Robuste Risiko-Optimierung mit Multi-Objective Neural Networks

Martin Christian Prescher

Der Fakultät für Elektrotechnik und Informationstechnik
der Universität der Bundeswehr München
zur Erlangung des akademischen Grades

Doktor-Ingenieur
(Dr.-Ing.)

vorgelegte Dissertation

Inhaltsverzeichnis

1 Einführung 1
1.1 Überblick & Einordnung der Arbeit 4
1.2 Konventionen und Notation 6

Teil I Deterministische Black-Box Modellierung

2 Nicht-parametrische Regression 11
2.1 Entwicklung nach Basisfunktionen 11
2.1.1 Beispiele für Basisfunktionen 12
2.2 Approximationseigenschaften von Neuronalen Netzen 20
2.2.1 Single-layer-SNNs 21
2.2.2 Multi-layer-SNNs 27
2.2.3 RBFNs 31

3 Redundanz & Komplexität bei Wavelet-Entwicklungen 35
3.1 Diskrete Wavelet-Transformation und Frames 39
3.2 Endliche Rekonstruktionen und quantisierte Phasenräume 76
3.3 Numerische Präzisierung der endlichen Rekonstruktionsformel 87
3.4 Robustheit, Redundanz & Komplexität 93

Teil II Stochastische Black-Box Modellierung

4 Hypothesenräume 99
4.1 Der stochastische Rahmen 99
4.2 Risiko-Minimierung 101
4.2.1 Empirische Risiko-Minimierung 107

4.2.2 Konsistenz 109
4.3 Modellkomplexität 115
4.3.1 Klassische Komplexitätskontrolle 118
4.3.2 Anwendung der VC-Theorie auf Neuronale Netze 123
4.4 Topologische Einschränkungen für Hypothesenräume 127
4.4.1 Kompaktheit & Konsistenz 128
4.4.2 Kompaktheit und das Bias-Variance-Problem 132
4.5 Konvergenz im Hypothesenraum 139

5 Robustheit von Lernproblemen 151
5.1 Stabilität von Algorithmen 151
5.1.1 Hypothesenstabilität 154
5.1.2 Regularisierte Lernprobleme 160
5.2 Einfluss von Ausreißern 162
5.3 Schlechte Kondition des Optimierungsproblems 168
5.4 Konstruktionsalgorithmen mit aufsteigender Komplexität 170
5.4.1 Hypothesen-Stabilität von Standard-greedy-Algorithmen 175

Teil III Anwendung

6 Multi-Objective Neuronale Netze 181
6.1 Run-Time-Robustheit & Kondition 182
6.1.1 Gleichgradig stetige Neuronale Netze 183
6.2 Optimale Neuronale Netze 200
6.2.1 Minimierung des Gesamtfehlers 206
6.2.2 Konditionskontrolle mit $\mathscr{F}_N^*$ 212
6.2.3 Ein neuer Konstruktionsalgorithmus 213
6.2.4 Erweiterung der Ergebnisse auf RBFNs und WNNs 219

7 Zusammenfassung und Ausblick 221

A Demonstratornetzwerk und Beispielprobleme 227

Literaturverzeichnis 239

Danksagung 245

1

Einführung

Ein grundlegendes Problem in verschiedensten Anwendungen ist die Approximation einer unbekannten Funktion

$$f : \mathbb{R}^d \longrightarrow \mathbb{R}^m$$

einzig und allein auf Grundlage von einzelnen Auswertungen (Messungen). In physikalischen, technischen oder auch ökonomischen Anwendungen tritt dieser Fall sehr häufig und immer dann ein, wenn der f zu Grunde liegende Mechanismus so kompliziert ist, dass der Zusammenhang der verschiedenen Einflussgrößen nicht analytisch, sondern nur empirisch ermittelt werden kann.
Zur Illustration ein im Rahmen dieser Dissertation untersuchtes Beispiel: Bestimmt werden soll der Massenfluss y durch ein technisches Bauteil (z.B. eine Kraftstoffpumpe) in Abhängigkeit von mehreren Stellgrößen (zusammengefasst in dem Vektor $\boldsymbol{x}$) wie z.B. Temperatur, Anstellwinkel eines Ventils und einigen Geometrieparamtern der Pumpe.
Eine detaillierte Untersuchung der beteiligten physikalischen Vorgänge ist in diesem Beispiel zwar möglich, aber nur mit Hilfe von aufwendigen Computersimulationen. Zudem sind die beteiligten Messdaten in der Realität immer auch mit Messfehlern behaftet, so dass sowohl y als auch die eingehenden Parameter $\boldsymbol{x}$ als Zufallsvariablen aufgefasst werden müssen.
Zusammengefasst ist diese Dissertation somit dem folgenden grundsätzlichen Problem gewidmet:

Problem 1.1 (Nicht-parametrische Regression). Betrachte das (Regressions-) Modell

$$Y = f(X) + E$$

mit einer unbekannten (und im Allgemeinen nicht-linearen) Funktion $f : \mathbb{R}^d \to \mathbb{R}^m$, wobei X und Y d- bzw. m-dimensionale komponentenweise unabhängig und identisch verteilte Zufallsvariablen bezeichnen (*input-output-Paar*). Die Störung E auf den Daten

sei ebenfalls eine m-dimensionale Zufallsvariable mit $\mathbb{E}[E] = 0$, die unabhängig von X ist. Offensichtlich ist demnach

$$f(X) = \mathbb{E}[Y|X]$$

und wir[1] schreiben

$$f(\boldsymbol{x}) = \mathbb{E}[Y|X = \boldsymbol{x}] \,.$$

Die Regressionsfunktion f sei nicht-linear und gehöre irgendeinem Funktionenraum an (z.B. C^1, L^2, etc.). Weiterhin sei

$$\mathcal{T} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$$

eine Stichprobe (Realisierung) des input-output-Paares (X, Y), den wir *Trainingsdatensatz* nennen. Vorgegeben ist auch ein Funktionenraum $\mathscr{F}$, der als Hypothesenraum bezeichnet wird. Finde nun basierend auf den Trainingsdaten $\mathcal{T}$ ein $\hat{f} \in \mathscr{F}$, so dass $\hat{f} : D \to \mathbb{R}^m$, $D \subset \mathbb{R}^d$, $\boldsymbol{x} \mapsto \hat{f}(\mathcal{T}, \boldsymbol{x})$, die Zielfunktion $f(\boldsymbol{x})$ bezüglich eines bestimmten Kriteriums (z.B. mittlere quadratische Abweichung) möglichst "gut" approximiert.

Ist $\mathscr{F}$ über eine feste Anzahl von unbekannten Parametern festgelegt, so spricht man von parametrischer, ansonsten von nicht-parametrischer Regression.
Weiterhin wird in der Regel zwischen zwei Herangehensweisen unterschieden, die *deterministic design* bzw. *random design* genannt werden. Ersteres bezeichnet die Situation, dass $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ fest gewählte, beziehungsweise regelmäßig angeordnete Design-Punkte, die Input Variablen $\boldsymbol{x}_i$ also nicht zufällig, sind. Von random design spricht man, wenn die Variablen $\boldsymbol{x}_i$ zufällig, unabhängig und identisch verteilt gewählt werden. Diese Dissertation konzentriert sich auf random-design-Modellierung.

Die typische Vorgehensweise in einem Regressions-Problem ist die Wahl einer parametrisierten Modellstruktur, innerhalb derer durch Anpassen der Parameter das "beste" Modell gesucht wird. Ein sehr einfaches Beispiel ist der Hypothesenraum der linearen Funktionen. Alle Mitglieder von $\mathscr{F}$ werden durch zwei Parameter gekennzeichnet: Steigung und Achsenabschnitt. Diese Parameter werden nun so angepasst, dass ein zuvor gewähltes Kriterium (z.B. kleinste quadratische Abweichung) ein optimales Modell innerhalb dieser Struktur identifiziert (z.B. durch ein Minimum). Die Wahl einer konkreten Modellstruktur ist also von a priori Informationen und Annahmen über den zu modellierenden Vorgang geprägt. Sind nur sehr wenige bis keine Informationen verfügbar, so spricht man von Black-Box-Modellen. Diese Modelle werden flexibel gehalten, so dass eine möglichst große Anzahl von verschiedenen Systemen approximiert werden kann.
Neuronale Netze haben sich als besonders erfolgreiche nicht-lineare Black-Box Modellstrukturen erwiesen. Der Grundgedanke ist der folgende: Neuronale Netze sollen die Informationsverarbeitung in Gehirnen nachahmen und so durch Beispiele (Messdaten) den Funktionszusammenhang f "erlernen". Diese Idee ist nicht neu, Warren McCulloch und

[1] Ein Kommentar bzgl. des benutzten Personalpronomens "wir" findet sich in Abschnitt 1.2.

Walter Pitts schlugen 1943 schon eine ähnliche Herangehensweise an maschinelles Lernen vor und passten Entwicklungsparamter anhand von “Lernbeispielen“ an. Die anfängliche Euphorie auf diese Weise künstliche Intelligenz “erzeugen“ zu können wich aber recht schnell der nüchternen mathematischen Realität und es musste erkannt werden, dass Neuronale Netze in Anwendungen des öfteren versagten.

In dieser Dissertation werden Neuronale Netze sowohl aus deterministischer als auch aus statistischer Sicht beschrieben und ihre Approximationsfähigkeiten (Risiko-Minimierung ist hier nur einer von mehreren Aspekten) modelltheoretisch analysiert. Hierbei spielt besonders der Begriff der *Komplexität* des Netzwerkes und das damit verbundene Bias-Variance-Dilemma auf modelltheoretischer Seite eine herausragende Rolle. Wir werden diesen Aspekt aus verschiedenen Blickwinkeln heraus beleuchten.
In der Praxis ist ein zusätzlicher Punkt von ganz entscheidender Bedeutung: *Robustheit* der Approximationsmethode gegenüber Störungen (z.B. Messfehler) in der Lernphase *und* im späteren Betrieb. Diese Eigenschaft der “Addaption“ an geänderte Umstände (wie z.B. kleine Störungen in den Daten) und die damit verbundene Analogie zu biologischen Systemen war einer der Grundgedanken und auch Namensgeber von Neuronalen Netzen. Alle klassischen Black-Box Modelle haben allerdings eines gemeinsam: Der “User“ hat keinen Einfluss auf die innere Struktur des Modells und muss nach Abschluss der Lernphase den Ergebnissen blind vertrauen. Das “learning-by-examples“-Paradigma hat in diesem Zusammenhang die große Schwäche, dass der erlernte Schätzer zwar an den Trainingsdatenpunkten die Funktion sehr gut und auch robust gegenüber Störungen in den Trainingsdaten approximiert, zwischen diesen Punkten allerdings *unkontrolliert* oszillieren kann. Solch ein Schätzer ist für praktische Anwendungen (in denen die Zielfunktion f immer als mehr oder weniger glatt angenommen werden kann) offensichtlich völlig ungeeignet. Wir werden in dieser Dissertation zeigen, dass Neuronale Netze in diesem Sinne nur vordergründig robust sind und der Aspekt der Sensitivität gegenüber Störungen in Trainingsdaten *und* Eingebadaten während der Anwendungsphase des Netzwerkes völlig anders als bislang behandelt werden muss. Teil III präsentiert einen neuen Konstruktionsalgorithmus für Neuronale Netze, der die aufgeworfenen Probleme löst und ein neues Paradigma des “eingeschränkten“ Lernens vorschlägt.

Im Rahmen dieser Dissertation spielt eine besondere Art von Neuronalem Netz eine hervorgehobene Rolle: Wavelet Neuronale Netze. Im Unterschied zu der verwandten Fourier-Transformation werden lokale Eigenschaften der zu approximierenden Funktion (z.B. Spitzen) in der Wavelet-Transformation nicht “verschmiert“, sondern bleiben präsent. Anders ausgedrückt gibt es einen Zusammenhang zwischen den lokalen Eigenschaften der Funktion und den errechneten Wavelet-Koeffizienten: Ist die Funktion glatt, so sind die entsprechenden Wavelet-Koeffizienten klein. Hat die Funktion eine Singularität, so steigt die Amplitude der Wavelet-Koeffizienten in diesem Bereich drastisch an. Die Fourier-Transformation ist zudem aufgrund des alternierenden Charakters der

Entwicklung nach trigonometrischen Funktionen sensitiv gegenüber Fehlern in den Phasenparametern. Dies ist für Wavelets nicht der Fall und es wird sich zeigen, dass Redundanzen in der Wavelet-Entwicklung eine gewisse *Robustheit* des Schätzers gegenüber Störungen in den Entwicklungskoeffizienten zur Folge haben.

Zusammengefasst können wir sagen, dass folgende Aspekte von Neuronalen Netzen und deren Zusammenspiel den wissenschaftlichen Fokus (sozusagen den "roten Faden") dieser Dissertation darstellen:

- ▷ Grundsätzliche Approximationsfähigkeiten von Neuronalen Netzen,
- ▷ Robustheit gegenüber Störungen in Eingabedaten und Netzwerkparametern,
- ▷ Konvergenz des Schätzers $\hat{f}$ gegen die Zielfunktion f für steigende Anzahl von Trainingsdatenpunkten,
- ▷ Komplexität des konstruierten Modells.

Wir konzentrieren uns somit auf die grundlegenden Eigenschaften von Neuronalen Netzen aus *lern-* und *modelltheoretischer* Sicht. Aus Gründen der Übersichtlichkeit wird der Aspekt der eigentlichen Optimierung der Netzwerkparameter[2] lediglich am Rande behandelt (obwohl z.B. Abschnitt 5.2 und Anhang A einige Details preisgeben). Eine detailliertere Darstellung der verwendeten und verbesserten Algorithmen findet sich in Pohl (2007), einer im Zuge dieser Dissertation vom Autor betreuten Diplomarbeit.

1.1 Überblick & Einordnung der Arbeit

Die Literaturlandschaft in Bezug auf Neuronale Netze kann als extrem heterogen bezeichnet werden. Ein Blick auf das Literaturverzeichnis dieser Dissertation macht deutlich, dass eine ganze Reihe von Journalen aus verschiedenen Gebieten der Mathematik, Informatik und Ingenieurswissenschaften Beiträge zu den theoretischen Eigenschaften von Neuronalen Netzen sowie deren Anwendungen publizieren. Diese sind im einzelnen:

- ▷ Neural Networks,
- ▷ Neural Computation,
- ▷ IEEE Trans. Neural Networks,
- ▷ Journal of Machine Learning,
- ▷ IEEE Trans. Information Theory,
- ▷ Computing,
- ▷ Information and Computation,
- ▷ Automatica,

[2] Zu diesem Thema wurden in den letzten 25 Jahren verschiedene Verfahren entwickelt (eines der bekanntesten ist sicherlich der Backpropagation-Algorithmus) und die Literaturlandschaft ist umfangreich.

- ▷ Journal of Computer and System Science,
- ▷ IEEE Trans. Signal Processing,
- ▷ IEEE Trans. Automatic Control,
- ▷ vereinzelte Beiträge in Annals of Statistics, Annals of Probability, Nature, Mathematics of Control, Signals and Systems, Statistics & Probability Letters, IEEE Trans. Systems, Man, Cybernetics.

Ziel dieser Dissertation ist neben der Beschreibung einer neuen Konstruktionsmethode für Neuronale Netze auch eine Vereinheitlichung der in der Literatur üblichen Notationen und Darstellungen. Publikationen in einem oder mehreren der oben genannten Journalen sind in Vorbereitung.

Teil I Deterministische Black-Box Modellierung
Teil I konzentriert sich auf die grundsätzlichen Approximationsfähigkeiten von Schätzern aus Hypothesenräumen, deren Elemente als Entwicklungen nach Basisfunktionen dargestellt werden können. Insbesondere wird die Approximation im deterministischen Sinne verstanden, also $\|f - \hat{f}\| \to 0$ in einer passenden Norm.
Kapitel 2 gibt hierbei zunächst einen allgemeinen Überblick über verschiedene Typen von Basisfunktionen wie Sigmoide Neuronale Netze, Radiale Basisfunktions-Neuronale Netze, Wavelet Neuronale Netze usw. Weiterhin werden die Netzwerke auf ihre prinzipiellen Approximationsfähigkeiten hin untersucht, d.h. Bedingungen gesucht, unter denen $f \in L^p$ mit beliebiger Genauigkeit reproduziert werden kann. Hierbei werden die in der Literatur verstreuten Resultate zusammengetragen, vereinheitlicht sowie an einigen Stellen auf allgemeinere Situationen erweitert.
Kapitel 3 beleuchtet das Approximationsproblem aus einem etwas anderen Blickwinkel und betrachtet so genannte Frames als redundante Basisfunktionssysteme. Hierbei steht besonders die Frage nach Eindeutigkeit, Redundanz und Komplexität der Frame-Entwicklungen im Vordergrund. Es werden Resultate aus der Literatur (z.B. Daubechies (1992)) erweitert, so dass u.A. eine numerisch leicht umsetzbare Version der endlichen Rekonstruktionsformel speziell für Wavelet-Frames präsentiert werden kann (Abschnitt 3.3). Auch die Ausführungen in 3.1 und 3.2 enthalten Erweiterungen der bestehenden Theorie und vor allem numerische Beispiele für verschiedene Wavelet-Typen. Abschnitt 3.4 stellt die Überleitung zu den wahrscheinlichkeitstheoretischen Betrachtungen in Teil II dar und analysiert den Zusammenhang von Redundanzen in Frames und Robustheit der Approximation.

Teil II Stochastische Black-Box Modellierung
Teil II betrachtet die Approximation von Funktionen durch Lernalgorithmen aus statistischer Sicht, so dass nun $\mathbb{E}[\|f - \hat{f}\|] \to 0$ ($\mathbb{E}[\cdot]$ bezeichnet den Erwartungswert) im Vordergrund steht. Alle beteiligten Größen werden als reelle und bezüglich eines Wahrscheinlichkeitsmaßes verteilte Zufallsvariablen verstanden. Diese Darstellungsform ist in der Literatur zu Neuronalen Netzen eher die Ausnahme.

Kapitel 4 legt hierbei die formalen Grundlagen: Das Modellierungsproblem wird stochastisch formuliert und eine Einführung in das Feld der (empirischen) Risiko-Optimierung gegeben. Abschnitt 4.3 führt den Begriff der Modellkomplexität aus mehreren Blickwinkeln heraus ein und zeigt wie er auf Neuronale Netze angewendet werden kann. Insbesondere wird die in der Lerntheorie bekannte SRM-Methode beschrieben. In Abschnitt 4.4 werden die Komplexitätsüberlegungen weitergeführt und der Hypothesenraum auf (konvexe) Kugeln eingeschränkt. Insbesondere wird in diesem Umfeld auf Grundlage der bewiesenen Sätze eine Lösung des Bias-Variance-Problems präsentiert. Abschnitt 4.5 diskutiert in völlig neuartiger Weise die Problematik der Risiko-Minimierung aus modelltheoretischer Sicht auf Grundlage von parametrisierten Hypothesenräumen.
Kapitel 5 befasst sich mit dem neben der Modellstruktur zweiten entscheidenden modelltheoretischen Aspekt von Black-Box-Approximation: Robustheit. Hierbei wird zunächst eine Methode vorgestellt, wie die Sensitivität von Lernalgorithmen auf Störungen in den Trainingsdaten quantifiziert werden kann (Stichwort Hypothesenstabilität). Weiterhin wird der Einfluss von Ausreißern auf Optimierungsalgorithmen diskutiert. Im letzten Abschnitt dieses Kapitels wird ein direkter Bezug zu Kapitel 4 hergestellt und eine Klasse von Algorithmen vorgestellt, die einen Schätzer rekursiv mit aufsteigender Komplexität der Hypothesenräume konstruieren. Die betrachteten Algorithmen werden hierbei in Hinblick auf ihre Hypothesenstabilität überprüft. Es gibt in der bestehenden Literatur keine vergleichbaren Ergebnisse. In diesem Abschnitt wird zudem ein neuer Algorithmus vorgestellt, der die Grundlage für den in Kapitel 6 vorgeschlagenen optimalen Konstruktionsalgorithmus für Neuronale Netze bildet.

Teil III Anwendung
Teil III wendet die Ergebnisse aus Teil I und II an und präsentiert eine völlig neue Klasse von Neuronalen Netzen, die mit den Ergebnissen aus Teil I und II als optimal in Hinblick auf Approximationsfähigkeit, Robustheit und Komplexität bezeichnet werden können.
Kapitel 6 führt zunächst gleichgradig stetige Neuronale Netze ein und zeigt die praktische Relevanz dieser Hypothesenraum-Klasse. Abschnitt 6.2 stellt dann auf dieser Grundlage und den theoretischen Ergebnissen aus Kapitel 2, 4 und 5 eine bisher nicht beschriebene Klasse von Neuronalen Netzen vor, die sowohl das Prognose-Risiko minimieren als auch alle geforderten Robustheitseigenschaften erfüllen. Insbesondere stellt der neue Konstruktionsalgorithmus eine neue Art der Anwendung des in Abschnitt 4.3 vorgestellten SRM-Prinzips dar und löst in dem beschriebenen Rahmen das Bias-Variance-Dilemma.
Anhang A beschreibt das im Rahmen dieser Dissertation entwickelte Wavelet Neuronale Netz und zeigt einige Beispielprobleme.

1.2 Konventionen und Notation

Im Zuge dieser Dissertation wird das Personalpronomen “wir“ in Kombination mit den zugehörigen Deklinationen verwendet. Hiermit soll *keine* Kollaboration mit anderen For-

schern ausgedrückt werden. Urheber dieser Arbeit ist einzig und allein der Autor und benutzte Resultate anderer Autoren sind eindeutig gekennzeichnet.

Weiterhin unterscheiden wir in dieser Dissertation zwischen Zufallsvariablen und deren Realisationen. Erstere werden mit Großbuchstaben (X, Y usw.) und letztere mit fett geschriebenen Kleinbuchstaben ($\boldsymbol{x}$, $\boldsymbol{y}$ usw.) bezeichnet. Für den Spezialfall des Prognose-Risikos $\mathit{\Lambda}$ (Teil II) unterscheiden wir zwischen Zufallsvariable und Realisation hingegen durch eine Übertilde ($\widetilde{\mathit{\Lambda}}$ ist die Zufallsvariable). In Teil I wird dieses Symbol auch als Kennzeichnung für Mutter-Basisfunktionen verwendet (z.B. $\widetilde{\Phi}$). Es werden an keiner Stelle Zweideutigkeiten auftreten.

Schätzfunktionen werden in der Regel durch einen "Hut" gekennzeichnet, z.B. $\hat{f}$. Da dies mit der häufig verwendeten Notation für die Fourier-Transformation einer Funktion kollidiert, verwenden wir für diese einen Überstrich $\overline{f}$.

Im Zuge dieser Dissertation wurde ein Wavelet Neuronales Netz als Demonstratornetzwerk entwickelt. Es wird des öfteren auf die Ergebnisse mit diesem Netzwerk Bezug genommen. Die algorithmischen Details stellen das Thema der vom Autor dieser Dissertation betreuten Diplomarbeit von Daniel Pohl dar (s. Pohl (2007)) .

Teil I

Deterministische Black-Box Modellierung

2

Nicht-parametrische Regression

Ein natürlicher Zugang zu der Problemstellung aus Kapitel 1 ist es, die (parametrisierte) Funktionenfamilie $\mathscr{F}$ als Entwicklung nach (i.A. nichtlinearen) Basisfunktionen Φ_i zu definieren.

2.1 Entwicklung nach Basisfunktionen

In diesem Ansatz hat $g \in \mathscr{F}$ die Form $g : \mathbb{R}^d \longrightarrow \mathbb{R}^m$ mit

$$g(\boldsymbol{x}) = g^{\boldsymbol{R}}(\boldsymbol{x}) = \sum_i u_i \, \Phi_i^{\boldsymbol{R}}(\boldsymbol{x}) \, .$$

Wir lassen die Grenzen der Summe zunächst offen und betrachten nur die generelle Struktur der Entwicklung. $\boldsymbol{R} \in \mathbb{R}^d \times \mathbb{R}^N$ bezeichnet eine Regressionsmatrix. Wie genau diese aus den Daten gewonnen wird lassen wir ebenfalls an dieser Stelle offen, zum Beispiel $\boldsymbol{R} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$. Das hochgestellte $\boldsymbol{R}$ soll verdeutlichen, dass die Basisfunktionen Φ_i vom Regressionsvektor abhängen, dieser aber zum Beispiel im Falle von Trainingsdaten im Allgemeinen fest gewählt wird. Wir werden daher im folgenden den Hinweis, dass die Basisfunktionen von $\boldsymbol{R}$ abhängen, weglassen.

Solch eine Entwicklung nach Basisfunktionen erinnert mit Absicht an eine Entwicklung in einem Funktionenraum. Es zeigt sich, dass dieser Ansatz sehr erfolgreich ist und weitreichende Sätze über die Approximationseigenschaften solcher Funkionenfamilien getroffen werden können. Hierzu aber mehr in 2.2.

Die meisten Black-Box-Modelle wählen als Basisfunktionen Parametrisierungen einer Grundfunktion, genannt *Mutter-Basisfunktion*

$$\Phi_i = \widetilde{\Phi}(a_i, t_i) \, ,$$

so dass sich in diesem Modell ein Schätzer für f in der Form

$$\begin{aligned} \hat{f} : \mathbb{R}^d &\longrightarrow \mathbb{R}^m \\ \boldsymbol{x} &\longmapsto u_0 + \sum_{i=1}^{M} u_i \, \Phi_i(\boldsymbol{x}) = u_0 + \sum_{i=1}^{M} u_i \, \widetilde{\Phi}(a_i, t_i; \boldsymbol{x}) \, . \end{aligned} \tag{2.1}$$

Oder etwas stilisiert:

$$\text{Output} = \text{Linearkombination}\Big(\text{nichtlineare Basisfunktionen}\big(\text{Input}\big)\Big) \, .$$

Eine mögliche und in unserem Fall häufig benutzte Parametrisierung im endimensionalen Fall ist

$$\widetilde{\Phi}(a_i, t_i; x) = \widetilde{\Phi}(a_i x + t_i) \, .$$

Auf die Wahl von M werden wir in Kapitel 4 und in Teil III noch eingehend zu sprechen kommen. M wählen heißt anhand der Trainingsdaten $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ die Zahl der möglichen Parameter

$$\begin{aligned} \boldsymbol{u} &= (u_1, ...)^T \\ \boldsymbol{a} &= (a_1, ...)^T \\ \boldsymbol{t} &= (t_1, ...)^T \end{aligned}$$

zu beschränken. Offensichtlicherweise stellt dies ein zentrales Problem in der Schätzung von f dar.

2.1.1 Beispiele für Basisfunktionen

Es können im eindimensionalen Fall zwei Grundformen von Basisfunktionen unterschieden werden. Dies sind *lokale* Basisfunktionen und *globale* Basisfunktionen. Salopp formuliert konzentrieren sich erstere auf ein Intervall, d.h. etwas genauer ausgedrückt, ihr Gradient hat kompakten Träger beziehungsweise verschwindet schnell im Unendlichen. Globale Basisfunktionen haben diese Eigenschaft nicht. Es ist einsichtig, dass lokale Basisfunktionen prinzipiell besser geeignet sind, um Funktionen mit lokalen Schwankungen oder Unstetigkeiten zu approximieren.

Im folgenden gehen wir auf einige wichtige Beispiele für Basisfunktionen ein.

Fourier-Reihen
Fourier-Reihen sind ein einfaches Beispiel für eine nicht-lokale, d.h. globale Approximationsmethode. Mit der Basisfunktion

$$\widetilde{\Phi}(x) = \cos(x)$$

erhält man als Schätzer das Fourier-Polynom

$$\hat{f}_n(t) = \frac{A_0}{2} + \sum_{k=1}^{n} A_k \cos(k\omega t - \varphi_k)$$

und mit $n \to \infty$ die Fourier-Reihe der Funktion f. Wir identifizieren also mit $(u_1, u_2, ...,)^T = (A_1, A_2, ...)^T$ die Amplituden, mit $(a_1, a_2, ...)^T = (\omega, 2\omega, ...)^T$ die Frequenzen und mit $(t_1, ...)^T = (\varphi_1, ...)^T$ die Phasen der Fourier-Entwicklung.

Im allgemeinen konvergiert beim Grenzübergang $n \to \infty$ die Fourier-Reihe einer Funktion f nicht gleichmäßig gegen f und auch nicht punktweise. Werden die $2n + 1$ Variablen $(A_0, ..., A_n; \varphi_1, ..., \varphi_n)$ allerdings so bestimmt, dass

$$A_k = \left[\underbrace{\left(\frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(kx)\, \mathrm{d}x \right)}_{a_k}^2 + \underbrace{\left(\frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(kx)\, \mathrm{d}x \right)}_{b_k}^2 \right]^{1/2}$$

und

$$\varphi_k = \begin{cases} \arctan(\frac{b_k}{a_k}), & a_k \geq 0, \\ \arctan(\frac{b_k}{a_k}) + \pi, & a_k < 0 \ , \end{cases}$$

so lässt sich die Konvergenz im quadratischen Mittel zeigen (s. z.B. Schwarz (1997)):

$$\lim_{n \to \infty} \| \hat{f}_n(x) - f(x) \|_2 = \left(\int_{-\pi}^{\pi} \left(\hat{f}_n(x) - f(x) \right)^2 \, \mathrm{d}x \right)^{1/2} = 0 \, .$$

Sigmoide Funktionen
Es sei $f : [a, b] \to \mathbb{R}$. Wir wählen nun die Indikatorfunktion als Basisfunktion:

$$\widetilde{\Phi}(x) = \mathbb{1}_{[0,1)}(x) = \begin{cases} 1, & 0 \leq x < 1, \\ 0, & \text{sonst.} \end{cases}$$

Weiterhin betrachten wir die äquidistante Zerlegung

$$x_k = x_0 + hk, \quad k = 0, ..., n, \quad h = \frac{b - a}{n} \, .$$

Dann können wir f mit Hilfe einer stückweise konstanten Funktion approximieren ($x_0 = 0$), also einem Polynom vom Grad 0:

$$\hat{f}_M(x) = \sum_{k=0}^{M} f\left(\frac{x_{k+1} + x_k}{2} \right) \cdot \widetilde{\Phi}\left(\frac{x - x_k}{x_{k+1} - x_k} \right) \, .$$

Dies entspricht, setzt man die Parametrisierung $\widetilde{\Phi} = \widetilde{\Phi}(a_i x + t_i)$ an, der Wahl $u_k = f\left(\frac{2k+1}{2}h\right)$, $a_k = \frac{1}{h}$ und $t_k = -k$.

Für einige weiterführende Ideen (wie zum Beispiel die neuronalen Netze) spielt die Heaviside-Funktion eine große Rolle. Sie ist definiert als

$$\widetilde{\Phi}(x) = \sigma(x) := \begin{cases} 0, & x < 0, \\ 1, & x > 0\,. \end{cases}$$

Die Indikatorfunktion ergibt sich dann einfach als Differenz zweier Heaviside-Funktionen.

Etwas allgemeiner sind sigmoide Funktionen erklärt als beschränkte messbare Funktionen mit $\sigma(x) \to 1$ für $x \to \infty$ und $\sigma(x) \to 0$ für $x \to -\infty$. Die sigmoiden Funktionen (und natürlich auch die Indikatorfunktion) sind Beispiele für *lokale* Funktionen.

In der Anwendung ist die glatte Version der Stufenfunktion weit verbreitet (Abb. 2.1):

$$\sigma(x) = \frac{1}{1 + e^{-x}}\,.$$

Ganz ähnliche Resultate liefert zum Beispiel $\sigma(x) = \tanh(x)$.

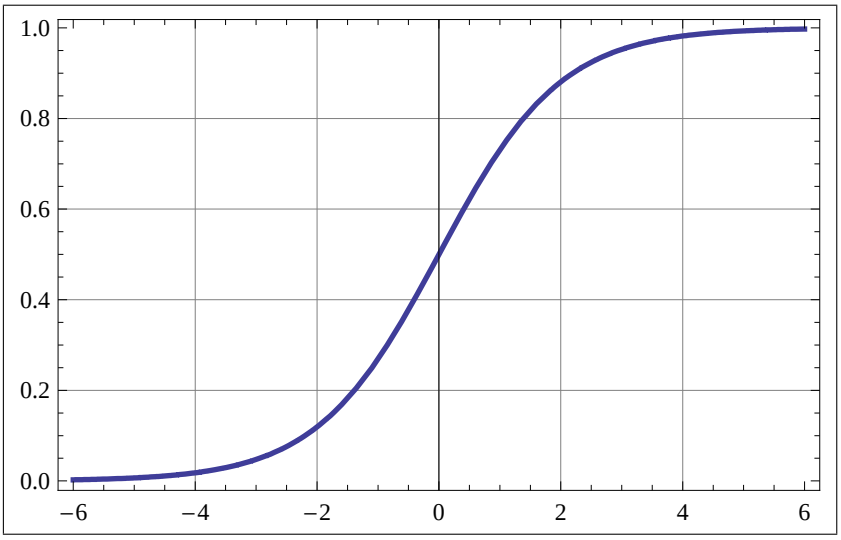


Abb. 2.1. Sigmoide Funktion

Mit Hilfe von sigmoiden Basisfunktionen können wir nun Funktionsapproximatoren der Form

$$\hat{f}_M(x) = u_0 + \sum_{i=1}^{M} u_i \, \sigma \left(a_i x + t_i \right) \tag{2.2}$$

betrachten. Auf die Approximationseigenschaften solcher Entwicklungen gehen wir näher in 2.2 ein. Von zentraler Bedeutung wird hier ein Resultat von Barron (1993) sein. Es zeigt, dass eine Funktionsapproximation in der Form von Glg. (2.2) deutlich bessere Konvergenzraten im Vergleich zu linearen Approximatoren besitzen *kann*. Denn es wird auch deutlich, dass es sehr von der zu approximierenden Funktion abhängt ob diese Aussage auch wirklich zutrifft.

Radiale Basisfunktionen
Radiale Basisfunktionen haben die Form

$$\widetilde{\Phi}(\boldsymbol{x}) = \widetilde{\Phi} \left(\|\boldsymbol{x}\|_{\boldsymbol{K}} \right)$$

mit $\boldsymbol{x} \in \mathbb{R}^d$ und

$$\|\boldsymbol{x}\|_{\boldsymbol{K}}^2 = \boldsymbol{x}^T \boldsymbol{K} \boldsymbol{x} \, .$$

$\boldsymbol{K} \in \mathbb{R}^d \times \mathbb{R}^d$ ist eine positiv definite Matrix von Skalierungsparametern, zum Beispiel

$$\boldsymbol{K} = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_d \end{pmatrix} \, .$$

Ein typisches Beispiel für radiale Basisfunktionen sind Gauss'sche Glocken:

$$\Phi_i(\boldsymbol{x}) = \exp \left(- \frac{\|\boldsymbol{x} - \boldsymbol{t}_i\|_2^2}{\sigma_i} \right) \, .$$

Und wieder erhalten wir als Schätzer

$$\hat{f}_M(x) = u_0 + \sum_{i=1}^{M} u_i \, \Phi_i(\boldsymbol{x}) \, .$$

Wir werden insbesondere auf Netzwerke dieser Art eingehen wenn wir die Robustheitseigenschaften von Black-Box-Modellen betrachten.

Wavelets
Nun verwenden wir als Mutter-Basisfunktionen so genannte Wavelets, ein Paradebeispiel für lokale Basisfunktionen (mit kompaktem Träger). Wavelets haben die Eigenschaft, dass sie parametrisiert sind durch Dilations- und Translationsparameter. Insbesondere existiert ein Mutter-Wavelet $\widetilde{\Phi}(x) = \psi(x)$, so dass die Familie

$$\left\{\psi_{a_i,t_j}(x) = a_i^{\frac{1}{2}}\psi(a_i x - t_j) \;:\; i,j \in \mathbb{Z}\right\}$$

ein (abzählbare) orthonormale Basis von $L^2(\mathbb{R})$ bildet. Die Diskretisierung der Basis wird in der Regel auf einem regelmäßigen Gitter vorgenommen, d.h.

$$\begin{aligned} a_i &= a_0^i \quad , \quad a_0 \in \mathbb{R} \\ t_j &= j\, t_0 \quad , \quad t_0 \in \mathbb{R}\,. \end{aligned}$$

Die skalaren Parameter a_0 und t_0 definieren die Schrittweiten der Dilation und Translation des Mutter-Wavelets. Oft wird $a_0 = 1/2$ und $t_0 = 1$ gewählt.

An dieser Stelle wollen wir lediglich die grundlegende (eindimensionale) Definition von Wavelets angeben und einige in der Anwendung wichtige Beispiele geben. Folgende Definition stammt von Morlet & Grossmann:

Definition 2.1 (Wavelet im Sinne von Morlet & Grossmann). *Eine Funktion $\psi : \mathbb{R} \to \mathbb{C}$ heißt Wavelet, wenn folgendes erfüllt ist:*

1. $\psi \in L^2$
2. $\|\psi\| = 1$
3. $C_\psi := 2\pi \displaystyle\int_{\mathbb{R}\setminus\{0\}} \frac{|\overline{\psi}(a)|^2}{|a|}\, da < \infty\,,$

wobei $\overline{\psi}$ die Fourier-Transformiert von ψ bezeichne.

Wavelets müssen nicht stetig sein, wie man am Beispiel des Haar-Wavelet sofort sieht:

$$\psi_{\text{Haar}} = \begin{cases} 1, & 0 \le x < 1/2, \\ -1, & 1/2 \le x < 1 \\ 0, & \text{sonst.} \end{cases}$$

Der folgende Satz ersetzt Bedingung 3:

Satz 2.1. *Für Funktionen $\psi \in L^2$ mit $\int |x|\,|\psi(x)|\, dx < \infty$ gilt:*

$$C_\psi := 2\pi \int_{\mathbb{R}\setminus\{0\}} \frac{|\overline{\psi}(a)|^2}{|a|}\, da < \infty \quad \Longleftrightarrow \quad \int_{-\infty}^{\infty} \psi(x)\, dx = 0 \quad \Longleftrightarrow \quad \overline{\psi}(0) = 0\,.$$

Ein Wavelet hat also Mittelwert 0.

Wir können diesen Satz verwenden um zu zeigen, dass eine beliebige Funktion $\psi \in L^2$ mit $\|\psi\| = 1$, Mittelwert 0 und kompaktem Träger ein Wavelet ist: Es sei $\psi(x) = 0$ für $|x| > x_0$. Die Funktion $h(x) := |x| \mathbb{1}_{[-x_0,x_0]}(x)$ liegt in L^2. Also ist auch

$$\int_{\mathbb{R}} |x|\,|\psi(x)|\, \mathrm{d}x = \langle h, |\psi| \rangle < \infty\,.$$

Und dann lässt sich obiger Satz anwenden.

In den späteren Anwendungen werden wir oft noch ein wenig mehr fordern, und zwar dass ψ normalisiert ist, d.h. dass neben

$$\|\psi\| = \int |\psi(x)|^2 \, \mathrm{d}x = 1 \quad \text{auch}$$
$$\int x|\psi(x)|^2 \, \mathrm{d}x = 0 \, .$$

Letztere Bedingung macht eine Aussage über den "Erwartungswert" des Mutter-Wavelets, also wo es zentriert ist.

Von ganz besonderer Bedeutung ist das Mexikanerhut-Wavelet (hier eindimensional):

$$\psi(x) = \underbrace{\frac{2}{\sqrt{3}}\pi^{-1/4}}_{=:\gamma}(1-x^2)\underbrace{e^{-x^2/2}}_{=:g(x)},$$

s. Abb. 2.2. Es ist offensichtlich $\psi(x) = -\gamma g''(x)$.

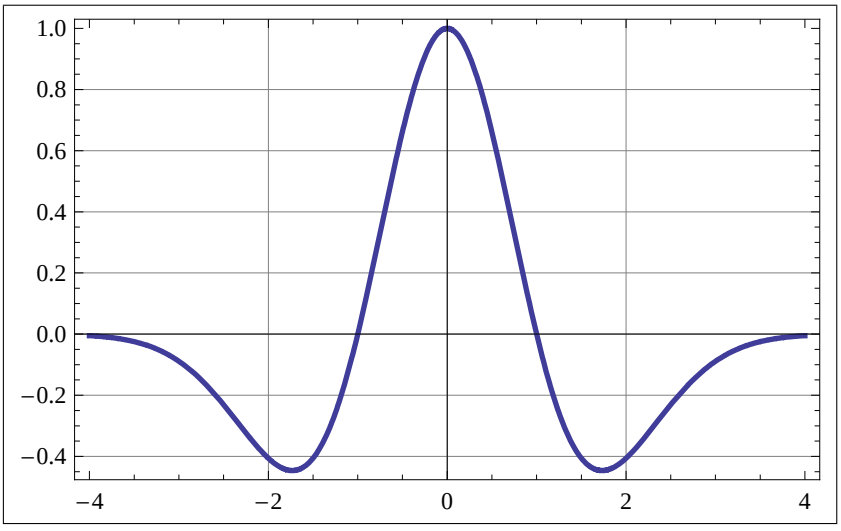


Abb. 2.2. Mexikanerhut-Wavelet

Das Morlet-Wavelet ist komplex und folgendermaßen definiert:

$$\psi(x) = \pi^{-1/4} \underbrace{\left(1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\sigma^2}\right)^{-1/2}}_{=:c_\sigma} \left(e^{i\sigma x} - \underbrace{e^{-\frac{1}{2}\sigma^2}}_{=:\kappa_\sigma}\right) e^{-x^2/2} .$$

Wie leicht an dem $e^{i\sigma x}$-Term zu sehen ist das Morlet-Wavelet eine durch eine Gauss-Glocke lokalisierte ebene Welle. κ_σ ist nur eine Translations-Konstante. Der Graph dieses Wavelets ist in Abb. 2.3 gezeigt.

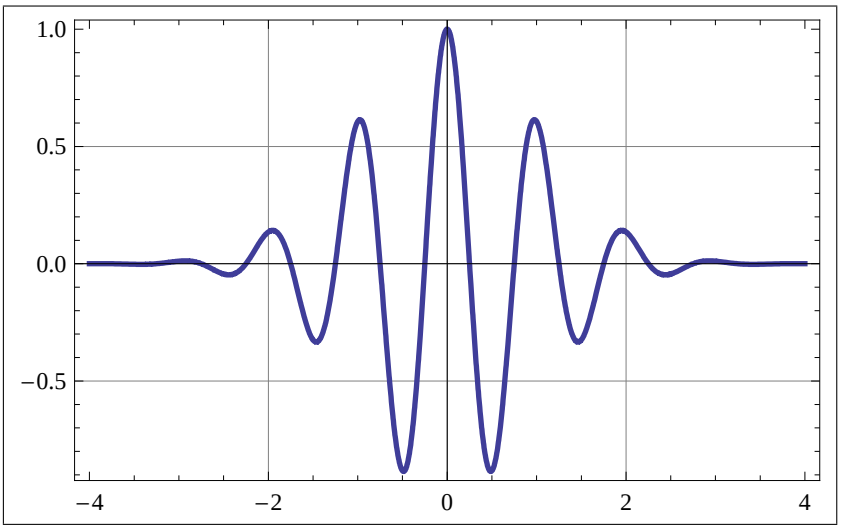


Abb. 2.3. Reeller Teil des Morlet-Wavelets

Weitere Basisfunktionen

Wir möchten der Vollständigkeit halber noch weitere Modellstrukturen ansprechen. Diese Dissertation konzentriert sich aber auf Anwendungen der Basisfunktions-Entwicklung in Form von Netzwerken, so dass Sigmoide Funtionen, Radiale Basisfunktionen und Wavelets besondere Bedeutung erlangen. Wir stellen hier auch nur die wichtigsten Konzepte dar, weitere Methoden wie zum Beispiel interpolations-Methoden können in Juditsky et al. (1995) sowie in Sjöberg (1995) nachgelesen werden.

Kernschätzer

Ein Spezialfall von lokalen Basisfunktions-Entwicklungen sind so genannte *Kernschätzer* (s. Nadaraya (1964); Watson (1969)). Die Gewichte werden über einen Kern und einen

Parameter (Bandbreite) $h > 0$ festgelegt. Die Kernfunktion ist eine bezüglich 0 symmetrische Wahrscheinlichkeitsdichte, die den Träger $[-1, 1]$ hat, beziehungsweise sehr schnell abfällt (z.B. Gauss'sche Normalverteilung). Es ist dann

$$\hat{f}_M(x) = \sum_{i=1}^{M} u_i y_i \widetilde{\Phi}\left(\frac{x - x_i}{h}\right) ,$$

wobei (x_i, y_i), $i = 1, ..., N$, wie zuvor die Realisierungen von N Trainingspunkten bezeichnet.

Ridge Construction
Ein Beispiel für eine Basisfunktions-Entwicklung, in der die Mutter-Basisfunktion zwar lokalen Träger hat, die Entwicklung aber unbeschränkten Träger aufweist, ist die "ridge" Konstruktion (vom Englischen "Bergrücken"): Hierbei ist ($\boldsymbol{a}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$)

$$\widetilde{\Phi} = \widetilde{\Phi}\left(\boldsymbol{a}_i^T \boldsymbol{x} + t_i\right) .$$

$\widetilde{\Phi}$ ist konstant in $\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{a}_i^T \boldsymbol{x} = \text{const}\}$, d.h. auch wenn $\widetilde{\Phi}$ beschränkten Träger hat, so haben die Basisfunktionen $\widetilde{\Phi}(\boldsymbol{a}_i, t)$ in diesem Unterraum unbeschränkten Träger. Die hieraus resultierende Basis ist somit "semi"-global. Im Gegensatz zu den radialen Basisfunktionen bieten die ridge Funktionen die Möglichkeit einer gewissen Richtungsabhängigkeit der Entwicklung an, was in manchen Anwendungen von Vorteil ist (auch wenn die Funktionen in diese Richtungen konstant sind).

Hinging Hyperplanes
Breiman (1993) beschreibt eine Alternative zu sigmoiden Funktionen als Basisfunktionen für neuronale Netze, das *hinging hyperplanes* Modell. Hier werden statt der sigmoiden Funktion σ so genannte "hinge" Funktionen gewählt. Diese haben die Form eines geöffneten Buches:

$$h(\boldsymbol{x}) = \pm \max\{\boldsymbol{a}^+ \boldsymbol{x} + t^+ \, , \, \boldsymbol{a}^- \boldsymbol{x} + t^-\} .$$

Wählt man als Basisfunktion

$$\widetilde{\Phi}(x) = \begin{cases} 0, & x < 0, \\ \pm x, & x > 0 , \end{cases}$$

so ergibt sich als Schätzer

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{M} \widetilde{\Phi}(\boldsymbol{a}_i^T \boldsymbol{x} + t_i) + \boldsymbol{\mu}^T \boldsymbol{x} + t_0 ,$$

wobei $\boldsymbol{\mu} \in \mathbb{R}^d$ ein Vektor von Parametern ist. Es kann allerdings gezeigt werden, dass dieses Modell überparametrisiert ist (s. Pucar & Sjöberg (1995)). Man kann leicht sehen, dass die hinging hyperplanes nichts anderes als eine ridge Konstruktion mit zusätzlichem linearen Term sind.

Projection Pursuit Regression
Ein weiteres Beispiel für eine spezielle ridge Konstruktion ist die Projection Pursuit Regression (Huber (1985)). Hier werden als Basisfunktionen $\widetilde{\Phi} : \mathbb{R}^d \to \mathbb{R}$

$$\widetilde{\Phi}(\boldsymbol{x}) = \widetilde{\Phi}(\boldsymbol{A}_i \boldsymbol{x} + \boldsymbol{t}_i)$$

verwendet, wobei $\boldsymbol{A} \in \mathbb{R}^d \times \mathbb{R}^q$ mit $d > q$. Diese Methode heißt Projection Pursuit, weil die Wahl von q die Projektionen in die Dimensionen des Regressor-Raumes festlegen, die am signifikantesten für die Systemeigenschaften sind.

2.2 Approximationseigenschaften von Neuronalen Netzen

Die Frage, die uns in diesem Kapitel interessiert, ist die nach der grundsätzlichen Fähigkeit von Neuronalen Netzen eine beliebige Funktion f zu approximieren. Wir werden hier Resultate präsentieren, die analog zu berühmten Approximationssätzen wie dem von Weierstrass sind. Zur Erinnerung:

Satz 2.2 (Weierstrass (1885)). *Es sei*

$$\Pi_n[a,b] := \left\{ p \in C[a,b] \ : \ p(x) = \sum_{i=n}^{n} a_i x^i \ , \ a_i \in \mathbb{R} \right\}$$

der Vektorraum der Polynome über $[a,b]$ mit Höchstgrad n. Weiterhin sei $f \in C[a,b]$. Dann existiert für alle $\varepsilon > 0$ ein $n \in \mathbb{N}$ und ein $p \in \Pi_n[a,b]$ mit $\|f - p\|_\infty \leq \varepsilon$. D.h. der Raum der Polynome liegt dicht in $C([a,b])$ bezüglich der Unendlichnorm.

Wir werden ähnliche Sätze für neuronale Netze zeigen. Allerdings beschränken wir uns nicht nur auf Funktionen aus $C(K)$, $K \subset \mathbb{R}^d$, sondern gehen weiter zu $L^1(K)$ und sogar $L^p(\mathbb{R}^d)$.

Besondere Bedeutung in dieser Dissertation haben die drei folgenden Netzwerktypen:

1. Superposition von sigmoiden Basisfunktionen (SNN),
2. Superposition von radialen Basisfunktionen (RBFN),
3. Superposition von Wavelets (WNN).

Ein wichtiger Aspekt der Frage nach der Approximationsfähigkeit dieser Netze ist, wie viele "Nodes" benötigt werden, beziehungsweise wie genau der Zusammenhang zwischen M aus Glg. (2.1) und der "Güte" der Approximation aussieht. Die meisten Autoren belassen es in diesem Punkt bei vagen Andeutungen, obwohl durchaus statistische Methoden zur Abschätzung von M entwickelt wurden, diese aber nie Einzug in die Theorie

von Neuronalen Netzen gehalten haben. Es ist also von größter Wichtigkeit darauf hinzuweisen, dass alle in diesem Abschnitt getroffenen Aussagen nur für frei wählbares M gelten. Somit besagen alle hier bewiesenen Sätze nur: Ein Neuronales Netz dieser oder jener Form kann *prinzipiell* eine Funktion aus $L^p(\mathbb{R}^d)$ beliebig genau approximieren. In der Praxis wird M oft im Voraus gewählt, so dass aber gerade hier eine starke Einschränkung gemacht wird, die die Approximationsfähigkeit der Neuronalen Netze erheblich mindern kann. In Teil II und III dieser Dissertation widmen wir gerade diesem Punkt große Aufmerksamkeit.

2.2.1 Single-layer-SNNs

Wir verallgemeinern nun die bisher angestellten Überlegungen und betrachten Basisfunktionen der Form

$$\widetilde{\Phi} = \sigma \circ A \,,$$

wobei $\sigma : \mathbb{R} \to \mathbb{R}$ eine *Aktivierungsfunktion* sei. Dies sind zunächst einfach messbare (und monotone) Funktionen. Oft wird $\lim_{x\to\infty} \sigma(x) = 1$ und $\lim_{x\to-\infty} \sigma(x) = 0$ gefordert (in diesem Fall spricht man von einer sigmoiden Funktion). Falls zusätzlich die Aktivierungsfunktion noch monoton steigend sein soll (also nicht-abfallend), dann nennt man σ eine "*squashing function*".
$A : \mathbb{R}^d \to \mathbb{R}$ sei ein affines Funktional auf $\mathbb{R}^d$ der Form

$$A(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + t \quad \text{mit} \quad \boldsymbol{a} \in \mathbb{R}^d \,,\, t \in \mathbb{R} \,. \tag{2.3}$$

Die Menge solcher Funktionale auf $\mathbb{R}^d$ nennen wir $\boldsymbol{A}(\mathbb{R}^d)$. Im folgenden betrachten in Hinblick auf eine mit der gängigen Literatur konsistenten Darstellung $\boldsymbol{A}(K)$ mit $K \subset \mathbb{R}^d$ und wählen noch spezieller $K = [0,1]^d$. Die Resultate sind sofort erweiterbar auf beliebige Kompakta in $\mathbb{R}^d$.

Definition 2.2 (Klasse der affinen Kompositionen, Teil 1). *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine Aktivierungsfunktion. $\Sigma^d(\sigma)$ bezeichne die Klasse von Funktionen $g : K \to \mathbb{R}$ der Form*

$$g(\boldsymbol{x}) = \sum_{i=1}^{M} u_i \, \sigma\left(A_i(\boldsymbol{x})\right)$$

mit $K \subset \mathbb{R}^d$, $u_i \in \mathbb{R}$, $A_i \in \boldsymbol{A}(K)$, $A_i(\boldsymbol{x}) = \boldsymbol{a}_i^T\boldsymbol{x} + t_i =$ und $M = 1, 2, ...$

Ist also eine Aktivierungsfunktion σ vorgegeben, so ist $\Sigma^d(\sigma)$ die Klasse der Funktionen auf K, die das neuronale Netz über die affine Komposition darstellen kann. Es zeigt sich (s. Cybenko (1989)), dass eine bestimmte Eigenschaft der Aktivierungsfunktion entscheidend ist in der Frage, ob für den Einheitswürfel $K = [0,1]^d$ die Menge $\Sigma^d(\sigma)$ dicht in $C([0,1]^d)$ liegt oder nicht:

Definition 2.3 (Diskriminatorische Aktivierungsfunktionen, Teil 1). *Es sei* $\mathbb{B}_{[0,1]^d}$ *die Menge aller signierten regulären*[1] *Borelmaße auf* $[0,1]^d$*. Eine stetige Aktivierungsfunktion* σ *heißt diskriminatorisch, wenn für ein* $\lambda \in \mathbb{B}_{[0,1]^d}$ *gilt:*

$$\int_{[0,1]^d} \sigma\left(A(\boldsymbol{x})\right)\, d\lambda(\boldsymbol{x}) = 0 \quad \forall A \in \boldsymbol{A}\left([0,1]^d\right) \quad \Longrightarrow \quad \lambda = \textit{Nullmaß}^2\,.$$

Ganz analog können wir definieren:

Definition 2.4 (Diskriminatorische Aktivierungsfunktionen, Teil 2). *Nun sei* $\sigma \in L^1(\mathbb{R})$*.* σ *heißt diskriminatorisch, wenn für* $h \in L^\infty\left(\mathbb{R}^d\right)$ *gilt:*

$$\int_{[0,1]^d} \sigma\left(A(\boldsymbol{x})\right)\, h(\boldsymbol{x})\, d\boldsymbol{x} = 0 \quad \forall A \in \boldsymbol{A}\left([0,1]^d\right) \quad \Longrightarrow \quad h(\boldsymbol{x}) = 0 \ \textit{fast-überall}\,.$$

Bevor wir zum Hauptresultat aus Cybenko (1989) kommen, ein nützliches Lemma, das den Bezug zu Neuronalen Netzen herstellt:

Lemma 2.1. *Jede stetige sigmoide Funktion ist diskriminatorisch.*

Beweis. Cybenko (1989) gibt den Beweis nur für $\sigma \in C(\mathbb{R})$, er kann aber leicht auf $L^1(\mathbb{R})$ erweitert werden. Kern des Beweises ist der Satz von Lebesgue.
Wir betrachten

$$\sigma_\tau = \sigma\left(\tau A(\boldsymbol{x}) + k\right) = \sigma\left(\tau\left(\boldsymbol{a}^T\boldsymbol{x} + t\right) + k\right)$$

mit $\tau, k \in \mathbb{R}$ und $A \in \boldsymbol{A}([0,1]^d)$. Es gilt für alle $\boldsymbol{x}, \boldsymbol{a}, t, k$, dass punktweise

$$\sigma_\tau(\boldsymbol{x}) \overset{\tau\to\infty}{\longrightarrow} \gamma(\boldsymbol{x}) = \begin{cases} 1, & A(\boldsymbol{x}) > 0 \\ 0, & A(\boldsymbol{x}) < 0 \\ \sigma(k), & A(\boldsymbol{x}) = 0\,. \end{cases}$$

Laut Voraussetzung ist $\int \sigma(A(\boldsymbol{x}))d\lambda = 0$ für alle A. Es ist aber $(\tau A(\boldsymbol{x}) + k) \in \boldsymbol{A}([0,1]^d)$, d.h. es gilt auch $\int \sigma_\tau(\boldsymbol{x})d\lambda = 0$ für alle τ. Die Funktionen der Folge $\{\sigma_\tau\}_{\tau\in\mathbb{R}}$ sind alle messbar auf $[0,1]^d \subset \mathbb{R}^d$ und es gibt ein M, so dass $|\sigma_\tau(\boldsymbol{x}| \leq M$ für alle τ und alle $\boldsymbol{x}$. Dies ist offensichtlich für eine sigmoide Funktion. Weiterhin gilt wie schon erwähnt $\sigma_\tau(\boldsymbol{x}) \to \gamma(\boldsymbol{x})$ für alle $\boldsymbol{x}$. Wir können also den Satz von der dominierten Konvergenz (Lebesgue) anwenden und erhalten für alle $\boldsymbol{a}, t, k$:

$$\begin{aligned} 0 = \lim_{\tau\to\infty} \int_{[0,1]^d} \sigma_\tau(\boldsymbol{x})\, \mathrm{d}\lambda(\boldsymbol{x}) &= \int_{[0,1]^d} \lim_{\tau\to\infty} \sigma_\tau(\boldsymbol{x})\, \mathrm{d}\lambda(\boldsymbol{x}) = \int_{[0,1]^d} \gamma(\boldsymbol{x})\, \mathrm{d}\lambda(\boldsymbol{x}) \\ &= \sigma(k)\, \lambda\left(H^{(0)}_{\boldsymbol{a},t}\right) + \lambda\left(H^{(+)}_{\boldsymbol{a},t}\right)\,, \end{aligned}$$

[1] Zur Erinnerung. Ein Maß $\mu : \mathscr{A} \to [0,\infty]$ auf der σ- Algebra $\mathscr{A}$ heißt regulär, falls für jedes $A \in \mathscr{A}$ gilt: $\mu(A) = \sup\{\mu(K) : K \subset A, K \text{ kompakt}\} = \inf\{\mu(U) : U \subset A, U \text{ offen}\}$.
[2] Das Nullmaß ist das Maß, das jeder Menge aus der σ-Algebra den Wert 0 zuordnet.

wobei $H^{(0)}_{\boldsymbol{a},t} := \{\boldsymbol{x} \,:\, \boldsymbol{a}^T\boldsymbol{x} + t = 0\}$ und $H^{(+)}_{\boldsymbol{a},t} := \{\boldsymbol{x} \,:\, \boldsymbol{a}^T\boldsymbol{x} + t > 0\}$.
Wäre λ ein positives Maß, so wäre der Beweis an dieser Stelle schon abgeschlossen, denn es würde sofort $\lambda = 0$ folgen. Leider ist dies nicht zwingend der Fall. Es gibt aber einen Ausweg:
Sei $\boldsymbol{a}$ fest gewählt und h eine beschränkte messbare Funktion. Definiere nun das lineare Funktional

$$F(h) := \int_{[0,1]^d} h\left(\boldsymbol{a}^T\boldsymbol{x}\right)\, \mathrm{d}\lambda(\boldsymbol{x})\,.$$

λ ist ein Borelmaß, d.h. F ist ein beschränktes Funktional auf $L^\infty(\mathbb{R})$. Wähle nun

$$h(x) = \mathbb{1}_{[t,\infty)}(x) = \begin{cases} 1, & x \geq t \\ 0, & x < t \end{cases},$$

also die Indikatorfunktion auf $[t, \infty)$. Damit wird dann

$$F(h) = \int_{[0,1]^d} \mathbb{1}_{[t,\infty)}\left(\boldsymbol{a}^T\boldsymbol{x}\right)\, \mathrm{d}\lambda(\boldsymbol{x}) = \lambda\left(H^{(0)}_{\boldsymbol{a},-t}\right) + \lambda\left(H^{(+)}_{\boldsymbol{a},-t}\right) = 0\,.$$

Ebenso ist $F(h) = 0$ auf (t, ∞). Da F linear ist gilt diese Aussage für alle (endlichen) Linearkombinationen von Indikatorfunktionen auf Intervallen. Solche Funktionen liegen dicht in $L^\infty(\mathbb{R})$, d.h. es folgt $F = 0$.
Wählen wir nun den Spezialfall (im mehrdimensionalen)

$$h(\boldsymbol{x}) = \cos(\boldsymbol{m}^T\boldsymbol{x}) + i\sin(\boldsymbol{m}^T\boldsymbol{x})\,,$$

so erhalten wir für die Fouriertransformierte von λ

$$\begin{aligned}\overline{\lambda} = F(h) &= \int_{[0,1]^d} \left[\cos(\boldsymbol{m}^T\boldsymbol{x}) + i\sin(\boldsymbol{m}^T\boldsymbol{x})\right]\, \mathrm{d}\lambda(\boldsymbol{x}) \\ &= \int_{[0,1]^d} \exp\left(i\boldsymbol{m}^T\boldsymbol{x}\right)\, \mathrm{d}\lambda(\boldsymbol{x}) = 0 \quad \forall \boldsymbol{m}\,.\end{aligned}$$

Hieraus folgt, dass λ selbst verschwinden muss (s. Rudin (1976)). Somit ist σ diskriminatorisch. ■

Nun kommen wir zum ersten zentralen Resultat:

Satz 2.3. *Sei σ eine diskriminatorische Funktion im Sinne von Def. 2.3. Dann liegt $\Sigma^d(\sigma)$ dicht in $C([0,1]^d)$. Mit anderen Worten: Für jedes $f \in C([0,1]^d)$ und $\varepsilon > 0$ existiert ein $g(\boldsymbol{x}) \in \Sigma^d(\sigma)$ mit*

$$\|g - f\|_\infty \leq \varepsilon\,,$$

wobei $\|\cdot\|_\infty$ die Supremumsnorm bezeichne, also $\|f\| = \|f\|_\infty = \sup_{\boldsymbol{x}\in[0,1]^d} |f(x)|$.

Beweis. Kern dieses Beweises ist der Satz von Hahn-Banach und der Riesz'sche Darstellungssatz.
Wie schon in Def. 2.2 beschrieben ist $\Sigma^d(\sigma)$ ein linearer Unterraum von $C([0,1]^d)$. Damit $\Sigma^d(\sigma)$ auch dicht in $C([0,1]^d)$ liegt, muss also der Abschluss $\overline{\Sigma}^d(\sigma)$ von $\Sigma^d(\sigma)$ gleich $C([0,1]^d)$ sein.
Ein Korollar des Satzes von Hahn-Banach besagt, dass wenn X ein normierter Raum ist und $U \subset X$, dann sind äquivalent[3]:

1. U liegt dicht in X,
2. $x^* \in X^*$ und $x^*|_U = 0 \Longrightarrow x^* = 0$.

(Beweis s. z.B. Werner (2006), Seite 347).

Angenommen $\Sigma^d(\sigma)$ liege nicht dicht in $C([0,1]^d)$. Dann gibt es ein $G^* \in C^*([0,1]^d)$ mit $G^* \neq 0$, obwohl $G^*(\Sigma^d(\sigma)) = G^*(\overline{\Sigma}^d(\sigma)) = 0$. Nun besagt der Riesz'sche Darstellungssatz[4], dass

$$G^*(h) = \int_{[0,1]^d} h(x)\,\mathrm{d}\lambda(x)$$

für ein endliches reguläres Borelmaß auf $[0,1]^d$ und alle $h \in C([0,1]^d)$. Nun ist aber

$$\int_{[0,1]^d} \sigma(A)\,\mathrm{d}\lambda = 0\,,$$

mit $A \in \boldsymbol{A}([0,1]^d)$, da nach Voraussetzung $G^*(\Sigma^d(\sigma)) = 0$. σ wurde aber als diskriminatorisch angenommen, d.h. es folgt sofort $\lambda = 0$. Dies widerspricht allerdings unserer Annahme, dass

$$G^*(h) = \int_{[0,1]^d} h(x)\,\mathrm{d}\lambda(x) \neq 0\,.$$

Also muss $\Sigma^d(\sigma)$ dicht in $C([0,1]^d)$ liegen. ■

Verallgemeinerungen
Dieses Resultat kann sofort verallgemeinert werden auf alle Funktionen in $L^1([0,1]^d)$, indem wir in der Definition einer diskriminatorischen Funktion von $C(\mathbb{R}^d)$ (Def. 2.3) zu

[3] X^* bezeichne den Dualraum, also die Menge aller linearen Funktionale auf X.
[4] Sei X ein lokal kompakter Hausdorff-Raum. Für jedes stetige lineare Funktional L auf $C^*(X)$ gibt es ein endliches reguläres Borelmaß λ auf X, so dass

$$L(h) = \int_X h(x)\,\mathrm{d}\lambda(x) \quad \forall h \in C(X)\,.$$

$L^1(\mathbb{R}^d)$ (Def. 2.4) übergehen und im Beweis von Satz 2.3 $\mathbb{B}_{[0,1]^d}$ durch $L^\infty([0,1]^d)$ ersetzen.
Die Ergebnisse von Cybenko (1989) können sogar, auch wenn es der Autor nicht anspricht, auf $L^p([0,1]^d)$ mit $p \in [1,\infty)$ erweitert werden. Hierzu muss zunächst Def. 2.4 modifiziert werden (statt L^1 einfach L^p). Im Beweis von Satz 2.3 benutzen wir dann den zugehörigen Dualraum $(L^p)^*([0,1]^d) = L^q([0,1]^d)$. Natürlich verwenden wir in der Definition der Dichtheit als Metrik die L^p-Norm $\|\cdot\|_p$. Somit erhalten wir durch Kombination von Satz 2.3 mit Lemma 2.1 mit den eben angestellten Überlegungen:

Korollar 2.1. *Neuronale Netze mit einem layer und einer beliebigen stetigen sigmoiden Aktivierungsfunktion können L^p-Funktionen auf Kompakta $K \in \mathbb{R}$ mit beliebiger Genauigkeit bezüglich der L^p-Norm $\|\cdot\|_p$ approximieren, falls keine Nebenbedingungen an die Anzahl der nodes M oder die Gewichte u_i gestellt werden.*

Wir können aber noch einen Schritt weitergehen. Es gilt folgender Satz:

Satz 2.4 (Lusin). *Sei $f \in C([a,b]^d)$ messbar und λ ein Borelmaß. Dann gibt es für jedes $\varepsilon > 0$ eine Funktion $g \in C([a,b]^d)$, so dass*

$$\lambda\left(\{\boldsymbol{x} : f(\boldsymbol{x}) \neq g(\boldsymbol{x})\}\right) \leq \varepsilon\,.$$

Das heißt jede messbare Funktion ist fast überall stetig.

Beweis. Siehe Lusin (1912).

Als Konsequenz dieses Satzen gelten alle bisher gemachten Aussagen auch für nichtstetige, aber *messbare* Funktionen.

Die bisherigen Ergebnisse gelten in ähnlicher Form auch für nicht-stetige Aktivierungsfunktionen wie z.B. die Stufenfunktion $\sigma(x) = 1$ für $x > 0$, $\sigma(x) = 0$ für $x < 0$. Für Funktionen dieser Art gibt es keine sehr guten Trainingsalgorithmen, weshalb sie in der praktischen Anwendung von Neuronalen Netzen kaum vertreten sind. Aber der Vollständigkeit halber sollen sie hier dennoch Erwähnung finden.

Satz 2.5. *Sei $\sigma \in L^1(\mathbb{R})$ mit*

$$\int_{\mathbb{R}} \sigma(x)\,dx \neq 0\,.$$

Dann liegt $\Sigma^d(\sigma)$ dicht in $L^1([0,1]^d)$.

Beweis. Zu zeigen ist, dass solche σ diskriminatorisch sind. Dann können wir Korollar 2.1 anwenden und erhalten das gewünschte Ergebnis.
In Beweis von Lemma 2.1 konnten wir auf die Tatsache zurückgreifen, dass σ_τ gegen ein geeignetes γ strebte für $t \to \infty$. Dies ist nun nicht mehr möglich. Diskriminatorisch heißt, dass[5]

[5] Mit der Notation $\mathrm{d}\boldsymbol{x}$ meinen wir das d-dimensionale Lebesgue-Maß $\mathrm{d}x_1 \times ... \times \mathrm{d}x_d$.

$$\int_{[0,1]^d} \sigma\left(A(\boldsymbol{x})\right)\, h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0 \quad \forall A \in \boldsymbol{A}\left([0,1]^d\right) \quad \Longrightarrow \quad h(\boldsymbol{x}) = 0 \text{ fast-überall}\,.$$

Analog zu zuvor definieren wir das lineare Funktional F auf $L^1(\mathbb{R})$ als

$$F(g) := \int_{[0,1]^d} g\left(A(\boldsymbol{x})\right) h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

wobei h wieder beschränkt und messbar sei. Da das Integral zudem über einem Kompaktum ausgeführt wird, existiert es.
Nun wählen wir $g = \sigma$ als Dilationen und Translationen von $\sigma \in L^1(\mathbb{R})$, d.h.

$$\sigma_{\kappa,\tau}\left(A(\boldsymbol{x})\right) = \sigma\left(\kappa\left(A(\boldsymbol{x})\right) + \tau\right)$$

mit $\kappa, \tau \in \mathbb{R}^d$ beliebig. Wir nehmen nun an, dass gilt

$$0 = F(\sigma_{\kappa,\tau}) = \int_{[0,1]^d} \sigma\left(\kappa\left(A(\boldsymbol{x})\right) + \tau\right) h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \quad \forall A \in \boldsymbol{A}\left([0,1]^d\right)\,.$$

Die Fouriertransformierte einer Linearkombination ist $\overline{f}(f + \alpha g) = \overline{f}(f) + \alpha\overline{f}(g)$ für $f, g \in L^1(\mathbb{R})$, $\alpha \in \mathbb{C}$, d.h.

$$\overline{\sigma_{\kappa,\tau}}(\omega) = \frac{\exp\left(i\tau\omega/\kappa\right)}{\kappa}\overline{\sigma}\left(\frac{\omega}{\kappa}\right)\,.$$

Da κ, τ beliebig gewählt wurden, verschwindet dieser Ausdruck nur, wenn $\omega = 0$. Nun ist allerdings

$$\overline{\sigma_{1,0}}(\omega) = \overline{\sigma}(\omega) = \int_{-\infty}^{\infty} \sigma(x)\, e^{-i\omega x}\mathrm{d}x \quad \Longrightarrow \quad \overline{\sigma_{1,0}}(0) = \int_{-\infty}^{\infty} \sigma(x)\,\mathrm{d}x\,.$$

Und wir haben angenommen, dass $\int_{\mathbb{R}} \sigma(x)\mathrm{d}x \neq 0$.
An dieser Stelle benötigen wir nun Cybenko (1989) folgend folgendes Theorem, das auf Wiener zurückgeht:

Satz 2.6 (Wieners allgemeines Tauber-Theorem). *Ein Unterraum $S \in L^1(\mathbb{R}^d)$ heißt translations-invariant, wenn $f(y - x) \in S$ für alle $f \in S$ und $x \in \mathbb{R}^d$. Sei $f \in L^1(\mathbb{R}^d)$ und sei S der kleinste abgeschlossene translations-invariante Unterraum mit $f \in S$. Dann ist $S = L^1(\mathbb{R}^d)$ genau dann wenn $\overline{f}(s) \neq 0$ für alle $s \in \mathbb{R}$.*

Beweis. Siehe Wiener (1932).

Mit diesem Tauber-Argument haben wir eine notwendige und hinreichende Bedingung dafür, dass ein abgeschlossener translations-invarianter Unterraum in $L^1(\mathbb{R}^d)$ gleich dem

gesamten Raum ist.
Mit diesem Argument ist also klar, dass der von $\sigma_{\kappa,\tau}$ aufgespannte Unterraum $S := \text{span}\{\sigma_{\kappa,\tau} : \kappa, \tau \in \mathbb{R}\}$ dicht in $L^1(\mathbb{R})$ liegt. Da $F(\sigma_{\kappa,\tau}) = 0$ angenommen wurde, folgt also $F(g) = 0$ für alle $g \in L^1(\mathbb{R})$. Nun knüpfen wir an den Beweis von Lemma 2.1 an und wählen für g wieder $g = \exp(i\boldsymbol{m}^T\boldsymbol{x})$ und folgern

$$F(g) = \int_{[0,1]^d} \exp(i\boldsymbol{m}^T\boldsymbol{x})h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$$

für alle $\boldsymbol{m}$. Hieraus folgt dann wieder, dass $g = 0$. Also ist σ diskriminatorisch. ■

2.2.2 Multi-layer-SNNs

Wir erweitern nun die Aussagen über die Approximationsfähigkeit der sigmoiden Neuronalen Netze von single-layer auf multi-layer Architekturen. Weiterhin werden wir die Einschränkung auf den Einheitswürfel $[0,1]^d$ fallen lassen und zu Argumenten aus beliebigen Kompakten Teilmengen von $\mathbb{R}^d$ übergehen. Die Ergebnisse aus Hornik, Stinchcombe & White (1989) enthalten allerdings neben diesen Erweiterungen einige unnötige Einschränkungen. So werden nur squashing-Funktionen betrachtet und nicht die allgemeineren sigmoiden Funktionen. Wir werden allerdings eine Verallgemeinerung diskutieren, so dass wir die Approximationsfähigkeit von single-layer Neuronalen Netzen im Sinne von Cybenko (1989) zurückerhalten.
Dieser Abschnitt ist also folgendermaßen zu verstehen: Zunächst gehen wir von single-layer zu multi-layer Neuronalen Netzen über und zeigen deren Approximationsfähigkeit im Raum der stetigen Funktionen und im Raum der messbaren Funktionen (in Bezug auf ein endliches Maß, so dass das Lebesgue-Maß zunächst ausgeschlossen ist, da aber alle Sätze auch auf Kompakta in $\mathbb{R}^d$ gelten sind die Ergebnisse auch korrekt für das Lebesgue-Maß auf solchen Kompakta). Da Wahrscheinlichkeitsmaße spezielle endliche Maße sind, sind die Resultate auch anwendbar auf "noisy data" (siehe Teil II). Wie erwähnt werden wir die Ergebnisse dann von squashing-Funktionen auf allgemeine sigmoide Funktionen erweitern und so multi-layer Neuronale Netze auf single-layer Netze zurückführen.

Wir erweitern die Definition aus 2.2.1 und benutzen die Notation aus Hornik, Stinchcombe & White (1989).

Definition 2.5 (Klasse der affinen Kompositionen, Teil 2). *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine Borel-messbare Funktion. $\Sigma\Pi^d(\sigma)$ bezeichne die Klasse von Funktionen $g : K \to \mathbb{R}$, $K \subset \mathbb{R}^d$, der Form*

$$g(\boldsymbol{x}) = \sum_{i=1}^{M} u_i \prod_{k=1}^{l_i} \sigma\left(A_{ik}(\boldsymbol{x})\right)$$

mit $u_i \in \mathbb{R}$, $l_i \in \mathbb{N}$, $A_{ik} \in \boldsymbol{A}(K)$ und $M = 1, 2, ...$

Anmerkung 2.1. Es ist offensichtlich, dass die Klasse Σ^d ein Spezialfall von $\Sigma\Pi^d$ ist für $l_j = 1$ für alle j. Die Klasse Σ^d ist *keine* Algebra von Funktionen. $\Sigma\Pi^d$ hingegen ist eine Funktionen-Algebra (Beweis siehe weiter unten) generiert von Σ^d. Im Beweis des folgenden Satzes können wir aufgrund dieser Eigenschaft das Theorem von Stone-Weierstrass anwenden. Zum Vergleich: Im Falle von Σ^d konnten wir den Satz von Hahn-Banach zu Rate ziehen als es um den Beweis der Dichtheit von Σ^d in $C^d([0,1]^d)$ ging.
Weiterhin ist noch erwähnenswert, dass $\Sigma\Pi^d$ allgemeiner ist als die Klasse von l-layer Neuronalen Netzen. Und zwar gilt für diese $l_j = l$ für alle j. Wir bezeichnen diese Klasse mit Σ_l^d.

Das Hauptresultat aus Hornik, Stinchcombe & White (1989) ist nun das folgende:

Satz 2.7. *Sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetige nicht-konstante Funktion. Dann liegt $\Sigma\Pi^d(\sigma)$ gleichmäßig dicht in Kompakta in $C^d(\mathbb{R}^d)$.*

Eine Menge $S \subset C^d(\mathbb{R}^d)$ liegt gleichmäßig dicht in Kompakta in $C^d(\mathbb{R}^d)$, wenn für alle kompakten Teilmengen $K \subset \mathbb{R}^d$ gilt, dass S ρ_K-dicht in $C^d(K)$ liegt. Hierbei bezeichnet $\rho_K = \sup_{\boldsymbol{x} \in K} |f(x) - g(x)|$ mit $f, g \in C^d$ wieder die Supremumsnorm.

Beweis. Der Beweis dieses Satzes benötigt das Theorem von Stone-Weierstrass, also eine Verallgemeinerung des Approximationssatzes von Weierstrass. Es besagt folgendes:

> **Satz 2.8 (Stone-Weierstrass (1885, 1937)).** *Sei P eine Unteralgebra der Algebra der reellen stetigen Funktionen A auf einer kompakten Menge K. Wenn A Punkte in K separiert (d.h. $\forall x \neq y \in K\ \exists g \in P : g(x) \neq g(y)$), in keinem Punkt verschwindet ($\forall x \in K\ \exists g \in P : g(x) \neq 0$) und bezüglich komplexer Konjugation abgeschlossen ist, so liegt P ρ_K-dicht in A.*
>
> *Beweis.* Siehe z.B. Werner (2007).

Zu zeigen ist also, dass $\Sigma\Pi^d$ solch eine Unteralgebra ist. Die entscheidende Eigenschaft von σ ist die nicht-Konstantheit. Die Stetigkeit von σ ist Voraussetzung für das Stone-Weierstrass-Theorem.
Sei $K \subset \mathbb{R}^d$ kompakt. Seien $x, y \in K$, $x \neq y$. Dann existiert ein $A \in \boldsymbol{A}$, so dass $\sigma(A(x)) \neq \sigma(A(y))$. Dies ist leicht einzusehen, denn für $a, b \in \mathbb{R}$, $a \neq b$ gibt es ein $A \in \boldsymbol{A}$ mit $A(x) = a \neq A(y) = b$. Weiterhin ist dann $\sigma(A(x)) \neq \sigma(A(y))$. Also ist $\Sigma\Pi^d$ separierend auf K.
Es gibt in jedem Fall mindestens ein $\sigma(A(\cdot)) = \text{const} \neq 0$. Um dies zu sehen, wähle $b \in \mathbb{R}$, so dass $\sigma(b) \neq 0$ und setze $A(x) = 0x + b$. Dann ist $G(A(x)) = G(b)$ für alle x. Also verschwindet $\Sigma\Pi^d$ in keinem Punkt.
$\Sigma\Pi^d$ ist offensichtlich abgeschlossen gegenüber der komplexen Konjugation, weil alle involvierten Größen reell sind.
Nun können wir den Satz von Stone-Weierstrass anwenden und erhalten sofort das Resultat, da K beliebig gewählt wurde. ■

Es ist möglich Satz 2.7 auf die Menge von Borel-messbaren Funktionen $\mathcal{M}^d(\mu)$ zu verallgemeinern. Hierbei ist μ ein endliches Maß, d.h. das Lebesgue-Maß ist hier ausgeschlossen! Um dies zu tun, muss allerdings zunächst eine Metrik ρ_μ auf $\mathcal{M}^d(\mu)$ definiert werden, so dass der Begriff Dichtheit in diesem Raum einen Sinn ergibt. Hornik, Stinchcombe & White (1989) verwenden hierfür $\rho_\mu : \mathcal{M}^d \times \mathcal{M}^d \to \mathbb{R}^+$ mit

$$\rho_\mu(f,g) := \inf\{\varepsilon > 0 : \mu\left(\{x : |f(x) - g(x)| > \varepsilon\}\right) < \varepsilon\} \,. \tag{2.4}$$

Weiterhin ist es möglich die Stetigkeitsforderung in Satz 2.7 fallen zu lassen und stattdessen die Aussage auf squashing functions zu verallgemeinern. Das führt dann zu folgendem Satz:

Satz 2.9. *Für jede squashing function σ und jedes endliche Borel Maß μ auf $\mathbb{R}^d$ liegt $\Sigma\Pi^d(\sigma)$ dicht in Kompakta in $C(\mathbb{R}^d)$ und dicht in $\mathcal{M}^d$ bezüglich ρ_μ.*

Satz 2.7 macht eine Aussage über die allgemeine Funktionenklasse $\Sigma\Pi^d$. Diese Aussagen lassen sich mit folgendem Lemma aus Hornik, Stinchcombe & White (1989) auf Σ^d erweitern, wobei auch die Stetigkeitsforderung an σ überflüssig wird:

Lemma 2.2. *Jede stetige squashing function f kann beliebig genau durch Superpositionen von Translationen und Dilationen von allgemeinen squashing functions g dargestellt werden, d.h. durch ein Element aus $\Sigma^1(g)$. Genauer: Für jedes $\varepsilon > 0$ gibt es ein $h \in \Sigma^1(g)$ mit $\sup_{x\in\mathbb{R}^d} |f(x) - h(x)| < \varepsilon$.*

Beweis. Kern dieses Beweises ist die Beschränktheit von f und g. Wähle zunächst ein beliebiges $\varepsilon > 0$. Wir müssen nun Konstanten u_i und affine Funktionen A_i finden, so dass

$$\sup_{x\in\mathbb{R}} \left| f(x) - \sum_{i=1}^{m-1} u_i g(A_i(x)) \right| < \varepsilon \,.$$

Wähle nun m so, dass $1/m < \varepsilon/2$ und setze $u_i = 1/m$. Wähle weiterhin $M > 0$, so dass $g(-M) < \varepsilon/(2m)$ und $g(M) > 1 - \varepsilon/(2m)$). Dieses M existiert, da g eine squashing function ist. Setze nun $r_i = \sup\{x : f(x) = i/m\}$ und $r_m = \sup\{x : f(x) = 1 - 1/(2m)\}$. Diese existieren, da f eine stetige squashing function ist. Wir definieren nun weiterhin für $r < s$ $A_{r,s} \in A$ die eindeutige affine Funktion mit $A_{r,s}(r) = M$ und $A_{r,s}(s) = -M$. Die gesuchte Approximation ist nun $h(x) = \sum_{i=1}^{m-1} u_i g(A_{r_i,r_{i+1}}(x))$. ■

An dieser Stelle können die Resultate von Hornik, Stinchcombe & White (1989) verallgemeinert werden, da die Sätze in ihren Arbeiten nur auf squashing functions bezogen sind und nicht auf die allgemeineren sigmoiden Aktivierungsfunktionen. Wir sehen aber, dass der Beweis von Lemma 2.2 nur die Beschränktheit von σ benötigt, so dass er ohne weiteres auf allgemeine sigmoide Aktivierungsfunktionen (die ja beschränkt sind) erweiterbar ist. Mit dieser Überlegung können wir nun das folgende Korollar zu Satz 2.7 formulieren, das von $\Sigma\Pi^d$ übergeht zu den single-layer Neuronalen Netzen Σ^d, im Unterschied zu den Ergebnissen im vorigen Abschnitt 2.2.1 allerdings wie angekündigt auf beliebigen kompakten Teilmengen von $\mathbb{R}^d$ und mit allgemeineren Aktivierungsfunktionen:

Korollar 2.2. *Für jede beschränkte nicht-konstante Aktivierungsfunktion σ und jedes endliche Borel Maß μ auf $\mathbb{R}^d$ liegt $\Sigma^d(\sigma)$ dicht in Kompakta in $C(\mathbb{R}^d)$ und dicht in $\mathcal{M}^d$ bezüglich ρ_μ.*

Beweis. Wählen wir die folgende stetige (squashing) Funktion

$$c(x) = \frac{1}{2}\left[1 + \cos(x + 3\pi/2)\right] \mathbb{1}_{[-\pi/2,\pi/2]} + \mathbb{1}_{[\pi/2,\infty)} ,$$

so ergibt sich mit Lemma 2.2, dass auf dem Intervall $[-T, T]$ die Kosinus-Funktion durch ein Element aus $\Sigma^1(\sigma)$ approximiert werden kann. Nun muss lediglich noch bermerkt werden, dass die trigonometrischen Polynome $\sum_{i=1}^m u_i \prod_{k=1}^{l_j} \cos(A_{ik}(\cdot))$ mit Hilfe der Identität $\cos a \cdot \cos b = \cos(a+b) - \cos(a-b)$ umgeschrieben werden können in die Form $\sum_{i=1}^n p_i \cos(A_i(\cdot))$. Mit Satz 2.7 folgt dann das Gewünschte. ■

Wir erweitern die gemachten Aussagen nun auf $L^p(\mathbb{R}^d, \mu)$, wobei μ ein endliches Maß sei. Wie schon mehrfach erwähnt ist das Lebesgue-Maß nicht endlich auf $\mathbb{R}^d$, so dass der folgende Satz für das Lebesgue-Maß nur auf Kompakta $K \subset \mathbb{R}^d$ gilt:

Korollar 2.3. *Sei σ eine beschränkte nicht-konstante Aktivierungsfunktion und μ ein endliches Borel-Maß auf $\mathbb{R}^d$. Gibt es ein $K \subset \mathbb{R}^d$ mit $\mu(K) = 1$, so liegt $\Sigma^d(\sigma)$ für jedes $p \in [1, \infty)$ dicht in $L^p(\mathbb{R}^d, \mu)$ bzgl. $\| \cdot \|_p$.*

Beweis. Wähle ein beliebiges $g \in L^p$ und ein $\varepsilon > 0$. Zu zeigen ist, dass es ein $f \in \Sigma^d(\sigma)$ gibt mit $\|f - g\|_p < \varepsilon$. Mit einem standard-$\varepsilon/3$-Argument (s. z.B. Werner (2006), S. 294 ff) kann man zeigen, dass es für jede Funktion $h \in L^p$ eine stetige Funktion $\tilde{f}$ gibt mit $\|\tilde{f} - h\|_p < \varepsilon/3$. Man wähle $M \in \mathbb{R}$ groß genug, so dass $h = g \cdot \mathbb{1}_{\{|g| \le M\}}$, und damit $\|g - h\|_p < \varepsilon/3$. Nun liegt $\Sigma^d(\sigma)$ dicht in Kompakta, d.h. es gibt ein $f \in \Sigma^d(\sigma)$ mit $\sup_{x \in K} |f(x) - \tilde{f}(x)|^p < (\varepsilon/3)^p$. Nach Voraussetzung ist $\mu(K) = 1$, d.h. $\|\tilde{f} - f\|_p < \varepsilon/3$. Wir fassen also die gemachten Überlegungen zusammen: $\|g - f\|_p \le \|g - h\|_p + \|h - \tilde{f}\|_p + \|\tilde{f} - f\|_p = \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon$. ■

Nun haben wir also multi-layer Neuronale Netze zurückgeführt auf single-layer Neuronale Netze. Fassen wir die in diesem und dem letzten Teil vorgelegten Resultate zusammen, so erhalten wir das folgende vorläufige Endresultat:

Korollar 2.4. *Neuronale Netze (single-layer und multi-layer) stellen universelle Approximatoren auf $L^p(\mathbb{R}^d)$ dar. Falls ein Neuronales Netz in einer technischen Anwendung versagt, so muss dies an unzureichendem "Lernen" der Parameter oder zu kleiner bzw. zu großer Anzahl an Nodes liegen.*

Zudem haben wir in diesen Überlegungen zwar die grundsätzliche Approximationsfähigkeit von Neuronalen Netzen dargestellt, aber noch keine Aussagen über die entsprechenden Konvergenzraten gemacht (hierfür s. Abschnitt 4.5).

2.2.3 RBFNs

Nun verlassen wir die sigmoiden Neuronalen Netze und wenden uns den Radialen Basisfunktions-Netzwerken zu. Eine recht umfassende Darstellung der Approximationseigenschaften dieses Typs ist in Park & Sandberg (1991) veröffentlicht worden.
Wir betrachten single-layer Neuronale Netze mit Basisfunktionen $\widetilde{\Phi} : \mathbb{R}^d \to \mathbb{R}$, die die Eigenschaft haben, dass $\widetilde{\Phi}(\boldsymbol{x}) = \widetilde{\Phi}(\boldsymbol{y})$ wenn $\|\boldsymbol{x}\| = \|\boldsymbol{y}\|$. Insbesondere interessieren uns, wie schon erwähnt, Kerne der Form

$$\Phi_i(\boldsymbol{x}) = \Phi_i\left(\frac{\boldsymbol{x} - \boldsymbol{t}_i}{\sigma_i}\right) = \varphi\left(\frac{\|\boldsymbol{x} - \boldsymbol{t}_i\|}{\sigma_i}\right) , \tag{2.5}$$

wobei $\varphi : [0, \infty) \to \mathbb{R}$ eine Aktivierungsfunktion sei, $\boldsymbol{t}_i \in \mathbb{R}^d$ der Zentrierungsvektor und $\sigma_i \neq 0$ ein Glättungsparameter. Besondere Bedeutung haben hierbei Gauss'sche Glockenkurven $\varphi(\cdot) = \exp(\cdot)$. Welche Norm wir verwenden lassen wir zunächst offen, z.B. $\|\cdot\| = \|\cdot\|_2$. Wir definieren zunächst wieder eine Funktionenklasse:

Definition 2.6 (Klasse der Kernfunktionen). *Es sei $\varphi : \mathbb{R}^d \to \mathbb{R}$ eine Kernfunktion. $\mathcal{K}^d(\varphi)$ bezeichne die Klasse von Funktionen $g : \mathbb{R}^d \to \mathbb{R}$ der Form*

$$g(\boldsymbol{x}) = \sum_{i=1}^{M} u_i \, \varphi\left(\frac{\|\boldsymbol{x} - \boldsymbol{t}_i\|}{\sigma}\right)$$

mit $u_i \in \mathbb{R}$, $\boldsymbol{t}_i \in \mathbb{R}^d$, $\sigma > 0$ und $M = 1, 2, ...$

Anmerkung 2.2. $\mathcal{K}^d(\varphi)$ stellt ein Netzwerk dar, das denselben Glättungsparameter für *alle* hidden nodes hat. Wir werden sehen, dass die entscheidende Eigenschaft der Kernfunktion, so dass $\mathcal{K}^d(\varphi)$ dicht in einem Funktionenraum liegt, $\int_{\mathbb{R}^d} \varphi \neq 0$ ist.

Wir schreiten nun fort und entwickeln Approximationsaussagen für $\mathcal{K}^d(\varphi)$. Die Originalveröffentlichung hierzu stammt wie erwähnt von Park & Sandberg (1991). Zunächst ein Lemma:

Lemma 2.3. *Sei $f \in L^p(\mathbb{R}^d)$ für $p \in [1, \infty)$ und sei $\phi \in L^1(\mathbb{R}^d)$, so dass $\int_{\mathbb{R}^d} \phi = 1$. Definiere für $\varepsilon > 0$ $\phi_\varepsilon = (1/\varepsilon^d)\phi(\boldsymbol{x}/\varepsilon)$. Dann gilt*

$$\lim_{\varepsilon \to 0} \|\phi_\varepsilon \star f - f\|_p = 0 \, ,$$

wobei $\star$ den Faltungsoperator bezeichne.

Beweis. Wir stellen zunächst fest, dass $\phi_\varepsilon \star f \in L^p(\mathbb{R}^d)$ (dies ergibt sich direkt aus der Anwendung der Hölder-Ungleichung). Dann ist

$$
\begin{aligned}
(\phi_\varepsilon \star f)(\boldsymbol{x}) - f(\boldsymbol{x}) &= \frac{1}{\varepsilon^d} \int_{\mathbb{R}^d} \phi(\boldsymbol{y}/\varepsilon)\, f(\boldsymbol{x}-\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} - f(\boldsymbol{x}) \\
&= \int_{\mathbb{R}^d} \phi(\boldsymbol{p})\, [f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x}]\ \mathrm{d}\boldsymbol{p}\,,
\end{aligned}
$$

wobei $\boldsymbol{p} = \boldsymbol{y}/\varepsilon$ umparametrisiert wurde, so das $\mathrm{d}\boldsymbol{y} = \varepsilon^d \mathrm{d}\boldsymbol{p}$. Weiterhin ist mit dem Satz von Fubini:

$$
\begin{aligned}
\|\phi_\varepsilon \star f - f\|_p &= \left[\int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d} \phi(\boldsymbol{p})\, [f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x}]\ \mathrm{d}\boldsymbol{p} \right|^p \ \mathrm{d}\boldsymbol{x} \right]^{1/p} \\
&\le \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} |\phi(\boldsymbol{p})|^p\, |f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x})|^p\ \mathrm{d}\boldsymbol{x} \right)^{1/p} \mathrm{d}\boldsymbol{p} \\
&\le \int_{\mathbb{R}^d} |\phi(\boldsymbol{p})| \left(\int_{\mathbb{R}^d} |f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x})|^p\ \mathrm{d}\boldsymbol{x} \right)^{1/p} \mathrm{d}\boldsymbol{p} \\
&= \int_{\mathbb{R}^d} |\phi(\boldsymbol{p})|\ \|f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x})\|_p\, \mathrm{d}\boldsymbol{p}
\end{aligned}
$$

Weil $\|f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x})\|_p \le 2\|f\|_p$ können wir den Satz von Lebesgue auf $|\phi(\boldsymbol{p})|\ \|f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x})\|_p$ anwenden:

$$
\lim_{\varepsilon\to 0} \int_{\mathbb{R}^d} |\phi(\boldsymbol{p})|\ \|f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x})\|_p\, \mathrm{d}\boldsymbol{p} = \int_{\mathbb{R}^d} \lim_{\varepsilon\to 0} |\phi(\boldsymbol{p})|\ \|f(\boldsymbol{x}-\varepsilon\boldsymbol{p}) - f(\boldsymbol{x})\|_p\, \mathrm{d}\boldsymbol{p}\,.
$$

Dieser Ausdruck verschwindet (Kontinuität der Norm). ■

Der folgende Satz stellt ein Hauptresultat dar:

Satz 2.10. *Sei die Kernfunktion $\varphi \in L^1(\mathbb{R}^d)$ beschränkt, stetig fast überall und $\int_{\mathbb{R}^d} \varphi \neq 0$. Dann liegt die Familie $\mathcal{K}^d(\varphi)$ dicht in $L^p(\mathbb{R}^d)$ für $p \in [1,\infty)$.*

Beweis. Der Raum der stetigen Funktionen mit kompaktem Träger liegt dicht in $L_p(\mathbb{R}^d)$, d.h. f kann bezüglich $\|\cdot\|_p$ approximiert werden durch solch ein f_c. Wir definieren nun $\overline{\varphi} := \varphi/\int_{\mathbb{R}^d} \varphi$. Aus Lemma 2.3 wissen wir, dass $\overline{\varphi}_\tau \star f_c$, mit $\overline{\varphi}_\tau$ definiert analog zu ϕ_ε, f_c beliebig gut approximiert in $L_p(\mathbb{R}^d)$ für genügend kleine τ.
Als nächstes diskretisieren wir die Faltung $\overline{\varphi}_\tau \star f_c$, und zwar mit Hilfe der Riemann-Summe als Annäherung an das Faltungsintegrals. $\overline{\varphi}_\tau \star f_c$ ist in der Tat Riemann integrierbar auf $[-T,T]^d$, da es fast überall stetig ist und beschränkt durch $\|\overline{\varphi}_\tau\|_\infty \|f_c\|_\infty < \infty$. Sei also y_j eine äquidistante Einteilung von $[-T,T]^d$, so dass

$$
R_N(x) = \sum_{i=1}^{N^d} \overline{\varphi}_\tau(\boldsymbol{x}-\boldsymbol{y}_j) f_c(\boldsymbol{y}_j) \left(\frac{2T}{N}\right)^d
$$

Es gilt $R_N(x) \to (\overline{\varphi}_\tau \star f_c)(x)$ punktweise. Wir müssen allerdings noch die Konvergenz in $L_p(\mathbb{R}^d)$ zeigen. Zu diesem Zweck stellen wir zunächst fest, dass $\|\overline{\varphi}_\tau \star f_c\|_p \to 0$ auf $\mathbb{R}^d \backslash [-T,T]^d$ für $T \to \infty$.
Nun kann mit Hilfe der Jensen Ungleichung angewandt auf $\left[(1/N^d) \sum |\overline{\varphi}_\tau(\boldsymbol{x} - \boldsymbol{y}_j)|\right]^p$ gezeigt werden (die Funktion ist in der Tat konvex, da $p \geq 1$), dass

$$\int_{\mathbb{R}^d \backslash [-T,T]^d} |R_N(\boldsymbol{x})|^p \, \mathrm{d}\boldsymbol{x} \to 0 \quad \text{für} \quad T \to \infty \, .$$

Zusammengefasst gilt also:

$$\int_{[-T,T]^d} |\overline{\varphi} \star f_c(\boldsymbol{x}) - R_N(\boldsymbol{x})|^p \, \mathrm{d}\boldsymbol{x} \to 0 \quad \text{für} \quad N \to \infty \, .$$

Hierfür haben wir den Satz von Lebesgue angewandt und die Tatsache, dass die Riemann-Summe gegen das Riemann-Integral konvergiert. Nun benötigen wir noch die Dreiecksungleichung:

$$\|R_N - \overline{\varphi}_\tau \star f_c\|_p \leq \|R_N - \overline{\varphi}_\tau \star f_c) \, \mathbb{1}_{[-T,T]^d}\|_p + \|(\overline{\varphi}_\tau \star f_c) \, \mathbb{1}_{\mathbb{R} \backslash [-T,T]^d}\|_p + \|R_N \, \mathbb{1}_{\mathbb{R} \backslash [-T,T]^d}\|_p \, .$$

Also konvergiert R_N in $L^p(\mathbb{R}^d)$ gegen $\overline{\varphi}_\tau \star f_c$ für große T, N. Außerdem approximiert $\overline{\varphi}_\tau \star f_c$ wie gezeigt f_c wenn τ klein genug ist. Und f_c approximiert f beliebig genau. Nun stellen wir noch fest, dass $R_N \in \mathcal{K}^d(\varphi)$ für $u_i = (2T/N\tau)^d f_c(y_i) / \int_{\mathbb{R}^d} \varphi$. ■

Es kann weiterhin gezeigt werden, dass RBFNs, genau wie SNNs (s. Korollar 2.2), dicht in $C(\mathbb{R}^d)$ liegen in Bezug auf ein endliches Maß:

Korollar 2.5. *Sei die Kernfunktion $\varphi \in L^1(\mathbb{R}^d)$ beschränkt und fast überall stetig mit $\int_{\mathbb{R}^d} \varphi \neq 0$. Sei weiterhin μ ein endliches Maß auf $\mathbb{R}^d$. Dann liegt die Familie $\mathcal{K}^d(\varphi)$ dicht in $C(\mathbb{R}^d)$ bezüglich ρ_μ (definiert in Glg. (2.4)).*

Beweis. Der Beweis läuft analog zu dem von Korollar 2.2 und kann im Detail nachgelesen werden in Park & Sandberg (1991).

Damit ein RBFN also ein universeller Approximator in $L^p(\mathbb{R})$ ist, muss die Kernfunktion die Eigenschaften erfüllen, dass sie beschränkt, fast überall stetig und im Mittel ungleich Null ist. φ muss also speziell nicht unbedingt eine radiale Basisfunktion sein. Wir konzentrieren uns aber in dieser Dissertation auf diesen Spezialfall.

Anmerkung 2.3. In Satz 2.5, bzw. im Beweis hierzu, haben wir gezeigt, dass die Eigenschaft $\int_{\mathbb{R}} \sigma \neq 0$ direkt impliziert, dass σ diskriminatorisch ist. Wir verwendeten hierzu Wieners allgemeines Tauber-Theorem. In Satz 2.10, der das Analogon zu diesem Satz für RBFNs darstellt, nun für RBFNs, fordern wir zusätzlich, dass die Kernfunktion φ beschränkt und zumindest fast überall stetig ist. Diese Forderung können wir auch nicht

ohne weiteres fallenlassen, da wir im Beweis sicherstellen müssen, dass die Riemann-Summe gegen das Faltungsintegral konvergiert. Die Frage, die sich nun stellt ist, ob ein φ definiert wie in Glg. (2.5) mit den zusätzlichen Voraussetzung von Satz 2.10 auch diskriminatorisch ist. Dann könnten wir zeigen, dass diese Eigenschaft einer Basisfunktion entscheidend für die Approximationsfähigkeiten des Netzes ist. Nur leider ergibt sich hierbei ein Problem: Eine Anpassung des Beweises von Satz 2.5 für Kernfunktionen der Form von Glg. (2.5) ist nicht möglich, da das allgemeine Tauber-Argument ein flexibles σ benötigt, analog zu $\text{span}\{\varphi_{\kappa,\tau} : \kappa, \tau \in \mathbb{R}\}$. Nun ist σ aber für $\mathcal{K}^d(\varphi)$ festgelegt, so dass $\overline{\varphi}_{\kappa,\tau} = 0$ nicht mehr gegeben wäre. Insofern stellt $\mathcal{K}^d(\varphi)$ mit festem σ keine Klasse von diskriminatorischen Funktionen dar. Definiert man hingegen $\mathcal{K}^d(\varphi)$ mit σ_i statt σ, $i = 1, ..., M$, so können wir den angesprochenen Beweis anpassen und erhalten in der Tat eine diskriminatorische Klasse.

3

Redundanz & Komplexität bei Wavelet-Entwicklungen

Wir haben in den vorangegangenen Abschnitten Funktionsklassen wie $\Sigma^d(\sigma)$ und $\mathcal{K}^d(\varphi)$ betrachtet, wobei wir zeigen konnten, dass letztere unter bestimmten Annahmen sogar dicht in $L^p(\mathbb{R}^d)$ liegen. Diese Klassen haben wir als sigmoide Neuronale Netze beziehungsweise radiale Basisfunktions-Netzwerke interpretiert. Worüber wir allerdings keine Aussage gemacht haben ist die Frage, ob diese Klassen eine Funktion auch eindeutig approximieren und ob die Vertreter orthogonal zueinander sind und eine Basis von $L^p(\mathbb{R}^d)$ bilden. Denn nur in diesem Fall kann man sichergehen, dass die Approximation *eindeutig* ist, also in gewissem Sinne optimal. Zudem machen die bisher gegebenen Sätze keinerlei Aussage über die Wahl der *Entwicklungskoeffizienten*, es wurden lediglich *Existenzbeweise* geführt. Carroll & Dickinson (1989) versuchen in ihrer Arbeit einen Bezug der Koeffizienten einer "besten" Approximation mit der Radon-Transformation herzustellen. Dies gelingt zwar nicht explizit, dafür kann auf Grund ihrer Arbeit ein konstruktiver Beweis von ähnlichen Approximationssätzen geführt werden. Weiterhin zeigen diese Autoren wie die Radon-Transformation zur Bestimmung der Entwicklungskoeffizienzen genutzt werden kann.

Ein alternativer Ansatz, der von einer orthogonalen Basis des zu approximierenden Funktionenraumes ausgeht, stellen Wavelet-Neuronale-Netze dar. Eine der ersten Veröffentlichung zu diesem Thema kam 1992 von Zhang & Benveniste. In erster Linie gilt es natürlich eines zu bewahren beim Übergang von RBFNs zu WNNs: Die universellen Approximations- bzw. Dichtheitseigenschaften des resultierenden Netzwerkes. Zusätzlich soll das Netzwerk aber konstruktiver Natur sein, das heißt wir streben nach einem expliziten Zusammenhang zwischen den Netzwerk-Koeffizienten und einer geeigneten Transformation. Die Radon-Transformation erwies sich in dieser Hinsicht als nicht ausreichend.

In diesem Kapitel werden wir zunächst einige Details aus der Wavelet-Theorie und deren Relevanz für unser Approximationsproblem darstellen. Die Wavelet-Transformation kann in Analogie zur Fourier- Transformation gesehen werden, denn in beiden Fällen werden eindimensionale zeitliche Signale $f(x)$ in ihr Frequenzspektrum zerlegt. Diese Zerlegung geschieht *ohne Informationsverlust*, d.h. das Ausgangssignal lässt sich aus seinem Frequenzspektrum wieder zurückgewinnen. Die Fourier-Transformation liefert die Frequenzanalyse ohne Information über den genauen Zeitpunkt, zu dem die einzelnen Frequenzen auftreten. Dennoch enthält die Fourier-Transformierte die volle Information, denn mit Hilfe der inversen Fourier-Transformation kann das Signal gemäß dem Umkehrsatz rekonstruiert werden. Die Wavelet-Transformation liefert die Information besser voneinander abgegrenzt: Man erhält sowohl die Frequenzanalyse als auch die Zeitpunkte des Auftretens der einzelnen Frequenzen. Der Grund hierfür ist, dass im Gegensatz zur Fourier-Transformation die Wavelet-Transformation auch lokale Unterschiede im Verhalten von f reflektiert: Die Fourier-Transformation basiert auf einer globalen Basisfunktion mit unendlichem Träger, Wavelets besitzen aber in der Regel kompakten Träger.

Definition 3.1 (kontinuierliche Wavelet-Transformation). *Sei $\psi : \mathbb{R} \to \mathbb{R}$ ein Wavelet im Morlet-Grossmann-Sinne (Definition 2.1). Dann heißt*

$$\begin{aligned} \mathcal{W}_\psi f : \mathbb{R}\backslash\{0\} \times \mathbb{R} &\longrightarrow \mathbb{C} \\ (a,t) &\longmapsto \sqrt{|a|} \int_{\mathbb{R}} f(x)\, \psi^*(ax-t)\ dx \end{aligned}$$

die Wavelet-Transformierte zu $f \in L^2(\mathbb{R})$. ψ^* bezeichne das komplex-konjugierte Wavelet zu ψ.

Die kontinuierliche Wavelet-Transformation kann also als eine Art Zeit-Frequenz- Lokalisation verstanden werden, wobei die Wavelet-Transformation im Gegensatz zur Fourier-Transformation eine Funktion von *zwei* Veränderlichen ist. Aus diesem Grund erwarten wir in der Rekonstruktion mit Wavelets größere Flexibilität im Vergleich zu Sinus- und Kosinus-Funktionen. Bei der Analyse von Funktionen $f \in L^2(\mathbb{R})$ bzw. durch Wavelets liefert Parameter a mit $|a| \ll 1$ breite "Fenster" zur Untersuchung langwelliger Schwingungsanteile des Signals, und Skalenwerte mit $|a| \gg 1$ schmale Fenster zur Untersuchung kurzwelliger bzw. hochfrequenter Schwingungsanteile.

Einschub: Eigenschaften der Wavelet-Transformation
Bevor wir fortfahren mit der Darstellung der Approximationsfähigkeit von Wavelet-Neuronalen Netzen ist es sinnvoll die Wavelet-Transformation etwas detaillierter zu studieren. Definieren wir die Familie

$$\Psi := \left\{ \psi_{a,t} \ : \ \psi_{a,t}(x) = \sqrt{a}\, \psi\left(ax - t\right) \ , \ (a,t) \in \mathbb{R}\backslash\{0\} \times \mathbb{R} \right\} \ , \tag{3.1}$$

so können wir die Wavelet-Transformierte auch als Skalarprodukt schreiben:

$$\mathcal{W}_\psi f(a,t) = \langle f, \psi_{a,t} \rangle \ ,$$

wobei das Skalarprodukt natürlich durch die entsprechende Norm induziert wird, in unserem Falle $\langle \cdot, \cdot \rangle = \| \cdot \|_2^2$. Mit der Schwarz'schen Ungleichung folgt somit auch sofort die Beschränktheit der Wavelet-Transformierten: $|\mathcal{W}_\psi f(a,t)| \leq \|f\|_2$ für alle $(a,t) \in \mathbb{R}\backslash\{0\} \times \mathbb{R}$.

Anmerkung 3.1. Die Mitglieder $\psi_{a,t}$ der durch Translation und Dilation aus dem Mutter-Wavelet ψ erzeugten Familie (3.1) lassen sich in etwas formalerem Zusammenhang beschreiben. $\mathcal{A}$ sei die Gruppe

$$\mathcal{A} = \big(\{(a,t) : (a,t) \in \mathbb{R}\backslash\{0\} \times \mathbb{R}\}, \circ\big) \ ,$$

wobei $(a_1,t_1) \circ (a_2,t_2) = (a_1 a_2, t_1 + t_2)$ ist $((a_1,t_1),(a_2,t_2) \in \mathcal{A})$. $\mathcal{A}$ ist lokalkompakt und topologisch. Wir betrachten nun die quadrat-integrierbare unitäre Darstellung dieser Gruppe[1]

$$\pi : \mathcal{A} \longrightarrow \mathcal{U}(L^2(\mathbb{R}))$$

mit

$$f_{a,t}(x) := (\pi(a,t)f)(x) = \sqrt{|a|}\, f\,(ax - t) \ .$$

Wir schreiben also im Speziellen:

$$\psi_{a,t}(x) = (\pi(a,t)\psi)(x) = \sqrt{|a|}\, \psi\,(ax - t) \ .$$

Die Wavelet-Transformation lässt sich dann umschreiben in

$$\mathcal{W}_\psi f(a,t) = \langle f, \pi(a,t)\psi \rangle \ .$$

Wir sind nun an einer Umkehrformel für die Wavelet-Transformation interessiert. Wir erwarten also Integrale der Form $\int \mathcal{W}_\psi f(a,t) \mathrm{d}a\mathrm{d}t$. Um über die Gruppe $\mathcal{A}$ zu integrieren, müssen wir übergehen zu dem (linksinvarianten) Haar-Maß auf $\mathcal{A}$, also $\mathrm{d}\mu = a^2 \mathrm{d}a\mathrm{d}t$. Der Hilbertraum $H = L^2(\mathbb{R}\backslash\{0\} \times \mathbb{R}, \mathrm{d}\mu)$ ist dann ausgestattet mit dem Skalarprodukt

$$\langle u, v \rangle_H := \int_{\mathcal{A}} u(a,t)\, v^*(a,t)\, \mathrm{d}\mu(a,t) \ .$$

Lemma 3.1. *Es gilt folgende Plancherel-Formel für die Wavelet-Transformation:*

$$\langle \mathcal{W}_\psi f, \mathcal{W}_\psi g \rangle_H = C_\psi \langle f, g \rangle_{L^2}$$

für $f, g \in L^2$.

Beweis. Der Beweis kann in Blatter (1998) nachgelesen werden.

Es gilt nun folgende Rekonstruktionsformel:

[1] $\mathcal{U}(L^2(\mathbb{R}))$ bezeichne die Gruppe aller unitären komplex linearen Abbildungen über $L^2(\mathbb{R})$.

Satz 3.1 (Kontinuierliche Rekonstruktionsformel). *Sei $f \in L^1$ stetig an der Stelle x. Dann gilt:*

$$\begin{aligned} f(x) &= \frac{1}{C_\psi} \int_{\mathbb{R}\setminus\{0\}\times\mathbb{R}} \mathcal{W}_\psi f(a,t)\, \psi_{a,t}(x)\, a^2 \mathrm{d}a\mathrm{d}t \\ &= \frac{1}{C_\psi} \int_{\mathcal{A}} \langle f, \pi(a,t)\psi\rangle_H\, (\pi(a,t)\psi)(x)\, \mathrm{d}\mu(a,t)\,. \end{aligned}$$

Beweis. Zum Beweis dieses Satzes reicht es zunächst zu bemerken, dass

$$\lim_{\sigma\downarrow 0} (f \star g_\sigma)(x) = f(x)\,,$$

wobei

$$g_\sigma(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)\,.$$

Wir schreiben die Plancherel-Formel aus Lemma 3.1 um zu

$$\langle f, g\rangle_{L^2} = \frac{1}{C_\psi} \int_{\mathbb{R}\setminus\{0\}\times\mathbb{R}} \mathcal{W}_\psi f(a,t)\, \langle \psi_{a,t}(x), g\rangle\, a^2 \mathrm{d}a\mathrm{d}t$$

und benutzen die Identität $(f\star g_\sigma)(x) = \langle f, T_x g_\sigma\rangle$, wobei mit T_x der Translationsoperator $T_x g_\sigma(t) = g_\sigma(x-t)$ gemeint ist. Wir erhalten dann

$$(f \star g_\sigma)(x) = \frac{1}{C_\psi} \int_{\mathbb{R}\setminus\{0\}\times\mathbb{R}} \mathcal{W}_\psi f(a,t)(\psi_{a,t} \star g_\sigma)(x)\, a^2 \mathrm{d}a\mathrm{d}t\,.$$

Mit $\sigma \downarrow 0$ ergibt sich dann die Behauptung. ■

Anmerkung 3.2. An dieser Stelle ist es wichtig darauf hinzuweisen, dass wir $\psi \in L^2$ schon in der Definition eines Wavelets gefordert haben. Von daher können wir nicht davon ausgehen Approximationssätze für L^p mit $p > 2$ zu erhalten wie für sigmoide oder RBFNs. Diese Einschränkung spielt aber in der Praxis meist keine Rolle und die positiven Eigenschaften des Wavelet-Ansatzes überwiegt sie bei Weitem.

Dieser Satz zeigt, dass das Ausgangssignal f als Überlagerung von Waveletfunktionen $\psi_{a,t}$ dargestellt werden kann, wobei die Wavelet-Transformation $\mathcal{W}_\psi f(a,t)$ die Koeffizienten liefert. Diese Zerlegung von f hat allerdings ein Problem: Die $\psi_{a,t}$ ($a \in \mathbb{R}\setminus\{0\}$, $t \in \mathbb{R}$) sind *nicht* linear unabhängig! Dieses Problem werden wir durch die Forderung der Orthonormalität lösen. Es stellt sich aber heraus, dass die Familie $\psi_{a,t}$ keine Basis, sondern lediglich einen so genannten "Frame" von H bilden.

3.1 Diskrete Wavelet-Transformation und Frames

Unser Ziel ist es Funktionen aus dem Hilbertraum L^2 zu approximieren. Die Umkehrformel zeigt wie eine L^2-Funktion dargestellt werden kann als Überlagerung der $\psi_{a,t}$, wobei die Wavelet-Transformation von f als Koeffizienten der Superposition dienen. Wir möchten nun diese stetige Transformation diskretisieren. In einem Hilbertraum findet sich natürlich immer eine Orthonormalbasis, in Bezug auf Wavelets stellt dies aber im Allgemeinen eine zu große Einschränkung dar und unter geeigneten Voraussetzungen enthält bereits eine diskrekte Folge $(a,t) \in \mathbb{R}\backslash\{0\} \times \mathbb{R}$ von Skalierungsparametern und Translationen die vollständige Information des Eingangssignals.
Die Frage ist also, welche Voraussetzungen ein Wavelet ψ, der Skalierungsfaktor-Faktor a und die Translations-Distanz t erfüllen müssen, damit die Wavelet-Transformation einer beliebigen Funktion $f \in L^2$ bereits durch die Folge ihrer *diskreten* Wavelet-Koeffizienten

$$(\mathcal{W}_\psi f(a_m, t_n))_{(m,n)\in\mathbb{Z}^2}$$

festgelegt ist? Wann lässt sich in diesem Falle das Ausgangssignal stetig aus seinen Wavelet- Koeffizienten rekonstruieren?

Definition 3.2 (Frame). *Eine Familie von Vektoren $(v_r)_{r\in R}$, R eine beliebige Menge, aus einem Hilbertraum H heißt Frame, falls es Konstanten $A, B > 0$ gibt, so dass für alle $x \in H$ gilt*

$$A\|x\|_H^2 \leq \sum_{r\in R} |\langle x, v_r\rangle_H|^2 \leq B\|x\|_H^2 \,.$$

Falls $A = B$, so heißt der Frame straff. Falls für jedes $k \in R$ die Familie $(v_r)_{r\in R\backslash\{k\}}$ kein Frame ist, so heißt der Frame $(v_r)_{r\in R}$ exakt.

Anmerkung 3.3. Wir haben für die Menge R bislang nichts voraussgesetzt. In den Anwendungen, die wir betrachten wird in der Regel $R \subseteq \mathbb{Z}^2$ sein. Damit ist der Frame abzählbar. Die Abzählbarkeit ist allerdings nicht unbedingt notwendig und wir werden im folgenden auch gelegentlich $R \subset M$ zulassen (s. Anmerkung 3.7), wobei M eine beliebige messbare Menge sei.

Anmerkung 3.4. Jeder Frame eines Hilbertraums ist ein Erzeugendensystem. Dies folgt aus der linken Seite der Frame-Ungleichung, sodass ($A > 0$)

$$\langle x, v_r\rangle_H = 0 \quad \forall r \in R \quad \Longrightarrow \quad x = 0\,.$$

Der Orthogonalraum besteht also nur aus dem Nullvektor.

Anmerkung 3.5. Sei $(w_j)_{j\in J}$ ein Orthonormalsystem in H. $(w_j)_{j\in J}$ ist genau dann eine Orthonormalbasis von H, wenn die Parseval'sche Gleichung

$$\|x\|_H^2 = \langle x, x\rangle_H = \sum_{j\in J} |\langle x, w_j\rangle_H|^2$$

für alle $x \in H$ gilt. Das heißt insbesondere, dass jede Orthonormalbasis ein straffer, exakter Frame mit $A = B = 1$ ist. Also ist der Begriff des Frames eine Verallgemeinerung einer Basis. Aber folgender Punkt kann nicht oft genug betont werden: Ein Frame, auch ein straffer Frame, ist i.A. *keine* Orthonormalbasis! In Lemma 3.6 formulieren wir die genauen Voraussetzungen, unter denen dies zutrifft. Zunächst aber ein Beispiel aus Daubechies (1992). Es illustriert gut, dass die Frame-Konstanten den Grad der *Redundanz* in der Rekonstruktion von f widerspiegeln und dass nur im Falle $A = B = 1$ der Frame auch eine Orthonormalbasis ist.

Beispiel 3.1. Sei $H = \mathbb{C}^2$ sowie $\boldsymbol{e}_1 = (0, 1)$, $\boldsymbol{e}_2 = (-\frac{\sqrt{3}}{2}, -\frac{1}{2})$ und $\boldsymbol{e}_3 = (\frac{\sqrt{3}}{2}, -\frac{1}{2})$. Für jedes $\boldsymbol{v} = (v_1, v_2) \in H$ gilt nun

$$\sum_{j=1}^{3} |\langle \boldsymbol{v}, \boldsymbol{e}_j\rangle|^2 = |v_2|^2 + \left|-\frac{\sqrt{3}}{2}v_1 - \frac{1}{2}v_2\right|^2 + \left|\frac{\sqrt{3}}{2}v_1 - \frac{1}{2}v_2\right|^2 = \frac{3}{2}\left[|v_1|^2 + |v_2|^2\right] = \frac{3}{2}\|\boldsymbol{v}\|^2 .$$

Die $\boldsymbol{e}_j$ bilden einen straffen Frame, aber definitiv keine orthonormale Basis! An dem Zahlenwert $A = 3/2$ können wir ablesen, dass wir drei Vektoren in zwei Dimensionen haben. Für $A = 1$ gibt es keine Redundanz, die Vektoren sind linear unabhängig, bilden also eine Basis.

Wir nehmen uns an dieser Stelle Zeit für einige Definitionen und/oder Lemmata, die im weiteren Verlauf nützlich sind.

Lemma 3.2. *Es sei die folgende lineare Abbildung auf H definiert*[2]*:*

$$\begin{aligned} T : H &\longrightarrow \ell^2(R) \\ x &\longmapsto (\langle x, v_r\rangle_H)_{r\in R} \end{aligned}$$

Dann gilt:

(i) *T ist stetig und injektiv, sowie $\|T\| \le \sqrt{B}$.*
(ii) *Die Umkehrabbildung $T^{-1} : T(H) \to H$ ist auch stetig mit $\|T^{-1}\| \le 1/\sqrt{A}$.*

Beweis. Mit der Definition der ℓ^p-Norm und der Frame Bedingung sehen wir, dass

$$\|Tx\|_{\ell^2(R)}^2 = \sum_{r\in R} |\langle x, v_r\rangle_H|^2 \le B\|x\|_H^2 < \infty .$$

[2] ℓ^p bezeichnet wie üblich den Folgenraum

$$\ell^p(\mathbb{N}) = \left\{ (a_n) \,:\, \sum_{n=1}^{\infty} |a_n|^p < \infty \right\} .$$

T ist beschränkt durch $\sqrt{B}$.
Es sei $x \neq 0$, dann ist

$$0 < A\|x\|_H^2 \leq \|Tx\|_{\ell^2(R)}^2 \quad \Longrightarrow \quad Tx \neq 0\,,$$

d.h. T ist injektiv. Und die Umkehrabbildung $T^{-1} : T(H) \to H$ erfüllt dann

$$\|T^{-1}y\|_H^2 = \|T^{-1}Tx\|_H^2 = \|x\|_H^2 \leq \frac{1}{A}\|Tx\|_{\ell^2(R)}^2 = \frac{1}{A}\|y\|_{\ell^2(R)}^2 \quad \forall y = Tx \in T(H)\,.$$

Da beide Operatoren beschränkt und linear sind, sind sie auch stetig. ■

Wir definieren nun für alle $x \in H$ und $y \in \ell^2(R)$ den zu T adjungierten Operator $T^* : \ell^2(R) \to H$ (er existiert, da T stetig ist) durch:

$$\langle Tx, y\rangle_{\ell^2(R)} = \langle x, T^*y\rangle_H\,. \tag{3.2}$$

T^* ist ebenfalls stetig.

Lemma 3.3 (Frame-Operator). *Sei $(v_r)_{r\in R}$ eine Familie von Vektoren aus dem Hilbertraum H. Der Operator $\mathscr{T} := T^* \circ T$ heißt Frame-Operator der Familie $(v_r)_{r\in R}$ und es gilt:*

$$\begin{aligned} \mathscr{T} : H &\longrightarrow H \\ x &\longmapsto \sum_{r\in R} \langle x, v_r\rangle_H v_r\,. \end{aligned}$$

Beweis. Dieses Lemma ergibt sich direkt aus der Definition von T und Glg. (3.2). ■

Anmerkung 3.6. Man sieht sofort, dass für den Fall einer orthogonalen Familie $(v_r)_{r\in R}$

$$v_k \mapsto \langle v_k, v_k\rangle_H v_k = \|v_k\|^2 v_k\,,$$

für eine orthonormierte Familie sogar $v_k \mapsto v_k$. In diesem Fall ist dann $\mathscr{T} = \mathbb{1}_H$.

Anmerkung 3.7. Wir können den Operator T auch kontinuierlich definieren, d.h. an die Stelle von $\ell^2(R)$ tritt $L^2(M,\mu)$, wobei M nun als abstrakte Menge und nicht mehr unbedingt abzählbar gewählt wird. μ sei ein Maß auf M und als σ-Algebra verwenden wir die Menge der messbaren Teilmengen von M. Damit ist $T' : H \to L^2(M,\mu)$ mit $x \mapsto (\langle x, v_m\rangle_H)_{m\in M}$. T' ist ebenfalls beschränkt und injektiv. Hiermit verallgemeinert sich der Frame-Begriff auf kontinuierliche Familien (insbesondere Wavelet-Frames). Wir lassen $R \subset M$ zu und formulieren folgendes Lemma:

Lemma 3.4. *Für jedes beliebige Wavelet ψ bildet die (kontinuierliche) Familie Ψ aus Glg. (3.1) einen straffen (kontinuierlichen) Wavelet-Frame mit der Frame-Konstanten C_ψ.*

Beweis. Dies folgt direkt aus der Überlegung, dass die Wavelet-Transformation nichts anderes ist als der Operator T' zugehörig der Familie Ψ. Es gilt nämlich nach Definition für die Familie $(v_r)_{r\in R}$, $R \subset M$,

$$(T'f)_r = \langle f, v_r\rangle_H \,.$$

Wähle nun $(v_r)_{r\in R} = (\psi_{a,t})_{(a,t)\in\mathbb{R}\setminus\{0\}\times\mathbb{R}}$. Dann ergibt sich:

$$(T'f)_{(a,t)} = \langle f, \psi_{a,t}\rangle_H = \mathcal{W}_\psi f(a,t)$$

und mit Hilfe von Lemma 3.1 und $H = L^2(\mathbb{R}\setminus\{0\}\times\mathbb{R}, \mathrm{d}\mu)$

$$\|\mathcal{W}_\psi f\|_H^2 = C_\psi \|f\|_{L^2}^2 \quad \forall f \in L^2\,,$$

wobei

$$C_\psi = 2\pi \int_{\mathbb{R}\setminus\{0\}} \frac{|\overline{\psi}(a)|^2}{|a|}\mathrm{d}a\,,$$

was zu beweisen war. ∎

Unter diesen Umständen ist

$$\mathscr{T}^{-1} = \frac{1}{C_\psi}\mathbb{1}$$

und $\mathscr{T}^{-1}\psi_{a,t} = \frac{1}{C_\psi}\psi_{a,t}$. Der duale Frame zu $(\psi_{a,t})_{(a,t)\in\mathbb{R}\setminus\{0\}\times\mathbb{R}}$ ist also gegeben durch

$$(\mathscr{T}^{-1}\psi_{a,t})_{(a,t)\in\mathbb{R}\setminus\{0\}\times\mathbb{R}} = \left(\frac{1}{C_\psi}\psi_{a,t}\right)_{(a,t)\in\mathbb{R}\setminus\{0\}\times\mathbb{R}}\,.$$

Wir notieren in folgendem Lemma nun einige Eigenschaften des Frame-Operators:

Lemma 3.5. *Es sei $\mathscr{T}$ der Frame-Operator wie oben definiert. Dann gilt*

$$A\mathbb{1} \le \mathscr{T} \le B\mathbb{1} \quad \textit{und}$$
$$\|\mathbb{1} - \mathscr{T}\| \le 1 - A\,,$$

wobei $\mathbb{1}$ die Identität in H bezeichne.

Beweis. Für beliebiges $x \in H$ ist die Folge

$$\left(s_N := \sum_{i=0}^{N} \langle x, v_r\rangle_H v_r\right)_{N\in\mathbb{N}}$$

eine Cauchy-Folge:

$$\|s_N - s_M\|^2 = \sup_{\|y\|=1} |\langle s_N - s_M, y\rangle_H|^2 = \sup_{\|y\|=1} \left| \sum_{i=N+1}^{M} \langle x, v_i\rangle_H \langle v_i, y\rangle_H \right|^2$$
$$\leq \sum_{i=N+1}^{M} |\langle x, v_i\rangle_H|^2 \sup_{\|y\|=1} \sum_{i=N+1}^{M} |\langle v_i, y\rangle_H|^2 \leq B \sum_{i=N+1}^{M} |\langle x, v_i\rangle_H|^2 \ .$$

$\sum_{i=N+1}^{M} |\langle x, v_i\rangle_H|^2$ wird nun beliebig klein, da $\sum_{i\in\mathbb{N}} |\langle x, v_i\rangle_H|^2 \leq B\|x\|^2$ konvergiert. Der Hilbertraum ist vollständig, d.h. $\sum_{i\in\mathbb{N}} \langle x, v_r\rangle_H v_r \in H$.
Sei nun $z \in H$ ein beliebiges Element aus dem Hilbertraum. Dann ist

$$\langle \mathscr{T}, z\rangle_H = \langle Tx, Tz\rangle_H = \langle (\langle x, v_i\rangle_H)_i, (\langle z, v_i\rangle_H)_i \rangle_H = \sum_{i\in\mathbb{N}} \langle x, v_i\rangle_H \overline{\langle z, v_i\rangle_H}$$
$$= \sum_{i\in\mathbb{N}} \langle x, v_i\rangle_H \langle v_i, z\rangle_H \ .$$

Da z beliebig gewählt wurde ergibt sich also direkt

$$\mathscr{T} x = \sum_{i\in\mathbb{N}} \langle x, v_i\rangle_H v_i \ .$$

Weiterhin gilt:

$$\langle \mathscr{T} x, x\rangle_H = \sum_{i\in\mathbb{N}} \langle x, v_i\rangle_H \langle v_i, x\rangle_H = \sum_{i\in\mathbb{N}} |\langle x, v_i\rangle_H|^2 \ .$$

Wir wenden nun die Frame-Definition an und erhalten:

$$A\|x\|_H^2 \leq \langle \mathscr{T} x, x\rangle_H \leq B\|x\|_H^2 \ ,$$

und somit

$$(1-B)\|x\|_H^2 \leq \langle (\mathbb{1} - \mathscr{T})x, x\rangle_H \leq (1-A)\|x\|_H^2 \quad \Longrightarrow \quad \|\mathbb{1} - \mathscr{T}\| \leq 1 - A \ .$$

Im letzten Schritt haben wir benutzt, dass $\|f\| = \sup_{\|x\|=1, \|y\|=1} |\langle f(x), y\rangle_H| = 1$ für eine stetige Abbildung $f : X \to Y$ zwischen zwei Hilberträumen. ■

Anmerkung 3.8. In Abschnitt 3.4 werden wir an dieser Stelle ansetzen und den Operator $\mathscr{T}$ in Hinblick auf seine Eigenwerte und Robustheit gegenüber Störungen genauer unter die Lupe nehmen.

Nun sind wir in der Lage den versprochenen Zusammenhang zwischen straffen Frames und Orthonormalbasen in einem Hilbertraum zu formulieren:

Lemma 3.6. *Sei $(v_r)_{r\in R}$ ein straffer Frame des Hilbertraums H mit $A = B = 1$. Zudem haben alle Mitglieder der Familie Norm 1. Dann bildet $(v_r)_{r\in R}$ eine Orthonormalbasis des Hilbertraums H. Es ist dann für alle $f \in H$*

$$f = \sum_{r\in R} \langle f, v_r\rangle_H v_r \ .$$

Beweis. Die Mitglieder der Familie haben nach Konstruktion alle Norm 1. Wir müssen nun also noch ihre Orthogonalität nachweisen. Dies ist allerdings trivial und folgt direkt aus der Straffheit. Wir wenden die Frame-Bedingung für $A = B = 1$ an:

$$\|v_r\|_H^2 = \sum_i |\langle v_r, v_i\rangle_H|^2 = \underbrace{|\langle v_r, v_r\rangle_{L^2}|^2}_{=\|v_r\|_H^4} + \sum_{\substack{i \\ i\neq r}} |\langle v_r, v_i\rangle_H|^2 \ .$$

Also, da $\|v_r\|_H = 1$,

$$\sum_{\substack{i \\ i\neq r}} |\langle v_r, v_i\rangle_H|^2 = 0 \quad \Longrightarrow \quad \langle v_r, v_i\rangle_H = 0 \quad \forall i \neq r \ .$$

Wir haben also ein Orthonormalsystem vorliegen. Nach Voraussetzung ist der Frame straff, d.h. Anmerkung 3.5 folgend bildet $(v_r)_{r\in R}$ eine Orthonormalbasis von H. Die Parseval'sche Gleichung für alle $f \in H$

$$\|f\|_H^2 = \sum_{r\in R} |\langle f, v_r\rangle_H|^2$$

liefert dann sofort die gewünschte Aussage für $f \in H$. ■

Im Speziellen bildet also ein straffer Frame von 1-normierten Funktionen aus L^2 eine Hilbertbasis von L^2. Bevor wir konkreter auf Wavelet-Familien eingehen, verallgemeinern wir die soeben gemachten Aussagen auf Fälle, in denen der Frame nicht straff ist:

Satz 3.2. *Sei $(v_r)_{r\in R}$ ein Frame in einem Hilbertraum H mit den Frame-Konstanten A und B. Dann gilt für den Frame-Operator $\mathscr{T}$:*

*(*i*) $\mathscr{T}$ ist invertierbar und $\frac{1}{B}\mathbb{1} \leq \mathscr{T}^{-1} \leq \frac{1}{A}\mathbb{1}$*

*(*ii*) Jedes $f \in H$ kann in der Form*

$$f = \sum_{r\in R} \langle f, v_r\rangle_H \mathscr{T}^{-1} v_r = \sum_{r\in R} \langle f, \mathscr{T}^{-1} v_r\rangle_H v_r \tag{3.3}$$

geschrieben werden.

Beweis. $\mathscr{T} = T^* \circ T$, d.h. $\mathscr{T}$ ist invertierbar, da nach Lemma 3.2 T invertierbar ist und damit auch T^* (s. Werner (2006)). Nach Lemma 3.5 ist dann $\frac{1}{B}\mathbb{1} \leq \mathscr{T}^{-1} \leq \frac{1}{A}\mathbb{1}$.
Weiterhin gilt für beliebiges $f \in H$

$$f = \mathscr{T}^{-1}\mathscr{T} f = \mathscr{T}^{-1}\left(\sum_{r\in R} \langle f, v_r\rangle_H v_r\right) = \sum_{r\in R} \langle f, v_r\rangle_H \mathscr{T}^{-1} v_r \ ,$$

da $\mathscr{T}^{-1}$ linear ist. ■

Anmerkung 3.9. Wir haben schon zuvor gezeigt, dass für einen *straffen* Frame jede L^2-Funktion *exakt* durch die Überlagerung der Mitglieder einer diskreten und auf 1 normierten Familie von L^2-Funktionen ausgedrückt werden kann. Es ist offensichtlich, dass in diesem Fall $\mathscr{T} = \mathbb{1}$ gilt und Satz 3.2 and Lemma 3.6 anschließt. Ist der Frame zwar straff, aber $A = B \neq 1$, so wird Glg. (3.3) offensichtlich zu

$$f = \frac{1}{A} \sum_{r \in R} \langle f, v_r \rangle_H v_r \, .$$

Anmerkung 3.10. $(\mathscr{T}^{-1} v_r)_{r \in R}$ bildet den so genannten *dualen* Frame zu $(v_r)_{r \in R}$ mit den Frame-Konstanten B^{-1} und A^{-1}. Ist also ein Hilbertraum H gegeben und ein Frame, so können wir jeden Vektor aus H darstellen durch Überlagerung der Elemente des dualen Frames, wobei als Koeffizienten die Skalarprodukte von f und den Frame-Vektoren dienen.

Der folgende Satz zeigt, dass die Wahl der Koeffizienten $\langle f, \mathscr{T}^{-1} v_r \rangle_H$ in gewisser Hinsicht "ökonomisch" ist:

Lemma 3.7. *Es sei $f = \sum_{r \in R} c_r \, v_r$ für eine Folge $(c_r)_{r \in R} \in \ell^2(R)$. Falls nicht alle c_r gleich $\langle f, \mathscr{T}^{-1} v_r \rangle_H$ sind, dann ist*

$$\sum_{r \in R} |c_j|^2 > \sum_{r \in R} |\langle f, \mathscr{T}^{-1} v_r \rangle_H|^2 \, .$$

Beweis. Der Beweis findet sich in Daubechies (1992).

Bevor wir die allgemeinen Frame-Betrachtungen verlassen kommen wir noch einmal auf die Approximation einer Funktion durch die Elemente des dualen Frames zu sprechen. In Beispiel 3.1 kamen wir kurz auf den Begriff der Redundanz zu sprechen: Je mehr A von 1 abweicht, desto redundanter wird der Frame. Es sei nun $\frac{B}{A} - 1 \ll 1$, d.h. A und B sind fast gleich. Die Abschätzungen $A\mathbb{1} \leq \mathscr{T} \leq B\mathbb{1}$ und $\frac{1}{B}\mathbb{1} \leq \mathscr{T}^{-1} \leq \frac{1}{A}\mathbb{1}$ zeigen dann, dass

$$\mathscr{T} \approx \frac{A+B}{2} \, \mathbb{1} \quad \text{und}$$
$$\mathscr{T}^{-1} \approx \frac{2}{A+B} \, \mathbb{1} \, .$$

Wir können also mit Satz 3.2 schreiben:

$$f = \frac{2}{A+B} \sum_{r \in R} \langle f, v_r \rangle_H v_r + Ef \, ,$$

wobei für den Fehleroperator folgendes gilt:

$$E = \mathbb{1} - \frac{2}{A+B}\mathscr{T} \quad \Longrightarrow \quad -\frac{B-A}{B+A}\,\mathbb{1} \leq E \leq \frac{B-A}{B+A}\,\mathbb{1}\,. \tag{3.4}$$

Aus dieser Abschätzung erhalten wir sofort

$$\|E\|_H \leq \frac{B-A}{B+A} = \frac{\frac{B}{A}-1}{\frac{B}{A}+1}\,.$$

Die Rekonstruktions-Formel ist also exakt bis auf den Fehlerterm

$$\|E\|_H \|f\|_H \leq \frac{\frac{B}{A}-1}{\frac{B}{A}+1}\|f\|_H\,.$$

Nun kommen wir aber auch zu Fällen, in denen B/A nicht nahe an 1 liegt, der Frame also ganz und gar nicht straff ist. Daubechies (1992) folgend notieren wir, dass

$$\mathscr{T} = \frac{A+B}{2}(\mathbb{1} - E)$$

und

$$\mathscr{T}^{-1}v_r = \frac{2}{A+B}\sum_{k=0}^{\infty} E^k v_r\,.$$

Die Reihe $\sum_{k=0}^{\infty} E^k$ konvergiert, da $\|E\|$ beschränkt ist. Nun schneiden wir die Reihe ab:

$$(\mathscr{T}^{-1}v_r)^N = \frac{2}{A+B}\sum_{k=0}^{N} E^k v_r = \mathscr{T}^{-1}v_r - \frac{2}{A+B}\sum_{k=N+1}^{\infty} E^k v_r = (\mathbb{1} - E^{N+1})\mathscr{T}^{-1}v_r\,.$$

Der Fehler bei dieser Approximation berechnet sich zu

$$\begin{aligned}\left\| f - \sum_{r\in R}\langle f, v_r\rangle_H (\mathscr{T}^{-1}v_r)^N \right\| &= \sup_{\|g\|_H=1} |\langle f, E^{N+1}g\rangle_H| \leq \|E^{N+1}\|_H \|f\|_H \\ &\leq \left(\frac{\frac{B}{A}-1}{\frac{B}{A}+1}\right)^{N+1} \|f\|_H\,.\end{aligned}$$

Weil $\frac{B/A-1}{B/A+1} < 1$, geht der Fehler mit wachsendem N exponentiell gegen 0. Weiterhin können die Elemente des dualen Frames durch einen iterativen Algorithmus berechnet werden:

$$(\mathscr{T}^{-1}v_r)^N = \sum_{l\in R} \alpha_{rl}^N v_l\,, \tag{3.5}$$

wobei

$$\alpha_{rl}^N = \frac{2}{A+B}\delta_{rl} + \alpha_{rl}^{N-1} - \frac{2}{A+B}\sum_{m\in R}\alpha_{rm}^{N-1}\langle v_m, v_l\rangle_H \,.$$

Ebenso kann f iterativ berechnet werden: $f = \lim_{N\to\infty} F_N$ mit

$$f_N = F_{N-1} + \frac{2}{A+B}\sum_{r\in R}(\langle f, v_r\rangle_H - \langle f_{N-1}, v_r\rangle_H)\, v_r \,. \tag{3.6}$$

Nach diesem Ausflug in die allgemeine Frame-Theorie betrachten wir nun als Spezialfall einen Frame aus Wavelets und fassen unsere soeben gefunden Ergebnisse (Satz 3.2, Lemma 3.6 und die vorangegangenen Überlegungen) in dem folgenden Rekonstruktionssatz zusammen. Er stellt die diskrete Version der Umkehrformel (Satz 3.1) aus der kontinuierlichen Wavelet-Theorie dar:

Satz 3.3 (Diskrete Rekonstruktionsformel). *Sei $\dot\Psi$ eine diskrete Version der Familie Ψ definiert durch*

$$\dot\Psi := \left\{\psi_{a_m,t_n} \,:\, \psi_{a_m,t_n}(x) = a_m^{-\frac{1}{2}}\,\psi\left(a_m^{-1}x - t_n\right)\,,\ (a_m,t_n)\in\mathbb{R}\backslash\{0\}\times\mathbb{R}\,,\ (m,n)\in\mathbb{Z}^2\right\},$$

also ein diskreter Wavelet-Frame festgelegt durch die Folge $(a_m,t_n)_{(m,n)\in\mathbb{Z}^2}$ von Dilations- und Translationsparametern. $\mathscr{T} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ bezeichne den Frame-Operator und $f \in L^2(\mathbb{R})$ eine Funktion. Weiterhin bezeichne

$$\mathcal{W}_\psi f(a_m,t_n) = \langle f, \psi_{a_m,t_n}\rangle_{L^2}$$

die diskrete Wavelet-Transformierte von f zu den Parametern a_m und t_n.

*(*i*) Ist der Wavelet-Frame eine Hilbert-Basis von L^2, so gilt:*

$$f = \sum_{(m,n)\in\mathbb{Z}^2}\mathcal{W}_\psi f(a_m,t_n)\,\psi_{a_m,t_n}\,.$$

*(*ii*) Ist der Wavelet-Frame straff mit Frame-Konstanten $A = B$, so ist*

$$f = \frac{1}{A}\sum_{(m,n)\in\mathbb{Z}^2}\mathcal{W}_\psi f(a_m,t_n)\,\psi_{a_m,t_n}\,.$$

*(*iii*) Ist der Frame nicht straff mit Frame-Konstanten $0 < A \le B$ lässt sich die Funktion folgendermaßen rekonstruieren:*

$$f = \sum_{(m,n)\in\mathbb{Z}^2}\mathcal{W}_\psi f(a_m,t_n)\,\mathscr{T}^{-1}\psi_{a_m,t_n}\,.$$

Die Elemente des dualen Frames $\mathscr{T}^{-1}\psi_{m,n}$ können folgendermaßen approximiert werden:

$$\mathscr{T}^{-1}\psi_{m,n} := \lim_{N\to\infty} \left(\mathscr{T}^{-1}\psi_{m,n}\right)^N = \frac{2}{A+B} \lim_{N\to\infty} \sum_{k=0}^{N} \left(\mathbb{1} - \frac{2}{A+B}\mathscr{T}\right)^k$$

und lassen sich mit Hilfe von Algorithmus (3.5) iterativ bestimmen. Die Konvergenzgeschwindigkeit der geometrischen Reihe hängt von dem Verhältnis $\frac{B-A}{B+A} < 1$ *ab:*

$$\left\| f - \sum_{(m,n)\in\mathbb{Z}^2} \mathcal{W}_\psi f(a_m, t_n) \left(\mathscr{T}^{-1}\psi_{m,n}\right)^N \right\| \leq \left(\frac{B-A}{B+A}\right)^{N+1} \|f\|_H \,.$$

Die Idee der Wavelet Neuronalen-Netze ist es nun, wie für die ursprünglichen Neuronalen Netze auch, die Parameter der zu Grunde liegenden Familie $\dot{\Psi}$ so zu optimieren, dass ein gegebenes Signal möglichst gut approximiert wird. Doch hier tut sich sofort ein massives Problem auf: $\dot{\Psi}$ hängt von zwar abzählbar, aber unendlich vielen Parametern ab! Um sich eine reelle Chance auf die numerische Umsetzung unserer Approximationsziele zu wahren, müssen wir die Familie $\dot{\Psi}$ also weiter konkretisieren und sie "griffiger" machen. Folgerichtig gehen wir über zu einer Familie mit möglichst wenigen Parametern, die aber noch flexibel genug ist um möglichst viele Signale aus L^2 rekonstruieren zu können. Eine, und die von uns bevorzugte, Möglichkeit ist es, $\dot{\Psi}$ durch einen festen "Zoom"-Parameter a_0 und einen Translationsparameter t_0 einzuschränken und $a_m = a_0^m$ sowie $t_n = nt_0$ zu wählen. Es entsteht dabei die folgende Familie:

$$\dot{\Psi}(a_0, t_0) := \left\{ \psi_{m,n} \,:\, \psi_{m,n}(x) = a_m^{-1/2}\, \psi\left(a_m^{-1}x - t_n\right) \,,\, a_m = a_0^m \,,\, t_n = nt_0 \,, \right. \qquad (3.7)$$
$$\left. (a_0, t_0) \in \mathbb{R}\backslash\{0\} \times \mathbb{R} \,,\, (m,n) \in \mathbb{Z}^2 \right\} .$$

Die entscheidende Frage ist nun: Unter welchen Voraussetzungen für ψ, a_0 und t_0 ist schon die *diskrete* Familie $\dot{\Psi}(a_0, t_0)$ ein Frame? In diesem Fall können wir mit Algorithmus (3.6) f iterativ rekonstruieren. Gelingt es uns sogar Bedingungen zu finden, so dass $\dot{\Psi}(a_0, t_0)$ ein straffer Frame ist, so können wir f *eindeutig* durch diese Familie darstellen.
Wir zeigen nun den folgenden Satz aus Daubechies (1992), der angibt unter welchen Voraussetzungen ein Wavelet ψ bei einem gegebenen Zoom-Parameter a_0 für verschiedene Translations-Parameter t einen diskreten Wavelet Frame erzeugt:

Satz 3.4. *Es sei* $a_0 > 1$ *vorgegeben sowie* $H = L^2(\mathbb{R})$.

*(*i*) Wir fordern:*

$$\kappa_\downarrow(\psi, a_0) := \inf_{\omega\neq 0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 > 0 \qquad \textit{und}$$
$$\kappa_\uparrow(\psi, a_0) := \sup_{\omega\neq 0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 < \infty \,.$$

(ii) *Weiterhin sei vorausgesetzt, dass es Kontanten $K > 0$ und $\alpha > 0$ gibt, so dass*

$$\beta(x) := \sup_{\omega} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}(a_0^m\omega + x)\right|$$

für $x \in \mathbb{R}$ die folgende Abschätzung erfüllt:

$$\beta(|x|) \leq \frac{K}{|x|^{1+\alpha}} \,.$$

Dann gibt es ein $t_{max} > 0$, so dass für alle $0 < t_0 < t_{max}$ die Familie $\dot{\Psi}(a_0, t_0)$ ein Frame ist mit den Frame Konstanten

$$A = \frac{2\pi}{t_0}\left(\kappa_\downarrow(\psi, a_0) - \sum_{k\in\mathbb{Z}\setminus\{0\}} \left[\beta\left(-\frac{2\pi}{t_0}k\right)\beta\left(\frac{2\pi}{t_0}k\right)\right]^{1/2}\right) \qquad \textit{und}$$

$$B = \frac{2\pi}{t_0}\left(\kappa_\uparrow(\psi, a_0) + \sum_{k\in\mathbb{Z}\setminus\{0\}} \left[\beta\left(-\frac{2\pi}{t_0}k\right)\beta\left(\frac{2\pi}{t_0}k\right)\right]^{1/2}\right) \,.$$

Beweis. Wir starten mit der Berechnung von $|\langle f, \psi_{m,n}\rangle|^2$ (wir lassen am Skalarprodukt den Index H bzw. L^2 in diesem Beweis weg, da immer L^2 gemeint ist). Die Fourier-Transformation ist isometrisch in L^2 (siehe z.B. Amann & Escher (2001) Theorem 9.23 oder Stein & Shakarchi (2003)), d.h. ($\overline{(\cdot)}$ bezeichne wieder die Fourier-Transformierte von $(\cdot)$ und * bezeichne die komplexe Konjugation)

$$\begin{aligned}|\langle f, \psi_{m,n}\rangle|^2 &= \left|\langle \overline{f}, \overline{\psi_{m,n}}\rangle\right|^2 = \langle \overline{f}, \overline{\psi_{m,n}}\rangle\langle\overline{\psi_{m,n}}, \overline{f}\rangle \\ &= \int_{\mathbb{R}} \overline{f}(\eta)\left(\overline{\psi_{m,n}}(\eta)\right)^* \,\mathrm{d}\eta \int_{\mathbb{R}} \overline{\psi_{m,n}}(\zeta)\left(\overline{f}(\zeta)\right)^* \,\mathrm{d}\zeta \,.\end{aligned}$$

Wir nutzen nun aus, dass allgemein für $\alpha > 0$ und $h \in \mathbb{R}$ gilt: $\overline{f(\alpha x)}(\omega) = \frac{1}{|a|}\overline{f}(\omega/\alpha)$ und $\overline{f(x+h)}(\omega) = e^{ih\omega}\overline{f}(\omega)$. Angewandt auf $\psi_{m,n} = a_0^{-m/2}\psi(a_0^{-m}x - t_n)$:

$$\begin{aligned}\overline{\psi_{m,n}}(\omega) &= a_0^{m/2}\exp\left[-ia_0^m t_0 n\omega\right]\overline{\psi}\left(a_0^m\omega\right) \quad \text{und} \\ \left(\overline{\psi_{m,n}}(\omega)\right)^* &= a_0^{m/2}\exp\left[ia_0^m t_0 n\omega\right]\left(\overline{\psi}\left(a_0^m\omega\right)\right)^* \,,\end{aligned}$$

also erhalten wir zunächst

$$|\langle f, \psi_{m,n}\rangle|^2 = a_0^m \int_{\mathbb{R}} \overline{f}(\eta)\left(\overline{\psi}\left(a_0^m\eta\right)\right)^* \Bigg[\int_{\mathbb{R}} \overline{\psi}\left(a_0^m\zeta\right)\cdot \left(\overline{f}(\zeta)\right)^* \exp\left[ia_0^m t_0 n(\eta-\zeta)\right] \,\mathrm{d}\zeta\Bigg] \,\mathrm{d}\eta$$

und durch Summation:

$$\sum_{n\in\mathbb{Z}} |\langle f, \psi_{m,n}\rangle|^2 = a_0^m \int_{\mathbb{R}} \overline{f}(\eta) \left(\overline{\psi}\left(a_0^m \eta\right)\right)^* \left[\int_{\mathbb{R}} \overline{\psi}\left(a_0^m(\eta-\zeta')\right) \cdot \left(\overline{f}(\eta-\zeta')\right)^* \left(\sum_{n\in\mathbb{Z}} \exp\left[i a_0^m t_0 n \zeta'\right]\right) \mathrm{d}\zeta'\right] \mathrm{d}\eta\,,$$

wobei zusätzlich noch die Variablentransformation $\eta - \zeta \to \zeta'$ vorgenommen wurde. Die Integrale sind symmetrisch, deswegen haben wir das Minuszeichen, welches bei der Transformation $\mathrm{d}\zeta \to \mathrm{d}\zeta'$ entstanden ist, weggelassen. Wir benötigen nun die Poisson'sche Summenformel (Higgins (1985) liefert eine gute Diskussion der Formel)

$$T \sum_{n\in\mathbb{Z}} f(nT) = \sqrt{2\pi} \sum_{k\in\mathbb{Z}} \overline{f}\left(\frac{2\pi k}{T}\right)\,,$$

wobei $T > 0$ und schreiben:

$$\begin{aligned}
\sum_{n\in\mathbb{Z}} \exp[i \underbrace{a_0^m t_0}_{:=T} n\zeta'] &= \frac{\sqrt{2\pi}}{T} \sum_{k\in\mathbb{Z}} \overline{\exp\left[i\zeta' n T\right]}\left(\frac{2\pi}{T}\right) \\
&= \frac{\sqrt{2\pi}}{T} \sum_{k\in\mathbb{Z}} \left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp[i\zeta' t]\, \exp\left[-i\frac{2\pi k}{T} t\right]\, \mathrm{d}t\right) \\
&= \frac{\sqrt{2\pi}}{T} \sum_{k\in\mathbb{Z}} \left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left[i\left(\zeta' - \frac{2\pi k}{T}\right) t\right]\, \mathrm{d}t\right) \\
&= \frac{2\pi}{T} \sum_{k\in\mathbb{Z}} \delta\left(\zeta' - \frac{2\pi k}{T}\right)\,.
\end{aligned}$$

Hierbei bezeichnet $\delta(x)$ die Dirac-Distribution, für die allgemein gilt:

$$\delta(x-a) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{it(x-a)}\, \mathrm{d}t\,.$$

In die Summe eingesetzt:

$$\sum_{n\in\mathbb{Z}} |\langle f, \psi_{m,n}\rangle|^2 = a_0^m \int_{\mathbb{R}} \overline{f}(\eta) \left(\overline{\psi}\left(a_0^m \eta\right)\right)^* \left[\int_{\mathbb{R}} \overline{\psi}\left(a_0^m(\eta-\zeta')\right) \cdot \left(\overline{f}(\eta-\zeta')\right)^* \frac{2\pi}{T} \sum_{k\in\mathbb{Z}} \left[\delta\left(\zeta' - \frac{2\pi k}{T}\right)\right] \mathrm{d}\zeta'\right] \mathrm{d}\eta$$

und zusammen mit der Faltungs-Eigenschaft der δ-Distribution

$$\int_{\mathbb{R}} f(x)\delta(x-a)\,\mathrm{d}x = f(a)$$

ergibt sich

$$\sum_{n\in\mathbb{Z}} |\langle f,\psi_{m,n}\rangle|^2 = \frac{2\pi}{t_0}\int_{\mathbb{R}} \overline{f}(\eta)\left(\overline{\psi}\left(a_0^m\eta\right)\right)^* \left[\sum_{k\in\mathbb{Z}} \overline{\psi}\left(a_0^m\eta - \frac{2\pi k}{t_0}\right)\cdot \left(\overline{f}\left(\eta - \frac{2\pi k}{a_0^m t_0}\right)\right)^*\right]\mathrm{d}\eta\,.$$

Nun noch inklusive der zweiten Summe:

$$\sum_{(m,n)\in\mathbb{Z}^2} |\langle f,\psi_{m,n}\rangle|^2 = \frac{2\pi}{t_0}\sum_{m\in\mathbb{Z}}\sum_{k\in\mathbb{Z}}\int_{\mathbb{R}} \overline{f}(\eta)\left(\overline{\psi}\left(a_0^m\eta\right)\right)^* \overline{\psi}\left(a_0^m\eta - \frac{2\pi k}{t_0}\right)\cdot \left(\overline{f}\left(\eta - \frac{2\pi k}{a_0^m t_0}\right)\right)^*\mathrm{d}\eta\,.$$

Wir extrahieren den Term für $k = 0$

$$\sum_{(m,n)\in\mathbb{Z}^2} |\langle f,\psi_{m,n}\rangle|^2 = \frac{2\pi}{t_0}\sum_{m\in\mathbb{Z}}\int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m\eta\right)\right|^2 \mathrm{d}\eta + \mathscr{E}$$

und schätzen ihn folgendermaßen ab[3]:

$$\begin{aligned}\sum_{m\in\mathbb{Z}}\int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m\eta\right)\right|^2 \mathrm{d}\eta &\le \int_{\mathbb{R}} \sup_{\eta}\left(\sum_{m\in\mathbb{Z}} \left|\overline{\psi}\left(a_0^m\eta\right)\right|^2\right)\left|\overline{f}(\eta)\right|^2 \mathrm{d}\eta \\ &= \kappa_\uparrow(\psi,a_0)\,\|f\|^2\,.\end{aligned}$$

Und ebenso für das Infimum, also

$$\kappa_\downarrow(\psi,a_0)\,\|f\|^2 \le \sum_{m\in\mathbb{Z}}\int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m\eta\right)\right|^2 \mathrm{d}\eta \le \kappa_\uparrow(\psi,a_0)\,\|f\|^2\,. \tag{3.8}$$

Nun müssen wir noch den Term $\mathscr{E}$ abschätzen. Dies geschieht mit der Cauchy-Schwarz-Ungleichung, angewendet auf das Integral:

$$|\mathscr{E}| \le \frac{2\pi}{t_0}\sum_{\substack{(m,k)\in\mathbb{Z}^2\\ k\neq 0}}\int_{\mathbb{R}} \left|\overline{f}(\eta)\left(\overline{\psi}\left(a_0^m\eta\right)\right)^* \overline{\psi}\left(a_0^m\eta - \frac{2\pi k}{t_0}\right)\cdot \left(\overline{f}\left(\eta - \frac{2\pi k}{a_0^m t_0}\right)\right)^*\right|\mathrm{d}\eta\,,$$

[3] $\eta = 0$ im Supremum stellt eine Nullmenge dar und spielt somit für das Integral keine Rolle. Wir müssen es aber korrekterweise ausschließen.

so dass

$$|\mathscr{E}| \leq \frac{2\pi}{t_0} \sum_{\substack{(m,k)\in\mathbb{Z}^2 \\ k\neq 0}} \left[\int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m \eta\right)\right| \left|\overline{\psi}\left(a_0^m \eta - \frac{2\pi k}{t_0}\right)\right| \,\mathrm{d}\eta \right]^{1/2} \cdot$$
$$\left[\int_{\mathbb{R}} \left|\overline{f}(\zeta')\right|^2 \left|\overline{\psi}\left(a_0^m \zeta'\right)\right| \left|\overline{\psi}\left(a_0^m \zeta' + \frac{2\pi k}{t_0}\right)\right| \,\mathrm{d}\zeta' \right]^{1/2} ,$$

wobei wieder eine Variablen-Substitution, und zwar $\zeta' = \eta - \frac{2\pi k a_0}{t_0}$ vorgenommen wurde. Nun noch einmal Cauchy-Schwarz für die Summe über m:

$$|\mathscr{E}| \leq \frac{2\pi}{t_0} \sum_{\substack{k\in\mathbb{Z} \\ k\neq 0}} \left\{ \left[\sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m \eta\right)\right| \left|\overline{\psi}\left(a_0^m \eta - \frac{2\pi k}{t_0}\right)\right| \,\mathrm{d}\eta \right]^{1/2} \cdot \right.$$
$$\left. \left[\sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \left|\overline{f}(\zeta')\right|^2 \left|\overline{\psi}\left(a_0^m \zeta'\right)\right| \left|\overline{\psi}\left(a_0^m \zeta' + \frac{2\pi k}{t_0}\right)\right| \,\mathrm{d}\zeta' \right]^{1/2} \right\} .$$

Für den ersten Faktor jedes Summanden der Summe über k gilt (Summe und Integral vertauschen wieder):

$$\sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m \eta\right)\right| \left|\overline{\psi}\left(a_0^m \eta - \frac{2\pi k}{t_0}\right)\right| \,\mathrm{d}\eta \leq \int_{\mathbb{R}} \sup_{\eta} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}\left(a_0^m \eta\right) \cdot \right.$$
$$\left. \overline{\psi}\left(a_0^m \eta - \frac{2\pi k}{t_0}\right)\right| \left|\overline{f}(\eta)\right|^2 \,\mathrm{d}\eta .$$

Mit Voraussetzung (ii) ergibt sich dann also (analog für den zweiten Faktor in der Summe über k):

$$\sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m \eta\right)\right| \left|\overline{\psi}\left(a_0^m \eta - \frac{2\pi k}{t_0}\right)\right| \,\mathrm{d}\eta \leq \beta\left(-\frac{2\pi k}{t_0}\right) \|f\|^2$$
$$\sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \left|\overline{f}(\zeta')\right|^2 \left|\overline{\psi}\left(a_0^m \zeta'\right)\right| \left|\overline{\psi}\left(a_0^m \zeta' + \frac{2\pi k}{t_0}\right)\right| \,\mathrm{d}\zeta' \leq \beta\left(\frac{2\pi k}{t_0}\right) \|f\|^2 .$$

Also:

$$|\mathscr{E}| \leq \frac{2\pi}{t_0} \|f\|^2 \sum_{\substack{k\in\mathbb{Z} \\ k\neq 0}} \left[\beta\left(\frac{2\pi k}{t_0}\right) \beta\left(\frac{-2\pi k}{t_0}\right) \right]^{1/2} .$$

Diese Reihe konvergiert nach Voraussetzung und es ist

$$\lim_{t_0\to 0} \sum_{\substack{k\in\mathbb{Z} \\ k\neq 0}} \left[\beta\left(\frac{2\pi k}{t_0}\right) \beta\left(\frac{-2\pi k}{t_0}\right) \right]^{1/2} = 0 .$$

Ergo:

$$\kappa_\downarrow(\psi, a_0) - \sum_{\substack{k\in\mathbb{Z}\\k\neq 0}} \left[\beta\left(\frac{2\pi k}{t_0}\right)\beta\left(\frac{-2\pi k}{t_0}\right)\right]^{1/2} > 0$$

für alle $0 < t_0 < t_{max}$. Wir kommen nun zurück zu

$$\sum_{(m,n)\in\mathbb{Z}^2} |\langle f, \psi_{m,n}\rangle|^2 = \frac{2\pi}{t_0} \sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \left|\overline{f}(\eta)\right|^2 \left|\overline{\psi}\left(a_0^m \eta\right)\right|^2 \, \mathrm{d}\eta + \mathscr{E}$$

und erhalten mit Glg. (3.8) die gewünschte Formel. ■

Anmerkung 3.11. In diesem Satz tritt des öfteren das Infimum über alle $\omega \neq 0$ auf. Der Fall $\omega = 0$ muss ausgeschlossen werden, da sogar für sehr glatte $\overline{\psi}(\omega)$ die Summe $\sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2$ an dieser Stelle unstetig ist, weil $\overline{\psi}(0) = 0$ für alle Wavelets. Für das Haar-Wavelet zum Beispiel ist wie wir in einem Beispiel etwas später sehen werden $\left|\overline{\psi}(\omega)\right| = \frac{4}{\sqrt{2\pi}}|\omega|^{-1}\sin^2\omega/4$, aber $\sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 = 1/(2\pi)$ für $\omega \neq 0$, aber 0 für $\omega = 0$. Statt $\omega = 0$ auszuschließen müssten wir streng genommen statt dem Infimum ess inf über alle ω betrachten:

$$\text{ess inf}_x f(x) := \sup\{a \in \mathbb{R} : \lambda(\{y : f(y) < a\}) = 0\} \,,$$

wobei λ das Lebesgue-Maß auf $\mathbb{R}$ bezeichne. Für stetige Funktionen fallen ess inf und inf allerdings zusammen.

Zusammengefasst können wir also sagen, dass wenn ψ schnell genug in Zeit und Frequenz abklingt, dann existieren eine ganze Reihe von Parametern a_0, t_0, so dass $\psi_{m,n}$ einen Frame bildet.

Dieser Satz bedarf aufgrund seiner Wichtigkeit einer ausgiebigen Diskussion. Voraussetzung (i) ist recht einleuchtend, wobei die Anforderung an das Infimum der Summe ausdrückt, dass der Träger von $\overline{\psi}$ nicht kleiner als ein Intervall $(b, a_0 b)$ (b beliebig) sein darf. Man sagt die Nullstellen von $\overline{\psi}$ "konspirieren" nicht. Weiterhin können wir anmerken, dass die Summe invariant ist gegenüber einer Streckung der Form $\omega \to a_0\omega$, d.h. es reicht das Supremum über $\omega \in [1, a_0]$ zu erstrecken.
(ii) hingegen ist schon etwas obskurer, besagt aber eigentlich nur das folgende:

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} \left[\beta\left(-\frac{2\pi}{t_0}k\right)\beta\left(\frac{2\pi}{t_0}k\right)\right]^{1/2} < \infty \,. \tag{3.9}$$

Im allgemeinen gilt, dass eine Reihe der Form

$$\sum_n a_n b_n$$

dann konvergiert, wenn die Folge (a_n) von endlicher Variation ist (bzw. im reellen monoton mit $\lim_{n\to\infty} a_n < \infty$) und $\sum_{k=1}^{\infty} b_k < \infty$ (Konvergenzkriterium von Abel). In unserem Fall heißt das also, dass

$$\sum_{k=0}^{\infty} \sqrt{\beta\left(\frac{2\pi}{t_0}k\right)} < \infty \quad \text{und} \quad \sum_{k=-\infty}^{0} \sqrt{\beta\left(\frac{2\pi}{t_0}k\right)} < \infty\,. \tag{3.10}$$

Hieraus ergibt sich schon (Umkehrung gilt natürlich nicht), dass

$$\lim_{k\to\pm\infty} \beta\left(\frac{2\pi}{t_0}k\right) = 0\,.$$

Wir benötigen aber zusätzlich noch, dass die Konvergenz monoton ist! Wir müssen für Satz 3.4 also eigentlich "nur" fordern, dass $\beta(|x|)$ mit $|x|$ schnell genug abfällt, so dass die involvierten Reihen konvergieren. Eine natürliche (und "minimale") Wahl ist die Forderung, dass

$$\beta(|x|) \le \frac{1}{|x|^{1+\alpha'}}$$

für eine Konstante $\alpha' > 0$ und alle $x \in \mathbb{R}$. Jedes langsamere Abklingen von β lässt die Reihen divergieren. Durch die Wurzeln in den Glgn. (3.10) erhalten wir aber zunächst

$$\beta(|x|) \le \frac{1}{|x|^{\frac{1}{2}+\frac{\alpha'}{2}}} = \frac{1}{|x|^{1+\frac{1}{2}(\alpha'-1)}}\,.$$

Wir sehen schon, dass nun $\alpha' > 1$ gewählt werden muss für Konvergenz. Doch dies ist zuviel, denn zurück in unserer Ausgangsreihe (3.9) verschwindet die Wurzel, da wir $|x|$ betrachten und es ergibt sich für Konvergenz abschließend die Forderung

$$\beta(|x|) \le \frac{1}{|x|^{1+\alpha}}$$

mit $\alpha > 0$.
Nun möchten wir allerdings gerne schon dem Mutter-Wavelet "ansehen" können, ob die Bedingungen des Satzes erfüllt sind. Hierbei hilft uns folgendes Lemma weiter[4]:

Lemma 3.8. *Es sei wieder $a_0 > 1$ und*

$$\kappa_{\downarrow}(\psi, a_0) = \inf_{\omega\in[1,a_0]} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 > 0\,.$$

Angenommen es gäbe nun Konstanten $\alpha > 0$, $\gamma > 0$ und C, so dass

$$\left|\overline{\psi}(\omega)\right| \le \begin{cases} C|\omega|^{\gamma}, & |\omega| \le 1, \\ C|\omega|^{-1-\alpha}, & |\omega| > 1. \end{cases} \tag{3.11}$$

Dann folgt:

[4] Vergleiche mit der deutlich stärkeren Forderung an $\overline{\psi}$ in Daubechies (1992).

*(*i*) Für alle $\omega \neq 0$ ist*

$$\kappa_\uparrow(\psi, a_0) = \sup_{\omega\in[1,a_0]} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 \leq \frac{1}{a_0^{2(1+\alpha)}-1} + \frac{1}{a_0^{2\gamma}-1} + 1\,.$$

*(*ii*) Für eine Konstante K ist*

$$\beta(x) \leq \frac{K}{|x|^{1+\alpha}}\,.$$

Anders ausgedrückt: $\overline{\psi}$ muss mindestens so schnell abfallen wie (3.11), damit die Voraussetzungen von Satz 3.4 erfüllt sind ("Minimalforderung").

Beweis. Ad (*i*):

$$\begin{aligned}\sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 &\leq \sup_{1\leq\omega\leq a_0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 \\ &\leq C^2 \left(\sum_{m=-\infty}^{-1} \left(a_0^{m+1}\right)^{2\gamma} + \sum_{m=0}^{\infty} \left(a_0^{m+1}\right)^{-2(1+\alpha)} \right) \\ &= \frac{a_0^{2\gamma}}{a_0^{2\gamma}-1} + \frac{1}{a_0^{2(1+\alpha)}-1}\,.\end{aligned}$$

Hierbei haben wir benutzt, dass wir uns für ω auf das Intervall $[1, a_0]$ beschränken können (Invarianz der Summe gegenüber Multiplikation mit a_0) und damit $a_0^m\omega \leq 1$ für $m < 0$ und entsprechend $a_0^m\omega > 1$ für $m \geq 0$. Im zweite Schritt wird die Voraussetzung eingesetzt und dann die geometrische Reihe ausgewertet.

Ad (*ii*):
Wir gehen analog vor und schreiben zunächst:

$$\begin{aligned}\beta(x) &= \sup_{1\leq\omega\leq a_0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}(a_0^m\omega + x)\right| \\ &= C^2 \sup_{1\leq\omega\leq a_0} \left\{ \sum_{m<0} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}(a_0^m\omega + x)\right| + \sum_{m\geq 0} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}(a_0^m\omega + x)\right| \right\}\,.\end{aligned}$$

1. Es sei zunächst $|x| \geq 2$.
   Für $m < 0$ ist immer $|a_0^m\omega + x| > 1$, da $a_0 > 1$ und $\omega \in [1, a_0]$ gewählt wurden, d.h. $|a_0^m\omega| \leq 1$. In diesem Bereich schätzen wir folglich ab:

$$\left|\overline{\psi}\left(|a_0^m\omega + x|\right)\right| \leq \left(\frac{1}{|a_0^m\omega + x|}\right)^{1+\alpha} \leq \left(\frac{|x|}{2}\right)^{-1-\alpha}\,,$$

weil wegen $|x| \geq 2$

$$|a_0^m \omega + x| \geq |x| - 1 \geq \frac{|x|}{2} \geq 1 \,.$$

Zurück zur Summe heißt das:

$$\sum_{m<0} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| \leq \sum_{m<0} |a_0^m \omega|^\gamma \left(\frac{|x|}{2}\right)^{-1-\alpha} = \frac{\omega^\gamma}{a_0^\gamma - 1} \left(\frac{|x|}{2}\right)^{-1-\alpha} \,.$$

Nun zu $m \geq 0$. In diesem Bereich müssen wir eine Unterscheidung treffen zwischen solchen x, für die $|a_0^m \omega + x| \leq 1$ und x für die das nicht zutrifft. Es ist klar, dass dies für $-1 - a_0^m \omega \leq x \leq 1 - a_0^m \omega$ gegeben ist. Es gibt also in Abhängigkeit von a_0 einige m, so dass x genau in dieses Intervall fällt. Diese m sind auf jeden Fall zusammenhängend, wir wissen allerdings nicht a priori wie viele es sind. Wir diskutieren diesen Punkt etwas später im Beweis weiter. An dieser Stelle nehmen wir einfach an, dass

$$-1 - a_0^m \omega \leq x \leq 1 - a_0^m \omega \quad \text{für} \quad m \in M := \{M_1, ..., M_2\} \subset \mathbb{N} \,.$$

Wir spalten nun die Summe auf:

$$\sum_{m\geq 0} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| = \sum_{\substack{m\geq 0 \\ m \notin M}} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| + \sum_{m \in M} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right|$$

und formen dies weiter um mit einem Trick und nutzen die Abschätzungen von oben:

$$\sum_{m\geq 0} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| \leq \sum_{m\geq 0} |a_0^m \omega|^{-\bar{\alpha}} \left(\frac{|x|}{2}\right)^{-\bar{\alpha}} - \sum_{m \in M} |a_0^m \omega|^{-\bar{\alpha}} \left(\frac{|x|}{2}\right)^{-\bar{\alpha}} + \sum_{m \in M} |a_0^m \omega|^{-\bar{\alpha}} |a_0^m \omega + x|^\gamma \,,$$

wobei $\bar{\alpha} = 1 + \alpha$. Der Term $|a_0^m \omega + x|$ ist ≤ 1 per Definition von M. Also schätzen wir weiter ab und berechnen die Summen:

$$\sum_{m\geq 0} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| \leq \frac{\omega^{-1-\alpha} a_0^{1+\alpha}}{a_0^{1+\alpha} - 1} \left(\frac{|x|}{2}\right)^{-1-\alpha} - \mathcal{K}\omega^{-1-\alpha} \left(\frac{|x|}{2}\right)^{-1-\alpha} + \mathcal{K}\omega^{-1-\alpha} \,.$$

Hierbei haben wir ausgenutzt, dass

$$\mathcal{K} := \sum_{m=M_1}^{M_2} |a_0^m \omega|^{-1-\alpha} = \frac{a_0^{(-1-\alpha)(M_1-1)} - a_0^{(-1-\alpha)M_2}}{a_0^{1+\alpha} - 1} = \text{const}\,.$$

Wir sehen, dass $\mathcal{K}$ stets positiv ist. Nun können wir die positive und negative Summe wieder zusammenfassen[5]:

$$\sum_{m<0}(\cdot) + \sum_{m\geq 0}(\cdot) \leq \left(\frac{|x|}{2}\right)^{-1-\alpha} \left[\frac{\omega^\gamma}{a_0^\gamma - 1} + \frac{\omega^{-1-\alpha} a_0^{1+\alpha}}{a_0^{1+\alpha}-1} - \mathcal{K}\omega^{-1-\alpha}\right] + \mathcal{K}\omega^{-1-\alpha}\,.$$

Die Funktion in der Klammer hat ihr Maximum bei $\omega = a_0$, d.h. wir erhalten:

$$\beta(x) \leq \left(\frac{|x|}{2}\right)^{-1-\alpha} C^2 \left[1 + \frac{1}{a_0^\gamma - 1} + \frac{1}{a_0^{1+\alpha}-1} - \frac{\mathcal{K}}{a_0^{1+\alpha}}\right] + \frac{\mathcal{K}}{a_0^{1+\alpha}}\,,$$

also wie gewünscht

$$\beta(|x|) \leq \frac{K}{|x|^{1+\alpha}}\,.$$

2. Nun sei $-1 \leq x < 0$.
Hier gilt für alle $m < 0$, dass $|a_0^m \omega + x| \leq 1$ und wir schreiben für die Summe:

$$\sum_{m<0}(\cdot) \leq \sum_{m<0} |a_0^m \omega|^\gamma = \frac{\omega^\gamma}{a_0^\gamma - 1}\,.$$

Nun zu $m \geq 0$ in diesem Bereich für x. Wir sehen sofort, dass wegen $a_0^m \omega \geq 1$ hier $M_1 = 0$ ist und wir spalten wieder auf:

$$\begin{aligned}\sum_{m\geq 0}(\cdot) &= \sum_{m=M_2+1}^{\infty} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}(a_0^m\omega + x)\right| + \sum_{m=0}^{M_2} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}(a_0^m\omega + x)\right| \\ &\leq \sum_{m=M_2+1}^{\infty} |a_0^m\omega|^{-1-\alpha} \left(\frac{|x|}{2}\right)^{-1-\alpha} + \mathcal{K}'\omega^{-1-\alpha} = \frac{a_0^{(-1-\alpha)M_2}\omega^{-1-\alpha}}{a_0^{1+\alpha}-1} + \mathcal{K}'\omega^{-1-\alpha}\,,\end{aligned}$$

wobei

$$\mathcal{K}' := \sum_{m=0}^{M_2} |a_0^m\omega|^{-1-\alpha} = \frac{a_0^{(1+\alpha)} - a_0^{(-1-\alpha)M_2}}{a_0^{1+\alpha}-1}\,.$$

Zusammengefasst erhalten wir also (das Supremum bestimmt sich zu $\omega = a_0$):

$$\beta(|x|) \leq C^2 \left[\frac{a_0^\gamma}{a_0^\gamma - 1} + \frac{a_0^{2(-1-\alpha)M_2}}{a_0^{1+\alpha}-1} + \frac{\mathcal{K}'}{a_0^{1+\alpha}}\right] = \text{const.}$$

Da $-1 \leq x < 0$ ist dies in jedem Fall kleiner als ein Ausdruck der Form $K'|x|^{-1-\alpha}$, wie gefordert.

[5] wir schreiben abkürzend $\sum_{m<0}(\cdot)$ für $\sum_{m<0} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}(a_0^m\omega + x)\right|$ und analog $\sum_{m\geq 0}(\cdot)$

3. Nun sei $0 < x < 1$.
   Für $m \geq 0$ gibt es keine Probleme, da hier immer $|a_0^m \omega + x| > 1$ ist, d.h. wir schätzen wie bei 1. ab:
   $$\sum_{m \geq 0} (\cdot) \leq \frac{\omega^{-1-\alpha} a_0^{1+\alpha}}{a_0^{1+\alpha} - 1} \left(\frac{|x|}{2} \right)^{-1-\alpha} .$$
   Für $m < 0$ allerdings gibt es ein $M_2 < 0$, so dass für alle $m \in \{-M_2, ..., 0\}$ $|a_0^m \omega + x| > 1$. Wir teilen in diesem Fall wie in den Fällen zuvor die Summe auf und erhalten wieder
   $$\beta(|x|) \leq \frac{K''}{|x|^{1+\alpha}} .$$
4. $1 \leq x < 2$:
   In diesem Bereich ist sowohl für $m \geq 0$ als auch für $m < 0$ der Ausdruck $|a_0^m \omega + x|$ stets > 1 und wir brauchen keine der Summen aufzuspalten, so dass wir mit derselben Abschätzung wie in 1. und 3. für $m < 0$ respektive $m \geq 0$ wieder
   $$\beta(|x|) \leq \frac{K'''}{|x|^{1+\alpha}}$$
   erhalten.
5. Der problematischste Fall ist $-2 < x < -1$.
   In diesem Bereich ist $M_1 < 0$ und $M_2 > 0$, d.h. wir spalten sowohl bei $m < 0$ als auch bei $m \geq 0$ auf:
   $$\sum_{m \geq 0} (\cdot) = \sum_{m=M_2+1}^{\infty} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| + \sum_{m=0}^{M_2} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right|$$
   und ebenso:
   $$\sum_{m < 0} (\cdot) = \sum_{m=-1}^{M_1} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| + \sum_{m=M_1-1}^{-\infty} \left|\overline{\psi}(a_0^m \omega)\right| \left|\overline{\psi}(a_0^m \omega + x)\right| .$$
   Wir schätzen die Summen nun ab wie in den Fällen zuvor, so dass wir wieder mit
   $$\beta(|x|) \leq \frac{K''''}{|x|^{1+\alpha}}$$
   enden.
6. Der Fall $x = 0$ ist trivial, da hier $K|x|^{-1-\alpha} \to \infty$, aber
   $$\beta(0) = \kappa_\uparrow \leq \text{const},$$
   wie wir in (i) gezeigt haben.

Ein kleines Detail lohnt noch der Diskussion. Und zwar die Frage, wie genau M von a_0 abhängt. Wir haben M so definiert, dass hier $-1 - a_0^m\omega \leq x \leq 1 - a_0^m\omega$, d.h. dieses Intervall hat offensichtlich die Länge 2. Es hängt aber von a_0 ab für wie viele m x tatsächlich in diesen Bereich fällt. Wir betrachten zunächst den Fall $m \geq 0$. Dann schiebt sich das Intervall immer weiter "nach links", d.h. in negativer Richtung. Wir vergleichen also $1 - a_0^{m+1}\omega$ und $-1 - a_0^m\omega$:

$$1 - a_0^{m+1}\omega + 1 + a_0^m\omega = 2 - a_0^m\omega(a_0 - 1)\,.$$

Wird dieser Ausdruck negativ, so gibt es folglich nur ein einziges m, so dass x in das Intervall fällt. D.h. im Minimalfall $\omega = 1$

$$a_0^m\omega(a_0 - 1) \geq 2 \Longleftrightarrow a_0^m \geq \frac{2}{a_0 - 1}\,.$$

Nach m aufgelöst erhalten wir die Bedingung:

$$m \geq \frac{\ln \frac{2}{a_0 - 1}}{\ln a_0}\,.$$

Wir sehen, dass dieser Ausdruck für $a_0 = 2$ zu 1 wird, d.h. für $a_0 > 2$ überschneiden sich die Intervalle nicht mehr für $m \geq 1$. Damit sich allerdings die Intervalle für $m = 0$ und $m = 1$ nicht überschneiden, muss a_0 mindestens 3 sein, denn dann wird der Ausdruck zu Null.
Für $m < 0$ stellt sich die Situation so dar, dass sich die Intervalle für alle m überschneiden, und zwar strebt $-1 - a_0^m\omega$ von "links" gegen -1 und $1 - a_0^m\omega$ von links gegen $+1$. Fällt ein x in den Bereich $(-1, 0)$, so liegt es in allen Intervallen (s. Unterscheidung 2.). Zwischen 0 und 1 hingegen gibt es wie in 3. argumentiert einen Punkt, ab dem x in allen nachfolgenden Intervallen liegt. Ebenso zwischen -1 und -2 (Unterscheidung 5.). ■

Wir werden sogleich zeigen, dass z.B. das Mexikanerhut-Wavelet dieses Lemma erfüllt, es aber auch durchaus Wavelets gibt, die dies nicht tun, wie z.B. das Haar-Wavelet.

Zunächst aber noch eine zweite Erweiterung zu Satz 3.4. Für einen *straffen* Frame, d.h. $A = B$ muss laut diesem Satz also gelten:

$$\inf_{\omega\neq 0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 - \sup_{\omega\neq 0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 = 2 \sum_{k\in\mathbb{Z}\setminus\{0\}} \left[\beta\left(-\frac{2\pi}{t_0}k\right)\beta\left(\frac{2\pi}{t_0}k\right)\right]^{1/2}\,.$$

Die rechte Seite dieser Gleichung ist stets positiv oder Null, d.h. wir folgern sofort, dass für $A = B$

$$\inf_{\omega\neq 0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 = \sup_{\omega\neq 0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 = A > 0$$

sowie

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} \left[\beta\left(-\frac{2\pi}{t_0}k\right)\beta\left(\frac{2\pi}{t_0}k\right)\right]^{1/2} = 0\,.$$

Offensichtlich zwei sehr starke Bedingungen, die uns direkt zu dem folgendem Korollar führen:

Korollar 3.1. *Für die Fourier-Transformierte $\overline{\psi}$ des Wavelets ψ gelte:*

*(*i*) $\overline{\psi}$ habe kompakten Träger im Intervall $[\omega_1,\omega_2]$, wobei $\omega_2 > \omega_1 > 0$,*

*(*ii*) für $1 \le \omega \le a_0$ sei*

$$\sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 = A' > 0\,.$$

Dann ist $\dot{\Psi}(a_0,t_0)$ mit

$$0 < t_0 \le t_{max} = \frac{2\pi}{\omega_2-\omega_1}$$

ein straffer Frame mit

$$A = B = \frac{2\pi}{t_0 \ln a_0}\int_0^\infty \frac{\left|\overline{\psi}(\omega)\right|^2}{\omega}\,\mathrm{d}\omega\,.$$

Beweis. Der erste Teil dieses Satzes findet sich schon in ähnlicher Form in Blatter (1998), allerdings mit gänzlich anderem Beweis. In unserem Fall ergibt sich die Straffheit des Frames direkt als Korollar aus Satz 3.4. Es ist nämlich durch Bedingung (ii) sichergestellt, dass

$$\inf_{\omega\in[1,a_0]} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 = \sup_{\omega\in[1,a_0]} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 > 0$$

und wegen Voraussetzung (i) erhalten wir, $k \in \mathbb{Z}^+\setminus\{0\}$ und o.B.d.A. $t_0 = t_{max}$,

$$\begin{aligned}
\beta\left(\frac{2\pi}{t_{max}}k\right) &= \sup_{\omega\in[1,a_0]} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right| \left|\overline{\psi}\left(a_0^m\omega + \frac{2\pi}{t_{max}}k\right)\right| \\
&= \sup_{\omega\in[1,a_0]} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right| \underbrace{\left|\overline{\psi}\left(a_0^m\omega + (\omega_2-\omega_1)k\right)\right|}_{=0} \\
&= 0\,,
\end{aligned}$$

weil nach Voraussetzung (i) $\overline{\psi}$ kompakten Träger $[\omega_1,\omega_2]$ hat und $a_0^m\omega > 0$ für alle $\omega \in [1,a_0]$ und alle $m \in \mathbb{Z}$. Für negatives k gilt entsprechend

$$\beta\left(-\frac{2\pi}{t_{max}}k\right) = 0\,,$$

d.h. wir erhalten für alle $0 < t_0 \le t_{max}$

$$\sum_{k\in\mathbb{Z}\setminus\{0\}} \left[\beta\left(-\frac{2\pi}{t_0}k\right)\beta\left(\frac{2\pi}{t_0}k\right)\right]^{1/2} = 0\,.$$

Für den zweiten Teil des Satzes müssen wir deutlich tiefer schürfen und benötigen den folgenden Hilfssatz, der schon in Daubechies (1992) auftaucht:

Lemma 3.9. *Falls die Familie $\dot{\Psi}(a_0, t_0)$ einen Frame für $L^2(\mathbb{R})$ mit Frame-Parametern A, B bildet, dann gilt*

$$\frac{b_0 \ln a_0}{2\pi} A \le \int_0^\infty \frac{\left|\overline{\psi}(\omega)\right|^2}{\omega}\,\mathrm{d}\omega \le \frac{b_0 \ln a_0}{2\pi} B \quad \textit{und}$$
$$\frac{b_0 \ln a_0}{2\pi} A \le \int_{-\infty}^0 \frac{\left|\overline{\psi}(\omega)\right|^2}{\omega}\,\mathrm{d}\omega \le \frac{b_0 \ln a_0}{2\pi} B\,.$$

Beweis. Siehe Daubechies (1992).

Aus diesem Lemma folgt dann mit $A = B$ sofort unsere Behauptung. ■

Dieses Korollar gibt uns also Voraussetzungen, unter denen ein gegebenes Wavelet in jedem Fall einen straffen Frame bildet. Die *Multiskalen-Analyse* geht umgekehrt vor und *konstruiert* Wavelets von vornherein so, dass sie eine orthonormale Basis des L^2 bilden (s. Blatter (1998) oder Daubechies (1992)).

Wir fassen nun zusammen. Die vorangegangenen Sätze haben uns nun einen Rahmen geschaffen, in den wir ein gegebenes Wavelet ψ einordnen können:

(a) Klingt die Fourier-Transformierte des Wavelets schneller als $1/x$ ab und konspirieren die Nullstellen von $\overline{\psi}$ nicht, dann gibt es ein t_{max}, so dass $\dot{\Psi}(a_0, t_0)$ ein Frame ist für $a_0 > 1$ und $0 < t_0 < t_{max}$. Damit lässt sich jede Funktion aus L^2 durch die Mitglieder aus $\dot{\Psi}(a_0, t_0)$ und die Elemente des dualen Frames darstellen, allerdings nicht eindeutig und i.A. *redundant*.

(b) Hat $\overline{\psi}$ kompakten Träger und konspirieren die Nullstellen nicht, so folgt aus (a), dass $\dot{\Psi}(a_0, t_0)$ ein Frame von L^2 ist mit einem bestimmten t_{max}. Dieser Frame ist sogar straff, also $A = B$. D.h. der Frame Operator ist ein Vielfaches der Identität in L^2, der Umkehroperator folglich ebenso, und jede Funktion aus L^2 lässt sich somit allein durch die Mitglieder aus $\dot{\Psi}(a_0, t_0)$ darstellen, allerdings nur eindeutig wenn auch $A = B = 1$. Dies erreicht man durch Normierung des Wavelets auf 1 bzgl. $\|\cdot\|_{L^2}$.

Wir weisen darauf hin, dass (a) und (b) nur Implikationen und keine "genau-dann-wenn"-Aussagen sind. Eine Erweiterung der Sätze auf ein aussagekräftigeres Kriterium fällt aber schwer wie das Beispiel des Haar-Wavelets zeigt, das kompakten Träger hat, weder (a) noch (b) erfüllt und trotzdem eine Orthonormalbasis von L^2 bildet. Wir werden nun ne-

ben dem Haar-Wavelet auch einige andere Beispiele diskutieren, wie das Mexikanerhut-Wavelet, welches (a) erfüllt, aber nicht (b), so dass kein straffer Frame entsteht. Das Meyer-Wavelet (Beispiel 2) erfüllt (b) und ist normiert auf 1, d.h. es bildet eine Orthonormalbasis von L^2. Alle diese Wavelet-Beispiele haben gemeinsam, dass entweder ψ selbst, oder $\overline{\psi}$ kompakten Träger hat. Es ist jedoch möglich straffe Frames aus anderen Wavelets zu konstruieren, z.B. aus Wavelets mit exponentiellem Abklingverhalten im Realraum als auch im Fourier-Raum. Ausgangspunkt hierfür ist die gefensterte Fourier-Transformation (Gabor Transformation). Details sind nachzulesen in Daubechies (1992).

Bevor wir nun zu den versprochenen Beispielen kommen bringen wir eine weitere allgemeine Beobachtung an, die uns in der Klassifikation eines gegebenen Wavelets weiterhilft. Wie in den vorangegangenen Sätzen gezeigt ist stets

$$A \leq \frac{2\pi}{t_0} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 \leq B$$

und je straffer der Frame wird (also je näher $B/A \geq 1$ an 1 herankommt), desto konstanter muss die Funktion $\sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2$ für $\omega \in [1, a_0]$ werden. Wir werden nun demonstrieren, dass es z.B. dem Mexikanerhut-Wavelet genau diese Eigenschaft fehlt und es deshalb keinen straffen Frame erzeugen kann.

Beispiel 1: Das Mexikanerhut-Wavelet
Das Mexikanerhut-Wavelet ist definiert als $\psi : \mathbb{R} \to \mathbb{R}$

$$\psi(x) = \frac{2}{\sqrt{3}\,\pi^{1/4}} \left(1 - x^2\right)\, e^{-x^2/2}\,, \quad \text{so dass}$$
$$\overline{\psi}(\omega) = \frac{2}{\sqrt{3}\,\pi^{1/4}}\, \omega^2\, e^{-\omega^2/2}\,.$$

Abb. 3.1 zeigt einige Mitglieder der von diesem Mutter-Wavelet erzeugten Familie $\dot{\Psi}(a_0, t_0)$, definiert in Glg. (3.15).

$\dot{\Psi}(a_0, t_0)$ ist nach Lemma 3.8 (und Satz 3.4) ein Frame, denn:

(a) Es gilt offensichtlich für $a_0 > 1$ und $\omega \in [1, a_0]$

$$\sum_{m\in\mathbb{Z}} \left|\overline{\psi}(a_0^m\omega)\right|^2 > 0\,,$$

(b) und

$$\left|\overline{\psi}(\omega)\right| \leq \begin{cases} C|\omega|^{\gamma}, & |\omega| \leq 1, \\ C|\omega|^{-1-\alpha}, & |\omega| > 1, \end{cases}$$

mit z.B. $\gamma = 2$; für jedes $\alpha > 0$ gibt es ein C, so dass $|\omega|^{-1-\alpha}$ schneller abfällt als $e^{-\omega^2}$.

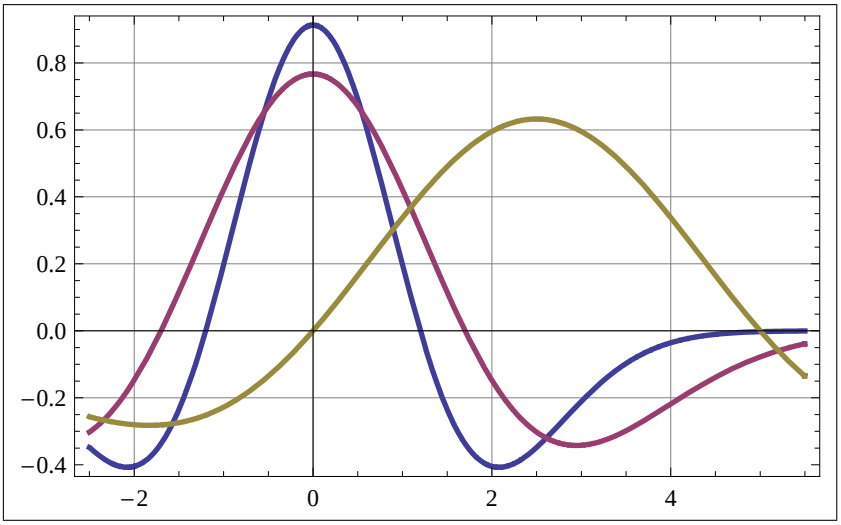


Abb. 3.1. Die beiden Wavelets links: $\psi_{1,0}$ mit $a_0 = 1.2$ und $\psi_{1,0}$ mit $a_0 = 1.7$. Das Wavelet rechts: $\psi_{1,1}$ mit $a_0 = 2.5$ und $t_0 = 1$.

Aber wann wird dieser Frame auch straff? Diese Funktion hat offensichtlich keinen kompakten Träger, d.h. wir können Korollar 3.1 nicht anwenden. Man sieht schon ohne Rechnung sofort, dass der Ausdruck

$$\sum_{m\in\mathbb{Z}} a_0^{4m}\omega^4\, e^{-a_0^{2m}\omega^2} \tag{3.12}$$

alles andere als konstant ist für $\omega \in [1, a_0]$ (für den a priori ausgeschlossenen Fall $a_0 = 1$ konvergiert die Reihe natürlich nicht einmal). Analytisch ist diese Reihe leider schwer zu lösen, aber numerisch für ein festes a_0 relativ problemlos. Abbildung 3.2 zeigt den Wert der Reihe für einige a_0 (numerische Details zur Approximation der Reihe siehe weiter unten). Man sieht deutlich, dass die Variation für $a_0 \to 1$ immer kleiner werden, der Frame also immer straffer. In der folgenden Tabelle sind einige Werte für A und B gegeben, berechnet aus den Formeln in Satz 3.4:

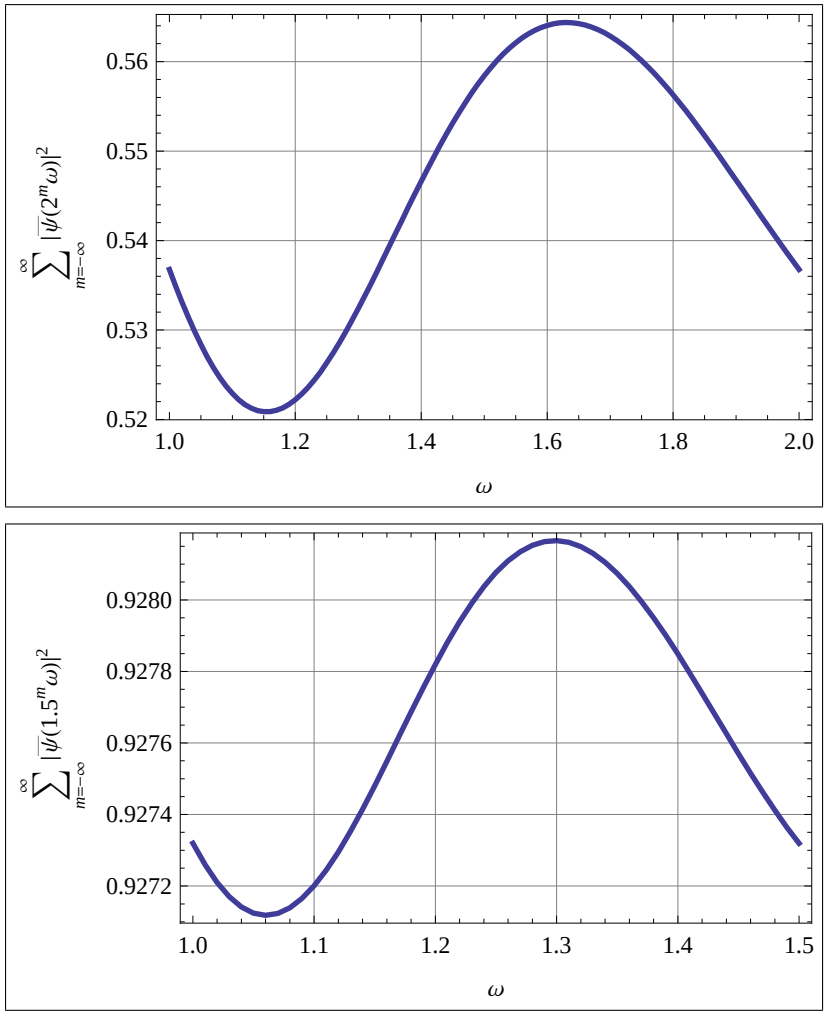


Abb. 3.2. Wert der Reihe (3.12) für $a_0 = 2$ und $a_0 = 1.5$.

	t_0	A	B	B/A
$a_0 = 2$	0.25	13.091	14.184	1.083
	0.50	6.545	7.092	1.083
	0.75	4.364	4.728	1.083
	1.00	3.273	3.546	1.083
	1.50	2.133	2.413	1.131
	2.00	1.248	2.162	1.732
	2.50	0.522	2.206	4.226
	3.00	0.091	2.182	23.978
	3.50	−0.160	2.108	—
	4.00	−0.342	2.047	—

	t_0	A	B	B/A
$a_0 = 1.5$	0.25	23.306	23.327	1.001
	0.50	11.653	11.664	1.001
	0.75	7.769	7.776	1.001
	1.00	5.826	5.832	1.001
	1.50	3.806	3.966	1.040
	2.00	2.274	3.555	1.563
	2.50	1.033	3.630	3.514
	3.00	0.311	3.575	11.495
	3.50	−0.105	3.436	—
	4.00	−0.408	3.323	—

Die numerischen Details dieser Rechnungen sind die folgenden: Der Ausdruck

$$a_0^{4m}\omega^4\, e^{-a_0^{2m}\omega^2}$$

hat sein Maximum (bzgl. ω) bei $\omega = \frac{\sqrt{2}}{a_0^m}$, d.h. abhängig von a_0 fällt dieses in das Intervall $[1, a_0]$ oder nicht. Für $m < 0$ ist das Maximum sogar immer größer als a_0. Wir schließen also:

$$a_0^{4m}\omega^4\, e^{-a_0^{2m}\omega^2} \leq \begin{cases} a_0^{4m}\, e^{-a_0^{2m}}, & m \geq \frac{\ln 2}{\ln a_0}, \\ a_0^{4(m+1)}\, e^{-a_0^{2(m+1)}}, & m < 0. \end{cases}$$

Wir vergleichen mit der Maschinengenauigkeit, so dass:

$$a_0^{4m}\omega^4\, e^{-a_0^{2m}\omega^2} \leq 10^{-15} \quad \text{für} \quad \begin{cases} m \geq 1.86902/\ln a_0, \\ m \leq 1 - 8.63469/\ln a_0. \end{cases}$$

Für z.B. $a_0 = 2$ ergibt sich also, dass die Summanden in Ausdruck (3.12) unter die Maschinengenauigkeit fallen für $m \geq 3$ und $m \leq -14$. Wir können also die in der Berechnung von A und B auftretenden Reihen problemlos numerisch approximieren als endliche Summen. Die Minimierung bzw. Maximierung der Summen bzgl. ω wurde mit einem einfachen Newton-Verfahren und zur Kontrolle mit einem Conjugate-Gradient-Verfahren durchgeführt. Wie in Abbildung 3.2 zu sehen erwarten wir auch bei der Bestimmung der Minima und Maxima keine numerischen Probleme, da die Reihe für $a_0 > 1$ eine gut-konditionierte Funktion ist.
Wir können analoge Überlegungen für die Reihe

$$\sum_{m\in\mathbb{Z}} \left[a_0^{2m}\omega^2\, e^{-\frac{1}{2}a_0^{2m}\omega^2} \left(\frac{2\pi k}{t_0} + a_0^m\omega \right)^2 e^{-\frac{1}{2}\left(\frac{2\pi k}{t_0} + a_0^m\omega\right)^2} \right], \quad k \in \mathbb{Z}\backslash\{0\}\,, \tag{3.13}$$

anstellen, wie sie in der Berechnung von $\beta\left(\pm\frac{2\pi k}{t_0}\right)$ auftaucht. Wir begnügen uns für diesen Teil mit einer numerischen Abschätzung anhand der folgenden Tatsachen: Ausdruck

(3.13) nimmt mit steigendem $t_0 > 0$ zu, aber mit steigenden $a_0 > 1$ ab. Wir wählen also maximales t_0, in unserem Fall $t_0 = 3.5$, und minimales a_0, z.B. $a_0 = 1.1$. (3.13) nimmt (sehr) schnell ab mit steigendem k, wir wählen also $k = 1$ (die Fälle $k < 0$ behandeln wir im Anschluss). Es stellt sich heraus ($\omega \in [1, a_0]$ beliebig), dass

$$\sum_{m=-\infty}^{\infty} (\,\cdot\,) - \sum_{m=-M}^{M} (\,\cdot\,), \quad M \in \mathbb{N}\,,$$

also der "Reihenabbruchs-Fehler" in (3.13), für $M \geq 200$ kleiner als die Maschinengenauigkeit ist. Weiterhin zeigt sich, dass der Wert von (3.13) für $k \geq 5$ ebenfalls schon $< 10^{-15}$ ist. Ebenso verhält es sich für $k < 0$. Zusammenfassend schließen wir guten Gewissens, dass alle auftretenden Reihen in der Berechnung des zweiten Terms von A bzw. B approximiert werden können durch endliche Reihen, wobei $m \in \{-200, ..., 200\}$ und $k \in \{-10, ..., 10\}\backslash\{0\}$ mehr als ausreichende Genauigkeit garantiert.

Die in der Tabelle grau unterlegten Zellen stellen die Werte von t_0 dar, für die A negativ wird. Für alle $t_0 \geq t_{max}$ ist $\dot{\Psi}(a_0, t_0)$ nicht mehr unbedingt ein Frame. Dieser abrupte Übergang von Frame zu nicht-Frame geschieht bei genauerer Betrachtung für $a_0 = 2$ bei $t_{max} \approx 3.1565$ und für $a_0 = 1.5$ bei $t_{max} \approx 3.3515$. Wir sehen weiterhin, dass der Frame als nahezu straff angesehen kann bis $t_0 \approx 1.00$. Diese beiden Rechnungen zeigen, dass es für $1.5 \leq a_0 \leq 2$ kein t_0 gibt, so dass $A = B = 1$. Aber mit einfachen Überlegungen sehen wir sofort, dass dies auch außerhalb dieses Bereiches nicht möglich ist: Wählen wir $a_0 > 2$, so weicht B/A immer mehr von 1 ab und wählen wir $1 < a_0 < 1.5$, so steigt B immer weiter, da je näher wir an $a_0 = 1$ herankommen der Wert der Summe (3.12) unbeschränkt steigt. Zusammengefasst:

Das Mexikanerhut-Wavelet kann also für $a_0 > 1$ keine Basis von L^2 bilden.

Wie schon in Anmerkung 3.5 zu der Frame-Definition 3.2 argumentiert, ist für straffe Frames der Wert $A = B$ eine Maß der *Redundanz* des Frames. Umgemünzt auf unsere diskrete Familie $\dot{\Psi}(a_0, t_0)$ können wir z.B. so argumentieren, dass wenn t_0 halbiert wird, so überlappen sich die Wavelets in der Familie (hier die "Mexikanerhüte") doppelt so stark, d.h. die Redundanz des Frames sollte sich verdoppeln. Die folgende Tabelle illustriert diese Idee und zeigt das Verhalten von A und B mit steigendem t_0:

t_0	A	B	$A \cdot t_0$	$B \cdot t_0$
0.25	13.091	14.184	3.273	3.546
0.50	6.545	7.092	3.273	3.546
0.75	4.364	4.728	3.273	3.546
1.00	3.273	3.546	3.273	3.546
1.50	2.133	2.413	3.200	3.620
2.00	1.248	2.162	2.496	4.324
2.50	0.522	2.206	1.305	5.515
3.00	0.091	2.182	0.273	6.546

Wir sehen, dass A und B in dem Bereich, in dem der Frame straff ist (bis $t_0 \approx 1.00$) invers proportional zu t_0 sind, verdoppelt sich also t_0, so halbiert sich $A \approx B$, wie erwartet. Abb. 3.3 illustriert diese Überlegung.

Das Mexikanerhut-Wavelet scheint also keine gut geeignete Basisfunktion zu sein, da es schließlich auch Wavelets gibt, die einen straffen Frame und normiert sogar eine Hilbertbasis von L^2 bilden. Rufen wir uns aber Teil (*iii*) des Rekonstruktionssatzs 3.3 in Erinnerung, so sehen wir, dass für $a_0 = 2$ und $t_0 = 1$

$$\frac{B-A}{B+A} \approx 0.040\,,$$

so dass die Rekonstruktion

$$f \approx \frac{2}{A+B} \sum_{(m,n)\in\mathbb{Z}^2} \mathcal{W}_\psi f(a_m, t_n) \sum_{k=0}^{N} \left(\mathbb{1} - \frac{2}{A+B} \mathscr{T} \right)^k \psi_{m,n}$$

schon für $N \geq 10$ besser als die Maschinengenauigkeit ist, also

$$\left\| f - \frac{2}{A+B} \sum_{(m,n)\in\mathbb{Z}^2} \mathcal{W}_\psi f(a_m, t_n) \sum_{k=0}^{N} \left(\mathbb{1} - \frac{2}{A+B} \mathscr{T} \right)^k \psi_{m,n} \right\| \leq 10^{-15} \|f\|\,.$$

Dieses Wavelet ist also sehr wohl in der Lage jedes beliebige Signal in L^2 nur mit minimalem numerischem Fehler zu approximieren.

Beispiel 2: Das Meyer-Wavelet
Das Meyer-Wavelet ist über seine Fourier-Transformierte definiert. In der Literatur (s. Meyer (1992)) wird zunächst die Hilfsfunktion $\nu : \mathbb{R} \to [0,1]$

$$\nu(x) := \begin{cases} 0, & x \leq 0, \\ 10x^3 - 15x^4 + 6x^5, & 0 \leq x \leq 1 \\ 1, & x \geq 1 \end{cases}$$

eingeführt, mit deren Hilfe dann die Funktion $w : \mathbb{R} \to [-1,1]$

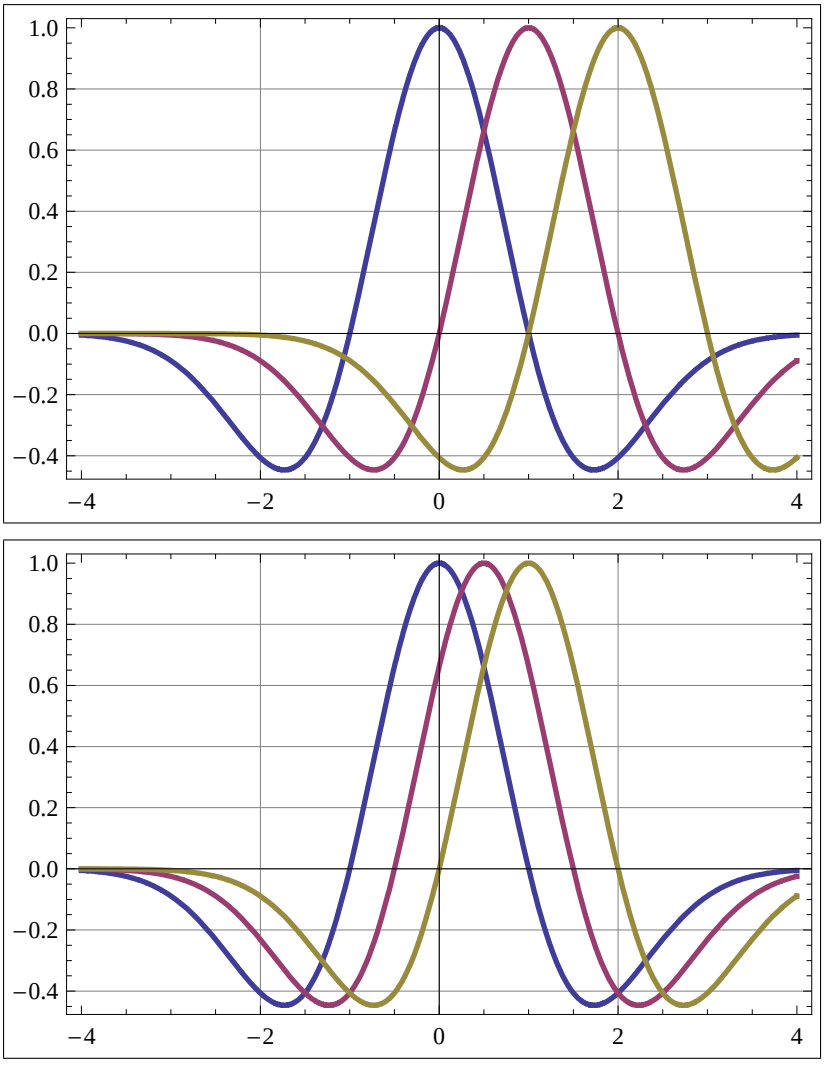


Abb. 3.3. Die Mexikanerhut-Wavelets $\psi_{0,0}(x)$, $\psi_{0,1}(x)$ und $\psi_{0,2}(x)$ für $t_0 = 1.0$ (oben) und $t_0 = 0.5$ (unten). Die Wavelets überlappen für halbiertes t_0 "doppelt so stark", d.h. die Redundanz in $\dot{\Psi}(a_0, t_0)$ ist "doppelt so groß".

$$w(\omega) := \begin{cases} \sin\left[\frac{\pi}{2}\nu\left(\frac{3\omega}{2\pi}-1\right)\right], & \frac{2\pi}{3} \le \omega \le \frac{4\pi}{3}, \\ \cos\left[\frac{\pi}{2}\nu\left(\frac{3\omega}{4\pi}-1\right)\right], & \frac{4\pi}{3} \le \omega \le \frac{8\pi}{3}, \\ 0, & \text{sonst} \end{cases}$$

und zu guter letzt die Fourier-Transformation des Meyer-Wavelets als

$$\begin{aligned} \overline{\psi} : \mathbb{R} &\longrightarrow \mathbb{C} \\ \omega &\longmapsto \frac{1}{\sqrt{2\pi}} \left[w(\omega) + w(-\omega)\right] \, e^{i\frac{\omega}{2}} \, . \end{aligned}$$

Das Meyer-Wavelet selbst ist dann definiert als

$$\begin{aligned} \psi : \mathbb{R} &\longrightarrow \mathbb{C} \\ x &\longmapsto \overline{\overline{\psi}}(-x) \, . \end{aligned}$$

$\overline{\psi}$ hat kompakten Träger:

$$\operatorname{supp} \overline{\psi} \subset \left[-\frac{8\pi}{3}\pi, -\frac{2}{3}\pi\right] \cup \left[\frac{2}{3}\pi, \frac{8}{3}\pi\right] \, .$$

Jetzt müssen wir nur noch die folgenden zwei Dinge zeigen.

(a) $\|\psi\|_{L^2} = 1$:

$$\begin{aligned} \|\psi\|_{L^2}^2 &= \|\overline{\psi}\|_{L^2}^2 \\ &= \frac{2}{2\pi} \int_{\frac{2\pi}{3}}^{\frac{4\pi}{3}} \sin^2\left[\frac{\pi}{2}\nu\left(\frac{3}{2\pi}x-1\right)\right] \, \mathrm{d}x + \frac{2}{2\pi} \int_{\frac{4\pi}{3}}^{\frac{8\pi}{3}} \cos^2\left[\frac{\pi}{2}\nu\left(\frac{3}{4\pi}x-1\right)\right] \, \mathrm{d}x \\ &= \frac{2}{3} \int_0^1 \sin^2\left[\frac{\pi}{2}\nu(y)\right] \, \mathrm{d}y + \frac{4}{3} \int_0^1 \cos^2\left[\frac{\pi}{2}\nu(y)\right] \, \mathrm{d}y \\ &= \frac{2}{3}\left[1 + \int_0^1 \cos^2\left[\frac{\pi}{2}\nu(y)\right] \, \mathrm{d}y\right] \, . \end{aligned}$$

Wegen $\nu(x) + \nu(1-x) = 1$ gilt:

$$\begin{aligned} \int_0^1 \cos^2\left[\frac{\pi}{2}\nu(y)\right] \, \mathrm{d}y &= \int_0^{\frac{1}{2}} \cos^2\left[\frac{\pi}{2}\nu(y)\right] \, \mathrm{d}y + \int_0^{\frac{1}{2}} \cos^2\left\{\frac{\pi}{2}\left[1-\nu\left(\frac{1}{2}-y\right)\right]\right\} \, \mathrm{d}y \\ &= \int_0^{\frac{1}{2}} \cos^2\left[\frac{\pi}{2}\nu(y)\right] \, \mathrm{d}y + \int_0^{\frac{1}{2}} \sin^2\left[\frac{\pi}{2}\nu(y)\right] \, \mathrm{d}y = \frac{1}{2} \, . \end{aligned}$$

(b) ψ hat Mittelwert 0, da $\overline{\psi}(0) = 0$.

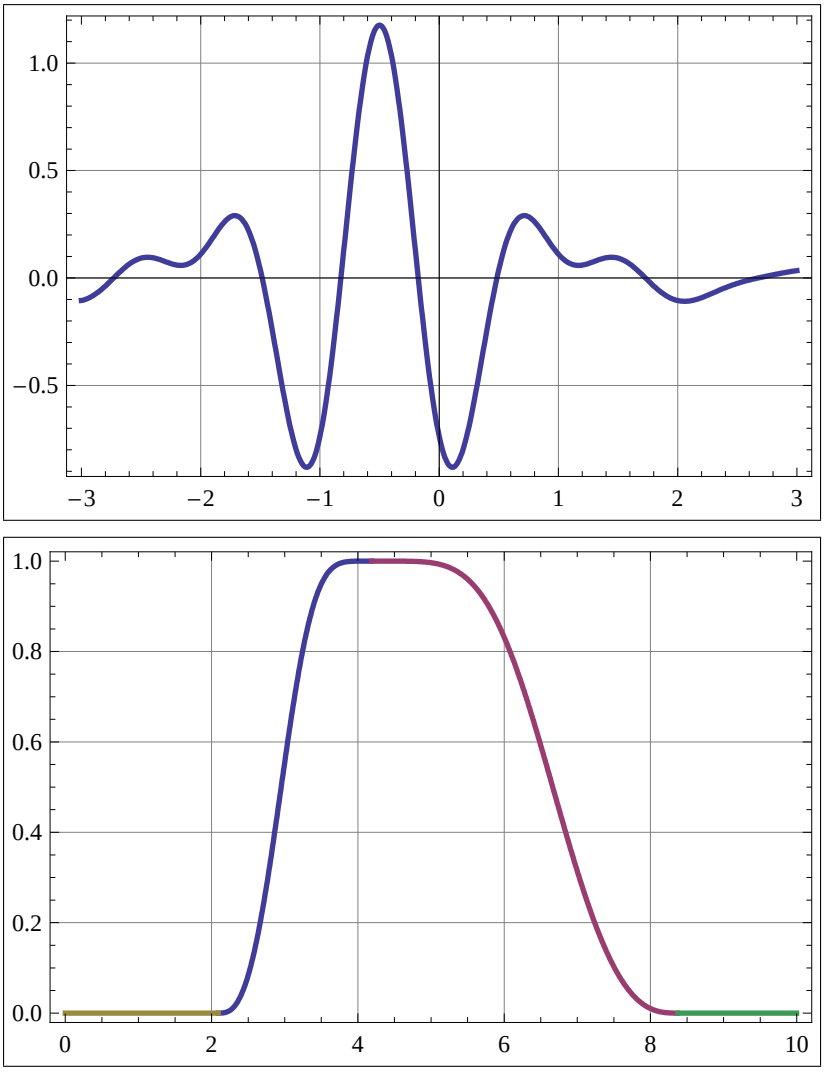


Abb. 3.4. Der Realteil des Meyer-Wavelets $\psi(x)$ (oben) sowie $\left|\overline{\psi}(\omega)\right|$ (unten).

Somit ist ψ nach Satz 2.1 in der Tat ein Wavelet. Das Meyer-Wavelet lässt sich analytisch nicht berechnen, natürlich aber numerisch. Abb. 3.4 zeigt den Realteil des (komplexen) Wavelets und den Betrag der Fourier-Transformierten.

Wir wissen schon aus Korollar 3.1, dass dieses Wavelet einen straffen Frame produziert, weil $\sum \left|\overline{\psi}(a_0^m\omega)\right|^2 > 0$ ist und $\overline{\psi}$ kompakten Träger hat. Das Korollar gibt uns auch den Wert der Frame-Schranken:

$$\begin{aligned} A = B &= \frac{2\pi}{t_0 \ln a_0} \int_0^\infty \frac{\left|\overline{\psi}(\omega)\right|^2}{\omega}\, d\omega \\ &= \frac{2\pi}{t_0 \ln a_0} \left\{ \frac{1}{2\pi} \int_{\frac{2\pi}{3}}^{\frac{4\pi}{3}} \frac{\sin^2\left[\frac{\pi}{2}\nu\left(\frac{3}{2\pi}\omega - 1\right)\right]}{\omega}\, \mathrm{d}\omega + \frac{1}{2\pi} \int_{\frac{4\pi}{3}}^{\frac{8\pi}{3}} \frac{\cos^2\left[\frac{\pi}{2}\nu\left(\frac{3}{4\pi}\omega - 1\right)\right]}{\omega}\, \mathrm{d}\omega \right\} \\ &= \frac{2\pi}{t_0 \ln a_0} \left\{ \frac{1}{2\pi} \int_0^1 \frac{\sin^2\left[\frac{\pi}{2}\nu(y)\right]}{y+1}\, \mathrm{d}y + \frac{1}{2\pi} \int_0^1 \frac{\cos^2\left[\frac{\pi}{2}\nu(y)\right]}{y+1}\, \mathrm{d}y \right\} \\ &= \frac{1}{t_0} \frac{\ln 2}{\ln a_0}\,. \end{aligned}$$

Wir sehen sofort, dass die Wahl $a_0 = 2$ und $t_0 = 1$ zu einem straffen Wavelet-Frame mit $A = B = 1$ führt, also zu einer Orthonormalbasis des Hilbertraums L^2. Korollar 3.1 hat sich somit als sehr hilfreich in der Klassifikation von Wavelet-Frames erwiesen! Wir werden etwas später eine Erweiterung dieser Aussage besprechen, die wichtige Konsequenzen hat für die praktische Umsetzung unseres Ziels eine beliebige Funktion $f \in L^2$ mit diskreten Wavelet-Netzwerken zu approximieren.

Beispiel 2*: Das Daubechies-Grossmann-Meyer-Wavelet
Das Daubechies-Grossmann-Meyer-Wavelet ist eine Verallgemeinerung des zuvor dargestellten Meyer-Wavelets. Seine Fourier-Transformierte $\overline{\psi} : \mathbb{R} \to \mathbb{R}$ ist nun definiert als

$$\overline{\psi}(\omega) := \sqrt{C} \begin{cases} \sin\left[\frac{\pi}{2}\nu\left(\frac{\omega-\omega'}{c_1\omega'-\omega'}\right)\right], & \omega' \le \omega \le c_1\omega', \\ \cos\left[\frac{\pi}{2}\nu\left(\frac{\omega-c_1\omega'}{c_1^2\omega'-\omega'}\right)\right], & c_1\omega' \le \omega \le c_1^2\omega', \\ 0, & \text{sonst} \end{cases}$$

mit

$$\omega' := \frac{2\pi}{(c_1^2 - c_2)}\,, \qquad c_1 > 1\,, c_2 > 0\,.$$

Das Meyer-Wavelet ist bis auf den Phasenfaktor $e^{i\omega/2}$ der Spezialfall $C = 1/2\pi$, $c_1 = 2$, $\omega' = 1$ des Daubechies-Grossmann-Meyer-Wavelets. $\overline{\psi}$ hat kompakten Träger $[\omega', c_1^2\omega']$ und somit ist wieder Korollar 3.1 anwendbar. Wir erhalten nach kurzer Rechnung analog zum Meyer-Wavelet:

$$A = B = \frac{2\pi C}{t_0} \frac{\ln c_1}{\ln a_0}\,.$$

Dieses Wavelet bildet also für $t_0 = 2\pi C$ und $a_0 = c_1$ eine orthonormale Basis des L^2.

Beispiel 3: Das Haar-Wavelet
Wir zeigen nun, dass das Haar-Wavelet weder (a) noch (b) erfüllt, aber trotzdem eine Orthonormalbasis von L^2 ist. Wir definieren wieder $\psi : \mathbb{R} \to [-1, 1]$

$$\psi_{\text{Haar}}(x) = \begin{cases} 1, & 0 \le x < 1/2, \\ -1, & 1/2 \le x < 1 \\ 0, & \text{sonst.} \end{cases}$$

ψ_{Haar} hat offensichtlich kompakten Träger, ist aber nicht differenzierbar an den Sprungstellen. Weiterhin gilt

$$\int_{-\infty}^{\infty} \psi_{\text{Haar}}(x)\,\mathrm{d}x = 0, \quad \text{und} \quad \int_{-\infty}^{\infty} |\psi_{\text{Haar}}|^2\,\mathrm{d}x = 1\,.$$

Die Fourier-Transformierte des Haar-Wavelets errechnet sich zu

$$\overline{\psi}_{\text{Haar}}(\omega) = \frac{i}{\sqrt{2\pi}} \frac{\sin^2(\omega/4)}{\omega/4}\, e^{-i\omega/2}\,. \tag{3.14}$$

Im "Zeitbereich", also bzgl. x, ist das Haar-Wavelet gut lokalisiert (es hat kompakten Träger), die Fourier-Transformierte hat aber keinen kompakten Träger. Trotzdem kann das Wavelet auch als relativ gut lokalisiert im "Frequenzbereich" angesehen werden (und zwar an ihrem ersten Maximum $\omega_0 \approx 4.66$), wie Abb. 3.5 illustriert. Versuchen wir aber wie für den Mexikanerhut die Gültigkeit von Lemma 3.8 für $\dot{\Psi}(a_0, t_0)$ mit $\psi = \psi_{\text{Haar}}$ als Mutterwavelet zu zeigen, so scheitern wir. Dies liegt offensichtlich daran, dass $\overline{\psi}_{\text{Haar}}(\omega)$ nicht schnell genug abklingt für $|\omega| > 1$, denn

$$\left|\overline{\psi}_{\text{Haar}}(\omega)\right| \le \frac{4}{\sqrt{2\pi}} \frac{1}{|\omega|}\,,$$

es gibt also kein $\alpha > 0$, so dass $\left|\overline{\psi}_{\text{Haar}}(\omega)\right| \le C|\omega|^{-1-\alpha}$ für $|\omega| > 1$.
Wir müssen also direkt auf Satz 3.4 zurückgreifen um zu zeigen, dass das Haar-Wavelet zunächst einen Frame bildet. Es ist für $\omega \in [1, a_0]$ und $a_0 > 1$

$$\sum_{m<0} a_0^{-2m}\omega^{-2} \sin^4\left(\frac{a_0^m\omega}{4}\right) = \sum_{m>0} \frac{a_0^{2m}}{\omega^2} \sin^4\left(\frac{\omega}{4a_0^m}\right) \le \sum_{m>0} \frac{\omega^2}{256\, a_0^{2m}} = \frac{\omega^2}{256(a_0^2-1)}\,,$$

wobei wir $\sin x < x$ für alle $x > 0$ ausgenutzt haben. Weiterhin:

$$\sum_{m\ge 0} a_0^{-2m}\omega^{-2} \sin^4\left(\frac{a_0^m\omega}{4}\right) \le \sum_{m\ge 0} \frac{1}{a_0^{2m}\omega^2} = \frac{a_0^2}{\omega^2(a_0^2-1)}\,,$$

so dass $\sup_{\omega\in[1,a_0]} \sum_{m\in\mathbb{Z}} \left|\overline{\psi}\,(a_0^m\omega)\right|^2 < \infty$. Die Bedingung $\inf_{\omega\in[1,a_0]} \sum [...] > 0$ ist ohnehin erfüllt, da "$= 0$" impliziert, dass $a_0^m - a^{m-1}$ konstant ist für alle $m \in \mathbb{Z}$, was aber

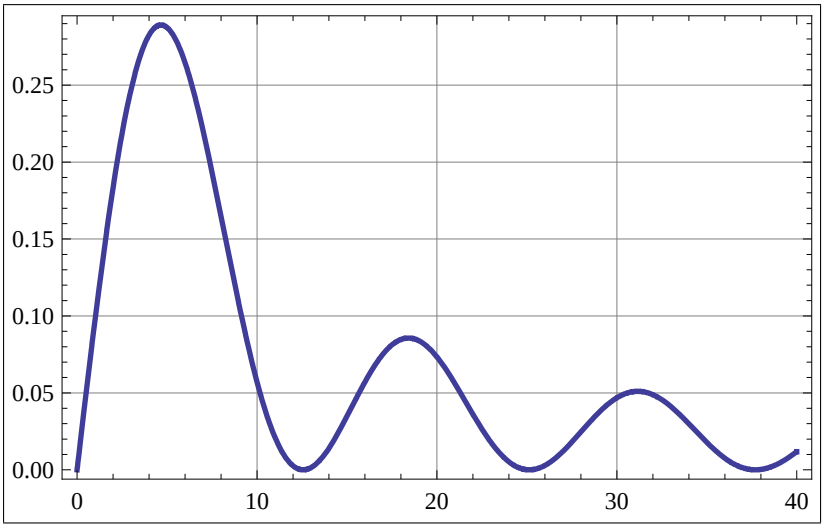


Abb. 3.5. $\left|\overline{\psi}_{\text{Haar}}(\omega)\right|$. Das erste Maximum liegt bei $\omega_0 \approx 4.66$, die weiteren Maxima nehmen für $\omega \to \infty$ wie $1/\omega$ ab.

wegen $a_0 > 1$ nicht zutreffen kann.
Voraussetzung (ii) an das Abklingverhalten von $\beta(|x|)$ lässt sich vermeintlich ebenso schnell zeigen. Es ist nämlich für $x > 0$

$$\sum_{m<0} \frac{\sin^2\left(\frac{a_0^m\omega}{4}\right)}{a_0^m\omega} \frac{\sin^2\left(\frac{a_0^m\omega+x}{4}\right)}{a_0^m\omega+x} \leq \sum_{m<0} \frac{a_0^m\omega}{16} \frac{\sin^2\left(\frac{a_0^m\omega+x}{4}\right)}{a_0^m\omega+x} \leq \frac{1}{x} \sum_{m<0} \frac{a_0^m\omega}{16} = \frac{\omega}{16(a_0-1)} \frac{1}{x}$$

und

$$\sum_{m\geq 0} \frac{\sin^2\left(\frac{a_0^m\omega}{4}\right)}{a_0^m\omega} \frac{\sin^2\left(\frac{a_0^m\omega+x}{4}\right)}{a_0^m\omega+x} \leq \sum_{m\geq 0} \frac{1}{a_0^m\omega(a_0^m\omega+x)} \leq \sum_{m\geq 0} \frac{1}{a_0^{2m}\omega^2} = \frac{a_0^2}{\omega^2(a_0^2-1)} \,,$$

unabhängig von $|x|$. Wir schließen, dass es eine Konstante K gibt, so dass $\beta(|x|) \leq K/|x|$ für alle $x \in \mathbb{R}\backslash\{0\}$. Aber was passiert für $a_0^m\omega = x$? Dieser Fall tritt immer ein, wie die folgende Überlegung zeigt: $a_0^m\omega = x$ impliziert $a_0^m = -x/\omega$. Da $a_0 > 1$ muss x also negativ sein und deshalb suchen wir ein $m \in \mathbb{Z}$, so dass $a_0^m = |x|/\omega$. Da $\omega \in [1, a_0]$ folgt $|x|/\omega \in [|x|/a_0, |x|]$, d.h. falls $a_0 \leq |x|$ gibt es in jedem Fall ein $m > 0$, so dass $a_0^m\omega + x = 0$. Da wir $x \in \mathbb{R}$ betrachten, divergiert $\beta(x)$ für ein geeignetes x. Weiterhin wird $x = -\frac{2\pi}{t_0}k$ mit $k \in \mathbb{N}$ beliebig groß, so dass $\sum_{k\neq 0} \beta\left(-2\pi k/t_0\right) \beta\left(2\pi k/t_0\right)$ stets divergiert.

Wir können also Satz 3.4 *nicht* auf das Haar-Wavelet anwenden!

Das heißt aber nicht, dass das Haar-Wavelet keinen Frame erzeugen kann. Um die Frame-Schranken zu berechnen formen wir die Reihe $\sum_{m\in\mathbb{Z}} \left|\overline{\psi}\left(a_0^m\omega\right)\right|^2$ zunächst unter zu Hilfenahme der Identität $\sin^4(x) = \frac{1}{8}[\cos(4x) - 4\cos(2x) + 3]$ um zu

$$\begin{aligned}\sum_{m\in\mathbb{Z}} a_0^{-2m}\omega^{-2}\sin^4\left(\frac{a_0^m\omega}{4}\right) &= \sum_{m\in\mathbb{Z}} a_0^{-2m}\omega^{-2}\frac{1}{8}\left[\cos(a_0^m\omega) - 4\cos\left(\frac{a_0^m\omega}{2}\right) + 3\right] \\ &= \frac{1}{8}\sum_{m\in\mathbb{Z}}\left\{\frac{\cos(a_0^m\omega)}{a_0^{2m}\omega^2} - 4\frac{\cos\left(a_0^m\omega/2\right)}{a_0^{2m}\omega^2} + 3a_0^{-2m}\omega^{-2}\right\} \\ &= \frac{1}{8}\sum_{m\in\mathbb{Z}}\left\{\frac{\cos(a_0^m\omega)}{(a_0^m\omega)^2} - \frac{\cos\left(a_0^m\omega/2\right)}{(a_0^m\omega/2)^2} + \frac{3}{(a_0^m\omega)^2}\right\}.\end{aligned}$$

Nun schließt sich eine Überlegung an, die in der Literatur nicht vorkommt, aber doch von ganz entscheidendem Charakter ist. Wir zeigen direkt für welches a_0 und t_0 das Haar-Wavelet überhaupt eine Basis des L^2 bilden kann:
Damit die ersten beiden Terme in der Summe eine Teleskopsumme der Form

$$\sum_{m=M}^{\infty}(a_m - a_{m-1}) = -a_M + \lim_{m\to\infty} a_m$$

bilden können (es sei $M \geq 0$ und $(a_m)_{m\in\mathbb{N}}$ eine Folge), muss $a_0^{m-1} = a_0^m/2$ für alle $m \in \mathbb{Z}$ gelten, d.h. es folgt sofort $a_0 = 2$!
Um nun den Wert der Reihe auszurechnen betrachten wir $m < 0$ und $m \geq 0$ getrennt und erhalten mit Hilfe der Teleskopsumme für $a_0 = 2$ im positiven Teil:

$$\begin{aligned}\frac{1}{8}\sum_{m\geq 0}\left\{\frac{\cos(2^m\omega)}{2^{2m}\omega^2} - \frac{\cos\left(2^{m-1}\omega\right)}{2^{2(m-1)}\omega^2} + \frac{3}{2^{2m}\omega^2}\right\} &= \frac{1}{8}\left(-\frac{\cos(\omega/2)}{2^{-2}\omega^2} + \lim_{m\to\infty}\frac{\cos(2^m\omega)}{2^{2m}\omega^2} + \frac{4}{\omega^2}\right) \\ &= -\frac{\cos(\omega/2)}{2\omega^2} + \frac{1}{2\omega^2}\end{aligned}$$

für alle $\omega \in [1, a_0]$. Nun zum negativen Teil:

$$
\begin{aligned}
&\frac{1}{8}\sum_{m<0}\left\{\frac{\cos(2^m\omega)}{2^{2m}\omega^2}-\frac{\cos\left(2^{m-1}\omega\right)}{2^{2(m-1)}\omega^2}+\frac{3}{2^{2m}\omega^2}\right\}\\
&\qquad=\frac{1}{8}\lim_{M\to\infty}\left\{\sum_{m=-M}^{-1}\left[\frac{\cos(2^m\omega)}{2^{2m}\omega^2}-\frac{\cos\left(2^{m-1}\omega\right)}{2^{2(m-1)}\omega^2}+\frac{3}{2^{2m}\omega^2}\right]\right\}\\
&\qquad=\frac{1}{8}\lim_{M\to\infty}\left\{-\frac{\cos(2^{-M-1}\omega)}{2^{2(-M-1)}\omega^2}+\frac{4^{M+1}-4}{\omega^2}\right\}+\frac{\cos(\omega/2)}{2\omega^2}\\
&\qquad=\frac{1}{8\omega^2}\lim_{M\to\infty}\left\{2^{2(M+1)}\left(1-\cos\left(\frac{\omega}{2^{M+1}}\right)\right)-4\right\}+\frac{\cos(\omega/2)}{2\omega^2}\\
&\qquad=\frac{1}{16}-\frac{1}{2\omega^2}+\frac{\cos(\omega/2)}{2\omega^2}\,.
\end{aligned}
$$

Zusammen erhalten wir also für $a_0 = 2$ (unabhängig von ω !) und mit dem zusätzlichen Vorfaktor $4/\sqrt{2\pi}$ aus Glg. (3.14)

$$\sum_{m\in\mathbb{Z}}\left|\overline{\psi}\left(a_0^m\omega\right)\right|^2=\frac{1}{2\pi}\,.$$

Somit gilt offensichtlich

$$\inf_{\omega\in[1,a_0]}\sum_{m\in\mathbb{Z}}\left|\overline{\psi}\left(a_0^m\omega\right)\right|^2=\sup_{\omega\in[1,a_0]}\sum_{m\in\mathbb{Z}}\left|\overline{\psi}\left(a_0^m\omega\right)\right|^2=\frac{1}{2\pi}\,,$$

d.h. wir erhalten für die Frame-Schranken sofort die Abschätzung

$$A\le\frac{1}{t_0}\le B\,.$$

Unser Ziel, $A = B = 1$, kann somit nur mit $t_0 = 1$ erreicht werden!
Wir konzentrieren uns also auf die Frage, ob die Familie

$$\psi_{m,n}(x)=2^{-m/2}\,\psi_{\text{Haar}}\left(2^{-m}x-n\right)\,,\qquad m,n\in\mathbb{Z}\,,\tag{3.15}$$

eine Orthonormalbasis des L^2 bildet. Diese Frage hat eine klare Antwort: Ja.

Beweis. Die Tatsache, dass das Haar-Wavelet eine Basis des L^2 bildet, genauer, dass jedes $f \in L^2(\mathbb{R})$ beliebig genau durch eine *endliche* Linearkombination von Haar-Wavelets dargestellt werden kann, beruht im Wesentlichen darauf, dass die Treppenfunktionen dicht liegen im (halbnormierten) Raum $L^p(\mathbb{R})$ (Beweis s. Werner (2006) Seite 252). Nun müssen wir nur noch die Haar-Wavelets als spezielle Treppenfunktionen auffassen und die Einteilung immer feiner machen. Jetzt gilt es nur noch zu zeigen, dass es für jede "Skala" eine Teilmenge der Familie $(\psi_{m,n})$ existiert, die eine Basis dieses Unterraums bildet. Diese Technik ist wie schon erwähnt bekannt als Multiskalen-Analyse (s. Daubechies (1992)).

3.2 Endliche Rekonstruktionen und quantisierte Phasenräume

In den allgemeinen Frame-Betrachtungen haben wir bislang keine speziellen Aussagen über die Dimension des zu Grunde liegenden Hilbertraumes gemacht. In der Frame-Definition 3.2 zum Beispiel nahmen wir die Familie aus Frame-Vektoren als R-dimensional an (zunächst abzählbar, aber auch überabzählbar wie in Anmerkung 3.3 erwähnt). In der numerischen Umsetzung müssen wir uns allerdings offensichtlich immer auf *endliche* Frames zurückziehen, d.h. $R \subsetneq \mathbb{Z}^2$ endlich. So arbeiten die im Rahmen dieser Dissertation praktisch umgesetzten Wavelet Neuronalen-Netze (s. Kapitel 6) immer mit einer endlichen Teilmenge der gewählten Familie $\dot{\Psi}(a_0, t_0)$. Eine perfekte Rekonstruktion jedes Signals $f \in L^2$ wird uns mit dieser Einschränkung natürlich niemals gelingen. Die Frage ist aber: Wie wählt man die Frame-Vektoren aus, so dass die Rekonstruktion immer noch möglichst gut ist? In Teil II werden wir mit statistischen Methoden an diese Frage herangehen, da wir nichts anderes vor uns haben als ein Standard-Problem der Statistik, und zwar das der Regressorauswahl.

Bevor wir einen allgemeinen Satz formulieren, der unsere soeben gestellen Fragen beantworten wird, stellen wir einige heuristische Überlegungen an. Seien zum Beispiel $|\psi|$ und $\left|\overline{\psi}\right|$ symmetrisch wie im Falle des Mexikanerhut-Wavelets. Weiterhin sei ψ normalisiert, also

$$\int |\psi(x)|^2 \, \mathrm{d}x = 1 \quad \text{und}$$
$$\int x|\psi(x)|^2 \, \mathrm{d}x = 0 \, .$$

Letztere Bedingung macht eine Aussage über den "Erwartungswert" des Mutter-Wavelets, also wo es zentriert ist. Man rechnet leicht nach, dass dies für das Mexikanerhut-Wavelet wie wir es bislang behandelt haben der Fall ist. Auf jeden Fall ist ψ nun zentriert bezüglich 0 in der x-Koordinate (wir nennen sie nun meist "Zeit", da wir gerade in technischen Anwendungen oft Signale betrachten, in denen x die Zeitkoordinate darstellt) und bezüglich $\pm\omega_0$ im Frequenzbereich, wobei

$$\omega_0 = \int_0^\infty \omega \, \left|\overline{\psi}(\omega)\right|^2 \, \mathrm{d}\omega \Big/ \int_0^\infty \left|\overline{\psi}(\omega)\right|^2 \, \mathrm{d}\omega \, .$$

Die durch Dilation und Translation von ψ entstandenen $\psi_{m,n} = a_0^{-m/2}\psi\left(a_0^{-m}x - nt_0\right)$ sind entsprechend lokalisiert um $a_0^m n t_0$ in der "Zeit" und um $\pm a_0^{-m}\omega_0$ im Frequenzbereich, denn:

$$\begin{aligned}\int_{-\infty}^{\infty} x\ |\psi_{m,n}(x)|^2\ \mathrm{d}x &= a_0^{-m} \int_{-\infty}^{\infty} x\left|\psi\left(a_0^{-m}x - nt_0\right)\right|^2\ \mathrm{d}x \\ &= \int_{-\infty}^{\infty} (a_0^m y + a_0^m nt_0)\,|\psi(y)|^2\ \mathrm{d}y \\ &= a_0^m \int_{-\infty}^{\infty} y\ |\psi(y)|^2\ \mathrm{d}y + a_0^m nt_0 \int_{-\infty}^{\infty} |\psi(y)|^2\ \mathrm{d}y \\ &= a_0^m nt_0\end{aligned}$$

und

$$\int_0^{\infty} \omega\ \left|\overline{\psi_{m,n}}(\omega)\right|^2\ \mathrm{d}\omega = a_0^{-m} \int_0^{\infty} \omega\left|\overline{\psi}(\omega)\right|^2\ \mathrm{d}\omega = a_0^{-m}\omega_0\ ,$$

falls wir zusätzlich o.B.d.A.

$$\int_{-0}^{\infty} \left|\overline{\psi}(\omega)\right|^2\ \mathrm{d}\omega = 1$$

fordern. Wir drücken den hier beschriebenen Sachverhalt noch etwas anders aus: Das Paar (m_0, n_0) entspricht einer Zeit x_0 und einer Frequenz ω_0. Wünschenswert ist es nun, dass die Eigenschaften der Funktion f in einem Frequenz- bzw. Zeitintervall um x_0 bzw. ω_0 möglichst gut durch die Koeffizienten $\langle f, \psi_{m,n}\rangle$ modelliert werden, wobei (m, n) nahe an (m_0, n_0) liegen sollen. Man könnte sagen, die Größe $\langle f, \psi_{m,n}\rangle$ repräsentiert den Informationsgehalt von f in der Nähe der Zeit $a_0^m nt_0$ und in der Nähe der Frequenzen $\pm a_0^{-m}\omega_0$.
Angenommen f selbst ist gut lokalisiert in $0 < T < \infty$ im Zeitraum und in $0 < \Omega_0 < \Omega_1 < \infty$ im Frequenzraum[6], d.h.

$$\begin{aligned}&\int_{x\in[-T,T]} |f(x)|^2\ \mathrm{d}x \geq (1-\delta)\,\|f\|^2 \quad \text{und} \\ &\int_{|\omega|\in[\Omega_0,\Omega_1]} \left|\overline{f}(\omega)\right|^2\ \mathrm{d}\omega \geq (1-\delta)\,\|f\|^2\end{aligned}$$

mit $\delta > 0$ klein. Entscheidend ist nun die folgende Überlegung: Sind die einzelnen Wavelets $\psi_{m,n}$ gut lokalisiert im Zeit-Frequenz-Raum (Phasenraum) um den Punkt $(a_0^m nt_0, a_0^{-m}\omega_0)$, dann wird das Skalarprodukt $\langle f, \psi_{m,n}\rangle$ klein, falls der Abstand im Phasenraum von $(a_0^m nt_0, a_0^{-m}\omega_0)$ zu dem Bereich $[-T, T] \times ([-\Omega_1, \Omega_0] \cup [\Omega_0, \Omega_1])$ groß ist. Wir schließen also, dass nur solche $m, n \in \mathbb{Z}$ zur Rekonstruktion von f herangezogen werden müssen, für die

$$(a_0^m nt_0, \pm a_0^{-m}\omega_0) \subset [-T, T] \times ([-\Omega_1, \Omega_0] \cup [\Omega_0, \Omega_1])$$

[6] Natürlich kann dies nie perfekt der Fall sein, da es keine Funktion gibt mit kompaktem Träger im Zeit- als auch im Frequenzraum.

oder die zumindest nahe an dieser Menge liegen.
Der folgende Satz inklusive Beweisidee stammt aus Daubechies (1990) und fasst unsere Überlegungen zusammen. Er stellt die endliche Version der diskreten Rekonstruktionsformel 3.3 dar:

Satz 3.5 (Endliche Rekonstruktionsformel). *Angenommen die Familie* $(\psi_{m,n})$, $(m,n) \in \mathbb{Z}^2$, *mit* $\psi_{m,n} = a_0^{-m/2}\psi\left(a_0^{-m}x - nt_0\right)$ *bildet einen Frame von* $L^2(\mathbb{R})$ *mit den Frame-Schranken* $0 < A \le B < \infty$. *Weiterhin gibt es* $\alpha > 1$, $\beta > 0$, $\gamma > 1$, *so dass*

$$|\psi(x)| \le \frac{C}{(1+x^2)^{\alpha/2}} \quad \textit{und}$$
$$\left|\overline{\psi}(\omega)\right| \le \frac{C\,|\omega|^{\beta}}{(1+\omega^2)^{(\beta+\gamma)/2}}\,.$$

Dann existiert für alle $\varepsilon > 0$ *eine endliche Menge* $B_\varepsilon = B_\varepsilon(\Omega_0, \Omega_1, T) \subset \mathbb{Z}^2$, $T > 0$, $0 < \Omega_0 < \Omega_1$, *so dass für alle* $f \in L^2(\mathbb{R})$

$$\left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n)\; \mathscr{T}^{-1}\psi_{m,n} \right\| \le \sqrt{\frac{B}{A}} \Bigg[\left(\int_{|\omega|\notin[\Omega_0,\Omega_1]} \left|\overline{f}(\omega)\right|^2 \,\mathrm{d}\omega \right)^{1/2} + \left(\int_{x\notin[-T,T]} |f(x)|^2 \,\mathrm{d}x \right)^{1/2} + \varepsilon\|f\| \Bigg]\,.$$

Beweis. Wir starten mit der Definition von B_ε in der folgenden Weise (Abb. 3.6 verdeutlicht dies):

$$B_\varepsilon(\Omega_0,\Omega_1,T) := \left\{(m,n)\in\mathbb{Z}^2;\; m_0 \le m \le m_1,\; |nt_0| \le a_0^{-m}T + h\right\}\,.$$

Die Bedeutung von m_0, m_1 und h ergibt sich im Fortgang des Beweises. Wir schätzen nun folgendermaßen ab:

$$\begin{aligned}
\left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n)\; \mathscr{T}^{-1}\psi_{m,n} \right\| &= \sup_{\|v\|=1} \left| \langle f, v\rangle - \sum_{(m,n)\in B_\varepsilon} \langle f, \psi_{m,n}\rangle\langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle \right| \\
&= \sup_{\|v\|=1} \left| \sum_{(m,n)\notin B_\varepsilon} \langle f, \psi_{m,n}\rangle\langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle \right| \\
&\le \sup_{\|v\|=1} \sum_{m\notin[m_0,m_1]} \sum_{n\in\mathbb{Z}} \left[|\langle P_{\Omega_0\Omega_1} f, \psi_{m,n}\rangle| + |\langle (1-P_{\Omega_0\Omega_1}) f, \psi_{m,n}\rangle| \right] \left|\langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle\right| \\
&+ \sup_{\|v\|=1} \sum_{m\in[m_0,m_1]} \sum_{|nt_0|>a_0^{-m}T+h} \left[|\langle Q_T f, \psi_{m,n}\rangle| + |\langle (1-Q_T) f, \psi_{m,n}\rangle| \right] \left|\langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle\right|,
\end{aligned}$$

wobei wir die folgenden zwei Projektionsoperatoren eingeführt haben:

$$(Q_T f)(x) = \begin{cases} f(x), & |x| \le T, \\ 0, & \text{sonst} \end{cases}$$

und

$$\overline{(P_{\Omega_0\Omega_1} f)}(\omega) = \begin{cases} \overline{f}(\omega), & \Omega_0 \le |\omega| \le \Omega_1, \\ 0, & \text{sonst.} \end{cases}$$

Nach Voraussetzung bilden die $\psi_{m,n}$ einen Frame mit Frame-Schranken A, B und die $\mathscr{T}^{-1}\psi_{m,n}$ den dualen Frame mit Schranken B^{-1}, A^{-1}, also:

$$A\|x\|^2 \le \sum_{(m,n)\in\mathbb{Z}^2} |\langle x, \psi_{m,n}\rangle|^2 \le B\|x\|^2 \quad \text{und}$$
$$B^{-1}\|x\|^2 \le \sum_{(m,n)\in\mathbb{Z}^2} \left|\langle x, \mathscr{T}^{-1}\psi_{m,n}\rangle\right|^2 \le A^{-1}\|x\|^2$$

für alle $x \in L^2(\mathbb{R})$. Hiermit schätzen wir ab:

$$\begin{aligned} &\sum_{m\notin[m_0,m_1]} \sum_{n\in\mathbb{Z}} |\langle (1-P_{\Omega_0\Omega_1})f, \psi_{m,n}\rangle| \left|\langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle\right| \\ &\qquad \le \left(\sum_{(m,n)\in\mathbb{Z}^2} |\langle (1-P_{\Omega_0\Omega_1})f, \psi_{m,n}\rangle|^2\right)^{\frac{1}{2}} \left(\sum_{(m,n)\in\mathbb{Z}^2} \left|\langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle\right|^2\right)^{\frac{1}{2}} \\ &\qquad \le \sqrt{\frac{B}{A}}\, \|(1-P_{\Omega_0\Omega_1})f\|\ \|v\| = \sqrt{\frac{B}{A}} \left(\int_{|\omega|\notin[\Omega_0,\Omega_1]} \left|\overline{f}(\omega)\right|^2\, \mathrm{d}\omega\right)^{\frac{1}{2}}, \end{aligned}$$

weil $\|v\| = 1$. Völlig analog ergibt sich:

$$\sum_{m\in[m_0,m_1]} \sum_{|nt_0|>a_0^{-m}T+h} |\langle (1-Q_T)f, \psi_{m,n}\rangle| \left|\langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle\right| \le \sqrt{\frac{B}{A}} \left(\int_{x\notin[-T,T]} |f(x)|^2\, \mathrm{d}x\right)^{\frac{1}{2}}.$$

Zurück zu der initialen Abschätzung zu Beginn des Beweises. Wir haben nun:

$$\begin{aligned}
\left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n)\, \mathscr{T}^{-1}\psi_{m,n} \right\| &= \sup_{\|v\|=1} \left| \langle f, v\rangle - \sum_{(m,n)\in B_\varepsilon} \langle f, \psi_{m,n}\rangle \langle \mathscr{T}^{-1}\psi_{m,n}, v\rangle \right| \\
&\le \sup_{\|v\|=1} \sum_{m\notin[m_0,m_1]} \sum_{n\in\mathbb{Z}} |\langle P_{\Omega_0\Omega_1} f, \psi_{m,n}\rangle| + \sup_{\|v\|=1} \sum_{m\in[m_0,m_1]} \sum_{|nt_0|>a_0^{-m}T+h} |\langle Q_T f, \psi_{m,n}\rangle| \\
&\quad + \sqrt{\frac{B}{A}} \left[\left(\int_{x\notin[-T,T]} |f(x)|^2 \,\mathrm{d}x \right)^{\frac{1}{2}} + \left(\int_{|\omega|\notin[\Omega_0,\Omega_1]} \left|\overline{f}(\omega)\right|^2 \,\mathrm{d}\omega \right)^{\frac{1}{2}} \right] \\
&\le \left[\left(\sum_{m\notin[m_0,m_1]} \sum_{n\in\mathbb{Z}} |\langle P_{\Omega_0\Omega_1} f, \psi_{m,n}\rangle|^2 \right)^{\frac{1}{2}} + \left(\sum_{m\in[m_0,m_1]} \sum_{|nt_0|>a_0^{-m}T+h} |\langle Q_T f, \psi_{m,n}\rangle|^2 \right)^{\frac{1}{2}} \right. \\
&\quad \left. + \left(\int_{x\notin[-T,T]} |f(x)|^2 \,\mathrm{d}x \right)^{\frac{1}{2}} + \left(\int_{|\omega|\notin[\Omega_0,\Omega_1]} \left|\overline{f}(\omega)\right|^2 \,\mathrm{d}\omega \right)^{\frac{1}{2}} \right] \sqrt{\frac{B}{A}}\,.
\end{aligned}$$

Zu zeigen ist also, dass die beiden ersten Terme in diesem Ausdruck für geeignete m_0, m_1 und h kleiner werden als $\varepsilon^2\|f\|^2/4$. Der erste Term kann mit Hilfe desselben Tricks abgeschätzt werden, den wir auch schon im Beweis von Satz 3.4 angewendet haben:

$$\begin{aligned}
\sum_{m\notin[m_0,m_1]} \sum_{n\in\mathbb{Z}} |\langle P_{\Omega_0\Omega_1} f, \psi_{m,n}\rangle|^2 &\le \frac{2\pi}{t_0} \sum_{m\notin[m_0,m_1]} \sum_{l\in\mathbb{Z}} \int\limits_{\substack{|\omega|\in[\Omega_0,\Omega_1] \\ \left|\omega - 2\pi l a_0^{-m} t_0^{-1}\right|\in[\Omega_0,\Omega_1]}} \left|\overline{f}(\omega)\right| \\
&\qquad \cdot \left|\overline{f}\left(\omega - a_0^{-m}\frac{2\pi}{t_0}l\right)\right| \left|\overline{\psi}\left(a_0^m\omega - \frac{2\pi}{t_0}l\right)\right| \left|\overline{\psi}\left(a_0^m\omega\right)\right| \,\mathrm{d}\omega \\
&\le \frac{2\pi}{t_0} \sum_{l\in\mathbb{Z}} \left[\int\limits_{\substack{|\eta|\in[\Omega_0,\Omega_1] \\ \left|\eta - 2\pi l a_0^{-m} t_0^{-1}\right|\in[\Omega_0,\Omega_1]}} \left|\overline{f}(\eta)\right|^2 \sum_{m\notin[m_0,m_1]} \left|\overline{\psi}\left(a_0^m\eta - \frac{2\pi}{t_0}l\right)\right|^\lambda \left|\overline{\psi}\left(a_0^m\eta\right)\right|^{2-\lambda} \right]^{\frac{1}{2}} \mathrm{d}\eta \\
&\qquad \cdot \left[\int\limits_{\substack{|\zeta|\in[\Omega_0,\Omega_1] \\ \left|\zeta - 2\pi l a_0^{-m} t_0^{-1}\right|\in[\Omega_0,\Omega_1]}} \left|\overline{f}(\zeta)\right|^2 \sum_{m\notin[m_0,m_1]} \left|\overline{\psi}\left(a_0^m\zeta - \frac{2\pi}{t_0}l\right)\right|^\lambda \left|\overline{\psi}\left(a_0^m\zeta\right)\right|^{2-\lambda} \right]^{\frac{1}{2}} \mathrm{d}\zeta\,.
\end{aligned}$$

λ werden wir gleich festlegen anhand von Konvergenzüberlegungen. Laut Voraussetzung können wir das Produkt der beiden Fourier-Transformierten von ψ abschätzen durch:

$$\begin{aligned}
\left|\overline{\psi}\left(a_0^m\eta\right)\right| \left|\overline{\psi}\left(a_0^m\eta - \frac{2\pi}{t_0}l\right)\right| &\le C^2 \left[1 + \left(a_0^m\eta\right)^2\right]^{-\frac{\gamma}{2}} \left[1 + \left(a_0^m\eta - \frac{2\pi}{t_0}l\right)^2\right]^{-\frac{\gamma}{2}} \\
&\le C_1 \left(1 + l^2\right)^{-\frac{\gamma}{2}}\,.
\end{aligned}$$

Wieder eingesetzt erhalten wir:

$$\sum_{m\notin[m_0,m_1]}\sum_{n\in\mathbb{Z}}|\langle P_{\Omega_0\Omega_1}f,\psi_{m,n}\rangle|^2 \le \frac{2\pi}{t_0}C_2\,\|P_{\Omega_0\Omega_1}f\|^2\sum_{l\in\mathbb{Z}}\left(1+l^2\right)^{-\frac{\lambda\gamma}{2}} \cdot \sup_{|\eta|\in[\Omega_0,\Omega_1]}\sum_{m\notin[m_0,m_1]}\left|\overline{\psi}\left(a_0^m\eta\right)\right|^{2(1-\lambda)}.$$

Die Summe über l konvergiert nur, wenn $\lambda > 1/\gamma$, d.h. wir wählen $\lambda = \frac{1}{2}(1+\gamma^{-1})$. Außerdem können wir für $|\eta| \in [\Omega_0, \Omega_1]$ abschätzen:

$$\sum_{m>m_1}\left|\overline{\psi}\left(a_0^m\eta\right)\right|^{2(1-\lambda)} \le C_3\sum_{m>m_1}\left(1+a_0^{2m}\Omega_0^2\right)^{-\gamma(1-\lambda)} \le C_4\Omega_0^{-2\gamma(1-\lambda)}a_0^{-2m_1\gamma(1-\lambda)} \quad \text{und}$$
$$\sum_{m<m_0}\left|\overline{\psi}\left(a_0^m\eta\right)\right|^{2(1-\lambda)} \le C_5\sum_{m<m_0}\left(a_0^m\Omega_1\right)^{2\beta(1-\lambda)} \le C_6\Omega_1^{2\beta(1-\lambda)}a_0^{2m_0\beta(1-\lambda)}.$$

Die Konstanten C_1 bis C_6 sind hierbei unabhängig von Ω_0, Ω_1, m_0 und m_1, aber trotzdem recht schwierig zu bestimmen. Wir setzen zurück ein (inklusive der speziellen Wahl für λ):

$$\sum_{m\notin[m_0,m_1]}\sum_{n\in\mathbb{Z}}|\langle P_{\Omega_0\Omega_1}f,\psi_{m,n}\rangle|^2 \le C_7\|f\|^2\left[\left(\Omega_0a_0^{m_1}\right)^{-(\gamma-1)}+\left(a_0^{m_0}\Omega_1\right)^{\beta(\gamma-1)/\gamma}\right].$$

Wir haben nun die Chance m_0 und m_1 festzulegen. Wählen wir nämlich

$$m_0 \le \frac{\frac{\gamma}{\beta(\gamma-1)}\ln\left(\frac{\varepsilon^2}{4C_7}\right)-\ln\Omega_1}{\ln a_0} = \frac{2\gamma}{\beta(\gamma-1)}\frac{\ln\varepsilon}{\ln a_0}-\frac{\ln\Omega_1}{\ln a_0}-\frac{\gamma}{\beta(\gamma-1)}\frac{\ln(4C_7)}{\ln a_0} \quad \text{und}$$
$$m_1 \ge \frac{\frac{1}{\gamma-1}\ln\left(\frac{4C_7}{\varepsilon^2}\right)-\ln\Omega_0}{\ln a_0} = -\frac{2}{\gamma-1}\frac{\ln\varepsilon}{\ln a_0}-\frac{\ln\Omega_0}{\ln a_0}+\frac{1}{\gamma-1}\frac{\ln(4C_7)}{\ln a_0},$$

so wird

$$\sum_{m\notin[m_0,m_1]}\sum_{n\in\mathbb{Z}}|\langle P_{\Omega_0\Omega_1}f,\psi_{m,n}\rangle|^2 \le \frac{\varepsilon^2}{4}\|f\|^2.$$

Für $\varepsilon \to 0$ wird das zur Rekonstruktion benötigte Fenster $[m_0, m_1]$ somit immer größer. Die Konstante C_7 ist für die in diesem Satz behandelten Wavelets schwer konkret zu bestimmen. Wir werden genauere Kontanten für schneller abfallende Wavelets in Lemma 3.11 erarbeiten, so dass wir in der Praxis m_0 und m_1 aus a_0, t_0 aus den Abklingparametern des Mutter-Wavelets numerisch leicht berechnen können.
Wir schätzen nun noch den verbliebenen Teil der anfänglichen Summe ab und erhalten am Ende eine untere Schranke für den Parameter h:

$$\sum_{|nt_0|>a_0^{-m}T+h} |\langle Q_T f, \psi_{m,n}\rangle|^2 \leq \sum_{|nt_0|>a_0^{-m}T+h} \|f\|^2 \|Q_T\psi_{m,n}\|^2$$
$$\leq \|f\|^2 \int_{x\in[-T,T]} a_0^{-m} \sum_{|nt_0|>a_0^{-m}T+h} |\psi(a_0^m x - nt_0)|^2 \, \mathrm{d}x$$

Wir spalten die Summe auf und erhalten zunächst mit $|a_0^m x - nt_0| = nt_0 - a_0^{-m}x \geq (n-n_1)t_0 + h + a_0^{-m}(T-x)$, wobei n_1 die kleinste Zahl größer als $t_0^{-1}\left(a_0^{-m}T+h\right)$ sei, und der Voraussetzung für das Abklingverhalten von $\psi(x)$:

$$\begin{aligned} & a_0^{-m} \int_{x\in[-T,T]} \sum_{|nt_0|>a_0^{-m}T+h} |\psi(a_0^m x - nt_0)|^2 \, \mathrm{d}x \\ & \leq a_0^{-m} \int_{x\in[-T,T]} \sum_{n=n_1}^{\infty} C_8 \left[1 + \left(h + (n-n_1)t_0 + a_0^{-m}(T-x)\right)^2\right]^{-\alpha} \mathrm{d}x \\ & \leq C_9 \sum_{l=0}^{\infty} \left[1 + (h + l\, t_0)^2\right]^{-\alpha} \leq C_{10}\, h^{-2\alpha} \,. \end{aligned}$$

Die Abschätzung für den Teil der Summe mit $n < -t_0^{-1}\left(a_0^{-m}T+h\right) < 0$ geschieht ganz analog, wobei n_2 die größte Zahl kleiner als $-t_0^{-1}\left(a_0^{-m}T+h\right)$ sei:

$$\begin{aligned} & a_0^{-m} \int_{x\in[-T,T]} \sum_{|nt_0|<-a_0^{-m}T+h} |\psi(a_0^m x - nt_0)|^2 \, \mathrm{d}x \\ & \leq a_0^{-m} \int_{x\in[-T,T]} \sum_{n=-\infty}^{n_2} C_8 \left[1 + \left(h + (n-n_1)t_0 + a_0^{-m}(T-x)\right)^2\right]^{-\alpha} \mathrm{d}x \\ & \leq C_{10}\, h^{-2\alpha} \,. \end{aligned}$$

Insgesamt erhalten wir also:

$$\sum_{m\in[m_0,m_1]} \sum_{|nt_0|>a_0^{-m}T+h} |\langle Q_T f, \psi_{m,n}\rangle|^2 \leq 2C_{10}(m_1 - m_1 + 1)\, h^{-2\alpha}\|f\|^2 \,.$$

Damit dieser Ausdruck kleiner als $\varepsilon^2\|f\|^2/4$ wird, müssen wir also nur

$$h \geq \left(\frac{8C_{10}(m_1 - m_0 + 1)}{\varepsilon^2}\right)^{1/(2\alpha)}$$

wählen. ■

Zusammengefasst heißt dieser Satz, dass falls das Signal f genügend gut konzentriert ist im Phasenraum, dann reicht für die Approximation von f eine endliche Anzahl von Wavelets aus, falls das Mutter-Wavelet selbst genügend gut konzentriert ist. Wir wünschen

uns also möglichst gut lokalisierte Wavelets! In Abb. 3.6 ist dieser Sachverhalt anschaulich dargestellt. Die Zeichnung zeigt den Zeit-Frequenz-Raum und das in diesem Raum auf zwei Rechtecke lokalisierte f. Ebenso ist die endliche Menge an Wavelet-Zentren B_ε eingezeichnet, die zur Rekonstruktion von f ausreicht. Es zeigt sich sofort, dass je größer

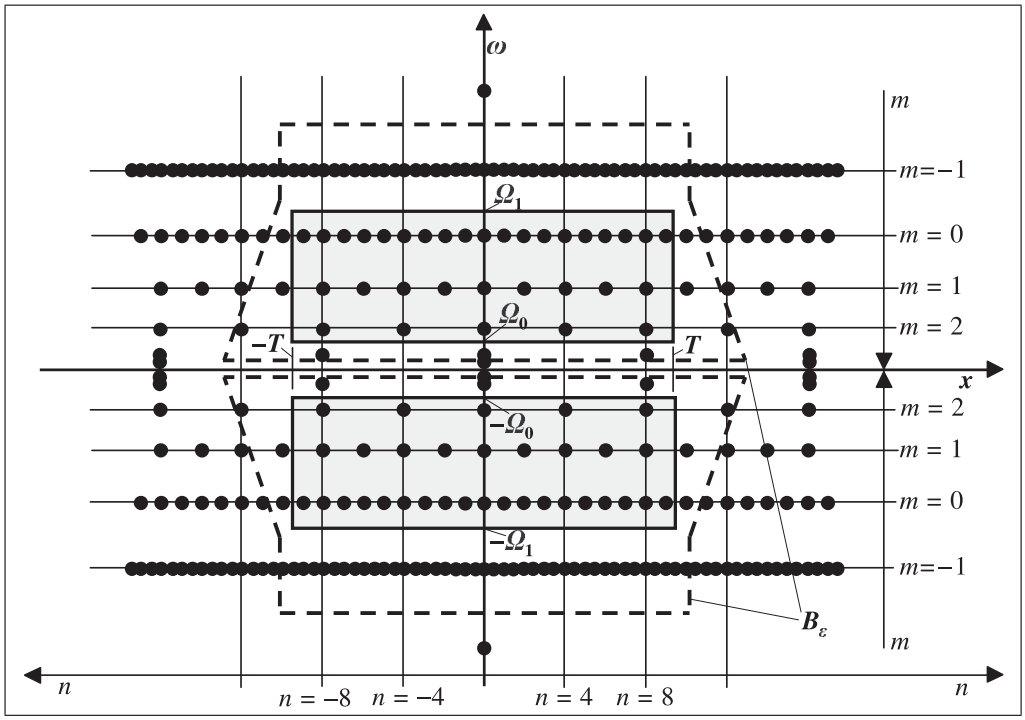


Abb. 3.6. Die Punkte in diesem Diagramm haben die Koordinaten $(a_0^m n t_0, \pm a_0^{-m}\omega_0)$, wobei $a_0 = 2$, $t_0 = 1$ und $\omega_0 = 1$. Jeder Punkt symbolisiert das Zentrum eines Wavelets aus der Familie $(\psi_{m,n})$. Die beiden grau unterlegten Rechtecke stellen die Region im Zeit-Frequenz-Raum dar, in dem das Signal f konzentriert ist. Die gestrichelten Boxen stehen für B_ε, d.h. die Punkte innerhalb dieser Region reichen aus um f bis zu einem Fehler von ε zu rekonstruieren. Diese Form der Darstellung von B_ε wurde angelehnt an die Darstellung in Daubechies (1990).

α, β, γ werden, desto näher kommen m_0, m_1, h an die folgenden Werte heran:

$$
\begin{aligned}
m_0 &\leq -\frac{\ln \Omega_1}{\ln a_0}\,, \\
m_1 &\geq -\frac{\ln \Omega_0}{\ln a_0}\,, \\
h &\geq 1\,.
\end{aligned}
$$

Für exponentiell abfallende Wavelets, wie z.B. das Mexikanerhut-Wavelet, können die Konstanten α, β und γ sehr groß gewählt werden, so dass wir für diese Wavelets m_0, m_1, h nahe an diesen Schranken erwarten. Das Beispiel weiter unten bestätigt dies.

Wir schätzen nun die Anzahl der in B_ε enthaltenen Punkte folgendermaßen ab: Für jedes $m \in [m_0, m_1]$ in B_ε gibt es zwei Mal (in jede Richtung, am besten verdeutlicht sich der Leser dies wiederum an Abb. 3.6) $t_0^{-1}\left(a_0^{-m}T + h\right)$ Punkte, die in der Menge liegen, also:

$$
\begin{aligned}
|B_\varepsilon(\Omega_0, \Omega_1, T)| &= 2\sum_{m=m_0}^{m_1} \frac{a_0^{-m}T + h}{t_0} = \frac{2}{t_0}\left[\frac{T}{a_0 - 1}\left(a_0^{1-m_0} - a_0^{-m_1}\right) + h\ (m_1 - m_0 + 1)\right] \\
&= \frac{2}{t_0}\Bigg\{\frac{T}{a_0 - 1}\left[a_0\Omega_1\left(\frac{\varepsilon^2}{4C_7}\right)^{-\frac{\gamma}{\beta(\gamma-1)}} - \Omega_0\left(\frac{\varepsilon^2}{4C_7}\right)^{\frac{1}{\gamma-1}}\right] \\
&\qquad - \frac{h}{\ln a_0}\left[\frac{\beta+\gamma}{\beta(\gamma-1)}\ln\left(\frac{\varepsilon^2}{4C_7}\right) + \ln\left(\frac{\Omega_0}{\Omega_1}\right)\right]\Bigg\}\,.
\end{aligned}
$$

Man sieht schnell, dass dieser Ausdruck gegen unendlich strebt, falls wir die Abmessungen der Box einzeln gegen unendlich gehen lassen, also $\Omega_1 \to \infty$, $T \to \infty$. Ebenso enthält B_ε auch unendlich viele Punkte bzw. Wavelets, wenn $\Omega_0 \to 0$. Je "unkonzentrierter" die Funktion f ist, desto mehr Wavelets brauchen wir zur Rekonstruktion, was nicht sehr überraschend. Interessanter ist die Betrachtung $\varepsilon \to 0$. Wir erhalten:

$$
\begin{aligned}
|B_\varepsilon(\Omega_0, \Omega_1, T)| &\propto \frac{2Ta_0}{t_0(a_0-1)}\frac{\Omega_1}{(4C_7)^{-\frac{\gamma}{\beta(\gamma-1)}}}\varepsilon^{-\frac{\gamma}{\beta(\gamma-1)}} - \frac{h}{\ln a_0}\frac{\beta+\gamma}{\beta(\gamma-1)}\ln\left(\frac{\varepsilon^2}{4C_7}\right) \\
&\xrightarrow{\varepsilon\to 0} \infty\,.
\end{aligned}
$$

Für unendliche Rekonstruktions-Präzision benötigen wir also unendlich viele Wavelets, bzw. Rekonstruktions-Koeffizienten $\langle f, \psi_{m,n}\rangle$!
Die Funktion f ist aber konzentriert in $[-T, T] \times ([-\Omega_1, \Omega_0] \cup [\Omega_0, \Omega_1])$. Diese Menge füllt eine Fläche von $4T(\Omega_1 - \Omega_0)$ im Phasenraum aus. Wir definieren konsequent eine *Phasenraumdichte* durch

$$
\rho = \lim_{\substack{T,\Omega_1\to\infty \\ \Omega_0\to 0}} \frac{|B_\varepsilon(\Omega_0, \Omega_1, T)|}{4T(\Omega_1 - \Omega_0)}\,.
$$

Diese Größe lässt sich verstehen, wenn wir die einzelnen $\psi_{m,n}$ als "Zustände" abgeleitet vom Mutter-Wavelet auffassen (wir verwenden diesen Begriff in Anlehnung an die Anwendung von Wavelets in der Quantenphysik. In der Quantenphysik spannt eine linear unabhängige Menge von Zuständen einen dichten Unterraum von $L^2(\mathbb{R})$ auf (man nennt diese Menge von Zuständen dann ein vollständiges Ensemble). Diese Zustände nehmen ein gewisses Volumen im Phasenraum ein. Ein Beispiel für einen Frame, der eine diskrete Phasenraumstruktur aufweist ist der so genannte *Weyl-Heisenberg-Frame*[7]:

$$g_{m,n}(x) = e^{imp_0 x} g(x - nq_0)\,,$$

wobei als Mutter-Funktion oft die Gauss-Funktion $g(x) = \pi^{-1/4}\exp(-x^2/2)$ genutzt wird. Wir sehen, dass für diesen Frame die Dilation als eine Art Fourier-Transformation der Ausgangsfunktion definiert wird. Es lässt sich nun zeigen, dass für ein in dieser Art hergestelltes Gitter die Phasenraumdichte[8]

$$\rho = \lim_{T,\Omega\to\infty} \frac{|B_\varepsilon(\Omega, T)|}{4T\Omega} = \frac{1}{p_0\, q_0}\,,$$

also *unabhängig* von ε! Es zeigt sich weiter, dass für alle Frames $\{g_{m,n} : m, n \in \mathbb{Z}\}$ stets $p_0 q_0 \leq 2\pi$ sei muss (notwendige Bedingung), d.h. die "optimale", d.h. kleinste Phasenraumdichte des Weyl-Heisenberg-Frames ist $1/(2\pi)$. Der Phasenraum im Weyl-Heisenberg-Fall ist also "*gequantelt*", weil wir eine Zusatzanforderung an p_0, q_0 stellen, und zwar $p_0 q_0 \leq 2\pi$ (wir sehen wieder die Analogie zur Quantenmechanik). Diese "magische" bzw. optimale Dichte ist nichts anderes als die *Nyquist Dichte* (s. Slepian & Pollak (1962)), die schon aus dem Shannon'schen Abtasttheorem bekannt ist, oft unter dem Namen Abtastdichte. Abhängig von der Wahl der Parameter p_0, q_0 wird der Phasenraum aber in unserem Fall überrepräsentiert, was daran liegt, dass die $g_{m,n}$ i.A. nicht orthonormal sind.

Für den Weyl-Heisenberg-Frame scheint also der Begriff des Phasenraums gut geeignet zu sein um z.B. den Grad der Redundanz des Frames und die Approximationsfähigkeit des Frames festzulegen. Nun aber zurück zu unseren Wavelet-Frames. Hier gibt es keine Nebenbedingungen für a_0 und t_0 (außer $a_0 > 1$, $t_0 > 0$) und wir haben gezeigt, dass prinzipiell ein Frame aus *jedem* Paar a_0, t_0 konstruiert werden kann. Dies liegt hauptsächlich daran, dass wenn $\{\psi_{m,n} : m, n \in \mathbb{Z}\}$ ein Frame ist mit Parametern a_0, t_0, dann ist auch $\{\psi_{\gamma;m,n} : m, n \in \mathbb{Z}\}$ mit $\psi_\gamma = \sqrt{\gamma}\psi(\gamma x)$ als Mutter-Wavelet ein Frame. Um diese "pathologischen" Frame-Konstruktionen auszuschließen können wir z.B. fordern, dass $C_\psi = 1$. Doch es ist trotzdem nicht möglich (a_0, t_0)-Paare bzw. ganze Regionen in der (a_0, t_0)-Ebene auszuschließen wie das folgende Gegenbeispiel zeigt:

Lemma 3.10. *Sei ψ das Meyer-Wavelet. Dann existiert ein $\varepsilon > 0$, so dass $\{\psi_{m,n} : m, n \in \mathbb{Z}\}$ mit*

[7] Er wird auch gefensterter Fourier-Fame genannt.
[8] Für diesen Frame kann stets $\Omega_0 = 0$ gewählt werden, bzw. wir benötigen nur ein Ω.

$$\psi_{m,n}(x) = 2^{-m/2}\psi(2^{-m}x - nt_0)$$

eine Basis von $L^2(\mathbb{R})$ *bildet, und zwar für jede Wahl* $t_0 \in (1-\varepsilon, 1+\varepsilon)$.

Beweis. Siehe Daubechies (1990), Appendix C.

Den Fall $\varepsilon = 0$ haben wir schon zuvor behandelt. Diese kleine Erweiterung sieht harmlos aus, aber sie zeigt, dass es keine Kurve in der (a_0, t_0)-Ebene geben *kann*, die den Bereich trennt in dem Frames möglich sind und den Bereich, in dem Frames nicht möglich sind. Denn wir wissen, dass die Familie für $t_0 = 1$ eine Basis bildet, d.h. schon "optimal" ist. D.h. für $t_0 > 1$ dürfte es keine Frames mehr geben (den Zusammenhang zwischen t_0 und der Redundanz des erzeugten Frames haben wir für das Beispiel des Mexikanerhut-Wavelets durchgerechnet, siehe auch Abb. 3.3). Für $t_0 < 1$ hingegen müsste der Frame redundanter werden. Das eben zitierte Lemma zeigt, dass diese Überlegungen nicht richtig sind. Wir schließen also, dass es für Wavelets keine Kurven gibt, die Frame- von nicht-Frame Regionen abgrenzen, so dass die Kurven selbst die Orthonormalbasen sind. Im Fall der Weyl-Heisenberg-Frames ist dies grundlegend anders. So trennt für diese Frames die Hyperbel $p_0 q_0 = 2\pi$ die Region, für die Frames unmöglich sind von der Region, in der straffe Frames möglich sind. Auf der Hyperbel selbst gibt es zwar Frames, aber nur mit schlechter Zeit-Frequenz-Lokalisation. Die Größe $p_0 q_0$ interpretieren wir somit als Zeit-Frequenz-Dichte ρ.
Im Wavelet-Fall gibt es allerdings wie soeben argumentiert keine Einschränkung für die $a_0 t_0$, bzw. für jedes Paar (a_0, t_0) lässt sich ein straffer Frame mit gute Zeit-Frequenz-Lokalisation konstruieren (Multiskalen-Analyse), d.h. wir erwarten für diese Größe eine Abhängigkeit von der Approximationsgenauigkeit ε. In der Tat ergibt sich im Wavelet-Fall:

$$\rho = \rho(\varepsilon) = \lim_{\substack{T,\Omega_1 \to \infty \\ \Omega_0 \to 0}} \frac{|B_\varepsilon(\Omega_0, \Omega_1, T)|}{4T(\Omega_1 - \Omega_0)} = \frac{a_0}{2t_0(a_0 - 1)} (4C_7)^{\frac{\gamma}{\beta(\gamma-1)}} \, \varepsilon^{-\frac{2\gamma}{\beta(\gamma-1)}} \, .$$

Dieser Ausdruck hängt also, weil $\frac{2\gamma}{\beta(\gamma-1)} > 0$, wie ε^{-c} mit einer positiven Konstante c von der Approximationsgenaugigkeit ab! Anders ausgedrückt: Je genauer ich f rekonstruieren möchte, desto *größer* wird die Phasenraumdichte der benötigten Wavelets.

Der Phasenraum ist für Wavelets also nicht wie im Weyl-Heisenberg-Fall "gequantelt"!

Der Grund warum wir trotzdem Wavelets als Basisfunktionen verwenden ist der, dass wir auf theoretischer Seite durch die Multiskalen-Anaylse abgesichtert sind, d.h. die Existenz von straffen Frames mit guter Zeit-Frequenz-Lokalisation ist sichergestellt, auch wenn der Phasenraum nicht quantisiert ist. Andererseits lässt sich auch zeigen, dass Wavelet-Basen für eine größere Klasse von Funktionen möglich sind als nur $L^2(\mathbb{R})$, was im Falle der gefensterten Fourier-Basen nur bedingt der Fall ist (s. Daubechies (1992)).

3.3 Numerische Präzisierung der endlichen Rekonstruktionsformel

Die endliche Rekonstruktionsformel betrachtet algebraische Wavelets und sie liefert relativ grobe Abschätzungen. Die Schranken für m_0 und m_1 sind relativ nutzlos für schneller abfallende Wavelets, da hier z.B. γ beliebig groß gewählt werden kann. Wir liefern nun in gewisser Hinsicht eine Präzisierung der endlichen Rekonstruktionsformel für *exponentiell* abfallende Wavelets, wie z.B. das Mexikanerhut-Wavelet. In diesem Fall lassen sich die Rekonstruktionsparameter m_0, m_1 und h für ein konkretes Mutter-Wavelet relativ leicht numerisch bestimmen. Als Beispiel nehmen wir wieder das in unseren Betrachtungen zentrale Mexikanerhut-Wavelet. Dieses Wavelet zeigt sowohl im Zeit- als auch im Frequenzraum ein exponentielles Abklingen (sogar wie e^{-x^2}). Wir formulieren:

Lemma 3.11. *Die Familie $(\psi_{m,n})_{(m,n)\in\mathbb{Z}^2}$ mit $\psi_{m,n} = a_0^{-m/2}\psi\left(a_0^{-m}x - nt_0\right)$ bilde wieder einen Frame von $L^2(\mathbb{R})$ mit den Frame-Schranken $0 < A \le B < \infty$. Weiterhin gibt es Konstanten C, $\alpha > 0$, $\beta > 0$ und $\gamma > 1$ so dass*

$$|\psi(x)| \le C\, e^{-\beta|x|}\,,$$
$$\left|\overline{\psi}(\omega)\right| \le C \begin{cases} |\omega|^{\gamma}, & |\omega| \le 1, \\ e^{-\alpha(|\omega|-1)}, & |\omega| > 1. \end{cases}$$

$\varepsilon > 0$ sei fest gewählt. Es sei weiterhin $1 < \Omega_0 < \Omega_1$ und

$$\chi = \chi(\varepsilon; t_0; C, \alpha) := \frac{e^{-\alpha}}{8\pi C^2}\, t_0\, \tanh\left[\frac{\alpha\,\pi}{2t_0}\right]\, \varepsilon^2\,.$$

Dann erreichen wir die Approximationsabschätzung von Satz 3.5 für

$$m_0 := \left\lceil -1 + \frac{\ln\left(\frac{a_0^{\gamma}-1}{\Omega_1^{\gamma}}\chi\right)}{\gamma \ln a_0} \right\rceil \le m \le m_1\,,$$

wobei m_1 festgelegt wird durch die Lösung von

$$m_1 = \max_{m^*\in\mathbb{Z}} \left\{ \sum_{m=m^*+1}^{\infty} e^{-\alpha(a_0^m \Omega_0)} \right\}$$

unter den Nebenbedingungen

$$\sum_{m=m^*+1}^{\infty} e^{-\alpha(a_0^m \Omega_0)} \le e^{-\alpha}\chi \quad \textit{und} \quad m_1 > m_0 \in \mathbb{Z}\,,$$

und

$$h = \left\lfloor -\frac{1}{2\beta} \ln\left(\frac{1-a_0}{8C^2} \frac{e^{2\beta t_0}}{e^{2\beta t_0}-1} \frac{\varepsilon^2}{a_0^{-m_1} - a_0^{1-m_0}} \right) \right\rfloor\,.$$

Beweis. Wir schließen an den Beweis der endlichen Rekonstruktionsformel an und notieren für alle $\omega \in [\Omega_0, \Omega_1]$, $a_0 > 1$, $m \in \mathbb{Z}$

$$\left|\overline{\psi}\left(a_0^m \omega\right)\right| \left|\overline{\psi}\left(a_0^m \omega - \frac{2\pi}{t_0} l\right)\right| \leq C^2 e^{2\alpha} e^{-\alpha \frac{2\pi}{t_0}} ,$$

so dass

$$\sum_{l \in \mathbb{Z}} \left|\overline{\psi}\left(a_0^m \omega\right)\right| \left|\overline{\psi}\left(a_0^m \omega - \frac{2\pi}{t_0} l\right)\right| \leq C^2 e^{2\alpha} \left(\frac{2}{e^{\alpha \frac{2\pi}{t_0}} - 1} + 1\right) = \frac{C^2 e^{2\alpha}}{\tanh\left(\frac{\alpha\,\pi}{t_0}\right)} .$$

Für den "sup"-Teil, also $\sup_{|\omega| \in [\Omega_0, \Omega_1]} \sum_{m \notin [m_0, m_1]} \left|\overline{\psi}\left(a_0^m \omega\right)\right|^{2(1-\lambda)}$, wählen wir o.B.d.A. $\lambda = 0.5$ und schätzen folgendermaßen ab:

$$\begin{aligned}
\sum_{m < m_0} \left|\overline{\psi}\left(a_0^m \omega\right)\right| &\leq C \sum_{m < m_0} \left|a_0^m \Omega_1\right|^{\gamma} = C\, \frac{a_0^{(m_0+1)\gamma}}{a_0^{\gamma} - 1} \Omega_1^{\gamma} , \\
\sum_{m > m_1} \left|\overline{\psi}\left(a_0^m \omega\right)\right| &\leq C\, e^{\alpha} \sum_{m > m_1} e^{-\alpha\left(a_0^m \Omega_0\right)} .
\end{aligned}$$

Wir setzen nun in den abzuschätzenden Ausdruck aus dem Beweis von Satz 3.5 ein:

$$\begin{aligned}
&\sum_{m \notin [m_0, m_1]} \sum_{n \in \mathbb{Z}} \left|\langle P_{\Omega_0 \Omega_1} f, \psi_{m,n} \rangle\right|^2 \\
&\leq \frac{2\pi}{t_0} \frac{C^2 e^{\alpha}}{\tanh\left(\frac{\alpha\,\pi}{2t_0}\right)} \|f\|^2 \left[\frac{a_0^{(m_0+1)\gamma}}{a_0^{\gamma} - 1} \Omega_1^{\gamma} + e^{\alpha} \sum_{m > m_1} e^{-\alpha\left(a_0^m \Omega_0\right)}\right] .
\end{aligned}$$

Nun ist nur noch der erste und zweite Term auf der rechten Seite separat mit $\varepsilon^2/4$ zu vergleichen und wir erhalten die Behauptung für m_0 und m_1.
Der Beweis für h verläuft ganz analog zu Satz 3.5 mit einigen Modifikationen. Es ist mit $|x| \leq T$

$$\begin{aligned}
\sum_{|nt_0| > a_0^{-m} T + h} \left|\psi(a_0^m x - n t_0)\right|^2 &\leq \sum_{n > n_1} \left|\psi\left[(n - n_1) t_0 + h + a_0^{-m}(T - x)\right]\right|^2 \\
&\quad + \sum_{n < n_2} \left|\psi\left[(n_2 - n) t_0 + h + a_0^{-m}(T - x)\right]\right|^2 \\
&\leq 2C^2 e^{-2\beta h} e^{-2\beta a_0^{-m}(T - x)} \sum_{l=0}^{\infty} e^{-2\beta t_0 l} \\
&\leq 2C^2 e^{-2\beta h} \frac{e^{2\beta t_0}}{e^{2\beta t_0} - 1} ,
\end{aligned}$$

so dass

$$\sum_{m\in[m_0,m_1]}\sum_{|nt_0|>a_0^{-m}T+h}|\langle Q_Tf,\psi_{m,n}\rangle|^2\le 2C^2e^{-2\beta h}\frac{e^{2\beta t_0}}{e^{2\beta t_0}-1}\frac{a_0^{-m_1}-a_0^{1-m_0}}{1-a_0}\,\|f\|^2\,.$$

Mit der Bedingung $\le\varepsilon^2\|f\|^2/4$ erhalten wir dann sofort die Behauptung. ■

Mit diesem Lemma lässt sich B_ε für in Zeit- und Frequenzraum mindestens exponentiell abfallende Wavelets recht genau approximieren. Wir sehen wie in den Sätzen zuvor, dass wir um Divergenz der Reihen $\sum|\psi(a_0^m\omega)|^\lambda$ für $m\to-\infty$ zu vermeiden die Einschränkung $|\psi(\omega)|\le C|\omega|^\gamma$ einbauen müssen.

Es ist an der Zeit eine Beispielrechnung durchzuführen. Das Mexikanerhut-Wavelet entspricht den Forderungen an $|\psi|$ und $|\overline{\psi}|$ für z.B. $C=1.02$, $\alpha=0.75$, $\beta=1$ und $\gamma=2$. Wir wählen weiterhin $\varepsilon=0.01$, $\Omega_0=10$, $\Omega_1=1000$ und $T=100$. Dann erhalten wir mit den Formeln aus dem vorangegangenen Lemma

$$m_0=-20\,,\qquad m_1=0\,,\qquad h=7\,.$$

Es ergibt sich:

$$|B_\varepsilon(\Omega_0,\Omega_1,T)|=2\sum_{m=m_0}^{m_1}\frac{a_0^{-m}T+h}{t_0}\approx 4.2\times10^8\,.$$

Wir benötigen also gut 400 Millionen Wavelets zur Rekonstruktion des auf das Interval $[-100,100]\times[-1000,-10]\cup[10,1000]\times$ konzentrierten Signals! Lemma 3.11 ist im Falle des Mexikanerhut-Wavelets immer noch eine recht grobe Abschätzung, da dieses Wavelets sogar wie e^{-x^2} abfällt. Es zeigt sich durch exakte Berechnung der auftretenden Summen (für eine Diskussion der numerischen Details wie z.B. Berechnung der Reihen siehe Beispiel 1 in Anschluss an Korollar 3.1) für $\varepsilon=0.01$, dass $m_0=-13$, $m_1=-7$ sowie $h=5$. Mit diesen Werten ergibt sich

$$|B_\varepsilon(\Omega_0,\Omega_1,T)|\approx 3.3\times10^6\,,$$

also um einen Faktor von 100 weniger als wir mit dem Lemma abgeschätzt hatten, aber immer noch sehr viele. An diesem Punkt sind wir an dem numerischen Wert von $|B_\varepsilon(\Omega_0,\Omega_1,T)|$ allerdings nur sekundär interessiert. Es geht vielmehr um eine qualitative Betrachtung des Verhaltens dieser Größe in Abhängigkeit von den an der Rechnung beteiligten Parametern, besonders den Frame-Parametern a_0 und t_0.
Wir sehen, dass $|B_\varepsilon|$ besonders stark von m_0 abhängt, kaum aber von m_1 und t. Abb. 3.7 zeigt die Abhängigkeit von m_0 von den anderen Parametern. Interessant ist die Beobachtung, dass die obere Schranke für $\varepsilon=0.01$ als $m_1=0$ gewählt werden kann, gehen wir über zu $\varepsilon=0.001$, so muss auch der Term $m=1$ hinzugenommen werden

für die Rekonstruktion. Dies ist aber nun ausreichend bis zu $\varepsilon = 10^{-6}$. Um Maschinengenauigkeit zu erreichen ist in diesem Beispiel $m_1 = 3$ nötig, was aber die Anzahl der zur Rekonstruktion benötigten Wavelets $|B_\varepsilon|$ praktisch nicht beeinflusst. Abb. 3.7 zeigt weiterhin, dass $|B_\varepsilon|$ mit steigendem a_0 immer kleiner wird, für $a_0 = 10$ zum Beispiel enthält diese Menge nur noch ungefähr 180 Millionen Wavelets. Dieser Zusammenhang ist auch nicht überraschend, denn die Redundanz des Frames nimmt mit steigendem a_0 ab, weil sich die einzelnen Mitglieder der Wavelet-Familie $(\psi_{m,n})$ mehr und mehr voneinander unterscheiden. Genau dasselbe gilt auch für den Frame-Parameter t_0. Die Straffheit des Frames hingegen nimmt mit wachsendem a_0, t_0 ebenfalls ab, d.h. wir entfernen uns immer weiter von einer Basis, die Güte der Approximation sinkt. Diese Tatsache findet sich in Satz 3.5 in dem Vorfaktor $\sqrt{B/A}$ wieder. Es gilt also das Optimum in a_0 und t_0 zu finden zwischen guter Approximation (kleines $\sqrt{B/A}$) und möglichst wenigen Rekonstruktions-Wavelets (kleines $|B_\varepsilon|$). Abb. 3.8 zeigt den Funktionsverlauf von $\sqrt{B/A}\,|B_\varepsilon|$ in Abhängigkeit von a_0 und t_0. Für das gewählte Beispiel liegt das Minimum bei $a_0 \approx 2.618$ bzw. $t_0 \approx 2.183$. Da a_0 und t_0 unkorreliert sind, kann die Optimierung für beide Parameter einzeln durchgeführt werden. Die so gewählten a_0 und t_0 sind also am besten geeignet zur Rekonstruktion des Signals f.

Mit dem folgenden Lemma präzisieren wir die endliche Rekonstruktionsformel für technische Anwendungen, in denen beides zusammenkommt: Einerseits wird f durch einen endlichen Frame rekonstruiert, was einen Fehler gemäß Satz 3.5 in der Rekonstruktion von f zur Folge hat. Andererseits müssen die Elemente des dualen Frames ebenfalls approximiert werden. Hierfür erinnern wir uns an Teil (*iii*) der diskreten Rekonstruktionsformel, also

$$\mathscr{T}^{-1}\psi_{m,n} \approx \frac{2}{A+B}\sum_{k=0}^{N}\left(\mathbb{1} - \frac{2}{A+B}\mathscr{T}\right)^k \psi_{m,n}\,,$$

und an den iterativen Algorithmus (3.5) zur praktischen Umsetzung dieser Approximation. Es ergibt sich dann folgendes Lemma für den Gesamtfehler der Approximation:

Lemma 3.12. *Unter denselben Voraussetzungen wie in Satz 3.5 gilt für jede Funktion* $f \in L^2(\mathbb{R})$

$$\left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n)\left(\mathscr{T}^{-1}\psi_{m,n}\right)^N \right\| \le \left(\frac{B-A}{B+A}\right)^{N+1}\|f\| + C\left[1 - \left(\frac{B-A}{B+A}\right)^{N+1}\right],$$

wobei die Konstante C dieselbe ist wie schon in Satz 3.5:

$$C := \sqrt{\frac{B}{A}}\left[\left(\int_{|\omega|\notin[\Omega_0,\Omega_1]} \left|\overline{f}(\omega)\right|^2 \, d\omega\right)^{1/2} + \left(\int_{x\notin[-T,T]} |f(x)|^2 \, dx\right)^{1/2} + \varepsilon\|f\|\right].$$

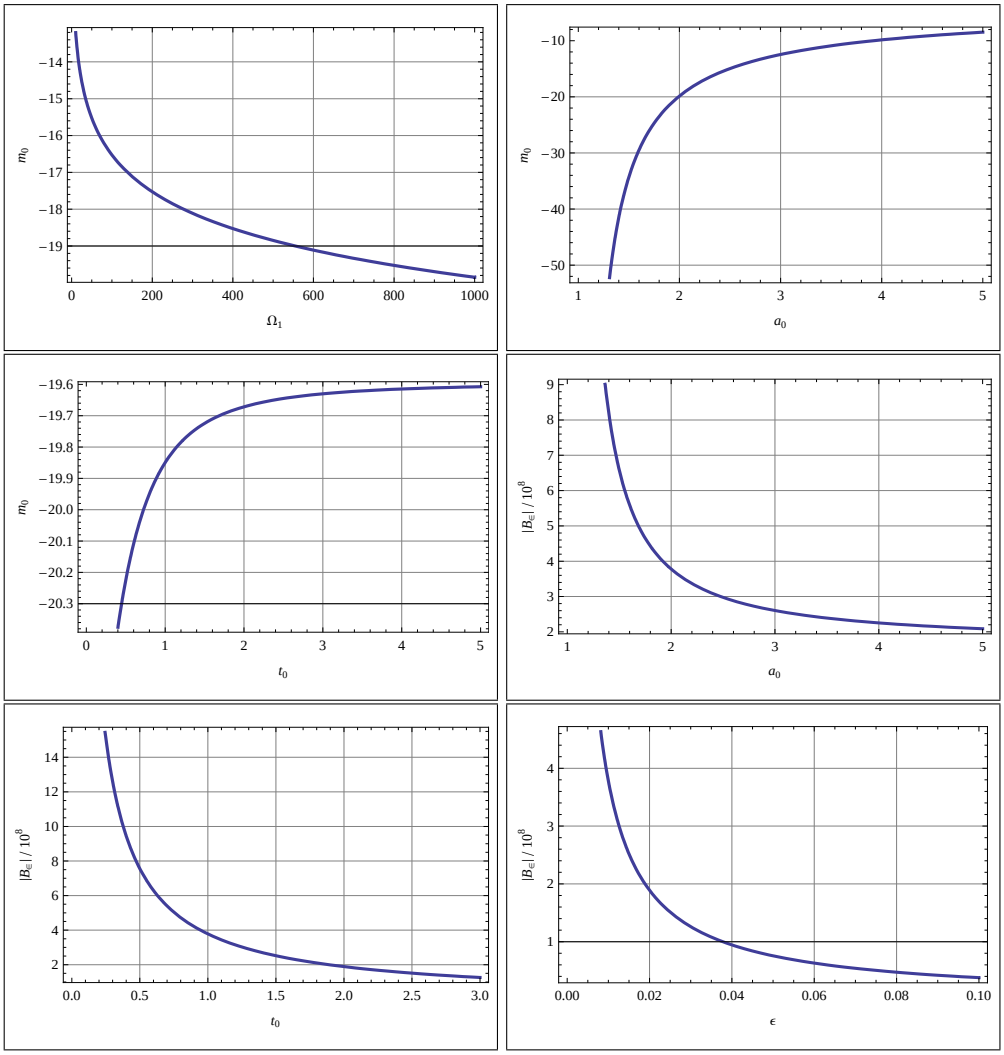


Abb. 3.7. Die ersten drei Zeichnungen zeigen m_0 aus Lemma 3.11 in Abhängigkeit von der Signal-Bandbreite Ω_1 und den Frame-Parametern a_0, t_0. Die Bandbreite im Zeitraum T spielt für m_0 keine Rolle und wurde festgelegt auf $T = 100$. Alle anderen Parameter sind festgelegt wie im Text angegeben. Die vierte und fünfte Zeichnung zeigt die Anzahl der zur Rekonstruktion eines Signals f benötigten Wavelets in Abhängigkeit von dem Frame-Parameter a_0 für $t_0 = 1$ bzw. in Abhängigkeit von t_0 für $a_0 = 2$. Die letzte Zeichnung zeigt diese Größe in Abhängigkeit von der Approximationsgenauigkeit ε.

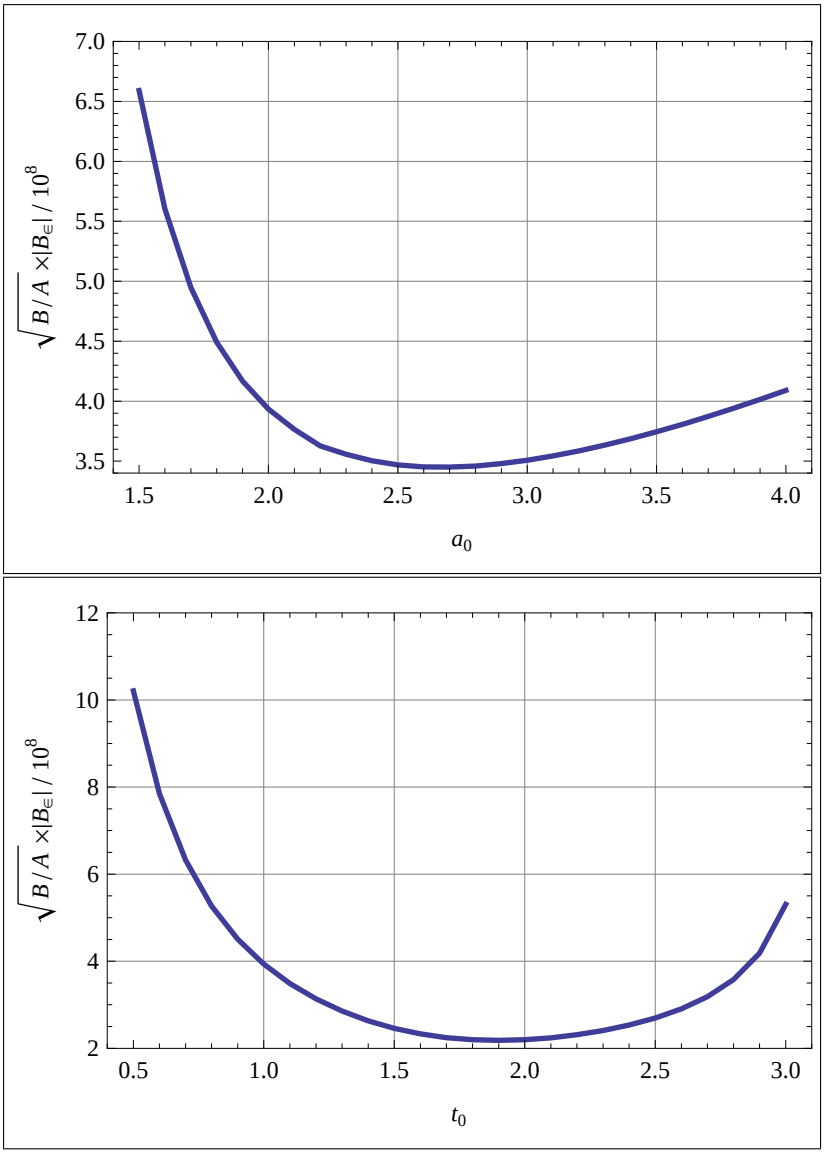


Abb. 3.8. Abhängigkeit der Größe $\sqrt{B/A}\ |B_\varepsilon|$ von den Frame-Parametern a_0 und t_0. In der ersten Zeichnung wurde $t_0 = 1$ gewählt, in der Zweiten $a_0 = 2$. Die restlichen Parameter sind dieselben wie in Abb. 3.7 bzw. wie im Text angegeben.

Beweis. Es gilt mit der Definition des Fehleroperators E aus Glg. (3.4)

$$\begin{aligned}
\left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n) \left(\mathscr{T}^{-1}\psi_{m,n}\right)^N \right\| &= \left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n) \left(\mathbb{1} - E^{N+1}\right) \mathscr{T}^{-1}\psi_{m,n} \right\| \\
&\leq \left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n) \mathscr{T}^{-1}\psi_{m,n} \right\| + \left\|E^{N+1}\right\| \left\| \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n) \mathscr{T}^{-1}\psi_{m,n} \right\| \\
&\leq C + \left\|E^{N+1}\right\| \left(\|f\| - \left\| f - \sum_{(m,n)\in B_\varepsilon} \mathcal{W}_\psi f(m,n) \mathscr{T}^{-1}\psi_{m,n} \right\| \right) \\
&= \left\|E^{N+1}\right\| \|f\| + C\left(1 - \left\|E^{N+1}\right\|\right) .
\end{aligned}$$

Mit der Abschätzung

$$E \leq \frac{B-A}{B+A}\,\mathbb{1}$$

aus Glg. (3.4) ergibt sich die Behauptung. ■

Anmerkung 3.12. Offensichtlich geht dieses Lemma für $N \to \infty$ in Satz 3.5 über, welcher sich wiederum in die diskrete Rekonstruktionsformel für $B_\varepsilon = \mathbb{Z}$ verwandelt (denn dann verschwinden die Integrale in C und es wird $C = \varepsilon\|f\|$).

Wir haben mit diesem Lemma einen konsistenten Rahmen zur numerischen Approximation eines Signals durch ein Wavelet-Netzwerk geschaffen, in dem wir alle auftretenden Approximationsfehler abschätzen können. In Kombination mit dem vorangegangenen Lemma und dem daraufhin erarbeiteten Verfahren zur Optimierung der Frame-Parameter in Hinblick auf die Minimierung des Approximationsfehlers steht nun einer algorithmischen Umsetzung in Form eines Modellierungsverfahrens nichts mehr im Wege.

Für eine praktische Umsetzung der hier vorgestellten theoretischen Ergebnisse (endliche Rekonstruktionsformel) verweisen wir den Leser an dieser Stelle zu Kapitel A. In Pohl (2007) wird in Kooperation mit dem Autor dieser Dissertation eine numerische Untersuchung der Rolle der Frame-Konstanten in der endlichen Rekonstruktionsformel 3.5 präsentiert. Wir kommen nun aber noch auf einen weiteren Punkt zu sprechen, der besonders in der technischen Anwendung von monumentaler Wichtigkeit ist, und zwar die "Robustheit" der Rekonstruktion.

3.4 Robustheit, Redundanz & Komplexität

In diesem Abschnitt führen wir unsere schon mehrfach angestellten Überlegungen zum Thema "Redundanz" in Frames weiter. Warum kümmern wir uns überhaupt um Frames? Und warum benutzen wir in der Rekonstruktion eines Signals f nicht immer eine Basis, z.B. die Haar-Basis? Die Antwort auf diese Fragen ist einfach: Wir brauchen

mehr Freiheitsgrade in der Darstellung eines Signals! Durch ihre Redundanz sind Frames (in so fern sie eben nicht straff und normiert sind) weniger beschränkt und stellen ein Signal-Analyse-Tool zur Verfügung, das widerstandsfähig ist gegen *additives Rauschen* und *Quantisierung* (z.B. Rundung). Zudem geht mit der Überrepräsentierung des Hilbertraums auch eine erhöhte *numerische Rekonstruktionsstabilität* und größere Freiheit in der Auflösung von *signifikanten Signal-Charakteristiken* einher.

Dass Statistik und Approximationstheorie in vielen Punkten aneinander grenzen haben wir schon mehrfach erwähnt und Kapitel 4 wird auf diesen Punkt näher eingehen. In Zusammenhang mit dem Begriff der Redundanz in Frames zeigt uns dies schon die folgende einfache Überlegung: Redundanzen in einer Familie von Vektoren sind nichts anderes als Korrelationen der Vektoren. Eine Regressorauswahl in der Statistik versucht genau diese Korrelationen zwischen Regressoren zu eliminieren. Die Begriffe Redundanz und Multikollinearität sind also in gewisser Hinsicht verwandt:

Redundanz in der Frame-Theorie $\longleftrightarrow$ Multikollinearität in der Regressionsanalyse.

In diesem Abschnitt schließen wir zunächst an unsere allgemeinen Betrachtungen an und zeigen, dass der Begriff der Redundanz in Frames sehr tief mit zwei anderen zentralen Begriffen dieser Arbeit verknüpft ist: *Kondition in der Modellbildung* und *Robustheit in Bezug auf Störungen*. Wir kehren zurück zu Lemma 3.2 und erinnern an den Operator

$$\begin{aligned} T : H &\longrightarrow \ell^2(R) \\ x &\longmapsto (\langle x, v_r\rangle_H)_{r\in R} \ . \end{aligned}$$

Ist der Frame sogar eine Basis von H, so ist $\mathrm{Im}(T) = H$. Enthält $(v_r)_{r\in R}$ aber Redundanzen, sind die Familienmitglieder also nicht unabhängig voneinander, so ist $\mathrm{Im}(T) \subsetneqq H$. Analog zu T definieren wir

$$\begin{aligned} \widetilde{T} : H &\longrightarrow \ell^2(R) \\ x &\longmapsto \left(\langle x, \mathscr{T}^{-1} v_r\rangle_H\right)_{r\in R} \ , \end{aligned}$$

so dass sich die Rekonstruktionsformel (3.3) in Satz 3.2 umschreiben lässt zu:

$$f = \widetilde{T}^* T f \ ,$$

denn $\widetilde{T}^* T = [T(T^*T)^{-1}]^* T = (T^*T)^{-1} T^* T = \mathbb{1}$. Also gilt $\widetilde{T}^* c = 0$ falls $c \perp \mathrm{Im}(T)$. Der "Clou" an der Rekonstruktionsformel

$$f = \sum_{r\in R} \langle f, v_r\rangle_H \mathscr{T}^{-1} v_r \tag{3.16}$$

war, dass zur Approximation des Signals f "nur" die Rekonstruktionskoeffizienten $\langle f, v_r\rangle_H$ berechnet werden müssen. Im Falle eines Wavelet-Frames sind diese Koeffizienten die Wavelet-Transformierte von f. Was passiert nun aber, wenn in den Koeffizienten Fehler auftreten, z.B. in der Form von Rundungsfehlern? Wir folgen Daubechies

(1992) und modellieren diesen Fall wie immer durch Addition einer $|R|$-dimensionalen Zufallsvariable E mit

$$\mathbb{E}[E_r] = 0 \ \forall r \in R \quad \text{und} \quad \mathrm{Cov}[E] = \sigma^2 \mathbb{1} \, .$$

E hat also komponentenweise einen Erwartungswert von 0 und die Komponenten von E sind unkorreliert. Das approximierte Signal ergibt sich dann zu:

$$\hat{f} = \sum_{r \in R} \left[\langle f, v_r \rangle_H + E_r \right] \mathscr{T}^{-1} v_r = \widetilde{T}^* \left(Tf + E \right) \, . \tag{3.17}$$

Mit Satz 3.5 ergibt sich:

$$\left\| f - \sum_{(m,n) \in B_\varepsilon} \mathcal{W}_\psi f(m,n) \ \mathscr{T}^{-1} \psi_{m,n} \right\| \leq \varepsilon \|f\| \, .$$

Angenommen der Frame sei fast straff, also $\mathscr{T}^{-1}\psi_{m,n} \approx A^{-1}\psi_{m,n}$. Wir fügen nun jedem $\mathcal{W}_\psi f(m,n)$ eine kleine Störung $E_{m,n}\eta$ hinzu mit der Annahme $\mathbb{E}[E_{m,n}, E_{m',n'}] = \delta_{mm'}\delta_{nn'}$ und $\mathbb{E}[E_{m,n}] = 0$. Dann ergibt sich:

$$\mathbb{E}\left[\left\| f - A^{-1} \sum_{(m,n) \in B_\varepsilon} \left(\mathcal{W}_\psi f(m,n) + E_{m,n}\eta \right) \mathscr{T}^{-1} \psi_{m,n} \right\| \right] \leq \varepsilon^2 \|f\|^2 + \eta^2 |B_\varepsilon| A^{-2} \, .$$

Halbiert man nun t_0, so verdoppelt sich A und $|B_\varepsilon|$ ebenso:

$$|B'_\varepsilon| A'^{-2} = \frac{1}{2} |B_\varepsilon| A^{-2} \, .$$

Wird also die Redundanz des Frames verdoppelt, so halbiert sich der Effekt von Störungen auf den Wavelet-Koeffizienten!

In der Theorie der Wavelet Neuronalen Netze werden die Wavelet-Koeffizienten durch erlernte Netzwerk-Gewichte ersetzt. Die Redundanz des Frames hat also direkte Auswirkungen auf die Approximationsfehler bei Störungen dieser Parameter. Wir werden auf diese Problematik in Kapitel 6 weiter diskutieren und auf die Robustheit von allgemeinen Neuronalen Netzen für zufällige Störungen im Gewichts-Parameterraum eingehen.
Zudem sehen wir in diesem Zusammenhang, dass die Redundanz eines Frames in gewisser Hinsicht ein Maß für die *Komplexität* des zu Grunde liegenden Hypothesenraumes $\mathscr{F}$ ist: Je mehr Basisfunktionen in der Entwicklung verwendet werden (also desto größer das Netzwerk ist), desto mehr Parameter müssen in der Optimierung des Netzwerkes angepasst werden. Redundanz ist also lediglich ein Maß für die Anzahl der verwendeten Parameter. Hinter dem Begriff der Komplexität eines Hypothesenraumes steckt allerdings mehr und der folgende Teil II dieser Dissertation ist hauptsächlich diesem Zusammenspiel zwischen Komplexität, Robustheit und Approximationsgenauigkeit eines allgemeinen Neuronalen Netzes gewidmet.

Teil II

Stochastische Black-Box Modellierung

4

Hypothesenräume

In diesem Teil stellen wir eine stochastische Sichtweise der Modellierung mit Neuronalen Netzen vor. Dieses Kapitel legt zunächst die Grundlagen für die in Teil III dieser Dissertation vorgestellten Anwendungen. Insbesondere werden wir wie zu Ende von Teil II angekündigt die Rolle und Struktur der zu Grunde liegenden Hypothesenräume en Detail analysieren.

4.1 Der stochastische Rahmen

Den meisten veröffentlichten Artikeln über Neuronale Netze und auch allgemein im Umfeld der Modellierungstheorie fehlt es sehr an der benötigten mathematischen Exaktheit.

Gegeben seien zwei Zufallsvariablen $X : \Omega_1 \to \mathbb{R}^d$ und $Y : \Omega_2 \to \mathbb{R}^m$, die die numerische Darstellung eines zu Grunde liegenden Phänomens darstellen. Wir nehmen immer an, dass X und Y endliche Erwartungswerte und endliche Varianzen haben. Die zugehörigen Maßräume seien $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ und $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$. Wir müssen an dieser Stelle nicht näher spezifizieren, in der Regel betrachten wir allerdings denselben Wahrscheinlichkeitsraum für beide Variablen. Als "Zielräume" verwenden wir wie schon angedeutet die beiden messbaren Räume $(\mathbb{R}^d, \mathrm{Bor}(\mathbb{R}^d))$ für X und $(\mathbb{R}^m, \mathrm{Bor}(\mathbb{R}^m))$ für Y. X hat somit die Verteilung $\mathbb{P}_X = \mathbb{P}_1 \circ X^{-1}$ und Y entsprechend $\mathbb{P}_Y = \mathbb{P}_2 \circ Y^{-1}$. Prinzipiell müssten wir nun grundlegend unterscheiden zwischen Situationen, in denen komplette Kontrolle über die Messwerte $\boldsymbol{x}$ (Realisierungen der Zufallsvariablen X) besteht und Situationen, in denen dies nicht der Fall ist. Ebenso gilt es zu trennen, ob $\boldsymbol{y}$ (Realisierungen der Zufallsvariablen Y) einzig und allein durch die Werte $\boldsymbol{x}$ festegelegt ist, oder ob noch andere Faktoren eine Rolle spielen. Es liegt auf der Hand, dass perfekte Messungen utopischer Natur sind und wir jeweils den zweiten Fall betrachten. Wir führen also eine neue Zufallsvariable $Z : \Omega_1 \times \Omega_2 \to \mathbb{R}^d \times \mathbb{R}^m$, $Z = (X, Y)$, mit dem Wahrscheinlichkeitsmaß $\mathbb{P}_1 \otimes \mathbb{P}_2$ ein. Die Verteilung von Z ist gegeben durch $\mathbb{P}_Z = \mathbb{P}_1 \otimes \mathbb{P}_2 \circ Z^{-1}$. Dieses Bildmaß

ist sozusagen die gemeinsame Verteilung von X und Y. Da wir annehmen, dass sich Y in irgendeiner Form aus X errechnet setzt sich $\mathbb{P}_Z$ zusammen aus der Verteilung $\mathbb{P}_X$ von X und einem konditionellen Zusammenhang $\mu : \mathbb{R}^m \to \mathbb{R}$, $\mu = \mu(\cdot \mid \boldsymbol{x})$, zwischen X und Y. Ausführlich geschrieben ($\Omega := \Omega_1 \times \Omega_2$):

$$\begin{aligned}\mu(A|\boldsymbol{x}) &:= \mathbb{P}_1 \otimes \mathbb{P}_2 \left(\{(\omega_1, \omega_2) \in \Omega : Y(\omega_2) \in A\} \mid \{(\omega_1, \omega_2) \in \Omega : X(\omega_1) = \boldsymbol{x}\}\right) \\ &= \frac{\mathbb{P}_1 \otimes \mathbb{P}_2 \left(\{(\omega_1, \omega_2) \in \Omega \,:\, Y(\omega_2) \in A \,\wedge\, X(\omega_1) = \boldsymbol{x}\}\right)}{\mathbb{P}_X(B)}\end{aligned}$$

für ein $A \in \mathbb{R}^m$, also speziell

$$\mathbb{P}_Z(\boldsymbol{x}, \boldsymbol{y}) = \mu(\boldsymbol{y}|\boldsymbol{x}) \cdot \mathbb{P}_X(\boldsymbol{x}) \,.$$

An sich ist diese Aufspaltung nichts anderes als eine Umschreibung des Satzes von Fubini: Sei $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ integrierbar, dann ist

$$\int_{\mathbb{R}^d \times \mathbb{R}^m} f(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\mathbb{P}_Z(\boldsymbol{x}, \boldsymbol{y}) = \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^m} f(\boldsymbol{x}, \boldsymbol{y}) \, \mathrm{d}\mu(\boldsymbol{y}|\boldsymbol{x}) \right) \mathrm{d}\mathbb{P}_X \,.$$

Angenommen Y ist komplett bestimmt durch X, also $Y = g(X)$ für eine Funktion g, dann ist $\mu(\{g(\boldsymbol{x})\}|\boldsymbol{x}) = \mathbb{P}_1 \otimes \mathbb{P}_2 (\Omega_1 \times \Omega_2) = 1$ für *alle* $\boldsymbol{x} \in \mathbb{R}^d$. Gibt es keinen exakten Zusammenhang zwischen X und Y wird es auch kein solches g geben.
Diese Funktion μ beinhaltet alle Informationen über den Zusammenhang zwischen X und Y. In der Praxis ist allerdings meist nur der Erwartungswert $f(X) := \mathbb{E}[Y|X]$ von Y für ein gegebenes X bekannt, bzw. gemessen wird "im Schnitt" $f(\boldsymbol{x}) := \mathbb{E}[Y|X = \boldsymbol{x}]$. Die eigentlich gemessene Realisierung $\boldsymbol{y} = Y(\omega)$, $\omega \in \Omega_2$, wird in der Praxis also von $f(\boldsymbol{x})$ abweichen, d.h. wir definieren den zufälligen Fehler als Zufallsvariable

$$E := Y - \mathbb{E}[Y|X]$$

und wegen $f(X) = \mathbb{E}[Y|X]$ erhalten wir das Modell

$$Y = f(X) + E \,.$$

Aus der Definition von E und den Eigenschaften des bedingten Erwartungswertes folgt sofort, dass $\mathbb{E}[E|X] = 0$. Wie schon beschrieben wird nun versucht, z.B. mit Hilfe eines Neuronalen Netzes, diese Funktion f zu erlernen, und zwar anhand einer Reihe von Messdaten. Es wird also nicht der explizite Zusammenhang zwischen X und Y approximiert, sondern vielmehr nur ein *Teilaspekt* des Problems, und zwar $\mathbb{E}[Y|X]$. Dass dies zu Problemen führen kann zeigt die folgende Überlegung: Angenommen wir haben $f(X)$ auf Grundlage der Messdaten gut approximiert und suchen nun den besten Schätzer für $g(Y)$, wobei g beliebig sei, z.B. eines der Momente von Y. Der offensichtliche Kandidat wäre $g(f(X)) = g(\mathbb{E}[Y|X])$. Doch wer garantiert uns, dass dies auch wirklich der beste Schätzer für den eigentlichen Erwartungswert $\mathbb{E}[g(Y)|X]$ ist?
Dieser Idee folgend betrachten Rossi & Conan-Guez (2005) funktionale Neuronale Netze (FNNs) auf Grundlage von Wavelet Neuronalen Netzes. FNNs können zumindest theoretisch den wirklichen konditionellen Zusammenhang zwischen X und Y erlernen.

4.2 Risiko-Minimierung

Wir kommen zurück zu der allgemeinen black-Box-Modellierung des Schätzers für die gesuchte Funktion f. Ausgangssituation ist wieder eine Stichprobe (N-Messungen)

$$\mathcal{T} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$$

der Zufallsvariablen (X, Y) und im speziellen Fall von $m = 1$, also eindimensionalen Messdaten, suchen wir eine eindimensionale-Basisfunktions-Entwicklung, d.h. einen Schätzer $\hat{f} : \mathbb{R}^d \times W \to \mathbb{R}$, so dass für $\boldsymbol{x} \in \mathbb{R}^d$, $w \in W = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, i = 1, ..., M\}$

$$f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x}; w) = \sum_{i=1}^{M} u_i \, \widetilde{\Phi}(\boldsymbol{a}_i, t_i; \boldsymbol{x}) \, . \tag{4.1}$$

W sei ganz allgemein der Raum der Gewichte (hier $W = \mathbb{R}^{M(d+2)}$, das Modell enthält also $M(d+2)$ numerische Parameter), dessen Struktur natürlich stark von der Architektur des Netzwerkes bzw. der Entwicklung abhängt. Für $\boldsymbol{y} \in \mathbb{R}^m$ ist entsprechend eine m-dimensionale Version der Basisfunktion $\widetilde{\Phi}$ zu wählen.
Im allgemeinen ist $\hat{f} \in \mathscr{F}(\Omega_1, \mathbb{R}^m)$ also eine Funktion der Form

$$\hat{f} : \Omega_1 \longrightarrow \mathbb{R}^m$$

und wir schreiben

$$\hat{f}(\boldsymbol{x}) := \hat{f}\left(X(\omega_1)\right), \quad \omega_1 \in \Omega_1 \, .$$

$\mathscr{F}$ bezeichne einen hier noch nicht näher spezifizierten Funktionenraum: Den *Hypothesenraum*. In der Regel wird $\hat{f}$ parametrisiert durch einen Parameter $w \in W$, also $\hat{f}(X) = \hat{f}(X, w)$. In den vorangegangenen Abschnitten haben wir gezeigt, dass Neuronale Netze universelle Approximatoren sind, d.h. in diesem Funktionenraum gibt es garantiert mindestens einen Vertreter, der jede Funktion aus L^p theoretisch beliebig genau approximieren kann. Wavelet Neuronale Netze haben diese Eigenschaft natürlich auch, sie besitzen aber bessere Konvergenzeigenschaften als herkömmlichen Neuronale Netze, so dass die Aussage "liegen dicht in" für Wavelets konkretisiert werden kann und am Ende sogar den Fehler bei der Approximation jedes $f \in L^2$ durch endliche Superposition von Wavelets unter bestimmten Voraussetzungen nach oben beschränkt ist (hierzu siehe Satz 3.5).
Nun kommen wir aber zurück zu der Frage: Unabhängig von der konkreten Struktur von $\hat{f}$, wie misst man die *Güte* eines Modells? Um dieser Frage auf den Grund zu gehen definieren wir zunächst als abstraktes "Performance"-Maß die Zufallsvariable[1]

$$\Xi : \Omega_1 \times \Omega_2 \times W \to \mathbb{R}$$

[1] Wie schon angedeutet ist in der Praxis zumeist $\Omega_1 = \mathbb{R}^d$ und $\Omega_2 = \mathbb{R}^m$ gemeint.

auf dem Wahrscheinlichkeitsraum $(\Omega_1 \times \Omega_2 \times W, \mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \mathcal{F}_W, \mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_W)$, so dass für gegebenen Input X und "target" Y die Schätz-Performance durch

$$\Xi(Y, \hat{f}(X, w))$$

gemessen wird. $w \in W \subset \Theta$, wobei Θ einen nicht näher spezifizierten Parameterraum bezeichne. Wir wählen in Hinblick auf praktische Anwendungen $\Theta = \mathbb{R}^s$, $s \in \mathbb{N}$, und $\mathcal{F}_W = \mathrm{Bor}(W)$. In der Praxis fordern wir zudem $\Xi(Y = \boldsymbol{y}, \hat{f}(w, X = \boldsymbol{x})) \geq 0$ und $\Xi(Y = \boldsymbol{y}, \hat{f}(w, X = \boldsymbol{x})) = 0$ genau dann wenn $\boldsymbol{y} = \hat{f}(w, \boldsymbol{x})$. Weiterhin sei Ξ eine L^1-Funktion in Bezug auf ein geeignetes Wahrscheinlichkeitsmaß. Ist Ξ *erwartungstreu*, so existiert für alle $w \in W$ die Funktion $\Lambda : W \to \mathbb{R}$

$$\Lambda_{\Xi,\hat{f}}(w) := \mathbb{E}\left[\Xi\left(Y, \hat{f}(X, w)\right)\right] = \int_{\Omega_1 \times \Omega_2 \times W} \Xi\Big(Y, \hat{f}(X, w)\Big) \, \mathrm{d}\left(\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_W\right) . \quad (4.2)$$

Λ misst die durchschnittliche bzw. erwartete Performance des Schätzers $\hat{f}$ und stellt somit eine *Prognose* dar. Man bezeichnet $\Lambda(w)$ auch als *Risiko-Funktional*. Sie hängt somit nicht von den konkreten Realisierungen $\boldsymbol{x}$ und $\boldsymbol{y}$ ab. Als Funktion ist Λ in dieser Definition allerdings nur von w abhängig, Ξ und $\hat{f}$ werden im voraus gewählt. Der Erwartungswert reduziert sich auf das Integral über die Wahrscheinlichkeitsmaße von Y und X, denn $w \in W$ ist deterministisch, also $(\mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \mathbb{P}_W)(A_1 \times A_2 \times A_3) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)\mathbb{P}_W(A_3) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$ für alle $A_1 \in \mathcal{F}_1$, $A_2 \in \mathcal{F}_2$, $A_3 \in \mathrm{Bor}(W)$. D.h.

$$\Lambda_{\Xi,\hat{f}}(w) = \int_{\Omega_1 \times \Omega_2 \times W} \Xi(Y, \hat{f}(X, w)) \, \mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) = \int_{\mathbb{R}^d \times \mathbb{R}^m} \Xi\left(\boldsymbol{y}, \hat{f}(\boldsymbol{x}, w)\right) \, \mathrm{d}\mathbb{P}_Z(\boldsymbol{x}, \boldsymbol{y}) , \quad (4.3)$$

also ein $\mathbb{R}^d \times \mathbb{R}^m$-dimensionales Lebesgue-Integral. Würden wir die gemeinsame Verteilung $\mathbb{P}_Z$ kennen, so könnte dieser Ausdruck direkt nach w umgeformt werden. Leider liegt uns diese gerade nicht vor, so dass wir vor der Aufgabe stehen (4.3) anhand von N identisch verteilten Samples $Z_1, ..., Z_N$ zu minimieren. Diesen Prozess bezeichnet man als *Lernproblem*. Ein Lernalgorithmus ist somit eine Funktion aus dem Raum der Stichproben in den Raum der Schätzer $\mathscr{F}$

$$\begin{aligned} A_{\Xi,\mathscr{F}} : \mathscr{T} &\longrightarrow \mathscr{F} \\ \mathcal{T} &\longmapsto \hat{f}_{\mathcal{T}} , \end{aligned}$$

wobei $\mathscr{T} = (\Omega_1 \times \Omega_2)^N$ der Raum der N-dimensionalen Stichproben sei. $\hat{f}_{\mathcal{T}} = A_{\Xi,\mathscr{F}}(\mathcal{T})$ bezeichnet man als Hypothese. Wir werden je nach Kontext nur $\mathcal{T}$ oder, falls die Länge der Stichprobe wichtig ist, $\mathcal{T}^N$ schreiben. Wir nehmen immer an, dass $A_{\Xi,\mathscr{F}}$ *symmetrisch* ist, d.h. es kommt nicht auf die Reihenfolge der Trainingsdaten an:

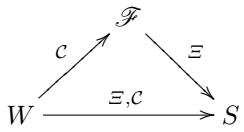
$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\Xi(Y, \hat{f}_{\mathcal{T}})\right] = \mathbb{E}_{\overline{\mathcal{T}}}\left[\Xi(Y, \hat{f}_{\mathcal{T}(\pi)})\right] ,$$

wobei $\mathcal{T}(\pi) = \{Z_{\pi(1)}, ..., Z_{\pi(n)}\}$ für jede Permutation π von $\{1, ..., n\}$. Der Erwartungswert $\mathbb{E}_{\mathcal{T}}$ wird gebildet über alle Trainings-Datensets der Länge N. Wir werden diese Notation in Abschnitt 4.2.1 präzisieren.

Ein Lernalgorithmus benötigt also vier "Zutaten":

1) Ein Funktionenraum[2] $\mathscr{F} = \{\hat{f} = \hat{f}(X, w) \in \mathcal{C} : w \in W\}$, der Raum der möglichen Schätzer, wobei $\mathcal{C} \subset C(\mathbb{R}^d, \mathbb{R}^m)$ eine Menge von stetigen Funktionen sei, die durch einen Parameter $w \in W$ parametrisiert werden: $\mathcal{C} = \{g \in C(\mathbb{R}^d, \mathbb{R}^m) : g(\boldsymbol{x}) = g(\boldsymbol{x}, w), w \in W\}$,
2) eine Performance-Funktion Ξ, die ein Maß für die Güte eines Schätzers aus $\mathscr{F}$ darstellt,
3) ein Verfahren, welches das Minimum des Erwartungswertes der Performance-Funktion in $S = \{\Xi = \Xi(Y, \hat{f}(X, w)) : \hat{f} \in \mathscr{F}\}$ findet.

Ein Schaubild illustriert die Situation:



Im endlichen Fall ist somit $|W_M| = |\mathscr{F}_M| = |S_M|$. $A_{\Xi,\mathscr{F}}$ könnte zum Beispiel der Backpropagation-Algorithmus in Verbindung mit der least-squares Performance-Funktion sein auf Grundlage des Funktionenraums $\mathscr{F} = \Sigma^d(\sigma)$. Der Algorithmus wählt einen Schätzer aus $\mathscr{F}$ so aus, dass er eine möglichst gute "Durchschnittsleistung" für die Trainingsdaten erbringt, und haben wir uns auf ein konkretes Modell festgelegt, also Ξ und $\hat{f}(\cdot, w)$ gewählt, so passen wir die Parameter von $\hat{f}$ so an, dass

$$\hat{w} = \operatorname*{argmin}_{w \in W} \Lambda_{\Xi,\hat{f}}(w) \,.$$

In dieser Überlegung sehen wir aber schon ein Grundproblem: Der Modellierungsprozess besteht nicht nur in der Wahl des richtigen $\hat{f} \in \mathscr{F}$, sondern auch in der Festlegung von Ξ. Je nachdem wie Ξ gewählt wird rückt ein anderer Aspekt (z.B. ein bestimmtes Moment der gemeinsamen Verteilung von X und Y) der Beziehung zwischen X und Y in den Vordergrund. Das ist eine herbe Einschränkung, da wir in einer konkreten Anwendung eventuell die Gewichte des Schätzer für *mehrere* Aspekte optimieren möchten, z.B. Erwartungswert *und* Varianz. Zur Illustration ein Beispiel. Wir wählen

[2] Die in dieser Dissertation betrachteten Hypothesenräume werden immer als parametrisiert angenommen. Dies ist in Hinblick auf Anwendungen wie z.B. Approximation durch Neuronale Netze auch durchaus sinnvoll. Viele der Ergebnisse sind aber auch auf allgemeinere Situationen verallgemeinerbar.

$$\Xi\left(Y, \hat{f}(X,w)\right) = \left(Y - \hat{f}(X,w)\right)^2 ,$$

bzw. auf Realisierungs-Ebene

$$\Xi\left(Y, \hat{f}(X,w)\right)(\omega) = \Xi\left(\boldsymbol{y}, \hat{f}(\boldsymbol{x},w)\right) = \left(Y - \hat{f}(X,w)\right)^2(\omega) = \left|\boldsymbol{y} - \hat{f}(\boldsymbol{x},w)\right|^2$$

für $\omega \in \Omega_1 \times \Omega_2 \times W$. Mit $f(X) = \mathbb{E}[Y|X]$ formen wir folgendermaßen um:

$$\begin{aligned}
\Lambda_{\Xi,\hat{f}}(w) &= \mathbb{E}\left[\left(Y - \hat{f}(X,w)\right)^2\right] \\
&= \mathbb{E}\left[\left(Y - \hat{f}(X,w) - f(X) + f(X)\right)^2\right] \\
&= \mathbb{E}\left[(Y - f(X))^2\right] + 2\mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)(Y - f(X))\right] + \mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)^2\right] .
\end{aligned}$$

Wegen $Y = f(X) + E$, $\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right]$ und den Eigenschaften von E folgt

$$\begin{aligned}
\mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)(Y - f(X))\right] &= \mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)E\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)E|X\right]\right] \\
&= \mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)\mathbb{E}\left[E|X\right]\right] \\
&= 0 ,
\end{aligned}$$

d.h.

$$\Lambda_{\Xi,\hat{f}}(w) = \mathbb{E}\left[(Y - f(X))^2\right] + \mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)^2\right] . \tag{4.4}$$

Falls ein $\hat{w}$ die Größe $\Lambda_{\Xi,\hat{f}}(w)$ minimiert, dann minimiert es somit auch

$$\mathbb{E}\left[\left(f(X) - \hat{f}(X,w)\right)^2\right] = \int_{\Omega_1}\left[f(X) - \hat{f}(X,w)\right]^2 \mathrm{d}\mathbb{P}_1 = \int_{\mathbb{R}^d}\left[f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x},w)\right]^2 \mathrm{d}\mathbb{P}_X(\boldsymbol{x}). \tag{4.5}$$

Für einen erwartungstreuen Schätzer, also $\mathbb{E}[\hat{f}] = f$, ist diese Größe aber nichts anderes als die Varianz des Schätzers:

$$\mathbb{V}\left[\hat{f}\right] = \mathbb{E}\left[\left(\hat{f} - \mathbb{E}\left[\hat{f}\right]\right)^2\right] = \mathbb{E}\left[\left(\hat{f} - f\right)^2\right] .$$

Wählen wir als Performance-Maß also speziell den quadratischen Fehler $\Xi(Y, \hat{f}(w,X)) = (Y - \hat{f}(X,w))^2$ und haben anhand von Trainingsdaten (oder auf andere Weise) einen Gewichts-Parameter $\hat{w}$ gefunden, so dass $\mathbb{E}\left[\Xi\right]$ minimal ist, dann ist $\hat{f}(X,\hat{w})$ auch der

beste Schätzer für den konditionellen Erwartungswert $f(X) = \mathbb{E}[Y|X]$ und hat minimale Varianz. Im Falle eines linearen Modells, also

$$\hat{f}(X, w) = Xw \,,$$

wobei $X \in \mathbb{R}^d \times \mathbb{R}^m$ eine Matrix und $w \in \mathbb{R}^m$ sei, ist dies nichts anderes als die Kernaussage des Satzes von Gauss-Markov (s. z.B. Meintrup & Schäffler (2005)).
Doch auch für nicht-erwartungstreue Schätzer lässt sich für die Performance-Funktion der quadratischen Abweichung allein aus $\mathbb{E}[(Y - \hat{f}(X, w))^2]$ ein Güte-Kriterium für den Schätzer ableiten, obwohl diese Größe *nicht* mehr auch direkt mit der Varianz des Schätzers verknüpft ist. Als Güte-Kriterium kann dann z.B. das Bayes-Risiko (s. wieder Meintrup & Schäffler (2005)) dienen. Ein Schätzer, der diese Güte optimiert ist aber i.A. (d.h. wenn keine weiteren Informationen über die Wahrscheinlichkeitsverteilung der Parameter w vorliegt) *nicht* Varianz-optimal[3]. Im Falle von z.B. Neuronalen Netzen liegen uns derartige Informationen i.A. nicht vor.

Im Prinzip gilt es also, ganz allgemein gesprochen, für gegebenes Ξ und $\hat{f}$ folgendes Zweizieloptimierungsproblem zu lösen:

$$\hat{w} = \underset{w \in W}{\operatorname{argmin}} \begin{pmatrix} \mathbb{E}[\Xi] \\ \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] \end{pmatrix} \,. \tag{4.6}$$

Die eigentlich gesuchte Größe $\mathbb{E}[(\hat{f} - f)^2]$ lässt sich nicht direkt optimieren, da f unbekannt ist. Aber wie gesagt, falls speziell $\Xi = (Y - \hat{f})^2$ gewählt wird, dann ergibt die Optimierung von $\mathbb{E}[\Xi]$ auch minimales $\mathbb{E}[(\hat{f} - f)^2]$, ob der Schätzer erwartungstreu ist oder nicht. Andersherum, falls Ξ allgemein belassen wird, der Schätzer aber erwartungstreu ist, so ist $\mathbb{V}[\hat{f}]$ minimal. Glg. (4.6) stellt somit den allgemeinen Fall dar und beinhaltet die betrachteten Spezialfälle.

Im Kontext des Lernproblems ist leicht einzusehen, dass $\hat{w}$ *nicht* eindeutig ist: Die "hidden" nodes des Neuronalen Netzwerkes sind vertauschbar und für hinreichend allgemeines W existieren somit immer multiple Lösungen. Hecht-Nielsen (1990) beschreibt eine mögliche Einschränkung für W (eine kegelförmige Struktur, Hecht-Nielsen-Cone genannt), so dass die Austauschbarkeit verloren geht und zumindest die Chance auf ein eindeutiges Minimum bestehen bleibt. Aus Sicht des Rechenaufwands beim Trainieren des Netzwerkes ist eine Einschränkung auf eine besondere Struktur von W allerdings problematisch. Wir werden in Abschnitt 6.1 auf eine einfachere Möglichkeit der Einschränkung für W eingehen.

[3] Falls doch Informationen über w in Form eines konkreten Wahrscheinlichkeitsmaßes auf einer σ-Algebra auf dem Parameterraum gegeben sind, so lässt sich der gleichmäßig beste Schätzer durch eine Bayes-Schätzfunktion bestimmen.

Ein für ein Performance-Maß Ξ gefundenes optimales $\hat{w}$ muss nicht auch für andere Performance-Maße $\Xi' \neq \Xi$ das Optimum darstellen. Und wie schon beschrieben reicht für eine vollkommen unvoreingenommene Modellierung eines physikalischen oder technischen Vorgangs die alleinige Optimierung der Modellparameter nicht aus, da Ξ und $\hat{f}$ schon a priori gewählt werden. Aber auch die Definition von Λ ist entscheidend. So spielt besonders in technischen Anwendungen die *Robustheit* des errechneten Schätzers gegenüber Messfehlern o.Ä. eine besondere Rolle. Wir gehen auf diese Problematik näher in Kapitel 6 ein. Wir fassen diese Überlegungen zusammen und bezeichnen von nun an das *optimierte Gesamtmodell* mit

$$\hat{\mathscr{M}} := \left(\Xi, \hat{f}, \hat{w}\right) = \left(\Xi, \hat{f}, \underset{w \in W}{\operatorname{argmin}} \begin{pmatrix} \mathbb{E}[\Xi] \\ \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] \end{pmatrix}\right) ,$$

und analog $\mathscr{M} = (\Xi, \hat{f}, w)$ für das noch nicht optimale Modell. Wir weisen darauf hin, dass diese Schreibweisen lediglich voraussetzen, dass $Y = f(X) + E$ mit $\mathbb{E}[E] = 0$ geschrieben werden kann. Alle weiteren Modellannahmen stecken dann in der Wahl von Ξ und $\hat{f}$ sowie in der Bestimmung von $\hat{w}$ gemäß (4.6). Den zu diesem Modell zugehörigen Parameterraum bezeichnen wir nun konsequenterweise mit W_{Ξ} bzw. eigentlich $W_{\hat{f}}$, denn er hängt nur von der gewählten Modellstruktur ab. Den Subskript lassen wir im folgenden aber zumeist weg. Anders für $\hat{w}$: Wird dieses auf Grundlage einer Stichprobe $\mathcal{T}$ bestimmt, dann verwandelt sich der Erwartungswert in eine Summe über die einzelnen Realisierungen und wir schreiben konsequenterweise $\hat{w}_{\mathcal{T}^N}$, so dass

$$\hat{f}_{\mathcal{T}^N} = \hat{f}\left(X, \hat{w}_{\mathcal{T}^N}\right) .$$

Die Güte des Modells sollte entgegen den gängigen Darstellungen dementsprechend auch nicht allein durch die mittlere Performance $\mathbb{E}[\Xi(Y, \hat{f}(X, \hat{w}))]$ (wie in der einschlägigen System-Identification Literatur verwendet, siehe für eine gute Darstellung z.B. Sjöberg (1995)), sondern vielmehr durch eine Kombination mehrerer Aspekte von Ξ definiert werden, z.B.

$$\text{Güte des Modells } \hat{\mathscr{M}} := \left\| \left(\mathbb{E}[\Xi], \mathbb{V}[\hat{f}], \text{Robustheit}[\hat{w}], ...\right)^T \right\|$$

in einer nicht näher spezifizierten Norm (der Begriff der Robustheit ist an dieser Stelle noch undefiniert). Durch "..." wollen wir andeuten, dass noch andere Gesichtspunkte von Bedeutung sein könnten, wie z.B. Maximierung der Glattheit des resultierenden Schätzers, ein Kriterium, das in der Konstruktion von Splines umgesetzt ist. Nowacki (1990) liefert eine gute Zusammenfassung über Glättungsverfahren von Kurven und Flächen.

Glgn. (4.4) und (4.5) machen einen weiteren Punkt deutlich, der von größter Wichtigkeit aus Modellierungssicht ist. Der erste Summand auf der rechten Seite von (4.4)

ist ein Integral über die gemeinsame Verteilung $\mathbb{P}_Z$ der Zufallsvariablen X und Y, genauso wie Λ selbst. Der zweite Summand hingegen ist ein Integral einzig und alleine über die Verteilung von X. Die Konsequenz ist, dass ein optimales $\hat{w}$, bestimmt in Bezug auf das Wahrscheinlichkeitsmaß $\mathbb{P}_Z$, im Allgemeinen *nicht* optimal ist für eine andere Verteilung $\mathbb{P}'_X \neq \mathbb{P}_X$. Konkret stellt sich uns also folgende Situation: Wenden wir den aus Trainingsdaten (deren X-Werte einer bestimmten Verteilung $\mathbb{P}_X$ folgen) gewonnenen Schätzer $\hat{f}(X, \hat{w})$ auf *neue* Daten mit $\mathbb{P}'_X$-verteilten X-Werten an, so wird der Schätzer i.A. sub-optimale Ergebnisse liefern. Ein Pre-Processing der Trainingsdaten bei dem die Verteilungen abgeändert werden bringt somit die erhebliche Gefahr mit sich, dass der resultierende Black-Box-Approximant in konkreten Anwendungssituationen versagt.

In der Literatur zum Thema Training von (Wavelet-) Neuronalen Netzen spielen diese Modellbildungs-Aspekte keine bzw. kaum eine Rolle (siehe grundlegende Arbeiten wie z.B. Zhang (1994) oder Zhang et al. (1995)[4]). Andere Autoren optimieren diesen speziellen Ansatz durch verbesserte least-squares-Methoden, um eine möglichst gute Immunität gegen Ausreißer bei den Trainingsdaten zu erreichen (s. z.B. Li & Leiss (2001)). In Artikeln wie Rivals & Personnaz (2003) werden statistische Methoden diskutiert, um den Konstruktionsprozess eines Neuronalen Netzes zu optimieren und z.B. Effekte wie schlecht konditionierte Jacobi-Matrizen bei der Optimierung der Parameter zu vermeiden. McKeown et al. (1997) steuerten schon erste Ansätze zu diesem Thema bei. Overfitting-Effekte (also zu viele Parameter, M wird zu groß gewählt) werden durch Regularisierungs-Methoden vermeintlich unter Kontrolle gehalten und M z.B. durch Informationskriterien wie AFPE (Akaike's final prediction error criterion) oder Cross Validation geschätzt (s. z.B. Anders & Korn (1999)).
Doch all diese Autoren verstehen den Modellbildungsprozess bei Neuronalen Netzen als *Konstruktionsprozess* des Netzwerkes, also Wahl der Parameter, Bildung der Verknüpfungen, etc. Dass die eigentlich Modellbildung schon viel früher ansetzt und schon allein in der Wahl von Ξ festgelegt wird welcher Aspekt des Modells "optimal" ist, wird in der Regel nicht beachtet.

4.2.1 Empirische Risiko-Minimierung

Wir kommen zurück auf die Dekomposition der erwarteten (mittleren) Performance des in Hinblick auf die Parameter optimierten Schätzers gemäß Glg. (4.4)

$$\Lambda_{\Xi,\hat{f}}(\hat{w}) = \mathbb{E}\left[E^2\right] + \mathbb{E}\left[\left(f(X) - \hat{f}(X, \hat{w})\right)^2\right] .$$

Der erste Term auf der rechten Seite ist unabhängig von $\hat{f}$ und den Trainingsdaten und kann somit nicht weiter minimiert werden, im Gegensatz zum zweiten Term. Man

[4] Man beachte: In dieser Dissertation sind mehrere Autoren mit dem Nachnamen Zhang vertreten.

beachte, dass der Erwartungswert lediglich über X genommen wird, also

$$\mathbb{E}\left[\left(f(X)-\hat{f}(X,\hat{w})\right)^2\right]=\int_{\Omega_1}\left(f(X)-\hat{f}(X,\hat{w})\right)^2\,\mathrm{d}\mathbb{P}_1\,.$$

Oft wird dieser Term *Generalisierungsfehler* des Modells genannt, allgemein

$$\text{Generalisierungsfehler}=\mathbb{E}\left[\Xi\left(f(X),\hat{f}(X,\hat{w})\right)\right]=\int_{\Omega_1}\Xi\left(f(X),\hat{f}(X,\hat{w})\right)\,\mathrm{d}\mathbb{P}_1\,.$$

In der Literatur wird allerdings etwas uneinheitlich zuweilen auch Λ selbst, also das Risiko als Generalisierungsfehler bezeichnet.

Angenommen es liegen perfekte Daten vor, also $E=0$. Dann müsste, falls das Modell $\hat{f}$ und die wirkliche Funktion f identische Strukturen haben, $\Lambda=0$ sein. Im Allgemeinen ist dies aber nicht der Fall, und der Grund hierfür liegt offensichtlich in der *prinzipiellen* Unfähigkeit des gewählten Modells den Funktionszusammenhang darzustellen. Man nennt dies *Bias-Fehler*. Nun kommt in der Praxis aber noch hinzu, dass uns in Form von $\mathcal{T}$ immer nur eine endliche Anzahl von Samples der Zufallsvariablen X und Y vorliegen, d.h wir erwarten von dieser Seite eine *zusätzliche* Fehlerquelle in der Rekonstruktion der Funktion f. Wir konkretisieren diese Überlegung und betrachten statt Λ selbst $\mathbb{E}_{\overline{\mathcal{T}}^N}[\Lambda]$. Hierbei bezeichne $\overline{\mathcal{T}}^N$ die Menge aller möglichen Trainingsdatensätze der Form $\mathcal{T}^N=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),...,(\boldsymbol{x}_N,\boldsymbol{y}_N)\}$, $N\in\mathbb{N}$ fest vorgegeben, wobei die Messdaten $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ Realisierungen derselben gemeinsamen Verteilung $\mathbb{P}_Z$ von $Z=(X,Y)$ seien. Vereinfacht ausgedrückt ist $\mathbb{E}_{\overline{\mathcal{T}}^N}[\Lambda]$ nichts anderes als der Mittelwert von Λ über alle möglichen Samples der Länge N aus der Verteilung $\mathbb{P}_Z$. Ist $\mathbb{P}_Z$ eine diskrete Verteilung, so ist $\mathbb{E}_{\overline{\mathcal{T}}^N}[\Lambda]=\Lambda$. Die Schreibweise $\hat{f}(X,\hat{w}_{\mathcal{T}^N})$ drückt wie zuvor aus, dass $\hat{w}$ von dem speziellen Trainingssatz $\mathcal{T}^N$ abhängt, also i.A. $\hat{w}_{\mathcal{T}^N}\neq\hat{w}$. Es gilt

$$\begin{aligned}\text{Generalisierungsfehler}&=\mathbb{E}_{\overline{\mathcal{T}}^N}\left[\mathbb{E}\left[\Xi\left(f(X),\hat{f}(X,\hat{w}_{\mathcal{T}^N})\right)\right]\right]\\&=\int_{(\Omega_1\times\Omega_2)^N}\left(\int_{\Omega_1}\Xi\left(f(X),\hat{f}(X,\hat{w}_{\mathcal{T}^N})\right)\,\mathrm{d}\mathbb{P}_1\right)\,\mathrm{d}\underbrace{(\mathbb{P}_1\otimes\mathbb{P}_2)^N}_{=:\mathbb{P}^N}\,.\end{aligned}$$

Konkret für den Spezialfall $\Xi=(Y-\hat{f})^2$:

$$\mathbb{E}_{\overline{\mathcal{T}}^N}[\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}^N})]=\mathbb{E}\left[E^2\right]+\mathbb{E}_{\overline{\mathcal{T}}^N}\left[\mathbb{E}\left[\left(f(X)-\hat{f}(X,\hat{w}_{\mathcal{T}^N})\right)^2\right]\right]\,.$$

Wir formen nun analog zu der Dekomposition (4.4) durch Addieren und gleichzeitiges Subtrahieren von f und $\mathbb{E}_{\overline{\mathcal{T}}^N}[\hat{f}(X,\hat{w}_{\mathcal{T}^N})]$ um, wobei auch in diesem Fall $\mathbb{E}_{\overline{\mathcal{T}}^N}[(f-\mathbb{E}_{\overline{\mathcal{T}}^N}[\hat{f}])(\mathbb{E}_{\overline{\mathcal{T}}^N}[\hat{f}]-\hat{f})]=0$, wie eine einfache Rechnung zeigt. Es ergibt sich die aus der Statistik bekannte Formel

$$\mathbb{E}_{\overline{\mathcal{T}}^N}[\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}^N})] = \underbrace{\mathbb{E}\left[E^2\right]}_{\text{Noise}} + \underbrace{\int_{\Omega_1}\left(f(X) - \mathbb{E}_{\overline{\mathcal{T}}^N}\left[\hat{f}(X,\hat{w}_{\mathcal{T}^N})\right]\right)^2 \mathrm{d}\mathbb{P}_1}_{\text{Bias}} \tag{4.7}$$
$$+ \underbrace{\mathbb{E}_{\overline{\mathcal{T}}^N}\left[\int_{\Omega_1}\left(\mathbb{E}_{\overline{\mathcal{T}}^N}\left[\hat{f}(X,\hat{w}_{\mathcal{T}^N})\right] - \hat{f}(X,\hat{w}_{\mathcal{T}^N})\right)^2 \mathrm{d}\mathbb{P}_1\right]}_{\text{Varianz}} .$$

Also:

$$\text{Generalisierungsfehler} = \text{bias} + \text{Variance} .$$

Der Vergleich dieser drei Terme mit den Korrespondierenden aus Glg. (4.4) führt sofort auf die folgende Frage: Wenn der Stichprobenraum $\overline{\mathcal{T}}$ nur "groß" genug gewählt wird, also unendlich große Samples betrachtet werden, wird dann

$$\mathbb{E}_{\overline{\mathcal{T}}^N}\left[\hat{f}(X,\hat{w}_{\mathcal{T}^N})\right] \xrightarrow{N\to\infty} \hat{f}(X,\hat{w}) \,?$$

Diese Frage werden wir etwas später im Rahmen der Konsistenzbetrachtungen näher beleuchten. Zunächst jedoch eine kurze Erläuterung von (4.7). Der erste Term (Bias) drückt aus, dass das Modell selbst im Mittel über alle möglichen Datensätze immer noch von den wahren Beobachtungen abweicht, es fehlt dem Modell also grundsätzlich an Parametern bzw. allgemeiner an Flexibilität. Für einen *erwartungstreuen* Schätzer verschwindet der Bias-Term. Der Varianz-Fehler ist die erwartete Fluktuation des aus einem Datensatz entstandenen Modells um das best-mögliche Modell derselben Struktur. Der Varianz-Fehler ist also eine Ausprägung der Tatsache, dass nur eine endliche Anzahl von Messpunkten vorliegen. Der Punkt ist nun der, dass eine Erhöhung der Flexibilität des Modells (z.B. mehr Parameter) den Bias-Fehler reduziert, im Gegenzug aber den Varianz-Fehler *erhöht*, weil sich das System immer besser an den *spezifischen* Datensatz $\mathcal{T}$ anpasst und immer mehr von der wahren Verteilung $\mathbb{P}_Z$ abweicht. Eine Reduktion des Varianz-Fehlers kann z.B. durch Glättung erreicht werden, also ein Zusammenfassen der Einflüsse von im Input-Raum benachbarter Punkte. Hierdurch wird allerdings im Gegenzug der Bias-Term vergrößert, weil lokale Eigenschaften der Funktion (Spitzen usw.) verschmiert werden. Grenander (1951) verglich dieses Verhalten mit der aus der Physik bekannten *Unschärferelation*. In der Statistik spricht man meist von *Bias-Variance-Dilemma*.
Verwendet man eine Trainings-Methode zur Optimierung der Parameter des Modells, so konvergiert der Bias-Fehler zu seinem Minimum, das Modell wird allerdings immer spezifischer in Hinblick auf $\mathcal{T}$, so dass der Varianz-Fehler steigt. Ultimativ wird die erwartete durchschnittliche Performance $\mathbb{E}_{\overline{\mathcal{T}}^N}[\Lambda]$ somit ab einem gewissen Punkt wieder ansteigen, man spricht in diesem Fall von *Overtraining*.

4.2.2 Konsistenz

Für die Praxis ist es sehr wichtig zu wissen, wie sich der optimale Parameter $\hat{w}_{\mathcal{T}^N}$ für eine immer größer werdende Stichprobe, also $N \to \infty$, verhält. Um dieser Frage nach zu

gehen müssen wir das in den vorangegangenen Abschnitten formulierte Optimierungsproblem auch im stichprobenabhängigen, also diskreten Fall exakt formulieren. Bisher fassten wir die Situation derart auf, dass die Verteilung $\mathbb{P}_Z$ unbekannt ist und wir durch wiederholte Messungen $\boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, ..., N$, der Zufallsvariablen $Z = (X, Y)$ empirische Informationen über $\mathbb{P}_Z$ sammeln können. Die $\boldsymbol{z}_i$ sind eigentlich Realisierungen der Zufallsvariablen $Z_i = (X_i, Y_i)$, wobei $(X_i)_{i=1}^N$ eine Folge unabhängig, identisch verteilter Zufallsvariablen ist und $(Y_i)_{i=1}^N$ die zugehörige Versuchsserie. Ausgehend von Glg. (4.3) definieren wir als Stichproben-Analogon zu $\mathbb{P}_1 \otimes \mathbb{P}_2$ ein zählendes Maß

$$\mu(S) := \frac{\text{Anzahl } \boldsymbol{z}_i \in S}{N}$$

für alle $(Z_1, ..., Z_N) \in (\Omega_1 \times \Omega_2)^N$ und $S \subset \Omega_1 \times \Omega_2$. Das Risiko-Funktional aus (4.3) wird nun für N Samples durch den (erwartungstreuen) Schätzer

$$\Lambda_{\Xi,\hat{f}}(w) \approx \Lambda^N_{\Xi,\hat{f}}(w) := \int_{\mathbb{R}^d \times \mathbb{R}^m} \Xi\left(\boldsymbol{y}, \hat{f}(\boldsymbol{x}, w)\right) \, \mathrm{d}\mu_Z(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N} \sum_{i=1}^N \Xi\left(\boldsymbol{y}_i, \hat{f}(\boldsymbol{x}_i, w)\right)$$

approximiert, den wir als *empirisches Risiko-Funktional* bezeichnen. Zu lösen ist dann

$$\hat{w}_{\mathcal{T}^N} = \operatorname*{argmin}_{w \in W} \Lambda^N_{\Xi,\hat{f}}(w) \,,$$

was wir zuweilen in Anlehnung an Vapnik (1999) verkürzt als *ERM-Problem* (empirische Risiko-Minimierung) bezeichnen. Es ist an dieser Stelle aber wichtig anzumerken, dass dieses Minimum in W nicht existieren muss. Das Infimum von $\Lambda^N_{\Xi,\hat{f}}(w)$ hingegen existiert immer, so dass wir wenn wir von der empirischen Risiko-Minimierung (ERM) sprechen, immer "fast"-Minimierung in folgendem Sinne meinen: Wähle $\hat{w}^\varepsilon_{\mathcal{T}^N} \in W$, so dass für jedes $\varepsilon > 0$

$$\Lambda^N_{\Xi,\hat{f}}(\hat{w}^\varepsilon_{\mathcal{T}^N}) \leq \inf_{w \in W} \Lambda^N_{\Xi,\hat{f}}(w) + \varepsilon \,.$$

Wir verallgemeinern und betrachten nun die Versuchsserie $(X_i, Y_i)_{i=1}^N \in (\Omega_1 \times \Omega_2)^N$ und die Zufallsvariable $\hat{W}_{\mathcal{T}^N}$, deren Realisierung $\hat{w}_{\mathcal{T}^N}$ ist. D.h. $\hat{W}_{\mathcal{T}^N} : (\Omega_1 \times \Omega_2)^N \to W$

$$\hat{W}_{\mathcal{T}^N} = \operatorname*{argmin}_{w \in W} \widetilde{\Lambda}^N_{\Xi,\hat{f}} \,,$$

wobei $\widetilde{\Lambda}^N_{\Xi,\hat{f}} : (\Omega_1 \times \Omega_2)^N \times W \to \mathbb{R}$,

$$\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot, w) = \frac{1}{N} \sum_{i=1}^N \Xi\left(Y_i, \hat{f}(X_i, w)\right) \,,$$

die zu $\Lambda^N_{\Xi,\hat{f}}(w)$ gehörige Zufallsvariable ist, deren Realisierung durch

$$\widetilde{\Lambda}^N_{\Xi,\hat{f}}((\omega_{11},\omega_{12}),...,(\omega_{N1},\omega_{N2})),w) = \frac{1}{N}\sum_{i=1}^{N}\Xi\left(Y_i(\omega_{i2}),\hat{f}(X_i(\omega_{i1}),w)\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\Xi\left(\boldsymbol{y}_i,\hat{f}(\boldsymbol{x}_i,w)\right)$$

gegeben ist. Die Zufallsvariable

$$\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,\hat{w}_{\mathcal{T}^N}) = \frac{1}{N}\sum_{i=1}^{N}\Xi(Y_i,\hat{f}(X_i,\hat{w}_{\mathcal{T}^N}))$$

bezeichnet konsequenterweise das Minimum des empirischen Risikos für die Stichprobe $\mathcal{T}$.

Fassen wir den Messprozess als Folge von Zufallsvariablen in N, also der Stichprobenlänge auf, so können wir Standard-Statistik-Werkzeuge wie z.B. das Gesetz der großen Zahlen oder den zentralen Grenzwertsatz anwenden. Es gilt allerdings zwischen verschiedenen Konsistenzbegriffen zu unterscheiden und wir müssen definieren in welchem Sinne Konvergenz zu verstehen ist. Neben der deterministischen Konvergenz (sei $\{a_n\} = (a_1, a_2, ...)$ ein Folge, dann konvergiert a_n gegen a, falls es ein $a \in \mathbb{R}$ gibt, so dass für jedes $\varepsilon > 0$ eine Zahl $N \in \mathbb{N}$ existiert, so dass $|a_n - a| < \varepsilon$ für alle $n \geq N$) benötigen wir die Begriffe der fast-sicheren Konvergenz (sei $\{A_n\}$ eine Folge von reellen Zufallsvariablen, dann konvergiert A_n gegen A fast sicher (geschrieben $A_n \xrightarrow{f.s.} A$) bzgl. $\mathbb{P}$, falls $\mathbb{P}[A_n \xrightarrow{n\to\infty} A] = \mathbb{P}[\{\omega \in \Omega : A_n(\omega) \xrightarrow{n\to\infty} A(\omega)\}] = 1$) und der stochastischen Konvergenz (sei $\{A_n\}$ eine Folge von reellen Zufallsvariablen, dann konvergiert A_n gegen A stochastisch, wenn für alle $\varepsilon > 0$ $\mathbb{P}[|A_n - A| < \varepsilon] \xrightarrow{n\to\infty} 1$). Die fast-sichere Konvergenz wird gelegentlich auch als starke Konvergenz bezeichnet. Ein vierter Konvergenzbegriff ist die schwache Konvergenz, auch Konvergenz in Verteilung (man sagt $A_n \xrightarrow{n\to\infty} A$, wenn dasselbe für die Verteilungen gilt, also $\mathbb{P}_{A_n} \xrightarrow{n\to\infty} \mathbb{P}_A$ punktweise).

Es stellen sich nun folgende für die "Theorie des Lernens" grundlegende Fragen:

(1) Ist das ERM-Problem konsistent? D.h.

$$\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}^N}) \xrightarrow{f.s.} \Lambda_{\Xi,\hat{f}}(\hat{w})\,,$$
$$\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w) \xrightarrow{f.s.} \Lambda_{\Xi,\hat{f}}(\hat{w}) \quad ?$$

Die Konvergenz bezieht sich hierbei auf $N \to \infty$ und das Wahrscheinlichkeitsmaß $\mathbb{P}^N = (\mathbb{P}_1 \otimes \mathbb{P}_2)^N$. Die erste Konvergenz sagt aus, dass die für jede Stichprobenlänge gefundenen Lösungen $\hat{w}_{\mathcal{T}^N}$ (bzw. $\hat{W}_{\mathcal{T}^N}$) mit Wahrscheinlichkeit 1 gegen die bestmögliche Lösung $\hat{w}$ konvergieren. Die zweite Konvergenz besagt, dass die Werte des empirischen Risikos gegen das minimale Risiko streben mit wachsender Stichprobenlänge.

(2) Wie schnell konvergiert die Folge von kleinsten empirischen Risiken gegen das minimale Risiko?

Natürlich können wir uns die Frage stellen, warum wir überhaupt die Konsistenz unserer Lösungen untersuchen, denn im Endeffekt werden in jeglicher Anwendung immer endlich viele Trainingsdaten vorliegen. Die Antwort ist, dass eine nicht-konsistente Theorie der Risiko-Minimierung völlig unvorhersehbare Ergebnisse bei Veränderung der Stichprobenlänge liefern kann, somit ist Konsistenz eine notwendige Bedingung für einen sinnvollen Optimierungsalgorithmus.

Zunächst zu Frage (1). Folgendes Theorem beweist, dass eine Lösung des Optimierungsproblems für eine unendliche Stichprobe existiert:

Satz 4.1. *Sei* $(\Omega, \mathcal{F}, \mathbb{P})$ *ein vollständiger Wahrscheinlichkeitsraum und* $\mathcal{T}^\infty := \{Z_i\}_{i=1}^\infty := \{Z_i : \Omega \to \mathbb{R}^g; i = 1, 2, ...\}$, $g \in \mathbb{N}$ [5], *eine Folge von identisch verteilten Zufallsvariablen und enstprechend* $\mathcal{T}^N := \{Z_i\}_{i=1}^N := (Z_i : \Omega \to \mathbb{R}^g; i = 1, 2, ..., N)$ *eine Teil-Stichprobe der Länge* N. *Sei weiterhin* $\Xi : \mathbb{R}^g \times W \to \mathbb{R}$ *eine Zufallsvariable, die für alle* $w \in W$ *mit* $W \subset \mathbb{R}^s$ *kompakt,* $s \in \mathbb{N}$, *Bor*$(\mathbb{R}^g)$*-messbar sei. Zusätzlich sei* Ξ *für alle* $\boldsymbol{z} \in \mathbb{R}^g$ *stetig auf* W. *Weiterhin existiere eine auf* W *integrierbare Zufallsvariable* $d : \mathbb{R}^g \to \mathbb{R}$, *die* Ξ *auf* W *dominiert, also für alle* $w \in W$ *gilt* $|\Xi(\boldsymbol{z}, w)| \leq d(\boldsymbol{z})$ *sowie* $\mathbb{E}[d(Z_i)] < \infty$ [6]. *Dann existiert für alle* $N \in \mathbb{N}$ *eine Lösung* $\hat{W}_{\mathcal{T}^N}$ *des Problems*

$$\operatorname*{argmin}_{w \in W} \widetilde{\Lambda}^N_{\Xi, \hat{f}}(\cdot, w) = \operatorname*{argmin}_{w \in W} \frac{1}{N} \sum_{i=1}^{N} \Xi\left(Y_i, \hat{f}(X_i, w)\right)$$

und es gilt[7]

$$\hat{W}_{\mathcal{T}^N} \xrightarrow{f.s.} D\,,$$

sowie

$$\sup_{w \in W} \left| \widetilde{\Lambda}^N_{\Xi, \hat{f}}(\cdot, w) - \Lambda_{\Xi, \hat{f}}(\hat{w}) \right| \xrightarrow{f.s.} 0\,,$$

wobei $D := \{\hat{w} \in W : \Lambda_{\Xi, \hat{f}}(\hat{w}) \leq \Lambda_{\Xi, \hat{f}}(w)\ \forall w \in W\}$.

Beweis. White (1989) gibt schon einen Beweis dieses Satzes. In Ljung (1989) findet sich ein ähnlicher Satz, allerdings nur für die spezielle Wahl $\Xi = (Y - \hat{f}(X, w))^2$ und *lineare* Modelle. In Ljung (1978) werden die Resultate verallgemeinert, aber unter geänderten

[5] Für $Z_i = (X_i, Y_i)$ wie bislang verwendet ist $g = m + d$.

[6] Man beachte wie immer: $d(\boldsymbol{z}) = d(Z(\omega))$ für das Ereignis $\omega \in \Omega$. Die $\hat{f}$-Abhängigkeit von Ξ lassen wir hier weg und schreiben $\Xi(Z, w)$ statt $\Xi(Y, \hat{f}(X, w))$.

[7] "fast sicher" verstehe man nun so: $\mathbb{P}[\{\omega \in \Omega : \hat{W}_{\mathcal{T}^n}(\omega) \xrightarrow{n \to \infty} D\}] = 1$, wobei Konvergenz in eine Menge hinein zu verstehen ist als $\inf_{w' \in D} \|\hat{w}_{\mathcal{T}^N} - w'\| \xrightarrow{N \to \infty} 0$ z.B. in Euclidischer Norm.

Forderungen für die Stichprobe. Der Beweis ist aber auch im allgemeinen Fall lediglich ein $\varepsilon/3$-Argument.
$\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot, w)$ ist stetig auf einem Kompaktum für jede Folge von Realisierungen $\{\boldsymbol{z}_i\}_{i=1}^N$ von Z, $N = 1, 2, ...$ Hieraus folgt sofort die Existenz von $\hat{W}_{\mathcal{T}^N}$. Durch die Existenz einer dominierenden Funktion d und die Kompaktheit von W lässt sich auch direkt auf die Stetigkeit von Λ selbst schließen (Stetigkeit des Integrals (Erwartungswert), s. ein Buch über Maßtheorie, z.B. Werner (2007)). Aus dem Gesetz der großen Zahlen (s. z.B. Meintrup & Schäffler (2005)) ergibt sich sofort (die Zufallsvariablen sind i.i.d.)

$$\sup_{w \in W} \left| \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot, w) - \Lambda_{\Xi,\hat{f}}(w) \right| \xrightarrow{f.s.} 0 \,.$$

Wähle nun in diesem Sinne eine konkrete Realisierung $\{\boldsymbol{z}_i\}_{i=1}^N$ und sei $\{\hat{W}_{\mathcal{T}^N}\}_{i=1}^N$ die Folge der Minimierungs-Parameter für das jeweilige $\{\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot, w)\}_{i=1}^N$, $N = 1, 2, ...$ Weil W als kompakt angenommen wurde, gibt es einen Punkt $\overline{w} \in W$ und eine Teilfolge $\{\hat{W}_{\mathcal{T}^{N_k}}\}_{k=1}^n$, die gegen $\overline{w}$ konvergiert. Durch die gleichmäßige Konvergenz in W und die Stetigkeit von Λ können wir weiterhin folgern:

$$\begin{aligned}\left| \widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, \hat{w}_{\mathcal{T}^{N_k}}) - \Lambda_{\Xi,\hat{f}}(\overline{w}) \right| \leq \left| \widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, \hat{w}_{\mathcal{T}^{N_k}}) - \Lambda^{N_k}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}^{N_k}}) \right| \\ + \left| \Lambda^N_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}^{N_k}}) - \Lambda_{\Xi,\hat{f}}(\overline{w}) \right| \leq 2\varepsilon\end{aligned}$$

für alle $\varepsilon > 0$ und N_k groß genug. Es gilt weiterhin unter dieser Voraussetzung:

$$\begin{aligned}\Lambda_{\Xi,\hat{f}}(\overline{w}) - \Lambda_{\Xi,\hat{f}}(w) = \left[\Lambda_{\Xi,\hat{f}}(\overline{w}) - \widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, \hat{w}_{\mathcal{T}^{N_k}}) \right] + \left[\widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, \hat{w}_{\mathcal{T}^{N_k}}) - \widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, w) \right] \\ + \left[\widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, w) - \Lambda_{\Xi,\hat{f}}(w) \right] \\ \leq 3\varepsilon \,,\end{aligned}$$

weil zusätzlich zu der zuvor verwendeten Dreiecksungleichung durch die Optimalität von $\hat{w}_{\mathcal{T}^{N_k}}$ $\widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, \hat{w}_{\mathcal{T}^{N_k}}) - \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot, w) \leq \varepsilon$ und durch die gleichmäßige Konvergenz auch $\widetilde{\Lambda}^{N_k}_{\Xi,\hat{f}}(\cdot, w) - \Lambda_{\Xi,\hat{f}}(w) \leq \varepsilon$. Diese Argumente gelten für beliebiges $w \in W$, d.h. $\overline{w} \in D$. Da ε beliebig gewählt wurde ist $\Lambda_{\Xi,\hat{f}}(\overline{w}) \leq \Lambda_{\Xi,\hat{f}}(w)$. Und zuletzt, weil auch die Folge $\{\hat{W}_{\mathcal{T}^N}\}_{i=1}^N$ beliebig ist, liegt jeder zu einer dieser Folgen gehörende Punkt $\overline{w}$ in D. Nehme nun an, dass $\inf_{\hat{w} \in D} \|\hat{w}_{\mathcal{T}^N} - \hat{w}\| \nrightarrow 0$. Dann existiert ein $\varepsilon > 0$ und eine Teilfolge $\{\hat{W}_{\mathcal{T}^{N_k}}\}_{k=1}^N$, so dass $\|\{\hat{w}_{\mathcal{T}^{N_k}}\}_{k=1}^N - \hat{w}\| \geq \varepsilon$ für alle N_k und alle $\hat{w} \in D$. Aber die Folge $\{\hat{W}_{\mathcal{T}^{N_k}}\}_{k=1}^N$ hat nach den vorstehenden Ausführungen einen Grenzpunkt $\overline{w}$, der in D liegt. Dies ist ein Widerspruch zu der Annahme $\inf_{\hat{w} \in D} \|\hat{w}_{\mathcal{T}^N} - \hat{w}\| \nrightarrow 0$. Es folgt die Behauptung. ■

Anmerkung 4.1. Entscheidend ist die Kompaktheit von W. Wir werden diesen Aspekt in Abschnitt 4.4 näher beleuchten.

Ein Neuronales Netz besitzt die Eigenschaft der universellen Approximation, d.h. jede Funktion aus L^2 kann prinzipiell beliebig genau approximiert werden. Zusammen mit Satz 4.1 ist nun endlich gesichert, dass die Methode eine beliebige Funktion durch Anlernen eines Neuronalen Netzwerkes (welcher Natur auch immer) auf Grundlage einer endlichen Zahl von Zufallsvariablen überhaupt von Erfolg gekrönt sein kann.

Wir kommen nun zu Frage (2). Die prinzipielle Konvergenz der Zufallsvariablen $\hat{W}_{\mathcal{T}^N}$ gegen die optimale Lösung lässt noch die Frage nach der Konvergenzgeschwindigkeit und der Grenzverteilung offen. In Satz 4.1 wurde Ξ als stetig auf W angenommen, eine starke, aber nötige Einschränkung. Auch für die folgenden Überlegungen muss Ξ hinreichend glatt sein in W, insbesondere benötigen wir nun auch mehrfache Differenzierbarkeit in bestimmtem Rahmen. Die Grundidee ist es, $\Lambda^N_{\Xi,\hat{f}}(w)$ in eine Taylorreihe um $\Lambda^N_{\Xi,\hat{f}}(\hat{w})$ zu entwickeln. Wir beschränken uns zunächst auf die Analyse in Bezug auf w, d.h. alle Ableitungen der Form $\partial\Lambda/\partial\Xi$ und $\partial\Lambda/\partial\hat{f}$ mögen verschwinden. Dieser Punkt ist wichtig, denn er impliziert, dass wir die *Struktur* des Modells "festhalten". Wir schreiben die Taylor-Reihe in den zugehörigen Zufallsvariablen:

$$\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w) = \Lambda_{\Xi,\hat{f}}(\hat{w}) + \left(\nabla_w \left.\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}_{\mathcal{T}^N}}\right)^T (w-\hat{w}_{\mathcal{T}^N}) + \frac{1}{2}(w-\hat{w}_{\mathcal{T}^N})^T \left.\nabla^2_w \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}_{\mathcal{T}^N}} (w-\hat{w}_{\mathcal{T}^N}) + \ldots$$

Hierbei bezeichnet ∇ den s-dimensionalen Gradienten und ∇^2 die $s \times s$-dimensionale Hesse-Matrix. Der zweite Term auf der rechten Seite verschwindet, weil $\widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)$ in $\hat{w}_{\mathcal{T}^N}$ ein Minimum hat. Wir nehmen zusätzlich an, dass sich die Menge D auf einen Punkt reduziert, d.h. das Minimum sei eindeutig. Wir könnten diese Annahme auch so verstehen, dass wir ein lokales Minimum (einen isolierten Punkt im Inneren von W) betrachten. Umgestellen und beide Seiten nach w an der Stelle des "wirklichen" Minimums $w=\hat{w}$ differenzieren (alle Ableitungen mögen existieren) ergibt:

$$\begin{aligned}\left.\nabla_w \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}} &\approx \frac{1}{2}\nabla_w \left\{ (w-\hat{w}_{\mathcal{T}^N})^T \left.\nabla^2_w \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}_{\mathcal{T}^N}} (w-\hat{w}_{\mathcal{T}^N}) \right\}\Bigg|_{w=\hat{w}} \\ &= \left.\nabla^2_w \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}_{\mathcal{T}^N}} (\hat{w}-\hat{w}_{\mathcal{T}^N}) \ ,\end{aligned}$$

also, falls die Hesse-Matrix nicht-singulär ist,

$$\hat{w}-\hat{w}_{\mathcal{T}^N} \approx \left.\nabla_w \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}} \left(\left.\nabla^2_w \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}_{\mathcal{T}^N}}\right)^{-1} . \tag{4.8}$$

Wir betrachten nun große N, wobei sich der formale Beweis, dass

$$\left.\nabla^2_w \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot,w)\right|_{w=\hat{w}_{\mathcal{T}^N}} \xrightarrow{\text{f.s.}} \left.\nabla^2_w \Lambda_{\Xi,\hat{f}}\right|_{w=\hat{w}}$$

aus Satz 4.1 ergibt (die Elemente der Matrix $\nabla^2 \Xi$ müssen alle auf W durch eine integrierbare Funktion dominiert werden). Es ist einsichtig, dass wir hinreichende Glattheit (mindestens 2-fach stetig differenzierbar) für Ξ fordern müssen, damit die Ableitung in das Integral des Erwartungswertes "hineingezogen" werden darf. Wir schreiben den ersten Term auf der rechten Seite von 4.8 unter diesen Voraussetzungen aus:

$$\nabla_w \left. \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot, w) \right|_{w=\hat{w}} = \frac{1}{N} \sum_{i=1}^{N} \nabla_w \Xi \left(Y_i(\cdot), \hat{f}(X_i(\cdot), w) \right) \Big|_{w=\hat{w}} .$$

Da Ξ bei $\hat{w}$ minimal ist verschwinden die Ausdrücke $\mathbb{E}[\nabla_w \Xi_i]$ für alle i in $w = \hat{w}$, wobei wir $\Xi_i = \Xi(Y_i, \hat{f}(X_i, w))$ schreiben. Es liegt also eine Summe von identisch verteilten Zufallsvariablen vor, falls (X_i, Y_i) für alle i identisch verteilt ist. Nimmt man nun noch an, dass die zugehörige Kovarianzmatrix

$$Q = \left(\nabla_w \left. \Lambda_{\Xi,\hat{f}} \right|_{w=\hat{w}} \right) \left(\nabla_w \left. \Lambda_{\Xi,\hat{f}} \right|_{w=\hat{w}} \right)^T$$

nicht-singulär ist, so, ergibt sich mit dem zentralen Grenzwertsatz sofort:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nabla_w \Xi \left(Y_i(\cdot), \hat{f}(X_i(\cdot), w) \right) \Big|_{w=\hat{w}} \xrightarrow{\text{d}} N(0, Q) ,$$

wobei $N(0, Q)$ die Normalverteilung mit Erwartungswert 0 und Kovarianzmatrix Q bezeichne. Zurück in Glg. (4.8) ergibt sich also:

$$\sqrt{N} \left(\hat{W}_{\mathcal{T}^N} - \hat{w} \right) \xrightarrow{\text{d}} N(0, C) \tag{4.9}$$

mit der asymptotischen Kovarianzmatrix $\hat{W}_{\mathcal{T}^N}$

$$C := \left(\nabla^2_w \left. \Lambda_{\Xi,\hat{f}} \right|_{w=\hat{w}} \right)^{-1} Q \left(\nabla^2_w \left. \Lambda_{\Xi,\hat{f}} \right|_{w=\hat{w}} \right)^{-1} .$$

4.3 Modellkomplexität

In diesem Abschnitt beginnen wir eine detaillierte Untersuchung das Bias-Variance-Dilemmas. Zu diesem Zweck präsentieren wir zunächst einen alternativen Zugang zu Fragen der Konsistenz von Schätzern innerhalb der Struktur eines Modells auf, der sehr interessante Einblicke in die tieferliegenden statistischen Zusammenhänge der "Theorie des Lernens" liefert und am Ende auf *verteilungsunabhängige* Konsistenzbedingungen führt. Wir haben in Satz 4.1 gesehen, dass die Lösbarkeit des ERM-Problems eine Einschränkung des Hypothesenraumes voraussetzt. Wir führen diese Überlegungen nun weiter und liefern eine erste exakte Definition des Begriff der Modellkomplexität und stellen den Zusammenhang von Komplexität und Konsistenz her.

Zunächst aber ein Hinweis zur Notation: Falls der Hypothesenraum durch einen Gewichtsparameter $w \in W$ parametrisiert werden kann, also $\mathscr{F} = \{f = f(X, w) \in \mathcal{C} \subset C : w \in W\}$, so wechseln wir je nach Bedarf zwischen den Notationen $\Xi(Z, w)$, $\Xi(Z, f)$ und $\Xi(Y, f(X, w))$. Analog für $\Lambda_\Xi(f)$ statt $\Lambda_{\Xi,f}(w)$ etc. Also gilt z.B.

$$\sup_{w \in W} \left| \widetilde{\Lambda}^N_\Xi(\cdot, w) - \Lambda_\Xi(w) \right| \equiv \sup_{f \in \mathscr{F}} \left| \widetilde{\Lambda}^N_\Xi(\cdot, f) - \Lambda_\Xi(f) \right| ,$$

wobei

$$\widetilde{\Lambda}^N_\Xi(\cdot, w) \equiv \widetilde{\Lambda}^N_\Xi(\cdot, f) = \frac{1}{N} \sum_{i=1}^{N} \Xi\left(Z_i, f\right) .$$

Klassische Modellauswahl-Kriterien (Informationskriterien) beschreiben das geschätzte Risiko als Funktion des empirischen Risikos:

$$\Lambda_\Xi(f) \approx h(k, N) \frac{k}{N} \Lambda^N_\Xi(f) ,$$

wobei k die Anzahl der Freiheitsgrade des Modells bezeichnet. Bekannte Beispiele für die Funktion h sind der Akaike final prediction error (FPE) (Akaike (1973))

$$h(k, N) = \frac{N + k}{N - k} ,$$

das Schwarz-Kriterium (Schwarz (1978))

$$h(k, N) = 1 + \frac{\ln N}{2} \frac{k}{N - k}$$

und generalisierte Kreuzvalidierung (cross-validation) (Craven & Wahba (1979)) mit

$$h(k, N) = \left(\frac{N}{N - k} \right)^2 .$$

Einen verteilungs*unabhängigen* Ansatz verfolgt die SRM-Methode (Structural Risk Minimization), auf die im folgenden unser Hauptaugenmerk gerichtet ist.

Vapnik & Chervonenkis (1991) geben ein zu Satz 4.1 analoges Theorem mit schwächeren Voraussetzungen an, welches Konsistenz mit der gleichmäßigen Konvergenz der Folge der empirischen Risiken identifiziert:

Satz 4.2. *Sei $\Xi(Z, w)$, $w \in W$ eine Zufallsvariable im Sinne von Satz 4.1. Weiterhin sei*

$$A \leq \mathbb{E}\left[\Xi\right] = \int_{\Omega_1 \times \Omega_2} \Xi\left(Z, w\right) \; d(\mathbb{P}_1 \otimes \mathbb{P}_2) \leq B$$

für alle $w \in W$. Dann ist die gleichmäßige Konvergenz der Folge $\widetilde{\Lambda}_{\Xi}^{N}(\cdot, w)$ gegen das eigentliche Risiko $\Lambda_{\Xi}(w)$ für alle $w \in W$ äquivalent zur Konsistenz des ERM-Problems:

$$\forall\, \varepsilon > 0 \;:\; \lim_{N\to\infty} \mathbb{P}^N \left[\left\{ \omega \in (\Omega_1 \times \Omega_2)^N \;:\; \sup_{w\in W} \left| \widetilde{\Lambda}_{\Xi}^{N}(\omega, w) - \Lambda_{\Xi}(w) \right| > \varepsilon \right\} \right] = 0\,. \quad (4.10)$$

Anmerkung 4.2. Man bemerke allerdings, dass dieser Satz nicht von fast-sicherer Konvergenz spricht, sondern nur von Konvergenz in Wahrscheinlichkeit. Im Gegensatz zu Satz 4.1 wird nämlich keine Kompaktheit des Hypothesenraumes gefordert. Es lässt sich aber unter Benutzung eines in Steele (1978) gezeigten Theorems auch die *fast-sichere* gleichmäßige Konvergenz von $\widetilde{\Lambda}_{\Xi}^{N}(\cdot, w)$ gegen das kontinuierliche Risiko unter den Voraussetzungen von Satz 4.2 beweisen. Steele benutzt hierfür ein Theorem von Kingman (1973) für ergodische Prozesse. Zusammen mit den Erweiterungen von Satz 4.2 in Vapnik & Chervonenkis (1981) auf Funktionen mit kompaktem Definitionsbereich lässt sich Steeles Argument auf den allgemeinen Fall übertragen (s. hierzu auch Talagrand (1987)).

Anmerkung 4.3. In den Sätzen bezüglich der Konsistenz des ERM-Problems in diesem Abschnitt und im darauf folgenden 4.4 wird stets die Beschränktheit von Ξ voraussgesetzt. Man vergleiche auch mit den Voraussetzungen in Satz 4.1. In Erinnerung an das Theorem von Arzelà-Ascoli (s. Satz 6.1) bemerkt man sofort, dass diese Bedingung durch die relative Kompaktheit von $\mathrm{Im}(\mathscr{F})$ ersetzt werden kann. Anders ausgedrückt muss die Zufallsvariable $f(X) - Y$ gleichmäßig beschränkt sein auf $\mathscr{F}$, bzw.

$$\mathbb{P}_1 \otimes \mathbb{P}_2 \left[|f(X) - Y| \le B \right] = 1$$

mit $0 < B < \infty$. Ist $\mathrm{Im}(\mathscr{F})$ beschränkt durch eine Zahl B, so folgt sofort

$$\mathbb{P}_1 \otimes \mathbb{P}_2 \left[|f(X) - Y| \le 2\max\{B, B'\} \right] = 1\,,$$

wobei $|f(x)| \le B'$ für alle $f \in \mathscr{F}$ und $x \in X$. Die Frage ist nun aber, unter welchen Voraussetzungen $|Y|$ beschränkt ist. Hier ist folgendes Theorem fundamental, das sich allerdings auf Performance-Funktionen der Form $\Xi = |f(X) - Y|^p$, $0 < p < \infty$ bezieht:

Satz 4.3. *Sei $\Xi = |f(X) - Y|^p$, $0 < p < \infty$. Falls*

$$\sup_{w\in W} \left| \left(\widetilde{\Lambda}_{\Xi}^{N}(\omega, w) \right)^{1/p} - \left(\Lambda_{\Xi}(w) \right)^{1/p} \right| \xrightarrow{f.s.} 0$$

für jede Verteilung von $Z = (X, Y)$, so dass Y mit Wahrscheinlichkeit 1 beschränkt ist, dann gilt

$$\mathbb{E}_{\overline{\mathcal{T}}} \left[\left| \hat{f}(X, \hat{w}_{\mathcal{T}^N}) - Y \right|^p \right] - \inf_{w\in W} \mathbb{E}\left[\left| f(X, w) - Y \right|^p \right] \xrightarrow{f.s.} 0$$

für jede Verteilung von Z, so dass insbesondere $\mathbb{E}[|Y|^p] < \infty$.

Beweis. Für einen Beweis in anderer Notation aber derselben Grundaussage siehe Lugosi & Zeger (1995).

Vapnik & Chervonenkis (1991) sowie Vapnik (1999) geben weiterführende Theoreme, die die Konsistenz des ERM-Problems mit der so genannten VC-Entropie (Vapnik-Chervonenkis-Entropie) $H^W(\mathcal{T}^N)$ verknüpfen. Diese Entropie ist definiert als der Logarithmus einer Größe $N^W(\mathcal{T}^N)$, die die Anzahl der möglichen Vektoren $\xi(w) := (\Xi(Z_1, w), ..., \Xi(Z_N, w))$, $w \in W$, angibt. H^W beschreibt also die "Verschiedenheit" der Performance-Maße $\Xi_i = \Xi(Z_i, w)$ im Raum der Gewichte W. Liegt Ξ zwischen Werte A und B, so beschreibt N die Anzahl der möglichen Ξ-Werte innerhalb eines N-dimensionalen Würfels der Kantenlänge $B - A$ auf Basis der verschiedenen Stichproben der Länge N. Vapnik & Chervonenkis (1991) zeigen nun, dass das ERM-Problem genau dann konsistent ist, wenn

$$\lim_{N\to\infty} \frac{H^W(\mathcal{T}^N)}{N} = 0\,.$$

Jede maschinelle Lernmethode muss diese Bedingung erfüllen wenn eine konsistente Minimierung des empirischen Risikos erreicht werden soll. Dieses wird als der erste Meilenstein der Lerntheorie bezeichnet. In Bezug auf die Konvergenzrate wird gezeigt, dass die Bedingung

$$\lim_{N\to\infty} \frac{G^W(\mathcal{T}^N)}{N} = 0\,,$$

wobei $G^W(N) = \ln \sup_{\mathcal{T}^N} N^W(\mathcal{T}^N)$ die "growth-function" der Menge der Performance-Maße $S := \{\Xi(Z, w) : w \in W\}$ bezeichnet[8], äquivalent zur Konsistenz des ERM-Problems ist. Aus dieser Bedingung folgt nun eine Konvergenzrate für die kontinuierlichen Risiken: Für $N > N_0$ und $c > 0$ gilt

$$\Lambda_\Xi(\hat{w}_{\mathcal{T}^N}) - \Lambda_\Xi(\hat{w}) < \exp\left(-c\,\varepsilon^2 N\right)\,. \tag{4.11}$$

Den Beweis werden wir in Abschnitt 4.4 führen und diese Aussage für den Spezialfall der quadratischen Performance-Funktion und eines kompakten sowie konvexen Hypothesenraumes präzisieren.

4.3.1 Klassische Komplexitätskontrolle

Das ERM-Problem des vorangegangenen Abschnitts betrachtet das Verhalten des Schätzers für große Stichprobenlängen. Unser Fokus liegt aber immer auf der für die Praxis sehr viel relevanteren Situation von relativ kleinen N. Um diese Fälle zu untersuchen zitieren wir zunächst sinngemäß folgenden Satz aus Vapnik (1999), im Zuge dessen die *Vapnik-Chervonenkis-Dimension* definiert wird:

[8] Zur Erinnerung: Für uns ist in der Regel $\Xi(Z, w) = \Xi(Y, \hat{f}(X, w))$ mit $\hat{f} \in \mathscr{F}$. Wir könnten in diesem Fall also auch schreiben $S = S_{\mathscr{F}} = \{\Xi(Y, \hat{f}(X, w)) : \hat{f} \in \mathscr{F}\}$.

Satz 4.4. *Für jede growth-function gilt entweder* $G^W(N) = N \ln 2$ *oder*

$$G^W(N) < h\left(\ln\frac{N}{h} + 1\right) ,$$

wobei $h \in \mathbb{N}$ *die VC-Dimension von* $\{\Xi(Z,w) : w \in W\}$ *bezeichne, für die* $G^W(h) = h > \ln 2$ *und* $G^W(h+1) \neq (h+1)\ln 2$ *gilt.*

Die growth-function ist also entweder linear oder durch eine logarithmische Funktion beschränkt. Man sagt die VC-Dimension sei unendlich im ersten Fall und gleich $h < \infty$ im zweiten Fall. h kann als die maximale Anzahl von Datenpunkten aufgefasst werden, die durch die Funktionenklasse auf 2^N Weisen in 2 Gruppen aufgeteilt werden können. So kann man z.B. durch eine lineare Funktion 2 Datenpunkte auf 4 Weisen in 2 Gruppen aufteilen. Bei 3 Datenpunkten funktioniert dies ebenfalls. Aber 4 Datenpunkte lassen sich nicht mehr auf $2^4 = 16$ Möglichkeiten aufteilen, von daher ist die VC-Dimension der Klasse der linearen Funktionen in zwei Dimensionen $h = 3$.

Beispiel 4.1. Die VC-Dimension der Menge von linearen Indikatorfunktionen

$$\left\{\Xi(Z,w) = \theta\left[\sum_{i=1}^{M}\alpha_i Z_i + \alpha_0\right] : w \in W\right\}$$

ist gleich $M+1$, wobei $\theta(x) = 0$ für $x < 0$ und $\theta(x) = 1$ für $x \geq 1$.

Beispiel 4.2. Für eine Menge $\{\Xi(Z,w) : w \in W\}$ von reell-wertigen Funktionen mit $A \leq \Xi(Z,w) \leq B$ ist die VC-Dimension gleich der VC-Dimension der folgenden Menge von Indikatorfunktionen:

$$I(Z,w,b) := \{\theta\left[\Xi(Z,w) - b\right] : w \in W\} ,$$

wobei $A < b < B$ eine Konstante ist.

Beispiel 4.3. Die Menge der Polynome vom Grad m hat die VC-Dimension $m+1$.

Für den weiteren Fortgang der Diskussion ist es sinnvoll das folgende vereinfachte Bild zu verwenden:
Die VC-Dimension ist eine Zahl, welche die "Schwierigkeit" des Lernprozesses innerhalb einer Modellstruktur misst. Sie hängt von der Anzahl der Netzwerk-Parameter und der Größe des Netzwerkes ab.

Die Endlichkeit der VC-Dimension ist äquivalent zur Konsistenz des ERM-Problems, und zwar unabhängig von der Wahl des Wahrscheinlichkeitsmaßes. Je kleiner die VC-Dimension, desto weniger Trainings-Samples werden für die Aufstellung der Hypothese benötigt um die Wahrscheinlichkeit der korrekten Klassifizierung von neuen Daten konstant zu halten. Anders formuliert: Ein Hypothesenraum mit unendlicher VC-Dimension

ist nicht geeignet für eine statistische Lernmethode, denn es kann für jedes gegebene Daten-Sample eine Hypothese aufgestellt werden und nach Carl Popper ist eine "Wissenschaft", in der es keine Regeln gibt für die Falsifizierung von Hypothesen, nicht gerechtfertigt.

Aus den vorangegangenen Überlegungen folgt aus der Endlichkeit der VC-Dimension auch sofort die exponentielle Konvergenz des empirischen Risikos. In Vapnik (1998) findet sich folgender Satz speziell für positive und beschränkte Performance-Funktionen:

Satz 4.5. *Es sei für alle* $w \in W$

$$0 \leq \Xi(Z, w) \leq B \,,$$

die Menge der Performance-Maße $S = \{\Xi(Z, w) : w \in W\}$ *auf dem Gewichts-Raum* W *ist also gleichmäßig beschränkt. Weiterhin sei* h *die zu dieser Menge gehörige VC-Dimension. Dann gilt mit Wahrscheinlichkeit* $1 - \eta$ *für alle* $\Xi \in S$

$$\Lambda_\Xi(w) \leq \Lambda_\Xi^N(w) + \frac{B\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4\Lambda_\Xi^N(w)}{B\varepsilon}}\right) \tag{4.12}$$

mit

$$\varepsilon = \frac{4}{N}\left[h\left(\ln\frac{2N}{h} + 1\right) - \ln\eta\right] \,.$$

Man bemerke, dass diese Abschätzung *unabhängig* von der Verteilungsfunktion von Z ist, also in jeder praktischen Situation gilt. Formt man den Ausdruck aus Satz (4.13 noch etwas um, so ergibt sich

$$\Lambda_\Xi^N(w) - \sqrt{\frac{h\ln\frac{2eN}{h} + \ln\frac{4}{\eta}}{N}} \leq \Lambda_\Xi(w) \leq \Lambda_\Xi^N(w) + \sqrt{\frac{h\ln\frac{2eN}{h} + \ln\frac{4}{\eta}}{N}} \,, \tag{4.13}$$

für alle $w \in W$ (bzw. $\hat{f} \in \mathscr{F}$). Für große N/h wird das empirische Risiko $\Lambda_\Xi^N(w)$ nahe am eigentlichen Risiko $\Lambda_\Xi(w)$ liegen. Für kleine Werte N/h müssen in Ungleichung (4.13) allerdings $\Lambda_\Xi^N(w)$ und h *simultan* minimiert werden.

Diese Simultan-Minimierung ist Gegenstand der SRM-Theorie, die voraussetzt, dass auf der Menge $S := \{\Xi(Z, w) : w \in W\}$ eine Struktur liegt, so dass

$$S_1 \subset S_2 \subset \cdots S_M \cdots$$

mit $S_i = \{\Xi(Z, w) : w \in W_i\}$, $S^* = \cup_i S_i$ und $S^* \subset S$. Um zu garantieren, dass diese Methode sinnvoll ist fordert man zusätzlich für diese Struktur:

1) S^* liegt dicht in S,

2) die VC-Dimension h_i von S_i ist endlich für alle $i \in \{1, 2, ...\}$ und
3) $0 \leq \Xi(Z, w) \leq B_i$ für alle $w \in W_i$ (jedes Element der Struktur ist also gleichmäßig beschränkt).

Hinter dieser Struktur steht nichts anderes als die Überlegung, die möglichen Modellstrukturen anhand ihrer *Komplexität* zu ordnen. Wir könnten auch sagen wir ordnen die Modellstrukturen anhand ihrer VC-Dimension:

$$h_1 \leq h_2 \leq \cdots \leq h_M \leq \cdots .$$

Das SRM-Prinzip besagt nun folgendes:

1) Wähle den Unterraum S_M basierend auf den Trainingsdaten $\mathcal{T}^N$, wobei $M = M(N)$ eine noch nicht näher spezifizierte Funktion sei.
2) Wähle innerhalb dieses Raumes S_M jenes $\hat{w}_{M(N)} \in W_{M(N)}$ aus, so dass das zugehörige $\Xi(Z, \hat{w}_{M(N)})$ die rechte Seite von (4.13) minimiert.

Das SRM-Prinzip versucht also zwischen der *Approximationsqualität* und *Approxmationskomplexität* abzuwägen. Schritt 1) entspricht der *Modellauswahl*, Schritt 2) der *Parameterschätzung* innerhalb der gewählten Modellkomplexität.

In Vapnik (1998) wird folgendes Theorem bewiesen, das die Bedeutung des SRM-Prinzips unterstreicht und die universelle Konsistenz der Methode belegt:

Satz 4.6. *Sei $\hat{w}_{M(N)}$ über das SRM-Prinzip bestimmt. Dann konvergiert $\Lambda_{\Xi}^{N}(\hat{w}_{M(N)})$ für jede Verteilung von Z gegen die bestmögliche Lösung mit Wahrscheinlichkeit 1.*

Es bleibt allerdings die Frage nach $M(N)$. Folgendes Theorem stellt eine Bedingung an $M(N)$, die "asymptotische" Konvergenz garantiert. Man sagt die Zufallsvariablen $X_1, X_2, ...$ konvergieren mit asymptotischer Rate $V(i)$ gegen die Zufallsvariable X_0, falls eine Konstante C existiert, so dass für $i \to \infty$

$$V^{-1}(i)\,|X_i - X_0| \xrightarrow{f.s.} C\,.$$

Es gilt nun folgender Satz:

Satz 4.7. *Eine Menge S sei mit einer Struktur ausgestattet, die den Punkten 1) - 3) genügt. Weiterhin sei $\Xi(Z, \hat{w}_{M(N)})$ auf Grundlage des SRM-Prinzips bestimmt, wobei $M(N)$ so gewählt wird, dass*

$$\lim_{N\to\infty} \frac{B_{M(N)} h_{M(N)} \ln N}{N} = 0\,.$$

Dann konvergiert die Folge von zugehörigen Risiken $\Lambda_{\Xi}(\hat{w}_{M(N)})$ gegen die beste Lösung $\Lambda(\hat{w})$ mit asymptotische Rate:

$$V(N) = r_{M(N)} + B_{M(N)} \sqrt{\frac{h_{M(N)} \ln N}{N}} \,,$$

wobei r_M die Approximationsgüte bezeichnet:

$$r_M := \inf_{w \in W_{M(N)}} \int_{\Omega_1 \times \Omega_2} \Xi(Z, w) \; d(\mathbb{P}_1 \otimes \mathbb{P}_2) - \inf_{w \in W} \int_{\Omega_1 \times \Omega_2} \Xi(Z, w) \; d(\mathbb{P}_1 \otimes \mathbb{P}_2) \,.$$

Es gilt also nach einer geeigneten Wahl der Performance-Funktion Ξ (es muss für dieses Ξ eine Struktur auf S existieren, die 1) bis 3) erfüllt) eine Funktion $M(N)$ zu finden, die den Voraussetzungen dieses Satzes genügt. Dann garantiert das SRM-Prinzip eine konsistente Minimierung des empirischen Risikos auf die optimale Lösung.

Beispiel 4.4 (Quadratische Performance-Funktion). Als wichtigstes Beispiel betrachten wir wieder $\Xi = (Y - \hat{f})^2$. In diesem Fall gilt für das Risiko-Funktional mit Wahrscheinlichkeit $1 - \eta$

$$\Lambda_\Xi(w) \leq \Lambda_\Xi^N(w) \cdot \underbrace{\left(1 - c\sqrt{\frac{h}{N} + \frac{h}{N} \ln\left(a\frac{h}{N}\right) - \frac{\ln \eta}{N}}\right)^{-1}}_{P\left(\frac{h}{N}, N\right)} \,, \tag{4.14}$$

wobei c eine Konstante ist, die von der Verteilung von Ξ abhängt (s. Vapnik (1999)). a ist eine weitere Konstante. Interessant ist, dass (4.14) *multiplikative Form* hat. P wird Penalty-Faktor genannt. Ebenfalls in Vapnik (1999) wird auf Grundlage von numerischen Experimenten gezeigt, dass die Wahl $a \approx 1$, $c \approx 1$ und $\eta = 1/\sqrt{N}$ sinnvoll ist.

Beispiel 4.5 (Quadratische Performance-Funktion & Polynominterpolation). Es sei wieder $\Xi = (Y - \hat{f})^2$. Als Approximatorenmenge wählen wir Polynome vom Grad m mit der Struktur $S_1 \subset S_2 \subset \cdots S_m \cdots$. S_m hat die VC-Dimension $m + 1$. Das Risiko-Funktional aus dem vorangehenden Beispiel vereinfacht sich dann zu:

$$\Lambda_\Xi(w) \leq \Lambda_\Xi^N(w) \cdot \left(1 - \sqrt{\frac{m+1}{N} - \frac{m+1}{N} \ln \frac{m+1}{N} + \frac{\ln N}{2N}}\right)^{-1} \,. \tag{4.15}$$

Wir kommen nun noch einmal auf das Bias-Variance-Problem (Minimierungsproblem für die quadratische Performance-Funktion) aus Abschnitt 4.2.1 zurück. In den vorangegangenen Abschnitten haben wir gezeigt, dass wenn der Funktionenraum S mit einer Struktur belegt ist, und über dieser Struktur eine Ordnung definiert werden kann bezüglich einer Komplexitätszahl M, so kann mit Hilfe der VC-Theorie die optimale Lösung ermittelt werden. Äquivalent hierzu kann auch der Raum der Approximanten $\mathscr{F}$ mit einer Ordnung versehen werden und diese impliziert eine Ordnung auf S, weil

$\Xi_M = \Xi(Y, \hat{f}_M)$, $\hat{f}_M \in \mathscr{F}_M$. Das Bias-Variance-Problem ist unter diesen Voraussetzungen nichts anderes als die Aufgabe M so zu bestimmen, dass $\Lambda_{\Xi,\hat{f}}(\hat{w})$ mit der Wahrscheinlichkeit $1-\delta$ minimal ist.

Wir fassen zusammen:
Prinzipiell besteht das Problem der Modell-Selektion anhand einer Komplexitätsanalyse aus drei Teilen:

1) Finde einen Komplexitätsindex, der eine Struktur auf dem Raum der Approximatoren induziert. Dies könnte z.B. die Anzahl der freien Parameter sein.
2) Das Vorhersage-Risiko muss anhand des empirischen Risikos abgeschätzt werden. Es wird das Modell gewählt, welches das Vorhersage-Risiko minimiert. Diese Aufgabe kann für nicht-lineare Schätzer sehr schwierig werden und führt in der Regel immer nur auf lokale Minima.
3) Die VC-Dimension des Hypothesenraumes muss bestimmt werden. Beispiel 4.1 zeigt, dass *lineare* Schätzer eine zu der Anzahl der freien Parameter proportionale VC-Dimension haben (für Polynome vom Grad n ist die VC-Dimension $n+1$). Für nicht-lineare Schätzer lässt sich h allerdings meist nicht geschlossen angeben, so dass die Modellauswahl anhand einer Komplexitätsanalyse im Stile von Vapnik in der Praxis sehr schwierig sein kann. I.A. ist die VC-Dimension bei nicht-linearen Schätzern größer als die Anzahl der freien Parameter.

Doch selbst wenn sich h gut bestimmt lässt kann die Komplexitätskontrolle in der Praxis durchaus versagen. In Cherkassky et al. (1999) wird eine konkrete Situation beschrieben, in denen der analytische Zugang über die VC-Komplexität schlechtere Resultate als eine Standardmethode wie cross-validation liefert (für Details s. Abschnitt 4.5). Die theoretische Überlegenheit einer Modellauswahl auf Grundlage einer VC-Komplexitäts-Kontrolle (im Spezialfall aus Beispiel 4.4) wurde allerdings in mehreren empirischen Studien gezeigt (siehe den soeben zitierten Artikel sowie Cherkassky & Shao (2001)).

4.3.2 Anwendung der VC-Theorie auf Neuronale Netze

Für Neuronale Netze wird wie schon angedeutet in der Regel als Performance-Funktion die quadratische Abweichung

$$\Xi_{LS}(Z, w) = \Big(Y - \hat{f}(X, w)\Big)^2$$

von den Messdaten verwendet, so dass das empirische Risiko-Funktional die folgende Form hat:

$$\Lambda^N(w) = \frac{1}{N} \sum_{i=1}^{N} \Big(\boldsymbol{y}_i - \hat{f}(\boldsymbol{x}_i, w)\Big)^2 \ .$$

Setzt man ein einzelnes "Neuron" als Schätzer an, also

$$\hat{f}(\boldsymbol{x}, w) = \widetilde{\Phi}\left(w^T \boldsymbol{x}\right) ,$$

$w \in \mathbb{R}^{d+1}, \boldsymbol{x} \in \mathbb{R}^d$, $x_0 = 0$, mit einer glatten sigmoiden Funktion (oder einer radialen Basisfunktion bzw. einem Wavelet) $\widetilde{\Phi}$, so ist das empirische Risiko-Funktional $\Lambda^N(w)$ glatt in w, weil σ hinreichend glatt ist. Somit kann ein gradienten-basiertes Verfahren zur Minimierung des empirischen Risikos verwendet werden, so dass im j-ten Iterationsschritt

$$w_{j+1} = w_j - \alpha_j \, \nabla_w \, \Lambda^N \big|_{w=w_j}$$

mit der Schrittweite α_j. Feedforward-Neuronale Netze verwenden dann eine Linearkombination von Neuronen gemäß Glg. (4.1), also

$$\hat{f}(\boldsymbol{x}, w) = \sum_{i=1}^{M} u_i \, \widetilde{\Phi}(\boldsymbol{a}_i, t_i; \boldsymbol{x}) \, , \tag{4.16}$$

mit $w \in W = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, i = 1, ..., M\}$. Rumelhart, Hinton & Williams (1986) führten den Backpropagation-Algorithmus als Verallgemeinerung des Gradientenverfahrens für einschichtige Neuronale Netze ein. Mit ihm werden alle Komponenten von w durch ein sukzessives Modifizieren (vom Output in Richtung Input zurückgehend, daher backpropagation) der Gewichte des Netzwerks bestimmt. Er ist der am weitesten verbreitete Lernalgorithmus für Neuronale Netze. Man sieht allerdings an der Struktur von $\hat{f}$ relativ leicht, dass das Minimierungsproblem recht schwer zu lösen sein wird:

1) $\Lambda^N(w)$ kann viele lokale Minima haben. So hängt die Güte des zuletzt gefundenen Minimums stark von den Startwerten des Algorithmus ab.
2) W ist ein hochdimensionaler Raum, hier $W \subset \mathbb{R}^{M(d+2)}$, so dass die Konvergenz des Backpropagation-Algorithmus sehr langsam sein kann.
3) Die Skalierungsfaktoren $\boldsymbol{a}_i$ beeinflussen die Approximationsqualität des Schätzers, da sie festlegen, wie sehr die Basisfunktion glättet. Die Frage aus welcher Grundmenge die $\boldsymbol{a}_i$ gewählt werden ist demnach direkt verknüpft mit der Approximationsgüte.

Wir wollen nun die SRM-Methode auf Neuronale Netze anwenden. Als Grundmenge liegt vor:

$$S = \{\Xi_{LS}(Z, w) : w \in W\}, \qquad W = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, i = 1, 2, ...\} \, .$$

Die Struktur auf S kann über M indiziert werden, d.h.

$$S_1 \subset S_2 \subset \cdots S_M \cdots$$

mit $S_M = \{\Xi_{LS}(Z,w) : w \in W_M\}$, $W_M = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, i = 1,2,...,M\}$. Somit ist $S^* = \cup_i S_i = S$ und liegt also auch dicht in S. Für beschränkte Basisfunktionen (wovon wir immer ausgehen) ist jedes $\Xi_{LS} \in S_M$ zudem gleichmäßig beschränkt, weil $\hat{f} \leq A \sum_{i=1}^{M} u_i \leq AM \max_i u_i$ wenn $\widetilde{\Phi}$ durch A beschränkt ist. Nun stellt sich noch die Frage nach der VC-Dimension von Approximatoren wie Neuronalen Netzen. Koiran & Sontag (1997) und Bartlett, Maiorov & Meir (1998) zeigen, dass die VC-Dimension für sigmoide Neuronale Netze superlinear mit der Anzahl der hidden layers steigt, Maass (1994) und Sakurai (1993) zeigen dies für SNNs mit fester layer-Anzahl (zwei- bzw. drei-layer sigmoide Netzwerke). Aus Maass (1994) stammt die Regel, dass für multi-layer Neuronale Netze mit einer Stufenfunktion als Aktivierungsfunktion $h \approx W \ln W$. Die Resultate für ein-layer-SNNs sind interessant in so fern, als dass einzelne Neuronen in der Regel eine mit der Anzahl der Modelparameter linear steigende VC-Dimension aufweisen (s. z.B. Haussler 1992 oder Anthony & Bartlett (1999)). Maass & Schmitt (1999) geben eine Ausnahme von dieser Regel für so genannte "spiking neurons".
Schmitt (2002) zeigt, dass Gauss'sche RBF-Netzwerke (also mit Gaussglocken als Basisfunktionen) eine VC-Dimension von *mindestens*

$$\left\lfloor \frac{k}{4} \right\rfloor \left\lfloor \log_2 \left(\frac{k}{4} \right) \right\rfloor \left(d - \left\lfloor \log_2 \left(\frac{k}{4} \right) \right\rfloor + 1 \right)$$

haben, wobei k die Anzahl von hidden nodes bezeichnet und d wie immer die Input-Dimension. Bezüglich der Gesamtanzahl von Netzwerke-Parametern W erfüllt die VC-Dimension für diese Netzwerke immer die Bedingung

$$h \geq \frac{W}{12} \log_2 \frac{k}{8} \, .$$

Dieser Artikel gibt noch weitere Sätze zur weiteren Eingrenzung der VC-Dimension für RBFNs und verwandte Netzwerke-Architekturen. Karpinski & Macintyre (1997) geben eine *obere* Schranke für die VC-Dimension von sigmoiden Neuronalen Netzen sowie RBF-Netzwerken, für letztere zum Beispiel $O(W^2k^2)$.
Die VC-Dimension von Wavelet Neuronalen Netzen ist aufgrund der Aktualität dieser Fragestellungen in der Lerntheorie-Forschung noch weitgehend ungeklärt. Es ist aber zu erwarten, dass auch Wavelet Neuronale Netze beschränkte VC-Dimension besitzen.
Das SRM-Prinzip ist somit zumindest prinzipiell auf Neuronale Netze anwendbar und der erste Schritt bestünde in der Wahl von M, d.h. einem Unterraum S_M von S mit der VC-Dimension $h_M < \infty$.

Innerhalb dieses Unterraums wird nun durch den Lernprozess der erste Term auf der rechten Seite von[9]

$$\Lambda_{\Xi}(w_{M(N)}) \leq \Lambda_{\Xi}^{N}(w_{M(N)}) + \Omega \left(\frac{h_M}{N} \ln \frac{2N}{h_M} \right)$$

[9] Ω bezeichnet als Analogon zum Landau-Symbol O die asympotisch untere Schranke.

minimiert. Die vorangegangenen Sätze über das SRM Prinzip garantieren die Konvergenz auf die beste Lösung innerhalb von S_M. Aber auch falls $\Lambda_{\Xi}^{N}(\hat{w}_{M(N)})$ gegen 0 konvergiert kann der Fehler zum eigentlichen Risiko $\Lambda_{\Xi}(\hat{w})$ immer noch *erheblich* sein, falls das Verhältnis h_M/N falsch gewählt wurde. Dies ist z.B. der Fall falls das Netzwerk zu komplex (großes h_M) für die gegebene Menge von Trainingsdaten N ist. In diesem Fall spricht man von *overfitting*. Um diesen Effekt zu vermeiden muss also ein Netzwerk mit kleiner VC-Dimension konstruiert werden. Für dieses Problem existiert bislang keine Lösung.

Es gibt allerdings doch einen Weg wie die VC-Theorie relativ problemlos auf Neuronale Netze angewendet werden kann. In Glg. (4.16) fassen wir die Entwicklung als *adaptiv* auf, d.h. die Baisisfunktionen $\widetilde{\Phi}$ sind nicht-linear in den Parametern $\boldsymbol{a}_i, t_i$ und der Optimierungsprozess umfasst die Anpassung aller Parameter anhand der Trainingsdaten. Fassen wir allerdings die Basisfunktionen als *fest vorgegeben* auf, so wird aus (4.16)

$$\hat{f}(\boldsymbol{x}, w) = \sum_{i=1}^{M} u_i \, \widetilde{\Phi}_i(\boldsymbol{x}) \,, \tag{4.17}$$

wir erstellen also eine *Bibliothek* an Basisfunktionen, für jedes S_M genau M-Stück.
Gerade bei Wavelet Neuronalen Netzen ist diese "Dictionary representation" sehr beliebt. So wird z.B. im Artikel von Zhang (1994) eine Wavelet-library auf Grund einer vorgegebenen Diskretisierung der Parameter erzeugt und dann anhand der Trainingsdaten sukzessive "ausgedünnt". Die Anzahl M der am Ende zur Rekonstruktion verwendeten Wavelets wird in diesen Ansätzen über Standard-Kriterien wie Akaike's final prediction error o.Ä. geschätzt. Es ist relativ deutlich, dass dieses Vorgehen von Zhang eine direkte Umsetzung der endlichen Rekonstruktionsformel von Daubechies für Wavelets (s. Abschnitt 3.2) darstellt. Einzig und allein die Koeffizienten der Linearkombination werden nicht wie in der Signal-Verarbeitung üblich über eine inverse Transformation (DWT für Wavelets, DFT für Fourier-Entwicklung) bestimmt, sondern anhand der Trainingsdaten erlernt.
Die Räume S_M haben nun zwar immer noch dieselbe Form, ihre Mitglieder werden aber a priori festgelegt und nicht als Teil des Algorithmus berechnet[10]. Die Räume S_M haben somit die VC-Dimension M und das SRM-Prinzip kann angewendet werden. Der große Vorteil der dictionary-Darstellung in Kombination mit der SRM-Methode liegt darin, dass die VC-Dimension unabhängig von den gestörten Daten Z_i ist. Nur so kann eine *robuste* Modellauswahl garantiert werden.

Es gibt andere Möglichkeiten eine Struktur auf S zu definieren. So wird in der Signal-Approximation oft so genanntes Wavelet-thresholding angewandt, wobei die Wavelet-

[10] Genau an dieser Stelle wird intuitiv klar warum die VC-Dimension für adaptive Methoden sehr viel größer sein muss als für lineare Methoden, man denke nur an die Definition der VC-Dimension als die maximale Anzahl an Datenpunkten, die in der Funktionenklasse in 2^N Gruppen aufgeteilt werden können.

Koeffizienten ihrer Größe nach geordnet werden (s. z.B. Donoho & Johnstone (1994)). Diese Methoden gehen alle nach demselben Muster vor:

1) Wende die diskrete Fourier- oder Wavelet-Transformation auf N Messdaten an und bestimme so die Fourier- bzw. Wavelet-Koeffizienten.
2) Ordne diese Koeffizienten der Größe nach.
3) Wähle aus dieser geordneten Menge die M "wichtigsten" Koeffizienten anhand eines Schwellwertes.
4) Wende die inverse DFT oder DWT auf die M gewählten Koeffizienten an und rekonstruiere das Signal.

Die verschiedenen existierenden Methoden unterscheiden sich nur in Schritt 3), also der Wahl der Schwelle (VISU-Methoden verwenden eine analytische Form von $M(N)$, SURE-Methoden benutzen den Stein-Schätzer für das empirische Risiko usw.). Es gibt auch Ansätze (s. Cherkassky & Shao (2001)), die in diesem Schritt M nicht über ein Standard-Kriterium schätzen, sondern diesen Parameter mit der VC-Dimension identifizieren und zusätzlich den Penalty-Faktor aus (4.14) mit einbeziehen. Die prinzipielle Schwäche aller dieser Methoden liegt aber darin, dass die Trainingsdaten schon die Ordnung der Struktur beeinflussen, so dass die VC-Dimension meist falsch bestimmt wird. Anders ausgerückt: Die Komplexität des Modells spielt bei diesen Methoden (wie natürlich auch für die Wavelet-library-reduction Methoden) erst im vorletzten Schritt, in der Schätzung von M ins Spiel und alle Methoden hängen von dem konkreten und a priori gewählten Signal-Rausch-Modell ab. Sie können also schon prinzipiell nicht *robust* sein.

4.4 Topologische Einschränkungen für Hypothesenräume

Die SRM-Theorie ist in der Praxis ungeeignet für Neuronale Netze. Wir verfolgen deshalb in diesem Abschnitt eine andere Strategie und schließen an Satz 4.1 an, der Voraussetzungen für die Lösbarkeit des ERM-Problems angibt:

Satz 4.8. *Es gelte folgendes:*

1) Gegeben ist $\mathcal{X} \subset \mathbb{R}^d$, eine kompakte Input-Menge und der Banachraum $C(\mathcal{X})$ (mit der Supremumsnorm $\|\cdot\|_\infty$). $\mathscr{F} \subset C(\mathcal{X})$ sei kompakt und

2) $\Xi = \Xi(\boldsymbol{y}, f(\boldsymbol{x}, w))$, $f \in \mathscr{F}$, strikt konvex und fast überall beschränkt (bzgl. beider Argumente) auf einem konvexen Definitionsbereich $\mathcal{D} \times \mathcal{Y}$, wobei $\mathcal{D} = \mathrm{Im}(\mathscr{F}) := \{\boldsymbol{y}' \in \mathbb{R}^m : \exists\, f \in \mathscr{F}, \exists\, \boldsymbol{x} \in \mathbb{R}^d, f(\boldsymbol{x}, w) = \boldsymbol{y}'\}$ und $\mathcal{Y} \subset \mathbb{R}^m$.

Dann existiert ein Minimum des Risikos $\Lambda_{\Xi,f}$ in $\mathscr{F}$. Ist $\mathscr{F}$ zusätzlich konvex, dann ist das Minimum auch eindeutig.

Beweis. Die Existenz des Minimums von $\Lambda_{\Xi,f}(w) = \mathbb{E}[\Xi(Y, f(X, w))]$ in $\mathscr{F}$ wird sichergestellt durch die Kompaktheit von $\mathscr{F}$ und die Stetigkeit von Ξ, die aus 2) folgt, da eine

auf einer konvexen Definitionsmenge konvexe und beschränkte Funktion im inneren des Definitionsbereichs auch stetig ist. Somit ist auch $\Lambda_{\Xi,f}$ stetig. Weiterhin hat eine stetige strikt konvexe Funktion auf einer kompakten konvexen Menge auf dieser Menge genau ein globales Minimum (s. ein Buch über nicht-lineare Optimierung, z.B. Jarre & Stoer (2003)). ■

Alternativ den Voraussetzungen von Satz 4.8 könnten wir auch folgendes für Ξ fordern:

1) Ξ ist konvex in Bezug auf das zweite Argument und
2) es gilt eine Lipschitz-Bedingung: Es gibt ein $K < \infty$, so dass

$$\forall\, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{D} \subset \mathbb{R}^m,\, \forall\, \boldsymbol{y} \in \mathcal{Y} \subset \mathbb{R}^m \,:\, |\Xi(\boldsymbol{y}, \boldsymbol{y}_1) - \Xi(\boldsymbol{y}, \boldsymbol{y}_2)| \leq K\, |\boldsymbol{y}_1 - \boldsymbol{y}_2| \ .$$

Für $\Xi = (Y - f)^2$ sind diese Bedingungen sofort erfüllt, falls $\mathscr{F}$ und $\mathcal{Y}$ fast überall beschränkt sind, d.h.

$$\exists\, M \ \ \forall\, f \in \mathscr{F} \,:\, \|f\|_\infty \leq M\,, \quad \forall\, \boldsymbol{y} \in \mathcal{Y} \,:\, |\boldsymbol{y}| \leq M\,,$$

$M < \infty$. Diese Bedingungen an Ξ werden in Abschnitt 5, also im Rahmen einer Stabilitätsbetrachtung des Lernproblems, eine große Rolle spielen.

Im folgenden Abschnitt werden wir zunächst auf den Zusammenhang zwischen Kompaktheit des Hypothesenraums und Konsistenz des Lernproblems eingehen. Abschnitt **??** kommt dann noch einmal detailliert auf das Bias-Variance-Problem in diesem Rahmen zurück.

4.4.1 Kompaktheit & Konsistenz

Die Kompaktheit des Hypothesenraumes hat neben der intuitiv recht offensichtlichen Implikation der Lösbarkeit des Bias-Variance-Problems noch weiter reichende Konsequenzen, denn sie ist hinreichend für die *Konsistenz* des Minimierungsproblems:

Satz 4.9. *Es sei $\mathscr{F} \subset C(\mathcal{X})$ kompakt, $\mathcal{X} \subset \mathbb{R}^d$ nicht notwendigerweise kompakt. Weiterhin gelte für Ξ eine Lipschitz-Bedingung: $f_1, f_2 \in \mathscr{F}$ und*

$$\begin{aligned}\forall\, \boldsymbol{x} \in \mathcal{X},\, \forall\, \boldsymbol{y} \in \mathcal{Y} \,:\, c_1 \big|\Xi(\boldsymbol{y}, f_1(\boldsymbol{x})) - \Xi(\boldsymbol{y}, f_2(\boldsymbol{x}))\big| &\leq \big|f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})\big| \\ &\leq c_2 \big|\Xi(\boldsymbol{y}, f_1(\boldsymbol{x})) - \Xi(\boldsymbol{y}, f_2(\boldsymbol{x}))\big|\,,\end{aligned}$$

$0 < c_1 < c_2$. Weiterhin sei Ξ beschränkt fast-überall: $0 \leq \Xi(Z, f) \leq M < \infty$. Dann ist das ERM-Problem konsistent im Sinne von Satz 4.2 bezüglich $S = \{\Xi(Y, f(X, w)) : f \in \mathscr{F}\}$

$$\forall\, \varepsilon > 0 \,:\, \lim_{N \to \infty} \mathbb{P}^N \left[\sup_{\Xi \in S} \big|\mathbb{E}_N\left[\Xi\right] - \mathbb{E}\left[\Xi\right]\big| > \varepsilon \right] = 0$$

und bezüglich $\mathscr{F}$:

$$\forall\, \varepsilon > 0 \;:\; \lim_{N\to\infty} \mathbb{P}^N \left[\sup_{f\in\mathscr{F}} \left| \mathbb{E}_N\left[f\right] - \mathbb{E}\left[f\right] \right| > \varepsilon \right] = 0\,,$$

wobei $\mathbb{E}_N[f] := N^{-1}\sum_{i=1}^N f(X_i)$ (eine Zufallsvariable!). Die Konvergenz ist in diesem Fall gleichmäßig und exponentiell:

$$\mathbb{P}^N \left[\sup_{f\in\mathscr{F}} \left| \mathbb{E}_N\left[\Xi(Z,f)\right] - \mathbb{E}\left[\Xi(Z,f)\right] \right| > \varepsilon \right] < 2\,\ell_{\mathscr{F}}\left(\frac{c_1}{4}\varepsilon\right) \exp\left(-\frac{N\varepsilon^2}{8\left(\sigma^2 + \frac{M^2\varepsilon}{6}\right)} \right)\,.$$

Hierbei bezeichnet

$$\ell_{\mathscr{F}}(r) := \min_{\ell\in\mathbb{N}} \left\{ (U_1,...,U_\ell) : \mathscr{F} = \bigcup_{i=1}^{\ell} U_i(f_i), f_i \in \mathscr{F}, U_i(f_i) = \{f\in\mathscr{F} : \|f - f_i\|_\infty \le r\} \right\}$$

die minimale Anzahl an Kugeln vom Radius r, die $\mathscr{F}$ überdecken und

$$\sigma^2 = \sup_{f\in\mathscr{F}} \mathbb{V}[\Xi(Z,f)] < \infty\,.$$

Beweis. Der Hypothesenraum $\mathscr{F}$ ist kompakt, d.h. er lässt sich durch eine endliche Anzahl von Mengen überdecken, z.B. ℓ Kugeln vom Radius $c_1\varepsilon/4$. Betrachte die i-te Kugel U_i um f_i, $i = 1,...,\ell$. Dann gilt für alle $f'\in U_i$ und alle $(Z_i)\in\mathcal{T}^N$

$$\left|\mathbb{E}_N[\Xi(Z,f')] - \mathbb{E}[\Xi(Z,f')] - \mathbb{E}_N[\Xi(Z,f_i)] + \mathbb{E}[\Xi(Z,f_i)]\right| \le \frac{2}{c_1}\|f - f_i\|_\infty \le \frac{2}{c_1}\frac{c_1}{4}\varepsilon = \frac{\varepsilon}{2}\,.$$

Also gilt:

$$\sup_{f'\in U_i} \left|\mathbb{E}_N[\Xi(Z,f')] - \mathbb{E}[\Xi(Z,f')]\right| \ge \varepsilon \Longrightarrow \left|\mathbb{E}_N[\Xi(Z,f_i)] - \mathbb{E}[\Xi(Z,f_i)]\right| \ge \frac{\varepsilon}{2}\,.$$

Wegen $0 \le \Xi(Z,f) \le M < \infty$ fast-überall gilt dasselbe für $|\Xi(Z_j,f) - \mathbb{E}[\Xi(Z,f)]|$ und $\mathbb{V}[\Xi(Z_j,f)] = \mathbb{V}[\Xi(Z_j,f) - \mathbb{E}[\Xi(Z,f)]] \le \sigma^2$, $j = 1,...,N$. Außerdem ist $\mathbb{E}[\Xi(Z_j,f) - \mathbb{E}[\Xi(Z,f)]] = 0$ für alle $j = 1,...,N$.
Also lässt sich die Bernstein-Ungleichung (s. z.B. Lugosi (2003)) auf dieses Problem anwenden:

$$\mathbb{P}^N \left[\left| \mathbb{E}_N\left[\Xi(Z,f_i)\right] - \mathbb{E}\left[\Xi(Z,f_i)\right] \right| > \frac{\varepsilon}{2} \right] < 2\,\exp\left(-\frac{N\varepsilon^2}{8\left(\sigma^2 + \frac{M^2\varepsilon}{6}\right)} \right)\,.$$

Und somit auch für alle $i = 1,...,\ell$

$$\mathbb{P}^N\left[\sup_{f'\in U_i}\left|\mathbb{E}_N\left[\Xi(Z,f')\right]-\mathbb{E}\left[\Xi(Z,f')\right]\right|>\varepsilon\right]<2\,\exp\left(-\frac{N\varepsilon^2}{8\left(\sigma^2+\frac{M^2\varepsilon}{6}\right)}\right).$$

Die Wahrscheinlichkeiten für jedes i müssen nun nur noch aufsummiert werden, weil

$$\mathbb{P}^N\left[\sup_{f\in\mathscr{F}}\left|\mathbb{E}_N\left[\Xi(Z,f)\right]-\mathbb{E}\left[\Xi(Z,f)\right]\right|>\varepsilon\right]$$
$$\leq\sum_{i=1}^{\ell}\mathbb{P}^N\left[\sup_{f'\in U_i}\left|\mathbb{E}_N\left[\Xi(Z,f')\right]-\mathbb{E}\left[\Xi(Z,f')\right]\right|>\varepsilon\right].$$

Die gleichmäßige Konvergenz der Minimierungsfolge ist offensichtlich, so dass mit Satz 4.2 die Konsistenz des ERM-Problems folgt.

Bislang haben wir nur die erste Abschätzung der Lipschitz-Bedingung ausgenutzt. Die Äquivalenz der Konsistenz bezüglich S und $\mathscr{F}$ folgt nun durch die gesamte Lipschitz-Bedingung, die sicherstellt, dass sich $|\mathbb{E}_N[\Xi(Z,f)]-\mathbb{E}[\Xi(Z,f)]|$ und $|\mathbb{E}_N[f(X)]-\mathbb{E}[f(X)|$ auf $\mathscr{F}$ lediglich durch eine positive Konstante unterscheiden. ■

Anmerkung 4.4. Zur Vereinfachung wurde in diesem Satz die Notation $\mathbb{E}_N[\Xi]$ statt $\widetilde{\Lambda}^N_\Xi(\cdot,w)$ verwendet. So ist in parametrisierten Hypothesenräumen

$$\sup_{\Xi\in S}\left|\mathbb{E}_N\left[\Xi\right]-\mathbb{E}\left[\Xi\right]\right|\equiv\sup_{w\in W}\left|\widetilde{\Lambda}^N_\Xi(\cdot,w)-\Lambda_\Xi(w)\right|\equiv\sup_{f\in\mathscr{F}}\left|\widetilde{\Lambda}^N_\Xi(\cdot,f)-\Lambda_\Xi(f)\right|.$$

Anmerkung 4.5. Die gleichmäßige Konvergenz ist sogar fast-sicher und nicht nur in Wahrscheinlichkeit. Hierzu muss lediglich der Beweis von Satz 4.1 rekapituliert werden.

Anmerkung 4.6. Aus der Konsistenz des ERM-Problems folgt aber nicht unbedingt die Kompaktheit von $\mathscr{F}$! (s. Tikhonov (1977)).

Auf Grundlage der endlichen Überdeckungszahl ℓ lässt sich die Vapnik-Konvergenzrate (4.11) für die spezielle Performance-Funktion $\Xi=(Y-\hat{f})^2$ präzisieren:

Satz 4.10. *Dieselben Voraussetzungen wie in Satz 4.9. Ist $\mathscr{F}$ zusätzlich konvex und $\Xi=(Y-\hat{f})^2$, so gilt für alle $\varepsilon>0$:*

$$\mathbb{P}^N\left[\left|\Lambda_\Xi(\hat{w}_{\mathcal{T}^N})-\Lambda_\Xi(\hat{w})\right|>\varepsilon\right]<\ell_{\mathscr{F}}\left(\frac{\varepsilon c_1^2}{12}\right)\,\exp\left(-\frac{Nc_1^2}{72}\varepsilon\right).$$

Beweis. Wir benutzen wieder die Bernstein-Ungleichung in der folgenden Form, $\varepsilon>0$, $0<\alpha\leq 1$, $f\in\mathscr{F}$:

$$\mathbb{P}^N\left[\frac{\left|\mathbb{E}\left[(f-\hat{f})^2\right]-\mathbb{E}\left[(f-\hat{f}_{\mathcal{T}^N})^2\right]\right|}{\mathbb{E}\left[(f-\hat{f})^2\right]+\varepsilon}>\alpha\right]<\exp\left(-\frac{N\alpha^2\left(\mathbb{E}\left[(f-\hat{f})^2\right]+\varepsilon\right)^2}{2\left(\sigma^2+\frac{M\alpha\left(\mathbb{E}\left[(f-\hat{f})^2\right]+\varepsilon\right)}{3}\right)}\right),$$

eine modifizierte Version von Lemma 7 in Cucker & Smale (2001). Wegen der Konvexität von $\mathscr{F}$ ist $\hat{f}$ eindeutig!
Für die Varianz gilt für alle $f \in \mathscr{F}$:

$$\sigma^2 = \mathbb{V}\left[\Xi(Z,f) - \Xi(Z,\hat{f})\right] \leq \mathbb{E}\left[\left(\Xi(Z,f) - \Xi(Z,\hat{f})\right)^2\right] \leq \frac{1}{c_1^2}\,\mathbb{E}\left[\left(f-\hat{f}\right)^2\right] .$$

Weiterhin gilt:

$$\frac{\varepsilon c_1^2}{2} \leq \frac{\left(\mathbb{E}\left[(f-\hat{f})^2\right] + \varepsilon\right)^2}{2\left(\sigma^2 + \frac{M\alpha\left(\mathbb{E}\left[(f-\hat{f})^2\right]+\varepsilon\right)}{3}\right)}$$

mit der obigen Abschätzung für σ^2. Also ergibt sich:

$$\mathbb{P}^N\left[\frac{\left|\mathbb{E}\left[(f-\hat{f})^2\right] - \mathbb{E}\left[(f-\hat{f}_{\mathcal{T}^N})^2\right]\right|}{\mathbb{E}\left[(f-\hat{f})^2\right] + \varepsilon} > \alpha\right] < \exp\left(-\frac{N\alpha^2 c_1^2}{2}\varepsilon\right)$$

für alle $f \in \mathscr{F}$. Nun wird argumentiert wie in Satz 4.9, d.h. die Kompaktheit von $\mathscr{F}$ wird ausgenutzt und $\mathscr{F}$ durch endlich viele Kugeln überdeckt. Wir nutzen das folgende Lemma: Sei für $f \in \mathscr{F}$, $\varepsilon > 0$, $0 < \alpha < 1$

$$\frac{\left|\mathbb{E}\left[(f-\hat{f})^2\right] - \mathbb{E}\left[(f-\hat{f}_{\mathcal{T}^N})^2\right]\right|}{\mathbb{E}\left[(f-\hat{f})^2\right] + \varepsilon} < \alpha\,,$$

dann gilt für alle $g \in \mathscr{F}$ mit $\|f-g\|_\infty \leq \frac{\alpha\varepsilon c_1}{2}$

$$\frac{\left|\mathbb{E}\left[(g-\hat{f})^2\right] - \mathbb{E}\left[(g-\hat{f}_{\mathcal{T}^N})^2\right]\right|}{\mathbb{E}\left[(g-\hat{f})^2\right] + \varepsilon} < 3\alpha\,.$$

(Der Beweis dieses Lemmas kann ganz analog zu dem Beweis von Lemma 8 in Cucker & Smale (2001) geführt werden). Somit ergibt sich sofort durch Summe über alle Kugeln:

$$\mathbb{P}^N\left[\sup_{f\in\mathscr{F}}\frac{\left|\mathbb{E}\left[(f-\hat{f})^2\right] - \mathbb{E}\left[(f-\hat{f}_{\mathcal{T}^N})^2\right]\right|}{\mathbb{E}\left[(f-\hat{f})^2\right] + \varepsilon} > 3\alpha\right] < \ell_{\mathscr{F}}\left(\frac{\alpha\varepsilon c_1}{2}\right)\,\exp\left(-\frac{N\alpha^2 c_1^2}{2}\varepsilon\right) .$$

Wird nun $\alpha = 1/3$ gewählt, so erhalten wir

$$\mathbb{P}^N\left[\sup_{f\in\mathscr{F}}\frac{\left|\mathbb{E}\left[(f-\hat{f})^2\right] - \mathbb{E}\left[(f-\hat{f}_{\mathcal{T}^N})^2\right]\right|}{\mathbb{E}\left[(f-\hat{f})^2\right] + \varepsilon} > \frac{1}{2}\right] < \ell_{\mathscr{F}}\left(\frac{\varepsilon c_1}{12}\right)\,\exp\left(-\frac{N c_1^2}{72}\varepsilon\right) .$$

Setzt man $f = \hat{f}_{\mathcal{T}^N}$, dann folgt

$$\mathbb{E}\left[(\hat{f}_{\mathcal{T}^N} - \hat{f})^2\right] < \varepsilon$$

und

$$\mathbb{P}^N\left[\mathbb{E}\left[(\hat{f}_{\mathcal{T}^N} - \hat{f})^2\right] > \varepsilon\right] < \ell_{\mathscr{F}}\left(\frac{\varepsilon c_1^2}{12}\right) \exp\left(-\frac{N c_1^2}{72}\varepsilon\right) .$$

Die Behauptung ergibt sich nun aus

$$\left|\Lambda_\Xi(\hat{w}_{\mathcal{T}^N}) - \Lambda_\Xi(\hat{w})\right| = \left|\mathbb{E}\left[(f - \hat{f}_{\mathcal{T}^N})^2\right] - \mathbb{E}\left[(f - \hat{f})^2\right]\right| ,$$

so dass

$$\left|\Lambda_\Xi(\hat{w}_{\mathcal{T}^N}) - \Lambda_\Xi(\hat{w})\right| > \varepsilon \Longrightarrow \mathbb{E}\left[(\hat{f}_{\mathcal{T}^N} - \hat{f})^2\right] > \varepsilon ,$$

was sich durch nachrechnen ergibt. ■

Zusammengefasst impliziert die Kompaktheit von $\mathscr{F}$ einerseits die Existenz des Minimums des erwarteten Risikos und falls $\mathscr{F}$ zusätzlich konvex ist auch dessen Eindeutigkeit. Andererseits folgt auch die Konsistenz des Lernproblems. In Kapitel 5 werden wir uns ausführlich mit dem Problem der Stabilität von Lernalgorithmen befassen. Im Vorgriff hierauf merken wir an dieser Stelle an, dass die Kompaktheit des Hypothesenraumen in Kombination mit Konvexität Hypothesenstabilität bezüglich der least-squares Performance-Funktion des Lernalgorithmus impliziert. Die Umkehrung gilt ebenfalls, allerdings folgt aus der Hypothesenstabilität des Algorithmus lediglich die Kompaktheit von .

4.4.2 Kompaktheit und das Bias-Variance-Problem

Wir konkretisieren nun die Aussagen aus Satz 4.8 und geben zunächst Cucker & Smale (2001) folgend einen allgemeinen "Hilbert-Rahmen" an, in dem von diesem Satz ausgehend das Bias-variance-Dilemma eindeutig lösbar ist:

Definition 4.1. *Es sei λ das Lebesgue-Maß auf dem kompakten Input-Raum $\mathcal{X} \subset \mathbb{R}^d$. Weiterhin sei $A : L^2(\lambda, \mathcal{X}) \to L^2(\lambda, \mathcal{X})$ ein kompakter und strikt positiver Operator. $s > 0$ sei fest vorgegeben und $\mathcal{E} := \{g \in L^2(\lambda, \mathcal{X}) : \|A^{-s} g\|_{L^2(\lambda,\mathcal{X})} < \infty\}$. Zusammen mit dem Skalaprodukt $\langle g, h\rangle_{\mathcal{E}} = \langle A^{-s} g, A^{-s} h\rangle_{L^2(\lambda,\mathcal{X})}$ ist $\mathcal{E}$ ein Hilbertraum und $A^{-s} : L^2(\lambda, \mathcal{X}) \to \mathcal{E}$ ein Hilbert-Isomorphismus. Weiterhin sei angenommen, dass $I : \mathcal{E} \hookrightarrow L^2(\lambda, \mathcal{X})$ als Verkettung der Form $I = K_C \circ J_{\mathcal{E}}$ darstellbar ist mit $K_C : C(\mathcal{X}) \hookrightarrow L^2(\lambda, \mathcal{X})$ und $J_{\mathcal{E}} : \mathcal{E} \hookrightarrow C(\mathcal{X})$. $J_{\mathcal{E}}$ sei zudem sogar eine kompakte Einbettung. Wenn zusätzlich B_R eine Kugel mit Radius R in $\mathcal{E}$ ist, dann betrachten wir als Hypothensenraum $\mathscr{F} = \mathscr{F}_{\mathcal{E},R} = \overline{J_{\mathcal{E}}(B_R)}$.*

Beispiel 4.6 (Reproducing kernel Hilbert Space (RKHS)). Das für die Anwendung wichtigste Beispiel eines solchen Hilbert-Rahmens sind Räume, die von einer Kern-Funktion erzeugt werden. Gegeben ist eine stetige und symmetrische Kern-Funktion $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Basierend auf einem Trainingsdatensatz $\boldsymbol{x}_1, ..., \boldsymbol{x}_N \subset \mathbb{R}^d$ wird nun die $N \times N$-Matrix $\mathcal{K}$ gebildet, wobei $\mathcal{K}_{ij} := K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. (Gelegentlich wird solch eine Funktion *Mercer-Kernel* genannt). In Aronszajn (1950) wird die Existenz eines Hilbertraumes H_K von stetigen Funktionen auf $\mathcal{X}$ nachgewiesen, so dass der lineare Operator $L_K^{1/2}$ ein Hilbert-Isomorphismus zwischen $L^2(\lambda, \mathcal{X})$ und H_K ist, wobei $L_K : L^2(\mathcal{X}) \to C(\mathcal{X})$,

$$(L_K(f))(x) := \int K(x,t) f(t)\, \mathrm{d}t \,.$$

Ist $K \in C^\infty$, so ist die Inklusion $I_K : H_K \hookrightarrow C(\mathcal{X})$ kompakt (s. Cucker & Smale (2001)). Der "reproducing" kernel K hat die folgende "Reproduktionseigenschaft", daher der Terminus RKHS:

$$\forall\, f \in H_K \quad f(\boldsymbol{x}) = \langle f(\boldsymbol{y}), K(\boldsymbol{y}, \boldsymbol{x}) \rangle_{H_K} \,.$$

Weiterhin kann K folgendermaßen definiert werden:

$$K(\boldsymbol{x}, \boldsymbol{y}) := \sum_{i=0}^{\infty} k_i\, \varphi_i(\boldsymbol{x}) \varphi_i(\boldsymbol{y}) \,,$$

wobei die Reihe gleichmäßig konvergiere. Solch ein K ist positiv definit. Mit der Reproduktionseigenschaft von K ergibt sich sofort, dass

$$f(\boldsymbol{x}) = \sum_{i=0}^{\infty} f_i\, \varphi_i(\boldsymbol{x}) \,,$$

$f_i \in \mathbb{R}$, $i = 1, 2, ...$ und

$$\|f\|_K^2 = \sum_{i=0}^{\infty} \frac{f_i^2}{k_i} \,.$$

(φ_i) ist eine Basis des RKHS (nicht unbedingt orthonormal) und der Kern K ist die Korrelationsmatrix zwischen diesen Basisfunktionen. Bemerkenswert ist in diesem Zusammenhang also die Äquivalenz der folgenden drei Vorgehensweisen:

- (*i*) Wähle den Reproducing-Hilbertraum H_K,
- (*ii*) wähle eine Menge von Basisfunktionen (φ_i) und die dazugehörigen Koeffizienten (k_i), $i = 1, 2, ...$ und
- (*iii*) wähle einen Kern K.

Die folgende Tabelle gibt einige Beispiele für Kern-Funktionen:

Kern-Funktion $K(\boldsymbol{x}, \boldsymbol{y}) =$	Netzwerk
$K(\|\boldsymbol{x} - \boldsymbol{y}\|)$	RBFN
$\exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2)$	Gauss'sches RBFN
$\tanh(\boldsymbol{x} \cdot \boldsymbol{y} - \theta)$	Multi-Layer Neuronales Netz für geeignete Wahl von θ
$(1 + \boldsymbol{x} \cdot \boldsymbol{y})^d$	Polynome vom Grad d
$B_{2n+1}(x - y)$	B-splines

Es ist auch möglich einen Wavelet-basierten RKHS zu konstruieren. Hierfür s. Gao, Chen & Shi (2004) und Pan et al. (2007).

Satz 4.11. *Im allgemeinen Rahmen von Definition 4.1 existiert für alle $N \in \mathbb{N}$ eine eindeutige Lösung R^* des Bias-Variance-Problems mit Wahrscheinlichkeit $1 - \delta$.*

Beweis. Cucker & Smale (2001) geben einen anderen Beweis (und wir werden auf den folgenden Seiten in ähnlicher Form argumentieren), das entscheidende ist aber an sich nur, dass ein *kompakter* und *konvexer* Hypothesenraum $\mathscr{F} = \mathscr{F}_{\mathcal{E},R} = \overline{J_{\mathcal{E}}(B_R)}$ konstruiert wird, weil $J_{\mathcal{E}}$ als kompakt angenommen wird und B_R beschränkt ist (sogar kompakt). Mit Satz 4.8 folgt dann die Behauptung, weil die Voraussetzungen an Ξ durch die spezielle Wahl $\Xi = (Y - f)^2$ erfüllt sind. ■

Um die Aussagen in diesem Satz zu quantifizieren benötigen wir zunächst folgendes Lemma:

Lemma 4.1. *Sei H ein Hilbertraum und $A : H \to H$ ein selbstadjungierter, strikt positiver und kompakter Operator. Dann gilt mit $s > r > 0$:*

*(*i*) Für $\gamma > 0$ und alle $x \in H$*

$$\min_{y \in H} \left(\|y - x\|^2 + \lambda \|A^{-s} y\|^2 \right) \leq \lambda^r \|A^{-sr} x\|^2$$

*(*ii*) und für $R > 0$ und alle $x \in H$*

$$\min_{\substack{y \in H \\ \|A^{-s} y\| \leq R}} \|y - x\| \leq \left(\frac{1}{R} \right)^{\frac{r}{s-r}} \|A^{-r} x\|^{\frac{s}{s-r}} .$$

In beiden Fällen existiert das Minimum $\hat{y}$ und ist eindeutig. Für (i) ist $\hat{y} = (\mathbb{1} + \lambda A^{-2s})^{-1} x$.

Beweis. Der Beweis stammt ursprünglich aus Smale & Zhou (2003). Zunächst wird der Fall $s = 1$ betrachtet und die Funktion

$$\phi(y) = \|y - x\|^2 + \lambda \|A^{-1} y\|^2 .$$

Ein Minimum von ϕ ist eine Nullstelle von $\partial_y \phi$, d.h. $x = (\mathbb{1} + \lambda A^{-2}) \hat{y}$. Der Operator in der Klammer ist invertierbar, weil λA^{-2} ein strikt positiver Operator ist. Bezeichnen nun $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ die Eigenwerte von A, so ergibt sich nach kurzer Rechnung

$$\phi(\hat{y}) = \lambda \sum_{j=1} \left(\frac{\lambda_j^{2r}}{\lambda_j^2 + \lambda} \right) \lambda_j^{-2r} x_j^2 \leq \lambda \left(\sup_{k \in \mathbb{R}^+} \frac{k^r}{k+\lambda} \right) \|A^{-r} x\|^2 \, .$$

Weiterhin:

$$\partial_k \left(\frac{k^r}{k+\lambda} \right) = \frac{r k^{r-1}}{k+\lambda} - \frac{k^r}{(k+\lambda)^2}$$

mit Nullstelle bei $\hat{k} = \sqrt{\lambda r/(1-r)}$, so dass

$$\frac{\hat{k}^r}{\hat{k}+\lambda} = \lambda^{r-1} r^r (1-r)^{1-r} \leq \lambda^{r-1} \, .$$

Behauptung (i) folgt nun aus

$$\phi(\hat{y}) \leq \lambda^r \|A^{-r} x\|^2 \, .$$

Für (ii) ist zunächst zu bemerken, dass falls $\|A^{-1}x\| \leq R$ das gesuchte Minimum Null wird, ein Trivialfall. Sei dies also nicht der Fall. Es ist klar, dass der Punkt $\hat{y}$, der $\|x-y\|$ in $\{y \in H : A^{-1}y \leq R\} \subset H$ minimiert auf dem Rand dieser Menge liegt, also in $\{y \in H : A^{-1}y = R\} \subset H$. Aus der Optimierungstheorie wissen wir, dass die Nebenbedingung in (ii) übersetzt werden kann in das Problem der Minimierung der zugehörigen Lagrange-Funktion $\mathcal{L} = \|y - x\|^2 + \lambda \|A^{-1}y\|^2$. Wir haben aber schon im Beweis von (i) gesehen, dass $\phi(\hat{y}) \leq \lambda^r \|A^{-r}x\|^2$, also $\lambda R^2 \leq \lambda^r \|A^{-r}x\|^2$. Außerdem ist aber, weil $\lambda \geq 0$,

$$\|\hat{y} - x\|^2 \leq \lambda^r \|A^{-r}x\|^2 \, .$$

Beide Ungleichungen zusammen ergeben also:

$$\lambda \leq \left(\frac{1}{R} \right)^{\frac{2}{1-r}} \|A^{-r}x\|^{\frac{2}{1-r}} \, .$$

Wird dies in die Ungleichung davor eingesetzt ergibt sich Behauptung (ii). ■

Mit Hilfe dieses Lemmas können wir eine obere Schranke für den Bias des Approximationsproblems angeben. In Glg. (4.7) haben wir gezeigt, dass sich der Generalisierungsfehler eines Modells aufspaltet in Bias und Varianz. Wir verwenden nun das vorangehende Lemma um den Generalisierungsfehler in diesem allgemeinen Rahmen zu beschränken[11]:

Satz 4.12. *Es sei der allgemeine Hilbertraum-Rahmen aus Def. 4.1 gegeben. Weiterhin sei $\mathscr{F} \subset L^2(\lambda, \mathcal{X})$ ein Hypothesenraum und $\hat{f}(X, \hat{w}) \in \mathscr{F}$, so dass*

$$\hat{f}(X,\hat{w}) = \operatorname*{argmin}_{g \in \mathscr{F}} \mathbb{E}\left[(Y-g)^2\right] \Longrightarrow \hat{f}(X,\hat{w}) = \operatorname*{argmin}_{g \in \mathscr{F}} \mathbb{E}\left[(f-g)^2\right] \, ,$$

wobei, zur Erinnerung, $Y = f(X) + E$. Dann gilt ($0 < r < s$):

[11] Man siehe Cucker & Smale (2001) für eine ähnliche Analyse des Bias-Variance-trade-offs.

$$\Lambda_\Xi(\hat{w}) = \mathbb{E}\left[(f(X) - \hat{f}(X,\hat{w}))^2\right] + \mathbb{E}[E^2] \leq \underbrace{\mathcal{D}^2_{\mathbb{P}_1\lambda}\left(\frac{1}{R}\right)^{\frac{2r}{s-r}} \|A^{-r}f\|^{\frac{2s}{s-r}}_{L^2(\lambda,\mathcal{X})} + \mathbb{E}[E^2]}_{=:\varepsilon_{bias}},$$

wobei $\mathcal{D}_{\mathbb{P}_1\lambda}$ die "Verzerrung" von $\mathbb{P}_1$ bezüglich λ bezeichnet, also die Operatornorm $\|J\|$, wobei J die Abbildung $L^2(\lambda,\mathcal{X}) \to L^2(\mathbb{P}_1,\mathcal{X})$ bezeichnet.

Beweis. A ist nach Voraussetzung kompakt und strikt positiv und somit auch selbstadjungiert in dem betrachteten Hilbertraum-Rahmen. Wir setzen in Lemma 4.1 $H = L^2(\lambda,\mathcal{X})$ und $x = f$. Zunächst bemerken wir, dass

$$\operatorname*{argmin}_{g\in B_R} \|f-g\|_{\mathbb{P}_1} \leq \mathcal{D}_{\mathbb{P}_1\lambda} \operatorname*{argmin}_{g\in B_R} \|f-g\|_\lambda .$$

Dann ergibt sich mit Teil (ii) von Lemma 4.1:

$$\mathbb{E}\left[(f(X) - \hat{f}(X,\hat{w}))^2\right] = \operatorname*{argmin}_{g\in B_R} \mathbb{E}\left[(f(X)-g(X))^2\right] \leq \left(\frac{1}{R}\right)^{\frac{r}{s-r}} \|A^{-r}f\|^{\frac{s}{s-r}}$$

und somit die Behauptung. ■

Anmerkung 4.7. Für $A = L_K^{1/2}$ und $s = 1$ ergibt sich direkt eine Abschätzung in einem RKHS (s. Beispiel 4.6). Cucker & Smale (2001) geben auch eine Anwendung von Satz 4.12 auf Sobolev-Räume an.

Dieser Satz gibt uns also eine obere Schranke für den "minimalen Abstand" von $\mathscr{F}$ zu $L^2(\lambda,\mathcal{X})$ an, besser kann in keinem Fall approximiert werden. Dieser Fehler ist i.A. gleich dem Bias-Fehler wenn das ERM-Problem konsistent ist. Nun kommt noch hinzu, dass nur empirische Daten vorliegen, d.h. wir kehren zu unserer Abschätzung in Satz **??** zurück:

$$\mathbb{P}^N\left[\left|\Lambda_\Xi(\hat{w}_{\mathcal{T}^N}) - \Lambda_\Xi(\hat{w})\right| > \varepsilon_{\text{sampling}}\right] \leq \ell_{\mathscr{F}}\left(\frac{\varepsilon_{\text{sampling}}c_1^2}{12}\right) \exp\left(-\frac{Nc_1^2}{72}\varepsilon_{\text{sampling}}\right). \tag{4.18}$$

Offensichtlich benötigen wir nun eine sinnvolle Abschätzung für die Überdeckungszahl $\ell_{\mathscr{F}}$. Williamson, Smola & Schölkopf (1998) geben eine interessante Bearbeitung dieses Themas mit Hilfe von Entropie-Zahlen, die folgendermaßen definiert sind: Sei S ein metrischer Raum, dann ist ($k \geq 1$)

$$e_k(S) := \inf\{\varepsilon > 0 : \exists\, D_1, ..., D_{2^k-1} \text{ abgeschlossen mit Radius } \varepsilon, \text{ die } S \text{ überdecken}\}$$

die k-te Entropie-Zahl von S. Wir möchten an dieser Stelle nicht tiefer in diese Fragestellungen eindringen, das für uns wichtigste Ergebnis bezieht sich auf den allgemeinen Hilbert-Rahmen aus Definition 4.1. Unter diesen fallen dann auch die Mercer-Kerne (RKHS). Es gilt:

Lemma 4.2. *Im allgemeinen Hilbert-Rahmen $\mathscr{F} = \mathscr{F}_{\mathcal{E},R}$ ist*

$$\ell_{\mathscr{F}}(\varepsilon) \leq \exp\left[\left(C_{\mathcal{E}}\frac{R}{\varepsilon}\right)^{1/l_{\mathcal{E}}}\right],$$

wobei $C_{\mathcal{E}}$ und $l_{\mathcal{E}}$ positive Konstanten sind, so dass

$$e_k(J_{\mathcal{E}}) \leq C_{\mathcal{E}} k^{-l_{\mathcal{E}}}.$$

Ungleichung (4.18) lautet umgeschrieben, dass mit Wahrscheinlichkeit $1-\delta$

$$\frac{Nc_1^2}{72}\varepsilon_{\text{sampling}} + \ln\delta - \ln\ell_{\mathscr{F}}\left(\frac{\varepsilon_{\text{sampling}}c_1^2}{12}\right) \leq 0,$$

bzw. mit Lemma 4.2

$$\frac{Nc_1^2}{72}\varepsilon_{\text{sampling}} + \ln\delta - \left(C_{\mathcal{E}}\frac{12R}{c_1^2\varepsilon_{\text{sampling}}}\right)^{1/l_{\mathcal{E}}} \leq 0. \tag{4.19}$$

Löst man die Gleichung

$$\frac{Nc_1^2}{72}\varepsilon_{\text{sampling}} + \ln\delta - \left(C_{\mathcal{E}}\frac{12R}{c_1^2\varepsilon_{\text{sampling}}}\right)^{1/l_{\mathcal{E}}} = 0 \tag{4.20}$$

nach $\varepsilon_{\text{sampling}}$ auf, so erhält man eine obere Schranke für den durch Sampling entstandenen Fehler. Analytisch ist die Lösung von (4.20) im allgemeinen Fall allerdings nicht ohne weiteres anzugeben, daher werten wir den Ausdruck im folgenden numerisch aus. Andererseits haben wir in Satz 4.12 eine obere Schranke für den prinzipiellen bias-Fehler in Abhängigkeit von R erhalten. Der *Gesamtfehler* lässt sich nun über die Summe dieser beiden Fehler abschätzen:

$$|\Lambda_{\Xi}(\hat{w}_{\mathcal{T}^N})| \leq |\Lambda_{\Xi}(\hat{w}_{\mathcal{T}^N}) - \Lambda_{\Xi}(\hat{w})| + |\Lambda_{\Xi}(\hat{w})| \leq \varepsilon_{\text{bias}}(R) + \varepsilon_{\text{sampling}}(R).$$

Es gilt also die rechte Seite von (4.4.2) zu minimieren. Abb. 4.1 zeigt beispielhaft den Verlauf dieser Funktion in Abhängigkeit von R. Man beachte, dass statt $\varepsilon_{\text{sampling}}(R)$ die Größe $\varepsilon'_{\text{sampling}}(R) := c_1^2\varepsilon_{\text{sampling}}(R)$ und statt $\varepsilon_{\text{bias}}(R)$ die Größe $\varepsilon'_{\text{bias}}(R) := \varepsilon_{\text{bias}}/\|A^{-r}f\|^{2s/(s-r)}_{L^2(\lambda,\mathcal{X})}$ aufgetragen wird. Wichtig ist, dass die Funktion $\varepsilon'_{\text{sampling}}(R) + \varepsilon'_{\text{bias}}(R)$ ein eindeutiges Minimum R^* hat, wie man folgendermaßen sieht: Es ist zunächst

$$\frac{\partial\varepsilon'_{\text{bias}}(R)}{\partial R} = \mathcal{D}^2_{\mathbb{P}_1\lambda}\frac{2r}{r-s}\left(\frac{1}{R}\right)^{\frac{s+r}{s-r}}\|A^{-r}f\|^{\frac{2s}{s-r}}_{L^2(\lambda,\mathcal{X})},$$

$$\frac{\partial^2\varepsilon'_{\text{bias}}(R)}{\partial R^2} = \mathcal{D}^2_{\mathbb{P}_1\lambda}\frac{2r(r+s)}{(r-s)^2}\left(\frac{1}{R}\right)^{\frac{-2s}{s-r}}\|A^{-r}f\|^{\frac{2s}{s-r}}_{L^2(\lambda,\mathcal{X})},$$

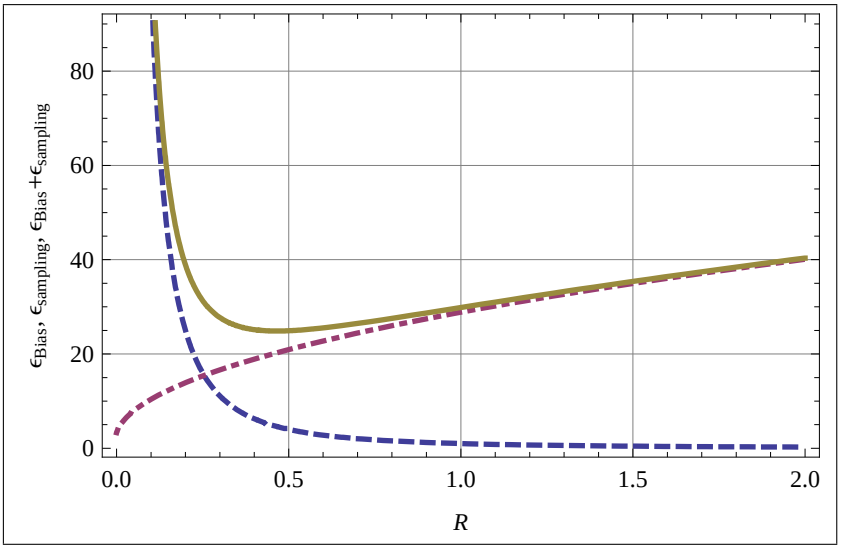


Abb. 4.1. Obergrenzen für Bias-Fehler (gestrichelt), sampling-Fehler (Punkt-gestrichelt) und den resultierenden Gesamtfehler in Abhängigkeit von dem Kugelradius R. Exemplarisch wurde gewählt: $N = 100$, $\delta = 0.01$, $C_{\mathcal{E}} = 512$, $l_{\mathcal{E}} = 1$, $\mathcal{D}_{\mathbb{P}_1\lambda} = 1$, $s = 1$, $r = 0.5$ und $\mathbb{E}[E^2] = 0$.

also ist $\varepsilon'_{\text{bias}}(R)$ wegen $s > r > 0$ eine positive streng monoton abnehmende Funktion. Wir sehen, dass $\partial\varepsilon'_{\text{bias}}(R)/\partial R \to 0$ für $R \to \infty$ und $\partial\varepsilon'_{\text{bias}}(R)/\partial R \to \infty$ für $R \to 0$.
Der Sampling-Fehler ist durch eine implizite Funktion gegeben und es ergibt sich die folgende Differentialgleichung:

$$\frac{\partial\varepsilon'_{\text{sampling}}(R)}{\partial R} = \left(\frac{l_{\mathcal{E}}}{72}NR\left(C_{\mathcal{E}}\frac{12R}{\varepsilon'_{\text{sampling}}}\right)^{-1/l_{\mathcal{E}}} + \frac{R}{\varepsilon'_{\text{sampling}}}\right)^{-1}.$$

Wegen $R, N, l_{\mathcal{E}}, C_{\mathcal{E}} > 0$ und auch $\varepsilon'_{\text{sampling}}(R) > 0$ für alle R ist somit $\varepsilon'_{\text{sampling}}(R)$ eine streng monoton steigende Funktion. Außerdem ist

$$\frac{\partial^2\varepsilon'_{\text{sampling}}(R)}{\partial R^2} = -72Nl_{\mathcal{E}}\left(C_{\mathcal{E}}\frac{12R}{\varepsilon'_{\text{sampling}}}\right)^{1/l_{\mathcal{E}}}\varepsilon'^2_{\text{sampling}} \cdot \frac{144\left(C_{\mathcal{E}}\frac{12R}{\varepsilon'_{\text{sampling}}}\right)^{1/l_{\mathcal{E}}} + N(l_{\mathcal{E}}-1)\varepsilon'_{\text{sampling}}}{R^2\left(72\left(C_{\mathcal{E}}\frac{12R}{\varepsilon'_{\text{sampling}}}\right)^{1/l_{\mathcal{E}}} + N\,l_{\mathcal{E}}\,\varepsilon'_{\text{sampling}}\right)^3}.$$

Also ergibt sich $\partial\varepsilon'_{\text{sampling}}(R)/\partial R \to \infty$ für $R \to \infty$ für alle $l_{\mathcal{E}}$ und $\partial\varepsilon'_{\text{sampling}}(R)/\partial R \to 0$ für $R \to 0$. Insgesamt erhalten wir also zusammen mit den Eigenschaften von

$\partial\varepsilon'_{\text{bias}}(R)/\partial R$, dass

$$\frac{\partial\varepsilon'_{\text{bias}}(R)}{\partial R} + \frac{\partial\varepsilon'_{\text{sampling}}(R)}{\partial R} = 0$$

eine eindeutige Lösung R^* besitzt und $\varepsilon'_{\text{bias}}(R^*)+\varepsilon'_{\text{sampling}}(R^*)$ das eindeutiges Minimum ist.

4.5 Konvergenz im Hypothesenraum

In Abschnitt 4.4 diskutiert wie sich die topologischen Eigenschaften des Hypothesenraumes auf den Approximationsfehler auswirken. Hierbei wurden speziell die Begriffe Kompaktheit und Konsistenz verknüpft. In diesem Abschnitt untersuchen wir nun die Frage der Konvergenz der Schätzer gegen die Zielfunktion genauer.

Die VC-Theorie stellt einen theoretischen Rahmen zur Verfügung, in dem eine analytische Komplexitätskontrolle während der Modellierungsphase möglich ist. Wir haben gezeigt, dass eine falsche Wahl von h/N in einem over- bzw. underfit resultieren kann. Das *Runge-Phänomen* ist nichts anderes als ein overfit an die Daten im Falle der Polynominterpolation: Je komplexer das Modell gewählt wird (steigender Polynomgrad), desto größer wird $\max_{x\in D}|f(x) - P_n(x)|$, $D \subset \mathbb{R}$. Der tieferliegende Grund hierfür liegt in der nicht-gleichmäßigen Konvergenz von P_n gegen f. Wir erinnern zunächst an den folgenden grundlegenden Satz von Faber (s. z.B. Deuflhard & Hohmann (2002)):

Satz 4.13 (Faber (1914)). *Für jede Folge $\{T_k\}$ von Stützstellen $T_k = \{t_{k,0}, ..., t_{k,n_k}\} \subset D$ existiert eine stetige Funktion $f \in C(D)$, so dass die Folge $\{P_k\}$ der zugehörigen Interpolationspolynome nicht gleichmäßig gegen f konvergiert.*

Andersherum zeigt der folgende Satz, dass die Stützstellen "nur" geschickt genug gewählt werden müssen:

Satz 4.14 (Marcinkiewicz). *Zu jeder Funktion $f \in C(D)$ gibt es eine Folge von Stützstellensätzen, so dass die zugehörige Folge von Interpolationspolynomen gleichmäßig gegen f konvergiert.*

Es lässt sich zeigen, dass die Identifikation der Stützstellen mit den Nullstellen der Tschebyscheff-Polynome genau dies leistet.
Eine etwas griffigere Darstellung der Interpolationsproblematik liefert die folgende Beobachtung (s. Schaback & Werner (1992)): Für eine k-fach *stetig differenzierbare* Funktion f ist

$$\max_{x\in D}|f(x) - P_n(x)| = o\left(\frac{\ln n}{n^{k-1}}\right) .$$

Es ist offensichtlich, dass für lediglich stetige Funktionen $f \in C(D)$ der Fehler unbeschränkt ist. Das von Carl Runge vorgeschlagene Beispiel

$$f(x) = \frac{1}{1+x^2} \quad x \in [-5,5]$$

illustriert dies.

Der Polynomgrad n ist im Wesentlichen nichts anderes als ein Maß für die Komplexität des Modells (die VC-Dimension des Raumes der Polynome vom Grad n ist wie erwähnt $n+1$). Von daher entspricht n der "Summenlänge" M der neuronalen Netze. Nach Barron (1993) gilt folgendes zum Satz von Faber analoges Theorem für sigmoide Neuronale Netze:

Satz 4.15. *Sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine sigmoide Funktion, f habe kompakten Träger und*

$$C_f = \int_{\mathbb{R}^d} |\omega| \, |\overline{f}(\omega)| \, d\omega < \infty \, .$$

Gewählt sei der Approximationsraum $\mathscr{F}_M(\mathbb{R}^d, \mathbb{R}) = \Sigma_M^d(\sigma)$ mit

$$\Sigma_M^d(\sigma) := \left\{ g_M : \mathbb{R}^d \to \mathbb{R} \; : \; g_M = \sum_{i=1}^{M} u_i \sigma(A_i);\; t_i \in \mathbb{R}, a_{ij} \in \mathbb{R}, u_i \in \mathbb{R}, \right.$$
$$\left. i \in \{1, ..., M\}, j \in \{1, ..., d\} \right\} .$$

Dann existiert ein $\hat{f}_M \in \Sigma_M^d$, so dass

$$\left\| \hat{f}_M - f \right\|_2 \leq 2\sqrt{d} \, C_f \, M^{-1/2} \, .$$

Anmerkung 4.8. Für andere Basisfunktion existieren ähnliche, aber z.T. sogar noch stärkere Existenzaussagen, wie z.B. in Zhang et al. (1995) für Wavelet Neuronale Netze. Siehe für eine sehr gute theoretische Ausarbeitung der Approximationseigenschaften von WNNs außerdem Delyon, Juditsky & Benveniste (1995).

Es ist klar, dass

$$\Sigma_M^d(\sigma) \subset \Sigma_{M+1}^d(\sigma) \subset \cdots \subset \Sigma^d(\sigma) \, .$$

Wir fassen nun die $\hat{f}_M$ für jedes M als Folge $(\hat{f}_M)_{M \in \mathbb{N}}$ auf. Offensichtlich impliziert Barron also, dass durch iterative Hinzunahme von Netzwerk-Knoten und anschließende Anpassung der Netzwerk-Parameter eine monotone L^2-Konvergenz der Folge $\hat{f}_M$ gegen die Zielfunktion möglich ist. Die Frage ist nun wie man diese Folge konstruiert. In Abschnitt **??** werden wir eine Klasse von rekursiven Algorithmen angeben, die die Konvergenz von $\hat{f}_M$ gegen f im Erwartungswert mit M^{-1} garantiert.
In Analogie zu der zuvor rekapitulierten Polyonominterpolation stellt sich aber auch hier die Frage der *Gleichmäßigkeit* der Konvergenz. Sie ist die entscheidende Eigenschaft wenn es um die Frage der Approximationsgüte "an den Zwischenpunkten" geht,

also wenn ein stark oszillatorisches Verhalten von $\hat{f}_M$ um f vermieden werden soll. In der Theorie der Neuronalen Netze und als Folge auch in der Anwendung gibt es hierzu überraschenderweise bislang noch keine Ergebnisse. Basierend auf den folgenden theoretischen Überlegungen präsentieren wir in Kapitel 6 eine Klasse von Neuronalen Netzen, mit denen "sichere Modellbildung" in sehr allgemeinem Rahmen möglich ist.

Im allgemeinen können wir nicht davon ausgehen, dass $\hat{f}_M$ eindeutig in $\mathscr{F}$ ist, bzw. der Optimierungsalgorithmus wird aufgrund der hohen Dimensionalität von $\mathscr{F}$ nur ein lokales Minimum $\hat{f}_M^{\text{lok}}$ finden, so dass die obige Konvergenz auch nur lokal zu verstehen ist:

$$\hat{f}_M^{\text{lok}} \xrightarrow{M\to\infty} \hat{f}^{\text{lok}} \,. \tag{4.21}$$

Grund hierfür ist die i.A. nicht gegebenen Konvexität des Hypothesenraumes. Die Hoffnung ist natürlich, dass $\hat{f}^{\text{lok}}$ möglichst nahe an f herankommt. Neuronale Netze benutzen für die Lösung des Minimierungsproblems wie erwähnt sehr häufig eine Variation des Backpropagation-Algorithmus. Yu, Onder Efe & Kaynak (2002) zeigen, dass die meisten in den Neuronalen Netzen genutzten Algorithmen nur Spezialfälle eines verallgemeinerten Backpropagation-Algorithmus sind. In diesem Artikel wird weiterhin gezeigt, dass für diese Klasse von Algorithmen keine globale Konvergenz zu erwarten ist:

> "*We will show that trapping into local minima is inherent with the learning algorithms based on the BP principle as they may only enable the weights to converge to global minima if it happens that either the initial weights are near a global minimum or the geometric distribution of weights enables the weights to converge to a global minimum.*"

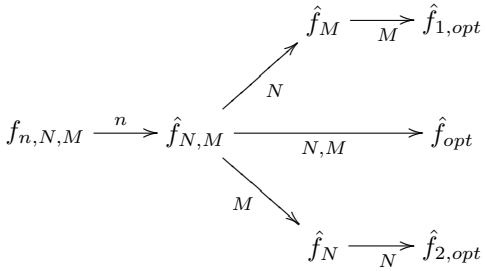
Diese Fragestellung soll allerdings nicht zentraler Gegenstand dieses Abschnittes sein und wir verwenden $\hat{f}^{\text{lok}}$ so, als sei dies die optimale Lösung.

Uns interessiert nun die Frage der *Gleichmäßigkeit* der Konvergenz in Glg. (4.21) und wie diese garantiert werden kann durch geeignete Einschränkungen von $\mathscr{F}_M$. Alle beteiligten Funktionen sind stetig und wir operieren in der Regel auf kompakten Definitionsbereichen, nach dem Satz von Dini können wir auf gleichmäßige Konvergenz allerdings nur zwingend schließen, wenn die Folge $(\hat{f}_M)$ auch monoton ist, was offensichtlich i.A. nicht bzw. nur für die L^2-Norm der Abweichung von der Zielfunktion gilt. Aber selbst wenn der Optimierungsalgorithmus für jeden Unterraum $\mathscr{F}_M$ ein globales Optimum findet, so ist immer noch nicht sichergestellt, dass diese für $M \to \infty$ auch gleichmäßig gegen f konvergiert.

Nun kommt noch hinzu, dass in der Praxis immer nur endlich viele Trainingsdaten zur Verfügung stehen. Wir betrachten den Minimierungsprozess $\hat{w} = \operatorname{argmin}_{w\in W_M} \Xi(Z, w)$ folglich als Folge $(f_{n,N,M})_{n\in\mathbb{N}}$ mit

$$f_{n,N,M} = f_M(X, w^n_{\mathcal{T}^N}) \,,$$

wobei $w^n_{\mathcal{T}^N} \to \hat{w}_{\mathcal{T}^N}$ und $f_{n,N,M} \to f_{N,M}$ für $n \to \infty$. Konsistenz betrifft die Frage $f_{n,N,M} \to f_M$ für $n, N \to \infty$. Und als letztes gilt es noch $f_{n,N,M} \to \hat{f}_N$ für $n, M \to \infty$ zu betrachten. Es ergibt sich für den Modellierungsprozess also folgendes Bild:



Hierbei haben wir nun

$$\hat{f}_{opt} = \underset{f \in \mathscr{F}}{\operatorname{argmin}} \, \Lambda_{\Xi}(f)$$

geschrieben, ob $\hat{f}_{opt}$ von f abweicht oder ein "Minimalabstand" in Form eines Bias-Terms besteht hängt von $\mathscr{F}$ ab (s. Abschnitt 4.4). Die Konvergenz in n lässt sich nur in Zusammenhang mit einem konkreten Minimierungs-Algorithmus behandeln. Bei einem Backpropagation-Verfahren werden die drei Parametersätze $u_i, \boldsymbol{a}_i, t_i$ in Richtung des jeweils negativen Gradienten angepasst:

$$\begin{aligned} \Delta u_i &= \alpha_u \frac{\partial \Lambda^N}{\partial u_i} \\ \Delta a_{ij} &= \alpha_a \frac{\partial \Lambda^N}{\partial a_{ij}} \\ \Delta t_i &= \alpha_t \frac{\partial \Lambda^N}{\partial t_i} \,, \end{aligned}$$

$i \in \{1, ..., M\}$, $j \in \{1, ..., d\}$. Die Literatur über Optimierungsmethoden ist extrem umfangreich. Da alle Parameter gleichzeitig abgestimmt werden ist ein Batch- einem Online-Learning-Verfahren vorzuziehen, so dass die Änderungen in den Parametern so überwacht werden können, dass gleichmäßige Konvergenz erreicht werden kann. Über die Anpassung der Schrittweiten $\alpha_{u,a,t}$ lassen sich Probleme wie z.B. das Überspringen des globalen Maximum oder das "stranden" auf Plateaus der Hyperfläche der Fehlerfunktion vermeiden. Ein Momentumterm kann zusätzlich dazu beitragen Störstellen zu überwinden. Eine gute Übersicht in Zusammenhang mit dem Vorschlag einer adaptiven Lernrate für den Backpropagation-Algorithmus, so dass eine größere Stabilität und des Algorithmus erreicht wird in Kombination mit einer Dämpfung von Oszillationen, liefern

z.B. Magoulas, Vrahatis & Androulakis (1999). In Kapitel 5, Abschnitt 5.3 werden wir diesen Aspekt näher beleuchten und auf das Problem der schlechten Kondition des Optimierungsproblems zu sprechen kommen.

Die Konvergenzen in N, M sind jeweils so zu verstehen, dass in jedem Schritt der Grenzwert $n \to \infty$ gebildet wird. Das Diagramm illustriert, dass die Gleichmäßigkeit der Konvergenzen von entscheidender Bedeutung ist, denn nur dann ist die Vertauschbarkeit der Grenzwerte sichergestellt und $\hat{f}_{1,opt} = \hat{f}_{2,opt} = \hat{f}_{opt}$. Zur Illustration ein Beispiel: Angenommen der Grenzwert

$$\lim_{M\to\infty} \left| \Lambda_\Xi(\hat{f}_M) - \Lambda(\hat{f}) \right| = 0$$

existiere, also $\hat{f}_M \to \hat{f}$ im Grenzfall $N \to \infty$. Dann ist, wenn die Konvergenz in N gleichmäßig fast-sicher ist,

$$\begin{aligned} \lim_{M\to\infty} \left| \Lambda_\Xi(\hat{f}_M) - \Lambda(\hat{f}) \right| &= \lim_{M\to\infty} \lim_{N\to\infty} \left| \widetilde{\Lambda}^N_\Xi(\hat{f}_{N,M}, \omega) - \Lambda(\hat{f}) \right| \\ &= \lim_{N\to\infty} \lim_{M\to\infty} \left| \widetilde{\Lambda}^N_\Xi(\hat{f}_{N,M}, \omega) - \Lambda(\hat{f}) \right| . \end{aligned}$$

Der Grenzwert $\lim_{M\to\infty} \left| \widetilde{\Lambda}^N_\Xi(\hat{f}_{N,M}, \omega) - \Lambda(\hat{f}) \right|$ muss nun aber nicht mehr existieren, wie man leicht an einem worst-case Beispiel der Form

$$\hat{f}_{N,M}(\omega) = \begin{cases} \hat{f}(\omega) & \text{für } Z(\omega) \in \mathcal{T}^N \\ \infty & \text{sonst} \end{cases}$$

erkennt.
Anders ausgedrückt erkennen wir auch den Zusammenhang zum Bias/Variance-Dilemma: Die Forderung der gleichmäßigen Konvergenz garantiert, dass der Varianz-Term in der Dekomposition (4.7) für $M \to \infty$ nicht unbeschränkt wächst. In Bezug auf $N \to \infty$ hat Vapnik die Äquivalenz der gleichmäßigen Konvergenz mit der Konsistenz des ERM-Problems gezeigt. Zusammengefasst wird also deutlich, dass die entscheidende Eigenschaft für ein "gutes" Modell die gleichmäßige Konvergenz der Funktionenfolgen im Hypothesenraum ist.

Die Komplexitätskontrolle nach Vapnik versucht den Einfluss der Diskretisierung mit der Einschränkung der Modellstruktur auf einen Unterraum (z.B. $\mathscr{F}_M$) zu verbinden. In Abschnitt 4.3 sind wir der Frage nach dem Zusammenhang zwischen Konsistenz, gleichmäßiger Konvergenz und VC-Dimension des Lernproblems nachgegangen. Es hat sich gezeigt, dass die Endlichkeit der VC-Dimension für ein bestimmtes M notwendig und hinreichend ist für gleichmäßige fast-sichere Konvergenz in N (entscheidend ist hierbei die Eigenschaft $\lim_{N\to\infty} G^W(\mathcal{T}^N)/N = 0$) [12]. Existiert also eine Struktur

[12] Zum Thema der fast-sicheren Konvergenz beachte man zusätzlich Anmerkung 4.2.

$\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}_M \subset \cdots$ auf $\mathscr{F}$, so dass alle $\mathscr{F}_M$ endliche VC-Dimension haben, so ist gleichmäßige Konvergenz $\hat{f}_{N,M} \to \hat{f}_M$ in N garantiert. Heraus kommt eine Komplexitätskontrolle, in der abhängig von N ein Optimum zwischen Komplexität (entspricht M) und Approximationsgenauigkeit zu erreichen. Die Komplexitätstheorie betrachtet also Folgen der Form

$$\left(\hat{f}_{N,h(M)}\right)_{N,M\in\mathbb{N}} .$$

Satz 4.5 zeigt, dass das empirische Risiko für $N \to \infty$ gegen das kontinuierliche Risiko konvergiert. Die Folge in h, also $h \to \infty$, lässt sich im Rahmen des SRM-Prinzips allerdings nicht behandeln (die Ausdrücke unter den Wurzelzeichen werden negativ). Zudem ist h als Funktion von M, wie schon erwähnt, für eine sehr große Klasse von Approximatoren (die Neuronalen Netze gehören dazu) nur sehr ungenau bekannt.

In Abschnitt 4.4 verfolgten wir einen anderen Ansatz. Hier wurde nicht die Komplexität von $\mathscr{F}$ eingeschränkt, sondern die Norm der enthalten Funktionen, man betrachtet also die Einschränkung von $\mathscr{F}$ auf kompakte Teilmengen. Die VC-Dimension bleibt hiervon jedoch interessanterweise unberührt wie z.B. in Williamson, Smola & Schölkopf (1998) gezeigt wird. Es zeigt sich in Abschnitt 4.4.1, z.B. Satz 4.9, dass die Kompaktheit des Hypothesenraumes zusammen mit einer Lipschitz-Stetigkeits-Bedingung für Ξ auch die gleichmäßige fast-sichere Konvergenz der empirischen Risiken gegen das wahre Risiko in N garantiert. Gleichmäßige Konvergenz ist hier zu verstehen als Supremum über den Raum der möglichen Schätzer $\mathscr{F}$. Dieser Satz entspricht Satz 4.1, in dem, bei genauer Betrachtung, ebenfalls Kompaktheit des Hypothesenraumes gefordert wurde, zusammen mit einer globalen Beschränktheit der Performance-Funktion[13]. Durch die resultierende Eindeutigkeit des Minimums ist somit auch die fast-sichere Konvergenz von $\hat{f}_N = \operatorname{argmin}_{f\in\mathscr{F}} \Lambda^N_\Xi(f)$ gegen $\hat{f}$ gesichert. Man bemerke, dass die VC-Theorie nur die gleichmäßige Konvergenz der Risiken behandelt (s. z.B. Satz 4.2).
In der Praxis lässt sich die Kompaktheit von $\mathscr{F}$ leicht umsetzen durch Nebenbedingungen der Art $\|f\| \leq R$ für alle $f \in \mathscr{F}$. Es ist wie gezeigt möglich ein Optimum im trade-off zwischen dem durch die Einschränkung auf die Kugel hervorgerufenen Bias-Fehler und dem durch Sampling entstandenen Fehler zu berechnen, also ein optimales R. Aber die Konvergenz von $\hat{f}$ in Bezug auf einen Komplexitätsparameter wie M oder $h(M)$ lässt dieser Ansatz prinzipiell *nicht* zu.

Zusammenfassend lässt sich sagen, dass wir zwei unterschiedliche Herangehensweisen an das allgemeine Problem des "Erlernens" einer unbekannten Funktion dargestellt haben, die beide in der gleichmäßigen fast-sicheren Konvergenz der *empirischen Risiken* in N münden. Garantiert wird also nur die gleichmäßige Konvergenz im Mittel. Dass die Konvergenz der Schätz-Funktionen selber sowohl in N als auch in M hingegen keineswegs

[13] Im Anschluss an Satz 4.1 haben wir zudem die Konvergenz in Verteilung gegen die Normalverteilung gezeigt.

gleichmäßig sein muss ist sogar schon anderen Autoren in Experimenten aufgefallen, die eigentlich zu gänzlich anderem Zweck durchgeführt wurden. Das grundlegende modelltheoretische Problem wurde aber nicht erkannt und weiter analysiert. Wir verweisen hier auf das schon erwähnte Beispiel in Cherkassky et al. (1999), Section IV: Die Autoren wählen für $\mathscr{F}_M$ den Raum der Polynome vom Grad M und bestimmen M einmal mittels klassischer Methoden (speziell cross-validation) und einmal mittels der Methoden aus der VC-Theorie, also Komplexitätskontrolle. Für letztere Methode beobachten sie das klassische Phänomen von Runge, also ein starkes Schwingen des Schätzers an den Grenzen des Intervalls. Die cross-validation-Methode liefert deutlich bessere Resultate, was aber nicht bedeutet, dass diese Methode gleichmäßige Konvergenz garantiert. In dieser Methode werden die Intervallgrenzen als Validierungsmenge verwendet, so dass für höher-gradige Polynome die Validierungs-Fehler in diesen Bereichen sehr viel größer sind als in anderen Samples und die Methode als Konsequenz den Polynomgrad klein wählt. Modifiziert man die cross-validation Methode so, dass nicht die Intervallgrenzen als Validierungsmenge verwendet werden, so zeigt sie ähnliche Oszillationen wie die VC-Methode. Dieses Beispiel zeigt, dass die VC-Theorie (und alle anderen verteilungsabhängigen Methoden) keine Möglichkeit beinhalten gleichmäßige Konvergenz gegen die wahre Lösung zu garantieren, somit ist ein guter Schätzer auch in diesem Rahmen scheinbar "*reine Glückssache*".

In Kapitel A geben wir weitere Beispiele für das von Cherkassky et al. (1999) beobachtete Verhalten, diesmal für Neuronale Netze. Im Rahmen dieser Dissertation wurde als Demonstrator ein Wavelet Neuronales Netz programmiert, das speziell die Untersuchung von Instabilitäten in der Modellauswahl erlaubt. In Kapitel A wird der entwickelte Code genauer beschrieben.

Wir konkretisieren nun unsere bislang ein wenig heuristisch anmutenden Argumente. Die Komplexitätstheorie und die Kompaktifizierungs-Theorie wie sie in dieser Dissertation dargestellt wurden, garantieren gleichmäßige Konvergenz der Risiken in N (gleichmäßig bezüglich der Funktionen im Hypothesenraum), erstere für allgemeine Performance-Funktionen, letztere nur für die quadratische Abweichung. In Abschnitt **??**, Satz 5.8, werden wir eine Klasse von Algorithmen beschreiben, die eine Folge in M konstruieren, so dass das die mittlere quadratische Abweichung monoton gegen die Zielfunktion konvergiert (modulo des unvermeidlichen Bias-Terms). Die Existenz einer solche Folge wurde wie erwähnt durch Barron (1993) für Neuronale Netze sichergestellt. Der Punkt ist allerdings, dass auch die so konstruierte Folge $(\hat{f}_M)_{M\in\mathbb{N}}$, $\hat{f}_M \in \mathscr{F}_M$ und $\mathscr{F}_M \subset \mathscr{F}$, nur im *quadratischen Mittel* gleichmäßig gegen $\hat{f}_{opt}$ konvergiert (im folgenden sei $\Xi = (Y - f(X))^2$), also

$$\lim_{M\to\infty} \mathbb{E}\left[\left|\hat{f}_M(X) - \hat{f}_{opt}(X)\right|^2\right] = \lim_{M\to\infty} \int_{\Omega_1} \left|\hat{f}_M(X) - \hat{f}_{opt}(X)\right|^2 \mathrm{d}\mathbb{P}_1 = 0\,.$$

Stochastisch ausgedrückt konvergiert also die Folge von Zufallsvariablen $\hat{f}_M(X)$ im 2-ten Mittel gegen $\hat{f}_{opt}(X)$. Wegen

$$\Lambda_\Xi(\hat{f}_M) \leq \mathbb{E}\left[\left|\hat{f}_M(X) - \hat{f}_{opt}(X)\right|^2\right] + \mathbb{E}\left[\left|\hat{f}_{opt}(X) - f(X)\right|^2\right] + \mathbb{E}\left[E^2\right]$$

entspricht dies also der Konvergenz der Differenz zwischen kontinuierlichem Risiko und Bias-Fehler:

$$\begin{gathered}\mathbb{E}\left[\left|\hat{f}_M(X) - \hat{f}_{opt}(X)\right|^2\right] \xrightarrow{M\to\infty} 0 \\ \Downarrow \\ \Lambda_\Xi(\hat{f}_M) - \underbrace{\left(\mathbb{E}\left[\left|\hat{f}_{opt}(X) - f(X)\right|^2\right] + \mathbb{E}\left[E^2\right]\right)}_{\text{Bias-Fehler}} \xrightarrow{M\to\infty} 0\,. \end{gathered} \tag{4.22}$$

Die Norm sei eine allgemeine Norm in einem nicht näher spezifizierten Hilbertraum. In Satz 4.12 haben wir den Bias-Fehler für einen kompakten Hypothesenraum nach oben abgeschätzt. (4.22) ist nichts anderes als

$$\lim_{M\to\infty}\lim_{N\to\infty}\left(\widetilde{\Lambda}^N_\Xi(\hat{f}_{N,M}) - \text{bias}\right) = 0\,.$$

Dass die Grenzwerte im allgemeinen nicht vertauscht werden können haben wir schon an einem Trivialbeispiel gesehen, aber auch der greedy-Algorithmus 5.8 zeigt, dass dies auch in der algorithmischen Praxis ein massives Problem darstellt. Wir greifen zur Illustration auf Satz 5.8 vor, der im Falle eines vereinfachten Neuronalen Netzes (Stufenfunktion statt regularisierte sigmoide Funktion) auf Lee, Bartlett & Williamson (1995) aufbauend folgende Fehlerabschätzung für den Fall $N, M < \infty$ liefert:

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_M - \hat{f}_{opt}\right\|^2 - \left\|\hat{f}_{opt} - \hat{f}\right\|^2\right] \leq C\left(\frac{R^2}{M} + \frac{MdR^2\ln(dN)}{N}\right)\,,$$

wobei $C > 0$ und $\|f\| \leq R$ für alle $f \in \mathscr{F}$ (für die genauen Voraussetzungen siehe Satz 5.8). Man sieht sofort, dass die Grenzprozesse auf der rechten Seite der Ungleichung nicht vertauschen. Die Divergenz ist nichts anderes als der schon des öfteren diskutierte Effekt des *Overfittings*.

Das folgende Lemma beweist, dass aus der Konvergenz im p-ten Mittel direkt die Konvergenz im Maße (Konvergenz in Wahrscheinlichkeit bzw. stochastische Konvergenz) folgt:

Lemma 4.3. *Sei $(\hat{f}_M)_{M\in\mathbb{N}}$ eine gemäß Satz 5.8 konstruierte Funktionenfolge in M und $(\hat{f}_N)_{N\in\mathbb{N}}$, $\hat{f}_N = \operatorname{argmin}_{f\in\mathscr{F}} \Lambda^N_\Xi(f)$, die Folge von Funktionen, die das empirische Risiko für N Trainingsdaten minimieren. Die Folgen seien jeweils als Folgen von Zufallsvariablen auf Ω_1 zu verstehen. Dann konvergiert $\hat{f}_{M/N}$ auch dem Maße nach gegen $\hat{f}_{opt}$, also*

$$\forall \varepsilon > 0 \; : \; \lim_{M/N \to \infty} \mathbb{P}_1 \left[\left\{ \omega \in \Omega_1 : \left| \hat{f}_{M/N}(\omega) - \hat{f}_{opt}(\omega) \right| > \varepsilon \right\} \right] = 0 \, .$$

Wir schreiben $\hat{f}_{M/N}$ um auszudrücken, dass diese Aussagen sowohl für die Folge in N als auch in M gelten. $\hat{f}_{opt}$ bezeichnet die jeweilige Grenzfunktion.

Beweis. Wir zeigen allgemeiner, dass aus L^p-Konvergenz auch Konvergenz dem Maße nach folgt. Sei $L^p(\Omega, \mathcal{F}, \mu)$ mit endlichem Maß μ gegeben, B ein Banach-Raum mit Norm $\| \cdot \|_B$ und $(f_n)_{n \in \mathbb{N}}$, $f_n : \Omega \to B$ messbar, eine Funktionenfolge mit

$$\lim_{n \to \infty} \|f_n - f\|_p = \lim_{n \to \infty} \left(\int_\Omega \|f_n - f\|_B^p \, \mathrm{d}\mu \right)^{1/p} = 0 \, .$$

Es genügt hierauf die Tschebyscheff-Ungleichung anzuwenden:

$$\mu \left\{ \omega \in \Omega : \left| f_n(\omega) - f(\omega) \right| > \varepsilon \right\} \leq \frac{1}{\varepsilon^p} \int_\Omega \|f_n - f\|_B^p \, \mathrm{d}\mu$$

für alle $\varepsilon > 0$. ■

Aus der Konvergenz im p-ten Mittel folgt aber keineswegs zwingend auch fast-sichere Konvergenz (maßtheoretisch Konvergenz fast-überall), also

$$\mathbb{P}_1 \left[\left\{ \omega \in \Omega_1 : \lim_{M/N \to \infty} \hat{f}_{M/N}(\omega) = \hat{f}_{opt}(\omega) \right\} \right] = 1 \, .$$

Es ist nämlich nicht schwer eine Funktionenfolge f_n zu konstruieren, die zwar im quadratischen Mittel, und damit auch dem Maße nach, gegen eine Funktion f konvergiert, aber nicht fast-sicher. Ein typisches Gegenbeispiel ist

$$f_n(X(\omega) = x) = \begin{cases} 1, \text{ für } \frac{m}{2^k} \leq x \leq \frac{m+1}{2^k} \\ 0, \text{ sonst,} \end{cases}$$

auf $([0,1], \mathrm{Bor}, \lambda)$, wobei n gemäß $n = 2^k + m$, $0 \leq m < 2^k$, eindeutig zerlegt wird. λ bezeichne das Lebesgue-Maß auf $\mathrm{Bor}([0,1])$. Diese Folge konvergiert sowohl im p-ten Mittel als auch stochastisch gegen $f(X(\omega)) = 0$, weil $\mathbb{E}[|f_n(X) - f(X)|^p] = 2^{-k}$. Nun nimmt aber $f_n(X(\omega))$ für ein ω unendlich oft den Wert 1 bzw. 0 an, so dass keine Konvergenz fast-überall vorliegt.
An sich zeigt dieses Beispiel nur, dass die Umkehrung im Continuous Mapping Theorem

$$X_n \xrightarrow{\text{f.s.}} X \quad \Longrightarrow \quad g(X_n) \xrightarrow{\text{f.s.}} g(X) \, , \tag{4.23}$$

$(X_n)_{n=1}^\infty$ Folge von k-dimensionalen Zufallsvariablen und $g : \mathbb{R}^k \to \mathbb{R}^l$ stetig, eben *nicht* gilt.

Um Äquivalenz in (4.23) zu garantieren benötigen wir also offensichtlich mehr als nur fast-sichere Konvergenz der Risiken. Es sei M fest gewählt. Wir betrachten die Konvergenz in N: Angenommen es sei für $\omega \in \Omega_1 \times \Omega_2$ und $f_M \in \mathscr{F}_M$ beliebig

$$\widetilde{\Lambda}^N(f_M, \omega) \geq \widetilde{\Lambda}^N(\hat{f}_{N,M}, \omega) \xrightarrow{\text{f.s.}} \Lambda(\hat{f}_{N,M}) \geq \Lambda(\hat{f}) \,,$$

wobei $\hat{f}_{N,M} = \operatorname{argmin}_{f \in \mathscr{F}_M} \widetilde{\Lambda}^N_{\Xi}(f)$. Es folgt sofort:

$$\widetilde{\Lambda}^N(f_M, \omega) - \Lambda(f_M) \geq \widetilde{\Lambda}^N(\hat{f}_{N,M}, \omega) - \Lambda(f_M) \geq \widetilde{\Lambda}^N(\hat{f}_{N,M}, \omega) - \Lambda(\hat{f}_{N,M})$$

und somit

$$\left|\widetilde{\Lambda}^N(\hat{f}_{N,M}, \omega) - \Lambda(f_M)\right| \leq \max\left\{\left|\widetilde{\Lambda}^N(f_M, \omega) - \Lambda(f_M)\right|, \left|\widetilde{\Lambda}^N(\hat{f}_{N,M}, \omega) - \Lambda(\hat{f}_{N,M})\right|\right\} \,.$$

Damit die rechte Seite der Abschätzung beschränkt ist, muss das Risiko *gleichmäßig fast-sicher* in N konvergieren, also

$$\mathbb{P}^\infty\left[\left\{\omega \in \Omega^\infty : \lim_{N\to\infty} \sup_{f_M \in \mathscr{F}_M} \left|\widetilde{\Lambda}^N(f_M, \omega) - \Lambda(f_M)\right| = 0\right\}\right] = 1 \,,$$

wobei Ω^∞ das Cartesische Produkt $\prod_{j=1}^\infty (\Omega_1 \times \Omega_2)$ bezeichnet und $\mathbb{P}^\infty$ das Wahrscheinlichkeitsmaß auf $(\Omega^\infty, \mathcal{F}^\infty)$, so dass $\mathbb{P}^\infty = \bigotimes_{j=1}^\infty \mathbb{P}^j$. Es folgt dann sofort

$$\left|\widetilde{\Lambda}^N(\hat{f}_{N,M}, \omega) - \Lambda(\hat{f}_M)\right| \leq \sup_{f_M \in \mathscr{F}_M} \left|\widetilde{\Lambda}^N(f_M) - \Lambda(f_M)\right| \xrightarrow{\text{f.s.}} 0 \,.$$

Ist das Minimum $\Lambda(\hat{f}_M)$ eindeutig, so konvergiert wie in Satz 4.1 gezeigt $(\hat{f}_{N,M})$ sogar fast-sicher gegen $\hat{f}_M$, ansonsten ist die Konvergenz lediglich in Wahrscheinlichkeit. Wir haben zur Genüge diskutiert, dass sich die gleichmäßige fast-sichere Konvergenz der Risiken z.B. durch einen kompakten Hypothesenraum garantieren lässt (Satz 4.9 und die nachfolgende Bermerkung). Man siehe auch Satz 4.2 und die hieran anschließende Bemerkung 4.2.
Nach dem Satz von Egorov (s. z.B. Werner (2006)) folgt aus der fast-sicheren Konvergenz auch die fast-gleichmäßige Konvergenz in N falls das Maß endlich ist, d.h. für jedes $\varepsilon > 0$ existiert eine Menge $A \in \mathcal{F}_1$ mit $\mathbb{P}_1[A] < \varepsilon$, so dass $(\hat{f}_{N,M})$ gleichmäßig auf $\Omega_1 \backslash A$ gegen $\hat{f}_M$ konvergiert.
In M ist genau dieser Schluss nicht zulässig und eine gleichmäßig fast-sichere Konvergenz der Risiken wie für N macht in diesem Fall auch keinen Sinn: Die Folge der Risiken $(\widetilde{\Lambda}^N(\hat{f}_{N,M}))_{M\in\mathbb{N}}$ ist keine Folge von unabhängigen Zufallsvariablen!

Zusammengefasst suchen wir im folgenden ein Verfahren, das folgendes leistet:

1) Eine Teilfolge $(\hat{f}_k)_{k\in I\subset\mathbb{N}}$ von $(\hat{f}_{N,M})_{N,M\in\mathbb{N}}$ wird so gewählt, dass das Risiko

$$\mathcal{A}(\hat{f}_k) - \text{bias} \xrightarrow{\text{f.s.}} 0$$

und sogar gleichmäßig.

2) Die Folge $(\hat{f}_k)_{k\in\mathbb{N}}$ selbst konvergiert fast-gleichmäßig auf einer kompakten Input-Menge gegen die Grenzfunktion $\hat{f}_{opt}$.

Zu Beginn von Kapitel 6 werden wir den Hypothesenraum weiter einschränken um unsere Forderungen erfüllen zu können. Abschnitt 6.1 beschreibt wie insbesondere Neuronale Netze so modifiziert werden können, dass das Runge Phänomen verhindert wird und eine inhärente Robustheit des Netzwerkes gegenüber Störungen in den Eingabedaten entsteht.

5

Robustheit von Lernproblemen

Bislang beschäftigte sich diese Dissertation mit den Approximationseigenschaften von Schätzern und den zu Grunde liegenden Hypothesenräumen, wenn eine endliche Anzahl von Trainingsdaten vorliegt. In diesem Kapitel bringen wir nun einen neuen Aspekt ins Spiel, und zwar den der Robustheit oder Stabilität eines Netzwerkes.

5.1 Stabilität von Algorithmen

In diesem Abschnitt präsentieren wir die Grundlagen für eine Sensitivitätsanalyse eines Systems und untersucht wie stark Perturbationen im Trainingsdatensatz den resultierenden Schätzer beeinflussen können. Insbesondere werden wir aufzeigen, in wie fern die Begriffe Konsistenz aus dem vorangegangenen Abschnitt, Stabilität und Generalisierung zusammenhängen.

Klassische Regularisierung
Wir möchten an dieser Stelle die enge Verknüpfung des Bias-Variance-Problems und dessen Lösbarkeit mit der weit verbreiteten Methode der Regularisierung aufzeigen. In diesem Ansatz wird die Stabilität eines Algorithmus und die Kondition des Lernproblems betrachtet. Die klassische Definition von Stabilität eines (linearen) Operators M ist, dass er stetig von den Input-Daten abhängt. Seien also zwei metrische Räume X, Y gegeben und $x \in X$, $y \in Y$, so dass $M : X \to Y$ und

$$y = Mx\,. \tag{5.1}$$

Damit das inverse Problem, also x zu bestimmen, wenn y gegeben ist, gut-gestellt ist, muss

1) M^{-1} stetig sein, also für alle $\varepsilon > 0$ gibt es ein $\delta > 0$, so dass für alle $y_1, y_2 \in Y$ mit $\|y_1 - y_2\|_Y < \delta$ auch $\|M^{-1}y_1 - M^{-1}y_2\|_X < \varepsilon$ und

2) eine eindeutige Lösung existieren.

Bei Lernproblemen ist, wie erwähnt, die Eindeutigkeit der Lösung ein massives Problem, kann aber durch Einführung von so genannten "almost minimizers" erreicht werden. Der grundlegende Gedanke der *Regularisierung* ist (s. z.B. Engl, Hanke & Neubauer (1999)) die Stabilität des inversen Problems durch geschickte Wahl von $\mathscr{F}$ zu erreichen. Im klassischen Ansatz der Tikhonov-Regularisierung (s. Tikhonov & Arsenin (1977)) wird das Problem $y = Mx$ durch das folgende ERM-Problem ersetzt:

$$\underset{x \in X}{\operatorname{argmin}} \|Mx - y\|_Y^2 \,.$$

In unserem Fall ist immer ein Trainingsdatensatz $\mathcal{T} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$ gegeben, so dass das klassische ERM-Problem die folgende Form annimmt:

$$\underset{f \in \mathscr{F}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \Xi \left(\boldsymbol{y}_i, f(\boldsymbol{x}_i) \right) \,,$$

wobei wir als Funktionenraum entsprechend Beispiel 4.6 einen RKHS wählen, also $\mathscr{F} = \overline{I_K(B_R)}$. B_R ist wie zuvor eine Kugel vom Radius R in H_K. Wie in dem Beispiel erwähnt ist I_K für glatte K kompakt. Das Minimierungsproblem hat somit eine eindeutige Lösung, und zwar

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} c_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) \,.$$

Im Fall $\Xi = (f(\boldsymbol{x}) - \boldsymbol{y})^2$ ist der Koeffizienten-Vektor $\boldsymbol{c}$ gegeben durch die Gleichung

$$\boldsymbol{y} = K\boldsymbol{c} \,,$$

wobei es nun von der Matrix K abhängt, ob das ERM-Problem gut- oder schlecht-gestellt ist. Um dies sicherzustellen wird in der Tikhonov-Regularisierung das ERM-Problem erweitert um einen Regularisierungsterm[1]:

$$\underset{f \in \mathscr{F}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{y}_i - f(\boldsymbol{x}_i) \right)^2 + \lambda \|f\|_K^2 \,.$$

Die Koeffizienten von f ergeben sich nun durch das Gleichungssystem

$$\boldsymbol{y} = (K + N\lambda \mathbb{1})\boldsymbol{c} \,.$$

Die Empfindlichkeit von f in Bezug auf Störungen in den Daten $\boldsymbol{y}$ kann nun über die Konditionszahl abgeschätzt werden:

[1] Man bemerke: K muss symmetrisch und positiv definit sein, so dass $\|\cdot\|_K$ eine Norm ist. Die Erweiterung auf allgemeinere Kern-Funktionen findet sich z.B. in Smola & Schölkopf (1998).

$$\kappa(\mathrm{ERM}) = \|K + N\lambda\mathbb{1}\|\|K + N\lambda\mathbb{1}\|^{-1} ,$$

falls die Matrix nicht-singulär ist.
Es ist interessant den allgemeinen Fall zu betrachten. Es gilt folgender Satz:

Satz 5.1. *$\mathscr{F}$ sei ein RKHS gemäß Beispiel 4.6 und $\mathcal{T} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$ ein Trainingsdatensatz. Dann ist die Lösung von*

$$\operatorname*{argmin}_{f \in \mathscr{F}} \frac{1}{N} \sum_{i=1}^{N} \Xi\left(\boldsymbol{y}_i, f(\boldsymbol{x}_i)\right) + \lambda \|f\|_K^2 \tag{5.2}$$

gegeben durch

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} c_i \, K(\boldsymbol{x}, \boldsymbol{x}_i) ,$$

wobei die Koeffizienten aber nicht mehr durch Lösen eines linearen Gleichungssystems gefunden werden können.

Beweis. Für einen Beweis dieses Satzes und die Angabe eines Algorithmus zur Bestimmung der Koeffizienten siehe Girosi (1998).

Die Regularisierung in (5.2) mit $\lambda > 0$ ist implizit nichts anderes als die Beschränkung der Lösung auf eine Hyperkugel im RKHS, also ein Optimierungsproblem mit Nebenbedingungen, was bekanntermaßen umgeschrieben werden kann in die Minimierung/Maximierung der zugehörigen Lagrange-Funktion, daher der Lagrange-Multiplikator λ (man vergleiche mit Lemma 4.1 und dessen Beweis). Das Lernproblem

$$\operatorname*{argmin}_{f \in \mathscr{F}} \frac{1}{N} \sum_{i=1}^{N} \Xi\left(\boldsymbol{y}_i, f(\boldsymbol{x}_i)\right)$$

unter der Nebenbedingung $\|f\|_K \leq R$ wird also umgewandelt in das Problem die Extrema der Lagrange-Funktion

$$\mathcal{L}(f, \lambda) = \frac{1}{N} \sum_{i=1}^{N} \Xi\left(\boldsymbol{y}_i, f(\boldsymbol{x}_i)\right) + \lambda \left(\|f\|_K^2 - R^2\right)$$

bezüglich f und λ zu finden. Es ist klar, dass λ die Empfindlichkeit der Lagrange-Funktion gegenüber Änderungen der Nebenbedingung "misst". Im folgenden Abschnitt werden wir detaillierter auf den Begriff der Stabilität von allgemeinen Lernproblemen eingehen.

5.1.1 Hypothesenstabilität

Der Tikhonov-Ansatz aus Abschnitt 4.4.2 ist erweiterbar auf allgemeine Lernprobleme. Das ERM-Problem definiert einen Algorithmus

$$\begin{aligned} A_{\Xi,\mathscr{F}} : \mathscr{T} &\longrightarrow \mathscr{F} \\ \mathcal{T} &\longmapsto \hat{f}_{\mathcal{T}}(X) = \hat{f}(X, \hat{w}_{\mathcal{T}}) \,, \end{aligned}$$

der einer Stichprobe einen Schätzer aus dem Hypothesenraum zuordnet. A entspricht M^{-1} in (5.1) und der (diskrete) Trainingsdatensatz $\mathcal{T}$ entspricht y. A ist im allgemeinen natürlich nicht linear. Die Frage ist nun wie A auf Änderungen in $\mathcal{T}$ reagiert. Ist $\mathscr{F}$ zuvor festgelegt, so minimiert $A_{\Xi,\mathscr{F}}$ für ein gegebenes Ξ und Datenpunkte $\mathcal{T}$ den Ausdruck $\mathbb{E}[\Xi]$ innerhalb von $S\{\Xi(Y, \hat{f}(X, w)) : w \in W\}$. Hat $\mathcal{T}$ die Länge N, so ist wie zuvor $\mathcal{T} = \{Z_1, ..., Z_N\}$. Die Z_i, $i = 1, ..., N$, haben alle dieselbe Verteilung $\mathbb{P}_Z$ und seien voneinander unabhängig. Störungen in der Trainingsdatenmenge verstehen wir in dreierlei Hinsicht:

1) Ersetzen von k Werten $Z_j, ..., Z_{j+k}$ durch geänderte, aber ebenfalls $\mathbb{P}_Z$-verteilte Werte $\tilde{Z}_j, ..., \tilde{Z}_{j+k}$:

$$\mathcal{T} = \{Z_1, ..., Z_N\} \longrightarrow \mathcal{T}^{\tilde{k}} := \{Z_1, ..., Z_{j-1}, \tilde{Z}_j, ..., \tilde{Z}_{j+k}, Z_{j+k+1}, ..., Z_N\} \,.$$

2) Ersetzen von k Werten durch beliebige Werte $\kappa_1, ..., \kappa_k \in \mathbb{R}$:

$$\mathcal{T} = \{Z_1, ..., Z_N\} \longrightarrow \mathcal{T}^{\kappa_k} := \{Z_1, ..., Z_{j-1}, \kappa_1, ..., \kappa_k, Z_{j+k+1}, ..., Z_N\} \,.$$

3) Entfernen von k Werten:

$$\mathcal{T} = \{Z_1, ..., Z_N\} \longrightarrow \mathcal{T}^{k'} := \{Z_1, ..., Z_{j-1}, Z_{j+k+1}, ..., Z_N\} \,.$$

Ist $k = 1$ so schreiben wir schlicht $\mathcal{T}^{\sim}$, $\mathcal{T}^{\kappa}$ und $\mathcal{T}'$. j bezeichne in diesem Fall den Index des entfernten/ersetzten/geänderten Wertes. In den Betrachtungen zur Konsistenz des Approximationsproblems war der Fokus gerichtet auf eine Abschätzung der Wahrscheinlichkeit, dass das auf Basis einer gezogenen Stichprobe der Länge N berechnete empirische Risiko von dem eigentlichen Risiko abweicht:

$$\mathbb{P}^N \left[\sup_{w \in W} \left| \widetilde{\Lambda}^N_{\Xi,\hat{f}}(\cdot, w) - \Lambda_{\Xi,\hat{f}}(w) \right| > \varepsilon \right] \,, \tag{5.3}$$

wobei das empirische Risiko gegeben war durch

$$\Lambda^N_{\Xi,\hat{f}}(w) = \frac{1}{N} \sum_{i=1}^{N} \Xi \left(\boldsymbol{y}_i, \hat{f}(\boldsymbol{x}_i, w) \right) \,.$$

$\mathbb{P}^N = (\mathbb{P}_1 \otimes \mathbb{P}_2)^N$ bezeichnete das Wahrscheinlichkeitsmaß auf dem Stichprobenraum $\mathscr{T} = (\Omega_1 \times \Omega_2)^N$. Wir verallgemeinern diese Notation im folgenden und schreiben

ganz analog $\mathbb{P}^{\mathcal{T}} = (\mathbb{P}_1 \otimes \mathbb{P}_2)^{|\mathcal{T}|}$. Die Konsistenz des ERM-Problems ist äquivalent zur gleichmäßigen Beschränktheit von (5.3) auf $S = \{\Xi(Y, \hat{f}(X, w)) : w \in W\}$ in Bezug auf das konkrete Wahrscheinlichkeitsmaß $\mathbb{P}^N$ (einen Hinweis wie eine von $\mathbb{P}^N$ unabhängige Darstellung konstruiert werden kann findet sich z.B. in Karandikar & Vidyasagar (2002)).

Wird allerdings die *Stabilität* des Algorithmus $A_{\Xi,\mathscr{F}}$ untersucht, so steht im Vordergrund wie stark der Output des Algorithmus in Abhängigkeit von der Stichprobe $\mathcal{T}$ schwankt. Dies lässt sich auf mehrere Arten messen, im Endeffekt kommt es darauf an, dass

$$\|\Xi(Y, \hat{f}_{\mathcal{T}}) - \Xi(Y, \hat{f}_{\mathcal{T}'})\|$$

in einer von der Definition der Stabilität abhängigen Norm punktweise oder gleichmäßig beschränkt ist. Gilt dies für einen Algorithmus, so ist er "stabil" in Bezug auf

1) die Bestimmung von $\hat{f}_{\mathcal{T}}$, d.h. wenn $\mathcal{T}$ variiert kommt trotzdem ein ganz ähnlicher Schätzer heraus,
2) in Bezug auf die Minimierung des Risikos.

Die folgende formale Definition findet sich in ähnlicher Form in Kearns & Ron (1999):

Definition 5.1 ((ε, β)-Hypothesen-Stabilität). *Ein Algorithmus $A_{\Xi,\mathscr{F}}$ hat Hypothesen-Stabilität (ε, β), wenn*

$$\mathbb{P}^{\mathcal{T}} \left[d_X \left(A_{\Xi,\mathscr{F}}(\mathcal{T}), A_{\Xi,\mathscr{F}}(\mathcal{T}') \right) \geq \varepsilon \right] \leq \beta \,,$$

wobei der Abstand zweier Funktionen h, h' bezüglich der Zufallsvariablen X gegeben ist durch

$$d_X(h, h') := \mathbb{P}_X \left[h(X) \neq h'(X) \right] \,.$$

Anmerkung 5.1. Die Hypothesen-Stabilität kann auch über die Performance-Funktion definiert werden:

$$\mathbb{P}^{\mathcal{T}} \left[d_Z \left(\Xi(\hat{f}_{\mathcal{T}}, Z), \Xi(\hat{f}_{\mathcal{T}'}, Z) \right) \geq \varepsilon \right] \leq \beta \,.$$

Ob die beiden Definition übereinstimmen hängt offensichtlich von der Performance-Funktion Ξ ab, es gilt aber allgemein

$$\mathbb{P}_X \left[\hat{f}_{\mathcal{T}}(X) \neq \hat{f}_{\mathcal{T}'}(X) \right] \geq \mathbb{P}_Z \left[\Xi(\hat{f}_{\mathcal{T}}, Z) \neq \Xi(\hat{f}_{\mathcal{T}'}, Z) \right] \,,$$

d.h. ist ein Algorithmus Hypothesen-stabil in Bezug auf die Performance-Funktion, so ist er es auch in Bezug auf die Schätzfunktion selbst.

Definition 5.2 (β-Hypothesen-Stabilität). *Ein Algorithmus $A_{\Xi,\mathscr{F}}$ heißt β-Hypothesen-stabil in Bezug auf die Performance-Funktion Ξ, wenn*

$$\mathbb{E}_{\overline{\mathcal{T}}} \left[\mathbb{E} \left[\left| \Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}'}, Z) \right| \right] \right] \leq \beta \,.$$

$\mathbb{E}_{\overline{\mathcal{T}}}$ bezeichnet wie zuvor den Erwartungswert über alle Stichproben der Länge $|\mathcal{T}|$ aus dem zu Grunde liegenden Stichprobenraum $\mathscr{T}$.

Bousquet & Elisseeff (2002) betrachten auf der Hypothesen-Stabilität aufbauend folgendes:

Definition 5.3 (β-Punktweise & β-gleichmäßige Stabilität). *Punktweise Stabilität eines Algorithmus $A_{\Xi,\mathscr{F}}$ in dem Trainingspunkt Z_j ist definiert als*

$$\forall\, j \in \{1,...,N\}\,, \quad \mathbb{E}_{\mathcal{T}}\left[\left|\Xi(\hat{f}_{\mathcal{T}}, Z_j) - \Xi(\hat{f}_{\mathcal{T}'}, Z_j)\right|\right] \leq \beta\,.$$

Die Stabilität von A heißt gleichmäßig, wenn

$$\forall\, \mathcal{T} \in \mathscr{T}\,, \quad \forall\, j \in \{1,...,N\}\,, \quad \left\|\Xi(\hat{f}_{\mathcal{T}}, \cdot) - \Xi(\hat{f}_{\mathcal{T}'}, \cdot)\right\|_{\infty} \leq \beta\,.$$

Kearns & Ron (1999) führen auch eine so genannte Fehler-Stabilität ein:

Definition 5.4 (β-Fehler-Stabilität). *Ein Algorithmus $A_{\Xi,\mathscr{F}}$ hat Fehler-Stabilität β in Bezug auf die Performance-Funktion Ξ, falls gilt:*

$$\forall\, \mathcal{T} \in \mathscr{T}\,, \quad \forall\, j \in \{1,...,N\}\,, \quad \left|\mathbb{E}\left[\Xi(\hat{f}_{\mathcal{T}}, Z)\right] - \mathbb{E}\left[\Xi(\hat{f}_{\mathcal{T}'}, Z)\right]\right| \leq \beta\,.$$

Anmerkung 5.2. Man beachte, dass in allen diesen Definitionen β von N abhängt, also $\beta = \beta(N) := \beta_N$. Wie genau diese Abhängigkeit aussieht spielt keine große Rolle, wichtig ist nur, dass für einen *stabilen* Algorithmus β nicht mit N zunehmen darf. Es muss also stets $\beta_{N+1} \leq \beta_N$ sein.

Zunächst zwei Lemmata:

Lemma 5.1. *Es gilt:*

$$\textit{Gleichmäßige Stabilität} \Longrightarrow \textit{Hypothesen-Stabilität} \Longrightarrow \textit{Fehler-Stabilität}\,.$$

Beweis. Der Beweis folgt direkt aus den Definitionen.

Lemma 5.2. *$A_{\Xi,\mathscr{F}}$ sei gleichmäßig stabil. Dann gilt:*

$$\forall\, \mathcal{T} \in \mathscr{T}\,, \quad \forall\, j \in \{1,...,N\}\,, \quad \left|\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}^{\sim}}, Z)\right| \leq 2\beta\,.$$

Beweis. Man notiere lediglich:

$$\left|\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}^{\sim}}, Z)\right| \leq \left|\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}'}, Z)\right| + \left|\Xi(\hat{f}_{\mathcal{T}'}, Z) - \Xi(\hat{f}_{\mathcal{T}^{\sim}}, Z)\right| \leq \beta + \beta\,,$$

weil $A_{\Xi,\mathscr{F}}$ gleichmäßig stabil ist. ■

Wir bezeichnen mit

$$\Lambda_{\Xi,\hat{f}}^{N\setminus j}(\hat{w}_{\mathcal{T}'}) = \frac{1}{|\mathcal{T}'|} \sum_{i\in\{1,\ldots,|\mathcal{T}'|\}} \Xi\left(\boldsymbol{y}_i, \hat{f}(\boldsymbol{x}_i, \hat{w}_{\mathcal{T}'})\right)$$

den empirischen Fehler für den Trainingssatz $\mathcal{T}'$ und ganz analog die empirischen Fehler $\Lambda_{\Xi,\hat{f}}^{N}(\hat{w}_{\mathcal{T}^\kappa})$ und $\Lambda_{\Xi,\hat{f}}^{N}(\hat{w}_{\mathcal{T}^\sim})$. Uns interessiert nun die Abweichung der empirischen Fehler vom minimalen Risiko $\Lambda_{\Xi,\hat{f}}(\hat{w})$. Die folgende Darstellung und die Notation ähnelt der Analyse aus Bousquet & Elisseeff (2002). Es gilt folgender Satz:

Satz 5.2. *$A_{\Xi,\mathscr{F}}$ sei ein β_1-hypothesen-stabiler und β_2-punktweise-stabiler Lernalgorithmus mit $0 \leq \Xi(\hat{f}_{\mathcal{T}}, Z) \leq M$ fast überall und für alle $\mathcal{T} \in \mathscr{T}$. Dann gilt für alle $N \geq 1$, $\delta \in (0,1)$ mit Wahrscheinlichkeit von mindestens $1-\delta$*

$$\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) \leq \Lambda_{\Xi,\hat{f}}^{N}(\hat{w}_{\mathcal{T}}) + \sqrt{\frac{M^2 + 12MN\beta_2}{2N\delta}}$$

und

$$\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) \leq \Lambda_{\Xi,\hat{f}}^{N\setminus j}(\hat{w}_{\mathcal{T}'}) + \sqrt{\frac{M^2 + 6MN\beta_1}{2N\delta}}\,.$$

Beweis. Der Beweis dieses Satzes fußt auf der Tschebyscheff-Ungleichung und dem folgenden Theorem von Steele (1986): Sei $F : \mathscr{T} \to \mathbb{R}$ eine messbare Funktion. Dann gilt für die Varianz

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\left(F(\mathcal{T}) - \mathbb{E}_{\overline{\mathcal{T}}}[F(\mathcal{T})]\right)^2\right] \leq \frac{1}{2}\sum_{i=1}^{N} \mathbb{E}_{\overline{\mathcal{T}}}\left[\left(F(\mathcal{T}) - F(\mathcal{T}^\sim)\right)^2\right]\,.$$

Lemma 9 in Bousquet & Elisseeff (2002) erweitert den Satz von Steele auf die folgenden Abschätzungen:

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[(\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda_{\Xi}^{N}(\hat{w}_{\mathcal{T}}))^2\right] \leq \frac{M^2}{2N} + 3M\,\mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_{\tilde{Z}_j}\left[\left|\Xi(\hat{f}_{\mathcal{T}}, Z_j) - \hat{f}_{\mathcal{T}^\sim}, Z_j)\right|\right]\right]\,,$$

sowie

$$\begin{aligned}\mathbb{E}_{\overline{\mathcal{T}}}\left[(\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda_{\Xi}^{N}(\hat{w}_{\mathcal{T}}))^2\right] \leq \frac{M^2}{2N} &+ M\mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_{\tilde{Z}_j}\left[\mathbb{E}_{Z}\left[\left|\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}^\sim}, Z)\right|\right]\right]\right] \\ &+ M\mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_{\tilde{Z}_j}\left[\left|\Xi(\hat{f}_{\mathcal{T}}, Z_j) - \Xi(\hat{f}_{\mathcal{T}^\sim}, Z_j)\right|\right]\right] \\ &+ M\mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_{\tilde{Z}_j}\left[\left|\Xi(\hat{f}_{\mathcal{T}}, Z_k) - \Xi(\hat{f}_{\mathcal{T}^\sim}, Z_k)\right|\right]\right]\end{aligned}$$

für alle $i, k \in \{1, ..., N\}$. $\mathcal{T}^\sim$ bezeichne zur Erinnerung den Datensatz, in dem Z_j durch $\tilde{Z}_j$ ersetzt wurde, wobei letzteres identisch verteilt ist. $\mathcal{T}'$ ist der Trainingsdatensatz, in dem

Z_j entfernt wurde. Um Verwirrung zu vermeiden haben wir die Erwartungswerte indiziert mit den Zufallsvariablen bezüglich derer sie zu verstehen sind. Devroye & Wagner (1979) gaben schon ein ähnliches Lemma für den Klassifikationsfall, also $\mathcal{Y} = \{-1, 1\}$. Weiterhin wird gezeigt, dass

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\left(\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda_{\Xi}^{N\setminus j}(\hat{w}_{\mathcal{T}'})\right)^2\right] \leq \frac{M^2}{2N} + 3M\mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_Z\left[\left|\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}'}, Z)\right|\right]\right]$$

sowie

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\left(\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda_{\Xi}^{N\setminus j}(\hat{w}_{\mathcal{T}'})\right)^2\right] \leq \frac{M^2}{2N} + 2M\mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_{\tilde{Z}_j}\left[\mathbb{E}_Z\left[\left|\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}^\sim}, Z)\right| + \left|\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}'}, Z)\right|\right]\right]\right] .$$

Nun muss nur noch die Tschebyscheff-Ungleichung

$$\mathbb{P}[X \geq \varepsilon] \leq \frac{\mathbb{E}[X^2]}{\varepsilon^2} ,$$

die umgeschrieben bedeutet, dass mit Wahrscheinlichkeit $1 - \delta$, $\delta > 0$,

$$X \leq \sqrt{\frac{\mathbb{E}[X^2]}{\delta}} ,$$

auf $\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda_{\Xi,\hat{f}}^N(\hat{w}_{\mathcal{T}})$ und $\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda_{\Xi,\hat{f}}^N(\hat{w}_{\mathcal{T}'})$ angewendet werden. ■

Anmerkung 5.3. Man beachte, dass in der Abschätzung von Satz 5.2

$$\mathbb{P}^{\mathcal{T}}\left[\left|\Lambda_{\Xi}(\hat{f}_{\mathcal{T}}) - \Lambda_{\Xi}^N(\hat{f}_{\mathcal{T}})\right| > \varepsilon\right]$$

betrachtet wird im Unterschied zu der in der Lerntheorie üblichen Vorgehensweise (wie z.B. in der Ungleichung (4.13)) nach Vapnik, wo

$$\mathbb{P}^{\mathcal{T}}\left[\sup_{f\in\mathscr{F}}\left|\Lambda_{\Xi}(f) - \Lambda_{\Xi}^N(f)\right| > \varepsilon\right]$$

im Fokus steht. Wir schließen somit nahtlos an die Abschätzung in Abschnitt 4.4 an. Diese Darstellung hat den Vorteil, dass keine Abschätzung über den *gesamten* Hypothesenraum $\mathscr{F}$ angegeben werden muss, denn dies kann sehr schwierig sein. Man betrachtet vielmehr nur die Abweichung im Optimum von $\mathscr{F}$.

Anmerkung 5.4. An Satz 5.2 sehen wir sofort, dass das ERM-Problem nur *konsistent* sein kann, wenn

$$\beta_{1/2} \xrightarrow{N\to\infty} 0 .$$

Satz 5.3 benötigt allerdings eine noch stärkere Bedingung.

Für β-gleichmäßige Stabilität gilt ein ähnlicher Satz:

Satz 5.3. *$A_{\Xi,\mathscr{F}}$ sei ein β-gleichmäßig-stabiler Lernalgorithmus mit $0 \leq \Xi(\hat{f}_{\mathcal{T}}, Z) \leq M$ fast überall und für alle $\mathcal{T} \in \mathscr{T}$. Dann gilt für alle $N \geq 1$, $\delta \in (0,1)$ mit Wahrscheinlichkeit von mindestens $1-\delta$*

$$\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) \leq \Lambda^{N}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) + 2\beta + (4N\beta + M)\sqrt{-\frac{\ln\delta}{2N}}$$

und

$$\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) \leq \Lambda^{N\setminus j}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}'}) + \beta + (4N\beta + M)\sqrt{-\frac{\ln\delta}{2N}}\,.$$

Beweis. Entscheidend für diesen Satz ist die Hoeffding-Ungleichung (Hoeffding (1963)), bzw. als Verallgemeinerung die McDiarmid-Ungleichung (McDiarmid (1989)): Wie im Satz von Steele sei $F : \mathscr{T} \to \mathbb{R}$ eine messbare Funktion. Außerdem existieren Konstanten c_j, $j = 1, ..., N$, so dass

$$\sup_{\mathcal{T}\in\mathscr{T},\tilde{Z}_j} \left|F(\mathcal{T}) - F(\mathcal{T}^{\sim})\right| \leq c_j\,,$$

dann gilt:

$$\mathbb{P}^{\mathcal{T}}\left[F(\mathcal{T}) - \mathbb{E}_{\overline{\mathcal{T}}}[F(\mathcal{T})] \geq \varepsilon\right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{j=1}^{N} c_j^2}\right)\,.$$

Bousquet & Elisseeff (2002) zeigen weiterhin, dass

$$\left|\Lambda^{N}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda^{N}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}^{\sim}})\right| \leq 2\beta + \frac{M}{N}\,,$$

so dass für die McDiarmid-Ungleichung lediglich $c_j = 4\beta + M/N$ gewählt werden muss. Bousquet & Elisseeff (2002) zeigen zudem, dass

$$\begin{aligned}
&\mathbb{E}_{\overline{\mathcal{T}}}\left[\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda^{N}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}})\right] = \mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_{\tilde{Z}_j}\left[\Xi(\hat{f}_{\mathcal{T}}, \tilde{Z}_j) - \Xi(\hat{f}_{\mathcal{T}^{\sim}}, \tilde{Z}_j)\right]\right]\,,\\
&\mathbb{E}_{\overline{\mathcal{T}}}\left[\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}'}) - \Lambda^{N\setminus j}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}'})\right] = 0\,,\\
&\mathbb{E}_{\overline{\mathcal{T}}}\left[\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda^{N\setminus j}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}'})\right] = \mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}_{Z}\left[\Xi(\hat{f}_{\mathcal{T}}, Z) - \Xi(\hat{f}_{\mathcal{T}^{\sim}}, Z)\right]\right]\,.
\end{aligned}$$

Dies führt direkt zu

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda^{N}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}})\right] \leq 2\beta\,,$$

so dass

$$\mathbb{P}^{\mathcal{T}}\left[\Lambda_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - \Lambda^{N}_{\Xi,\hat{f}}(\hat{w}_{\mathcal{T}}) - 2\beta_N \geq \varepsilon\right] \leq \exp\left(-\frac{2N\varepsilon^2}{(4N\beta_N + M)^2}\right) =: \delta\,.$$

Es ergibt sich sofort die erste Behauptung. Die Zweite folgt in gleicher Weise. ■

Anmerkung 5.5. Dieser Satz benötigt für Konsistenz des Lernproblems sogar

$$\frac{\beta N}{\sqrt{N}} \xrightarrow{N\to\infty} 0 \,.$$

In Abschnitt 6.2.3 werden wir auf diese Bedingung zurückkommen.

5.1.2 Regularisierte Lernprobleme

In diesem Unterabschnitt werden wir als Zielfunktion des Lernalgorithmus im Speziellen regularisierte empirische Risiken der Form

$$\Lambda^N_{\Xi,R}(f) = \frac{1}{N}\sum_{i=1}^{N} \Xi(f, Z_i) + \lambda R(f) \,, \tag{5.4}$$

$$\Lambda^{N\setminus j}_{\Xi,R}(f) := \frac{1}{N-1}\sum_{i\in\{1,\dots,N\}\setminus\{j\}} \Xi(f, Z_i) + \lambda R(f) \tag{5.5}$$

betrachten, wobei der Stabilisierungsterm eine Funktion $R : \mathscr{F} \to \mathbb{R}_+$ ist. Für allgemeine R existiert bislang in der Literatur keine Stabilitätsbetrachtung. Wir werden allerdings in Abschnitt 6.2.3 einen neuen Konstruktionsalgorithmus für Neuronalen Netze präsentieren und diesen auch in Hinblick auf seine Hypothesen-Stabilität untersuchen.
Für den Spezialfall eines RKHS hingegen, also $R = \|\cdot\|_K$, wobei K die Kern-Funktion bezeichnet, lässt sich ein Zusammenhang zwischen β und λ direkt angeben (s. Bousquet & Elisseeff (2002)):

Satz 5.4. *Sei $\mathscr{F}$ ein RKHS gemäß Beispiel 4.6, so dass für alle $x \in \mathcal{X}$ $K(x,x) \le \kappa^2 < \infty$. Weiterhin gelte für Ξ eine Lipschitzbedingung entsprechend Satz 4.9. Dann ist ein Lernalgorithmus mit*

$$\hat{f}_{\mathcal{T}} = \operatorname*{argmin}_{f\in\mathscr{F}} \frac{1}{N}\sum_{i=1}^{N} \Xi(f, Z_i) + \lambda\|f\|_K^2$$

gleichmäßig Hypothesen-stabil mit

$$\beta \le \frac{\kappa^2}{2c_1^2\lambda N} \,.$$

Beweis. Für den Beweis wird das folgende allgemeine Lemma benötigt:

Lemma 5.3. *Es sei Ξ konvex. Weiterhin sei $R : \mathscr{F} \to \mathbb{R}_+$ so gewählt, dass die empirischen Risiken (5.4) und (5.5) jeweils ein Minimum $\hat{f}^N$ bzw. $\hat{f}^{N\setminus j}$ in $\mathscr{F}$ zumindest besitzen. Dann gilt für alle $k \in [0,1]$ und alle $i \in \{1,\dots,N\}$:*

$$\left\{R\left(\hat{f}^N\right) - R\left(\hat{f}^N + k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right)\right\} + \left\{R\left(\hat{f}^{N\setminus j}\right) - R\left(\hat{f}^{N\setminus j} - k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right)\right\} \le \frac{k}{\lambda c_1 N}\left|\hat{f}^{N\setminus j}(X_i) - \hat{f}^N(X_i)\right| \,.$$

Beweis. Weil Ξ konvex ist, gilt dasselbe auch für $\Lambda_\Xi^{N\setminus j}(f)$ (definiert analog zu (5.5) für $R = 0$), d.h.

$$\Lambda_\Xi^{N\setminus j}\left(\hat{f}^N + k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right) - \Lambda_\Xi^{N\setminus j}\left(\hat{f}^N\right) \\ \leq k\left(\Lambda_\Xi^{N\setminus j}\left(\hat{f}^{N\setminus j}\right) - \Lambda_\Xi^{N\setminus j}\left(\hat{f}^N\right)\right).$$

Eine ganz analoge Ungleichung erhalten wir, wenn die Rollen von $\hat{f}^{N\setminus j}$ und $\hat{f}^N$ vertauscht werden. Durch Addition dieser und der ursprünglichen Ungleichung erhalten wir:

$$\Lambda_\Xi^{N\setminus j}\left(\hat{f}^N + k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right) - \Lambda_\Xi^{N\setminus j}\left(\hat{f}^N\right) \\ + \Lambda_\Xi^{N\setminus j}\left(\hat{f}^{N\setminus j} - k\left(\hat{f}^N - \hat{f}^{N\setminus j}\right)\right) - \Lambda_\Xi^{N\setminus j}\left(\hat{f}^{N\setminus j}\right) \leq 0\,. \quad (5.6)$$

Nach Voraussetzung ist

$$\Lambda_{\Xi,R}^N\left(\hat{f}^N\right) - \Lambda_{\Xi,R}^N\left(\hat{f}^N + k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right) \leq 0\,, \\ \Lambda_{\Xi,R}^{N\setminus j}\left(\hat{f}^{N\setminus j}\right) - \Lambda_{\Xi,R}^{N\setminus j}\left(\hat{f}^N - k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right) \leq 0\,,$$

so dass die Summe dieser beiden Ungleichungen zusammen mit (5.6) folgendes ergibt:

$$\Xi\left(\hat{f}^N, Z_i\right) - \Xi\left(\hat{f}^N, Z_i\left(\hat{f}^N + k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right), Z_i\right) \\ + \lambda N\left\{R\left(\hat{f}^N\right) - R\left(\hat{f}^N + k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right)\right. \\ \left. + R\left(\hat{f}^{N\setminus j}\right) - R\left(\hat{f}^{N\setminus j} - k\left(\hat{f}^{N\setminus j} - \hat{f}^N\right)\right)\right\} \leq 0$$

für alle $i \in \{1, ..., N\}$. Jetzt muss nur noch die Lipschitz-Bedingung für Ξ angewendet werden. ■

Wir notieren nun, dass

$$2\|\hat{f}^{N\setminus j} - \hat{f}^N\|_K^2 \leq \frac{1}{\lambda c_1 N}\left|\hat{f}^{N\setminus j}(X_i) - \hat{f}^N(X_i)\right|,$$

so dass wegen $|f(x)| \leq \|f\|_K\sqrt{K(x,x)}$ für alle $f \in \mathscr{F}$ und alle $x \in \mathcal{X}$

$$\left|\hat{f}^{N\setminus j}(X_i) - \hat{f}^N(X_i)\right| \leq \kappa\|\hat{f}^{N\setminus j} - \hat{f}^N\|_K$$

$$\|\hat{f}^{N\setminus j} - \hat{f}^N\|_K \leq \frac{\kappa}{2c_1\lambda N} \,.$$

Mit der Lipschitz-Bedingung für Ξ ergibt sich nun die Behauptung. ■

Anmerkung 5.6. Insbesondere lässt sich Satz 5.4 auf die kleinste-Quadrate-Regression anwenden. Dieser Algorithmus ist also gleichmäßig Hypothesen-stabil.

Anmerkung 5.7. Wie zu Ende von Abschnitt 5.1 angedeutet besteht ein direkter Zusammenhang zwischen dem Lagrange-Parameter λ und dem Kugelradius R aus Abschnitt 4.4: $\lambda \propto R$. In Satz 5.4 wurde soeben $\beta \propto \lambda^{-1}$ im Spezialfall eines RKHS gezeigt. Konsequenterweise ließe sich also für ein RKHS in allen Abschätzungen aus Abschnitt 4.4 R durch β^{-1} ersetzen, so dass der Hypothesenraum $\mathscr{F}$ in Hinblick auf seine Stabilitätseigenschaften optimiert werden kann! Wir kommen auf diese Beobachtung im Rahmen der rekursiven Algorithmen noch einmal zurück.

5.2 Einfluss von Ausreißern

Die in den vorangegangenen Abschnitten definierte Hypothesen-Stabilität betrachtet die Sensibilität der Performance-Funktion Ξ auf Störungen in der Trainingsdatenmenge und zwar im Erwartungswert über alle Trainingsdatensätze. In diesem Abschnitt betrachten wir den bislang vernachlässigten Fall 2) der zu Beginn von 5.1.1 definierten gestörten Trainingsdatensätze und analysieren

$$\mathbb{P}_1 \left[\sup_{\mathcal{T}^{\kappa_k}} \left| \hat{f}_{\mathcal{T}^{\kappa_k}} - \hat{f}_{\mathcal{T}} \right| > \varepsilon \right] \,. \tag{5.7}$$

Uns interessiert nun also analog zu der Frage nach Konvergenz im Mittel im Unterschied zu gleichmäßiger Konvergenz nicht die mittlere Sensibilität des Algorithmus, sondern die Maximale.

In Kapitel 4 haben wir bereits bemerkt, dass die Wahl der Performance-Funktion ganz entscheidenden Einfluss auf die Güte der Approximation (Güte zu verstehen als Genauigkeit in Kombination mit Robustheit) hat. Die meisten Autoren missachten dieses äußerst wichtige Detail flächendeckend und betrachten praktisch ausschließlich die für robuste Neuronale Netze ungeeignete least-squares Performance-Funktion, weil hinlänglich bekannt ist, dass $\Xi = (Y - \hat{f})^2$ anfällig für Ausreißer ist (durch Quadrieren gehen große Werte stärker in den Erwartungswert ein).

Die Idee einen im Supremum robusten Trainingsprozess durch die Verwendung einer "robusten" Performance-Funktion zu erreichen ist somit recht naheliegend. Eine Möglichkeit (5.7) zu analysieren stellt der in der Statistik bekannte *breakdown point* dar. Aufbauend auf Rousseeuw & Leroy (1987) geben wir nun eine stochastische Version der Störungs-Immunität eines Neuronalen Netzes:

Definition 5.5. *Es sei mit der in Abschnitt 5.1.1 eingeführten Notation*

$$\mathcal{T} = \{Z_1, ..., Z_N\} \longrightarrow \mathcal{T}^{\kappa_k} := \{Z_1, ..., Z_{j-1}, \kappa_1, ..., \kappa_k, Z_{j+k+1}, ..., Z_N\}$$

ein Datensatz, in dem k Werte durch beliebige Werte $\kappa_1, ..., \kappa_k \in \mathbb{R}$ ersetzt wurden. Wir definieren nun die Störungs-Immunität des Netzwerkes als

$$\xi := \min_k \left\{ \frac{k}{N} \ : \ \mathbb{P}_1 \left[\sup_{\mathcal{T}^{\kappa_k}} \left| \hat{f}_{\mathcal{T}^{\kappa_k}} - \hat{f}_{\mathcal{T}} \right| = \infty \right] > 0 \right\} .$$

ξ wird breakdown-point (Bruchpunkt) genannt.

Es ist offensichtlich, dass $|\hat{f}_{\mathcal{T}^{\kappa_k}} - \hat{f}_{\mathcal{T}}|$ beschränkt ist, wenn der Vektor der Gewichte $\boldsymbol{u}$ beschränkt ist. Für RBFNs hingegen (bzw. für alle Basisfunktionen mit *kompaktem Träger*) ergibt sich das folgende interessante Detail, das in ähnlicher Form in Li & Leiss (2001) allerdings nur deterministisch vorkommt:

Satz 5.5. *RBFNs und WNNs sind störungs-immun gegen eine beliebige Anzahl von horizontalen Ausreißern, d.h.*

$$Z = (X, Y) \longrightarrow (\kappa, Y) ,$$

wenn die Basisfunktionen hinreichend gut lokalisiert sind.

Beweis. Der "worst-case" tritt sicherlich ein, wenn alle Daten gestört sind, also $k = N$. Es gilt nun mit Wahrscheinlichkeit 1

$$\forall\, i \in \{1, ..., M\} \ \forall \boldsymbol{x} \in \mathcal{T} \ : \ \phi(\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{t}_i\|^2}{\sigma_i} \right) \approx 0 ,$$

falls σ_i klein genug ist. Dies entspricht also der Forderung, dass die radialen Basisfunktionen hinreichend gut lokalisiert sind. Dann haben horizontale Ausreißer offensichtlich keinen unbeschränkten Einfluss, weil

$$\hat{f}_{\mathcal{T}^{\kappa_k}}(\boldsymbol{x}) \approx 0$$

und somit $\left|\hat{f}_{\mathcal{T}^{\kappa_k}} - \hat{f}_{\mathcal{T}}\right|$ fast überall beschränkt ist. ■

Satz 5.6. *RBFNs und WNNs sind hingegen nicht störungs-immun gegen vertikale Ausreißer, d.h.*

$$Z = (X, Y) \longrightarrow (X, \kappa) .$$

Beweis. Da κ nach Voraussetzung beliebig groß wird ergibt sich sofort, dass auch $\left|\hat{f}_{\mathcal{T}^{\kappa_k}} - \hat{f}_{\mathcal{T}}\right|$ auf mehr als einer $\mathbb{P}_1$-Nullmenge beliebig ist. ■

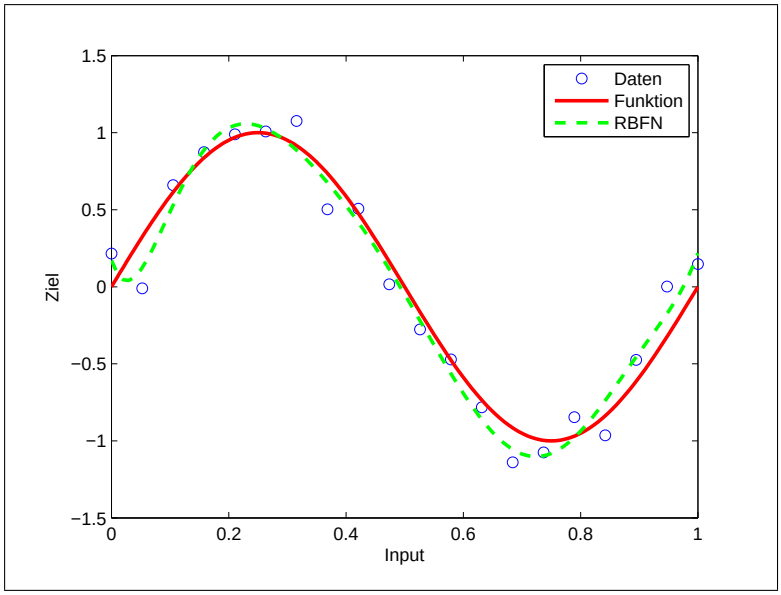


Abb. 5.1. Approximation der Sinus-Funktion durch ein RBFN. Gezeigt sind 20 Datenpunkte mit zufälligen Störungen (Standardabweichung 0.2). Gewählt wurde ein RBFN mit 7 nodes und Gauss'schen Glockenkurven als Basisfunktionen.

Abb. 5.1, 5.2 und 5.3 zeigen diesen Sachverhalt für ein numerisches Beispiel. Als Grundlage dient das im Rahmen dieser Dissertation erstellte Wavelet Neuronale Netzwerk, das für dieses Beispiel auf 2D und Gauss'sche radiale Basisfunktionen spezialisiert wurde. Man erkennt sofort die Anfälligkeit für vertikale Ausreißer.

Der tiefer liegende Grund für die Anfälligkeit der Netzwerke gegen Ausreißer ist die gewählte Performance-Funktion. Es ist klar, dass die least-squares Methode durch Quadrieren große Werte in der Bildung des Erwartungswertes stärker bewertet als kleine. Rousseeuw & Leroy (1987) zeigen hingegen, dass die Methode der *least-trimmed-squares* (LTS) den höchsten breakdown-point besitzt. Statt den Durchschnitt aller quadrierten Residuen zu betrachten, werden in der LTS-Methode nur die q kleinsten Residuen berücksichtigt, so dass

$$A^N_{\mathrm{LTS}}(\hat{f}) = \frac{1}{2}\sum_{i=1}^{q} r_i^2 \,,$$

wobei $r_i = \boldsymbol{y}_i - \hat{f}(\boldsymbol{x}_i)$ und $r_1^2 \leq r_2^2 \leq ... \leq r_q^2 \leq ... \leq r_N^2$. Die Wahl von q ist offensichtlich nicht trivial. Li & Leiss (2001) zeigen folgenden Satz:

Satz 5.7. *Wird bei einem RBFN die LTS-Methode für die Risiko-Minimierung verwandt und*

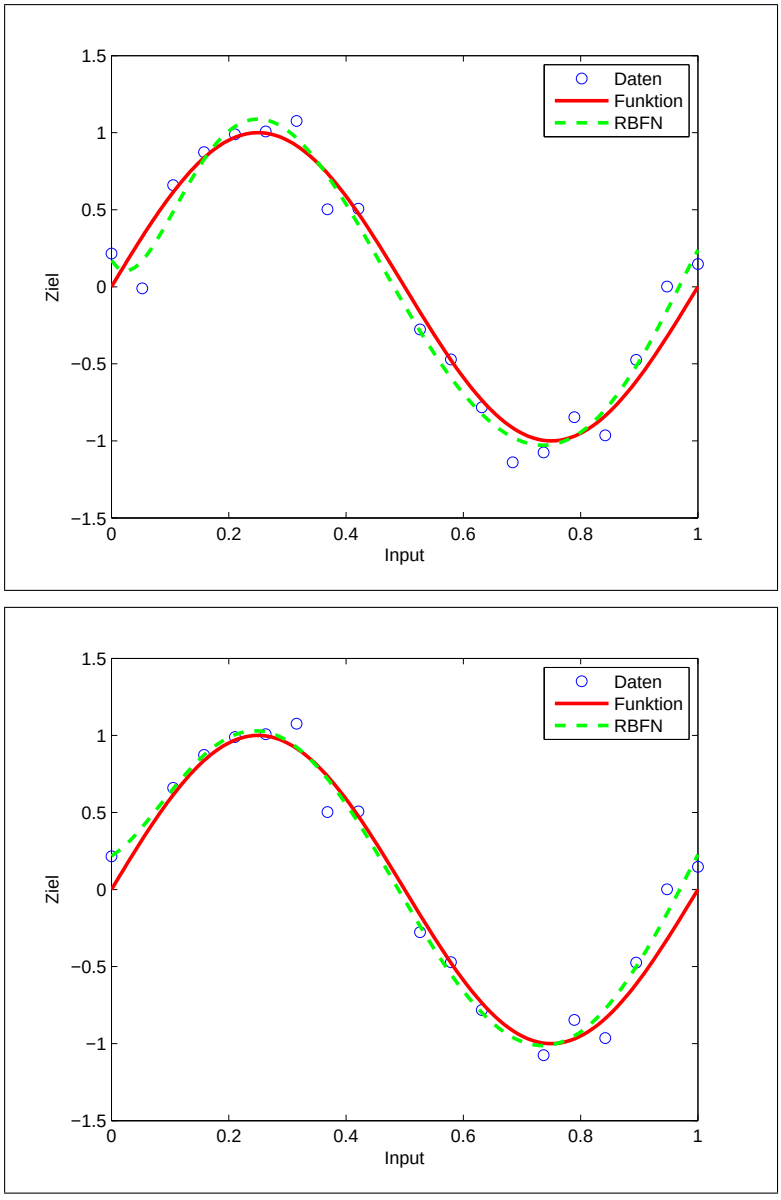


Abb. 5.2. Dasselbe Netzwerk wie in Abb. 5.1, nun wurden allerdings horizontale Ausreißer eingefügt. In der ersten Zeichnung wurde $x_{10} = 2$ gesetzt, in der Zweiten $x_2 = 2$, $x_{10} = 2$ und $x_{14} = 2$.

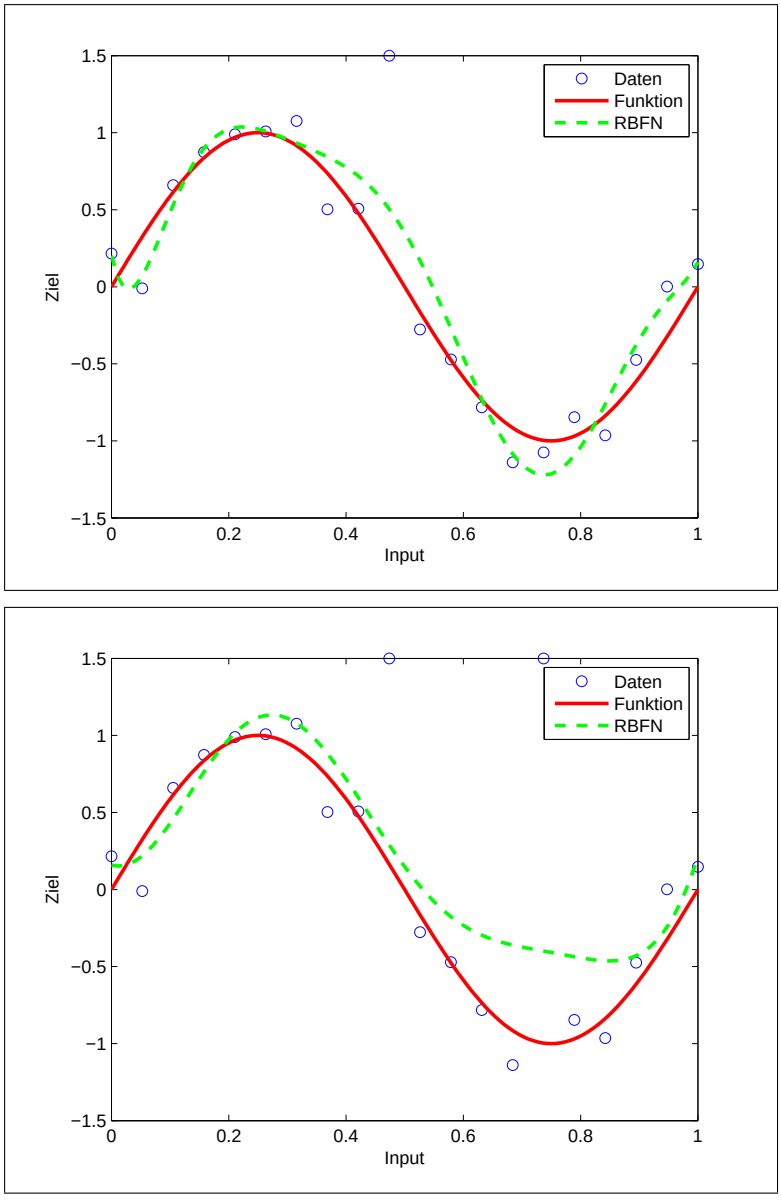


Abb. 5.3. Dasselbe Netzwerk wie in Abb. 5.1, nun mit vertikalen Ausreißern. In der ersten Zeichnung wurde $y_{10} = 1.5$ gesetzt, in der Zweiten $y_{10} = 1.5$ und $y_{15} = 1.5$.

$$q = \left\lfloor \frac{N}{2} \right\rfloor + \left\lfloor \frac{M+1}{2} \right\rfloor$$

gewählt, so ist

$$\xi = \frac{1}{N} \left(\left\lfloor \frac{N-M-1}{2} \right\rfloor + 1 \right) .$$

Die im Rahmen dieser Dissertation erstellte Software wendet die LTS-Methode auf ein Wavelet Neuronales Netz an. Abb. 5.4 zeigt die resultierende LTS-Approximation bei zwei vertikalen Ausreißern. Man erkennt sofort die verbesserte Approximation.

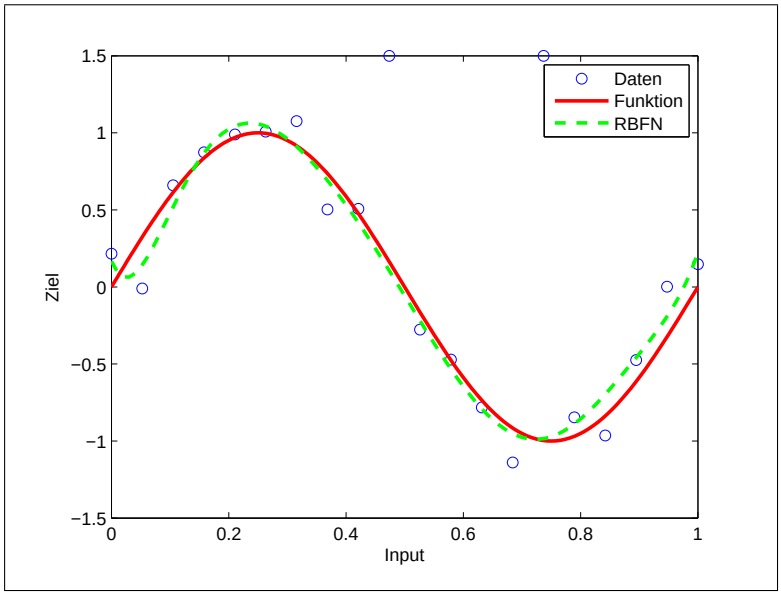


Abb. 5.4. Approximation der Sinus-Funktion durch ein LTS-RBFN mit 7 nodes und 2 Ausreißern. Der Vergleich mit den Abb. 5.1 und 5.3 zeigt die Wirkung der LTS-Methode.

Rousseeuw & Leroy (1987) weisen allerdings auf die schlechte Effizienz der LTS-Methode für normalverteilte Störungen hin und es entsteht durch das Ordnen der Residuen ein Overhead im Rechenaufwand von $O(N \log N)$. Li & Leiss (2001) geben einen alternativen Algorithmus zur Steigerung der Effizienz (adaptive RBF Lernmethoden mit LTS). Zudem zeigen aber Ergebnisse von Rousseeux & Bassett Jr. (1991), dass die Zielfunktion für die LTS-Methode wie für die meisten Methoden mit hohem breakdown-point nicht konvex ist.

Es wurden in den letzten Jahren auch andere robuste Trainingsprozeduren vorgeschlagen wie z.B. informationstheoretische Cluster-Algorithmen (Böhm et al. (2006)). Ein anderer Ansatz findet sich in Bors & Pitas (2001). Hier wird das Risiko nicht über den Erwartungswert, sondern den Median berechnet (Median RBFNs).

Das in dieser Dissertation erstellte Wavelet-Neuronale Netz verfolgt einen etwas anderen Ansatz. In Kooperation mit Pohl wurde eine Anwendung von Cluster-Algorithmen in der Erstellung der Wavelet-*Bibliothek* erarbeitet. Hierdurch können zusätzlich Ausreißer identifiziert und gegebenenfalls ausgeschlossen werden. Für Details s. Pohl (2007).

Die in diesem Abschnitt diskutierten Methoden beziehen sich allesamt auf die Minderung der Effekte von Ausreißern in den Trainingsdaten, ein Standard-Feld der Statistik. Ein robuster Trainingsprozess garantiert aber nicht die gleichmäßige Konvergenz des Schätzers gegen die Zielfunktion. Hierfür muss "das Netzwerk selbst" robust sein, d.h. die Optimierung muss unter Nebenbedingungen durchgeführt werden (*Multi-Objective Neuronale Netze*), so dass der gesamte Prozess der Modellbildung robust ist. Ein Hinweis wie solche Nebenbedingungen aussehen können ist schon in Satz 5.5 versteckt. Dort fordern wir eine "ausreichende" Lokalisation der Basisfunktionen, was mit einer Einschränkung für σ verknüpft ist. Die triviale Abschätzung

$$\left|\hat{f}_{\mathcal{T}^{\kappa_k}} - \hat{f}_{\mathcal{T}}\right| \leq \sum_{i=1}^{M} \left|u_i^{\mathcal{T}^{\kappa_k}}\right| + \sum_{i=1}^{M} \left|u_i^{\mathcal{T}}\right|$$

zeigt, dass für Neuronale Netze die Summe der Gewichte in irgendeiner Weise beschränkt sein muss. Diese Ideen werden wir in Kapitel 6 weitertreiben und in Abschnitt 6.2.3 einen neuen Konstruktionsalgorithmus auf dieser Basis vorstellen.

5.3 Schlechte Kondition des Optimierungsproblems

In diesem Abschnitt beschäftigen uns mit der dritten Komponente des Lernproblems: Finden des Minimums in $\mathscr{F}$. In Abschnitt 4.5 haben wir die Schwierigkeit dieses Problems aufgrund der i.A. nicht garantierbaren Konvexität des Hypothesenraums bereits angedeutet.

Der Großteil an Literatur über Neuronale Netze verwendet als Optimierungsmethode eine Variante von gradient descent. Aufgrund der Komplexität des Hypothesenraumes zeigen die meisten Algorithmen ein langsames Konvergenzverhalten (s. Abschnitt 5.4 für einen alternativen Ansatz). Nun kommt allerdings noch hinzu, dass der Trainingsprozess sogar schlecht konditioniert sein kann: Saarinen, Bramley & Cybenko (1993) zeigen, dass während der Optimierung schlecht konditionierte Hesse-Matrizen auftreten und das sogar recht häufig. Eine Erhöhung der Ordnung des Verfahrens erhöht somit nicht die

Effizienz des Algorithmus.

In Abschnitt 5.2 haben wir den Einfluss der Performance-Funktion auf die Robustheit des Schätzers dargestellt. In der Praxis wird für Neuronale Netze in der Regel die least-squares Risiko-Minimierung verwendet, also

$$\hat{w}_{\mathcal{T}^N} = \operatorname*{argmin}_{w \in W} \Lambda^N_\Xi(w) = \operatorname*{argmin}_{w \in W} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i, w))^2 = \operatorname*{argmin}_{w \in W} \frac{1}{N} \sum_{i=1}^{N} \Xi_i^2(w) \,.$$

Der Gradient von $\Lambda^N(w)$ ist gegeben durch

$$\nabla \Lambda^N(w) = J(w)^T \Xi(w) \,,$$

wobei $[J_{ij}] = \partial \Xi_i / \partial w_j$ die Jacobi-Matrix von Ξ bezeichnet. Die Hesse-Matrix von $\Lambda^N(w)$ ist dann

$$H(w) = J(w)^T J(w) + \sum_{i=1}^{N} \Xi_i(w) H_i(w) \,.$$

$H_i(w)$ bezeichnet die Hesse-Matrix von $\Xi_i(w)$. Optimierungsverfahren unterscheiden sich nun in der Suchrichtung und der Festlegung der Schrittweite. Die folgende Tabelle fasst die Suchrichtungen für verschiedene Optimierungsverfahren zusammen. Es wird stets angenommen, dass J vollen Rang hat.

Algorithmus	Suchrichtung
Steepest Descent	$-J^T f$
Conjugate Gradient	$-J^T f + \beta \tilde{p}$, β skalar, $\tilde{p}$ Suchrichtung im letzten Schritt
Newton	$-(J^T J + \sum_{i=1}^{N} \Xi_i H_i)^{-1} J^T f$
Gauss-Newton	$-(J^T J)^{-1} J^T f$
Levenberg-Marquardt	$-(J^T J + \rho_k \mathbb{1})^{-1} J^T f$
Quasi-Newton	$-(J^T J + B_k)^{-1} J^T f$, B_k erfüllt Quasi-Newton Bedingung

Wir fassen die Eigenschaften der Methoden kurz zusammen (für detaillierte Informationen s. z.B. Jarre & Stoer (2003) oder Riedmüller & Ritter (1992)):
Steepest Descent Algorithmen haben eine q-lineare Konvergenzrate mit einer asymptotischen Fehler Konstante proportional zu $(\kappa - 1)/(\kappa + 1)$, wobei κ die Konditionszahl der Hesse-Matrix ist. Conjugate Gradient Methoden konvergieren i.A. linear, Quasi-Newton Methoden sogar superlinear. Hierbei darf $H(\hat{w})$ allerdings nicht singulär sein und $J^T J + B_k$ muss $H(\hat{w})$ entlang der Suchrichtung approximieren. Newton-Methoden konvergieren quadratisch, die Hesse-Matrix darf aber nicht singulär sein bei $\hat{w}$. Die Konvergenzrate der Gauss-Newton und Levenberg-Marquardt-Methode hängen von der Größe des Residuums bei $\hat{w}$ ab. Ist es Null, so konvergiert Gauss-Newton quadratisch, in so fern $J(\hat{w})$ vollen Rang hat. Für Levenberg-Marquardt gilt dies für $\rho_k = 0$. Ist das Residuum

groß, so haben beide Methoden lineare Konvergenzraten. Die Größen ρ_k und B_k werden so gewählt, dass die Suchrichtungen der beiden Methoden stets wohldefiniert sind, so dass die Levenberg-Marquardt und Quasi-Newton-Methode trotz Singularitäten von H oder Rangdefizit von J konvergieren.

Saarinen, Bramley & Cybenko (1993) zeigen, dass für Neuronale Netzwerk-Probleme eine große Anzahl von Spalten der Jacobi-Matrix leicht (fast) linear abhängig sein können, so dass J rangdefizitär in der 2-Norm wird. Auch andere Autoren weisen auf die Rang-Problematik der Jacobi-Matrix (s. McKeown, Stella & Hall (1997)) und auf die damit verbundenen Konsequenzen für den Modellierungsprozess hin (s. Rivals & Personnaz (2004)).

Das im Rahmen dieser Dissertation erstellte Demonstrator-Wavelet-Neuronale Netz beinhaltet u.a. eine robustifizierte Version des Levenberg-Marquardt-Algorithmus: In den Experimenten zeigt sich, dass das in Abschnitt 4.5 beschriebene Runge-Phänomen unter anderem durch einen in den LM-Algorithmus eingefügten Konditions-Penalty-Term weitgehend vermieden werden kann (für numerische Beispiele und weitere Details s. Kapitel A).

5.4 Konstruktionsalgorithmen mit aufsteigender Komplexität

In diesem Abschnitt wollen wir unsere Ergebnisse auf eine spezielle Klasse von Algorithmen anwenden.

Die Menge der Neuronalen Netzen bildet i.A. keine Orthonormalbasis des Raumes der zu approximierenden Funktionen. Eine Optimierung aller $M(d+2)$ Parameter ist offensichtlich für steigendes M immer aufwendiger und für große M mit erheblichen Rechenzeiten verbunden. Es scheint daher ein sinnvoller Ansatz bei einem kleinen M mit der Optimierung zu beginnen und die Optimallösung für $M+1$ auf dieser Lösung aufzubauen, so dass eine rekursive Vorschrift der Art

$$\hat{f}_{M+1} = \text{Funktion}(\hat{f}_M)$$

entsteht. Der Komplexität des Hypothesenraums wird also Schritt für Schritt vergrößert. Eine diese Strategie verfolgende Klasse von Algorithmen wird unter dem Sammelbegriff “greedy“-Algorithmus zusammengefasst. Sie haben einerseits den Vorteil den Rechenaufwand bei der Optimierung dramatisch zu reduzieren. Andererseits erlauben sie eine Abschätzung des Approximationsfehlers (bias-Fehler). In Lee, Bartlett & Williamson (1995) wird $\hat{f}_{M+1}$ auf Grundlage von $\hat{f}_M$ so konstruiert, dass die Approximationsgenauigkeit aus Barron (1993) (s. Satz 4.15) erreicht wird. Allerdings betrachten die Autoren in Bezug auf Neuronale Netze nur die Heaviside-Funktion $\sigma(x) = 1$ für $x > 0$, $\sigma(x) = 0$

sonst als Aktivierungsfunktion. Es wird jedoch folgender Satz in allgemeinem Rahmen bewiesen:

Satz 5.8. *Sei H ein Hilbert-Raum mit Norm $\|\cdot\|$. Sei $G \subset H$ mit $\|g\| \leq b$ für alle $g \in G$. Bezeichne weiterhin $d_f := \inf_{g' \in co(G)} \|g' - f\|$. Im ersten Schritt wähle $\hat{f}_1$ so, dass*

$$\|\hat{f}_1 - f\|^2 \leq \inf_{g \in G} \|g - f\|^2 + \varepsilon_1 \, .$$

Konstruiere dann rekursiv $\hat{f}_k$ so, dass

$$\|\hat{f}_k - f\|^2 \leq \inf_{g \in G} \|\alpha \hat{f}_{k-1} + \overline{\alpha} g - f\|^2 + \varepsilon_k \, ,$$

wobei $\alpha = 1 - 2/(k+1)$, $\overline{\alpha} = 1 - \alpha$, $c \geq b^2$ und $\varepsilon_k \leq 4(c - b^2)/(k+1)^2$. Dann gilt für alle $k \geq 1$

$$\|\hat{f}_k - f\|^2 - d_f^2 \leq \frac{4c}{k} \, .$$

Insbesondere ist dieser Algorithmus also nur konsistent, wenn f in der konvexen Hülle von G liegt!
Vor Kurzem publizierten Barron et al. (2008) einen rekursiven Algorithmus, der auch das Erlernen von Funktionen außerhalb der konvexen Hülle erlaubt. Burger & Hofinger (2005) erweiterten die Ergebnisse auch in Hinblick auf Konvergenz in stärkeren Normen, insbesondere Sobolev-Normen. Leider erlauben die Ergebnisse dieser Autoren keine Erweiterung auf $W^{1,\infty}$, was für die in Kapitel 6 diskutierten Räumen notwendig wäre.
Lee, Bartlett & Williamson (1995) geben auf Grundlage von Satz 5.8 eine Abschätzung für den Gesamtfehler der Approximation an. Die Autoren betrachten allerdings lediglich Netzwerke der Form

$$\mathscr{N}_{B,k} = \Bigg\{ g : \mathbb{R}^d \to \mathbb{R} \, : \, g = \sum_{i=1}^{k} u_i H(A_i); \, , t_i \in \mathbb{R}, \boldsymbol{a}_i \in \mathbb{R}^d, \\ \sum_{i=1}^{k} |u_i| \leq B, \text{ maximal } D \text{ Komponenten } a_{ij} \text{ von } \boldsymbol{a}_i \text{ sind ungleich Null} \} \Bigg\} \, ,$$

wobei H die Heaviside-Funktion bezeichnet. Weiterhin sei $\mathscr{N}_B := \bigcup_{k=1}^{\infty} \mathscr{N}_{B,k}$.

Satz 5.9. *Es bezeichne $f_{opt} = \inf_{g \in \mathscr{N}_B} \|g - f\|^2$. $\hat{f}_k$ bezeichne den durch Algorithmus 5.8 produzierten Schätzer im k-ten Schritt. Dann gilt:*

$$\mathbb{E}_{\overline{\mathcal{T}}} \left[\|\hat{f}_k - f\|^2 - \|f_{opt} - f\|^2 \right] = O\left(\frac{B^2}{k} + \frac{kDB^2 \ln(dN)}{N} \right) \, . \tag{5.8}$$

Beweis. Der Beweis dieses Satzes kann analog zu der Analyse in Abschnitt 4.4.2 geführt werden. Lee, Bartlett & Williamson (1995) geben einen ähnlichen Beweis. ■

Anmerkung 5.8. Dieser Satz beschreibt einen trade-off zwischen der Approximationstiefe k, der Anzahl der Trainingsdaten N und dem Kugelradius B. Man erinnere sich nun an Anmerkung 5.7, wo argumentiert wurde, dass im Spezialfall eines RKHS R im Wesentlichen durch den Stabilitätsparameter β^{-1} ersetzt werden kann. Somit kann (5.8) in diesem Fall auch herangezogen werden um den trade-off zwischen Approximationstiefe k, Anzahl der Trainingsdaten N und *Stabilität des Algorithmus* zu beschreiben.

Lemma 5.4. *Für die Umsetzung von Algorithmus 5.8 werden* $O(2^D D^3 k d^{2D} N^{2(D+1)})$ *Rechenoperationen benötigt.*

Beweis. Der Beweis nutzt in erster Linie aus, dass durch die Wahl der Heaviside-Funktion als Aktivierungsfunktion in jedem Schritt ein Klassifikationsproblem gelöst werden muss, d.h. es genügt die Anzahl der möglichen Dichotomien der gegebenen Trainingsmenge zu berechnen. In jedem Schritt muss ein Satz von maximal $D+1$ linearen Gleichungen gelöst werden, die maximal $D+1$ Variablen enthalten. Weiterhin müssen in jedem der $d^D 2^{D+1} N^{D+1}$ Schritte $O(d^D N^{D+1})$ Vergleiche mit dem Trainingsdatensatz vorgenommen werden. Für die Details siehe Siehe Lee, Bartlett & Williamson (1995). ■

Wir geben nun einen neuen greedy-Algorithmus auf Basis von Satz 5.8 mit $\sqrt{k}^{-1}$-Konvergenz für Räume aufsteigender Komplexität an:

Satz 5.10. *Sei H ein Hilbert-Raum mit Norm $\|\cdot\|$ und $G_k \subset H$ mit $\|g\| \leq b_k$ für alle $g \in G_k$. Bezeichne weiterhin $d_f := \inf_{g' \in co(\overline{G})} \|g' - f\|$, wobei $\overline{G} = \cup_{k=1}^{\infty} G_k$. $co(\overline{G})$ bezeichnet die konvexe Hülle von $\overline{G}$. Im ersten Schritt wähle f_1 so, dass*

$$\|f_1 - f\|^2 \leq \inf_{g \in G_1} \|g - f\|^2 + \varepsilon_1$$

mit $\varepsilon_1 = b_1^2$. Konstruiere dann rekursiv f_k so, dass

$$\|f_k - f\|^2 \leq \inf_{g \in G_k} \|\alpha f_{k-1} + \overline{\alpha} g - f\|^2 + \varepsilon_k \,,$$

wobei $\alpha = 1 - 2/(k+1)$, $\overline{\alpha} = 1 - \alpha$ und

$$\varepsilon_k \leq 4b_k^2 \left(k^{-\frac{1}{2}} - (k+1)^{-2}\right) - 4b_{k-1}^2 \left((k-1)^{-\frac{1}{2}} - 2(k+1)^{-1}(k-1)^{-\frac{1}{2}}\right) .$$

Weiterhin sei die Folge $(b_k)_{k \in \mathbb{N}}$ so gewählt, dass so gewählt wird, dass

$$\varepsilon_k < \varepsilon_{k-1}$$

und $\varepsilon_k \to 0$ für $k \to \infty$. Dann gilt für alle $k \geq 1$

$$\|f_k - f\|^2 - d_f \leq \frac{4b_k^2}{\sqrt{k}} .$$

Beweis. Der Beweis kann ganz analog zu Satz 5.8 mit vollständiger Induktion geführt werden. Es sei $h_k \in \mathrm{co}(G_k)$ mit $\|h_k - f\| \leq d_f + \delta$, wobei $\delta > 0$. Dann lässt sich h_k in der Form $h_k = \sum_{i=1}^{p} \gamma_i g_i$ mit $g_i \in g_k$, $\gamma_i \geq 0$ und $\sum_{i=1}^{p} \gamma_i = 1$ für genügend großes p schreiben. Dann gilt für alle $\alpha \in [0,1]$:

$$\|\alpha f_{k-1} + \overline{\alpha} g - f\|^2 = \|\alpha f_{k-1} + \overline{\alpha} g - h_k\|^2 + \|h_k - f\|^2 + 2\langle \alpha f_{k-1} + \overline{\alpha} g - h_k, h_k - f\rangle \,.$$

Also ist:

$$\begin{aligned}
\|\alpha f_{k-1} + \overline{\alpha} g - f\|^2 - \|h_k - f\|^2 &= \|\alpha f_{k-1} + \overline{\alpha} g - h_k\|^2 + 2\langle \alpha f_{k-1} + \overline{\alpha} g - h_k, h_k - f\rangle \\
&= \|\alpha(f_{k-1} - h_k) + \overline{\alpha}(g - h_k)\|^2 + 2\langle \alpha f_{k-1} + \overline{\alpha} g - h_k, h_k - f\rangle \\
&= \alpha^2 \|f_{k-1} - h_k\|^2 + \overline{\alpha}^2 \|g - h_k\|^2 \\
&\quad + 2\alpha\overline{\alpha}\langle f_{k-1} - h_k, g - h_k\rangle + 2\langle \alpha f_{k-1} + \overline{\alpha} g - h_k, h_k - f\rangle \,.
\end{aligned}$$

Wir bilden nun den Durschnittswert von

$$\|\alpha \hat{f}_{k-1} + \overline{\alpha} g - f\|^2 - \|h_k - f\|^2 \,,$$

wobei g aus der Menge $\{g_1, ..., g_p\}$ mit der Wahrscheinlichkeit $\mathbb{P}[g = g_i] = \gamma_i$ unabhängig gezogen wird:

$$\begin{aligned}
&\sum_{i=1}^{p} \gamma_i \big[\alpha^2 \|f_{k-1} - h_k\|^2 + \overline{\alpha}^2 \|g_i - h_k\|^2 + 2\alpha\overline{\alpha}\langle f_{k-1} - h_k, g_i - h_k\rangle \\
&\qquad\qquad + 2\langle \alpha f_{k-1} + \overline{\alpha} g_i - h_k, h_k - f\rangle\big] \\
&= \alpha^2 \|f_{k-1} - h_k\|^2 + \overline{\alpha}^2 \sum_{i=1}^{p} \gamma_i \|g_i - h_k\|^2 + 2\alpha\overline{\alpha} \sum_{i=1}^{p} \gamma_i \langle f_{k-1} - h_k, g_i - h_k\rangle \\
&\qquad\qquad + 2\sum_{i=1}^{p} \gamma_i \langle \alpha f_{k-1} + \overline{\alpha} g_i - h_k, h_k - f\rangle \\
&= \alpha^2 \|f_{k-1} - h_k\|^2 + \overline{\alpha}^2 \left(\sum_{i=1}^{p} \gamma_i \left(\|g_i\|^2 - 2\langle g_i, h_k\rangle + \|h_k\|^2 \right) \right) \\
&\qquad\qquad + 2\sum_{i=1}^{p} \gamma_i \langle \alpha f_{k-1} + g_i - \alpha g_i - h_k, h_k - f\rangle \\
&= \alpha^2 \|f_{k-1} - h_k\|^2 + \overline{\alpha}^2 \left(\sum_{i=1}^{p} \gamma_i \|g_i\|^2 - \|h_k\|^2 \right) + 2\alpha\langle f_{k-1} - h_k, h_k - f\rangle \\
&\leq \alpha^2 \|f_{k-1} - h_k\|^2 + \overline{\alpha}^2 b_k^2 + 2\alpha\langle f_{k-1} - h_k, h_k - f\rangle
\end{aligned}$$

Es gibt also ein $g \in \{g_1, ..., g_p\}$, so dass wegen $0 \leq \alpha \leq 1$

$$\begin{aligned}\|\alpha f_{k-1} + \overline{\alpha} g - f\|^2 - \|h - f\|^2 &\leq \alpha^2 \|f_{k-1} - h\|^2 + \overline{\alpha}^2 b_k^2 + 2\alpha \langle f_{k-1} - h, h - f\rangle \\ &= \alpha\big[\alpha \|f_{k-1} - h\|^2 + 2\langle f_{k-1} - h, h - f\rangle\big] + \overline{\alpha}^2 b_k^2 \\ &\leq \alpha\big[\|f_{k-1} - h\|^2 + 2\langle f_{k-1} - h, h - f\rangle\big] + \overline{\alpha}^2 b_k^2 \,.\end{aligned}$$

Wegen

$$\|f_{k-1} - f\|^2 = \|f_{k-1} - h\|^2 + \|h - f\|^2 + 2\langle f_{k-1} - h, h - f\rangle$$

erhalten wir

$$\|f_{k-1} - f\|^2 - \|h - f\|^2 = \|f_{k-1} - h\|^2 + 2\langle f_{k-1} - h, h - f\rangle \,,$$

so dass mit $\delta \to 0$

$$\inf_{g \in G_k} \|\alpha f_{k-1} + \overline{\alpha} g - f\|^2 - d_f^2 \leq \alpha \left[\|f_{k-1} - f\|^2 - d_f^2\right] + \overline{\alpha}^2 b_k^2 \,.$$

Setzt man $k = 1$, $\alpha = 0$ und $f_0 = 0$, so erhält man:

$$\inf_{g \in G_1} \|g - f\|^2 - d_f^2 \leq b_1^2 \,.$$

Der Satz ist also für $k = 1$ bestätigt. Nun kommt der Induktionsschritt. Angenommen der Satz sei korrekt für $k - 1$, also

$$\|f_{k-1} - f\|^2 - d_f \leq \frac{4b_{k-1}^2}{\sqrt{k-1}} \,.$$

Dann gilt

$$\begin{aligned}\inf_{g \in G_k} \|\alpha f_{k-1} + \overline{\alpha} g - f\|^2 - d_f^2 + \varepsilon_k \leq \alpha \frac{4b_{k-1}^2}{\sqrt{k-1}} + \overline{\alpha}^2 b_k^2 + 4b_k^2 \Big(k^{-\frac{1}{2}} \\ -(k+1)^{-2}\Big) - 4b_{k-1}^2 \left((k-1)^{-\frac{1}{2}} - 2(k+1)^{-1}(k-1)^{-\frac{1}{2}}\right) \,.\end{aligned}$$

Mit $\alpha = 1 - 2/(k+1)$ und $\overline{\alpha} = 1 - \alpha$ ergibt sich durch nachrechnen sofort:

$$\inf_{g \in G_k} \|\alpha f_{k-1} + \overline{\alpha} g - f\|^2 - d_f^2 + \varepsilon_k \leq \frac{4b_k^2}{\sqrt{k}} \,.$$

Damit ist der Beweis komplett. ■

Anmerkung 5.9. Die Anforderungen an die Folge $(b_k)_{k \in \mathbb{N}}$ sind nicht sehr hoch, allerdings muss b_k in jedem Fall sublinear ansteigen. Wählt man z.B. $b_k = C_1 + \ln(C_2 + k/(C_3)$ mit positiven Konstanten C_1, C_2, C_3 und $C_2 + 1/C_3 \geq 1$ so sind die Anforderungen aus dem Satz erfüllt.

Anmerkung 5.10. Man bemerke, dass die Konvergenzgeschwindigkeit von $k^{-1/2}$ der aus Koiran (1994) entspricht. Die verbesserte Konvergenzrate aus Lee, Bartlett & Williamson (1995) bzw. Barron et al. (2008) lässt sich für den hier vorgestellten Algorithmus eventuell auch erreichen bei sehr geschickter Wahl von ε_k bzw. α und $\overline{\alpha}$. Wir lassen dieses Problem an dieser Stelle offen.

5.4.1 Hypothesen-Stabilität von Standard-greedy-Algorithmen

Wir analysieren nun den im vorangegangenen Abschnitt vorgestellten greedy-Algorithmus in Hinblick auf seine Hypothesen-Stabilität. Es gilt folgender Satz:

Satz 5.11. *Sei H ein Hilbertraum mit Norm $\|\cdot\|$ (induziert durch das Skalarprodukt $\langle\cdot,\cdot\rangle$ in H). Weiterhin sei $\mathscr{F}_B \subset H$ wie in Satz 5.9 mit $\|f\| \leq B$, $B > 0$, für alle $f \in \mathscr{F}_B$. Weiterhin sei $|\boldsymbol{y}| \leq B$ für alle $\boldsymbol{y} \in \mathcal{Y}$. Als Performance-Funktion sei $\Xi(f, Z) = (Y - f(X))^2$ gewählt. Dann gilt für den rekursiven greedy Algorithmus 5.8 im M-ten Schritt, $M \geq 1$, die folgende Abschätzung für die Hypothesen-Stabilität:*

$$\begin{aligned}\beta_M^{RG} &\leq \max\left\{KCB^2\left[\frac{M(M+1)}{2}\left(\frac{d\ln(dN)}{N}+\frac{d\ln(d(N-1))}{N-1}\right)+2\sum_{k=1}^{M}\frac{1}{k}\right],16B^4\right\}\\ &\approx \max\left\{KCB^2\left[\frac{M(M+1)}{2}\left(\frac{d\ln(dN)}{N}+\frac{d\ln(d(N-1))}{N-1}\right)+2(\ln M+\gamma)\right],16B^4\right\}\end{aligned} \tag{5.9}$$

mit $K, C > 0$. γ bezeichnet die Euler-Mascheroni-Konstante.

Beweis. Wir bezeichnen mit $\hat{f}_{M,\mathcal{T}}$ die Funktion in $\mathscr{F}_B$, die das empirische Risiko

$$\Lambda_{\Xi}^{N}(g) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i - g(\boldsymbol{x}_i))^2$$

für den Datensatz $\mathcal{T}$ unter der Nebenbedingung $\|g\| \leq B$ minimiert. Entsprechend bezeichne $\hat{f}_{M,\mathcal{T}'}$ die Funktion, die

$$\Lambda_{\Xi}^{N\setminus\{j\}}(g) = \frac{1}{N-1}\sum_{i\neq j}(\boldsymbol{y}_i - g(\boldsymbol{x}_i))^2$$

mit $g \in \mathscr{F}_B$, $\|g\| \leq B$ minimiert. Weiterhin bezeichnen $f \in H$ die Zielfunktion, also $f = \mathbb{E}[Y|X]$, und f_{opt} die bestmögliche Approximation innerhalb von $\mathscr{F}_B$, also $f_{opt} = \operatorname{argmin}_{f\in\mathscr{F}_B}\mathbb{E}[(\boldsymbol{y} - f(\boldsymbol{x}))^2]$. f_{opt} ist also die Projektion von f auf die konvexe Hülle von $\mathscr{F}_B$. Wir wollen nun den Term

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{M,\mathcal{T}} - \hat{f}_{M,\mathcal{T}'}\right\|^2\right],$$

$M \geq 1$, abschätzen.
Wir beginnen mit $M = 1$. Nach Definition des rekursiven greedy-Algorithmus (RGA) gilt für $\hat{f}_{M,\mathcal{T}}$ im ersten Schritt:

$$\left\|\hat{f}_{1,\mathcal{T}} - f_{opt}\right\|^2 \leq \inf_{g\in\mathscr{F}_B}\|g - f_{opt}\|^2 + \varepsilon_1\,.$$

Statt f haben wir also f_{opt} im Algorithmus verwendet. Wir benutzen Satz 5.9 um den Approximationsfehler im ersten Schritt des greedy-Algorithmus abzuschätzen:

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{1,\mathcal{T}} - f_{opt}\right\|^2\right] \leq C\left(B^2 + dB^2\frac{\ln(dN)}{N}\right) .$$

$C > 0$ ist eine Konstante. Wir schätzen ab:

$$\begin{aligned}\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{1,\mathcal{T}} - \hat{f}_{1,\mathcal{T}'}\right\|^2\right] &\leq \mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{1,\mathcal{T}} - f_{opt}\right\|^2 + \left\|\hat{f}_{1,\mathcal{T}'} - f_{opt}\right\|^2\right] \\ &\leq CB^2\left(2 + \frac{d\ln(dN)}{N} + \frac{d\ln(d(N-1))}{N-1}\right) .\end{aligned}$$

$\mathscr{F}_B$ ist total beschränkt, ebenso wie $\mathcal{Y}$, so dass Ξ die Lipschitzbedingung

$$\forall\, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{D} \subset \mathbb{R}^m,\ \forall\, \boldsymbol{y} \in \mathcal{Y} \subset \mathbb{R}^m\ :\ |\Xi(\boldsymbol{y}, \boldsymbol{y}_1) - \Xi(\boldsymbol{y}, \boldsymbol{y}_2)| \leq K\,|\boldsymbol{y}_1 - \boldsymbol{y}_2|$$

erfüllt. Wegen

$$\left\|\hat{f}_{1,\mathcal{T}} - \hat{f}_{1,\mathcal{T}'}\right\|^2 = \mathbb{E}\left[\left|\hat{f}_{1,\mathcal{T}}(X) - \hat{f}_{1,\mathcal{T}'}(X)\right|^2\right]$$

folgern wir also:

$$\mathbb{E}_{\overline{\mathcal{T}}}\left[\mathbb{E}\left[\left|\Xi(Y, \hat{f}_{1,\mathcal{T}}(X)) - \Xi(Y, \hat{f}_{1,\mathcal{T}'}(X))\right|^2\right]\right] \leq K\,\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{1,\mathcal{T}} - \hat{f}_{1,\mathcal{T}'}\right\|^2\right] .$$

Im ersten Schritt ist der RG-Algorithmus also β_1^{hyp}-Hypothesen-Stabil gemäß Definition 5.2 mit

$$\beta_1^{hyp} \leq KCB^2\left(2 + \frac{d\ln(dN)}{N} + \frac{d\ln(d(N-1))}{N-1}\right) .$$

Wir sehen aber auch, dass der RGA für einen nicht näher spezifizierten Hilbertraum H i.A. *nicht* punktweise oder sogar gleichmäßig stabil ist!

Nun der Schritt von $M-1$ nach M. Es ist nach Definition eines rekursiven greedy Algorithmus:

$$\left\|\hat{f}_{M,\mathcal{T}} - f\right\|^2 \leq \inf_{g \in \mathscr{F}_B}\left\|\alpha_M \hat{f}_{M-1,\mathcal{T}} + (1-\alpha_M)g - f\right\|^2 + \varepsilon_M .$$

Wir schätzen ab:

$$\begin{aligned}
\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{M,\mathcal{T}}-\hat{f}_{M,\mathcal{T}'}\right\|^2\right] &\leq \mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{M,\mathcal{T}}-f_{opt}\right\|^2\right]+\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{M,\mathcal{T}'}-f_{opt}\right\|^2\right] \\
&\quad +\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{M-1,\mathcal{T}}-\hat{f}_{M-1,\mathcal{T}'}\right\|^2\right] \\
&\leq \ldots \leq \sum_{k=1}^{M}\left\{\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{k,\mathcal{T}}-f_{opt}\right\|^2\right]+\mathbb{E}_{\overline{\mathcal{T}}}\left[\left\|\hat{f}_{k,\mathcal{T}'}-f_{opt}\right\|^2\right]\right\} \\
&\leq C\sum_{k=1}^{M}\left\{\frac{2B^2}{k}+dkB^2\left(\frac{\ln(dN)}{N}+\frac{\ln(d(N-1))}{N-1}\right)\right\}.
\end{aligned}$$

Aufgrund der "trivialen" Abschätzung

$$\Xi(Y,\hat{f}_{M,\mathcal{T}}(X)) \leq 4B^2$$

ergibt sich sofort der zweite Teil der max(·)-Aussage und die Behauptung ist gezeigt. ■

Wie erwähnt gibt dieser Satz lediglich eine Abschätzung der Hypothesen-Stabilität eines rekursiven greedy Algorithmus. Gleichmäßige Stabilität wäre nur garantiert für den speziellen Rahmen der stetigen Funktionen mit $\|\cdot\| = \|\cdot\|_\infty$.

Der Ausdruck (5.9) für die Hypothesen-Stabilität eines greedy-Algorithmus muss nun diskutiert werden. Im Wesentlichen ist die obere Schranke für β_M von der Form

$$\overline{\beta}_{M,B} = cB^2\left[\frac{M(M+1)}{2}h(N)+2(\ln M+\gamma)\right]. \tag{5.10}$$

Wir können dies so interpretieren, dass der Algorithmus also von Schritt zu Schritt instabiler werden kann! Wichtig ist das Verhalten von β für große Stichprobenlängen N. Prinzipiell erwartet man, dass ein Algorithmus für $N \to \infty$ eine Hypothesenstabilität von 0 aufweist. Für allgemeine rekursive greedy Algorithmen ergibt sich aber vielmehr

$$\overline{\beta}_{M,B} \longrightarrow 2cB^2(\ln M+\gamma), \tag{5.11}$$

was immer noch von M abhängt. Greedy-Algorithmen in der bislang kennen gelernten Form sind also alles andere als Hypothesen-stabil!

Teil III

Anwendung

6

Multi-Objective Neuronale Netze

Neuronale Netze werden in der Regel aufgrund ihrer Analogie zu biologischen Systemen als inhärent robust betrachtet (s. z.B. Ham & Kostanic (2001)), sowohl in Architektur als auch Output-Verhalten. Diese Eigenschaft wird aus der Tatsache abgeleitet, dass die Trainingsdaten mit Störungen versehen sind und somit das resultierende Netzwerk auch bei neuen gestörten Eingaben (in so fern sie die gleiche statistische Verteilung wie die Trainigsdaten haben, s. Abschnitt 4.2.1) gutmütig reagieren sollte. Einige Autoren (z.B. Phatak & Koren (1995)) haben allerdings das Gegenteil festgestellt und für Hardware Implementierungen auch experimentell bestätigt. In Phatak & Koren (1995), Séquin & Clay (1990) und Emmerson & Damper (1993) werden Methoden beschrieben, wie auftretende Hardware-Fehler (z.B. Gewicht bei Null eingefroren) durch Redundanzen in der Netzwerkstruktur und intelligentes Re-Training aufgefangen werden können, so dass ein Fehler-tolerantes Netzwerk entsteht. In Abschnitt 3.4 werden die Zusammenhänge zwischen Redundanzen innerhalb des Netzwerkes und Fehler-Toleranz in Bezug auf Wavelet Neuronale Netze auf theoretischer Seite diskutiert. Derartige Methoden, d.h. Nodes und die assoziierten Gewichte schlicht zu replizieren, sind aber nichts anderes als ein "Robustifizieren" eines bestehenden, d.h. bereits trainierten, Netzwerkes. Auf theoretischer Seite ist diese "Un-Robustheit" des Netzwerkes zurückzuführen auf die fehlende gleichgradige Stetigkeit des Netzwerkes, also auf unbeschränkte Variation des Schätzers. Auf diesen Punkt werden wir in 6.1 ausführlich eingehen. Experimentell wurde dieser Sachverhalt in Chandra & Singh (2003) für sigmoide Neuronale Netze bereits bestätigt.

Robustheit in Zusammenhang mit Neuronalen Netzen ist allerdings in zwei verschiedenen Zusammenhängen zu verstehen. Wir müssen unterscheiden zwischen Robustheit des Netzwerkes gegenüber Störungen der Eingabedaten

1) *während des Kontruktionsprozesses* (Training) und
2) Sensitivität des Netzwerk-Outputs im *laufenden Betrieb*.

Punkt 1) bezieht sich auf den verwendeten Risiko-Minimierungs-Algorithmus und Punkt 2) auf den verwendeten Hypothesenraum. Robuste Modellbildung mit Neuronalen Netzen bedeutet also zusammengefasst folgendes:

Nur wenn der Optimierungs-Algorithmus gut konditioniert (Punkt 1)) und die Konvergenz im Hypothesenraum gleichmäßig ist (Punkt 2)) können wir von einer sicheren Modellbildung sprechen.

Um die Kondition während des Trainingsprozesses zu kontrollieren wurden in der Vergangenheit verschiedene Strategien verfolgt. Besonders ist hier auf die numerische Untersuchung von Pohl & Prescher (s. Pohl (2007)) zu verweisen, die von dem Autor dieser Dissertation fachlich betreut wurde.
Autoren wie Minnix (1991), Matasuoka (1992), Holmstrom & Koistinen (1992) oder Murray & Edwards (1994) versuchen während Trainingsprozess die Inputs mit künstlichen Störungen zu kontaminieren. Chiu et al. (1994), Hammadi & Ito (1997) und Hammadi & Ito (1998) zeigen wie der Backpropagation-Algorithmus durch "smoothing"-Kriterien (Gewichte ausgeglichen verteilt) robustifiziert werden kann. In Kapitel 5 sind wir eingehend auf die verschiedenen Aspekte der Stabilität eines Lernalgorithmus eingegangen. Im folgenden beschäftigen wir uns nun mit Punkt 2) und schließen somit an Kapitel 4 an.

6.1 Run-Time-Robustheit & Kondition

Gerade in technischen Anwendungen ist es von entscheidender Bedeutung, dass das mühsam trainierte Neuronale Netz auch im laufenden Betrieb, also wenn es auf neue Daten angewendet werden soll, einen stabilen Output produziert. Und in Bezug auf die technische Umsetzung von Neuronalen Netzen, z.B. mit Hilfe von spezieller Hardware[1], ergibt sich noch eine weitere Frage: Was passiert, wenn ein Neuron ausfällt, nicht korrekt funktioniert oder ein Gewicht nahe oder gleich Null während des Trainigsprozesses zugeordnet bekommen hat? Resultiert hieraus ein schlecht konditionierter Approximator?
Chandra & Singh (2004) haben ähnliche Fragen für feedforward Neuronale Netze analysiert. Wir werden die in diesem Artikel gefundenen Ergebnisse erweitern allgemeine Neuronale Netze. Es stellt sich heraus, dass die Eigenschaft der gleichgradigen Stetigkeit Robustheit gegenüber Störungen der Eingabedaten und/oder der Gewichte während des laufenden Betriebes garantiert. Die gleichgradige Stetigkeit drückt genau das aus, was wir von einem stabilen Approximationsverfahren während der Laufzeit (also für neue Inputs) erwarten: Bei kleinen Störungen des Inputs ändern sich die Outputs (Funktionswerte) auch nur wenig.

[1] Hardware-Neuronale Netze werden meist als ANN - Artificial Neural Network - bezeichnet.

Aus dieser Forderung werden wir Nebenbedingungen für die Gewichte des Netzwerkes ableiten können, die wiederum in den Konstruktionsprozess integriert werden können.

6.1.1 Gleichgradig stetige Neuronale Netze

Die Existenz einer zu Ende des Kapitel 4 geforderten Teilfolge wird für einen kompakten Hypothesenraum durch den Satz von Arzelà-Ascoli sichergestellt. Zunächst eine Standard-Definition:

Definition 6.1 (Gleichgradige Stetigkeit). *Gegeben seien zwei metrischer Räume (X, d_X) und (Y, d_Y) und $F \subset C_b(X, Y)$ (Menge der beschränkten stetigen Funktionen $f : X \to Y$). Die Familie F heißt gleichgradig stetig im Punkt $x \in X$, wenn gilt:*

$$\forall \varepsilon > 0 \ \exists \delta > 0 \ : \ \forall f \in F \ : \ d_X(x, x') \leq \delta \Longrightarrow d_Y \left(f(x), f(x') \right) \leq \varepsilon \, .$$

Satz 6.1 (Arzelà-Ascoli). *Sei X ein kompakter metrischer Raum und Y ein Banachraum. Weiterhin sei $\mathscr{F} \subset C_b(X, Y)$ und $\mathscr{F}$ gleichgradig stetig und $Im(\mathscr{F})$ sei relativ kompakt in Y für alle $x \in X$. Dann ist $\mathscr{F}$ relativ kompakt in $C_b(X, Y)$.*

Beweis. Siehe jedes Analysis-Buch, z.B. Amann & Escher (2001).

Korollar 6.1. *Jede Folge aus $\mathscr{F}$ besitzt eine auf X gleichmäßig konvergente Teilfolge.*

Anmerkung 6.1. Man erinnere sich an Anmerkung 4.3 bezüglich der Beschränktheit von $\mathrm{Im}(\mathscr{F})$. Diese lässt sich im Rahmen der empirischen Risiko-Minimierung stets ohne Einschränkungen annehmen.

Korollar 6.2. *Unter den Voraussetzungen von Satz 6.1 sei eine Folge $(f_n)_{n \in \mathbb{N}} \subset \mathscr{F}$ von Funktionen $f_n : X \to Y$ punktweise konvergent gegen eine Grenzfunktion f auf X. Dann konvergiert die Folge auch gleichmäßig auf X.*

Beweis. Der Beweis funktioniert über ein Standard-$\varepsilon/3$-Argument, Details s. z.B. Dieudonné (1969).

Anmerkung 6.2. Insbesondere gelten die Korollare also für gleichmäßig beschränkte Folgen von differenzierbaren Funktionen mit gleichmäßig beschränkten Ableitungen.

Anmerkung 6.3. In stochastischem Rahmen sind punktweise (fast-sichere) Konvergenz und fast-gleichmäßige Konvergenz wie schon erwähnt äquivalent, wenn das Wahrscheinlichkeitsmaßes endlich ist (Satz von Egorov). Entscheidend für Punkt 2) unserer Forderungen ist also, dass die konstruierte Folge fast-sicher konvergiert.

Es liegt also nahe aus diesen Überlegungen heraus die Eigenschaft der gleichgradigen Stetigkeit für die Funktionenfamilie $\mathscr{F}$ zu fordern. Insbesondere wird sich herausstellen, dass derart konstruierte Neuronale Netze besonders robust gegenüber Störungen der Input-Daten sind.

Wir schließen zunächst an Abschnitt 2.2.1 an, wo die Klasse der affinen Kompositionen $\Sigma^d(\sigma)$ (Def. 2.2) der Form

$$g_M(\boldsymbol{x}) = \sum_{i=1}^{M} u_i \, \sigma\left(A_i(\boldsymbol{x})\right) \ , \quad \boldsymbol{x} \in D \subset \mathbb{R}^d \, ,$$

definiert werden, wobei $M \in \mathbb{N}$. $\Sigma^d(\sigma)$ stellt somit die Menge aller Netzwerke mit *linearem* Output dar. Wie zuvor ist $A_i : D \to \mathbb{R}$, $A_i(\boldsymbol{x}) = \boldsymbol{a}_i^T \boldsymbol{x} + t_i$, $a_{ij} \in \mathbb{R}$, $t_i \in \mathbb{R}$, $i \in \{1, ..., M\}$, $j \in \{1, ..., d\}$, ein affines Funktional auf D. In Satz 2.3 zeigen wir, dass diese Klasse für sigmoide Aktivierungsfunktionen prinzipiell in der Lage ist jede stetige Funktion beliebig genau zu approximieren (dieser Satz ist allerdings nicht-konstruktiv). In diesem Abschnitt gehen wir nun einen Schritt weiter und erweitern Σ^d auf Netzwerke mit *nicht-linearem* Output, also

$$\sigma\left(\Sigma^d(\sigma)\right) := \Bigg\{ \sigma\left(g_M\right) : D \to \mathbb{R} \ : \ \sigma\left(g_M\right) = \sigma\left(\sum_{i=1}^{M} u_i \sigma(A_i) \right) ; \ M \in \mathbb{N}, t_i \in \mathbb{R}, a_{ij} \in \mathbb{R}, \\ u_i \in \mathbb{R}, \ i \in \{1, ..., M\}, j \in \{1, ..., d\} \Bigg\} .$$

Weiterhin werden die folgenden Netzwerke mit Nebenbedingungen eingeführt:

$$\Sigma^d|_{B_1}(\sigma) := \Bigg\{ g_M : D \to \mathbb{R} \ : \ g_M = \sum_{i=1}^{M} u_i \sigma(A_i); M \in \mathbb{N}, t_i \in \mathbb{R}, a_{ij} \in \mathbb{R}, \\ \max\{|u_i| : 1 \le i \le M\} \le B, \ i \in \{1, ..., M\}, j \in \{1, ..., d\} \Bigg\} ,$$

$$\Sigma^d|_{B_2}(\sigma) := \Bigg\{ g_M : D \to \mathbb{R} \ : \ g_M = \sum_{i=1}^{M} u_i \sigma(A_i); \ M \in \mathbb{N}, t_i \in \mathbb{R}, \\ \max\{|a_{ij}|, |u_i| : 1 \le i \le M, 1 \le j \le d\} \le B, \ i \in \{1, ..., M\}, j \in \{1, ..., d\} \Bigg\} ,$$

$$\Sigma^d|_{B_3}(\sigma) := \Bigg\{ g_M : D \to \mathbb{R} \ : \ g_M = \sum_{i=1}^{M} u_i \sigma(A_i); \ M \in \mathbb{N}, t_i \in \mathbb{R}, \boldsymbol{a}_i \in S^d, \\ \max\{|u_i| : 1 \le i \le M\} \le B, \ i \in \{1, ..., M\}, j \in \{1, ..., d\} \Bigg\} ,$$

wobei $B \in (0, \infty]$. Hierbei bezeichnet S^d in B_4 die d-dimensionale Einheitskugel, also $\|\boldsymbol{a}_i\| \leq 1$ für alle $i \in \{1, ..., M\}$. Es ist klar, dass $\Sigma^d|_{B_{1,2}}(\sigma) = \Sigma^d(\sigma)$ für $B = \infty$. In Stinchcombe & White (1990) wird folgender Satz gezeigt:

Satz 6.2. *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma(x) = (1 + e^{-x})^{-1}$ als Aktivierungsfunktion gewählt. Dann liegt die Menge $\Sigma^d|_{B_2}(\sigma)$ dicht in $C(D)$ für jedes $B < \infty$, wobei $D \subset \mathbb{R}^d$ kompakt.*

Anmerkung 6.4. Dass $\Sigma^d|_{B_1}(\sigma)$ ebenfalls dicht in $C(D)$ für jedes $B < \infty$ liegt, folgt nun sofort.

Anmerkung 6.5. Stinchcombe & White (1990) zeigen diesen Satz etwas allgemeiner, und zwar für superanalytische Aktivierungsfunktionen. $\sigma \in C(\mathbb{R})$ heißt superanalytisch in $x \in \mathbb{R}$ mit Konvergenzradius $r > 0$, falls es eine unendliche Folge von reellen Zahlen $\{c_n\}$, $n \geq 0$ und $c_n \neq 0$ für $n \geq 1$ gibt, so dass für $|a - x| < r$ $\sum_{n=0}^{\infty} c_n(a - x)^n$ konvergiert und $\sigma(a) = \sum_{n=0}^{\infty} c_n(a - x)^n$. Funktionen wie Sinus, Kosinus und die logistische Funktion sind superanalytisch fast überall. Lemma 2.1 zeigt, dass die Klasse der diskriminatorischen Funktionen erwartungsgemäß viel größer ist also die der superanalytischen Funktionen. Der Approximationssatz von Weierstrass zeigt aber, dass "weniger" als syperanalytisch nicht ausreicht für Satz 6.2. Intuitiv ist dies auch einleuchtend, denn schränken wir gegenüber Satz 6.2 den Parameterraum ein, so benötigen wir für universelle Approximation in C im Gegenzug stärkere Einschränkungen für σ. Im allgemeinen sind Neuronale Netzwerke mit beschränktem Parameterraum somit auch größer als unbeschränkte Netze, wenn dieselbe Approximationsgenauigkeit erreicht werden soll.

Stinchcombe & White (1990) zeigen weiterhin:

Satz 6.3. *Es sei wieder die logistische Funktion als Aktivierungsfunktion gewählt. Dann liegt die Menge $\Sigma^d|_{B_3}(\sigma)$ dicht in $C(D)$ für jedes $B < \infty$, wobei $D \subset \mathbb{R}^d$ kompakt.*

Anmerkung 6.6. Dieser Satz ist allerdings *nicht* gültig für alle superanalytische Aktivierungsfunktionen. Damit $\Sigma^d|_{B_3}(\sigma)$ dicht liegt in C auf kompakten Teilmengen von $\mathbb{R}^d$, benötigen wir zusätzlich, dass $\text{span}\{\sigma^{(k)}|_{(-r,r)} : k \geq 0\} = C(\mathbb{R})|_{(-r,r)}$ für alle $B \geq 1$, wobei $\sigma^{(k)}$ die k-te Ableitung von σ und $f|_{(-r,r)}$, $f \in C(\mathbb{R}^d)$, die Einschränkung von f auf das Intervall $(-r, r)$ bezeichnet.

Anmerkung 6.7. Wie in Abschnitt 2.2.1 auch, können wir diese Ergebnisse auch allgemeiner formulieren für (Borel-) messbare statt stetige Funktionen. Der Beweis hierfür fußt wie zuvor auf dem Satz von Lusin.

Die Sätze 6.2 und 6.3 sind sehr interessant, weil sie die Wahl der Parameter aus einer *kompakten Menge* erlauben, in der Praxis natürlich ein unschätzbarer Vorteil. Das Auftreten der logistischen Funktion macht deutlich warum gerade diese Funktion große Beliebtheit in der Modellierung mit Neuronalen Netzen erfahren hat.

Die vorangegangenen Sätze sind erweiterbar auf Netzwerke mit nicht-linearem Output, allerdings muss die Aktivierungsfunktion nun stetig sein. Es gilt folgender Satz als Korollar zu den Sätzen in Abschnitt 2.2.1 [2]:

Korollar 6.3. *$\sigma(\Sigma^d(\sigma))$ liegt dicht in $C(K)$ mit $K \subset \mathbb{R}^d$ kompakt, falls σ eine stetige squashing Funktion ist (s. Abschnitt 2.2.1 für eine Definition). Somit liegen auch die Mengen $\sigma(\Sigma^d(\sigma))|_{B_1}$, $\sigma(\Sigma^d(\sigma))|_{B_2}$ und $\sigma(\Sigma^d(\sigma))|_{B_3}$ dicht in $C(K)$.*

Ausgehend von diesen Resultaten lässt sich nun zeigen, dass diese Aussagen auch für Multi-Layer-Neuronale Netze (s. Abschnitt 2.2.2) gelten, d.h.:

Korollar 6.4. *$\sigma(\Sigma\Pi^d(\sigma))$ liegt dicht in $C(K)$ mit $K \subset \mathbb{R}^d$ kompakt, falls σ eine stetige squashing- Funktion ist. Somit liegen auch die Mengen $\sigma(\Sigma\Pi^d(\sigma))|_{B_1}$, $\sigma(\Sigma\Pi^d(\sigma))|_{B_2}$ und $\sigma(\Sigma\Pi^d(\sigma))|_{B_3}$ dicht in $C(K)$.*

Beweis. Dieses und das vorangegangene Resultat ergeben sich direkt aus Satz 2.3 und Satz 2.7. Es ist leicht zu zeigen, dass $\sigma^{-1} \circ f$ für $f \in C(K)$ approximiert werden kann durch ein SNN im Sinne von Abschnitt 2.2.1 bzw. 2.2.2 (K ist kompakt und f stetig auf K, d.h. f ist auch beschränkt auf K). Die Stetigkeit von σ garantiert uns die Existenz von σ^{-1}. Diese Eigenschaft von σ macht dann aber auch deutlich, dass auch f selbst durch dasselbe Netzwerk approximiert werden kann. Die Resultate für die eingeschränkten Mengen zeigen sich mit Hilfe der Sätze 6.2 und 6.3. ■

Wir passen uns nun der Notation aus Abschnitt 4.2 an und bemerken zunächst, dass die Aussage "liegt dicht in" nichts anderes bedeutet, als dass für jede (stetige) Funktion f und jede Genauigkeit ε ein Neuronales Netz $\hat{f}$ (z.B. $\in \Sigma^d(\sigma)$) als Schätzer existiert, das f beliebig genau approximiert:

$$\begin{aligned} \hat{f} : \mathbb{R}^d \times W &\longrightarrow \mathbb{R} \\ (\boldsymbol{x}; w) &\longmapsto \sum_{i=1}^{M} u_i \, \sigma\left(\boldsymbol{a}_i^T \boldsymbol{x} + t_i\right) \ . \end{aligned}$$

Wie schon zuvor bemerkt, ist W somit ein $M(d+2)$-dimensionaler Raum. Die Frage ist nun, wie der Schätzer auf Störungen der Werte (a) im Input-Raum $\mathbb{R}^d$ und (b) im Gewichts-Raum W reagiert. Um einen stabilen Schätzer zu garantieren liegt es also nahe die Eigenschaft der *gleichgradigen Stetigkeit* für die Familie Σ^d (bzw. $\sigma(\Sigma^d)$) zu fordern. Da die Mitglieder dieser Familie Funktionen von $\mathbb{R}^d$ nach $\mathbb{R}$ sind, betrachten wir somit also gleichgradige Stetigkeit im Input-Raum. Etwas später werden wir Σ^d umdefinieren, so dass die enthaltenen Funktionen Argumente aus W akzeptieren und die Familie über $\boldsymbol{x} \in \mathbb{R}^d$ parametrisiert ist. Dies erlaubt uns gleichgradige Stetigkeit im Gewichts-Raum zu studieren.

[2] Castro, Mantas & Benítez (2000) haben schon ein schwächeres Resultat ohne Parametereinschränkungen gezeigt.

Netzwerke mit festem $M \in \mathbb{N}$
In der Praxis wird der Parameter M in der Regel zu Beginn des Modellierungsprozesses willkürlich festgelegt, durch z.B. statistische Verfahren geschätzt oder während des Trainingsprozesses so angepasst, dass Effekte wie Overparametrization o.Ä. vermieden werden. In jedem Fall ist er aber während der Laufzeit des Netzwerkes *nicht mehr veränderbar*. Konsequenterweise führen wir also unsere Beweise zunächst für Netzwerke mit *festem* $M \in \mathbb{N}$ durch:

$$\Sigma_M^d(\sigma) := \left\{ g_M : D \to \mathbb{R} \ : \ g_M = \sum_{i=1}^{M} u_i \sigma(A_i);\ t_i \in \mathbb{R}, a_{ij} \in \mathbb{R}, u_i \in \mathbb{R}, \right. \\ \left. i \in \{1, ..., M\}, j \in \{1, ..., d\} \right\}.$$

Es gilt nun folgender Satz:

Satz 6.4. *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare (Aktivierungs)-Funktion. Weiterhin sei $G \subset \mathbb{R}^d$ konvex und nicht-leer. Dann sind die Familien $\Sigma_M^d(\sigma)$ und $\sigma(\Sigma_M^d(\sigma))$ für alle $M \in \mathbb{N}$ nicht gleichgradig stetig in allen Punkten $\boldsymbol{x} \in G$.*

Beweis. Chandra & Singh (2004) geben einen eingeschränkten Satz für den Spezialfall der logistischen Funktion. Zudem ist die in dem Artikel erhaltenen Formel nicht korrekt, die in diesem Artikel präsentierten Ergebnisse stimmen aber zumindest qualitativ mit den Formeln aus dieser Dissertation überein.
Es seien $\boldsymbol{x}, \boldsymbol{y} \in G$, $\boldsymbol{x} \neq \boldsymbol{y}$, wobei die Verbindungsstrecke $\overline{\boldsymbol{xy}}$ von $\boldsymbol{x}$ und $\boldsymbol{y}$ somit wiederum in G liegt (G ist konvex). Nach dem Mittelwertsatz der Differentialrechnung existiert somit ein $\boldsymbol{\xi} \in \overline{\boldsymbol{xy}}$, $\boldsymbol{\xi} \notin \{\boldsymbol{x}, \boldsymbol{y}\}$, mit

$$\begin{aligned} |g_M(\boldsymbol{x}) - g_M(\boldsymbol{y})| &= \left| \sum_{i=1}^{M} u_i \nabla\sigma\left(A_i\right)(\boldsymbol{\xi}) \right| |\boldsymbol{x} - \boldsymbol{y}| = \left| \sum_{i=1}^{M} u_i \nabla\sigma\left(A_i\right)(\boldsymbol{\xi}) \right| d(\boldsymbol{x}, \boldsymbol{y}) \\ &= \left| \sum_{i=1}^{M} u_i \sigma'\left(A_i(\boldsymbol{\xi})\right) \boldsymbol{a}_i \right| d(\boldsymbol{x}, \boldsymbol{y}) \,. \end{aligned}$$

σ ist nach Voraussetzung überall stetig differenzierbar, d.h. Lipschitz-stetig auf jedem Kompaktum $K \subset \mathbb{R}$ mit der Lipschitz-Konstanten L, d.h. σ' ist inbesondere beschränkt auf jedem $K \subset \mathbb{R}$. Wir definieren $L := \max_{i \in \{1,...,M\}} L_i$, wobei L_i die Lipschitz-Konstante von σ auf dem zu A_i gehörenden kompakten Intervall K_i bezeichnet[3]:

[3] Topologisch könnte man so argumentieren, dass die Funktion stetig differenzierbar ist und deshalb lokal Lipschitz-stetig auf ganz $\mathbb{R}$. Jedes Kompaktum kann durch endlich viele offene Mengen überdeckt werden, und auf all diesen Mengen hat die Funktion somit eine Lipschitz-Konstante, und man setze L_i gleich dem Maximum dieser endlich vielen Konstanten.

$$K_i := \left\{ \lambda x' + (1-\lambda)y' + t_i \ : \ x' = \boldsymbol{a}_i^T \boldsymbol{x}, y' = \boldsymbol{a}_i^T \boldsymbol{y}, \lambda \in \mathbb{R}, 0 \leq \lambda \leq 1 \right\} ,$$

also:

$$|g_M(\boldsymbol{x}) - g_M(\boldsymbol{y})| \leq L(g_M) \left| \sum_{i=1}^{M} u_i \boldsymbol{a}_i \right| d(\boldsymbol{x}, \boldsymbol{y}) .$$

Fordern wir nun gleichgradige Stetigkeit für die Familie $\Sigma_M^d(\sigma)$, so muss also für $\varepsilon > 0$

$$\delta = \frac{\varepsilon}{L(g_M) \left| \sum_{i=1}^{M} u_i \boldsymbol{a}_i \right|} \tag{6.1}$$

gewählt werden. Wir schreiben $L(g_M)$, weil die maximale Lipschitzkonstante von dem speziellen g_M abhängt. Wäre σ sogar gleichmäßig Lipschitz-stetig, gäbe es eine maximale Lipschitz-Konstante auf ganz $\mathbb{R}$, so fiele dies weg. Ein Beispiel für diesen Sachverhalt ist die logistische Funktion (s. nachfolgendes Beispiel). Der Punkt ist aber, dass dieses δ von u_i und $\boldsymbol{a}_i$ abhängt, die im Falle von $\Sigma_M^d(\sigma)$ aber beliebig sind. Weiterhin wurden $\boldsymbol{x}$ und $\boldsymbol{y}$ beliebig in G gewählt. Somit ist $\Sigma_M^d(\sigma)$ nicht gleichgradig stetig auf G.
Wir verfahren ebenso für $\sigma(\Sigma_M^d(\sigma))$. Nach dem Mittelwertsatz gilt ganz analog:

$$\begin{aligned} |\sigma(g_M(\boldsymbol{x})) - \sigma(g_M(\boldsymbol{y}))| &= \left| \sigma' \left(\sum_{i=1}^{M} u_i \sigma(A_i(\boldsymbol{\xi})) \right) \sum_{i=1}^{M} u_i \sigma'(A_i(\boldsymbol{\xi})) \boldsymbol{a}_i \right| d(\boldsymbol{x}, \boldsymbol{y}) \\ &\leq L^2(g_M) \left| \sum_{i=1}^{M} u_i \boldsymbol{a}_i \right| d(\boldsymbol{x}, \boldsymbol{y}) , \end{aligned}$$

also

$$\delta = \frac{\varepsilon}{L^2(g_M) \left| \sum_{i=1}^{M} u_i \boldsymbol{a}_i \right|} \tag{6.2}$$

und mit derselben Argumentation wie zuvor ist $\sigma(\Sigma_M^d(\sigma))$ somit ebenfalls nicht gleichgradig stetig auf G. ■

Beispiel 6.1. Nehmen wir als Beispiel die für Neuronale Netze häufig verwendete logistische Funktion $\sigma : \mathbb{R} \to \mathbb{R}$,

$$\sigma(x) = \frac{1}{1 + e^{-x}} .$$

Sie ist überall differenzierbar. Um zu untersuchen, ob sie auch Lipschitz-stetig ist betrachten wir:

$$\left| \frac{\sigma(x_1) - \sigma(x_2)}{x_1 - x_2} \right| = \frac{1}{2} \left| \tanh\left(\frac{x_1}{2}\right) - \tanh\left(\frac{x_2}{2}\right) \right| \Big/ |x_1 - x_2| .$$

Dieser Ausdruck wird maximal für $x_1 \to x_2$:

$$\lim_{x_1 \to x_2} \left| \frac{\sigma(x_1) - \sigma(x_2)}{x_1 - x_2} \right| = \frac{1}{2\,[1 + \cosh(x_2)]}$$

und nimmt wiederum ein *globales* Maximum von $1/4$ bei $x_2 = 0$ an. Somit können wir als globale Lipschitz-Konstante $L = 1/4$ wählen und erhalten

$$\delta = \frac{4\varepsilon}{\left| \sum_{i=1}^{M} u_i \boldsymbol{a}_i \right|} .$$

Beispiel 6.2. Als zweites Beispiel betrachten wir die Funktion tanh, die zweite Standard-Aktivierungsfunktion in der Theorie der Neuronalen Netze. Es ist

$$\tanh'(x) = \frac{1}{\cosh^2(x)} .$$

Diese Funktion hat ein globales Maximum von 1 bei $x = 0$ und ansonsten Singularitäten oder Unstetigkeitsstellen. Somit wählen wir $L = 1$, d.h.

$$\delta = \frac{\varepsilon}{\left| \sum_{i=1}^{M} u_i \boldsymbol{a}_i \right|} .$$

Anmerkung 6.8. Für Neuronale Netze benutzt man für σ im allgemeinen *squashing Funktionen*, also man fordert (1) $\lim_{x \to \infty} \sigma(x) = 1$ und $\lim_{x \to -\infty} \sigma(x) = 0$ und (2) $\sigma(x)$ nicht-abnehmend (bzw. monoton). Wir nennen eine squashing Funktion, bei der diese Konvergenz gleichmäßig ist eine *gleichmäßige squashing Funktion*. Wir können für diese Funktionenklasse folgendes Lemma formulieren:

Lemma 6.1. *Sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare gleichmäßige squashing Funktion. Dann ist σ global (gleichmäßig) Lipschitz-stetig.*

Beweis. Durch die gleichmäßige Konvergenz ist die Vertauschung von Grenzwert und Ableitung gesichert. Da σ zudem monoton und überall differenzierbar ist, ist σ' beschränkt und das Maximum liegt in einem Kompaktum um den Nullpunkt. Dieses Maximum kann dann als Lipschitz-Konstante gewählt werden. ■

Dieses Lemma zeigt, dass für solche Aktivierungsfunktionen jedes $g_M \in \Sigma_M^d(\sigma)$ eine globale Lipschitz-Konstante L besitzt. Beispiele für gleichmäßige squashing Funktionen sind tanh und die logistische Funktion.

Folgendes Korollar zeigt, dass auch die eingeschränkten Familien $\Sigma_M^d|_{B_1}(\sigma)$ und $\sigma(\Sigma_M^d|_{B_1}(\sigma))$ nicht gleichgradig stetig auf G sind:

Korollar 6.5. *Mit denselben Voraussetzungen wie in Satz 6.4 sind die Familien $\Sigma_M^d|_{B_1}(\sigma)$ und $\sigma(\Sigma_M^d|_{B_1}(\sigma))$ für alle $M \in \mathbb{N}$ nicht gleichgradig stetig in allen Punkten $\boldsymbol{x} \in G$.*

Beweis. Es gilt nun $u_i \leq B$ für $i \in \{1, ..., M\}$, d.h. wir setzen in Glg. (6.1) bzw. (6.2) ein und erhalten

$$\delta = \frac{\varepsilon}{L(g_M)B\left|\sum_{i=1}^{M} \boldsymbol{a}_i\right|}$$

bzw.

$$\delta = \frac{\varepsilon}{L^2(g_M)B\left|\sum_{i=1}^{M} \boldsymbol{a}_i\right|} .$$

Wie zuvor können wir nun argumentieren, dass $\delta = \delta(g_M)$. ■

Man sieht allerdings sofort, dass wir für die Einschränkung B_2 bzw. B_3 allerdings gleichgradige Stetigkeit erhalten:

Korollar 6.6. *Mit denselben Voraussetzungen wie in Satz 6.4 sind die Familien $\Sigma_M^d|_{B_2}(\sigma)$, $\sigma(\Sigma_M^d|_{B_2}(\sigma))$, $\Sigma_M^d|_{B_3}(\sigma)$ und $\sigma(\Sigma_M^d|_{B_3}(\sigma))$ für alle $M \in \mathbb{N}$ gleichgradig stetig in allen Punkten $\boldsymbol{x} \in G$, falls σ zusätzlich noch global Lipschitz-stetig bzw. spezieller eine gleichmäßige squashing Funktion ist.*

Beweis. Wir setzen wieder in Glgn. (6.1) und (6.2) ein und erhalten für B_2:

$$\delta = \frac{\varepsilon}{L(g_M)M\sqrt{d}B^2}$$

bzw.

$$\delta = \frac{\varepsilon}{L^2(g_M)M\sqrt{d}B^2} .$$

Ist σ global Lipschitz-stetig (bzw. spezieller eine gleichmäßige squashing Funktion, siehe Lemma 6.1) fällt die g_M-Abhängigkeit von L weg, so dass wir ein eindeutiges δ auf G für die gesamte Funktionenfamilie und damit gleichgradige Stetigkeit erhalten. Dieselben Argumente können auf $\Sigma_M^d|_{B_3}(\sigma)$ und $\sigma(\Sigma_M^d|_{B_3}(\sigma))$ angewendet werden, so dass im ersten Fall

$$\delta = \frac{\varepsilon}{L(g_M)MB}$$

bzw.

$$\delta = \frac{\varepsilon}{L^2(g_M)MB}$$

für $\sigma(\Sigma_M^d|_{B_3}(\sigma))$. ■

Als nächstes möchten wir die Robustheit der verschiedenen Σ_M^d in Hinblick auf Störungen im Gewichtsraum behandeln. Hierfür betrachten wir diese Familien als Funktionen in W, also:

$$\Sigma_M^d(\sigma) = \Big\{ g_M : W \to \mathbb{R} \ : \ g_M(w) = \sum_{i=1}^{M} u_i \sigma\left(\boldsymbol{a}_i^T \boldsymbol{x} + t_i\right) ; \ \boldsymbol{x} \in \mathbb{R}^d, \\ w = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, i \in \{1, ..., M\}\} \Big\} .$$

So ausgerüstet zeigen wir nun den zu 6.4 analogen Satz:

Satz 6.5. *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare und beschränkte (Aktivierungs)-Funktion. Weiterhin sei $K \subset W \subset \mathbb{R}^{M(d+2)}$ konvex und nicht-leer. Dann sind die Familien $\Sigma_M^d(\sigma)$ und $\sigma(\Sigma_M^d(\sigma))$ für alle $M \in \mathbb{N}$ nicht gleichgradig stetig in allen Punkten $w \in K$.*

Beweis. Der Beweis führt sich wie zuvor. Betrachte $w_1, w_2 \in K$, $w_1 := \{(u_i^1, \boldsymbol{a}_i^1, t_i^1), i \in \{1, ..., M\}\}$, $w_2 := \{(u_i^2, \boldsymbol{a}_i^2, t_i^2), i \in \{1, ..., M\}\}$, und einen Punkt $\xi = \{(\upsilon_i, \boldsymbol{\alpha}_i, \tau_i), i \in \{1, ..., M\}\} \in K$ auf der Verbindungsstrecke der beiden, der, weil K konvex ist, existiert. Dann gilt nach dem Mittelwertsatz

$$|g_M(w_1) - g_M(w_2)| = |\nabla_w g_M(\xi)| \ d_W(w_1, w_2) ,$$

wobei d_W eine geeignete, hier zunächst nicht näher spezifizierte Metrik im Raum der Gewichte sei. Weiterhin ist der Gradient bezüglich eines $w \in W$ gegeben durch

$$\nabla_w = \left(\frac{\partial}{\partial u_1}, ..., \frac{\partial}{\partial u_M}, \frac{\partial}{\partial a_{11}}, ..., \frac{\partial}{\partial a_{1d}}, ..., \frac{\partial}{\partial a_{M1}}, ..., \frac{\partial}{\partial a_{Md}}, \frac{\partial}{\partial t_1}, ..., \frac{\partial}{\partial t_M} \right)^T .$$

Wegen

$$\begin{aligned} \frac{\partial g_M}{\partial u_i}(\xi) &= \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right) , \\ \frac{\partial g_M}{\partial a_{ij}}(\xi) &= \upsilon_i \sigma'\left(\boldsymbol{\alpha}_i^T \boldsymbol{x} + \tau_i\right) x_j , \\ \frac{\partial g_M}{\partial t_i}(\xi) &= \upsilon_i \sigma'\left(\boldsymbol{\alpha}_i^T \boldsymbol{x} + \tau_i\right) , \end{aligned}$$

$i \in \{1, ..., M\}$, $j \in \{1, ..., d\}$, gilt somit

$$|g_M(w_1) - g_M(w_2)| = \Bigg[\sum_{i=1}^{M} \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right)^2 + \sum_{i=1}^{M} \sum_{j=1}^{d} \left(\upsilon_i \sigma'\left(\boldsymbol{\alpha}_i^T \boldsymbol{x} + \tau_i\right) x_j\right)^2 \\ + \sum_{i=1}^{M} \left(\upsilon_i \sigma'\left(\boldsymbol{\alpha}_i^T \boldsymbol{x} + \tau_i\right)\right)^2 \Bigg]^{1/2} d_W(w_1, w_2) .$$

Nach Voraussetzung ist σ beschränkt (sagen wir $\sigma(x) \leq A$ für alle $x \in \mathbb{R}$) und wegen der stetigen Differenzierbarkeit σ' ebenfalls. Wir verfahren wie zuvor und nehmen für diese Schranke die maximale Lipschitz-Konstante $L := \max L_i$ über alle Kompakta, mit denen

$$K_i := \left\{ \lambda x' + (1-\lambda) y' \,:\, x' = (\boldsymbol{a}_i^1)^T \boldsymbol{x} + t_i^1, y' = (\boldsymbol{a}_i^2)^T \boldsymbol{x} + t_i^2, \lambda \in \mathbb{R}, 0 \leq \lambda \leq 1 \right\} ,$$

überdeckt werden kann. Wie zuvor hängt dieses maximale L von der Funktion g_M ab, d.h. hier von dem speziellen $\boldsymbol{x}$. Falls σ global Lipschitz-stetig wäre, so gäbe es ein L, so dass die Lipschitz-Bedingung für alle $\boldsymbol{x} \in \mathbb{R}$, also alle $g_M \in \Sigma_M^d(\sigma)$ erfüllt wäre. Es ist also:

$$|g_M(w_1) - g_M(w_2)| \leq \left[MA^2 + L^2(g_M) \sum_{i=1}^{M} \upsilon_i^2 \sum_{j=1}^{d} x_j^2 + \sum_{i=1}^{M} \upsilon_i^2 \right]^{1/2} d_W(w_1, w_2) ,$$

so dass

$$\delta = \frac{\varepsilon}{\sqrt{MA^2 + L^2(g_M)\,|\boldsymbol{\upsilon}|\,(|\boldsymbol{x}|+1)}}$$

gewählt werden muss. Dieser Ausdruck hängt, wie zuvor, von dem speziell betrachteten g_M und implizit sogar von w_1, w_2 ab, so dass $\Sigma_M^d(\sigma)$ nicht gleichgradig stetig ist. In derselben Weise ergibt sich für $\sigma(\Sigma_M^d(\sigma))$:

$$\delta = \frac{\varepsilon}{L(g_M)\sqrt{MA^2 + L^2(g_M)\,|\boldsymbol{\upsilon}|\,(|\boldsymbol{x}|+1)}}$$

mit derselben Konsequenz. ■

Anmerkung 6.9. Die Beschränktheit von σ kann, da immer nur endliche Summen über σ auftauchen, auch fallengelassen werden. In diesem Fall wäre in der Endformel A^2 zu ersetzen durch $\max_{i \in \{1,\ldots,M\}} \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right)^2$.

Wir sehen an dieser Formel, das für gleichgradige Stetigkeit zwei Dinge gegeben sein müssen: Einerseits muss $|\boldsymbol{\upsilon}|$ beschränkt sein, was z.B. durch eine Beschränkung von $|\boldsymbol{u}|$ erreicht werden kann (denn $\boldsymbol{\upsilon}$ liegt auf der Verbindungsstrecke zwischen $\boldsymbol{u}^1$ und $\boldsymbol{u}^2$). Andererseits muss aber auch $|\boldsymbol{x}|$ beschränkt sein. Betrachtet man $\boldsymbol{x}$ auf einem Kompaktum $\subset \mathbb{R}^d$, so ist dies gegeben. In der Praxis könnte man so argumentieren, dass die Familie $\Sigma_M^d(\sigma)$ an sich immer diskreter Natur ist, d.h. die Familie umfasst nicht alle $\boldsymbol{x} \in \mathbb{R}^d$, sondern nur die N Trainingsdaten, also $\boldsymbol{x} \in \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$. Zusammengefasst formulieren wir folgendes Korollar:

Korollar 6.7. *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare (Aktivierungs)-Funktion. Weiterhin sei $K \subset W \subset \mathbb{R}^{M(d+2)}$ konvex und nicht-leer. Dann sind die Familien*

$$\Sigma_M^d|_{B_1}(\sigma) := \Bigg\{ g_M : K \to \mathbb{R} \,:\, g_M(w) = \sum_{i=1}^{M} u_i \sigma\left(\boldsymbol{a}_i^T \boldsymbol{x} + t_i\right);\ \boldsymbol{x} \in D \subset \mathbb{R}^d \text{ kompakt},$$
$$w = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, i \in \{1, ..., M\}, \max\{|u_i| : 1 \leq i \leq M\} \leq B\} \Bigg\},$$

$$\Sigma_M^d|_{B_2}(\sigma) := \Bigg\{ g_M : K \to \mathbb{R} \,:\, g_M(w) = \sum_{i=1}^{M} u_i \sigma\left(\boldsymbol{a}_i^T \boldsymbol{x} + t_i\right);\ \boldsymbol{x} \in D \subset \mathbb{R}^d \text{ kompakt},$$
$$w = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, i \in \{1, ..., M\},$$
$$\max\{|a_{ij}|, |u_i| : 1 \leq i \leq M, 1 \leq j \leq d\} \leq B\} \Bigg\},$$

$$\Sigma_M^d|_{B_3}(\sigma) := \Bigg\{ g_M : K \to \mathbb{R} \,:\, g_M(w) = \sum_{i=1}^{M} u_i \sigma\left(\boldsymbol{a}_i^T \boldsymbol{x} + t_i\right);\ \boldsymbol{x} \in D \subset \mathbb{R}^d \text{ kompakt},$$
$$w = \{(u_i, \boldsymbol{a}_i, t_i) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}, \boldsymbol{a}_i \in S^d, i \in \{1, ..., M\},$$
$$\max\{|u_i| : 1 \leq i \leq M\} \leq B\} \Bigg\}$$

und die entsprechenden $\sigma(\Sigma_M^d|_{B_{1,2,3}}(\sigma))$ für alle $M \in \mathbb{N}$ gleichgradig stetig in allen Punkten $w \in K$, falls σ zusätzlich noch beschränkt und global Lipschitz-stetig bzw. spezieller eine gleichmäßige squashing Funktion ist.

Beweis. Der Beweis dieses Korollars ist im Prinzip schon geschehen, wir notieren zunächst, dass für B_1, B_2 und B_3

$$\delta = \frac{\varepsilon}{\sqrt{MA^2 + L^2 B(T+1)}},$$

wobei $T := \max_D |\boldsymbol{x}|$ und L wieder die globale Lipschitz-Konstante von σ ist. Für eine gleichmäßige squashing Funktion ist zudem $A = 1$. Für $\sigma(\Sigma_M^d|_{B_{1,2,3}}(\sigma))$ ist dann entsprechend

$$\delta = \frac{\varepsilon}{L\sqrt{MA^2 + L^2 B(T+1)}}.$$

Somit sind diese Familien alle gleichgradig stetig in K. ■

Netzwerke mit beliebigem M

Im vorangegangenen Abschnitt haben wir Netzwerke mit festem M untersucht. Die "Dichtheits"-Aussagen für SNNs beziehen sich aber immer auf Netzwerke mit beliebigem M, d.h. es stellt sich die Frage in wie fern $\Sigma^d(\sigma)$ eine gleichgradig stetige Familie ist. Wir notieren zunächst folgendes Lemma:

Lemma 6.2. *Zwischen den bisher betrachteten Familien gelten folgende Beziehungen:*

$$\Sigma_M^d(\sigma) \subset \Sigma^d(\sigma)\,,$$
$$\Sigma_M^d(\sigma)|_{B_{1,2,3}} \subset \Sigma^d(\sigma)|_{B_{1,2,3}}$$

und ebenso für die entsprechenden $\sigma(\Sigma_M^d(\sigma))$*. Weiterhin ist*

$$\Sigma^d(\sigma) = \lim_{M\to\infty} \Sigma_M^d(\sigma)\,,$$
$$\Sigma^d(\sigma)|_{B_{1,2,3}} = \lim_{M\to\infty} \Sigma_M^d(\sigma)|_{B_{1,2,3}}$$

sowie

$$\Sigma^d(\sigma)|_{B_3} \subsetneqq \Sigma^d(\sigma)|_{B_2} \subsetneqq \Sigma^d(\sigma)|_{B_1} \subsetneqq \Sigma^d(\sigma)$$

und ebenso für die entsprechenden $\sigma(\Sigma_M^d(\sigma))$*.*

Beweis. Die Aussagen folgen direkt aus den Definitionen. Für $\Sigma^d(\sigma) = \lim_{M\to\infty} \Sigma_M^d(\sigma)$ reicht es zu bemerken, dass in $\lim_{M\to\infty} \Sigma_M^d(\sigma)$ auch alle endlichen Summen enthalten sind, es müssen nur ab dem gewünschten M alle weiteren Koeffizienten gleich Null gesetzt werden. ■

In Hinblick auf die Eigenschaft der gleichgradigen Stetigkeit im Input-Raum gilt nun folgender Satz:

Satz 6.6. *Es sei* $\sigma : \mathbb{R} \to \mathbb{R}$ *eine stetig differenzierbare (Aktivierungs)-Funktion. Weiterhin sei* $G \subset \mathbb{R}^d$ *konvex und nicht-leer. Dann sind die Familien* $\Sigma^d(\sigma)$*,* $\Sigma^d(\sigma)|_{B_1}$*,* $\Sigma^d(\sigma)|_{B_2}$*,* $\Sigma^d(\sigma)|_{B_3}$ *und die entsprechenden* $\sigma(\Sigma^d|_{B_{1,2,3}}(\sigma))$ *nicht gleichgradig stetig in allen Punkten* $\boldsymbol{x} \in G$*.*

Beweis. Für $\Sigma^d(\sigma)$ erinnern wir an die Formeln (6.1) und (6.2) für δ:

$$\delta = \frac{\varepsilon}{L(g_M)\left|\sum_{i=1}^M u_i \boldsymbol{a}_i\right|}$$

bzw.

$$\delta = \frac{\varepsilon}{L^2(g_M)\left|\sum_{i=1}^M u_i \boldsymbol{a}_i\right|}\,.$$

Selbst wenn σ global Lipschitz-stetig ist (also L nicht von g_M abhängt), so tritt in der Formel M als obere Summationsgrenze auf, d.h. δ hängt wiederum i.A. von dem speziellen g_M ab. Ebenso verhält es sich mit $\Sigma^d(\sigma)|_{B_1}$, $\Sigma^d(\sigma)|_{B_2}$ und $\Sigma^d(\sigma)|_{B_3}$ sowie den entsprechenden $\Sigma^d(\sigma)|_{B_{1,2,3}}$, hierzu siehe die Korollare 6.5 und 6.6. ■

Ebenso verhält es sich mit gleichgradiger Stetigkeit im Gewichts-Raum:

Satz 6.7. *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare und beschränkte (Aktivierungs)-Funktion. Weiterhin sei $K \subset W \subset \mathbb{R}^{M(d+2)}$ konvex und nicht-leer. Dann sind die Familien $\Sigma^d(\sigma)$, $\Sigma^d(\sigma)|_{B_1}$, $\Sigma^d(\sigma)|_{B_2}$, $\Sigma^d(\sigma)|_{B_3}$ und die entsprechenden $\sigma(\Sigma^d|_{B_{1,2,3}}(\sigma))$ nicht gleichgradig stetig in allen Punkten $w \in K$.*

Beweis. Wir greifen auf die Formeln für δ aus Satz 6.5 und Korollar 6.7 zurück. δ hängt in beiden Fällen direkt von M ab, was zeigt, dass die entsprechenden Familien nicht gleichgradig stetig im Raum der Gewichte sind. ■

Das Problem ist also die M-Abhängigkeit von δ. Heuristisch argumentiert lässt sich diese nur durch eine Beschränkung von $|\boldsymbol{u}|$ unabhängig von M verhindern, wenn $\boldsymbol{a}_i$ gleichzeitig auf der Einheitskugel liegt. Wir erinnern daran, dass die Approximationsaufgabe darin besteht für jedes $f \in C(D)$, $D \subset \mathbb{R}^d$ kompakt, ein g_M zu finden, dass

$$\|f(\boldsymbol{x}) - g_M(\boldsymbol{x})\|_\infty \leq \varepsilon \,,$$

$\boldsymbol{x} \in D$, also

$$\left\| \sum_{i=1}^{M} u_i \sigma\left(A_i(\boldsymbol{x})\right) \right\|_\infty \leq \|f(\boldsymbol{x})\|_\infty + \varepsilon \,.$$

Ist σ beschränkt mit $\sigma(x) \leq A$ für alle $x \in \mathbb{R}$, und o.B.d.A. $A = 1$, so reduziert sich dies auf

$$\left| \sum_{i=1}^{M} u_i \right| \leq \|f(\boldsymbol{x})\|_\infty + \varepsilon \,.$$

Diese Ungleichung müssen wir bei der Beschränkung von $|\boldsymbol{u}|$ also beachten wollen wir die universelle Approximationsfähigkeit (also Dichtheit im Raum der zu approximierenden Funktionen) der gesamten Familie nicht verlieren. Wir erinnern uns nun zurück an Korollar 6.3, in dem die Dichtheit von $\sigma(\Sigma^d(\sigma))$ und den Einschränkungen $\sigma(\Sigma^d|_{B_{1,2,3}}(\sigma))$ in $C(D)$ gezeigt wird für stetige squashing Funktionen und alle $B < \infty$. Nach dem Satz vom Maximum/Minumum nimmt f auf D eben selbiges an, d.h. wir können die Approximationsaufgabe o.B.d.A. auf die Approximation von Funktionen mit $\|f(\boldsymbol{x})\|_\infty \leq \mathcal{F} < \infty$ zurückführen. D.h. es ergibt sich sofort die Bedingung

$$\left| \sum_{i=1}^{M} u_i \right| \leq \mathcal{F} + \varepsilon \,.$$

Zusammengefasst scheint also die folgende Menge ein hoffnungsvoller Kandidat für eine gleichgradige Familie im Input- und Gewichtsraum zu sein:

$$\Sigma^d|_{B_4}(\sigma) := \Big\{ g_M : D \to \mathbb{R} \,:\, g_M = \sum_{i=1}^{M} u_i \sigma(A_i);\; M \in \mathbb{N}, t_i \in \mathbb{R}, \boldsymbol{a}_i \in S^d, \\ |\boldsymbol{u}| \leq \mathcal{F} + \varepsilon,\; i \in \{1, ..., M\} \Big\} \,, \quad (6.3)$$

wobei ε die Approximationsgenaugkeit bezeichnet. Wir formulieren auf Basis der vorangegangenen Überlegungen folgenden Satz:

Satz 6.8. *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare (Aktivierungs)-Funktion, die außerdem beschränkt und global Lipschitz-stetig bzw. spezieller eine gleichmäßige squashing Funktion ist. Weiterhin seien $G \subset \mathbb{R}^d$ und $K \subset W \subset \mathbb{R}^{M(d+2)}$ beliebig, aber konvex und nicht-leer. Dann sind die Familien $\Sigma_M^d|_{B_4}(\sigma)$ und $\sigma(\Sigma_M^d|_{B_4}(\sigma))$ gleichgradig stetig für alle $M \in \mathbb{N}$ in allen Punkten $\boldsymbol{x} \in G$ sowie $w \in K$. Weiterhin sind die Familien $\Sigma^d|_{B_4}(\sigma)$ und $\sigma(\Sigma^d|_{B_4}(\sigma))$ ebenfalls gleichgradig stetig im Raum der Inputs und liegen dicht in $C(D)$ für jedes $B < \infty$ und jede kompakte Menge $D \subset \mathbb{R}^d$. $\Sigma^d|_{B_4}(\sigma)$ und $\sigma(\Sigma^d|_{B_4}(\sigma))$ (definiert analog zu den Familien in Korollar 6.7) sind allerdings nicht gleichgradig stetig im Raum der Gewichte.*

Beweis. Wir kehren zurück zu Glg. (6.1):

$$\delta = \frac{\varepsilon_1}{L(g_M)\left|\sum_{i=1}^M u_i \boldsymbol{a}_i\right|}$$

und sehen sofort, dass mit $|\boldsymbol{a}_i| = 1$ für alle $i \in \{1, ..., M\}$, $|\boldsymbol{u}| \leq \mathcal{F} + \varepsilon$ und der globalen Lipschitz-Stetigkeit von σ

$$\delta \geq \frac{\varepsilon_1}{L\sqrt{\sum_{i=1}^M u_i^2 |\boldsymbol{a}_i|^2}} = \frac{\varepsilon_1}{L\,|\boldsymbol{u}|} \geq \frac{\varepsilon_1}{L(\mathcal{F}+\varepsilon)}\,,$$

wobei ε_1 die für die gleichgradige Stetigkeit gewählte Konstante ist (wir können z.B. $\varepsilon_1 = \varepsilon$ wählen). Somit ist $\Sigma_M^d|_{B_4}(\sigma)$ gleichgradig stetig für alle $\boldsymbol{x} \in G$, und zwar unabhängig von M, so dass auch die gesamte Familie $\Sigma^d|_{B_4}(\sigma)$ auf G gleichgradig stetig ist. Dieselben Argumente können für $\sigma(\Sigma_M^d|_{B_4}(\sigma))$ und $\sigma(\Sigma^d|_{B_4}(\sigma))$ angewendet werden mit

$$\delta \geq \frac{\varepsilon_1}{L^2(\mathcal{F}+\varepsilon)}\,.$$

Wegen

$$|\boldsymbol{u}| \leq |f(\boldsymbol{x})| + \varepsilon \Longrightarrow \left|\sum_{i=1}^M u_i\right| \leq |f(\boldsymbol{x})| + \varepsilon$$

für alle $\boldsymbol{x} \in D$ und der Beschränktheit von σ folgt dann sofort aus Korollar 6.3 (dieses ist anwendbar aufgrund der im Vergleich sogar noch verschärften Voraussetzungen an σ), dass $\Sigma^d|_{B_4}(\sigma)$ dicht in $C(D)$ liegt.
Wir benutzen für den Gewichtsraum die Formel für δ aus dem Beweis von Satz 6.5:

$$\delta = \frac{\varepsilon}{\sqrt{\sum_{i=1}^M \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right)^2 + L^2(g_M)\,|\boldsymbol{v}|\,(|\boldsymbol{x}|+1)}}$$

und reduzieren sie gemäß den Voraussetzungen (man beachte, dass der Input-Vektor $\boldsymbol{x}$ gemäß den Definitionen in Korollar 6.7 auf ein Kompaktum $\subset \mathbb{R}^d$ eingeschränkt ist, und $T := \max_D |\boldsymbol{x}|$ sowie $|\boldsymbol{u}| \leq (\mathcal{F} + \varepsilon) \Rightarrow |\boldsymbol{v}| \leq (\mathcal{F} + \varepsilon)$) auf

$$\delta \geq \frac{\varepsilon_1}{\sqrt{\sum_{i=1}^{M} \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right)^2 + L^2(\mathcal{F} + \varepsilon)\,(T+1)}}\,,$$

bzw.

$$\delta \geq \frac{\varepsilon_1}{L\sqrt{\sum_{i=1}^{M} \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right)^2 + L^2(1 + \varepsilon)\,(T+1)}}\,,$$

für $\sigma(\Sigma^d|_{B_4}(\sigma))$. Allerdings hängt der Term $\sum_{i=1}^{M} \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right)^2$ von M ab, so dass weder $\Sigma^d|_{B_4}(\sigma)$ noch $\sigma(\Sigma^d|_{B_4}(\sigma))$ gleichgradig stetig im Raum der Gewichte sind. ■

Anmerkung 6.10. Dass $\Sigma^d|_{B_4}(\sigma)$ bzw. $\sigma(\Sigma^d|_{B_4}(\sigma))$ nicht gleichgradig stetig im Raum der Gewichte sind liegt nur an der M-Abhängigkeit von $\sum_{i=1}^{M} \sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right)^2$. Es ist allerdings (s. Beweis von Satz 6.5)

$$\sigma\left(\boldsymbol{\alpha}_i^T \boldsymbol{x}\right) = \frac{\partial g_M}{\partial u_i}(\xi)\,,$$

so dass dieser Term (und die gesamte Summe) verschwinden, falls alle u_i, $i \in \{1, ..., M\}$, als Konstanten betrachtet werden. In diesem Fall wären $\Sigma^d|_{B_4}(\sigma)$ bzw. $\sigma(\Sigma^d|_{B_4}(\sigma))$ auch gleichgradig stetig im Raum der Gewichte. Ein Netzwerk dieser Form hätte also konstante bzw. nicht-fehlerbehaftete Koeffizienten in der Linearkombination der hidden-nodes.

Robustheit und Kondition von NN für Input und Gewichte

Nachdem nun die Fragen nach der gleichgradigen Stetigkeit von Neuronalen Netzen ausreichend geklärt sind stellt sich die Frage, ob wir durch diese Forderung eine gewisse Fehlertoleranz des Netzwerkes erreichen können. Wir fassen noch einmal kurz zusammen in welcher Situation wir uns befinden. Durch den Trainingsprozess wurde aufgrund von N Messungen das Optimierungsproblem

$$\hat{w} = \underset{w \in W}{\operatorname{argmin}}\, \Lambda\left(\Xi, \hat{d}, w\right)$$

gelöst, wobei $\hat{f} \in \Sigma^d(\sigma)$ gewählt wurde. Das Ergebnis ist somit der Schätzer $\hat{f} : \mathbb{R}^d \to \mathbb{R}$

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{M} \hat{u}_i\, \sigma\left(\hat{\boldsymbol{a}}_i^T \boldsymbol{x} + \hat{t}_i\right)\,.$$

Der "Meta-Parameter" M wird im Vorfeld geschätzt (z.B. durch ein Informationskriterium) bzw. auch noch während des Trainingsprozesses angepasst. Dieser Schätzer soll nun auf neue Daten angewendet werden, sagen wir auf eine Folge von unabhängigen identisch

verteilten Zufallsvariablen $(\widetilde{X}_i)$, $X_i : \Omega_1 \to \mathbb{R}^d$, jeweils mit Erwartungswert μ und Varianz σ^2. In Abschnitt 4.2 (s. Glgn. (4.4) und (4.5)) wurde schon argumentiert, dass die neuen Inputs dieselbe Verteilung wie die Trainings-Daten haben müssen, um eine systematische sub-Optimalität des Schätzers zu vermeiden. Wir setzen also voraus, dass z.B. kein pre-processing der Daten vorgenommen wird, welches die beteiligten Verteilungen verändern könnte. Wir modellieren diese Situation ganz analog zu der in Abschnitt 4.1 und führen eine Zufallsvariable E ein mit

$$E_X = \widetilde{X} - X \,. \tag{6.4}$$

Wir nehmen wie zuvor an, dass $\mathbb{E}[E_X] = 0$. Wir berechnen nun wie stark $\hat{f}(\widetilde{X})$ "im Mittel der Quadrate" von $\hat{f}(X)$ abweicht (MSE):

$$\mathbb{E}\left[\left(\hat{f}(\widetilde{X}) - \hat{f}(X)\right)^2\right] \leq L^2(\hat{f}) \left|\sum_{i=1}^{M} \hat{u}_i \hat{\boldsymbol{a}}_i\right|^2 \mathbb{E}\left[\left(\widetilde{X} - X\right)^2\right] = L^2(\hat{f}) \left|\sum_{i=1}^{M} \hat{u}_i \hat{\boldsymbol{a}}_i\right|^2 \mathbb{E}\left[E_X^2\right] \,,$$

wobei wir Glgn. (6.1) und (6.4) verwendet haben und L wieder das Maximum der lokalen Lipschitzkonstanten sei. Ist E_X z.B. $N(0, \sigma^2)$-verteilt, so ergibt sich sofort

$$\mathbb{E}\left[E_X^2\right] = \mathbb{V}\left[E_X\right] = \sigma^2 \,.$$

Etwas anders aufgeschrieben zeigt sich sofort der Zusammenhang zur numerischen Kondition des Problems:

$$\left(\hat{f}(\widetilde{X}) - \hat{f}(X)\right)^2 (\omega) = \left|\hat{f}(\widetilde{\boldsymbol{x}}) - \hat{f}(\boldsymbol{x})\right|^2 \leq L^2(\hat{f}) \left|\sum_{i=1}^{M} \hat{u}_i \hat{\boldsymbol{a}}_i\right|^2 |\widetilde{\boldsymbol{x}} - \boldsymbol{x}|^2 \,,$$

mit $\omega \in \Omega_1$, es ist also

$$\kappa_{\text{abs}}(\hat{f}) := \left|L(\hat{f})\right| \left|\sum_{i=1}^{M} \hat{u}_i \hat{\boldsymbol{a}}_i\right|$$

die absolute Kondition des Problems (der Funktion $\hat{f}$). Die relative Kondition des Schätzers kann analog folgendermaßen definiert werden:

$$\kappa_{\text{rel}}(\hat{f}; \boldsymbol{x}) := \left|\frac{\sum_{i=1}^{M} \hat{u}_i \nabla\sigma\left(\hat{A}_i\right)(\boldsymbol{x})}{\sum_{i=1}^{M} \hat{u}_i \sigma\left(\hat{A}_i(\boldsymbol{x})\right)}\right| |\boldsymbol{x}| \leq \left|L(\hat{f})\right| \left|\sum_{i=1}^{M} \hat{u}_i \hat{\boldsymbol{a}}_i\right| \left|\frac{\boldsymbol{x}}{\sum_{i=1}^{M} \hat{u}_i \sigma\left(\hat{A}_i(\boldsymbol{x})\right)}\right| \,.$$

Wir können hiervon ausgehend nun die maximale Kondition einer Netzwerkfamilie Σ^d definieren:

$$\kappa_{\text{abs}}^{\max}\left(\Sigma^d(\sigma)\right) := \sup_{\hat{f} \in \Sigma^d(\sigma)} \kappa_{\text{abs}}(\hat{f})$$

und ebenso für κ_{rel}. Ganz analog können wir auch eine Kondition im Gewichtsraum definieren:

$$\kappa_{\text{abs},W}(\hat{f}) := \sqrt{\sum_{i=1}^{M} \sigma\left(\hat{\boldsymbol{a}}_i^T \boldsymbol{x}\right)^2 + L^2(g_M)\,|\hat{\boldsymbol{u}}|\,(|\boldsymbol{x}|+1)}\,.$$

Die Verbindung zu der Eigenschaft der gleichgradigen Stetigkeit der Netzwerkfamilie ist direkt ersichtlich:

Satz 6.9. *Ist die Netzwerkfamilie gleichgradig stetig (im Input- oder Gewichtsraum), so ist die Kondition jeder enthaltenen Funktion nach oben beschränkt. Die Gegenrichtung gilt ebenso.*

Wir sehen sofort, dass κ zum einen direkt mit dem maximalen Betrag der Ableitung von σ zusammenhängt. Dies ist nicht überraschend. Andererseits hängt κ aber auch mit M zusammen, und zumindest im Raum der Gewichte steigt die Kondition mit steigendem M.
Als Korollar zu den Sätzen über die im vorangegangenen Abschnitt definierten Netzwerkfamilien ergibt erkennen wir zunächst, dass die Familien $\Sigma^d(\sigma)$, $\sigma(\Sigma^d(\sigma))$ und die entsprechenden Einschränkungen $B_{1,2,3}$ keine beschränkte Kondition sowohl im Input- als auch im Gewichtsraum besitzen. Anders verhält es sich für die Familien, die sich als gleichgradig stetig herausgestellt haben:

Satz 6.10. *Es sei $\sigma : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare (Aktivierungs)-Funktion, die außerdem beschränkt und global Lipschitz-stetig bzw. spezieller eine gleichmäßige squashing Funktion ist. Weiterhin seien $G \subset \mathbb{R}^d$ und $K \subset W \subset \mathbb{R}^{M(d+2)}$ beliebig, aber konvex und nicht-leer. Dann gelten die folgenden oberen Schranken für die Kondition im Input-Raum:*

$$\begin{aligned}
\kappa_{abs}\left(\Sigma_M^d|_{B_2}(\sigma)\right) &\leq LM\sqrt{d}B^2\,,\\
\kappa_{abs}\left(\Sigma_M^d|_{B_3}(\sigma)\right) &\leq LMB\,,\\
\kappa_{abs}\left(\Sigma_M^d|_{B_4}(\sigma)\right) &\leq L(\mathcal{F}+\varepsilon)\,,
\end{aligned}$$

wobei $\varepsilon > 0$ die gewünschte maximal mögliche Approximationsgenauigkeit des Netzwerkes bezeichne. Weiterhin gilt somit ebenfalls

$$\kappa_{abs}\left(\Sigma^d|_{B_4}(\sigma)\right) \leq L(\mathcal{F}+\varepsilon)\,.$$

Im Raum der Gewichte gelten die folgenden Schranken:

$$\begin{aligned}
\kappa_{abs,W}\left(\Sigma_M^d|_{B_{1,2,3}}(\sigma)\right) &\leq \sqrt{MA^2 + L^2B(T+1)}\,,\\
\kappa_{abs,W}\left(\Sigma_M^d|_{B_4}(\sigma)\right) &\leq \sqrt{MA^2 + L^2(\mathcal{F}+\varepsilon)(T+1)}\,,
\end{aligned}$$

wobei $\sigma(x) \leq A < \infty$ für alle $x \in \mathbb{R}$ und $T = \max_D |\boldsymbol{x}|$. Alle Schranken müssen mit $1/L$ multipliziert werden für die entsprechenden $\sigma(\cdot)$-Versionen.

Für alle anderen Familien lässt sich keine obere Schranke für die Kondition der Mitglieder unabhängig von den spezifischen Netzwerk-Parametern angeben.

Eine wichtige Konsequenz aus diesem Satz ist, dass $\Sigma^d|_{B_4}(\sigma)$ der einzige nicht-M-abhängige Hypothesenraum mit beschränkter Kondition ist, zumindest im Raum der Inputs. Weiterhin hängt die Kondition dieser Familie auch für festes M *nicht* von diesem Parameter ab. Dies ist ein ganz entscheidender Punkt: In Abschnitt 4.5 haben wir einen Satz von Barron (1993) rezitiert, der zeigt, dass die Approximationsgenauigkeit des Neuronalen Netzes proportional zu $M^{-1/2}$ ist, d.h. verkürzt ausgedrückt je größer M, desto "besser" ist das Netzwerk. Wird aber die Kondition als *zusätzliches Optimierungsziel* während des Konstruktionsprozesses eingebracht, so hängt diese für alle Familien *außer* B_4 so von M ab, dass beide Optimierungsziele konkurrieren. Die *Mehrzieloptimierung* für B_4 reduziert sich somit wiederum zu einem einfach-Optimierungsproblem.

6.2 Optimale Neuronale Netze

In diesem Abschnitt wenden wir die theoretischen Ergebnisse aus Kapitel 4 auf die Kandidatenmenge $\Sigma^d|_{B_4}(\sigma)$ an und betrachten ihre (Risiko-)Konvergenzeigenschaften für $N \to \infty$ und $M \to \infty$. Zunächst einmal gilt es zu bemerken, dass keine der hier beschriebenen eingeschränkten Mengen *konvex* ist. Sie sind aber alle kompakt. Wir kehren somit zurück zu Satz 4.9, müssen aber die Voraussetzungen ein wenig anpassen. Wie in Anmerkung 4.3 argumentiert können wir immer $|Y| \leq L < \infty$ fast-sicher annehmen. Wir setzen also voraus, dass

$$\big|f(X) - Y\big| \leq \mathcal{F} + L$$

fast überall, wenn $\|f\|_\infty \leq M$. Damit ist im Falle von $\Xi = (f(X) - Y)^2$ auch Ξ beschränkt fast-überall. In Abschnitt 4.5 wurde argumentiert, dass wir eine Teilfolge von $(\hat{f}_{N,M})$ konstruieren müssen, so dass diese gleichmäßig konvergiert. Wir schlagen daher auf Grundlage von $\Sigma^d|_{B_4}(\sigma)$ folgenden Hypothesenraum vor:

$$\mathscr{F}_N^\varepsilon := \Sigma^d_{\mathcal{M}(N)}(\sigma) := \Big\{ g : D \to \mathbb{R} \;:\; g = \sum_{i=1}^{\mathcal{M}(N)} u_i\, \sigma(A_i);\; \mathcal{M}(N) \in \mathbb{N}, t_i \in \mathbb{R}, \\ \boldsymbol{a}_i \in S^d,\; \sum_{i=1}^{\mathcal{M}(N)} |u_i| \leq \mathcal{F}(N) + \varepsilon,\; i \in \{1, ..., \mathcal{M}(N)\} \Big\}$$

mit $D \subset \mathbb{R}^d$. $\mathcal{M}(N)$ und $\mathcal{F}(N)$ sind jeweils Funktionen von $N \in \{1, 2, ...\}$ mit $\mathcal{M}, \mathcal{F} \to \infty$ (streng) monoton für $N \to \infty$. Weiterhin gilt offensichtlich

$$\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}\,,$$

wobei $\mathscr{F} = \Sigma^d|_{\boldsymbol{a}_i \in S^d \cup \{0\}}(\sigma)$. Da nach Korollar 6.3 $\Sigma^d|_{B_3}(\sigma)$ dicht in $C(D)$ liegt, tut dies also auch $\mathscr{F}$.

Lemma 6.3. *$\mathscr{F} = \bigcup_{N=1}^{\infty} \mathscr{F}_N^{\varepsilon}$ liegt dicht in $L^p(D,\mu)$ für ein beliebiges Maß.*

Beweis. Dieses Lemma folgt mit den Ergebnissen aus Abschnitt 2.2.1, insbesondere Korollar 2.1 und der Beobachtung, dass die Gewichte $\boldsymbol{u}$ nicht eingeschränkt sind, weil $\mathscr{F} = \Sigma^d|_{\boldsymbol{a}_i \in S^d}(\sigma)$. ■

Anmerkung 6.11. Man beachte, dass mit den Ergebnissen aus 2.2.3 sogar Dichtheit in $L^p(\mathbb{R}^d,\mu)$ garantiert ist, falls σ eine Kernfunktion ist, die die Voraussetzungen von Satz 2.10 erfüllt.

Wir schreiben suggestiv N als Index, denn wir meinen in der Tat die Anzahl der Datenpunkte. Die Idee hinter dieser Menge ist, dass je mehr Datenpunkte zur Verfügung stehen, desto komplexer sollte das Netzwerk werden. Den genauen Funktionszusammenhang werden wir etwas später diskutieren. Der Raum $\mathscr{F}$ ist nicht mehr gleichgradig stetig, aber jeder Unterraum $\mathscr{F}_N$, $N = 1, 2, ...$

Es ist klar, dass die Einschränkung $\sum_{i=1}^{\mathcal{M}(N)} |u_i| \leq \mathcal{F}(N) + \varepsilon$ sowohl impliziert, dass $|\boldsymbol{u}| \leq \mathcal{F}(N) + \varepsilon$, als auch, dass $\|g\|_\infty \leq \mathcal{F}(N) + \varepsilon$, $g \in \mathscr{F}_N^{\varepsilon}$, weil $\sigma(x) \leq 1$ für alle $x \in \mathbb{R}$. Wir haben schon in Abschnitt 4.5 darauf hingewiesen, dass die alleinige Beschränkung von $\|g\|_\infty$ keinen Einfluss auf die VC-Dimension des Hypothesenraumes hat. $\mathscr{F}_N^{\varepsilon}$ wurde nun allerdings so konstruiert, dass mit steigendem N auch $\mathcal{M}$ steigt, so dass sich in der Tat eine Ordnung der Modellstrukturen anhand ihrer Komplexität[4] ergibt:

$$h_1 \leq h_2 \leq \cdots \leq h_{\mathcal{M}} \leq \cdots$$

ergibt.

Es ist somit möglich das in Abschnitt 4.3 eingeführte SRM-Prinzip auf $\mathscr{F}^{\varepsilon}$ anzuwenden.

Doch auch für $\mathscr{F}^{\varepsilon}$ sind sowohl die VC- als auch die fat-shattering Dimension nur sehr ungenau bekannt, was diese Methode lediglich aus theoretischer Sicht interessant macht.

Sampling-Fehler

Für jeden N-Wert wird also ein Element $\hat{f}_{N,M(N)}$ in $\mathscr{F}_N^{\varepsilon}$ gefunden, welches das empirische Risiko

$$\widetilde{A}^N(f, \mathscr{F}_N^{\varepsilon}) = \frac{1}{N} \sum_{i=1}^{N} \left| f(X_i) - Y_i \right|^2,$$

$f \in \mathscr{F}_N^{\varepsilon}$, minimiert. Wir zeigen nun, wie das empirische Risiko abgeschätzt werden kann:

[4] Die Ordnung auf $\mathscr{F}^{\varepsilon}$ könnte alternativ auch über einen anderen Komplexitätsbegriff geschehen, die so genannte fat-shattering Dimension (s. Bartlett (1998)). Diese ist nicht unabhängig von $\|g\|_\infty$ und steigt in unserem Fall mit N und $\mathcal{M}$, so dass es keinen Unterschied macht welcher Dimensionsbegriff der Ordnung zu Grunde liegt.

Satz 6.11. *Es sei $\mathscr{F}_N^\varepsilon$ wie soeben definiert mit $\mathcal{M}(N), \mathcal{F}(N) \to \infty$ streng monoton. Dann gilt, falls $|Y| < L$ fast-überall und N groß genug, dass $\mathcal{F}(N) > L$:*

$$\mathbb{P}^N \left[\sup_{f \in \mathscr{F}_{\mathcal{M}(N)}^\varepsilon} \left| \widetilde{\Lambda}^N(Z,f) - \Lambda(f) \right| > \alpha \right] < \left(\frac{128e(\mathcal{M}+1)(\mathcal{F}+\varepsilon)^2}{\alpha} \right)^{\mathcal{M}(2d+3)+1} \cdot 2\exp\left(-\frac{3N}{8(\mathcal{F}+\varepsilon)^4} \frac{\alpha^2}{8\alpha+3} \right) .$$

Beweis. Wie schon erwähnt können wir o.B.d.A. $|Y| < L$ fast-überall annehmen (Anmerkung 4.3), so dass diese Voraussetzung eigentlich sogar weggelassen werden könnte. Wie erwähnt ist der Hypothesenraum nicht konvex, d.h. wir wenden Satz 4.9 an. Aus der Definition von $\mathscr{F}_N^\varepsilon$ folgt, dass $\|f\|_\infty \leq \mathcal{F}+\varepsilon$ für alle $f \in \mathscr{F}_N^\varepsilon$, so dass $|f(X)-Y| \leq \mathcal{F}+\varepsilon+L$. Wir ersetzen also in diesem Satz c_1 durch $(2\mathcal{F}+2L)^{-1}$ (man überzeuge sich hiervon durch eine kurze Rechnung für $\Xi = (Y-f)^2$), schätzen σ^2 durch $(\mathcal{F}+\varepsilon+L)^2$ ab und nutzen $\mathcal{F}(N) > L$:

$$\mathbb{P}^N \left[\sup_{f \in \mathscr{F}_{\mathcal{M}(N)}^\varepsilon} \left| \widetilde{\Lambda}^N(Z,f) - \Lambda(f) \right| > \alpha \right] < 2\,\ell_{\mathscr{F}_{\mathcal{M}(N)}^\varepsilon} \left(\frac{\alpha}{8(\mathcal{F}+\varepsilon+L)} \right) \cdot \exp\left(-\frac{3N}{8(\mathcal{F}+\varepsilon)^4} \frac{\alpha^2}{8\alpha+3} \right) . \tag{6.5}$$

$\mathcal{F}(N)$ ist streng monoton steigend, d.h. es gibt ein N, so dass $\mathcal{F}(N) > L$, so dass wir (6.5) weiter nach oben abschätzen (man beachte: $a < b \Rightarrow \ell(b) < \ell(a)$):

$$\mathbb{P}^N \left[\sup_{f \in \mathscr{F}_{\mathcal{M}(N)}^\varepsilon} \left| \widetilde{\Lambda}^N(Z,f) - \Lambda(f) \right| > \alpha \right] < 2\,\ell_{\mathscr{F}_{\mathcal{M}(N)}^\varepsilon} \left(\frac{\alpha}{32(\mathcal{F}+\varepsilon)} \right) \cdot \exp\left(-\frac{3N}{8(\mathcal{F}+\varepsilon)^4} \frac{\alpha^2}{8\alpha+3} \right) . \tag{6.6}$$

Nun geht es wieder um die Abschätzung der (durch die Kompaktheit des Hypothesenraumes) endlichen Überdeckungszahl. Hier taucht nun wieder die Komplexität des Hypothesenraumes in Form seiner VC-Dimension auf. Es gilt mit Haussler (1992):

$$\ell_{\mathscr{F}_{\mathcal{M}(N)}^\varepsilon} \left(\frac{\alpha}{32(\mathcal{F}+\varepsilon)} \right) \leq 2 \left(\frac{64e\mathcal{F}^2(1+\varepsilon)^2}{\alpha} \ln \frac{64e(\mathcal{F}+\varepsilon)^2}{\alpha} \right)^h . \tag{6.7}$$

Die folgende Argumentationskette geht im Prinzip auf Lugosi & Zeger (1995) zurück. Man definiere die folgenden drei Mengen:

$$G_1 = \left\{ \boldsymbol{a}^T \boldsymbol{x} + b; \boldsymbol{a} \in \mathbb{R}, b \in \mathbb{R} \right\}$$
$$G_2 = \left\{ \sigma\left(\boldsymbol{a}^T \boldsymbol{x} + b \right); \boldsymbol{a} \in \mathbb{R}, b \in \mathbb{R} \right\}$$
$$G_3 = \left\{ c\sigma\left(\boldsymbol{a}^T \boldsymbol{x} + b \right); \boldsymbol{a} \in \mathbb{R}, b \in \mathbb{R}, c \in [-(\mathcal{F}(N)+\varepsilon), (\mathcal{F}(N)+\varepsilon)] \right\} .$$

Nach Cover (1965) ist

$$\ell_{G_1}(\alpha) = d + 1$$

und weil G_2 ein Verkettungen von Funktionen beinhaltet, ist $\ell_{G_2}(\alpha) \leq d + 1$ (s. Nolan & Pollard (1987)). Mit (6.7) erhalten wir somit:

$$\ell_{G_2}(\alpha) \leq 2\left(\frac{2e}{\alpha}\right)^{2(d+1)} ,$$

was zusammen mit der Beobachtung (Pollard (1990)), dass sich die Überdeckungszahlen bei Mengen, die aus dem Produkt aus Funktionen aus zwei verschiedenen Funktionenräumen bestehen, berechnet als das Produkt der Einzel-Überdeckungszahlen, folgendes ergibt für $\mathcal{F} > 2/e$:

$$\ell_{\mathscr{F}^{\varepsilon}_{\mathcal{M}(N)}}\left(\frac{\alpha}{32\mathcal{F}(1+\varepsilon)}\right) \leq \left(\frac{128e(\mathcal{M}(N)+1)(\mathcal{F}(N)+\varepsilon)^2}{\alpha}\right)^{\mathcal{M}(N)(2d+3)+1} .$$

In (6.6) eingesetzt folgt die Behauptung. ■

Nach Satz 6.11 gilt also mit Wahrscheinlichkeit $1-\delta$ für den Sampling-Fehler α:

$$\frac{3}{8}\frac{N}{(\mathcal{F}+\varepsilon)^4}\frac{\alpha^2}{8\alpha+3} - (\mathcal{M}(2d+3)+1)\ln\left(\frac{128e(\mathcal{M}+1)(\mathcal{F}+\varepsilon)^2}{\alpha}\right) - \ln 2 + \ln\delta \leq 0 . \quad (6.8)$$

Wir lassen die N-Abhängigkeit von $\mathcal{M}$ zunächst weg. Setzen wir die linke Seite gleich 0 und lösen nach α, so erhalten wir den minimalen Sampling-Fehler (man vergleiche mit Glg. (4.20) und der anschließenden Analyse). Im Unterschied zu (4.20) ist aber nun im wesentlichen

$$\alpha = \alpha(\mathcal{M}, \mathcal{F}, N) ,$$

d.h. der Sampling-Fehler hängt von der Wahl der Funktion $\mathcal{M}$ und von $\mathcal{F}$ ab.

Gleichung (6.8) lautet umgeschrieben:

$$c_1\frac{\alpha^2}{8\alpha+3} = -(c_2\mathcal{M}+1)\ln\left(\frac{\alpha}{c_3(\mathcal{M}+1)}\right) + c_4 \quad (6.9)$$

mit positiven Konstanten $c_1, ..., c_4$ (wobei für c_4 angenommen wird, dass $0 \leq \delta \leq 2$), die nur von $\mathcal{F}(N)$ und N abhängen. Diese Gleichung hat offensichtlich eine eindeutige positive Lösung. Wir analysieren (6.8) nun in Hinblick auf die Abhängigkeit von $\mathcal{M}$. Umgeschrieben erhalten wir:

$$\alpha = c_3(\mathcal{M}+1)\,\exp\left(-\frac{c_1\frac{\alpha^2}{8\alpha+3} - c_4}{c_2\mathcal{M}+1}\right)$$

Für große α wird $\alpha^2/(8\alpha+3) \approx \alpha/8$, so dass

$$\alpha \approx c_3(\mathcal{M}+1)\,\exp\left(-\frac{c_1\frac{\alpha}{8}-c_4}{c_2\mathcal{M}+1}\right)\,.$$

Mit $\bar{\alpha}=\alpha/(\mathcal{M}+1)$ erhalten wir die Lösung

$$\bar{\alpha} = \frac{8c_2}{c_1} L\left[\frac{c_1c_3}{8c_2}\exp\left(\frac{c_4}{c_2(\mathcal{M}+1)}\right)\right]\,, \tag{6.10}$$

wobei $L[z]$ die Lösung der Differentialgleichung

$$\frac{dx}{dz} = \frac{x}{z(x+1)}$$

bezeichnet (auch Lambert-Funktion genannt), die sich numerisch leicht und mit beliebiger Genauigkeit bestimmen lässt. (6.10) zeigt, dass

$$\bar{\alpha} \xrightarrow{\mathcal{M}\to\infty} \frac{8c_2}{c_1} L\left[\frac{c_1c_3}{8c_2}\right] =: \alpha_\infty(\mathcal{F},N)\,.$$

α_∞ hat somit ein Konvergenzverhalten wie

$$\alpha_\infty \propto \frac{\mathcal{F}^4}{N} L\left[\frac{N}{\mathcal{F}^2}\right] \longrightarrow 0\,,$$

weil $L[z\to\infty]\to\infty$ und $\mathcal{F}^4$ nach Voraussetzung langsamer steigt als N. Abb. (6.1) zeigt den Verlauf von α_∞ für ein Beispiel. Es ist wichtig auf die Bedeutung von α_∞ hinzuweisen. Diese Größe stellt die Steigung des Sampling-Fehlers dar, der entsteht, wenn eine endliche Anzahl Daten N zur Verfügung steht und ein Neuronales Netz mit unendlich vielen Nodes und der Nebenbedingung

$$\sum_{i=1}^{\infty}|u_i| \le \mathcal{F}(N)+\varepsilon$$

verwendet wird. Je größer N, desto "konstanter" bleibt der Sampling-Fehler mit steigendem $\mathcal{M}$, d.h. desto weniger variiert α auf einem $\mathcal{M}$-Intervall. Außer für $N\to\infty$ verstärkt also eine Erhöhung von $\mathcal{M}$ den durch die endliche Anzahl von Trainingsdaten enstandenen Fehler. Abb. 6.2 zeigt $\bar{\alpha}$ als Funktion von $\mathcal{M}$ mit der klar zu erkennenden Asymptote α_∞.
Das folgende Korollar führt Satz 6.11 fort und zeigt unter welchen Bedingungen eine Erhöhung der Trainings-Menge auch zu einer besseren Approximation des Netzwerkes führt:

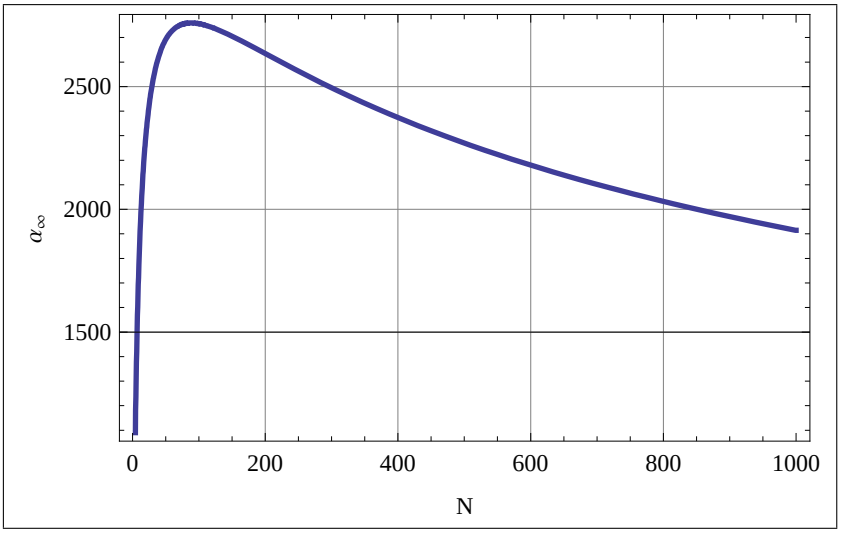


Abb. 6.1. α_∞ für die Wahl $\mathcal{F}(N) = 1 + \ln N$. Die übrigen Parameter sind dieselben wie in den vorangegangenen Abbildungen.

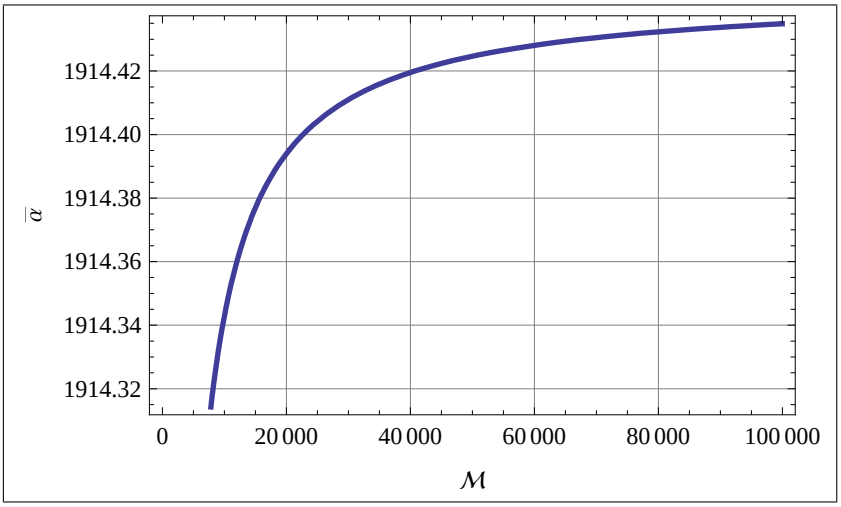


Abb. 6.2. $\bar{\alpha}$ für $N = 1000$ und $\mathcal{F}(N) = 1+\ln N$. Die Asymptote ist in diesem Fall $\alpha_\infty = 1914.45$. Die weiteren Parameter sind dieselben wie in den vorangegangenen Abbildungen.

Korollar 6.8. *Die Wahrscheinlichkeit in Satz 6.11 konvergiert gegen Null, falls zusätzlich zu* $\mathcal{M}(N), \mathcal{F}(N) \to \infty$

$$\mathcal{K} := \frac{\mathcal{M}(N)\mathcal{F}^4(N) \ln\left[\mathcal{M}(N)\mathcal{F}^2(N)\right]}{N} \longrightarrow 0\,.$$

Die Konvergenz ist sogar fast-sicher, falls ein $\eta > 0$ *existiert, so dass*

$$\mathcal{F}^4(N) N^{\eta-1} \longrightarrow 0 \quad .$$

Beweis. Wegen

$$\left(\frac{128e(\mathcal{M}+1)(\mathcal{F}+\varepsilon)^2}{\alpha}\right)^{\mathcal{M}(2d+3)+1} \exp\left(-\frac{3N}{8(\mathcal{F}+\varepsilon)^4}\frac{\alpha^2}{8\alpha+3}\right)$$
$$= \exp\left[(\mathcal{M}(2d+3)+1)\ln\left(\frac{128e(\mathcal{M}+1)(\mathcal{F}+\varepsilon)^2}{\alpha}\right) - \frac{3N}{8(\mathcal{F}+\varepsilon)^4}\frac{\alpha^2}{8\alpha+3}\right]$$

erhalten wir die erste Behauptung, denn die rechte Seite kann nur unter der ersten Bedingungen gegen 0 konvergieren. Der Beweis, dass die Konvergenz mit der Zusatzbedingung auch fast-sicher ist benutzt das Lemma von Borel-Cantelli (s. Lugosi & Zeger (1995)). Und zwar lässt sich

$$\sum_{N=1}^{\infty} \exp\left[(\mathcal{M}(2d+3)+1)\ln\left(\frac{128e(\mathcal{M}+1)(\mathcal{F}+\varepsilon)^2}{\alpha}\right) - \frac{3N}{8(\mathcal{F}+\varepsilon)^4}\frac{\alpha^2}{8\alpha+3}\right] < \infty$$

nur garantieren, falls

$$\int_1^{\infty} \exp\left(-C\frac{N}{\mathcal{F}^2}\right)$$

existiert (Integralkriterium). Falls es kein $\eta > 0$ gibt, was die Forderung erfüllt, so wird die Konvergenz der Reihe beliebig langsam. ■

Abb. 6.3 zeigt den Verlauf von α für das Beispiel $\mathcal{M}(N) = \mathcal{F}(N) = 1 + \ln N$, das offensichtlich die Bedingungen von Korollar 6.8 erfüllt. In der Zeichnung wurde α mit $4\mathcal{F}^2$ normiert.

6.2.1 Minimierung des Gesamtfehlers

Nun schätzen wir den Fehler ab, der durch die Einschränkung des Hypothesenraumes auf $\mathscr{F}_N^{\varepsilon}$ entsteht. In Satz 5.10 hatten wir gezeigt wie sich eine Folge in M konstruieren lässt, so dass

$$\|\hat{f}_M - f\|^2 \leq \frac{4\mathcal{F}^2}{\sqrt{M}} = \alpha_{\text{bias}}\,.$$

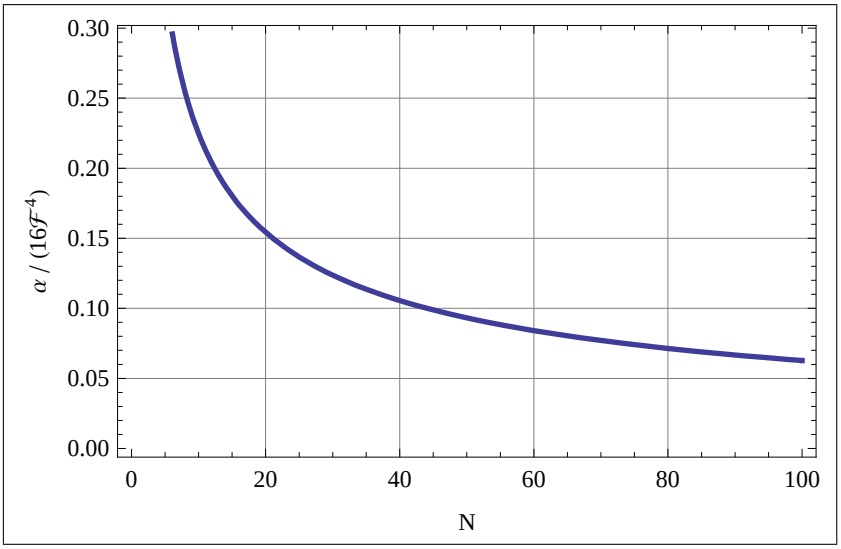


Abb. 6.3. Der Sampling-Fehler (normiert) für das Beispiel: $\delta = 0.01$, $\varepsilon = 0.01$, $d = 3$ und $\mathcal{M}(N) = \mathcal{F}(N) = 1 + \ln N$.

Der Algorithmus sucht in den Räumen $\mathscr{F}_N^\varepsilon$ für $N = 1, 2, ...$ entsprechend der Vorgabe ein fast-Minimum auf Grundlage des in Schritt $N-1$ gefundenen fast-Minimums. Der Gesamtfehler ist nach oben beschränkt durch die Summe von Bias- und Sampling-Fehler. Geben wir nun eine Funktion $\mathcal{F}(N)$ vor, z.B. wie oben $1 + \ln N$, und wählen ein N fest, so gibt es nur noch einen "Freiheitsgrad" in der Abschätzung des Gesamtfehlers, und zwar der Funktionszusammenhang $\mathcal{M}(N)$. Wir schlagen dementsprechend folgendes Vorgehen vor: Für festes N und gewählte Funktion $\mathcal{F}(N)$ bestimme das Minimum $\mathcal{M}_N^*$ von $\alpha_{\text{bias}} + \alpha_{\text{sampling}}$ in $\mathcal{M}$ und betrachte die Hypothesenraum-Folge

$$\mathscr{F}_N^* = \Bigg\{ g : D \to \mathbb{R} \,:\, g = \sum_{i=1}^{[\mathcal{M}_N^*]} u_i \sigma(A_i);\ , t_i \in \mathbb{R}, \boldsymbol{a}_i \in S^d \cup \{0\}, \sum_{i=1}^{\mathcal{M}(N)} |u_i| \le \mathcal{F}(N),\ i \in \{1, ..., [\mathcal{M}_N^*]\} \Bigg\}.$$

Hierbei bezeichnet $[\mathcal{M}_N^*]$ das auf die nächste ganze Zahl ≥ 1 gerundete $\mathcal{M}_N^*$.

Proposition 6.1. *Man betrachte Satz 6.11 mit der Sequenz von Hypothesenräumen $(\mathscr{F}_N^*)_{N\in\mathbb{N}}$ wie soeben definiert. Falls zusätzlich zu $\mathcal{F}(N) \to \infty$ auch*

$$\frac{S^{-\frac{2}{3}} \mathcal{F}^{\frac{20}{3}} \ln[S^{-\frac{2}{3}} \mathcal{F}^{\frac{14}{3}}]}{N} \longrightarrow 0\,,$$

wobei

$$S(N) = \frac{8c_2[1 - \ln \frac{\alpha_\infty}{c_3}]\alpha_\infty}{8c_2 + c_1\alpha_\infty} ,$$

dann gilt

$$\mathbb{P}^N \left[\sup_{f \in \mathscr{F}^{\varepsilon}_{\mathcal{M}(N)}} \left| \widetilde{\Lambda}^N(Z, f) - \Lambda(f) \right| > \alpha \right] \longrightarrow 0 .$$

Die Konvergenz ist sogar fast-sicher, falls ein $\eta > 0$ existiert, so dass

$$\mathcal{F}^4(N) N^{\eta - 1} \longrightarrow 0 \quad .$$

Der Beweis dieser Proposition benötigt etwas Vorlauf. Wie beschrieben bestimmen wir zu einem gegebenen N die Lösung

$$\mathcal{M}^*_N := \min_{\mathcal{M} \in \mathbb{N}} \left\{ \frac{4\mathcal{F}^2}{\sqrt{\mathcal{M}}} + \frac{\alpha_{\text{sampling}}(\mathcal{M})}{4\mathcal{F}^2} \right\} , \tag{6.11}$$

wobei α_{sampling} die Lösung der Gleichheit in Glg. (6.8) bezeichnet und $4\mathcal{F}^2$ als Normierungsfaktor für α_{sampling} hinzugefügt wurde unter der Bedingung, dass $L = \max_{f \in \mathscr{F}^*_N} \|f\|_\infty = \mathscr{F}(N)$ (man vergleiche mit der Bias-Variance-Abschätzung in Abschnitt 4.4.2), was aber schon in Satz 6.11 vorausgesetzt wurde. Glg. (6.8) kann mit einem Standard-Newton-Verfahren numerisch stabil gelöst werden. Abb. 6.4 illustriert (6.11) für drei Beispiele. Man beachte, dass in der Zeichnung der Gesamtfehler mit $4\mathcal{F}^2$ normiert wurde.

Satz 6.12. *Das Minimierungsproblem (6.11) hat eine eindeutige Lösung.*

Beweis. Für ein Minimum $\mathcal{M}^*_N$ ist es notwendig, dass

$$\frac{\partial \alpha_{\text{bias}}}{\partial \mathcal{M}}(\mathcal{M}^*_N) = -(4\mathcal{F}^2)^{-1} \frac{\partial \alpha_{\text{sampling}}}{\partial \mathcal{M}}(\mathcal{M}^*_N) . \tag{6.12}$$

Durch Differenzieren von (6.9) auf beiden Seiten erhalten wir für den Sampling-Fehler folgende Differentialgleichung:

$$\frac{\partial \alpha}{\partial \mathcal{M}} = -\frac{(3 + 8\alpha)^2 \left[-1 - c_2\mathcal{M} + c_2(\mathcal{M} + 1) \ln(\bar{\alpha}/c_3)\right] \bar{\alpha}}{9 + 9c_2\mathcal{M} + 2\alpha(3 + 4\alpha)(8 + 8c_2\mathcal{M} + c_1\alpha)} . \tag{6.13}$$

Weil

$$\frac{\partial \alpha_{\text{bias}}}{\partial \mathcal{M}} = -2\frac{\mathcal{F}^2}{M^{3/2}}$$

strikt negativ ist, ist für die Lösbarkeit von (6.12) folgende Bedingung notwendig:

$$\frac{\bar{\alpha}}{c_3} < \exp\left[\frac{1 + c_2\mathcal{M}}{c_2(\mathcal{M} + 1)}\right] , \tag{6.14}$$

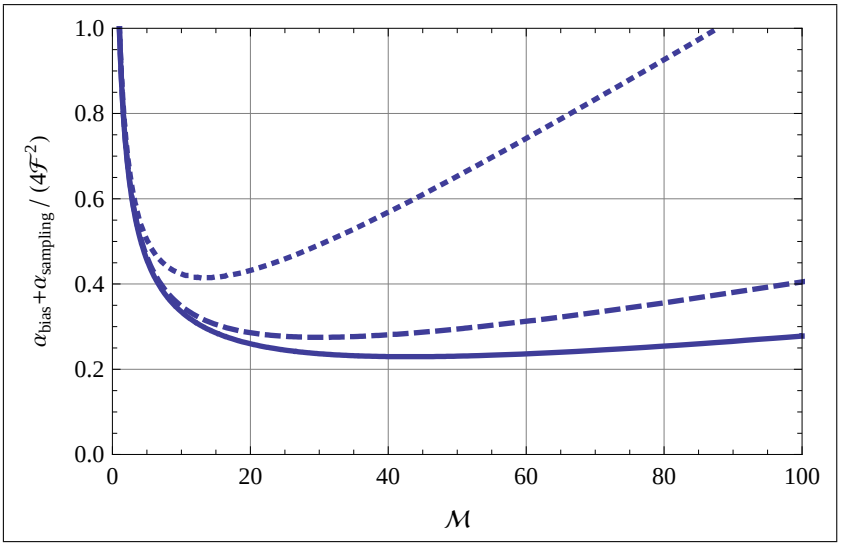


Abb. 6.4. Die obere Schranke für den Gesamtfehler (normiert) in Abhängigkeit von $\mathcal{M}$ für die Fälle $N = 100$ (dotted), $N = 500$ (gestrichelt) und $N = 1000$. Als Parameter wurden $\delta = 0.1$ und $d = 3$ gewählt. Die Lösung von Glg. (6.8) wurde mit dem Standard-Newton-Verfahren ermittelt.

wobei noch zu beachten ist, dass $\alpha(\mathcal{M})$ strikt positiv ist. Abb. 6.5 zeigt den Verlauf von $\bar{\alpha}/c_3$. Mit (6.10) erhalten wir

$$\frac{\bar{\alpha}}{c_3} = \frac{8c_2}{c_1c_3} L\left[\frac{c_1c_3}{8c_2}\exp\left(\frac{c_4}{c_2(\mathcal{M}+1)}\right)\right] , \tag{6.15}$$

was von der allgemeinen Form $c_2H^{-1}(N)L[H(N)G(\mathcal{M})]$ ist, wobei $H(N) \geq 1$ eine superlinear in N streng monoton steigende und $G(\mathcal{M})$ eine in $\mathcal{M}$ streng monoton asymptotisch auf 1 fallende Funktion bezeichnet. Diese Funktion fällt streng monoton auf 0 und umso schneller je größer $\mathcal{M}$ ist. Es stellt sich also die Frage, ob es für alle $N \geq 1$ ein $\mathcal{M} \geq 0$ gibt, so dass

$$L\left[\frac{c_1c_3}{8c_2}\exp\left(\frac{c_4}{c_2(\mathcal{M}+1)}\right)\right] < \frac{c_1c_3}{8c_2}\exp\left[\frac{1+c_2\mathcal{M}}{c_2(\mathcal{M}+1)}\right] . \tag{6.16}$$

Die linke Seite dieser Ungleichung ist streng monoton fallend, die rechte streng monoton steigend. Damit die Ungleichung für alle $\mathcal{M}$ nicht erfüllt ist, muss somit insbesondere

$$L\left[\frac{c_1c_3}{8c_2}\right] > e\frac{c_1c_3}{8c_2}$$

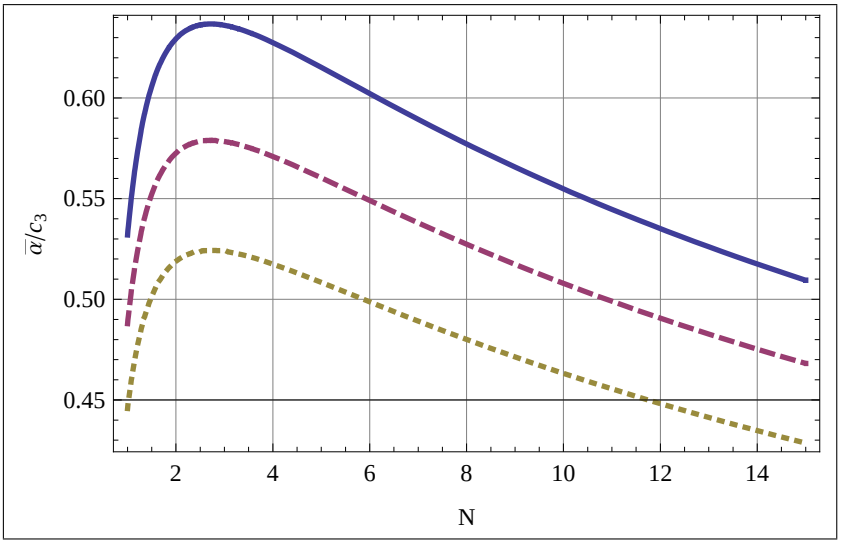


Abb. 6.5. $\bar{\alpha}/c_3$ für $\mathcal{M} = 0$ (durchgezogene Kurve), $\mathcal{M} \to \infty$ (dotted) und $\mathcal{M} = 2$ (gestrichelt). Es wurden dieselben Parameter wie für die anderen Abbildungen verwendet. Weil $1 < \exp[(1 + c_2\mathcal{M})/(c_2(\mathcal{M}+1))] < \exp(1/c_2)$ und $c_2 \geq 5$ existiert somit für alle $N \geq 1$ eine Lösung von (6.11).

gelten, was aber aufgrund der Eigenschaften von L nicht möglich ist. Es gibt also für alle $N \geq 1$ in jedem Fall eine Lösung $\mathcal{M}$ von (6.12), unabhängig von der Wahl der Parameter d, δ und $\mathcal{F}$.
Die Eindeutigkeit der Lösung ergibt sich folgendermaßen. Durch Nachrechnen erhält man:

$$\frac{\partial \alpha}{\partial \mathcal{M}} \xrightarrow{\mathcal{M}\to\infty} \frac{8c_2[1 - \ln \frac{\alpha_\infty}{c_3}]\alpha_\infty}{8c_2 + c_1\alpha_\infty} > 0$$

für jedes feste N, $\alpha(\mathcal{M})$ ist also im Grenzfall eine linear in $\mathcal{M}$ ansteigende Funktion. Hieraus ergibt sich sofort die Eindeutigkeit der Lösung, denn die Ableitung des Bias-Fehlers strebt asymptotisch gegen Null für $\mathcal{M} \to \infty$ und $\alpha_{\text{bias}} \to 0$ für $\mathcal{M} \to \infty$. ■

Wir kommen nun zurück zu Proposition 6.1. Es gilt zu zeigen, unter welchen Voraussetzungen Korollar 6.8 für die aus den Minima von (6.11) gebildete Folge $(\mathcal{M}_N^*)$ erfüllt ist. Wie gezeigt verhält sich $\alpha(\mathcal{M})$ für $N \geq 1$ praktisch linear in $\mathcal{M}$ mit der Steigung

$$S(N, \mathcal{F}) = \frac{8c_2[1 - \ln \frac{\alpha_\infty}{c_3}]\alpha_\infty}{8c_2 + c_1\alpha_\infty} .$$

Somit betrachten wir Näherungsweise

$$\mathcal{M}_N^* \approx \min_{\mathcal{M}\in\mathbb{N}} \left\{ \frac{4\mathcal{F}^2}{\sqrt{\mathcal{M}}} + \frac{S\mathcal{M}+C}{4\mathcal{F}^2} \right\} ,$$

mit der Lösung

$$\mathcal{M}_N^* \approx 4S(N,\mathcal{F})^{-\frac{2}{3}} \mathcal{F}(N)^{8/3} . \tag{6.17}$$

Abb. 6.6 zeigt den exakten Verlauf von $\mathcal{M}_N^*$ im Vergleich mit der Näherung und $\mathcal{F}(N) = 1 + \ln N$. Man erkennt sofort die Exaktheit der Näherung. Setzt man (6.17)

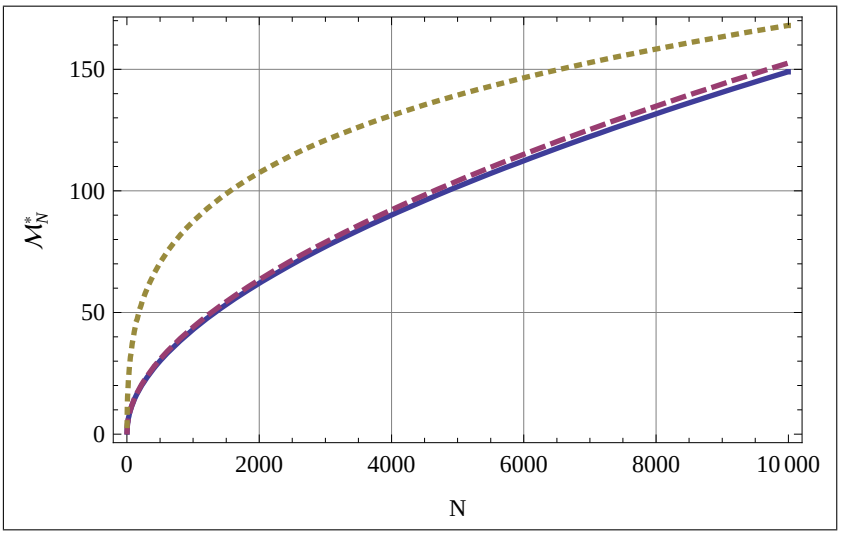


Abb. 6.6. Der Verlauf des numerisch berechneten $\mathcal{M}_N^*$ (durchgezogen) im Vergleich mit der Näherung $\mathcal{M}_N^* \approx 4S^{-\frac{2}{3}}\mathcal{F}^{8/3}$ (gestrichelt) und $\mathcal{F}(N) = 1 + \ln N$ (dotted). Die Parameterwahl ist dieselbe wie in Abb. 6.4.

in Korollar 6.8 ein, so ergibt sich der erste Teil der Proposition. Die fast-sichere Konvergenz ergibt sich wie in dem Korollar. Abb. 6.7 bestätigt die Proposition auch numerisch und zeigt den Verlauf von $\mathcal{K}$ für große N im Vergleich mit einer Funktion für $\mathcal{M}$, die Korollar 6.8 sicher erfüllt, wie z.B. $\mathcal{M}(N) = \ln N$. Man sieht, dass die Folge schneller als logarithmisch abfällt.

Wir kommen nun zum entscheidenden Punkt:

Korollar 6.9. *Es sei*

$$\hat{f}_N^* := \operatorname*{argmin}_{f\in\mathscr{F}_N^*} \widetilde{A}^N(f) .$$

Erfüllt $\mathcal{F}$ die Voraussetzungen von Proposition 6.1, so konvergiert die Folge $(\hat{f}_N^*)_{N\in\mathbb{N}}$ punktweise fast-überall und sogar fast-gleichmäßig.

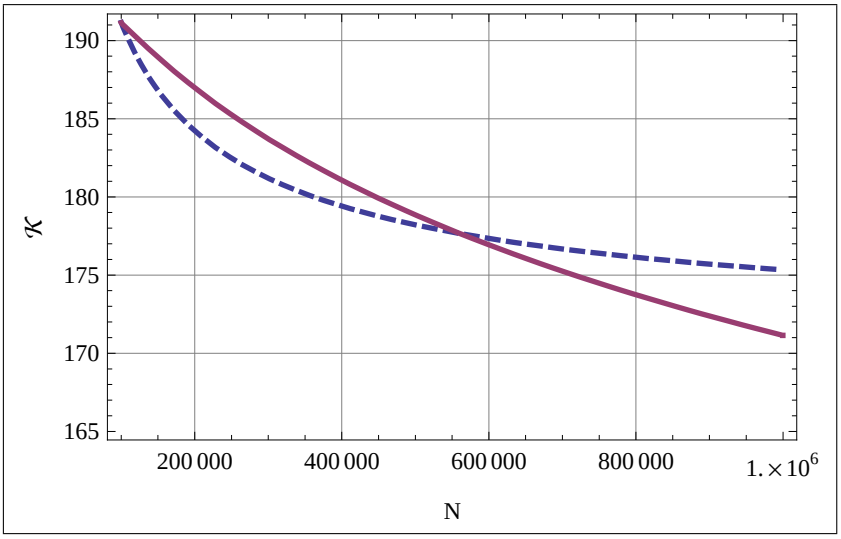


Abb. 6.7. Der Verlauf von $\mathcal{K} = \mathcal{M}(N)\mathcal{F}^4(N) \ln\left[\mathcal{M}(N)\mathcal{F}^2(N)\right] N^{-1}$ für $\mathcal{M}(N) = \mathcal{M}_N^*$ und für $\mathcal{M}(N) = 170 + \ln N$ (gestrichelt). Die Parameter sind dieselben wie in den Abb. 6.6 und 6.4.

Beweis. Die gleichmäßig fast-sichere Konvergenz der Folge der empirischen Risiken folgt wie gezeigt mit Proposition 6.1. Für die fast-sichere (punktweise fast-überall) Konvergenz der Funktionenfolge $(\hat{f}_N^*)_{N\in\mathbb{N}}$ ist wie in Satz 4.1 das Gesetz der großen Zahlen anzuwenden. Die fast-gleichmäßige Konvergenz folgt dann wegen der Endlichkeit des Maßes aus dem Satz von Egorov. ■

Wie haben somit die zu Ende von Abschnitt 4.5 formulierten Ziele erreicht.

6.2.2 Konditionskontrolle mit $\mathscr{F}_N^*$

Bislang haben wir in diesem Abschnitt aber noch nicht die Einschränkung von $\boldsymbol{a}_i$ auf S^d beachtet: Alle gemachten Aussagen gelten auch für beliebiges $\boldsymbol{a}_i$. Wir erinnern an die Definition von $\Sigma_M^d|_{B_4}(\sigma)$ (6.3) und Satz 6.10. Es ist klar, dass $\mathscr{F}_N^*$ von der Klasse $\Sigma_M^d|_{B_4}(\sigma)$ ist, nur dass $\mathcal{F}$ von N abhängt. Von daher ist jedes $\mathscr{F}_N^*$, $N = 1, 2, ...$, gleichgradig stetig. Unter der Voraussetzung, dass σ global Lipschitz-stetig mit der Lipschitz-Konstanten L ist, schließen wir also mit Satz 6.10, dass für alle $f \in \mathscr{F}_N^*$

$$\kappa_{\text{abs}}\left(\mathscr{F}_N^*\right) \leq L\mathcal{F}(N)\,.$$

Man spricht in der Regel von einem gut-konditioniertem Problem, wenn $\kappa_{\text{abs}} \leq 1$. Wählen wir also z.B.

$$\mathcal{F}(N) = \frac{1}{L \ln 3} \ln\left(2 + \frac{N}{N^*}\right) , \tag{6.18}$$

so ist für eine fest vorgegebene Trainingsdatenmenge N^*

$$\kappa_{\text{abs}}\left(\mathscr{F}^*_{N^*}\right) \leq 1 .$$

Weiterhin erfüllt $\mathcal{F}(N)$ die Anforderungen der Proposition 6.1 und damit gilt Korollar 6.9. Die Lipschitz-Konstante für die logistische Funktion als Aktivierungsfunktion ist wie in Beispiel 6.1 gezeigt $1/4$ und für tanh ist $L = 1$ (Beispiel 6.2). Weiterhin gilt mit Satz 6.10, dass die Kondition für alle $f \in \sigma(\mathscr{F}^*_{N^*})$ durch $1/L$ beschränkt ist.
Mit Satz 6.8 wird ersichtlich, dass $\mathscr{F}^*_{N^*}$ und $\sigma(\mathscr{F}^*_{N^*})$ auch gleichgradig stetig im Raum der Gewichte sind, wobei die Kondition nach Satz 6.10 durch

$$\kappa_{\text{abs},W} \leq \sqrt{\mathcal{M}^*_{N^*} A^2 + L(T+1)}$$

beschränkt ist, wobei A gegeben ist durch $\sigma(x) \leq A$ für alle $x \in \mathbb{R}$ und $T = \max_{\mathcal{X}} |\boldsymbol{x}|$. Je größer N^* gewählt wird, desto mehr wirken sich Störungen in den Gewichten auf den Output des Netzwerkes aus.

6.2.3 Ein neuer Konstruktionsalgorithmus

Wir schlagen nun auf Grundlage der in Abschnitt 6.2 und 5.4 erarbeiteten Ergebnisse folgenden praktischen Konstruktionsalgorithmus für Regressionsprobleme vor. Wir spezialisieren Algorithmus 5.10 auf Neuronale Netze und wählen

$$G_k = \{\boldsymbol{x} \mapsto u\sigma(\boldsymbol{a} + t) : |u| = b_k, \boldsymbol{a} \in S^d\} \cup \{\boldsymbol{x} \mapsto u : |u| = b_k, \boldsymbol{a} \in S^d \cup \{0\}\} ,$$

so dass

$$\text{co}(G_k) = \left\{ g : \mathbb{R}^d \to \mathbb{R} \ : \ g = \sum_{i=1}^{\infty} u_i \sigma(A_i);\ , t_i \in \mathbb{R}, \boldsymbol{a}_i \in S^d, \sum_{i=1}^{\infty} |u_i| \leq b_k \right\} .$$

Gegeben seien eine Aktivierungsfunktion σ mit Lipschitzkonstante L sowie N^* Datenpunkte $\mathcal{T}^{N^*} : \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{N^*}, y_{N^*})\}$ mit $\boldsymbol{x}_i \in \mathbb{R}^d$. Weiterhin bezeichne $\|\cdot\|_N$ die empirische Norm im zu Grunde liegenden Hilbertraum, also

$$\|f(X) - Y\|_N := \frac{1}{N} \sum_{i=1}^{N} |f(\boldsymbol{x}_i) - y_i|^2 .$$

1: **procedure** Optimales Neuronales Netz($\mathcal{T}^{N^*}$)
2: **Initialisierung:**

3: Wähle

$$b_k = \frac{1}{L \ln C} \ln\left[C - 1 + h\left(\frac{k}{N^*}\right)\right]$$

mit $C \geq 2$ und einer Funktion h so, dass Proposition 6.1 erfüllt ist und

$$\lim_{x\to 0} h(x) = 0\,, \quad \lim_{x\to\infty} h(x) = \infty\,, \quad h(1) = 1\,.$$

4: Finde die eindeutige Lösung des Minimierungsproblems

$$\mathcal{M}^*_{N^*} := \min_{\mathcal{M}\in\mathbb{N}} \left\{ \frac{4b^2_{N^*}}{\sqrt{\mathcal{M}}} + \frac{\alpha_{\text{sampling}}(\mathcal{M})}{4b^2_{N^*}} \right\}\,,$$

wobei α_{sampling} durch die Lösung von (6.8) gegeben ist.

5: Wähle

$$\varepsilon_1 \leq b_1^2\,.$$

6: Bestimme ein $f_1 := g$, so dass

$$\|f_1 - Y\|^2_{N^*} \leq \inf_{g\in G_1} \|g - Y\|^2_{N^*} + \varepsilon_1\,.$$

7: **Iteration:**

8: **for** $k = 2$ **to** $\mathcal{M}^*_{N^*}$ **do**

9: Wähle

$$\varepsilon_k \leq 4b_k^2 K_1 - 4b^2_{k-1} K_2\,,$$

wobei $K_1 = k^{-\frac{1}{2}} - (k+1)^{-2}$ und $K_2 = (k-1)^{-\frac{1}{2}} - 2(k+1)^{-1}(k-1)^{-\frac{1}{2}}$.

10: Wähle $\alpha = 1 - 2/(k+1)$ und $\overline{\alpha} = 1 - \alpha$.

11: Bestimme $f_k := \alpha f_{k-1} + \overline{\alpha} g$, so dass

$$\|f_k - Y\|^2_{N^*} \leq \inf_{g\in G_k} \|\alpha f_{k-1} + \overline{\alpha} g - Y\|^2_{N^*} + \varepsilon_k\,.$$

12: **end for**

13: **return** $f_{\mathcal{M}^*_{N^*}}$

14: **end procedure**

$\text{co}(\overline{G})$ liegt nach Lemma 6.3 dicht in L^p, so dass für $f \in L^p$ $d_f = \inf_{g'\in\text{co}(\overline{G})} \|g' - Y\|_{N^*} = 0$, wobei $\overline{G} = \bigcup_{k=1}^{\infty} G_k$. Der mittlere quadratische Fehler des nach diesem Algorithmus erhaltenen Schätzers beträgt somit nach Satz 5.10

$$\|f_{\mathcal{M}^*_{N^*}} - Y\|^2 \leq \frac{4\mathcal{F}(N^*)^2}{\sqrt{\mathcal{M}^*_{N^*}}} = \frac{4}{L^2\sqrt{\mathcal{M}^*_{N^*}}} \approx 2\sqrt[3]{\frac{S(N^*, \mathcal{F}(N^*))}{L^2}}\,,$$

wobei S in Proposition 6.1 definiert ist. Dieser Ausdruck konvergiert offensichtlich für N^* gegen Null und hängt nun lediglich von der Input-Dimension d und der Lipschitz-Konstanten L ab. Abb. 6.8 illustriert dies für einige Beispiele. Der Gesamtfehler hängt

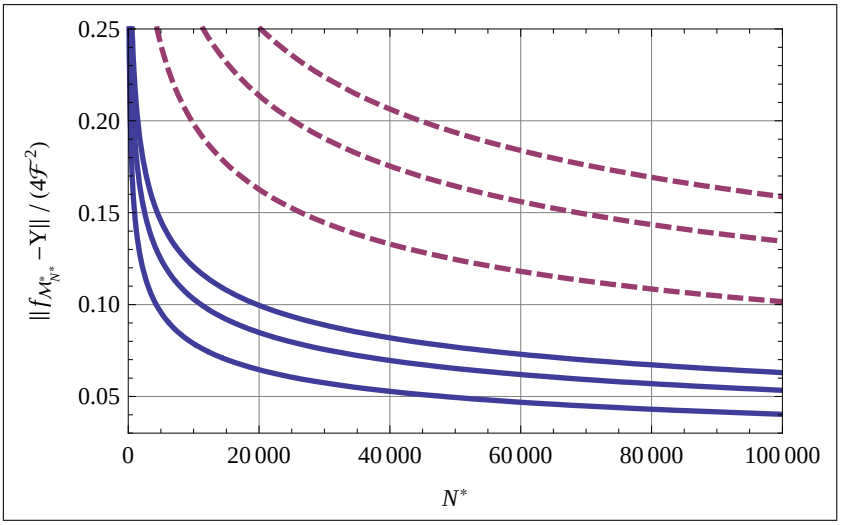


Abb. 6.8. Der Verlauf der oberen Schranke des MSE $\|f_{\mathcal{M}^*_{N^*}} - Y\|^2$ für $L = 1$ (durchgezogene Linien) wie z.B. $\sigma = \tanh$ und für $L = 0.25$ (gestrichelt). Gezeichnet sind jeweils die Fälle $d = 1$ (die jeweils untersten der drei Linien), $d = 5$ (mittlere Linien) und $d = 10$ (oberste Linien).

nun einzig und allein von der gegebenen Menge von Trainingspunkten N^* ab:

$$(\alpha_{\text{bias}} + \alpha_{\text{sampling}})(N^*) \approx 2\sqrt[3]{\frac{S(N^*)}{L^2}} + \frac{1}{4}\left(L^2\alpha^*_{\text{sampling}}(N^*)\right) ,$$

wobei $\alpha^*_{\text{sampling}}$ die Lösung von

$$\frac{3}{8}NL^4\frac{\alpha^2}{8\alpha+3} - \left(4S^{-\frac{2}{3}}L^{-\frac{8}{3}}(2d+3)+1\right)\ln\left(\frac{128e(4S^{-\frac{2}{3}}L^{-\frac{8}{3}}+1)}{L^2\alpha}\right) - \ln 2 + \ln\delta = 0$$

bezeichnet. Abb. 6.9 zeigt den Gesamtfehler in Abhängigkeit von N^* für drei Lipschitz-Konstanten.

Satz 6.13. *Gegeben seien N^* Datenpunkte. Dann ist der Algorithmus* Optimales Neuronales Netz *gleichmäßig Hypothesen-stabil mit*

$$\beta^{glm}_{\mathcal{M}^*_N} \leq 16\, b^4_{\mathcal{M}^*_{N^*}} . \tag{6.19}$$

Beweis. Wir bezeichnen mit $g_{k,\mathcal{T}}$ die Funktion in G_k, die das empirische Risiko

$$\Lambda^N_\Xi(g) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{y}_i - g(\boldsymbol{x}_i))^2$$

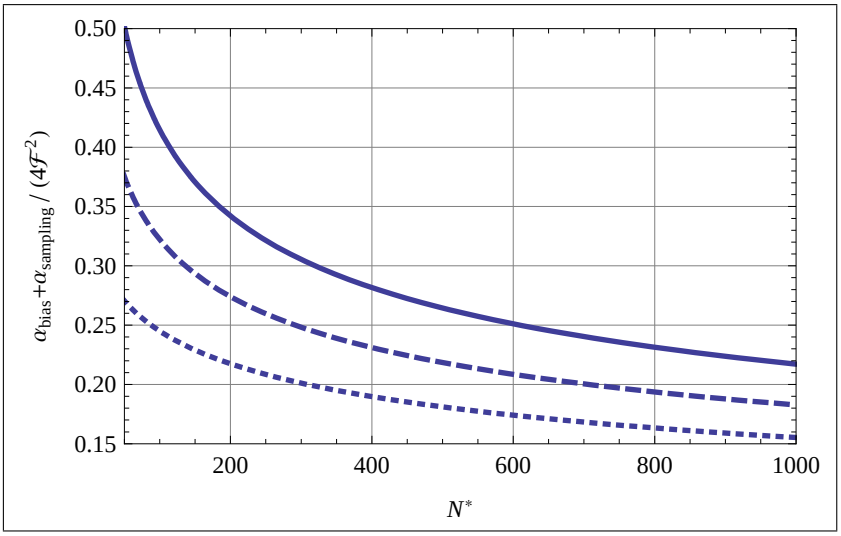


Abb. 6.9. Der Verlauf der des Gesamtapproximationsfehlers für $L = 1$ (durchgezogene Linie), $L = 0.25$ (gestrichelt) und $L = 0.1$. Es wurden $d = 3$ und $\delta = 0.1$ gewählt.

für den Datensatz $\mathcal{T}$ unter der Nebenbedingung $\|g\| \leq b_k$ minimiert. Entsprechend bezeichne $g_{k,\mathcal{T}'}$ die Funktion, die

$$\Lambda_{\Xi}^{N\setminus\{j\}}(g) = \frac{1}{N-1}\sum_{i\neq j}(\boldsymbol{y}_i - g(\boldsymbol{x}_i))^2$$

mit $g \in G_k$, $\|g\| \leq b_k$ minimiert. Für die gleichmäßige Hypothesen-Stabilität schätzen wir folgendermaßen ab:

$$\begin{aligned}\mathbb{E}_{\mathcal{T}}\left[\|\Xi(Y, g_{k,\mathcal{T}}(X)) - \Xi(Y, g_{k,\mathcal{T}'}(X))\|_\infty^2\right] &\leq 4b_k(N)^2\,\mathbb{E}_{\mathcal{T}}\left[\|g_{k,\mathcal{T}} - g_{k,\mathcal{T}'}\|_\infty^2\right]\\ &\leq 16b_k(N)^4\,.\end{aligned}$$

Damit ist die Behauptung gezeigt. ■

Aus der gleichmäßigen Hypothesen-Stabilität folgt direkt auch die punktweise. Diese lässt sich auch im Stile von Satz 5.11 direkt abschätzen:

Lemma 6.4. *Gegeben seien N^* Datenpunkte. Dann ist der Algorithmus* Optimales Neuronales Netz *punktweise Hypothesen-stabil mit*

$$\beta_{\mathcal{M}_N^*}^{pkt} \leq 4b_{\mathcal{M}_N^*}(N)\sum_{k=1}^{[\mathcal{M}_N^*]}\left\{\frac{4}{\sqrt{k}}\left[b_k(N)^2 + b_k(N-1)^2\right] + \frac{\alpha(k,N)}{4b_k(N)^2} + \frac{\alpha(k,N-1)}{4b_k(N-1)^2}\right\}, \tag{6.20}$$

wobei $\alpha \equiv \alpha_{sampling}$.

Beweis. Der Beweis verläuft analog zu dem von Satz 5.11. Es bezeichne $f \in H$ die Zielfunktion, also $f = \mathbb{E}[Y|X]$, und f_{opt} die bestmögliche Approximation innerhalb von $\mathrm{co}(\overline{G})$, also $f_{opt} = \mathrm{argmin}_{f \in \mathrm{co}(\overline{G})} \mathbb{E}[(\boldsymbol{y} - f(\boldsymbol{x}))^2]$. f_{opt} ist also die Projektion von f auf die konvexe Hülle von $\overline{G}$. Wir wollen nun den Term

$$\mathbb{E}_{\overline{\mathcal{T}}} \left[\|g_{k,\mathcal{T}} - g_{k,\mathcal{T}'}\|_N^2 \right] ,$$

$k \geq 1$, abschätzen.
Wir beginnen mit $k = 1$. Nach Definition gilt für $g_{k,\mathcal{T}}$ im ersten Schritt:

$$\|g_{1,\mathcal{T}} - f_{opt}\|_N^2 \leq \inf_{g \in G_1} \|g - f_{opt}\|_N^2 + \varepsilon_1 .$$

Wir schätzen den Approximationsfehler gemäß Abschnitt 6.2.1 ab:

$$\mathbb{E}_{\overline{\mathcal{T}}} \left[\|g_{1,\mathcal{T}} - f_{opt}\|_N^2 \right] \leq 4b_1(N)^2 + \frac{\alpha_{\text{sampling}}(1,N)}{4b_1(N)^2} ,$$

so dass

$$\begin{aligned} \mathbb{E}_{\overline{\mathcal{T}}} \left[\|g_{1,\mathcal{T}} - g_{1,\mathcal{T}'}\|_N^2 \right] &\leq \mathbb{E}_{\overline{\mathcal{T}}} \left[\|g_{1,\mathcal{T}} - f_{opt}\|_N^2 + \|g_{1,\mathcal{T}'} - f_{opt}\|_N^2 \right] \\ &\leq 4 \left[b_1(N)^2 + b_1(N-1)^2 \right] + \frac{\alpha_{\text{sampling}}(1,N)}{4b_1(N)^2} + \frac{\alpha_{\text{sampling}}(1,N-1)}{4b_1(N-1)^2} . \end{aligned}$$

Wegen

$$\mathbb{E}_{\overline{\mathcal{T}}} \left[\|\Xi(Y, g_{1,\mathcal{T}}(X)) - \Xi(Y, g_{1,\mathcal{T}'}(X))\|_N^2 \right] \leq 4b_1(N)^2 \, \mathbb{E}_{\overline{\mathcal{T}}} \left[\|g_{1,\mathcal{T}} - g_{1,\mathcal{T}'}\|_N^2 \right] \qquad (6.21)$$

ist der Algorithmus im ersten Schritt β_1^{pkt}-Hypothesen-Stabil gemäß Definition 5.2 mit

$$\beta_1^{pkt} \leq 16 \left[b_1(N)^3 + b_1(N) b_1(N-1)^2 \right] + \frac{\alpha_{\text{sampling}}(1,N)}{b_1(N)} + \frac{b_1(N)}{b_1(N-1)^2} \alpha_{\text{sampling}}(1,N-1).$$

Nun betrachten wir den Fall $k = \mathcal{M}_N^*$:

$$\begin{aligned} \mathbb{E}_{\overline{\mathcal{T}}} \left[\left\| g_{\mathcal{M}_N^*,\mathcal{T}} - g_{\mathcal{M}_N^*,\mathcal{T}'} \right\|_N^2 \right] &\leq \mathbb{E}_{\overline{\mathcal{T}}} \left[\left\| g_{\mathcal{M}_N^*,\mathcal{T}} - f_{opt} \right\|_N^2 \right] + \mathbb{E}_{\overline{\mathcal{T}}} \left[\left\| g_{\mathcal{M}_N^*,\mathcal{T}'} - f_{opt} \right\|_N^2 \right] \\ &\quad + \mathbb{E}_{\overline{\mathcal{T}}} \left[\left\| g_{\mathcal{M}_{N-1}^*,\mathcal{T}} - g_{\mathcal{M}_{N-1}^*,\mathcal{T}'} \right\|_N^2 \right] \\ &\leq \ldots \leq \sum_{k=1}^{\mathcal{M}_N^*} \left\{ \mathbb{E}_{\overline{\mathcal{T}}} \left[\|g_{k,\mathcal{T}} - f_{opt}\|_N^2 \right] + \mathbb{E}_{\overline{\mathcal{T}}} \left[\|g_{k,\mathcal{T}'} - f_{opt}\|_N^2 \right] \right\} \\ &\leq \sum_{k=1}^{\mathcal{M}_N^*} \left\{ \frac{4}{\sqrt{k}} \left[b_k(N)^2 + b_k(N-1)^2 \right] + \frac{\alpha_{\text{sampling}}(k,N)}{4b_k(N)^2} + \frac{\alpha_{\text{sampling}}(k,N-1)}{4b_k(N-1)^2} \right\} \end{aligned}$$

und die Behauptung ist gezeigt. ■

Eine numerische Auswertung des Ausdrucks für $\beta^{pkt}_{\mathcal{M}^*_N}$ zeigt allerdings, dass die "triviale" Abschätzung

$$\beta^{pkt}_{\mathcal{M}^*_N} \leq \beta^{glm}_{\mathcal{M}^*_N}$$

sogar eine bessere Obergrenze für die punktweise Hypothesen-Stabilität darstellt (s. Abb. 6.10 für ein Beispiel). Für

$$b_k = \frac{1}{L \, \ln C} \, \ln \left[C - 1 + h\left(\frac{k}{N^*}\right) \right]$$

wird nun ersichtlich, dass wegen $\lim_{x \to 0} h(x) = 0$

$$\beta^{glm}_{\mathcal{M}^*_{N^*}} \xrightarrow{N^* \to \infty} \text{const} =: b_\infty \tag{6.22}$$

mit

$$b_\infty = \frac{\ln(C-1)}{L \, \ln C} \, .$$

Der Algorithmus ist also für jede Datenmenge N^* gleichmäßig Hypothesen-stabil!

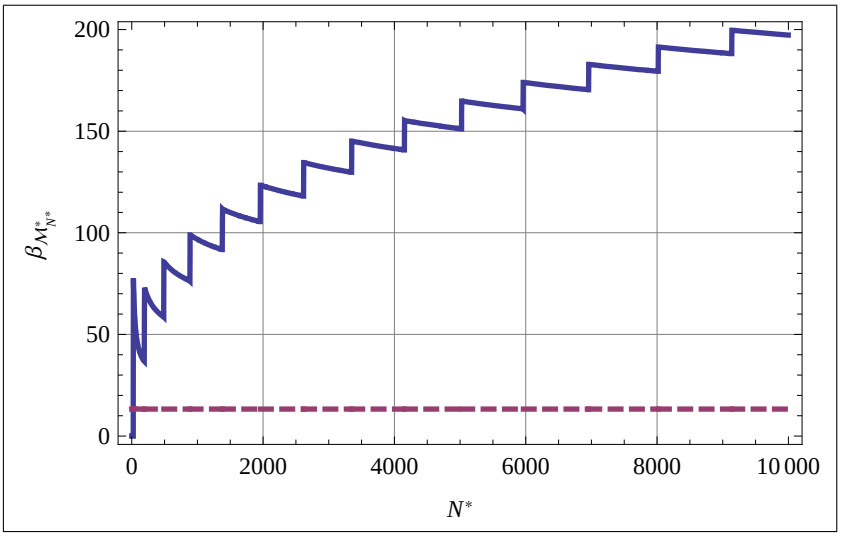


Abb. 6.10. Der Verlauf der gleichmäßigen Hypothesen-Stabilitätskonstante (6.19) (gestrichelt) im Vergleich mit der punktweisen (6.20) für $h(x) = x$. Als Konstanten wurden $d = 3$, $\delta = 0.1$ und $L = 1$ gewählt. In diesem Beispiel ist (6.19)<(6.20) für $N^* > 23$, wobei $\beta^{glm} = b_\infty = 13.2665$.

Wir wenden nun Satz 5.3 aus Abschnitt 5.1 an. Zur Erinnerung: Dieser Satz besagt, dass der Sampling-Fehler folgendermaßen abgeschätzt werden kann für einen β-gleichmäßig-stabilen Lernalgorithmus:

$$\Lambda(\hat{f}_{\mathcal{T}}) - \Lambda^N(\hat{f}_{\mathcal{T}}) \leq 2\beta^{glm}_{\mathcal{M}^*_{N^*}} + \left(4N^*\beta^{glm}_{\mathcal{M}^*_{N^*}} + 4b_k^2\right)\sqrt{-\frac{\ln\delta}{2N^*}}\,, \tag{6.23}$$

wobei $0 < \delta < 1$. In Abschnitt 5.4.1, z.B. Glg. (5.11), haben wir gezeigt, dass allgemeine rekursive greedy-Algorithmen nicht als global Hypothesen-stabil bezeichnet werden können, weil $N\beta$ nicht gegen Null konvergierte.
Der Algorithmus Optimales Neuronales Netz hingegen ist global gleichmäßig Hypothesen-stabil für geeignetes C und h, wie die folgende Überlegung zeigt:

$$N^*\beta^{glm}_{\mathcal{M}^*_{N^*}} \leq 16N^*\, b^4_{\mathcal{M}^*_{N^*}} = \frac{16N^*}{L\,\ln C}\,\ln\left[C - 1 + h\left(\frac{k}{N^*}\right)\right]\,.$$

Weil die Funktion $x\ln(1+x^{-1-\delta})$ für alle $\delta > 0$ gegen Null konvergiert, setzen wir $C = 2$ und h so, dass es schneller als x ansteigt. Dann konvergiert $N^*\beta^{glm}_{\mathcal{M}^*_{N^*}}$ und die ganze rechte Seite von (6.23) gegen Null. Wir fassen diese Überlegung in einem Satz zusammen:

Satz 6.14. *Es sei $C = 2$ und $h(x) = x^{1+\delta}$ mit $\delta > 0$. Dann ist der Algorithmus* Optimales Neuronales Netz *global gleichmäßig Hypothesen-stabil in jedem Schritt $k \geq 1$, d.h.*

$$\Lambda(\hat{f}_{\mathcal{T}}) - \Lambda^N(\hat{f}_{\mathcal{T}}) \xrightarrow{N^*\to\infty} 0$$

für alle $k \geq 1$.

6.2.4 Erweiterung der Ergebnisse auf RBFNs und WNNs

Alle in diesem Kapitel vorgestellten Ergebnisse lassen sich prinzipiell auch auf andere Basisfunktionen anwenden, in sofern diese Lipschitz-stetig mit Lipschitz-Konstante L sind. Für Radiale Basisfunktionen und Wavelets ist dies natürlich der Fall. Ein wichtiger Punkt in Algorithmus Optimales Neuronales Netz ist allerdings, dass der Hypothesenraum in jedem Konstruktionsschritt gleichgradig stetig ist. Für sigmoide Aktivierungsfunktionen führte dies direkt auf die Zusatzeinschränkung, dass $\boldsymbol{a}_i$, $i = 1, ..., \mathcal{M}$, auf der Einheitskugel liegt (s. Satz 6.4). Für Basisfunktionen $\psi : \mathbb{R} \to \mathbb{R}$ mit $\psi = \psi(\boldsymbol{a}^T\boldsymbol{x} - t)$ lässt sich Satz 6.4 somit uneingeschränkt anwenden, wobei L wie zuvor von der konkreten Aktivierungsfunktion abhängt. Für das Mexikanerhut-Wavelet

$$\psi(x) = \frac{2}{\sqrt{3}}\pi^{-1/4}\left(1 - x^2\right)\, e^{-x^2/2}$$

zum Beispiel ist $L = 2\sqrt{3}\pi^{-1/4} \approx 2.6$. Alle folgenden Sätze wie z.B. 6.6, 6.8 und 6.10 fordern für die Aktivierungsfunktion lediglich Beschränktheit und globale Lipschitz-Stetigkeit und sind somit auch für andere Basisfunktionen anwendbar. Für radiale Basisfunktionen der Form $\psi = \psi(\|\boldsymbol{x} - \boldsymbol{t}\|/\sigma)$ ergibt sich in gleicher Weise sofort, dass der Ausdruck

$$\sum_{i=1}^{M} \left| \frac{u_i}{\sigma_i} \right|$$

beschränkt sein muss, damit das resultierende Netzwerk gleichgradig stetig ist, d.h. wir fordern $\boldsymbol{\sigma}^{-1} = (\sigma_1^{-1}, ..., \sigma_M^{-1}) \in S^d$.

7

Zusammenfassung und Ausblick

Ziel dieser Dissertation ist es die Schwierigkeiten bei der Modellierung mit Black-Box Modellen aufzuzeigen, zu quantifizieren und Lösungsvorschläge zu bieten. Als primäres Beispiel für Black-Box-Approximatoren werden Neuronale Netze in verschiedenen Ausprägungen gewählt. Insbesondere stellt diese Arbeit Neuronale Netze aus zwei verschiedenen Sichtweisen (deterministisch und stochastisch) dar und ermöglicht so die simultane Untersuchung mehrerer Aspekte dieser Approximationsmethoden. Im Vordergrund stehen hierbei:

- ▷ Die grundsätzliche Approximationsfähigkeit von Neuronalen Netzen,
- ▷ aus modelltheoretischer Sicht die Komplexität der zu Grunde liegenden Hypothesenräume und der damit verbundene trade-off zwischen prinzipieller Approximationsgenauigkeit (Bias-Fehler) und Varianz der Schätzer,
- ▷ die Konvergenz der Neuronalen Netze gegen die Zielfunktion für steigende Anzahl von Trainingsdatenpunkten und wachsende Komplexität des Netzwerkes,
- ▷ aus praktischer Sicht die Robustheit der resultierenden Schätzfunktionen gegenüber Störungen in den Daten während der Trainings- *und* Anwendungsphase sowie den Netzwerk-Gewichten.

Teil I konzentriert sich auf die Approximationsfähigkeit von Neuronalen Netzen in Hinblick auf "Dichtheitsaussagen" der Form "die Klasse der single-layer Neuronalen Netze mit einer diskriminatorischen Aktivierungsfunktion liegt dicht in $C(D)$ mit $D \subset \mathbb{R}^d$ kompakt". Bestehende Ergebnisse werden an mehreren Stellen auf allgemeinere Rahmen erweitert. Zudem wird in Kapitel 3 die Rolle von Redundanzen in Wavelet-Entwicklungen und die Verbindung zu Wavelet Neuronalen Netzen sowie deren Robustheit gegenüber Störungen in den Wavelet-Koeffizienten untersucht. Dieser Sachverhalt wird numerisch an den für die Praxis wichtigsten Beispielen in bislang nicht erreichter Exaktheit untersucht. Insbesondere werden die Frame-Parameter und der Grad der Redundanz für das Mexikanerhut-Wavelet, das Haar-Wavelet, das Meyer-Wavelet und das

Daubechies-Grossmann-Meyer-Wavelet für verschiedene Zoom Parameter a_0 und Translationsparamter t_0 numerisch exakt bestimmt. Es zeigt sich z.B., dass das weit verbreitete Mexikanerhut-Wavelet für $a_0 > 1$ unter keinen Umständen eine Basis des L^2 bildet. Zudem werden bislang unbekannte Erweiterungen der bestehenden Theorie wie die numerische Präzisierung der endlichen Rekonstruktionsformel von Daubechies aufgezeigt. Als weitere Neuerung weisen wir auf die Verknüpfung der Begriffe Redundanz, Robustheit und Komplexität hin, so dass sich sich dieses Kapitel sinnvoll in die allgemeine Thematik dieser Dissertation einordnet.

Teil II stellt das geeignete Fundament für die Analyse der Robustheit von Black-Box Modellen gegenüber in der Praxis unvermeidlicher Störungen zur Verfügung. Dieser Ansatz unterscheidet sich von den meisten Darstellungen in der Literatur (einige Ansätze finden sich z.B. in White (1989)). In diesem *stochastischem* Rahmen werden die folgenden Aspekte aus verschiedenen Sichtweisen heraus beleuchtet und auf völlig neue Weise verknüpft:

- ▷ Empirische Risiko-Minimierung mit Black-Box Approximatoren in einem stochastischem Rahmen.
- ▷ Konsistenz in der empirischen Risiko-Minimierung.
- ▷ Topologische Einschränkungen von Hypothesenräumen und die Verbindung zu modelltheoretischen Problemstellungen (Bias-Variance-Dilemma).
- ▷ Die gleichmäßige Konvergenz der Schätzer gegen die Zielfunktion als Hauptproblem bei der praktischen Anwendung von Neuronalen Netzen. Hierbei werden vor allem sowohl Konvergenz in N (Anzahl Trainingsdatenpunkte) als auch in M (entspricht Modellkomplexität) betrachtet.
- ▷ Verschiedene Definitionen von "Robustheit" für Black-Box Approximatoren (Hypothesenstabilität, Ausreißerstabilität, Kondition des Optimierungsproblems).
- ▷ Konstruktionsalgorithmen für Neuronale Netze auf Grundlage von der Komplexität nach geordneten Modellstrukturen.
- ▷ Robustheits- und Stabilitätseigenschaften dieser Algorithmen.

Weiterhin diskutieren wir ausgiebig die Rolle der Performance-Funktion aus verschiedenen Blickwinkeln heraus. Insbesondere erweist sich dies als zentraler Punkt bezüglich der Hypothesenstabilität von Lernalgorithmen.

In Teil III werden die in Teil I und II erarbeiteten Zusammenhänge angewendet, so dass ein Konstruktionsalgorithmus angegeben werden kann, der sich bezüglich mehrerer Aspekte (Hypothesen-Stabilität, Komplexität, Konsistenz, Ausreißer-Immunität und Kondition) als optimal erweist. Neuronale Netze werden somit auf neuartige Weise als ganzheitliches Modellierungsproblem präsentiert, denn ein "gutes Modell" ist nur durch die simultane Optimierung verschiedener Aspekte erreichbar (Multi-Objective Modellierung). Ein zentraler Punkt dieser Dissertation ist in diesem Zusammenhang die Abschätzung des Bias-Fehlers mit Hilfe eines erweiterten greedy-Algorithmus. Es gelingt

den Gesamtfehler der Approximation in Abhängigkeit von Modellkomplexität und Anzahl der Trainingsdatenpunkte analytisch nach oben abzuschätzen und zu minimieren. Das resultierende Netzwerk ist zudem in dem in Teil II definierten Rahmen robust. Wir fassen im folgenden das neue Konstruktionsverfahren etwas detaillierter zusammen.

In klassischen Ansätzen wie dem SRM-Prinzip wird als Komplexitätsmaß die so genannte VC-Dimension des Hypothesenraumes verwendet, die für nicht-lineare Schätzer wie Neuronale Netze bislang allerdings nicht berechenbar ist. Auf Grundlage der Trainingsdaten wird des weiteren eine "optimale" Modellstruktur identifiziert. Es lässt sich zeigen, dass mit dieser Methode die fast-gleichmäßige Konvergenz der *empirischen Risiken* in N (Anzahl Trainingsdatenpunkte) garantiert ist. In Abschnitt 4.5 weisen wir allerdings auf das Problem hin, dass das SRM-Prinzip somit lediglich die *stochastische Konvergenz* der eigentlichen Schätzfunktionen gegen die Zielfunktion f sicher stellt und somit keine Möglichkeit bietet unkontrollierte "Spitzenbildung" zu vermeiden. Wir belegen diese Überlegung numerisch (s. Abschnitt 4.5 und Kapitel A).
Der in dieser Dissertation vorgeschlagene Algorithmus Optimales Neuronales Netz verwendet zwar ebenfalls Hypothesenräume aufsteigender Komplexität, garantiert aber sogar fast-sichere Konvergenz der *Schätzer*, nicht bloß der empirischen Risiken. Vereinfacht ausgedrückt wird somit eine Schätzfunktion garantiert, die (praktisch) keine Spitzen aufweist und die Zielfunktion glatt approximiert. Grundgedanke ist die Koppelung der Parameter N und M (entspricht der Modellkomplexität), so dass im Sinne des Satzes von Arzelà-Ascoli die Teilfolge der Folge in N und M identifiziert wird, die *fast-gleichmäßig* gegen f konvergiert.
In Kombination mit den in Satz 6.14 präsentierten Erweiterungen leistet der Algorithmus zusammengefasst folgendes:

1) Der **Gesamtfehler** der Approximation wird **minimiert**. Das Minimum ist eindeutig.
2) Fast-sichere und damit fast-gleichmäßige Konvergenz der Schätzer in N und M. Hieraus folgt, dass der Schätzer auch zwischen den Trainingsdatenpunkten von **beschränkter Variation** ist (also "glatt" ist bis auf eine Menge von Maß Null).
3) Jeder Hypothesenraum ist gleichgradig stetig, so dass ein produzierter Schätzer auch während der Laufzeit **robust gegenüber Störungen in den Eingabedaten und Netzwerkgewichten** ist.
4) Der Lernalgorithmus ist gleichmäßig hypothesen-stabil, und somit **robust gegenüber Störungen im Trainingsdatensatz**.
5) Jeder Schätzer ist numerisch **gut konditioniert**.
6) Durch Verwendung des LTS-Algorithmus als Optimierungsverfahren ist der Schätzer **maximal robust gegenüber einzelnen Ausreißern**.

Ausblick

Es gibt viele Stellen in dieser Dissertation, an denen sich eine detailliertere Analyse der beschriebenen Sachverhalte lohnen würde. Diese sind an den entsprechenden Stellen im Text gekennzeichnet und mit Referenzen versehen. Auf Grundlage der präsentierten Ergebnisse lassen sich aber auch in größerem Rahmen weitere Studien in verschiedene Richtungen durchführen.

Besonders interessant ist die Frage nach der Rolle der Konvexität der beteiligten Hypothesenräume. Die in dieser Dissertation angegebenen Räume sind i.A. nicht konvex, so dass ein eindeutiges Minimum nicht garantierbar ist. Es gilt Bedingungen für die Netzwerk-Parameter zu finden, so dass die resultierenden Funktionenräume konvex sind. Abschnitt 4.4.1 gibt in diesem Zusammenhang schon einige Abschätzungen.
Weiterhin lohnt sich sicherlich eine eingehendere Untersuchung für spezielle Verteilungen des "Rauschens". Bezüglich eines normalverteilten Fehlerterms gibt Kapitel 6 schon einige Abschätzungen.
Lediglich erwähnt wurde die Möglichkeit Neuronale Netze unter der Nebenbedingung einer minimalen Verformungsenergie des resultierenden Schätzers zu betrachten, wie es im Falle von Splines umgesetzt ist. Es stellt sich die Frage ob diese Bedingung auf ähnliche Einschränkungen wie die in dieser Dissertation Beschriebenen führt.
Eine weitere Untersuchung der Rolle der Redundanz in auf Grundlage einer Transformation (Radon, Wavelet usw.) konstruierten Schätzern kann an Kapitel 3 angeschlossen werden.

Die Tatsache, dass viele der präsentierten Analysemethoden auch anwendbar sind, wenn die Kugelradien der einzelnen Hypothesenräume durch die Konstante der Hypothesen - Stabilität ersetzt werden, wurde im Text angedeutet. Dies entspricht einer Verschiebung des Fokus von Approximationsgenauigkeit auf Robustheit. Es gilt diesen Aspekt weiter zu untersuchen. Allgemein gesprochen muss diese Dissertation in Hinblick auf die Verknüpfung der Konzepte Hypothesenstabilität, Robustheit, Komplexität, Approximationsgenauigkeit und Konvergenzeigenschaften von Black-Box Methoden als explorativ betrachtet werden. Weitere Forschungsarbeit ist unerlässlich und in Vorbereitung.

Ein Hauptfokus dieser Dissertation liegt auf der Untersuchung der modelltheoretischen Komplexität der zu Grunde liegenden Hypothesenräume. En Detail dargestellt wurde die VC-Dimension als klassisches Komplexitätsmaß. Wie beschrieben lässt sich diese Größe nicht ohne weiteres auf Neuronale Netze anwenden, so dass auch das damit verbundene klassische SRM-Prinzip von Vapnik eher theoretischen Wert besitzt. Um diese Schwierigkeiten zu umgehen wird in dieser Arbeit wie beschrieben in der Konstruktion des Netzwerkes als Komplexitätsmaß eine Kopplung von "Summenlänge" M und Kugelradius $\mathcal{F}$ vorgeschlagen. Bezüglich dieses Maßes ergibt sich dann die gewünschte

Ordnung der Hypothesenräume. Als Erweiterung der entwickelten Methode wäre eine Untersuchung von anderen Komplexitätsmaßen und möglichen Kopplungen interessant. Bereits erwähnt wurde neben der VC-Dimension hierbei z.B. die fat-shattering Dimension.

Eine wichtige Erweiterung der Darstellungen in dieser Dissertation wurde schon zu Ende von Abschnitt 4.1 angesprochen. Auf Grundlage der Arbeiten von Rossi & Conan-Guez (2005) können die Ergebnisse sicherlich auf funktionale Neuronale Netze (FNNs) erweitert werden. FNNs haben den theoretischen Vorteil, dass sie den "wirklichen" konditionellen Zusammenhang zwischen X und Y erlernen können und nicht bloß einzelne Momente wie den Erwartungswert.

Ein weiterer in dieser Dissertation bislang nicht näher beleuchtete Aspekt ist die Einbeziehung von "externem Wissen" in den Modellierungsprozess. Autoren wie Selonen, Lampinen & Ikonen (1996) beschreiben in ihren Arbeiten Methoden, um Wissen, das auf anderem Wege (z.B. physikalische Bedingungen) erworben wurde, in den Modellierungsprozess mit einzubeziehen.
Interessant ist hierbei sicherlich auch die Verbindungen von Neuronalen Netzen mit entscheidungstheoretischen Methoden wie fuzzy-Sets. Mit den in dieser Dissertation erarbeiteten Ergebnissen könnten somit Netzwerke konstruiert werden, die robuste Risiko-Minimierung auch in einem Umfeld garantieren, in dem Variablen nur unscharf formuliert sind.

In dieser Dissertation liegt ein Fokus auf der Robustheit der resultierenden Schätzer gegenüber Störungen in den Eingabedaten und der Netzwerk-Parameter. In der Praxis können aber auch andere Nebenbedingungen eine Rolle spielen. Es gilt dies in konkreten Situationen zu untersuchen und in den gegebenen Rahmen zu integrieren.

Abschließend lässt sich sagen, dass sich durch die konsequente stochastische Formulierung von Black-Box Modellen große Möglichkeiten für die Weiterentwicklung dieser Approximationsmethoden bieten. Wie in dieser Arbeit gezeigt reicht dies für einen erfolgreichen praktischen Einsatz von Neuronalen Netzen aber noch nicht aus. Entscheidend ist, dass die empirische Risiko-Minimierung nicht das einzige Ziel in der Modelloptimierung sein kann. Black-Box Approximatoren sind vielmehr ein komplexes multi-objective Optimierungsproblem. Diese Dissertation liefert in diesem Sinne wertvolle neue Methoden um Licht in den "schwarzen Kasten" zu bringen und wahrlich intelligente Lernalgorithmen zu konstruieren.

A

Demonstratornetzwerk und Beispielprobleme

Die in dieser Dissertation gezeigten numerischen Resultate sind das Ergebnis eines selbstprogrammierten Neuronalen Netzes. In Pohl (2007) wird zudem ein im Rahmen dieser Dissertation entstandenes Wavelet Neuronales Netz präsentiert. Wir werden im folgenden die algorithmischen Details recht knapp beschreiben (für Details s. Pohl (2007)) und einige weitere Beispielprobleme als "proof of principle" vorstellen.

Erstellung einer Wavelet-Bibliothek
Wir erinnern an dieser Stelle an die endliche Rekonstruktionsformel 3.5 die zeigt, dass schon mit einer endlichen Anzahl von Wavelets eine beliebig genaue Approximation möglich ist. Die Erstellung des Wavelet Netzwerkes läuft nun in drei Schritten ab: Zunächst wird eine Bibliothek erstellt, dann werden die für das konkrete Problem signifikantesten Regressoren (also Wavelets in der Bibliothek) ausgewählt und im letzten Schritt werden die Netzwerk-Parameter optimiert.
Pohl (2007) gibt einen Überblick über zwei in der Praxis verwendete Standardalgorithmen zur Auswahl der Regressoren in Schritt 2: Sukzessive Regressorauswahl mit schrittweiser Orthogonalisierung und Rückwärtselemination von Regressoren. Man beachte, dass die Verfahren im Vergleich mit Zhang (1994) etwas abgewandelt wurden.

Regressoroptimierung
Die Standard-Optimierungsmethode für Probleme mit möglicherweise rangdefizitären Jacobi-Matrizen (s. Abschnitt 5.3) stellt die Levenberg-Marquardt-Methode dar. In der Umsetzung des WNN wurde allerdings ein Konditions-Penalty-Term eingefügt, so dass sich Effekte, die durch nicht-gleichmäßige Konvergenz des Schätzers entstehen abfangen lassen. Die "Strafe" ist klein für schlecht lokalisierte Wavelets, die viele Trainingsdaten abdecken und groß für sehr "spitze" Wavelets.

Cluster-Algorithmus
Pohl (2007) beschreibt zusätzlich eine komplett neue Methode der Konstruktion eines WNN, und zwar durch Anwendung eines Cluster-Algorithmus. Hierbei werden die Zentren der Wavelets mit Hilfe eines Cluster-Algorithmus aus den Trainingsdaten bestimmt.

Beispiel 1
In diesem Problem wird die Funktion $f(x_1, x_2) = (x_1^2 - x_2^2)\sin(5x_1)$ approximiert. Die 441 Vektoren x_i liegen dabei auf einem regulären Gitter mit einem Gitterabstand von 0.1 auf dem Quadrat $[-1, 1] \times [-1, 1]$. Abb. A.1 zeigt die Trainingsdaten.

Die Wavelet-Bibliothek wurde auf Grundlage der soeben zusammengefassten Algorithmen erstellt. Als Konstanten wurden $\sigma = 2$ und $\beta = 1$ gewählt. Der Parameter M wurde mit Hilfe der Standard-Akaike-FPE-Methode auf 6 gesetzt und die Bibliothek enthält nach Regressorauswahl 67 Wavelets. Das Resultat ist in Abb. A.2 gezeigt. Die Wurzel des mittleren quadratischen Fehlers (empirisches Risiko)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\boldsymbol{x}_i) - y_i\right)^2}$$

beträgt ca. 0.008. Die Funktion sieht augenscheinlich glatt aus und ist offensichtlich eine gute Näherung an die Trainingsdaten, bis auf die vier markanten Spitzen. Diese vier Spitzen sind wie beschrieben ein Ausdruck schlechter Kondition bei der Modellbildung. Abb. A.3 zeigt das Ergebnis nach 50 Iterationen des robusten Levenberg-Marquardt-Algorithmus. Es ist zu erkennen, dass sich die Glattheit der Kurve etwas verbessert hat. RMSE ≈ 0.002, d.h. dieses Kriterium hat sich noch einmal deutlich verbessert. Zwei der Spikes sind fast ganz zurück gegangen, aber die anderen beiden sind geblieben. Auch bei noch mehr Iterationsschritten ändert sich daran grundsätzlich nichts. Grund hierfür ist die in Abschnitt 4.5 beschriebene nicht-gleichmäßige Konvergenz des Schätzers gegen die wahre Funktion.

Abb. A.4 und A.5 zeigen die Ergebnisse der robusten Algorithmen. Der mittlere Fehler ist auf RMSE ≈ 0.002 gesunken.

Beispiel 2
Beispielproblem 2 zeigt die Approximation der Funktion $f(x_1, x_2) = y = \sin(x_1^2)\cos x_2$, wobei ein Trainingsdatensatz $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_{500}, y_{500})\}$ mit normalverteilten Störungen als Realisierungen von fünf Zufallsvariablen $X_1, ..., X_5$ gewählt wurden, $\boldsymbol{x}_1, ..., \boldsymbol{x}_{100}$ als Realisierungen von X_1, $\boldsymbol{x}_{101}, ..., \boldsymbol{x}_{200}$ als Realisierungen von X_2 usw. X_1 ist hierbei $N((0,0), \mathbb{1})$-verteilt, $X_2 \sim N((0,1), \mathbb{1})$, $X_3 \sim N((0,-1), \mathbb{1})$, $X_4 \sim N((1,0), \mathbb{1})$ und $X_5 \sim N((-1,0), \mathbb{1})$. ($\mathbb{1}$ bezeichnet in diesem Fall die 2×2-Einheitsmatrix). Die Fehlerterme $E_1, ..., E_{500}$ sind ebenfalls normalverteilt: $E_i \sim N\left((0,0), 0.1^2\mathbb{1}\right)$. Abb. A.6 zeigt

die hieraus resultierende Punktwolke.
Abb. A.7 zeigt das Ergebnis der Approximation mittels des entwickelten Wavelet Netzwerkes. Um das Modellierungsproblem sichtbar zu machen wurden reine Standardverfahren angewendet. Der erwartete Bias-Fehler des Problems liegt bei 0.1, der berechnete mittlere quadratische Fehler in der gezeigten Approximation beträgt 0.12. Die Approximation ist bezüglich dieses Maßes also durchaus gut.
Die Wavelet-Bibliothek wurde auf Grundlage der soeben zusammengefassten Algorithmen erstellt. Als Konstanten wurden $\sigma = 2$, $\beta = 1$ gewählt. Der Parameter M wurde mit Hilfe der Standard-Akaike-FPE-Methode auf 45 gesetzt. Je näher wir nun dem Minimum des empirischen Risikos kommen (also je höher n gewählt wird), desto schlimmer wird die Spitzen-Bildung, wie Abb. A.8 zeigt. Das Optimierungsproblem ist offensichtlich schlecht konditioniert. Hier wurden 50 Levenberg-Marquardt-Schritte verwendet, was den mittleren quadratischen Fehler (empirisches Risiko) nochmals um 50% senkt. Die maximale Spitzen-Höhe erhöht sich allerdings auf -40.
Abb. A.9, A.10 und A.11 zeigen die Ergebnisse der robusten Algorithmen. Der mittlere Fehler ist auf RMSE ≈ 0.1 gesunken.

Beispiel 3
Dieses Beispiel approximiert

$$f(x_1, x_2) = \begin{cases} 0 & \text{für} \quad x_1 + x_2 \leq 0 \\ 1 & \text{sonst} \end{cases}$$

Die 441 Vektoren $\boldsymbol{x}_i$ liegen dabei genau wie in Problem 1 auf einem regulären Gitter mit einem Gitterabstand von 0.1 auf dem Quadrat $[-1, 1] \times [-1, 1]$. Abb. A.12 zeigt die Trainingsdaten. Auf Grundlage der Trainingsdaten wurde, genau wie in Problem 1 und 2, eine Wavelet-Bibliothek mit den vorgegebenen Konstanten $\sigma = 2$, $\beta = 1$ und 6 nodes erzeugt. Genau wie in Problem 1 und 2 erfolgte danach eine Vorwärtsauswahl von Regressoren. Dabei wurden 68 Wavelets ausgewählt. Das Resultat zeigt Abb. A.13. Die Wurzel des mittleren quadratischen Fehlers ergibt sich zu RMSE ≈ 0.032. Mit einem besonders guten Resultat war bei diesem Problem nicht zu rechnen, weil f eine Sprungfunktion ist und $\hat{f}_{\mathcal{T}}$ immer eine stetige Funktion. Das Resultat ist deutlich schlechter als erhofft. Man kann extremes Spiking an der Sprungkante beobachten. Auf den Halbebenen auf denen f kostant ist, wird die Funktion durch das Wavelet Netzwerk $\hat{f}_{\mathcal{T}}$ erwartungsgemäß gut beschrieben.
Führen wir jetzt wieder, genau wie in Problem 1 beschrieben, 50 Iterationen des Levenberg-Marquqrdt-Algorithmus aus, so erhalten wir das Ergebnis aus Abb. A.14. RMSE ≈ 0.01. Damit hat sich die Anpassung an die Zielfunktion um Faktor 3 verbessert, das grundlegende Problem der nicht-gleichmäßigen Konvergenz ist allerdings wie erwartet geblieben.
Abb. A.15 und A.16 zeigen die Ergebnisse der robusten Algorithmen. Der mittlere Fehler ist auf RMSE ≈ 0.006 gesunken.

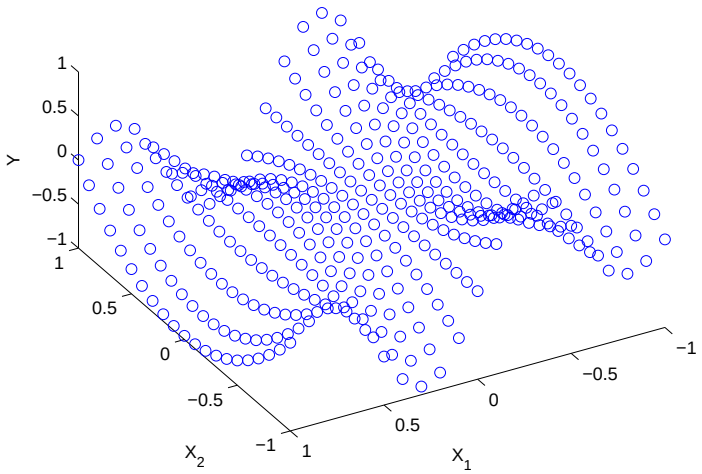


Abb. A.1. Die Trainingsdaten von Problem 1

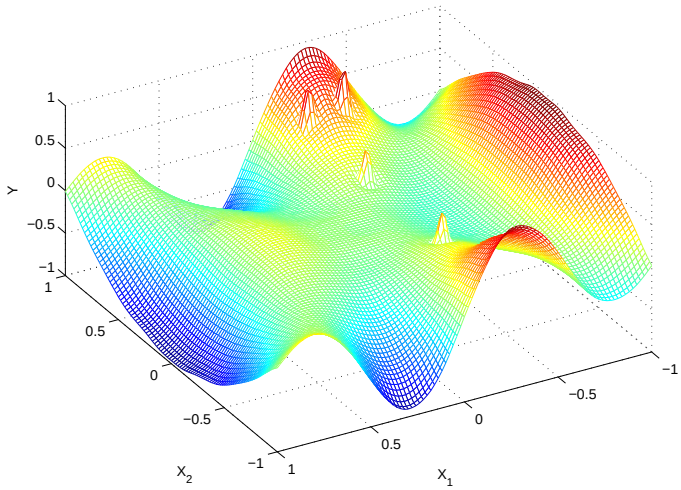


Abb. A.2. Problem 1: $\hat{f}_{\mathcal{I}}$ nach Standard-Vorwärtsregressorauswahl.

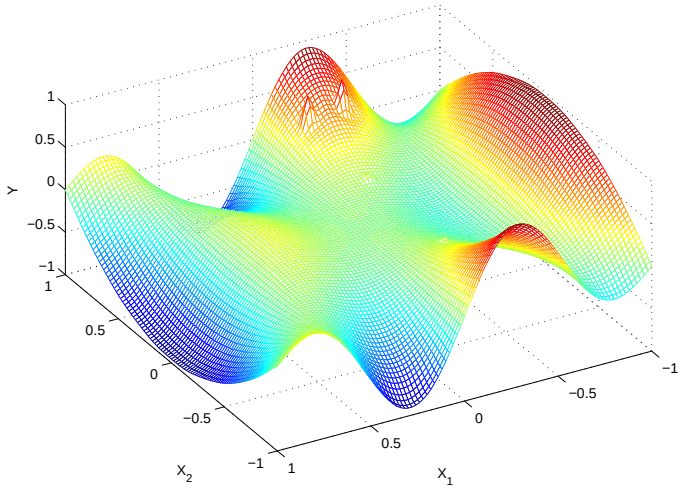


Abb. A.3. Problem 1: $\hat{f}_{\mathcal{T}}$ nach Standard-Vorwärtsregressorauswahl und 50 Levenberg-Marquardt-Schritten

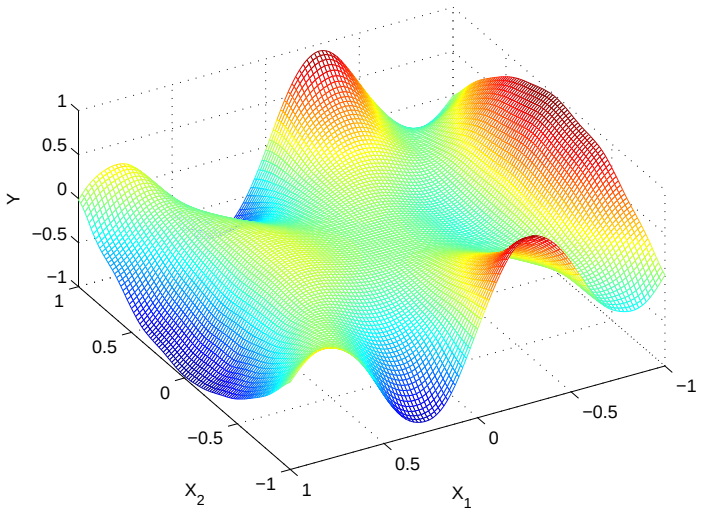


Abb. A.4. Problem 1: $\hat{f}_{\mathcal{T}}$ nach robuster Vorwärtsregressorauswahl.

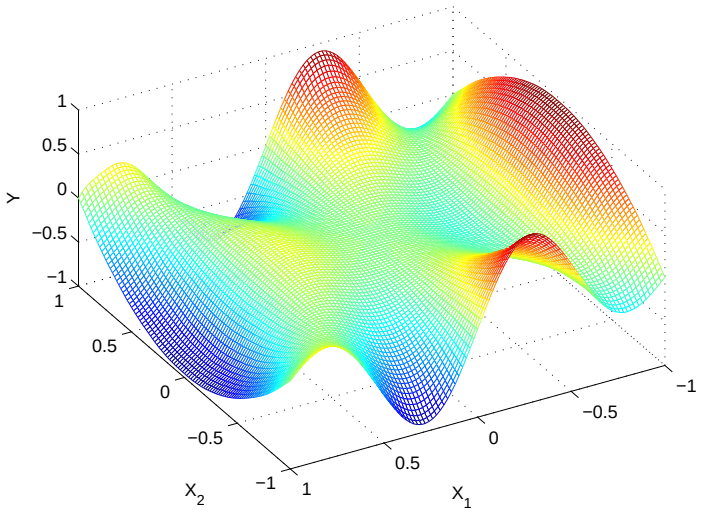


Abb. A.5. Problem 1: $\hat{f}_{\mathcal{T}}$ nach robuster Vorwärtsregressorauswahl, Rückwärtselemination und 50 robusten Levenberg-Marquardt-Schritten.

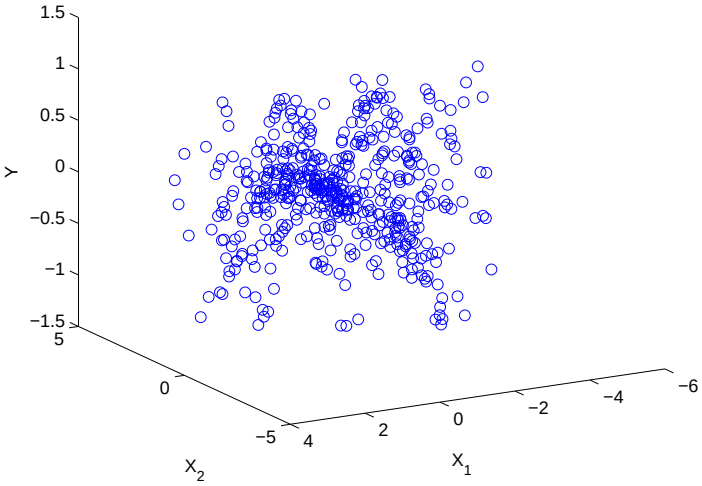


Abb. A.6. Trainingsdaten von Problem 2 mit normalverteilten Störungen.

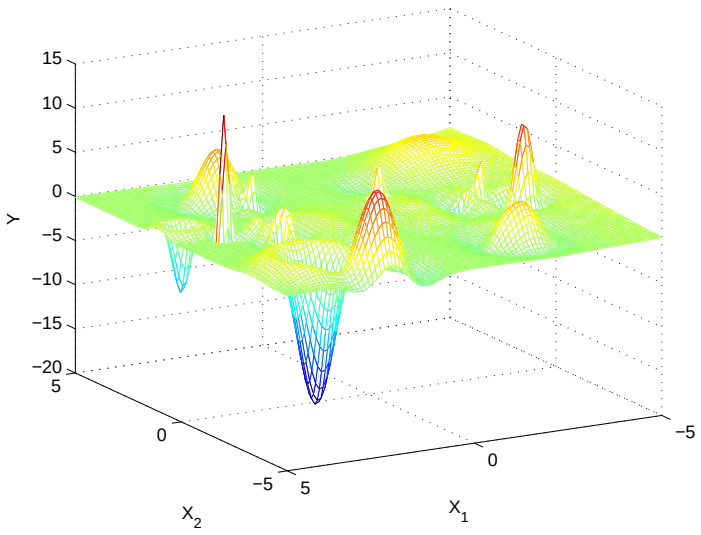


Abb. A.7. Problem 2: Die Approximation durch das Wavelet-Netzwerk bei Standard-Vorwärtsregressorauswahl.

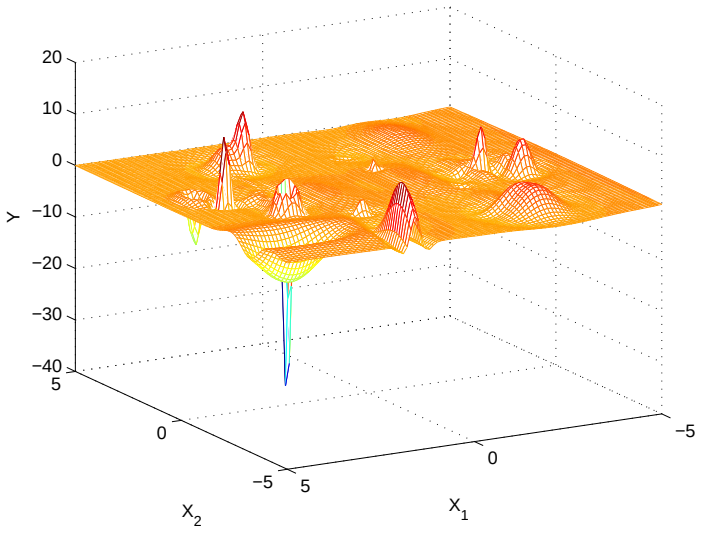


Abb. A.8. Problem 2: Die Approximation durch das Wavelet-Netzwerk bei Standard-Vorwärtsregressorauswahl und 50 Levenberg-Marquardt-Schritten.

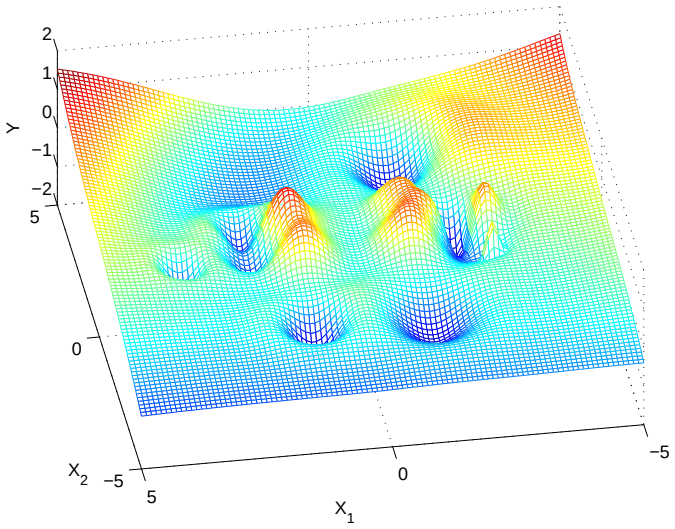


Abb. A.9. Problem 2: $\hat{f}_{\mathcal{T}}$ nach robuster Vorwärtsregressorauswahl und Rückwärtselimination.

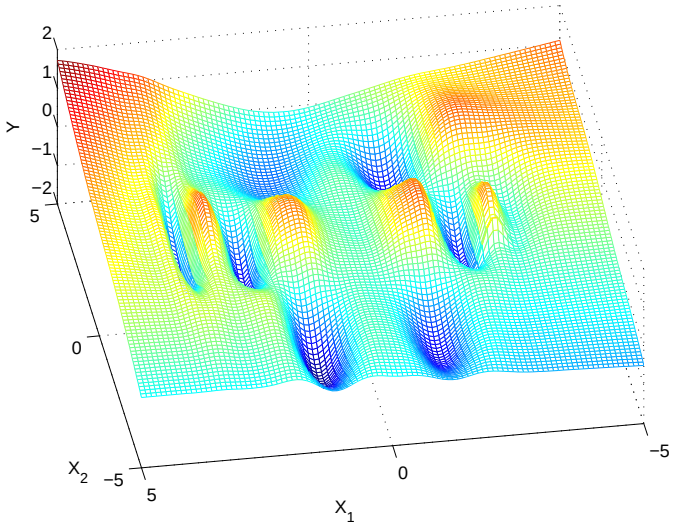


Abb. A.10. Problem 2: $\hat{f}_{\mathcal{T}}$ nach robuster Vorwärtsregressorauswahl, Rückwärtselimination und 50 verbesserten Levenberg-Marquardt-Schritten.

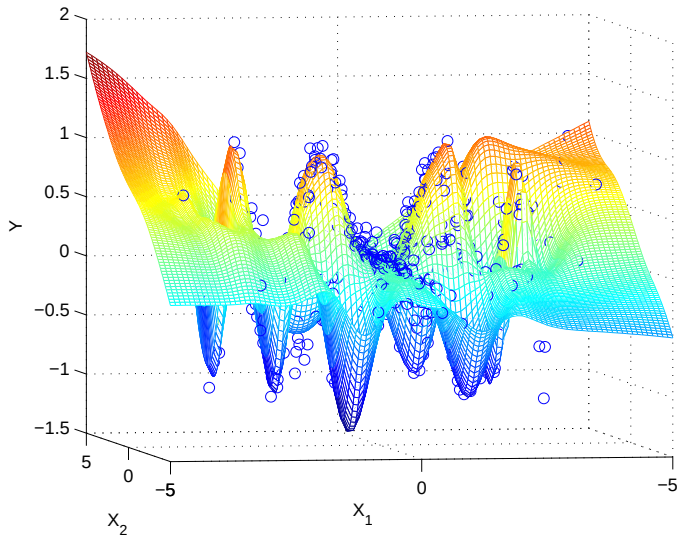


Abb. A.11. Problem 2: $\hat{f}_{\mathcal{T}}$ nach robuster Vorwärtsregressorauswahl, Rückwärtselimination und 50 verbesserten Levenberg-Marquardt-Schritten mit Trainingsdaten.

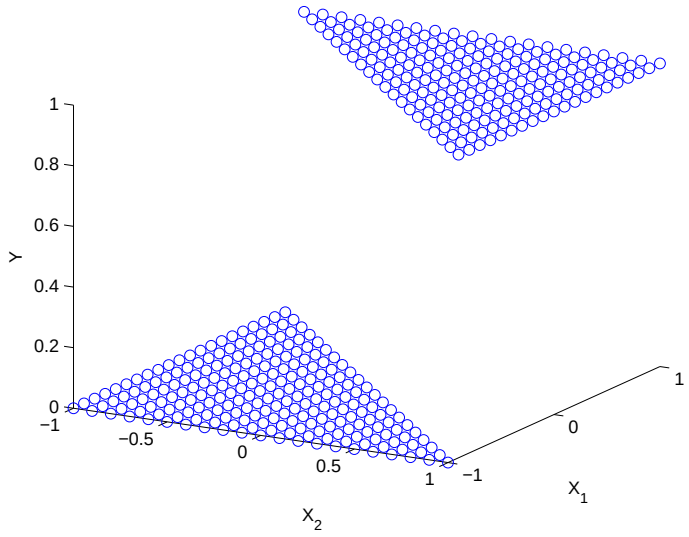


Abb. A.12. Trainingsdaten von Problem 3

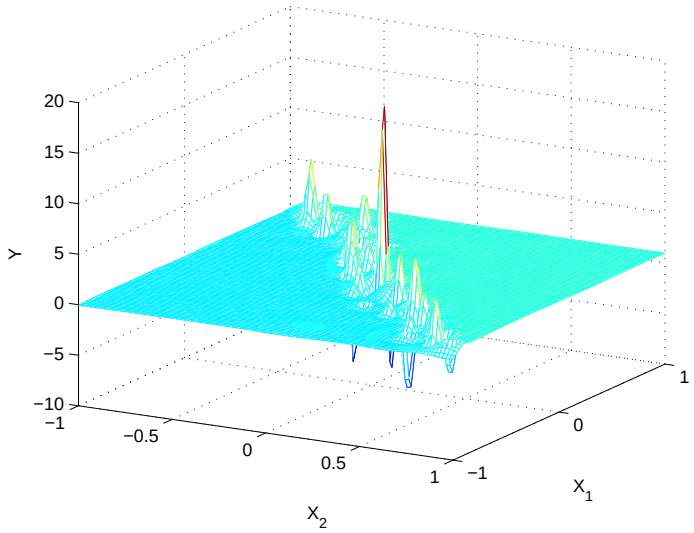


Abb. A.13. Problem 3: $\hat{f}_{\mathcal{T}}$ nach Standard-Vorwärtsregressorauswahl.

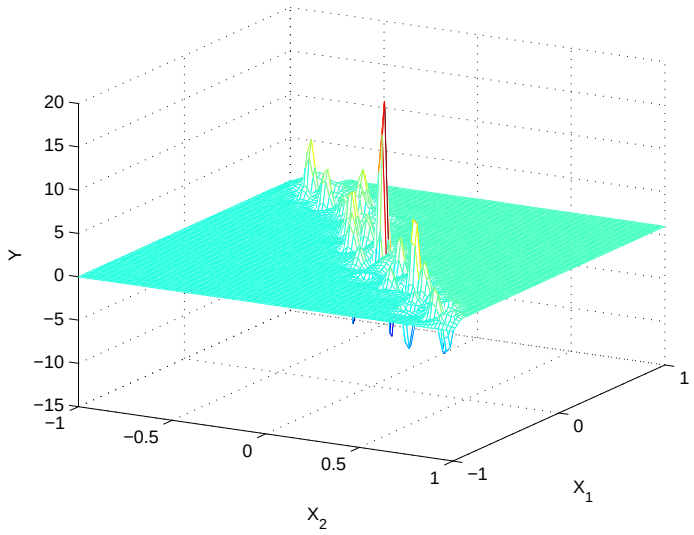


Abb. A.14. Problem 3: $\hat{f}_{\mathcal{T}}$ nach Standard-Vorwärtsregressorauswahl und 50 Levenberg-Marquardt-Schritten

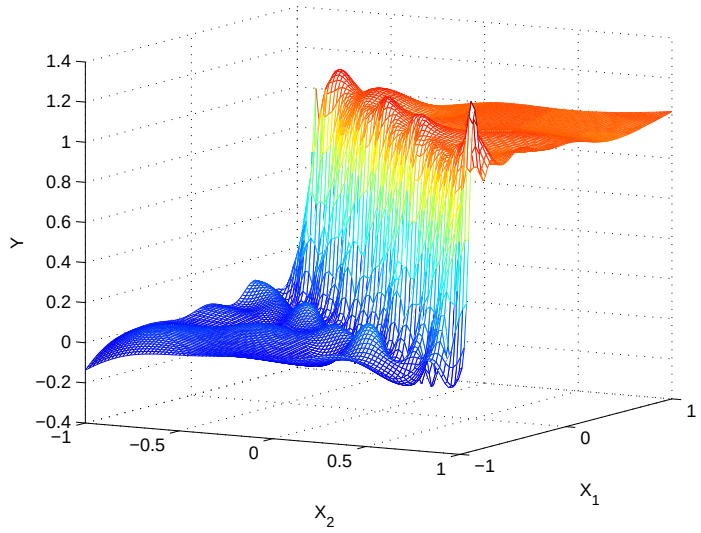


Abb. A.15. Problem 3: $\hat{f}_{\mathcal{T}}$ nach robuster Vorwärtsregressorauswahl und Rückwärtselimination

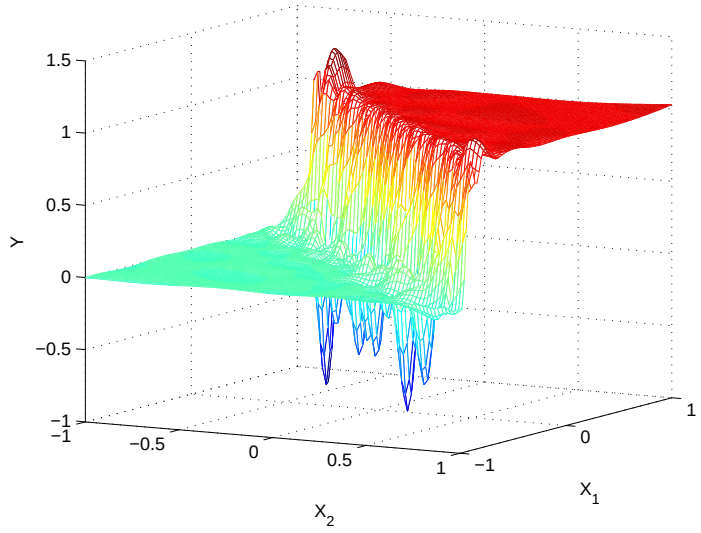


Abb. A.16. Problem 3: $\hat{f}_{\mathcal{T}}$ nach robuster Vorwärtsregressorauswahl, Rückwärtselimination und 50 verbesserten Levenberg-Marquardt-Schritten.

Literaturverzeichnis

[Akaike, H. (1973)] *Information theory and an extension of the maximum likelihood principle.* Proc. of the Second International Symposium on Information Theory Budapest: Akademiai Kiado, 267-281, 1973

[Amann, H., Escher, J. (2002)] *Analysis I.* 2. Auflage, Birkhäuser Verlag, Basel, 2002

[Amann, H., Escher, J. (2001)] *Analysis III.* Birkhäuser Verlag, Basel, 2001

[Anders, U., Korn, O. (1999)] *Model selection in neural networks.* Neural Networks, 12:309-323, 1999

[Anthony, M., Bartlett, P.L. (1999)] *Neural Network Learning: Theoretic Foundations.* Cambridge University Press, Cambridge

[Aronszajn, N. (1950)] *Theory of reproducing kernels.* Trans. AMS, 68:337-404, 1950

[Barron, A. (1993)] : *Universal approximation bounds for superpositions of a sigmoidal function.* IEEE Trans. Information Theory, 39:930-945, 1993

[Barron, A., Cohen, A., Dahmen, W., Devore, R. (2008)] *Approximation and Learning by greedy Algorithms.* Annals of Statistics, 36:64-94, 2008

[Bartlett, P.L., Maiorov, V. & Meir, R. (1998)] *Almost linear VC dimension bounds for piecewise polynomial networks.* Neural Computation, 5:45-60

[Bartlett, P.L. (1998)] *The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network.* IEEE Trans. Information Theory, 44:525-536, 1998

[Blatter, C. (1998)] : *Wavelets - Eine Enführung.* Advanced Lectures in Mathematics, Vieweg, Braunschweig, 1998

[Böhm, C., Faloutsos, C., Pan, J-Y., Plant, C. (2006)] : *Robust Information-theoretic Clustering.* Research Track Paper, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006

[Bors, A.G., Pitas, I. (2001)] : *Robust RBF Network.* Howlett R.J, Jain, L.C. (Hrsg.): Radial basis function neural networks: design and applications, Physica-Verlag, Heidelberg, 125-153, 2001

[Bousquet, O., Elisseeff, A. (2002)] *Stability and Generalization.* Journal of Machine Learning Research, 2:499-526, 2002

[Breiman, L. (1993)] : *Hinging hyperplanes for regression, classification and function approximation.* IEEE Trans. Information Theory, 39:999-1013, 1993

[Burger, M., Hofinger, A. (2005)] *Regularized Greedy Algorithms for Neural Network Training with Data Noise.* Computing, 74:1-22, 2005

[Carroll, S.M., Dickinson, B.W. (1989)] : *Construction of Neural Nets using the Radon-Transform*. Proc. IJCNN, Vol. 1, 607-611, 1989

[Castro, J.L., Mantas, C.J., Benítez, J.M. (2000)] *Neural networks with a continuous squashing function in the output are universal approximators*. Neural Networks, 13:561-563, 2000

[Chandra, P., Singh, Y. (2003)] *Fault tolerance of feedforward artificial neural networks - A framework of study*. Proc. IJCNN, 489-494, 2003

[Chandra, P., Singh, Y. (2004)] *Feedforward Sigmoidal Networks-Equicontinuity and Fault-Tolerance Properties*. IEEE Trans. Neural Networks, 15:1350-1366, 2004

[Cherkassky, V., Shao, X., Mulier, F., Vapnik, V. (1999)] *Model Complexity Control for Regression Using VC Generalization Bounds*. IEEE Trans. Neural Networks, 10:1075-1089, 1999

[Cherkassky, V., Shao, X. (2001)] *Signal estimation and denoising using VC-Theory*. Neural Networks, 14:37-52, 2001

[Chiu, C.T., Mehrotra, K., Mohan, C.K., Ranka, S. (1994)] : *Training techniques to obtain fault tolerant neural networks*. Proc. Int. Symp. Fault Tolerant Computing, Autin, Texas, 360-369, 1994

[Craven, P., Wahba, G. (1979)] *Smoothing Noisy Data with Spline Functions*. Numer. Math., 31:377-403, 1979

[Cucker, F., Smale, S. (2001)] *On the Mathematical Foundations of Learning*. Bulletin AMS, 39:1-49, 2001

[Cybenko, G. (1989)] : *Approximation by superpositions of a sigmoidal function*. Mathematics of Control, Signals, and Systems, 2:303-314, 1989

[Daubechies, I. (1990)] *The Wavelet Transform, Time-Fequency Localization and Signal Analysis*. IEEE Trans. Information Theory, 36:961-1005, 1990

[Daubechies, I. (1992)] *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, 1992

[Deuflhard, P., Hohmann, A. (2002)] : *Numerische Mathematik I*. de Gruyter-Verlag, Berlin, 2002

[Delyon, B., Juditsky, A., Benveniste, A. (1995)] : *Accuracy Analysis for Wavelet Approximations*. IEEE Trans. Neural Networks, 6:332-348, 1995

[Devroye, L.P., Wagner, T.J. (1979)] *Distribution-free performance bounds for potential function rules*. IEEE Trans. Information Theory, 25:601-604, 1979

[Dieudonné, J. (1969)] : *Foundations of Modern Analysis*. Academic Press, New York, 1969

[Donoho, D., Johnstone, I. (1994)] *Ideal spatial adaptation by wavelet shrinkage*. Biometrika, 81:425-455, 1994

[Emmerson, M.D., Damper, R.I. (1993)] *Determining and improving the fault tolerance of multilayer perceptrons in a pattern recognition application*. IEEE Trans. Neural Networks, 4:788-793, 1993

[Engl, H., Hanke, M., Neubauer, A. (1999)] *Regularization of Inverse Problems*. SIAM Review, 41:386-387, 1999

[Gao, J., Chen, F., Shi, D. (2004)] *On the Construction of Support Wavelet Network*. IEEE Intern. Conf. Systems, Man and Cybernetics, 4:3204-3207, 2004

[Girosi, F. (1998)] *An equivalence between sparse approximation and Support Vector Maschines*. Neural Computation, 7:219-269, 1998

[Grenander, U. (1951)] *On empirical spectral analysis of stochastic processes*. Ark. Math., 1:503-531, 1951

[Ham, F.M., Kostanic, I. (2001)] *Principles of Neurocomputing for Science and Engineering*. McGraw-Hill, Singapore, 2001

[Hammadi, N.C., Ito, H. (1997)] : *A learning algorithm for fault tolerant feedforward networks*. Proc. IEICE Trans. Information Systems, Vol. E80-D, 21-27, 1997

[Hammadi, N.C., Ito, H. (1998)] : *On the activation function and fault tolerance in feedforward artificial neural networks*. Proc. IEICE Trans. Information Systems, Vol. E81-D, 66-72, 1998

[Haussler, D. (1992)] *Decision theoretic generalizations of the PAC model for neural net and other learning applications*. Information and Computation, 100:78-150

[Hecht-Nielsen, R. (1990)] *Theory of the back-propagation Neural Network*. Caudill, M. (Hrsg.): International Joint Conference on Neural Networks, Volume I, Washington, DC, 593-601, 1990

[Higgins, J. (1985)] : *Five Short Stories about the Cardinal Series*. Bull. AMS, 12:45-89, 1985

[Hoeffding, W. (1963)] : *Probability inequalities for sums of bounded random variables*. Journal of the American Statistical Association, 58:13–30, 1963

[Holmstrom, L., Koistinen, P. (1982)] : *Using additive noise in backpropagation training*. IEEE Trans. Neural Networks, 3:24-28, 1992

[Hornik, K., Stinchcombe, M., & White, H. (1989)] : *Multilayer feedforward networks are universal approximators*. Neural Networks, 2(5):359-366, 1989

[Huber, P. (1985)] : *Projection Pursuit*. Annals of Statistics, 13:435-475, 1985

[Jarre, F., Stoer, J. (2003)] : *Optimierung*. Springer Verlag, Berlin Heidelberg, 2003

[Juditsky, A., Hjalmarsson, H., Benveniste, A., Deylon, B. et al (1995)] *Nonlinear black box models in system identication*. Mathematical Foundations. Automatica., 1995

[Karandikar, R.L., Vidyasagar, M. (2002)] : *Rates of uniform convergence of empirical means with mixing processes*. Statistics & Probability Letters, 58:297-307, 2002

[Karpinski, M. & Maxintyre, A. (1997)] *Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks*. Journal of Computer and System Sciences, 54:169-176

[Kearns, M., Ron, D. (1999)] *Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation*. Neural Computation, 11:1427-1453, 1999

[Kingman, J.F. (1973)] : *Subadditive ergodic theory*. Annals of Probability, 1:883–909, 1973

[Koiran, P. (1994)] : *Efficient Learning of Continuous Neural Networks*. Proc. Seventh Annual Conference on Computational Learning Theory, New Brunswick, New Jersey, 348-355, 1994

[Koiran, P., Sontag, E.D. (1997)] *Neural Networks with quadratic VC-Dimension*. Journal of Computer and System Sciences, 54:190-198

[Lee, W.S., Bartlett, P.L., Williamson, R.C. (1995)] : *Efficient Agnostic Learning of Neural Networks with Bounded Fan-in*. 6th Australian Conference on Neural Networks, Sydney, 6.-8. Februar, 1995

[Li, S.T., Leiss, E.L. (2001)] : *On noise-immune RBF networks*. Radial basis function networks 1: recent developments in theory and applications, Physica Verlag Rudolf Liebing KG, 95-124, 2001

[Ljung, L. (1978)] : *Convergence Analysis of Parametric Identification Models*. IEEE Trans. Automatic Control, 23:770-783, 1978

[Ljung, L. (1987)] : *System Identification - Theory for the User*. Prenctice Hall Inc., New Jersey, 1987

[Lugosi, G. (2003)] : *Concentration-of-measure inequalities*. Lecture Notes, http://www.econ.upf.es/ lugosi/anu.pdf, 2003

[Lugosi, G., Zeger, K. (1995)] : *Nonparametric Estimation via Empirical Risk Minimization*. IEEE Trans. Information Theory, 41:677-687, 1995

[Lusin, N. (1912)] : *Sur les propriétés des fonctions mesurables*. Comptes Rendus Acad. Sci. Paris, 154:1688-1690, 1912

[Maass, W. (1994)] *Neural nets with super-linear VC-dimension*. Neural Computation, 6:877-884, 1994

[Maass, W., Schmitt, M. (1999)] *On the complexity of learning for spiking neurons with temporal coding*. Information and Computation, 153:26-46

[Magoulas, G.D., Vrahatis, M.N., Androulakis, G.S. (1999)] *Improving the Convergence of the Backpropagation Algorithm Using Learning Rate Adaptation Methods*. Neural Computation, 11:1769-1796

[Matasuoka, K. (1992)] : *Noise injection into inputs in backpropagation learning*. IEEE Trans. Systems, Man, Cybernetics, 22:436-440, 1992

[McDiarmid, C. (1989)] *On the Method of Bounded Differences*. Surveys of Combinatorics, Cambridge University Press, Cambridge, 148-188, 1989

[McKeown, J., Stella, F., Hall, G. (1997)] : *Some Numerical Aspects of the Training Problem for Feed-Forward Neural Nets*. Neural Networks, 10:1455-1463, 1997

[Meintrup, D., Schäffler, S. (2005)] *Stochastik - Theorie und Anwendungen*. Springer-Verlag, Berlin Heidelberg, 2005

[Melody, J. (1999)] : *On Universal Approximation Using Neural Networks*. Technical Report ECE 480, 1999

[Meyer, Y. (1992)] : *Wavelets and Operators*. Cambridge University Press, 1992. [Englische Version von *Ondelettes et opérateurs*. Hermann, 1990]

[Minnix, J.I. (1991)] : *Fault tolerance of backpropagation neural networks trained with noisy data*. Proc. IJCNN, Vol. 1, 703-708, 1991

[Murray, A.F., Edwards, P.J. (1994)] : *Synaptic weight noise during MLP training: Enhanced MLP performance and fault tolerance resulting from synaptic noise during training*. IEEE Trans. Neural Networks, 5:792-802, 1994

[Nolan, D., Pollard, D. (1987)] : *U-processes: Rates of convergence*. Annals of Statistics, 15:780-799, 1987

[Nowacki, H. (1990)] *Mathematische Verfahren zum Glätten von Kurven und Flächen*. Encarnacao, J.L., Hoschek, J., Rix, J. (Hrsg.): Geometrische Verfahren der Graphischen Datenverarbeitung, Springer, 22-45, 1990

[Pan, G-F., He, P., Zhou, Y-T., Li, J-H. (2007)] : *Constructing a Wavelet-Bases RKHS and its associated Scaling Kernel for Support Vector Approximation*. Proc. Intern. Conf. Wavelet Analysis and Pattern Recognition, Beijing, China, 1403-1407, 2007

[Park, J., Sandberg, W. (1991)] : *Universal approximation using radial-basis-functions networks*. Neural Computation, 3:246-257, 1991

[Phatak, D., Koren, I. (1995)] : *Complete and Partial Fault Tolerance of Feedforward Neural Nets*. IEEE Trans. Neural Networks, 6:446-456, 1995

[Pohl, D. (2007)] : *Robuste Wavelet Netzwerke*. Diplomarbeit am Fachbereich für Elektrotechnik und Informationstechnik der Universität der Bundeswehr München, 2007, betreut von Martin Prescher, Kontakt: daniel.pohl@unibw.de oder mail@martinprescher.de

[Pollard, D. (1990)] : *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward, CA, Alexandria, VA, 1990

[Pucar, P. & Sjöberg, J. (1995)] : *On the hinge finding algorithm for hinging hyperplanes*. Technical Report LiTH-ISY-R-1720, Dep. of EE, Linköping University, 1995

[Riedmüller, B., Ritter, K. (1992)] : *Lineare und quadratische Optimierung*. Schriftenreihe des Instituts für Angewandte Mathematik und Statistik der Technischen Universität München, 1992

[Rivals, I., Personnaz, L. (2003)] : *Neural-Network Construction and Selection in Nonlinear Modeling*. IEEE Trans. Neural Networks, 14:804-819, 2003

[Rivals, I., Personnaz, L. (2004)] : *Jacobian ConditioningAnalysis for Model Validation*. Neural Computation, 16:401-418, 2004

[Rossi, F., Conan-Guez, B. (2005)] : *Functional Multi-Layer Perceptron: a Nonlinear Tool for Functional Data Analysis*. Neural Networks, 18:45-60, 2005

[Rousseeuw, P.J., Leroy, A.M. (1987)] : *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, 1987

[Rousseeuw, P.J., Bassett Jr., G.W. (1991)] : *Robustness of the p-subset algorithm for regression with high breakdown point*. Stahel, W., Weisberg, S. (Hrsg.): Directions in Robust Statistics and Diagnotsics Part II, Springer-Verlag, New York, 185-194, 1991

[Rudin, W. (1976)] : *Functional Analysis*. McGraw-Hill, St. Louis, 1976

[Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986)] : *Learning representations by back-propagating errors*. Nature, 323:533-536, 1986

[Saarinen, S., Bramley, R., Cybenko, G. (1993)] : *Ill-Conditioning in Neural Network Training Problems*. SIAM Journal on Scientific Computing, 14:693-714, 1993

[Sakurai, A. (1993)] *Tighter bounds on the VC-dimension of three-layer networks*. Proc. World Congress on Neural Networks, Vol. 3, Erlbaum, Hillsdale, New Jersey, 540-543

[Schaback, R., Werner, H. (1992)] : *Numerische Mathematik*. Springer-Verlag, Berlin, 1992

[Schwarz, G. (1978)] : *Estimating the Dimension of a Model*. Annals of Statistics, 2:461-464, 1978

[Schwarz, H.R. (1997)] : *Numerische Mathematik*. B.G. Teubner Stuttgart, Stuttgart, 1997

[Selonen, A., Lampinen, J., Ikonen, L. (1996)] : *Using External Knowledge in Neural Network Models*. Casasent, D.P. (Hrsg.): Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling, Proc. SPIE 2904, 239-249, 1996

[Séquin, C.H., Clay, R.D. (1990)] *Fault Tolerance in Artificial Neural Networks*. Proc. IJCNN, Vol. 1, 703-708, 1990

[Sjöberg, J. (1995)] : *Non-Linear System Identification with Neural Networks*. Dissertation, University of Linköpings Tryckeri AB 1995.28, 1995

[Slepian, D., Pollak, H. (1962)] *Prolate spheroidal wave functions, Fourier analysis and uncertainty*. Teil I: Bell. Syst. Tech. Journal, 40:43-64, 1961; Teil II: Landau, H., Pollak, H., Bell. Syst. Techn. Journal, 40:65-84, 1961; Teil III: Landau, H., Pollak, H., Bell. Syst. Techn. Journal, 41:1295-1336, 1962

[Smale, S., Zhou, D-X. (2003)] : *Estimating the Approximation Error in Learning Theory*. Annals of Applied Probability (Singap.) 1, 1:17-41, 2003

[Smola, A., Schölkopf, B. (1998)] *On a kernel-based method for pattern recognition, regression, approximation and operator inversion*. Algorithmica, 22:211-231, 1998

[Steele, J.M. (1978)] *Empirical Discrepancies And Subadditive Processes*. Annals of Probability, 6:118-127, 1978

[Steele, J.M. (1986)] *An Efron-Stein inequality for nonsymmetric statistics*. Annals of Statistics, 14:753-758, 1986

[Stein, E., Shakarchi, R. (2003)] *Fourier Analysis*. Princeton University Press, Princeton and Oxford, 2003

[Stinchcombe, M., White, H. (1990)] : *Approximating and Learning Unknown Mappings Using Multilayer Feedforward Networks with Bounded Weights*. Proc. IJCNN, Vol. III, 7-16, 1990

[Talagrand, M. (1987)] *The Glivenko–Cantelli problem*. Annals of Probability, Annals of Probability, 15:837–870, 1987

[Tikhonov, A.N., Arsenin, V.Y. (1977)] : *Solution of Ill-posed Problems*. Winston & Sons, Washington, 1977

[Vapnik, V.N., Chervonenkis, A.J. (1981)] : *Necessary and sufficient conditions for the uniform convergence of means to their expectations*. Theory Probab. Appl. 26:532–553

[Vapnik, V.N., Chervonenkis, A.J. (1991)] : *The necessary and sufficient conditions for consistency in the empirical risk minimization method*. Pattern Recognition and Image Analysis, 1:284-305, 1991

[Vapnik, V.N. (1998)] : *An Overview of Statistical Learning Theory*. IEEE Trans. Neural Networks, 10:988-999, 1999
[Vapnik, V.N. (1999)] : *Statistical Learning Theory*. Wiley, New York, 1998
[Watson, G. (1969)] : *Smooth regression analysis*. Sankhya Series, A(26):359-372, 1969
[Werner, D. (2006)] : *Einführung in die höhere Analysis*. Springer-Verlag, Heidelberg, 2006
[Werner, D. (2007)] : *Funktionalanalysis*. Springer-Verlag, Heidelberg, 2007
[White, H. (1989)] : *Learning in Artificial Neural Networks: A Statistical Perspective*. Neural Computation, 1:425-464, 1989
[Wiener, N. (1932)] : *Tauberian Theorem*. Annals of Mathematics 33, 1–100, 1932
[Williamson, R.C., Smola, A., Schölkopf, B. (1998)] : *Generalization performance of regularization networks and support vector maschines via entropy numbers of compact operators*. Tech. Report NC2-TR-1998-019, NeuroCOLT2, 1998
[Yu, X., Onder Efe, M., Kaynak, O. (2002)] : *A General Backpropagation Algorithm for Feedforward Neural Networks Learning*. IEEE Trans. Neural Networks, 13:251-259, 2002
[Zhang, J., Walter, G.G., Miao, Y., Lee, W. (1995)] : *Wavelet Neural Networks for Function Learning*. IEEE Trans. Sign. Proc., 43:1485-1497, 1995
[Zhang, Q. (1994)] : *Using Wavelet Networks in Nonparametric Estimation*. Technical Report 833, IRISA
[Zhang, Q., Benveniste, A. (1992)] : *Wavelet Networks*. IEEE Trans. Neural Networks, 3:889-898, 1992

Danksagung

Mein größter Dank gilt ohne Zweifel meinem Doktorvater und Mentor Professor Dr. Dr. Stefan Schäffler. Seine überragende fachliche Kompetenz gepaart mit seiner Menschlichkeit werden mir mein Leben lang ein Vorbild sein.
Weiterhin bedanke ich mich besonders bei Professor Dr. Albert Gilg. Ohne seine Unterstützung wäre diese Dissertation nicht möglich gewesen. Ich freue mich auf eine erfolgreiche Zusammenarbeit in der Zukunft. In diesem Zusammenhang danke ich auch für die finanzielle Unterstützung der Siemens AG und speziell dem Fachzentrum Corporate Technology PP2 und Prof. Dr. Jörg Schulze für viele Anregungen und eine spannende und lehrreiche Zeit im Zuge der gemeinsam durchgeführten Projekte.
Wichtiger Dank gebührt auch meinen Zweitgutachter Prof. Dr. Jochen Schein und dem Vorsitzenden Prof. Dr. Dieter Gerling für ihre Unterstützung.

Besonderer Dank für eine schöne Zeit und viele interessante und anregende Gespräche gilt den Mitarbeitern und Besuchern am Lehrstuhl für Mathematik & Operations Research an der Universität der Bundeswehr: Daniela, F.D., Harald, Ina, Katharina, Mathias, Michael, Robert, Sabine, Steffi und allen, die ich unabsichtlich vergessen habe. Zudem danke ich Leutnant Daniel Pohl für seinen Scharfsinn. Ich freue mich auch in Zukunft auf eine angenehme und erfolgreiche Zusammenarbeit.

Völlig selbstverständlich gebührt unendlicher Dank meinen Freunden und besonders meiner Familie. Ich kann Euch leider an dieser Stelle nicht einzeln danken ohne jeden Rahmen zu sprengen.