frac{1}{4} + \left(\frac{1}{2}\right)^{2^j} \leq \frac{1}{2}.\end{aligned} \tag{6.15}$$

Jetzt wird die Größe von $\Phi_{2^{j+1};\epsilon}^{\frac{1}{2},d}$ abgeschätzt. Mit der Induktionshypothese und Lemma 3.6(a)(i) erhält man

$$L\left(\Phi_{2^{j+1};\epsilon}^{\frac{1}{2},d}\right) = L\left(\Phi_{2^j;\frac{\epsilon}{4}}^{1,d}\right) + L\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right).$$

Wende jetzt A.2 (i) und A.3 (i) an, dann kann dies abgeschätzt werden mit

$$\leq C_{sq} \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + log_2(4)\right) + C_{sq} j \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d)\right) + 2C_{sq}(j-1)$$

wenn jetzt C_{sq} ausgeklammert, die Gleichung zusammengefasst und $log_2(4) = 2$ genutzt wird, bekommt man

$$= C_{sq}\left((j+1)\, log_2\left(\frac{1}{\epsilon}\right) + (j+1)\, log_2(d) + 2j\right)$$

und damit ist (i) bewiesen. Für (ii) wird Lemma 3.6(a)(ii) zusammen mit der Induktionshypothese benutzt und es gilt

$$M\left(\Phi_{2^{j+1};\epsilon}^{\frac{1}{2},d}\right) \leq M\left(\Phi_{2;\frac{\epsilon}{4}}^{1,d}\right) + M\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right) + M_1\left(\Phi_{2;\frac{\epsilon}{4}}^{1,d}\right) + M_{L\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right).$$

Das lässt sich mit Hilfe von Definition A.2 (ii), (iii) und Proposition A.3 (ii) und (iii) abschätzen durch

$$\begin{aligned}\leq\ & C_{sq} d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + log_2(4)\right) + C_{sq} j d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d)\right) \\ & + 4C_{sq} \cdot (j-1)d^3 + 2C_{sq} d^3\end{aligned}$$

um den Beweis für (ii) zu beenden, muss nur noch ausgeklammert und zusammengefasst werden, dann gilt

$$= C_{sq} d^3 \cdot \left((j+1) log_2 \left(\frac{1}{\epsilon} \right) + (j+1) log_2(d) + 4j \right).$$

Schlussendlich folgt mit Lemma 3.6(a)(iii) zusammen mit der Induktionshypothese genau so wie mit 3.6(a)(iv), dass mit der Definition von $\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}$ gilt

$$M_1 \left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon} \right) = M_1 \left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon} \right) \leq C_{sq} d^3,$$

sowie auch um den Beweis zu beenden

$$M_{L\left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon}\right)} \left(\Phi^{\frac{1}{2},d}_{2^{j+1};\epsilon} \right) = M_{L\left(\Phi^{1,d}_{2;\frac{\epsilon}{4}}\right)} \left(\Phi^{1,d}_{2;\frac{\epsilon}{4}} \right).$$

□

Jetzt wird gezeigt, wie ein NN gebaut wird, welches die Abbildung $\mathbf{A} \mapsto \mathbf{A}^k$ für ein beliebiges $k \in \mathbb{N}_0$ imitiert. Es wird weiter angenommen, dass die Norm der Matrizen durch $\frac{1}{2}$ beschränkt ist.

Proposition A.4. Sei $d \in \mathbb{N}, k \in \mathbb{N}_0$, und $\epsilon \in \left(0, \frac{1}{4}\right)$. Dann existiert ein NN $\Phi^{\frac{1}{2},d}_{k;\epsilon}$ mit d^2-dimensionalem In- und Output, welche die folgenden Eigenschaften erfüllt:

(i)

$$\begin{aligned} L\left(\Phi^{\frac{1}{2},d}_{k;\epsilon}\right) &\leq \lfloor log_2(max\{k,2\}) \rfloor L\left(\Phi^{1,d}_{mult,\frac{\epsilon}{4}}\right) + L\left(\Phi^{\frac{1}{2},d}_{2^{\lfloor log_2(max\{k,2\}) \rfloor};\epsilon}\right) \\ &\leq 2C_{sq} \lfloor log_2(max\{k,2\}) \rfloor \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 2 \right), \end{aligned}$$

(ii)

$$\begin{aligned} M\left(\Phi^{\frac{1}{2},d}_{k;\epsilon}\right) \leq \frac{3}{2} C_{sq} d^3 \cdot \lfloor log_2(max\{k,2\}) \rfloor \cdot (\lfloor log_2(max\{k,2\}) \rfloor + 1) \\ \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4 \right), \end{aligned}$$

(iii)

$$M_1\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right) \leq C_{sq} \cdot (\lfloor log_2(max\{k,2\})\rfloor + 1)d^3 \; sowie$$
$$M_{L\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right) \leq C_{sq}d^3,$$

(iv)

$$\sup_{\mathbf{vec}(\mathbf{A})\in K_d^{\frac{1}{2}}} \left\| \mathbf{A}^k - \mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \leq \epsilon,$$

(v) für alle $\mathbf{vec}(\mathbf{A}) \in K_d^{\frac{1}{2}}$ haben wir

$$\left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \leq \epsilon + \left\|\mathbf{A}^k\right\|_2 \leq \frac{1}{4} + \|\mathbf{A}\|_2^k \leq \frac{1}{4} + \left(\frac{1}{2}\right)^k.$$

Nun zum Beweis. Per Induktion über $k \in \mathbb{N}_0$ werden die Ergebnisse gezeigt. Für $k = 0, k = 1$, ist dies trivial, wenn die NNs wie folgt definiert werden

$$\Phi_{0;\epsilon}^{\frac{1}{2},d} := \left(\left(\mathbf{0}_{\mathbb{R}^{d^2}x\mathbb{R}^{d^2}}, \mathbf{vec}\left(\mathbf{Id}_{\mathbb{R}^d}\right)\right)\right), \quad \Phi_{1;\epsilon}^{\frac{1}{2},d} := \left(\left(\mathbf{Id}_{\mathbb{R}^{d^2}}, \mathbf{0}_{\mathbb{R}^{d^2}}\right)\right).$$

Für die Hypothese wird angenommen, dass die Ergebnisse für alle $k' \leq k \in \mathbb{N}$ gelten. Wenn $k = 2$ ist, gelten die Ergebnisse mit Proposition A.3, somit kann ohne Verlust der Allgemeinheit vermutet werden, dass $k \neq 2$ ist. Definiere $j := \lfloor log_2(k)\rfloor$, so dass für $t := k - 2^j$ gilt $0 < t < 2^j$. Das impliziert $A^k = A^{2^j}A^t$, was zu späterem Zeitpunkt im Beweis noch notwendig ist. Mit A.3 und der Induktionshypothese weiß man, dass ein NN $\Phi_{2^j;\epsilon}^{\frac{1}{2},d}$ existiert, welches die Eigenschaften von A.3 erfüllt und ein NN $\Phi_{t;\epsilon}^{\frac{1}{2},d}$, welches die Eigenschaften von A.4 erfüllt. Definiere nun das NN

$$\Phi_{k;\epsilon}^{\frac{1}{2},d} := \Phi_{mult;\frac{\epsilon}{4}}^{1,d,d,d} \odot P\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}, \Phi_{t;\epsilon}^{\frac{1}{2},d}\right) \bullet \left(\left(\begin{pmatrix}\mathbf{Id}_{\mathbb{R}^{d^2}}\\ \mathbf{Id}_{\mathbb{R}^{d^2}}\end{pmatrix}, \mathbf{0}_{\mathbb{R}^{2d^2}}\right)\right).$$

Um (iii) zu beweisen, wird die Konstruktion und Lemma 3.6(a)(iv) verwendet, dann ist leicht zu sehen

$$M_{L\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right) = M_{L\left(\Phi_{mult;\frac{\epsilon}{4}}^{1,d,d,d}\right)}\left(\Phi_{mult;\frac{\epsilon}{4}}^{1,d,d,d}\right) \leq C_{sq}d^3.$$

Außerdem, mit der Induktionshypothese und mithilfe der Eigenschaften aus Lemma 3.6(a)(iii) und (b)(iv) gilt

$$M_1\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right) = M_1\left(P\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}, \Phi_{t;\epsilon}^{\frac{1}{2},d}\right)\right) = M_1\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right) + M_1\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)$$

Nun verwende A.3 (iii) und A.4 (iii), damit kann das abgeschätzt werden mit

$$\leq C_{sq}d^3 + (j+1)C_{sq}d^3 = (j+2)C_{sq}d^3.$$

und damit ist (iii) der Behauptung bewiesen. Um (iv) zu beweisen gehe genau so vor wie bei (6.11). Es wird erst +0 gerechnet und dann nutze die Definition $A^k = A^{2^j}A^t$, um dann die Dreiecksungleichung verwenden zu können

$$\begin{aligned}
&\left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^k\right\|_2 \\
&\leq \left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^{2^j}\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \\
&+ \left\|\mathbf{A}^{2^j}\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^{2^j}\mathbf{A}^t\right\|_2.
\end{aligned} \tag{6.16}$$

Mit der Konstruktion von $\Phi_{k;\epsilon}^{\frac{1}{2},d}$ und Proposition 3.7(iv) lässt sich sagen, dass

$$\begin{aligned}
&\left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right. \\
&\left. - \mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right)\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \leq \frac{\epsilon}{4}.
\end{aligned}$$

Dadurch, dass die Matrizen mit $\frac{1}{2}$ beschränkt sind, ist dies $\leq \frac{\epsilon}{4}$.

Mit (6.16) können wir folgende Abschätzung vornehmen

$$\begin{aligned}
&\left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^k \right\|_2 \\
&\leq \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) \right\| \\
&+ \left\| \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^k \right\|_2
\end{aligned}$$

so wie Φ konstruiert ist und mit der Definition von k ergibt sich

$$\begin{aligned}
&\leq \frac{\epsilon}{4} + \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) \right. \\
&\quad \left. - \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) \right\|_2 \\
&\quad + \left\| \mathbf{A}^{2^j}\,\mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^k \right\|_2 .
\end{aligned}$$

Mit der Submultiplikativität der Norm resultiert dann

$$\begin{aligned}
&\leq \frac{\epsilon}{4} + \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) \right\|_2 \\
&\quad \cdot \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^{2^j} \right\|_2 \\
&\quad + \left\| \mathbf{A}^{2^j} \right\|_2 \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^t \right\|_2 \\
&\quad := \frac{\epsilon}{4} + I + II
\end{aligned}$$

Mit der vorherigen Konstruktion

$$\Phi_{1;\epsilon}^{\frac{1}{2},d} := \left(\left(\mathbf{Id}_{\mathbb{R}^{d^2}}, \mathbf{0}_{\mathbb{R}^{d^2}} \right) \right)$$

lässt sich sagen sagen, dass II gleich 0 ist für $t = 1$. Das heißt es gilt

$$\left\|\mathbf{A}^{2^j}\right\|_2 \left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^t\right\|_2 = 0$$

und somit dann auch

$$\begin{aligned}&\left\|\mathbf{A}^{2^j}\right\|_2 \cdot \left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\left(\mathbf{Id}_{\mathbb{R}^{d^2}}, \mathbf{0}_{\mathbb{R}^{d^2}}\right)\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^1\right\|_2 \\ &\quad \leq \frac{1}{4}\left\|\mathbf{matr}\left(R_\varrho^{K_d^{\frac{1}{2}}}\left(\left(\mathbf{Id}_{\mathbb{R}^{d^2}}, \mathbf{0}_{\mathbb{R}^{d^2}}\right)\right)(\mathbf{vec}(\mathbf{A}))\right) - \mathbf{A}^1\right\|_2\end{aligned}$$

Mit Lemma 3.3 kann gezeigt werden, dass der erste Teil der Norm $\mathbf{Id}_{\mathbb{R}^{d^2}}$ entspricht. Mit dieser Eigenschaft und damit, dass $\mathbf{vec}(\mathbf{A}) \in K_d^{\frac{1}{2}}$, gilt

$$\begin{aligned}&= \frac{1}{4}\left\|\mathbf{matr}\left(\mathbf{Id}_{\mathbb{R}^{d^2}} \cdot \mathbf{vec}(\mathbf{A})\right) - \mathbf{A}\right\|_2 \\ &= \frac{1}{4}\left\|\mathbf{A} - \mathbf{A}\right\|_2 = 0,\end{aligned}$$

dass $II = 0$ ist und damit

$$\frac{\epsilon}{4} + I + II = \frac{\epsilon}{4} + I \leq \frac{\epsilon}{4} + \|\mathbf{A}\|_2 \epsilon \leq \frac{3\epsilon}{4} \leq \epsilon.$$

Durch die Beschränktheit der Matrizen, A.3(iv) und (6.16) lassen sich I und II für $t \geq 2$ wie folgt abschätzen

$$\begin{aligned}&\frac{\epsilon}{4} + I + II \\ &\leq \frac{\epsilon}{4} + \epsilon\left(\epsilon + \|\mathbf{A}\|^t + \left\|\mathbf{A}^{2^j}\right\|\right) \\ &\leq \frac{\epsilon}{4} + \left(\frac{1}{4} + \left(\frac{1}{2}\right)^t + \left(\frac{1}{2}\right)^{2^j}\right) \\ &\leq \frac{\epsilon}{4} + \frac{3\epsilon}{4} = \epsilon.\end{aligned}$$

Damit ist der Beweis für (iv) beendet. Um (v) zu zeigen, wird wieder die Dreiecksungleichung verwendet

$$\begin{aligned}
&\left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A})) \right) - \mathbf{A}^k \right\|_2 \leq \epsilon \\
&\Rightarrow \left\| \mathbf{matr}\left(R_{\varrho}^{K_d^{\frac{1}{2}}}\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right)(\mathbf{vec}(\mathbf{A})) \right) \right\| \leq \epsilon + \left\| \mathbf{A}^k \right\|_2 \\
&\leq \epsilon + \|\mathbf{A}\|_2^k \leq \frac{1}{4} + \left(\frac{1}{2}\right)^k
\end{aligned}$$

Und damit ist (v) der Behauptung bewiesen.

Nun wird die Größe von $\Phi_{k;\epsilon}^{\frac{1}{2},d}$ analysiert. Wenn nun Lemma 3.6(a)(i) und (b)(i) kombiniert werden und die Konstruktion Verwendung findet, gibt der erste Schritt

$$L\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right) \leq L\left(\Phi_{mult;\epsilon}^{1,d,d,d}\right) + max\left\{L\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right), L\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right)\right\}.$$

Für die nächste Abschätzung verwende nun die Induktionsvoraussetzung, also die Aussage von Proposition A.4 (i) für $k = 2^j$ genau so wie für $k = t$. Somit ergibt sich folgende Abschätzung

$$\begin{aligned}
&\leq L\left(\Phi_{mult;\epsilon}^{1,d,d,d}\right) \\
&+ max\left\{(j-1)L\left(\Phi_{mult;\epsilon}^{1,d,d,d}\right) + L\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right), (j-1)L\left(\Phi_{mult;\epsilon}^{1,d,d,d}\right) + L\left(\Phi_{2^{j-1};\epsilon}^{\frac{1}{2},d}\right)\right\} \\
&\leq jL\left(\Phi_{mult;\epsilon}^{1,d,d,d}\right) + L\left(\Phi_{2^j;\epsilon}^{\frac{1}{2},d}\right)
\end{aligned}$$

nutze nun A.3(i) und A.2(i), dann gilt

$$\begin{aligned}
&\leq C_{sq} j \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 2\right) + C_{sq} j \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d)\right) + 2C_{sq} \cdot (j-1) \\
&\leq 2C_{sq} j \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 2\right),
\end{aligned}$$

und damit ist der Beweis von (i) abgeschlossen.

Nun zu den Gewichten des NNs, welche ungleich 0 sind. Zu erst gilt mit Lemma

3.6(a)(ii), dass

$$\begin{aligned} M\left(\Phi_{k;\epsilon}\frac{1}{2},d\right) &\leq \left(M\left(\Phi^{1,d,d,d}_{mult;\epsilon}\right)+M_1\left(\Phi^{1,d,d,d}_{mult;\epsilon}\right)\right)+M\left(P\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon},\Phi^{\frac{1}{2},d}_{t;\epsilon}\right)\right) \\ &\quad + M_{L\left(P\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon},\Phi^{\frac{1}{2},d}_{t;\epsilon}\right)\right)}\left(P\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon},\Phi^{\frac{1}{2},d}_{t;\epsilon}\right)\right) \\ &:= I' + II'(a) + II'(b). \end{aligned}$$

Mit den Eigenschaften von $\Phi^{1,d,d,d}_{mult;\epsilon}$, A.2(ii),(iii) kann I' wie folgt abgeschätzt werden

$$\begin{aligned} I' &= M\left(\Phi^{1,d,d,d}_{mult;\epsilon}\right)+M_1\left(\Phi^{1,d,d,d}_{mult;\epsilon}\right) \\ &\leq C_{sq}d^3\left(log_2\left(\frac{1}{\epsilon}\right)+log_2(d)+2\right)+C_{sq}d^3 \\ &= C_{sq}d^3\left(log_2\left(\frac{1}{\epsilon}\right)+log_2(d)+3\right) \end{aligned}$$

Sei nun $L \leq 2C_{sq}j\left(log_2\left(\frac{1}{\epsilon}\right)+log_2(d)+2\right)$, und mit der Definition für Parallelisierungen gilt

$$II'(a) = M\left(P\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon},\Phi^{\frac{1}{2},d}_{t;\epsilon}\right)\right)$$

nun verkette das erste NN mit $\Phi^{Id}_{d^2,L}$, damit die NNs die gleiche Anzahl an Schichten haben und somit 3.6(b)(iii) benutzt werden kann, dass gibt uns dann

$$\begin{aligned} &= M\left(P\left(\Phi^{Id}_{d^2,L}\odot\Phi^{\frac{1}{2},d}_{2^j;\epsilon},\Phi^{\frac{1}{2},d}_{t;\epsilon}\right)\right) \\ &= M\left(\Phi^{Id}_{d^2,L}\odot\Phi^{\frac{1}{2},d}_{2^j;\epsilon}\right)+M\left(\Phi^{\frac{1}{2},d}_{t;\epsilon}\right) \\ &\leq M\left(\Phi^{Id}_{d^2,L}\right)+M_1\left(\Phi^{Id}_{d^2,L}\right)+M_{L\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon}\right)}\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon}\right)+M\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon}\right)+M\left(\Phi^{\frac{1}{2},d}_{t;\epsilon}\right) \\ &\leq 2d^2(L+1)+C_{sq}d^3+M\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon}\right)+M\left(\Phi^{\frac{1}{2},d}_{t;\epsilon}\right) \end{aligned}$$

wobei für die vorletzte Abschätzung 3.6(a)(ii) und für die letzte A.3(iii) verwendet wurde.
Mit der Definition von Parallelisierungen zweier NNs mit unterschiedlich vielen Schichten erhält man

$$II'(b) = M_{L\left(P\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon},\Phi^{\frac{1}{2},d}_{t;\epsilon}\right)\right)}\left(P\left(\Phi^{\frac{1}{2},d}_{2^j;\epsilon},\Phi^{\frac{1}{2},d}_{t;\epsilon}\right)\right) \leq d^2+C_{sq}d^3.$$

Wenn man jetzt alle 3 kombiniert, kann folgende Abschätzungen vorgenommen werden

$$\begin{aligned} M\left(\Phi_{k;\epsilon}^{\frac{1}{2},d}\right) &\leq C_{sq}d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 3\right) + 2d^2 \cdot (L+1) + d^2 + C_{sq}d^3 \\ &\quad + M\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right) + M\left(\Phi_{2^j,\epsilon}^{\frac{1}{2},d}\right) \\ &\leq C_{sq}d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4\right) + 2d^2 \cdot (L+2) \\ &\quad + M\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right) + M\left(\Phi_{2^j,\epsilon}^{\frac{1}{2},d}\right). \end{aligned}$$

Mit der Abschätzung

$$M\left(\Phi_{2^{j+1};\epsilon}^{\frac{1}{2},d}\right) \leq C_{sq}d^3\left((j+1)log_2\left(\frac{1}{\epsilon}\right) + (j+1)log_2(d) + 4j\right)$$

aus A.3 und der Abschätzung für L kann die zuvorstehende Gleichung abgeschätzt werden mit

$$\begin{aligned} &\leq C_{sq}(j+1)d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4\right) + 2d^2 \cdot (L+2) + M\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right) \\ &\leq C_{sq}(j+1)d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4\right) + 2C_{sq}jd^2 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 2\right) \\ &+ 4d^2 + M\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right) \end{aligned}$$

das kann mit

$$\begin{aligned} &2C_{sq}jd^2\left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 2\right) + 4d^2 \\ &\qquad \leq 2C_{sq}d^3(j+1)\left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4\right) \end{aligned}$$

abgeschätzt werden und das gibt mit der Tatsache, dass $\Phi_{t;\epsilon}^{\frac{1}{2},d}$ (i)-(v) erfüllt und A.4 (ii)

$$\begin{aligned} &\leq 3C_{sq}d^3(j+1)\left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4\right) + M\left(\Phi_{t;\epsilon}^{\frac{1}{2},d}\right) \\ &\leq 3C_{sq}d^3\left(j+1+\frac{j(j+1)}{2}\right) \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4\right) \\ &= \frac{3}{2}C_{sq} \cdot (j+1) \cdot (j+2)d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + 4\right) \end{aligned}$$

und beendet den Beweis. □

Bis jetzt ging es nur um Konstruktionen von Netzen, dessen Normen durch $\frac{1}{2}$ beschränkt waren. Jetzt kommt eine Konstruktion, in der die Realisierung des NNs $\Phi_{k;\epsilon}^{Z,d}$ die Abbildung $\mathbf{A} \mapsto \mathbf{A}^k$ approximiert, dessen Normen alle durch ein beliebiges $Z > 0$ beschränkt sind.

Korollar A.5. Es existiert eine Konstante $C_{pow} > C_{sq}$, so dass für alle $Z > 0, d \in \mathbb{N}$ und $k \in \mathbb{N}_0$ ein NN $\Phi_{k;\epsilon}^{Z,d}$ existiert, so dass folgende Eigenschaften gelten:

(i) $L\left(\Phi_{k;\epsilon}^{Z,d}\right) \leq C_{pow} log_2\left(max\{k,2\}\right) \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + k log_2\left(max\{1,Z\}\right)\right)$,

(ii) $M\left(\Phi_{k;\epsilon}^{Z,d}\right) \leq C_{pow} log_2^2\left(max\{k,2\}\right) d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(d) + k log_2\left(max\{1,Z\}\right)\right)$,

(iii) $M_1\left(\Phi_{k;\epsilon}^{Z,d}\right) \leq C_{pow} log_2\left(max\{k,2\}\right) d^3$, so wie auch $M_{L\left(\Phi_{k;\epsilon}^{Z,d}\right)}\left(\Phi_{k;\epsilon}^{Z,d}\right) \leq C_{pow} d^3$,

(iv) $\sup\limits_{\mathbf{vec}(\mathbf{A}) \in K_d^Z} \left\| \mathbf{A}^k - \mathbf{matr}\left(\mathcal{R}_\varrho^{K_d^Z}\left(\Phi_{k;\epsilon}^{Z,d}\right)\left(\mathbf{vec}(\mathbf{A})\right)\right)\right\|_2 \leq \epsilon$,

(v) für alle $\mathbf{vec}(\mathbf{A}) \in K_d^Z$ bekommen wir

$$\left\|\mathbf{matr}\left(\mathcal{R}_\varrho^{K_d^Z}\left(\Phi_{k;\epsilon}^{Z,d}\right)\left(\mathbf{vec}(\mathbf{A})\right)\right)\right\|_2 \leq \epsilon + \left\|\mathbf{A}^k\right\|_2 \leq \epsilon + \|\mathbf{A}\|_2^k .$$

Nun zum Beweis. Sei $\left(\left(\mathbf{A}_1, \mathbf{b}_1\right), ..., \left(\mathbf{A}_L, \mathbf{b}_L\right)\right) := \Phi_{k;\frac{\epsilon}{2max\{1,Z^k\}}}^{\frac{1}{2},d}$ um sich auf A.4 zu beziehen. Dann erfüllt folgendes NN die genannten Eigenschaften

$$\Phi_{k;\epsilon}^{Z,d} := \left(\left(\frac{1}{2Z}\mathbf{A}_1, \mathbf{b}_1\right), \left(\mathbf{A}_2, \mathbf{b}_2\right), ..., \left(\mathbf{A}_{L-1}, \mathbf{b}_{L-1}\right), \left(2Z^k\mathbf{A}_L, 2Z^k\mathbf{b}_L\right)\right).$$

Es wurden schon NNs konstruiert, welche eine Matrix als Input nehmen und eine Potenz dieser berechnet. Mit dem Wissen kann nun Theorem 3.8 bewiesen werden.

Beweis. Mit den Eigenschaften von Partialsummen und der Neumannreihe gilt für $m \in \mathbb{N}$ und für jedes $\mathbf{vec}(\mathbf{A}) \in K_d^{1-\delta}$

$$\begin{aligned}\left\|\left(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A}\right)^{-1} - \sum_{k=0}^{m} \mathbf{A}^k\right\|_2 &= \left\|\left(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A}\right)^{-1} \mathbf{A}^{m+1}\right\|_2 \leq \left\|\left(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A}\right)^{-1}\right\|_2 \|\mathbf{A}\|_2^{m+1} \\ &\leq \frac{1}{1-(1-\delta)} \cdot (1-\delta)^{m+1} = \frac{(1-\delta)^{m+1}}{\delta}.\end{aligned}$$

Im nächsten werden nur die zuvor definierte Notation und die Rechenregeln für Logarithmen angewendet und dann gilt

$$m(\epsilon,\delta) = \left\lceil log_{1-\delta}(2) log_2\left(\frac{\epsilon\delta}{2}\right)\right\rceil = \left\lceil \frac{log_2(\epsilon)+log_2(\delta)-1}{log_2(1-\delta)}\right\rceil \geq \frac{log_2(\epsilon)+log_2(\delta)-1}{log_2(1-\delta)}$$

womit folgendes gilt

$$\begin{aligned}\left\|(\mathbf{Id}_{\mathbb{R}^d}-\mathbf{A})^{-1} - \sum_{k=0}^{m(\epsilon,\delta)} \mathbf{A}^k\right\|_2 &= \left\|(\mathbf{Id}_{\mathbb{R}^d}-\mathbf{A})^{-1}\mathbf{A}^{m(\epsilon,\delta)+1}\right\|_2 \\ &\leq \left\|(\mathbf{Id}_{\mathbb{R}^d}-\mathbf{A})^{-1}\right\|_2 \|\mathbf{A}\|_2^{m(\epsilon,\delta)+1} \\ &\leq \frac{1}{1-(1-\delta)}\cdot(1-\delta)^{m(\epsilon,\delta)+1} \\ &= \frac{(1-\delta)^{m(\epsilon,\delta)+1}}{\delta} \leq \frac{\epsilon}{2}\end{aligned}$$

diese Behauptung gilt, da sich mit den Rechenregeln für Logarithmen zeigen lässt

$$\begin{aligned}\frac{(1-\delta)^{m+1}}{\delta} &\leq \frac{(1-\delta)^{\log_{1-\delta}(2)\log_2(\epsilon\delta/2)}}{\delta} \\ &= \frac{2^{\log_2(\epsilon\delta/2)}}{\delta} = \frac{\epsilon\delta}{2\delta} = \frac{\epsilon}{2}.\end{aligned}$$

Nun sei

$$\begin{aligned}&((\mathbf{A}_1,\mathbf{b}_1),...,(\mathbf{A}_L,\mathbf{b}_L)) \\ &\qquad := \left(\left(\left(\mathbf{Id}_{\mathbb{R}^{d^2}}|...|\mathbf{Id}_{\mathbb{R}^{d^2}}\right),\mathbf{0}_{\mathbb{R}^{d^2}}\right)\right) \\ &\qquad \odot P\left(\Phi^{1,d}_{1;\frac{\epsilon}{2(m(\epsilon,\delta)-1)}},...,\Phi^{1,d}_{m(\epsilon,\delta);\frac{\epsilon}{2(m(\epsilon,\delta)-1)}}\right) \\ &\qquad \bullet \left(\left(\begin{pmatrix}\mathbf{Id}_{\mathbb{R}^{d^2}}\\ \vdots \\ \mathbf{Id}_{\mathbb{R}^{d^2}}\end{pmatrix},\mathbf{0}_{\mathbb{R}^{2m(\epsilon,\delta)d^2}}\right)\right),\end{aligned}$$

wobei $\left(\left(\left(\mathbf{Id}_{\mathbb{R}^{d^2}}|...|\mathbf{Id}_{\mathbb{R}^{d^2}}\right),\mathbf{0}_{\mathbb{R}^{d^2}}\right)\right) \in \mathbb{R}^{d^2 \mathtt{x} m(\epsilon,\delta)\cdot d^2}$. Nun definiere

$$\Phi^{1-\delta,d}_{inv;\epsilon} := ((\mathbf{A}_1,\mathbf{b}_1),...,(\mathbf{A}_L,\mathbf{b}_L+\mathbf{vec}(\mathbf{Id}_{\mathbb{R}^d}))).$$

Um (iv) zu beweisen, wird genau wie in den Abschnitten zuvor vorgegangen, man rechnet erst +0 um dann die Dreiecksungleichung zu benutzen. Hier wird wieder vor-

ausgesetzt, dass $\mathbf{vec}(\mathbf{A}) \in K_d^{1-\delta}$. Daraus lässt sich dann wie zuvor analog (v) beweisen.

$$\begin{aligned}
&\left\|(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A})^{-1} - \mathbf{matr}\left(\mathcal{R}_\varrho^{K_d^{1-\delta,d}}\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \\
&\leq \left\|(\mathbf{Id}_{\mathbb{R}^d} - \mathbf{A})^{-1} - \sum_{k=0}^{m(\epsilon,\delta)} \mathbf{A}^k\right\|_2 \\
&+ \left\|\sum_{k=0}^{m(\epsilon,\delta)} \mathbf{A}^k - \mathbf{matr}\left(\mathcal{R}_\varrho^{K_d^{1-\delta,d}}\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \\
&\leq \frac{\epsilon}{2} + \sum_{k=2}^{m(\epsilon,\delta)} \left\|\mathbf{A}^k - \mathbf{matr}\left(\mathcal{R}_\varrho^{K_d^{1-\delta,d}}\left(\Phi_{k;\frac{\epsilon}{2(m(\epsilon,\delta)-1}}^{1,d}\right)(\mathbf{vec}(\mathbf{A}))\right)\right\|_2 \\
&\leq \frac{\epsilon}{2} + (m(\epsilon,\delta)-1)\frac{\epsilon}{2(m(\epsilon,\delta)-1)} = \epsilon,
\end{aligned}$$

wobei hierfür genutzt wurde, dass die Summe für $k = 0, 1$ null ist. Damit ist der Beweis für (iv) komplett.
Jetzt geht es um die Größe des entstehenden NNs. Zu erst bekommt man durch Verwendung von Lemma 3.6(b)(i) und unserer Konstruktion zu Anfang die erste Gleichung. In der ersten Ungleichung wird A.5(i) verwendet, da unser Φ alle Eigenschaften erfüllt, danach wird $m(\epsilon, \delta)$ eingesetzt und es gilt folgende Abschätzung

$$\begin{aligned}
L\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right) &= \max_{k=1,\dots,m(\epsilon,\delta} L\left(\Phi_{k;\frac{\epsilon}{2(m(\epsilon,\delta)-1}}^{1,d}\right) \\
&\leq C_{pow} log_2(m(\epsilon,\delta)-1) \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + 1 + log_2(m(\epsilon,\delta)-1) + log_2(d)\right) \\
&\leq C_{pow} log_2\left(\frac{log_2(0.5\epsilon\delta)}{log_2(1-\delta)}\right) \\
&\cdot \left(log_2\left(\frac{1}{\epsilon}\right) + 1 + log_2\left(\frac{log_2(0.5\epsilon\delta)}{log_2(1-\delta)}\right) + log_2(d)\right),
\end{aligned}$$

damit ist (i) gezeigt. (ii) wird bewiesen, indem zu erst 3.6(b)(ii) benutzt und im zweiten Schritt die Monotonie des Logarithmus und A.5(ii) verwendet wird, daraus ergibt sich

dann

$$
\begin{aligned}
M\left(\Phi_{inv;\epsilon}^{1-\delta,d}\right) &\leq 3 \cdot \left(\sum_{k=1}^{m(\epsilon,\delta)} M\left(\Phi_{k;\frac{\epsilon}{2(m(\epsilon,\delta)-1}}^{1,d}\right)\right) \\
&\quad + 4C_{pow} m(\epsilon,\delta) d^2 log_2(m(\epsilon,\delta)) \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + 1 + log_2(m(\epsilon,\delta)) + log_2(d)\right) \\
&\leq 3C_{pow} \cdot \left(\sum_{k=1}^{m(\epsilon,\delta)} log_2^2(max\{k,2\})\right) d^3 \\
&\quad \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + 1 + log_2(m(\epsilon,\delta)) + log_2(d)\right) \\
&\quad + 5m(\epsilon,\delta) d^2 C_{pow} log_2(m(\epsilon,\delta)) \\
&\quad \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + 1 + log_2(m(\epsilon,\delta)) + log_2(d)\right) =: I.
\end{aligned}
$$

Da $\sum_{k=1}^{m(\epsilon,\delta)} log_2^2(max\{k,2\}) \leq m(\epsilon,\delta) log_2^2(m(\epsilon,\delta))$ ergibt sich für eine Konstante $C_{inv} > C_{pow}$, dass gilt

$$
I \leq C_{inv} m(\epsilon,\delta) log_2^2(m(\epsilon,\delta)) d^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2(m(\epsilon,\delta)) + log_2(d)\right).
$$

Das beweist (ii). Der dritte Teil der Behauptung lässt sich aus der zweiten ableiten. □

6.4 Erster Teil Beweis von Theorem 4.3

In diesem Beweis gibt es wenig mathematische Schritte, die hinzugefügt werden können, da fast überall die genaue Erläuterung der Schritte dabei steht. Ich werde versuchen, falls es sinnige Schritte gibt, diese zu ergänzen und die angewendeten Theoreme, Definitionen und Annahmen direkt auf die einzelnen Schritte zu beziehen, damit nicht der Überblick verloren wird. Nun weiter mit dem Beweis von Theorem 4.3.

Zuerst wird eine Beschränkung für $\left\| \mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon})}} - \alpha \mathbf{B}_{y,\tilde{\epsilon}}^{rb} \right\|_2$ eingeführt.
Proposition B.1 Für alle $\alpha \in (0, \frac{1}{C_{cont}})$ und $\delta := \alpha C_{coer} \in (0,1)$ gilt

$$
\left\| \mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon})}} - \alpha \mathbf{B}_{y,\tilde{\epsilon}}^{rb} \right\|_2 \leq 1 - \delta < 1 \ \forall \ y \in \mathcal{Y}, \tilde{\epsilon} > 0.
$$

Beweis. Da $\mathbf{B}_{y,\tilde{\epsilon}}^{rb}$ symmetrisch ist gilt nun

$$\left\|\mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon})}} - \alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right\|_2 = \max_{\mu\in\sigma\left(\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)} |1-\alpha\mu| \leq \max_{\mu\in[C_{coer},C_{cont}]} |1-\alpha\mu|$$
$$= 1 - \alpha C_{coer} = 1 - \delta < 1,$$

für alle $y \in \mathbb{Y}, \tilde{\epsilon} > 0$.
Mit einer Approximation an die parameterabhängige Steifigkeitsmatrix bezüglich der reduzierten Basis kann mit Annahme 4.1 eine Konstruktion der Realisierung eines NNs formuliert werden, dessen Approximation die Abbildung $y \mapsto \left(\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}$ darstellt. Zu erst werden folgende Bemerkungen benötigt.

Bemerkung B.2. Es ist nicht schwer zu sehen, dass wenn $\left(\left(\mathbf{A}_{\tilde{\epsilon},\epsilon}^1, \mathbf{b}_{\tilde{\epsilon},\epsilon}^1\right), ..., \left(\mathbf{A}_{\tilde{\epsilon},\epsilon}^L, \mathbf{b}_{\tilde{\epsilon},\epsilon}^L\right)\right)$ $:= \Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{B}}$ das NN aus Annahme 4.1 ist, dann

$$\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{B},\ \mathbf{Id}} := \left(\left(\mathbf{A}_{\tilde{\epsilon},\epsilon}^1, \mathbf{b}_{\tilde{\epsilon},\epsilon}^1\right), ..., \left(-\mathbf{A}_{\tilde{\epsilon},\epsilon}^L, -\mathbf{b}_{\tilde{\epsilon},\epsilon}^L + \mathbf{vec}\left(\mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon})}}\right)\right)\right)$$

bekommt man folgende Abschätzung, welche später noch relevant wird

$$\sup_{y\in\mathbb{Y}} \left\|\mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon})}} - \alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb} - \mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{B},\ \mathbf{Id}}\right)(y)\right)\right\|_2 \leq \epsilon,$$

genau so wie $M\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{B},\ \mathbf{Id}}\right) \leq B_M(\tilde{\epsilon},\epsilon) + d(\tilde{\epsilon})^2, L\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{B},\ \mathbf{Id}}\right) = B_L(\tilde{\epsilon},\epsilon)$. Nun wird sich eine Konstruktion angeschaut, die die Abbildung $y \mapsto \left(\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}$ darstellt.

Proposition B.3. Sei $\tilde{\epsilon} \geq \hat{\epsilon}, \epsilon \in \left(0, \frac{\alpha}{4} \cdot min\{1, C_{coer}\}\right)$ und $\epsilon' := \frac{3}{8}\epsilon\alpha C_{coer}^2 < \epsilon$. Angenommen, dass die Annahme 4.1 gilt. Dann definiere ein NN

$$\Phi_{inv;\tilde{\epsilon},\epsilon}^{\mathbf{B}} := \left(\left(\alpha\mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon})}}, \mathbf{0}_{\mathbb{R}^{d(\tilde{\epsilon})}}\right)\right) \bullet \Phi_{inv;\frac{\epsilon}{2\alpha}}^{\frac{1-\delta}{2},d(\tilde{\epsilon}} \odot \Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B},\mathbf{Id}},$$

mit p-dimensionalem Input und $d(\tilde{\epsilon})^2$ dimensionalem Output. Dann existiert eine Konstante $C_B = C_B\left(C_{coer}, C_{cont}\right) > 0$, so dass folgende Eigenschaften gelten

(i)

$$L\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right) \leq C_B log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right) \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right) + log_2\left(d\left(\tilde{\epsilon}\right)\right)\right) + B_L(\tilde{\epsilon}, \epsilon^{'})$$

(ii)

$$M\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right) \leq C_B log_2\left(\frac{1}{\epsilon}\right) log_2^2\left(log_2\left(\frac{1}{\epsilon}\right)\right) d(\tilde{\epsilon})^3 \cdot \left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right) + log_2\left(d\left(\tilde{\epsilon}\right)\right)\right) + 2B_M(\tilde{\epsilon}.\epsilon^{'})$$

(iii)

$$\sup_{y\in\mathcal{Y}} \left\| \left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1} - \mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right)(y)\right)\right\|_2 \leq \epsilon,$$

(iv)

$$\sup_{y\in\mathcal{Y}} \left\| \mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}} \cdot \left(\left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1} - \mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right)(y)\right)\right)\right\|_2 \leq \epsilon,$$

(v)

$$\sup_{y\in\mathcal{Y}} \left\| \mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right)(y)\right)\right\|_2 \leq \epsilon\frac{1}{C_{coer}},$$

(vi)

$$\sup_{y\in\mathcal{Y}} \left\| \mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}}\mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right)(y)\right)\right\|_2 \leq \epsilon\frac{1}{C_{coer}}.$$

Nun zum Beweis. Zu erst gilt, dass die Matrix $\mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{\tilde{\epsilon},\epsilon'}\right)(y)\right)$ für alle $y \in \mathcal{Y}$ invertierbar ist. Das kann aus folgender Abschätzung abgeleitet werden.

$$\left\|\alpha\mathbf{B}^{rb}_{y,\tilde{\epsilon}} - \mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{\tilde{\epsilon},\epsilon'}\right)(y)\right)\right\|_2 \leq \epsilon^{'} < \epsilon \leq \frac{\alpha min\{1, C_{coer}\}}{4} \leq \frac{\alpha C_{coer}}{4}. \quad (6.17)$$

Tatsächlich kann abgeschätzt werden, dass mit der umgekehrten Dreiecksungleichung gilt

$$\begin{aligned}
&\min_{\mathbf{z}\in\mathbb{R}^{d(\tilde{\epsilon})}ohne\{0\}} \frac{\left|\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\mathbf{z}\right|}{|\mathbf{z}|} \\
&\geq \min_{\mathbf{z}\in\mathbb{R}^{d(\tilde{\epsilon})}ohne\{0\}} \frac{\left|\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\mathbf{z}\right|}{|\mathbf{z}|} - \max_{\mathbf{z}\in\mathbb{R}^{d(\tilde{\epsilon})}ohne\{0\}} \frac{\left|\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\mathbf{z} - \mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\mathbf{z}\right|}{|\mathbf{z}|}
\end{aligned}$$

für die nächsten beiden Schritte nutze die Definition der 2-Norm und setze danach $\tilde{\mathbf{z}} := \left(\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)\mathbf{z}$. Daraus resultiert dann, dass dies

$$\begin{aligned}
&\geq \left(\max_{\mathbf{z}\in\mathbb{R}^{d(\tilde{\epsilon})}ohne\{0\}} \frac{|\mathbf{z}|}{\left|\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\mathbf{z}\right|}\right)^{-1} - \left\|\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb} - \mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right\|_2 \\
&\geq \left(\max_{\tilde{\mathbf{z}}\in\mathbb{R}^{d(\tilde{\epsilon})}ohne\{0\}} \frac{\left|\left(\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}\tilde{\mathbf{z}}\right|}{|\tilde{\mathbf{z}}|}\right)^{-1} - \left\|\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb} - \mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right\|_2 \\
&\geq \left\|\left(\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}\right\|_2^{-1} - \left\|\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb} - \mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right\|_2.
\end{aligned}$$

wenn nun die Definition von (6.17) und (2.9) genutzt wird, gibt das schlussendlich

$$\geq \alpha C_{coer} - \frac{\alpha C_{coer}}{4} \geq \frac{3}{4}\alpha C_{coer}.$$

Dann folgt damit

$$\left\|\left(\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right)^{-1}\right\|_2 \leq \frac{4}{3}\frac{1}{C_{coer}\alpha}.$$

Durch das addieren von 0 und der Dreiecksungleichung gilt dann

$$\begin{aligned}
&\left\|\frac{1}{\alpha}\left(\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1} - \mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{inv;\frac{\epsilon}{2\alpha}}^{1-\frac{\delta}{2},d(\tilde{\epsilon})} \odot \Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B},\mathbf{Id}}\right)(y)\right)\right\|_2 \\
&\leq \left\|\frac{1}{\alpha}\left(\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1} - \left(\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right)^{-1}\right\|_2 \\
&+ \left\|\left(\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right)^{-1} - \mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{inv;\frac{\epsilon}{2\alpha}}^{1-\frac{\delta}{2},d(\tilde{\epsilon})} \odot \Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B},\mathbf{Id}}\right)(y)\right)\right\|_2 =: I + II.
\end{aligned}$$

Mit der Eigenschaft für zwei invertierbare Matrizen **M,N** gilt

$$\left\|\mathbf{M}^{-1}-\mathbf{N}^{-1}\right\|_2=\left\|\mathbf{M}^{-1}\left(\mathbf{N}-\mathbf{M}\right)\mathbf{N}^{-1}\right\|_2\leq\left\|\mathbf{M}-\mathbf{N}\right\|_2\left\|\mathbf{M}^{-1}\right\|_2\left\|\mathbf{N}^{-1}\right\|_2$$

und für I gibt das durch das Benutzen von Annahme 4.1, (2.9) und der Gleichung für zwei invertierbare Matrizen

$$\begin{aligned}I&\leq\left\|\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}-\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right\|_2\left\|\left(\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}\right\|_2\left\|\left(\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right)^{-1}\right\|_2\\&\leq\frac{3}{8}\epsilon\alpha C_{coer}^2\frac{1}{\alpha C_{coer}}\frac{4}{3}\frac{1}{C_{coer}\alpha}=\frac{\epsilon}{2\alpha}.\end{aligned}$$

Jetzt berechne II. Als erstes kann mit der Dreiecksungleichung und Bemerkung B.2 angenommen werden, dass für alle $y\in\mathcal{Y}$ gilt

$$\begin{aligned}\left\|\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)\right\|_2&\leq\left\|\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon'}^{\mathbf{B}}\right)(y)\right)-\left(\mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon}}}-\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)\right\|_2\\&\quad+\left\|\mathbf{Id}_{\mathbb{R}^{d(\tilde{\epsilon}}}-\alpha\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right\|_2\\&\leq\epsilon^{'}+1-\delta\leq1-\delta+\frac{\alpha C_{coer}}{4}\leq1-\delta+\frac{\alpha C_{cont}}{4}\\&\leq1-\delta+\frac{\delta}{2}=1-\frac{\delta}{2}.\end{aligned}$$

Da nun $\frac{\epsilon}{2\alpha}=\frac{\frac{\alpha C_{coer}}{4}}{2\alpha}=\frac{C_{coer}}{8}<\frac{1}{8}<\frac{1}{4}$. Somit wird mit Hilfe des Beweises von Korollar A.5(iv) gezeigt, dass $II\leq\frac{\epsilon}{2\alpha}$ ist. Zusammengesetzt gibt das dann $I+II\leq\frac{\epsilon}{\alpha}$. Durch die Konstruktion von Φ kann direkt daraus geschlossen werden, dass (iii) gilt. Nutze nun Gleichung (2.7) und multipliziere sie mit (iii), daraus resultiert

$$\sup_{y\in\mathcal{Y}}\left\|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}}\left(\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}-\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}}\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{inv;\tilde{\epsilon},\epsilon}^{\mathbf{B}}\right)(y)\right)\right\|_2\leq\epsilon,$$

und somit ist (iv) bewiesen. Für jedes $y\in\mathcal{Y}$ gilt nun folgende Abschätzung

$$\begin{aligned}&\left\|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}}\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{inv;\tilde{\epsilon},\epsilon}^{\mathbf{B}}\right)(y)\right)\right\|_2\\&\leq\left\|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}}\cdot\left(\left(\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}-\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{inv;\tilde{\epsilon},\epsilon}^{\mathbf{B}}\right)(y)\right)\right)\right\|_2\\&+\left\|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}}\left(\mathbf{B}_{y,\tilde{\epsilon}}^{rb}\right)^{-1}\right\|_2\leq\epsilon+\frac{1}{C_{coer}}.\end{aligned}$$

und haben damit (vi) bewiesen. (v) beweist man einfach indem man (iii) + $\left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1}$ rechnet und umstellt wie in den Abschnitten zuvor. Nun zu (i) und (ii). Mit der Konstruktion für Φ gilt $L\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right) = L\left(\Phi^{1-\frac{\delta}{2},d(\tilde{\epsilon})}_{inv;\frac{\epsilon}{2\alpha}} \odot \Phi^{\mathbf{B},\mathbf{Id}}_{\tilde{\epsilon},\epsilon'}\right), M\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right) = M\left(\Phi^{1-\frac{\delta}{2},d(\tilde{\epsilon})}_{inv;\frac{\epsilon}{2\alpha}} \odot \Phi^{\mathbf{B},\mathbf{Id}}_{\tilde{\epsilon},\epsilon'}\right)$. Außerdem, wenn nun erst Lemma 3.6(a)(i) benutzt und dann Theorem 3.8(i) und Bemerkung B.2 verwendet wird, lässt sich (i) der Behauptung beweisen.

$$\begin{aligned} L\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right) &\leq L\left(\Phi^{1-\frac{\delta}{2},d(\tilde{\epsilon})}_{inv;\frac{\epsilon}{2\alpha}}\right) + L\left(\Phi^{\mathbf{B},\mathbf{Id}}_{\tilde{\epsilon},\epsilon'}\right) \\ &\leq C_{inv} log_2\left(m\left(\frac{\epsilon}{2\alpha},\frac{\delta}{2}\right)\right) \\ &\quad \cdot \left(log_2\left(\frac{2\alpha}{\epsilon}\right) + log_2\left(m\left(\frac{\epsilon}{2\alpha},\frac{\delta}{2}\right)\right) + log_2(d(\tilde{\epsilon}))\right) + B_L(\tilde{\epsilon},\epsilon'). \end{aligned}$$

Um (ii) der Behauptung zu beweisen, gehe genau so vor, mit dem Unterschied, dass immer die zweite Annahme von 3.6 und 3.8 genutzt wird.

$$\begin{aligned} M\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon}\right) &\leq M\left(\Phi^{1-\frac{\delta}{2},d(\tilde{\epsilon})}_{inv;\frac{\epsilon}{2\alpha}}\right) + M\left(\Phi^{\mathbf{B},\mathbf{Id}}_{\tilde{\epsilon},\epsilon'}\right) + M_{L\left(\Phi^{\mathbf{B},\mathbf{Id}}_{\tilde{\epsilon},\epsilon'}\right)}\left(\Phi^{\mathbf{B},\mathbf{Id}}_{\tilde{\epsilon},\epsilon'}\right) \\ &\leq 2M\left(\Phi^{1-\frac{\delta}{2},d(\tilde{\epsilon})}_{inv;\frac{\epsilon}{2\alpha}}\right) + 2M\left(\Phi^{\mathbf{B},\mathbf{Id}}_{\tilde{\epsilon},\epsilon'}\right) \\ &\leq 2C_{inv} m\left(\frac{\epsilon}{2\alpha},\frac{\delta}{2}\right) log_2^2\left(m\left(\frac{\epsilon}{2\alpha},\frac{\delta}{2}\right)\right) d(\tilde{\epsilon})^3 \\ &\quad \cdot \left(log_2\left(\frac{2\alpha}{\epsilon}\right) + log_2\left(m\left(\frac{\epsilon}{2\alpha},\frac{\delta}{2}\right)\right) + log_2(d(\tilde{\epsilon}))\right) + 2d(\tilde{\epsilon})^2 + 2B_M(\tilde{\epsilon},\epsilon'). \end{aligned}$$

Zusätzlich durch die Definition von $m(\epsilon,\delta)$ in Theorem 3.8 gilt für eine Konstante $\tilde{C} > 0, m\left(\frac{\epsilon}{2\alpha},\frac{\delta}{2}\right) \leq \tilde{C} log_2\left(\frac{1}{\epsilon}\right)$. Somit folgt, wenn eine passende Konstante $C_B = C_B\left(C_{coer}, C_{cont}\right) > 0$ gewählt wird, resultiert dadurch (ii) der Behauptung. □

Beweis von Theorem 4.3

Zu erst wird (i) bewiesen, nur der Part für $\Phi^{\mathbf{u},h}_{\tilde{\epsilon},\epsilon}$, denn für $\Phi^{\mathbf{u},rb}_{\tilde{\epsilon},\epsilon}$ funktioniert dies genau so, nur ohne dass wir (2.7) nutzen. Also, für $y \in \mathcal{Y}$ und durch zweimaliges $+0$ rechnen (mit Hilfe der Konstruktion von Φ aus Bemerkung 4.4), anwenden der

Dreiecksungleichung und der Definition aus Kapitel 2 für $\tilde{\mathbf{u}}$ bekommt man

$$
\begin{aligned}
&\left|\tilde{\mathbf{u}}^h_{y,\tilde{\epsilon}} - \mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{u},h}_{\tilde{\epsilon},\epsilon}\right)(y)\right|_{\mathbf{G}} \\
&= \left|\mathbf{G}^{\frac{1}{2}} \cdot \left(\mathbf{V}_{\tilde{\epsilon}}\left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1}\mathbf{f}^{rb}_{y,\tilde{\epsilon}} - \mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{u},h}_{\tilde{\epsilon},\epsilon}\right)(y)\right)\right| \\
&\leq \left|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}} \cdot \left(\left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1}\mathbf{f}^{rb}_{y,\tilde{\epsilon}} - \left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1}\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{f}}_{\tilde{\epsilon},\epsilon''}\right)(y)\right)\right| \\
&+ \left|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}} \cdot \left(\left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1}\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{f}}_{\tilde{\epsilon},\epsilon''}\right)(y) - \mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'}\right)(y)\right)\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{f}}_{\tilde{\epsilon},\epsilon''}\right)(y)\right)\right| \\
&+ \left|\mathbf{G}^{\frac{1}{2}} \cdot \left(\mathbf{V}_{\tilde{\epsilon}}\mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'}\right)(y)\right)\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{f}}_{\tilde{\epsilon},\epsilon''}\right)(y) - \mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{u},h}_{\tilde{\epsilon},\epsilon}\right)(y)\right)\right| \\
&=: I + II + III.
\end{aligned}
$$

Nun berechne diese drei Gleichungen. Mit (2.7), (2.9), Annahme 4.2 und der Definition von $\epsilon^{''}$ für alle $y \in \mathcal{Y}$ gilt

$$I \leq \frac{1}{C_{coer}}\frac{\epsilon C_{coer}}{3} = \frac{\epsilon}{3}.$$

Nun lässt sich schnell sehen, dass mit Annahme 4.2 und der Definition von f folgende Abschätzung gelten muss

$$\sup_{y\in\mathcal{Y}}\left|\mathbb{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{f}_{\tilde{\epsilon},\epsilon}\right)\right| \leq \epsilon + C_{rhs}. \tag{6.18}$$

Wegen der Definition $\epsilon^{'} = \frac{\epsilon}{max\{6,C_{rhs}\}} \leq \epsilon$. Durch Benutzen von Annahme 4.1 und (6.18) gibt das in Kombination mit Proposition B.3(i) und der Submultiplikativitätseigenschaft

$$
\begin{aligned}
II &= \left|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}} \cdot \left(\left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1}\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{f}}_{\tilde{\epsilon},\epsilon''}\right)(y)\right.\right. \\
&\left.\left. - \mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'}\right)(y)\right)\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{f}}_{\tilde{\epsilon},\epsilon''}\right)(y)\right)\right| \\
&\leq \left\|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}} \cdot \left(\left(\mathbf{B}^{rb}_{y,\tilde{\epsilon}}\right)^{-1} - \mathbf{matr}\left(\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'}\right)(y)\right)\right)\right\|_2 \left|\mathcal{R}^{\mathcal{Y}}_{\varrho}\left(\Phi^{\mathbf{f}}_{\tilde{\epsilon},\epsilon''}\right)(y)\right| \\
&\leq \epsilon^{'} \cdot \left(C_{rhs} + \frac{\epsilon C_{coer}}{3}\right) \leq \frac{\epsilon C_{rhs}}{max\{6,C_{rhs}\}} + \frac{\epsilon C_{coer}}{max\{6,C_{rhs}\}}\frac{\epsilon}{3} \leq \frac{2\epsilon}{6} = \frac{\epsilon}{3},
\end{aligned}
$$

wobei $C_{coer}\epsilon < C_{coer}\frac{\alpha}{4} < 1$ genutzt wurde. Um die dritte Gleichung abzuschätzen, wurde die Konstruktion von Φ aus Bemerkung 4.4 verwendet. Wenn noch zusätzlich

Proposition B.3(v) benutzt wird, ergibt das

$$\left\|\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{inv;\tilde{\epsilon},\epsilon}^{\mathbf{B}}\right)(y)\right)\right\|_2 \leq \epsilon + \frac{1}{C_{coer}} \leq 1 + \frac{1}{C_{coer}} \leq \mathcal{K}$$

und mit (6.18)

$$\left|\mathbb{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{f}\right)\right| \leq \epsilon^{''} + C_{rhs} \leq \epsilon C_{coer} + C_{rhs} \leq 1 + C_{rhs} \leq \mathcal{K}.$$

Und damit kann dann, wenn die beiden Abschätzungen zuvor genommen werden, III berechnet werden

$$\begin{aligned}
III &= \left|\mathbf{G}^{\frac{1}{2}} \cdot \left(\mathbf{V}_{\tilde{\epsilon}}\mathbf{matr}\left(\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{inv;\tilde{\epsilon},\epsilon^{'}}^{\mathbf{B}}\right)(y)\right)\mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon^{''}}^{\mathbf{f}}\right)(y) - \mathcal{R}_{\varrho}^{\mathcal{Y}}\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}\right)(y)\right)\right| \\
&\leq \left|\mathbf{G}^{\frac{1}{2}} \cdot (\mathbf{V}_{\tilde{\epsilon}}\mathcal{K}\mathcal{K}) - \mathcal{K}\right| \\
&\leq \left\|\mathbf{G}^{\frac{1}{2}}\mathbf{V}_{\tilde{\epsilon}}\right\|_2 \|\mathcal{K}\mathcal{K} - \mathcal{K}\|_2 .
\end{aligned}$$

Wenn $\mathcal{K}$ nun richtig gewählt wird, gibt dies wie gewünscht $III \leq \frac{\epsilon}{3}$ und damit ist (i) bewiesen. Eigenschaft (v) der Behauptung lässt sich ganz leicht aus (i) beweisen und wird deshalb hier nicht weiter ausgeführt. Nun wieder zur Größe des NNs. Zu erst (ii). Mit der Definition von $\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}, \Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}$ sowie Lemma 3.6 (a)(i), Proposition 3.7 und Prop B.3(i) im dritten Schritt und einer passenden Konstante $C_L^u = C_L^u(\mathcal{K}, \epsilon^{'}, C_B) = C_L^u(C_{rhs}, C_{coer}, C_{cont}) > 0$, erhalten wir

$$\begin{aligned}
L\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},rb}\right) \leq L\left(\Phi_{\tilde{\epsilon},\epsilon}^{\mathbf{u},h}\right) &\leq 1 + L\left(\Phi_{mult;\frac{\epsilon}{3}}^{\mathcal{K},d(\tilde{\epsilon}),d(\tilde{\epsilon}),1}\right) + L\left(P\left(\Phi_{inv;\tilde{\epsilon},\epsilon^{'}}^{\mathbf{B}}, \Phi_{inv;\tilde{\epsilon},\epsilon^{''}}^{\mathbf{f}}\right)\right) \\
&\leq 1 + C_{mult} \cdot \left(log_2\left(\frac{3}{\epsilon}\right) + \frac{3}{2}log_2(d(\tilde{\epsilon})) + log_2(\mathcal{K})\right) \\
&+ max\left\{L\left(\Phi_{inv;\tilde{\epsilon},\epsilon^{'}}^{\mathbf{B}}\right), F_L\left(\tilde{\epsilon}, \epsilon^{''}\right)\right\} \\
&\leq C_L^{\mathbf{u}} max\left\{log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right)\left(log_2\left(\frac{1}{\epsilon}\right) + log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right) + log_2(d(\tilde{\epsilon}))\right)\right. \\
&\left. + B_L\left(\tilde{\epsilon}, \epsilon^{'''}\right), F_L\left(\tilde{\epsilon}, \epsilon^{''}\right)\right\}.
\end{aligned}$$

Halte fest, dass wenn (iii) bewiesen ist, (iv) direkt daraus folgt. Um jetzt (iii) zu beweisen verwendet man Lemma 3.6(a)(ii) in Kombination mit Proposition 3.7

$$\begin{aligned} M\left(\Phi^{\mathbf{u},rb}_{\tilde{\epsilon},\epsilon}\right) &\leq 2M\left(\Phi^{\mathcal{K},d(\tilde{\epsilon}),d(\tilde{\epsilon}),1}_{mult;\frac{\epsilon}{3}}\right) + 2M\left(P\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'},\Phi^{\mathbf{f}}_{inv;\tilde{\epsilon},\epsilon''}\right)\right) \\ &\leq 2C_{mult}d(\tilde{\epsilon})^2\cdot\left(log_2\left(\frac{3}{\epsilon}\right)+\frac{3}{2}log_2(d(\tilde{\epsilon}))+log_2(\mathcal{K})\right) \\ &+ 2M\left(P\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'},\Phi^{\mathbf{f}}_{inv;\tilde{\epsilon},\epsilon''}\right)\right). \end{aligned}$$

Der zweite Teil der Gleichung wird wie folgt berechnet. Zu erst wird Lemma 3.6(b)(ii) in Kombination mit Proposition B.3 benutzt und dann die Annahmen 4.1 und 4.2, dann gilt

$$\begin{aligned} &M\left(P\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'},\Phi^{\mathbf{f}}_{inv;\tilde{\epsilon},\epsilon''}\right)\right) \\ &\leq M\left(\Phi^{\mathbf{B}}_{inv;\tilde{\epsilon},\epsilon'}\right)+M\left(\Phi^{\mathbf{f}}_{inv;\tilde{\epsilon},\epsilon''}\right) \\ &+ 8d(\tilde{\epsilon})^2 max\{C^{\mathbf{u}}_L log_2\left(log_2\left(\frac{1}{\epsilon'}\right)\right)\left(log_2\left(\frac{1}{\epsilon'}\right)+log_2\left(log_2\left(\frac{1}{\epsilon'}\right)\right)+log_2(d(\tilde{\epsilon}))\right) \\ &+ B_L\left(\tilde{\epsilon},\epsilon^{'''}\right),F_L\left(\tilde{\epsilon},\epsilon^{''}\right)\} \\ &\leq C_B log_2\left(\frac{1}{\epsilon'}\right)log_2^2\left(log_2\left(\frac{1}{\epsilon'}\right)\right)d(\tilde{\epsilon})^3\cdot\left(log_2\left(\frac{1}{\epsilon'}\right)+log_2\left(log_2\left(\frac{1}{\epsilon'}\right)\right)+log_2(d(\tilde{\epsilon}))\right) \\ &+ 8d(\tilde{\epsilon})^2 max\{C^{\mathbf{u}}_L log_2\left(log_2\left(\frac{1}{\epsilon'}\right)\right)\left(log_2\left(\frac{1}{\epsilon'}\right)+log_2\left(log_2\left(\frac{1}{\epsilon'}\right)\right)+log_2(d(\tilde{\epsilon}))\right) \\ &+ B_L\left(\tilde{\epsilon},\epsilon^{'''}\right),F_L\left(\tilde{\epsilon},\epsilon^{''}\right)\}+B_M\left(\tilde{\epsilon},\epsilon^{'''}\right),F_M\left(\tilde{\epsilon},\epsilon^{''}\right) \\ &\leq C^{\mathbf{u}}_M d(\tilde{\epsilon})^2\cdot\left(d(\tilde{\epsilon})log_2\left(\frac{1}{\epsilon}\right)log_2^2\left(log_2\left(\frac{1}{\epsilon}\right)\right)\left(log_2\left(\frac{1}{\epsilon}\right)+log_2\left(log_2\left(\frac{1}{\epsilon}\right)\right)\right)\right. \\ &\left.+\ log_2(d(\tilde{\epsilon}))\right)...+B_L\left(\tilde{\epsilon},\epsilon^{'''}\right)+F_L\left(\tilde{\epsilon},\epsilon^{''}\right)\Big)+2B_M\left(\tilde{\epsilon},\epsilon^{'''}\right)+F_M\left(\tilde{\epsilon},\epsilon^{''}\right) \end{aligned}$$

indem wie zuvor eine passende Konstante
$C^{\mathbf{u}}_M = C^{\mathbf{u}}_M\left(\epsilon^{'},C_B,C^{\mathbf{u}}_L\right) = C^{\mathbf{u}}_L\left(C_{rhs},C_{coer},C_{cont}\right) > 0$ gewählt wird. Das zeigt letztendlich die Behauptung und beendet auch diese Arbeit. □

7 Quellen

[1] M. Ohlberger and S. Rave. Reduced basis methods: Success, limitations and future challenges. Proceedings of the Conference Algoritmy, pages 1–12, 2016.
[2] G. Rozza, D. B. P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations: application to transport and continuum mechanics. Arch. Comput. Methods Eng., 15(3):229–275, 2008.
[3] C. Prud'Homme, D. Rovas, K. Veroy, L. Machiels, Y. Maday, A. Patera, and G. Turinici. Reduced–basis output bound methods for parametrized partial differential equations. In Proceedings SMA Symposium, volume 1, page 1, 2002.
[4] D. Yarotsky. Error bounds for approximations with deep ReLU networks. Neural Netw., 94:103–114, 2017.
[5] S. Mallat. Understanding deep convolutional networks. Phil. Trans. R. Soc. A, 374(2065):20150203, 2016.
[6] H. Mhaskar, Q. Liao, and T. Poggio. Learning functions: when is deep better than shallow. arXiv preprint arXiv:1603.00988, 2016.
[7] C. Canuto, T. Tonn, and K. Urban. A posteriori error analysis of the reduced basis method for nonaffine parametrized nonlinear PDEs. SIAM J. Numer. Anal., 47(3):2001–2022, 2009.
[8] M. Bachmayr, A. Cohen, and G. Migliorati. Sparse polynomial approximation of parametric elliptic PDEs. Part I: Affine coefficients. ESAIM Math. Model. Numer. Anal., 51(1):321–339, 2017.
[9] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. Neural Netw., 2(5):359–366, 1989.
[10] Barbara Burke Hubbard: Wavelets: Die Mathematik der kleinen Wellen. 1. Auflage. Birkhäuser Verlag, 1997.
[11] J. Hesthaven, G. Rozza, and B. Stamm. Certified Reduced Basis Methods for Parametrized Partial Differential Equations. Springer Briefs in Mathematics. Springer, Switzerland, 1 edition, 2015.
[12] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. Acta Numer., 24:1–159, 2015.
[13] A. Quarteroni, A. Manzoni, and F. Negri. Reduced basis methods for partial differential equations, volume 92 of Unitext. Springer, Cham, 2016. An introduction, La Matematica per il 3+2.
[14] W. Dahmen. How to best sample a solution manifold? In Sampling theory, a re-

naissance, Appl. Numer. Harmon. Anal., pages 403–435. Birkh¨auser/Springer, Cham, 2015.
[15] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. SIAM J. Math. Anal., 43(3):1457–1472, 2011.
[16] P. C. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. Neural Netw., 180:296–330, 2018
[17] D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab. DNN expression rate analysis of high-dimensional PDEs: Application to option pricing. arXiv preprint arXiv:1809.07669, 2018.
[18] https://rocketloop.de/de/blog/kuenstliche-neuronale-netze/
[19] J. He, L. Li, J. Xu, and C. Zheng. ReLU deep neural networks and linear finite elements. arXiv preprint arXiv:1807.03973, 2018
[20] M. Telgarsky. Neural networks and rational functions. In 34th International Conference on Machine Learning, ICML 2017, volume 7, pages 5195–5210. International Machine Learning Society (IMLS), 1 2017
[21] I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org
[22] N. Dal Santo, S. Deparis, and L. Pegolotti. Data driven approximation of parametrized PDEs by Reduced Basis and Neural Networks. arXiv preprint arXiv:1904.01514, 2019.
[23] V. Strassen. Gaussian elimination is not optimal. Numer. Math., 13(4):354–356, 1969
[24] F. Cucker and S. Smale. On the mathematical foundations of learning. Bull. Am. Math. Soc., 39:1–49, 2002
[25] https://arxiv.org/pdf/1904.00377.pdf
[26] https://datascience.eu/de/maschinelles-lernen/relu-aktivierungsfunktion/
[27] https://en.wikipedia.org/wiki/Gram_matrix
[28] https://www.nld.rwth-aachen.de/go/id/gyzo
[29] https://de.wikipedia.org/wiki/Schauderbasis
[30] https://en.wikipedia.org/wiki/Statistical_learning_theory

8 Eigenständigkeitserklärung

Hiermit versichere ich, Niek Maurits Jung, dass die vorliegende Arbeit „*Eine theoretische Analyse von komplexen neuronalen Netzen und parametrischen partiellen Differentialgleichungen*“ selbständig von mir und ohne fremde Hilfe verfasst worden ist, dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt worden sind und dass die Stellen der Arbeit, die anderen Werken wie auch elektronischen Medien, dem Wortlaut oder Sinn nach entnommen wurden, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht worden sind. Mir ist bekannt, dass es sich bei einem Plagiat um eine Täuschung handelt, die gemäß der Prüfungsordnung sanktioniert werden kann.

Ich erkläre mich mit einem Abgleich der Arbeit mit anderen Texten zwecks Auffindung von Übereinstimmungen sowie mit einer zu diesem Zweck vorzunehmenden Speicherung der Arbeit in einer Datenbank einverstanden.

Ich versichere, dass ich die vorliegende Arbeit oder Teile daraus nicht anderweitig als Prüfungsarbeit abgegeben habe.

Ort, Datum	Unterschrift