Kapitel 1: Einleitung

1.1 Überblick

Neuronale Netzwerke — deren Dynamik die vorliegende Arbeit untersucht — stellen eine völlig neue Art der Datenverarbeitung dar. Alle bislang gebräuchlichen Computer sind Derivate der von Neumannschen Rechnerarchitektur: ein einziges komplexes Rechenwerk, gegebenenfalls verbunden mit diversen Speichermedien. Die Datenverarbeitung geschieht in solchen Rechnern vorwiegend seriell; ein Programmschritt wird nach dem anderen abgearbeitet. Dies gilt selbst, im übertragenen Sinne, für so ausgefeilte Multiprozessorsysteme wie die CRAY Y–MP und ähnliche Computer.

Neuronale Netzwerke zeichnen sich hingegen durch eine drastische Abkehr von dieser traditionellen Rechnerarchitektur aus: hier interagieren sehr viele, aber einfachste Prozessoreinheiten, die durch synaptische Kopplungen miteinander verbunden sind. Im Schema der synaptischen Kopplungen ist sowohl die Wissensbasis als auch die Art der Datenverarbeitung festgelegt. Obwohl das einzelne Prozessorelement dabei nur bescheidene algorithmische Fähigkeiten besitzt, findet durch die große Anzahl der Neuronen ein Qualitätssprung statt, der das Neuronale Netz als Ganzes in die Lage versetzt, Datenverarbeitungsaufgaben zu lösen, an denen ein einzelnes Prozessorelement (oder auch eine CRAY Y–MP) kläglich scheitern würde.

Erste, erfolgversprechende Arbeiten auf dem Gebiet der Neuronalen Netze wurden schon in den sechziger Jahren unternommen [Ro62, Ca61, Bl62], doch erst durch die Arbeiten von W. Little [Li74, Li78] und J.J. Hopfield [Ho82] erlebte die Theorie Neuronaler Netze Anfang der achtziger Jahre einen neuen, unerwarteten Aufschwung. Hopfield stellte eine Analogie zwischen Neuronalen Netzwerken und Modellsystemen von Spingläsern her, die es ermöglichte, mächtige mathematische Werkzeuge der Spinglas–Physik auf Neuronale Netzwerke anzuwenden. Die Erfolge waren anfänglich beeindruckend; so konnte u.a. die maximale Speicherkapazität von Neuronalen Netzwerken mit den Methoden der statistischen Physik bestimmt werden [Am85, Ka87, Ga88].

Die auf diese Weise in den letzten Jahren gewonnenen Erkenntnisse haben allerdings einen entscheidenden Nachteil: es sind Aussagen, die sich nur auf die *statischen* Eigenschaften der Netzwerke beziehen. Versuche, die Dynamik, also die eigentliche Datenverarbeitung der Neuronalen Netze zu klären, waren wenig erfolgreich. Für eines der einfachsten Netzwerke, das Hopfieldmodell, konnte die Dynamik lediglich *zwei* Zeitschritte weit exakt berechnet werden [Ga87]. Darüberhinaus ist man bislang auf möglicherweise wenig aussagekräftige, numerische Simulationen angewiesen [Fo88, Kr90].

Ziel der vorliegenden Arbeit ist es , einen erweiterten Einblick in die Art und Weise zu geben, wie ein Neuronales Netzwerk durch kollektive Interaktion seiner Neuronen eine bestimmte Datenverarbeitungsaufgabe löst.

1.2 Definition des Modells

Neuronale Netzwerke sind ein wesentlicher Bestandteil aller höherentwickelten Lebewesen. Diese Netzwerke, die fast alle Steuerungs- und Regelungsaufgaben

innerhalb des Organismus übernehmen, sind im allgemeinen äußerst komplex. Daneben gibt es noch die abstrakten Modelle Neuronaler Netzwerke, mit denen wir uns in dieser Arbeit beschäftigen werden.

Diese Modelle neuronaler Netzwerke sind heute weit davon entfernt, auch nur annähernd der biologischen Realität zu entsprechen. Sie können allenfalls als extrem vereinfachte Modelle biologischer Netzwerke interpretiert werden, was jedoch in zweierlei Hinsicht keinen Nachteil darstellt. So zeigen selbst extrem abstrahierte Neuronale Netzwerke beachtliche Datenverarbeitungseigenschaften, die unmittelbar zu technischer Nutzung führen können. Gerade bei technischen Anwendungen ist man nicht an schwierig zu implementierenden, komplexen Algorithmen interessiert. Ferner scheint das Funktionieren einfachster Modellsysteme Neuronaler Netze die Vermutung nahezulegen, daß viele der komplexen biochemischen Prozesse, die in biologischen Nervenzellen stattfinden, für die eigentlichen Denkprozesse nicht unbedingt relevant sind.

1.2.1 Die Prozessoreinheiten

Die biologische Nervenzelle*

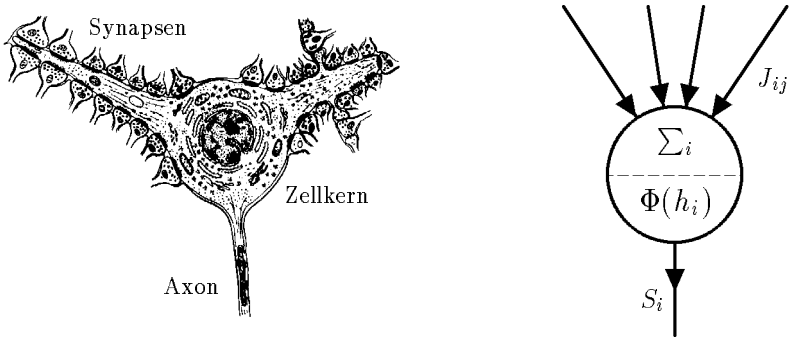
Biologische Nervenzellen bilden die Vorlage für die formalen Neuronen heutiger Modelle Neuronaler Netzwerke. Die biologischen Nervenzellen zeigen ein äußerst differenziertes Erscheinungsbild, denn sie sind vielfach den speziellen Prozessen, welche sie innerhalb der Organismen steuern, optimal angepaßt. Aus diesem Grunde unterscheiden sich beispielsweise Neuronen in der Retina des Auges erheblich von Motoneuronen, welche den Bewegungsapparat steuern; trotzdem sind bei biologischen Nervenzellen funktionelle Gemeinsamkeiten festzustellen, die in ein formalisiertes Modellneuron Eingang finden können.

Nervenzellen besitzen, wie alle Zellen, einen Zellkern mit den dazugehörigen Substrukturen, welche für den Stoffwechsel der Zelle verantwortlich sind und deren Energieversorgung aufrechterhalten. Der auffallendste Unterschied gegenüber normalen Körperzellen sind vom Zellkörper der Neuronen wegführende wurzelartige Fortsätze, die Dendriten und eine einzige, lange Zellfaser, das Axon. Die Dendriten sind, neben dem Zellkörper, Rezeptoren für Signale von anderen Nervenzellen, während das Axon zur Übertragung der Information hin zu anderen Nervenzellen dient. Die Verbindungstelle zwischen zwei Nervenzelle ist die Synapse. Dies sind kleine Plättchen, die an den Enden des sich in mehrere Fortsätze verzweigenden Axons sitzen. Die Synapsen haften an den Dendriten und am Zellkörper anderer Neuronen (Abbildung 1.1, links).

Formen und Größen von Nervenzellen können sehr unterschiedlich sein, und nicht alle Nervenzellen besitzen Dendriten. Im menschlichen Nervensystem werden Axonlängen zwischen wenigen Mikrometern und weit über einem Meter gefunden.

Die Datenverarbeitung biologischer Nervenzellen läuft auf elektro–chemischer Basis ab. Ionenpumpen erhalten im Ruhezustand der Nervenzelle im Protoplasma, dem Innern der Nervenzelle, ein negatives Potential von etwa -70mV gegenüber der Umgebung aufrecht. Signale, die an den Synapsen eintreffen, veranlassen diese, eine große Anzahl von chemischen Neurotransmittern (beispielsweise Acetylcholin) auszuschütten, die zu einer vorübergehenden Depolarisation der Nervenzelle führen können. Übersteigt diese Depolarisation eine gewisse Schwelle von etwa -50mV, entsteht ein kurzer, positiver Impuls von $+30\text{mV}$.

*Die Informationen dieses Abschnittes wurden im wesentlichen aus [Ku84, Ka86, Sch87] zusammengestellt.



biologische Nervenzelle formalisiertes Neuron

Abbildung 1.1: Auf einer rudimentären Ebene besitzen biologische Nervenzellen (links, aus [KA86]) und formalisierte Neuronen die gleiche Funktionalität.

Dieser Aktionspotential genannte Impuls hat einen charakteristischen zeitlichen Verlauf und breitet sich sofort längs des Axons der Nervenzelle als lokalisierte, solitäre Welle aus: das Neuron feuert.

Neben den eben beschriebenen exzitatorischen Synapsen gibt es auch inhibitorische Synapsen, welche das Neuron daran hindern, einen Impuls auszusenden. Man unterscheidet dabei zwei verschiedene Mechanismen der Inhibition. Bei der postsynaptischen Hemmung haften die inhibitorischen Synapsen ebenso wie die exzitatorischen Synapsen am Zellkörper bzw. an den Dendriten des Neurons. Ob eine Synapse inhibitorisch oder exzitatorisch wirkt, wird dabei nicht durch unterschiedliche Neurotransmitter, sondern durch die speziellen Eigenschaften der subsynaptischen Membran, also des Teils der Nervenzelle, an der die Synapse anliegt, entschieden. Dabei sitzen die hemmenden Synapsen meistens in der Nähe des Axons, da dort die inhibitorische Wirkung am größten ist.

Eine funktionell ganz andere Art der Inhibition ist die präsynaptische Hemmung. Hier liegen die inhibitorischen Synapsen nicht am Neuron, sondern an exzitatorischen Synapsen an und hindern diese gegebenenfalls am Ausschütten von Neurotransmittern. Dieses Abblocken von synaptischen Signalen findet man allerdings im Zentralnervensystem von Wirbeltieren eher selten.

Im Prinzip kann schon die Einwirkung einer einzigen exzitatorischen Synapse zur Depolarisation und dem anschließenden 'Feuern' des Neurons führen, aber normalerweise sind dazu mehrere gleichzeitig aktive Synapsen erforderlich. Der Zellkörper der Nervenzelle mit seinen Dendriten wirkt als eine Art von Summationsvorrichtung, der alle einkommenden Signale, exzitatorische und inhibitorsche, aufsummiert. Erst wenn der kombinierte Einfluß mehrerer aktiver Synapsen auf die Zelle überwiegt und dabei die Aktivitätsschwelle von -50mV überschritten wird, feuert diese.

Es ist anzumerken, daß das Aussenden eines einzelnen Impulses entlang des Axons in biologischen Netzwerken, von primitiven Nervensystemen einmal abgesehen, kein echtes Signal darstellt. Erst mehrere, zeitlich kurz hintereinanderfolgende Impulse sind ein echtes Signal. Dabei kodiert offensichtlich die Frequenz der Impulse die Stärke des ausgesendeten Signals; biologische Neuronen sind also, obwohl die Entscheidung zum Aussenden eines einzelnen Impulses eine Ja/Nein–Entscheidung ist, keine binären Schaltelemente, sondern eher analoge

Abbildung 1.2: Eine sigmoidale Übertragungsfunktion verknüpft internes Feld und Zustandsvariable $S_i = \Phi(h)$ der formalen Neuronen. Die Übertragungsfunktion modelliert dabei unterschiedliche Aspekte des Verhaltens biologischer Neuronen, je nachdem, ob von einer binären oder kontinuierlichen Zustandsvariablen S_i des Modellneurons ausgegangen wird (normale oder kursive Schrift).

Prozessoren mit der Möglichkeit einer differenzierten Antwort auf die eintreffenden Informationen.

Das formale Neuron

Formale Neuronen können in verschiedenster Komplexität definieren werden. Sowohl vom technischen als auch vom erkenntnistheoretischen Standpunkt (wie weit darf man die Neuronen vereinfachen, um noch genügend Funktionalität zu behalten?) erscheinen aber gerade die extrem vereinfachten Modelle am interessantesten.

Die formalisierten Neuronen, mit denen wir uns im folgenden beschäftigen werden, wurden im wesentlichen schon in den vierziger Jahren dieses Jahrhunderts von W. McCulloch and W. Pitts definiert [Mc43]. Sie besitzen eine Signalvariable S_i, die bei einfachen McCulloch–Pitts–Neuronen die diskreten Werte ± 1 annimmt (ursprünglich 0 und 1, in Anlehnung an die boolsche Algebra). Sie zeigt damit das Feuern $(+1)$ oder Nichtfeuern (-1) des Neurons an. Man kann aber auch kontinuierliche Werte für die Signalvariable S_i zulassen, typischerweise mit $S_i \in [-1, +1]$. Die Zustandsvariable S_i ist dann ein Maß für die Aktivität des Neurons, wobei (-1) zum Ruhezustand des Neurons mit dem Aussenden von keinen oder nur wenigen Impulsen, $(+1)$ zur maximal möglichen Aktivität des Neurons mit dicht aufeinanderfolgenden Neuronenimpulsen korrespondiert. In beiden Fällen ist S_i eine nichtlineare Funktion

$$S_i(t+1) = \Phi\big(h_i(t)\big)$$

des postsynaptischen Potentials oder internen Feldes h_i. Dieses wird durch

$$h_i(t) = \sum_i J_{ij} S_j(t) \ ,$$

der Summe der mit den synaptischen Kopplungen J_{ij} gewichteten Signale S_j von anderen Neuronen, erzeugt.

Die Nichtlinearität der Übertragungsfunktion $\Phi(h)$ ist dabei ein essentieller Bestandteil des Modellneurons. Eine lineare Übertragungsfunktion würde zu einem trivialen Verhalten des Neuronalen Netzwerkes führen. Bei kontinuierlichen Signalvariablen S_i wählt man als Übertragungsfunktion eine abgeflachte Stufenfunktion, beispielsweise $\Phi(h) = \tanh \beta h$, wodurch automatisch der Wert der Signalvariable S_i auf Werte zwischen -1 und $+1$ begrenzt wird. Grob wird durch diese sigmoidale Funktion eine Ansprechschwelle, ein mehr oder weniger linearer Zwischenbereich und die Sättigung biologischer Neuronen modelliert. Bei binären McCulloch–Pitts–Neuronen hat $\Phi(h)$ die Form einer Stufenfunktion $\Phi(h) = \mathrm{sign}(h)$, wodurch direkt Ruhepotential, Schwellwert und Aktionspotential biologischer Neuronen simuliert werden (Abbildung 1.2). Wir beschränken uns im folgenden auf diese binären McCulloch–Pitts–Neuronen.

1.2.2 Universalität der formalen Neuronen

Universelle Informationsverarbeitung kann als Fähigkeit definiert werden, durch die internen Verarbeitungsmechanismen beliebige Übertragungsfunktionen zwischen den Eingangs– und Ausgangssignalen des Netzwerkes zu realisieren. Es stellt sich die Frage, ob mittels der formalen Neuronen überhaupt eine solche universelle Datenverarbeitung möglich ist. Bei zellulären Automaten findet man beispielsweise vier verschiedene Klassen von Automaten, unter denen aber lediglich eine zu universellen Berechnungen in der Lage ist [Wo84].

Wenn unsere primitiven formalen Neuronen zur universellen Datenverarbeitung geeignet sein sollen, müssen sie zumindest die drei elementaren logischen Funktionen, NEGATION, UND– sowie ODER–Verknüpfung, darstellen können. Aus diesen ist es dann bekanntermaßen möglich, durch entsprechende Verknüpfungen

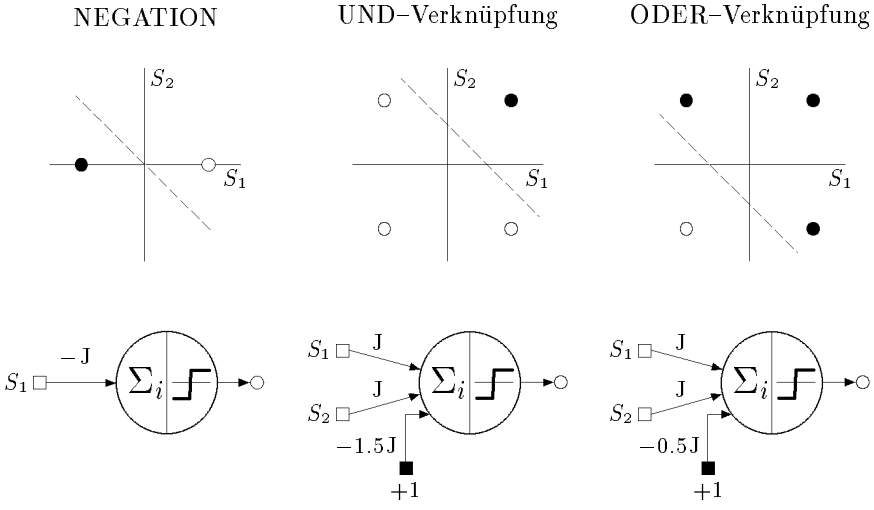


Abbildung 1.3: Durch entsprechende Wahl der Kopplungen sind Modellneuronen in der Lage, NEGATION, UND– sowie ODER–Verknüpfungen zu realisieren. Aus ihnen lassen sich dann beliebige binäre Funktionen aufbauen.

beliebige binäre Funktionen zu erzeugen.

Nun kann ein einzelnes, formales Neuron in der Tat alle drei elementaren logischen Verknüpfungen realisieren. Dazu sind lediglich die synaptischen Kopplungen J_{ij} entsprechend zu wählen (vergl. Abbildung 1.3, unten). Die drei elementaren Logikfunktionen gehören nämlich alle zu den sogenannten linear separablen Funktionen, bei denen im Raum der Eingangszustände eine Ebene bzw. eine Hyperebene genügt, um die Bereiche mit aktiven und passiven Ausgangszuständen abzugrenzen. In Abbildung 1.3, oben, ist dies für die drei elementaren Logikfunktionen exemplarisch im Falle zweier Eingangssignale S_1 und S_2 dargestellt. Die aktiven Ausgangszustände des Neurons werden durch dunkle, die passiven durch weiße Punkte angezeigt. Offensichtlich lassen sich beide Bereiche durch eine Hyperebene, in diesem Fall durch die gestrichelt dargestellte Linie, trennen.

Unsere formalen Neuronen können *nur* linear separabele Funktionen realisieren. Sie sind damit allerdings sehr viel universeller als herkömmliche Logikbausteine. Dazu ein Beispiel: Soll erkannt werden, ob von 20 Eingangsleitungen mindestens 10 aktiv sind, so sind mit herkömmlicher Technologie, also normalen Logikbausteinen, bei zwei Schichten etwa 10^5 UND–Gatter und ein ODER–Gatter erforderlich [Vo86]. Da dieses Problem allerdings ein linear separabeles Problem ist, genügt ein *einziges* unserer formalen Neuronen, um die gleiche Aufgabe zu lösen!

1.2.3 Die Topologie von Neuronalen Netzen

Natürlich existieren auch binäre Funktionen, die nicht zu der Klasse der linear separabelen Funktionen gehören. Beschränkt man sich auf zwei Eingangsleitungen, so sind von den dann möglichen, insgesamt $2^3 = 8$ verschiedenen binären Funktionen zwei, die EXKLUSIV–ODER–Verknüpfung und die ÄQUIVALENZ–Operation, nicht linear separabel, also auch nicht durch ein einziges Neuron darstellbar [Mi69]. Diese Tatsache führte Ende der sechziger Jahre zu einem dramatischen, aber wohl etwas voreiligen Halt der damaligen Neuronalen Netzwerk–Forschung.

Man kann nicht erwarten, daß ein einzelnes Neuron zu universellen Berechnungen in der Lage ist. Wie bei herkömmlicher Computerhardware, wo sehr viele elementare Logikbausteinen in einer komlexen Struktur miteinander wechselwirken, muß man für universelle Berechnungen auch in Neuronalen Netzwerken zu komplexeren Topologien übergehen.

Den einfachen Perceptron–Typ eines Neuronalen Netzes (Abbildung 1.4.A) findet man deshalb in technischen Anwendungen nur noch selten. Aber bereits eine zusätzliche Zwischenschicht (Abbildung 1.4.B) ermöglicht die Realisation *beliebiger* Übertragungsfunktionen. Dies folgt zwanglos aus der Darstellbarkeit jeder binären Funktion als disjunktive Normalform — dies ist eine ODER–Verknüpfung von vielen UND-Verknüpfungen der Eingangssignale — und der Universalität unserer formalen Neuronen. Mehrschichtnetzwerke mit mehr als einer Zwischenschicht sind an sich nicht nötig, doch können Netze mit mehreren Zwischenschichten zu einer Verringerung der Anzahl der benötigten Neuronen führen.

Sowohl ein einfaches Perceptron als auch Mehrschichtnetzwerke kondensieren Information von vielen Eingangskanälen auf wenige Ausgangskanäle. Strukturen mit solcher Funktionalität findet man in biologischen Nervensystemen typischerweise in der Vorverarbeitung von Sinneseindrücken. So komprimieren in der Retina des Auges weit über 100 Millionen Nervenzellen ihre Informatio-

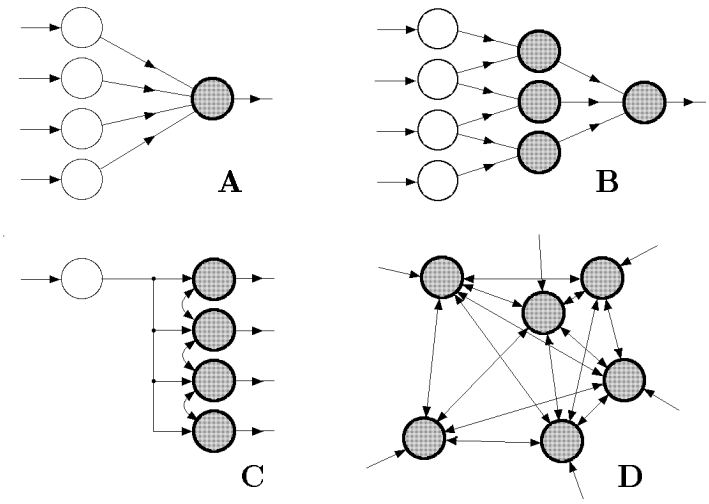


Abbildung 1.4: Die wesentlichen Topologien Neuronaler Netzwerke. Dabei ist zwischen einfachen Input–/Output–Neuronen und wirklichen datenverarbeitenden Neuronen (dunkel dargestellt) zu unterscheiden. Neben dem einfachen Perceptron (A) und dem Mehrschicht–Netzwerk (B) sind noch die Kohonen–Karte (C) und die autoassoziative Netzwerke (D) bekannt.

nen über die Welt auf die nur etwa 1 Million Nervenfasern im Sehnerv. Diese komprimierten Informationen müssen in höheren Neuronenschichten natürlich wieder aufgeschlüsselt werden, und auch die dazugehörigen Topologien sind aus der Neurophysiologie wohlbekannt; ihr technisches Pendant wird nach einem der Pioniere der Neuronalen Netzwerkforschung als Kohonen–Karte (Abbildung 1.4.C) bezeichnet [Ko84].

Bei der Kohonen–Karte wird ein einkommendes Signal auf eine räumlich ausgedehnte Schicht von Neuronen verteilt, wobei durch laterale, inhibitorische Verbindungen innerhalb der Neuronenschicht erreicht wird, daß nur wenige, im Extremfall nur ein einzelnes Neuron auf den Eingangsreiz ansprechen. Unterschiedliche Eingangssignale werden dadurch auf unterschiedliche räumliche Bereiche der Neuronenschicht abgebildet. Durch eine wohldefinierte *Dynamik der Synapsen* wird eine geordnete Abbildung der Eingangsdaten auf die im allgemeinen zweidimensionale Topologie der Neuronenschicht erreicht. Ähnliche Eingangssignale aktivieren dabei benachbarte Neuronenareale. Solche Abbildungen werden in sehr vielen Bereichen der Großhirnrinde gefunden, typischerweise dort, wo einkommende Sinneseindrücke (Auge, Hör– und Tastsinn) verarbeitet werden. Es sei angemerkt, daß möglicherweise die Datenreduktion in den Sinnesorganen und die anschließend notwendige Reinterpretation der komprimierten Daten in der Großhirnrinde für eine Vielzahl bekannter Sinnestäuschungen verantwortlich ist.

Als vierten Typ von Neuronalen Netzwerken, mit gänzlich anderer Topologie wie die bisher vorgestellten, kennt man die Autoassoziativen Netzwerke (Abbildung 1.4.D). Ein Autoassoziatives Netzwerk besitzt keine vorgegebene Topologie und damit zunächst auch keine vorgegebene Funktionalität. In der Regel wechselwirkt jedes Neuron mit jedem anderen. Dadurch entsteht eine Rückkopplung des Neuronalen Netzes auf sich selbst, was, zusammen mit der nichtlinearen Übertragungsfunktion der Neuronen, zu einem äußerst komplexen dynamischen

Verhalten des Netzwerkes führt. Es ist diese Eigendynamik, welche ein Autoassoziatives Netzwerk deutlich von den anderen elementaren Netzwerk–Topologien unterscheidet. Während Mehrschichtnetzwerke und Kohonen–Karten gewissermaßen einander komplementäre Strukturen mit einfachstem dynamischen Übertragungsverhalten sind, stellen die Autoassoziativen Netzwerke mit ihrer komplexen Funktionalität möglicherweise die Strukturen dar, in denen die eigentlichen Denkprozesse ablaufen. Wir werden uns im Folgenden auf diese vollvernetzten, autoassoziative Netzwerke konzentrieren.

1.3 Autoassoziative Netzwerke

1.3.1 Synchrone und asynchrone Dynamiken

Die Dynamik eines einzelnen unserer formalen Neuronen ist durch

$$S_i(t+1) = \mathrm{sign}\big(\, h_i(t) \,\big) \tag{1.1}$$

festgelegt. Bei Mehrschichtnetzwerken und Kohonen–Karten, die gerichtete Netzwerkstrukturen sind, wird durch deren Topologie letztendlich die Dynamik des gesamten Netzwerkes bestimmt. Neuronen in höheren Schichten dieser Netzwerke können ihren Zustand erst dann bestimmen, wenn die unteren Eingangsschichten ihren Zustand bereits festgelegt haben. Bei den in sich selbst rückgekoppelten Autoassoziativen Netzwerken ist dies nicht der Fall. Hier wird die Reihenfolge wichtig, mit der die Neuronen nach (1.1) ihren neuen Zustand bestimmen. Man unterscheidet prinzipiell zwei verschiedene Dynamiken:

- **asynchrone Dynamik:**
  Bei der asynchronen Dynamik schaltet jeweils ein Neuron nach dem anderen. Die Reihenfolge wird dabei zufällig oder in einem festen Zyklus vorgegeben. Diese Art der Dynamik wurde in numerischen Simulationen zur Spinglas–Physik verwendet und dann auf Neuronale Netzwerke übertragen. Da die Neuronen ihren neuen Zustand nacheinander bestimmen, läßt sich diese Art der Dynamik auf herkömmlichen, seriellen Rechenmaschinen recht einfach implementieren; aber mit Sicherheit arbeiten biologische Neuronale Netze *nicht* seriell.

- **synchrone Dynamik:**
  Bei der synchronen Dynamik schalten alle Neuronen gleichzeitig. Dies ist von einem technologischen Standpunkt aus gesehen sicherlich die attraktivste Dynamik. Zum einen lassen sich synchrone Schaltungen wesentlich einfacher realisieren als asynchrone Schaltungen. Zum anderen läßt sich nur mit synchroner Dynamik die Parallelität Neuronaler Netzwerke voll nutzen.

Sieht man einmal von der asynchronen Dynamik mit zufälliger Schaltreihenfolge ab, sind Autoassoziative Neuronale Netzwerke mit einer endlichen Anzahl von Neuronen endliche zelluläre Automaten; damit endet die Dynamik irgendwann in einem Grenzzyklus. Dies ist eine unmittelbare Folge des begrenzten Zustandraumes des Systems und der deterministischen Dynamik. Geht aber die Anzahl der Neuronen $N \to \infty$, kann sich das System sogar auf chaotischen Trajektorien durch den Zustandsraum bewegen. Damit das Neuronale Netz sinnvoll arbeitet, müssen die synaptischen Kopplungen J_{ij} entsprechend gewählt werden.

1.3.2 Prinzipielle Anwendung

Wie der Name schon nahelegt, werden die vollvernetzten Anordnungen von Neuronen in der Regel als Assoziativ–Speicher eingesetzt. Dabei soll das Netzwerk

aus wenigen, gegebenenfalls verrauschten Hinweisen möglichst die vorher gelernte vollständige Information assoziieren. Eine typische Anwendung findet sich in der Bildrestauration, bei der den Neuronen des autoassoziativem Netzwerkes ein verrauschtes Bild aufgeprägt wird, welches sich dann durch die intrinsische Dynamik des Netzwerkes weiterentwickelt. Im Idealfall wird dabei das zuvor gelernte Muster wiedererkannt.

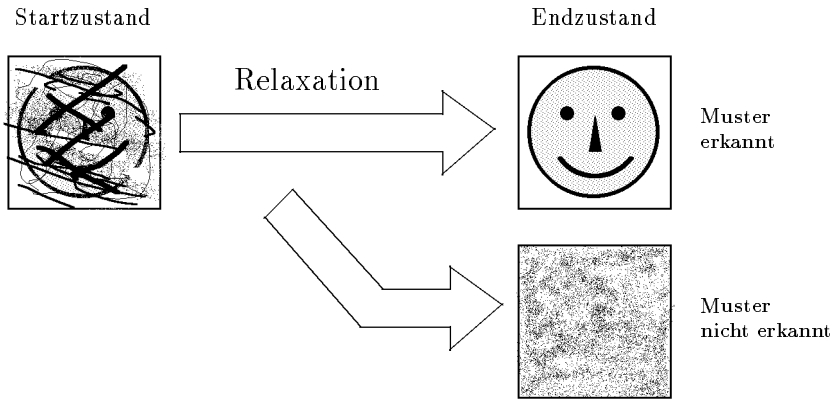


Abbildung 1.5: Verrauschte oder nur teilweise vorhandene Informationen kann ein autoassoziatives Netzwerk wieder restaurieren. Ist der Verrauschungsgrad zu stark, findet allerdings keine Restauration statt.

A priori ist nicht klar, ob ein Neuronales Netzwerk überhaupt in der Lage ist, assoziativ zu arbeiten; dazu ist ein Zugang zur Beschreibung der Dynamik Neuronaler Netzwerke erforderlich. Im wesentlichen kennt man zwei eng miteinander verwandte Analysemethoden, die aber, streng genommen, wegen der komplexen Interaktion der Neuronen in der Assoziationsphase, eine direkte Behandlung der Dynamik des Neuronalen Netzes vermeiden: Lyapunov–Funktionen und Freie Energie–Funktionale.

1.3.3 Lyapunov–Funktionen

Lyapunov–Funktion sind reelle Funktionen der Zustandsvariablen eines Neuronalen Netzwerkes, die während der dynamischen Entwicklung des Systems kontinuierlich ihren Wert veringern. Die Minima der Lyapunov–Funktion sind damit notwendigerweise mit den Endzuständen der Dynamik identisch.

Unter der Voraussetzung symmetrischer Kopplungen, also $J_{ij} = J_{ji}$, ist für die asynchrone Dynamik die Funktion

$$L_a\big[\{S_i(t)\}\big] = -\sum_i S_i(t)h_i(t) = -\sum_{i,j} S_i(t)J_{ij}S_j(t) \tag{1.2}$$

eine Lyapunov–Funktion [Ho82]. In der Tat ändert sich, falls vom Zeitpunkt t zum Zeitpunkt $t+1$ das i–te Neuron seinen Zustand $S_i(t)$ nach Gleichung (1.1) invertiert, $L_a\big[\{S_i(t)\}\big]$ um

$$\Delta L = -2 \cdot S_i(t+1)h_i(t) = -2 \cdot |h_i(t)| \leq 0\,.$$

$L_a\big[\{S_i(t)\}\big]$ nimmt also während der Dynamik des Neuronalen Netzes nie zu. Schließt man Nullfelder, also interne Felder mit $h_i(t) = 0$, aus, gilt sogar streng

$\Delta L < 0$. Damit kann die Dynamik nur in einem Zustand enden, in dem keines der Neuronen seinen Zustand ändert — andernfalls müßte ja die Lyapunov–Funktion weiter abnehmen. Es folgt, daß, unter der obigen Voraussetzung symmetrischer Kopplungen J_{ij}, bei der asynchronen Dynamik nur zeitlich konstante Bitmuster als Attraktoren der Dynamik auftreten. Für sie gelten die Stabilitätsgleichungen

$$S_i = \mathrm{sign}\Big(\sum_j J_{ij} S_j\Big) \; . \tag{1.3}$$

Grenzzyklen oder die im Limes $N \to \infty$ möglichen chaotischen Attraktoren kommen nicht vor.

Bei der synchronen Dynamik kann eine Lyapunov–Funktion — auch hier wieder unter der Voraussetzung symmetrischer Kopplungen — durch

$$\begin{aligned} L_s\big[\{S_i(t)\}\big] &= -\sum_{i \neq j} S_i(t) J_{ij} S_j(t-1) \\ &= -\sum_i \Big| \sum_{j \neq i} J_{ij} S_j(t) \Big| \end{aligned} \tag{1.4}$$

definiert werden [Go87].

Für die Änderung dieser Funktion nach einem Zeitschritt erhält man

$$L_s\big[\{S_i(t+1)\}\big] - L_s\big[\{S_i(t)\}\big] = -\sum_i (S_i(t+1) - S_i(t-1)) h_i(t) \; . \tag{1.5}$$

Ist $S_i(t+1)$ ungleich $S_i(t-1)$, ergibt sich für das Vorzeichen der Differenz zwischen beiden Zuständen das Vorzeichen von $S_i(t+1)$. Dies ist aber, wegen der Neuronendynamik (1.1), identisch dem Vorzeichen von $h_i(t)$. Damit liefert jeder Summand der rechten Seite, falls er nicht verschwindet, einen negativen Beitrag. $L_s\big[\{S_i(t)\}\big]$ nimmt also während der dynamischen Evolution des Netzwerkes monoton ab, und ist somit, wie behauptet, eine Lyapunov–Funktion.

Gleichung (1.5) liefert auch die möglichen dynamischen Attraktoren; $L_s\big[\{S_i(t)\}\big]$ ändert sich nur dann nicht, falls

$$\forall_i \colon \; S_i(t+1) = S_i(t-1)$$

gilt. Dies ist zunächst bei allen Punktattraktoren der asynchronen Dynamik der Fall, bei denen ja die Werte $S_i(t)$ für alle Neuronen zeitlich konstant sind. Daneben können aber noch Attraktoren auftreten, in denen einige oder alle Neuronen in jedem Zeitschritt ihr Vorzeichen ändern,

$$\ldots = S_i(t+1) = -S_i(t) = S_i(t-1) = \ldots \; .$$

Hier reproduziert sich das Bitmuster also jeden zweiten Zeitschritt, wir haben einen Zweierzyklus vorliegen. Die Zweierzyklen sind eine Besonderheit der synchronen Dynamik [Fr86].

1.3.4 Freie Energien

Einen anderen Zugang zur Dynamik autoassoziativer Netzwerke kann über die statistische Physik erfolgen. Man macht sich dabei zunutze, daß eine entsprechend definierte, stochastische Dynamik für lange Zeiten $t \to \infty$ zu Gleichgewichtszuständen führt, die durch ein Freie Energie–Funktional beschrieben werden können. Dadurch hat man zur Analyse Neuronaler Netzwerke alle Standardmethoden der statistischen Physik zur Verfügung.

Das Prinzip ist recht einfach und beruht auf der Theorie Markovscher Ketten [Pe84]. Man nutzt folgende Tatsache aus: In einem Markovschen System gibt es Übergangswahrschweinlichkeiten $w(I|J)$ für den Übergang von einem Zustand J zu einem Zustand I. Gilt nun für diese Übergangswahrscheinlichkeiten

$$w(I|J) \cdot g(J) = w(J|I) \cdot g(I) \ ,$$

eine Bedingung, die als 'detailed balance' bekannt ist [Me53], so folgt zwingend* für große Zeiten $t \to \infty$, daß das System mit der Wahrscheinlichkeit

$$p(I; t \to \infty) \propto g(I)$$

im Zustand I zu finden ist.

Wählen wir nun für $g(I)$ eine Boltzmann–Verteilung,

$$g(I) = \exp\big(-\beta H(I)\big) \ ,$$

womit ein Hamiltonian $H(I)$ und eine inverse Temperatur $\beta = 1/T$ definiert ist, liegt eine wohlbekannte Situation der statistischen Physik vor: Die von der Zustandssumme

$$Z = \sum_I \exp\big(-\beta H(I)\big) \tag{1.6}$$

abgeleitete Freie Energie

$$F = -T \ln(Z) \tag{1.7}$$

wird thermodynamisches Potential und die Gleichgewichtszustände des Systems korrespondieren zu Minima von F. Damit ist die Untersuchung des Netzwerkverhaltens für lange Zeiten wiederum auf das Studium einer Funktion, in diesem Falle der Freien Energie F, zurückgeführt.

Bei Neuronalen Netzwerken erfüllt — mit entsprechend definierten Hamiltonian — die Glauberdynamik [Gl63] die Forderung nach 'detailed balance'. Bei der Glauber–Dynamik schaltet das Neuron mit internem Feld $h_i(t)$ jeweils mit der Wahrscheinlichkeit

$$p(S_i(t+1)) = \frac{\exp\big(\beta h_i(t) S_i(t+1)\big)}{\exp\big(\beta h_i(t)\big) + \exp\big(-\beta h_i(t)\big)}$$

in den Zustand $S_i(t+1)$. Der Parameter β kontrolliert die Stochastizität des dynamischen Prozesses. Für $\beta \to \infty$ erhalten wir wieder das alte, deterministische Schaltgesetz $S_i = \mathrm{sign}(h_i)$ zurück.

Unter der Voraussetzung symmetrischer Kopplungen, $J_{ij} = J_{ji}$, sind die entsprechenden Hamiltonian für die asynchrone Dynamik [Pe84] durch

$$H_a\big[\{S_i\}\big] = -\frac{1}{2} \sum_{i \neq j} S_i J_{ij} S_j$$

und für die synchrone Dynamik [Fo87] durch

$$H_s\big[\{S_i\}\big] = -\frac{1}{\beta} \sum_i \ln\Big\{2\cosh\Big(\beta \sum_j J_{ij} S_j\Big)\Big\}$$

*Man muß hierbei noch die Ergodizität des Systems voraussetzen: Der Raum der Systemzustände I darf nicht in Bereiche zerfallen, zwischen denen keine Übergänge möglich sind. Gerde bei Neuronalen Netzwerken tritt aber, zumindest im Limes $N \to \infty$, eine Ergodizitätsbrechung auf (vergl. [Am89]).

gegeben. Die Freie Energie folgt natürlich aus

$$F = -\frac{1}{\beta} \ln \Big(\sum_{\{S_i\}} \exp \big(-\beta H_{a/s}[\{S_i\}] \big) \Big) \ .$$

Beide Hamiltonian sind, bei der synchronen Dynamik nur für $\beta \longrightarrow 0$, im wesentlichen mit den korrespondierenden Lyapunov–Funktionen identisch.

Die Freie Energie–Funktionale ermöglichen lediglich die Beschreibung des Langzeitverhaltens Neuronaler Netzwerke. Der dynamische Prozess der Assoziation wird ja durch Betrachtungen im thermodynamischen Gleichgewicht ersetzt. Trotzdem sind die wesentlichen Erkenntnisse über autoassoziative Netzwerke bislang ausschließlich über diesen thermodynamischen Zugang erarbeitet worden.

Die Kopplungsmatrizen

Bei Autoassoziativen Netzwerken sollen die Kopplungen so gewählt werden, daß aus wenigen, verrauschten Hinweisen vollständige Informationen assoziiert werden können. Konkret werden in einem Netz mit N Neuronen p verschiedene Bitmuster $\xi_i^\mu = \pm 1$ (mit $\mu = 1, \ldots, p$ und $i = 1, \ldots, N$) ("Assoziationen" oder "Bilder") so eingespeichert, daß sie Attraktoren der Dynamik werden. Diese Forderung ist synonym damit, daß die Bitmuster mindestens lokale Minima der Lypunov–Funktionen oder der Freien Energie–Funktionale werden bzw. die Stabilitätgleichungen (1.3) erfüllen. Dies läßt sich, bei gegebener Neuronenzahl N, natürlich nicht für beliebig hohe Musterzahlen p erreichen und die Speicherdichte $\alpha = \frac{p}{N}$ ist eine wichtige Kenngröße des Netzwerkes; für jede Kopplungsmatrix gibt es eine kritische Speicherdichte α_c, oberhalb derer ein weiteres Einspeichern von Mustern nicht möglich ist.

Man kennt im wesentlichen drei verschiedene Typen von Kopplungsmatrizen:

- **Die Hopfield–Kopplungsmatrix:**
  Sie ist die klassische Kopplungsmatrix, und gegeben durch

  $$J_{ij} = N^{-1} \sum_\mu \xi_i^\mu \xi_j^\mu \ .$$

  Die einfache analytische Form der Hopfield–Matrix ermöglichte es, mit Mitteln der statistischen Physik eine Reihe interessanter Ergebnisse zu erhalten [Am85]. Der größte Nachteil dieser Kopplungsmatrix liegt in der geringen maximalen Speicherkapazität von $\alpha_c = 0.14$. Trotz der einfachen Gestalt der Kopplungsmatrix konnte die Dynamik des dazugehörigen Neuronalen Netzwerkes lediglich zwei Zeitschritte weit exakt berechnet werden [Ga87].

- **Die Perceptron–Kopplungsmatrix:**
  Herausragendstes Merkmal der Perceptron–Kopplungsmatrix ist ihre wesentlich höhere Speicherkapazität von $\alpha_c = 2$ [Ga88]. Ihr größter Nachteil ist ihre, bei vorgegebenen Mustern, unbekannte analytische Form. Dies verhindert eine analytische Behandlung der Dynamik des Netzwerkes mit dieser Kopplungsmatrix. Die Perceptron–Matrix ist außerdem die numerisch am aufwendigsten zu berechnende Kopplungsmatrix und damit auch für numerische Untersuchungen zur Dynamik Neuronaler Netze eher unattraktiv.

- **Die pseudoinverse Kopplungsmatrix:**

  Die pseudoinverse Kopplungsmatrix [Ko84, Pe86] hat eine etwas geringere maximale Speicherkapazität von $\alpha_c = 1$ [Ka87]. Sie ist, bis auf die fehlende Diagonale, identisch dem Projektionsoperator in den Raum der Muster (siehe Anhang B). Mit der Korrelationsmatrix der Muster,

  $$C_{\mu\nu} = N^{-1} \sum_i \xi_i^\mu \xi_i^\nu \ ,$$

  ergibt sich die pseudoinverse Kopplungsmatrix zu

  $$J_{ij} = N^{-1} \sum_{\mu\nu} \xi_i^\mu C_{\mu\nu}^{-1} \xi_j^\nu \ .$$

  Sie stellt sich damit als eine orthogonalisierte Form der Hopfield–Matrix dar. Die pseudoinverse Kopplungsmatrix besitzt für die Analytik wesentlich günstigere Eigenschaften als Hopfield– oder Perceptron–Matrix (Anhang B). Wichtige dynamische Kenngrößen von pseudoinverser Kopplungsmatrix und Perceptron–Matrix unterscheiden sich nur wenig und das dynamische Verhalten beider Matrizen ist sehr ähnlich (Anhang C).

Sowohl pseudoinverse Kopplungsmatrix als auch die Perceptron–Matrix stabilisieren die gelernten Muster perfekt, während die Hopfield–Matrix stabile Konfigurationen in der Nähe der Muster liefert. Bei allen drei Matrizen werden aber noch, neben den Mustern, eine Vielzahl anderer Bitkombinationen erzeugt, die dynamisch ebenfalls stabil sind [Ba87]. Diese metastabilen Niveaus sind zwar immer nur lokale Minima der Lyapunov–Funktionen, wirken aber während des Assoziationsvorgangs als dynamische Fallen. Wegen ihrer großen Häufigkeit haben die metastabilen Niveaus einen erheblichen Einfluß auf das dynamische Verhalten der Netzwerke.

1.4 Zusammenfassung und Problemstellung

Als einzige der bekannten Netzwerktopologien besitzen Autoassoziative Netzwerke ein eigenständige, interessante Neuronendynamik. Durch die extreme Rückkopplung des Netzwerkes, im Zusammenspiel mit den Nichtlinearitäten der Neuronen, ergibt sich ein äußerst komplexes dynamisches Verhalten.

Die bei Autoassoziativen Netzwerken verwendeten Kopplungsmatrizen erzeugen neben den gelernten Mustern eine große Anzahl von unerwünschten, metastabilen Niveaus als dynamische Attraktoren. Über diese metastabilen Niveaus und ihre Auswirkungen auf den Assoziationsvorgang ist bisher wenig bekannt, nicht zuletzt deswegen, weil die Dynamik Neuronaler Netzwerke bislang nicht genau untersucht werden konnte — dies ist der Gegenstand dieser Arbeit.

Dazu werden wir zunächst Netzwerke bei niedriger Speicherkapazität untersuchen, da hier der Zustandsraum sehr regelmäßig strukturiert ist. Das Aufweichen dieser Struktur bei Erhöhung der Speicherdichte α wird in einem folgenden Kapitel analysiert. Den Hauptteil der Arbeit stellt die Untersuchung des dynamischen Verhaltens des Neuronalen Netzwerkes mit pseudoinverser Kopplungsmatrix dar.