Unüberwachtes Lernen mit informationstheoretischen Kriterien in konnektionistischen Modellen

Dissertation der Fakultät für Physik
der Ludwig-Maximilian-Universität München

vorgelegt von

Lucas Cristobal Parra

San Miguel de Tucuman, Argentinien

18. Oktober 1995
Princeton, USA

Erstgutachter: Professor Paul Tavan

Zweitgutachter: Professor Herbert Spohn

Tag der mündlichen Prüfung: 17. July 1996

Inhaltsangabe

Kapitel 1

Zusammenfassung

Ein konnektionistisches Modell verwirklicht eine Relation zwischen einer Eingabe und einer Ausgabe mittels einer vernetzten Struktur. Lernen in konnektionistischen Modellen geschieht durch Adaptieren der Kopplungskoeffizienten dieser Netzstruktur. Beim unüberwachten Lernen wird dem System die Ausgabe, die im einzelnen erzeugt werden soll, nicht vorgegeben. Statt dessen wird ein Optimalitätskriterium definiert, das keine externe Vorgabe benötigt. In dieser Arbeit werden verschiedene Informationskriterien für unüberwachtes Lernen untersucht.

Zunächst wird das Prinzip der *maximalen Transinformation* im thermodynamisch motivierten Modell der Boltzmann-Maschine betrachtet. Das bekannte Prinzip der *maximalen Entropie* wird dadurch auf ein stochastisches, rückgekoppeltes System angewandt. Für binäre Eingabesignale werden mit dieser Lernregel die Aufgaben der Rekodierung und Datenkompression gelöst. Das vorgestellte Konzept kann auch zur selbständigen oder unüberwachten Klassifikation von kontinuierlichen Signalen verwendet werden. Dies wird an einem realen Problem der medizinischen Diagnose demonstriert. Die Lernregel erzeugt aus beobachteten Symptomen selbständig korrekte Diagnosekategorien.

Weiter wird das Kriterium der maximalen Transinformation bei *minimaler gemeinsamer Information* an der Ausgabe betrachtet. Das resultierende System ist in der Lage, statistisch unabhängige binäre Merkmale der Eingabedaten zu extrahieren. In Experimenten mit binären Eingabesignalen wird Datenkompression und Rekodierung in einer redundanzfreien Darstellung der Eingabeinformation erfolgreich demonstriert. Es wird ein einfaches Modell einer visuellen Retina als einschichtige Boltzmann-Maschine verwirklicht. Die Optimierungskriterien bewirken in diesem Modell die Bildung von räumlichen rezeptiven Feldern, die experimentellen Befunden in biologischen Netzen entsprechen.

Durch ein neues thermodynamisches Modell mit mehr-dimensionalen, kontinuierlichen Zuständen wird versucht, die Beschränkung der Boltzmann-Maschine auf binäre Ausgabesignale aufzuheben. Mit Hilfe der Mittelfeldtheorie wird eine deterministische Dynamik für die Zustandsmittelwerte abgeleitet. Hiermit läßt sich die für den kontinuierlichen Fall aufwendige stochastische Dynamik umge-

hen. Das Kriterium der *minimalen relativen Entropie* kann somit für kontinuierliche Signalverteilungen exakt formuliert werden. Die Konvergenzeigenschaften des neuen Modells werden genauer untersucht und in Experimenten verifiziert. Es wird nachgewiesen, daß das System zur Funktionsapproximation verwendet werden kann. Insbesondere weist es die Fähigkeit auf, stückweise stetige Funktionen zu modellieren.

Zuletzt wird ein neuer Ansatz zum unüberwachten Lernen bei kontinuierlichen Signalverteilungen formuliert: die *Minimierung der gemeinsamen Information bei konstanter Informationsübertragung*, was auch als *informationserhaltende Faktorisierung* bezeichnet werden könnte. Statistische Merkmale werden extrahiert, indem man eine faktorielle Darstellung der gemeinsamen Wahrscheinlichkeitsverteilung der Signale generiert. Zusätzlich garantiert man mit volumenerhaltenden Transformationen perfekte Informationsübertragung.

Es werden verschiedene adaptive, volumenerhaltende Transformationen vorgestellt. Die Nebenbedingung der Informationserhaltung vereinfacht die Faktorisierung. Es genügt die Einzelentropien der Ausgabekoordinaten zu minimieren. Mit Statistik zweiter Ordnung läßt sich ein einfaches und effizientes Minimierungskriterium definieren. Es werden verbesserte Messungen der Entropie vorgestellt, die mittels Kumulanten Statistik höherer Ordnung erfassen. Zusammenfassend läßt sich das Verfahren als eine nichtlineare Komponentenanalyse verstehen, die im Gegensatz zur linearen Hauptkomponentenanalyse Statistik höherer Ordnung und nichtlineare Transformationen verwendet. Diese Faktorisierungsmethode bietet außerdem einen neuen Lösungsansatz zur klassischen Aufgabe der Dichteschätzung. Zuletzt wird mit der vorgeschlagenen Dichteschätzungsmethode das Problem der Neuheitsdetektion bearbeitet. Insbesondere wird das Verfahren auf ein technologisches Problem der Motorausfallsvorhersage angewendet. Weiter ermöglicht die Methode ein neues Verfahren zur Schätzung der bedingten Wahrscheinlichkeit. Die Vielfältigkeit der Anwendungen verdeutlicht, daß hier ein Lösungsansatz für eine grundlegende Aufgabe der informationstheoretischen Signalverarbeitung vorgestellt wird: die Faktorisierung einer Wahrscheinlichkeitsverteilung.

Die Arbeit, die in den Kapiteln 4, 6 und in Teilen von Kapitel 7 vorgestellt wird, ist im wesentlichen in den Veröffentlichungen (Parra, 1995; Parra und Deco, 1995, 1996; Parra, Deco, und Miesbach, 1995; Deco und Parra, 1995) erschienen.

Kapitel 2

Einleitung

Bei der Erforschung von biologisch neuronalen Netzen lassen sich zwei Hauptcharakteristika erkennen: der hohe Grad an Konnektivität zwischen den Neuronen einerseits und die Adaptionsfähigkeit dieser Verknüpfungen andererseits. Obschon sich das Forschungsgebiet der neuronalen Netze weit über diese Grundlagen hinausentwickelt hat, bilden diese beiden Elemente nach wie vor die Basis für lernfähige, konnektionistische Modelle.

Die Funktion eines Netzes hängt von den Kopplungen, d.h. von den synaptischen Verbindungen ab. Der Prozeß der Anpassung dieser Verbindungen wird als Lernen bezeichnet. Die Forschung untersucht diesen Lernprozeß sowie die Struktur von Kopplungen und die Dynamik einzelner Neuronen. Dabei taucht die entscheidende Frage nach den zugrundeliegenden Lernkriterien auf.

Die Entwicklung künstlicher neuronaler Netze oder konnektionistischer Modelle hat sich an den Fragestellungen biologischer Modelle orientiert. Insbesondere der Lernprozeß zur Verarbeitung sensorischer Signale ruft Interesse hervor. Wie ist es möglich, daß sich die noch kaum ausgebildete Verarbeitung visueller Signale, wie beim Neugeborenen, selbständig entwickelt? Lediglich die visuellen Signale, die von der Umwelt aufgenommen werden, bestimmen die zwischen den Neuronen ausgebildeten Kopplungen. Anscheinend wird eine Repräsentation der Umwelt, die erlaubt, Objekte oder Eigenschaften zu erkennen, erst in Abhängigkeit von Signalen der Umgebung erlernt. Solches Lernen - ohne Vorgaben über die Art der Verarbeitung der Signale - wird als *unüberwachtes* Lernen bezeichnet. Welche Dynamik leitet diesen unüberwachten Lernprozeß und welche Kriterien liegen zugrunde? Welche Repräsentation sensorischer Signale erweist sich als sinnvoll?

Zur Beantwortung dieser Fragen wird man sich hier aber nicht auf biologisch motivierte Modelle beschränken. In der Tat wird das grundlegende Problem hier allgemeiner definiert. Die Umgebung manifestiert sich durch ein sensorisches Signal als Eingabe $\mathbf{x}$. Das adaptive System produziert eine Ausgabe $\mathbf{y}$ in Abhängigkeit von Kopplungsparameter $\mathbf{w}$ (siehe Abbildung 2.1). Unüberwachtes Lernen bedeutet, daß dem Lernprozeß nicht explizit vorgegeben wird, welche Ausgabe $\mathbf{y}$ einer Eingabe $\mathbf{x}$ zugeordnet werden soll. Statt dessen verwendet man Optima-

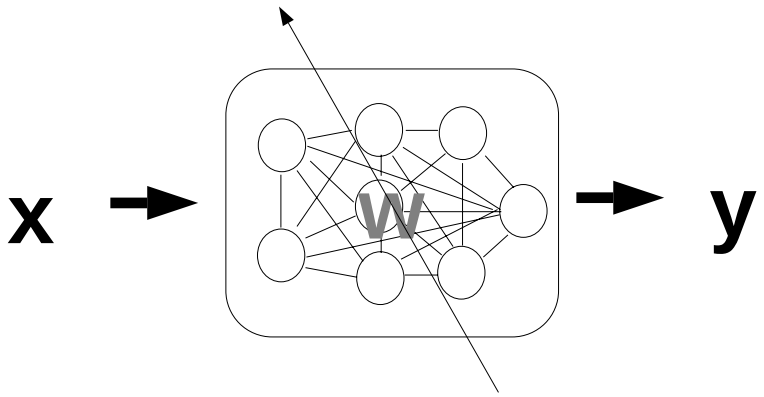

Abbildung 2.1: Die Eingabe-/Ausgaberelation des konnektionistischen Modells hängt von den Kopplungsparametern $\mathbf{w}$ ab. Der diagonale Pfeil soll andeuten, daß die Kopplungsparameter adaptiert, d.h. gelernt, werden können.

litätskriterien, die eine gewünschte Eigenschaft der Abbildung hervorrufen.

Ziel dieser Arbeit ist es, verschiedene Konzepte aus der Informationstheorie zur Definition von Lernkriterien zu verwenden. Es wird die Hypothese aufgestellt, daß Begriffe wie *relative Entropie*, *gemeinsame Information* oder *Transinformation* zu nützlichen Kriterien für unüberwachtes Lernen führen können. Die Aufgabe besteht darin, geeignete Begriffe aus der Informationstheorie als Optimalitätskriterien für das unüberwachte Lernen zu definieren. Es sollen konnektionistische Modelle untersucht werden, in denen sich diese Kriterien anwenden lassen, um dann ihren potentiellen Nutzen zu untersuchen.

Zunächst werden die zentralen Begriffe *Information* und *Entropie* eingeführt.

2.1 Information und Entropie

Was versteht man unter der *Information* eines sensorischen Signals? Zur Definition der Information konzentriert man sich zunächst auf diskrete Signale $\mathbf{x} \in \mathbf{\Xi}$, wobei $\mathbf{\Xi}$ den diskreten Raum der möglichen Signale benennt. Man bezeichne mit $P(\mathbf{x})$ die Wahrscheinlichkeit, mit der man ein gewisses Signal $\mathbf{x}$ beobachtet. In dieser Sichtweise handelt es sich bei den Signalen also um diskrete Zufallsvariablen. Shannon (1948) definiert die Information $I[P(\mathbf{x})]$ einer Wahrscheinlichkeitsverteilung $P(\mathbf{x})$ als

$$I[P(\mathbf{x})] = \sum_{\mathbf{x}\in\mathbf{\Xi}} P(\mathbf{x}) \ln P(\mathbf{x}) \tag{2.1}$$

Dieser Ausdruck erinnert unmittelbar an die Definition der *Entropie* $H[P(\mathbf{x})]$

in der Thermodynamik

$$H[P(\mathbf{x})] = -k_B \sum_{\mathbf{x}\in\Xi} P(\mathbf{x}) \ln P(\mathbf{x}) \tag{2.2}$$

Hier gibt $P(\mathbf{x})$ die Wahrscheinlichkeit an, ein thermodynamisches System in einem Zustand $\mathbf{x}$ anzutreffen. In dieser Arbeit werden Zustände eines thermodynamischen Modells Signalen zugeordnet und umgekehrt. Je nach Kontext wird dem einen oder anderen Begriff Vorzug gegeben. In (2.2) bezeichnet k_B die Boltzmannkonstante. Wir werden im folgenden $k_B = 1$ setzen. In der statistischen Thermodynamik wird Entropie auch als $H = k_B ln\Omega$ definiert, wobei Ω die Anzahl Zustände bezeichnet, die einem thermodynamischen System bei gegebener Energie zugänglich sind (siehe z.B. Reif, 1965). Auf diese Definition stützt sich das intuitive Verständnis von Entropie als einer Größe, die "Unsicherheit" quantifiziert. Je größer die Anzahl möglicher Zustände, desto größer die Unsicherheit in welchem Zustand sich das System gerade befindet. Die Parallelität von (2.1) und (2.2) führt zu der inzwischen vielleicht als selbstverständlich erachteten Äquivalenz des Begriffes Entropie mit "Mangel" an Information oder Unsicherheit. Diese Analogie der Definitionen hat der Informationstheorie den Zugang zu Methoden der statistischen Thermodynamik eröffnet.

Auch in konnektionistischen Modellen haben sich Konzepte der Thermodynamik niedergeschlagen: Das bekannte Hopfield-Modell stützt sich auf theoretische Modelle von Spingläser (Hopfield, 1982a); das Simulated-Annealing als thermodynamisch motiviertes Optimierungsverfahren (Geman und Geman, 1984); die Mittelfeldnäherung für die Boltzmann-Maschine (Peterson und Anderson, 1987). Dies sind nur einige der bekanntesten und frühen Konzepte, die in Anlehnung an die theoretische Thermodynamik vorgeschlagen wurden. Die Liste der analytischen Methoden, die aus diesem Feld der Physik in konnektionistischen Modellen angewendet werden, hat sich beständig verlängert.

Auch in den ersten Kapiteln dieser Arbeit wird diese Strategie verfolgt. Man wendet die vorgeschlagenen informationstheoretischen Lernkriterien auf das thermodynamisch motivierte Modell der *Boltzmann-Maschine* an. Die Boltzmann-Maschine wurde von Ackley, Hinton, und Sejnowski (1985) eingeführt. Als ein stochastisches System hat sie die wichtige Eigenschaft, daß die Wahrscheinlichkeitsverteilung für ihre binären Zustände durch eine explizite Definition analytisch gegeben ist. Man erwartet daher, daß Informationbegriffe wie *relative Entropie*, *maximale Transinformation* und *minimale Redundanz*, die auf der Wahrscheinlichkeitsverteilung aufbauen, bei der Boltzmann-Maschine analytisch zugänglich sind. Im Kapitel 3 wird eine ausführliche Einführung in dieses stochastische, rückgekoppelte System gegeben.

Im folgenden sollen diese informationstheoretischen Lernkriterien ausgehend von einfachen Überlegungen zu den Signalverarbeitungsaufgaben für ein Lebewesen begründet werden. In den folgenden Kapiteln werden diese auf verschiedene Modelle angewandt.

2.2 Minimale relative Entropie

Bei einem Lebewesen ist die erste und grundlegende Aufgabe der Informationsverarbeitung das Erkennen seiner sensorischen Umwelt. Man kann davon ausgehen, daß die Umwelt durch die Wahrscheinlichkeit gekennzeichnet wird, mit der Ereignisse oder Zustände eintreten. Noch bevor ein kognitiver Prozeß stattfinden kann (wie das Erkennen von Regelmäßigkeiten oder Unregelmäßigkeiten, die ein Indiz auf eine Struktur in der Umgebung sein können), ist es wichtig, zu wissen, wie wahrscheinlich ein gewisses sensorisches Signal ist. In dieser Sichtweise bedeutet also, "die Umwelt zu kennen", daß man ein Modell der Wahrscheinlichkeitsverteilung der sensorischen Stimuli gebildet hat. Diese Aufgabe bezeichnet man im allgemeinen als *Verteilungs-* oder *Dichteschätzung.*

Nehmen wir an, man verfügt über ein adaptives System, das einem Eingabesignal $\mathbf{x} \in \Xi$ eine gewisse Wahrscheinlichkeit $P(\mathbf{x})$ zuordnen kann, wobei diese Wahrscheinlichkeit von der Wahl des Adaptionsparameter $\mathbf{w}$ abhängt, d.h. $P(\mathbf{x}) = P(\mathbf{x}; \mathbf{w})$. [1] Weiter sei $R(\mathbf{x})$ eine feste Wahrscheinlichkeitsverteilung, die den Signalen in der Umwelt entspricht. Diese ist dem Beobachter allerdings nicht bekannt. In der Praxis liegt nur ein Satz von Datenpunkten vor, die gemäß dieser Verteilung beobachtet wurden. Dichteschätzung bedeutet, daß man mit Hilfe dieser Datenpunkte ein Modell $P(\mathbf{x}; \mathbf{w})$ der Verteilung $R(\mathbf{x})$ bildet. Der erste Begriff, der uns zu einem Lernkriterium führt, ist die *relative Entropie.* Die relative Entropie $H[P, R]$ mißt den Abstand zweier Wahrscheinlichkeitsverteilungen $P(\mathbf{x})$ und $R(\mathbf{x})$

$$H[P(\mathbf{x}), R(\mathbf{x})] = \sum_{\mathbf{x} \in \Xi} P(\mathbf{x}) \ln \frac{P(\mathbf{x})}{R(\mathbf{x})} \tag{2.3}$$

Die bekannte Ungleichung $0 \leq H[P, R]$ (siehe Cover und Thomas (1991)), wobei die Gleichheit auch $P(\mathbf{x}) = R(\mathbf{x})$ impliziert, stellt die Grundlage für diesen Abstandsbegriff dar. Hiermit läßt sich eine Schätzung der Wahrscheinlichkeitsverteilung gewinnen. Man optimiert das adaptive System, so daß man ein Parameter $\mathbf{w}^*$ mit *minimaler relativer Entropie* findet (siehe Abbildung 2.2). Hat man ein gutes Minimum gefunden, so modelliert die vom System erzeugte Verteilung $P(\mathbf{x}; \mathbf{w}^*)$ die zu lernende Verteilung $R(\mathbf{x})$.

Das Lernkriterium der minimalen relativen Entropie bezeichnet man als *überwachtes Lernen*, da man eine Zielverteilung $R(\mathbf{x})$ vorgegeben hat. Im Gegensatz hierzu steht das *unüberwachte Lernen*, bei dem man dem konnektionistischen und adaptiven System lediglich eine Eingabe zur Verfügung stellt. Es bleibt dem System dann frei überlassen, welche Repräsentation es findet, um ein gewisses Lernkriterium zu erfüllen.

[1] In dieser Arbeit werden Parameter einer Verteilung durch ein Semikolon von den Zufallsvariablen getrennt. Zur Nomenklatur dieser Arbeit siehe Kapitel 8

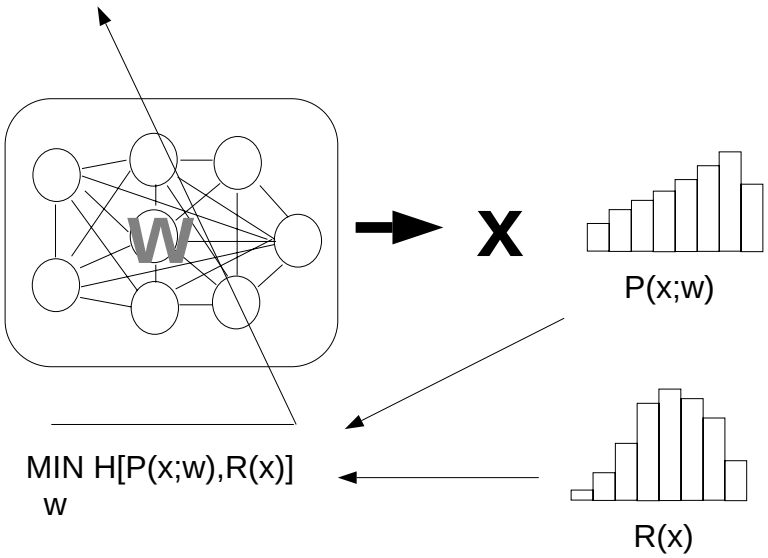

Abbildung 2.2: Relative Entropie $H[P(\mathbf{x};\mathbf{w}), R(\mathbf{x})]$ mißt den Abstand der erzeugten Verteilung $P(\mathbf{x};\mathbf{w})$ von der Zielverteilung $R(\mathbf{x})$ und kann für die Adaptation der Kopplungsparameter $\mathbf{w}$ verwendet werden.

2.3 Maximale Transinformation

Aber welches Kriterium für das unüberwachte Lernen könnte hilfreich sein? Wie angenommen wurde, ist es für ein Lebewesen eine grundlegende Aufgabe, seine Umgebung zu kennen. Man kann nun weiter annehmen, daß es grundlegend ist, welche Repräsentation von der Umgebung gebildet wird. Entsprechend der Information, die aus dem Signal extrahiert werden soll, kann es verschiedene sinnvolle Repräsentationen geben. Ist es z.B. wichtig, "ähnlichen" sensorischen Signalen eine "benachbarte" Repräsentation zuzuordnen, so findet man Konzepte wie die "selbstorganisierenden Merkmalskarten" von Kohonen (1982). Ohne Anhaltspunkte allerdings, wie die Information weiter verarbeitet wird, erscheint es am sinnvollsten, eine Repräsentation zu finden, die möglichst viel Information der Umwelt enthält.

Eine Größe, die angibt, wieviel Information eine Zufallsvariable $\mathbf{y} \in \mathbf{\Upsilon}$ über eine Zufallsvariable $\mathbf{x} \in \mathbf{\Xi}$ enthält, ist die *gemeinsame Information* $I[P(\mathbf{x},\mathbf{y})]$

$$I[P(\mathbf{x},\mathbf{y})] = H[P(\mathbf{y})] + H[P(\mathbf{x})] - H[P(\mathbf{x},\mathbf{y})] \tag{2.4}$$

Hier sind die ersten zwei Terme die Entropien der Wahrscheinlichkeitsverteilungen der jeweiligen Zufallsvariablen. Der dritte Term mißt die Information, die in der *gemeinsamen Wahrscheinlichkeitsverteilung* $P(\mathbf{x},\mathbf{y})$ der beiden Zufallsvariablen enthalten ist

$$H[P(\mathbf{x},\mathbf{y})] = -\sum_{\mathbf{y}\in\Upsilon}\sum_{\mathbf{x}\in\Xi} P(\mathbf{x},\mathbf{y})\ln P(\mathbf{x},\mathbf{y}) \tag{2.5}$$

Im diskreten Zustandsraum ist die gemeinsame Information $I[P(\mathbf{x},\mathbf{y})]$ stets positiv und nicht größer als die Entropie einer der beiden Verteilungen $P(\mathbf{x})$ oder $P(\mathbf{y})$ (Cover und Thomas, 1991). Wir werden in Kapitel 4 und Kapitel 5 verschiedene Interpretationen der gemeinsamen Information (2.4) betrachten. Versteht man die beiden Zufallsvariablen als Ein- und Ausgabe eines stochastischen Systems, so wird dieser Ausdruck auch *Transinformation* genannt. Man verwendet das Prinzip der *maximalen Transinformation*, falls man an der Ausgabe eine Repräsentation mit möglichst viel Information über die Eingabe erhalten möchte. Die obere Schranke macht dieses Prinzip zu einem quantitativ aussagekräftigen Optimalitätskriterium. Es hat sich als fruchtbares Kriterium für unüberwachtes Lernen in konnektionistischen Modellen erwiesen. Dieses Konzept wurde von mehreren Autoren vorgeschlagen und auf verschiedene Systeme angewendet (Becker, 1992; Bridle, 1989; Bridle, MacKay, und Heading, 1991; Linsker, 1988, 1989, 1992; Atick und Redlich, 1990). Kapitel 4 erklärt einige dieser Ansätze. Diese Modelle sind in der Lage, statistisch signifikante Merkmale zu extrahieren (Atick und Redlich, 1990). Im Fall von linear korrelierten Eingabesignalen ist dieses Kriterium mit der linearen *Hauptkomponentenanalyse* (PCA) verwandt (Földiák, 1989). Alle diese Modelle beschränken sich allerdings auf lineare Transformationen. In den zwei jüngsten Arbeiten von (Nadal und Parga, 1994) bzw. (Bell und Sejnowski, 1995) wird das Prinzip der maximalen Transinformation (auch als *Infomax* bekannt) in einem nichtlinearen deterministischen Fall betrachtet.

In Kapitel 4 wird untersucht, ob das Prinzip *maximaler Transinformation* auch auf ein stochastisches, rückkoppelndes, nichtlineares System wie die Boltzmann-Maschine anwendbar ist. Verwendet man an der Ausgabe einen kleineren Zustandraum, so erwartet man, daß die wesentliche Information der Eingabe an der reduzierten Ausgabe erhalten bleibt. In Abbildung 2.3 ist dieses Konzept für eine dreidimensionale binäre Eingabe und eine zweidimensionale binäre Ausgabe dargestellt. Es wird untersucht, ob das System die hier suggerierte, adaptive Datenkompression verwirklichen kann.

2.4 Statistisch unabhängige Merkmale und gemeinsame Information

Neben einer getreuen Informationsübertragung erwartet man von einem sensorischen System auch das Erkennen und Extrahieren von Merkmalen der Umwelt.

An dieser Stelle soll die informationstheoretische Definition des Begriffes *statistisches Merkmal* und des verwandten Begriffes der *Redundanz* einführt wer-

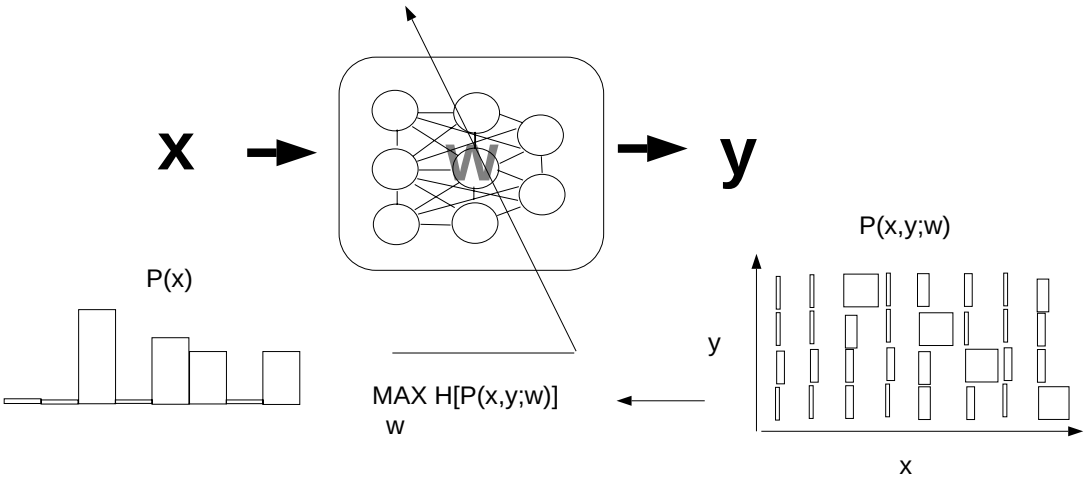

Abbildung 2.3: Die Eingabe $\mathbf{x}$ besteht aus drei binären Elementen. Es können somit acht Zustände kodiert werden. Die dargestellte Eingabeverteilung $P(\mathbf{x})$ ist so gewählt, daß im wesentlichen nur vier dieser Zustände vorkommen. Die zweidimensionale binäre Ausgabe kann nur vier Zustände kodieren. Die gemeinsame Verteilung dieser Zustände $P(\mathbf{x}, \mathbf{y}; \mathbf{w})$ hängt von den Kopplungsparametern $\mathbf{w}$ ab. Mit der gemeinsamen Information $I[P(\mathbf{x}, \mathbf{y}; \mathbf{w})]$ läßt sich die Information, die Zufallsvariable $\mathbf{y}$ über $\mathbf{x}$ enthält, durch Adaptieren der Kopplungsparameter $\mathbf{w}$ maximieren. Hier wird eine gemeinsame Verteilung $P(\mathbf{x}, \mathbf{y}; \mathbf{w})$ angezeigt, die eine gute Lösung für das Problem darstellt. (Die Häufigkeit, mit der eine bestimmte $(\mathbf{x}, \mathbf{y})$ Kombination auftritt, ist in der Größe der entsprechenden Kästchen kodiert.) Jedem der vier möglichen Ausgabewerte entspricht im wesentlichen ein bestimmter Eingabewert, d.h. die Ausgabe enthält die wesentliche Information über die Eingabe.

den. Die statistische Unabhängigkeit fungiert hier als Schlüsselidee. Als *statistisches Merkmal* versteht man im allgemeinen eine Eigenschaft, die statistisch unabhängig von anderen Eigenschaften ist. Man nehme an, daß verschiedene Eigenschaften durch verschiedene Koordinaten $y_1, .., y_n$ eines mehrdimensionalen Zustandsraumes gemessen werden. Läßt sich die gemeinsame Wahrscheinlichkeitsverteilung $P(y_1, y_2, .., y_n)$ als Produkt der Wahrscheinlichkeitsverteilungen der Einzelkoordinaten $P(y_i)$ darstellen

$$P(\mathbf{y}) = \prod_{i}^{n} P(y_i) \tag{2.6}$$

so bezeichnet man die y_i als statistisch unabhängig. Es ist interessant anzumerken, daß die Definition (2.4) der gemeinsamen Information als die relative Entropie zwischen der rechten und linken Seite der Bedingung (2.6) für den Fall zweier mehrdimensionaler Zufallsvariablen verstanden werden kann. Das bedeutet, daß die gemeinsame Information angibt, wie "faktoriell" eine gemeinsame Wahrscheinlichkeitsverteilung ist. Mit (2.6) kann man die gemeinsame Information $MI[P(\mathbf{y})]$ (mutual information) für n eindimensionale Variablen definieren

$$MI[P(\mathbf{y})] = \sum_{i}^{n} H[P(y_i)] - H[P(\mathbf{y})] \tag{2.7}$$

Die gemeinsame Information ist eine nicht negative Größe, wobei $MI[P(\mathbf{y})] = 0$ die Gültigkeit von (2.6) impliziert. Von Redundanz spricht man, wenn in einer diskreten Kodierung ein Ziffernwert y_i Information über den Wert einer anderen Ziffer y_j enthält - sie sind also statistisch abhängig. Redundanz $R[P(\mathbf{y})]$ kann somit durch gemeinsame Information gemessen werden. Sie wird mit $H[P(\mathbf{y})]$ normiert

$$R[P(\mathbf{y})] = \frac{MI[P(\mathbf{y})]}{H[P(\mathbf{y})]} \tag{2.8}$$

Das Finden von statistisch unabhängigen Merkmalen wurde unter anderen von Zipf (1949) und von Attneave (1954) als Grundlage der sensorischen Signalverarbeitung des Nervensystems vorgeschlagen. Barlow (1989) identifiziert dieses Prinzip als fundamentales Ziel des unüberwachten Lernens. Die Arbeiten von Redlich (1993a), Redlich (1993b) und Atick und Redlich (1992) implementieren Barlows Merkmalsextraktion durch Redundanzreduktion. Auch in diesem Gebiet haben sich die Arbeiten bisher hauptsächlich auf lineare Abhängigkeit zwischen den Signalen beschränkt. Um analytische Resultate zu erreichen, hat man dabei oft normalverteilte Eingabesignale angenommen.

In Kapitel 5 wird untersucht, ob sich das Prinzip der *minimalen gemeinsamen Information* in dem stochastischen, rückgekoppelten, nichtlinearen System der Boltzmann-Maschine verwirklichen läßt. Dabei sollen keine Annahmen über die Signalverteilung an der Eingabe gemacht werden. Ziel ist es, eine binäre

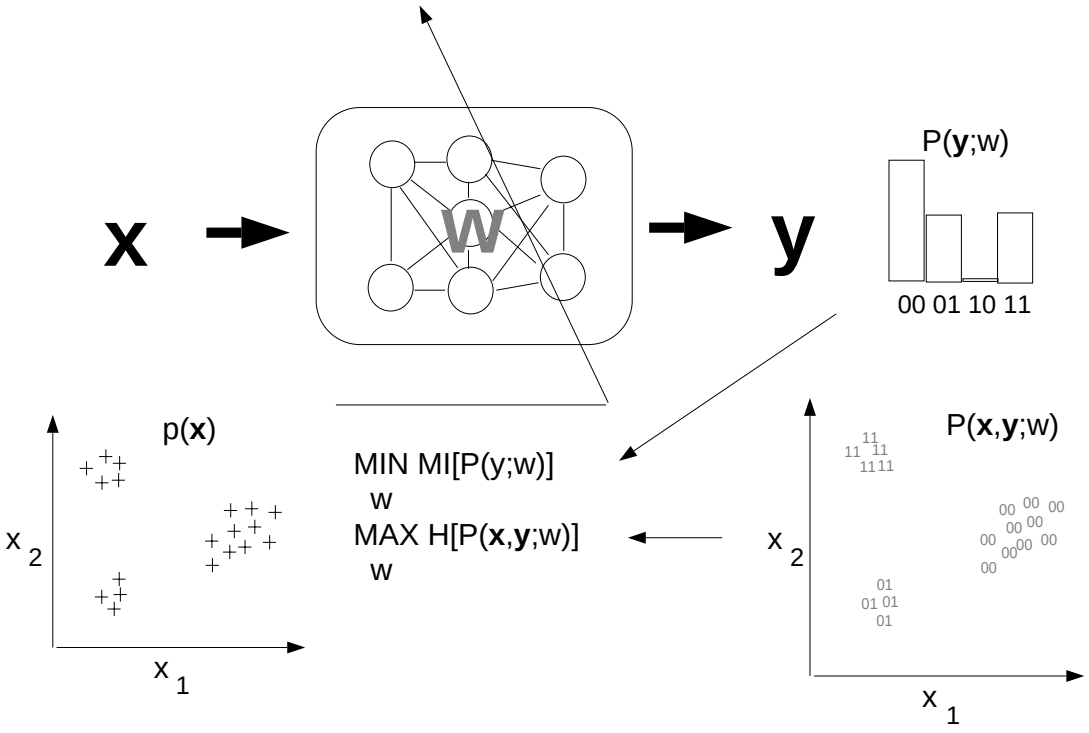

Abbildung 2.4: Die Eingabe $\mathbf{x}$ ist ein zweidimensionaler kontinuierlicher Vektor. Die Ausgabe soll die wichtigsten Merkmale der Eingabeverteilung in zwei binären Elementen kodieren. Die gemeinsame Verteilung $P(\mathbf{x}, \mathbf{y}; \mathbf{w})$ hängt von den Kopplungsparametern $\mathbf{w}$ ab. Man erwartet, daß maximale Transinformation $H[P(\mathbf{x}, \mathbf{y}; \mathbf{w})]$ diese Aufgabe erfüllt. Die extrahierten binären Merkmale sollen unabhängig sein. Minimale gemeinsame Information $MI[P(\mathbf{y}; \mathbf{w})]$ an der Ausgabe soll dieses Ziel erreichen. Hier wird eine gemeinsame Verteilung $P(\mathbf{x}, \mathbf{y}; \mathbf{w})$ angedeutet, die eine Lösung des Problems darstellt. Punkte innerhalb einer Punktwolke erzeugen die gleiche Ausgabe (11, 01, 00), während verschiedene Punktwolken verschiedene Ausgaben erzeugen. Die Punkte wurden in Klassen eingeteilt. Die Optimierungsaufgabe repräsentiert somit eine unüberwachte Klassifikation. Jedes der beiden Ausgabeelemente bezeichnet eine andere Eigenschaft der Eingabe: oben/unten bzw. links/rechts.

Merkmalsextraktion sowie redundanzfreie binäre Rekodierung zu verwirklichen. In Abbildung 2.4 wird ein einfaches Beispiel für eine binäre Merkmalsextraktion gegeben. Aus einer zweidimensionalen kontinuierlichen Verteilung (vertikale und horizontale Lage eines Punktes) sollen zwei binäre Eigenschaften extrahiert werden. In diesem Beispiel erwartet man, daß die binären Merkmale oben/unten und links/rechts extrahiert werden. Das bedeutet, daß jeder Eingabewert einer gewissen Klasse zugeordnet wird.

2.5 Boltzmann-Maschine im kontinuierlichen Zustandsraum

Die Boltzmann-Maschine hat bisher eine grundlegende Einschränkung: der Zustandsraum ist binär. In der konnektionistischen Informationsverarbeitung und ihren Anwendungen ist es wichtig, kontinuierliche Merkmale erzeugen zu können.

Der analytische Ausdruck für Wahrscheinlichkeiten in der Boltzmann-Maschine erlaubt die informationstheoretischen Maße in Lernkriterien umzusetzen. Es wird daher in Kapitel 6 versucht, eine Erweiterung für kontinuierliche Zustände zu formulieren. Es wird untersucht, welches der bisher vorgeschlagenen Lernkriterien auf solch ein Modell angewandt werden kann.

Wie sich zeigt, beschränkt die Stochastizität der thermodynamischen Modelle im kontinuierlichen Fall die Implementation der theoretischen Ergebnisse. In den folgenden Kapiteln dieser Arbeit wird daher ein völlig neuer Ansatz gewählt. Man wird von vornherein nur ein deterministisches System zulassen, für das man keinen analytischen Ausdruck der Wahrscheinlichkeitsdichten besitzt. Daraus ergibt sich die neue Aufgabe, eine gute Schätzung der Entropien aus den gemessenen Signalen zu gewinnen.

2.6 Informationsbegriffe im kontinuierlichen Zustandsraum

Die informationstheoretischen Lernkonzepte, die bisher eingeführt wurden, basieren auf der Definition der Entropie. Die Hauptaufgabe besteht nun darin, die Entropie einer kontinuierlichen *Wahrscheinlichkeitsdichte* $p(\mathbf{x})$ zu messen. Bislang wurde von Entropien einer diskreten Wahrscheinlichkeitsverteilung $P(\mathbf{x})$ gesprochen. Es sollen nun keine diskreten Zustände, sondern alle möglichen kontinuierlichen Zustände $\mathbf{x} \in \Re^n$ zugelassen sein. Die Definition der Entropie einer *Wahrscheinlichkeitsdichte* $p(\mathbf{x})$ von Shannon (1948) ist analog zu (2.2)

$$H[p(\mathbf{x})] = -\int_{\Re^n} \mathbf{x} p(\mathbf{x}) \ln p(\mathbf{x}) \tag{2.9}$$

Hat man jedoch keinen analytischen Ausdruck für das Integral in (2.9), so wird es schwierig, die darauf aufbauenden Lernkriterien auszuwerten. Eine numerische Integration über einen hochdimensionalen Raum ist praktisch nicht durchführbar.

Während man im Diskreten Information durch einfaches Zählen von Ereignissen messen kann, ist man im Kontinuierlichen mit Problemen wie Meßgenauigkeit oder Rauschen konfrontiert. Im Gegensatz zum diskreten Fall ändert eine beliebige Skalierung einer Wahrscheinlichkeitsdichte die zugehörige Entropie. Wie in Abschnitt 7.1 deutlich wird, ändert eine Volumenzerrung den Entropiegehalt einer Signaldichte. Diese Änderung des Informationsmaßes ist aber nicht von Bedeutung, sondern lediglich auf die Normierungsbedingung für eine Wahrscheinlichkeitsdichte zurückzuführen. Wie man im gleichen Abschnitt sieht, wird aus diesem Grund für kontinuierliche Signale der Zusammenhang zwischen Information und Entropie strenger als Entropiedifferenz vor und nach einer Messung definiert.

2.7 Konstante Informationsübertragung

Im kontinuierlichen Fall ist es notwendig, das Kriterium der maximalen Transinformation zu überdenken. Die Definition der Transinformation einer kontinuierlichen Wahrscheinlichkeitsdichte $p(\mathbf{x}, \mathbf{y})$ ist analog zu (2.4). Während die Transinformation im diskreten Fall beidseitig beschränkt ist, gilt das im Kontinuierlichen nicht mehr. Für eine deterministische Abbildung ist die Transinformation unendlich groß. Dies bringt zum Ausdruck, daß mit einer Abbildung unendlicher Genauigkeit unendlich viel Information übertragen werden könnte. Die Maximierung der Transinformation macht für eine deterministische Abbildung keinen Sinn mehr.

In dieser Arbeit wird vorgeschlagen, statt dessen *konstante Informationsübertragung* als neues Prinzip zu verwenden:

$$H[p(\mathbf{x})] \stackrel{!}{=} H[p(\mathbf{y}; \mathbf{w})] \tag{2.10}$$

Ein berechtigter Einwand könnte sein, daß man im Signal enthaltenes Rauschen nicht vollständig übertragen möchte. Signal und Rauschen lassen sich allerdings nicht sinnvoll voneinander trennen, ohne zuvor ein Modell der beobachteten Signale gebildet zu haben bzw. Annahmen über die Rauschstatistik gemacht zu haben. Solange man jedoch kein Modell der Signalverteilung zur Verfügung hat, ist es sinnvoll, die Information der Eingabesignale konstant auf die Ausgabe abzubilden.

Ziel des unüberwachten Lernens ist es, eine neue Darstellung der Signale zu gewinnen. Anders als in Modellen biologisch neuronaler Netze, gibt es in einer künstlichen, konnektionistischen Struktur keinen Grund, stochastische Abbildungen zu betrachten. Das Rauschen, das von einer stochastischen Abbildung an der

Ausgabe generiert wird, vermindert die übertragene Information auf eine künstliche Art und Weise. Es liegt also nahe, deterministische Abbildungen $\mathbf{y} = \mathbf{f}(\mathbf{x}; \mathbf{w})$ zwischen den Eingabesignalen und der neuen Ausgabe zu verwenden. Wie man in Kapitel 7 aus Ungleichung (7.1) ersieht, ist Bedingung (2.10) für volumenerhaltende Abbildungen erfüllt

$$\left| \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right| = 1 \tag{2.11}$$

Zumeist beschränkt sich die Literatur in diesem Forschungsgebiet auf die Untersuchung von einfachen Abbildungen, die durch lineare Transformationen aus den beobachteten Signalen hervorgehen. Im Gegensatz dazu beabsichtigt man in dieser Arbeit, Klassen von *nichtlinearen* Transformationen zu verwenden, die die Anwendung der vorgeschlagenen Informationskriterien erlauben.

Das führt zu den *symplektischen* Abbildungen, einer Klasse von volumenerhaltenden, nichtlinearen Transformationen (siehe Abraham und Marsden (1978)). Es werden in Kapitel 7.1.1 und Kapitel 7.1.2 implizite und explizite Definitionen von symplektischen Funktionen betrachtet. Es wird untersucht, wie diese Abbildungen in adaptiven konnektionistischen Strukturen verwendet werden können.

2.8 Minimale gemeinsame Information

Es ist wichtig zu bemerken, daß für die korrekte Definition von Volumenerhaltung im Ein- und Ausgaberaum die gleiche Dimension gewählt werden muß. Das heißt, man wird nicht ohne weiteres eine Dimensionsreduktion erzielen können. Statt dessen wird eine neue Darstellung der Eingabeverteilung an der Ausgabe erzeugt. Das Kriterium soll, wie im ersten Teil der Arbeit, die Extraktion von statistisch unabhängigen Merkmalen sein. Man minimiert also die gemeinsame Information in den Ausgabekoordinaten.

Die Definition der gemeinsamen Information im Kontinuierlichen ist wieder analog zu (2.7). Da die Entropie bei der Abbildung erhalten bleibt, muß man zur Reduktion der gemeinsame Information lediglich die Entropien $H[p(y_i)]$ der Einzelkoordinaten y_i minimieren

$$MI[p(\mathbf{y})] = \sum_i^n H[p(y_i)] - H[p(\mathbf{y})] = \sum_i^n H[p(y_i)] - H[p(\mathbf{x})] \tag{2.12}$$

Die Entropie $H[p(\mathbf{x})]$ der Eingabe ist konstant und hängt nicht von den zu optimierenden Kopplungsparametern ab. Zur Minimierung der gemeinsamen Information ist es also nicht mehr notwendig, die statistische Abhängigkeit der verschiedenen Koordinaten zu beachten, was eine immense Reduktion der Komplexität der Optimierungsaufgabe verspricht. Minimale gemeinsame Information mit volumenerhaltenden Abbildungen ist kein vollkommen neues Konzept.

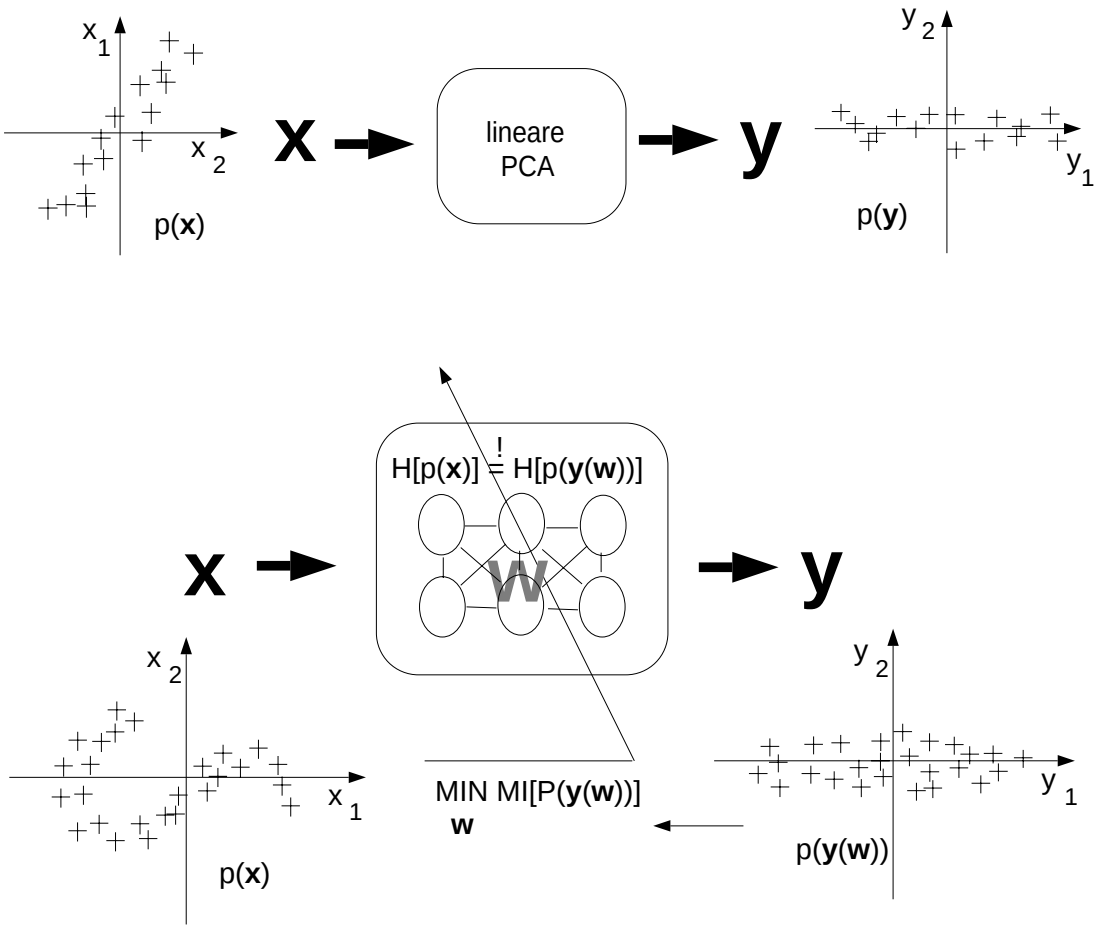

Abbildung 2.5: Oben: Das Konzept der linearen PCA wird verdeutlicht. Das Finden der linearen Hauptkomponenten einer Verteilung kann auch als Rotation verstanden werden, bei der die Hauptkomponente auf eine der neuen Koordinaten projiziert wird. Unten: Eine nichtlineare Komponentenanalyse transformiert die nichtlineare Struktur der Eingabeverteilung $p(\mathbf{x})$ in eine Ausgabeverteilung $p(\mathbf{y}(\mathbf{w}))$ mit statistisch unabhängigen Koordinaten. Man beachte die Schreibweise. Die Abhängigkeit $\mathbf{y}(\mathbf{w})$ macht deutlich, daß die Verteilung mittels $\mathbf{y}$ nur implizit von den Kopplungsparametern $\mathbf{w}$ abhängt. Statistische Unabhängigkeit erreicht man durch Minimierung der gemeinsamen Information $MI[p(\mathbf{y}(\mathbf{w}))]$ der Ausgabekoordinaten. Man garantiert gleichzeitig konstante Informationsübertragung $H[p(\mathbf{x})] = H[p(\mathbf{y}(\mathbf{w}))]$.

Tatsächlich läßt sich die bekannte Hauptkomponentenanalyse (PCA) auf dieses Konzept zurückführen. Die Rotation, die in der PCA verwendet wird, stellt eine lineare, volumenerhaltende Transformation dar. Nimmt man an, daß die Eingabesignale normal verteilt sind, reduziert sich die minimale gemeinsame Information auf lineare Dekorrelation, wie sie in der PCA verwendet wird. Für allgemeinere Verteilungen allerdings impliziert Dekorrelation keine statistische Unabhängigkeit der Koordinaten. Ausgehend von dem Prinzip der minimalen gemeinsamen Information, leiten Deco und Brauer (1994) Kriterien für statistische Unabhängigkeit durch Mehrkoordinaten-Statistik höherer Ordnung ab. Comon (1994) macht einen ähnlichen Ansatz mit Kumulanten höherer Ordnung, indem er den Abstand der Ausgabeverteilung zu einer Normalverteilung betrachtet (standardisierte gemeinsame Information). Er beschränkt sich dabei auf lineare Transformationen.

Die vorliegende Arbeit beabsichtigt somit eine Verallgemeinerung der PCA in zweifacher Weise. Zum einen sollen *nichtlineare* Transformationen verwendet werden, zum anderen möchte man auch *nicht-normalverteilte* Ein- oder Ausgabekoordinaten zulassen. Die Idee der nichtlinearen Komponentenanalyse wird in Abbildung 2.5 verdeutlicht. Dieses Konzept wird in Kapitel 7 untersucht.

2.9 Entropiemessung mit Statistik höherer Ordnung

Zur Minimierung der gemeinsamen Information der Ausgabekoordinaten benötigt man eine Methode zur Messung oder Schätzung der Entropien der Einzelkoordinaten $H[p(y_i)]$. Diese sind aber nicht genau bestimmbar, da die Ausgabeverteilung nur über einen Satz von Datenpunkten gegeben ist. Darüber hinaus gehen die Parameter, die optimiert werden sollen, nur implizit mittels der Ausgabewerte in die Verteilung ein. Allerdings ist eine einfache obere Grenze für die Entropie bekannt. Diese läßt sich durch die meßbare Varianz σ der Wahrscheinlichkeitsdichte $p(y)$ analytisch ausdrücken

$$H[p(y)] \leq \frac{1}{2} ln(2\pi e) + \frac{1}{2} ln\sigma^2 \tag{2.13}$$

Sie ermöglicht eine numerisch effiziente Optimierung des Lernkriteriums der minimalen gemeinsamen Information (2.12). Dieser Ansatz wird in Abschnitt 7.2.1 diskutiert und für die Anwendungen der vorgestellten nichtlinearen Merkmalsextraktion in Abschnitt 7.3.2 und 7.3.3 verwendet.

Die Minimierung der Entropie mittels der Varianz erfaßt allerdings keine Statistik höherer Ordnung an der Ausgabeverteilung. Das bedeutet nicht, daß die Eingabeverteilung auf Statistik zweiter Ordnung beschränkt bleiben muß. Um darüber hinaus Statistik höherer Ordnung an der Ausgabe zu betrachten, wird hier die Edgeworth-Entwicklung zur Schätzung einer Wahrscheinlichkeitsdichte

verwendet. Diese Messung basiert auf der Bestimmung von Kumulanten höherer Ordnung und hat den Vorteil, daß man einen analytischen Ausdruck für die Entropie und somit für die Kostenfunktion erhält. Man erhofft damit, Gradientenmethoden für die Optimierung der konnektionistischen Strukturen zu gewinnen. Dieses Konzept wird in Abschnitt 7.2 analysiert.

2.10 Dichteschätzung - bedingte Wahrscheinlichkeit

Hat man schließlich eine faktorielle Darstellung der Wahrscheinlichkeitsdichte gefunden, kann dies auch für die klassische Dichteschätzung verwendet werden. Statt der Schätzung einer Dichte im n-dimensionalen Zustandsraum, kann man jetzt die verschiedenen Koordinaten unabhängig voneinander betrachten. Das Problem reduziert sich somit auf die wesentlich einfachere Schätzung von eindimensionalen Dichten. Das spezielle Konzept der konstanten Informationsübertragung und die Architekturen mit denen die volumenerhaltenden Transformationen verwirklicht werden, legen zwei Anwendungen dieses vorgeschlagenen Konzeptes zur Dichteschätzung durch minimale gemeinsame Information nahe: *Neuheitsdetektion* und Schätzung von *bedingten Wahrscheinlichkeiten* bzw. Funktionenapproximation, beides für kontinuierliche Signale.

Bei der Schätzung einer *bedingten Wahrscheinlichkeitsverteilung* will man die Wahrscheinlichkeit $p(\mathbf{y}|\mathbf{x})$ modellieren, die einem Signal $\mathbf{y}$ zukommt, wenn ein $\mathbf{x}$ beobachtet wurde. Im allgemeinen steht für diese Aufgabe die Zielverteilung in Form von Wertepaaren $(\mathbf{y}, \mathbf{x})$ zur Verfügung. Diese sind gemäß der gemeinsamen Dichte $p(\mathbf{y}, \mathbf{x})$ gezogen worden. Man kann nun den kombinierten Raum der beiden Zufallsvariablen betrachten und die gemeinsame Verteilung $p(\mathbf{y}, \mathbf{x})$ schätzen (siehe Abbildung 2.6). Die bedingte Wahrscheinlichkeitsdichte erhält man aus

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{y}, \mathbf{x})}{p(\mathbf{x})} \tag{2.14}$$

Man modelliert hierbei einen probabilistischen Zusammenhang zwischen der Eingabe und der Ausgabe. Es lassen sich daraus verschiedene statistische Aussagen ableiten. So kann man die Sicherheit angeben, mit der ein Ausgabewert in einem bestimmten Intervall zu beobachten ist. Dies ist bei Prognoseaufgaben wünschenwert, bei denen man wissen möchte, mit welcher Wahrscheinlichkeit eine bestimmte Vorhersage eintrifft. Eine Lösung für diese Aufgabenstellung wird in Abschnitt 7.3.1 ausgearbeitet.

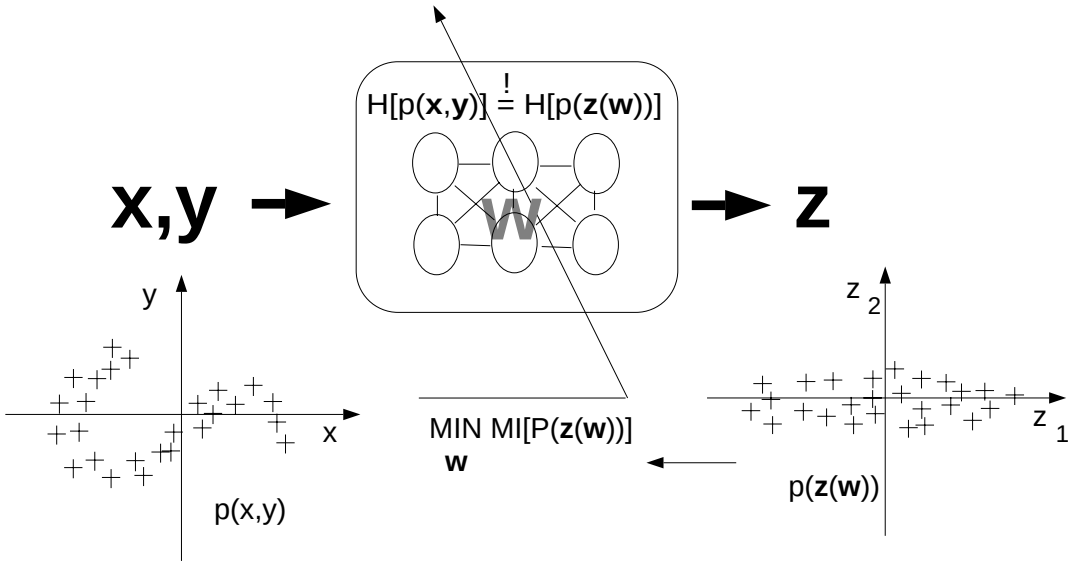

Abbildung 2.6: Die Eingabe des Systems besteht nun aus den kombinierten Zufallsvariablen $(\mathbf{y}, \mathbf{x})$. Die Ausgabe wird hier mit $\mathbf{z}$ bezeichnet. Es wird eine einfache Verteilung $p(\mathbf{z}(\mathbf{w}))$ mit statistisch unabhängigen Koordinaten an der Ausgabe erzeugt. Diese liefert eine Schätzung der Dichte $p(\mathbf{y}, \mathbf{x})$. Auf ähnliche Weise kann man eine Schätzung von $p(\mathbf{x})$ gewinnen. Mit Gleichung (2.14) erhält man dann eine Schätzung für die bedingte Dichte $p(\mathbf{y}|\mathbf{x})$.

2.11 Dichteschätzung - Neuheitsdetektion

In dieser Arbeit wird die mit der Dichteschätzung eng verwandte Aufgabe der *Neuheitsdetektion* betrachtet. Eine Dichte ist gegeben durch einen Satz von bekannten Datenpunkten. In der Neuheitsdetektion muß entschieden werden, ob ein neuer Datenpunkt zur geschätzten Verteilung gehört oder nicht. Die Frage lautet daher: Wie wahrscheinlich ist ein beobachteter neuer Datenpunkt entsprechend der bisher beobachten Datenpunkte? Für eine gegebene Entscheidungsschwelle möchte man somit die entsprechende Konturlinie der Dichte der bekannten Datenpunkte bestimmen. Benötigt man die Kontur für jede beliebige Entscheidungsschwelle, so muß die gesamte Dichte geschätzt werden. Die Aufgabe der Neuheitsdetektion mit Hilfe einer vorhergehenden Faktorisierung der Wahrscheinlichkeitsdichte wird in Abbildung 2.7 veranschaulicht. Dieses Konzept wird in Abschnitt 7.3.2 diskutiert.

In Abschnitt 7.3.3 schließlich wird dieses Konzept an einem technischen Problem angewandt. Es soll der Ausfall eines Elektromotors vorhergesagt werden, indem man Unregelmäßigkeiten in einem gemessenen Stromspektrum erkennt.

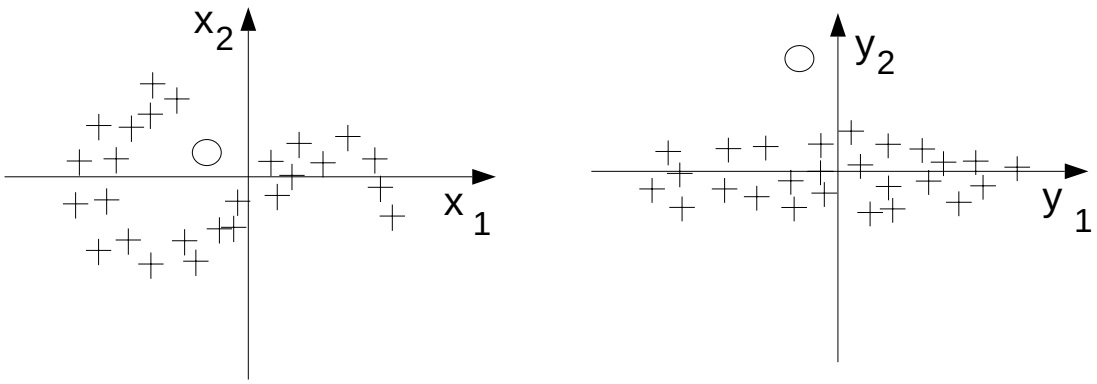

Abbildung 2.7: Der Punkt, der durch ein Kreis repräsentiert wird und als neu detektiert werden soll, ist nach der Transformation einfacher von dem Rest der Verteilung zu trennen sein. Die Transformation ist wieder die vorgeschlagene informationserhaltende Faktorisierung, so wie sie in Abbildung 2.5 unten dargestellt wurde.

Der hoch-dimensionale Merkmalsraum (200-1000 Frequenzwerte) macht eine Dimensionsreduktion erforderlich. In der Praxis wird hierzu oft lineare Hauptkomponentenanalyse verwendet. Da das vorgestellte Konzept als eine nichtlineare Komponentenanalyse verstanden werden kann, erwartet man gute Resultate. Im Gegensatz zu anderen verwendeten Techniken liefert diese Methode zudem eine informationstheoretische Interpretation der Ergebnisse.

2.12 Zusammenfassung

In dieser Arbeit werden die Konzepte der *maximalen Transinformation*, *minimalen gemeinsamen Information* und *konstanten Informationsübertragung* analysiert und als Grundlage für das unüberwachte Lernen von konnektionistischen Modellen verwendet. Für diskrete Ausgabesignale verwendet man das stochastische, thermodynamisch motivierte Modell der Boltzmann-Maschine, das für kontinuierliche Signale erweitert wird. Das neue Konzept der Faktorisierung unter konstanter Information wird auf deterministische Modelle angewandt und analysiert. Es werden die verschiedenen Aufgaben demonstriert, die von den vorgestellten Konzepten gelöst werden können.

Kapitel 3

Relative Entropie im Diskreten - klassische Boltzmann-Maschine

Dieses Kapitel erfüllt zwei Funktionen. Einerseits ist es als Einführung in die Boltzmann-Maschine gedacht, andererseits soll es darlegen, wie das Lernkriterium der minimalen relativen Entropie zur Optimierung eines konnektionistischen Modells verwendet werden kann. Die Beschreibung der neuen unüberwachten Lernkriterien (Kapitel 4 und 5) und der Erweiterung der Boltzmann-Maschine auf einen kontinuierlichen Zustandsraum (Kapitel 6) baut hierauf auf.

Die Boltzmann-Maschine ist eines der ersten konnektionistischen Modelle, das mit dem Begriff der relativen Entropie ein informationstheoretisches Lernkriterium verwendet (Ackley et al., 1985). Es stellt eine Generalisierung des bekannten Hopfield-Modells dar (Hopfield, 1982a). Wie auch beim Hopfield-Modell werden die binären Zustände des stochastischen Systems mit den binären Signalen, die verarbeitet werden sollen, identifiziert. Das Hopfield-Modell wird erweitert, indem man freie Zustände einführt. Diese werden nicht direkt mit den äußeren, vorgegebenen Signalen identifiziert. Die inneren Zustände können so eine Repräsentation der gewünschten Zusammenhänge der äußeren oder "sichtbaren" Zustände verwirklichen. Die Probleme, die aus der inhärenten Stochastizität des thermodynamischen Modells entspringen, werden durch die Anwendung der Mittelfeldtheorie durch Peterson und Anderson (1987) überwunden. Die stochastische Dynamik der binären Zustände wird dabei durch deterministische Gleichungen der Mittel dieser Zustände ersetzt. Die Boltzmann-Maschine ist für verschiedene kombinatorische Aufgaben wie Mustervervollständigung und Kodierung angewendet worden. Im wesentlichen kann man diese in zwei Kategorien unterteilen: Autoassoziation und Modellierung binärer Funktionen.

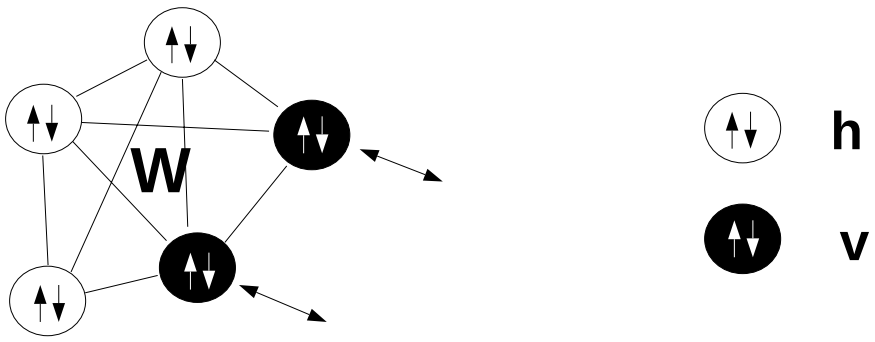

Abbildung 3.1: In der klassischen Boltzmann-Maschine werden die binären Zustände $\mathbf{s}$ in sichtbare $\mathbf{v}$ und innere Zustände $\mathbf{h}$ unterteilt. Sie wechselwirken mittels der Kopplungsmatrix W. Die sichtbaren Zustände dienen als Ein- und Ausgabe.

3.1 Die klassische Boltzmann-Maschine

3.1.1 Definition

Die Boltzmann-Maschine besteht aus binären stochastischen Elementen $\mathbf{s} \in \{1, -1\}^n$. Diese stehen in Wechselwirkung durch eine symmetrische Kopplungsmatrix $W \in \Re^n \times \Re^n$, $w_{ij} = w_{ji}$. Wie angedeutet unterscheidet man zwischen sichtbaren Elementen $\mathbf{v}$ (visible) und inneren Elementen $\mathbf{h}$ (hidden). Die entsprechende Notation $\mathbf{s} = (\mathbf{v}, \mathbf{h})$ sollte man sich vor Augen halten. Dem Zustand $\mathbf{s}$ wird die quadratische Energie $E(\mathbf{s}; W)$ zugeordnet

$$E(\mathbf{s}; W) = \sum_{ij} s_i w_{ij} s_j \tag{3.1}$$

Die zugrundeliegende Idee der Funktionsweisen ist für die Boltzmann-Maschine und das Hopfield-Modell gleich. Die Zustände minimaler Energie sollen die zu reproduzierenden binären Muster darstellen, wobei diese im Fall der Boltzmann-Maschine nur mit den sichtbaren Zuständen $\mathbf{v}$ identifiziert werden (siehe Abbildung 3.1). Man fordert für die Zustände $(\mathbf{v}, \mathbf{h})$ eine Boltzmann-Verteilung

$$P(\mathbf{v}, \mathbf{h}; W) = e^{-\beta E(\mathbf{v}, \mathbf{h}; W)} / Z \tag{3.2}$$

$$Z = \sum_{\mathbf{vh}} e^{-\beta E(\mathbf{v}, \mathbf{h}; W)} \tag{3.3}$$

wobei Z die Zustandssumme und $\beta = 1/T$ eine inverse Temperatur darstellt.

Die Summe ist über den gesamten Zustandsraum $\{1,-1\}^n$ zu erstrecken. [1] Man beachte, daß bei niedrigen Temperaturen die Verteilung (3.2) Zustände niedriger Energie bevorzugt. Das bedeutet, daß den Zuständen, die von dem System reproduziert werden sollen, eine hohe Wahrscheinlichkeit zukommt.

Obwohl diese Wahrscheinlichkeiten sich explizit aus (3.1)-(3.2) berechnen lassen, ist das für große n in der Praxis nicht realisierbar, da sich die Zustandssumme in (3.3) über 2^n Zustände erstreckt. Kappen (1994) reduziert den effektiven Zustandraum erheblich, indem er Kopplungen einführt, die einem Großteil der Zustände verschwindende Wahrscheinlichkeit zuordnet. Diese eingeschränkte Boltzmann-Maschine ähnelt dem bekannten Perceptron-Modell (Rosenblatt, 1962; Rumelhart, Hinton, und Williams, 1988). Will man aber den Zustandsraum nicht einschränken, so ist man gezwungen, die Verteilung durch eine stochastische Dynamik zu simulieren.

3.1.2 Stochastische Dynamik

Um eine Boltzmann-Verteilung zu gewährleisten, muß das stochastische, dynamische System dem Prinzip der detailierten Balance für die Übergangswahrscheinlichkeiten $W(s \to s')$ genügen (Fischer und Herz, 1991). Dieses Prinzip wird von der sogenannten Glauber-Dynamik erfüllt (Glauber, 1963)

$$W(s_i \to -s_i) = \frac{1}{2}(1 - s_i \tanh(\beta \sum_j w_{ij} s_j)) \tag{3.4}$$

Es soll hervorgehoben werden, daß für die Ableitung dieses Ausdruckes die Symmetrie-Eigenschaft der Kopplungen w_{ij} verwendet wurde. Diese Bedingung ist nicht zwingend und wird in (Schürmann, 1989) abgeschwächt. Eine weitere stochastische Dynamik, die im Fall kontinuierlicher Zustände zu einer Boltzmann-Verteilung führt, wird durch die Langevin-Gleichungen beschrieben (Skohorod, 1985). Sie ist eng mit der Glauber-Dynamik verwandt. Im wesentlichen beschreiben die Langevin-Gleichungen einen Gradientenabstieg auf der Energiefunktion der kontinuierlichen thermodynamischen Zustände. Ein additiven Rauschterm simuliert dabei die thermische Stochastizität.

3.1.3 Abkühlung

Zur Generierung einer Boltzmann-Verteilung verwendet man also im Fall großer n die Glauber-Dynamik. Jede Ausgangsverteilung von Zuständen in einem Ensemble gleicher Systeme konvergiert unter Glauber-Dynamik (3.4) gegen die stationäre Boltzmann-Verteilung. Im stationären Zustand ist das System ergodisch

[1] Es gibt eine deutliche Analogie zu den physikalischen Ising-Spinmodellen, mit denen Spingläser beschrieben werden (Fischer und Herz, 1991). Es soll in dieser Arbeit aber nur beschränkt auf diese Analogie eingegangen werden, da das Hauptinteresse in den informationstheoretischen Lernkriterien liegt.

und man kann die Ensemblestatistik durch die Zeitstatistik ersetzen. Man kann sich somit auf ein einziges Exemplar des Ensembles beschränken. Ausgehend von einem beliebigen Zustand berechnet man gemäß (3.4) einen neuen Zustand. Die Folge von Zuständen genügt dann im Stationären einer Boltzmann-Verteilung. Dieses Verfahren ist auch als Metropolis-Algorithmus bekannt (Metropolis, Rosenblut, Rosenblut, Teller, und Teller, 1953). Die so erzeugte Verteilung kann zur Berechnung der Zustandsmittelwerte verwendet werden. Wie schon angedeutet, ist man daran interessiert, Zustände hoher Wahrscheinlichkeit zu finden. Man möchte also die Verteilung bei niedrigen Temperaturen generieren. Für kleinere Temperatur benötigt die Dynamik länger bis zur Konvergenz ins Stationäre. Daher wird man die Glauber-Dynamik zunächst bei hohen Temperaturen iterieren und dann quasistationär abkühlen.[2]

3.2 Lernen mit minimaler relativer Entropie

Der Lernprozeß in der Boltzmann-Maschine soll die geeignete Energiefunktion $E(\mathbf{s}; W)$ finden, die den zu lernenden binären Mustern eine hohe Wahrscheinlichkeit zuordnet. Allgemeiner gesagt, sollte das Lernen die Kopplungen W adaptieren, so daß einem sichtbaren Zustand $\mathbf{v}$ eine vorgegebene Wahrscheinlichkeit $R(\mathbf{v})$ zugeordnet wird. Im Gegensatz zum Hopfield-Modell, kann das System hierzu eine interne Repräsentation $\mathbf{h}$ eines sichtbaren Musters $\mathbf{v}$ in den inneren Elementen erzeugen. Der Abstand der vorgegebenen Zielverteilung $R(\mathbf{v})$ von der durch die Boltzmann-Maschine erzeugten Zustandsverteilung der sichtbaren Elemente $P(\mathbf{v}; W) = \sum_{\mathbf{h}} P(\mathbf{v}, \mathbf{h}; W)$, läßt sich durch die relative Entropie messen

$$H[P(\mathbf{v}; W), R(\mathbf{v})] = \sum_{\mathbf{v}} P(\mathbf{v}; W) \ln \frac{P(\mathbf{v}; W)}{R(\mathbf{v})} \tag{3.5}$$

Dieser Ausdruck wird nun als Kostenfunktion eines in den Kopplungen W adaptiven Systems betrachtet. Zur Optimierung wird ein Gradientenabstieg in W angewendet. Die folgende Lernregel wird beispielsweise in (Herz, Krogh, und Palmer, 1990) abgeleitet

$$\Delta w_{ij} = -\varepsilon \frac{\partial H}{\partial w_{ij}} = \frac{\varepsilon}{T} \left[\langle s_i s_j \rangle_{\mathbf{v}} - \langle s_i s_j \rangle \right] \tag{3.6}$$

Diese Lernregel verwendet lediglich die Mittelwerte der paarweisen Korrelationen der zu einer Kopplung gehörigen Zustandsvariablen. Diese Korrelationsterme werden als Hebb'sches und Anti-Hebb'sches Lernen bezeichnet. Sie sind insbesondere interessant, da sie nur Information verwenden, die lokal zur Verfügung steht.

[2] Geman und Geman (1984) geben als theoretische Grenzgeschwindigkeit für eine quasistationäre Abkühlung $T(k) \geq \frac{c}{ln(1+k)}$ an, wobei k die Nummer des Iterationsschrittes und c eine noch zu bestimmende Konstante sind.

Hebb (1949) stellt als erster die Hypothese auf, daß Lernen in einem biologischen neuronalen Netz auf der Adaptation von synaptischen Kopplungen, gemäß der Korrelation von presynaptischer und postsynaptischer Aktivität, basiert. Die Mittelwerte

$$\langle f(\mathbf{s})\rangle_{\mathbf{v}} = \sum_{\mathbf{vh}} P(\mathbf{h}|\mathbf{v};W)R(\mathbf{v})f(\mathbf{s})$$

$$\langle f(\mathbf{s})\rangle = \sum_{\mathbf{vh}} P(\mathbf{h},\mathbf{v};W)f(\mathbf{s})$$

nennt man auch *fixiertes* und *freies* Mittel. Gemeint ist, daß das thermische Mittel bei fest vorgegebenen bzw. freien sichtbaren Zuständen $\mathbf{v}$ durchgeführt wird. Die bedingte Wahrscheinlichkeit $P(\mathbf{h}|\mathbf{v};W)$ bezeichnet die Verteilung der inneren Zustände $\mathbf{h}$ bei gegebenem sichtbarem Zustand $\mathbf{v}$ und ist bestimmt durch $P(\mathbf{h}|\mathbf{v};W) = P(\mathbf{v},\mathbf{h};W)/P(\mathbf{v};W)$. Diese Mittel können für kleine n explizit berechnet werden. Für große n bevorzugt man den Metropolis-Algorithmus, wobei die Glauber-Dynamik für freie bzw. für fixierte sichtbare Elemente iteriert wird. Die sichtbaren Elemente werden dabei gemäß der Zielwahrscheinlichkeit $R(\mathbf{v})$ auf die entsprechenden Zustände fixiert. Beide Methoden können natürlich sehr rechenintensiv werden. Aus diesem Grund ist es sinnvoll, die stochastische Dynamik durch eine deterministische Dynamik zu ersetzen.

3.3 Deterministische Dynamik

3.3.1 Mittelfeldgleichungen

Man beachte, daß bedingt durch die einfache Struktur der Energiefunktion (3.1), lediglich Mittel der Zustandsvariablen in die Lerngleichungen (3.6) eingehen. Es wird versucht, diese Mittelwerte zu berechnen, ohne die Boltzmann-Verteilung explizit simulieren oder berechnen zu müssen. Hierzu zieht man die Mittelfeldtheorie zu Hilfe. [3] Peterson und Anderson (1987) sowie Peterson und Hartman (1989) schlagen vor, die kostenintensive, stochastische Dynamik durch eine deterministische Dynamik der Mittelwerte zu ersetzen. Es gibt verschiedene Methoden, zu den entsprechenden Mittelfeldgleichungen zu gelangen. In Kapitel 6 wird die Sattelpunktnäherung angesprochen. In der Mittelfeldnäherung approximiert man das Mittel des Feldes $\mathbf{f} = \nabla_{\mathbf{s}} E(\mathbf{s})$ durch das Feld der mittleren Zustände. Für Spinzustände bedeutet das

$$\langle \tanh(\beta f_i)\rangle \approx \tanh(\beta \langle f_i\rangle)$$

[3] Die Mittelfeldtheorie wird allgemeiner im Kontext kombinatorischer Optimierungsaufgaben verwendet. Yuille und Kosowsky (1994) geben einen Überblick über die verschiedenen Ansätze in diesem Gebiet.

Man erhält damit Gleichungen für die mittleren Zustandsvariablen

$$\langle s_i \rangle = \tanh \left(\beta \sum_{j=1}^{n} w_{ij} \langle s_j \rangle \right) \tag{3.7}$$

Statt die stochastische Dynamik auszuführen, kann man jetzt Lösungen von (3.7) suchen. Für Temperaturen oberhalb einer kritischen Temperatur T_c hat diese Gleichung nur die triviale Lösung $\langle \mathbf{s} \rangle = 0$. Andere Lösungen können durch Fixpunktiteration gefunden werden. In der Tat ist die Gleichung 3.7 stationäre Lösung der partiellen Differenzialgleichung erster Ordnung

$$\frac{d}{dt} \langle s_i \rangle = - \langle s_i \rangle + \tanh \left(\beta \sum_{j=1}^{n} w_{ij} \langle s_j \rangle \right) \tag{3.8}$$

Die Fixpunktiteration kann als diskrete Integration dieser Gleichung mit einer Zeitdiskretisierung von $\Delta t = 1$ verstanden werden. Für diese Differentialgleichung gibt Hopfield (1982b) eine Lyapunov-Funktion an, die die Konvergenz in eine stationäre Lösung für symmetrische Kopplungen garantiert.

Für die Anwendung der Lernregel (3.6) ist eine weitere Näherung notwendig. Die Reihenfolge von thermischem Mittel und Korrelation wird vertauscht,

$$\langle s_i s_j \rangle = \langle s_i \rangle \langle s_j \rangle \tag{3.9}$$

Mit den Mittelwertgleichungen (3.7) und dieser letzten Näherung läßt sich die relative Entropie mit der Lernregel (3.6) minimieren.

3.3.2 Kritische Temperatur der Mittelfeldgleichungen

Es soll hervorgehoben werden, daß es in einem Spinsystem mit endlichem n keinen Phasenübergang gibt, der eine physikalisch bedeutungsvolle kritische Temperatur definiert. Die Mittelfeldgleichungen (3.7) werden aber für den Grenzwert $n \to \infty$ abgeleitet. Für die Gleichungen (3.7) gibt es daher mindestens zwei ausgezeichnete Phasen, die eine sinnvolle kritische Temperatur definieren: Die "paramagnetische" Phase, in der es nur eine stabile Lösung mit verschwindender "Magnetisierung" $\langle \mathbf{s} \rangle$ gibt; und eine zweite Phase, bei der vielfältige, stabile und metastabile Zustände möglich sind. Es soll hier nicht die Gültigkeit des Mittelwertansatzes diskutiert werden. Allerdings soll darauf hingewiesen werden, daß die Mittelfeldtheorie gerade in der Nähe der kritischen Temperatur eine unzureichende Beschreibung von physikalischen Spinsystemen liefert.

Die Bedeutung der kritischen Temperatur soll hier veranschaulicht werden. Da für $T \gg T_c$, $\langle \mathbf{s} \rangle = 0$ gilt, ist es zulässig, die Gleichungen (3.7) für hohe Temperaturen um den Nullpunkt zu linearisieren: $\langle \mathbf{s} \rangle = \frac{1}{T} w \langle \mathbf{s} \rangle$. Betrachtet man diese nun als Fixpunktgleichung

$$\langle \mathbf{s} \rangle (t+1) = \frac{1}{T} W \langle \mathbf{s} \rangle (t)$$

so läßt sich die Temperatur bestimmen, bei der die Lösung $\langle \mathbf{s} \rangle = 0$ instabil wird und die Magnetisierung anwächst. Die Stabilitätsbedingung lautet hierfür: $||W/T|| < 1$. Die Norm einer symmetrischen Matrix ist aber nichts anderes als der betragsmäßig größte Eigenwert λ_{max}. Der Phasenübergang tritt also bei $T_c = \lambda_{max}$ ein. Eine ähnliche Grenztemperatur findet man in einer Verallgemeinerung dieses binären Modells auf ein System mit kontinuierlichen Zuständen in Kapitel 6.

3.4 Binäre Autoassoziation

An der Aufgabe der Autoassoziation soll die Funktionsweise der deterministischen Boltzmann-Maschine verdeutlicht werden. Unter Autoassoziation versteht man die Rekonstruktion eines Signals aus einem gestörten oder verrauschten Signal. Man gibt dabei einen festen Satz zulässiger Signale vor, die abgerufen werden können. Im diskreten Zustandsraum handelt es sich im allgemeinen um binäre Muster. Man identifiziert in der Boltzmann-Maschine ein Muster mit einer festen Konfiguration $\mathbf{v}$ der sichtbaren Elemente. Den Satz binärer Muster repräsentiert man mit der zu lernenden Verteilung $R(\mathbf{v})$. Die relative Entropie als Lernkriterium adaptiert die Energiefunktion, so daß die Boltzmann-Verteilung $P(\mathbf{v})$ die vorgegebene Verteilung $R(\mathbf{v})$ möglichst gut wiedergibt, d.h. vorgegebenen Zuständen wird eine hohe Wahrscheinlichkeit zugeordnet. Die Boltzmann-Maschine relaxiert bei abnehmenden Temperaturen auf Zustände hoher Wahrscheinlichkeit. Nach dem Lernen wird sie also gegen diese vorgegebenen Zuständen konvergieren. Auch die deterministischen Gleichungen (3.7) der Mittelwerte werden gegen die gelernten Zuständen konvergieren. Neben den gelernten Mustern kann es aber noch andere stabile Zustände geben.[4] Die genaue Form des Anziehungsgebietes der gelernten Muster wird von der quadratischen Energiefunktion bestimmt. Diese wurde aber auch mit dem Ziel trainiert, nicht vorgegebenen Mustern eine verschwindende Wahrscheinlichkeit zuzuordnen und damit die Anziehungsgebiete der vorgegebenen Muster zu vergrößern. Zur Autoassoziation wird das zu rekonstruierende, gestörte Muster als Anfangswert der sichtbaren Zustände $\mathbf{v}$ der Relaxation der deterministischen Gleichungen (3.7) verwendet. Das System konvergiert zu dem nächstliegenden Fixpunkt, der gleichzeitig dem ähnlichsten der gelernten Muster entspricht.

Die Lernregeln (3.6) enthalten freie und fixierte Mittel. Diese können mit Mittelfeldgleichungen (3.7) bestimmt werden. Die fixierten Mittel erhält man, indem die sichtbaren Elemente auf einen der zu lernenden Zustände $\mathbf{v}$ fixiert werden,

[4] Viele dieser Zustände sind unter einer stochastischen Dynamik in einem System mit endlichem n lediglich metastabil. Das heißt, sie verschwinden für abnehmende Temperatur. Aufgrund der deterministischen Mittelwertdynamik, die exakt nur für die Grenzwerte $n \to \infty$ oder $T \to 0$ gilt, werden diese metastabilen Zustände in einem endlichen System stabil. Zu den hier angesprochenen Grenzwerten siehe Abschnitt 6.3.

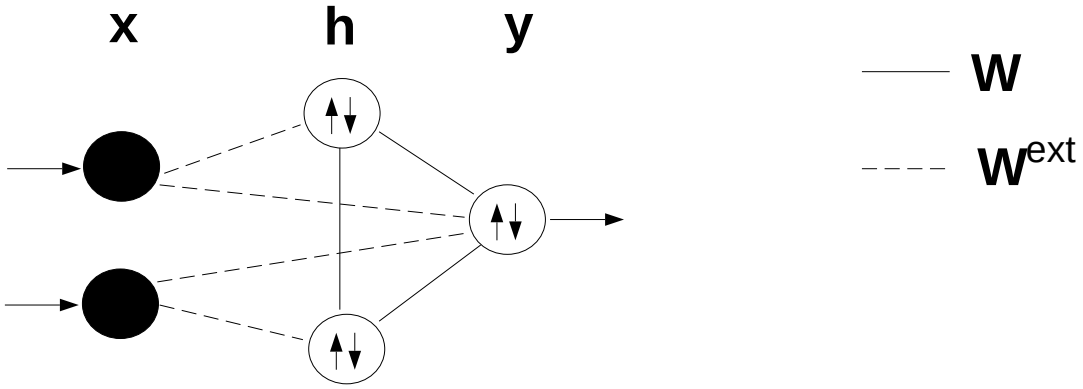

Abbildung 3.2: In der Standard-Boltzmann-Maschine unterscheidet man zwischen einer möglicherweise kontinuierlichen Eingabe $\mathbf{x}$ und den binären Zuständen, die einer Boltzmann-Verteilung genügen. Die Eingabe wird mittels einer Kopplungsmatrix W^{ext} als externes Feld an das System angebunden. Die binären Zustände werden oft weiter in Ausgabe $\mathbf{y}$ und innere Zustände $\mathbf{h}$ unterteilt.

während die restlichen Zustände gemäß der Mittelfelddynamik (3.7) relaxiert werden. Dabei wird die in Abschnitt 3.1.3 angesprochene Abkühlung simuliert. Ausgehend von einer Temperatur oberhalb der kritischen Temperatur relaxiert das System bei gleichzeitig abnehmender Temperatur. Die resultierenden Mittelwerte für die verschiedenen Lernmuster $\mathbf{v}$ werden dann entsprechend $R(\mathbf{v})$ gemittelt. Zur Bestimmung der freien Mittel wird Gleichung 3.7 mit kleinen zufälligen Startwerten initialisiert und dann ohne Einschränkungen relaxiert.

3.5 Standard-Boltzmann-Maschine

Die klassische Boltzmann-Maschine wurde von Hopfield (1987) erweitert, indem er ein konstantes externes Feld $\mathbf{f}^{ext}$ zuläßt. Über diesem läßt sich mit einer Kopplungsmatrix $w^{ext} \in \Re^n \times \Re^m$ eine kontinuierliche Eingabe $\mathbf{x} \in \Re^n$ in die Boltzmann-Maschine einführen, $f_i^{ext} = \sum_j w_{ij}^{ext} x_j$. Die sichtbaren Zustände werden jetzt als reine Ausgabe betrachtet und mit $\mathbf{y}$ benannt.

Die Zustandsvariable $\mathbf{s}$ setzt sich also aus inneren Zuständen und den Ausgabezuständen zusammen, das heißt $\mathbf{s}=(\mathbf{h},\mathbf{y})$. Die entsprechende Energiefunktion mit äußerem Feld als Eingabe lautet

$$E(\mathbf{s}|\mathbf{x}; W, W^{ext}) = \sum_{ij} s_i w_{ij} s_j + \sum_{ij} s_i w_{ij}^{ext} x_j$$

Man beachte die Schreibweise der Argumente der Energie. Die Kopplungsstärken sind Parameter der Energie. Die Eingabe $\mathbf{x}$ ist keine Zustandva-

riable des Systems. Sie genügt nicht einer Boltzmann-Verteilung sondern einer fest vorgegeben externen Verteilung $R(\mathbf{x})$. Um präziser zu sein: die Verteilung $R(\mathbf{x}, \mathbf{y})$ ist durch ein Satz von N Wertepaaren $(\mathbf{x}^1, \mathbf{y}^1), ..., (\mathbf{x}^N, \mathbf{y}^N)$ gegeben. [5]

Die Verteilung, die diese Boltzmann-Maschine generieren wird, ist jetzt eine bedingte Wahrscheinlichkeit $P(\mathbf{s}|\mathbf{x}; W, W^{ext})$. Summiert man über die inneren Zustände $\mathbf{h}$, so erhält man die Verteilung, die zur Modellierung benutzt wird, $P(\mathbf{y}|\mathbf{x}; W, W^{ext}) = \sum_{\mathbf{h}} P(\mathbf{s}|\mathbf{x}; W, W^{ext})$. Die Verteilung, die mit dem Kriterium der minimalen relativen Entropie modelliert werden soll, ist die bedingte Wahrscheinlichkeit $R(\mathbf{y}|\mathbf{x})$, einen Zustand $\mathbf{y}$ bei gegebenem $\mathbf{x}$ zu erhalten. Der Abstand dieser beiden Verteilungen wird wieder mit der relativen Entropie gemessen. Die relative Entropie wird als Kostenfunktion betrachtet und die Kopplungen w_{ij} und w_{ij}^{ext} mit einem Gradientenabstieg optimiert. Die Lerngleichungen für die inneren Kopplungen sind analog zu (3.6). Auch für die externen Kopplungen erhält man entsprechendes

$$\Delta w_{ij} = \frac{\varepsilon}{T}\left[\langle s_i s_j\rangle_{\mathbf{y},\mathbf{x}} - \langle s_i s_j\rangle_{\mathbf{x}}\right] \tag{3.10}$$

$$\Delta w_{ij}^{ext} = \frac{\varepsilon}{T}\left[\langle s_i x_j\rangle_{\mathbf{y},\mathbf{x}} - \langle s_i x_j\rangle_{\mathbf{x}}\right] \tag{3.11}$$

Die thermischen Mittel werden jetzt für feste Eingaben $\mathbf{x}$ gebildet und dann zusätzlich über die Eingabeverteilung $R(\mathbf{x})$ gemittelt

$$\langle f(\mathbf{s},\mathbf{x})\rangle_{\mathbf{y},\mathbf{x}} = \sum_{\mathbf{x}\in\Xi}\sum_{\mathbf{yh}} P(\mathbf{h}|\mathbf{y},\mathbf{x}; W, W^{ext}) R(\mathbf{y},\mathbf{x}) f(\mathbf{s})$$

$$\langle f(\mathbf{s},\mathbf{x})\rangle_{\mathbf{x}} = \sum_{\mathbf{x}\in\Xi}\sum_{\mathbf{yh}} P(\mathbf{h},\mathbf{y}|\mathbf{x}; W, W^{ext}) R(\mathbf{x}) f(\mathbf{s})$$

Die vorgegebene gemeinsame Verteilung $R(\mathbf{y},\mathbf{x}) = R(\mathbf{y}|\mathbf{x})R(\mathbf{x})$ definiert die zu lernende Relation zwischen Eingabe und Ausgabe. Der Relaxationsprozeß unterscheidet sich vom bisher beschriebenen Modell. Die Eingabe in das System bleibt im Relaxationsprozeß stets konstant, während bisher die Eingabe lediglich für das fixierte Mittel festgehalten wurde. Die Ausgabe wird wie bisher für das

[5] Die Eingabedaten $\mathbf{x}^i$ mögen gemäß einer kontinuierlichen Wahrscheinlichkeitsdichte $r(\mathbf{x})$ gezogen worden sein. In der deterministischen Boltzmann-Maschine benötigt man Summen über die gegebenen Verteilungen. Da man diese Dichte nicht vollständig gegeben hat, sondern nur einen Satz von Datenpunkten $\Xi = \{\mathbf{x}^1, ..., \mathbf{x}^N\}$, können diese Summen näherungsweise als Summe über die gegebenen Datenpunkte ersetzt werden.

$$< f(\mathbf{x}) >= \int r(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \approx 1/N \sum_{\mathbf{x}\in\Xi} f(x)$$

Analog hierzu kann man aber auch die diskrete Verteilung $R(\mathbf{x}) = 1/N$ mit $\mathbf{x} \in \Xi$ als diskrete Näherung von $r(\mathbf{x})$ betrachten. Man kann somit die kontinuierliche Eingabe als diskret verteilt behandeln.

fixierte Mittel auf die zu lernenden Zustände festgehalten. Für das freie Mittel hingegen soll die Ausgabe der generierten Boltzmann-Verteilung genügen, und man überläßt sie der Dynamik der Mittelfeldgleichungen. Die Mittelfeldgleichungen für eine feste externe Eingabe ergeben sich jetzt zu

$$\langle s_i \rangle = \tanh \left(\beta \sum_{j=1}^{n} w_{ij} \langle s_j \rangle + \beta \sum_{j=1}^{n} w_{ij}^{ext} x_j \right) \tag{3.12}$$

Das Mittel über $\mathbf{x}$ erhält man, indem (3.12) für alle $\mathbf{x} \in \Xi$ auswertet und gemäß $R(\mathbf{x})$ mittelt.

Die Eingabe bestimmt in diesem Modell die Energiefunktion und nicht, wie bisher, den Startwert der Relaxation. Dieser grundlegend verschiedene Ansatz macht die Abrufphase dieser Version der Boltzmann-Maschine wesentlich stabiler. Darin liegt wohl der Grund, daß sich diese Interpretation der Boltzmann-Maschine mit kontinuierlicher Eingabe und binärer Ausgabe zum Standard entwickelt hat.

Benutzt man die beschriebene deterministische Dynamik gemäß (3.12), wird diese Boltzmann-Maschine für eine präsentierte Eingabe nur einen einzigen Zustand an der Ausgabe erzeugen. Es macht somit keinen Sinn, der deterministischen Boltzmann-Maschine nicht-deterministische Relationen als Lernziel vorzugeben. Man wird sich vielmehr auf deterministische Abbildungen $f: \Re^m \rightarrow \{0,1\}^n$ beschränken. Das System wird somit eine Funktionenapproximation vom kontinuierlichen in den binären Zustandsraum verwirklichen.

In der konnektionistischen Forschung bezeichnet man die Funktionsapproximation als *überwachtes Lernen*, da die Ausgabe, die das System erzeugen soll, explizit vorgegeben ist. Im Gegensatz hierzu steht das *unüberwachte Lernen*, bei dem man dem konnektionistischen und adaptiven System lediglich eine Eingabe zur Verfügung stellt. Es bleibt dem System dann frei überlassen, welche Repräsentation es findet, um ein gewisses Lernkriterium zu erfüllen. Mit Konzepten zum unüberwachten Lernen beginnt im nächsten Kapitel die Beschreibung der in dieser Arbeit neu entwickelten Lernmethoden. Diese werden hier erstmalig in einem rückgekoppelten, nichtlinearen System wie der Boltzmann-Maschine angewendet.

Kapitel 4

Maximale Transinformation im Diskreten

Welches Kriterium für das unüberwachte Lernen ist von Nutzen? In Abschnitt 2.3 wurde dargelegt, daß ein Ziel für das unüberwachte Lernen die Modellierung der Verteilung von beobachteten Signalen sein kann. In dieser Arbeit wird angenommen, daß man keinerlei Anhaltspunkte über die Struktur der beobachteten Signale hat. Man hat insbesondere keine Angaben über das Rauschen oder die Statistik eines ungestörten Signals. Unter diesen Bedingungen ist es am sinnvollsten, bei der Modellierung eine Repräsentation zu finden, die maximale Information der Umwelt enthält.

Informationstheoretische Konzepte wurden gleichzeitig von mehreren Autoren eingeführt mit dem Ziel, unüberwachte Lernmethoden zu untersuchen (Becker, 1992; Bridle, 1989; Bridle et al., 1991; Linsker, 1988, 1989, 1992). Das Optimierungsprinzip von Linsker (1988, 1989, 1992) - "Infomax" genannt - verwirklicht die eben vorgeschlagene Idee. Linsker adaptiert die Kopplungskoeffizienten unter gewissen Grenzbedingungen, so daß die Transinformation zwischen der Eingabe und Ausgabe maximiert wird. (Atick und Redlich, 1990) zeigen, daß herausragende Merkmale sich aus verrauschten Eingabesignalen mit dem Prinzip der maximalen Transinformation extrahieren lassen. Es wurden einige Algorithmen zur Maximierung von Transinformation in probabilistischen linearen Systemen (Linsker, 1992) und für das sogenannte "winner-take-all" Netzwerk (Linsker, 1989) entwickelt. Földiák (1989) zeigt, daß im Fall von linearen deterministischen Transformationen (d.h. kein Rauschen an der Ausgabe) und für unkorreliertes Gaußsches Rauschen in der Eingabe das "Infomax"-Prinzip äquivalent zur Hauptkomponentenanalyse (PCA) ist. Das Ziel in diesem Teil der Arbeit ist, das Prinzip der maximalen Transinformation auf die Boltzmann-Maschine anzuwenden. Man verwirklicht dadurch nichtlineare Merkmalsextraktion in einem rückgekoppelten konnektionistischen System, das eine interne Repräsentation der extrahierten Merkmale zu Hilfe nehmen kann.

Der resultierende Lernalgorithmus enthält wie in der ursprünglichen

Boltzmann-Maschine einen Hebb'schen und Anti-Hebb'schen Term. Diese werden jetzt mit einem Informationsausdruck gewichtet. Der Algorithmus wird an einem Kodierungsproblem erfolgreich getestet. Man erhält hierbei optimale Datenkompression. Mit der vorgestellten Methode können auch kontinuierliche Signale unüberwacht klassifiziert werden. Das System unterteilt den Eingaberaum in Klassen, so daß die diskrete Ausgabe maximale Information enthält. Auch Bridle et al. (1991) präsentiert das Kriterium der maximalen Transinformation zur unüberwachten Klassifikation. Der Vorteil der hier vorgeschlagenen Implementation in der Boltzmann-Maschine liegt in der exponentiell reduzierten Anzahl der benutzten Kopplungsparameter. Die von Bridle et al. (1991) vorgeschlagene Architektur benötigt für die Klassifikation von m-dimensionalen Eingabesignalen in c Klassen cm Kopplungskoeffizienten. Hier benötigt man dazu nur $m \log_2(c)$ Koeffizienten.

4.1 Das probabilistische Modell

Es wird die Standard-Boltzmann-Maschine benutzt, so wie sie im vorhergehenden Kapitel in Abschnitt 3.5 beschrieben wurde. Die binäre Zustandsvariable $\mathbf{s} \in \{1,-1\}^n$ setzt sich aus inneren Zuständen $\mathbf{h} \in \{1,-1\}^{n_h}$ und den Ausgabezuständen $\mathbf{y} \in \{1,-1\}^{n_y}$ zusammen, d.h. $\mathbf{s} = (\mathbf{h},\mathbf{y})$ mit $n = n_h + n_y$. Das Eingabesignal ist durch die Datenpunkte $\mathbf{\Xi} = \{\mathbf{x}^1,...,\mathbf{x}^N\}$ gegeben. Diese sind möglicherweise kontinuierliche m-dimensionale Signale, $\mathbf{x} \in \Re^m$, die gemäß einer Wahrscheinlichkeitsdichte $r(\mathbf{x})$ gemessen wurden. Diese Wahrscheinlichkeitsdichte kann näherungsweise durch eine diskretisierte Verteilung $R(\mathbf{x})$ mit $\mathbf{x} \in \mathbf{\Xi}$ charakterisiert werden (siehe Fußnote in Seite 32).

Die folgende Energiefunktion koppelt mittels der Kopplungskoeffizienten $W^{ext} \in \Re^n \times \Re^m$ die Eingabe als äußeres Feld an das thermodynamische System, das von den Wechselwirkungskoeffizienten $W \in \Re^n \times \Re^n$ bestimmt wird.[1]

$$E(\mathbf{s}|\mathbf{x}) = \sum_{ij} s_i w_{ij} s_j + \sum_{ij} s_i w^{ext}_{ij} x_j \tag{4.1}$$

Die Verteilung der binären Zustände $\mathbf{s}$ soll bei einer Eingabe $\mathbf{x}$ weiterhin gegeben sein durch

$$\begin{aligned} P(\mathbf{s}|\mathbf{x}) &= e^{-\beta E(\mathbf{s}|\mathbf{x})}/Z \\ Z &= \sum_{\mathbf{yh}} e^{-\beta E(\mathbf{s},\mathbf{h})} \end{aligned} \tag{4.2}$$

[1] Im folgenden wird die explizite Abhängigkeit der Energie und somit der Verteilungen und Informationsausdrücke von den Kopplungsparametern W und W^{ext} in der Notation vernachlässigt. Dies soll die Lesbarkeit der Ausdrücke erleichtern.

Die Verteilung der Ausgabe $\mathbf{y}$ wird nicht wie bisher vorgegeben. Statt dessen wird jetzt ein Optimalitätskriterium für die Verteilung der Ausgabezustände festgelegt. Dem System bleibt frei, eine Repräsentation an der Ausgabe zu finden, die dieses Kriterium optimiert.

4.2 Lernen mit maximaler Transinformation

Die Kopplungskoeffizienten sollen so adaptiert werden, daß die Transinformation zwischen Eingabe $\mathbf{x}$ und Ausgabe $\mathbf{y}$ maximiert wird. Dazu formuliert man zunächst die Transinformation, die in Abschnitt 2.3 definiert wurde als Funktion der bedingten Wahrscheinlichkeiten $P(\mathbf{s}|\mathbf{x})$. Die Definition 2.4 läßt sich wie folgt umschreiben,

$$\begin{aligned} I[P(\mathbf{x},\mathbf{y})] &= H[P(\mathbf{y})] + H[R(\mathbf{x})] - H[P(\mathbf{y},\mathbf{x})] && (4.3)\\ &= H[P(\mathbf{y})] - H[P(\mathbf{y}|\mathbf{x})] && (4.4)\\ &= -\sum_{\mathbf{xy}} P(\mathbf{y},\mathbf{x}) \ln P(\mathbf{y}) + \sum_{\mathbf{xy}} P(\mathbf{y},\mathbf{x}) \ln P(\mathbf{y}|\mathbf{x}) && (4.5) \end{aligned}$$

Die Summe in $\mathbf{x}$ erstreckt sich über die gegebenen Datenpunkte $\boldsymbol{\Xi}$. Der erste Term in (4.4) stellt die Entropien der Ausgabe dar. Der zweite Term mißt im wesentlichen die Entropie, die bei der Transformation erzeugt wird. Die von der Eingabe zur Ausgabe transferierte Information ist offensichtlich die Differenz dieser beiden Ausdrücke. Hat man eine deterministische Abbildung, bei der einem Eingabewert $\mathbf{x}$ stets der gleiche Ausgabewert $\mathbf{y}$ zugeordnet wird, so verschwindet der Entropieterm $H[P(\mathbf{y}|\mathbf{x})]$, der in (4.4) die Unsicherheit der Abbildung mißt. Die Transinformation (4.4) hat somit als obere Grenze die Entropie der Ausgabe. Der Ausdruck für die Transinformation in (4.3) ist symmetrisch in $\mathbf{x}$ und $\mathbf{y}$. Es ist von daher leicht einsichtig, daß auch die Entropie der fest vorgegebenen Eingabe die obere Grenze darstellen kann.

$$I[P(\mathbf{x},\mathbf{y})] \leq I[R(\mathbf{x})] \qquad (4.6)$$

Man konsultiere hierzu das Kapitel über gemeinsame Information in (Papoulis, 1991). Die Umformung von (4.3) nach (4.4) basiert auf der Definition der Entropie $H[P(\mathbf{y}|\mathbf{x})]$ einer bedingten Wahrscheinlichkeit $P(\mathbf{y}|\mathbf{x})$

$$\begin{aligned} H[P(\mathbf{y}|\mathbf{x})] &= -\sum_{\mathbf{xy}} P(\mathbf{y},\mathbf{x}) \ln P(\mathbf{y}|\mathbf{x})\\ &= +\sum_{\mathbf{xy}} P(\mathbf{y},\mathbf{x}) \ln R(\mathbf{x}) - \sum_{\mathbf{xy}} P(\mathbf{y},\mathbf{x}) \ln P(\mathbf{y},\mathbf{x})\\ &= -H[R(\mathbf{x})] + H[P(\mathbf{y},\mathbf{x})] \end{aligned}$$

Die feste obere Grenze (4.6) macht die Maximierung der Transinformation zu einem sinnvollen Maximalitätsprinzip. Mit den folgenden Ersetzungen läßt sich die Kostenfunktion der relativen Entropie vollständig als Funktion der Verteilung $P(\mathbf{s}|\mathbf{x})$ darstellen.

$$\begin{aligned} P(\mathbf{y},\mathbf{x}) &= P(\mathbf{y}|\mathbf{x})R(\mathbf{x}) && (4.7)\\ P(\mathbf{y}) &= \sum_{\mathbf{x}} P(\mathbf{y},\mathbf{x}) && (4.8)\\ P(\mathbf{y}|\mathbf{x}) &= \sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h}|\mathbf{x}) && (4.9) \end{aligned}$$

Mit den Ersetzungen (4.7) und (4.8) ergibt sich die Transinformation (4.5) zu

$$I[P(\mathbf{x},\mathbf{y})] = -\sum_{\mathbf{xy}} R(\mathbf{x})P(\mathbf{y}|\mathbf{x}) \ln \sum_{\mathbf{x}} R(\mathbf{x})P(\mathbf{y}|\mathbf{x}) + \sum_{\mathbf{xy}} R(\mathbf{x})P(\mathbf{y}|\mathbf{x}) \ln P(\mathbf{y}|\mathbf{x}) \tag{4.10}$$

Man beachte, daß $P(\mathbf{s}|\mathbf{x})$ gemäß Gleichungen (4.1) und (4.2) von den Kopplungskoeffizienten abhängt. Nach (4.9) ist $P(\mathbf{s}|\mathbf{x})$ und somit auch die Transinformation 4.10 eine explizite Funktion der Kopplungskoeffizienten

$$I[P(\mathbf{x},\mathbf{y})] = I[P(\mathbf{s}|\mathbf{x}), R(\mathbf{x})] = I[R(\mathbf{x}); w, w^{ext})]$$

Diese explizite Definition der Wahrscheinlichkeitsverteilung als Funktion der zu adaptierenden Systemparameter macht es möglich, informationstheoretische Konzepte, wie das der maximalen Transinformation, durch eine analytische Optimierungsfunktion darzustellen. Das erlaubt, zur Optimierung der Parameter Gradientenmethoden zu verwenden. Das einfachste Gradientenverfahren ist in diesem Fall der Gradientenaufstieg

$$\begin{aligned} \Delta w_{ij} &= \epsilon \frac{\partial I[R(\mathbf{x}); w, w^{ext})]}{\partial w_{ij}}\\ \Delta w_{ij}^{ext} &= \epsilon \frac{\partial I[R(\mathbf{x}); w, w^{ext})]}{\partial w_{ij}^{ext}} \end{aligned}$$

In jedem Lernschritt werden die Kopplungsparameter in Richtung des Gradienten mit einer Schrittweite proportional zu einer Lernkonstante ϵ um Δw_{ij} bzw. Δw_{ij}^{ext} inkrementiert. Für die Ableitung der Transinformation (4.10) nach einer beliebigen Kopplungsstärke w ergibt sich

$$\frac{\partial I[R(\mathbf{x}); w, w^{ext})]}{\partial w} = -\sum_{\mathbf{xy}} R(\mathbf{x}) \frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x}) \ln \sum_{\mathbf{x}} R(\mathbf{x})P(\mathbf{y}|\mathbf{x}) \tag{4.11}$$

$$- \sum_{\mathbf{xy}} R(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) \frac{\sum_{\mathbf{x}} R(\mathbf{x}) \frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x})}{\sum_{\mathbf{x}} R(\mathbf{x}) P(\mathbf{y}|\mathbf{x})} \tag{4.12}$$

$$+ \sum_{\mathbf{xy}} R(\mathbf{x}) \frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x}) \ln P(\mathbf{y}|\mathbf{x}) \tag{4.13}$$

$$+ \sum_{\mathbf{xy}} R(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) \frac{\frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y}|\mathbf{x})} \tag{4.14}$$

$$= \sum_{\mathbf{xy}} R(\mathbf{x}) \frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x}) \ln \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})} \tag{4.15}$$

Die Ableitungen innerhalb der Logarithmen haben sich für den ersten und den zweiten Term in (4.10) aufgehoben. Aus (4.9), (4.2) und (4.1) erhält man die Ableitungen der bedingten Wahrscheinlichkeit nach einem bestimmten Wechselwirkungskoeffizienten w_{ij}. Diese sind für die Standard-Boltzmann-Maschine bereits berechnet worden (siehe Herz et al., 1990)

$$\frac{\partial}{\partial w_{ij}} P(\mathbf{y}|\mathbf{x}) = \beta \sum_{\mathbf{h}} P(\mathbf{s}|\mathbf{x}) \left(s_i s_j - \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}) s_i s_j \right) \tag{4.16}$$

Analog gilt für die externen Kopplungskoeffizienten w_{ij}^{ext},

$$\frac{\partial}{\partial w_{ij}^{ext}} P(\mathbf{y}|\mathbf{x}) = \beta \sum_{\mathbf{h}} P(\mathbf{s}|\mathbf{x}) \left(s_i x_j - \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}) s_i x_j \right) \tag{4.17}$$

Man behalte die Äquivalenz $\mathbf{s} = (\mathbf{h}, \mathbf{y})$ vor Augen. Kombiniert man die Gleichungen (4.15), (4.16), (4.17) und verwendet die Schreibweise für das freie Mittel, das im vorhergehenden Kapitel in Abschnitt 3.5 eingeführt wurde, so ergeben sich die Lerngleichungen zu:

$$\Delta w_{ij} = \epsilon \beta \sum_{\mathbf{x}} R(\mathbf{x}) \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}) \ln \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})} \left(s_i s_j - \langle s_i s_j \rangle_{\mathbf{x}} \right) \tag{4.18}$$

$$\Delta w_{ij}^{ext} = \epsilon \beta \sum_{\mathbf{x}} R(\mathbf{x}) \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}) \ln \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})} \left(s_i x_j - \langle s_i x_j \rangle_{\mathbf{x}} \right) \tag{4.19}$$

Die abgeleiteten Lerngleichungen haben eine interessante Interpretation. Ein Hebb'scher Term $s_i s_j$ wird durch die unmittelbare Korrelation zweier Zustände gegeben. Der zweite Anti-Hebb'sche Term $-\langle s_i s_j \rangle_{\mathbf{x}}$ liefert die mittlere Korrelation zwischen zwei Zuständen. Beide Terme werden über alle möglichen Zustände $\mathbf{s}$ und verschiedene Eingabedaten $\mathbf{x}$ summiert und mit einer Messung der Information, die von einem Eingabezustand auf einen Ausgabezustand übertragen wird, gewichtet.

Die Hebb'schen Term erinnern an die Lerngleichungen (3.10) der überwachten Standard-Boltzmann-Maschine. Die zu berechnenden Zustandssummen weichen

jedoch entscheidend voneinander ab. Der Hebb'sche und Anti-Hebb'sche Term liefern in der Standard-Boltzmann-Maschine den Unterschied in den Korrelationen, die für die fixierten und freien Phase gemittelt werden. Diese Unterscheidung von fixierter (überwachter) und freier Phase existiert hier nicht. Bei dem vorliegenden unüberwachten Lernen, müssen die Ausgabezustände nicht künstlich auf bestimmte Werte fixiert werden. Es handelt sich bei beiden Hebb'schen Termen um ein freies Mittel über die Verteilung aller möglichen Zustände $P(\mathbf{s})$.

4.3 Simulationen und Resultate

4.3.1 Implementation und Komplexität des Modells

Man beachte, daß die Gewichtungsterme in den Lerngleichungen (4.18) und (4.19) jetzt die explizite Berechnung der verschiedenen Wahrscheinlichkeiten notwendig machen. In der Abrufphase, d.h. nach dem Lernen, kann man zur Berechnung der Zustände weiterhin die Mittelfeldgleichungen (3.12) verwenden. Während dem Lernen kann man aus den ohnehin benötigten Wahrscheinlichkeiten leicht den exakten Mittelwert bestimmen. Man muß die Wahrscheinlichkeiten $P(\mathbf{s}|\mathbf{x})$ für alle 2^n möglichen Zustände $\mathbf{y}$ kennen. Eine stochastische Dynamik zur Simulation von $P(\mathbf{s}|\mathbf{x})$ ist nicht sinnvoll, da man diese direkt gemäß (4.9), (4.2) und (4.1) bestimmen kann.

Mit einer m-dimensionalen Eingabe benötigt man zur Berechnung der Energie eines Zustands n^2+nm Operationen. Bei einem Satz von N Datenpunkten ist die Komplexität zur Berechnung eines Gradientenschrittes $O((n^2+nm)N2^n)$. Dies ist äquivalent zur Standard-Boltzmann-Maschine ohne der Mittelwertdynamik.

Das unüberwachte System kann zu Klassifikation von m-dimensionalen Signalen in c verschiedenen Klassen verwendet werden. Das "maximum likelihood"-Netzwerk oder das unüberwachte Modell von Bridle et al. (1991) benötigen für diese Aufgabe eine c-dimensionale Ausgabe. Diese ergibt mc Kopplungskoeffizienten und resultiert in einem Algorithmus der Komplexität $O(mcN)$. Das vorliegende Modell benötigt zur Kodierung der c Klassen nur $\log_2(c)$ binäre Ausgabezustände. Die resultierende Komplexität des Algorithmus ist etwas höher $O(\log_2(c)mcN)$. Die Anzahl der benötigten Kopplungskoeffizienten ist allerdings nur $m\log_2(c)$.

Die Kopplungskoeffizienten w der binären Zustände realisieren eine Rückkopplung zwischen den Ausgabezuständen und den inneren Zuständen. Sie vermitteln aber auch eine Wechselwirkung zwischen den Ausgabezuständen. Als Spezialfall ist es auch möglich, die inneren Zustände oder Wechselwirkung an der Ausgabe zu eliminieren ($n_h = 0$ bzw. $w_{ij} = 0$ für $i,j \leq n_y$). Verwendet man keine Rückkopplung ($w_{ij} = 0$ für beliebige i,j), so vereinfachen sich die Mittelfeldgleichungen in der Abrufphase zu einem einschichtigen Perceptron-Modell mit sigmoiden Ausgabeelementen (siehe Herz et al., 1990). Für die Grenztemperatur

$T \to 0$ wird die Boltzmann-Maschine insofern deterministisch, als die Mittelwerte gegen $+1$ oder -1 konvergieren. Dies entspricht einem Perceptron-Modell mit einer Stufenfunktion an den Ausgabeelementen. Die Anwendungsbeispiele, die nun folgen, benötigten keine inneren Zustände ($n_h = 0$). Zumeist wird aber eine Wechselwirkung an der Ausgabe zugelassen.

4.3.2 Rekodierung und Kompression binärer Signale

Um das Lernkriterium und die abgeleiteten Lerngleichungen zu überprüfen, wird zunächst das Kodierungsproblem, das auch von Ackley et al. (1985) in der klassischen Boltzmann-Maschine benutzt wurde, untersucht. Obwohl sich hier die Eingabe auf binäre Signale beschränkt, ist dieses Experiment interessant, um zu überprüfen, ob das Lernkriterium zu den theoretisch bestmöglichen Ergebnissen konvergiert. Im $N \to n$ Kodierungsproblem sollen N verschiedene binäre Muster mit je N Bits, von denen nur eines den Wert $+1$ annimmt, auf einen binären Kode mit n Bits abgebildet werden. Im allgemeinen ist $n < N$. Es werden gleiche Wahrscheinlichkeiten für die verschiedenen Eingabemuster angenommen, d.h. $R(x) = 1/N$. Hier wird ein System mit n nicht wechselwirkenden Ausgabezuständen ($w_{ij} = 0$) und ohne innere Zustände ($n_h = 0$) gewählt. Es wurde eine inverse Temperatur $\beta = 1$ und eine Lernrate $\epsilon = 0.1$ für alle nun folgenden Experimente verwendet.

Für das $4 \to 2$ und $5 \to 3$ Kodierungsproblem konvergiert die Transinformation nach etwa 1000 Gradientenschritten. Man beginnt das Gradientenverfahren mit kleinen, zufällig gewählten Kopplungsstärken ($-0.001 < w_{ij}^{ext} < 0.001$). Zu Beginn sind alle möglichen Ausgabezustände auf Grund der kleinen Kopplungsstärken annähernd gleichverteilt, $P(y|x) = P(y) = 1/2^n$. Wie in Abbildung 4.1 ersichtlich, ist die Transinformation daher anfangs gleich Null und steigt nach etwa 1000 Gradientenschritte asymptotisch zu den theoretischen maximalen Werten $\log_2(2)$ und $\log_2(3)$.

Es ist interessant festzustellen, daß nach dem Lernen eine perfekte Datenkompression erreicht wurde. Die gefundenen Kodierungen und die entsprechenden Wahrscheinlichkeiten vor und nach dem Lernen für das $5 \to 3$ Problem sind in Tabelle 4.1 zusammengefaßt.

Erlaubt man Wechselwirkung zwischen den Ausgabezuständen ($w_{ij} \neq 0$), so erhält man ein qualitativ gleiches Resultat. Man kann diese Wechselwirkung als *laterale Inhibition* interpretieren. In diesem Experiment verbessert diese die Konvergenzgeschwindigkeit zu dem Wert der maximalen Transinformation. Das $5 \to 5$, $8 \to 3$ und $40 \to 6$ Kodierungsproblem wurde mit Wechselwirkung zwischen den Ausgabezuständen perfekt gelöst. Wird eine redundante Anzahl von Zustände an der Ausgabe verwendet, wie im $5 \to 5$ Problem, so versucht das Kriterium der maximalen Transinformation, eine verteilte Darstellung der Signale zu finden. Die Signale werden also in jedem Fall rekodiert. Ein ähnlicher Effekt wurde von Ackley et al. (1985) in der klassischen Boltzmann-Maschine für die

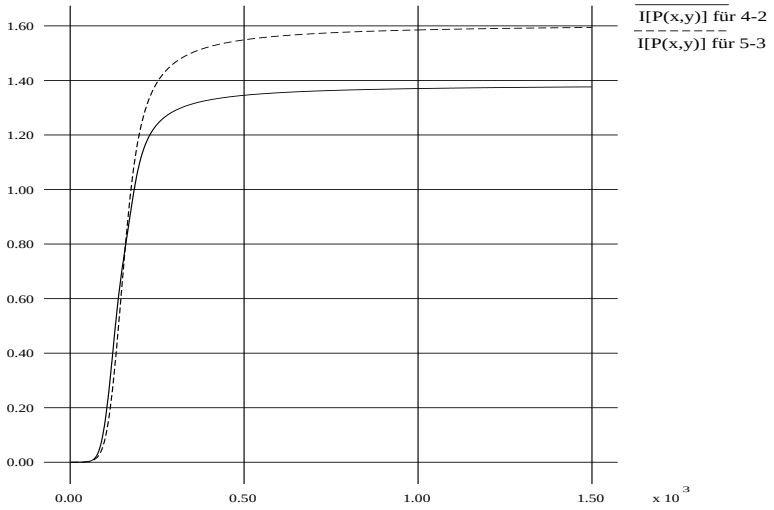

Abbildung 4.1: Entwicklung der Transinformation während dem Lernen mit maximaler Transinformation für das Kodierungsproblem $4 \rightarrow 2$ und $5 \rightarrow 3$. Die vertikale Achse kennzeichnet Transinformation. Die horizontale Achse kennzeichnet die Anzahl Gradientenschritte.

P(x)	x	y	P(y) vorher	P(y/x) nachher	P(y) nachher
0.2	$---\,-+$	$+--$	0.124	0.99	1.0
0.2	$---+-$	$+++$	0.126	1.0	1.0
0.2	$--+--$	$-+-$	0.125	0.98	1.0
0.2	$-+---$	$--+$	0.125	1.0	1.0
0.2	$+----$	$-++$	0.124	1.0	1.0
		$---$	0.124	0.0	0.0
		$+-+$	0.126	0.0	0.0
		$++-$	0.125	0.0	0.0

Tabelle 4.1: Kodierungen und die entsprechenden Wahrscheinlichkeiten vor und nach dem Lernen für das $5 \rightarrow 3$ Kodierungsproblem. Die Vorzeichen '+' und '−' kennzeichnen die binären Zustände $+1$ und '-1 für die fünf Eingabe- und drei Ausgabeelemente.

inneren Zustände diskutiert.

Um die Effizienz der vorgestellten Methode zu beurteilen, kann man die Ergebnisse für eines der größeren Beispiele mit denen der überwachten Boltzmann-Maschine vergleichen. In (Ackley et al., 1985) wird bei dem $40 \rightarrow 10$ Problem eine Fehlerrate von 1.2% erreicht. Mit dem vorliegenden Modell war es möglich, die maximale Kompression $10 \rightarrow 6$ mit einer vernachlässigbaren Fehlerrate von 0.03% zu verwirklichen.

4.3.3 Kompression von nicht gleichverteilten Daten

Ziel dieses Abschnitts ist es, eine nichtlineare, umkehrbare Abbildung zwischen einer redundanten Kodierung und einer komprimierten Kodierung zu finden. Die verschiedenen Muster sollen dabei mit verschiedener Häufigkeit auftreten, d.h. $R(\mathbf{x}) \neq 1/N$. Diese Nebenbedingung erweist sich bei der Datenkompression im Empirischen als entscheidende Erschwernis. Die Übertragung der gesamten Information wird für einen diskreten Kode durch die Umkehrbarkeit der Abbildung garantiert. Man verwendet hier wieder einen redundanten Kode mit *lokaler* Repräsentation. Das bedeutet, daß in den binären Eingabemustern nur ein einziges Element auf $+1$ und alle anderen auf -1 gesetzt sind. Da bei der Kompression die Information erhalten bleiben soll, erwartet man, daß die relevanten Eigenschaften des Signals extrahiert werden.

Merkmalsextraktion wird ausdrücklich erst im nächsten Kapitel behandelt. Es wird aber gezeigt, daß auch dieses System bereits Merkmalsextraktion realisiert. Wie in Abschnitt 2.3 angesprochen, kann man unter Merkmalen statistisch unabhängige Eigenschaften verstehen. Eine Messung für statistische Unabhängigkeit ist, wie dargelegt wurde, die *gemeinsame Information*. Für binäre Signale verwenden manche Autoren statt dessen den Begriff *Redundanz*. In dieser Arbeit versteht man unter Redundanz die "normierte" gemeinsame Information (2.8)

$$R[P(\mathbf{y})] = \frac{\sum_i^n H[P(y_i)] - H[P(\mathbf{y})]}{H[P(\mathbf{y})]} \tag{4.20}$$

Im Fall binärer Zustände bezeichnet man die Summe der Entropien der Einzelzustände als *Bit-Entropie*. Diese Nomenklatur geht auf Barlow, Kaushal, und Mitchison (1989) zurück. [2] Die Bit-Entropie läßt sich als Funktion der Wahrscheinlichkeiten P_i ausdrücken. P_i kennzeichnet die Wahrscheinlichkeit, das $i$te Element auf dem Zustand $+1$ anzutreffen.

[2] Der Begriff "normiert" mag irreführend sein. Die obere Grenze für die Redundanz, so wie hier definiert, ist nicht konstant in der Dimension. Man betrachte zum Beispiel n binäre Zustände, die vollkommen korreliert sind. Die maximale Bit-Entropie ist dann $n \ln(2)$ und die Gesamtentropie $\ln(2)$. Die Redundanz ist in diesem Fall $R = (n-1)/n$ und variiert daher mit der Dimension n.

Kodierung	Bit-Entropie	Entropie	Redundanz
lokal am Eingang	4.33	2.99	0.45
komprimiert am Ausgang	2.99	2.99	0.0

Tabelle 4.2: Informationsgehalt und Redundanz nach dem Maximieren der Transinformation für eine $8 \rightarrow 3$ Kompression von binären Signalen, die gemäß geometrischer Progression verteilt sind. Die gesamte Entropie der Eingabe (2.99) wurde auf die Ausgabe übertragen.

$$\sum_{i}^{n} H[P(y_i)] = -\sum_{i} P_i \ln(P_i) - \sum_{i}(1-P_i)\ln(1-P_i) \tag{4.21}$$

Aus der Definition der Transinformation (4.4) wird deutlich, daß zur Maximierung der Transinformation auch die Entropie an der Ausgabe vergrößert werden kann. Dies wiederum erniedrigt gemäß (2.7) die gemeinsame Information und somit die Redundanz des Ausgabekodes.

In den nächsten zwei Abschnitten wird Datenkompression für binäre Signale mit verschiedenen Verteilungen $R(\mathbf{x})$ gezeigt. Die resultierende Redundanz des komprimierten Kodes wird angegeben.

Gemäß geometrischer Progression verteilte binäre Signale

Die Verteilung der redundanten, lokalen Kodierung soll jetzt einer geometrischen Progression genügen: $P(\mathbf{x}_i) = ka^i$. Es wird $a = 0.95$ und $i = 1, .., 8$ gewählt. Das sind also acht Eingabezustände, und die Normierungskonstante k ist entsprechend zu wählen. An der Ausgabe wird die minimale Anzahl von drei Zuständen zugelassen. Hentschel und Barlow (1991) untersuchen diese Verteilung im gleichen Kontext. Tabelle 4.2 zeigt, daß die Boltzmann-Maschine nach dem Lernen an den drei Ausgabeelementen die gesamte Information der acht Eingabeelement reproduziert, nun allerdings in einer nicht redundanten Kodierung. Die vollständige Reduktion der Redundanz ergibt sich nur aufgrund der Datenkompression. Die gefundene Rekodierung ist in Tabelle 4.3 zu sehen.

Gemäß Potenzgesetz verteilte binäre Signale

Wie Hentschel und Barlow (1991) hervorheben, taucht in der Praxis auch eine Potenzgesetzverteilung häufig auf. Die Häufigkeit der Benutzung von Worten aus der englischen Sprache z.B. genügt einer solchen Verteilung. Die Eingabezustände sollen jetzt also gemäß $P(\mathbf{x}_i) = ki^{-2}$ mit der entsprechenden Normierungskonstante k verteilt sein. Die Resultate für eine $4 \rightarrow 2$ Kompression sind in Tabelle 4.4 und Tabelle 4.5 zu sehen.

lokale Kodierung am Eingang	komprimierte Kodierung am Ausgang
− − − − − − −+	− + −
− − − − − − +−	+ + +
− − − − − + −−	− + +
− − − − + − −−	− − +
− − − + − − −−	− − −
− − + − − − −−	+ − +
− + − − − − −−	+ + −
+ − − − − − −−	+ − −

Tabelle 4.3: Binäre Datenkompression $8 \rightarrow 3$. Die gefundene Abbildung ist umkehrbar. (Jedem Ausgabewert läßt sich eindeutig eine, und nur eine, Eingabe zuordnen). Es wurde somit maximale Transinformation erreicht.

Kodierung	Bit-Entropie	Entropie	Redundanz
lokal am Eingang	2.20	2.99	0.72
komprimiert am Ausgang	1.28	1.28	0.05

Tabelle 4.4: Informationsgehalt und Redundanz nach der Maximierung der Transinformation für eine $4 \rightarrow 2$ Kompression von binären Signalen, die gemäß einem Potenzgesetz verteilt sind.

lokale Kodierung am Eingang	komprimierte Kodierung am Ausgang
− − −+	−−
− − +−	++
− + −−	+−
+ − −−	−+

Tabelle 4.5: Binäre umkehrbare Datenkompression $4 \rightarrow 2$

4.3.4 Unüberwachte Klassifikation von kontinuierlichen Signalen

Die bisherigen Experimente wurden auch aufgrund ihrer Anschaulichkeit präsentiert und haben sich daher auf binäre Eingabesignale beschränkt. Jetzt soll die Fähigkeit zur unüberwachten Klassifikation von kontinuierlichen Eingabesignalen demonstriert werden. Die Methode liefert eine unüberwachte Klassifikation mit sigmoiden Funktionen, die als Vorverarbeitung einer überwachten Lernmethode verwendet werden könnte. Der auffälligste Unterschied zu anderen konnektionistischen Klassifikationsmethoden ist, daß hier eine Klasse durch einen binären Kode, und nicht etwa durch einen einzelnen aktiven Zustand, dargestellt wird.

In Abbildung 4.2 ist die Klassifikation von vier Gauß-verteilten Punktwolken im zweidimensionalen Eingaberaum (Koordinaten x_1 und x_2) dargestellt. Eine Boltzmann-Maschine mit zwei Eingabeelementen und zwei Ausgabezuständen wurde mit diesen Eingabedaten trainiert. Die vier verschiedenen Markierungen der Punkte repräsentieren die verschiedenen Zustände, die nach dem Lernen am Ausgang erzeugt werden ($--$, $+-$, $-+$ und $++$). Die verschiedenen Punktwolken werden in diesem einfachen Fall offensichtlich perfekt klassifiziert. Die Entwicklung der Kopplungskoeffizienten w_1^{ext} und w_2^{ext} für die Kopplung von der Eingabe zum ersten und zweiten Ausgabezustand wird in der gleichen Abbildung durch eine Sequenz von Punkten dokumentiert. Die zugehörigen Ortsvektoren geben die Richtungen der Koeffizienten w_{1j}^{ext} und w_{2j}^{ext} an. Die zwei Trennlinien der sigmoiden Funktionen liegen orthogonal zu diesen Ortsvektoren und verlaufen durch den Ursprung. Der Betrag der Kopplungsstärken liefert die Schärfe der sigmoiden Trennfunktion. Die exakten Lösungen liegen auf der Winkelhalbierenden der Koordinatenachsen.

Selbstständige Klassifikation von klinischer Diagnoseinformation

Unüberwachte Klassifikation kann ein wichtiges Element zur Generierung einer Diagnose aus einem Satz von beobachteten Symptomen sein. Insbesondere wenn noch keine Diagnosekriterien bekannt sind, ist die Fähigkeit eines adaptiven Systems, selbstständig die Symptome in getrennten Diagnosekategorien einzuordnen, von großer Bedeutung. Mit einem Beispiel, bei dem die Kombinationen von Symptomen bereits von einem Experten in verschiedene Diagnoseklassen unterteilt wurden, soll die Fähigkeit zum Auffinden der richtigen Klassen überprüft werden. Fünf aufgenommene Thyroid-Symptome sollen möglichst einer der drei bekannten Diagnosen Euthyroidmus, Hypothyroidmus oder Hyperthyroidmus zugeordnet werden (Coomans, Broeckaert, Jonckheer, und Massart, 1983). Die aufgenommenen Symptome basieren auf einer ausgiebigen medizinischen Untersuchung von 100 Patienten. Aufgenommen wurden die Werte von folgenden fünf Tests: T3-Resin Aufnahmetest, gesamt Serum-Thyroxin, gesamt Serum-Triiodothyronine, Grundwert von Thyroid-Stimulationshormon (TSH), maxima-

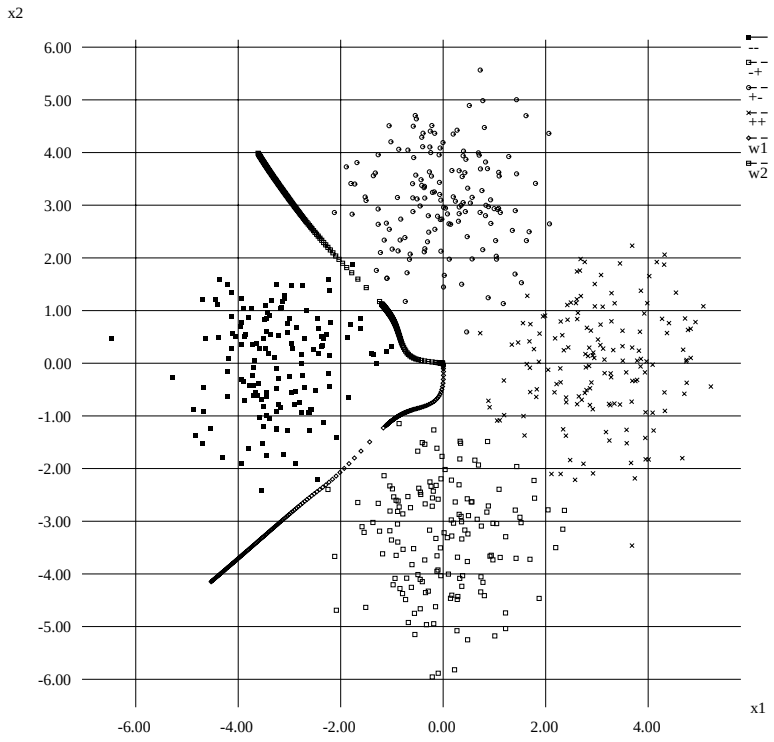

Abbildung 4.2: Resultat der Klassifikation nach der Maximierung der Transinformation. Die verschiedenen Markierungen kennzeichnen den Zustand, der einem Punkt des Eingaberaumes an der Ausgabe zugeordnet wurde. Die zwei Punktsequenzen, die sich annähernd entlang der Winkelhalbierenden des Koordinatenkreuzes befinden, dokumentieren die Entwicklung der Kopplungskoeffizienten während des Lernens. Die Unterbrechung in dieser Sequenz ist lediglich ein Artefakt der Datenaufnahme.

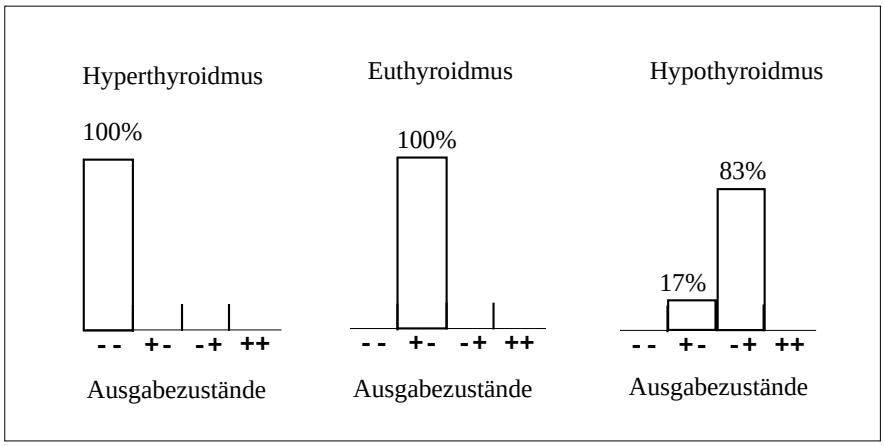

Abbildung 4.3: Häufigkeit (in Prozent), mit der Symptome der drei bekannten Diagnosen den vier möglichen Ausgabezuständen der Boltzmann-Maschine zugeordnet werden. Hyperthyrodismus und Euthyrodismus wird in allen Fällen einem einzigen Zustand zugeordnet. Die Symptome, die dem Hypothyroidmus entsprechen, sind zu 83% richtig zugeordnet. Eine vierte Diagnose in den möglichen Zustand $+-$ ist nicht gefunden worden. Die Diagnosen sind somit ohne Klasseninformation richtig erzeugt und insgesamt zu 95% richtig zugeordnet worden.

le absolute Differenz im Vergleich zum Grundwert von THS nach Injektion von $200mg$ Thyrotropin-Hormone.

Das adaptive System besteht hier aus fünf Eingabeelementen und zwei Ausgabezuständen mit entsprechend vier möglichen Klassen. Nach der Optimierung werden nur drei Zuständen eine von Null verschiedene Wahrscheinlichkeit zugeordnet. Dem Zustand $+-$ wird keiner der 100 Datenpunkte zugeordnet. Man beachte nochmals, daß die Information über die Klassen nicht zum Lernen verwendet wird. Das heißt, das System erkennt richtig, daß in den Daten nur drei getrennte Diagnoseklassen vorhanden sind. In Abbildung 4.3 wird das Klassifikationsergebnis mit Diagnosen verglichen, die von Experten bestimmt wurden. Sie zeigt die Prozentzahl richtiger Klassifikationen für die drei verschiedenen Klassen. Insgesamt ordnet das System 95% der Symptome der richtigen Diagnose zu. Zwei Klassen werden zu 100% erkannt, während die dritte zu 86% richtig bestimmt wird. Um das Resultat beurteilen zu können, ist ein überwachtes "Back-Propagation"-Netz trainiert worden (Rumelhart et al., 1988). Es wurden fünf Eingabeelemente, 15 innere und drei Ausgabeelemente verwendet. Die bekannte Klasseninformation wurde zum Lernen verwendet. Man erhält eine zu 96% richtige Klassifikation. Das Resultat, das mit der unüberwachten Boltzmann-Maschine erreicht wurde, ist daher erstaunlich gut.

4.4 Schlußfolgerungen

In diesem Kapitel wurde gezeigt, wie das Kriterium der maximalen Transinformation in einem adaptiven konnektionistischen System für unüberwachtes Lernen verwendet werden kann. Das "Infomax"-Prinzip wird dadurch auf ein stochastisches, rückgekoppeltes System erweitert. Die Hebb'sche und Anti-Hebb'sche Lernregeln der Boltzmann-Maschine werden dabei mit einem Maß der übertragenen Information gewichtet. Für binäre Eingabesignale wurden mit dieser Lernregel in Simulationen die Aufgaben der Rekodierung und Datenkompression gelöst. Das vorgestellte Konzept kann auch zur selbständigen oder unüberwachten Klassifikation von kontinuierlichen Signalen verwendet werden. Dies wurde an einem realen Problem zur Erzeugung einer medizinischen Diagnose, basierend auf beobachten Symptomen, demonstriert. Insbesondere bei der Datenkompression wird deutlich, daß zur Maximierung der Transinformation an der Ausgabe Kodierungen mit niedriger Redundanz erzeugt werden. Für die Extraktion statistisch unabhängiger Eigenschaften wird im nächsten Kapitel ein strengeres Kriterium zur Minimierung statistischer Abhängigkeit eingeführt.

Die Boltzmann-Maschine extrahiert statistisch relevante Merkmale aus folgendem Grund: Ein wesentlich kleinerer Ausgabezustandsraum muß maximale Information des Eingabezustandsraums darstellen. Es müssen die relevanten Merkmale abgebildet werden, um die Information zu erhalten. Der Ausgaberaum der neuen, unüberwachten Boltzmann-Maschine kann aus zweierlei Gründen "kleiner" sein als der Eingaberaum. Die Dimension der Ausgabe kann frei bestimmt werden. Ist man nun daran interessiert, von einem redundanten, binären Eingabesignal eine kompaktere Darstellung zu gewinnen, so wird man eine kleinere Ausgabedimension wählen. Zum anderen kann der Eingabezustandsraum der Boltzmann-Maschine auch kontinuierlich sein, während die Ausgabe auf binäre Zustände beschränkt bleibt. Dies stellt offensichtlich eine massive Reduktion des Zustandsraums dar.

Kapitel 5

Minimale gemeinsame Information im Diskreten

Das Auffinden von statistischen Koinzidenzen aus einer Kombination von sensorischen Signalen ist gemäß (Barlow, 1989) eine der wichtigsten Aufgaben eines kognitiven Systems. Es ist wichtig zu wissen, ob eine gewisse Kombination von sensorischen Signalen oder sensorischen Ereignissen ein erwartetes, gewöhnliches und normales Ereignis ist, oder ob es eher unerwartet auftritt. Auf dieser Grundlage ist es dem kognitiven System eines Lebewesens möglich, kausale Abhängigkeiten zwischen sensorischen Ereignissen und Körperbewegungen und deren Wirkung zu finden. Unter anderem wurde von Zipf (1949) und Attneave (1954) vermutet, daß das Nervensystem versucht, unabhängige Merkmale aus den sensorischen Signalen zu extrahieren. Barlow (1989) verbindet diese Aufgabe mit dem generellen Problem des unüberwachten Lernens. Ein Wahrnehmungssystem erzeugt innere Zustände zur Darstellung von externen Ereignissen einer unbekannten Umgebung. Das Ziel des unüberwachten Lernens ist es, einen Satz von Merkmalen zu finden, die unabhängig voneinander auftreten, d.h. die extrahierten Merkmale sind nicht redundant. Dies wird als *faktorielles Lernen* bezeichnet. Es bedeutet, daß die gemeinsame Wahrscheinlichkeitsverteilung der inneren Zustände als Produkt der Verteilungen der einzelnen Zustände dargestellt werden kann. Das Ergebnis dieses Lernens ist eine faktorielle Darstellung der sensorischen Eingabe. Zu erwähnen sind die Arbeiten von Redlich (1993a, 1993b) und Atick und Redlich (1990), die sich auf Barlows ursprüngliche Idee der Merkmalsextraktion durch Redundanzreduktion konzentrieren. Schmidhuber (1992) definiert eine informationstheoretische Heuristik, die eine binäre und faktorielle Darstellung eines Eingabesignals finden soll. In der vorliegenden Arbeit hingegen wird ein streng informationstheoretisches Kriterium verwendet.

In diesem Kapitel soll zur Implementation von Barlows Prinzip zum unüberwachten Lernen wieder die Boltzmann-Maschine benutzt werden. Ein Thema wird daher erneut die Extraktion von binären Merkmalen sein. Im Kapitel 7 wird man auf kontinuierliche Merkmale zurückkommen.

Im vorhergehenden Kapitel wurde gezeigt, wie das Prinzip der maximalen Transinformation zur Erzeugung von Kodierungen mit niedriger Redundanz verwendet werden kann. In diesem Kapitel soll das Kriterium der Redundanzreduktion explizit dem Kriterium der maximalen Transinformation hinzugefügt werden. An den binären Ausgabezuständen der Boltzmann-Maschine gewinnt man dadurch eine verlustfreie Darstellung der sensorischen Information in einer faktoriellen Kodierung. Auf diese Weise wird Barlows Prinzip in einem nichtlinearen, rückgekoppelten, stochastischen System implementiert. In Experimenten wird faktorielle Rekodierung und faktorielle Datenkompression verwirklicht. Schließlich liefert der Algorithmus eine informationstheoretische Erklärung für die Entstehung von *rezeptiven Feldern* im visuellen Kortex in einem einfachen Retinamodell.

5.1 Lernen mit minimaler gemeinsamer Information

Es wird das gleiche Modell wie im vorhergehenden Kapitel verwendet. Die Beschreibung von Abschnitt 4.1 ist komplett übertragbar.

Die Kopplungskoeffizienten sollen so adaptiert werden, daß die Transinformation zwischen Eingabe $\mathbf{x}$ und Ausgabe $\mathbf{y}$ maximiert und gleichzeitig die Redundanz der Ausgabe minimiert wird. Die Redundanzreduktion wird durch Minimierung der gemeinsamen Information der verschiedenen Ausgabeelemente verwirklicht.

Vor der Definition einer geeigneten Kostenfunktion, die beide Kriterien kombiniert, sollte man sich einige Grenzsätze vor Augen führen. Wie in Abschnitt 2.4 bemerkt wird, ist die Transinformation (4.3) beidseitig beschränkt,

$$0 \leq I[P(\mathbf{x}, \mathbf{y})] \leq H[R(\mathbf{x})] \tag{5.1}$$

So wie die gemeinsame Information, ist auch die Transinformation als relative Entropie zwischen einer gemeinsamen Verteilung $P(\mathbf{x}, \mathbf{y})$ und dem Produkt einzelner Verteilungen $P(\mathbf{x})$ und $P(\mathbf{y})$ zu verstehen: Es gilt $I[P(\mathbf{x}, \mathbf{y})] = H[P(\mathbf{x}, \mathbf{y}), P(\mathbf{x})P(\mathbf{y})]$. Wie bereits angedeutet, ist die relative Entropie eine nicht negative Größe. Dies erklärt die untere Grenze der Transinformation. Die obere Grenze wurde im vorhergehenden Kapitel in Abschnitt 4.2 durch Ungleichung (4.6) erklärt.

Die gemeinsame Information $MI[P(\mathbf{y})]$, die durch (2.7) definiert ist, mißt die statistische Abhängigkeit unter den Ausgabeelementen. Sie ist als relative Entropie wieder eine nicht negative Größe,

$$0 \leq MI[P(\mathbf{y})] \tag{5.2}$$

Die untere Grenze $MI[P(\mathbf{y})] = 0$ garantiert statistische Unabhängigkeit (2.6). Die Kostenfunktion K kann jetzt definiert werden als, [1]

$$\begin{aligned} K[P(\mathbf{x},\mathbf{y})] &= MI[P(\mathbf{y})] - (I[P(\mathbf{x},\mathbf{y})] - H[R(\mathbf{x})]) && (5.3)\\ &= \sum_i^n H[P(y_i)] - 2H[P(\mathbf{y})] + H[P(\mathbf{x},\mathbf{y})] && (5.4) \end{aligned}$$

Es wird die gemeinsame Information $MI[P(\mathbf{y})]$ an der Ausgabe minimiert und die Transinformation $I[P(\mathbf{x},\mathbf{y})]$ bis auf ihre obere Grenze $H[R(\mathbf{x})]$ maximiert.

Im vorhergehenden Kapitel wurde die gemeinsame Information $MI[P(\mathbf{y})]$ für die binären Ausgangszustände bereits als Bit-Entropie (4.21) bezeichnet. Sie wird dort als Funktion von P_i definiert, wobei P_i die Wahrscheinlichkeit angibt, mit der das i-te Element auf dem Zustand $+1$ zu finden ist,

$$P_i = P(y_i = +1) = \sum_{\mathbf{y}} \delta_{+,y_i} P(\mathbf{y}) = \sum_{\mathbf{xy}} \delta_{+,y_i} P(\mathbf{y}|\mathbf{x}) R(\mathbf{x}) \qquad (5.5)$$

wobei das Kroneckersymbol δ_{+,y_i} garantiert, daß nur über Ausgabezuständen mit $y_i = +1$ summiert wird. Die Kostenfunktion läßt sich ausschreiben zu,

$$\begin{aligned} K[P(\mathbf{x},\mathbf{y})] &= -\sum_i P_i \ln(P_i) - \sum_i (1-P_i)\ln(1-P_i) && (5.6)\\ &\quad +2\sum_{\mathbf{xy}} P(\mathbf{y}|\mathbf{x})R(\mathbf{x}) \ln \sum_{\mathbf{x}} P(\mathbf{y}|\mathbf{x})R(\mathbf{x}) && (5.7)\\ &\quad -\sum_{\mathbf{xy}} P(\mathbf{y}|\mathbf{x})R(\mathbf{x}) \ln P(\mathbf{y}|\mathbf{x}) + H[R(\mathbf{x})] && (5.8) \end{aligned}$$

Die Lernregeln werden jetzt als Gradientenabstieg auf der Kostenfunktion definiert. Die Ableitung der Kostenfunktion nach einem Kopplungskoeffizienten w berechnet sich wie folgt:

$$\begin{aligned} \frac{\partial K[P(\mathbf{x},\mathbf{y})]}{\partial w} &= -\sum_l \frac{\partial}{\partial w} P_l \{\ln(P_l) - \ln(1-P_l)\} && (5.9)\\ &\quad +2\sum_{\mathbf{xy}} R(\mathbf{x}) \frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x}) \{\ln P(\mathbf{y}) + 1\} && (5.10)\\ &\quad -\sum_{\mathbf{xy}} R(\mathbf{x}) \frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x}) \{\ln(P(\mathbf{y}|\mathbf{x})R(\mathbf{x})) + 1\} && (5.11) \end{aligned}$$

[1] Wie auch im vorhergehenden Kapitel wird hier zu Gunsten einer besseren Lesbarkeit die Abhängigkeit der Wahrscheinlichkeiten und Informationsausdrücke von den Kopplungskoeffizienten in der Notation vernachlässigt.

Die Ableitung der bedingten Wahrscheinlichkeit $P(\mathbf{y}|\mathbf{x})$ ist in (4.16) bzw. (4.17) gegeben. Die Ableitung von P_l ergibt sich aus (5.5) zu,

$$\frac{\partial}{\partial w} P_l = \sum_{\mathbf{xy}} \delta_{+,y_l} R(\mathbf{x}) \frac{\partial}{\partial w} P(\mathbf{y}|\mathbf{x}) \tag{5.12}$$

Zusammen liefert das,

$$\frac{\partial K}{\partial w} = \sum_{\mathbf{xy}} R(\mathbf{x}) \frac{\partial P(\mathbf{y}|\mathbf{x})}{\partial w} \left(2 \ln P(\mathbf{y}) + 1 - \ln P(\mathbf{y}|\mathbf{x}) - \sum_l \delta_{+,y_l} \{ \ln(P_l) - \ln(1-P_l) \} \right) \tag{5.13}$$

Alle Terme werden hier mit der Ableitung der bedingten Wahrscheinlichkeit $\partial P(\mathbf{y}|\mathbf{x})/\partial w$ gewichtet und summiert. Man beachte, daß der Hebb'sche und der Anti-Hebb'sche Term sich in dieser Ableitung (4.16) bzw. (4.17) aufheben, falls die Summanden in (5.13) von $\mathbf{y}$ unabhängig sind. Die Konstante $+1$ verschwindet und man kann gleichzeitig einen konstanten Ausdruck $-\sum_l ln(1-P_l)$ ergänzen. Man beachte, daß $1 - \delta_{+,y_l} = \delta_{-,y_l}$

$$\begin{aligned} \frac{\partial K}{\partial w} &= \sum_{\mathbf{xy}} R(\mathbf{x}) \frac{\partial P(\mathbf{y}|\mathbf{x})}{\partial w} \\ &\quad \left(2 \ln P(\mathbf{y}) - \ln P(\mathbf{y}|\mathbf{x}) - \sum_l \{ \delta_{+,y_l} \ln(P_l) + \delta_{y_l,-} \ln(1-P_l) \} \right) \end{aligned}$$

Zuletzt kann man noch die Reihenfolge der Summe über l mit dem Logarithmus vertauschen,

$$\sum_l \{ \delta_{+,y_l} \ln(P_l) + \delta_{-,y_l} \ln(1-P_l) \} = \ln \prod_l \{ \delta_{+,y_l} P_l + \delta_{-,y_l} (1-P_l) \}$$

Insgesamt liefert der Gardientenabstieg die folgenden Lerngleichungen für die Kopplungskoeffizienten w_{ij} und w_{ij}^{ext},

$$\Delta w_{ij} = \epsilon \beta \sum_{\mathbf{x}} R(\mathbf{x}) \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}) (*) \left(s_i s_j - \langle s_i s_j \rangle_{\mathbf{x}} \right) \tag{5.14}$$

$$\Delta w_{ij}^{ext} = \epsilon \beta \sum_{\mathbf{x}} R(\mathbf{x}) \sum_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}) (*) \left(s_i x_j - \langle s_i x_j \rangle_{\mathbf{x}} \right) \tag{5.15}$$

$$(*) = \left(\ln \frac{P(\mathbf{y}|\mathbf{x})}{P(\mathbf{y})} - \ln \left(\frac{P(\mathbf{y})}{\prod_l \{ \delta_{+,y_l} P_l + \delta_{y_l,-} (1-P_l) \}} \right) \right)$$

Diese Lerngleichungen haben eine ähnliche Interpretation wie die Lerngleichungen (4.19) und (4.19) des vorhergehenden Kapitels. Man erkennt deutlich die zwei Kriterien der Lernfunktion. Der erste Term des Klammerausdrucks (*)

mißt die Transinformation und ist identisch zu dem bisherigen Resultat. Der zweite Term wird von diesem abgezogen und ist ein Maß für die Redundanz an der Ausgabe. Diese Informationsmaße gewichten wieder die "freien" Hebb'schen Terme.

5.2 Simulationen und Resultate

5.2.1 Implementation und Komplexität des Modells

Es wurden die Lerngleichungen (5.14) und (5.15) für ein Modell ohne innere Zustände implementiert ($n_h = 0$). Das benutzte Modell gleicht dem System des vorhergehenden Kapitels. Lediglich die Lerngleichungen haben sich geändert. Die Überlegungen zur Implementation und Komplexität können also unverändert aus Abschnitt 4.3.1 übernommen werden. Für das Lernen müssen nun zusätzlich die Wahrscheinlichkeiten P_i nach Gleichung (5.5) bestimmt werden. Dies vergrößert die Komplexität des Algorithmus zur Bestimmung eines Gradientenschrittes jedoch nicht.

Ziel der nächsten Experimente ist es zunächst, die Funktionsfähigkeit der abgeleiteten Lerngleichungen zu überprüfen. Zum Vergleich der Ergebnisse verwendet man die gleichen Experimente wie in Abschnitt 4.3.2 und 4.3.3. Danach wird mit der Boltzmann-Maschine ein einfaches Modell der Retina simuliert. Die Redundanzreduktion mit maximaler Transinformation liefert dabei eine informationstheoretische Erklärung für die Entstehung von rezeptiven Feldern im visuellen Kortex.

5.2.2 Rekodierung und Komprimierung binärer Signale

Die Wirkung der zusätzlichen Minimierung der gemeinsamen Information kann an einem einfachen Kodierungsproblem beobachtet werden: die Dimension des binären Kodes soll nicht verkleinert werden. Maximale Transinformation erzeugt in diesem Fall keine faktorielle Kodierung. Die verschiedenen Eingabemuster sollen nicht gleichverteilt sein. Tabelle 5.1 zeigt die Eingabemuster $\mathbf{x}$ und die zugehörigen Wahrscheinlichkeiten für eine $3 \rightarrow 3$ Rekodierung. Die Boltzmann-Maschine wurde zum Vergleich zunächst nur mit dem Kriterium der maximalen Transinformation (max. I) und danach mit dem zusätzlichen Kriterium der minimalen gemeinsamen Information (min. MI) trainiert. Die Kodierungen, die man für diese Fälle erhält, sind ebenfalls in Tabelle 5.1 angegeben. In Abbildung 5.1 erkennt man deutlich die unterschiedliche Entwicklung der gemeinsamen Information $MI[P(\mathbf{y})]$ während des Lernens. Die Entropien der Eingabe und der Ausgabe nach dem Lernen sind in Tabelle 5.2 zusammengefaßt. Die Redundanz $R[P(\mathbf{y})]$ wird gemäß (4.20) bestimmt. Die Entropie der Eingabekodierung ist $H[P(\mathbf{x})] = 1.97$. Für beide Optimierungskriterien wird die gesamte Information

$P(\mathbf{x})$	$\mathbf{x}$	$\mathbf{y}$ max. I	$\mathbf{y}$ min. MI
0.2	$--+$	$--+$	$+-+$
0.3	$-++$	$-++$	$+++$
0.3	$++-$	$-+-$	$-+-$
0.2	$+++$	$++-$	$++-$

Tabelle 5.1: Eingabekodierung und Verteilung für eine $3 \rightarrow 3$ Rekodierung. Die beiden rechten Spalten geben den erzeugten Ausgabekode nach dem Lernen mit maximaler Transinformation (max. I) und zusätzlicher minimaler gemeinsamer Information (min. MI) an.

Kodierung	Bit-Entropie	MI	I	R
am Eingang	2.85	0.88	-	0.45
am Ausgang mit max. I	2.60	0.63	1.97	0.32
am Ausgang mit min. MI	1.97	0.0	1.97	0.0

Tabelle 5.2: Informationsgehalt und Redundanz für eine $3 \rightarrow 3$ Rekodierung nach dem Lernen mit maximaler Transinformation (max. I) und zusätzlicher minimaler gemeinsamer Information (min. MI).

der Eingabe auf die Ausgabe übertragen. Eine faktorielle Kodierung erreicht man allerdings nur, falls auch die gemeinsame Information $MI[P(\mathbf{y})]$ minimiert wird.

Das nächste Beispiel ist aus (Schmidhuber, 1992) entnommen. Eine lokale Kodierung mit vier binären Eingabezuständen und einer nicht-uniformen Verteilung soll auf zwei Bits komprimiert werden ($4 \rightarrow 2$). Schmidhuber (1992) suggeriert, daß diese spezielle Wahl der Verteilung der Eingabemuster nur eine einzige faktorielle Darstellung in zwei Dimensionen besitzt. Es existieren hingegen viele umkehrbare - jedoch redundante - Kodierungen. Die Boltzmann-Maschine findet diese faktorielle Darstellung, falls sie mit minimaler gemeinsamer Information und maximaler Transinformation optimiert wird. In Tabelle 5.3 und Tabelle 5.4 sind die Resultate zusammengefaßt.

Die Experimente aus Abschnitt 4.3.3 mit ungleichverteilten Eingabedaten wurden hier wiederholt. Aufgrund der Datenkompression erhält man bereits dort eine gute Faktorisierung. Für die nach einem Potenzgesetz verteilten Daten wurde jetzt eine genaue faktorielle Kodierung gefunden (siehe Tabelle 5.5).

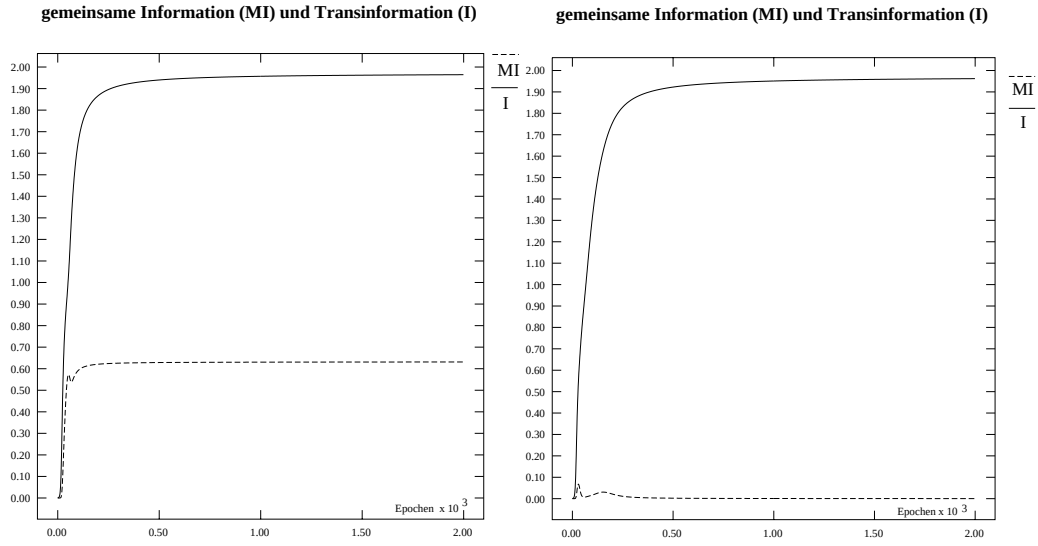

Abbildung 5.1: Entwicklung der Transinformation und der gemeinsamen Information während des Lernens für das Rekodierungsproblem $3 \rightarrow 3$. Links mit maximaler Transinformation (max. I) und rechts mit zusätzlicher minimaler gemeinsamer Information (min. MI) gelernt.

$P(\mathbf{x})$	$\mathbf{x}$	$\mathbf{y}$ max. I	$\mathbf{y}$ min. MI
$1/9$	$---+$	$+-$	$++$
$2/9$	$--+-$	$++$	$-+$
$2/9$	$-+--$	$-+$	$+-$
$4/9$	$+---$	$--$	$--$

Tabelle 5.3: Eingabekodierung und Verteilung für eine $4 \rightarrow 2$ Komprimierung. Die beiden rechten Spalten geben den erzeugten Kode nach dem Lernen mit maximaler Transinformation (max. I) und zusätzlicher minimaler gemeinsamer Information (min. MI) an.

Kodierung	Bit-Entropie	MI	I	R
am Eingang	3.03	1.19	-	0.45
am Ausgang mit max. I	1.91	0.07	1.84	0.04
am Ausgang mit min. MI	1.84	0.0	1.84	0.0

Tabelle 5.4: Informationsgehalt und Redundanz für eine $4 \rightarrow 2$ Komprimierung nach dem Lernen mit maximaler Transinformation (max. I) und zusätzlicher minimaler gemeinsamer Information (min. MI). Die Entropie des Eingabekodes ist $H[P(\mathbf{x})] = 1.84$. In beiden Fällen wird die gesamte Information der Eingabe auf die Ausgabe übertragen

Kodierung	Bit-Entropie	MI	I	R
am Eingang	2.02	0.63	-	0.45
am Ausgang mit max. I	1.31	0.07	1.28	0.05
am Ausgang mit min. MI	1.28	0.0	1.28	0.0

Tabelle 5.5: Informationsgehalt und Redundanz für eine $4 \rightarrow 2$ Komprimierung von binären Signalen, verteilt gemäß einem Potenzgesetz.

5.2.3 Entstehung von rezeptiven Feldern in einem Retinamodell

Ziel der nun folgenden Simulationen ist es, zu zeigen, wie das von Barlow vorgeschlagene Kriterium zur Merkmalsextraktion tatsächlich in der Lage ist, biologische Befunde zur Signalverarbeitung im visuellen Kortex zu simulieren. Wie von Rubner und Schulten (1990) bemerkt, ist das Verständnis der Verarbeitung von räumlicher Information im visuellen Kortex für lange Zeit ein ungelöstes Problem gewesen. Hubel und Wiesel (1962) beschreiben die Zellen im primären visuellen Kortex (Area 17) als Merkmalsdetektoren. Diese Neuronen führen die erste Verarbeitung räumlicher Information durch. Es stellt sich die folgende Frage: Werden dort zuerst räumlich lokale Merkmale extrahiert oder wird vielmehr eine Zerlegung in räumliche Frequenzen, ähnlich einer Fouriertransformation, durchgeführt (siehe Campbell und Robson, 1968; MacKay, 1981; Pollen, Lee, und Taylor, 1971). Rubner und Schulten (1990) beschreiben einen Mechanismus zur Bildung von räumlichen Merkmalsdetektoren für den Fall von Neuronen mit linearer Antwortfunktion. Dieser Mechanismus führt zu einer herkömmlichen Hauptkomponentenanalyse. Durch das vorliegende Modell erhält man eine nichtlineare, räumliche Merkmalsextraktion.

Das Lernkriterium bildet räumliche rezeptive Felder in einem einfachen Modell einer Retina. Man stelle sich ein gleichmäßiges 10×10 Gitter von Eingabeneu-

ronen vor. Die kontinuierliche Eingabe $\mathbf{x}$ ist somit 100-dimensional. Diese Eingabe ist durch die synaptischen Kopplungen w_{ij}^{ext} zunächst nur mit zwei binären Ausgabeneuronen verbunden. Die Rückkopplung unter den Ausgabezuständen erwies sich als nicht notwendig, daher setzt man $w_{ij} = 0$. Jedes Eingabesignal, das zum Lernen verwendet wurde, kann man sich als ein Gaußsches Lichtsignal vorstellen, das auf die Retina trifft. Die Schwerpunkte der Gaußschen Signale werden gleichmäßig über die Retina verteilt. Um Randeffekte zu vermeiden, sind sie jedoch stets mindestens zwei Neuronen vom Rand entfernt. Die Redundanzreduktion führt zu einer Dekorrelation der synaptischen Verbindungen. Minimale gemeinsame Information wird dabei durch die Bildung von rezeptiven Feldern erreicht. Dies wird ersichtlich, wenn man die Kopplungsstärken der verschiedenen Ausgabeneuronen vergleicht. In Abbildung 5.2 sieht man die gelernten Kopplungsstärken, die je zwei Eingabeneuronen mit den zwei Ausgabezuständen verbinden. Aus diesen Abbildungen wird deutlich, daß die Ausgabeneuronen für Signale aus verschiedenen Bereichen im Eingaberaum reagieren. Für starke positive (exitatorische) Kopplungsstärken nimmt das Neuron bei einem Signal aus diesem Bereich einen Wert $+1$ an, für negative (inhibierende) Kopplungen entsprechend -1. Es gibt offensichtlich vier verschiedene Bereiche (rezeptive Felder), in denen ein räumlich lokalisiertes Signal eine der vier verschiedenen Ausgabekonfigurationen $(++, --, +-, -+)$ erzeugen wird. Ein Signal bei $x_1 = 4$ wird im linken Neuron einen Zustand $+1$ bewirken, während ein Signal bei $x_1 = 7$ einen Zustand -1 hervorrufen wird. Entsprechend liefert ein Signal bei $x_1 = 4$ im linken Neuron ein Zustand -1 und einen Zustand $+1$ bei $x_2 = 7$. Dieses einfache Experiment zeigt, daß die Bildung von räumlich selektiven Zellen durch das Prinzip der minimalen Redundanz bei vollständiger Informationsübertragung erklärt werden kann.

Ein interessantes Resultat erhält man, wenn man an der Ausgabe sechs Neuronen verwendet. Die entstandenen rezeptiven Felder der verschiedenen Neuronen sind in Abbildungen 5.3 und 5.4 zu sehen. Wie bereits für den linearen Fall von Rubner und Schulten (1990) analysiert, ergibt das vorliegende Modell orthogonale und räumlich oszillierende rezeptive Felder. Jones und Palmer (1987) wie auch Jones, Stepnoski, und Palmer (1987) beobachten im visuellen Kortex von Katzen die gleichen oszillierenden Muster. Die gebildeten rezeptiven Felder haben exitatorische und inhibitorische Bereiche und stellen einfache Zellen dar, die selektiv auf Kanten oder Linien bestimmter Orientierung reagieren. Folgt man also dem Kriterium der Redundanzreduktion ohne Verlust von Information, so bilden sich Detektoren für unabhängige Merkmale des Eingabesignals.

5.3 Schlußfolgerungen

In diesem Kapitel wurde das Kriterium der maximalen Transinformation zwischen Eingabe und Ausgabe mit dem Kriterium der minimalen gemeinsamen Informa-

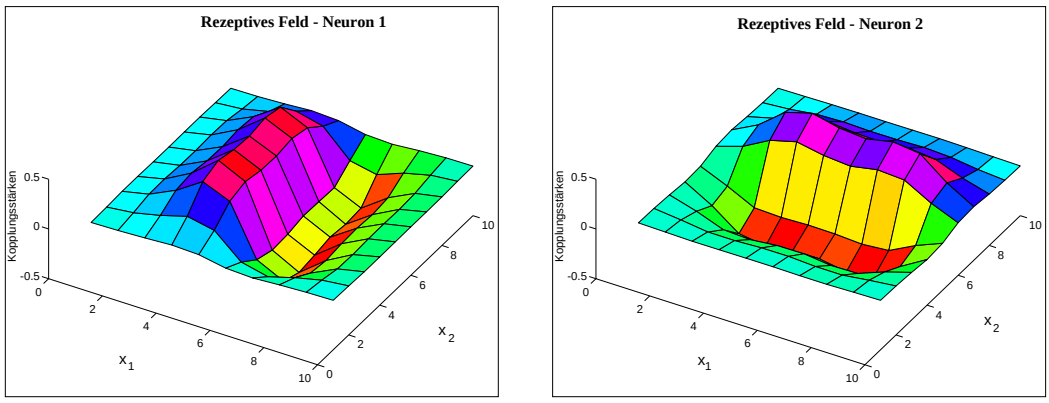

Abbildung 5.2: Gelernte Kopplungsstärken der Eingabeneuronen mit den zwei Ausgabezuständen. Die horizontale Ebene gibt die Lage der Eingabeneuronen auf dem Gitter (Retina) wieder. In vertikaler Richtung ist die zugehörige Kopplungsstärke aufgetragen.

tion an der Ausgabe ergänzt. Dies stellt das Prinzip der Redundanzreduktion von Barlow (1959) dar. Durch die explizite Definition der Wahrscheinlichkeitsverteilung der diskreten Zustandsvariablen läßt sich dieses Kriterium als analytische Kostenfunktion definieren. Mit einem Gradientenverfahren können die adaptiven Kopplungsstärken optimiert werden, und man erhält neue Gleichungen für unüberwachtes Lernen in der Boltzmann-Maschine. Das resultierende System ist in der Lage, statistisch unabhängige, binäre Merkmale der Eingabedaten zu extrahieren. In Experimenten mit binären Eingabesignalen wurden Datenkompression und Rekodierung verwirklicht. Man erzeugt dabei eine redundanzfreie Darstellung der Eingabeinformation. Es wurde ein einfaches Modell einer visuellen Retina als eine einschichtige Boltzmann-Maschine dargestellt. Die Optimierungskriterien bewirken die Bildung von räumlichen rezeptiven Feldern, ähnlich derer, die in einem biologischen System gefunden wurden.

Zuletzt sei angemerkt, daß die Struktur der Standard-Boltzmann-Maschine, die ein nichtlineares, rückgekoppeltes System mit inneren Elementen darstellt, die Formulierung dieser Optimalitätskriterien ermöglicht. Andererseits aber stellen die binären Zustände eine ernsthafte Einschränkung dar. Man möchte ggf. auch kontinuierliche Merkmale extrahieren. Die Arbeit, die im nächsten Kapitel dargestellt wird, hat das Ziel, ein analoges stochastisches System im kontinuierlichen Zustandsraum zu definieren.

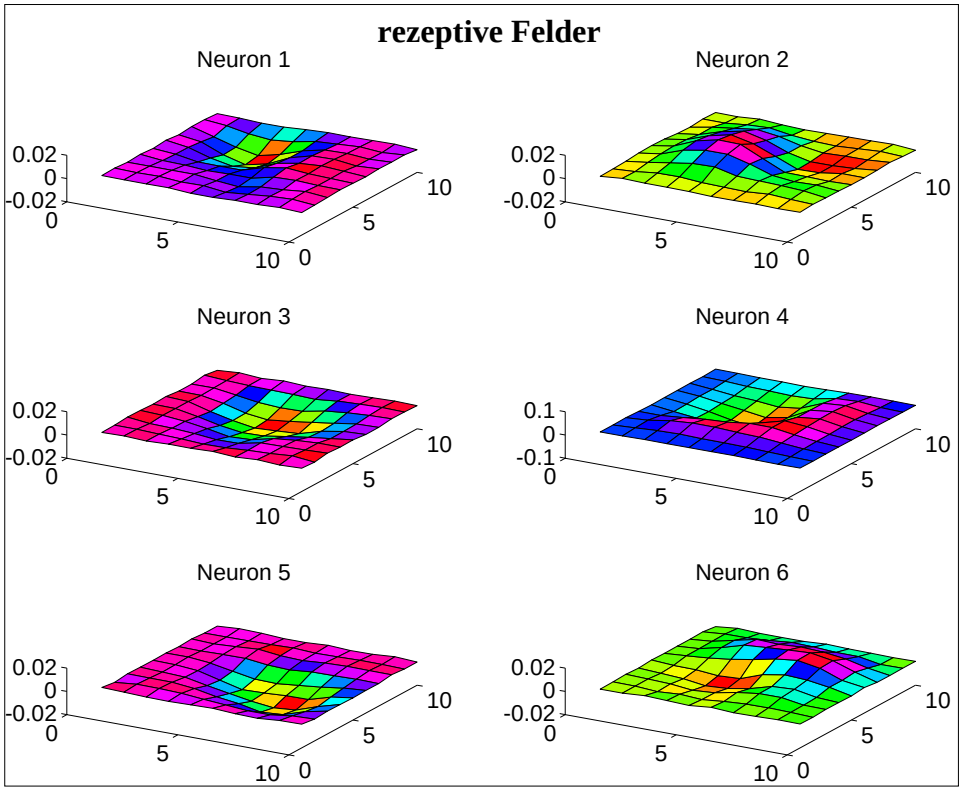

Abbildung 5.3: Kopplungsstärken der Eingabeneuronen mit sechs Ausgabeneuronen nach der Optimierung von Redundanz und Transinformation. Die horizontale Ebene gibt die Lage der Eingabeneuronen auf dem Gitter (Retina) wieder. In vertikaler Richtung ist die zugehörige Kopplungsstärke aufgetragen.

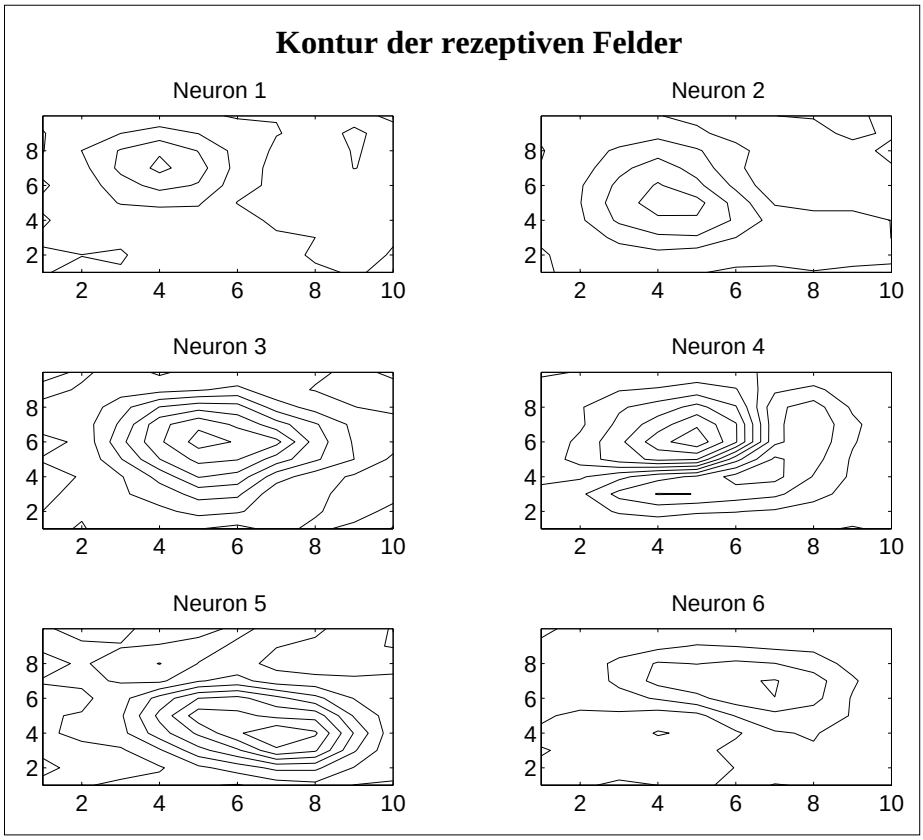

Abbildung 5.4: Kopplungsstärken durch Höhenlinien dargestellt.

Kapitel 6

Relative Entropie im Kontinuierlichen

In diesem Kapitel wird ein kontinuierliches System vorgestellt, das die Formulierung eines informationstheoretischen Kriteriums ermöglicht. Im wesentlichen wird eine Boltzmann-Maschine mit kontinuierlichen Zuständen definiert. In diesem System läßt sich das Lernkriterium der relativen Entropie analog zur Standard-Boltzmann-Maschine verwenden. Es werden zunächst einige bekannte Ansätze diskutiert.

Hopfield (1982b) stellt eine kontinuierliche Version seines ursprünglichen Modells vor. Er verwendet die Gleichungen (3.12) als Definition eines kontinuierlichen und deterministischen Systems. Diese Fixpunktgleichungen konvergieren tatsächlich zu kontinuierlichen Zuständen im Intervall $[-1, 1]$. Pineda (1987) fügt diesem Modell innere Elemente hinzu und legt damit die Grundlage für zeitabhängige, rückgekoppelte Netze. Bei diesen deterministischen, nichtlinearen Systemen gibt es allerdings keine naheliegende Interpretation einer Wahrscheinlichkeit. Informationstheoretische Kriterien lassen sich daher nicht ohne weiteres anwenden.

Die Arbeit von Kosmatopoulos und Christodoulou (1993) behandelt kontinuierliche Elemente, deren stochastische Dynamik durch Langevin-Gleichungen beschrieben wird. Die Autoren basieren die Lerndynamik auf die Arbeit von Ioannou und Sun (1995), die keine Informationskriterien betrachten, sondern den quadratischen Abstand der generierten zu einer vorgegebenen Verteilung minimieren. Sie beziehen Kopplungen höherer Ordnung mit ein, die es erlauben, die Einschränkung auf quadratische Energiefunktionen zu überwinden. Movellan und McClelland (1991) erweitern ein deterministisches Modell ähnlich dem von Hopfield (1982b) indem sie thermisches Rauschen und innere Neuronen einführen. Die Klasse der Dichtefunktionen, die modelliert werden können, ist nicht klar zu erkennen, da die Stochastik nicht rigoros eingeführt wird und daher auch die erzeugte Statistik nicht explizit bestimmt wurde. In einer vorläufigen Arbeit beschreibt Movellan (1993) ein stochastisches System mit quadratischer Energie-

funktion und symmetrischen Kopplungen auf der Grundlage von Diffusions- bzw. Langevin-Gleichungen. Er ist in den Simulationen jedoch gezwungen, die Stochastik numerisch auszuführen. Der resultierende Lernalgorithmus sowie auch das Abrufen von gelernten Assoziationen sind somit extrem rechenintensiv.

Bei informationstheoretischen Lernkriterien wird man damit konfrontiert, Integrale über den Zustandsraum berechnen zu müssen. Das ergibt sich bereits aus der Definition der Entropie (2.9). Die numerische Berechnung von Zustandssummen im kontinuierlichen Raum, entweder direkt oder mittels Monte-Carlo-Simulationen, ist sehr rechenintensiv. Es ist wichtig, einen analytischen Ausdruck für die zum Lernen benötigten Zustandssummen zu gewinnen. Eine analytische Näherung, wie sie die Mittelfeldtheorie liefert, ist im diskreten Fall hilfreich, aber nicht zwingend. Für den kontinuierlichen Fall erscheint sie allerdings unumgänglich.

Deshalb konzentriert sich diese Arbeit auf ein kontinuierliches Modell, das die Anwendung der Mittelfeldtheorie analog zu der von Peterson und Anderson (1987) ermöglicht. Man könnte versuchen, direkt stochastische und kontinuierliche Elemente zu betrachten. Mit einer definierten Verteilung formuliert man wie bisher die Informationsbegriffe und leitet die entsprechende Lernregeln ab. Indem man die Mittelfeldnäherung anwendet, erhält man eine deterministische Dynamik. Diese Strategie wird vorliegenden Kapitel in Abschnitt 6.6 verfolgt. Die resultierenden Mittelfeldgleichungen werden jedoch für die Grenztemperatur $T \to 0$ instabil, und man erhält wieder ein System mit nur zwei stabilen diskreten Zuständen.

Hier wird als Lösung eine Generalisierung des binären Spin-Modells verwendet. Gislén, Peterson, und Södeberg (1992) präsentieren Mittelfeldgleichungen für mehrdimensionale, kontinuierliche und normierte Zustände, die *Rotoren* genannt werden. Die Normierung zwingt zwar, die Zustände mehrdimensional zu wählen, erlaubt aber zugleich, die Analysis durchzuführen, die für die *Mittelfeldtheorie* notwendig ist. Ähnliche Zustände, genannt *Phasoren*, wurden von Noest (1988) eingeführt. Diese komplexwertigen Elemente sind auf den zweidimensionalen, komplexen Raum beschränkt; auch das konnektionistische Modell ist auf eine einschichtige Struktur ohne innere Zustände limitiert. Mozer, Zemel, Behrmann, und Williams (1992) betrachten ebenfalls zweidimensionale Elemente. Die vorliegende Arbeit stellt eine mehrdimensionale Erweiterung dieses Modells dar. Man nennt die Zustände Rotoren, da sie Werte auf der mehrdimensionalen Einheitssphäre annehmen können. Sie sind natürlicherweise für zyklische Probleme geeignet, bei denen zwei, drei oder mehrere Dimensionen zu behandeln sind.

6.1 Mittelfeldtheorie für Rotoren

Im allgemeinen wird in dieser Arbeit der Schwerpunkt auf die informationstheoretischen Lernkriterien gelegt. Hier wird allerdings ein neues thermodynamisches

System als konnektionistische Architektur vorgeschlagen. Aus diesem Grund sollen die dynamischen Eigenschaften des neuen System etwas genauer untersucht werden. Dazu wird hier die Mittelfeldtheorie von Gislén et al. (1992) für Rotoren kurz dargelegt. Die Autoren führen die Mittelfeldtheorie auf eine sehr allgemeine Weise ein. Als Rotoren bezeichnen sie die mehrdimensionalen Einheitsvektoren,

$$\mathbf{s}_i = \Re^d \quad |\mathbf{s}_i| = 1 \quad i = 1...n \tag{6.1}$$

Sie betrachten die Aufgabe der Minimierung einer Energiefunktion $E(\mathbf{s}_1, ..., \mathbf{s}_n)$. Die einzelnen d-dimensionalen normierten Zustände sollen mit $\mathbf{s}_i$ bezeichnet werden, alle n Zustände gemeinsam werden als $\mathbf{S}$ notiert, d.h. $\mathbf{S} = (\mathbf{s}_1, ..., \mathbf{s}_n)$. Ausgangspunkt für die Mittelfeldtheorie ist die Zustandssumme des thermodynamischen Systems,

$$Z = \int_\Phi e^{-\frac{1}{T}E(\mathbf{S})} d\mathbf{S} \tag{6.2}$$

Die Integration erstreckt sich über alle n d-dimensionalen Einheitssphären, die hier kurz mit Φ benannt werden. Aus der Zustandssumme lassen sich verschiedene thermodynamische Mittelwerte ableiten. Insbesondere ist man an den Zustandsmittelwerten interessiert. Da die numerische Auswertung dieser Summe im Kontinuierlichen nicht durchführbar ist, versucht man eine Näherung für Z zu gewinnen. Bevor man dieses Integral durch einen Sattelpunktsansatz nähert, führt man neue sogenannte Mittelfeldvariablen $\mathbf{u}_i$ und $\mathbf{v}_i$ ein, und die Integration in $\mathbf{s}_i$ wird ausgeführt.[1] Die neuen Mittelfeldvariablen sind nicht auf die Einheitssphären beschränkt. Sie sind für den gesamten kontinuierlichen Raum zugelassen, d.h. $\mathbf{v}_i, \mathbf{u}_i \in \Re^d$. Auf diese Weise wird der Raum $|\mathbf{s}_i| = 1$ mit einem neuen Raum der Mittelfeldvariablen ersetzt. Die so gewonnene Darstellung kann nun wieder als Zustandssumme der neuen Variablen $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_n)$ und $\mathbf{U} = (\mathbf{u}_1, ..., \mathbf{u}_n)$ interpretiert werden, auf die ein neues, effektives Potential $E_{eff}(\mathbf{V}, \mathbf{U})$ wirkt,

$$Z \propto \int_{\Re^{nd}} \int_{\Re^{nd}} e^{-\frac{1}{T}E_{eff}(\mathbf{V},\mathbf{U})} d\mathbf{V} d\mathbf{U} \tag{6.3}$$

$$E_{eff}(\mathbf{V}, \mathbf{U}) = E(\mathbf{V}) - T \sum_i \mathbf{v}_i \cdot \mathbf{u}_i + T \sum_i G(|\mathbf{u}_i|) \tag{6.4}$$

wobei $G(u)$ durch die modifizierten Bessel'schen Funktionen $I_m(u)$ gegeben ist,

[1] Die neuen Variablen werden mit einer Dirac'schen "δ-Funktion" nach einer Standardmethode der Mittelfeldtheorie eingeführt. Sei $s \in \Phi$,

$$\int_\Phi f(s)ds = \int_\Phi \int_{-\infty}^{\infty} f(s)\delta(s-v)dsdv = \int_\Phi \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(s)e^{\imath(s-v)u}dsdvdu = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{f}(v,u)dvdu$$

wobei $\tilde{f}(v,u)$ das Resultat der Integration über s im Raum Φ darstellt.

$$G(u) = \ln I_{\frac{d-2}{2}}(u) - \frac{d-2}{2}\ln(u) \tag{6.5}$$

Die Zustandssumme (6.3) charakterisiert ein thermodynamisches System für die Variablen $\mathbf{V}$ und $\mathbf{U}$. Nun ist man in der Lage eine Sattelpunktsnäherung durchzuführen. Gislén et al. (1992) sind bezüglich der Gültigkeit dieser Näherung an dieser Stelle etwas ungenau. Deshalb soll die Sattelpunktsnäherung hier etwas genauer dargelegt werden. Grundidee ist es, die Zustandssumme (6.3) durch eine Freie Energie darzustellen. Diese wird zu einer Funktion des Sattelpunkts $\mathbf{V}^0, \mathbf{U}^0$ der effektiven Energie. [2] Der Sattelpunkt ist durch die Gleichungen $\nabla_{\mathbf{V}} E(\mathbf{V}^0, \mathbf{U}^0) = 0$, und $\nabla_{\mathbf{U}} E(\mathbf{V}^0, \mathbf{U}^0) = 0$ festgelegt. Man identifiziert die Sattelpunkte mit der mittleren Zustandsvariablen des ursprünglichen Systems. In diesem Fall heißt das $\mathbf{V}^0 = \langle \mathbf{S} \rangle$. Das soll hier etwas genauer gezeigt werden, um bei der Gelegenheit die Grenzfälle, für die die Näherung gültig ist, zu diskutieren. Das thermische Mittel der Zustandsvariablen $\mathbf{S}$, d.h. das Mittel über die Boltzmann-Verteilung ist gegeben durch,

$$\langle \mathbf{S} \rangle = \frac{1}{Z}\int_{\Phi} \mathbf{S} e^{-\frac{1}{T}E(\mathbf{S})} d\mathbf{S} = \frac{\int_{\Re^{nd}}\int_{\Re^{nd}} \mathbf{V} e^{-\frac{1}{T}E_{eff}(\mathbf{V},\mathbf{U})} d\mathbf{V} d\mathbf{U}}{\int_{\Re^{nd}}\int_{\Re^{nd}} e^{-\frac{1}{T}E_{eff}(\mathbf{V},\mathbf{U})} d\mathbf{V} d\mathbf{U}} \tag{6.6}$$

Der Exponent wird nun um einen Sattelpunkt $\mathbf{V}^0, \mathbf{U}^0$ der effektiven Energie entwickelt. Der Term nullter Ordnung ist eine Konstante in der Integration

[2] Dies basiert auf folgende Näherung (siehe Herz et al., 1990; Fischer und Herz, 1991). Der Einfachheit halber sei die Zustandsvariable vorübergehend $\mathbf{s} \in \Re^n$, und man möchte das Funktional Z approximieren

$$Z = \left(\frac{n/T}{2\pi}\right)^{n/2} \int_{\Re^n} e^{-\frac{1}{T}nf(\mathbf{s})} d\mathbf{s}$$

Die Normierung im Klammerausdruck ist nicht weiter von Bedeutung. Der Exponent in der Zustandssumme läßt sich in dieser Form nur schreiben, falls die Energiefunktion des zugehörigen physikalischen Systems linear mit der Systemgröße n, d.h. der Anzahl freier Zustände, skaliert. Für $T \to 0$ bzw. $n \to \infty$ gilt in Näherung zweiter Ordnung,

$$Z = |J[f(\mathbf{s}^0)]|^{-1/2} e^{-\frac{1}{T}nf(\mathbf{s}^0)}$$

wobei der Sattelpunkt $\mathbf{s}^0$ aus der Sattelpunktsgleichung $\nabla_{\mathbf{s}} f(\mathbf{s}^0) = 0$ zu bestimmen ist. Hier ist $|J[f(\mathbf{s}^0)]|$ die Jacobi-Determinante von $f(\mathbf{s})$ am Sattelpunkt. Diese Näherung gilt allerdings nur falls die Jacobi-Matrix J am Sattelpunkt positiv definit ist, d.h. ein echtes Minimum vorliegt. Die Freie Energie F ist in der Thermodynamik als Funktion der Zustandssumme Z darstellbar,

$$F = -\frac{1}{T}\ln Z = nf(\mathbf{s}^0) + const.$$

Die Freie Energie wird eine Funktion der Mittelfeldvariablen $\mathbf{s}^0$. Heuristisch sagt man daher, daß der Sattelpunkt die minimale freie Energie des thermodynamischen Systems angibt. In unserem Fall entspricht die freie Energie der effektive Energie E_{eff} in (6.3).

und kürzt sich mit dem entsprechenden Term im Nenner. Der Term erster Ordnung verschwindet, da um den Sattelpunkt entwickelt wurde. Die zweite Ordnung ergibt eine mehrdimensionale Gaußfunktion, bei der die Varianz mit $\sqrt{T}$ skaliert. Diese Gaußfunktionen führen für den Grenzfall $T \to 0$ zu Dirac'schen "δ-Funktionen". Vernachlässigt man Terme höherer Ordnung, so ergibt sich für diesen Grenzwert

$$\langle \mathbf{S} \rangle_{T \to 0} = \frac{\int_{\Re_{nd}} \int_{\Re_{nd}} \mathbf{V} \delta(\mathbf{V} - \mathbf{V}^0) \delta(\mathbf{U} - \mathbf{U}^0) d\mathbf{V} d\mathbf{U}}{\int_{\Re_{nd}} \int_{\Re_{nd}} \delta(\mathbf{V} - \mathbf{V}^0) \delta(\mathbf{U} - \mathbf{U}^0) d\mathbf{V} d\mathbf{U}} = \mathbf{V}^0 \tag{6.7}$$

Es sei angemerkt, daß die gleiche Näherung durchgeführt werden kann, falls die effektive Energiefunktion linear mit der Systemgröße n skaliert, d.h. $E_{eff}(\mathbf{S}) = nf(\mathbf{S})$, wobei $f(\mathbf{S})$ als die freie Energie des Einzelzustandes zu interpretieren ist. Der Grenzfall $n \to \infty$ ist sodann identisch zum eben betrachteten. Die Mittelfeldgleichungen sind dann sowohl für große Systeme als auch für kleine Temperaturen näherungsweise gültig. Man identifiziert also die Sattelpunktsvariable $\mathbf{V}^0$ mit dem thermischen Mittel der Zustände. Bis hier reicht die Ergänzung zur Arbeit von Gislén et al. (1992). Die Autoren geben folgende Sattelpunktsgleichungen für Rotoren an

$$\mathbf{u}_i^0 = -\frac{1}{T} \nabla_{\mathbf{v}_i} E(V_1^0, ..., V_n^0) \tag{6.8}$$

$$\mathbf{v}_i^0 = \frac{\mathbf{u}_i^0}{|\mathbf{u}_i^0|} F(|\mathbf{u}_i^0|) \tag{6.9}$$

Hier ist $F(u) = \frac{\partial}{\partial u} G(u)$ und hat die Form einer sigmoiden Funktion. Mit diesen Gleichungen als Grundlage ist man nun in der Lage, ein neues deterministisches System im kontinuierlichen Raum zu definieren, bei dem es eine fundamentale Interpretation der Wahrscheinlichkeitsdichte gibt.

6.2 Kontinuierliche Boltzmann-Maschine

Mit diesen Mittelfeldgleichungen (6.8) und (6.9) wird in der vorliegenden Arbeit die kontinuierliche Version der Boltzmann-Maschine durch geeignete Wahl der Energiefunktion definiert. Es wird eine quadratische Funktion analog zur Standard-Boltzmann-Maschine gewählt.[3] Man erlaubt unabhängige Wechselwir-

[3]Sowohl hier als auch in der klassischen Boltzmann-Maschine sind Energiefunktionen mit Gliedern höherer Ordnung denkbar, so wie sie von Kosmatopoulos und Christodoulou (1993) in ihrem Langevin-Modell verwendet werden. Es wurde allerding kürzlich von Kappen (1994) gezeigt, daß eine deterministische Boltzmann-Maschine mit quadratischer Energiefunktion sogar unter einer stark eingeschränkten Konnektivität ähnliche Abbildungsfähigkeiten wie der klassische Back-Propagation-Algorithmus besitzt (Rumelhart et al., 1988).

kung der einzelnen Dimensionen mit denen der anderen Rotor-Zustände.[4]

$$E = -\frac{1}{2}\sum_{ijkl} s_{ik} w_{ikjl} s_{jl} = -\frac{1}{2}\sum_{ij} \mathbf{s}_i \cdot \mathbf{w}_{ij} \cdot \mathbf{s}_j \tag{6.10}$$

Die Indizes $i, j = 1, .., n$ numerieren die verschiedenen Rotoren. Die Indizes $k, l = 1, .., d$ beziehen sich auf die verschiedenen Dimensionen innerhalb eines Rotors. Zur Notation siehe Tabelle 8.3. Um die Konvergenz der Dynamik zu garantieren, muß man Symmetrie in i, j und k, l fordern, d.h. $w_{ikjl} = w_{jlik}$. Die Sattelpunktsgleichungen (6.8) und (6.9) sind hiermit,

$$\mathbf{u}_i = -\frac{1}{T}\sum_j \mathbf{w}_{ij} \cdot \mathbf{v}_j \tag{6.11}$$

$$\mathbf{v}_i = \frac{\mathbf{u}_i}{|\mathbf{u}_i|} F(|\mathbf{u}_i|) = \mathbf{f}(\mathbf{u}_i) \tag{6.12}$$

Dies sind die Definitionsgleichungen für eine deterministische und kontinuierliche Boltzmann-Maschine. Für $d = 1$ vereinfacht sich $F(u)$ zu einem Tangens hyperbolicus, und die Gleichungen (6.11) und (6.12) reduzieren sich auf die bekannten Mittelfeldgleichungen der Form (3.7). Es sei an der Stelle angemerkt, daß man hier im Gegensatz zur Standard-Boltzmann-Maschine kein externes Feld verwendet. Es ist unproblematisch, ein solches hinzuzufügen. Für die Kopplungsstärken reicht dafür ein Tensor dritter Stufe. Man hat zugunsten der einfacheren Diskussion der Konvergenzeigenschaften in Abschnitt (6.3) darauf verzichtet. Statt dessen verwendet man als Eingabe einige der Rotor-Zustände, so wie das im ursprünglichen Modell von Ackley et al. (1985) geschieht.

Man erhält hier ein kontinuierliches, deterministisches System, bei dem die Wahrscheinlichkeitsdichte der Zustände weiterhin eine grundlegende Bedeutung hat. Zur Lösung dieser deterministischen Gleichungen kann man wieder eine Fixpunktiteration verwenden. Es sollen im nächsten Abschnitt die Konvergenzeigenschaften des Gleichungssystems untersucht werden.

6.3 Konvergenzeigenschaften der deterministischen Dynamik

Die Gleichungen (6.11) und (6.12) können wie die Mittelfeldgleichungen (3.7) für die klassische Boltzmann-Maschine wieder als stationäre Lösung einer partiellen Differenzialgleichung erster Ordnug betrachtet werden

[4] Obwohl die mehrdimensionalen Rotoren Analogien zu den Heisenberg-Spins aufweisen (siehe Fischer und Herz, 1991), unterscheidet sich hier die Energiefunktion von der Energie mehrdimensionaler Spin-Systeme.

$$\frac{d\mathbf{u}_i}{dt} = -\mathbf{u}_i + \frac{1}{T}\sum_j \mathbf{w}_{ij} \cdot \mathbf{f}(\mathbf{u}_j) \tag{6.13}$$

Die Fixpunktiteration zur Suche von Lösungen von (6.11) und (6.12) kann als Zeitdiskretisierung dieser Gleichung mit einem Zeitschritt $\Delta t = 1$ verstanden werden.

6.3.1 Lyapunov-Funktion

Um die Konvergenz von (6.13) zu den Fixpunkten (6.11) und (6.12) zu beweisen, zeigt man analog zum Ansatz von Hopfield (1982b), daß eine Lyapunov-Funktion L existiert, mit $\partial L/\partial t < 0$. Man betrachte,

$$L = -\frac{1}{2T}\sum_{ij} \mathbf{v}_i \cdot \mathbf{w}_{ij} \cdot \mathbf{v}_j + \sum_i \int_0^{\mathbf{v}_i} \mathbf{f}^{-1}(\mathbf{v}) \cdot d\mathbf{v} \tag{6.14}$$

wobei $\mathbf{f}^{-1} = \mathbf{v}/|\mathbf{v}|F^{-1}(|\mathbf{v}|)$ existiert, da stets $F' > 0$ gilt. Das Pfadintegral kann über eine beliebige Kurve ausgeführt werden, da der Integrand rotationsfrei ist. Man muß nun zeigen, daß die so definierte Funktion L eine stets negative Zeitableitung besitzt.

$$\begin{aligned}
\frac{dL}{dt} &= -\sum_i \frac{d\mathbf{v}_i}{dt} \cdot \left(\frac{1}{T}\sum_j \mathbf{w}_{ij} \cdot \mathbf{v}_j - \mathbf{f}^{-1}(\mathbf{v}_i)\right) && (6.15)\\
&= -\sum_i \frac{d\mathbf{v}_i}{dt} \cdot \frac{d\mathbf{u}_i}{dt} && (6.16)\\
&= -\sum_{ij} \frac{d\mathbf{v}_i}{dt} \cdot \nabla_{\mathbf{v}_j}\mathbf{f}_i^{-1} \cdot \frac{d\mathbf{v}_j}{dt} < 0 && (6.17)
\end{aligned}$$

Die Ungleichung ist gültig, falls $\nabla_{\mathbf{v}_j}\mathbf{f}_i^{-1}$ positiv definit ist. Man beachte, daß für (6.15) die Symmetrie der Kopplungsstärken verwendet wurde. Mithilfe des Theorems von Gershgorin zeigt man, daß der Tensor vierter Stufe positiv definit ist. Um Mißverständnisse zu vermeiden, wird im folgenden die vollständige Indizierung verwendet. Es wird mit $v_i = |\mathbf{v}_i|$ und $g_i = F^{-1}(|\mathbf{v}_i|)$ abgekürzt.

$$h_{ik}^{jl} \equiv \frac{\partial f_{ik}^{-1}}{\partial v_{jl}} = \delta_{ik}^{jl}\frac{g_i}{v_i} - \frac{v_{ik}v_{jl}\delta_j^i}{v_i^2}\left(-\frac{g_i}{v_i} + g_i'\right) \tag{6.18}$$

hier ist δ_{ik}^{jl} das entsprechende Kronecker-Symbol, mit $\delta_{ik}^{jl} = 1$ nur für $ik = jl$, sonst $\delta_{ik}^{jl} = 0$. Zunächst wird gezeigt, daß die Diagonalelemente h_{ik}^{ik} positiv sind:

$$h_{ik}^{ik} = \frac{g_i}{v_i} + \frac{v_{ik}^2}{v_i^2}\left(-\frac{g_i}{v_i} + g_i'\right) = \frac{g_i}{v_i}\left(1 - \frac{v_{ik}^2}{v_i^2}\right) + \frac{v_{ik}^2}{v_i^2}g_i' \geq 0 \tag{6.19}$$

Dies gilt, da g_i und g_i' für das positive Argument $|v_i|$ positiv sind. Weiter muß man zeigen, daß

$$h_{ik}^{ik} > \sum_{jl \neq ik} \left| h_{ik}^{jl} \right| \tag{6.20}$$

Die rechte Seite kann geschrieben werden als,

$$\sum_{jl \neq ik} \left| \frac{v_{ik} v_{jl} \delta_j^i}{v_i^2} \left(-\frac{g_i}{v_i} + g_i' \right) \right| = \sum_{l \neq k} \left| \frac{v_{ik} v_{jk}}{v_i^2} \left(-\frac{g_i}{v_i} + g_i' \right) \right| = \frac{|v_{il}|}{v_i^2} \left| -\frac{g_i}{v_i} + g_i' \right| \sum_{l \neq k} |v_{il}| \tag{6.21}$$

Betrachtet man (6.19), (6.20) und (6.21) so muß man zeigen,

$$\frac{g_i}{v_i} - \frac{v_{ik}^2}{v_i^2} \left(-\frac{g_i}{v_i} + g_i' \right) - \frac{|v_{ik}|}{v_i^2} \left| -\frac{g_i}{v_i} + g_i' \right| \sum_{l \neq k} |v_{il}| > 0 \tag{6.22}$$

was äquivalent ist zu,

$$|v_{ik}| \left(-\frac{g_i}{v_i} + g_i' \right) - \left| -\frac{g_i}{v_i} + g_i' \right| \sum_{l \neq k} |v_{il}| < g_i \frac{v_i}{|v_{ik}|} \tag{6.23}$$

Mit

$$-\frac{g_i}{v_i} + g_i' < 0 \tag{6.24}$$

läßt sich (6.23) schreiben als,

$$\left(\frac{g_i}{v_i} - g_i' \right) \sum_{l} |v_{il}| < g_i \frac{v_i}{|v_{ik}|} \tag{6.25}$$

Da für einen beliebigen Vektor $|v_k| \leq \sum_l |v_l| \leq |\mathbf{v}|$ gilt, reicht es zu überprüfen ob,

$$\left(\frac{g_i}{v_i} - g_i' \right) v_i < g_i \tag{6.26}$$

Das ist gemäß den entsprechenden Definitionen gültig, da

$$0 < g_i' v_i \tag{6.27}$$

6.3.2 Fixpunktiteration

Zur Berechnung der Lösung der deterministischen Gleichungen (6.11) und (6.12) kann man auch folgende Fixpunktiteration verwenden:

$$\mathbf{v}_i(t+1) = \mathbf{f}\left(-\frac{1}{T}\sum_j \mathbf{w}_{ij}\cdot\mathbf{v}_j(t)\right) \tag{6.28}$$

Neben den eben bewiesenen Konvergenzeigenschaften für den Fall kontinuierlicher Zeit, kann man hier explizit die lokale Konvergenz dieser Iteration nachweisen. Man wird erkennen, daß oberhalb einer kritischen Temperatur das System auch hier in die "antimagnetische" Phase mit verschwindender mittlerer Magnetisierung $\mathbf{V} = \langle\mathbf{S}\rangle = 0$ konvergiert. Gemäß dem Banachschen Fixpunktsatz ist lokale Konvergenz garantiert, falls

$$\left\|\frac{\partial f_{ik}}{\partial v_{jl}}\right\| < 1 \tag{6.29}$$

$$\left\|\frac{\partial f_{ik}}{\partial v_{jl}}\right\| = \left\|\sum_{nm}\frac{\partial f_{ik}}{\partial u_{nm}}\frac{\partial u_{nm}}{\partial v_{jl}}\right\| = \left\|\nabla_{\mathbf{v}}\mathbf{f}\cdot\frac{\mathbf{W}}{T}\right\| \tag{6.30}$$

Es werden die Hilfsgrößen h und g jetzt neu definiert. Es sei nun $h = \nabla_{\mathbf{v}}\mathbf{f}$ der neue Tensor vierter Stufe. Man verwende jetzt die Abkürzungen $g_i = F(|\mathbf{u}_i|)$ und $u_i = |\mathbf{u}_i|$. Ersetzt man in Gleichung (6.18) $\mathbf{V}$ durch $\mathbf{U}$, so gilt diese Gleichung mit den neuen g_i analog. Da die Bedingungen (6.24) und (6.27) entsprechend gelten, sind die Ungleichungen analog zu (6.19) und (6.20) ebenfalls gültig. Der neue Tensor h ist daher positiv definit mit positiven Eigenwerten

$$\lambda_{ik} \in \left[h^{ik}_{ik} + \sum_{ik\neq jl}|h^{jl}_{ik}|\ ,\ h^{ik}_{ik} - \sum_{ik\neq jl}|h^{jl}_{ik}|\right] \tag{6.31}$$

Nun läßt sich folgende obere Grenze für die Norm angeben

$$\begin{aligned}
\|\nabla_{\mathbf{v}}\mathbf{f}\| &= \max_{ik}|\lambda_{ik}| \leq \max_{ik}\left(h^{ik}_{ik} + \sum_{ik\neq jl}|h^{jl}_{ik}|\right) && (6.32)\\
&< \max_{ik}(2h^{ik}_{ik}) && (6.33)\\
&= 2\max_{ik}\left(\frac{F_i}{u_i} + \frac{u^2_{ik}}{u^2_i}\left(-\frac{F_i}{u_i} + F'_i\right)\right) && (6.34)\\
&\leq 2\max_{i}\left(\frac{F_i}{u_i} + \left(-\frac{F_i}{u_i} + F'_i\right)\right) = \frac{2}{d} && (6.35)
\end{aligned}$$

wobei $1/d$ die maximale Steigung von F am Nullpunkt angibt. Schließlich erhält man aus (6.29) eine Bedingung für lokale Konvergenz:

$$\left\|\frac{\partial f_{ik}}{\partial v_{jl}}\right\| = \left\|\nabla_{\mathbf{v}}\mathbf{f}\cdot\frac{\mathbf{W}}{\mathbf{T}}\right\| \leq \|\nabla_{\mathbf{v}}\mathbf{f}\|\left\|\frac{\mathbf{W}}{\mathbf{T}}\right\| < \frac{2}{Td}\|\mathbf{W}\| < 1 \tag{6.36}$$

Für Temperaturen oberhalb einer Grenztemperatur $T_g = 2\|\mathbf{W}\|/d$ läßt sich die Konvergenz der Fixpunktiteration also explizit nachweisen. Die berechnete Grenze ergibt sich für die Stelle $\langle \mathbf{S} \rangle = 0$. Das bedeutet, daß oberhalb dieser Temperatur die verschwindende "Magnetisierung" eine stabile Lösung der Mittelfeldgleichungen ist. Man beachte allerdings, daß einige grobe Abschätzungen gemacht wurden, wodurch dem Faktor 2 keine besondere Signifikanz gegeben werden kann. Von einer kritischen Temperatur im eigentlichen Sinne kann somit nicht die Rede sein. Jedenfalls zeigt sich in Experimenten, daß die Fixpunktiteration (6.28) für beliebige Temperaturen innerhalb weniger Schritte konvergiert. Für Temperaturen, die deutlich über dieser Grenze liegen ($T >> T_g$), konvergieren die Gleichungen gegen die triviale Lösung. Bei Temperaturen um T_g nimmt der Betrag der Zustandsmittelwerte zu. Die Zustandsmittelwerte konvergieren dann für Temperaturen mit $T << T_g$ gegen stabile Lösungen auf der Einheitssphäre.

6.4 Lernen im Kontinuierlichen - Funktionsapproximation

Für das Lernen verwendet man das Kriterium der minimalen relativen Entropie. Man weiß von der klassischen Boltzmann-Maschine, daß sich die resultierenden Lerngleichungen durch Mittel der Zustandsvariablen darstellen lassen. Dies ist wichtig, da sich hier im Gegensatz zum diskreten Zustandsraum die Zustandssummen nicht mehr explizit berechnen lassen.

Zur korrekten Formulierung der Lerngleichungen sowie deren Ableitung ist es ausreichend, in der Ableitung der Lerngleichungen von Abschnitt 3.2 (Seite 27), die Summe über diskrete Zustände mit den entsprechenden Integralen über kontinuierliche Zustände zu ersetzen:

$$\sum_{\mathbf{s}} \rightarrow \int_{\Phi} d\mathbf{S}$$

wobei die Summen entsprechend über die verschiedenen Räume der inneren und sichtbaren Zustände $\mathbf{S}^h$ bzw $\mathbf{S}^v$ [5] zu erstrecken sind. Das Ausschreiben all dieser Definitionen und Ableitungen weist keine analytischen Hindernisse auf und bedürfte hier lediglich einer umständlichen Notation. Es wird daher direkt das Resultat angegeben:

$$\Delta w_{ikjl} = \frac{\varepsilon}{T} \left[\langle s_{ik} s_{jl} \rangle_{fixiert} - \langle s_{ik} s_{jl} \rangle_{frei} \right] \tag{6.37}$$

[5] Die Schreibweise variiert hier etwas vom Rest der Arbeit. Konsequenterweise müßte man für die inneren und sichtbaren Zustände jetzt $\mathbf{V}$ und $\mathbf{H}$ schreiben. Dies würde aber zu Verwechslungen mit den Mittelfeldvariablen $\mathbf{V}$ oder sogar der Entropie H führen.

Die Indizierung *fixiert* und *frei* kennzeichnet die verschiedenen Mittel, die man auch in Abschnitt 3.2 findet. Diese Mittel sind bei festen bzw. freien, sichtbaren Zuständen $\mathbf{S}^v$ zu bestimmen. Man verwendet wieder die zusätzliche Näherung

$$\langle s_{ik} s_{jl} \rangle = \langle s_{ik} \rangle \langle s_{jl} \rangle$$

Auch in diesem Kapitel ist man an einer funktionalen Abhängigkeit zwischen Eingabe und Ausgabe interessiert. Die Ausgabe ist nun ebenfalls kontinuierlich. Man trennt dazu die sichtbaren Elemente weiter in Eingabe $\mathbf{X}$ und Ausgabe $\mathbf{Y}$. Die Lerngleichungen sind dann analog zu denen der Standard-Boltzmann-Maschine (3.10), wobei man hier keine Unterscheidung zwischen externen Kopplungen und den rückgekoppelten Wechselwirkungen macht. Die Lerngleichungen schreiben sich mit den Mittelfeldvariablen wie folgt:

$$\Delta w_{ikjl} = \frac{\varepsilon}{T} \left[\langle v_{ik} v_{jl} \rangle_{\mathbf{Y},\mathbf{X}} - \langle v_{ik} v_{jl} \rangle_{\mathbf{X}} \right] \tag{6.38}$$

Die Mittelwerte sind gegeben durch

$$\langle f(\mathbf{S}) \rangle_{\mathbf{Y},\mathbf{X}} = \int_{\Phi} d\mathbf{S} P(\mathbf{S}^h | \mathbf{Y}, \mathbf{X}) R(\mathbf{Y}, \mathbf{X}) f(\mathbf{S})$$

$$\langle f(\mathbf{S}) \rangle_{\mathbf{X}} = \int_{\Phi} d\mathbf{S} P(\mathbf{S}^h, \mathbf{Y} | \mathbf{X}) R(\mathbf{X}) f(\mathbf{S})$$

Die Zielverteilung $R(\mathbf{Y}, \mathbf{X})$ ist durch einen Satz von Datenpunkten gegeben. Die Integration über diese Verteilung reduziert sich auf die Summe über alle diese Datenpunkte. Zur Bestimmung der *fixierten* und *freien* Mittel wird die deterministische Dynamik, die durch die Fixpunktgleichungen (6.28) gegeben ist, in den zwei verschiedenen Relaxationsphasen iteriert. Beim *fixierten* Mittel $\langle v_{ik} v_{jl} \rangle_{\mathbf{Y},\mathbf{X}}$ werden Eingabe und Ausgabe auf die vorgegebenen Lerndaten fixiert und die Ergebnisse der Relaxation über alle Lernmuster gemittelt. Beim *freien* Mittel $\langle v_{ik} v_{jl} \rangle_{\mathbf{X}}$ werden während der Relaxation lediglich die Eingabezustände auf die vorgegebenen Datenpunkte fixiert, um das Resultat dann über alle verschiedenen Eingabemuster zu mitteln. Ein Funktionswert der modellierten Abbildung kann nach dem Lernen abgerufen werden, indem man die Eingabeelemente auf die gewünschte Eingabe fixiert und das System bei kleinen Startwerten $V_i \approx 0$ relaxiert. Bei all diesen Relaxationen verwendet man die bekannte Abkühlung.

Die Lerngleichungen, die für das unüberwachte Lernen in den Kapiteln 4 und 5 abgeleitet wurden, lassen sich völlig analog für die Rotor-Boltzmann-Maschine formulieren. Man hat nun aber die Schwierigkeit, daß die Integrale über die kontinuierlichen Zustände explizit ausgeführt werden müssen. Es ist nun nicht mehr möglich, diese Gleichungen allein durch Mittelwerte der Zustände zu formulieren. Die Hauptschwierigkeit für eine analytische Lösung ist dabei der Logarithmus von Wahrscheinlichkeiten, der in den Lerngleichungen (4.18) bzw. (5.14) auftaucht. Die Integrale lassen sich natürlich durch Monte-Carlo-Simulationen berechnen,

indem man die entsprechende Glauber-Dynamik, so wie sie in Abschnitt 3.1.2 erklärt wurde, durchführt. Dies ist analog zu den Modellen, die zur Simulation der thermischen Stochastik Langevin-Gleichungen verwenden. Diese Ansätze sind extrem rechenaufwendig und daher für die Anwendungen von konnektionistischen Modellen uninteressant. Man ist dazu gezwungen, die Stochastik mit Hilfe der Mittelfeldgleichungen zu eliminieren. Aus diesem Grund wird man sich also auf das Lernen mit dem Kriterium der minimalen relativen Entropie beschränken müssen.

6.5 Simulationen

Zunächst soll in Simulationen überprüft werden, ob das vorgeschlagene System die erwarteten Konvergenzeigenschaften aufweist. In Vorexperimenten zeigte sich, daß die Fixpunktgleichungen (6.28) unabhängig von der Temperatur in wenigen Zyklen konvergieren. Das gilt auch für den Fall von unsymmetrischen Kopplungen. Bei einer hohen Temperatur konvergieren die Gleichungen zum Ursprung $V_i \approx 0$. Für abnehmende Temperaturen nimmt die Norm der Lösungen zu, bis eine Grenztemperatur erreicht wird. Dort sind $V_i \approx 1$ und die Werte bleiben konstant. Man beobachtet in den Experimenten, daß die Grenztemperatur T_g mit der Kopplungsstärke korreliert ist ($\|W\|/T_g$ konstant der Größenordnung 1). Dies bestätigt das theoretische Resultat von Abschnitt 6.3.2. Es gibt eine Richtlinie für die Abkühlung des Systems. Man beginnt etwas oberhalb der Grenztemperatur und kühlt allmählich ab. In den Experimenten wurde bei einer Temperatur von $T = 1.0$ begonnen und mit einem Abkühlungsfaktor von 0.85 auf eine Endtemperatur von 0.001 abgekühlt. Prinzipiell muß bei jeder Temperatur die Fixpunktiteration bis zur Konvergenz durchgeführt werden. Erst dann kann eine neue, niedrigere Temperatur eingestellt werden. Man erhält jedoch identische Resultate, falls Abkühlung und Iteration der Fixpunktgleichungen gleichzeitig durchgeführt werden. Das bewirkt natürlich eine entscheidende Beschleunigung des Relaxationsalgorithmus.

6.5.1 Zuordnung der Signale an die kontinuierlichen Zustände

Eine kontinuierliche Abbildung erfordert mindestens zwei-dimensionale Rotoren. Aufgrund der Normalisierungsbedingung (6.1) für die Rotorzustände benötigt man einen $(d+1)$-dimensionalen Rotor, um ein d-dimensionales Signal mit einem einzigen Ausgabeelement zu repräsentieren. Man kann aber auch n zweidimensionale Rotoren verwenden, um ein n-dimensionales Ausgabesignal darzustellen. Für die Eingabeelemente gilt diese Einschränkung nicht. Der Betrag der Koordinate, die zur Normierung verwendet wird, ist durch die Normierungsbedingung bestimmt. Das Vorzeichen kann frei gewählt werden. Man legt sich hier in allen

Fällen auf positive Werte fest. Diese Wahl ist zufällig und hat keinen Einfluß auf die Abbildung, die modelliert werden soll. Man bemerke, daß das vorliegende Modell leicht auf ein System mit Rotoren verschiedener Dimensionen erweitert werden kann.

In zwei einfachen Vorexperimenten wurden zweidimensionale Zustände verwendet. Man bestätigte mit dem klassischen XOR-Problem die Funktionsfähigkeit der Lerngleichungen und erhält ähnliche Ergebnisse wie in der ursprünglichen Arbeit von Peterson und Anderson (1987). Weiter wurde die Fähigkeit bestätigt, eine einfache kontinuierliche Abbildung wie die Sinusfunktion zu lernen. Das System wurde mit 20 Datenpunkten trainiert und nähert die Funktion nahezu perfekt. (0.9% quadratischer Fehler auf den Lerndaten im Verhältnis zum Fehler der trivialen Lösung $V_i = 0$)

6.5.2 Eindimensionale, stückweise stetige Abbildung

Es soll die inhärente Fähigkeit des Systems, stückweise stetige Abbildungen zu modellieren, erklärt und demonstriert werden. Unstetigkeit bedeutet, daß an einer bestimmten Stelle eine kleine Änderung in der Eingabe zu einer drastischen Änderung der Ausgabewerte führt. Man beobachtet das, falls das System für leicht veränderte Eingabewerte zu völlig neuen Fixpunkten konvergiert. Man erinnere sich, daß eine Lyapunov-Funktion existiert und ein Fixpunkt als Minimum eines Energietals verstanden werden kann. Die Fixpunktiteration ist äquivalent zu einem Gradientenabstieg auf dieser Energiefläche mit einer Schrittweite von 1.0. Dies wurde in Abschnitt 6.3 erklärt. Die Eingabe wird während der Relaxation konstant gehalten. Sie spielt die Rolle eines konstanten externen Feldes, das die Energiefunktion parametrisiert. Die Energielandschaft hängt also kontinuierlich von der Eingabe ab. Beginnt die Relaxation bei stets der gleichen Konfiguration der inneren und äußeren Ausgabezustände, so wird sie immer zum gleichen Fixpunkt konvergieren. Ändert sich die Eingabe ein wenig, so wird sich die Lage des Fixpunktes, zu dem das System konvergiert, ebenfalls nur wenig verändern. Das gilt unter der Voraussetzung, daß die Startkonfiguration bei der geänderten Eingabe weiterhin im gleichen Anziehungsgebiet dieses Fixpunktes verbleibt. Andernfalls konvergiert das System zu einem völlig neuen Fixpunkt, mit u.U. drastisch verschiedenen Ausgabewerten. Man erwartet also, daß das System stückweise kontinuierliche Abbildungen verwirklichen kann. Tatsächlich zeigt das System in Simulationen die Tendenz, stückweise stetige Abbildungen zu erzeugen. Wie in nächsten Beispiel demonstriert wird, kann die Lage der Unstetigkeit gelernt werden.

Es wird eine willkürliche, stückweise stetige, eindimensionale Abbildung mit einer Unstetigkeit am Ursprung als Zielfunktion gewählt: $f(x) = sig(x)exp(-|x|)$. Abbildung 6.2 zeigt den quadratischen Fehler, der während des Lernens aufgenommen wurde. Die Lage der erzeugten Unstetigkeit hängt empfindlich von den Kopplungsstärken ab. Das erklärt die Ausschläge des quadratischen Fehlers.

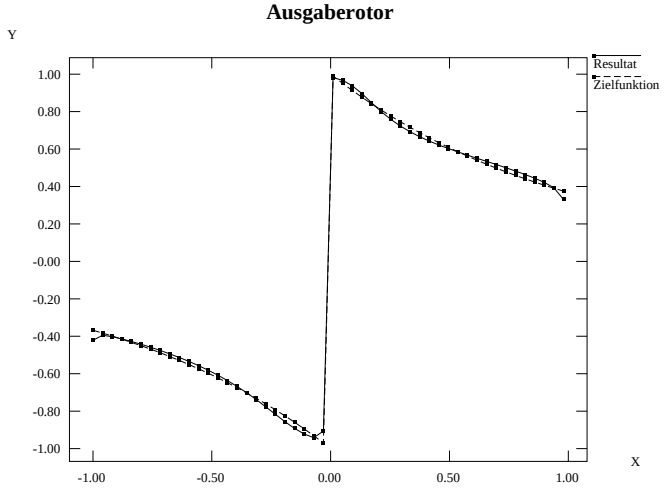

Abbildung 6.1: Eine Rotor-Boltzmann-Maschine mit fünf zweidimensionalen Elementen (drei innere Rotoren, eine Ausgabe und eine Eingabe) wurde mit dem Kriterium der minimalen relativen Entropie optimiert. Die abgebildete stückweise stetige Funktion wird erfolgreich modelliert. "x" und "y" kennzeichnen jeweils eine Koordinate des Eingabe- bzw. des Ausgaberotors. Die zweite Dimension ergibt sich gemäß der Normierungsbedingung für Rotoren.

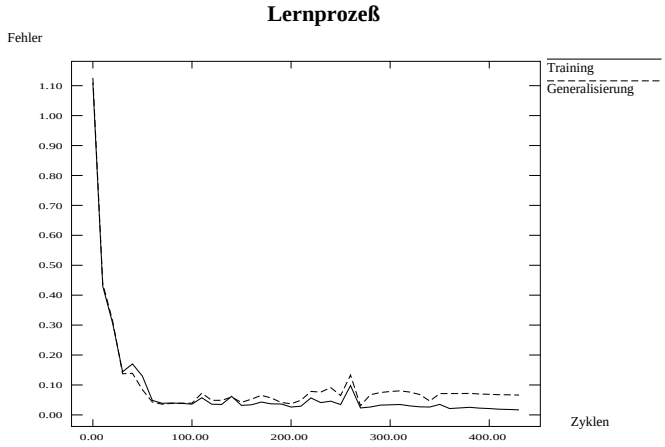

Abbildung 6.2: Entwicklung des quadratischen Fehlers während des Lernens für das Beispiel der vorhergehenden Abbildung. Ein Zyklus bedeutet, daß man die Lerngleichungen für alle 50 vorgegebenen Datenpunkte berechnet. Die Kurven zeigen den quadratischen Fehler für die Trainingsdaten sowie den Generalisierungsfehler für Zwischenwerte, die nicht in den Trainingsdaten enthalten sind. Nach 400 Zyklen erreicht man ein Fehler von 2.0%.

Für einen Gradientenabstieg mit ausreichend kleiner Lernrate ϵ sollte der Fehler monoton absteigen. Um diese Ausschläge zu kompensieren ohne die Lerngeschwindigkeit zu kompromittieren, wurde die Lernratensteuerung von Silva und L. (1990) verwendet. Das Ergebnis ist in Abbildung 6.1 zu sehen.

6.5.3 Modellierung einer stückweise stetigen Kontrollfunktion

Eine interessante Anwendung für die Modellierung von stückweise stetigen Funktionen findet man bei Kontrollproblemen. Gelegentlich trifft man auf das Problem, verschiedene Kontrollstrategien kombinieren zu müssen. Innerhalb verschiedener Bereiche des Raums, in dem das System gesteuert werden muß, ist das Kontrollsignal zumeist eine stetige Funktion des Systemzustands. An den Grenzen dieser Bereiche, an denen sich die Kontrollstrategie verändern soll, weist die Kontrollfunktion Unstetigkeiten auf. Für diese Art von Kontrollproblemen ist die Berechnung der Kontrollfunktion an einem bestimmten Punkt oftmals ein numerisch extrem aufwendiger Prozeß. In Anwendungen, bei denen das Kontrollsignal in Echtzeit zur Verfügung stehen muß, wünscht man sich daher eine Methode zur Modellierung der stückweise stetigen Kontrollfunktion. Das Modell kann dann schneller Näherungswerte zur Verfügung stellen.

Hier wird ein Beispiel aus der Differenzialspiel-Theorie verwendet, das von Breakwell (1977) vorgestellt wurde. Bei dem "cornered rat"-Problem stellt sich folgende Aufgabe: Man stelle sich eine Maus vor, die versucht einer Katze zu entweichen, indem sie in eines von zwei Löchern in einem rechteckigen Spielfeld gelangt. Etwas allgemeiner ausgedrückt, versucht die Maus ihre Lebenszeit zu verlängern. Wenn die Maus voraussehen kann, daß sie keines der zwei Löcher erreichen kann, so läuft sie in die Richtung, bei der die Katze eine maximale Zeit benötigt, sie zu fangen, also zu einer Spielfeldecke. Je nach aktueller Position wird die Maus verschiedene Richtungen wählen. Die optimalen Richtungen werden in Abbildung 6.3 ersichtlich. Man möchte nun die entsprechende Kontrollfunktion für die optimale Richtung in Abhängigkeit zur aktuellen Position modellieren. Es soll nicht das Optimalitätsproblem an sich gelöst werden. Es soll nur die numerisch berechnete Kontrollfunktion modelliert werden. Gabler, Miesbach, Breitner, und Pesch (1993) berechnen die Lösung des Problems numerisch mit dem sogenannten Überlebensmengen-Algorithmus. Tatsächlich sind die Richtungen, die in Abbildung 6.3 abgebildet sind, aus dieser Arbeit entnommen. Die gleiche stückweise stetige Funktion ist auch in Abbildung 6.4 zu sehen. Diese wird als Zielfunktion für das Lernen verwendet.

Es wurden vier innere, zweidimensionale Rotoren verwendet, zwei Eingabeelemente für die Position der Maus und ein Ausgaberotor, der direkt die Kontrollrichtung angeben soll. Der Wertebereich der zwei Eingabekoordinaten $[0, 10] \times [0, 6]$ wurde auf $[-1, +1] \times [-1, +1]$ linear transformiert. Als Trainings-

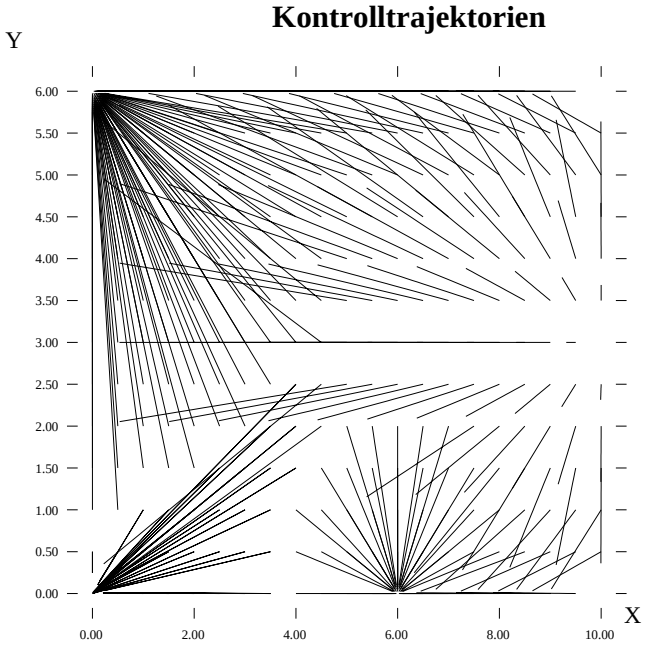

Abbildung 6.3: Von Gabler et.al. (1993) berechnete Bahnen für die Maus im "cornered rat"-Differenzialspiel. Ausgehend von verschiedenen Positionen versucht eine Maus einer Katze zu entfliehen, die in Position (10,3) sitzt. Die Mauslöcher sind bei (0,6) und (6,0).

daten wurden 13×21 Stichproben der Zielfunktion verwendet, die in Abbildung 6.4 zu sehen sind. Nach 5000 Zyklen reduzierte sich der mittlere Winkelfehler auf 5.45^o. Die resultierende Kontrollfunktion ist in Abbildung 6.5 zu sehen und der resultierende verbleibende Fehler an den Trainingspunkten in Abbildung 6.6. Die Übereinstimmung der Unstetigkeitsstellen ist auffällig. Zum Vergleich wurde ein Standardnetzwerk, das Multi-Layer-Perceptron (MLP) von Rumelhart et al. (1988), mit sieben inneren Neuronen und einer Ausgabe mit sinusförmiger Aktivierungsfunktion trainiert. Nach 6000 Zyklen mit einem "online"-Gradientenabstieg erreicht man einen mittleren Fehler von 16.28^o. Das Resultat ist in Abbildung 6.7 zu sehen. Basierend auf den gleichen Trainingsdaten wurde auch ein Netzwerk mit normierten, radialen Basisfunktionen und linearem Ausgabeelement trainiert (Moody und Darken, 1989). Das beste Resultat mit einem mittleren Fehler von 9.19^o erhielt man für acht innere Neuronen und der Initialisierung der Gewichte mit dem "K-nearest-neighbor"-Algorithmus gemäß (Moody und Darken, 1989; Duda und Hart, 1973). Vergleicht man Abbildungen 6.5 und 6.7, so erkennt man die Schwierigkeit des MLP, die Unstetigkeiten nachzubilden. Im Gegensatz dazu modelliert die Rotor-Boltzmann-Maschine diese problemlos.

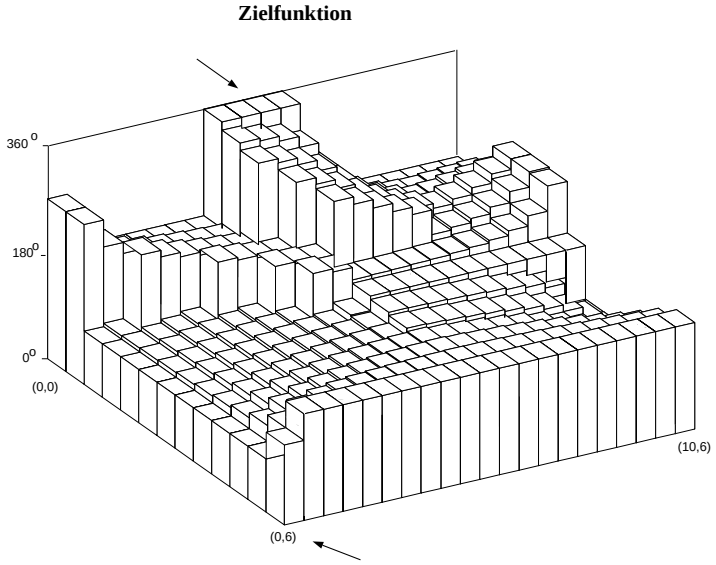

Abbildung 6.4: Von Gabler et.al. (1993) berechnete optimale Kontrollrichtungen in Abhängigkeit der Position der Katze. Die vertikale Achse repräsentiert den Winkel der Richtungen der vorhergehenden Abbildung. Die Pfeile zeigen auf die Position der Mauslöcher.

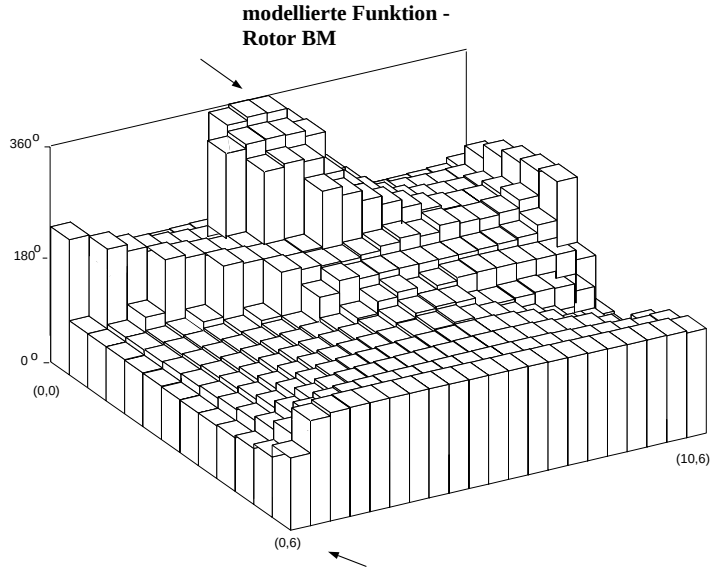

Abbildung 6.5: Modell der Kontrollfunktion, die von der Rotor-Boltzmann-Maschine nach 5000 Lernschritten mit einem abschließenden mittleren Fehler von 5.48^o erzeugt wurde.

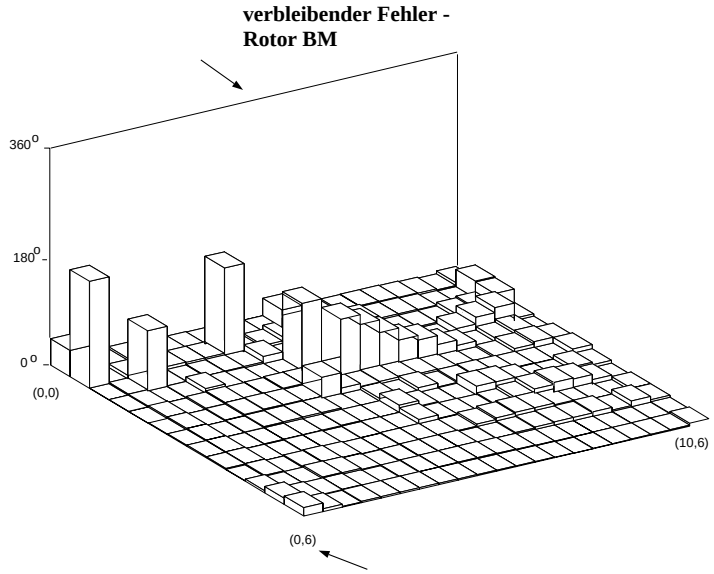

Abbildung 6.6: Verbleibender Fehler der Rotor-Boltzmann-Maschine. Lediglich die genaue Lage der Unstetigkeit ist an manchen Stellen etwas verfehlt.

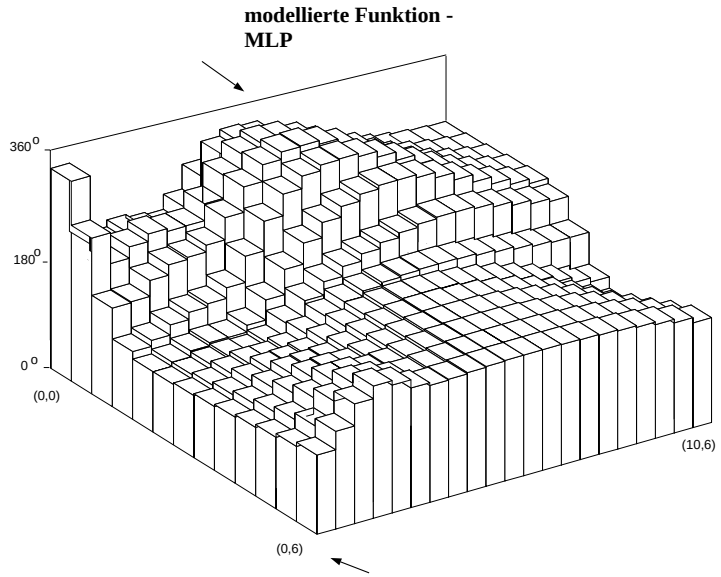

Abbildung 6.7: Modell der Kontrollfunktion, die von einem Multi-Layer-Perceptron nach 6000 Lernschritten mit einem abschließenden mittleren Fehler von 16.28^o erzeugt wurde.

6.6 Mittelfeldgleichungen für einfache kontinuierliche Zustände

Nachdem man dieses System kennengelernt hat, stellt sich die Frage, ob man die mehrdimensionalen Rotoren wirklich benötigt. Prinzipiell lassen sich auch eindimensionale, kontinuierliche, stochastische Elemente definieren. Es ist denkbar, mit der Mittelfeldnäherung deterministische Gleichungen der Mittelwerte zu erhalten. Die Mittelfeldnäherung stößt hier aber auf ein analytisches Problem. Beschränkt man den Zustandsraum Φ nicht, so hat man keine Garantie für die Beschränktheit des Integrals der Zustandssumme. Man kann nun den Zustandsraum ohne Beschränkung der Allgemeinheit auf $s_i \in [-1,+1]$ limitieren.[6] Der gesamte Zustandsraum aller s_i ist jetzt $\Phi = [-1,+1]^n$. Die Zustandssumme wandelt sich bei Einführung der Mittelfeldvariablen v_i und u_i wie folgt um,

$$Z = \int_{\Phi} e^{-\beta E(\mathbf{s})d\mathbf{s}} \propto \int_{\Re^n}\int_{\Re^n} d\mathbf{v}d\mathbf{u}e^{-\frac{1}{T}E(\mathbf{v})-\mathbf{v}\cdot\mathbf{u}+\sum_i \ln F(u_i)}$$

wobei sich für die Integration von $\mathbf{s}$ in Φ folgendes ergibt:

$$F(u) = \ln \int_{[-1,+1]} e^{us} ds = \ln\left(\frac{e^{us}}{u}\right)\Big|_{-1}^{1} = \ln\left(\frac{2\sinh(u)}{u}\right)$$

Die Ableitungen der effektiven Energie nach v_i und u_i ergeben die entsprechenden Sattelpunktsgleichungen

$$v_i = f\left(-\frac{1}{T}\frac{\partial E(\mathbf{v})}{\partial v_i}\right)$$
$$f(u) = \frac{1}{\tanh(u)} - \frac{1}{u}$$

Diese Gleichungen sind qualitativ gleich dem Resultat mit binären Spin-Zuständen (3.7). Lediglich die Aktivierungsfunktion $f(u)$ ist verschieden. Sie hat aber nach wie vor die sigmoide Struktur des Tangens hyperbolicus. Für die Grenztemperatur $T \to 0$ liefert eine Fixpunktiteration nur zwei mögliche stabile Lösungen $\{-1,+1\}$. Das bedeutet allerding nicht, daß es keine anderen reellwertigen Lösungen gibt. Im binären Fall hatten diese Lösungen nur keine sinnvolle Bedeutung. Nichtsdestotrotz ist dieses Resultat nicht praktikabel, da keine einfachen rekursiven Gleichungen zur Berechnung der kontinuierlichen Lösung mehr gegeben sind.

[6] Dies wurde in einem persönlichen Gespräch von Sompolinsky (1994) vorgeschlagen.

6.7 Schlußfolgerungen

Das Ziel dieses Kapitels war die Formulierung eines kontinuierlichen Systems, das die Umsetzung eines informationstheoretischen Lernkriteriums, wie das der minimalen relativen Entropie, erlaubt. Die Boltzmann-Verteilung ermöglicht das, indem sie einen expliziten, analytischen Ausdruck für die Wahrscheinlichkeitsdichte von kontinuierlichen Zuständen angibt. Die für den kontinuierlichen Fall extrem aufwendige stochastische Dynamik wird umgangen, indem man eine deterministische Dynamik für die Mittelwerte ableitet. Die zugrundeliegende Mittelfeldtheorie zwingt jedoch zur Betrachtung mehrdimensionaler Zustände. Man erhält eine kontinuierliche Erweiterung der deterministischen Boltzmann-Maschine. Die Arbeit in diesem Kapitel verallgemeinert ähnliche zweidimensionale Modelle auf beliebige Dimensionen. Die Konvergenzeigenschaften dieses neuen Modells wurden genauer untersucht und in Experimenten verifiziert. Trainiert man das System mit dem Kriterium der minimalen relativen Entropie, so kann es zur Funktionenapproximation verwendet werden. Insbesondere zeigt dieses System die Fähigkeit stückweise stetige Funktionen zu modellieren.

Leider lassen sich die Lerngleichungen, die mit den Kriterien der maximalen Transinformation und der minimalen gemeinsamen Information für eine kontinuierliche Boltzmann-Maschine ableitbar sind, nicht praktikabel umsetzen. Die Boltzmann-Maschine war bis zu dieser Stelle nützlich, indem sie die klare Formulierung von informationstheoretischen Lernkriterien als differenzierbare Kostenfunktionen erlaubt hat. Die thermische Stochastik war dabei ein Schlüsselelement. Gleichzeitig erschwert eben diese Stochastik die Umsetzung dieser Kriterien für kontinuierliche Signale. Im nächsten Kapitel wird daher ein völlig neuer Ansatz gewählt. Man wird von vornherein nur ein deterministisches System zulassen und keinen analytischen Ausdruck für Wahrscheinlichkeitsdichten besitzen. Eine der Hauptaufgaben ist dann, eine gute Schätzung der Entropien aus den gemessenen Signalen zu gewinnen.

Kapitel 7

Minimale gemeinsame Information und Volumenerhaltung

Annahmen und Zielsetzung

Das Ziel dieser Arbeit ist es, unüberwachtes Lernen für kontinuierliche Signale zu definieren.

Für diskrete Signale wurde bisher maximale Transinformation und minimale gemeinsame Information an der Ausgabe vorgeschlagen. Es wurde ein probabilistisches System untersucht, welches die Ableitung von Lernkriterien in geschlossener Form ermöglichte. Die Stochastik erschwerte andererseits die praktische Umsetzung der gewonnenen Lernregeln. Daher konzentriert sich das folgende Kapitel auf deterministische Transformationen von kontinuierlichen Signalen.

Es wird angenommen, daß das lernende System keinerlei Information über den Ursprung der Signale hat. Insbesondere hat man keine Information über Signal- oder Rauschstatistik. Bevor man ein Modell der Signaldichten gebildet hat, ist es schwierig, Signal und Rauschen voneinander zu trennen. Dem verwendeten deterministischen System wird daher die Aufgabe gestellt, die Information des Eingabesignals vollständig auf die Ausgabe zu übertragen.

Wie in Abschnitt 7.1 diskutiert, ist die Transinformation für eine deterministische kontinuierliche Abbildung kein geeignetes Maß. Statt dessen verwendet man hier das Prinzip der *konstanten Informationsübertragung*, indem man lediglich volumenerhaltende Transformationen zuläßt. Es werden mehrere volumenerhaltende, adaptive, konnektionistische Abbildungen vorgestellt. Merkmalsextraktion wird auch für kontinuierliche Signale durch minimale gemeinsame Information an der Ausgabe verwirklicht. Die konstante Informationsübertragung erlaubt dabei eine entscheidende Vereinfachung der Optimierungsaufgabe.

Die Hauptschwierigkeit ist nun die Messung der Entropien an den einzelnen Ausgabekoordinaten, da kein analytischer Ausdruck für die Signalverteilungen

vorliegt. Hierzu werden mehrere Näherungsmethoden vorgestellt, die gemessene Kumulanten der Signalverteilung verwenden.

Das vorgestellte System kann zur Dichteschätzung und insbesondere zur Detektion unerwarteter Ereignisse verwendet werden. Man wendet diese Methode im Fall des technologischen Problems der Fehlervorhersage bei der automatisierten Überwachung von Elektromotoren an.

Verwandte Fragestellungen

Bevor dieses Konzept entwickelt wird, soll die relevante Literatur aus verwandten Forschungsgebieten vorgestellt werden.

Die Aufgabe der Merkmalsextraktion ist eng mit der linearen Hauptkomponentenanalyse (PCA) verwandt. Das Ziel der klassischen PCA ist es, Merkmale linear zu dekorrelieren, indem die Koordinaten des Merkmalsraums rotiert werden. In den Arbeiten über Neuronale Netze gibt es zahlreiche Veröffentlichungen, die sich mit der linearen PCA beschäftigen. Der größte Teil dieser Arbeiten zeigt, wie verschiedene konnektionistische Strukturen PCA verwirklichen und wie diese aus informationstheoretischen Prinzipien abgeleitet werden können. Letztere sind insbesondere wichtig, um ein Verständnis über die verschiedenen Lernalgorithmen zu gewinnen. Die Ergebnisse sind allerdings auf normalverteilte Eingabesignale beschränkt.

Unabhängig voneinander zeigen Baldi und Hornik (1989) und Bourlard und Kamp (1988), daß Back-Propagation in einem linearen Autoassoziationsnetz in der inneren Schicht eine orthogonale Projektion der Eingabesignale auf einen Unterraum generiert. Dieser Unterraum wird von den Eigenvektoren der ersten Hauptkomponenten der Autokorrelationsmatrix der Eingabedaten aufgespannt. Oja (1989) formuliert eine Hebb'sche Lernregel, die die ersten Hauptkomponenten extrahiert und zu einer orthogonalen Transformation führt. Rubner und Tavan (1989) zeigen, daß ein Netzwerk mit linearen Neuronen in der Lage ist, eine Hauptkomponentenanalyse durchzuführen. Das Netz verwendet dabei eine Hebb'sche Lernregel für die Eingabe-/Ausgabe-Kopplungen und eine Anti-Hebb'schen Lernregel für die lateralen Kopplungen zwischen den Ausgabeneuronen. Ein ähnliches Konzept wurde von Földiák (1989) vorgeschlagen. Die Hebb'schen und Anti-Hebb'schen Lernregeln können auch aus informationstheoretischen Kriterien abgeleitet werden (Linsker, 1988; Kuehnel und Tavan, 1990). Obradovic und Deco (1995) kombinieren Dekorrelation mit der allgemeinsten Klasse von linearen, volumenerhaltenden und somit informationserhaltenden Abbildungen. Fügt man die Bedingung der Normerhaltung hinzu, so ergibt sich eine Hauptkomponentenanalyse. In einer parallelen Arbeit formulieren Deco und Brauer (1994) das Prinzip der minimalen gemeinsamen Information für beliebige Signaldichten als eine spezielle Form der Diagonalisierung von multivariaten Momenten höherer Ordnung. Sie schlagen dabei eine bestimmte Klasse von nichtlinearen volumenerhaltende Abbildungen vor. Die Berechnung und Optimierung

von multi-variaten Momenten höherer Ordnung ist allerding numerisch sehr aufwendig.

In der jüngsten Literatur der Signalverarbeitung wurde das hiermit eng verwandte Problem der "blind separation of sources" häufig behandelt. Man will dabei statistisch unabhängige Signale rekonstruieren, die zuvor linear überlagert wurden. In einer hervorragenden Arbeit formalisiert Comon (1994) die lineare unabhängige Komponentenanalyse ("Independent Component Analysis" kurz ICA), die er in (Comon, Jutten, und Herault, 1991) einführte. Während in der PCA lediglich versucht wird, Signale zu dekorrelieren, möchte man in der ICA statistisch unabhängige Signale gewinnen. Lediglich im Fall normalverteilter Signale sind beide Kriterien gleichbedeutend. Comon beschränkt sich auf lineare Transformationen, löst aber das Problem für nicht normalverteilte Signale, indem er die gemeinsame Information mit Kumulanten höherer Ordnung näherungsweise mißt. In der vorliegenden Arbeit werden Korrekturen zu diesen Entropiemessungen vorgeschlagen. Die bisher zu diesem Thema im nichtlinearen Fall veröffentlichten Arbeiten sind entweder von sehr eingeschränkter Gültigkeit oder mangelhafter theoretischer Grundlage (Karhunen und Joutsensalo, 1994; Bell und Sejnowski, 1995; Burel, 1992). Schließlich kann noch die Arbeit von Hastie und Stuetzle (1989) erwähnt werden, die einen statistischen Algorithmus zum Auffinden von Hauptkurven ("Principal Curves") vorstellt. Dieser Algorithmus liefert allerdings keine analytische Darstellung der gefundenen Kurven.

So werden in diesem Kapitel zwei Verallgemeinerungen der PCA vorgenommen. Statt Dekorrelation wird statistische Unabhängigkeit mittels minimaler gemeinsamer Information verwirklicht. Zur Messung der gemeinsamen Information wird dabei Statistik höherer Ordnung verwendet, um auch nicht normalverteilte Signale zuzulassen. Soweit entspricht das der linearen ICA. Darüber hinaus werden statt Rotationen nichtlineare, volumenerhaltende Transformationen betrachtet.

7.1 Konstante Informationsübertragung

Man sollte zunächst die Definitionen der Entropie, Transinformation und gemeinsamen Information im Kontinuierlichen überprüfen.

Die Entropie (2.2) einer diskreten Verteilung ist stets positiv, somit nach unten beschränkt. Die Entropie einer Dichte (2.9) allerdings kann auch negativ sein und hat keine untere Grenze. Im Gegensatz zum diskreten Fall ändert eine beliebige Skalierung einer Wahrscheinlichkeitsdichte die zugehörige Entropie. Etwas allgemeiner läßt sich das durch die nachfolgende Ungleichung formulieren. Betrachtet man eine beliebige Abbildung $\mathbf{y} = f(\mathbf{x})$ im $\Re^n$, so transformiert die Entropie $H[p(\mathbf{x})]$ des Eingaberaums zur Entropie des Ausgaberaums $H[p(\mathbf{y})]$ durch (siehe Papoulis, 1991)

$$H[p(\mathbf{y})] \leq H[p(\mathbf{x})] + \int_{\Re^n} p(\mathbf{x}) \ln \left| \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right| \tag{7.1}$$

Ist $f(\mathbf{x})$ umkehrbar, so gilt die Gleichheit. Der zweite Term stellt das Mittel über den Logarithmus der Jacobideterminante $|\partial \mathbf{y}/\partial \mathbf{x}|$ der Abbildung $f(\mathbf{x})$ dar. Ist die Jacobideterminante stets gleich Eins (d.h. die Abbildung ist volumenerhaltend), so verschwindet dieser Term. Die Entropien der Ein- und Ausgabedichten sind dann gleich. Andernfalls generiert oder vermindert die Abbildung Information, wenn auch nur durch einfache Skalierung der Koordinaten. Diese Änderung des Informationsgehaltes aufgrund einer Volumenzerrung folgt aus der Normierungsbedingung einer Wahrscheinlichkeitsdichte und ist nicht von Bedeutung. Man möchte Volumenzerrung daher vermeiden.

Im kontinuierlichen Fall ist es zudem notwendig, das Kriterium der maximalen Transinformation zu überdenken. Die Definition der Transinformation einer kontinuierlichen Wahrscheinlichkeitsdichte $p(\mathbf{x}, \mathbf{y})$ ist analog zu (2.4),[1]

$$\begin{aligned} I[p(\mathbf{x}, \mathbf{y})] &= H[p(\mathbf{y})] + H[p(\mathbf{x})] - H[p(\mathbf{x}, \mathbf{y})] && (7.2) \\ &= H[p(\mathbf{y})] - H[p(\mathbf{y}|\mathbf{x})] && (7.3) \end{aligned}$$

Während die Transinformation im diskreten Fall beidseitig beschränkt ist, gilt das im Kontinuierlichen nicht mehr. Für eine deterministische Abbildung ist sogar $H[p(\mathbf{y} = f(\mathbf{x})|\mathbf{x})] = -\infty$. Läßt man allerdings ein verschwindend kleines additives Rauschen in der Abbildung zu, so ist $H[p(\mathbf{y}|\mathbf{x})]$ nach unten beschränkt. Man kann bei festem Rauschen diesen Term als konstant betrachten und zur Maximierung der Transinformation auf die Maximierung der Entropie $H[p(\mathbf{y})]$ der Ausgabedichte konzentrieren. Im wesentlichen handelt es sich hier wieder um das bekannte "Infomax"-Prinzip. Dies wurde unter anderem von Nadal und Parga (1994), Atick (1992) oder Bell und Sejnowski (1995) vorgeschlagen. Die Schwierigkeit der Skalierungsabhängigkeit der Entropie wird dort durch Fixierung einer

[1] Die Definition (7.3) wird oft auch als Informationszuwachs einer Messung verstanden. Man interpretiert dabei die Zufallsvariable $\mathbf{y}$ als einen gemessenen, mit Rauschen behafteten Wert einer zugrundeliegenden tatsächlichen Größe $\mathbf{x}$. Der erste Term $H[p(\mathbf{y})]$ gibt die Entropie der gemessenen Variablen an, d.h. die Ungewißheit über den Ausgang der Messung. Der zweite Term $H[p(\mathbf{y}|\mathbf{x})]$ gibt die Ungewißheit über den zu messenden Wert $\mathbf{y}$ an, falls $\mathbf{x}$ vorgegeben ist. Dieser letzte Ausdruck gibt im wesentlichen die Ungewißheit an, die durch die Meßungenauigkeit eingeführt wird. Wäre man in der Lage, eine kontinuierliche Variable mit beliebiger Genauigkeit zu messen, so würde diese Messung unendlich viel Information enthalten. Dies gibt dem oben angesprochenen Grenzwert eine anschauliche Bedeutung. Es sei noch hinzugefügt, daß diese Definition nicht die Skalierungsabhängigkeit der Entropie (2.9) besitzt, da die Differenz der Entropien den Skalierungsterm eliminiert. Aus diesen Gründen wird (7.3) oft als die angemessene Definition der Information einer kontinuierlichen Zufallsvariablen betrachtet. Diese Definition beinhaltet allerdings den Meßprozeß, von dem man zumeist keine einfache analytische Beschreibung hat und der hier als unbekannt vorrausgesetzt wurde.

Skala überbrückt, indem die Abbildungen auf einen bestimmten Wertebereich beschränkt werden. In einem biologischen System mag die Annahme von additivem Rauschen und einer beschränkten Dynamik der Abbildung zutreffend sein. Hier werden diese Elemente aber lediglich zur Überbrückung analytischer Probleme in der Definition der Transinformation im Kontinuierlichen eingeführt.

Es wurden für diesen Abschnitt zwei Vorgaben gewählt: es soll eine deterministische Abbildung betrachten werden und es ist kein *a priori* Wissen über die Signal- oder Rauschdichte gegeben. In Anbetracht dieser Vorgaben wird hier ein neues Konzept vorgeschlagen. Man erachtet die Signale als virtuell rauschfrei. Die Skalierungsabhängigkeit der Entropie und die Unbeschränktheit der Transinformation einer deterministischen Abbildung erfordern, daß das Konzept der maximalen Transinformation mit dem Konzept der *konstanten Informationsübertragung* ersetzt wird. Die Information der Eingabesignale soll konstant auf die Ausgabe abgebildet werden:

$$H[p(\mathbf{x})] = H[p(\mathbf{y})] \tag{7.4}$$

Wie man in Ungleichung (7.1) erkennt, ist Bedingung (7.4) erfüllt, falls die Abbildung volumenerhaltend ist:

$$\left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| = 1 \tag{7.5}$$

Es sei angemerkt, daß die Rotation, die in der Hauptkomponentenanalyse verwendet wird, ein Spezielfall der volumenerhaltenden linearen Abbildungen ist (Obradovic und Deco, 1995). Man interessiert sich in dieser Arbeit für nichtlineare, adaptive Transformationen, die Bedingung (7.5) strikt erfüllen. In den nächsten drei Abschnitten werden einige Möglichkeiten hierfür vorgestellt. Zunächst werden die *symplektischen* Abbildungen - eine Klasse von volumenerhaltenden nichtlinearen Transformationen (Abraham und Marsden, 1978) - behandelt. Es werden implizite und explizite Definitionen von symplektischen Funktionen betrachtet. Indem man adaptive konnektionistische Strukturen in deren Definitionen einführt, erhält man Architekturen, die man als *rückgekoppelt* bzw. als *vorwärts vermittelnd* verstehen kann. Eine weitere Klasse von vorwärts vermittelnden Transformationen, diesmal mit dreiecksförmigen Jacobideterminanten, wird in Abschnitt 7.1.3 vorgeschlagen.

7.1.1 Implizite symplektische Abbildung

Die symplektischen Abbildungen wurden im Zusammenhang mit der Theorie multivariater Funktionen erstmalig von Siegel (1943) eingeführt. Sie basieren auf der symplektischen Geometrie in einem gerade-dimensionalen Raum. In der symplektischen Geometrie entsprechen gewisse Flächenelemente einer Länge im Euklidischen Raum. Transformationen, die diese Flächenelemente erhalten, bezeichnet

man als symplektisch. Symplektische Transformationen erhalten auch das Volumen. Die Umkehrung gilt allerdings nicht, d.h. Volumenerhaltung ist nicht hinreichend für Symplektizität.

Es sollen hier zunächst die impliziten symplektischen Abbildungen betrachtet werden. Eine interessante und für unsere Zwecke wichtige Tatsache ist, daß jede nicht reflektierende[2], symplektische Abbildung implizit durch eine skalare Funktion $S(\mathbf{z})$ ausgedrückt werden kann, und auch umgekehrt definiert jede skalare Funktion $S(\mathbf{z})$ eine symplektische Abbildung $\mathbf{y} = \mathbf{f}(\mathbf{x})$ im R^n:

$$\mathbf{y} = \mathbf{x} + J\frac{\partial}{\partial \mathbf{z}} S\left(\frac{\mathbf{x}+\mathbf{y}}{2}\right) \qquad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} = -J^{-1} \tag{7.6}$$

wobei I die Einheitsmatrix in $R^{n/2}$ darstellt. Der Gradient bezieht sich auf das Argument $\mathbf{z} = (\mathbf{x}+\mathbf{y})/2$ von $S(\mathbf{z})$. Man beachte, daß eine quadratische Funktion $S(\mathbf{z})$ in (7.6) zu einer linearen Transformation führt.

Für den Beweis der Volumenerhaltung der Abbildung (7.6) konsultiere man die mathematische Literatur zu den symplektischen Transformationen. [3]

Parametrisierung mit adaptiver Netzstruktur

Es wird für die skalare Funktion $S(\mathbf{z})$ ein dreischichtiges Multi-Layer-Perceptron (Rumelhart et al., 1988) mit einem linearen Ausgabeelement verwendet. Diese adaptive Struktur hat sich als allgemeiner Funktionsapproximierer bewährt (Hornik, Stinchcombe, und White, 1989). Man erhält dadurch eine allgemeine symplektische Transformation:

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + J\tfrac{\partial}{\partial \mathbf{z}} S\left(\tfrac{\mathbf{x}+\mathbf{y}}{2}, \mathbf{w}, W\right) \\ S(\mathbf{z}, \mathbf{w}, W) &= \mathbf{w} \cdot g(W\mathbf{z}) \end{aligned} \tag{7.7}$$

Die generierende Funktion $S(\mathbf{z}, \mathbf{w}, W)$ ist jetzt mit den Kopplungskoeffizienten $\mathbf{w} \in R^m$ und $W \in R^m \times R^n$ parametrisiert. Damit J in (7.6) wohl definiert ist, muß n eine gerade Zahl sein. Das ist keine entscheidende Einschränkung, da man jederzeit eine neutrale Dimension hinzufügen kann. Die Aktivierungsfunktion g wirkt auf jede Komponente des Vektors $W\mathbf{z}$ und kann wie üblich als eine sigmoide Funktion gewählt werden.

---

[2] Eine Transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ bezeichnet man als nicht reflektierend, falls $det\left(I - \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right) \neq 0$, d.h. wenn die Jacobideterminante von $\mathbf{f}(\mathbf{x})$ keine Eigenwerte -1 besitzt.

[3] Die Darstellung (7.6) der symplektischen Abbildung ist ein Spezialfall der Theorie der generierenden Funktionen, die in voller Allgemeinheit von Feng und Qin (1985) entwickelt wird. Ein Beweis der Darstellung (7.6) und eine Diskussion über ihre Rolle in der numerischen Integration von Hamiltonschen Systemen kann man in (Miesbach und Pesch, 1992) finden.

Numerische Lösung der impliziten Gleichungen

Nun will man für eine Eingabe $\mathbf{x}$ die zugehörige Ausgabe $\mathbf{y}$ aus (7.7) berechnen. Es gibt verschiedene Möglichkeiten die Gleichungen (7.7) zu lösen. Der schnellste Algorithmus wäre Newtons Nullstellensuche für die Funktion $F(\mathbf{y}) = -\mathbf{y} + \mathbf{x} + J\partial S((\mathbf{x}+\mathbf{y})/2)/\partial \mathbf{z}$. Eine andere naheliegende Methode wäre die Gleichungen (7.7) als Fixpunktiteration zu formulieren:

$$\mathbf{y}(t+1) = \mathbf{x} + J\frac{\partial}{\partial \mathbf{z}} S\left(\frac{\mathbf{x}+\mathbf{y}(t)}{2}, \mathbf{w}, W\right) \tag{7.8}$$

Schließlich könnte man auch Gradientenmethoden wie den konjugierten Gradienten zur Minimierung von $\|F(\mathbf{y})\|^2$ verwenden. All diese Verfahren sollten bei $\mathbf{y}(0) = \mathbf{x}$ starten, da dies zu den Lösungen führen wird, die der identischen Abbildung am "nächsten" sind. Die Metrik, die den Begriff "Nähe" definiert, wird durch (7.7) festgelegt und ist schwer zu analysieren. Die Iteration wird erfolgreich sein, falls der angegebene Startpunkt im Anziehungsgebiet des Suchverfahrens liegt. Im allgemeinen wird dieses Anziehungsgebiet kleiner sein für zunehmende Dimension n und wachsendes $\|\mathbf{w}\|$ sowie für wachsendes $\|W\|$, falls g eine monoton steigende Funktion ist.

In Simulationen erwiesen sich Fixpunktiteration und konjugierte Gradientenmethode als recht effizient, während Newtons Nullstellensuche Konvergenzprobleme aufweist. Die Fixpunktiteration konvergiert zumeist innerhalb 30 Iterationen und ist daher schnell. Die Konvergenz versagt im Laufe des Lernens, falls die Norm der Kopplungsstärken zunimmt. Man beachte die Bedingung für lokale Konvergenz $\|\partial \mathbf{y}(t+1)/\partial \mathbf{y}(t)\| \leq 1$. Die konjugierte Gradientenmethode konvergiert zumeist innerhalb von 10 Iterationen, aber nicht notwendigerweise zum gewünschten globalen Maximum. Ein stabileres und global konvergentes Verfahren ist das Homotopie-Verfahren (Stoer und Bulirsch, 1993). Man startet bei einer bekannte Lösung und modifiziert die nichtlinearen Gleichungen graduell, während man aufeinanderfolgend die entsprechend neuen Lösungen findet. Für gewöhnlich beginnt man bei der identischen Abbildung und führt die Nichtlinearität graduell mit einem Parameter α ein. In unserem Fall muß man die Lösungen von

$$\left\| -\mathbf{y} + \mathbf{x} + \alpha J\frac{\partial}{\partial \mathbf{z}} S\left(\frac{\mathbf{x}+\mathbf{y}(t)}{2}, \mathbf{w}, W\right)\right\|^2 = 0 \tag{7.9}$$

für gegebenes α suchen. Beim Homotopie-Verfahren wächst $\alpha = 0 \rightarrow 1$ graduell mit einer Schrittweite $\Delta\alpha$, während man die Lösungen der jüngsten Relaxation $\mathbf{y}(\infty, \alpha)$ als Startwerte $\mathbf{y}(0, \alpha + \Delta\alpha)$ der nächsten Minimierung verwendet:

$$\mathbf{y}(0, \alpha + \Delta\alpha) = \mathbf{y}(\infty, \alpha) \tag{7.10}$$

Dieses Verfahren kann u.U. die Konvergenz des gesamten Suchverfahrens beschleunigen, da für jeden neuen Parameterwert die Lösung von (7.9) in wenigen

Iterationen gefunden wird. Das allmähliche Anwachsen der Nichtlinearität hilft andererseits, im Anziehungsbereich der globalen Lösung zu verbleiben. Man gewinnt aber dadurch keine Garantie für die Konvergenz der Verfahrens. Das Verfahren kann auf einen Bifurkationspunkt stoßen und dem Pfad der degenerierenden Lösung folgen, d.h. die Lösung entwickelt sich zu einem relativen Minimum. Für den Lernprozeß hat man ein natürliches Verfahren, um im Anziehungsgebiet der globalen Lösung zu verbleiben. Man folgt den Lösungen von (7.9) für $\alpha = 1$, während man die Kopplungsparameter allmählich gemäß der Lernregel adaptiert. Die Rolle von α wird jetzt von $\mathbf{w}$ und W übernommen. Man beginnt die Suche der Lösung von (7.9) mit neuen Kopplungsstärken $(\mathbf{w} + \Delta\mathbf{w}, W + \Delta W)$ bei den vorhergehenden Lösungen $\mathbf{y}(\infty, \mathbf{w}, W)$:

$$\mathbf{y}(0, \mathbf{w} + \Delta\mathbf{w}, W + \Delta W) = \mathbf{y}(\infty, \mathbf{w}, W) \tag{7.11}$$

Dieses Homotopie-Verfahren wurde für das Lernen verwendet, da ein natürlicher Zugang zum Pfad der Kopplungsstärken vorliegt. Das erste Homotopie-Verfahren wird statt dessen für Datenpunkte, die nicht zum ursprünglichen Satz von Lerndaten gehören, benötigt.

Optimierung der adaptiven symplektischen Abbildung

Für das Optimieren der symplektischen Abbildung wird ein Gradientenverfahren verwendet. Das macht insbesondere Sinn, falls der Gradient der zu optimierenden Kostenfunktion bezüglich der Adaptionsparameter leicht zu berechnen ist. Entsprechend der Kostenfunktion (7.28), die in Abschnitt (7.2.1) noch vorgestellt wird, benötigt man die Ableitungen der Ausgabekoordinaten nach den Adaptionsparametern, die hier generisch mit p bezeichnet werden:

$$\Delta p = -\frac{\epsilon}{2}\frac{\partial}{\partial p}\left\langle \|\mathbf{y} - \langle\mathbf{y}\rangle\|^2 \right\rangle = -\epsilon\left(\left\langle \mathbf{y}\cdot\frac{\partial\mathbf{y}}{\partial p}\right\rangle - \langle\mathbf{y}\rangle\cdot\left\langle\frac{\partial\mathbf{y}}{\partial p}\right\rangle\right) \tag{7.12}$$

Der Parameter p repräsentiert die Komponenten von $\mathbf{w}$ bzw. W. Wie üblich wurde eine Schrittweite ϵ eingeführt. Leitet man die implizite Gleichung (7.7) nach den Parametern ab, so erhält man

$$\frac{\partial\mathbf{y}}{\partial p} = -J^{-1}\left(\left.\frac{\partial^2 S}{\partial\mathbf{z}\partial\mathbf{z}}\right|_{\frac{\mathbf{x}+\mathbf{y}}{2}}\frac{\partial\mathbf{y}}{\partial p} + \left.\frac{\partial^2 S}{\partial\mathbf{z}\partial p}\right|_{\frac{\mathbf{x}+\mathbf{y}}{2}}\right) \tag{7.13}$$

Hier ist $\partial^2 S/\partial\mathbf{z}^2$ die Hessematrix von $S(\mathbf{z})$ an der Stelle $\mathbf{z} = (\mathbf{x}+\mathbf{y})/2$. Dieses lineare Gleichungssystem liefert die benötigte Ableitung bei gegebenen $\mathbf{x}$ und $\mathbf{y}$. Die expliziten Ableitungen der generierenden Funktion werden hier ausgelassen, da sie keine besondere Einsicht in das Verfahren geben und daher die Darstellung nur unnötig erschweren.

Hier wurde der Gradient für die Kostenfunktion (7.28) angegeben. Tatsächlich sind das auch die Gleichungen, die in den Experimenten zu Abschnitt 7.2.1 verwendet werden, sowie auch bei den Anwendungen der entsprechenden Kostenfunktion in den Abschnitten 7.3, 7.3.2 und 7.3.3. Prinzipiell wird sich jede Kostenfunktion, die als Funktion der Ausgabekoordinaten analytisch gegeben ist, auf diese Weise ableiten lassen.

Die numerische Komplexität zur Lösung von (7.7) ist der Ordnung $O(nmt)$, wobei t die Anzahl Iterationen benennt, die man bis zur Konvergenz benötigt. In der Praxis sind das $10 < t < 200$. Die numerische Komplexität zur Lösung von (7.13) für alle $n^2 + m$ Parameter und das Produkt in (7.12) ist der Ordnung $O((n^2 + m)(1 + n + n^2/2) + n^3)$. Man könnte an Näherungen von (7.12) denken, mit dem Ziel die Parameter nach jedem der N Datenpunkte zu adaptieren. Es wird hier der Einfachheit halber der exakte Gradientenabstieg durchgeführt, was einen zusätzlichen Faktor von N impliziert.

7.1.2 Explizite symplektische Abbildung

Wie man sehen konnte, führt die *implizite* Definition der symplektischen Funktion zu einem aufwendigen und numerisch kostspieligen Verfahren. Für die Berechnung eines jeden Funktionswertes ist die Relaxation eines nichtlinearen Gleichungssystems notwendig. Es kann als rückgekoppeltes Relaxationsnetzwerk verstanden werden. In der Praxis ist man bei diesem Ansatz auf Abbildungen von nicht mehr als 10 Dimensionen beschränkt.

Hier soll nun eine *explizite* Definition einer symplektischen Funktion vorgestellt werden, die uns zu einer volumenerhaltenden, vorwärts vermittelnden und somit effizienten Abbildung führen wird. Zu diesem Zwecke denke man sich den Eingabe- und Ausgaberaum in zwei Teilräume getrennt, d.h. $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ und $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$, wobei $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2 \in \Re^{n/2}$.

$$\mathbf{y}_1 = \mathbf{x}_1 - \frac{\partial P(\mathbf{x}_2)}{\partial \mathbf{x}_2} \quad , \quad \mathbf{y}_2 = \mathbf{x}_2 + \frac{\partial Q(\mathbf{y}_1)}{\partial \mathbf{y}_1} \, . \tag{7.14}$$

Die funktionale Abhängigkeit der Ausgabe von der Eingabe wird in Abbildung 7.1 verdeutlicht. Es können zwei skalare Funktionen $P : \Re^{n/2} \mapsto \Re$ und $Q : \Re^{n/2} \mapsto \Re$ frei gewählt werden. Man beachte wieder, daß für quadratische generierende Funktionen Gleichung 7.14 eine lineare Transformation ergibt. Man verwendet für diese skalaren generierenden Funktionen wieder ein dreischichtiges MLP mit nichtlinearen inneren Elementen und jeweils einer linearen Ausgabe:

$$p(\mathbf{x}_2) = \mathbf{w}_2 \cdot g(W_2 \mathbf{x}_2) \quad , \quad Q(\mathbf{y}_1) = \mathbf{w}_1 \cdot g(W_1 \mathbf{y}_1) \tag{7.15}$$

Die skalaren Funktionen P und Q sind nun mit den Kopplungsstärken $\mathbf{w}_1, \mathbf{w}_2 \in R^m$ und $W_1, W_2 \in R^m \times R^{n/2}$ parametrisiert. Die nichtlineare Aktivierungsfunktion g der inneren Elemente wirkt auf jede Komponente des Vektors

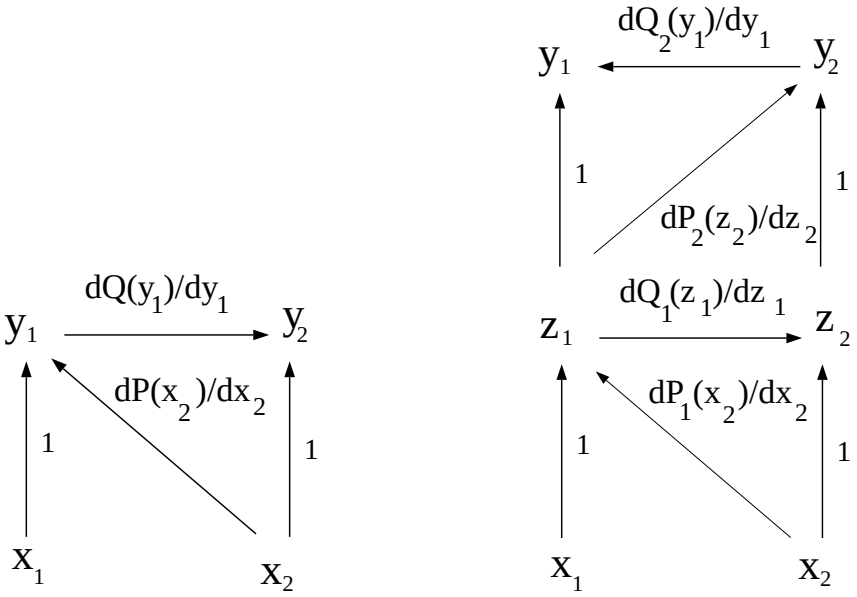

Abbildung 7.1: Links: Struktur der expliziten symplektischen Abbildung. Die Größen $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2$ sind hier vektorwertig. Rechts: Zwei aufeinanderfolgende explizite symplektische Abbildungen. Die gesamte Transformation ist weiterhin volumenerhaltend. P und Q können beliebige skalare Funktionen sein. Hier werden dafür dreischichtige MLP mit nichtlinearen inneren Elementen und einer linearen (skalaren) Ausgabe verwendet.

$W_1\mathbf{y}_1$ und $W_2\mathbf{x}_2$. Sie kann wie üblich als sigmoide Funktion gewählt werden. Aufgrund der Struktur der Gleichungen hängen die Ausgabekoordinaten $\mathbf{y}_1$ nur additiv von den Eingabekoordinaten $\mathbf{x}_1$ ab. Um eine allgemeinere Abbildung zu erhalten, wird eine weitere symplektische Transformation hinzugefügt.

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{x}_1 - \frac{\partial P_1(\mathbf{x}_2)}{\partial \mathbf{x}_2} \quad , \quad \mathbf{z}_2 = \mathbf{x}_2 + \frac{\partial Q_1(\mathbf{z}_1)}{\partial \mathbf{z}_1} \quad , \\ \mathbf{y}_1 &= \mathbf{z}_1 - \frac{\partial P_2(\mathbf{y}_2)}{\partial \mathbf{y}_2} \quad , \quad \mathbf{y}_2 = \mathbf{z}_2 + \frac{\partial Q_2(\mathbf{z}_1)}{\partial \mathbf{z}_1} \quad . \end{aligned} \tag{7.16}$$

Jetzt hat man für jede Schicht je zwei skalare Funktionen. Das sind P_1, Q_1 und entsprechend P_2, Q_2, die analog zu Gleichung 7.16 definiert werden (siehe Abbildung 7.1). Die numerische Komplexität zur Auswertung dieser Abbildung ist nun $O(nm)$. Das ist wesentlich weniger als die rückgekoppelte Struktur, die im vorhergehenden Abschnitt vorgeschlagen wurde (siehe Seite 90). Der numerische Aufwand zur Auswertung der Funktion liegt nicht wesentlich über dem einer linearen Abbildung. Für gewöhnlich ist das $O(n^2)$.

Diese Abbildung wird in den Experimenten zum Abschnitt 7.2.2 eingesetzt. Dort werden stochastische Optimierungsverfahren verwendet. Daher werden in dieser Arbeit keine Gradienten der expliziten Abbildung benötigt. Die Berechnung von Gradientenausdrücken ist jedoch unproblematisch.

7.1.3 Dreiecksförmige Jacobimatrizen

In dieser Arbeit wurden adaptive, volumenerhaltende und somit informationserhaltende Transformationen nur mit den symplektischen Transformationen implementiert. Es soll trotzdem eine weitere interessante volumenerhaltende Transformation vorgestellt werden.

Der Gradient in den symplektischen Abbildungen verkompliziert die Struktur der Transformation. Hier wird eine Abbildung vorgeschlagen, die den Gradienten vermeidet, indem eine bestimmte eingeschränkte Kopplungsstruktur verwendet wird.

Die Struktur ist in Abbildung 7.2 dargestellt. Es sei $\mathbf{g}(\mathbf{y})$ eine Funktion im $\Re^n$ mit einer Jacobimatrix $L := \partial\mathbf{g}/\partial\mathbf{y}$ linker, unterer Dreiecksgestalt, d.h. $L_{ij} = 0$ für $i \leq j$. Entsprechend sei $\mathbf{f}(\mathbf{x})$ eine Funktion im $\Re^n$ mit $R := \partial\mathbf{f}/\partial\mathbf{x}$ rechter, oberer Dreiecksgestalt, d.h. $R_{ij} = 0$ für $i \ \geq j$. Folgende Transformation $\mathbf{x} \mapsto \mathbf{y}$ ist explizit und volumenerhaltend:

$$\mathbf{g}(\mathbf{y}) + \mathbf{y} = \mathbf{x} + \mathbf{f}(\mathbf{x}) \tag{7.17}$$

Beispielsweise für $n = 3$ schreibt sich (7.17) genauer

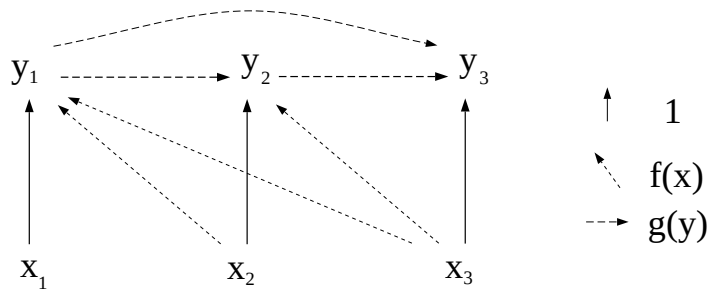

Abbildung 7.2: Struktur einer volumenerhaltenden expliziten Transformation mit Dreiecksgestalt in den Jacobimatrizen $\partial\mathbf{g}/\partial\mathbf{y}$ und $\partial\mathbf{f}/\partial\mathbf{x}$, hier für $n = 3$ dargestellt.

$$\begin{array}{lcccccc} y_1 & = & & & x_1 & + & f_1(x_2, x_3) \\ y_2 & = & g_2(y_1) & + & x_1 & + & f_2(x_3) \\ y_3 & = & g_3(y_1, y_2) & + & x_1 & + & \end{array} \tag{7.18}$$

Diese Gleichungen definieren eine explizite Abbildung, da man, beginnend mit der ersten Koordinaten y_1 durch aufeinanderfolgendes Einsetzen von $y_1, .., y_i$ in die Gleichung von y_{i+1}, eine explizite Abhängigkeit der Ausgabekoordinaten $\mathbf{y}$ von den Eingabekoordinaten $\mathbf{x}$ erhält.

Die Volumenerhaltung erkennt man leicht, indem man die Jacobideterminante von (7.17) wie folgt berechnet:

$$\frac{\partial\mathbf{g}}{\partial\mathbf{x}} + \frac{\partial\mathbf{y}}{\partial\mathbf{x}} = I + \frac{\partial\mathbf{f}}{\partial\mathbf{x}} \tag{7.19}$$

$$\left(\frac{\partial\mathbf{g}}{\partial\mathbf{y}} + I\right)\frac{\partial\mathbf{y}}{\partial\mathbf{x}} = I + \frac{\partial\mathbf{f}}{\partial\mathbf{x}} \tag{7.20}$$

$$|L + I|\,|\partial\mathbf{y}/\partial\mathbf{x}| = |I + R| \tag{7.21}$$

$$|\partial\mathbf{y}/\partial\mathbf{x}| = 1 \tag{7.22}$$

Zunächst wurde die Definitionsgleichung (7.17) nach $\mathbf{x}$ differenziert. In der vierten Zeile wurde zu beiden Seiten die Determinante angewendet. Die Matrix $|I+R|$ und $|L+I|$ haben Dreiecksgestalt mit Diagonalelementen gleich Eins. Für die Determinanten gilt: $|I + R| = |L + I| = 1$.

Die generierenden Funktionen sind hier nicht mehr auf skalare Funktionen beschränkt. Man vermeidet auch die Komplikationen der Struktur, die durch einen Gradienten hervorgerufen werden. Auch die willkürlich erscheinende Trennung des Raums, die für die explizit symplektische Abbildung notwendig war, wird hier vermieden. Die generierenden Funktionen $\mathbf{g}$ und $\mathbf{f}$ lassen sich vollkommen

frei wählen. Es können dadurch wieder adaptive, konnektionistische Strukturen in die Abbildung eingeführt werden.

Diese Transformation wurde, unabhängig hiervon, etwas früher von Deco und Brauer (1994), Deco und Schürman (1994) auf einer andere Weise formuliert. Für die generierende Funktion $\mathbf{g}(\mathbf{x})$ werden dort Polynome bis zur vierten Ordnung verwendet. Die Struktur, die sich dabei ergibt, ist für die Modellierung von zeitlichen Abhängigkeiten interessant, falls man den Index der Koordinaten als diskreten Zeitschritt interpretiert. Dabei wird $L = 0$ gesetzt. Die Dreiecksgestalt von R bringt zum Ausdruck, daß die Ausgabe nur von den "früheren" Eingabekoordinaten abhängig ist.

7.2 Minimale gemeinsame Information im Kontinuierlichen

Nachdem die verschiedenen volumenerhaltenden Abbildungen vorgestellt wurden, kann man sich jetzt auf das unüberwachte Lernkriterium konzentrieren.

Es ist wichtig zu bemerken, daß für die korrekte Definition von Volumenerhaltung,im Eingabe- und Ausgaberaum die gleiche Dimension gewählt werden muß. Das heißt, man kann nicht ohne weiteres eine Dimensionsreduktion erzielen. Statt dessen wird man eine neue Darstellung der Eingabedichte an der Ausgabe erzeugen. Das Kriterium für diese neue Darstellung soll wie im ersten Teil der Arbeit, die Extraktion von statistisch unabhängigen Merkmalen sein, also die Minimierung der gemeinsamen Information in den Ausgabekoordinaten.

Man erinnere sich an die Ausführungen zur statistischen Unabhängigkeit für diskrete Zufallsvariablen in Abschnitt 2.4. Für kontinuierliche Zufallsvariablen $\mathbf{y}$ in R^n gelten die gleichen Aussagen. Von statistischer Unabhängigkeit spricht man, falls die gemeinsamen Wahrscheinlichkeitsdichten $p(\mathbf{y})$ faktorisieren,

$$p(\mathbf{y}) = \prod_i^n p(y_i) \tag{7.23}$$

Die Definition der gemeinsamen Information basiert auf dieser Definition von Unabhängigkeit. Wenn zwei Zufallsvariablen unabhängig sind, so erwartet man, daß sie keine gemeinsame Information tragen. Ein Maß, das diesem genügt, ist die relative Entropie zwischen der Wahrscheinlichkeitsdichte zur rechten und zur linken Seite von (7.23). Ist die Gleichung nicht gültig, so ergibt die relative Entropie einen positiven Wert. Diesen definiert man als die gemeinsame Information $MI[p(\mathbf{y})]$ der mehrdimensionalen Zufallsvariablen $\mathbf{y}$

$$0 \leq MI[p(\mathbf{y})] = -H[p(\mathbf{y})] + \sum_i^n H[p(y_i)] \tag{7.24}$$

Man beachte, daß die Bedingung (7.23) erfüllt ist, falls $MI[p(\mathbf{y})] = 0$ (Atick und Redlich, 1992).

Wie verhält sich das nun zur linearen Dekorrelation, die in der Hauptkomponentenanalyse verwendet wird? Im Fall von normalverteilten Eingabesignalen erweist sich das Diagonalisieren der Korrelationsmatrix der Ausgabe als äquivalent zur Minimierung der gemeinsamen Information (Papoulis, 1991). Für allgemeine Dichten bedingt Dekorrelation keine statistische Unabhängigkeit. Dies wird insbesondere in dem Beispiel von Abschnitt 7.2.2 deutlich, bei dem Optimierungskriterien mit Statistik zweiter Ordnug mit Kriterien höherer Ordnung verglichen werden.

In dieser Arbeit wird also das allgemeinere Kriterium der minimalen gemeinsamen Information verwendet, indem man das mit dem Prinzip der konstanten Informationsübertragung kombiniert. Das ist also eine neue Version der Merkmalsextraktion, so wie sie von Barlow (1959, 1989) vorgeschlagen wurde.

Um eine faktorielle Darstellung der Signaldichte an der Ausgabe zu finden benutzt man hier die gemeinsame Information (7.24) als Kostenfunktion. Da für volumenerhaltende Abbildungen $H[p(\mathbf{x})] = H[p(\mathbf{y})]$ gilt, verbleibt lediglich die Aufgabe der Minimierung der Entropien $H[p(y_i)]$ der einzelnen Koordinaten. Auf diese Weise vermeidet man die kostenintensive Berechnung von Mehrkoordinatenstatistik. Andererseits benötigt man einen analytischen Ausdruck oder zumindest eine Messung der Entropien der einzelnen Koordinaten. Da man nur einen Satz von Datenpunkten kennt, die gemäß der Wahrscheinlichkeitsdichte gezogen wurden, bleibt das weiterhin eine schwierige Aufgabe. Hilfreich ist die Tatsache, daß man eine einfache obere Grenze für diese Entropien kennt.

7.2.1 Varianz als obere Grenze

Die explizite Messung der Entropie läßt sich mit Hilfe des zweiten Theorems von Gibbs umgehen. Dieses Theorem gibt eine obere Grenze der Entropie als Funktion der Varianz an. Die Minimierung der Entropien der Einzelkoordinaten $H[p(y_i)]$ vereinfacht sich zur Minimierung der Varianzen der Ausgabekoordinaten σ_i. Bevor man zum Gibbs-Theorem übergeht, wird hier eine verwandte Ungleichung abgeleitet.

$$
\begin{aligned}
MI[p(\mathbf{y})] &= -H[p(\mathbf{x})] - \sum_i^n \int_R p(y_i) \ln p(y_i)\, dy_i && (7.25)\\
&\leq -H[p(\mathbf{x})] - \sum_i^n \int_R p(y_i) \ln g(y_i)\, dy_i && (7.26)\\
&= -H[p(\mathbf{x})] + \frac{n}{2}\ln(2\pi) + \frac{1}{2}\sum_i^n \int_R p(y_i)(y_i - \langle y_i\rangle)^2\, dy_i && (7.27)\\
&= -H[p(\mathbf{x})] + \frac{n}{2}\ln(2\pi) + \frac{1}{2}\left\langle \|y_i - \langle y_i\rangle\|^2 \right\rangle && (7.28)
\end{aligned}
$$

wobei das Mittel gegeben ist durch $\langle y_i \rangle = \int_R p(y_i)\, y_i\, dy_i$. Die Ungleichung in (7.26) verwendet die Tatsache, daß die relative Entropie zweier Dichten stets positiv oder Null ist. In diesem Fall sind es $p(y_i)$ und $g(y_i)$, wobei man eine Normalverteilung $g(y) = 1/\sqrt{\pi}\exp(-(y - \langle y \rangle)^2)$ gewählt hat. Die zwei ersten Terme in (7.27) sind konstant. D.h. für konstante Informationsübertragung kann man das Problem der Faktorisierung auf die Minimierung der Varianzen der Ausgabekoordinaten reduzieren. Man könnte auch Normalverteilungen $g(y_i)$ mit verschiedenen Varianzen für jede Koordinate verwenden, um genauer zu sein, die Varianzen der Dichten $p(y_i)$. Das ist die Idee des Gibbs-Theorems, welches die Entropie der Normalverteilung mit der Varianz von $p(y_i)$ als obere Grenze für die Entropie von $p(y_i)$ angibt. In diesem Fall ist die obere Grenze die Summe der Logarithmen der Varianzen:

$$MI[p(\mathbf{y})] \leq -H[p(\mathbf{x})] + \frac{n}{2}\ln(2\pi e) + \frac{1}{2}\sum_i^n \ln\left(\int_R p(y_i)(y_i - \langle y_i \rangle)^2\, dy_i\right) \qquad (7.29)$$

Diese zwei verschiedenen Kostenfunktionen ergeben zwei leicht unterschiedliche Lernziele. Die direkte Summe der Varianzen in (7.27) begünstigt gleiche Varianzen in den verschiedenen Ausgabekoordinaten. Die Summe der Logarithmen der Varianzen in (7.29) hingegen bevorzugt keine spezielle Skalierung zwischen den Varianzen. In allen Experimenten, die auf der Minimierung der Varianzen basieren, hat man die Kostenfunktion (7.27) verwendet.

Es erscheint als eine starke Einschränkung, nur Momente zweiter Ordnung für die Minimierung der Entropie zu verwenden. Statt dessen möchte man auch Momente höherer Ordnung miteinbeziehen, um so eine genauere Schätzung der Dichte und somit der entsprechenden Entropie zu erhalten. Dieser Ansatz wird in Abschnitt 7.2.2 behandelt. Aber zunächst beschränkt man sich auf den numerisch effizienteren Ansatz.

Die Kostenfunktion (7.27) hat zudem eine nützliche Eigenschaft: Ist die volumenerhaltende Transformation $\mathbf{y} = \mathbf{f}(\mathbf{x})$ flexibel genug, so erzeugt diese Kostenfunktion an der Ausgabe eine Normalverteilung. Das läßt sich mit dem folgenden Variationsansatz beweisen. Angenommen, die verwendete Transformation ist in der Lage, für eine gegebene Eingabeverteilung beliebige Ausgabeverteilungen $p(\mathbf{y})$ zu generieren. In diesem Fall läßt sich die Kostenfunktion (7.27) als ein zu minimierendes Funktional von $p(\mathbf{y})$ betrachten. Man führt die Nebenbedingungen der Norm $\int_{\Re^n} p(\mathbf{y})\, d\mathbf{y} = 1$ und der konstanten Information $\int_{\Re^n} p(\mathbf{y})\, \ln p(\mathbf{y}) d\mathbf{y} = const$ mit Lagrange-Parametern λ_1 und λ_2 ein. Das Funktional - nun $J[p(\mathbf{y})]$ genannt - ergibt sich zu:

$$J[p(\mathbf{y})] = \int_{\Re^n} \left(p(\mathbf{y})\, \|\mathbf{y}\|^2 + \lambda_1 p(\mathbf{y}) + \lambda_2 p(\mathbf{y}) \ln p(\mathbf{y})\right) d\mathbf{y} \qquad (7.30)$$

Die zugehörige Euler-Lagrange Gleichung ist dann

$$\|\mathbf{y}\|^2 + \lambda_1 + \lambda_2(\ln p(\mathbf{y}) + 1) = 0 \tag{7.31}$$

was zusammen mit der Normalisierungsbedingung eine symmetrische Normalverteilung ergibt:

$$p(\mathbf{y}) = \frac{1}{\sqrt{\pi\,\lambda_2}}\, e^{-\|\mathbf{y}\|^2/\lambda_2} \tag{7.32}$$

Die Normalverteilung ist daher optimal, und das Lernen wird versuchen, die Eingabedichte in eine Normalverteilung zu verwandeln. Das wird für die Dichteschätzung, die in Abschnitt 7.3 behandelt wird, nützlich sein.

Dieses Ergebnis soll an einem Experiment demonstriert werden. Als volumenerhaltende Abbildung wurde eine zweidimensionale, implizite, symplektische Transformation verwendet. Die adaptive Netzwerkstruktur hat hier zwei innere Knoten ($n = 2$, $m = 2$). Die Abbildung hat insgesamt sechs adaptive Kopplungsparameter $\mathbf{w} \in \Re^2$ und $W \in \Re^2 \times \Re^2$. Man optimiert die Abbildung mit dem Gradientenverfahren, das in Abschnitt 7.1.1 beschrieben wurde. Abbildung 7.3 zeigt das Resultat.

7.2.2 Näherungen der Entropie mit Kumulanten

Für die Gültigkeit dieses Ansatzes ist es allerding entscheidend, daß die Abbildung $\mathbf{y} = f(\mathbf{x})$ in der Lage ist, die gegebene Eingabedichte $p(\mathbf{x})$ in eine Normalverteilung an der Ausgabe $\mathbf{y}$ zu transformieren. Ein einfaches - aber wichtiges - Gegenbeispiel ist die Gleichverteilung. Um eine Gleichverteilung in eine Normalverteilung zu transformieren, muß die Abbildung an den Grenzen der Gleichverteilung unbeschränkt sein. Das widerspricht allerdings der Volumenerhaltung. Etwas allgemeiner ausgedrückt, wird eine volumenerhaltende Abbildung nicht in der Lage sein, eine unstetige oder disjunkte Dichte in eine Normalverteilung zu transformieren. Das Versagen der Varianz als Maß zur Bestimmung statistischer Unabhängigkeit erkennt man sehr leicht im Fall einer rotierten zweidimensionalen Gleichverteilung (siehe 7.5). Die korrekte Transformation, die eine statistische unabhängige Darstellung dieser Dichte liefert, ist schlicht die Rotation, die das Quadrat wieder in die achsenparallele Richtung bringt. Aber die Varianzen, ja sogar die gesamte Kovarianzmatrix, sind unabhängig von der Orientierung der quadratischen Verteilung. Eine Kostenfunktion, die nur Statistik zweiter Ordnung beinhaltet, wird keine faktorielle Darstellung finden. An dieser Stelle wird auch der Unterschied zwischen Hauptkomponentenanalyse und der Analyse unabhängiger Achsen (Comon et al., 1991) klar. Nur für normalverteilte Ausgabesignale sind diese beiden Ansätze gleich.

Die Minimierung der Entropie mittels der Varianz erfaßt also keine Statistik höherer Ordnung der Ausgabedichte. Das bedeutet keineswegs, daß die Eingabedichte auf Statistik zweiter Ordnung beschränkt bleiben muß. Um aber darüber

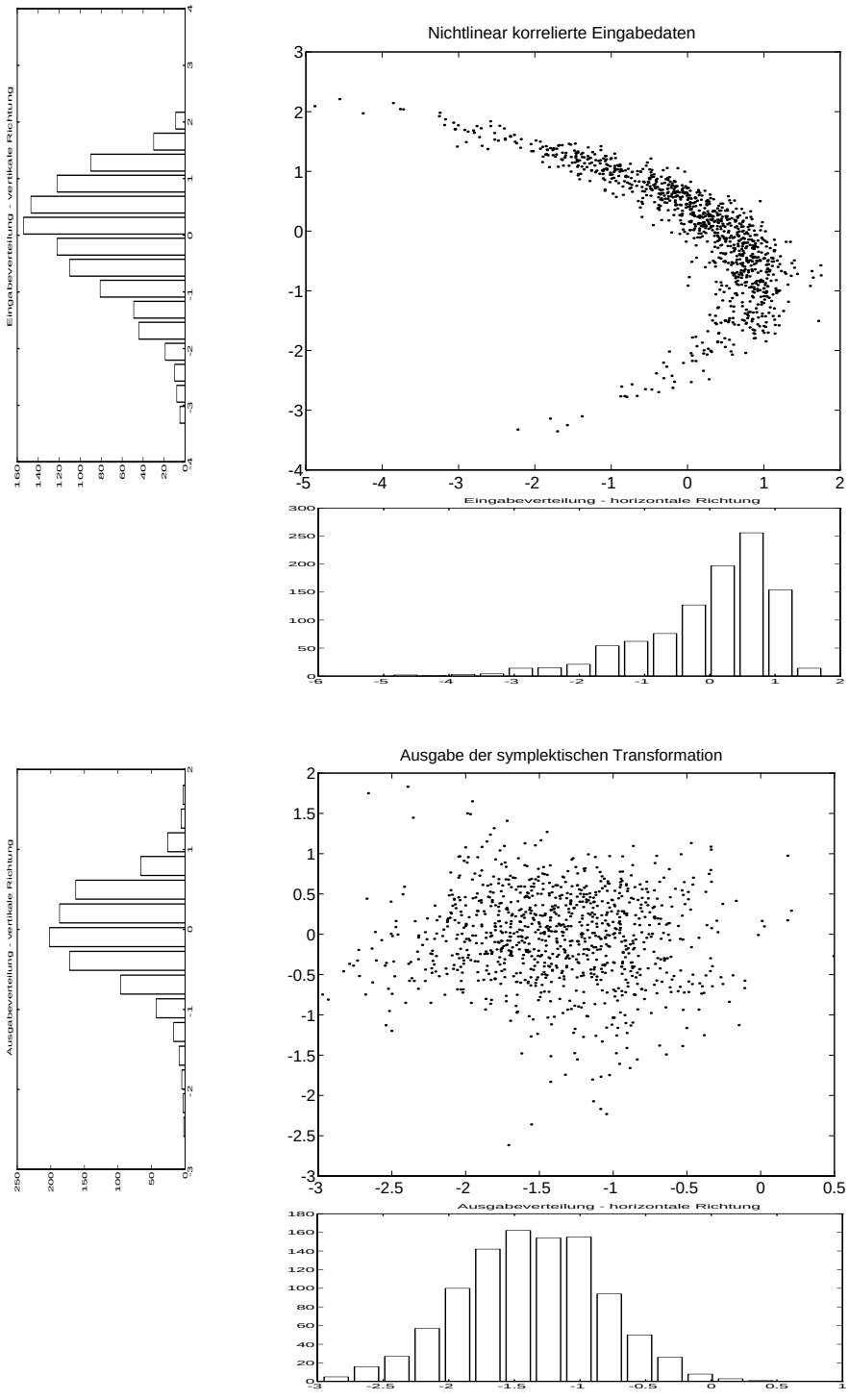

Abbildung 7.3: Oben: Nichtlinear korrelierte Eingabedichte. Unten: Nach dem Lernen wurde die Ausgabedichte zufriedenstellend in eine statistisch unabhängige Normalverteilung umgewandelt. Der Wert der Kostenfunktion wurde nach 300 Lernzyklen um 68% reduziert.

hinaus Statistik höherer Ordnung an der Ausgabe zu betrachten, wird hier die Edgeworth-Entwicklung zur Schätzung einer Wahrscheinlichkeitsdichte verwendet. Diese wurde im Kontext von statistischer Unabhängigkeit durch lineare Transformationen für den Fall von nicht normalverteilten Signalen von Comon (1994) vorgeschlagen.

Messung der Entropie mit Kumulanten höherer Ordnung

Die Entropie kann näherungsweise bestimmt werden, indem man meßbare Momente höherer Ordnung miteinbezieht. Comon (1994) schlägt vor, die Edgeworth-Entwicklung zu verwenden, da diese zu einem analytischen Ausdruck der Entropie als Funktion von meßbaren Kumulanten führt.

Zunächst soll die Definition von Momenten und Kumulanten angegeben werden. Eine Dichtefunktion ist im Prinzip durch ihre Momente m_α oder Kumulanten c_α festgelegt. Der Mittelwert m und die Varianz σ^2 entsprechen zum Beispiel der ersten und zweiten Kumulanten ($m = c_1, \sigma^2 = m_2$). Für eine Dichtefunktion mit Mittelwert $m = 0$ sind die Momente gegeben durch:

$$m_\alpha = \int_{-\infty}^{\infty} y^\alpha \, dy \tag{7.33}$$

Die Momente einer experimentell beobachteten Verteilung können durch Mittelung über gemessene Stichproben $\{y_1, .., y_N\}$ bestimmt werden:

$$m_\alpha \approx \frac{1}{N} \sum_{i=1}^{N} y_i^\alpha \tag{7.34}$$

Daher geben Momente einen einfachen Zugang zur Messung der Eigenschaften einer beobachteten Verteilung. Die Kumulanten können als Funktion der Momente der Verteilung ausgedrückt werden. Man beschränkt sich hier auf die ersten fünf Kumulanten. Wieder für eine Dichtefunktion mit Mittelwert $m = 0$ schreibt man:

$$c_2 = m_2 \quad , \quad c_3 = m_3 \quad , \quad c_4 = m_4 - 3m_2^2 \quad , \quad c_5 = m_5 - 10 m_3 m_2 \tag{7.35}$$

Die Entwicklung von Edgeworth macht deutlich, daß sich eine Dichtefunktion als Funktion ihrer Kumulanten formulieren läßt und somit durch diese festgelegt wird. Edgeworth entwickelt den multiplikativen Korrektionsterm zur besten Gauß'schen Näherung in der orthogonalen Basis von Hermite-Polinomen $h_\alpha(y)$. Die Entwicklungskoeffizienten sind im wesentlichen die aus den Datenpunkten ermittelten Kumulanten c_α der Ordnung α, (siehe Kendall und Stuart, 1969). Die Entwicklung für eine Dichte mit Mittel $m = 0$ und Varianz σ^2 schreibt sich,

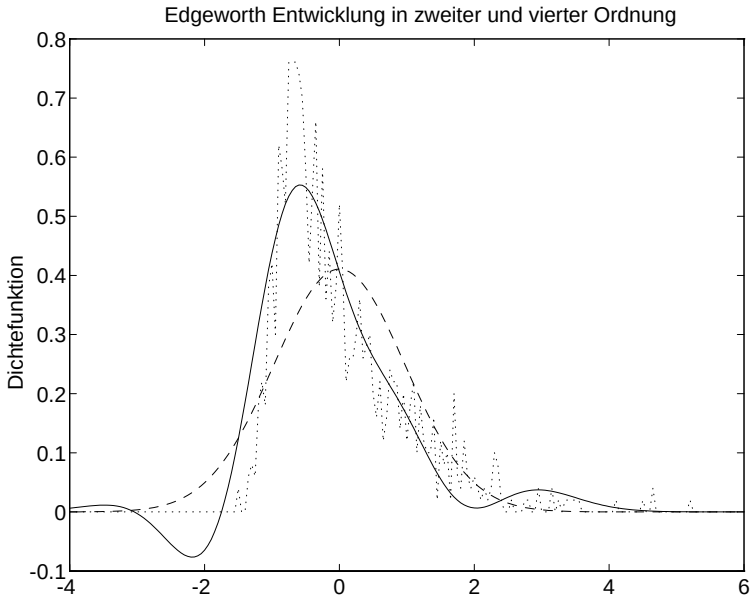

Abbildung 7.4: Punktierte Linie: Simulierte Exponentialverteilung mit additivem Gauß'schen Rauschen bei einer Stichprobengröße von 1000. (Rauschvarianz/Abfallskonstante = 0.2). Gestrichelte Linie: Gauß'sche Näherung, äquivalent zu Edgeworth-Näherung in zweiter Ordnung. Durchgezogene Linie: Edgeworth-Näherung mit Termen bis zur vierten Ordnung.

$$\begin{aligned} p(y) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} f(y) \\ f(y) &= 1 + \frac{c_3}{6\sigma^3} h_3(\frac{y}{\sigma}) + \frac{c_4}{24\sigma^4} h_4(\frac{y}{\sigma}) + \frac{c_5}{120\sigma^5} h_5(\frac{y}{\sigma}) + ... \end{aligned} \tag{7.36}$$

Bricht man diese Entwicklung bei einer bestimmten Ordnung ab, so erhält man eine Näherung $p_{est}(y)$, die auch negative Werte annehmen kann. Abbildung 7.4 zeigt einen Stichprobensatz einer Exponentialverteilung mit additivem Gauß'schem Rauschen. Es wird die Edgeworth-Näherung in zweiter und vierter Ordnung gezeigt. Dieses Beispiel ist für diese Entwicklung insbesondere problematisch, da bei einer Exponentialverteilung Momente höherer Ordnung theoretisch ungebunden sind.

Für manche Verteilungen nehmen die Terme mit steigender Ordnung ab.[4]

[4]Das gilt für Gram-Charlier-Dichten vom Typ "A". Das sind Dichten, die vollkommen durch ihre Momente determiniert sind. Eine Dichte mit exponentiellem Abfall genügt dieser Bedingung nicht . Abgesehen davon, sind gemessene Momente einer Ordnung höher als vier oft unzuverlässig auf Grund von Fluktuationen der Stichproben speziell für kleine N.

Mit dieser Begründung bricht Comon (1994) die Entwicklung 7.36 bei der vierten Ordnung ab. Wenn man weiterhin den Logarithmus der Entropie in einer Potenzreihe entwickelt und einige Integrationsregeln für Hermitepolynome verwendet, gewinnt man eine Näherung für die Entropie, die nur noch von den meßbaren Kumulanten der Dichte abhängt (Comon, 1994):

$$H[p(y)_{est}] \approx \frac{1}{2}\ln(2\pi e) + \ln\sigma - \frac{1}{12}\frac{c_3^2}{\sigma^6} - \frac{1}{48}\frac{c_4^2}{\sigma^8} - \frac{7}{48}\frac{c_3^4}{\sigma^{12}} + \frac{1}{8}\frac{c_3^2}{\sigma^6}\frac{c_4}{\sigma^4} \tag{7.37}$$

Dieser Ausdruck kann direkt für die Minimierung der Entropie in der Definition der gemeinsamen Information (7.24) verwendet werden. Er ermöglicht die Verwendung von Gradientenmethoden für die Optimierung der volumenerhaltenden Strukturen. Allerdings muß dazu die Entwicklung (7.36) bei der vierte Ordnung abgebrochen werden. Man ist somit auf Statistik vierter Ordnung beschränkt, die u.U. für eine zufriedenstellende Näherung nicht ausreichend sein könnte.

Messung der Entropie durch Stichprobenmittelung

Die von Comon (1994) vorgeschlagene zusätzliche Näherung des Logarithmus ist nur für kleine Korrekturen der Gauß'schen Approximation zufriedenstellend, d.h. für $f(y) \approx 1$. Für Dichten mit Nicht-Gauß'schem Abfall ergeben sich beträchtliche Korrekturterme, die, wie oben angemerkt, sogar negativ sein können. In diesem Fall versagt die Näherung des Logarithmus vollkommen. Daher wird hier vorgeschlagen, die Entropie direkt durch Mittelung des Logarithmus der geschätzten Verteilung $\ln p_{est}(y)$ über die Lerndaten zu bestimmen:

$$H[p(y)] \approx -\int_{-\infty}^{\infty} p(y)\ln p_{est}(y)\,dy \approx -\frac{1}{N}\sum_{i}^{N}\ln p_{est}(y_i) \tag{7.38}$$

wobei man die Dichte $p(y)$ durch die an den Stichproben $\{y_1, .., y_N\}$ diskretisierte Verteilung ersetzt hat.

$$p(y) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(y - y_i) \tag{7.39}$$

Hier ist $\delta(y)$ die Dirac'sche Punktdichte. Mit der Edgeworth-Entwicklung kann man die Entropie in einen Entropieterm für die Gauß'sche Näherung und einen Entropieterm für den Korrekturterm $f(y)$ trennen

$$H[p(y)] \approx \frac{1}{2}\ln(2\pi e) + \ln\sigma - \frac{1}{N}\sum_{i}^{N}\ln f(y) \tag{7.40}$$

Weiterhin läßt sich jetzt die in (7.36) abgebrochene Entwicklung p_{est} korrigieren, indem man für alle $f_{est}(y_i) < 0$, $f_{est}(y_i) \rightarrow 0$ setzt. Für die Messung der Entropie gemäß dieser Stichprobenmittelung gibt es im Prinzip keine Beschränkung auf eine bestimmte Ordnung. Man kann an dieser Stelle auch jede andere, nicht-parametrische Dichteschätzung verwenden, z.B. Parson-Dichten, die auf Momenten basieren. Für eine detailierte Diskussion über die Eigenschaften von verschiedenen Schätzungsmethoden auf der Grundlage von Momenten und Kumulanten siehe Kendall und Stuart (1969). Statt Näherungen auf der Grundlage von Kumulanten zu betrachten, könnte man auch das Parzens-Schätzungsverfahren verwenden, (siehe Duda und Hart, 1973). Dieses nicht-parametrische Verfahren ist insbesondere für die Fälle interessant, bei denen nur ein kleiner Stichprobensatz für das Lernen vorliegt. Einerseits geben die gemessenen Kumulanten für kleine N die wahren Kumulanten nur ungenau wieder, andererseits wird die Parzens-Schätzungsmethode rechnerisch zugänglich.

Nummerischer Vergleich der Entropiemessungen

In Tabelle 7.1 werden die verschiedenen Maße für die Entropie verglichen. Die Werte in der Zeile "Partitionierung" werden durch Zählen der $n(i)$ Datenpunkte, die in äquidistante Intervalle i mit einer Breite Δy fallen, bestimmt. Man summiert $-p(i)\Delta y \ln p(i)$ über alle Intervalle, wobei $p(i)\Delta y = n(i)/N$. Das liefert im Vergleich zu den theoretischen Werten nur für relativ große Stichprobenanzahl N ein gutes Resultat. Diese Werte werden hier angegeben, um eine verläßlichere Schätzung der Entropie im Fall der Exponentialverteilung zu haben, bei der Kumulantenmethoden im allgemeinen versagen. Die Resultate für die Exponentialverteilung weisen auf die Schwierigkeit der von Comon (1994) vorgeschlagenen Messung hin, während die Messung, die in 7.40 vorgestellt wird, auch für (unzuverlässige) Kumulanten fünfter und sechster Ordnung stabil ist. Die Resultate für die symmetrische Dreiecksverteilung und für die Gleichverteilung verdeutlichen nochmal die Insensibilität der Gauß'schen oberen Grenze für das Beispiel von Abbildung 7.5. In der Projektion auf die Abszisse und Ordinate lassen sich in diesem Beispiel eine Dreiecksverteilung für den Winkel $\alpha = \Pi/4$ und eine Gleichverteilung für $\alpha = 0$ beobachten. Die Entropie der Gauß'schen Näherung ist für beide Fälle gleich. Die Maße, die Kumulanten höherer Ordnung verwenden, bestimmen hingegen korrekt die minimale Entropie bei konstanter Gesamtinformation für die Gleichverteilung bei $\alpha = 0$.

Optimierung bei Kumulanten höherer Ordnung

Im Optimierungskriterium der gemeinsamen Information in Gleichung (7.24) setzt man die Näherungen für die Entropie der Einzelkoordinaten (7.37) oder (7.40) ein. Diese Näherungen sind mittels der Momente als Funktion der gemessenen Datenpunkte an der Ausgabe ausgedrückt. Siehe dazu auch die De-

Gemessene Entropie für simulierte Dichten	normalverteilt	gleichverteilt	symmetrisch dreiecksverteilt	exponentiell + Gauß'sches Rauschen
Partitionierung	1.35 ± 0.02	0.024 ± 0.006	0.14 ± 0.02	1.31 ± 0.03
Gauß'sche obere Grenze (7.29)	1.415 ± 0.02	0.18 ± 0.016	0.18 ± 0.02	1.53 ± 0.04
Comon (7.37)	1.414 ± 0.02	0.14 ± 0.015	0.17 ± 0.02	3.0 ± 2.5
Mittel (7.40) bis vierte Ordnung	1.414 ± 0.02	0.13 ± 0.015	0.17 ± 0.02	1.39 ± 0.05
Mittel (7.40) bis sechste Ordnung	1.414 ± 0.02	0.092 ± 0.001	0.16 ± 0.02	1.3 ± 0.5
Theoretischer Wert	1.419	0.0	0.1534	

Tabelle 7.1: Entropie für verschiedene Dichten, die mit $N = 1000$ Stichproben simuliert wurden. Gemessen wurde mit den verschiedenen Näherungsmethoden, die im Text erklärt werden. Die angegebenen Standardabweichungen wurden numerisch durch mehrfache Wiederholung der Experimente bestimmt.

finitionsgleichungen für Kumulanten und Momente (7.33) und (7.35) und die Messung (7.34).

Der Gradient dieser Optimierungskriterien nach den Ausgabekoordinaten kann demnach explizit bestimmt werden. Die Ausgabekoordinaten wiederum können nach den adaptiven Parametern der Abbildung differenziert werden. Man kann also im Prinzip den Gradienten der Kostenfunktion berechnen und somit Gradientenmethoden verwenden. Obwohl die numerische Komplexität im Vergleich zum einfachen Kriterium der minimalen Varianz nicht wesentlich vergrößert wird, erschwert sich die Komplexität des resultierenden Algorithmus erheblich. Daher wählt man hier ein Optimierungsverfahren, das lediglich den Wert der Kostenfunktion verwendet. Man umgeht damit die aufwendige Berechnung von Gradienteninformation. Ein einfaches stochastisches und paralelles Optimierungsverfahren wurde kürzlich von (Unnikrishnan und Venugopal, 1994) unter dem Namen ALOPEX eingeführt. Das Verfahren basiert auf der Korrelation der Änderung der Kostenfunktion mit der Änderung der Adaptationsparameter. Man adaptiert alle Parameter der Abbildung gleichzeitig, nachdem man die Kostenfunktion für alle Datenpunkte berechnet hat. Dieses einfache Optimierungsverfahren ermöglicht, die eher komplexen Kostenfunktionen mit Kumulanten höherer Ordnung gemeinsam mit den vorgestellten volumenerhaltenden und adaptiven Abbildungen einfach zu optimieren.

Vorteil von Statistik höherer Ordnung - Experimente

Der Vorteil von Statistik höherer Ordnug zur Optimierung der gemeinsamen Information gemäß Näherung (7.40) wird in zwei einfachen Experimenten demonstriert. Als volumenerhaltende Abbildung wir hier die explizit symplektische Abbildung, so wie sie in Abschnitt 7.1.2 beschrieben wird, verwendet. Der numerische Aufwand ist im Vergleich zu den Experimenten mit den implizit symplektischen Transformationen minimal.

Wie bereits erklärt, können Techniken, die lediglich Statistik zweiter Ordnung verwenden, bei einer rotierten, quadratischen, gleichverteilten Dichte die statistisch unabhängigen Koordinaten nicht korrekt bestimmen. Die statistisch un-

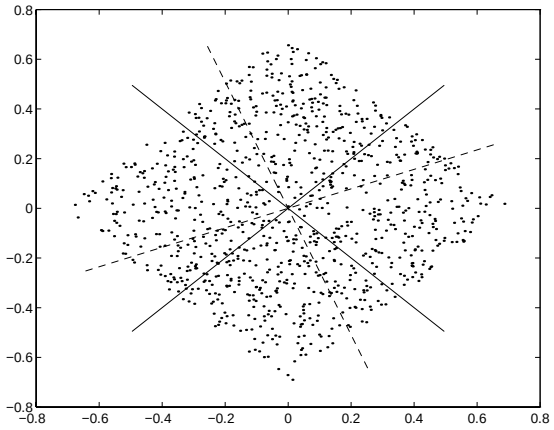

Abbildung 7.5: Simulierte gleichverteilte Dichte rotiert um $\pi/4$. Durchgezogene Linien repräsentieren die Richtungen, die durch ein Kriterium höherer Ordnung gefunden wurden. Gestrichelte Linien repräsentieren Richtungen, die durch Hauptkomponentenanalyse berechnet wurden. Diese sind völlig zufällig und schwanken stochastisch von Experiment zu Experiment.

abhängigen Koordinaten sind einfach die Achsen parallel zu den Kanten der Dichte (siehe Abbildung 7.5). Eine Rotation, d.h. eine lineare Transformation, genügt für diese Aufgabe. Aber die Kovarianzmatrix ist diagonal für jede Orientierung des Quadrats und gibt von daher keinerlei Information über die richtige Orientierung. Es ist bekannt, daß die Hauptkomponentenanalyse für nicht-normalverteilte Daten keine statistisch unabhängigen Koordinaten findet. Ebenso versagt an diesem Beispiel das Optimierungskriterium der Varianz als oberer Grenze. Statt dessen war es mit jedem der oben erklärten Kriterien höherer Ordnung möglich, für beliebige, linear transformierte, zweidimensionale, gleichverteilte Dichten die richtige Orientierung zu finden. Bei der symplektischen Abbildung (7.16) hat man sich in diesem Experiment auf quadratische generierende Funktionen beschränkt, da eine lineare Transformation für diese Aufgabe ausreicht.

Das zweite Experiment zeigt, daß die in diesem Abschnitt vorgeschlagene Methode tatsächlich nichtlineare Abhängigkeit zwischen der Eingabe und der Ausgabe finden kann. Ein eindimensionales Signal, das gemäß der Dichte von Abbildung 7.4 verteilt ist, wird nichtlinear in ein zweidimensionales Signal transformiert und mit additivem Gauß'schem Rauschen gestört. Die resultierende Signaldichte, die als Eingabe verwendet wird, ist in Abbildung 7.6 (links) zu sehen. Für die Aufgabe, statistisch unabhängige Koordinaten in dieser Dichte zu finden, wurde eine explizite symplektische Abbildung (7.16) mit $n = 2$ und $m = 6$ verwendet. Auf der rechten Seite von Abbildung 7.6 sind die verschiedenen Resultate der Optimierung mit dem Kriterium der Gauß'schen oberen Grenze (7.29) und

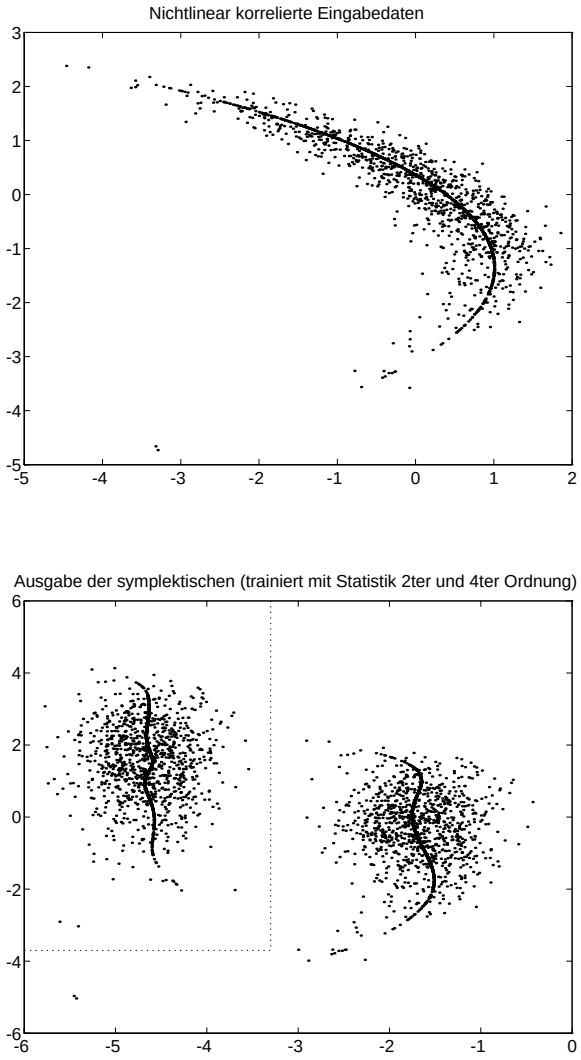

Abbildung 7.6: Eine implizit symplektische Transformation wurde mit Kriterien zweiter und vierter Ordnung trainiert, um statistisch unabhängige Koordinaten zu finden. Oben: Eingabedichte; die Linie durch die Mitte der Datenpunkte zeigt das nichtlinear transformierte, noch ungestörte Signal, das gemäß der Dichte in Abbildung 7.2 verteilt ist. Unten: Ausgabedichten, die von der symplektischen Abbildung nach dem Training mit den verschiedenen Kostenfunktionen generiert wurden. Sie werden in dem gleichen Rahmen gezeigt, um besser verglichen werden zu können. Die Verteilung rechts entspricht dem Kriterium zweiter Ordnung, während die Verteilung links sich mit Kumulanten bis zur vierten Ordnung ergibt. Man beachte die unterschiedliche Transformation der zentralen Linie.

dem Kriterium der genäherten Entropie (7.40) zu sehen. Offensichtlich verbessert die Statistik höherer Ordnung das Ergebnis, indem es die bessere Darstellung der nichtlinearen Abhängigkeit findet.

7.3 Dichteschätzung mit nichtlinearer Faktorisierung

Die resultierende faktorielle Darstellung der Daten an der Ausgabe ermöglicht eine einfache Dichteschätzung. In Abschnitt 7.3.2 wird diese Methode insbesondere zur Neuheitsdetektion verwendet.

Weiß man, daß eine Wahrscheinlichkeitsdichte faktorisiert, so reduziert sich die Aufgabe der Schätzung einer n-dimensionalen gemeinsamen Dichte auf die Aufgabe der Schätzung von n eindimensionalen Wahrscheinlichkeitsdichten. Hat man eine Schätzung an der Ausgabe gefunden, so läßt sich die Dichte an der Eingabe jederzeit berechnen, falls man eine umkehrbare und differenzierbare Abbildung verwendet hat. Gemäß Papoulis (1991) gilt:

$$p(\mathbf{x}) = \left. \frac{p(\mathbf{y})}{det\left(\frac{\partial \mathbf{f}^{-1}}{\partial \mathbf{y}}\right)} \right|_{\mathbf{y}=\mathbf{f}(\mathbf{x})} = \left. \frac{\prod_i^n p(y_i)}{det\left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)} \right|_{\mathbf{y}=\mathbf{f}(\mathbf{x})} = \left. \prod_i^n p(y_i) \right|_{\mathbf{y}=\mathbf{f}(\mathbf{x})} \tag{7.41}$$

Die zweite Gleichheit gilt, da man volumenerhaltende Transformationen benutzt. Verwendet man die Gauß'sche obere Grenze als Optimierungskriterium, so hat man in Abschnitt 7.2.1 gesehen, daß das Training eine normalverteilte Dichte als optimale Lösung besitzt. Falls das Lernen erfolgreich ist, wird man die Dichte mit der naheliegenden Annahme von normalverteilten Ausgabedaten nähern können:

$$p(\mathbf{x}) = \prod_i^n G(y_i) = \prod_i^n \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(y_i - \langle y_i \rangle)^2}{2\sigma_i^2} \right) \tag{7.42}$$

Eine mögliche Einschränkung der symplektischen Transformation ist bereits angesprochen worden und wird hier in Abbildung 7.7 verdeutlicht. Eine volumenerhaltende stetige Abbildung ist nicht in der Lage, disjunkte Signalverteilungen auf eine zusammenhängende Dichte zu transformieren. Die zwei getrennten '+' Punktwolken in Abbildung 7.7 können nicht in eine zusammenhängende Normalverteilung transformiert werden, ohne eine Diskontinuität im Zwischenraum zu haben. Falls die Gauß'sche Näherung wie hier nicht ausreichend ist, kann das Faktorisierungsverfahren zur Dichteschätzung dennoch verwendet werden. Man muß bei der Dichteschätzung der Einzelkoordinaten flexiblere Verfahren verwenden, so wie im vorhergehenden Abschnitt dargelegt.

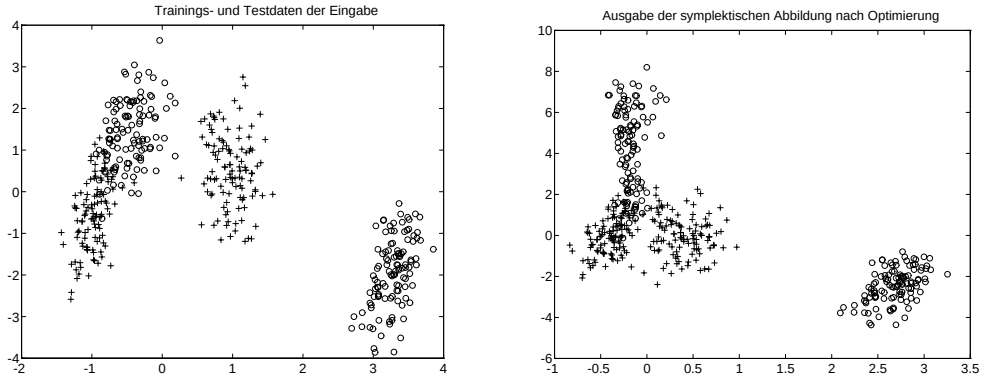

Abbildung 7.7: '+' Trainingsdaten. 'o' Testdaten. Links: Eingabesignale; Rechts: Ausgabesignale einer trainierten, impliziten, symplektischen Transformation. Die symplektische Transformation bildet die zwei getrennten '+' Punktwolken so gut wie möglich in einer Normalverteilung ab. Die Abbildung besitzt sechs adaptive Kopplungsparameter ($n = m = 2$). Datenpunkte 'o', die nicht zu den Trainingsdaten gehören, werden von dem Gebiet zwischen den zwei Trainingspunktwolken "wegtransformiert". Ellipsen kennzeichnen mögliche Klassifikationsgrenzen, um Trainingsdaten '+' von Testdaten 'o' zu trennen.

7.3.1 Schätzung der bedingten Dichte - Funktionsapproximation

In diesem Abschnitt wird ein Konzept zur Schätzung der bedingten Wahrscheinlichkeitsdichte vorgeschlagen. Es kann als Grundlage für zukünftige Arbeiten verwendet werden. In vielen Anwendungen für die Modellierung von funktionalen Abhängigkeiten ist man mit der Situation konfrontiert, keine rein deterministische Abhängigkeit in den vorgegebenen Trainingsdaten zu haben. Bei der Vorhersage von stochastischen Zeitreihen, wie z.B. Wechselkursvorhersage oder Stromverbrauchsvorhersage für Kraftanlagen, ist man nicht nur an einem Erwartungswert interessiert. Vielmehr will man auch eine Angabe darüber, mit welcher Wahrscheinlichkeit sich der prognostizierte Wert in einem gewissen Intervall befinden wird. Man will also aus einem Satz von Lerndaten $(\mathbf{y}_1, \mathbf{x}_1), .., (\mathbf{y}_N, \mathbf{x}_N)$ eine bedingte Wahrscheinlichkeitsdichte $p(\mathbf{y}|\mathbf{x})$ gewinnen, die angibt, mit welcher Wahrscheinlichkeit ein Signal $\mathbf{y}$ bei gegebenem $\mathbf{x}$ in einem bestimmten Intervall gefunden wird.

Eine Möglichkeit hierzu ist die Modellierung der gemeinsamen Wahrscheinlichkeitsdichte $p(\mathbf{y}, \mathbf{x})$. Aus dieser läßt sich dann die bedingte Dichte $p(\mathbf{y}|\mathbf{x})$ bestimmen:

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{y}, \mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{y}, \mathbf{x})}{\int p(\mathbf{y}, \mathbf{x})\, d\mathbf{y}} \tag{7.43}$$

Dieses Konzept zur Modellierung einer stochastischen Abhängigkeit wurde in

verschiedenen adaptiven konnektionistischen Modellen angewandt (Bishop, 1994; Redlich, 1993b). Die Hauptschwierigkeit bei diesem Ansatz liegt im allgemeinen in der Berechnung des Integrals im Nenner von (7.43).

Hier wird vorgeschlagen, zur Schätzung der bedingten Dichte, eine faktorielle Repräsentation wie in (7.41) für $p(\mathbf{y}, \mathbf{x})$ und gleichzeitig für $p(\mathbf{x})$ zu gewinnen. Die volumenerhaltende Transformation, die in Abschnitt 7.1.3 vorgestellt wurde, ermöglicht dies für eine spezielle Wahl der generierenden Funktionen. Die Struktur der verwendeten Transformation ist in Abbildung 7.8 zu sehen. Der Eingaberaum der Volumenerhaltenden sei nun der gemeinsame Raum $(\mathbf{x}, \mathbf{y}) \in \Re^{(n+m)}$ und die Ausgabe sei $\mathbf{z} = (\xi, \eta) \in \Re^{(n+m)}$. Man betrachte die Transformation

$$\begin{array}{rcl} \mathbf{g}_\xi(\eta) \quad +\xi & = & \mathbf{x} + \mathbf{f}_\xi(\mathbf{x}) \\ \mathbf{g}_\eta(\xi, \eta) \quad +\eta & = & \mathbf{y} + \mathbf{f}_\eta(\mathbf{y}) \end{array} \tag{7.44}$$

Man identifiziert diese Funktionen mit denen in Gleichung (7.17) durch $\mathbf{g}(\mathbf{z}) := (\mathbf{g}_\xi(\eta), \mathbf{g}_\eta(\xi, \eta))$ und $\mathbf{f}(\mathbf{x}, \mathbf{y}) := (\mathbf{f}_\xi(\mathbf{x}), \mathbf{f}_\eta(\mathbf{y}))$. Die Jacobimatrix der Funktion $\mathbf{g}(\mathbf{z}) : \Re^{(n+m)} \mapsto \Re^{(n+m)}$ muß wieder linke, untere Dreiecksgestalt haben. Die Funktion $\mathbf{f}(\mathbf{x}, \mathbf{y}) : \Re^{(n+m)} \mapsto \Re^{(n+m)}$ wählt man so, daß die Untermatrix $\partial \mathbf{f}_\xi / \partial \mathbf{y} = 0$. Die Untermatrizen $\partial \mathbf{f}_\xi / \partial \mathbf{x}$ und $\partial \mathbf{f}_\eta / \partial \mathbf{y}$ haben weiterhin rechte, obere Dreiecksgestalt. Man überprüft leicht, daß diese Abbildung der oben angegebenen Volumenerhaltenden entspricht. Hier handelt es sich allerdings um einen Spezialfall, bei dem der Teil $\xi \in \Re^n$ des Ausgaberaums nur von der Variablen $\mathbf{x}$, nicht aber von $\mathbf{y}$, abhängig ist (siehe Abbildung 7.8). Betrachtet man lediglich die Abbildung $\mathbf{x} \in \Re^n \mapsto \xi \in \Re^n$, so ist diese für sich genommen volumenerhaltend. Die Ausgabevariable $\eta \in \Re^m$ hängt hingegen von allen $(\mathbf{x}, \mathbf{y})$ Variablen ab. Man wird also das vorgestellte Faktorisierungsverfahren gleichzeitig für $\mathbf{x}$ und $(\mathbf{x}, \mathbf{y})$ durchführen können.

Durch eine der oben angegebenen Methoden wird die gemeinsame Information im gesamten Raum von $\mathbf{z}$ minimiert. Sind die verwendeten adaptiven Funktionen $\mathbf{g}(\mathbf{z})$ und $\mathbf{f}(\mathbf{x}, \mathbf{y})$ allgemein genug, so erhält man eine faktorielle Darstellung für $p(\mathbf{x})$ und $p(\mathbf{x}, \mathbf{y})$

$$p(\mathbf{x}) = \prod_i^n p(\xi_i) \tag{7.45}$$

$$p(\mathbf{x}, \mathbf{y}) = \prod_i^{n+m} p(\mathbf{z}_i) = \prod_i^n p(\xi_i) \prod_i^m p(\eta_i) \tag{7.46}$$

Hiermit vereinfacht sich die bedingte Wahrscheinlichkeitsdichte (7.43) zu

$$p(\mathbf{y}|\mathbf{x}) = \prod_i^m p(\eta_i(\mathbf{x}, \mathbf{y})) \tag{7.47}$$

Verwendet man zur Minimierung der gemeinsamen Information von $p(\mathbf{z})$ die Gauß'sche obere Grenze, so sind nach erfolgreichem Training die Dichten $p(\eta_i)$

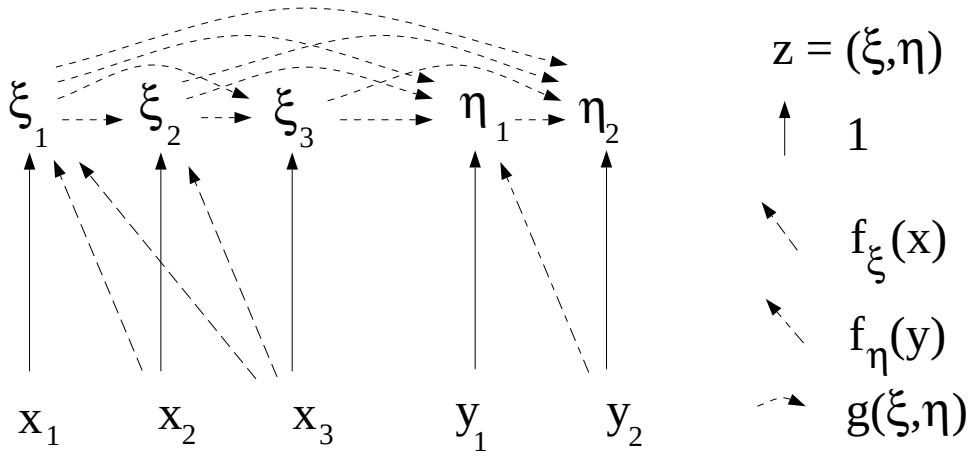

Abbildung 7.8: Diese Transformation ist in $(\mathbf{x}, \mathbf{y}) \mapsto (\xi, \eta)$ als auch in $\mathbf{x} \mapsto \xi$ volumenerhaltend. Sie wird hier für $n = 3$ und $m = 2$ dargestellt. Diese Transformation kann zur gleichzeitigen Faktorisierung in beiden Räumen verwendet werden. Es läßt sich hiermit eine Schätzung der bedingten Wahrscheinlichkeit $p(\mathbf{y}|\mathbf{x})$ verwirklichen.

der Ausgabekoordinaten durch Gaußfunktionen gegeben. Ihr Maximum liegt bei $\langle\eta\rangle$ und ist leicht zu bestimmen. Man kann jetzt hiermit die Größe $\mathbf{y}^*$ bestimmen, bei der die bedingte Wahrscheinlichkeitsdichte $p(\mathbf{y}^*|\mathbf{x})$ maximal ist. Das ist der wahrscheinlichste Wert von $\mathbf{y}$ bei gegebenem $\mathbf{x}$. Dieser ist gegeben durch das Maximum der Normalverteilung $\eta(\mathbf{x}, \mathbf{y}^*) = \langle\eta\rangle$. Setzt man dieses maximale η in (7.44) ein, so erhält man:

$$\begin{array}{rcrcl} \mathbf{g}_\eta(\xi, \langle\eta\rangle) & + & \langle\eta\rangle & = & \mathbf{y}^* + \mathbf{f}_\eta(\mathbf{y}^*) \\ \mathbf{g}_\xi(\langle\eta\rangle) & + & \xi & = & \mathbf{x} + \mathbf{f}_\xi(\mathbf{x}) \end{array} \tag{7.48}$$

Diese Gleichung definiert wieder eine explizite Abbildung zwischen einem gegebenem $\mathbf{x}$ und dem wahrscheinlichsten $\mathbf{y}^*$. Dies stellt somit eine spezielle Form der Funktionsapproximation dar. So wie in Abschnitt 7.1.3 wird jede Ausgabekoordinate nacheinander berechnet. Man gewinnt aus $\mathbf{x}$ die Werte für ξ. Mit diesen wird dann nacheinander jede Koordinate von $\mathbf{y}^*$ aus ξ und $\langle\eta\rangle$ entsprechend bestimmt. Man beachte, daß obwohl η normalverteilt ist, die Variable $\mathbf{y}$ keinesfalls einer Normalverteilung genügen muß.

Ebenso wichtig wie die funktionale Abhängigkeit, die man hier gewinnt, ist der einfache Ausdruck für die gesamte bedingte Wahrscheinlichkeitsdichte $p(\mathbf{y}|\mathbf{x})$ in (7.47). Man kann daraus die Wahrscheinlichkeit berechnen, einen Punkt $\mathbf{y}$ in einem Intervall $\Delta\mathbf{y}$ bei gegebenem $\mathbf{x}$ zu finden

$$W(\mathbf{y} \in [\mathbf{y}, \mathbf{y} + \Delta\mathbf{y}]) = \int_{\eta(\mathbf{x},\mathbf{y})}^{\eta(\mathbf{x},\mathbf{y}+\Delta\mathbf{y})} \prod_{i}^{m} (p(\eta_i)\, d\eta)$$

Dies gilt für alle der oben angegebenen Kostenfunktionen. Wählt man natürlich die Gauß'sche obere Grenze, so wird dies zu einem Standardintegral:

$$W(\mathbf{y} \in [\mathbf{y}, \mathbf{y} + \Delta\mathbf{y}]) = \prod_{i}^{m} \sigma_i(\Phi(\eta_i(\mathbf{x}, \mathbf{y} + \Delta\mathbf{y})/\sigma_i) - \Phi(\eta_i(\mathbf{x}, \mathbf{y})/\sigma_i))$$

hier ist Φ die Stammfunktion der Gaußfunktion und σ_i die meßbare Standardabweichung der Variablen η_i.

Grundlegend für dieses Verfahren ist, daß die Faktorisierungen (7.45) und (7.46) erzeugt werden können. Die Funktionenklasse, die hier zugelassen wird, ist wesentlich allgemeiner als die symplektischen Transformationen. Es ist daher zu erwarten, daß man zumindest gleich gute Ergebnisse der Faktorisierung erhält. Die Kaskadierung der Abhängigkeiten mag zwar die Berechnung der Funktionswerte, insbesondere in hohen Dimensionen, erschweren; die Allgemeinheit der Funktionenklasse wird aber dadurch nicht beeinträchtigt. Auch Gradientenverfahren, die zur Optimierung verwendet werden können, dürften durch die Kaskadierung keine besonderen Schwierigkeiten aufweisen. Aus der Erfahrung der Error-Back-Propagation weiß man, daß eine Rückpropagierung über mehr als zwei Stufen vernachlässigt werden kann.

Dieses Konzept stellt also eine vielversprechende Grundlage für zukünftige Arbeiten dar.

7.3.2 Neuheitsdetektion

Schließlich soll das vorgestellte Verfahren auf das Problem der Neuheitsdetektion angewendet werden. Die Aufgabe der Neuheitsdetektion ist mit der Dichteschätzung eng verwandt. Mit Hilfe eines gegebenen Satzes von Stichproben einer vorgegebenen Dichte soll entschieden werden, ob ein neuer Datenpunkt zu der bekannten Dichte gehört oder nicht? Für eine gegebene Entscheidungsschwelle entspricht das der Frage nach der zugehörigen Kontur der Wahrscheinlichkeitsdichte der beobachteten Datenpunkte. Wird die Kontur für beliebige Entscheidungsschwellen benötigt, so ist die gesamte Wahrscheinlichkeitsdichte gesucht. Als Lösung dieser Aufgabe wird hier die bereits vorgestellte symplektische Faktorisierung mit der einfachen Gauß'schen Dichteschätzung (7.42) vorgeschlagen. Die Entscheidungsfläche für die Neuheitsdetektion ist dann lediglich ein Hypersphäre im Ausgaberaum der symplektischen Transformation nach der Minimierung der gemeinsamen Information für die gegebenen Stichproben.

Abbildung 7.7 verdeutlicht diese Idee. Eine implizite, adaptive, symplektische Abbildung wurde mit den Datenpunkten '+' trainiert. Die Datenpunkte

'o' sollen von diesen getrennt werden können, d.h. als nicht zu dieser Verteilung gehörend erkannt werden. Offensichtlich liefert (nach Normierung der Varianz) eine kreisförmige Entscheidungsgrenze an der Ausgabe eine brauchbare Lösung. In diesem Beispiel könnte man natürlich auch mit einer "gaussian mixture"-Methode (Duda und Hart, 1973) mit zwei Gauß'schen Elementen ein gutes Resultat erhalten. Trotzdem soll für dieses einfache Beispiel zugunsten des Verständnisses von späteren Experimenten das Ergebnis der Klassifizierung präsentiert werden.

Eine mögliche quantitative Beurteilung einer binären Klassifizierungsmethode stellen die Rate der "verfehlten" Klassifizierung (misclassification) und die Rate der "falschen" Alarme (false-alarms) dar. Das wird in Abbildung 7.9 verdeutlicht. Eine Klassifizierung bezeichnet man als verfehlt, falls ein "neues" Signal fälschlicherweise als "normales" Signal bewertet wird. Den entgegengesetzten Fall, bei dem ein "normales" Signal als "neu" klassifiziert wird, nennt man entsprechend falschen Alarm (siehe Abbildung 7.9, links).

Neuheitsdetektion ist explizit unüberwachtes Lernen. Es kann auch als eine binäre Klassifizierung verstanden werden, bei der nur eine Klasse bekannt ist. Während des Trainings ist es nicht erlaubt, die Verteilung der "neuen" Datenpunkte zu verwenden, da auch in der Praxis diese Verteilung nicht bekannt sein wird. Stünde diese Information zur Verfügung, könnten verschiedenste überwachte Klassifizierungstechniken verwendet werden. Trotzdem benötigt man während der Entwicklungsphase eines Neuheitsdetektionsalgorithmus die Verteilung der "neuen" Signale, um die Qualität des Algorithmus überprüfen zu können.

Die Transinformation entfernt die nichtlineare, statistische Abhängigkeit in den Trainingsdaten, indem sie normalverteilte Ausgabedichten erzeugt. Hat man eine gute Lösung hierfür erhalten, so läßt sich mit einem unüberwachten Kriterium nichts weiter verbessern. An dieser Stelle hat man eventuell eine perfekte Darstellung der bekannten Signalverteilung. Nun benötigt man Information über die Verteilung der "neuen" Signale. Dieses Wissen kann implizit sein, z.B. ein Qualitätskriterium wie die Fehlerraten von Abbildung 7.9. In diesem Beispiel liefern die Fehlerkurven Information darüber, welche der beiden extrahierten statistisch unabhängigen Koordinaten zur Trennung der Signale besser geeignet sind (siehe Abbildung 7.9 rechts). Eine Koordinate des Merkmalsraums (Abszisse) enthält für die Entscheidung, die man treffen will, im wesentlichen nur Rauschen, während die zur Trennung wichtige Information in der Ordinate enthalten ist. Dieses anschauliche Beispiel zeigt, wie Rauschen und nützliche Information als solche erst erkannt werden können, wenn ein Entscheidungskriterium, also eine Form von Überwachung, vorliegt.

7.3.3 Motorausfallsvorhersage

In diesem Abschnitt soll gezeigt werden, daß das vorgeschlagene Verfahren zur Neuheitsdetektion gute Resultate für ein reales technisches Problem in einem hochdimensionalen Raum liefert. Ziel in der Motorausfallsvorhersage ist es, Un-

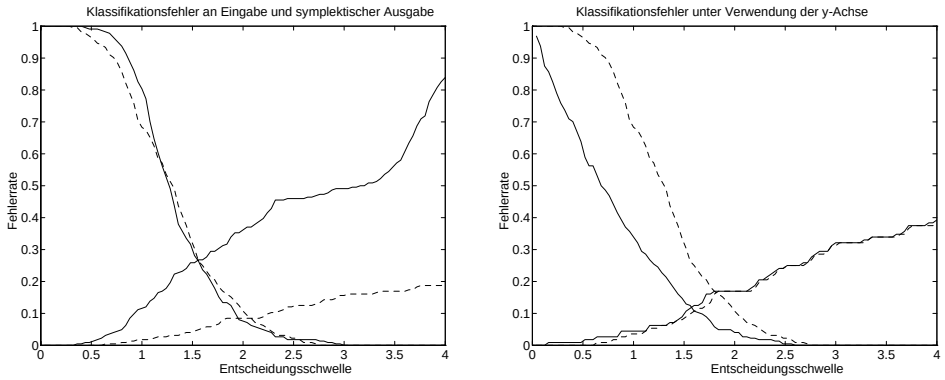

Abbildung 7.9: Fehlerrate von verfehlter Klassifikation und falschem Alarm (misclassification und false-alarms) für das Beispiel der vorhergehenden Abbildung. Die Trainingspunkte '+' sollen als "normal" und die Testpunkte 'o' als "neu" erkannt werden. Man benutzte für die Eingabe und Ausgabe eine elliptische Abstandsmessung, d.h. man klassifiziert alle Punkte, die innerhalb der Ellipse um den Schwerpunkt liegen, als "normal", sonst als "neu". Die abfallende Fehlerrate gibt die falschen Alarme an, während die steigende Kurve die Rate der verfehlten Klassifikationen wiedergibt. Der Schnittpunkt dieser beiden Kurven kann als Qualitätskriterium des Klassifikationsalgorithmus verwendet werden. Links: durchgezogene und gestrichelte Linien entsprechen der Entscheidungsgrenze in der Eingabe bzw. in der Ausgabe, so wie in der vorhergehenden Abbildung angedeutet. Offensichtlich ist die Klassifizierung an der faktorisierten Ausgabe wesentlich besser. Rechts: Man betrachte nun die zwei Koordinaten der erzeugten Ausgabe. Man hat für diesen Vergleich die Punktwolke 'o' rechts unten ignoriert. Jetzt entsprechen durchgezogene Linien der Abstandsmessung in beiden Koordinaten. Die gestrichelten Linien hingegen entsprechen der Messung des Abstands vom Schwerpunkt nur in Richtung der Ordinaten. Auf diese Weise wird ersichtlich, welche der extrahierten Koordinaten für die gewünschte Entscheidung die meiste Information enthält.

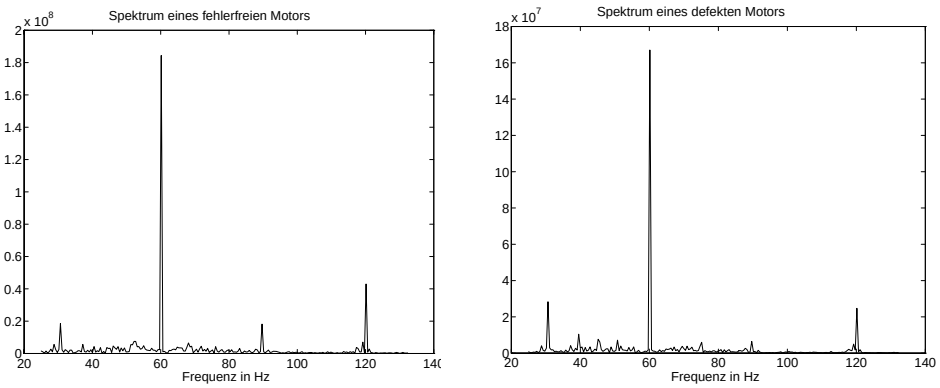

Abbildung 7.10: Links: Spektrum eines fehlerfrei arbeiten Elektromotors zwischen 20Hz und 130Hz. Rechts: Spektrum eines defekten Motors. Es sind keine einfachen Entscheidungskriterien auszumachen.

regelmäßigkeiten in Elektromotoren durch Überwachung des elektrischen Stromes zu erkennen. Das Frequenzspektrum des Stromes wird als Merkmalsvektor verwendet, (siehe Abbildung 7.10). Der Fehlerdetektor wird mit den Daten des fehlerfrei funktionierenden Motors trainiert. Der Detektor soll einen bevorstehenden Ausfall rechtzeitig erkennen.

Typischerweise hat man hier mindestens 100 und höchstens 1000 Dimensionen. Die Experimente wurden zunächst mit der impliziten symplektischen Transformation durchgeführt. Aufgrund der numerischen Komplexität schränkt das natürlich die Anzahl Dimensionen, die bearbeitet werden können, ein. Andererseits sind nicht alle 100 (oder mehr) Koordinaten nichtlinear miteinander korreliert. Wahrscheinlicher ist, daß die meisten Koordinaten - falls überhaupt - nur linear korreliert sind. Das Spektrum wird daher zunächst mit der linearen Hauptkomponentenanalyse transformiert. Es werden 230 Spektralkomponenten zwischen 20Hz und 130Hz verwendet.

Man beobachtet, daß nur wenige der ersten Hauptkomponenten paarweise nichtlinear korreliert sind. Es konnte keine paarweise nichtlineare Struktur für die Komponenten höher als die ersten 10 oder 15 erkannt werden. Alle anderen Koordinaten sind durch die Hauptkomponentenanalyse linear dekorreliert worden. Diese sind symmetrisch um einen Schwerpunkt verteilt. Sie können recht gut durch Normalverteilungen genährt werden, (siehe Abbildung 7.11, rechts). Für normalverteilte Daten reicht die lineare Dekorrelation zur Minimierung der gemeinsamen Information aus. Die Aufgabe ist somit für diese Hauptkomponenten im wesentlichen gelöst. Abbildungen 7.11, links und Mitte, zeigen wie zwei der ersten 10 Hauptkomponenten mit einer adaptiven impliziten Abbildung transformiert wurden ($n = 10$, $m = 20$). Die Struktur wurde mit 800 Datenpunkten aus 10 verschiedenen Versuchsreihen mit Motoren bei unterschiedlichen Belastungs-

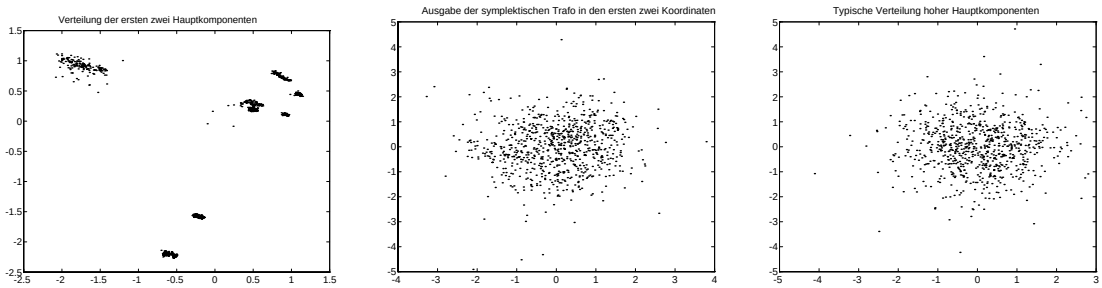

Abbildung 7.11: Links: Datenpunkte der ersten zwei Hauptkomponenten zeigen eine deutlich nichtlineare Abhängigkeit. Mitte: Resultierende Dichte der gleichen zwei Komponenten nach nichtlinearer Faktorisierung der ersten 10 Hauptkomponenten. Rechts: Für Komponenten höher als die 15te Hauptkomponente konnte keine paarweise Abhängigkeit beobachtet werden. Hier werden zufällig ausgewählte Komponenten 50 und 100 dargestellt.

bedingungen trainiert.

Man verwendet die erhaltene Transformation um "gute" von "schlechten" Motoren gemäß (7.42) unterscheiden zu können. Da die Resultate für die verschiedenen vorkommenden Fehlerfälle (Umwucht, Lagerringloch, Rotorbruch) variieren können, werden die Ergebnisse getrennt angegeben. In Abbildung 7.13 sieht man die Ergebnisse einer einfachen Maximumsnorm ($\max_i |y_i - \langle y_i \rangle|$) im gesamten 230-dimensionalen Raum der Hauptkomponenten (links) und die Gauß'sche Schätzung in den ersten 10 nichtlinear transformierten Hauptkomponenten (Mitte).

Es wurde auch analysiert, inwieweit man mit einer bestimmten Komponente einen "guten" Motor von einem "schlechten" Motor unterscheiden kann. Man kann dazu das Verhältnis der entsprechenden Varianzen vergleichen. Ein Maß für die Trennbarkeit der Dichten in einer bestimmten Koordinate i könnte folgendes sein:

$$\Delta_i = \frac{\langle (y_i^{test} - \langle y_i^{train} \rangle)^2 \rangle}{\langle (y_i^{train} - \langle y_i^{train} \rangle)^2 \rangle} \tag{7.49}$$

Der Vergleich der Trennbarkeit der verschiedenen Koordinaten liefert in dieser konkreten Problemstellung eine nützliche Erkenntnis. In Abbildung 7.12 wird Δ_i für alle linearen Hauptkomponenten dargestellt. Die linearen Hauptkomponenten kleiner Varianz werden im allgemeinem nach einer Hauptkomponentenanalyse verworfen, da man diese als reines Rauschen erachtet. Wie sich der Abbildung 7.12 entnehmen läßt, haben gerade diese tendenziell ein hohes Δ_i. Diese Koordinaten ermöglichen somit eine gute Trennung. Sie tragen für die Klassifizierungsaufgabe entscheidende Information. In eben diesen Richtungen steigt die Amplitude des Spektrums für fehlerhafte Motoren. Dieses Phänomen zeigt sich

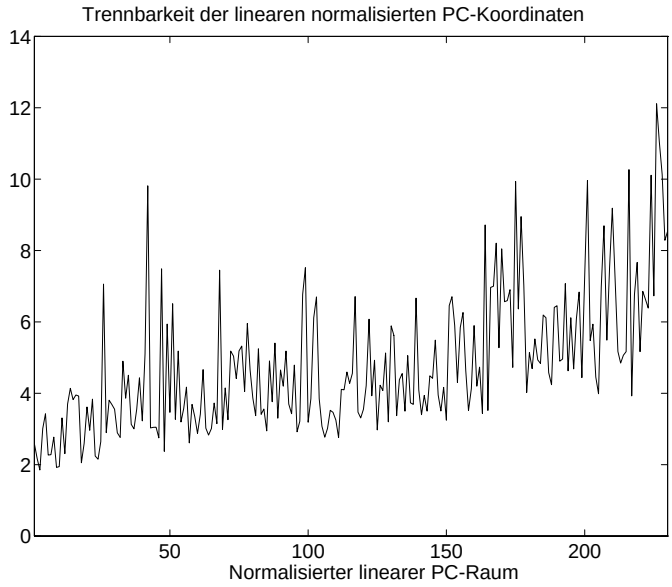

Abbildung 7.12: Maß Δ_i für die Trennbarkeit in den linearen Hauptkomponenten (PC) so wie sie von der Hauptkomponentenanalyse mit vorhergehender Normalisierung der Varianz aus dem Spektrum der Trainings- und Testdaten erzeugt werden.

am deutlichsten, wenn das Spektrum vor der Hauptachsentransformation in jeder Frequenz auf eine normierte Varianz skaliert wird. Man bezeichnet das hier als *normierte* Hauptkomponentenanalyse. Anschaulich bedeutet diese Beobachtung, daß der Abfall der diskreten Frequenzen (Abbildung 7.10) bei fehlerhaften Motoren breiter wird. Diese Tatsache wird somit nicht auf diese spezifischen Datensätzen beschränkt sein. Man wird also die normalisierten Hauptkomponenten niedriger Varianz in die Klassifikationsmessung miteinbeziehen. Dies geschieht am einfachsten, indem man diese als zusätzliche Koordinaten in die Euklidische Norm hinzufügt. Die besten Resultate erhält man, indem man zwischen fünf und 20 Hauptkomponenten niedriger Varianz hinzufügt (siehe Abbildung 7.13, rechts).

Ein weiteres Bewertungskriterium eines Klassifikationsalgorithmus ist der Klassifikationsfehler bei der optimalen Entscheidungsschwelle. Die vorgeschlagene Methode erreicht einen optimalen Klassifikationsfehler von 10.0%. Dieses Ergebnis ist vergleichbar zu anderen bekannten Methoden, die bei SCR[5] zu diesem Problem angewandt wurden; unter anderem MLP-Autoassozierer (11%), radiale Basisfunktionen-Autoassozierer (10%) , Nächster-Nachbar-Algorithmus (18%-32%), Hypersphären-Klusterung (10%), lineare Hauptkomponentenanalyse (12%)

[5]Siemens Corporate Research, Princeton, NJ, USA

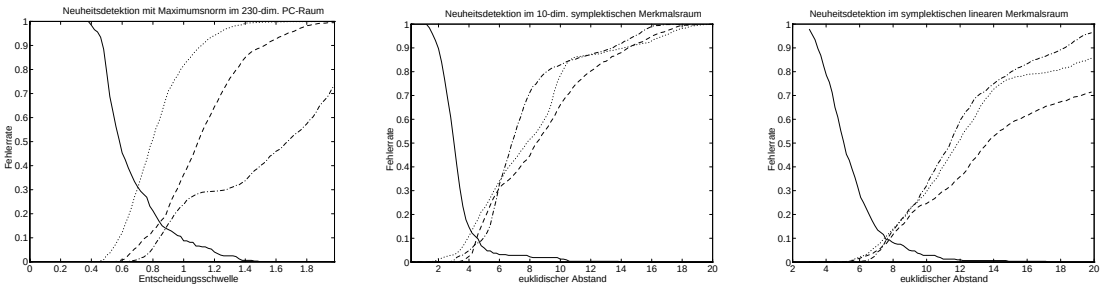

Abbildung 7.13: Links: Maximumsnorm im 230-dimensionalen Raum der Hauptkomponenten. Mitte: Sphärische Abstandsmessung in den 10 symplektisch transformierten ersten Hauptkomponenten. Die Transformation reduziert in 650 Zyklen die Varianz um 65%. Rechts: Kombinierter symplektischer und linearer Merkmalsraum: 10 symplektisch transformierte Hauptkomponenten und sieben der normalisierten linearen Komponenten niedriger Varianz. Die abfallende Kurve liefert die Rate der falschen Alarme. Jede der drei steigenden Kurven gibt die Rate an, mit der fehlerhafte Motoren unerkannt blieben. Es sind die drei verschiedenen Fehlerfälle getrennt angegeben ($\cdot\cdot$ Lagerringloch, $--$ Umwucht, $\cdot-$ Rotorbruch).

oder Maximumsnorm von 2000 Frequenzen (11%).

7.4 Schlußfolgerungen

In diesem Kapitel wurde ein neues informationstheoretisches Konzept für unüberwachtes Lernen bei kontinuierlichen Signalen vorgestellt: Man extrahiert statistische Merkmale, indem man eine faktorielle Darstellung der gemeinsamen Wahrscheinlichkeitsverteilung der Signale generiert. Dabei garantiert man mit volumenerhaltenden Transformationen die Erhaltung der Information des Eingabesignals.

Es wurden verschiedene adaptive, volumenerhaltende Transformationen vorgestellt. Die Faktorisierung reduziert sich durch die Nebenbedingung der Informationserhaltung auf die Minimierung der Entropie der Einzelkoordinaten an der Ausgabe.

Durch Statistik zweiter Ordnung läßt sich ein einfaches und effizientes Minimierungskriterium definieren. Für den Fall, daß Varianzen nicht ausreichend sind, wurden verbesserte Messungen der Entropie auf der Grundlage von Kumulanten vorgestellt.

Zusammenfassend läßt sich das Verfahren als eine nichtlineare Komponentenanalyse verstehen, die im Gegensatz zur linearen Hauptkomponentenanalyse Statistik höherer Ordnung und nichtlineare Transformationen verwendet.

Diese Faktorisierungsmethode bietet aber auch einen neuen Lösungsansatz zur klassischen Aufgabe der Dichteschätzung. Am Ende dieses Kapitels wird mit dieser Dichteschätzungsmethode das Problem der Neuheitsdetektion bearbeitet. Insbesondere wird das Verfahren auf ein technisches Problem der Motorausfallsvorhersage angewandt.

Die Methode ermöglicht aber auch, ein neues Verfahren zur Schätzung der bedingten Wahrscheinlichkeit zu definieren.

Die Vielfältigkeit der Anwendungen verdeutlicht, daß hier ein Lösungsansatz für eine grundlegende Aufgabe der informationstheoretischen Signalverarbeitung vorgestellt wurde: die Faktorisierung einer Wahrscheinlichkeitsverteilung.

Kapitel 8

Nomenklatur

Tabelle 8.1: **Nomenklatur aus der Einleitung**

$\mathbf{x}$	mehrdimensionale Eingabe oder Zustandsvariable
$\mathbf{y}$	mehrdimensionale Ausgabe oder Zustandsvariable
Ξ	diskreter Zustandsraum der Eingabe
Υ	diskreter Zustandsraum der Ausgabe
$P(\mathbf{x})$, $P(\mathbf{x};W)$	Verteilung der diskreten Zustandsvariablen $\mathbf{x}$ mit den Kopplungskoeffizienten W parametrisiert
$I[P(\mathbf{x})]$	Information einer diskreten Verteilung $P(\mathbf{x})$
$H[P(\mathbf{x})]$	Entropie einer diskreten Verteilung $P(\mathbf{x})$
$R(\mathbf{x})$	feste vorgegebene diskrete Verteilung
$H[P(\mathbf{x}),R(\mathbf{x})]$	relative Entropie zweier diskreter Verteilungen
$I[P(\mathbf{x},\mathbf{y})]$	gemeinsame Information der Zufallsvariablen $\mathbf{x}$ und $\mathbf{y}$ bzw. Transinformation der Eingabe zur Ausgabe
$H[P(\mathbf{x},\mathbf{y})]$	Entropie der gemeinsamen Verteilung $P(\mathbf{x},\mathbf{y})$
$MI[P(\mathbf{y})]$	gemeinsame Information der Einzelkoordinaten y_i von $\mathbf{y}$
$R[P(\mathbf{y})]$	Redundanz in einer diskreten Zufallsvariablen $\mathbf{y}$
$p(\mathbf{x})$	Dichtefunktion einer kontinuierlichen Wahrscheinlichkeitsverteilung oder kurz: kontinuierliche Wahrscheinlichkeitsverteilung
$H[p(\mathbf{x})]$	Entropie der Dichtefunktion $p(\mathbf{x})$
$MI[p(\mathbf{x})]$	gemeinsame Information der Dichtefunktion $p(\mathbf{x})$
σ	Varianz
$p(\mathbf{y}\|\mathbf{x})$	Dichtefunktion einer kontinuierlichen bedingten Wahrscheinlichkeitsverteilung oder kurz: kontinuierliche bedingte Wahrscheinlichkeitsverteilung

Tabelle 8.2: **Diskrete Boltzmann-Maschine**

s_i	$i$te binärer Zustand
$\mathbf{s}$	diskreter Zustandsvektor
w_{ij}	Wechselwirkungskoeffizienten der binären Zustände
W	Matrix der Wechselwirkungskoeffizienten
$\mathbf{h}$	innere Zustände
$\mathbf{v}$	sichtbare Zustände
$\mathbf{y}$	binäre Ausgabezustände
$\mathbf{x}$	kontinuierlicher Eingabevektor
w_{ij}^{ext}	Wechselwirkungskoeffizienten der binären Zustände mit der externen Eingabe $\mathbf{x}$
$E(\mathbf{s}), E(\mathbf{s}, W)$	Energie des Zustands $\mathbf{s}$ abhängig von Parameter W
$P(\mathbf{s}), P(\mathbf{s}, W)$	Boltzmann-Verteilung der binären Zustände
Z	Zustandssumme
$\beta = 1/T$	inverse Temperatur
$W(s_i \rightarrow -s_i)$	Übergangswahrscheinlichkeit für den $i$ten Zustand
$\langle\rangle$	freies Mittel
$\langle\rangle_{\mathbf{v}}$	fixiertes Mittel
T_c	kritische Temperatur
λ_{max}	größter Eigenwert der Kopplungsmatrix w_{ij}
$E(\mathbf{s}\|\mathbf{x})$	Energie des Zustands $\mathbf{s}$ bei externer Eingabe $\mathbf{x}$
$P(\mathbf{s}\|\mathbf{x})$	bedingte Wahrscheinlichkeit des Zustand $\mathbf{s}$ bei Eingabe $\mathbf{x}$
$R(\mathbf{x}, \mathbf{y})$	durch Datenpunkte ($\mathbf{x}$,$\mathbf{y}$) vorgegebene Verteilung.
$\langle\rangle_{\mathbf{x}}$	freies Mittel für fixierte Eingabe gemittelt über $R(\mathbf{x})$
$\langle\rangle_{\mathbf{y},\mathbf{x}}$	fixiertes Mittel für fixierte Eingabe gemittelt über $R(\mathbf{x})$

Tabelle 8.3: **Rotor-Boltzmann-Maschine**

s_{ik}	$k$te Dimension des $i$ten Rotors
$\mathbf{s}_i$	$i$te Rotor
$\mathbf{S}$	Vektor von Rotoren
$E(\mathbf{S})$	Energie der Rotor-Zustände
$v_{ik}, \mathbf{v}_i, \mathbf{V}$	entsprechende Mittelfeldvariablen
$u_{ik}, \mathbf{u}_i, \mathbf{U}$	konjugierte Mittelfeldvariablen
$E_{eff}(\mathbf{V}, \mathbf{U})$	effektive Energie der Mittelfeldvariablen oder freie Energie
$\mathbf{V}^0, \mathbf{U}^0$	Sattelpunkte der effektiven Energie
w_{ikjl}	Kopplungstärke der Koordinate s_{ik} mit s_{jl}
$\mathbf{w}_{ij}$	Kopplungsmatrix der $i$ten mit dem $j$ten Rotor
$\mathbf{W}$	gesamte Kopplungsmatrix
$\mathbf{f}(\mathbf{s}_i)$	mehrdimensionale Verallgemeinerung der Sigmoiden Aktivierungsfunktion bei Rotoren
$\mathbf{S}^h$	innere Rotorzustände
$\mathbf{S}^v$	sichtbare Rotorzustände
L	Lyapunov-Funktion
T_g	Grenztemperatur

Literaturverzeichnis

Abraham, R., und Marsden, J. (1978). *Foundations of Mechanics*. The Benjamin-Cummings Publishing Company, Inc., London.

Ackley, D., Hinton, G., und Sejnowski, T. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science*, *9*, 147–169.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing. *Network*, *3*, 213–251.

Atick, J., und Redlich, A. (1990). Towards a theory of early visual processing. *Neural Computation*, *2*, 308–320.

Atick, J., und Redlich, A. (1992). What Does the Retina Know about Natural Scenes. *Neural Computation*, *4*, 196–210.

Attneave, F. (1954). Informational Aspects of Visual Perception. *Psychological Review*, *61*, 183–193.

Baldi, P., und Hornik, K. (1989). Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minima. *Neural Networks*, *2*, 53–58.

Barlow, H. (1959). Sensory Mechanism, the Reduction of Redundancy, and Intelligence. In *National Physical Laboratory Symposium*, Vol. 10. Her Majesty's Stationery Office, London. The Mechanization of Thought Processes.

Barlow, H. (1989). Unsupervised Learning. *Neural Computation*, *1*(1), 295–311.

Barlow, H., Kaushal, T., und Mitchison, G. (1989). Finding Minimum Entropy Codes. *Neural Computation*, *1*(3), 412–423.

Becker, S. (1992). *An Information-theoretic Unsupervised Learning Algorithm for Neural Networks*. Ph.D. thesis, University of Toronto.

Bell, A., und Sejnowski, T. (1995). An information-maximasation appoach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159. submitted.

Bishop, Chris, M. (1994). Mixture Density Networks. Tech. rep. NCRG/4288, Neural Computing Research Group , Department of Computer Science, Aston University, Birmingham, UK.

Bourlard, H., und Kamp, Y. (1988). Auto-Association by Multilayer Perceptron and Singular Value Decomposition.. *Biological Cybernetics*, *58*, 291–294.

Breakwell, J. (1977). Zero-Sum Differential Games with Piece-Wise Continuous Trajectories. *Lecture Notes of Control and Information Science*, *3*.

Bridle, J. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Neural Information Processing Systems 2*, pp. 211–217. Morgan Kaufmann.

Bridle, J., MacKay, D., und Heading, A. (1991). Unsupervised Classifiers, Mutual Information and 'Phantom Targets'. In *Neural Information Processing Systems 4*, pp. 1096–1101. Morgan Kaufmann.

Burel, G. (1992). Blind Separation of Sources: A Nonlinear Neural Algorithm. *Neural Networks*, *5*, 937–947.

Campbell, F., und Robson, J. (1968). Application of Fourier analysis to the sisibility of gratings. *Journal of Physiology*, *197*, 551–566.

Comon, P. (1994). Independent component analysis, A new concept. *Signal Processing*, *36*, 287–314.

Comon, P., Jutten, C., und Herault, J. (1991). Blind separation of sources, Part II: Problem statement. *Signal Processing*, *24*, 11–20.

Coomans, D., Broeckaert, M., Jonckheer, M., und Massart, D. (1983). Comparison of Multivariate Discriminant Techniques for Clinical Data - Application to the Thyroid Functional State. *Meth. Inform. Med.*, *22*, 93–101.

Cover, Thomas, M., und Thomas, Joy, A. (1991). *Elements of information theory.* Wiley.

Deco, G., und Brauer, W. (1994). Higher Order Statistical Decorrelation by Volume Concerving Nonlinear Maps. *Neural Networks*, *?* submitted.

Deco, G., und Parra, L. (1995). Unsupervised Learning for Boltzmann Machines. *Network*, *6*(3), 437–448.

Deco, G., und Schürman, B. (1994). Learning Time Series Evolution by Unsupervised Extraction of Correlations. *Physical Review E*, *?* submitted.

Duda, R., und Hart, P. (1973). *Pattern Classification and Scene Analysis.* Wiley - Interscience.

Feng, K., und Qin, M.-z. (1985). The Symplectic Methods for the Computation of Hamiltonian Equations. In Zhu You-lan, G. B.-y. (Ed.), *Numerical Methods for Partial Differential Equations.* Proceedings of a Conference held in Shanghai, 1987. Lecture Notes in Mathematics. Vol. 1297, pp. 1-35. Springer, Berlin Heidelberg New York.

Fischer, K., H., und Herz, J. A. (1991). *Spin glasses.* Cambridge University Press.

Földiák, P. (1989). Adaptive network for optimal linear feature extraction. In *IEEE/INNS International Joint Conference on Neural Networks*, pp. 401–405.

Gabler, T., Miesbach, S., Breitner, H., und Pesch, H. (1993). Synthesis of Optimal Strategies for Differential Games by Neural Networks. Report 468, Deutsche Forschungsgemeinschaft. Schwerpunktprogramm: Anwendungsbezogene Optimierung und Steuerung.

Geman, S., und Geman, D. (1984). Stochstic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gislén, L., Peterson, C., und Södeberg, B. (1992). Rotor Neurons: Basic Formalism and Dynamics. *Neural Computation*, *4*, 737–745.

Glauber, R. (1963). Time-Dependent Statistics of the Ising Model. *Journal of Mathematical Physics*, *4*, 294–307.

Hastie, T., und Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association*, *84*(406), 502–516.

Hebb, D., O. (1949). *The Organization of Behavior*. Wiley, New York. Partially reprinted in Anderson and Rosenberg 1988.

Hentschel, H., und Barlow, H. (1991). Minimum-entropy coding with Hopfield networks. *Network*, *2*, 135–148.

Herz, J., Krogh, A., und Palmer, R. (1990). *Introduction to The Theory of Neural Computation*. Addison-Wesley.

Hopfield, J. (1982a). Neural Networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences,*, *79*, 2554–2558.

Hopfield, J. (1982b). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences,*, *81*, 3088–3092.

Hopfield, J. (1987). Learning Algorithms and Probability Distribution in Feed-Forward and Feed-Back Networks. *Proceedings of the National Academy of Sciences,*, *84*, 8429–8433.

Hornik, K., Stinchcombe, M., und White, H. (1989). Multilayer Feedforward Neural Networks are Universal Approximators. *Neural Networks*, *2*, 359–366.

Hubel, D., und Wiesel, T. (1962). Receptive Fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Ioannou, P. A., und Sun, J. (1995). *Stable and robust adaptive control*. Englewood Cliffs, Prentice Hall, NJ.

Jones, J., P., und Palmer, L., A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*, 1187–1211.

Jones, J., P., Stepnoski, A., und Palmer, L., A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*, 1112–1232.

Kappen, Hilbert, J. (1994). Deterministic learning rules for Boltzmann Machines. *Neural Networks*, *?*, ? to appear.

Karhunen, J., und Joutsensalo, J. (1994). Representation and Separation of Signal Using Nonlinear PCA Type Learning. *Neural Network*, *7*(1), 113–127.

Kendall, M. G., und Stuart, A. (1969). *The Advanced Theory of Statistics* (3 edition)., Vol. 1. Charles Griffin and Company Limited, London.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, *43*, 59–69.

Kosmatopoulos, Elias, B., und Christodoulou, Manolis, A. (1993). The Boltzmann g-RHONN: A Learning Maschine for Estimating Unknown Probability Distributions. *Neural Networks*, *7*(2), 271–278.

Kuehnel, H., und Tavan, P. (1990). The anti-Hebb Rule derived from Information Theory. In R. Eckmiller, G. H., und Hauske, G. (Eds.), *Parallel processing in neural systems and computers*, pp. 187–190. North-Holland: Elselvier Science, Amsterdam.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*, 105–.

Linsker, R. (1989). How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, *1*, 402–411.

Linsker, R. (1992). Local Synaptic Learning Rules Suffice to Maximize Mutual Information in a Linear Network. *Neural Computation*, *4*, 691–702.

MacKay, D. (1981). Strife over visual cortical functions. *Nature*, *289*, 117–118.

Metropolis, N., Rosenblut, A., W., Rosenblut, M., N., Teller, A., H., und Teller, E. (1953). Equation of state calculations by fast computing maschines. *Journal of Chemical Physics*, *21*, 1087–1091.

Miesbach, S., und Pesch, H. (1992). Symplectic phase flow approximation for the numerical integration of canonical systems. *Numerical Mathematics*, *61*, 501–521.

Moody, und Darken (1989). Fast Learning in Network of locally-tuned Processing Units. *Neural Computation*, *1*(2).

Movellan, Javier, R. (1993). Covariance Learning Rules for Stocastic Neural Networks. In *International Joint Confonference on Neural Networks*, Vol. 3, pp. 376–381 San Diego.

Movellan, Javier, R., und McClelland, James, L. (1991). Learning Continuous Probability Distributions with the Constrastive Hebbian Algorithm. Tech. rep. NCRG/4288, Parallel Distributed Processing and Cognitive Neuroscience, Department of Psychology, Carnegie Mellon University, Pittsburgh, Pansilvania.

Mozer, M., Zemel, R., Behrmann, M., und Williams, C. (1992). Learning to Segment Images Using Dynamic Feature Binding. *Neural Computation*, *4*(5), 650–665.

Nadal, J.-P., und Parga, N. (1994). Non linear neurons in the low noise limit: a factorial code maximizes information transfer. *Network*, *5*(4), 565–581.

Noest, A. (1988). Associative Memory in Sparse Phasor Neural Networks. *Europhysics Letters*, *6*(6), 469–474.

Obradovic, D., und Deco, G. (1995). Generalized Linear Features Extraction: An Information Theory Approach. *Neural Computation*, *?*, ? submitted.

Oja, E. (1989). Neural Networks, Principal Components, and Subspaces. *International Journal of Neural Systems*, *1*(1), 61–68.

Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes.* Third Edition, McGraw-Hill, New York.

Parra, L., und Deco, G. (1995). Continuous Boltzmann Machine with Rotor Neurons. *Neural Networks*, *8*(3), 375–385.

Parra, L., und Deco, G. (1996). Statistical Independence with Information Preserving Nonlinear Maps. *Neural Computation*, *8*(2), 262–271.

Parra, L., Deco, G., und Miesbach, S. (1995). Redundancy reduction with information-preserving nonlinear maps. *Network*, *6*(1), 61–72.

Parra, L. (1995). Symplectic Nonlinear Component Analysis. In *Neural Information Processing Systems 8*, p. to appear. Morgan Kaufmann.

Peterson, C., und Anderson, J. (1987). A Mean Field Theory Algorithm for Neural Networks. *Complex Systems*, *1*, 995–1019.

Peterson, C., und Hartman, E. (1989). Exploration of the Mean Field Theory Learning Algorithm. *Neural Networks*, *2*, 475–494.

Pineda, F., J. (1987). Generalization of Back-Propagation to recurrent Neural Networks. *Physical Review Letters*, *59*, 2229–2232.

Pollen, D., Lee, J., und Taylor, J. (1971). How does the striate cortex begin the reconstruction of the visual world. *Science*, *173*, 74–77.

Redlich, A. (1993a). Redundancy Reduction as a Strategy for Unsupervised Learning. *Neural Computation*, *5*(2), 289–304.

Redlich, A. (1993b). Supervised Factorial Learning. *Neural Computation*, *5*, 750–766.

Reif (1965). *A mathematical theory of communication.* McGraw-Hill.

Rosenblatt, F. (1962). *Principles of Neurodynamics.* Spartan, New York.

Rubner, J., und Schulten, K. (1990). Development of Feature Detectors by Self-Organization. *Biological Cybernetics*, *62*, 193–199.

Rubner, J., und Tavan, P. (1989). A Self-Organization Network for Principal-Component Analysis. *Europhysics Letters*, *10*, 693–698.

Rumelhart, D., E., Hinton, G., E., und Williams, R., J. (1988). Learning Representations by Back-Propagating Errors. *Nature*, *323*, 533–536.

Schmidhuber, J. (1992). Learning Factorial Codes by Predictability Minimization. *Neural Computation*, *4*(6), 863–879.

Schürmann, B. (1989). Stability and Adaption in Artificial Neural Systems. *Physical Review A*, *40*(5), 2681–2688.

Shannon (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Siegel, C. (1943). Symplectic Geometry. *American Journal Mathematics*, *65*, 1–86.

Silva, F., und L., A. (1990). Speeding up Backpropagation. *Advanced Neural Computers*, *?*, 151–158.

Skohorod, A., V. (1985). *Studies in the theory of random processes.* Addison-Wesley.

Stoer, J., und Bulirsch, R. (1993). *Introduction to Numerical Analysis.* Springer, New York.

Unnikrishnan, K., P., und Venugopal, K., P. (1994). Alopex: A Correlation-Based Learning Algorithm for Feedforward and Recurrent Neural Networks. *Neural Computation*, *6*(3), 469–490.

Yuille, A., L., und Kosowsky, J., J. (1994). Statistical Physics Algorithms That Converge. *Neural Computation*, *6*(3), 341–356.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Cambridge, Mass.

Danksagung

Besonderer Dank gilt meinem Betreuer Gustavo Deco. Er hat mich geleitet, unterstützt und Wissen vermittelt, das über den Rahmen dieser Arbeit hinausgeht.

Mein Dank gilt auch Professor Paul Tavan. Seine Hinweise haben der Arbeit eine klare Ausrichtung gegeben.

Ohne die Unterstützung der Siemens AG wäre diese Arbeit nicht möglich gewesen. Ich bedanke mich in diesem Zusammenhang insbesondere bei Professor Bernd Schurmann, der mir diese unschätzbare Möglichkeit eröffnet hat.

Stefan Miesbach bin ich für die mathematischen Werkzeuge, die er mir nahegebracht hat, zu Dank verpflichtet. Auch die Diskussionen mit Michael Haft haben mir geholfen, Ideen zu überprüfen.

Schließlich danke ich meiner lieben Frau, die bei der Erstellung dieses Manuskripts eine wichtige Rolle gespielt hat und meiner kleinen Tochter Vera, trotz derer diese Arbeit zu Ende geführt werden konnte.

Lebenslauf

Name:	Parra
Vorname:	Lucas, Cristobal
Geburtsdatum:	16 August 1966
Geburtsort:	Tucuman in Argentina
Familienstand:	verheiratet, 1 Kind
Staatsbürgerschaft:	Deutsch, Argentinisch
Sprachen:	Deutsch, Englisch, Spanisch

Schulausbildung:

1972 - 1976	Grundschule in Tucuman, Argentinien
1976 - 1978	Grundschule in München
1978 - 1982	Realschule in München
Juli 1982	Mittlere Reife
1982 - 1984	Fachoberschule München
Juli 1984	Fachabitur

Studium:

1984 - 1985	Physik an der Fachhochschule München
Juli 1985	Vordiplom
1985 - 1992	Physik mit Nebenfach Informatik an der Ludwig-Maximilian-Universität München
Februar 1992	Hauptdiplom

Doktorarbeit:
April 1992 - März 95: Doktorand der Gruppe für Neuronale Netze von Professor Schürmann in der Zentralen Forschung und Entwicklung der Siemens AG unter Aufsicht von Dr. Gustavo Deco.
Februar 94 - März 95: Gast im Learning-Department der Siemens Corporate Research, Princeton, USA.
Die Arbeit wurde von Professor Paul Tavan der Medizinischen Optik in der Sektion Physik der Ludwig-Maximilian-Universität betreut.