Speicherkapazitäten verdünnter neuronaler Netzwerke

Inaugural–Dissertation
zur
Erlangung des Doktorgrades
der Naturwissenschaftlichen Fachbereiche
der Justus–Liebig–Universität Gießen
— Fachbereich Physik —

vorgelegt von

Peter Kuhlmann

aus Ehringshausen

Gießen, im Mai 1993

D26

Dekan: Prof. Dr. Arthur Scharmann

I. Berichterstatter: Prof. Dr. Wolfgang Kinzel

II. Berichterstatter: Prof. Dr. Werner Scheid

Inhaltsverzeichnis

1 Die lineare Verdünnung des Perzeptrons optimaler Stabilität 4

1.1 Einleitung: Neuronale Netzwerke 4
1.2 Das Modell 6
1.2.1 Das verallgemeinerte Ising–Modell 6
1.2.2 Attraktornetzwerke 7
1.2.3 Einschichtnetzwerke 10
1.3 Das Perzeptron optimaler Stabilität 11
1.3.1 Zufallsmuster und Zufallsausgaben 12
1.3.2 Die maximale Speicherkapazität des Perzeptrons optimaler Stabilität 12
1.3.3 Die Verallgemeinerungsfähigkeit des Perzeptrons optimaler Stabilität 15
1.4 Die lineare Verdünnung des Perzeptrons optimaler Stabilität . . . 17
1.5 Inhaltsangabe 20

2 Ein neues Verfahren zur Berechnung der Speicherkapazität des Perzeptrons optimaler Stabilität 21

2.1 Der Rechenansatz 21
2.1.1 Die Komplementaritätsbedingung 21
2.1.2 Die Zahl der Kuhn–Tucker–Lösungen 24
2.2 Die Berechnung der gemittelten Zahl der Lösungen 24
2.2.1 Die Jacobi–Determinante 25
2.2.2 Die Mittelung über die Muster 25
2.2.3 Die Entkopplung der Variablen 26
2.2.4 Das Ergebnis für f 27
2.2.5 Die Lösung der Sattelpunktgleichungen 28
2.3 Ergebnisse 28
2.4 Lernalgorithmen 29

3 Die optimale Verdünnung des Perzeptrons 31

3.1 Die obere Schranke für die maximale Speicherkapazität 31
3.2 Der Ansatz zur Berechnung der optimalen Speicherkapazität . . 34
3.2.1 Ausgeglühte und eingefrorene Verdünnung 35
3.2.2 Das allgemeine Ergebnis für natürliche n 35
3.3 Die replikasymmetrische Näherung 36

3.4 Die Replika–Symmetriebrechung erster Stufe für die optimale Verdünnung . . . 40
3.4.1 Das SK–Modell und der Parisi–Ansatz . . . 40
3.4.2 Die Anwendung der RSB1 auf die optimale Verdünnung . 43
3.4.3 Der Ansatz im Fall $q_1 \to 1$ . . . 45
3.5 Die Ergebnisse der RSB1–Näherungsrechnung . . . 46

4 Der Quersummenverdünnungsalgorithmus 51
4.1 Die Definition des Algorithmus . . . 51
4.1.1 Motivation . . . 51
4.1.2 Der Algorithmus . . . 52
4.2 Die analytische Rechnung . . . 53
4.2.1 f als Funktion von w . . . 53
4.2.2 Der Ansatz für die Gardner–Rechnung und die Mittelung über die Muster . . . 54
4.2.3 Das Ergebnis für allgemeine n . . . 55
4.2.4 Die Annahme der Replika–Symmetrie und die Lösung der Sattelpunktgleichungen . . . 57
4.2.5 Das Ergebnis am kritischen Punkt . . . 58
4.2.6 Der Limes $f \to 0$ . . . 59
4.3 Ergebnisse . . . 60

5 Das Schneideverfahren zur Verdünnung des Perzeptrons optimaler Stabilität 64
5.1 Der Algorithmus . . . 64
5.2 Die Gardner–Rechnung zum Einschrittverfahren . . . 66
5.2.1 f als Funktion von w . . . 66
5.2.2 Der Ansatz für die Gardner–Rechnung . . . 68
5.2.3 Der Rechenweg zur Gewinnung des allgemeinen Ergebnisses für $\Phi_R(m,n)$ . . . 70
5.2.4 Die Annahme der Replika–Symmetrie . . . 71
5.2.5 Die Limites $p \to 1$ und $q \to 1$ . . . 74
5.2.6 Die Ergebnisse der analytischen Rechnung . . . 76
5.3 Ergebnisse . . . 77
5.3.1 Die Überprüfung der analytischen Ergebnisse durch die Computersimulation . . . 77
5.3.2 Die Ergebnisse für α_{eff} bei Einschritt– und Mehrschrittalgorithmus . . . 80
5.3.3 Die Verteilung der Kopplungen bei Mehrschrittalgorithmus und optimaler Verdünnung . . . 83

6 Die Verallgemeinerungsrate eines verdünnten Perzeptrons 85
6.1 Problemstellung . . . 85
6.2 Die allgemeine Verdünnungsformel . . . 87
6.3 f_c als Funktion von f_l, α und B_0 . . . 89
6.4 Die Berechnung der Verallgemeinerungsrate . . . 91
6.4.1 Die Rechnung bis zum allgemeinen Ergebnis . . . 91

6.4.2 Die replikasymmetrische Annahme 93
6.5 Ergebnisse 95

7 Zusammenfassung und Ausblick 104
7.1 Zusammenfassung 104
7.2 Ausblick 105
7.2.1 Ergänzungen technischer Natur 105
7.2.2 Verdünnte Mehrschichtnetzwerke 106

A Häufig verwendete Formeln 108
A.1 δ–Funktion, θ–Funktion und Kronecker–δ 108
A.2 Gauß–Integrale 109
A.3 Die Φ–Funktion 109
A.4 Die allgemeine Gauß–Formel 110

B Eine Formel zur Replika–Symmetriebrechung erster Stufe 112

C Die Sattelpunktgleichungen und die Kopplungsverteilung zur RSB1 des optimal verdünnten Perzeptrons 114
C.1 Die Durchführung des Limes $q_1 \to 1$ in den Sattelpunktgleichungen 114
C.2 Das Endergebnis für die RSB1 der optimalen Verdünnung des Perzeptrons 116
C.3 Die Wahrscheinlichkeitsdichte der Kopplungen in RSB1–Näherung 119

D Die Sattelpunktgleichungen zur Bestimmung der Verallgemeinerungsrate des Quersummenalgorithmus 121

Literaturverzeichnis 123

Kapitel 1

Die lineare Verdünnung des Perzeptrons optimaler Stabilität

1.1 Einleitung: Neuronale Netzwerke

Neuronale Netzwerke haben sich in den letzten Jahren zum Gegenstand der Forschung und Anwendung entwickelt, weil sie neuartige Lösungsansätze für verschiedenste Probleme aufzeigen. Ihr Hauptmerkmal ist eine der Biologie entlehnte Struktur aus Schaltelementen (Neuronen) und Verbindungen (Synapsen oder Kopplungen). Ihre Vorzüge sind vor allem Fehlertoleranz und adaptive Belernbarkeit, die auch bei biologischen Systemen auftreten. Die wichtigsten Anwendungsgebiete neuronaler Netze sind die Sprachverarbeitung [WaLe90] und die industrielle Bildverarbeitung [Sch+90], [Re+82], [HM92], [Ct92], [IEEE92], bei denen oft nach umfangreicher klassischer Vorverarbeitung ein neuronales Netz die abschließende Klassifikation vornimmt. Zur Erfüllung dieser Aufgabe werden meist komplizierte Netzwerke verwandt, die aus mehreren Schichten von Neuronen bestehen und die nach dem Backpropagation–Verfahren belernt werden [Ru86].

Die Probleme dieser Netzwerke sind:

1. Bei großen Netzwerken [FI91] treten große Lernzeiten auf.

2. Es fehlt ein Algorithmus, der die notwendige Zahl von Zwischenschichtneuronen ermittelt.

3. Das Problem der Verallgemeinerung ist nicht geklärt, d.h. die Frage wie das neuronale Netzwerk auf neue, noch nicht gelernte Eingaben („Muster“) reagiert, ist nicht beantwortet.

Die in der statistischen Physik behandelten neuronalen Netze offenbaren hingegen einen anderen Zugang zur Lösung komplizierter Klassifikationsprobleme. Die zugrunde liegende Idee ist dabei, ein Mehrschichtnetzwerk aus vollständig erforschten Einschichtnetzwerken aufzubauen. Die Konstruktion des Mehrschicht-

netzwerkes erfolgt kachelförmig [1], d.h. es werden Schritt für Schritt Zwischenschichtneuronen hinzugefügt, bis das Klassifikationsproblem gerade gelöst ist. Es werden also nur die unbedingt benötigten Zwischenschichtneuronen belernt (Problem 2) [MeNa89], [BiOp91], [Bi91], [Ru90]. Als Grundbestandteil wird das Perzeptron optimaler Stabilität [Ga88] vorgeschlagen. Für dieses Einschichtnetzwerk sind die verbleibenden oben angesprochenen Probleme 1 und 3 geklärt: Zum einen existieren schnelle Lernalgorithmen ([KrMe87], [AnBi90], [Ru91]), zum anderen besitzt das Perzeptron optimaler Stabilität eine hohe Fähigkeit zum Verallgemeinern ([Op+90], [Ne91], [Wa+92]). Parallel zur Entwicklung praktikabler Mehrschichtnetzwerke werden einfache Modelle derselben theoretisch untersucht. Dabei wird sowohl das Problem der maximalen Speicherkapazität [En+92], [BiOp91] als auch das Problem der Verallgemeinerung mit analytischen Rechnungen behandelt [GyTi90], [Wa+92], [Sw91], [Sn92].

Nachdem die Frage nach der minimalen Zahl von Zwischenschichtneuronen in einem Mehrschichtnetzwerk geklärt ist, bleibt die Aufgabe, die einzelnen Einschichtnetzwerke möglichst klein zu machen. Es muß also ein Verfahren gefunden werden, mit dem man einen gewissen Prozentsatz der Kopplungen eines Einschichtnetzwerkes einsparen kann. Die Vorteile dieser Verdünnung des Netzwerkes liegen in einer Beschleunigung des Lern– als auch des Klassifikationsprozesses und in einem geringen Speicherbedarf. Außerdem eröffnet die Verdünnung die Möglichkeit, unwichtige Komponenten der Eingangsmuster zu erkennen, indem man darauf achtet, an welchen Komponenten die Kopplungen des neuronalen Netzwerkes entfernt worden sind. Ein solches verdünntes Mehrschichtnetzwerk kann also eine Auskunft darüber geben, welche Komponenten der Vorverarbeitung von Bedeutung sind.

Zum Verständnis der statistischen Mechanik verdünnter Mehrschichtnetzwerke ist ein ähnliches Vorgehen notwendig, wie es oben beschrieben wurde. Zunächst muß die statistische Mechanik verdünnter Einschichnetzwerke behandelt werden. Dies soll in der vorliegenden Arbeit geschehen. Die Konstruktion verdünnter Mehrschichtnetzwerke bleibt der zukünftigen Forschung vorbehalten. Neben einer naiven Verallgemeinerung der obigen kachelartigen Verfahren ([MeNa89], [BiOp91], [Bi91]) wird vor allem Gegenstand der Forschung sein, wie die Verdünnung der verschiedenen Einschichtnetzwerke aufeinander abgestimmt werden kann. Das bedeutet, daß wohl ein Mehrschichtnetzwerk aus unterschiedlich verdünnten Bestandteilen zu erwarten sein wird.

In den verbleibenden Abschnitten des ersten Kapitels wird das zu behandelnde Einschichtnetzwerk, das Perzeptron optimaler Stabilität, vorgestellt. Es wird herausgestellt, daß die statistische Mechanik eine analytische Behandlung dieses neuronalen Netzes erlaubt, wenn zwei Vorbedingungen erfüllt sind: Zum einen müssen die Muster durch einen wohldefinierten Zufallsprozeß ermittelt werden, zum anderen muß der Grenzwert der Zahl der Neuronen $N \to \infty$ betrachtet werden. Bevor schließlich in einer Inhaltsangabe ein Überblick über die vorliegende Arbeit gegeben wird, werden die Arten der Verdünnung des Perzeptrons vorgestellt.

[1] Ins Englische übersetzt heißt Kachel „Tiling“.

1.2 Das Modell

In diesem Abschnitt wird beschrieben, wie sich neuronale Einschichtnetzwerke in die statistische Physik einordnen. Dazu werden zunächst in zwei Unterpunkten die wichtigsten Modelle der statistischen Physik neuronaler Netze vorgestellt.

1.2.1 Das verallgemeinerte Ising–Modell

Das verallgemeinerte Ising–Modell ist ein System aus N Spins

$$S_i \in \{-1, +1\}$$

mit reellen Kopplungen J_{ij}, die nicht notwendigerweise symmetrisch sein müssen ([Is25], [Am89a], siehe auch [Ku90]). Mit der Glauber–Dynamik [Gl63] steht eine Zeitentwicklung für das Spinsystem bereit, die für jeden Zeitschritt eines diskreten Prozesses die Wahrscheinlichkeit

$$W(\underline{S}(t + \Delta t) \leftarrow \underline{S}(t))$$

für den Übergang vom Zustand

$$\underline{S}(t) = (S_1(t), \ldots, S_N(t))^T$$

zum Zustand $\underline{S}(t + \Delta t)$ angibt. Die Glauber–Dynamik kann seriell oder parallel definiert werden (siehe [Am89b], [Pe84], [Ku90]). Im folgenden wird die serielle Version vorgestellt.

In jedem Zeitschritt wird ein einzelner Spin $S_i \in \{-1, +1\}$ seriell oder zufällig ausgewählt. Er wird nach der Wahrscheinlichkeit

$$W(\underline{S}(t + \Delta t) \leftarrow \underline{S}(t)) = \frac{\exp(\beta S_i(t + \Delta t) h_i(t))}{\exp(\beta h_i(t)) + \exp(-\beta h_i(t))} \tag{1.1}$$

in $S_i(t + \Delta t) \in \{-1, +1\}$ umgewandelt. $h_i(t)$ ist dabei das innere Feld am Platz i:

$$h_i(t) = \sum_{j=1}^{N} J_{ij} S_j(t) \tag{1.2}$$

$\beta = 1/T$ ist der Kehrwert der Temperatur T eines Wärmebades [Am89b]. Der Zusammenhang mit der Temperatur in der statistischen Mechanik [Huang], [Reif] wird klar, wenn man die Frage nach einer Gleichgewichtsverteilung stellt: Für sehr große Zeiten (idealisiert $t \to \infty$) ist nämlich die Wahrscheinlichkeit, daß ein Zustand $\underline{S}$ angelaufen wird, durch die Gleichgewichtsverteilung

$$P^{eq}(\underline{S}) = \frac{\exp(-\beta E(\underline{S}))}{Z} \tag{1.3}$$

gegeben. Dies gilt, wenn J_{ij} symmetrisch ist und keine Selbstkopplungen vorliegen ($J_{ii} = 0 \ \forall i$) [Am89b], [Pe84]. $E(\underline{S})$ ist dabei die Energiefunktion

$$E(\underline{S}) = -\frac{1}{2} \sum_{i,j(i \neq j)} J_{ij} S_i S_j \tag{1.4}$$

Der Normierungsfaktor Z ist die Zustandssumme

$$Z = \sum_{\underline{S}} \exp\left(-\beta E(\underline{S})\right) \tag{1.5}$$

P^{eq} ist die wohlbekannte Boltzmann–Verteilung für die Wahrscheinlichkeit eines Zustandes $\underline{S}$, den ein System im Wärmebad der Temperatur T annimmt. Die obige Glauber–*Dynamik* liefert also die aus rein statischen Überlegungen gewonnene Boltzmann–Verteilung für Zustände in einem kanonischen Ensemble. Dies eröffnet zwei Möglichkeiten, die Gleichgewichtseigenschaften verallgemeinerter Ising–Modelle (mit symmetrischen Kopplungen und $J_{ii} = 0\ \forall i$) zu untersuchen. Zum einen kann man den dynamischen Prozeß aus Gl.(1.1) auf dem Rechner simulieren („Monte–Carlo–Simulation"), zum anderen kann man in einer Rechnung die lokalen und globalen Minima der statistischen Größe freie Energie pro Spin

$$f = -\frac{T}{N} \ln Z \tag{1.6}$$

berechnen. Dies ist analytisch nur für unendlich große Systeme („thermodynamischer Limes $N \to \infty$") möglich. Führt man des weiteren den Grenzübergang $T \to 0$ durch, so erhält man die lokalen Minima und die globalen Minima (Grundzustände) der Energie:

$$f \to \frac{E}{N} \quad (T \to 0)$$

In einer Computersimulation ist dieser Grenzübergang $T \to 0$ oft sehr aufwendig. Die Bereitstellung geeigneter Techniken zur Durchführung des Grenzübergangs sind Aufgabe des Forschungsgebietes „simulated annealing" [Me+87], [NR88].

1.2.2 Attraktornetzwerke

Ein neuronales Attraktornetzwerk ist ein verallgemeinertes Ising–Modell, bei dem die Kopplungen J_{ij} so gewählt sind, daß p vorgegebene Muster

$$\xi_i^\mu \in \{-1, +1\},\ i = 1, \ldots, N,\ \mu = 1, \ldots, p \tag{1.7}$$

thermodynamisch stabil werden. Diese Muster können dann von dem System mit Hilfe des oben geschilderten dynamischen Glauber–Prozesses wiedererkannt werden. Die Spins heißen hier Neuronen, die Kopplungen heißen Synapsen. Es ist gezeigt worden, daß J_{ij} für große Systeme gefunden werden kann, wenn die Zahl der Muster proportional zur Zahl der Neuronen ist:

$$p = \alpha N \tag{1.8}$$

wobei die Proportionalitätskonstante α Speicherkapazität genannt wird (siehe Amit et al. in [Am+87]).

Amit et al. untersuchten die sogenannten Hebb–Kopplungen [He49]

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \tag{1.9}$$

und fanden heraus, daß unterhalb einer kritischen Speicherkapazität $\alpha_c \sim 0.14$ Zustände in der Nähe der Muster stabile Grundzustände sind. Voraussetzung für die analytische Berechnung der freien Energie pro Spin f war die zufällige Auswahl der Muster ($p(\xi_i^\mu = \pm 1) = \frac{1}{2}$) und der thermodynamische Limes $N \to \infty$. Die Hebb–Kopplungen waren schon zuvor von Hopfield mit Monte–Carlo–Simulationen untersucht worden [Ho82]. Seine Veröffentlichung war der Ursprung des Forschungszweiges „Statistische Physik neuronaler Netze".

Im Laufe der Zeit sind weitere Modelle von Attraktornetzwerken untersucht worden. Zunächst ist das neuronale Netzwerk mit Projektorkopplungen anzusprechen [Ko88], [Pe+85], [Pe+86]. Sind die Muster zufällig und unabhängig voneinander ausgewählt mit $p(\xi_i^\mu = \pm 1) = \frac{1}{2}$, so ist für $\alpha < 1$ die Korrelationsmatrix

$$C_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu \tag{1.10}$$

im thermodynamischen Limes $N \to \infty$ positiv definit. Folglich existiert die Kopplungsmatrix

$$J_{ij} = \frac{1}{N} \sum_{\mu,\nu} \xi_i^\mu (C_{\mu\nu})^{-1} \xi_j^\nu \tag{1.11}$$

für $\alpha < 1$. J_{ij} ist die Projektormatrix in den Raum der Muster: Betrachtet man die Zerlegung eines Zustandes

$$\underline{S} = \sum_{\mu=1}^{p} a_\mu \underline{\xi}^\mu + \underline{\delta S} \tag{1.12}$$

wobei $\underline{\delta S}$ orthogonal zu allen Mustern ist, so gilt

$$\sum_{j=1}^{N} J_{ij} S_j = \sum_{\mu=1}^{p} a_\mu \xi_i^\mu \quad \forall i \tag{1.13}$$

Die thermodynamischen Eigenschaften dieses neuronalen Netzwerks sind von Kanter und Sompolinsky untersucht worden [KaSo87]. Sie haben die Energiefunktion

$$E(\underline{S}) = -\frac{1}{2} \sum_{i,j(i \neq j)} J_{ij} S_i S_j \tag{1.14}$$

betrachtet und die freie Energie pro Spin f im thermodynamischen Limes $N \to \infty$ berechnet. Die Durchführung des Grenzübergangs $T \to 0$ zeigte dann, daß die Muster die stabilen Grundzustände sind, falls $\alpha < 1$. Möchte man dieses Ergebnis in einem Prozeß des „simulated annealing" bestätigen, so ist zu beachten, daß die oben ebenfalls definierten Selbstkopplungen J_{ii} während der Monte–Carlo–Simulation nicht in die inneren Felder eingehen. Die inneren Felder in Gl.(1.2) sind also

$$h_i(t) = \sum_{j(\neq i)} J_{ij} S_j(t) \tag{1.15}$$

Ist man schließlich bei $T = 0$ in einem lokalen oder globalen Minimum der Energie angelangt, so wird die Glauber–Dynamik deterministisch. Aus Gl.(1.1) ergibt sich nämlich im Limes $T \to 0$

$$S_i(t + \Delta t) = \operatorname{sign} h_i(t) = \operatorname{sign} \left(\sum_{j(\neq i)} J_{ij} S_j(t) \right) \tag{1.16}$$

Endzustände dieser Dynamik erfüllen generell die Bedingung

$$S_i = \operatorname{sign}(h_i) \quad \forall i \tag{1.17}$$

Man nennt solche Zustände im allgemeinen metastabile Zustände [2].

Die lokalen Energien

$$\lambda_i := S_i h_i \tag{1.18}$$

sind insbesondere für den *Grundzustand* an allen Plätzen i positiv

$$\lambda_i > 0 \quad \forall i \tag{1.19}$$

und stellen Maße für die Stabilität des Grundzustandes gegen Spinflips dar. Die Projektormatrix ist gerade so konstruiert, daß für die lokalen Energien der Muster

$$\begin{aligned} \lambda_i^\mu &= \xi_i^\mu \left(\sum_{j=1}^{N} J_{ij} \xi_j^\mu - J_{ii} \xi_i^\mu \right) \\ &= (1 - J_{ii}) \to 1 - \alpha \quad (N \to \infty) \end{aligned} \tag{1.20}$$

gilt (siehe [KaSo87], [Ku90]). Dies ist ein Hinweis darauf, daß die Muster für alle $\alpha < 1$ Grundzustände sind.

Nachdem man bereits festgestellt hat, daß die Projektormatrix die Muster für $\alpha < 1$ stabilisiert, kann man sich die Frage nach einer optimalen Kopplungsmatrix stellen. Wenn man auf die Forderung nach einer symmetrischen Kopplungsmatrix verzichtet, kann man das Optimierungsproblem für jede Zeile i getrennt formulieren:

Maximiere

$$\lambda_i^\mu = \xi_i^\mu \sum_{j(\neq i)} J_{ij} \xi_j^\mu \tag{1.21}$$

bei festgehaltener Norm $\sum_{j(\neq i)} {J_{ij}}^2$. Die einzelnen Zeilen der so bestimmten Kopplungsmatrix des Attraktornetzwerkes sind die Kopplungsvektoren des „Perzetrons optimaler Stabilität“. Auf die Bestimmung seiner maximalen Speicherkapazität $\alpha = 2$ wird unten eingegangen (Abschnitt 1.3).

[2] Viele dieser metastabilen Zustände werden bei Spinglasmodellen als unerwünschte Endzustände angelaufen, wenn man statt des korrekten „simulated annealing“ die deterministische Dynamik (1.16) verwendet [FiHe91], [BrMo80], [Ga86]. Metastabile Zustände sind also im allgemeinen Endzustände eines schockartigen Einfrierprozesses, wohingegen globale Minima der Energie Endzustände eines „simulated annealing“–Prozesses sind.

1.2.3 Einschichtnetzwerke

Die Stabilitätsbedingung

$$\lambda_i^\mu = \xi_i^\mu \sum_{j(\neq i)} J_{ij} \xi_j^\mu > 0 \quad \forall i, \mu \tag{1.22}$$

beim obigen Attraktornetzwerk führt auf die Frage nach der Konstruktion eines geeigneten Kopplungsvektors in jeder Zeile i der Kopplungsmatrix. Setzt man $\xi_i^\mu = S^\mu$, wobei S^μ als eine vorgegebene Ausgabe betrachtet werden kann, so lautet die Aufgabe:

Gegeben seien $p = \alpha N$ Muster–Vektoren

$$\underline{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)^T, \quad \mu = 1, \ldots, p \tag{1.23}$$

Zu jedem Vektor gehört eine binäre Ausgabe $S^\mu \in \{-1, +1\}$ [3]. Gesucht ist ein Kopplungsvektor

$$\underline{J} = (J_1, \ldots, J_N)^T, \tag{1.24}$$

so daß

$$S^\mu = \text{sign}\left(\sum_{j=1}^N J_j \xi_j^\mu\right) \forall \mu \iff S^\mu \sum_{j=1}^N J_j \xi_j^\mu > 0 \ \forall \mu \tag{1.25}$$

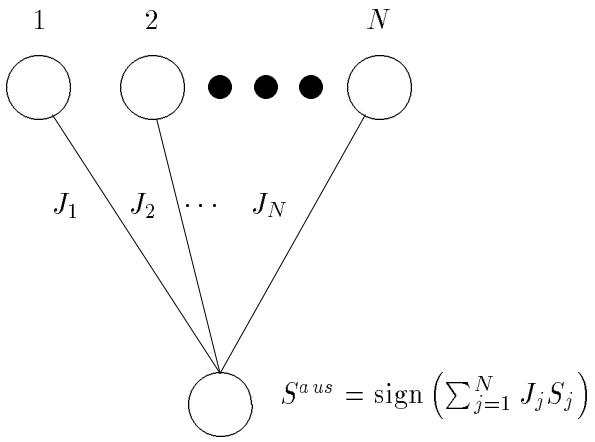
Falls ein solcher Kopplungsvektor $\underline{J}$ existiert, wird er Perzeptron–Vektor genannt. Das zugehörige Netzwerk heißt Perzeptron. $\underline{J}$ ist Normalenvektor einer Hyperebene durch den Ursprung, die die Muster in zwei Klassen separiert. Der Mustersatz heißt dann linear separabel. Wenn ein Perzeptron existiert, so besagt das Perzeptron–Konvergenz–Theorem, daß man den Perzeptron–Vektor mit einem Algorithmus in endlich vielen Schritten lernen kann [Ri+91].

In der bisherigen Darstellung wurde das Perzeptron im Zusammenhang mit der statistischen Physik von Attraktornetzwerken eingeführt. Die obige Formulierung des Problems eröffnet jedoch eine andere Sichtweise, die in der vorliegenden Arbeit vorherrschen wird. Man kann nämlich das Perzeptron als Einschichtnetzwerk auffassen. In einer Eingabeschicht wird ein reeller Zustand $\underline{S}$ präsentiert. Das Perzeptron führt die Operation

$$S^{aus} = \text{sign}\left(\sum_{j=1}^N J_j S_j\right)$$

aus und zeigt in der Ausgabeschicht das binäre S^{aus} an.

[3] Man beachte, daß die Muster nicht mehr notwendigerweise binär sein müssen



Es liegt ein einfaches Netzwerk mit vorwärtsgerichteter Informationsverarbeitung vor („feedforward–network"), das Zustände in zwei Klassen einteilt.

Rosenblatt, der Erfinder des Perzeptrons, beabsichtigte bereits einfache Hardware–Versionen von Perzeptronen für Klassifikationsaufgaben zu verwenden [Ro58], [HeNi91]. Nach Erscheinen des Buches von Minsky und Papert im Jahre 1969 [MiPa69] war jedoch die gängige Meinung, daß Perzeptrone in der Praxis nicht brauchbar seien, um nicht linear separable Mustersätze zu klassifizieren. Im Zuge der Erforschung der statistischen Physik neuronaler Netze sind hingegen Mehrschichtnetzwerke aus Perzeptronen entwickelt worden, die nicht linear separable Probleme lernen können [MeNa89], [BiOp91], [Bi91]. Wie in Abschnitt 1.1 beschrieben, lohnt es sich also, den Baustein Perzeptron gut zu verstehen.

1.3 Das Perzeptron optimaler Stabilität

Existiert ein Perzeptron–Vektor $\underline{J}$ für einen vorgegebenen Satz von Mustern $\underline{\xi}^\mu$ und Ausgaben S^μ, so definiert

$$\kappa = \frac{\min_\mu \left\{ \frac{S^\mu}{\sqrt{N}} \sum_{j=1}^N J_j \xi_j^\mu \right\}}{\sqrt{\frac{1}{N} \sum_{j=1}^N {J_j}^2}} > 0 \tag{1.26}$$

die Stabilität des Perzeptrons. κ ist ein Maß für die Fehlertoleranz des Einschichtnetzwerkes. Die Definition von κ erlaubt die Formulierung eines Optimierungsproblems:

Finde das Perzeptron optimaler Stabilität, d.h. finde einen Perzeptron–Vektor mit minimalem Quadrat der Norm

$$Q = \frac{1}{N} \sum_{j=1}^N {J_j}^2 \tag{1.27}$$

unter Erfüllung der p Zwangsbedingungen

$$\frac{1}{\sqrt{N}} S^\mu \sum_{j=1}^{N} J_j \xi_j^\mu \geq c > 0 \ \forall \mu \tag{1.28}$$

c ist hier eine positive Konstante. Setzt man $c = 1$, so errechnet sich die Stabilität aus Gl.(1.26) zu

$$\kappa = \frac{1}{\sqrt{Q}} \tag{1.29}$$

Die Minimierung des quadratischen Ausdrucks Q unter linearen Nebenbedingungen stellt ein quadratisches Optimierungsproblem dar. Da die zu minimierende Funktion Q streng konvex ist, ist die Lösung der Optimierungsaufgabe eindeutig [Fl87]. Es existiert also nur ein Perzeptron optimaler Stabilität.

1.3.1 Zufallsmuster und Zufallsausgaben

Um das Problem im Zusammenhang mit der statistischen Mechanik behandeln zu können, nehmen wir im folgenden an, daß die ξ_i^μ unabhängig voneinander zufällig ausgewürfelt sind mit Mittelwert $\langle \xi_i^\mu \rangle = 0$ und $\left\langle (\xi_i^\mu)^2 \right\rangle = 1$. Da später der thermodynamische Limes $N \to \infty$ gebildet wird, können wir neben gleichverteilten binären Mustern $\xi_i^\mu = \pm 1$ auch unabhängige gaußverteilte Muster mit Wahrscheinlichkeitsdichte

$$p(\xi_i^\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\xi_i^\mu)^2\right) \ \forall i, \mu \tag{1.30}$$

betrachten. Für analytische Rechnungen gilt, daß binäre Muster wie gaußverteilte Muster zu behandeln sind [HeKu91]. Möchte man die Ergebnisse mit Simulationen überprüfen, treten bei binären Mustern oft größere finite size–Effekte auf. In dieser Arbeit werden deshalb stets gaußverteilte Muster angenommen. Die zu lernenden binären Ausgaben S^μ seien gleichverteilt mit Wahrscheinlichkeit $p(S^\mu = \pm 1) = \frac{1}{2}$. Dann sind die *transformierten Muster*

$$\sigma_i^\mu := S^\mu \xi_i^\mu \tag{1.31}$$

gaußverteilt (mit Mittelwert 0 und Streuung 1). Die Mittelwerte über die Verteilung der σ_i^μ schreiben wir im folgenden als

$$\langle\langle f(\{\sigma_i^\mu\})\rangle\rangle_{\{\sigma_i^\mu\}} = \left(\prod_{i,\mu} \int_{-\infty}^{+\infty} \frac{d\sigma_i^\mu}{\sqrt{2\pi}} \ \exp\left(-\frac{1}{2}(\sigma_i^\mu)^2\right) \right) f(\{\sigma_i^\mu\}) \tag{1.32}$$

1.3.2 Die maximale Speicherkapazität des Perzeptrons optimaler Stabilität

Liegen $p = \alpha N$ Zufallsmuster vor, so stellt sich die Frage, wie stabil sie mit einem Perzeptron gespeichert werden können. Umgekehrt ausgedrückt: Wie groß ist die maximale (oder kritische) Speicherkapazität α_c, bei der noch ein Perzeptron mit Stabilität κ existiert?

Diese Frage ist von E. Gardner im Rahmen eines Ansatzes aus der statistischen Physik beantwortet worden [Ga88]. Die im folgenden skizzierte Rechnung und ähnliche Rechnungen werden deshalb Gardner–Rechnungen genannt. Man faßt die Menge aller normierten Kopplungen $\underline{J}$ als Phasenraum auf und betrachtet den Bruchteil V der Oberfläche dieser „Gardner–Kugel", in dem mit den Nebenbedingungen verträgliche Perzeptron–Vektoren liegen:

$$V = \frac{\left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} dJ_j\right) \delta\left(\sum_{j=1}^{N} {J_j}^2 - N\right) \cdot \prod_{\mu=1}^{p} \theta\left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_j \sigma_j^\mu - \kappa\right)}{\left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} dJ_j\right) \delta\left(\sum_{j=1}^{N} {J_j}^2 - N\right)} \tag{1.33}$$

Die θ–Funktion ist definiert als

$$\theta(x) = \begin{cases} 1 & \text{für} \quad x \geq 0 \\ 0 & \text{für} \quad x < 0 \end{cases}$$

Die δ–Funktion erzeugt die Zwangsbedingung der normierten Kopplungsvektoren.

V hängt noch von der Unordnung, d.h. von den Zufallszahlen σ_j^μ ab. Um eine analytische Rechnung durchführen zu können, muß man geeignet mitteln. Von der statistischen Physik der Spingläser ist bekannt, daß intensive Größen wie Entropie (pro Spin) und freie Energie (pro Spin) für $N \to \infty$ *selbstmittelnd* sind [BiYo86], [EnHe84]. In unserer Formulierung kann man V mit einer Zahl erlaubter Zustände vergleichen und die Entropie

$$s = \frac{1}{N} \ln V \tag{1.34}$$

definieren. Die Selbstmittelungseigenschaft drückt sich dann durch die Gleichung

$$\lim_{N\to\infty} \frac{1}{N} \ln V = \lim_{N\to\infty} \left\langle \left\langle \frac{1}{N} \ln V \right\rangle \right\rangle_{\{\sigma_i^\mu\}} \tag{1.35}$$

aus, wobei der Mittelwert nach Gl.(1.32) zu bilden ist. Gl.(1.35) „ist nichts anderes als das Gesetz der großen Zahlen" [Am89b].

Das Problem der Mittelung über $\ln V$ wird mit Hilfe der Replika–Methode gelöst [HePa79]. Sie basiert auf der Identität

$$\ln V = \lim_{n\to 0} \frac{V^n - 1}{n} \tag{1.36}$$

Auf die erheblichen mathematischen Schwierigkeiten bei der Anwendung der Methode auf die Berechnung der Entropie soll hier nicht eingegangen werden (siehe dazu [HePa79]). Wir geben stattdessen eine Rechenvorschrift an, die Grundlage aller Replika–Rechnungen in der statistischen Physik ist und [HePa79] entstammt:

1. Berechne

$$\Phi_R(n) = \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle V^n \rangle\rangle_{\{\sigma_i^\mu\}} \tag{1.37}$$

   für positives ganzzahliges n.

2. Setze geeignet analytisch nach $n = 0$ fort.

3. Bilde
$$\left.\frac{\partial \Phi_R}{\partial n}\right|_{n=0} \tag{1.38}$$
als Näherung für $\lim_{N\to\infty} \frac{1}{N}\langle\langle \ln V\rangle\rangle$.

Punkt 3 machen wir uns plausibel, indem wir bei endlichen N

$$\Phi_R^{(N)}(n) = \frac{1}{N} \ln \langle\langle V^n \rangle\rangle$$

betrachten

$$\Longrightarrow \left.\frac{\partial \Phi_R^{(N)}}{\partial n}\right|_{n=0} = \left\langle\left\langle \frac{1}{N} \ln V \right\rangle\right\rangle$$

Die richtige analytische Fortsetzung nach $n = 0$ in Schritt 2 hängt vom jeweiligen Problem ab. Im Falle des optimalen Perzeptrons stellt sich heraus, daß die „replikasymmetrische" Fortsetzung die korrekte Entropie liefert [GaDe88]. E. Gardner erhält das Ergebnis

$$\alpha_c(\kappa) = \left[\left(1+\kappa^2\right)\Phi(\kappa) + \frac{\kappa}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\kappa^2\right)\right]^{-1} \tag{1.39}$$

wobei die Φ–Funktion als

$$\Phi(x) = \int_{-\infty}^{x} \frac{d\lambda}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\lambda^2}$$

definiert ist [4] (siehe Gl.(A.11)).

Für $\kappa = 0$ erhält man $\alpha_c(\kappa = 0) = 2$, dies ist die maximale Speicherkapazität, unterhalb der für die betrachteten Zufallsmuster ein Perzeptron gefunden werden kann. Größere Anzahlen von Zufallsmustern sind nicht mehr linear separabel. Dieses Ergebnis ist schon früher durch Abzählen der erlaubten Ausgabevektoren

$$\vec{S} = \left(S^1, \ldots, S^p\right)^T$$

gefunden worden [Wi63], [Co65], [MiDu89]. T. Cover beweist $\alpha_c = 2$ unter der Voraussetzung der allgemeinen Lage („general position") der Muster.

Definition:
p Muster
$$\underline{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)^T,\ \mu = 1, \ldots, p$$
heißen in allgemeiner Lage („general position"), wenn jede N–elementige Untermenge der p Mustervektoren linear unabhängig ist.

Zieht man bei **endlichen** N und p die Zufallsmuster gemäß einer **kontinuierlichen** Verteilung, etwa einer Gauß–Verteilung, so ist die Voraussetzung der allgemeinen Lage mit Wahrscheinlichkeit 1 erfüllt.

Anhand der Arbeit von T. Cover formuliere ich den folgenden Satz:

[4] Man unterscheide die Φ–Funktion stets von der obigen Hilfsgröße $\Phi_R^{(N)}$ der Replika–Rechnung.

Satz:
Vorgelegt seien p Muster

$$\underline{\xi}^{\mu} = (\xi_1^{\mu}, \ldots, \xi_N^{\mu})^T, \ \mu = 1, \ldots, p$$

in allgemeiner Lage. Die Zahl der binären Ausgabevektoren

$$\vec{S} = \left(S^1, \ldots, S^p\right)^T$$

für die die Muster linear separabel sind, ist

$$C(p,N) = \begin{cases} 2^p & \text{für} \quad p \leq N \\ 2\sum_{i=0}^{N-1} \binom{p-1}{i} & \text{für} \quad p > N \end{cases} \tag{1.40}$$

Die Wahrscheinlichkeit, daß ein Mustersatz linear separabel ist, ist

$$W_s = \frac{C(p,N)}{2^p} \tag{1.41}$$

Für eine Folge von Mustern (in allgemeiner Lage), deren Zahl proportional zur Zahl der Komponenten gemäß $p = \alpha N$ (mit konstantem α) wächst, gilt dann

$$\lim_{N\to\infty} W_s = \theta(2-\alpha) \tag{1.42}$$

Im Fall $\alpha > 2$ erhält man mit Hilfe der Stirling–Formel insbesondere

$$\lim_{N\to\infty} \frac{1}{N} \ln W_s = \ln\alpha - (\alpha-1)\ln\left(1 - \frac{1}{\alpha}\right) - \alpha \ln 2 \tag{1.43}$$

Für $\alpha < \alpha_c = 2$ existiert also mit Wahrscheinlichkeit 1 ein Perzeptron. Diese kritische Speicherkapazität des Perzeptrons optimaler Stabilität kann auch mit einem neuen Verfahren ohne Verwendung der Replika–Methode berechnet werden. Dies ist Gegenstand des zweiten Kapitels.

1.3.3 Die Verallgemeinerungsfähigkeit des Perzeptrons optimaler Stabilität

Betrachtet man statt zufälliger Ausgaben S^{μ} von einem Lehrer $\underline{B}$ vorgegebene, so stellt sich die Frage, wie gut das Perzeptron optimaler Stabilität $\underline{J}$ ein neues Muster verallgemeinert. Der Lehrer

$$\underline{B} = (B_1, \ldots, B_N)^T \tag{1.44}$$

erzeugt also die Ausgaben

$$S^{\mu} = \operatorname{sign}\left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} B_j \xi_j^{\mu}\right) \tag{1.45}$$

für eine Folge gaußverteilter Zufallsmuster. Diese Folge linear separabler Zufallsmuster wird von einem Schüler $\underline{J}$ gelernt. Sind $p = \alpha N$ Muster vorgelegt

worden, so versteht man unter der Verallgemeinerungsrate G die Wahrscheinlichkeit, daß die Ausgaben von Lehrer– und Schüler–Perzeptron für ein neues, vom Schüler noch nicht gelerntes Muster übereinstimmen:

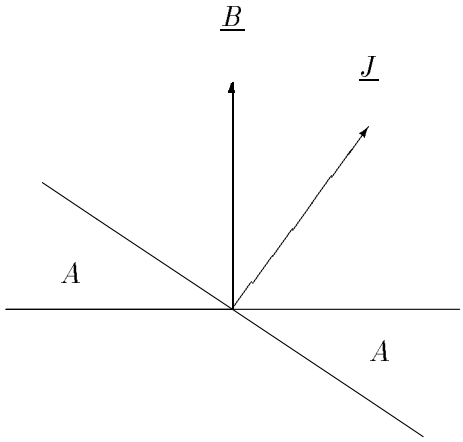
G ist die Wahrscheinlichkeit, daß

$$\text{sign}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j \xi_j^{p+1}\right) = \text{sign}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} B_j \xi_j^{p+1}\right) \tag{1.46}$$

Unter

$$\varepsilon = 1 - G \tag{1.47}$$

versteht man den Verallgemeinerungsfehler. Betrachtet man die aus den beiden Vektoren $\underline{J}$ und $\underline{B}$ gebildete Hyperebene, so kann man G aus einem einfachen geometrischen Argument gewinnen:



Von $\underline{J}$ werden diejenigen Muster falsch klassifiziert, deren Projektionen in der Fläche A im obigen Diagramm liegen. Folglich gilt

$$\varepsilon = \frac{1}{2\pi} \cdot 2\angle(\underline{B}, \underline{J}) = \frac{1}{\pi} \arccos \frac{\underline{B} \cdot \underline{J}}{|\underline{B}| \cdot |\underline{J}|} \tag{1.48}$$

und

$$G = 1 - \frac{1}{\pi} \arccos \frac{\underline{B} \cdot \underline{J}}{|\underline{B}| \cdot |\underline{J}|} \tag{1.49}$$

Dieses Ergebnis kann man auch analytisch gewinnen, indem man

$$G = \left\langle \left\langle \theta\left(\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j \xi_j^{p+1}\right) \cdot \left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} B_j \xi_j^{p+1}\right)\right)\right\rangle\right\rangle_{\{\xi_j^{p+1}\}} \tag{1.50}$$

berechnet, wobei $\langle\langle \ldots \rangle\rangle_{\{\xi_j^{p+1}\}}$ den Mittelwert nach Gl.(1.30) über das neue Muster angibt (siehe [Sw91]).

Um G in Abhängigkeit von α ausrechnen zu können, muß man jedoch wissen, wie der Vektor $\underline{J}$ zu $\underline{B}$ steht, wenn p Muster gelernt sind. Dieses Problem kann mit Hilfe der statistischen Physik gelöst werden ([Op+90], [Ne91], siehe auch den Review [Wa+92]), indem man in einer Gardner–Rechnung das anteilige Phasenraumvolumen analog Gl.(1.33) berechnet.

Die Überlappung

$$R = \frac{\underline{B} \cdot \underline{J}}{|\underline{B}| \cdot |\underline{J}|} \tag{1.51}$$

und damit die Verallgemeinerungsrate G ergeben sich dann als Funktionen von α. Es zeigt sich, daß das Perzeptron optimaler Stabilität ein hervorragendes Verallgemeinerungsverhalten hat.

1.4 Die lineare Verdünnung des Perzeptrons optimaler Stabilität

Wir wenden uns der Frage der Verdünnung des Perzeptrons zu und führen die in dieser Arbeit verwendete Notation ein. Vorgelegt sei ein Satz von gaußverteilten Mustervektoren

$$\underline{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)^T, \quad \mu = 1, \ldots, p,$$

die N Komponenten haben. Es seien p zufällige binäre Ausgaben $S^\mu = \pm 1$ mit $p(S^\mu = \pm 1) = \frac{1}{2}$ vorgegeben. Ist $\alpha < 2$, so existiert (für $N \to \infty$) ein Perzeptronvektor

$$\underline{T} = (T_1, \ldots, T_N)^T$$

mit

$$S^\mu = \text{sign}\left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} T_j \xi_j^\mu\right) \quad \forall \mu$$

Die Frage ist nun, welche der N Eingangsneuronen des Perzeptrons man weglassen kann, so daß der Mustersatz immer noch linear separabel ist. Der auf den verbleibenden Neuronen wirkende Perzeptron–Vektor wird sich dabei in der Regel von dem ausgedünnten ursprünglichen $\underline{T}$–Vektor unterscheiden, ein *Nachlernen* des neuen Perzeptron–Vektors ist nötig.

Aus der Struktur der Cover–Wahrscheinlichkeit W_s in Gl.(1.41) ist klar, daß ein Nachlernen nur dann Erfolg verspricht, wenn die Zahl der verbleibenden Neuronen von der Ordnung der Zahl der Muster, d.h. von der Ordnung N ist [5]. Es wird deshalb angenommen, daß $f \cdot N$ Eingangsneuronen verbleiben. Dies bezeichne ich im folgenden als **lineare Verdünnung**. Um die Existenzbedingungen für das verdünnte Perzeptron korrekt formulieren zu können, wird die folgende Notation eingeführt.

Die Größen

$$c_j \in \{0, 1\}, \quad j = 1, \ldots, N \tag{1.52}$$

[5] Betrachte etwa

$$w = \binom{N}{N^\gamma} / 2^N \Longrightarrow \lim_{N \to \infty} w = 0$$

für $0 < \gamma < 1$.

zeigen an, ob das Neuron j entfernt wurde ($c_j = 0$) oder noch vorhanden ist ($c_j = 1$). Dann gilt

$$f = \frac{1}{N} \sum_{j=1}^{N} c_j \tag{1.53}$$

Bei einem gegebenen f gibt es

$$\binom{N}{Nf} \tag{1.54}$$

Möglichkeiten, einen Verdünnungsvektor

$$\underline{c} = (c_1, \ldots, c_N)^T \tag{1.55}$$

zu bilden. Der Index k numeriere alle verbleibenden Neuronen j so durch, daß $k(j)$ das k–te $c_j = 1$ angibt. Dann geben die

$$\vartheta_k^\mu$$

die Muster ξ_j^μ auf den verbleibenden Plätzen an. Des weiteren seien die

$$\eta_k^\mu = S^\mu \vartheta_k^\mu \tag{1.56}$$

die transformierten Muster auf den verbleibenden Plätzen.

Der Vektor

$$\underline{J}_{Nf} = (J_1, \ldots, J_{Nf})^T \tag{1.57}$$

des verdünnten Perzeptrons muß folglich die Existenzbedingungen

$$S^\mu = \text{sign}\left(\frac{1}{\sqrt{Nf}} \sum_{k=1}^{Nf} J_k \vartheta_k^\mu \right) \quad \forall \mu \tag{1.58}$$

erfüllen.

Die Stabilität des verdünnten Perzeptrons wird analog zu Gl.(1.26) definiert:

$$\kappa = \frac{\min_\mu \left\{ \frac{1}{\sqrt{Nf}} \sum_{k=1}^{Nf} J_k \eta_k^\mu \right\}}{\sqrt{\frac{1}{Nf} \sum_{k=1}^{Nf} {J_k}^2}} \tag{1.59}$$

Das Hauptproblem bei der Verdünnung ist die Suche nach einem geeigneten Verdünnungsvektor $\underline{c}$. Ist dieser gefunden, so kann man auf ihn das Perzeptron optimaler Stabilität lernen.

Ermittelt man den Verdünnungsvektor $\underline{c}$ durch naives Auswürfeln und lernt dann nach, so liegt ein Gardner–Problem mit $p = \frac{\alpha}{f} \cdot fN$ Zufallsmustern auf fN Plätzen vor. Gl.(1.39) lautet dann einfach

$$\frac{\alpha}{f} = \left[\left(1 + \kappa^2\right) \Phi(\kappa) + \frac{\kappa}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \kappa^2 \right) \right]^{-1} \leq 2 \tag{1.60}$$

Offenbar spielt die Zahl der maximal speicherbaren Muster pro Neuron eine entscheidende Rolle. Ich definiere die effektive Speicherkapazität

$$\alpha_{eff} = \frac{\alpha_c(\kappa, f)}{f} \tag{1.61}$$

α_{eff} ist ein Maß dafür, inwiefern unwichtige Komponenten j der Eingangsmuster entfernt wurden. In der vorliegenden Arbeit wird gezeigt, daß effektive Speicherkapazitäten $\alpha_{eff} > 2$ durch intelligente Verdünnungsverfahren erzeugt werden können.

1.5 Inhaltsangabe

Nachdem im ersten Kapitel die wichtigsten Ergebnisse für das Perzeptron optimaler Stabilität vorgestellt wurden und in die Begriffswelt der linearen Verdünnung des Perzeptrons optimaler Stabilität eingeführt wurde, befaßt sich das zweite Kapitel zunächst mit einem neuen Rechenverfahren. Es wird gezeigt, daß die Speicherkapazität des Perzeptrons optimaler Stabilität auch ohne die Replika–Methode berechnet werden kann. Da das Rechenverfahren sehr verwandt mit den wichtigsten Lernalgorithmen des Perzeptrons optimaler Stabilität ist, werden diese kurz vorgestellt.

Im dritten Kapitel wird das Problem der maximalen effektiven Speicherkapazität α_{eff} des optimalen (aber nicht praktikablen) Verdünnungsverfahrens bearbeitet. Nach der Gewinnung einer oberen Schranke für α_{eff} wird eine Rechnung in erster Stufe der Replika – Symmetriebrechung durchgeführt, um das optimale α_{eff} anzunähern. Im vierten Kapitel wird ein praktikabler Algorithmus vorgestellt, der $\alpha_{eff} > 2$ erzeugt. Die Gardner–Rechnung für diesen Algorithmus wird mit numerischen Simulationen überprüft. Da dieser „Quersummenalgorithmus“ weit unterhalb der optimalen Kurve für α_{eff} liegt, befaßt sich das fünfte Kapitel mit dem derzeit besten Verdünnungsalgorithmus, dem Mehrschritt – Schneideverfahren. Für eine einfache Version des Algorithmus läßt sich eine Gardner–Rechnung durchführen. Die analytischen Ergebnisse werden wieder durch numerische Simulationen überprüft.

Das sechste Kapitel befaßt sich mit der Frage der Verallgemeinerungsfähigkeit verdünnter Perzeptrone. Hier werden analytische Ergebnisse des optimalen Verfahrens mit denen des Quersummenalgorithmus verglichen. Die Arbeit schließt im siebten Kapitel mit einer Zusammenfassung und einem Ausblick auf verdünnte Mehrschichtnetzwerke.

Kapitel 2

Ein neues Verfahren zur Berechnung der Speicherkapazität des Perzeptrons optimaler Stabilität

Aufgabe dieses Kapitels ist, das Ergebnis (1.39) für die Speicherkapazität des Perzeptrons optimaler Stabilität ohne Replika herzuleiten. Es wird ein neues Verfahren bereitgestellt, das allgemein zur Berechnung von Speicherkapazitäten und zur Lösung von Verallgemeinerungsproblemen verwendbar ist. In Abschnitt 2.1 wird der Rechenansatz vorgestellt. Er fußt auf dem Abzählen der Zahl $\mathcal{N}$ der Kuhn–Tucker–Lösungen des Optimierungsproblems zum Perzeptron optimaler Stabilität. In Abschnitt 2.2 wird der Mittelwert von $\mathcal{N}$ über die Zahl der Muster bestimmt. Die Ergebnisse der Rechnung werden in Abschnitt 2.3 zusammengefaßt. Schließlich stellen wir in Abschnitt 2.4 zwei wichtige Lernalgorithmen des Perzeptrons optimaler Stabilität vor.

2.1 Der Rechenansatz

2.1.1 Die Komplementaritätsbedingung

Vorgelegt sei die in Abschnitt 1.3 definierte Aufgabe der Bestimmung des Perzeptrons optimaler Stabilität.

$$\text{Minimiere} \quad f(\underline{J}) = \frac{1}{2}\sum_{j=1}^{N} {J_j}^2 \quad \text{unter NB} \quad \lambda_\mu = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j \sigma_j^\mu \geq 1 \quad \forall \mu \quad (2.1)$$

Die λ_μ stellen dabei die lokalen Energien der Muster dar. Es handelt sich um ein Problem der quadratischen Optimierung (siehe [Fl87] für einen Überblick). Da die Zielfunktion f eine positiv definite Hesse–Matrix hat, ist die Lösung $\underline{J}$ des Optimierungsproblems eindeutig.

Die zugehörige Lagrange–Funktion mit Lagrange–Multiplikatoren x_μ lautet

$$L(\underline{J}, \vec{x}) = \frac{1}{2}\sum_{j=1}^{N} {J_j}^2 - \sum_{\mu=1}^{p} x_\mu \left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_j \sigma_j^\mu - 1 \right) \tag{2.2}$$

Wenn die (eindeutige) Lösung $\underline{J}$ existiert, so besagt der Satz von Kuhn und Tucker:

Es gibt Lagrange–Multiplikatoren x_μ, so daß

$$\frac{\partial L}{\partial J_j} = 0 \quad \forall j \Longleftrightarrow J_j = \frac{1}{\sqrt{N}} \sum_{\mu=1}^{p} x_\mu \sigma_j^\mu \quad \forall j \tag{2.3}$$

$$\lambda_\mu = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_j \sigma_j^\mu \geq 1 \quad \forall \mu \tag{2.4}$$

$$x_\mu \geq 0 \quad \forall \mu \tag{2.5}$$

$$(\lambda_\mu - 1) x_\mu = 0 \quad \forall \mu \tag{2.6}$$

Es bleibt also das Problem, die Lagrange–Multiplikatoren des Problems zu finden. Hat man ein solches $\vec{x}$ gewonnen, so erhält man die eindeutige Lösung $\underline{J}$ als Linearkombination (Gl. 2.3). Die Komponenten x_μ geben an, inwiefern die Muster zum Lösungsvektor beitragen. Die x_μ heißen deshalb auch *Einbettungsstärken*. Mit der Korrelationsmatrix

$$C_{\mu\nu} = \frac{1}{N} \sum_{j=1}^{N} \sigma_j^\mu \sigma_j^\nu \tag{2.7}$$

kann man die lokalen Energien der Muster als

$$\lambda_\mu = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_j \sigma_j^\mu = \sum_{\nu=1}^{p} C_{\mu\nu} x_\nu \tag{2.8}$$

schreiben. Das Minimum der Zielfunktion folgt aus

$$Q = \frac{1}{N} \sum_{j=1}^{N} {J_j}^2 = \frac{1}{N} \sum_{\mu,\nu} x_\mu C_{\mu\nu} x_\nu \tag{2.9}$$

Der Satz von Kuhn und Tucker stellt uns die zum Ausgangsproblem äquivalente Aufgabe, Einbettungsstärken x_μ zu finden, die die folgende **Komplementaritätsbedingung** erfüllen

$$(x_\mu = 0 \text{ und } \lambda_\mu \geq 1) \text{ oder } (x_\mu > 0 \text{ und } \lambda_\mu = 1) \quad \forall \mu \tag{2.10}$$

wobei nach Gl.(2.8) $\lambda_\mu = \sum_{\nu=1}^{p} C_{\mu\nu} x_\nu$. Ist diese Komplementaritätsbedingung von irgendeinem $\vec{x}$ erfüllt, so ist der Kopplungsvektor $\underline{J}$ eindeutig durch

$$J_j = \frac{1}{\sqrt{N}} \sum_{\mu=1}^{p} x_\mu \sigma_j^\mu \quad \forall j$$

gegeben. Wir nennen ein solches $\vec{x}$ eine Kuhn–Tucker–Lösung. Über die Eindeutigkeit der Lösung $\vec{x}$ sagt der Satz von Kuhn und Tucker zunächst nichts aus.

Die Komplementaritätsbedingung teilt die Muster in zwei Klassen. Die Muster mit $x_\mu > 0$ heißen eingebettete Muster, weil sie zu $\underline{J}$ beitragen. Sie liegen auf den Rändern der von den Zwangsbedingungen gebildeten Punktmengen. Die Muster mit $x_\mu = 0$ hingegen werden einfach mitgelernt und automatisch richtig klassifiziert. Die eingebetteten Muster sind also die zur Lösung des Problems relevanten Muster.

Um eine Aussage über die Eindeutigkeit der Einbettungsstärken zu erhalten, muß man zwei Zusatzannahmen machen. Wir formulieren den folgenden
Satz:
Die Muster seien in allgemeiner Lage (siehe Abschnitt 1.3.2). Es existiere ein Perzeptron optimaler Stabilität mit Vektor $\underline{J}$. Es gebe eine Kuhn–Tucker–Lösung $\vec{x}_a$ mit höchstens N vielen Komponenten $x_\mu^{(a)} \neq 0$. Es sei

$$\lambda_\mu^{(a)} > 1 \quad \forall \Big(\mu \quad \text{mit} \quad x_\mu^{(a)} = 0 \Big)$$

Dann gibt es keine andere Kuhn–Tucker–Lösung $\vec{x}_b \neq \vec{x}_a$ mit höchstens N vielen Komponenten $x_\mu^{(b)} \neq 0$.

Beweis:
Annahme: Es gibt ein solches $\vec{x}_b$, $\vec{x}_b \neq \vec{x}_a$.

Fallunterscheidung:

1.
$$x_\mu^{(a)} > 0 \Longleftrightarrow x_\mu^{(b)} > 0 \quad \forall \mu$$

   Da $\vec{x}^{(a)}$ eine eindeutige Koordinatendarstellung in einer Basis aus linear unabhängigen Mustervektoren ist, muß gelten $\vec{x}_a = \vec{x}_b$, Widerspruch.

2. Es gibt ein μ mit $x_\mu^{(b)} > 0$, aber $x_\mu^{(a)} = 0$. Daraus folgt $\lambda_\mu^{(b)} = 1$, da $x_\mu^{(b)} > 0$. Nach Voraussetzung ist aber $\lambda_\mu^{(a)} > 1$, Widerspruch.

Die Annahme muß also falsch sein.

Wenn wir gemäß einer kontinuierlichen Wahrscheinlichkeits–Verteilung gezogene Muster betrachten, so sind die Voraussetzung des obigen Satzes bei **endlichen** N und p mit Wahrscheinlichkeit 1 erfüllt. Darüber hinaus ist hier die Wahrscheinlichkeit 0, daß Kuhn–Tucker–Lösungen mit mehr als N Komponenten $x_\mu \neq 0$ auftreten, denn für solche Kuhn–Tucker–Lösungen stellen die Bedingungen

$$\lambda_\mu = 1 \quad \mu = 1, \ldots, N+1$$

überbestimmte lineare Gleichungssysteme dar.

Für die in dieser Arbeit betrachteten gaußverteilten Muster gilt also: Wenn ein Perzeptron–Vektor existiert, so gibt es mit Wahrscheinlichkeit 1 einen eindeutig bestimmten Vektor $\vec{x}$, der die Kuhn–Tucker–Bedingungen erfüllt und den

Vektor des *optimalen* Perzeptrons darstellt. Der Vektor $\vec{x}$ weist höchstens N viele Komponenten $x_\mu \neq 0$ auf. Dies gilt bei endlichen N und p; der Grenzübergang $N \to \infty$ wird erst im nächsten Abschnitt in Zusammenhang mit der Berechnung des Grenzwertes der Cover–Wahrscheinlichkeit benötigt.

2.1.2 Die Zahl der Kuhn–Tucker–Lösungen

Die Zahl $\mathcal{N}$ aller Vektoren $\vec{x}$, die die Komplementaritätsbedingung erfüllen, wird abgezählt, indem man für jede Klasseneinteilung der Muster (in zwei disjunkte Mengen) die Bedingungen für die lokalen Energien λ_μ berücksichtigt. In jeder Klasseneinteilung K numerieren die Indizes $\mu_1(K)$ die eingebetteten Muster sowie die Indizes $\mu_0(K)$ die Muster mit $x_{\mu_0} = 0$.

Um das Problem mit den Methoden der statistischen Physik behandeln zu können, gehen wir im folgenden davon aus, daß ein fester Mustersatz gaußverteilter Muster vorgelegt sei. Nach den obigen Überlegungen läuft dann der Index μ_1 nur über höchstens N viele Muster. Außerdem gibt es dann höchstens eine Kuhn–Tucker–Lösung.

Wir zählen eine solche Lösung, wenn wir setzen [1]

$$\begin{aligned}\mathcal{N} \;=\; \sum_K \left(\int_0^\infty \prod_{\mu_1} dx_{\mu_1} \right) |\mathrm{Det}(C_{\mu_1\nu_1})| \cdot \left(\prod_{\mu_1} \delta\left(\sum_{\nu_1} C_{\mu_1\nu_1} x_{\nu_1} - 1 \right) \right) \cdot \\ \cdot \left(\prod_{\mu_0} \theta\left(\sum_{\nu_1} C_{\mu_0\nu_1} x_{\nu_1} - 1 \right) \right) \qquad (2.11)\end{aligned}$$

Dabei sind die Korrelationsmatrizen wie in Gl.(2.7) definiert. Die Korrelationsmatix der eingebetteten Muster ist positiv definit, da die eingebetteten Muster mit Wahrscheinlichkeit 1 linear unabhängig sind. Die zugehörige (Jacobi–) Determinante dient zur Normierung der Zwangsbedingung der eingebetteten Muster. Den Ungleichungen für die lokalen Energien λ_{μ_0} der Muster mit $x_{\mu_0} = 0$ wird durch die θ–Funktionen Rechnung getragen.

An dieser Stelle sei betont, daß duch den obigen Ansatz mit einem gaußverteilten Mustersatz gewährleistet wird, daß

$$\text{entweder} \quad \mathcal{N} = 1 \quad \text{oder} \quad \mathcal{N} = 0$$

gilt.

2.2 Die Berechnung der gemittelten Zahl der Lösungen

Um das Problem mit Methoden der statistischen Physik lösen zu können, stellen wir uns nun vor, daß der Reihe nach Mustersätze gezogen werden. Die Zahl $\mathcal{N}$

[1] Ein ähnlicher Rechenansatz wurde bereits von S. Diederich und M. Opper bei einem anderen Problem, der Replikator–Dynamik, verwendet [DiOp89].

gibt dann jeweils an, ob ein Perzeptron existiert. Für einen solchen Zufallsprozeß ist aber der Erwartungswert von $\mathcal{N}$ über die gaußverteilten Muster nichts anderes als die Cover–Wahrscheinlichkeit W_s aus Gl.(1.41):

$$W_s = \langle\langle \mathcal{N} \rangle\rangle_{\{\sigma_j^\mu\}} \tag{2.12}$$

Die analytische Berechnung von $\langle\langle \mathcal{N} \rangle\rangle$ ist in unserem Fall nur möglich, wenn wir den Limes $N \to \infty$ durchführen und die Sattelpunktmethode [Br61] anwenden. Wir berechnen

$$f = \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle \mathcal{N} \rangle\rangle \tag{2.13}$$

2.2.1 Die Jacobi–Determinante

Die Determinante $\mathrm{Det}(C_{\mu_1\nu_1})$ könnte mit Grassmann–Variablen dargestellt werden (siehe [FuSh90]); es ist jedoch bekannt, daß die Determinante einer positiv definiten Korrelationsmatrix im betrachteten Grenzfall $N \to \infty$ selbstmittelnd ist und als

$$\mathrm{Det}(C_{\mu_1\nu_1}) \propto \exp\left(N[-\eta\alpha - (1-\eta\alpha)\ln(1-\eta\alpha)]\right) \tag{2.14}$$

geschrieben werden kann [Ku90], [KaSo87]. Dabei ist $\eta \in \left[0, \frac{1}{\alpha}\right]$, und

$$r = \eta p = \eta\alpha N \tag{2.15}$$

ist die Zahl der eingebetteten Muster für die jeweilige Klasseneinteilung. Da die eingebetteten Muster linear unabhängig sind, gilt $\eta\alpha \leq 1$.

2.2.2 Die Mittelung über die Muster

Der Mittelwert über die Muster läßt sich in Formel (2.11) an der Summe über die Klasseneinteilungen vorbeiziehen. Dann können die Muster umgeordnet werden, so daß $\mu_1 = 1, \ldots, r$ und $\mu_0 = r+1, \ldots, p$. Wir verwenden die Fourier–Darstellungen der δ–Funktion und der θ–Funktion (Gln.(A.1) und (A.5)) und erhalten

$$\begin{aligned}
\langle\langle \mathcal{N} \rangle\rangle = & \qquad (2.16)\\
& \sum_{r=1}^{N} \binom{p}{r} \left(\prod_{\mu_1=1}^{\eta p} \int_0^\infty dx_{\mu_1} \int_{-\infty}^{+\infty} \frac{dk_{\mu_1}}{2\pi} \right) \mathrm{Det}(C_{\mu_1\nu_1}) \cdot \\
& \cdot \left\langle\!\!\left\langle \exp\left(i \sum_{\mu_1=1}^{\eta p} k_{\mu_1} \left(1 - \sum_{\nu_1=1}^{\eta p} \frac{1}{N} \sum_{j=1}^{N} \sigma_j^{\mu_1} \sigma_j^{\nu_1} x_{\nu_1} \right) \right) \right. \right. \cdot \\
& \cdot \left(\prod_{\mu_0=\eta p+1}^{p} \int_0^\infty da_{\mu_0} \int_{-\infty}^{+\infty} \frac{db_{\mu_0}}{2\pi} \right) \cdot \\
& \left. \left. \cdot \exp\left(i \sum_{\mu_0=\eta p+1}^{p} b_{\mu_0} \left(a_{\mu_0} - \sum_{\nu_1=1}^{\eta p} \frac{1}{N} \sum_{j=1}^{N} \sigma_j^{\mu_0} \sigma_j^{\nu_1} x_{\nu_1} + 1 \right) \right) \right\rangle\!\!\right\rangle
\end{aligned}$$

Als nächstes werden die Muster durch die Einführung von

$$J_j = \frac{1}{\sqrt{N}} \sum_{\nu_1=1}^{\eta p} \sigma_j^{\nu_1} x_{\nu_1}$$

unter Verwendung von Gl.(A.2) entkoppelt. Führt man wieder die Fourier–Darstellung der δ–Funktion ein, so erhält man

$$\begin{aligned}
&\langle\langle \mathcal{N} \rangle\rangle = \qquad (2.17)\\
&\sum_{r=1}^{N} \binom{p}{r} \left(\prod_{\mu_1=1}^{\eta p} \int_0^{\infty} dx_{\mu_1} \int_{-\infty}^{+\infty} \frac{dk_{\mu_1}}{2\pi} \right) \left(\prod_{\mu_0=\eta p+1}^{p} \int_0^{\infty} da_{\mu_0} \int_{-\infty}^{+\infty} \frac{db_{\mu_0}}{2\pi} \right) \cdot \\
&\cdot \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} dJ_j \int_{-\infty}^{+\infty} \frac{d\hat{J}_j}{2\pi} \right) \mathrm{Det}(C_{\mu_1\nu_1}) \cdot \\
&\cdot \exp\left(i \sum_{\mu_1=1}^{\eta p} k_{\mu_1} + i \sum_{\mu_0=\eta p+1}^{p} b_{\mu_0}(a_{\mu_0}+1) + i \sum_{j=1}^{N} \hat{J}_j J_j \right) \cdot \\
&\cdot \left\langle\left\langle \exp\left[-\frac{i}{\sqrt{N}} \sum_{j=1}^{N} \left(\left(\sum_{\mu_1=1}^{\eta p} k_{\mu_1} J_j + x_{\mu_1} \hat{J}_j \right) \sigma_j^{\mu_1} + \sum_{\mu_0=\eta p+1}^{p} b_{\mu_0} J_j \sigma_j^{\mu_0} \right) \right] \right\rangle\right\rangle
\end{aligned}$$

Die Mustermittelung wird mit Hilfe der Hubbard–Stratonovich–Identität [Hu59] (siehe Gl.(A.8)) ausgeführt, und man erhält

$$\begin{aligned}
\langle\langle \exp[\ldots] \rangle\rangle \quad = \quad & \exp\left[-\frac{1}{2N} \sum_{j,\mu_1} \left({k_{\mu_1}}^2 {J_j}^2 + 2 k_{\mu_1} x_{\mu_1} \hat{J}_j J_j + {x_{\mu_1}}^2 {\hat{J}_j}^2 \right) + \right. \\
& \left. -\frac{1}{2N} \sum_{j,\mu_0} {b_{\mu_0}}^2 {J_j}^2 \right] \qquad (2.18)
\end{aligned}$$

2.2.3 Die Entkopplung der Variablen

Zur Entkopplung der Variablen mit Index j von den Variablen mit Indizes μ_1 und μ_0 werden die Sattelpunktvariblen

$$X = \frac{1}{N} \sum_{j=1}^{N} {J_j}^2 \qquad (2.19)$$

$$Y = \frac{1}{N} \sum_{j=1}^{N} J_j \hat{J}_j \qquad (2.20)$$

$$Z = \frac{1}{N} \sum_{j=1}^{N} {\hat{J}_j}^2 \qquad (2.21)$$

mit konjugierten Größen $\hat{x}, \hat{y}, \hat{z}$ nach Formel (A.3) eingeführt. Die Summe über die Zahl r der eingebetteten Muster wird durch ein Integral über den Parameter

$\eta = r/p$ ersetzt, so daß später die Sattelpunktmethode angewendet werden kann.

$$\sum_{r=1}^{p} \to N \int_0^{1/\alpha} d\eta \tag{2.22}$$

Die Determinante $\mathrm{Det}(C_{\mu_1\nu_1})$ wird mit Hilfe von Gl.(2.14) ersetzt. Nach der Entkopplung der Variablen kann die Sattelpunktmethode [Br61] angewendet werden, und man erhält

$$\begin{aligned} f &= \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle N \rangle\rangle \\ &= \mathrm{sattel}_{\mathcal{M}} \\ &\quad -\alpha\eta \ln \eta - \alpha(1-\eta)\ln(1-\eta) - \eta\alpha - (1-\eta\alpha)\ln(1-\eta\alpha) + Y + \\ &\quad +\hat{x}X + \hat{y}Y + \hat{z}Z + \ln\left(\int_{-\infty}^{+\infty} dJ \int_{-\infty}^{+\infty} \frac{d\hat{J}}{2\pi} \exp\left(-\hat{x}J^2 - i\hat{y}J\hat{J} - \hat{z}\hat{J}^2 \right) \right) + \\ &\quad +\alpha\eta \ln\left(\int_0^{\infty} dx \int_{-\infty}^{+\infty} \frac{dk}{2\pi} \exp\left(-\frac{1}{2}Xk^2 + ik(xY+1) - \frac{1}{2}Zx^2 \right) \right) + \\ &\quad +\alpha(1-\eta)\ln\left(\int_0^{\infty} da \int_{-\infty}^{+\infty} \frac{db}{2\pi} \exp\left(-\frac{1}{2}Xb^2 + ib(a+1) \right) \right) \end{aligned} \tag{2.23}$$

Dabei steht $\mathrm{sattel}_{\mathcal{M}}$ für den Sattelpunkt bezüglich des Satzes von Variablen

$$\mathcal{M} = \{\hat{x}, \hat{y}, \hat{z}, X, Y, Z, \eta\}$$

2.2.4 Das Ergebnis für f

Die Gauß–Integrale in Gl.(2.23) können mit Gl.(A.8) gelöst werden bzw. in Φ–Funktionen umgewandelt werden (siehe Gl.(A.11)). Die Sattelpunktgleichungen für die konjugierten Variablen $\hat{x}, \hat{y}, \hat{z}$ werden dann algebraisch, und man erhält nach Einsetzen

$$\hat{x}X + \hat{y}Y + \hat{z}Z = 1 \tag{2.24}$$

und

$$\hat{y}^2 + 4\hat{x}\hat{z} = \frac{1}{Y^2 + XZ} \tag{2.25}$$

Die konjugierten Variablen können vollständig eliminiert werden, so daß schließlich

$$\begin{aligned} f &= \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle N \rangle\rangle \\ &= \mathrm{sattel}_{\{X,Y,Z,\eta\}} \\ &\quad -\alpha\eta \ln \eta - \alpha(1-\eta)\ln(1-\eta) - \eta\alpha - (1-\eta\alpha)\ln(1-\eta\alpha) + \\ &\quad +1 + Y + \frac{1-\eta\alpha}{2} \ln\left(Y^2 + XZ\right) - \frac{\alpha\eta}{2} \cdot \frac{Z}{Y^2+XZ} + \\ &\quad +\alpha\eta \ln \Phi\left(-\frac{Y/\sqrt{X}}{\sqrt{Y^2+XZ}} \right) + \alpha(1-\eta)\ln\Phi\left(-\frac{1}{\sqrt{X}} \right) \end{aligned} \tag{2.26}$$

2.2.5 Die Lösung der Sattelpunktgleichungen

Startet man mit dem Ansatz $Z = 0$, so gilt

$$\frac{\partial f}{\partial Y} = 0 \Longrightarrow Y = -(1 - \eta\alpha) < 0 \tag{2.27}$$

$$\frac{\partial f}{\partial \eta} = 0 \Longrightarrow \eta = \Phi\left(\frac{1}{\sqrt{X}}\right) \tag{2.28}$$

$$\frac{\partial f}{\partial X} = 0 \quad \text{ist erfüllt}$$

Es bleibt, den Ansatz $Z = 0$ zu bestätigen

$$\left.\frac{\partial f}{\partial Z}\right|_{Z=0} = \frac{1 - \eta\alpha}{2} \cdot \frac{X}{Y^2} - \frac{\alpha\eta}{2Y^2} + \alpha\eta \cdot \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{X}\right)}{\Phi\left(+\frac{1}{\sqrt{X}}\right)} \cdot \frac{\sqrt{X}}{2Y^2} \tag{2.29}$$

Setzt man Y in die Gleichung ein und berücksichtigt, daß aufgrund der Gln. (2.19) und (1.29) für die Stabilität

$$\kappa = \frac{1}{\sqrt{X}} \tag{2.30}$$

gilt, so liefert Gl.(2.29) gerade das Ergebnis (1.39) von E. Gardner für $\alpha < 2$.

Nach Einsetzen der Lösung in den Exponenten f in Gl.(2.26) erhält man einfach

$$f(\alpha < 2) = 0 \tag{2.31}$$

Setzen wir die Sattelpunktvariablen im Falle $\alpha \geq 2$ fort und beachten wir, daß dann N eingebettete Muster vorliegen müssen, der Parameter η also an seiner oberen Schranke $\eta = \frac{1}{\alpha}$ angelangt ist, so gilt mit

$$\kappa = 0,\ Y = 0,\ Z = 0,\ \eta = \frac{1}{\alpha}$$

nach Einsetzen

$$f(\alpha \geq 2) = \ln\alpha - (\alpha - 1)\ln\left(1 - \frac{1}{\alpha}\right) - \alpha \ln 2$$

2.3 Ergebnisse

Die Berechnung des Mittelwertes $\langle\langle \mathcal{N} \rangle\rangle$ der Zahl der Kuhn–Tucker–Punkte $\vec{x}$ bestätigt die Ergebnisse der replikasymmetrischen Gardner–Rechnung zum Perzeptron optimaler Stabilität. Zudem wird der Bruchteil η der eingebetteten Muster ermittelt. Die Lösung der Sattelpunktgleichungen zu

$$f = \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle \mathcal{N} \rangle\rangle$$

ergibt:

1. Die Gleichung (1.39) wird bestätigt

$$\alpha_c(\kappa) = \left[\left(1+\kappa^2\right)\Phi(\kappa) + \frac{\kappa}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\kappa^2\right)\right]^{-1} \tag{2.32}$$

liefert die optimale Stabilität bei vorgegebenem $\alpha < 2$.

2. Das von M. Opper mit Replika–Methoden berechnete Ergebnis für den Bruchteil der eingebetteten Muster wird bestätigt ([Op88], [KiOp89], [Wen91]).

Der Bruchteil der eingebetteten Muster ist für $\alpha < 2$

$$\eta = \Phi(\kappa) \tag{2.33}$$

Damit gilt $\eta\alpha < 1$. [2]

3. Nach Einsetzen der Lösungen in die Sattelpunktfunktion erhält man für $\alpha < 2$

$$f(\alpha < 2) = 0 \tag{2.34}$$

Wir erhalten also das gleiche Ergebnis wie Cover: Im Grenzfall $N \to \infty$ und für $\alpha < 2$ gibt es mit Wahrscheinlichkeit 1 ein Perzeptron.

Für $\alpha \geq 2$ stellen wir uns vor, daß die Zahl der eingebetteten Muster stets N ist. η ist also auf den Wert $\eta = \frac{1}{\alpha}$ eingefroren. Die Fortsetzung der mit der Sattelpunktmethode gewonnenen Lösungen ergibt

$$f(\alpha \geq 2) = \ln\alpha - (\alpha-1)\ln\left(1-\frac{1}{\alpha}\right) - \alpha\ln 2 \tag{2.35}$$

Die Wahrscheinlichkeit, daß ein Perzeptron existiert, ist hier also exponentiell klein. Das Ergebnis ist identisch mit dem Grenzwert der Cover–Wahrscheinlichkeit W_s (siehe Gl.(1.43)): Es gilt

$$f(\alpha \geq 2) = \lim_{N\to\infty}\frac{1}{N}\ln W_s \tag{2.36}$$

2.4 Lernalgorithmen

Das in Abschnitt 2.1 vorgestellte Problem der quadratischen Optimierung kann mit effizienten numerischen Verfahren gelöst werden. Die bekanntesten Algorithmen zur Bestimmung des Perzeptrons optimaler Stabilität sind der MinOver–Algorithmus [KrMe87], das AdaTron [AnBi90], [Bi91] und die direkte quadratische Optimierung [Ru91].

Das AdaTron bringt die Einbettungsstärken gemäß der Formel

$$x_\mu(t+1) = x_\mu(t) + \max\{\gamma(1-\lambda_\mu),\ -x_\mu\} \tag{2.37}$$

[2] $(\eta\alpha)$ ist eine streng monoton fallende Funktion von κ mit $(\eta(0)\alpha(0)) = 1$ und $(\eta\alpha) \to 0$ $(\kappa \to \infty)$

auf den neuesten Stand. Dies kann sequentiell oder parallel in den Mustern erfolgen. γ ist eine geeignet zu wählende Schrittweite. λ_μ ist die lokale Energie (Gl.(2.8))

$$\lambda_\mu = \sum_{\nu=1}^{p} C_{\mu\nu} x_\nu \tag{2.38}$$

Durch die obige Definition wird gewährleistet, daß stets $x_\mu \geq 0$ gilt. Die Abbruchbedingung des Algorithmus ist die Komplementaritätsbedingung (2.10).

Bei der direkten quadratischen Optimierung wird einfach das Ausgangsproblem (Gl.(2.1)) betrachtet. Das Verfahren sucht dabei nach den aktiven Zwangsbedingungen, für die die lokalen Energien gerade 1 sind [Ru91], [GoId83], [IMSL].

Kapitel 3

Die optimale Verdünnung des Perzeptrons

In diesem Kapitel wird das Problem der maximalen Speicherkapazität bei der optimalen linearen Verdünnung des Perzeptrons behandelt. Dabei wird die in Abschnitt 1.4 eingeführte Notation verwendet. Es wird gezeigt, daß α_{eff}, die Speicherkapazität pro Neuron, viel größer als 2 sein kann.

In Abschnitt 3.1 wird zunächst eine obere Schranke für α_{eff} berechnet. Dies geschieht mit Hilfe des Satzes von Cover aus Abschnitt 1.3.2.

Im Rest des Kapitels wird eine Näherungsrechnung für α_{eff} vorgestellt, die auf Replika–Methoden beruht: Nach der Vorstellung des Rechenansatzes in Abschnitt 3.2 folgen zwei Abschnitte, die sich mit der analytischen Rechnung beschäftigen. In Abschnitt 3.3 wird die bereits von Bouten et al. [Bo+90] durchgeführte replikasymmetrische Rechnung vorgestellt. In Abschnitt 3.4 folgt die Brechung der Replika–Symmetrie, die, wie sich zeigen wird, qualitativ und quantitativ neue Ergebnisse hervorbringt. Die Ergebnisse der Rechnung werden schließlich in Abschnitt 3.5 vorgestellt.

3.1 Die obere Schranke für die maximale Speicherkapazität

Wir formulieren unser Problem der optimalen Verdünnung unter Verwendung der Notation aus Abschnitt 1.4:
Unter den $\binom{N}{Nf}$ Möglichkeiten, das Netzwerk so zu verdünnen, daß Nf Neuronen übrigbleiben, sind diejenigen Verdünnungsvektoren $\underline{c}$ zu finden, auf denen das Perzeptron optimaler Stabilität die größte Stabilität überhaupt hat.

Für große N hätte man also *exponentiell* viele $\underline{c}$– Vektoren zu betrachten. Für jeden $\underline{c}$–Vektor müßte man überprüfen, ob ein Perzeptron existiert. Mit einem Lernalgorithmus wäre dann die optimale Stabilität $\kappa(\underline{c})$ zu bestimmen. Schließlich wären das Maximum

$$\kappa_{max} = \max_{\{\underline{c}\}} \{\kappa(\underline{c})\} \tag{3.1}$$

zu bilden und die zugehörigen $\underline{c}^{max}$ abzuspeichern. Ein solches Vorgehen ist

nicht praktikabel, da der Rechenaufwand exponentiell in N ist. Wir sehen jedoch, daß das Problem zwei Gesichter hat. Der „Janus“ besteht in einer diskreten Optimierung in den c_i und in einer quadratischen Optimierung in den Kopplungen, die zu einem Verdünnungsvektor $\underline{c}$ gehören.

Zur Berechnung der oberen Schranke der Speicherkapazität verwenden wir den Satz von Cover (siehe Gl.(1.40)). f sei vorgegeben. Wir betrachten

$$\alpha_{eff} = \frac{p}{Nf} = \frac{\alpha}{f} > 2 \tag{3.2}$$

Dann ist die Zahl der binären Ausgabevektoren, für die die Muster linear separabel sind, für jeden Verdünnungsvektor nach Gl.(1.40) als

$$C(p, Nf) = 2 \sum_{i=0}^{Nf-1} \binom{p-1}{i} \tag{3.3}$$

gegeben. Da $Nf < \frac{p}{2}$ ist, folgt aus der Struktur des Pascal–Dreiecks

$$C(p, Nf) \leq 2 \cdot Nf \cdot \binom{p}{Nf} \tag{3.4}$$

Damit die Muster auf irgendeinem der $\binom{N}{Nf}$ vielen $\underline{c}$–Vektoren linear separabel sind, muß der vorgegebene Ausgabevektor $\vec{S}$ einer von z möglichen Ausgabevektoren sein. Wenn man die Annahme macht, daß die Ausgaben S^μ unabhängig voneinander gemäß der Wahrscheinlichkeit $p(S^\mu = \pm 1) = \frac{1}{2}$ ausgewürfelt werden, so gilt für die Wahrscheinlichkeit für lineare Separabilität

$$\begin{aligned} W_s &= \frac{z}{2^p} \leq \sum_{\underline{c}} C(p, Nf)/2^p \\ W_s &\leq \binom{N}{Nf} \cdot 2 \cdot Nf \cdot \binom{p}{Nf} / 2^p \end{aligned} \tag{3.5}$$

Die Ungleichung beruht darauf, daß Ausgabevektoren mehrfach gezählt werden können. Mit Hilfe der Stirling–Formel und $p = \alpha N$ folgt analog zu Gl.(2.36)

$$\lim_{N\to\infty} \frac{1}{N} \ln W_s \leq \tag{3.6}$$
$$-f \ln f - (1-f) \ln (1-f) + f \ln \frac{\alpha}{f} - (\alpha - f) \ln \left(1 - \frac{f}{\alpha}\right) - \alpha \ln 2 =: s(f, \alpha)$$

Die Nullstelle α_0 dieser Funktion s ist also eine obere Schranke für die maximale Speicherkapazität α_c bei gegebenem f. Da das verdünnte Perzeptron ein kleineres α_c als das vollvernetzte haben muß, gilt

$$\alpha_c(f) \leq \min \{\alpha_0(f), 2\} \tag{3.7}$$

Diese obere Schranke ist in Abbildung 3.1 dargestellt.

Abschließend diskutieren wir die Nullstellen von s für $f \to 1$ und $f \to 0$.

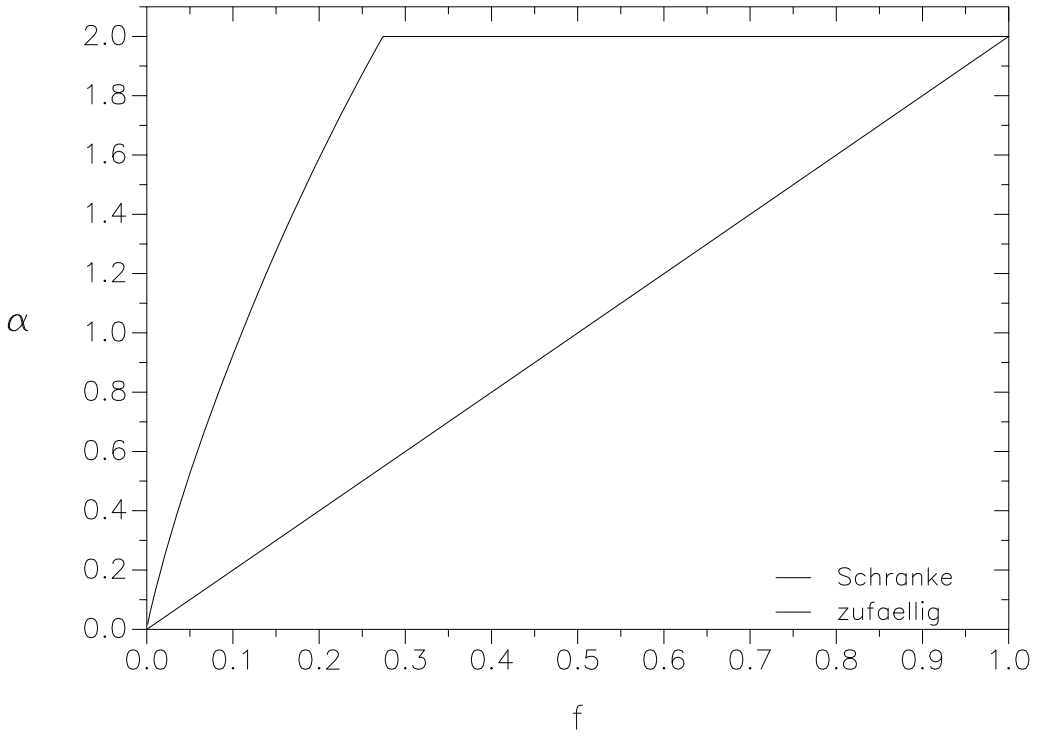


Abbildung 3.1: Die obere Schranke aus der Ungleichung (3.7) im Vergleich mit der Speicherkapazität $\alpha = 2f$ der zufälligen Verdünnung

Für $f \to 1$ erhalten wir einfach

$$s(f,\alpha) \to -\alpha \ln 2 + \ln \alpha - (\alpha - 1) \ln \left(\frac{\alpha - 1}{\alpha} \right) \quad (f \to 1) \tag{3.8}$$

Also gilt $s(f = 1, \alpha = 2) = 0$

Für $f \to 0$ erhalten wir

$$\frac{s(f,\alpha)}{f} = -\ln f + 1 - \frac{\alpha}{f} \ln 2 + \ln \frac{\alpha}{f} - \left(\frac{\alpha}{f} - 1 \right) \ln \left(1 - \frac{f}{\alpha} \right) + \mathcal{O}(f) \tag{3.9}$$

Damit eine Nullstelle α_0 existiert, muß

$$\frac{\alpha_0}{f} = -C \cdot \ln f, \quad C = \mathcal{O}(1)$$

gelten. Daraus folgt

$$\frac{s(f,\alpha_0)}{-f \ln f} \to 1 - C \ \ln 2 \quad (f \to 0) \tag{3.10}$$

Die obere Schranke für die Speicherkapazität divergiert also logarithmisch, und es gilt

$$\lim_{f \to 0} \frac{\alpha_0}{-f \ln f} = \frac{1}{\ln 2} \tag{3.11}$$

bzw. mit dem dualen Logarithmus

$$\lim_{f \to 0} \frac{\alpha_0}{-f \ \mathrm{ld} f} = 1 \tag{3.12}$$

3.2 Der Ansatz zur Berechnung der optimalen Speicherkapazität

Wie schon in Abschnitt 3.1 beschrieben, besteht das Problem der optimalen Verdünnung aus zwei Teilproblemen. Die optimalen Verdünnungsvektoren $\underline{c}$ müssen gefunden werden, und auf ihnen muß das Perzeptron optimaler Stabilität gelernt werden. Bouten et al. [Bo+90] trugen dieser Tatsache Rechnung, indem sie die anteiligen Phasenraumvolumina für alle zugelassenen Verdünnungsvektoren aufsummierten:

$$V_{ges} = \sum_{\underline{c}} V(\underline{c}) \cdot \delta_{Kr} \left(Nf - \sum_{j=1}^{N} c_j \right) \tag{3.13}$$

wobei δ_{Kr} das Kronecker–δ aus Gl.(A.6) ist. $V(\underline{c})$ ist das anteilige Gardner–Volumen (Gl.(1.33)) auf den jeweils verbleibenden Nf Plätzen. Man erhält

$$\begin{aligned} V_{ges} \;=\; & \sum_{\underline{c}} \delta_{Kr} \left(Nf - \sum_{j=1}^{N} c_j \right) \cdot \\ & \cdot \frac{\left(\prod_{k=1}^{Nf} \int_{-\infty}^{+\infty} \frac{dJ_k}{\sqrt{2\pi}} \right) \delta\left(\sum_{k=1}^{Nf} {J_k}^2 - Nf \right) \cdot \prod_{\mu=1}^{p} \theta\left(\frac{1}{\sqrt{Nf}} \sum_{k=1}^{Nf} J_k \eta_k^\mu - \kappa \right)}{\left(\prod_{k=1}^{Nf} \int_{-\infty}^{+\infty} \frac{dJ_k}{\sqrt{2\pi}} \right) \delta\left(\sum_{k=1}^{Nf} {J_k}^2 - Nf \right)} \end{aligned} \tag{3.14}$$

Dabei numerieren die Indizes $k(\underline{c})$ die verbleibenden Neuronen. Die η_k^μ sind die zugehörigen transformierten Muster. Um die Abhängigkeit $k(\underline{c})$ zu eliminieren, multipliziert man Zähler und Nenner der obigen Brüche mit

$$\left(\prod_{k=Nf+1}^{N} \int_{-\infty}^{+\infty} \frac{dJ_k}{\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum_{k=Nf+1}^{N} {J_k}^2 \right) \tag{3.15}$$

und erhält unter Verwendung von Gl.(A.7)

$$\begin{aligned} V_{ges} \;=\; & \sum_{\{\underline{c}\}} \int_{-i\pi}^{i\pi} \frac{d\psi}{2\pi i} \exp\left(\psi \left(Nf - \sum_{j=1}^{N} c_j \right) \right) \cdot \\ & \cdot \frac{1}{C_{norm}} \cdot \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} \frac{dT_j}{\sqrt{2\pi}} \right) \delta\left(\sum_{j=1}^{N} c_j {T_j}^2 - Nf \right) \exp\left(-\frac{1}{2} \sum_{j=1}^{N} (1 - c_j) {T_j}^2 \right) \cdot \\ & \cdot \prod_{\mu=1}^{p} \theta\left(\frac{1}{\sqrt{Nf}} \sum_{j=1}^{N} c_j T_j \sigma_j^\mu - \kappa \right) \end{aligned} \tag{3.16}$$

Dabei sind die T_j die neuen Integrationsvariablen. Die Normierungskonstante für die Oberfläche der Gardner–Kugel ist

$$C_{norm} = \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} \frac{dT_j}{\sqrt{2\pi}} \right) \delta\left(\sum_{j=1}^{N} c_j {T_j}^2 - Nf \right) \exp\left(-\frac{1}{2} \sum_{j=1}^{N} (1 - c_j) {T_j}^2 \right)$$

3.2.1 Ausgeglühte und eingefrorene Verdünnung

In der obigen Gleichung (3.16) kann man V_{ges} wieder als die Zahl erlaubter Zustände auffassen. Die Zwangsbedingungen sind dabei gegeben durch

1. das Feld ψ, welches die Verdünnung f festhält und
2. die Θ–Funktionen für die lokalen Energien der Muster bezüglich der verdünnten Perzeptrone

Summiert wird über alle Verdünnungsvektoren $\underline{c}$ und alle Vektoren $\underline{T}$. Somit sind neben den Kopplungsvektoren $\underline{T}$ auch die Verdünnungsvektoren $\underline{c}$ thermodynamische Variablen. Bouten et al. bezeichnen deshalb die optimale Verdünnung auch als **ausgeglühte** Verdünnung („annealed dilution“).

Im Gegensatz dazu bezeichnen sie ein Verfahren, in dem der Verdünnungsvektor feststeht und nur bezüglich der Kopplungen optimiert wird, als **eingefrorene** Verdünnung („quenched dilution“). Das in Abschnitt 1.4 angesprochene Verfahren, zufällig zu verdünnen und dann nachzulernen (siehe Gl.(1.60)) ist ein einfaches Beispiel für die eingefrorene Verdünnung. In den Kapiteln 4 und 5 werden wir uns effektiveren Verfahren der eingefrorenen Verdünnung zuwenden.

3.2.2 Das allgemeine Ergebnis für natürliche n

Man nimmt nun an, daß die Entropie

$$s = \frac{1}{N} \ln V_{ges} \tag{3.17}$$

bezüglich der Mittelung über *alle* transformierten Muster selbstmittelnd ist. Die zugehörige Replika–Rechnung findet sich in [Bo+90] (siehe auch [Gc92]). Wie in Abschnitt 1.3.2 beschrieben, ist bei der Replika–Methode

$$\Phi_R(n) = \lim_{N\to\infty} \frac{1}{N} \ln \left\langle \left\langle {V_{ges}}^n \right\rangle \right\rangle_{\{\sigma_j^\mu\}} \tag{3.18}$$

aus Gl.(1.37) zu berechnen.

Man erhält

$$\begin{aligned} \Phi_R(n) = & \qquad (3.19) \\ & \mathrm{sattel}_{\{\psi_\rho, Q_{\rho\sigma}, t_{\rho\sigma}\}} \quad f \sum_{\rho=1}^{n} \psi_\rho - \sum_{\rho\le\sigma} Q_{\rho\sigma} t_{\rho\sigma} + \\ & + \ln \left(\sum_{\{c_\rho\}} \left(\prod_{\rho=1}^{n} \int_{-\infty}^{+\infty} DJ_\rho \right) \exp\left(\frac{1}{f} \sum_{\rho\le\sigma} J_\rho c_\rho t_{\rho\sigma} J_\sigma c_\sigma - \sum_{\rho=1}^{n} c_\rho \psi_\rho \right) \right) + \\ & + \alpha \ln \left(\left(\prod_{\rho=1}^{n} \int_{\kappa}^{\infty} \frac{d\lambda_\rho}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{dr_\rho}{\sqrt{2\pi}} \right) \exp\left(-\frac{1}{2} \sum_{\rho,\sigma} r_\rho Q_{\rho\sigma} r_\sigma + i \sum_{\rho=1}^{n} r_\rho \lambda_\rho \right) \right) \end{aligned}$$

Der Ausdruck „sattel“ bedeutet wieder, daß der Sattelpunkt bezüglich der angegebenen Variablen gesucht werden muß. Der Ausdruck DJ_ρ ist in Gl.(A.10)

erklärt. $\rho, \sigma = 1, \ldots, n$ sind Replika–Indizes. Die (symmetrische) Matrix der Ordnungsparameter

$$Q_{\rho\sigma} = \frac{1}{Nf} \sum_{j=1}^{N} c_j^\rho T_j^\rho c_j^\sigma T_j^\sigma \tag{3.20}$$

gibt die Überlappung zweier Lösungsvektoren an, die zumindest eine Stabilität κ besitzen. Wegen der Zwangsbedingung der Gardner–Kugel ist $Q_{\rho\rho} = 1 \; \forall \rho$. Die symmetrische Matrix der Hilfsvariablen $t_{\rho\sigma}$ ist die zu $Q_{\rho\sigma}$ konjugierte Matrix. Sie wurde bei einer Entkopplung der Form (A.3) eingeführt.

Um die Entropie s berechnen zu können, müssen wir nun $\Phi_R(n)$ analytisch nach $n = 0$ fortsetzen. In Abschnitt 3.3 wird gezeigt, daß die naive replikasymmetrische Fortsetzung von Φ_R keine zufriedenstellenden Ergebnisse liefert. In Abschnitt 3.4 wird die Replikasymmetriebrechung erster Stufe (RSB1) behandelt. Das bedeutet, daß eine analytische Fortsetzung von $\Phi_R(n)$ in RSB1 gebildet wird.

3.3 Die replikasymmetrische Näherung

Die naive replikasymmetrische Fortsetzung von Φ_R in Gl.(3.19) basiert auf der Annahme

$$\begin{aligned} Q_{\rho\sigma} &= q, \quad t_{\rho\sigma} = t, \quad \rho \neq \sigma \\ t_{\rho\rho} &= T, \psi_\rho = \psi, \quad \rho = 1, \ldots, n \end{aligned} \tag{3.21}$$

für die Sattelpunktvariablen $Q_{\rho\sigma}$ und ψ_ρ und analog für $t_{\rho\rho}$. Setzt man dies in Gl.(3.19) ein und entwickelt Φ_R bis zur ersten Ordnung in n, so erhält man nach [Bo+90] (siehe auch [Gc92] für Einzelheiten der Rechnung):

$$\begin{aligned} s_{sym} &= \left. \frac{\partial \Phi_R^{(sym)}}{\partial n} \right|_{n=0} \\ &= \mathrm{sattel}_{\{q,\psi,t,T\}} \qquad f\psi - T + \frac{1}{2}qt + \\ &\quad + \int_{-\infty}^{+\infty} Dz \; \ln\left(1 + \frac{\exp(-\psi)}{\sqrt{1 + \frac{t-2T}{f}}} \cdot \exp\left(\frac{1}{2}z^2 \cdot \frac{t/f}{1 + \frac{t-2T}{f}}\right)\right) + \\ &\quad + \int_{-\infty}^{+\infty} Dz \; \ln \Phi\left(-\frac{\kappa + z\sqrt{q}}{\sqrt{1-q}}\right) \end{aligned} \tag{3.22}$$

Dabei ist die Funktion Φ in Gl.(A.11) erklärt.

Bei der Lösung der vier Sattelpunktgleichungen für q, ψ, t, T beobachtet man, daß q gegen 1 strebt, wenn die Stabilität κ größer wird. Dies ist konsistent mit der obigen Annahme (3.21) für die Überlappungsmatrix $Q_{\rho\sigma}$: Alle Perzeptron–Lösungen haben bei gegebener Stabilität κ die gleiche Überlappung

$$q = \frac{1}{Nf} \sum_{j=1}^{N} c_j^\rho T_j^\rho c_j^\sigma T_j^\sigma \tag{3.23}$$

Erhöht man κ, so werden sich die Lösungen immer ähnlicher. Schließlich bleibt am kritischen Punkt nur *ein* einziger Lösungsvektor übrig. Da $q = 1$ gilt, haben wir also hier die Vorstellung, daß nur *ein einziger Verdünnungsvektor* $\underline{c}$ bei vorgegebenem f die optimale Stabilität liefert.

Bouten et al. erhalten im Limes $q \to 1$ das folgende Endergebnis.

Seien κ und f gegeben, so gilt in replikasymmetrischer Näherung für die maximale Speicherkapazität

$$\alpha = \frac{f + w\sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2}w^2\right)}{(1+\kappa^2)\Phi(\kappa) + \frac{\kappa}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\kappa^2\right)} \tag{3.24}$$

wobei w die Lösung der Gleichung

$$f = 2\Phi(-w) \tag{3.25}$$

ist. Dies beinhaltet das Gardner–Ergebnis für das vollvernetzte Perzeptron mit $f = 1$. Setzen wir $\kappa = 0$, so beobachten wir, daß die effektive Speicherkapazität

$$\alpha_{eff} = \frac{\alpha}{f} > 2$$

ist. In Abbildung 3.2 sieht man eine deutliche Verbesserung des Wertes

$$\alpha = 2f$$

der zufälligen Verdünnung. In Abbildung 3.2 wird α aus Gleichung (3.24) ebenfalls mit der in Abschnitt 3.1 gewonnenen oberen Schranke verglichen. Die ebenfalls dargestellte „AT–Linie“ wird unten erklärt.

Im Falle kleiner f stellt sich heraus, daß die obere Schranke verletzt wird, dies ist in Abbildung 3.3 dargestellt.

Für die obere Schranke erhält man den Grenzwert

$$\lim_{f\to 0} \frac{\alpha_s}{-f \ln f} = \frac{1}{\ln 2} \sim 1.4427 \tag{3.26}$$

aus Gl.(3.11).

Zur Herleitung des Grenzwertes im Fall der optimalen Verdünnung betrachten wir Gl.(3.25) für große w. Wegen Gl.(A.14) gilt

$$f = \frac{2}{w} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) \cdot \left(1 + \mathcal{O}\left(\frac{1}{w^2}\right)\right) \tag{3.27}$$

Setzt man dies in Gl.(3.24) ein, so gilt

$$\frac{\alpha(\kappa = 0)}{f} = 2 \cdot w^2 + \mathcal{O}(1) \tag{3.28}$$

Durch die Bildung des Logarithmus in Gl.(3.27) und anschließender Division folgt schließlich aus Gl.(3.28) das von Bouten et al. angegebene Endergebnis

$$\lim_{f\to 0} \frac{\alpha(\kappa = 0)}{-f \ln f} = 4 \tag{3.29}$$

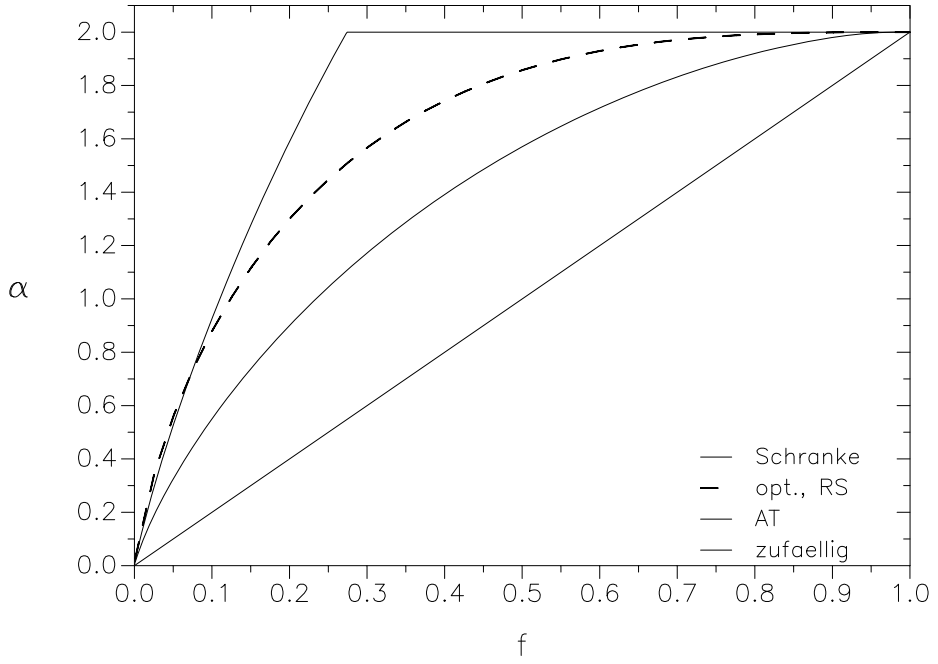


Abbildung 3.2: Der Vergleich der oberen Schranke für die Speicherkapazität des optimal verdünnten Perzeptrons mit der replikasymmetrischen Näherung von Bouten et al. [Bo+90]. Von oben nach unten sind dargestellt: obere Schranke, replikasymmetrische Näherung, AT–Linie und zufällige Verdünnung

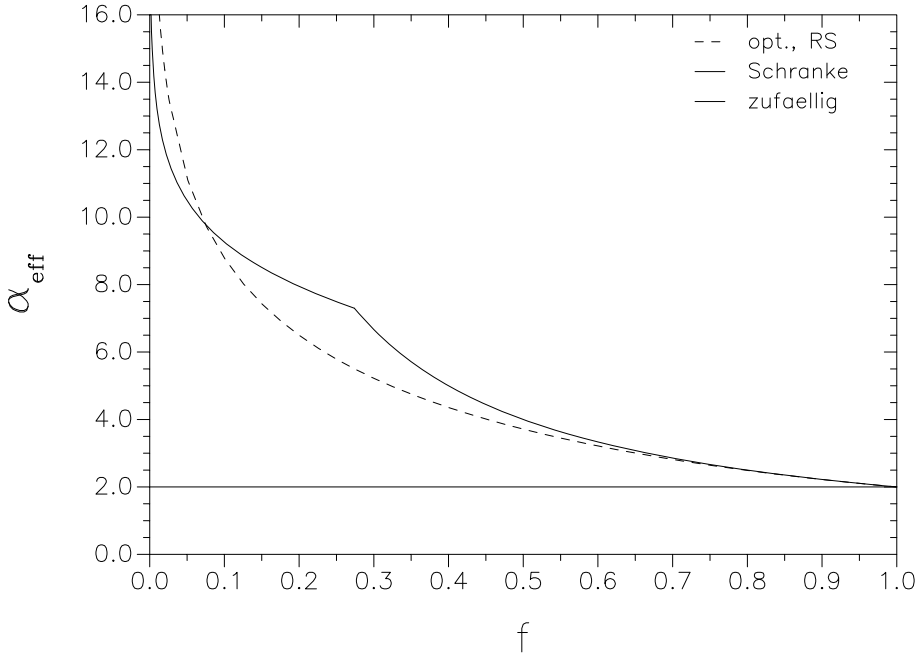


Abbildung 3.3: Die effektive Speicherkapazität $\alpha_{eff} = \frac{\alpha}{f}$ mit α aus Abbildung 3.2. Die replikasymmetrische Näherung verletzt die obere Schranke bei kleinen Werten von f.

Die in diesem Abschnitt besprochene replikasymmetrische Rechnung kann folglich nicht korrekt sein. Einen Hinweis auf diese Tatsache erhält man, wenn man eine Stabilitätsuntersuchung der replikasymmetrischen Lösung vornimmt. Es handelt sich hierbei um ein von de Almeida und Thouless entwickeltes Verfahren, mit dem man im Rahmen der Replika–Methode die Stabilität des gewonnenen Sattelpunktes untersucht [AT78], [Bo92]. Für das vorliegende Problem hat M. Wong diese Rechnung durchgeführt [Wo91]. Die zugehörige de Almeida–Thouless– Linie (AT–Linie) ist in Abbildung 3.2 zu sehen. Sie gibt an, bis zu welchem Punkt die replikasymmetrische Lösung stabil ist. Die Lösung (3.22) ist also für jedes f nur bis zu einem maximalen $q < 1$ stabil, und die $q = 1$–Lösung liegt im Bereich der Instabilität.

3.4 Die Replika–Symmetriebrechung erster Stufe für die optimale Verdünnung

Wie in Abschnitt 3.3 aufgezeigt wurde, muß man über den replikasymmetrischen Ansatz (3.21) hinausgehen. Das Problem ist dabei, die korrekte analytische Fortsetzung für den Limes $n \to 0$ des Replika–Index zu finden. Von dem Sherrington–Kirkpatrick–Modell (SK–Modell) der Spingläser ist ein solcher Ansatz bekannt. Er wurde von G. Parisi entwickelt. Es handelt sich um das Parisi–Schema der Replika–Symmetriebrechung (RSB). Die vollständige Parisi–Lösung besteht aus einem komplizierten Iterationsverfahren, das zufriedenstellende Ergebnisse liefert. Den ersten Schritt dieses Iterationsverfahrens stelle ich in Abschnitt 3.4.1 in Zusammenhang mit dem SK–Modell vor. Diese Replika–Symmetriebrechung erster Stufe (RSB1) stellt dort bereits eine erhebliche Verbesserung des replika–symmetrischen Ergebnisses dar. In Abschnitt 3.4.2 wenden wir die RSB1 auf das Problem der optimalen Verdünnung an. Die maximale Speicherkapazität wird schließlich in Abschnitt 3.4.3 berechnet.

3.4.1 Das SK–Modell und der Parisi–Ansatz

Betrachtet sei ein verallgemeinertes Ising–Modell (siehe Abschnitt 1.2) mit Energiefunktion

$$E(\underline{S}) = -\frac{1}{2} \sum_{i,j\,(i \neq j)} J_{ij} S_i S_j \tag{3.30}$$

Die Kopplungen J_{ij} seien symmetrisch. Bei dem auf Sherrington und Kirkpatrick zurückgehenden Modell ([SK75], [SK78], [BiYo86], [FiHe91], [Me+87]) sind die Kopplungen gaußverteilt mit einer Streuung $\sigma = \frac{J}{\sqrt{N}}$. Für die Wahrscheinlichkeitsdichte von J_{ij} ($i < j$, da $J_{ij} = J_{ji}$) gilt

$$P(J_{ij}) = \sqrt{\frac{N}{2\pi J^2}} \exp\left(-\frac{N {J_{ij}}^2}{2\pi J^2}\right) \tag{3.31}$$

Diese Unordnung in den Kopplungen bewirkt, daß die freie Energie pro Spin f sehr viele lokale Minima aufweist. Bei der Berechnung von

$$f = -\frac{T}{N} \ln Z \tag{3.32}$$

mit der Zustandssumme

$$Z = \sum_{\{\underline{S}\}} \exp\left(-\beta E(\underline{S})\right) \tag{3.33}$$

muß man die Selbstmittelungseigenschaft

$$\lim_{N\to\infty} f = \lim_{N\to\infty} \langle\langle f \rangle\rangle_{\{J_{ij}\}} \tag{3.34}$$

voraussetzen. Man wendet die Replika–Methode zur Berechnung von

$$\Phi_R(n) = \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle Z^n \rangle\rangle_{\{J_{ij}\}} \tag{3.35}$$

an. Dabei führt man den Edwards–Anderson–Ordnungsparameter

$$Q_{\rho\sigma} = \frac{1}{N}\sum_{j=1}^{N} S_j^{\rho} S_j^{\sigma}, \quad \rho,\sigma = 1,\ldots,n \tag{3.36}$$

ein, der die Überlappung zweier thermodynamischer Zustände des n–fach replizierten Spinsystems angibt. Φ_R ist schließlich eine Funktion der Überlappungsmatrix $Q_{\rho\sigma}$. Der in der Arbeit von van Hemmen und Palmer [HePa79] zitierte Satz von Elliott Lieb besagt, daß für eine natürliche Zahl n, $n \geq 1$, die replikasymmetrische Lösung vorliegt. In der Arbeit von Kondor [Ko83] wird jedoch gezeigt, daß die naive replikasymmetrische Fortsetzung nach $n = 0$ falsch ist, weil sie von einem gewissen n_c an, $0 < n_c < 1$, instabil wird. Außerdem wurde schon in der Arbeit von Sherrington und Kirkpatrick [SK78] für die replikasymmetrische Näherung die negative Entropie

$$s(T=0) = -\frac{df}{dT} = -\frac{\partial f}{\partial T} \sim -0.16 \tag{3.37}$$

am Temperaturnullpunkt berechnet [1]. Dies widerspricht dem dritten Hauptsatz der Thermodynamik [Reif], [Huang], nach dem $s(T=0) = 0$ gelten muß. Die Zahl der Grundzustände ist freilich größer als 1, so daß s nicht negativ sein kann. Analog zu den Attraktornetzwerken können wir darüber hinaus davon ausgehen, daß die Zahl $\mathcal{N}_G$ der Grundzustände höchstens polynomial in N ist, so daß stets

$$s(T=0) = \lim_{N\to\infty} \frac{1}{N} \ln \mathcal{N}_G = 0 \tag{3.38}$$

gilt. Im Zusammenhang mit dem SK–Modell wurde Gl.(3.38) von van Hemmen und van Enter bewiesen [EnHe84].

Die von Parisi vorgeschlagene Parametrisierung der Überlappungsmatrix $Q_{\rho\sigma}$ [Pa80a], [Pa80b], [Me+87] beseitigt beide oben angesprochenen Unstimmigkeiten. Die vollständige Parisi–Lösung ist marginal stabil und weist die Entropie $s = 0$ auf. Bereits die Replika–Symmetriebrechung erster Stufe, die im folgenden vorgestellt wird, verbessert die Entropie vom obigen Wert $s = -0.16$ auf $s = -0.01$.

Die Replika–Symmetriebrechung erster Stufe besteht in der folgenden Parametrisierung der Matrix $Q_{\rho\sigma}$.

$$(Q_{\rho\sigma}) = \begin{pmatrix} \begin{pmatrix} 1 & q_1 & q_1 \\ q_1 & 1 & q_1 \\ q_1 & q_1 & 1 \end{pmatrix} & q_0 & q_0 \\ q_0 & \begin{pmatrix} 1 & q_1 & q_1 \\ q_1 & 1 & q_1 \\ q_1 & q_1 & 1 \end{pmatrix} & q_0 \\ q_0 & q_0 & \begin{pmatrix} 1 & q_1 & q_1 \\ q_1 & 1 & q_1 \\ q_1 & q_1 & 1 \end{pmatrix} \end{pmatrix} \tag{3.39}$$

[1] Globale Ableitung und partielle Ableitung nach der Temperatur sind gleich, weil f am Sattelpunkt bezüglich der Ordnungsparameter $Q_{\rho\sigma}$ betrachtet wird.

Die n Replika sind in $\frac{n}{m}$ Gruppen von m Replika aufzuteilen.

Man setzt

$$Q_{\rho\sigma} = 1, \quad \text{wenn} \quad \rho = \sigma$$

$$Q_{\rho\sigma} = q_1,$$

wenn $\rho \neq \sigma$ und beide derselben Gruppe angehören.

$$Q_{\rho\sigma} = q_0,$$

wenn ρ und σ verschiedenen Gruppen angehören.

Im vorliegenden Beispiel ist $n = 9$ und $m = 3$. Es gibt $\frac{n}{m} = 3$ Gruppen von 3 Replika.

Die physikalische Interpretation des obigen Ansatzes, die in [Me+87] ausführlich dargelegt ist, gebe ich hier kurz wieder. Im thermodynamischen Gleichgewicht weist die freie Energie viele lokale Minima auf, die durch Barrieren der Höhe N voneinander getrennt sind. Im Limes $N \to \infty$ gibt es dann viele globale Minima der freien Energie pro Spin f. Der Phasenraum zerfällt vollständig in sogenannte reine Phasen. Jede reine Phase ist eineindeutig einem globalen Minimum von f zugeordnet [EnHe84]. Bei endlichen N kann man sich eine solche reine Phase als Menge von Zuständen vorstellen, in die eine serielle Glauber–Dynamik hineinläuft. Je nach Wahl des Anfangszustandes wird die Dynamik eine andere Phase ansteuern. Da Barrieren der Größenordnung N die reinen Phasen voneinander trennen, verharrt das System in der Phase, wenn die Beobachtungszeit nicht allzu groß wird.

Im Zusammenhang mit der Replika–Symmetriebrechung erster Stufe stellen wir uns die reinen Phasen als gleichartig vor. Zwischen zwei Zuständen a und b in einer reinen Phase liegt die mittlere Überlappung $q_{ab} = q_1$ vor, während Zustände c und d verschiedener reiner Phasen die mittlere Überlappung $q_{cd} = q_0$ aufweisen. Die Wahrscheinlichkeitsverteilung der Überlappungen

$$q_{ab} = \frac{1}{N} \sum_{j=1}^{N} S_j^a S_j^b \tag{3.40}$$

besteht aus zwei δ–Spitzen, die mit Hilfe des Parameters m gewichtet werden. Im Replika–Limes $n \to 0$ liegt m im Intervall $[0, 1]$ (siehe [Me+87]). Für die Wahrscheinlichkeitsverteilung gilt

$$P(q) = m\delta(q - q_0) + (1 - m)\delta(q - q_1) \tag{3.41}$$

Zusammenfassend können wir sagen, daß der Ansatz der Replika – Symmetriebrechung der Kompliziertheit des Phasenraums im thermodynamischen Gleichgewicht Rechnung trägt. Die Wahrscheinlichkeitsverteilung der Überlappung q_{ab} zweier Zustände a und b ergibt sich aus dem Ansatz für die Matrix $Q_{\rho\sigma}$ der Überlappung zweier Replika ρ und σ. Im Falle eines Modells mit Replika–Symmetrie wie dem Perzeptron ist $q_0 = q_1$. Hier ist der Phasenraum der Kopplungen von einfacher Struktur, es gibt nur eine reine Phase. Beim SK–Modell hingegen ist der Phasenraum im thermodynamischen Gleichgewicht von so komplizierter Struktur, daß man über das einfache Schema der RSB1 hinausgehen muß. Es

ist hier nötig, die obige Parametrisierung für die Blöcke der Länge m zu wiederholen. Setzt man die Parametrisierung der neu entstandenen Blöcke immer weiter fort, so erhält man im Grenzfall die komplette Parisi–Lösung. Sie fußt im Endeffekt auf der Annahme, daß der Ordnungsparameter für Spingläser die Wahrscheinlichkeitsverteilung $P(q)$ der Überlappungen ist. Diese Annahme ist ständiger Gegenstand der Forschung [EnHe84], [En+92].

3.4.2 Die Anwendung der RSB1 auf die optimale Verdünnung

Wir wenden Parisis Parametrisierungsansatz auf das Problem der optimalen Verdünnung an. In die Matrizen $Q_{\rho\sigma}$ und $t_{\rho\sigma}$ in Gl.(3.19) gehen dann die Variablen q_0, q_1 sowie T, t_0, t_1 ein. T steht dabei in der Hauptdiagonale von $t_{\rho\sigma}$. Der Parameter m ist bei beiden Matrizen gleich, da $t_{\rho\rho}$ die zu $Q_{\rho\sigma}$ konjugierte Variable ist (siehe [Bo+90]). Für ψ_ρ machen wir wieder den Ansatz

$$\psi_\rho = \psi, \quad \rho = 1, \ldots, n \tag{3.42}$$

Um den Ansatz in Gl.(3.19) einzusetzen, verwenden wir die folgende Formel, die im Anhang B bewiesen wird. **Formel zur RSB1:**

Sei $\mathrm{Tr}_{\{S^\rho\}}$ eine Spur über Größen S^ρ. Sie faktorisiere gemäß

$$\mathrm{Tr}_{\{S^\rho\}} = \prod_{\rho=1}^{n} \mathrm{Tr}_{S^\rho}$$

Dann gilt in RSB1:

$$\ln\left(\mathrm{Tr}_{\{S^\rho\}} \exp\left(\gamma \sum_{\rho\leq\sigma} Q_{\rho\sigma} S^\rho S^\sigma\right)\right) = \tag{3.43}$$

$$n \cdot \int_{-\infty}^{+\infty} Dz \cdot \frac{1}{m} \cdot$$

$$\ln\left[\int_{-\infty}^{+\infty} Dy \cdot \left(\mathrm{Tr}_S \exp\left(S\left(y\sqrt{\gamma(q_1-q_0)} + z\sqrt{\gamma q_0}\right) + \gamma\left(Q - \frac{1}{2}q_1\right)S^2\right)\right)^m\right] +$$

$$+\mathcal{O}\left(n^2\right) \tag{3.44}$$

Nach Anwendung der Formel auf Gl.(3.19) erhält man mit

$$\sum_{\rho\leq\sigma} Q_{\rho\sigma} t_{\rho\sigma} = nT - \frac{1}{2}n((1-m)q_1 t_1 + m q_0 t_0) + \mathcal{O}\left(n^2\right) \tag{3.45}$$

das Ergebnis $\Phi_R(n) = \mathcal{O}(n)$. Die Entropie

$$s = \left.\frac{\partial \Phi_R}{\partial n}\right|_{n=0} \tag{3.46}$$

ergibt sich dann zu

$$
\begin{aligned}
&s(\alpha,\kappa,f) = \qquad (3.47)\\
&\quad \mathrm{sattel}_{\{q_0,q_1,T,t_0,t_1,m,\psi\}} \quad f\psi - T + \frac{1}{2}((1-m)q_1t_1 + mq_0t_0) + \\
&\quad + \int_{-\infty}^{+\infty} Dz \frac{1}{m} \ln \left[\int_{-\infty}^{+\infty} Dy \cdot \right. \\
&\quad \left. \cdot \left(1 + e^{-\psi} \frac{1}{\sqrt{1-\frac{2T-t_1}{f}}} \cdot \exp\left(\frac{1}{2} \cdot \frac{(y\sqrt{t_1-t_0} + z\sqrt{t_0})^2/f}{1-\frac{2T-t_1}{f}} \right) \right)^m \right] + \\
&\quad + \alpha \cdot \int_{-\infty}^{+\infty} Dz \frac{1}{m} \ln \left[\int_{-\infty}^{+\infty} Dy \left(\Phi\left(-\frac{\kappa + y\sqrt{q_1-q_0} + z\sqrt{q_0}}{\sqrt{1-q_1}} \right) \right)^m \right]
\end{aligned}
$$

Zur Vereinfachung führen wir folgende Variablentransformation durch

$$E = 1 - \frac{2}{f}T \qquad (3.48)$$

E ersetzt T vollständig. Anschließend transformieren wir den Variablensatz (ψ, t_0, t_1) auf (η, s_0, s_1) gemäß

$$
\begin{aligned}
\frac{\eta}{2} &= -\psi - \frac{1}{2}\ln\left(E + \frac{t_1}{f}\right) \\
s_1 &= \frac{t_1/f}{E + t_1/f} \\
s_0 &= \frac{t_0/f}{E + t_1/f} \qquad (3.49)
\end{aligned}
$$

Dadurch sind die Integrale nicht mehr von E abhängig. Die Sattelpunktgleichung der transformierten Funktion s nach E läßt sich dann einfach auflösen zu

$$\frac{\partial s}{\partial E} = 0 \longrightarrow E = \frac{1 - s_1}{1 - s_1 + (1-m)q_1s_1 + mq_0s_0} \qquad (3.50)$$

Damit erhält man das Endergebnis

$$
\begin{aligned}
&s(\alpha,\kappa,f) = \qquad (3.51)\\
&\quad \mathrm{sattel}_{\{q_0,q_1,s_0,s_1,m,\eta\}} \quad -\frac{f}{2}\eta + \frac{f}{2}\ln\left(1 - s_1 + (1-m)q_1s_1 + mq_0s_0\right) + \\
&\quad + \alpha \cdot \int_{-\infty}^{+\infty} Dz \frac{1}{m} \ln \left[\int_{-\infty}^{+\infty} Dy \left(\Phi\left(-\frac{\kappa + y\sqrt{q_1-q_0} + z\sqrt{q_0}}{\sqrt{1-q_1}} \right) \right)^m \right] + \\
&\quad + \int_{-\infty}^{+\infty} Dz \frac{1}{m} \ln \left[\int_{-\infty}^{+\infty} Dy \left(1 + \exp\left(\frac{1}{2} \cdot (y\sqrt{s_1-s_0} + z\sqrt{s_0})^2 + \frac{\eta}{2} \right) \right)^m \right]
\end{aligned}
$$

Wie man leicht sieht, enthalten die 6 Sattelpunktgleichungen bezüglich

$$q_0, q_1, s_0, s_1, m, \eta$$

ebenfalls Doppelintegrale über y und z. s_0, s_1 und η sind Hilfsvariablen. Das bedeutet, daß sie bei gegebenen q_0, q_1, m eindeutig bestimmt sind aus der Lösung des nichtlinearen Gleichungssystems der drei Gleichungen

$$\left(\frac{\partial s}{\partial s_0}, \frac{\partial s}{\partial s_1}, \frac{\partial s}{\partial \eta}\right) = (0,0,0) \tag{3.52}$$

Die Entropie s muß dann bezüglich der drei verbleibenden Variablen q_0, q_1, m optimiert werden. Diese Variablen sind dabei als Variationsparameter aufzufassen. Dadurch wird sichergestellt, daß die RSB1 der vollständigen Parisi–Lösung nahe kommt. Analog zum SK–Modell zeigt sich, daß wegen des Limes $n \to 0$ nicht das Maximum, sondern das *Minimum* der Entropie s bezüglich q_0, q_1 und m gesucht wird.

Bei der Lösung des obigen Problems treten erhebliche numerische Schwierigkeiten auf. Es stellt sich zwar heraus, daß die Replika–Symmetrie gebrochen ist, jedoch ist die Entropie als Funktion von q_0, q_1, m so flach, daß größere Fluktuationen in q_0, q_1 und m nur Unterschiede der Größenordnung 10^{-4} in s hervorrufen. Eine vollständige Lösung des Gleichungssystems aller sechs Sattelpunktvariablen

$$\left(\frac{\partial s}{\partial q_0}, \frac{\partial s}{\partial q_1}, \frac{\partial s}{\partial m}, \frac{\partial s}{\partial s_0}, \frac{\partial s}{\partial s_1}, \frac{\partial s}{\partial \eta}\right) = (0,0,0,0,0,0) \tag{3.53}$$

stößt auf die gleichen Schwierigkeiten [2]. Wir behelfen uns deshalb am kritischen κ mit einem Skalierungsansatz, um die Doppelintegrale im obigen Gleichungssystem in Einfachintegrale zu überführen.

3.4.3 Der Ansatz im Fall $q_1 \to 1$

Der folgende Skalierungsansatz im Fall $q_1 \to 1$ ist analog zu Skalierungsansätzen bei anderen Modellen neuronaler Netzwerke [En+92], [Br+92]. Man nimmt an, daß $m \to 0$ ($q_1 \to 1$), wobei

$$c = \frac{m}{1 - q_1} \tag{3.54}$$

eine Konstante ist. Die Überlappung verschiedener Phasen hat für $m \to 0$ weiterhin einen Grenzwert q_0, $0 < q_0 < 1$. Für die Hilfsvariablen s_0, s_1, η macht man im Limes $m \to 0$ den Ansatz [3]

$$s_0 \cdot m \to t_0, \;\; s_1 \cdot m \to t_1, \;\; \eta \cdot m \to h \tag{3.55}$$

Betrachtet man nun das Gleichungssystem (3.53), so vereinfachen sich im Limes $m \to 0$ die Doppelintegrale zu Einfachintegralen. Dies wird in Anhang C.1

[2] Die rechten Seiten des Gleichungssystems sind von der Größenordnung 10^{-5}. Das bedeutet, daß man die Doppelintegrale auf der linken Seite des Gleichungssystems stets auf eine Mindestgenauigkeit von etwa 10^{-7} berechnen müßte, um die Konvergenz eines Lösungsalgorithmus für nichtlineare Gleichungssysteme zu ermöglichen.

[3] t_0 und t_1 sind in der obigen Gleichung neu eingeführte Variablen, sie haben nichts mit bereits beseitigten Variablen gemein.

besprochen. Nach Durchführung des Limes wird das entstehende Gleichungssystem numerisch stabil. Es liefert die Variablen c, q_0, t_0, t_1 und h. Die Gleichung

$$\frac{\partial s}{\partial m} = 0$$

ist dabei die Bestimmungsgleichung für das kritische α. Das Endergebnis für das Gleichungssystem im Limes $q_1 \to 1$ ist aus Gründen der Übersichtlichkeit in Anhang C.2 wiedergegeben.

3.5 Die Ergebnisse der RSB1–Näherungsrechnung

Mit der oben geschilderten analytischen Rechnung haben wir die RSB1–Näherung des SK–Modells erfolgreich auf das Problem der optimalen Verdünnung angewandt. Die RSB1–Näherung geht dabei von einer einfachen Situation im Phasenraum aus.

Der Phasenraum $\mathcal{P}$ ist die Menge aller Kopplungsvektoren

$$\underline{T} \in R^N$$

die $N(1-f)$ viele Komponenten $T_j = 0$ aufweisen und auf $\sum_{j=1}^{N} {T_j}^2 = Nf$ normiert sind. Wir können $\mathcal{P}$ als Vereinigung von $\binom{N}{Nf}$ vielen Nf–dimensionalen Gardner–Kugeln mit Radius $\sqrt{Nf}$ auffassen. Zu jedem zugelassenen Verdünnungsvektor

$$\underline{c} = (c_1, \ldots, c_N)^T, \quad c_j \in \{0, 1\} \ \forall j$$

gehört eine solche Gardner–Kugel.

Unterhalb der kritischen Speicherkapazität tragen sehr viele Gardner–Kugeln Perzeptron–Vektoren bei, die die Klassifikationsaufgabe erfüllen. Fordert man ein höheres κ, so schrumpfen die zugelassenen Gebiete auf den Oberflächen der einzelnen Gardner–Kugeln. Einige Kugeln scheiden vollständig aus. Die Überlappung

$$q = \frac{1}{Nf} \sum_{j=1}^{N} c_j^a T_j^a c_j^b T_j^b \tag{3.56}$$

zweier Perzeptron–Vektoren hat im allgemeinen eine komplizierte Wahrscheinlichkeitsverteilung $P(q)$. Wenden wir die Parisi–Theorie des SK–Modells auf unser Problem an, so gehen wir von der Annahme aus, daß $P(q)$ die thermodynamischen Eigenschaften unseres Problems vollständig beschreibt. Genauer ausgedrückt bedeutet dies, daß die Entropie s (siehe Gl.(3.17)) nur von dem „Ordnungsparameter“ $P(q)$ abhängt. Wenn wir die komplizierte Theorie auf die RSB1–Näherung einschränken, so haben wir von der einfachen Form

$$P(q) = m\delta(q - q_0) + (1 - m)\delta(q - q_1) \tag{3.57}$$

auszugehen. Der Phasenraum gliedert sich in mehrere reine Phasen, wobei die Überlappung von zwei Zuständen in einer reinen Phase q_1 beträgt. Die Überlappung von Zuständen verschiedener reiner Phasen ist q_0. Für ein unterkritisches

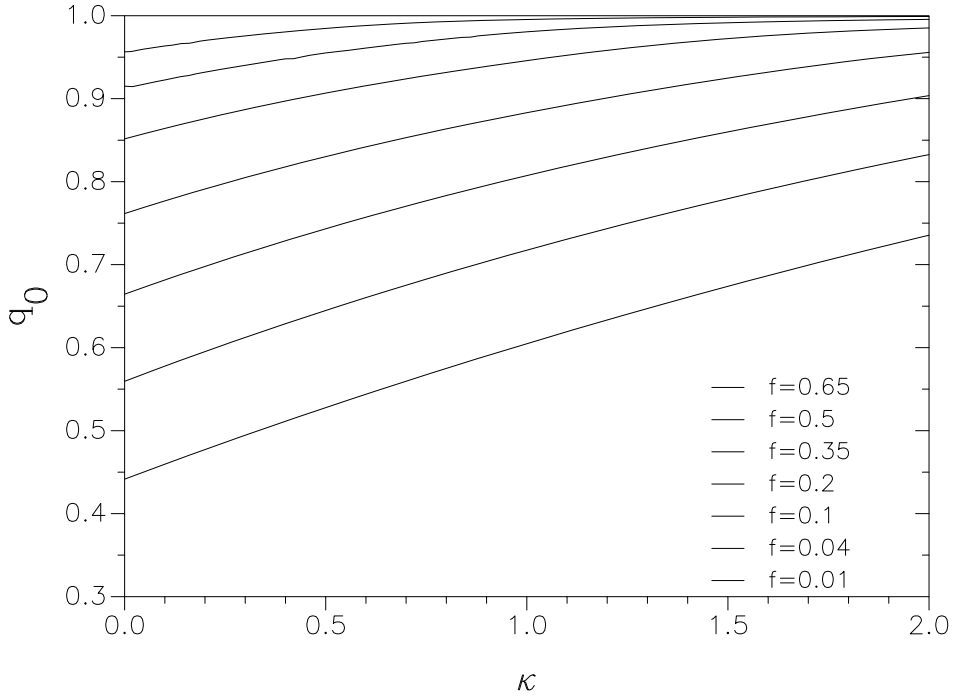


Abbildung 3.4: Der Ordnungsparameter q_0 der RSB1–Näherung für $f = 0.01$, 0.04, 0.1, 0.2, 0.35, 0.5, 0.65 in aufsteigender Reihenfolge

κ können dabei durchaus Zustände aus *verschiedenen* Gardner–Kugeln eine reine Phase bilden. Ist aber die kritische Stabilität κ_c erreicht, so gibt es in jeder Gardner–Kugel höchstens eine Lösung. Unsere Theorie sagt aus, daß $q_1 = 1$, d.h. jede Phase besteht nur noch aus einem Vektor. Wegen $m \to 0$ gilt aber auch

$$P(q) = 0 + 1 \cdot \delta(q - 1)$$

Die Theorie behauptet also darüber hinaus, daß es am kritischen Punkt nur einen Perzeptron–Vektor gibt, es gibt nur einen Vektor der optimalen Verdünnung. Die Überlappung q_0 friert dabei auf einem von f abhängigen Wert ein, siehe Bild 3.4.

Für größere κ–Werte und höhere f–Werte sehen wir, daß $q_0 \to 1$. Das bedeutet, daß in dieser Region die Vorhersage der replikasymmetrischen Lösung vertrauenswürdiger ist. In Bild 3.5 sieht man, daß die $\kappa(\alpha)$–Kurven für große f und kleine α sehr steil werden. Wir erwarten deshalb, daß sich die RSB1–Näherung besonders im flachen Bereich, das heißt im Bereich kleiner κ und kleiner f, stark von der replikasymmetrischen Näherung unterscheidet. Dies wird besonders deutlich, wenn wir α bzw. α_{eff} der RSB1–Näherung mit den in den vorherigen Abschnitten gewonnenen Ergebnissen am Punkt $\kappa = 0$ vergleichen. In den Abbildungen 3.6 und 3.7 ist der Vergleich mit der oberen Schranke, der replikasymmetrischen Näherung, der AT–Linie und der Geraden $\alpha = 2f$ des zufälligen Verdünnens dargestellt. Die RSB1–Näherung liegt dabei wie erwartet oberhalb der AT–Linie der replikasymmetrischen Näherung.

Um einen Eindruck davon zu gewinnen, wie gut die RSB1–Näherung mit

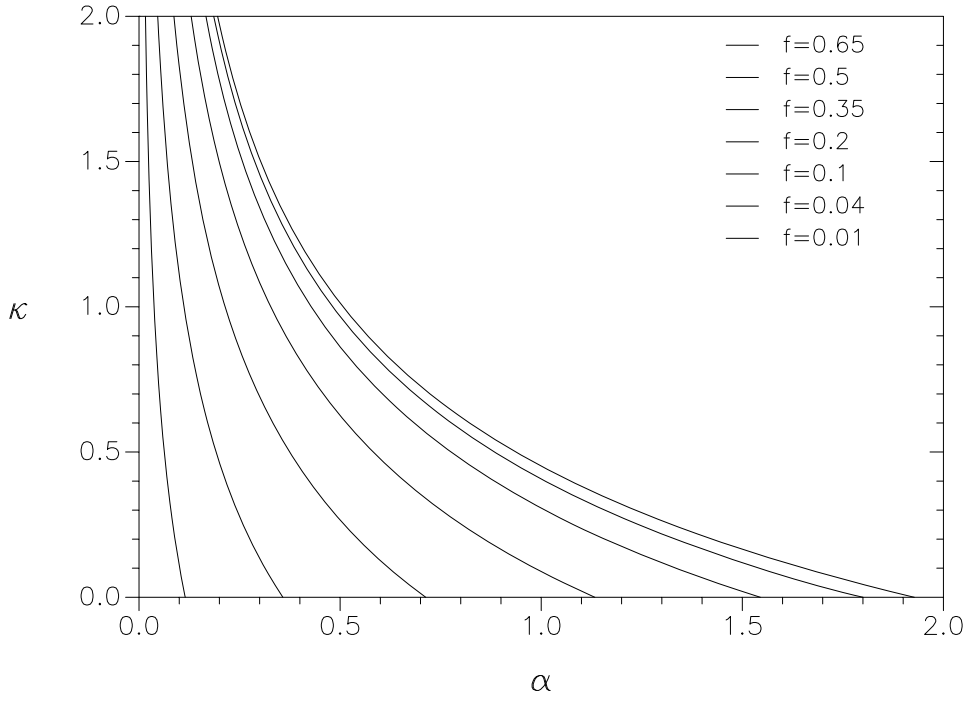


Abbildung 3.5: Die Stabilität κ in RSB1–Näherung für $f = 0.01, 0.04, 0.1, 0.2, 0.35, 0.5, 0.65$ in aufsteigender Reihenfolge

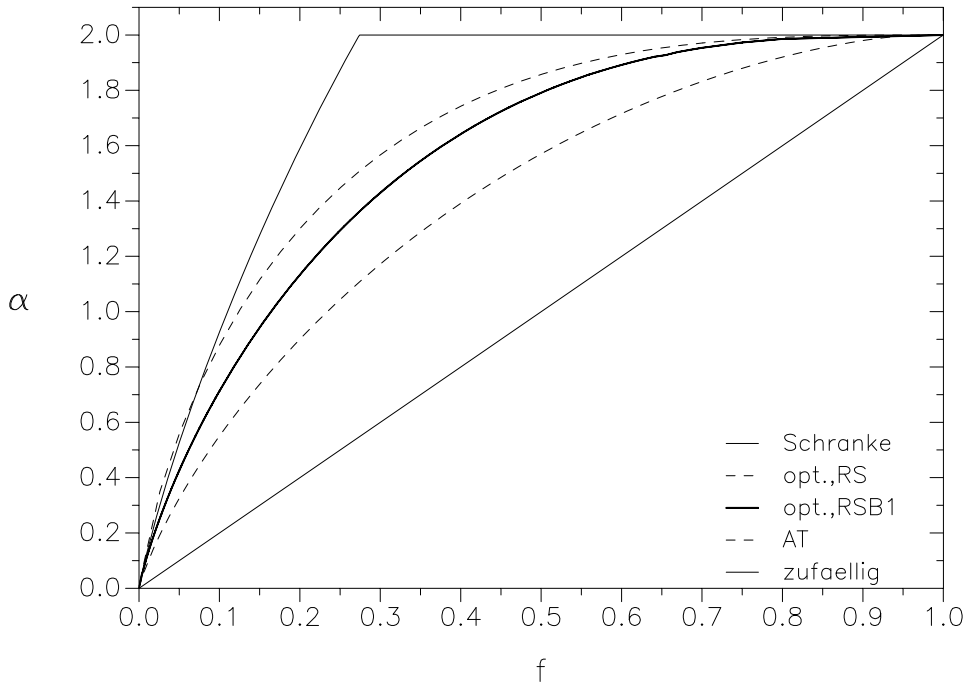


Abbildung 3.6: Die Speicherkapazität $\alpha(\kappa = 0)$ in RSB1–Näherung im Vergleich mit dem replikasymmetrischen Ergebnis von Bouten et al. [Bo+90].

dem Ansatz für $q_1 \to 1$ ist, wird in Abbildung 3.8

$$\frac{\alpha(\kappa = 0)}{-f \ln f}$$

der RSB1–Näherung für sehr kleine f betrachtet. Im Gegensatz zur replikasymmetrischen Näherung genügt die RSB1–Näherung für alle in der numerischen Rechnung erreichten Werte von f der oberen Schranke; wir beobachten sogar eine Annäherung der beiden Kurven, wenn f kleiner wird. Der Grenzwert der oberen Schranke

$$\lim_{f \to 0} \frac{\alpha(\kappa = 0)}{-f \ln f} \leq \frac{1}{\ln 2}$$

ist im Diagramm 3.8 eine konstante Funktion (siehe Gl.(3.11)).

In diesem Kapitel wurde also eine Näherung für die maximale Speicherkapazität eines optimal verdünnten Perzeptrons gewonnen, die mit allen bisherigen Ergebnissen konsistent ist. Sie liegt oberhalb der AT–Linie der replikasymmetrischen Näherung, das heißt, sie liegt wie erwartet in dem Bereich, in dem die replikasymmetrische Lösung instabil ist. Des weiteren genügt sie für alle f einer oberen Schranke und stimmt im Limes $f \to 0$ mit dieser überein. Es ist jedoch nicht bekannt, inwiefern weitere Stufen der Replika–Symmetriebrechung das gewonnene α weiter verringern. Des weiteren ist zu betonen, daß die in diesem Kapitel durchgeführte Rechnung darauf beruht, daß eine Analogie des optimal verdünnten Perzeptrons zum SK–Modell angenommen wurde. Obwohl diese Annahme sehr oft zu akzeptablen Ergebnissen führt, ist sie jedoch nicht zwingend.

Die Näherung wird in den nächsten beiden Kapiteln mit praktikablen Verdünnungsalgorithmen des Perzeptrons optimaler Stabilität verglichen. Der Vorteil dieser Algorithmen ist, daß zugehörige analytische Ergebnisse mit Simulationen überprüfbar sind. Außerdem wird sich zeigen, daß die den Verdünnungsalgorithmen zugrunde liegende Theorie einfacher ist.

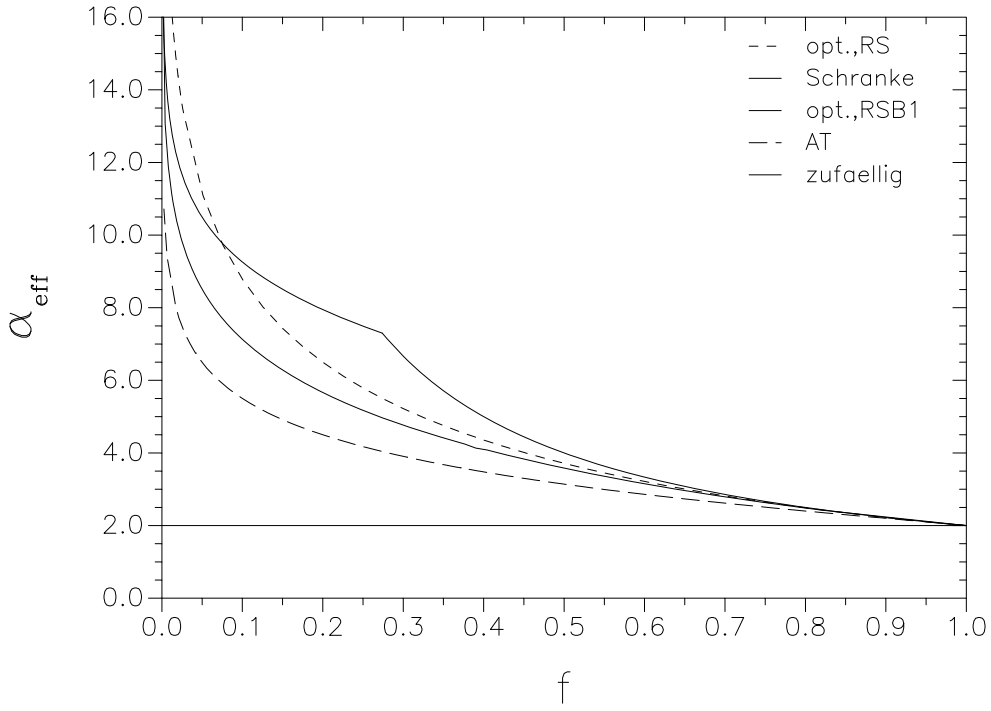


Abbildung 3.7: Die effektiven Speicherkapazitäten $\alpha_{eff}(\kappa = 0)$ im Vergleich (vgl. Abb. 3.6).

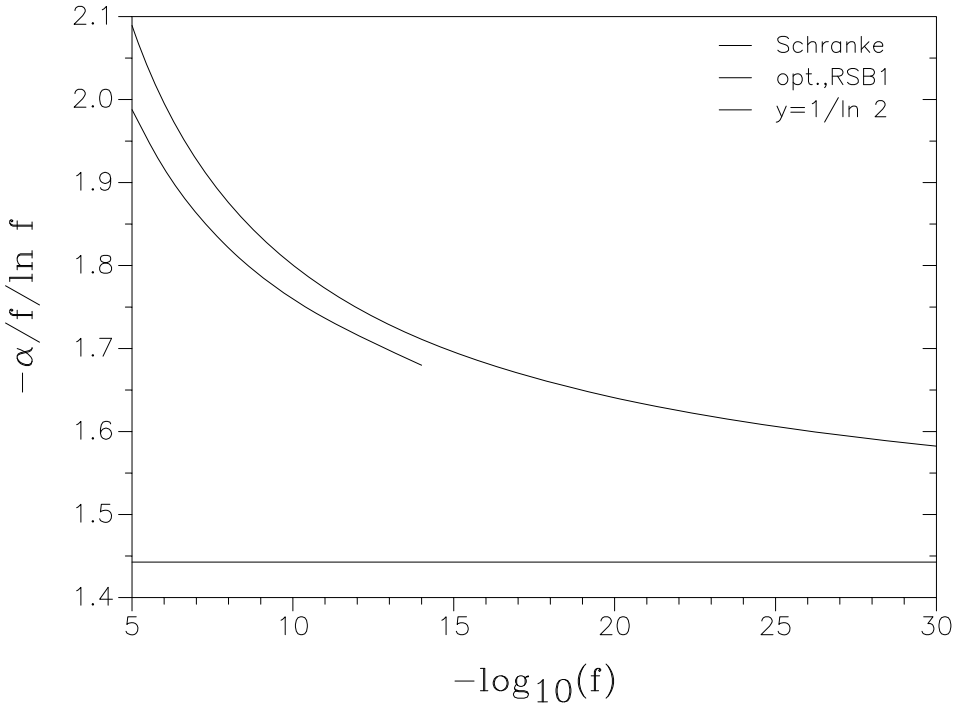


Abbildung 3.8: $y = -\frac{\alpha(\kappa=0)}{f \ln f}$ der RSB1–Näherung im Vergleich mit der oberen Schranke. Die Näherung genügt für alle in einer numerischen Rechnung erreichbaren f–Parameter der oberen Schranke. Die Gerade $y = \frac{1}{\ln 2}$ stellt dabei den Grenzwert der oberen Schranke dar (Gl.(3.11)).

Kapitel 4

Der Quersummen-verdünnungsalgorithmus

Nachdem im vorherigen Kapitel Überlegungen darüber angestellt wurden, welche effektiven Speicherkapazitäten α_{eff} maximal möglich sind, wird nun ein praktikabler Verdünnungsalgorithmus vorgeschlagen. Er besteht aus zwei Schritten. Im ersten Schritt werden die zu entfernenden Neuronen ermittelt, indem man einen geeignete Quersumme über die Muster betrachtet. Auf den verbleibenden Neuronen wird das Perzeptron optimaler Stabilität nachgelernt. Da die Wahl der Verdünnung sehr schnell vonstatten geht, ist der Rechenaufwand vergleichbar mit dem Rechenaufwand effektiver Lernverfahren des Perzeptrons optimaler Stabilität. Dies erlaubt die Überprüfung analytischer Ergebnisse durch numerische Simulationen.

In Abschnitt 4.1 wird der Verdünnungsalgorithmus explizit definiert, wobei wieder die Notation aus Abschnitt 1.4 vorausgesetzt wird. Dabei wird herausgestellt, daß ein Verfahren der „eingefrorenen Verdünnung“ vorliegt. In Abschnitt 4.2 wird die analytische Rechnung vorgestellt. Das Ergebnis der Rechnung wird in Abschnitt 4.3 mit Computersimulationen verglichen. Es wird gezeigt, daß mit dem Quersummenverdünnungsalgorithmus effektive Speicherkapazitäten $\alpha_{eff} > 2$ erreicht werden.

4.1 Die Definition des Algorithmus

4.1.1 Motivation

Für $\alpha < 2$ existiert für $N \to \infty$ mit Wahrscheinlichkeit 1 ein Vektor des Perzeptrons optimaler Stabilität, der die Zwangsbedingungen mit einer Stabilität κ erfüllt:

$$\frac{1}{\sqrt{\sum_{j=1}^{N} {J_j}^2}} \sum_{j=1}^{N} J_j \sigma_j^\mu \geq \kappa \quad \forall \mu \tag{4.1}$$

Wir stellen uns die Aufgabe, diejenigen Komponenten j der transformierten Muster $\left\{\sigma_j^\mu\right\}$ zu bestimmen, die die obigen Zwangsbedingungen am wenigsten beeinflussen. Wenn dies obendrein ohne Kenntnis des ursprünglichen Perzeptron–

Vektors $\underline{J}$ geschehen soll, ist der naheliegendste Ansatz, die Plätze j wegzulassen, welche vom Betrag her die kleinsten *Quersummen*

$$H_j = \frac{1}{\sqrt{N\alpha}} \sum_{\mu=1}^{p} \sigma_j^\mu \tag{4.2}$$

aufweisen [1]. Diese Quersummen sind aber nichts anderes als die auf einem Einschichtnetzwerk mit Mustern $\underline{\xi}^\mu$ und Ausgaben $\{S^\mu\}$ definierten *Hebb–Kopplungen* (siehe Gl. (1.9) mit $\xi_i^\mu = S^\mu$ für das „Ausgabeneuron" i).

Schneideverfahren mit Hebb–Kopplungen sind bisher ausschließlich in Zusammenhang mit Attraktornetzwerken behandelt worden [So86], [Do+89], [Mu92]. Wir verwenden das Schneideverfahren im folgenden als Vorstufe zu einem Nachlernprozeß, um alle in ein Einschichtnetzwerk eingespeisten Muster perfekt zu klassifizieren.

4.1.2 Der Algorithmus

Vorgelegt seien p transformierte Muster $\underline{\sigma}^\mu$. Das Netzwerk soll so verdünnt werden, daß $N \cdot f$ Neuronen übrig bleiben.

1. Berechne als Quersummen für die zu entfernenden Plätze j die Hebb–Kopplungen

   $$H_j = \frac{1}{\sqrt{N\alpha}} \sum_{\mu=1}^{p} \sigma_j^\mu \tag{4.3}$$

   Ordne die Beträge $|H_j|$ in aufsteigender Reihenfolge und ordne die Platzvektoren

   $$\vec{\sigma}_j = \left(\sigma_j^1, \ldots, \sigma_j^p\right)^T \tag{4.4}$$

   dementsprechend um mit neuen Indizes $m(j)$.

2. Entferne alle Plätze j mit den kleinsten Beträgen $|H_j|$, d.h. entferne alle $\vec{\sigma}_m$ mit

   $$m \leq N(1-f) \tag{4.5}$$

   Dabei sei die Schranke w für die Beträge der Hebb–Kopplungen als

   $$w = |H_m| \tag{4.6}$$

   definiert.

3. Lerne auf den verbleibenden Nf Plätzen das Perzeptron optimaler Stabilität, um sicherzustellen, daß alle Muster wieder korrekt klassifiziert werden.

Nach der Verdünnungsprozedur liegt also der Verdünnungsvektor

$$\underline{c} = (c_1, \ldots, c_N)^T \tag{4.7}$$

[1] Die Skalierung mit $\frac{1}{\sqrt{N\alpha}}$ wurde gewählt, um zu gewährleisten, daß $H_j = \mathcal{O}(1)$.

fest. Er ist wegen der obigen Definition der Schranke w gegeben durch

$$c_j = \Theta(|H_j| - w), \quad j = 1, \ldots, N \tag{4.8}$$

Da die Kopplungen also erst optimiert werden, *nachdem* der Verdünnungsvektor feststeht, liegt eine „eingefrorene Verdünnung“ vor (siehe Abschnitt 3.2.1). Der Aufwand bei der Auswahl der c_j wird sich dabei in dem Ergebnis $\alpha_{eff} > 2$ widerspiegeln.

4.2 Die analytische Rechnung

4.2.1 f als Funktion von w

Im thermodynamischen Limes $N \to \infty$ läßt sich bei gegebenem Verdünnungsparameter f die zugehörige Schranke w leicht ermitteln. Für f gilt nämlich einfach

$$f = \frac{1}{N}\sum_{j=1}^{N} c_j = \frac{1}{N}\sum_{j=1}^{N} \Theta\left(\left|\frac{1}{\sqrt{N\alpha}}\sum_{\mu=1}^{p}\sigma_j^\mu\right| - w\right) \tag{4.9}$$

Man sieht leicht, daß f im Limes $N \to \infty$ selbstmittelnd ist [2]. Wir berechnen der Einfachheit halber den Mittelwert von f und schreiben

$$\begin{aligned}\langle\langle f\rangle\rangle_{\{\sigma_j^\mu\}} = & \\ & \left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty} dH_j\right)\frac{1}{N}\sum_{j=1}^{N}\Theta(|H_j| - w)\cdot\left\langle\left\langle\prod_{j=1}^{N}\delta\left(H_j - \frac{1}{\sqrt{N\alpha}}\sum_{\mu=1}^{p}\sigma_j^\mu\right)\right\rangle\right\rangle_{\{\sigma_j^\mu\}} \\ = & \left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty} DH_j\right)\frac{1}{N}\sum_{j=1}^{N}\Theta(|H_j| - w) = 2\int_{-\infty}^{-w} DH\end{aligned} \tag{4.10}$$

Im Limes $N \to \infty$ können wir dann mit Formel (A.11) setzen

$$f = 2\Phi(-w) \tag{4.11}$$

Wir wissen also, bei welcher Schwelle w wir die Beträge der Hebb–Kopplungen schneiden müssen, um einen Verdünnungsparameter f zu erhalten.

[2] Die etwas längliche Berechnung der Wahrscheinlichkeitsdichte

$$P(f) = \left\langle\left\langle\delta\left(f - \frac{1}{N}\sum_{j=1}^{N} c_j\right)\right\rangle\right\rangle_{\{\sigma_j^\mu\}}$$

liefert nichts anderes als eine Bestätigung des zentralen Grenzwertsatzes für einen Random Walk [Reif], [Fe68].

4.2.2 Der Ansatz für die Gardner–Rechnung und die Mittelung über die Muster

Um das Nachlernen auf den verbleibenden Plätzen k, $k = 1, \ldots, Nf$, zu erfassen, formuliere ich das zugehörige Gardner–Volumen

$$V = \frac{1}{C_{norm}} \left(\prod_{k=1}^{Nf} \int_{-\infty}^{+\infty} \frac{dJ_k}{\sqrt{2\pi}} \right) \delta \left(\sum_{k=1}^{Nf} {J_k}^2 - Nf \right) \cdot \prod_{\mu=1}^{p} \theta \left(\frac{1}{\sqrt{Nf}} \sum_{k=1}^{Nf} J_k \eta_k^\mu - \kappa \right) \tag{4.12}$$

Dabei ist die Normierungskonstante C_{norm} gegeben durch

$$C_{norm} = \left(\prod_{k=1}^{Nf} \int_{-\infty}^{+\infty} \frac{dJ_k}{\sqrt{2\pi}} \right) \delta \left(\sum_{k=1}^{Nf} {J_k}^2 - Nf \right)$$

Die η_k^μ sind die transformierten Muster σ_j^μ auf den verbleibenden Plätzen. Um das Verdünnungsverfahren korrekt in die Rechnung einzubeziehen, multipliziere ich V mit dem Faktor

$$1 = \left(\prod_{l=Nf+1}^{N} \int_{-\infty}^{+\infty} \frac{dJ_l}{\sqrt{2\pi}} \right) \exp \left(-\frac{1}{2} \sum_{l=Nf+1}^{N} {J_l}^2 \right) \tag{4.13}$$

Die Komponenten des Verdünnungsvektors unterscheiden dann gemäß

$$c_j = \Theta(|H_j| - w) = \Theta \left(\left| \frac{1}{\sqrt{N\alpha}} \sum_{\mu=1}^{p} \sigma_j^\mu \right| - w \right) \tag{4.14}$$

zwischen den Indizes k und l.

Für das Phasenraumvolumen gilt also

$$\begin{aligned} V \;=\; & \frac{1}{C_{norm}} \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} \frac{dT_j}{\sqrt{2\pi}} \right) \delta \left(\sum_{j=1}^{N} c_j {T_j}^2 - Nf \right) \cdot \exp \left(-\frac{1}{2} \sum_{j=1}^{N} (1 - c_j) {T_j}^2 \right) \cdot \\ & \cdot \prod_{\mu=1}^{p} \theta \left(\frac{1}{\sqrt{Nf}} \sum_{j=1}^{N} c_j T_j \sigma_j^\mu - \kappa \right) \end{aligned} \tag{4.15}$$

Im Vergleich mit dem Phasenraumvolumen für die optimale Verdünnung (V_{ges} aus Gl.(3.16)) stellen wir nocheinmal fest, daß nicht über die c_j summiert wird. Die c_j sind hier eingefroren, da sie nur von den *vorgegebenen* Mustern abhängen. V ist das Phasenraumvolumen für *ein* Perzeptronproblem auf wohldefinierten Plätzen k. Demzufolge ist die Entropie

$$s = \lim_{N \to \infty} \frac{1}{N} \ln V \tag{4.16}$$

wieder selbstmittelnd, es ist gleich, ob wir einen festen Mustersatz mit großem N betrachten, oder ob wir für große N über die gaußverteilten Muster mitteln:

$$\lim_{N \to \infty} s = \lim_{N \to \infty} \langle \langle s \rangle \rangle_{\{\sigma_j^\mu\}} \tag{4.17}$$

Man könnte nun einwenden, daß das Phasenraumvolumen in Gl.(4.12) nur Nf viele Muster enthält und somit bei einer Berechnung von s nur das Ergebnis des zufälligen Verdünnens herauskommt. Dem ist entgegenzuhalten, daß die Plätze k in Gl.(4.12) von dem gesamten Mustersatz abhängen. Dieser Korrelation wird nur Rechnung getragen, wenn man in der Darstellung (4.15) beachtet, daß auch die Muster auf den weggenommenen Plätzen in die c_j eingehen.

Ich wende wieder die Replika–Methode aus Abschnitt 1.3.2 an und berechne

$$\Phi_R(n) = \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle V^n \rangle\rangle_{\{\sigma_j^\mu\}} \tag{4.18}$$

Dazu führe ich die Fourierdarstellung der θ–Funktion für die Zwangsbedingungen ein. Die Hebb–Kopplungen H_j in den c_j werden analog zu Gl.(A.3) ersetzt. Die $\hat{h}_j$ sind dabei die zu H_j konjugierten Variablen. Es gilt

$$\begin{aligned}
&\langle\langle V^n \rangle\rangle = \\
&\left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} dH_j \int_{-\infty}^{+\infty} \frac{d\hat{h}_j}{2\pi}\right) \exp\left(i \sum_{j=1}^{N} \hat{h}_j H_j\right) \left(\prod_{j,\rho} \int_{-\infty}^{+\infty} DT_j^\rho\right) \cdot \\
&\cdot \left(\prod_{\rho=1}^{n} \delta\left(\sum_{j=1}^{N} c_j \left(T_j^\rho\right)^2 - Nf\right)\right) \left(\prod_{\mu,\rho} \int_{\kappa}^{\infty} dX_\mu^\rho \int_{-\infty}^{+\infty} \frac{dx_\mu^\rho}{2\pi}\right) \exp\left(i \sum_{\mu,\rho} x_\mu^\rho X_\mu^\rho\right) \cdot \\
&\cdot \left\langle\left\langle \exp\left[-\frac{i}{\sqrt{N}} \sum_{j,\mu} \left(\sum_{\rho=1}^{n} x_\mu^\rho \cdot \frac{1}{\sqrt{f}} c_j T_j^\rho + \hat{h}_j \cdot \frac{1}{\sqrt{\alpha}}\right) \sigma_j^\mu\right] \right\rangle\right\rangle
\end{aligned} \tag{4.19}$$

Dabei sind die c_j Funktionen der Integrationsvariablen H_j gemäß

$$c_j(H_j) = \theta(|H_j| - w) \tag{4.20}$$

Die Mustermittelung liefert

$$\begin{aligned}
&\langle\langle \exp[\ldots] \rangle\rangle = \\
&\exp\left[-\frac{1}{2N} \sum_{j,\mu} \left(\sum_{\rho,\sigma} x_\mu^\rho \frac{1}{f} c_j T_j^\rho T_j^\sigma x_\mu^\sigma + \frac{2}{\sqrt{\alpha f}} \sum_{\rho=1}^{n} x_\mu^\rho c_j T_j^\rho \hat{h}_j + \frac{1}{\alpha} \hat{h}_j{}^2\right)\right]
\end{aligned} \tag{4.21}$$

4.2.3 Das Ergebnis für allgemeine n

Zur Entkopplung der Indizes j von den Indizes μ führe ich den Ordnungsparameter

$$Q_{\rho\sigma} = \frac{1}{Nf} \sum_{j=1}^{N} c_j T_j^\rho T_j^\sigma \tag{4.22}$$

und das Feld

$$\hat{K}_\rho = \frac{1}{N} \sum_{\mu=1}^{p} x_\mu^\rho \tag{4.23}$$

mit konjugierten Größen $t_{\rho\sigma}$, k_ρ analog zu Gl.(A.3) ein. $Q_{\rho\sigma}$ stellt wieder die mittlere Überlappung zweier Lösungen der Klassifizierungsaufgabe auf den verbleibenden Plätzen dar. Wegen der Normierung des Phasenraumvolumens ist wieder $Q_{\rho\rho} = 1\ \forall\rho$.

Nach der Berechnung der Gauß–Integrale in den $\hat{h}_j$ nach Formel (A.8) kann die Sattelpunktmethode zur Berechnung von $\Phi_R(n)$ angewendet werden. Nach der Transformation

$$\hat{K}_\rho \to \frac{\hat{K}_\rho}{i} \tag{4.24}$$

erhält man das Ergebnis für allgemeine n:

$$\begin{aligned}
\Phi_R(n) \;=\;& \operatorname{sattel}_{\{Q_{\rho\sigma},t_{\rho\sigma},\hat{K}_\rho,k_\rho\}} \\
& -\sum_{\rho\le\sigma} t_{\rho\sigma}Q_{\rho\sigma} + \sum_{\rho=1}^{n} k_\rho\hat{K}_\rho - \frac{1}{2\alpha}\sum_{\rho,\sigma}\hat{K}_\rho Q_{\rho\sigma}\hat{K}_\sigma + \\
& + \ln\left[1 - f + \int_{-\infty}^{+\infty} DH\theta(|H| - w)\left(\prod_{\rho=1}^{n}\int_{-\infty}^{+\infty} DT_\rho\right)\cdot\right. \\
& \left.\cdot\exp\left(\frac{1}{f}\sum_{\rho\le\sigma} T_\rho t_{\rho\sigma}T_\sigma - \frac{H}{\sqrt{\alpha f}}\sum_{\rho=1}^{n}\hat{K}_\rho T_\rho\right)\right] + \\
& + \alpha\ln\left[\left(\prod_{\rho=1}^{n}\int_{\kappa}^{\infty}\frac{dX_\rho}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{dx_\rho}{\sqrt{2\pi}}\right)\cdot\right. \\
& \left.\cdot\exp\left(-\frac{1}{2}\sum_{\rho,\sigma} x_\rho Q_{\rho\sigma}x_\sigma + i\sum_{\rho=1}^{n} x_\rho(X_\rho - k_\rho)\right)\right]
\end{aligned} \tag{4.25}$$

4.2.4 Die Annahme der Replika–Symmetrie und die Lösung der Sattelpunktgleichungen

Man setzt das Ergebnis mit der Annahme der Replika–Symmetrie nach $n = 0$ fort:

$$\begin{aligned} Q_{\rho\sigma} &= q, \quad \rho, \sigma = 1, \ldots, n, \quad \rho \neq \sigma \\ k_\rho &= k, \quad \rho = 1, \ldots, n \end{aligned} \tag{4.26}$$

und analog für $t_{\rho\sigma}$, $t_{\rho\rho}$, $\hat{K}_\rho$.

Die Mittelwerte über die Matrizen $t_{\rho\sigma}$ und $Q_{\rho\sigma}$ lassen sich nach Formel (B.1) berechnen, wenn man in dem RSB1–Ergebnis einfach $q_0 = q_1$ setzt. Der Parameter m kürzt sich dann heraus. Man erhält als Näherung für die Entropie

$$\begin{aligned} s \;=\; & \operatorname{sattel}_{\{q,t,T,k,\hat{K}\}} \\ & -T + \frac{1}{2}qt + k\hat{K} - \frac{1}{2\alpha}\hat{K}^2(1-q) - \frac{f}{2}\ln\left(1 + \frac{t-2T}{f}\right) + \\ & + \frac{1}{2}\cdot\frac{\hat{K}^2 E}{\alpha f\left(1+\frac{t-2T}{f}\right)} + \frac{1}{2}\cdot\frac{t}{1+\frac{t-2T}{f}} + \alpha\int_{-\infty}^{+\infty} Dz \ln\Phi\left(-\frac{\kappa - k + z\sqrt{q}}{\sqrt{1-q}}\right) \end{aligned} \tag{4.27}$$

Dabei ist E durch die Schranke w gegeben. Es gilt

$$E = 2\cdot\int_w^\infty DH\cdot H^2 = f + \sqrt{\frac{2}{\pi}}w\exp\left(-\frac{1}{2}w^2\right) \tag{4.28}$$

nach Formel (A.21).

Die Sattelpunktgleichungen bezüglich der Hilfsvariablen $\hat{K}, t, T$, sind algebraisch. Sie lassen sich auflösen zu

$$\hat{K} = \frac{k\alpha}{1-q}\cdot\frac{1}{1-\frac{E}{f}} \tag{4.29}$$

$$t = f\cdot\frac{1}{(1-q)^2} - f\cdot\frac{1}{1-q} - \frac{\hat{K}^2 E}{\alpha f} \tag{4.30}$$

$$T = \frac{1}{2}f\cdot\frac{1}{(1-q)^2} - f\cdot\frac{1}{1-q} + \frac{1}{2}f - \frac{\hat{K}^2 E}{2\alpha f} \tag{4.31}$$

Nach dem Einsetzen in die Gl.(4.27) erhält man

$$\begin{aligned} s \;=\; & \operatorname{sattel}_{\{q,k\}} \frac{1}{1-q}\cdot \\ & \left\{\frac{1}{2}f - \frac{f}{2}(1-q) + \frac{\alpha}{2}\cdot\frac{k^2}{1-\frac{E}{f}} - \frac{f}{2}(1-q)\ln(1-q) + \right. \\ & \left. + \alpha(1-q)\int_{-\infty}^{+\infty} Dz\ln\Phi\left(-\frac{\kappa - k + z\sqrt{q}}{\sqrt{1-q}}\right)\right\} \end{aligned} \tag{4.32}$$

Dieses Ergebnis kann man im Grenzübergang $q \to 1$ noch weiter vereinfachen. Die Ableitung der Entropie nach q liefert dann nämlich die Bestimmungsgleichung für das kritische α. Es stellt sich sogar heraus, daß die Lösung der Sattelpunktgleichung bezüglich q auf die Suche nach der Nullstelle des Grenzwertes

$$l = \lim_{q \to 1} (1-q)s \tag{4.33}$$

hinausläuft, wobei s die obige Entropie darstellt [En+92].

Das Integral über die Φ–Funktion wird im Limes $q \to 1$ mit Hilfe der Entwicklung (A.14) und der Gauß–Formel (A.21) berechnet. Zudem wird die Variable k durch die Variable a ersetzt:

$$a = \kappa - k \tag{4.34}$$

Man erhält

$$l = \lim_{q \to 1} (1-q)s = \tag{4.35}$$
$$\frac{f}{2} + \frac{\alpha}{2} \cdot \frac{(\kappa - a)^2}{1 - \frac{E}{f}} - \frac{\alpha}{2}\left(\left(1 + a^2\right)\Phi(a) + \frac{a}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}a^2\right)\right)$$

Die kritische Speicherkapazität α und die Sattelpunktvariable a werden schließlich aus dem folgenden System zweier Gleichungen gewonnen.

$$\begin{aligned} l(a, \alpha) &= 0 \\ \frac{\partial l(a, \alpha)}{\partial a} &= 0 \end{aligned} \tag{4.36}$$

4.2.5 Das Ergebnis am kritischen Punkt

Sind die Stabilität κ und der Verdünnungsparameter f für das Perzeptron auf den verbleibenden Plätzen vorgegeben, so erhalten wir die kritische Speicherkapazität aus der folgenden Prozedur:

Berechne die Schranke w aus der Gleichung

$$f = 2\Phi(-w) \tag{4.37}$$

Dann ist die Konstante E gegeben als

$$E = 2 \cdot \int_w^\infty DH \cdot H^2 = f + \sqrt{\frac{2}{\pi}} w \exp\left(-\frac{1}{2}w^2\right) \tag{4.38}$$

Löse anschließend die Gleichung für a

$$\kappa - a = \left(\frac{E}{f} - 1\right) \cdot \left[a \cdot \Phi(a) + \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}a^2\right)\right] \tag{4.39}$$

Schließlich ist das kritische α gegeben durch

$$\alpha = f \cdot \left[-\frac{(\kappa - a)^2}{1 - \frac{E}{f}} + \left(1 + a^2\right)\Phi(a) + \frac{a}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}a^2\right)\right]^{-1} \tag{4.40}$$

4.2.6 Der Limes $f \to 0$

Um zu zeigen, daß die effektive Speicherkapazität α_{eff} divergiert, entwickelt man die Φ–Funktionen in den Gleichungen des obigen Abschnitts gemäß Formel (A.14). Es gilt im Limes $f \to 0$

$$f = 2\Phi(-w) = \frac{1}{w}\sqrt{\frac{2}{\pi}} \cdot \exp\left(-\frac{1}{2}w^2\right)\left(1 + \mathcal{O}\left(\frac{1}{w^2}\right)\right) \tag{4.41}$$

Daraus folgt

$$\ln f = -\frac{1}{2}w^2 + \mathcal{O}(\ln w) \tag{4.42}$$

und

$$1 - \frac{E}{f} = w^2\left(1 + \frac{1}{w^2} + \mathcal{O}\left(\frac{1}{w^4}\right)\right) \tag{4.43}$$

Setzt man diese Entwicklungen in die Sattelpunktgleichung (4.39) für a ein, so folgt daraus die Divergenz

$$a \to -\infty \quad (f \to 0)$$

Man erhält

$$(\kappa - a) \cdot a^2\left(1 + \mathcal{O}\left(\frac{1}{a^2}\right)\right) \cdot \sqrt{2\pi} \cdot \exp\left(+\frac{1}{2}a^2\right) = -2\ln f \tag{4.44}$$

Wegen des exponentiellen Faktors $\exp\left(\frac{1}{2}a^2\right)$ in dieser Gleichung divergiert a sehr langsam. Für die effektive Speicherkapazität gilt dann

$$\begin{aligned} \alpha_{eff} = \frac{\alpha}{f} &= -a\sqrt{2\pi}\exp\left(+\frac{1}{2}a^2\right)\left(1 + \mathcal{O}\left(\frac{1}{a}\right)\right) \\ \alpha_{eff} &= -\frac{2\ln f}{a^2}\left(1 + \mathcal{O}\left(\frac{1}{a}\right)\right) \end{aligned} \tag{4.45}$$

α_{eff} divergiert hier also schwächer als im optimalen Fall, eine logarithmische Divergenz in $\ln f$ wird nicht erreicht.

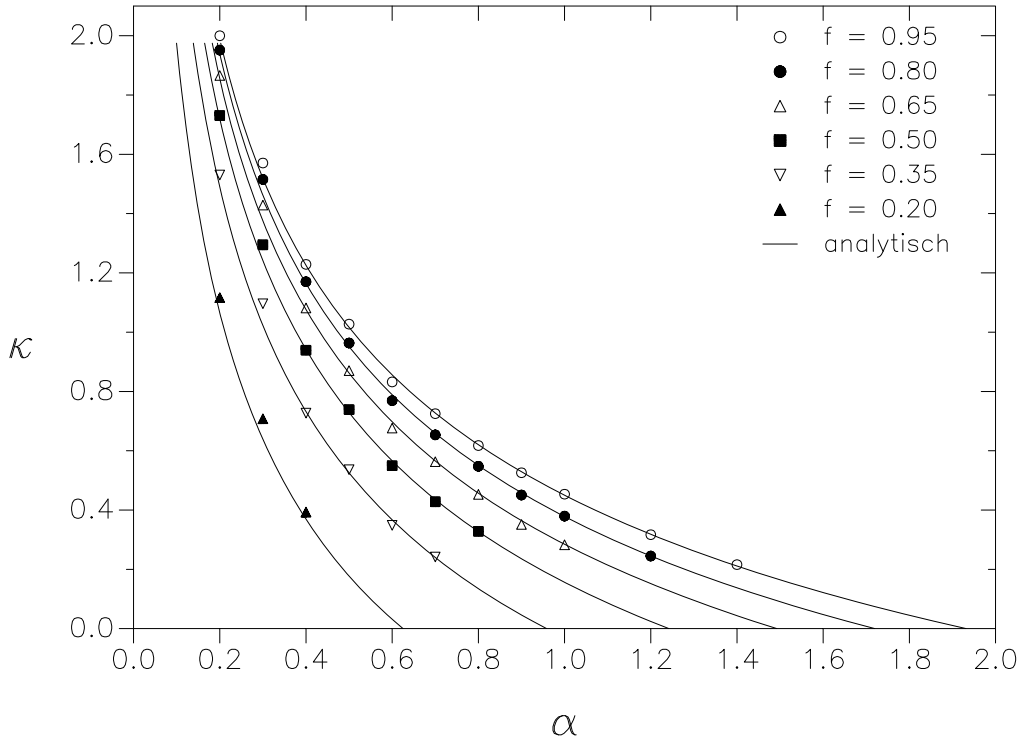


Abbildung 4.1: Die Stabilität $\kappa(\alpha)$ beim Quersummenalgorithmus. Die Computersimulationen wurden mit $N = 200$ Neuronen bei 50 Wiederholungen durchgeführt. Die Simulationen sind durch die Symbole wiedergegeben, wobei die Streuungen kleiner als die Symbolgrößen sind. Die durchgezogene Linie stellt die analytische Rechnung dar.

4.3 Ergebnisse

Das Ergebnis der Replika–Rechnung kann mit Computersimulationen überprüft werden. Dazu gibt man die ursprüngliche Zahl der Neuronen N, die Zahl $p = \alpha N$ der Muster und die Zahl Nf der verbleibenden Neuronen vor. Die gaußverteilten transformierten Muster werden mit einem Zufallszahlengenerator bestimmt. Die Gauß–Verteilung wird dabei mit dem Box–Muller–Algorithmus erzeugt [NR88]. Nach der Verdünnung des Netzwerks gemäß der Beträge der Hebb–Kopplungen lernt man auf den verbleibenden Nf Plätzen das Perzeptron optimaler Stabilität mit Hilfe des AdaTron–Algorithmus.

In Abbildung 4.1 beobachtet man eine sehr gute Übereinstimmung der analytischen Ergebnisse mit der Computersimulation. Die Simulation wurde dabei mit $N = 200$ Neuronen durchgeführt, wobei eine Mittelung über 50 Läufe erfolgte [Gc92], [Ku+92].

Man erreicht in der Computersimulation nicht die Stabilität $\kappa = 0$, weil bei niedrigen Stabilitäten auch die Wahrscheinlichkeit sinkt, daß ein Perzeptron–Vektor existiert. Die zugehörigen Läufe, in denen der Algorithmus nicht konvergiert, fallen aus der Mittelung heraus. Dies verschiebt den Mittelwert der Stabilitäten nach oben. Bei höheren κ bzw. kleineren α ist die Wahrscheinlich-

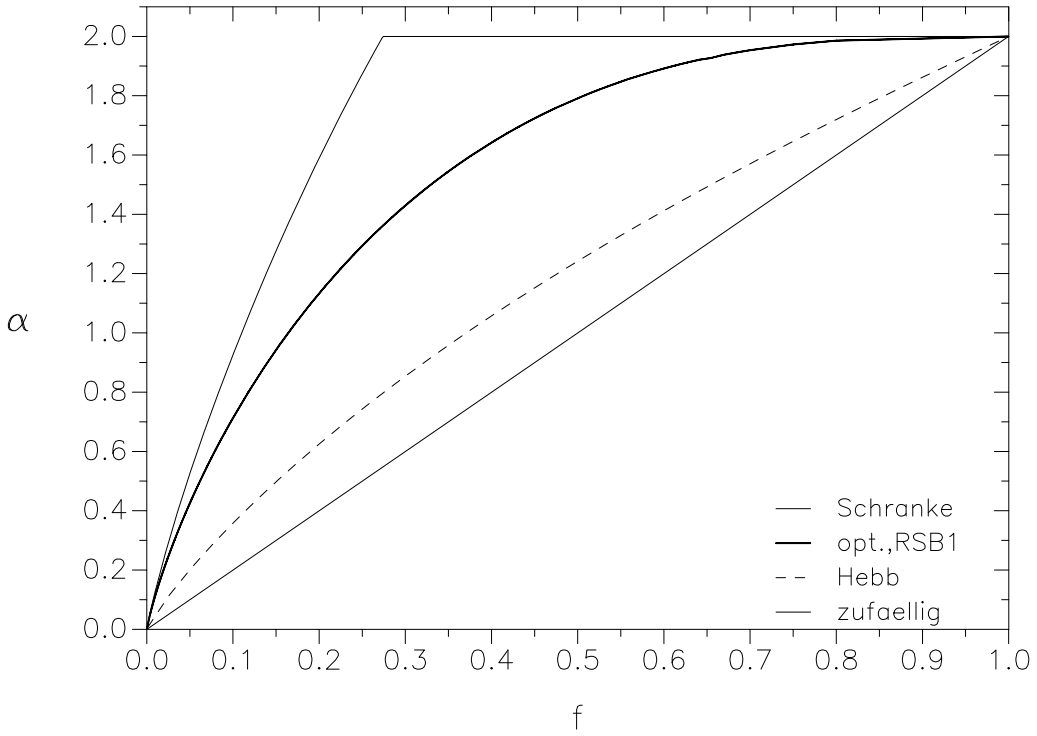


Abbildung 4.2: Die bisherigen Ergebnisse für die Speicherkapazitäten $\alpha(\kappa = 0)$ im Vergleich. Der Quersummenalgorithmus („Hebb“) stellt eine Verbesserung der zufälligen Verdünnung dar.

keit, daß ein Perzeptron existiert, jedoch nahezu 1, und die Mittelwerte sind sehr zuverlässig.

Die analytischen Ergebnisse für die Speicherkapazität α bei $\kappa = 0$ sind in Abb. 4.2 aufgeführt. Sie werden mit den bisher gewonnenen Ergebnissen verglichen: RSB1–Näherung, obere Schranke und Gerade $\alpha = 2f$ der zufälligen Verdünnung. Wie erwartet liegt die Speicherkapazität beim Quersummenverdünnungsalgorithmus zwischen der Näherung für den optimalen Fall und der Geraden. Das wichtigste Ergebnis dieses Kapitels sehen wir nocheinmal in Abbildung 4.3. Die effektive Speicherkapazität ist beim Quersummenverdünnungsalgorithmus größer als 2.

$$\alpha_{eff} = \frac{\alpha(\kappa = 0)}{f} > 2 \tag{4.46}$$

Dieses Ergebnis konnte bei höheren κ–Werten mit Computersimulationen belegt werden. $\alpha_{eff} > 2$ kann also auch in einem sehr schnellen Algorithmus erzeugt werden, es tritt nicht nur bei der optimalen Verdünnung auf. Im Limes $f \to 0$ beobachten wir auch hier eine Divergenz in α_{eff}. Diese Divergenz ist jedoch schwächer als im optimalen Fall (siehe Gl.(4.45)).

Gegen das Ergebnis $\alpha_{eff} > 2$ könnte man einwenden, daß es doch eigentlich sehr unwahrscheinlich sein müßte, einen Verdünnungsvektor $\underline{c}$ mit $\alpha_{eff} > 2$ durch einen solch einfachen Auswahlprozeß wie dem Quersummenverdünnungsalgorithmus zu finden. Wir begegnen dem Einwand, indem wir einen Vergleich

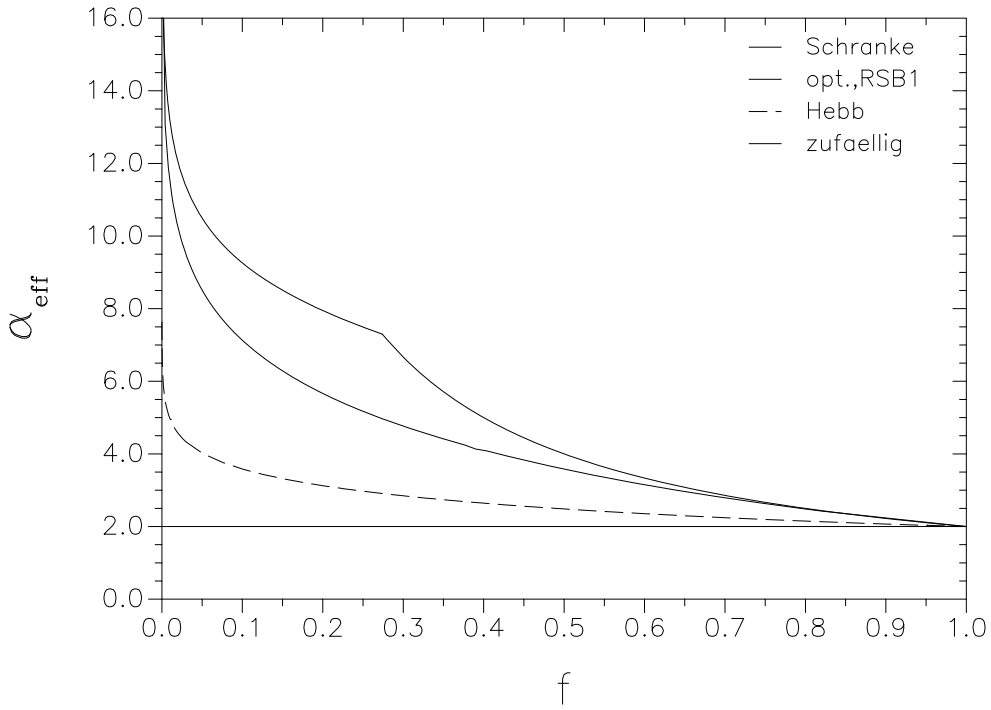


Abbildung 4.3: Die bisherigen Ergebnisse für die effektiven Speicherkapazitäten $\alpha_{eff}(\kappa = 0)$ im Vergleich. α_{eff} des Quersummenalgorithmus („Hebb“) divergiert schwächer als die obere Schranke und die RSB1–Näherung der optimalen Verdünnung.

mit einer einfachen Kette nicht wechselwirkender Spins anstellen. Dort sind Magnetisierungen $m \neq 0$ zwar sehr unwahrscheinlich, doch ist es sehr einfach, einen Zustand mit $m \neq 0$ anzugeben. Im Fall des Problems der Verdünnung ist die Menge M aller Verdünnungsvektoren $\underline{c}$ mit $\alpha_{eff} > 2$ zwar sehr klein, jedoch landet der Quersummenverdünnungsalgorithmus für $N \to \infty$ mit Wahrscheinlichkeit 1 in dieser Menge M.

In einem anderen Einwand gegen $\alpha_{eff} > 2$ könnte man mit dem Satz von Cover argumentieren: Nach dem Verdünnen gemäß der Beträge der Hebb–Kopplungen liegen p Muster auf Nf Plätzen vor. Diese sind zwar wegen des Auswahlprozesses korreliert, doch befinden sie sich immer noch in allgemeiner Lage. Dann ist aber die Wahrscheinlichkeit, daß ein Perzeptron existiert, durch die Cover–Wahrscheinlichkeit W_s (Gl.(1.41)) gegeben, und es folgt $\alpha_{eff} = 2$ im Limes $N \to \infty$.

Man entkräftet diesen Einwand mit dem Hinweis auf die Auswahl der möglichen Ausgaben. In die Hebb–Kopplungen der *transformierten* Muster gehen die vorgegebenen Ausgaben ein. Das bedeutet freilich, daß der Verdünnungsvektor $\underline{c}$ schon unter Kenntnis der Ausgaben bestimmt wurde. Zwar gibt es auf den verbliebenen Plätzen $C(p, Nf)$ viele mögliche Ausgaben (siehe Gl.(1.40)), jedoch wurde durch den Auswahlprozeß bewerkstelligt, daß sich die gewünschte Ausgabe mit erhöhter Wahrscheinlichkeit P unter diesen $C(p, Nf)$ vielen Ausgaben

befindet:

$$P > \frac{1}{C(p, Nf)} \tag{4.47}$$

Die Berechnung von P wurde mit Hilfe der Gardner–Rechnung umgangen. Dies ist ein eindruckvolles Beispiel für die Mächtigkeit der Methoden der statistischen Physik auf dem Gebiet der neuronalen Netzwerke.

Wenden wir uns schließlich unserer Gardner–Rechnung zu. Hier ist die Frage nach der Richtigkeit der Annahme der Replika–Symmetrie noch offen. Ein mathematischer Beweis der Richtigkeit unserer Rechnung könnte auf zweierlei Art erfolgen. Zum einen könnte man den von van Hemmen und Palmer [HePa79] angegebenen Satz von Carlsson als Grundlage für einen Beweis benutzen, zum anderen könnte man das in Kapitel 2 geschilderte Verfahren benutzen, um ohne eine Replika–Rechnung die Richtigkeit der angegebenen Formeln nachzuweisen.

Wir begnügen uns an dieser Stelle mit einer Betrachtung des Phasenraums unseres Verdünnungsproblems. Im Gegensatz zur optimalen Verdünnung besteht hier der Phasenraum einfach aus einer Gardner–Kugel, die von Kopplungen auf den verbleibenden Nf Plätzen gebildet wird. Das zugehörige Optimierungsproblem ist weiterhin ein Problem der quadratischen Optimierung. Da die Verdünnung eingefroren ist, entfällt eine diskrete Optimierung bezüglich der Verdünnungsvektoren. Es liegt also kein Grund dafür vor, weshalb die Replika–Symmetrie gebrochen sein sollte, und die Computersimulationen bestätigen unsere einfache replikasymmetrische Theorie in allen zugänglichen Parameterbereichen.

Kapitel 5

Das Schneideverfahren zur Verdünnung des Perzeptrons optimaler Stabilität

In Kapitel 4 wurde aufgezeigt, daß es prinzipiell möglich ist, mit einem einfachen Verdünnungsalgorithmus $\alpha_{eff} > 2$ zu erzeugen. Nichtsdestotrotz ist das erzielte Ergebnis noch weit von demjenigen der RSB1–Näherung der optimalen Verdünnung entfernt. Demzufolge befaßt sich dieses Kapitel mit der Frage, wie nahe man mit einem praktikablen Algorithmus an das α_{eff} der optimalen Verdünnung herankommen kann.

Der in Abschnitt 5.1 vorgestellte Algorithmus fußt darauf, Perzeptron–Kopplungen zur Festlegung des Verdünnungsvektors zu verwenden. Statt der Beträge der Hebb–Kopplungen werden die Beträge der Kopplungen des optimalen Perzeptrons dazu verwendet, die zur Lösung der Klassifikationsaufgabe relevanten Neuronen auszuwählen. Man kann dieses Verfahren auch mehrmals anwenden, um eine höhere Stabilität des verdünnten Perzeptrons zu erzeugen. Während dieses Mehrschrittverfahren nur auf dem Computer simuliert werden kann, ist das zum Quersummenalgorithmus analoge Einschrittverfahren durch eine Replika–Rechnung zugänglich. Diese analytische Rechnung wird in Abschnitt 5.2 dargestellt.

Die Ergebnisse der replikasymmetrischen Theorie werden in Abschnitt 5.3 mit Computersimulationen ([Gc92], [Gc+92]) verglichen.

5.1 Der Algorithmus

Analog dem Quersummenalgorithmus aus Abschnitt 4.1.2 formuliere ich das folgende *Einschrittverfahren*:

Vorgelegt seien p transformierte Muster $\underline{\sigma}^{\mu}$. Das Netzwerk soll so verdünnt werden, daß Nf Neuronen übrigbleiben. Wenn auf diesen Nf Neuronen eine perfekte Klassifizierung erfolgen soll, muß zumindest das Perzeptron optimaler Stabilität auf allen N Plätzen existieren.

1. Lerne den zugehörigen Vektor

$$\underline{J} = (J_1, \ldots, J_N)^T$$

des Perzeptrons optimaler Stabilität auf dem unverdünnten Einschichtnetzwerk. $\underline{J}$ habe die Stabilität κ_1. Ordne die Beträge $|J_j|$ der Perzeptron–Kopplungen in aufsteigender Reihenfolge und ordne die Platzvektoren

$$\vec{\sigma}_j = \left(\sigma_j^1, \ldots, \sigma_j^p\right)^T$$

dementsprechend um mit neuen Indizes $m(j)$.

2. Entferne alle Plätze j mit den kleinsten Beträgen $|J_j|$, d.h. entferne alle $\vec{\sigma}_m$ mit

$$m \leq N(1-f)$$

Dabei sei die Schranke w für die Beträge der Perzeptron–Kopplungen als

$$w = |J_m|$$

definiert.

3. Lerne auf den verbleibenden Nf Plätzen das Perzeptron optimaler Stabilität mit Kopplungen

$$\underline{L} = (L_1, \ldots, L_{Nf})^T$$

um sicherzustellen, daß alle Muster wieder korrekt klassifiziert werden. Dieses zweite Perzeptron habe die optimale Stabilität κ_2.

Nach der Verdünnungsprozedur liegt also der Verdünnungsvektor

$$\underline{c} = (c_1, \ldots, c_N)^T$$

fest. Er ist wegen der obigen Definition der Schranke w gegeben durch

$$c_j = \Theta(|J_j| - w), \quad j = 1, \ldots, N$$

Es handelt sich wieder um eine eingefrorene Verdünnung. Eine Verbesserung des Verfahrens kann dadurch erreicht werden, daß man die obige Prozedur in mehreren Schritten durchführt: Man verdünnt zunächst auf Nf_1 Neuronen, lernt nach, verdünnt auf Nf_2 und so weiter, bis man schließlich bei der geforderten Zahl Nf anlangt und ein letztes Mal nachlernt. Wir wollen im folgenden unter dem Mehrschrittalgorithmus den optimalen Fall, d.h. das Nachlernen für jedes einzelne Neuron verstehen:

1. Lerne das Perzeptron optimaler Stabilität mit Vektor

$$\underline{J} = (J_1, \ldots, J_N)^T$$

2. Entferne nur den Platz j, der den kleinsten Betrag $|J_j|$ aufweist.

3. Lerne das Perzeptron optimaler Stabilität auf den verbleibenden Plätzen nach.

4. Iteriere die Schritte 2 und 3, bis die gewünschte Verdünnung erreicht ist.

Ein ähnlicher Mehrschrittalgorithmus ist bereits von Janowsky im Rahmen von Attraktornetzwerken untersucht worden. Janowsky definiert die Speicherkapazität jedoch in Zusammenhang mit einem Glauber–Prozeß. Außerdem verwendet er statt dem optimalen das einfache Perzeptron [Ja89].

5.2 Die Gardner–Rechnung zum Einschrittverfahren

Das Einschrittverfahren kann in einer Gardner–Rechnung behandelt werden, wenn man die Komponenten

$$c_j = \Theta(|J_j| - w), \quad j = 1, \ldots, N$$

des Verdünnungsvektors korrekt in die Rechnung einbezieht. Dies geschieht mit Hilfe einer Replika–Mittelung über alle zugelassenen Kopplungsvektoren $\underline{J}$ des ersten Perzeptrons. In Abschnitt 5.2.1 wird diese Methode bereits im Zusammenhang mit der Berechnung der Schranke w für das Schneideverfahren vorgestellt. In Abschnitt 5.2.2 wird dann der Ansatz für die Gardner–Rechnung eingehend erklärt. Es wird dargelegt, daß es sich um eine Replika–Rechnung mit zwei Replika–Indizes handelt [1].

Da die bei Gardner–Rechnungen angewendeten Methoden bereits in Kapitel 4 vorgestellt wurden, wird in Abschnitt 5.2.3 aus Gründen der Übersichtlichkeit der Rechenweg bis zur Gewinnung des allgemeinen Ergebnisses skizziert. Anschließend wird in Abschnitt 5.2.4 die replikasymmetrische Annahme gemacht. Die Ergebnisse der analytischen Rechnung vereinfachen sich erheblich, wenn man fordert, daß sowohl das erste Perzeptron (mit Kopplungen $\underline{J}$) als auch das zweite (mit Kopplungen $\underline{L}$) optimale Stabilität besitzen (κ_1 bzw. κ_2). Diese Grenzübergänge werden in Abschnitt 5.2.5 durchgeführt. Die Ergebnisse der analytischen Rechnung werden in Abschnitt 5.2.6 vorgestellt.

5.2.1 f als Funktion von w

Um den Verdünnungsparameter

$$f(\underline{J}) = \frac{1}{N}\sum_{j=1}^{N} c_j = \frac{1}{N}\sum_{j=1}^{N} \Theta(|J_j| - w) \tag{5.1}$$

berechnen zu können, müssen die Kopplungen J_j des ersten Perzeptrons in die Rechnung einbezogen werden. Dies geschieht durch eine Replika–Mittelung über das erste Perzeptron. Wenn die Stabilität κ_1 unterhalb der kritischen Stabilität ist, gibt es mehrere Perzeptron–Vektoren $\underline{J}$, die die Klassifikationsaufgabe erfüllen. Die $\underline{J}$ nehmen ein Gebiet auf der Gardner–Kugel ein, das gleich

[1] Man bezeichnet eine solche Rechnung auch oft als „doppelte Gardner–Rechnung“

dem Gardner–Phasenraum–„Volumen“ V ist. Der Mittelwert der Funktion $f(\underline{J})$ über alle Perzeptrone im ersten Schritt ist dann

$$\langle f(\underline{J})\rangle_{\{\underline{J}\}} = \tag{5.2}$$
$$\frac{\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty} dJ_j\right)\delta\left(\sum_{j=1}^{N} {J_j}^2 - N\right)\left(\prod_{\mu=1}^{p}\theta\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j\sigma_j^\mu - \kappa_1\right)\right) f(\underline{J})}{\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty} dJ_j\right)\delta\left(\sum_{j=1}^{N} {J_j}^2 - N\right)\left(\prod_{\mu=1}^{p}\theta\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j\sigma_j^\mu - \kappa_1\right)\right)}$$

Da später über die Muster gemittelt werden soll, muß man die rechte Seite in eine Replika–Mittelung über das erste Perzeptron überführen. Dies ist ein bekanntes Verfahren aus der Spinglasphysik [Me+87]. Eingeführt wird zunächst ein Replika–Index

$$\beta = 1, \ldots, m$$

Später wird dann formal der Limes $m \to 0$ durchgeführt. Multipliziert man die obige Gleichung mit

$$1 = \lim_{m\to 0} V^m \tag{5.3}$$

wobei V das Phasenraumvolumen, also den Nenner in Gl.(5.2) darstellt, so gilt

$$\begin{aligned}\langle f(\underline{J})\rangle_{\{\underline{J}\}} \;=\; & \lim_{m\to 0}\left(\prod_{\beta=1}^{m}\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty} dJ_j^\beta\right)\delta\left(\sum_{j=1}^{N}\left(J_j^\beta\right)^2 - N\right)\right)\cdot \\ & \cdot\left(\prod_{\beta=1}^{m}\prod_{\mu=1}^{p}\theta\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} J_j^\beta\sigma_j^\mu - \kappa_1\right)\right) f\left(\underline{J}^1\right)\end{aligned} \tag{5.4}$$

Man sieht hier, daß im Gegensatz zu den bisher vorgestellten Replika–Rechnungen in dieser Rechnung ein Replikon ausgezeichnet ist; ich wähle auch im folgenden immer das erste Replikon.

Um f als Funktion von w mit den Methoden der statistischen Physik berechnen zu können, müssen wir also zweierlei voraussetzen

1. f sei selbstmittelnd bezüglich der Mittelung über alle Perzeptron–Vektoren des ersten Lernschritts.
2. f sei selbstmittelnd über alle Muster.

In Formeln ausgedrückt heißt dies

$$\lim_{N\to\infty} f = \lim_{N\to\infty}\left\langle\!\left\langle\ \langle f(\underline{J})\rangle_{\{\underline{J}\}}\ \right\rangle\!\right\rangle_{\{\sigma_j^\mu\}} \tag{5.5}$$

Da wegen der Mustermittelung kein Platz mehr ausgezeichnet ist, liefert jeder Summand in Gl.(5.1) das gleiche Ergebnis. Ich wähle den N–ten Summanden. Es gilt

$$\begin{aligned}\left\langle\!\left\langle\ \langle f(\underline{J})\rangle_{\{\underline{J}\}}\ \right\rangle\!\right\rangle &= \left\langle\!\left\langle\ \langle c_N\rangle_{\{\underline{J}\}}\ \right\rangle\!\right\rangle \\ &= \int_{-\infty}^{+\infty} dJ\ \theta(|J| - w)\left\langle\!\left\langle\ \langle\delta(J - J_N)\rangle_{\{\underline{J}\}}\ \right\rangle\!\right\rangle\end{aligned} \tag{5.6}$$

Die Wahrscheinlichkeitsdichte der Perzeptron–Kopplungen

$$P(J) = \lim_{N\to\infty} \left\langle\!\left\langle\ \langle \delta(J - J_N)\rangle_{\{\underline{J}\}}\ \right\rangle\!\right\rangle \tag{5.7}$$

ist dann mit Hilfe einer Gardner–Rechnung zu bestimmen. Da die Technik der Replika–Mittelung noch in den nächsten Unterpunkten vorgestellt wird, greife ich hier auf die Rechnung in [Bo+90] zurück. Bouten et al. erhalten eine Gauß–Verteilung, die von der Stabilität unabhängig ist.

$$P(J) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}J^2\right) \tag{5.8}$$

Dieses Ergebnis ist anschaulich verständlich, wenn wir die Normierungsbedingung für die Perzeptron–Kopplungen betrachten. Nach dem Gesetz der großen Zahlen gilt nämlich

$$\overline{J^2} = \frac{1}{N}\sum_{j=1}^{N} {J_j}^2 = 1 \tag{5.9}$$

Mit der obigen Wahrscheinlichkeitsdichte folgt dann einfach unser Endergebnis

$$f = 2\int_{-\infty}^{-w} DJ = 2\Phi(-w) \tag{5.10}$$

Analog zur Verdünnung der Hebb–Kopplungen haben wir also auch im vorliegenden Fall den mittleren Teil (d.h. alle J mit $|J| < w$) aus einer Gaußglocke herausgeschnitten und die verbleibende Fläche berechnet.

5.2.2 Der Ansatz für die Gardner–Rechnung

Wir gehen davon aus, daß im ersten Schritt ein Perzeptron mit Kopplungen $\underline{J}$ und Stabilität κ_1 gelernt worden sei. κ_1 sei unterhalb der kritischen Stabilität, die zu dem vorgegebenen α gehört (Gl.(1.39)), so daß eine Replika–Mittelung über alle Perzeptrone mit Stabilität $\kappa > \kappa_1$ möglich ist. Wir verdünnen die Neuronen gemäß der obigen Schranke w und lernen auf den verbleibenden Plätzen Perzeptron–Kopplungen

$$\underline{L}_{Nf} = (L_1, \ldots, L_{Nf})^T \tag{5.11}$$

nach, wobei wir mindestens eine Stabilität κ_2 fordern. Das Phasenraumvolumen für das zweite Perzeptron ist dann analog zu den Gleichungen (4.12) und (4.15)

$$\begin{aligned} V &= \frac{\left(\prod_{k=1}^{Nf}\int_{-\infty}^{+\infty} dL_k\right)\delta\left(\sum_{k=1}^{Nf} {L_k}^2 - Nf\right)\prod_{\nu=1}^{p}\Theta\left(\frac{1}{\sqrt{Nf}}\sum_{k=1}^{Nf} L_k\eta_k^\nu - \kappa_2\right)}{\left(\prod_{k=1}^{Nf}\int_{-\infty}^{+\infty} dL_k\right)\delta\left(\sum_{k=1}^{Nf} {L_k}^2 - Nf\right)} \\ &= \frac{1}{C_{\text{norm}}}\cdot\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty}\frac{dT_j}{\sqrt{2\pi}}\right)\delta\left(\sum_{j=1}^{N} c_j {T_j}^2 - Nf\right)\exp\left(-\frac{1}{2}\sum_{j=1}^{N}(1-c_j){T_j}^2\right)\cdot \\ &\quad\cdot\prod_{\nu=1}^{p}\Theta\left(\frac{1}{\sqrt{Nf}}\sum_{j=1}^{N} c_j T_j \sigma_j^\nu - \kappa_2\right) \end{aligned} \tag{5.12}$$

mit der Normierung des Phasenraumvolumens

$$C_{\text{norm}} = \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} \frac{dT_j}{\sqrt{2\pi}} \right) \delta \left(\sum_{j=1}^{N} c_j {T_j}^2 - Nf \right) \cdot \exp \left(-\frac{1}{2} \sum_{j=1}^{N} (1 - c_j) {T_j}^2 \right)$$

Um die

$$c_j = \Theta(|J_j| - w), \quad j = 1, \ldots, N \tag{5.13}$$

in die Rechnung einzubeziehen, müssen wir die Selbstmittelungseigenschaft der Entropie

$$s(\underline{J}) = \frac{1}{N} \ln V(\underline{J}) \tag{5.14}$$

bezüglich der Mittelung über die Perzeptrone $\underline{J}$ des ersten Schritts und bezüglich der Mittelung über die transformierten Muster voraussetzen. Analog zu Gl.(5.5) berechnen wir also

$$\lim_{N \to \infty} s = \lim_{N \to \infty} \left\langle \left\langle \ \langle s(\underline{J}) \rangle_{\{\underline{J}\}} \ \right\rangle \right\rangle_{\{\sigma_j^\mu\}} \tag{5.15}$$

Dabei muß wieder die Replika–Methode angewandt werden, um die Mittelungen über den Logarithmus zu ermöglichen. Dazu gehört ein Replika–Index

$$\rho = 1, \ldots, n$$

Zur Durchführung der Mittelung über die Perzeptronkopplungen $\underline{J}$ muß ein zweiter Replika–Index

$$\beta = 1, \ldots, m$$

eingeführt werden.

Zur korrekten Formulierung der Replika–Methode für diese doppelte Gardner–Rechnung betrachten wir

$$\Phi_R^{(N)}(m, n) = \frac{1}{N} \ln \langle \langle U(m, n) \rangle \rangle \tag{5.16}$$

wobei die Abkürzung $U(m, n)$ für die Replika–Mittelung über V^n steht:

$$\begin{aligned} U(m,n) \ = \ & \left(\prod_{\beta=1}^{m} \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} dJ_j^\beta \right) \delta \left(\sum_{j=1}^{N} \left(J_j^\beta \right)^2 - N \right) \right) \cdot \\ & \cdot \left(\prod_{\beta=1}^{m} \prod_{\mu=1}^{p} \theta \left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_j^\beta \sigma_j^\mu - \kappa_1 \right) \right) \left(V \left(\underline{J}^1 \right) \right)^n \end{aligned} \tag{5.17}$$

Leitet man $\Phi_R^{(N)}$ partiell nach n ab, so erhält man wie gewünscht an der Stelle $n = 0$

$$\begin{aligned} & \lim_{m \to 0} \left. \frac{\partial \Phi_R^{(N)}(m, n)}{\partial n} \right|_{n=0} = \\ & \lim_{m \to 0} \frac{1}{N} \cdot \frac{1}{\langle \langle U(m, 0) \rangle \rangle} \left\langle \left\langle \left(\prod_{\beta=1}^{m} \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} dJ_j^\beta \right) \delta \left(\sum_{j=1}^{N} \left(J_j^\beta \right)^2 - N \right) \right) \cdot \right. \right. \end{aligned}$$

$$\cdot\left(\prod_{\beta=1}^{m}\prod_{\mu=1}^{p}\theta\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}J_j^\beta\sigma_j^\mu-\kappa_1\right)\right)\ln V\Bigg\rangle\Bigg\rangle$$

$$= \left\langle\left\langle \quad \left\langle\frac{1}{N}\ln V(\underline{J})\right\rangle_{\{\underline{J}\}} \quad\right\rangle\right\rangle_{\{\sigma_j^\mu\}} \tag{5.18}$$

Um die analytische Rechnung durchführen zu können, müssen wir annehmen, daß der Grenzwert $N\to\infty$ mit den Grenzwerten $m\to 0$ und $n\to 0$ vertauscht. Wir setzen also

$$\Phi_R(m,n)=\lim_{N\to\infty}\Phi_R^{(N)}(m,n) \tag{5.19}$$

und berechnen

$$\lim_{N\to\infty}s=\lim_{m\to 0}\left.\frac{\partial\Phi_R(m,n)}{\partial n}\right|_{n=0} \tag{5.20}$$

5.2.3 Der Rechenweg zur Gewinnung des allgemeinen Ergebnisses für $\Phi_R(m,n)$

Wir betrachten $U(m,n)$ aus Gl.(5.17), setzen das Phasenraumvolumen (5.12) ein und führen analog zur Rechnung aus Kapitel 4 (siehe Gl.(4.19)) die Fourierdarstellung für alle Zwangsbedingungen ein. Man erhält

$$\begin{aligned}
U\propto\;&\left(\prod_{j,\beta}\int_{-\infty}^{+\infty}dJ_j^\beta\right)\left(\prod_{\beta=1}^{m}\delta\left(\sum_{j=1}^{N}\left(J_j^\beta\right)^2-N\right)\right)\cdot\\
&\left(\prod_{\mu,\beta}\int_{\kappa_1}^{\infty}dX_\mu^\beta\int_{-\infty}^{+\infty}\frac{dx_\mu^\beta}{2\pi}\right)\exp\left(i\sum_{\mu,\beta}x_\mu^\beta X_\mu^\beta\right)\left(\prod_{\mu,\rho}\int_{\kappa_2}^{\infty}dZ_\mu^\rho\int_{-\infty}^{+\infty}\frac{dz_\mu^\rho}{2\pi}\right)\exp\left(i\sum_{\mu,\rho}z_\mu^\rho Z_\mu^\rho\right)\cdot\\
&\cdot\left(\prod_{j,\rho}\int_{-\infty}^{+\infty}\frac{dT_j^\rho}{\sqrt{2\pi}}\right)\left(\prod_{\rho=1}^{n}\delta\left(\sum_{j=1}^{N}c_j\left(T_j^\rho\right)^2-Nf\right)\right)\cdot\exp\left(-\frac{1}{2}\sum_{j,\rho}(1-c_j)\left(T_j^\rho\right)^2\right)\cdot\\
&\cdot\exp\left[-\frac{i}{\sqrt{N}}\sum_{j,\mu}\sigma_j^\mu\left(\sum_{\beta=1}^{m}x_\mu^\beta J_j^\beta+\frac{1}{\sqrt{f}}\sum_{\rho=1}^{n}z_\mu^\rho c_j T_j^\rho\right)\right]
\end{aligned} \tag{5.21}$$

Man beachte, daß c_j von der Perzeptronkopplung J_j^1 (des ersten Replikons $\beta=1$) wie folgt abhängt:

$$c_j=\theta\left(\left|J_j^1\right|-w\right),\quad j=1,\ldots,N \tag{5.22}$$

Mittelt man U über die gaußverteilten transformierten Muster, so entstehen die Überlappungen

$$P_{\beta\gamma} = \frac{1}{N}\sum_{j=1}^{N}J_j^\beta J_j^\gamma \tag{5.23}$$

$$R_{\beta\rho} = \frac{1}{N}\sum_{j=1}^{N}J_j^\beta c_j T_j^\rho \tag{5.24}$$

$$Q_{\rho\sigma} \quad = \quad \frac{1}{Nf}\sum_{j=1}^{N} c_j T_j^{\rho} T_j^{\sigma} \tag{5.25}$$

Dabei sind $\beta, \gamma \in \{1, \ldots, m\}$ und $\rho, \sigma \in \{1, \ldots, n\}$.

Die Überlappungen werden als Ordnungsparameter gemäß Gl.(A.3) mit konjugierten Variablen $\hat{P}_{\beta\gamma}$, $\hat{R}_{\beta\rho}$ und $\hat{Q}_{\rho\sigma}$ eingeführt. Es gelingt dadurch, die Indizes μ von den Indizes j zu entkoppeln, und man kann $\Phi_R(m,n)$ mit Hilfe der Sattelpunktmethode berechnen.

Man erhält

$$\begin{aligned}
&\Phi_R(m,n) = \\
&\quad \text{sattel}_{\{P_{\beta\gamma},\hat{P}_{\beta\gamma},R_{\beta\rho},\hat{R}_{\beta\rho},Q_{\rho\sigma},\hat{Q}_{\rho\sigma}\}} \\
&\quad \sum_{\beta\le\gamma} P_{\beta\gamma}\hat{P}_{\beta\gamma} + \sum_{\beta,\rho} R_{\beta\rho}\hat{R}_{\beta\rho} + \sum_{\rho\le\sigma} Q_{\rho\sigma}\hat{Q}_{\rho\sigma} + \\
&\quad + \ln\left[\left(\prod_{\beta=1}^{m}\int_{-\infty}^{+\infty}\frac{dJ_\beta}{\sqrt{2\pi}}\right)\exp\left(-\sum_{\beta\le\gamma} J_\beta\hat{P}_{\beta\gamma}J_\gamma\right)\cdot\right. \\
&\qquad \left.\cdot\left(\prod_{\rho=1}^{n}\int_{-\infty}^{+\infty} DT_\rho\right)\cdot\exp\left(-\sum_{\beta,\rho} J_\beta\hat{R}_{\beta\rho}c_1T_\rho - \frac{1}{f}\sum_{\rho\le\sigma} c_1T_\rho\hat{Q}_{\rho\sigma}T_\sigma\right)\right] + \\
&\quad + \alpha\ln\left[\left(\prod_{\beta=1}^{m}\int_{\kappa_1}^{\infty} dX_\beta\int_{-\infty}^{+\infty}\frac{dx_\beta}{2\pi}\right)\left(\prod_{\rho=1}^{n}\int_{\kappa_2}^{\infty} dZ_\rho\int_{-\infty}^{+\infty}\frac{dz_\rho}{2\pi}\right)\cdot\right. \\
&\qquad \cdot\exp\left[-\frac{1}{2}\sum_{\beta,\gamma} x_\beta P_{\beta\gamma}x_\gamma - \frac{1}{\sqrt{f}}\sum_{\beta,\rho} x_\beta R_{\beta\rho}z_\rho - \frac{1}{2}\sum_{\rho,\sigma} z_\rho Q_{\rho\sigma}z_\sigma + \right. \\
&\qquad \left.\left. + i\sum_{\beta=1}^{m} x_\beta X_\beta + i\sum_{\rho=1}^{n} z_\rho Z_\rho\right]\right]
\end{aligned} \tag{5.26}$$

Dabei ist der Verdünnungskoeffizient c_1 nur von der Integrationsvariablen J_1 abhängig:

$$c_1 = \theta(|J_1| - w) \tag{5.27}$$

$P_{\beta\gamma}$ stellt die Matrix der Überlappungen der Kopplungsvektoren des ersten Perzeptrons dar. $Q_{\rho\sigma}$ gehört zum zweiten Perzeptron. Die Rechteckmatrix $R_{\beta\rho}$ gibt die Überlappungen der Kopplungen des ersten Perzeptrons mit denen des zweiten Perzeptrons wieder.

5.2.4 Die Annahme der Replika–Symmetrie

Betrachtet man die Integrale über die Perzeptron–Kopplungen $\{J_\beta\}$ und $\{T_\rho\}$, so fällt auf, daß das Replikon $\beta = 1$ ausgezeichnet ist. Dies liegt an unserem obigen Ansatz für die Replika–Mittelung über die Kopplungen des ersten Perzeptrons. In dem Mischausdruck

$$-\sum_{\beta,\rho} J_\beta\hat{R}_{\beta\rho}c_1T_\rho \tag{5.28}$$

kommt dadurch eine Asymmetrie zustande. Ich mache deshalb die Annahme, daß die zu dem Mischausdruck gehörenden Ordnungsparameter $R_{\beta\rho}$ und $\hat{R}_{\beta\rho}$ ebenfalls eine Asymmetrie bezüglich $\beta = 1$ aufweisen. Für das erste Perzeptron liegt hingegen Replika–Symmetrie vor, denn es wurde einfach vorgelernt. Der spätere Verwendungszweck der Kopplungen $\underline{J}$ hat das Lernen in keiner Weise kausal beeinflußt. Auch für das zweite Perzeptron machen wir analog zu Abschnitt 4.2.4 den replikasymmetrischen Ansatz.

Wir setzen

$$\begin{aligned} P_{\beta\gamma} &= p, \quad \beta,\gamma = 1,\ldots,m \text{ mit } \beta \neq \gamma && (5.29)\\ R_{\beta\rho} &= r_1, \quad \beta = 1 \text{ und } \rho = 1,\ldots,n && (5.30)\\ R_{\beta\rho} &= r, \quad \beta = 2,\ldots,m \text{ und } \rho = 1,\ldots,n && (5.31)\\ Q_{\rho\sigma} &= q, \quad \rho,\sigma = 1,\ldots,n \text{ mit } \rho \neq \sigma && (5.32) \end{aligned}$$

Es gilt

$$P_{\beta\beta} = 1 \;\forall\beta \quad \text{und} \quad Q_{\rho\rho} = 1 \;\forall\rho$$

Für die konjugierten Variablen ist der Ansatz analog, wobei noch gesetzt wird

$$\begin{aligned} \hat{P}_{\beta\beta} &= \hat{P}, \quad \beta = 1,\ldots,m && (5.33)\\ \hat{Q}_{\rho\rho} &= \hat{Q}, \quad \rho = 1,\ldots,n && (5.34) \end{aligned}$$

Der Ansatz für $R_{\beta\rho}$ und $\hat{R}_{\beta\rho}$ muß später mit Hilfe von Computersimulationen bestätigt werden.

Mit dem Ansatz läßt sich die Funktion $\Phi_R(m,n)$ um $m = 0$ und $n = 0$ entwickeln. Es gilt

$$\begin{aligned} \Phi_R(m,n) &= \text{sattel}_{\{p,\hat{p},P,\hat{P},r_1,\hat{r_1},r,\hat{r},q,\hat{q},\hat{Q}\}} \\ &\quad 0 + m\cdot A + n\cdot B + m\cdot n\cdot C + \mathcal{O}\left(m^2\right) + \mathcal{O}\left(n^2\right) \qquad (5.35) \end{aligned}$$

Man wendet wieder die Formel (B.2) für den replikasymmetrischen Fall an und erhält

$$\begin{aligned} A &= \hat{P} - \frac{1}{2}\hat{p}p + \alpha \int_{-\infty}^{+\infty} Du\; \Phi\left(-\frac{\kappa_1 + \sqrt{p}u}{\sqrt{1-p}}\right) + \\ &\quad -\frac{1}{2}\ln\left(2\hat{P} - \hat{p}\right) - \frac{1}{2}\cdot\frac{\hat{p}}{2\hat{P} - \hat{p}} \qquad (5.36) \end{aligned}$$

Da $n = 0$ gesetzt wurde, ist A nichts anderes als die Entropie

$$s = \frac{1}{N}\ln V$$

für das erste Perzeptron mit dem Phasenraumvolumen V aus Gl.(1.33). Betrachten wir die Vorschrift (5.20) zur Berechnung der Entropie. Sie besagt, daß der Sattelpunkt an der Stelle $n = 0$, aber bei zunächst allgemeinem m zu bilden ist. Die Ordnungsparameter $p, \hat{p}$ und $\hat{P}$ erhält man also aus den Sattelpunktgleichungen bezüglich der Größe A. Das ist anschaulich klar, denn das

erste Perzeptron hat noch keine Informationen vom zweiten Perzeptron. Die Ordnungsparameter des ersten Perzeptrons dürfen deshalb nicht von den Parametern r_1, r und q abhängen.

Es gilt nach [Ga88]

$$\hat{p} = -\frac{1}{(1-p)^2} + \frac{1}{1-p} \tag{5.37}$$

$$\hat{P} = -\frac{1}{2}\frac{1}{(1-p)^2} + \frac{1}{1-p} \tag{5.38}$$

Der Ordnungsparameter p löst schließlich die Sattelpunktgleichung

$$\frac{\partial A}{\partial p} = 0 \tag{5.39}$$

Die Ordnungsparameter $p, \hat{p}$ und $\hat{P}$ sind also als Funktionen von α und κ_1 bekannt.

Setzen wir nun Gleichung (5.35) in Gl.(5.20) ein, so ist die Entropie s einfach durch den Term B gegeben. Man wendet Formel $B.2$ für den replikasymmetrischen Fall auf $\Phi_R(0,n)$ (siehe Gl.(5.26)) an. Bei der Behandlung der konstanten Terme ist zu beachten, daß gilt:

$$\lim_{m\to 0} \sum_{\beta,\rho} \hat{R}_{\beta\rho} R_{\beta\rho} = \lim_{m\to 0} n \cdot (\hat{r}_1 r_1 + (m-1)\hat{r}r) = n(\hat{r}_1 r_1 - \hat{r}r) \tag{5.40}$$

Die Rechnung ist ansonsten analog zu derjenigen in Abschnitt 4.2.4. Auch hier lassen sich die konjugierten Größen eliminieren. Man erhält schließlich

$$\begin{aligned} s = \; & \mathrm{sattel}_{\{r,r_1,q\}} \\ & -\frac{1}{2}\cdot\frac{{r_1}^2}{E(1-q)} + \frac{1}{2}\cdot\frac{(pr_1 - r)^2}{f(1-p)^2(1-q)} + \frac{1}{2}\cdot\frac{f}{1-q} - \frac{f}{2} + \frac{f}{2}\ln(1-q) + \\ & +\alpha \int_{-\infty}^{+\infty} d\lambda \int_{-\infty}^{+\infty} \frac{d\mu}{2\pi} \exp(i\lambda\mu) \int_{-\infty}^{+\infty} Du \exp\left(i\frac{u\lambda r}{\sqrt{pf}} + \frac{1}{2}\cdot\frac{\lambda^2 r^2}{pf} - \frac{1}{2}q\lambda^2\right) \cdot \\ & \cdot \ln\Phi\left(-\frac{\kappa_2 - \mu}{\sqrt{1-q}}\right) \cdot \frac{\Phi\left(-\frac{\kappa_1 + \sqrt{p}u + i\frac{\lambda}{\sqrt{f}}(r_1 - r)}{\sqrt{1-p}}\right)}{\Phi\left(-\frac{\kappa_1 + \sqrt{p}u}{\sqrt{1-p}}\right)} \end{aligned} \tag{5.41}$$

Dabei ist wieder

$$E = f + \sqrt{\frac{2}{\pi}} w \exp\left(-\frac{1}{2}w^2\right) \tag{5.42}$$

analog zu Gl.(4.28). Die Φ–Funktion des komplexen Arguments ist hier wohldefiniert: Für eine komplexe Konstante $d = d_1 + id_2$ gilt einfach

$$\Phi(d_1 + id_2) = \int_{-\infty}^{0} \frac{d\lambda}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\lambda + d_1 + id_2)^2\right) \tag{5.43}$$

5.2.5 Die Limites $p \to 1$ und $q \to 1$

Durch den Grenzübergang $p \to 1$ legen wir das erste Perzeptron auf den kritischen Punkt fest. Die ursprüngliche Mittelung über alle Perzeptrone des ersten Schritts wird dann scharf. Bei einem vorgegebenem α errechnet sich die kritische Stabilität des ersten Perzeptrons aus der Gleichung von E. Gardner (1.39). Des weiteren vereinfachen sich die Mehrfach–Integrale in Gl.(5.41) erheblich, wenn man den Ansatz

$$r_1 - r = d(1-p) \tag{5.44}$$

macht, wobei auch im Limes $p \to 1$

$$d = \mathcal{O}(1)$$

gelten soll. d ersetzt als neue Sattelpunktvariable r_1. Man kann dann die L'Hospitalsche Regel anwenden, um den Quotienten der beiden Φ–Funktionen zu berechnen [Wen91]. Man erhält mit $\varepsilon = 1 - p$

$$\lim_{\varepsilon \to 0} \frac{\Phi\left(-\frac{1}{\sqrt{\varepsilon}}\left(\kappa_1 + u\sqrt{1-\varepsilon} + i\frac{\lambda}{\sqrt{f}}d \cdot \varepsilon\right)\right)}{\Phi\left(-\frac{1}{\sqrt{\varepsilon}}\left(\kappa_1 + u\sqrt{1-\varepsilon}\right)\right)} =$$

$$\theta((-\kappa_1) - u) + \theta(u - (-\kappa_1)) \cdot \exp\left(-(\kappa_1 + u) \cdot i\frac{\lambda}{\sqrt{f}}d\right) \tag{5.45}$$

Anschließend werden die Integrationen über λ mit Formel (A.8) ausgeführt. Die Integrale über u werden mit Hilfe der allgemeinen Gauß–Formel (A.21) in Φ–Funktionen überführt. Um das Ergebnis in möglichst komprimierter Form darzustellen, transformiere ich noch

$$a = \frac{r}{\sqrt{f}}, \quad b = \frac{r-d}{\sqrt{f}} \tag{5.46}$$

Das Ergebnis am kritischen Punkt ($p \to 1$) des ersten Perzeptrons ist dann

$$\begin{aligned} s \;=\; & \mathrm{sattel}_{\{a,b,q\}} \qquad\qquad (5.47)\\ & \frac{1}{1-q}\left\{-\frac{1}{2}f \cdot \frac{a^2}{E} + \frac{1}{2}b^2 + \frac{f}{2}q + \frac{f}{2}(1-q)\ln(1-q) + \right.\\ & +\alpha(1-q)\int_{-\infty}^{+\infty}\frac{dz}{\sqrt{2\pi}}\ln\Phi\left(-\frac{\kappa_2 + z\sqrt{q}}{\sqrt{1-q}}\right)\cdot\\ & \cdot\left[\exp\left(-\frac{1}{2}z^2\right)\cdot\Phi\left(\frac{1}{\sqrt{q-a^2}}(-\kappa_1\sqrt{q} - za)\right) + \right.\\ & +\sqrt{\frac{q}{q-a^2+b^2}}\cdot\exp\left(-\frac{1}{2}\cdot\frac{\left(z\sqrt{q} + \kappa_1(a-b)\right)^2}{q-a^2+b^2}\right)\cdot\\ & \left.\left.\cdot\Phi\left(\sqrt{\frac{q-a^2+b^2}{q-a^2}}\left(\kappa_1 + \frac{b}{q-a^2+b^2}(z\sqrt{q} + \kappa_1(a-b))\right)\right)\right]\right\} \end{aligned}$$

E ist durch Gl.(5.42) gegeben. κ_1 ist die kritische Stabilität für das erste Perzeptron bei gegebenem α.

Schließlich führen wir den Grenzübergang $q \to 1$ für das zweite Perzeptron durch. Dann ist die Sattelpunktgleichung

$$\frac{\partial s}{\partial q} = 0$$

wie in Abschnitt 4.2.4 gleichbedeutend mit der Gleichung

$$l = \lim_{q \to 1} (1 - q)s = 0$$

Diese Gleichung bestimmt die kritische Stabilität $\kappa_2(\alpha)$ bei gegebenem α. Um die Rechnung etwas zu vereinfachen, geben wir im folgenden κ_2 vor. Gesucht ist dann die kritische Stabilität des ersten Perzeptrons.

5.2.6 Die Ergebnisse der analytischen Rechnung

Gegeben sei der Parameter f für den Einschritt – Verdünnungsalgorithmus. Es wird gefordert, daß das zweite Perzeptron optimal sei mit Stabilität κ_2. Gesucht ist die optimale Stabilität κ_1 des ersten Perzeptrons. Das kritische α des Modells bestimmt sich schließlich aus der Gardner–Gleichung für das erste Perzeptron

$$\alpha(\kappa_1) = \left[\left(1+\kappa_1^2\right)\Phi(\kappa_1) + \frac{\kappa_1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\kappa_1^2\right)\right]^{-1} \tag{5.48}$$

Das Problem wird durch die Betrachtung des Grenzwerts der Entropie gelöst:

$$\begin{aligned} l(\kappa_1,\kappa_2,f) = \lim_{q\to 1}(1-q)\; s = \qquad & (5.49)\\ \text{sattel}_{\{a,b\}} & \\ -\frac{1}{2}\alpha(\kappa_1)\int_{-\kappa_2}^{\infty}\frac{dz}{\sqrt{2\pi}}(\kappa_2+z)^2\cdot & \\ \cdot\left[\exp\left(-\frac{1}{2}z^2\right)\Phi\left(-\frac{\kappa_1+za}{\sqrt{1-a^2}}\right)+\frac{1}{\sqrt{1-a^2+b^2}}\exp\left(-\frac{1}{2}\cdot\frac{(z+\kappa_1(a-b))^2}{1-a^2+b^2}\right)\cdot\right. & \\ \left.\cdot\Phi\left(\sqrt{\frac{1-a^2+b^2}{1-a^2}}\left(\kappa_1+\frac{b}{1-a^2+b^2}(z+\kappa_1(a-b))\right)\right)\right] + & \\ -\frac{fa^2}{2E}+\frac{b^2}{2}+\frac{f}{2} & \end{aligned}$$

Dabei ist $E = f + \sqrt{\frac{2}{\pi}}w\exp\left(-\frac{1}{2}w^2\right)$, w ist die Schranke für die Kopplungen, es löst die Gleichung

$$f = 2\Phi(-w)$$

Die Stabilität κ_1 ist die Nullstelle von l; man hat also das nichtlineare Gleichungssystem

$$\begin{aligned} l(\kappa_1,a,b,\kappa_2,f) &= 0\\ \frac{\partial l}{\partial a} &= 0\\ \frac{\partial l}{\partial b} &= 0 \end{aligned} \tag{5.50}$$

zu lösen und anschließend

$$\alpha(\kappa_2,f) = \alpha(\kappa_1)$$

aus Gl.(5.48) zu berechnen.

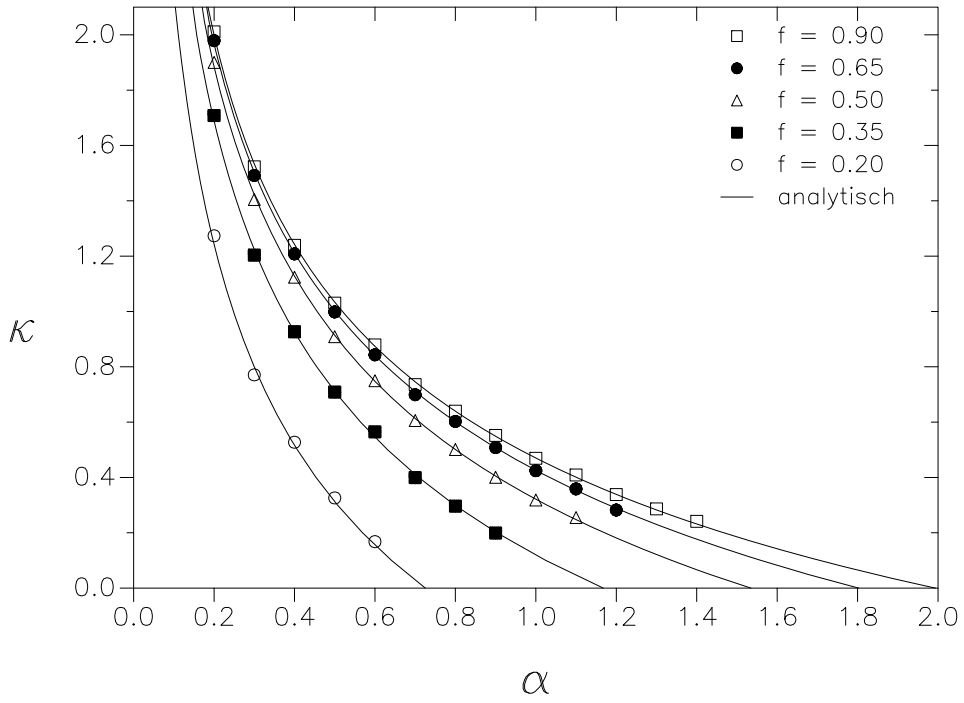


Abbildung 5.1: Die Stabilität $\kappa(\alpha)$ beim Einschrittalgorithmus. Die Computersimulationen wurden mit $N = 200$ Neuronen bei 50 Wiederholungen durchgeführt. Die Simulationen sind durch die Symbole wiedergegeben, wobei die Streuungen kleiner als die Symbolgrößen sind. Die durchgezogene Linie stellt die analytische Rechnung dar.

5.3 Ergebnisse

5.3.1 Die Überprüfung der analytischen Ergebnisse durch die Computersimulation

Die in Abschnitt 5.2 vorgestellte Replika–Rechnung kann wieder durch Computersimulationen überprüft werden. Die Simulationen entsprechen genau dem in Abschnitt 5.1 beschriebenen Algorithmus. Im ersten Schritt wird das Perzeptron optimaler Stabilität mit dem AdaTron–Algorithmus gelernt, und $N(1-f)$ Plätze werden gemäß den Beträgen der Kopplungen entfernt. Im zweiten Schritt wird auf den verbliebenen Plätzen mit dem AdaTron–Algorithmus nachgelernt. Die Simulation wurde mit $N = 200$ durchgeführt, wobei über 50 Läufe gemittelt wurde. In Abbildung 5.1 sehen wir eine gute Übereinstimmung der analytischen Ergebnisse mit den Computersimulationen. Wieder werden in der Simulation keine Stabilitäten $\kappa = 0$ erreicht, weil bei endlichen N auch die Wahrscheinlichkeit endlich ist, daß das zweite Perzeptron nicht existiert (siehe Abschnitt 4.3).

Die Simulationsergebnisse bestätigen jedoch wiederum

$$\alpha_{eff} = \frac{\alpha}{f} > 2 \tag{5.51}$$

Eine weitere Bestätigung der Replika–Rechnung sehen wir, wenn wir den Ordnungsparameter

$$a = \frac{r}{\sqrt{f}} \tag{5.52}$$

simulieren (siehe Gl.(5.46)). Dabei gibt r gemäß unserer replikasymmetrischen Theorie die Überlappung des ersten mit dem zweiten Perzeptron an (siehe Gln.(5.29) – (5.32) und (5.24)):

$$a = \frac{1}{\sqrt{f}N} \sum_{j=1}^{N} c_j J_j T_j \tag{5.53}$$

mit $c_j = \theta(|J_j| - w)$.

Wegen den Normierungen von $\underline{J}$ und $\underline{T}$ folgt aus der Cauchy–Schwarzschen Ungleichung noch

$$a \leq 1 \tag{5.54}$$

a ist ein Maß für die Wichtigkeit des Nachlernens. In Abbildung 5.2 sind die Werte aus den Simulationen und die Werte der analytischen Rechnung in guter Übereinstimmung [2].

Wir sehen, daß das Nachlernen für kleine f und große α wichtig wird. Dies ist anschaulich klar: Bei kleinen f sind die Kopplungen so stark verdünnt, daß die verbleibenden Kopplungen zu viele Fehler in der Klassifikation der Muster erzeugen. Die Zahl der falsch klassifizierten Muster steigt natürlich auch mit dem Parameter α.

[2] Die leichte Diskrepanz bei kleinen f ist auf „finite size“–Effekte zurückzuführen, da die gesamte Simulation bei $N = 200$ durchgeführt wurde. Die Simulationen wurden für kleine f bei $N = 400$ wiederholt. Es zeigte sich hier eine genaue Bestätigung der analytischen Ergebnisse. Die Stabilität κ hingegen erwies sich als weniger anfällig gegenüber “finite size“–Effekten; sie erreichte schon bei kleineren N ihren endgültigen Wert.

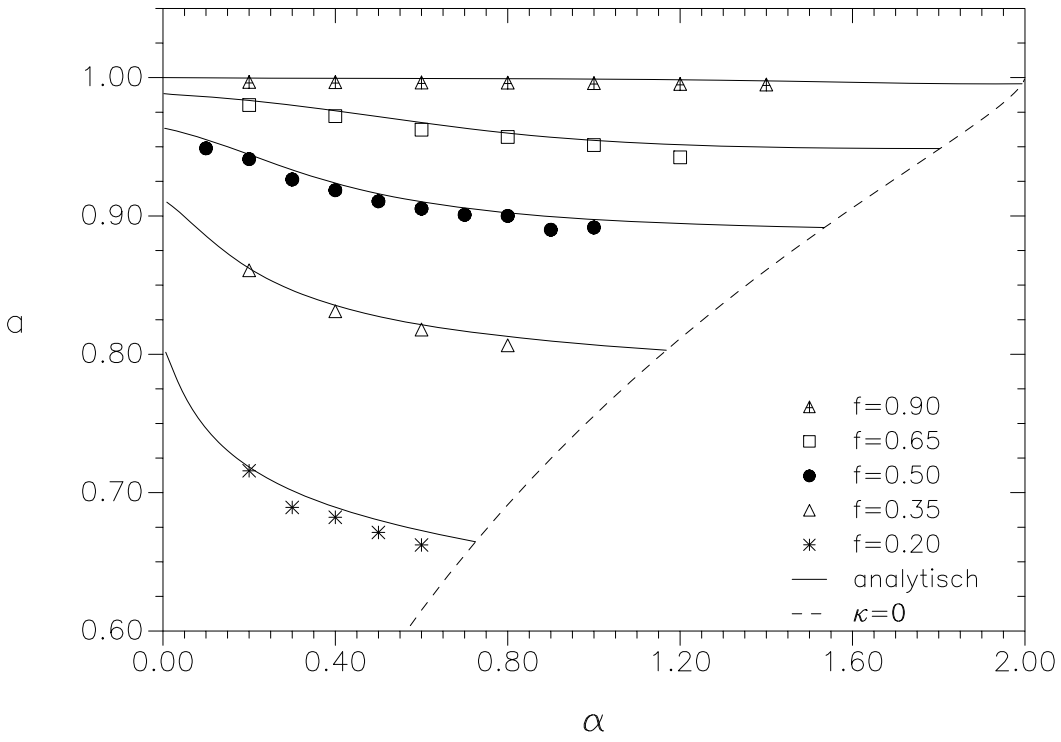


Abbildung 5.2: Die Überlappung a zwischen den Perzeptronen des Verdünnungsschritts und des Nachlernschritts beim Einschrittalgorithmus. Die Computersimulationen (Symbole) werden wieder mit der analytischen Rechnung (durchgezogene Linien) verglichen. Die gestrichelte Linie $a(\kappa = 0)$ stellt die unteren Schranken für a bei gegebenen f dar.

5.3.2 Die Ergebnisse für α_{eff} bei Einschritt– und Mehrschrittalgorithmus

Für den Mehrschrittalgorithmus ist nur eine Computer–Simulation möglich. Die zum Diagramm 5.1 analogen Ergebnisse müssen dann zunächst interpoliert werden. Es zeigt sich (siehe [Gc92] und [Gc+92]), daß die analytischen Kurven aus Abschnitt 5.3.1 dies bewerkstelligen. Dabei gehören zu Fitparametern f_F des Einschrittalgorithmus selbstverständlich kleinere Werte des Verdünnungsparameters f des Mehrschrittalgorithmus. In Abbildung 5.3 sieht man, daß Simulationsergebnisse nur für höhere Werte von κ gewonnen werden können. Dies liegt wieder an der endlichen Wahrscheinlichkeit für die Nichtseparabilität der Muster. Da im vorliegenden Fall $N = 200$ ist und $N(1-f)$ viele Läufe des AdaTron–Algorithmus durchgeführt werden, steigt diese Fehlerwahrscheinlichkeit beträchtlich. Dies ist auf die endliche Systemgröße zurückzuführen, denn für $N \to \infty$ erhalten wir für die Cover–Wahrscheinlichkeit [He+91]

$$W_s = 1 - \mathcal{O}(\exp(-\text{const} \cdot N)) \tag{5.55}$$

Wir erwarten, daß das gleiche Verhalten auch für spätere Verdünnungsschritte vorliegt, daß also für die Gesamtwahrscheinlichkeit

$$W_{ges} \sim {W_s}^N \to 1$$

gilt.

In den Abbildungen 5.4 bzw. 5.5 sehen wir α bzw. α_{eff} für die Stabilität $\kappa = 0$ des letzten Perzeptrons. Wir stellen den Vergleich zu allen bisherigen Rechnungen her. Wie erwartet verbessert der Einschrittalgorithmus den Quersummenalgorithmus erheblich. Beide zeigen jedoch für $f \to 0$ das gleiche divergente Verhalten in α_{eff}.

Der Mehrschrittalgorithmus stellt eine weitere Verbesserung des Einschrittalgorithmus dar und kommt der RSB1–Approximation sehr nahe.

Wegen des erheblichen Rechenaufwands ist der Mehrschrittalgorithmus jedoch nicht für praktische Anwendungen geeignet. Hier sollte der Einschrittalgorithmus oder ein Zweischrittalgorithmus (Verdünnen – Nachlernen – Verdünnen – Nachlernen) vorgezogen werden. In einer analytischen Rechnung für den Zweischrittalgorithmus, die analog zu Abschnitt 5.2 durchzuführen ist, könnte man dabei den Verdünnungsparameter des Zwischenschritts zur Optimierung der Stabilität des letzten Perzeptrons benutzen [3].

Der Mehrschrittalgorithmus ist eher von theoretischer Bedeutung im Hinblick auf die maximal erreichbare Speicherkapazität bei vorgegebenem f. Wir haben die tatsächliche Kurve nun auf den Bereich zwischen der oberen Schranke und der Kurve $\alpha(\kappa = 0)$ des Mehrschrittalgorithmus eingeengt. Die RSB1–Näherung liegt (für alle f) in diesem Bereich.

[3] Eine technische Besonderheit dieser Rechnung ist, daß alle Ordnungsparameter des ersten Schritts schon aus Abschnitt 5.2 bekannt sind

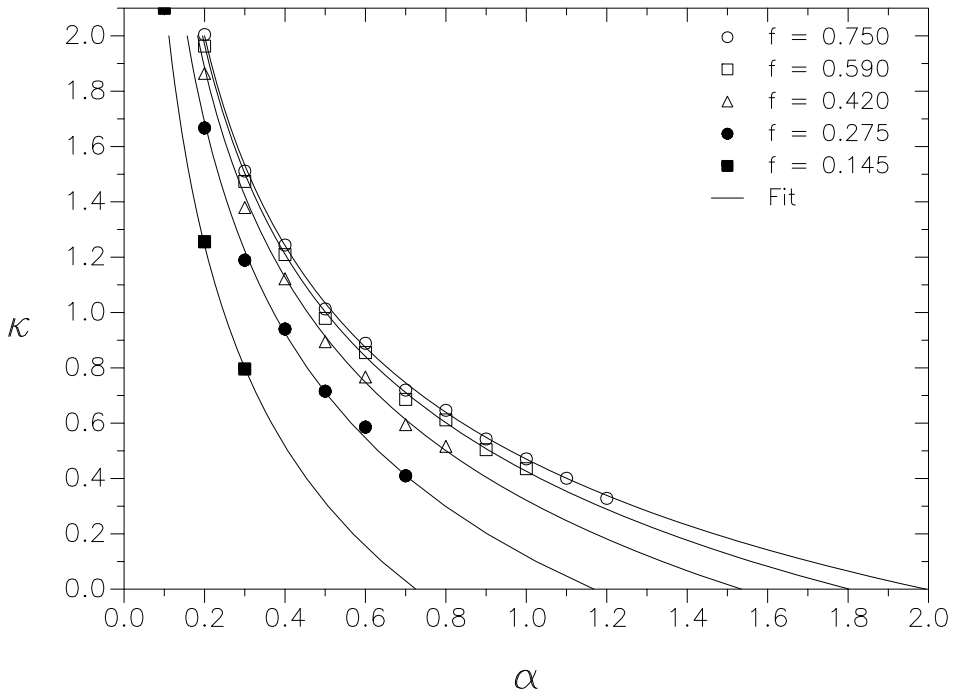


Abbildung 5.3: Die Stabilität $\kappa(\alpha)$ beim Mehrschrittalgorithmus. Die Computersimulationen wurden mit $N = 200$ Neuronen bei 50 Wiederholungen durchgeführt. Die Simulationen sind durch die Symbole wiedergegeben, wobei die Streuungen in der Größenordnung der Symbolgrößen sind. Man beachte, daß die durchgezogene Linie hier eine Inter– und Extrapolation der Daten ist (siehe Text).

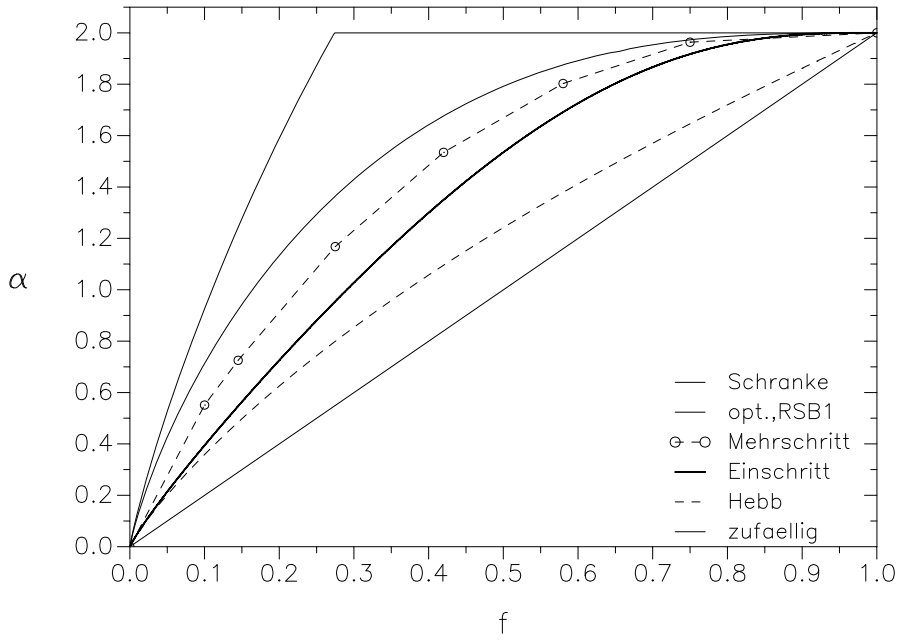


Abbildung 5.4: Alle Ergebnisse für die Speicherkapazitäten $\alpha(\kappa = 0)$ im Vergleich. In aufsteigender Reihenfolge sind abgebildet: zufällige Verdünnung, Quersummenalgorithmus (Hebb), Einschrittalgorithmus, Mehrschrittalgorithmus, RSB1 – Näherung der optimalen Verdünnung und obere Schranke

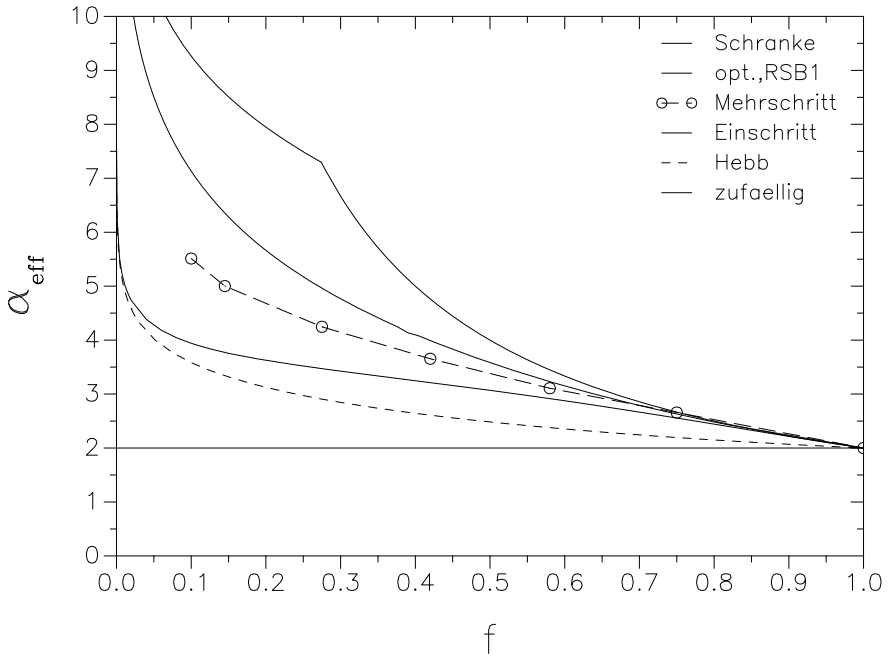


Abbildung 5.5: Alle Ergebnisse für die effektiven Speicherkapazitäten $\alpha_{eff}(\kappa = 0)$ im Vergleich (siehe Abb. 5.4).

5.3.3 Die Verteilung der Kopplungen bei Mehrschrittalgorithmus und optimaler Verdünnung

Die RSB1–Näherung und das Ergebnis des Mehrschrittalgorithmus stimmen nicht nur quantitativ in der Speicherkapazität fast überein. Es zeigt sich auch eine Ähnlichkeit in der Verteilung $P(J)$ der Kopplungen nach dem letzten Nachlernen. In den Abbildungen 5.6 und 5.7 sehen wir sowohl in der Simulation als auch in der analytischen Rechnung eine Höckerstruktur. Kleine Beträge der Kopplungen treten nach dem letzten Nachlernen beim Mehrschrittalgorithmus mit geringer Wahrscheinlichkeit auf. Die RSB1–Näherung liefert ein ähnliches Ergebnis. Im Gegensatz zu der replikasymmetrischen Näherung von Bouten et al. [Bo+90] ist $P(J)$ in der RSB1–Näherung für positive J nicht streng monoton fallend. Offenbar werden auch im optimalen Fall kleine Beträge der Kopplungen benötigt, um alle Muster zu stabilisieren. Die analytische Rechnung zur Gewinnung der Wahrscheinlichkeitsdichte $P(J)$ ist analog zu der RSB1–Rechnung in Abschnitt 3.4.2. Die zugehörigen Formeln sind in Anhang C.3 aufgeführt. Wir erwarten, daß die in der Verteilung auftretende Schranke bei höheren Stufen der Replika–Symmetriebrechung verschwindet.

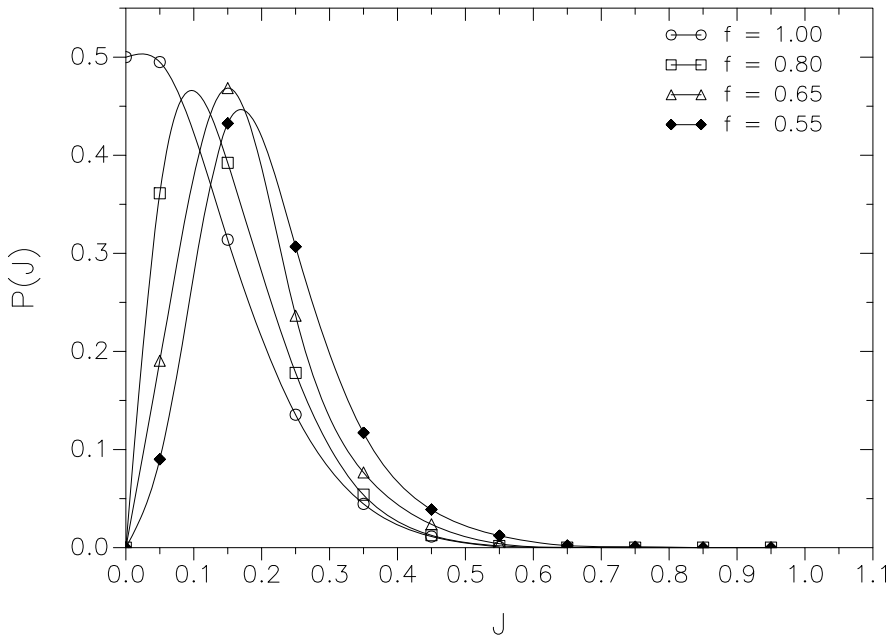


Abbildung 5.6: Die Verteilung der Kopplungen beim Mehrschrittalgorithmus (nach [Gc92]).

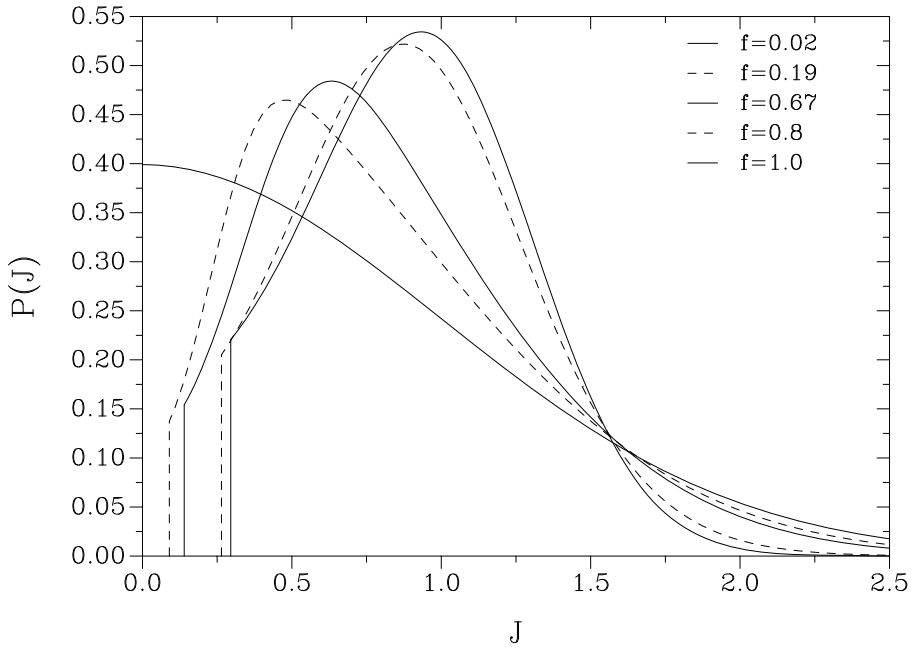


Abbildung 5.7: Die Wahrscheinlichkeitsdichte der Kopplungen bei der optimalen Verdünnung in RSB1–Näherung. Wiedergegeben ist das RSB1–Ergebnis am kritischen Punkt ($q_1 \to 1$) bei der Stabilität $\kappa = 0$. $P(J)$ ist achsensymmetrisch.

Kapitel 6

Die Verallgemeinerungsrate eines verdünnten Perzeptrons

Nachdem Speicherkapazitäten verdünnter Perzeptrone ausgiebig behandelt worden sind, befaßt sich dieses Kapitel mit deren Verallgemeinerungsraten. Unsere Aufgabe ist dabei, mit einem verdünnten Schüler die Ausgaben eines verdünnten Lehrers zu lernen. Da wir vordringlich an qualitativen Aussagen interessiert sind, behandeln wir den analytisch verhältnismäßig einfachen Fall des Quersummenalgorithmus aus Kapitel 4. Das Problem wird in Abschnitt 6.1 ausführlich dargestellt. In Abschnitt 6.2 wird die Schranke w für das Schneiden der Hebb–Kopplungen in Abhängigkeit der vorgegebenen Verdünnungen f_s des Schülers und f_l des Lehrers berechnet. Dieses mathematisch exakte Ergebnis gibt uns einen ersten Aufschluß über die Fähigkeit des Algorithmus, die relevanten Neuronen des Lehrers herauszufinden. In Abschnitt 6.3 sind die zugehörigen Diagramme aufgeführt.

In Abschnit 6.4 befindet sich die replikasymmetrische Rechnung zur Verallgemeinerungsfähigkeit. Die Ergebnisse werden in Abschnitt 6.5 vorgestellt und diskutiert. Dabei wird ein Vergleich zur Verallgemeinerungsfähigkeit eines bezüglich der Stabilität optimal verdünnten Perzeptrons angestellt [Mu92], [KuMu93].

6.1 Problemstellung

In dem in Abschnitt 1.3.3 angesprochenen Verallgemeinerungsproblem geben wir einen verdünnten Lehrer

$$\underline{B} = (B_1, \ldots, B_N)^T \tag{6.1}$$

vor, der auf den letzten $N(1 - f_l)$ Komponenten Nullen hat:

$$B_j = 0 \text{ für } j = Nf_l + 1, \ldots, N \tag{6.2}$$

f_l ist dabei der Verdünnungsparameter des Lehrers. Auf den ersten Nf_l Plätzen erfülle der Lehrer die Normierungsbedingung

$$\frac{1}{Nf_l} \sum_{j=1}^{Nf_l} {B_j}^2 = 1 \tag{6.3}$$

Die Normierungsbedingung ist im Limes $N \to \infty$ erfüllt, wenn wir die Lehrerkopplungen unabhängig voneinander gemäß der Gaußverteilung

$$p(B_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \cdot \frac{(B_j - B_0)^2}{\sigma^2}\right), \quad j = 1, \ldots, N f_l \tag{6.4}$$

auswürfeln, wobei der Mittelwert B_0 und die Streuung σ Punkte auf einem Kreis darstellen: Wegen des Gesetzes der großen Zahlen folgt nämlich aus der Normierungsforderung

$$1 = \int_{-\infty}^{+\infty} dB \; p(B) \cdot B^2 = \sigma^2 + {B_0}^2 \tag{6.5}$$

Dabei ist der Fall $B_0 = 1$, d.h.

$$p(B) = \delta(B - 1) \tag{6.6}$$

ausdrücklich zugelassen.

Dem Schüler werden nun gaußverteilte Muster $\underline{\xi}^\mu$ der Reihe nach vorgelegt. Die zugehörigen Ausgaben S^μ sind dabei vom verdünnten Lehrer vorgegeben

$$S^\mu = \text{sign}\left(\frac{1}{\sqrt{N f_l}} \sum_{j=1}^{N f_l} B_j \xi_j^\mu\right) \tag{6.7}$$

Wäre der Schüler unverdünnt, so würde die Verteilung der Lehrerkopplungen keine Rolle spielen und wir erhielten das bekannte Ergebnis für die Verallgemeinerungsfähigkeit des optimalen Perzeptrons [Ne91], [Op+90]. Fordern wir aber, daß der Schüler einen Verdünnungsparameter f_s besitzt, so entsteht das Problem der Wahl der günstigen Plätze. Gehen wir in Analogie zu den unverdünnten Perzeptronen davon aus, daß hohe Stabilitäten κ auch hohe Verallgemeinerungsraten bedingen, dann ist das bezüglich der Stabilität optimal verdünnte Perzeptron zu studieren [Mu92], [KuMu93]. Um zu belegen, daß die dort auftretenden Effekte auch durch einen praktikablen Algorithmus erzielt werden können, betrachten wir im folgenden den Quersummenalgorithmus:

Die Plätze des Schülers werden hier gemäß

$$c_j = \theta(|H_j| - w) \tag{6.8}$$

ausgewählt, wobei die H_j die Hebb–Kopplungen sind

$$H_j = \frac{1}{\sqrt{N\alpha}} \sum_{\mu=1}^{p} S^\mu \xi_j^\mu \tag{6.9}$$

Man beachte, daß die Ausgaben S^μ nach der obigen Gl.(6.7) durch den Lehrer gegeben sind. Die Form des Lehrers ist dem Schüler dabei in keiner Weise bekannt.

Nach der Auswahl der Plätze wird das Perzeptron optimaler Stabilität gelernt. Der Perzeptronvektor soll wieder der Normierungsbedingung

$$\sum_{k=1}^{Nf_s} {J_k}^2 = \sum_{j=1}^{N} c_j {T_j}^2 = Nf_s \tag{6.10}$$

genügen.

Die Formel (1.51) für die Überlappung von Lehrer und Schüler lautet dann mit den entsprechenden Normierungen

$$R = \frac{1}{N\sqrt{f_l f_s}} \sum_{j=1}^{N} c_j B_j T_j \tag{6.11}$$

Gemäß Gl.(1.49) gilt für die Verallgemeinerungsrate

$$G = 1 - \frac{1}{\pi} \arccos R \tag{6.12}$$

Die Funktion $G(\alpha)$ wird in Abschnitt 6.4 mit Hilfe einer Gardner–Rechnung gewonnen.

Zunächst berechnen wir in Abschnitt 6.2 die Schranke w als Funktion der vorgegebenen Parameter α, B_0, f_l und f_s. Wir gewinnen dabei einen ersten Hinweis auf die Verallgemeinerungsfähigkeit des Quersummenalgorithmus.

6.2 Die allgemeine Verdünnungsformel

Um die Bestimmungsgleichung für die Schranke w im Limes $N \to \infty$ zu erhalten, formulieren wir f_s zunächst als Funktion von w, f_l, B_0 und α. Es gilt

$$f_s = \frac{1}{N} \sum_{j=1}^{N} c_j \tag{6.13}$$

Analog zu Abschnitt 4.2.1 gehen wir wieder davon aus, daß f_s selbstmittelnd bezüglich der Mustermittelung ist:

$$\begin{aligned} \lim_{N\to\infty} f_s &= \lim_{N\to\infty} \langle\langle f_s \rangle\rangle_{\{\xi_j^\mu\}} \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{r=1}^{N} \langle\langle \theta(|H_r| - w) \rangle\rangle \end{aligned} \tag{6.14}$$

Wir berechnen $A_r = \langle\langle \theta(|H_r| - w) \rangle\rangle$ mit den bekannten Methoden

$$\begin{aligned} A_r = & \\ & \left\langle \left\langle \int_{-\infty}^{+\infty} dH_r \int_{-\infty}^{+\infty} \frac{d\hat{h}_r}{2\pi} \left(\prod_{\mu=1}^{p} \int_{-\infty}^{+\infty} dU_\mu \int_{-\infty}^{+\infty} \frac{d\hat{u}_\mu}{2\pi} \right) \cdot \theta(|H_r| - w) \cdot \right. \right. \\ & \left. \left. \cdot \exp\left[i\hat{h}_r \left(H_r - \frac{1}{\sqrt{N}\alpha} \sum_{\mu=1}^{p} \mathrm{sign}\,(U_\mu) \xi_r^\mu \right) + i \sum_{\mu=1}^{p} \hat{u}_\mu \left(U_\mu - \frac{1}{\sqrt{Nf_l}} \sum_{j=1}^{N} B_j \xi_j^\mu \right) \right] \right\rangle \right\rangle \end{aligned}$$

$$= \int_{-\infty}^{+\infty} \frac{dH}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} D\hat{h}\ \exp\left(i\hat{h}H\right)\ \theta(|H|-w)\cdot$$
$$\left(\int_{-\infty}^{+\infty} \frac{dU}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(U + i\,\mathrm{sign}\,(U)\cdot\frac{\hat{h}B_r}{N\sqrt{\alpha f_l}}\right)^2\right]\right)^p \tag{6.15}$$

Daraus folgt für f_s

$$f_s = \int_{-\infty}^{+\infty} \frac{dH}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} D\hat{h}\ \exp\left(i\hat{h}H\right)\ \theta(|H|-w)\cdot$$
$$\lim_{N\to\infty} \frac{1}{N}\sum_{r=1}^{N}\ 2^{\alpha N}\cdot \exp\left[\alpha N \ln\Phi\left(-\frac{i\hat{h}B_r}{N\sqrt{\alpha f_l}}\right)\right] \tag{6.16}$$

Wir wenden Formel (A.12) zur Entwicklung der Φ–Funktion an. Anschließend ersetzen wir wegen des Gesetzes der großen Zahlen die Summe über alle Kopplungen durch einen Erwartungswert bezüglich der Verteilung (6.4).

$$\lim_{N\to\infty} \frac{1}{N}\sum_{r=1}^{N} g(B_r) = f_l \int_{-\infty}^{+\infty} dB\ p(B)\ g(B) + (1-f_l)g(0) \tag{6.17}$$

Wir setzen also für die allgemeine Funktion g in der obigen Gleichung den Integranden aus Gl.(6.16) ein. Nach Ausführung des Integrals über $\hat{h}$ erhalten wir

$$f_s = f_l \int_{-\infty}^{+\infty} dB\ p(B) \int_{-\infty}^{+\infty} \frac{dH}{\sqrt{2\pi}}\ \theta(|H|-w)\cdot\exp\left[-\frac{1}{2}\left(H-\sqrt{\frac{2\alpha}{\pi f_l}}B\right)^2\right] +$$
$$+(1-f_l)\cdot 2\Phi(-w) \tag{6.18}$$

Setzt man die Gaußverteilung (6.4) ein, so folgt das Endergebnis

$$f_s = f_c + (1-f_l)\cdot 2\Phi(-w) \tag{6.19}$$

mit

$$f_c = f_l \cdot \left[\Phi\left(-\frac{w}{\sqrt{1+\frac{2\alpha}{\pi f_l}(1-{B_0}^2)}} + \frac{B_0\sqrt{\frac{2\alpha}{\pi f_l}}}{\sqrt{1+\frac{2\alpha}{\pi f_l}(1-{B_0}^2)}}\right) + \right.$$
$$\left. +\Phi\left(-\frac{w}{\sqrt{1+\frac{2\alpha}{\pi f_l}(1-{B_0}^2)}} - \frac{B_0\sqrt{\frac{2\alpha}{\pi f_l}}}{\sqrt{1+\frac{2\alpha}{\pi f_l}(1-{B_0}^2)}}\right)\right] \tag{6.20}$$

Die Schranke w folgt bei gegebenen f_s, f_l, α und B_0 als Lösung der Gleichung für f_s [1].

[1] In dem Ergebnis ist auch der Spezialfall $B_0 = 1$ enthalten.

Der erste Term f_c in der Gleichung für f_s gibt dabei an, wieviele relevante Plätze [2] des Lehrers der Schüler herausgefunden hat:

$$f_c = \frac{1}{N} \sum_{j=1}^{N f_l} c_j \tag{6.21}$$

f_c ist also der Prozentsatz der Zahl der Plätze, auf denen die Verdünnungsvektoren von Lehrer und Schüler übereinstimmen. Wir bezeichnen f_c deshalb als die normierte Zahl der übereinstimmenden Plätze von Lehrer und Schüler.

6.3 f_c als Funktion von f_l, α und B_0

Bei gegebenem f_l geben die Kurven $f_c(\alpha)$ eine Auskunft darüber, wie schnell der Schüler die relevanten Plätze des Lehrers herausfindet. Wir erwarten, daß die Verallgemeinerungsrate einen ähnlichen Verlauf hat.

Im Grenzübergang $\alpha \to \infty$ muß ein intelligenter Verdünnungsalgorithmus natürlich alle Plätze des Lehrers herausfinden, falls $f_s \geq f_l$. Falls der Schüler im Falle $f_s < f_l$ zu wenig Plätze zur Verfügung hat, so sollten sich die aktiven Neuronen des Schülers unter den aktiven Neuronen des Lehrers befinden. Wir fordern also

$$\lim_{\alpha \to \infty} f_c = \min(f_s, f_l) \tag{6.22}$$

Dies ist nicht selbstverständlich, denn bei einer zufälligen Auswahl der c_j gilt stets

$$f_c^{(z)}(\alpha) = f_s \cdot f_l \tag{6.23}$$

Berechnen wir also nach Gl.(6.19) die Schranke w und anschließend $f_c(\alpha)$, so beobachten wir, daß tatsächlich der obige Grenzwert angestrebt wird. Dies gilt für alle $B_0 \in [0, 1]$. In Abbildung 6.1 sind $f_c(\alpha)$–Kurven bei festen Werten der Verdünnungsparameter ($f_l = 0.2$ und $f_s = 0.4$) für verschiedene B_0–Werte zu sehen. Wir beobachten, daß sich die f_c–Kurven zum Parameter $B_0 = 1$ durch sehr schnelle Konvergenz für $\alpha \to \infty$ auszeichnen. Den gleichen Effekt beobachtet man auch im Fall $f_s < f_l$ in Abbildung 6.2. Offenbar nähert der Schüler den Lehrer umso besser an, je schwächer das Rauschen in den B_j–Kopplungen ist. Da für alle B_0–Werte das gleiche qualitative Verhalten vorliegt, setzen wir in den folgenden Abschnitten

$$B_0 = 1$$

um die auftretenden Effekte möglichst stark auszuprägen.

[2] Der Index c steht für „coinciding“

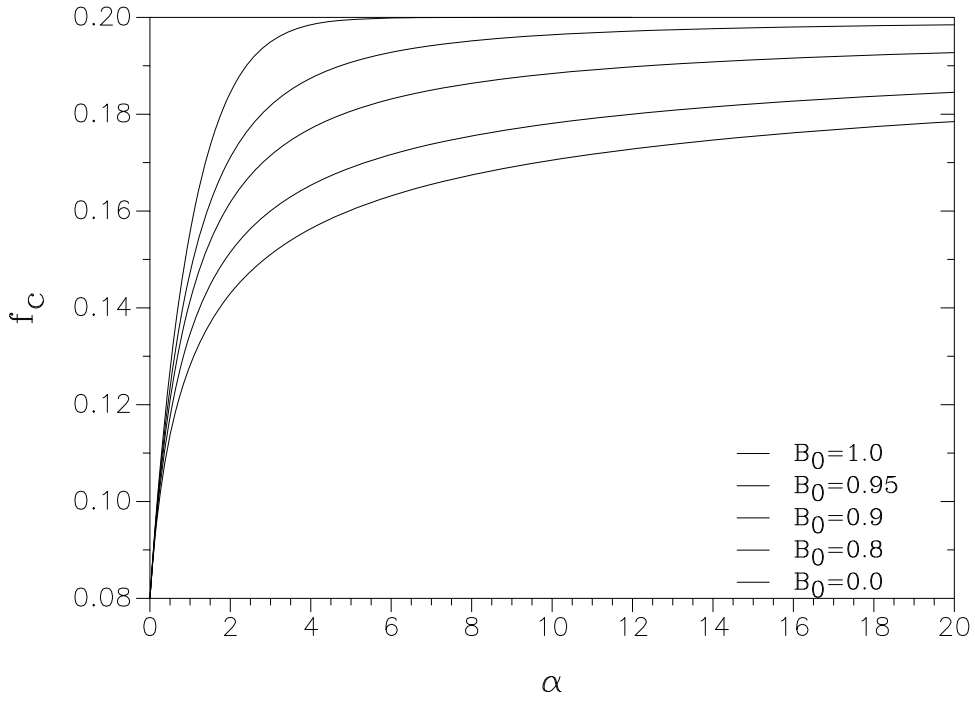


Abbildung 6.1: Die normierte Zahl f_c der übereinstimmenden Plätze von Lehrer und Schüler bei festen Parametern $f_s = 0.4$ und $f_l = 0.2$, d.h. $f_s > f_l$. Für niedrige Mittelwerte B_0 des Lehrers strebt f_c seinen Grenzwert $f_l = 0.2$ nur sehr langsam an.

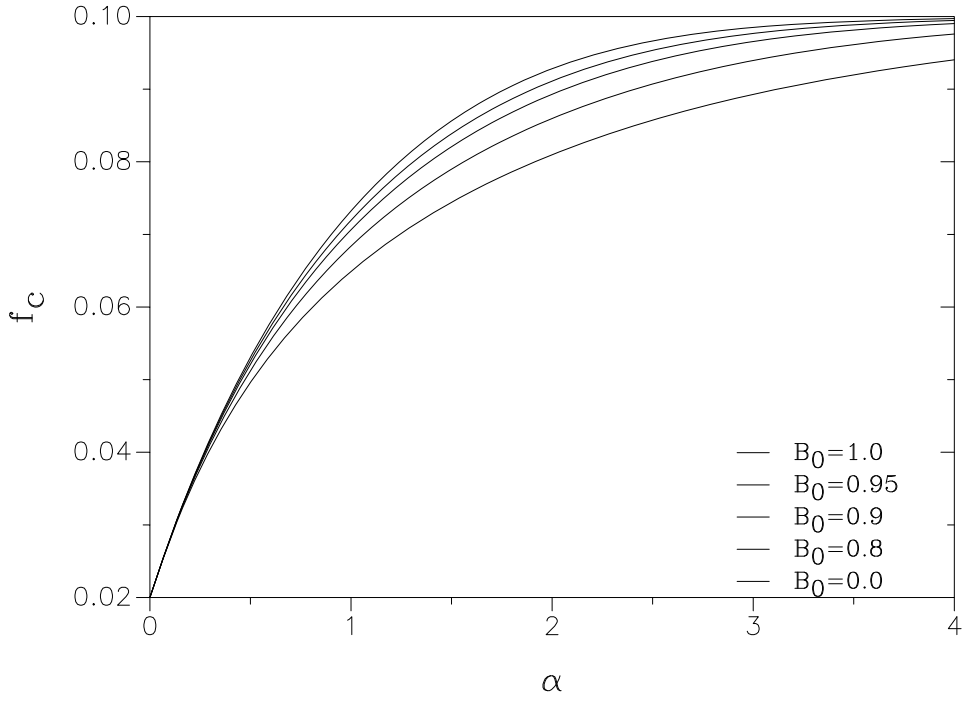


Abbildung 6.2: f_c bei festen Parametern $f_s = 0.1$ und $f_l = 0.2$, d.h. $f_s < f_l$. f_c konvergiert in diesem Fall für alle B_0 schnell gegen den Grenzwert $f_s = 0.1$.

6.4 Die Berechnung der Verallgemeinerungsrate

6.4.1 Die Rechnung bis zum allgemeinen Ergebnis

Wie in Abschnitt 1.3.3 beschrieben, ist eine Gardner–Rechnung analog zu Abschnitt 4.2 durchzuführen. Die Überlappung R zwischen Lehrer und Schüler tritt dabei als zusätzlicher Ordnungsparameter auf. Die Ausgaben sind nicht mehr zufällig, sondern sie sind durch einen Lehrer gegeben.

$$S^\mu = \mathrm{sign}\left(\frac{1}{\sqrt{N f_l}} \sum_{j=1}^{N f_l} B_j \xi_j^\mu\right) \quad \forall \mu \tag{6.24}$$

Wir führen die zugehörigen inneren Felder des Lehrers

$$U_\mu = \frac{1}{\sqrt{N f_l}} \sum_{j=1}^{N f_l} B_j \xi_j^\mu \quad \forall \mu \tag{6.25}$$

mit konjugierten Variablen $\hat{u}_\mu$ ein und setzen die Rechnung nach der Formel (4.19) für $\langle\langle V^n \rangle\rangle$ fort ($\sigma_j^\mu = S^\mu \xi_j^\mu$). Wir erhalten nach der Mittelung über die Muster

$$\begin{aligned}
&\langle\langle V^n \rangle\rangle = \qquad (6.26)\\
&\left(\prod_{j=1}^{N}\prod_{\rho=1}^{n}\int_{-\infty}^{+\infty} DT_j^\rho\right)\left(\prod_{\rho=1}^{n}\delta\left(\sum_{j=1}^{N} c_j\left(T_j^\rho\right)^2 - N f_s\right)\right)\left(\prod_{j=1}^{N}\int_{-\infty}^{+\infty} dH_j \int_{-\infty}^{+\infty}\frac{d\hat{h}_j}{2\pi}\right)\cdot\\
&\cdot\exp\left(i\sum_{j=1}^{N}\hat{h}_j H_j\right)\left(\prod_{\mu=1}^{p}\prod_{\rho=1}^{n}\int_{\kappa}^{\infty} dX_\mu^\rho \int_{-\infty}^{+\infty}\frac{d\hat{x}_\mu^\rho}{2\pi}\right)\exp\left(i\sum_{\mu,\rho}\hat{x}_\mu^\rho X_\mu^\rho\right)\cdot\\
&\cdot\left(\prod_{\mu=1}^{p}\int_{-\infty}^{+\infty} dU_\mu \int_{-\infty}^{+\infty}\frac{d\hat{u}_\mu}{2\pi}\right)\exp\left(i\sum_{\mu=1}^{p}\hat{u}_\mu U_\mu\right)\cdot\\
&\cdot\exp\left[-\frac{1}{2N}\sum_{j,\mu}\left(\sum_{\rho=1}^{n}\hat{x}_\mu^\rho\cdot\frac{\mathrm{sign}\,(U_\mu)}{\sqrt{f_s}}c_j T_j^\rho + \frac{\hat{u}_\mu}{\sqrt{f_l}}\cdot B_j + \frac{\hat{h}_j}{\sqrt{\alpha}}\,\mathrm{sign}\,(U_\mu)\right)^2\right]
\end{aligned}$$

Es gilt wieder

$$c_j = \theta(|H_j| - w)$$

Nach der Ausführung des Quadrats in der obigen Formel steht wieder die Entkopplung der j–Variablen von den μ–Variablen an. Wir führen dazu die folgenden Ordnungsparameter ein:

$$Q_{\rho\sigma} = \frac{1}{N f_s}\sum_{j=1}^{N} c_j T_j^\rho T_j^\sigma \tag{6.27}$$

$$R_\rho = \frac{1}{N\sqrt{f_s f_l}}\sum_{j=1}^{N} c_j T_j^\rho B_j \tag{6.28}$$

$$\hat{K}_\rho = \frac{1}{N}\sum_{\mu=1}^{p} \hat{x}_\mu^\rho \tag{6.29}$$

$$S = \frac{1}{N}\sum_{\mu=1}^{p} \hat{u}_\mu \,\mathrm{sign}\,(U_\mu) \tag{6.30}$$

Die zugehörigen konjugierten Variablen sind $\hat{Q}_{\rho\sigma}$, $\hat{R}_\rho$, k_ρ und $\hat{s}$. R_ρ ist der Ordnungsparameter, der die Überlappung von Lehrer und Schüler angibt. Sein Wert am Sattelpunkt dient zur Berechnung der Verallgemeinerungsrate nach Gl.(6.12).

Wir lösen die Gauß–Integrale über $\hat{h}_j$ und $\hat{u}_\mu$, fassen die Terme in den j–Variablen bzw. in den μ–Variablen zusammen und bilden

$$\Phi_R(n) = \lim_{N\to\infty} \frac{1}{N} \ln \langle\langle V^n \rangle\rangle \tag{6.31}$$

Für $n \to 0$ erhalten wir in den Sattelpunktgleichungen für S und $\hat{s}$

$$\lim_{n\to 0} \frac{\partial \Phi_R}{\partial S} = 0 \Longrightarrow \hat{s} = 0 \tag{6.32}$$

$$\lim_{n\to 0} \frac{\partial \Phi_R}{\partial \hat{s}} = 0 \Longrightarrow S - i\alpha \left.\frac{\Phi'(-i\hat{s})}{\Phi(-i\hat{s})}\right|_{\hat{s}=0} = 0$$

$$\Longrightarrow \quad S = i\alpha\sqrt{\frac{2}{\pi}} \tag{6.33}$$

Damit ist wie erwartet

$$\Phi_R(n) = \mathcal{O}(n)$$

und wir erhalten das allgemeine Endergebnis

$$\begin{aligned}
&\Phi_R(n) = \qquad\qquad (6.34)\\
&\quad \mathrm{sattel}_{\{Q_{\rho\sigma},\hat{Q}_{\rho\sigma},R_\rho,\hat{R}_\rho,\hat{K}_\rho,k_\rho\}}\\
&\quad \sum_{\rho\le\sigma} Q_{\rho\sigma}\hat{Q}_{\rho\sigma} + \sum_{\rho=1}^{n} R_\rho\hat{R}_\rho + \sum_{\rho=1}^{n} k_\rho\hat{K}_\rho - \frac{1}{2\alpha}\sum_{\rho,\sigma}\hat{K}_\rho Q_{\rho\sigma}\hat{K}_\sigma + \sqrt{\frac{2}{\pi}}\sum_{\rho=1}^{n} R_\rho\hat{K}_\rho +\\
&\quad f_l \int_{-\infty}^{+\infty} dB\; p(B)\; \ln I(B) + (1-f_l)\int_{-\infty}^{+\infty} dB\; \delta(B)\; \ln I(B) +\\
&\quad +\alpha \ln\left[\left(\prod_{\rho=1}^{n}\int_{\kappa}^{\infty} dX_\rho \int_{-\infty}^{+\infty}\frac{d\hat{x}_\rho}{2\pi}\right)\int_{-\infty}^{+\infty} DU\cdot\right.\\
&\qquad \left.\cdot\exp\left[-\frac{1}{2}\sum_{\rho,\sigma}\hat{x}_\rho(Q_{\rho\sigma}-R_\rho R_\sigma)\hat{x}_\sigma + i\sum_{\rho=1}^{n}\hat{x}_\rho(X_\rho - k_\rho - |U|\,R_\rho)\right]\right]
\end{aligned}$$

Für das Kopplungsintegral $I(B)$ gilt

$$I(B) = \left(\prod_{\rho=1}^{n}\int_{-\infty}^{+\infty} DJ_\rho\right)\int_{-\infty}^{+\infty} DH\cdot \tag{6.35}$$

$$\cdot \exp\left[-\frac{1}{f_s} \sum_{\rho \leq \sigma} c J_\rho \hat{Q}_{\rho\sigma} J_\sigma - \frac{1}{\sqrt{f_s f_l}} \sum_{\rho=1}^{n} \hat{R}_\rho c J_\rho B + \right.$$
$$\left. - \sum_{\rho=1}^{n} \hat{K}_\rho \cdot \frac{1}{\sqrt{\alpha f_s}} c J_\rho H + \sqrt{\frac{2\alpha}{\pi f_l}} H \cdot B \right]$$

Dabei ist der Verdünnungskoeffizient im H–Integral

$$c = \theta(|H| - w) \tag{6.36}$$

Die Verteilung der Kopplungen

$$p(B) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \cdot \frac{(B - B_0)^2}{\sigma^2} \right) \tag{6.37}$$

wurde nach dem Gesetz der großen Zahlen gemäß der Gl.(6.17) eingeführt. Die Bedingung $B_j = 0$ für $j = N f_l + 1, \ldots, N$ wird durch die δ–Funktion

$$\delta(B - 0)$$

gewährleistet.

6.4.2 Die replikasymmetrische Annahme

Die Herleitung des Endergebnisses verläuft analog zum Abschnitt 4.2.4. Zur Durchführung der vollständigen Rechnung sind jedoch umfangreiche algebraische Umformungen vorzunehmen, bei denen alle in dieser Arbeit behandelten Techniken angewandt werden. Ich skizziere im folgenden den weiteren Verlauf der Rechnung.

Für die eingeführten Ordnungsparameter $Q_{\rho\sigma}, R_\rho, k_\rho$ und die konjugierten Variablen wird die replikasymmetrische Annahme gemacht.

$$Q_{\rho\sigma} = q, \quad \rho, \sigma = 1, \ldots, n, \ \rho \neq \sigma$$

und

$$k_\rho = k, \ R_\rho = R, \ \hat{Q}_{\rho\rho} = \hat{Q}, \quad \rho = 1, \ldots, n$$

Außerdem gilt wieder für die Norm der Perzeptron–Kopplungen

$$Q_{\rho\rho} = 1, \quad \rho = 1, \ldots, n$$

Die Berechnung der Entropie

$$s = \left. \frac{\partial \Phi_R}{\partial n} \right|_{n=0} \tag{6.38}$$

erfolgt wieder mit Hilfe der replikasymmetrischen Version von Gl.(B.2). Die Sattelpunktgleichungen bezüglich der Hilfsvariablen $\hat{q}, \hat{Q}, \hat{K}$ und $\hat{R}$ sind algebraisch, sie lassen sich auflösen. Man erhält das replikasymmetrische Ergebnis

$s = \text{sattel}_{\{q,k,R\}}$

$$\begin{aligned}
&\frac{f_s}{2\,(1-q)} - \frac{f_s}{2} + R\hat{R} + k\hat{K} - \frac{1}{2\alpha}\hat{K}^2\,(1-q) + \sqrt{\frac{2}{\pi}}\hat{K}R + \\
&+ \frac{f_s}{2}\ln(1-q) + \frac{f_l}{2}\,(1-q)\cdot\left(\frac{\hat{R}^2}{f_s f_l}I_1 + 2\frac{\hat{R}\hat{K}}{f_s\sqrt{\alpha f_l}}I_2 + \frac{\hat{K}^2}{\alpha f_s}I_3\right) + \qquad (6.39)\\
&+ \frac{1-f_l}{2}\,(1-q)\,\frac{\hat{K}^2}{\alpha f_s}I_4 + 2\alpha\int_0^\infty Du\int_{-\infty}^\infty Dz\ln\Phi\left(-\frac{\kappa - k - uR + z\sqrt{q-R^2}}{\sqrt{1-q}}\right)
\end{aligned}$$

Dabei sind die Integrale $I_1,\ldots,I_4$ Funktionen der Schranke w. Die Hilfsvariablen $\hat{R}$ und $\hat{K}$ sind am Sattelpunkt Funktionen der Parameter und der Ordnungsparameter q, k, R. Die Hilfsgrößen werden im folgenden Schritt für Schritt angegeben.

1. Die Schranke w wird als Lösung der Gl.(6.19) gewonnen.

2. Dann gilt für die Integrale $I_1,\ldots,I_4$

$$I_1 = \left\langle B^2\right\rangle_{\{B,H\}} \qquad (6.40)$$

$$I_2 = \langle B\cdot H\rangle_{\{B,H\}} \qquad (6.41)$$

$$I_3 = \left\langle H^2\right\rangle_{\{B,H\}} \qquad (6.42)$$

$$I_4 = \int_{-\infty}^{+\infty} DH\ \theta(|H|-w)\cdot H^2 \qquad (6.43)$$

Der Mittelwert über die Integrationsvariablen B (Lehrer) und H (Hebb–Kopplung) ist in den Integralen I_1,I_2 und I_3 wie folgt definiert

$$\langle\,(\ldots)\,\rangle_{\{B,H\}} = \int_{-\infty}^{+\infty} dB\ p(B)\int_{-\infty}^{+\infty}\frac{dH}{\sqrt{2\pi}}\,\theta(|H|-w)\exp\left(-\frac{1}{2}\left(H-\sqrt{\frac{2\alpha}{\pi f_l}}B\right)^2\right)\cdot(\ldots)$$

3. Wir bilden die Abkürzungen C und D

$$C = -\sqrt{\frac{f_l}{{f_s}^2\alpha}}I_2\left[-\frac{1}{\alpha} + f_l\frac{I_3}{\alpha f_s} + (1-f_l)\frac{I_4}{\alpha f_s}\right]^{-1} \qquad (6.44)$$

$$D = \left(-h - \sqrt{\frac{2}{\pi}}R\right)\left[-\frac{1}{\alpha} + f_l\frac{I_3}{\alpha f_s} + (1-f_l)\frac{I_4}{\alpha f_s}\right]^{-1} \qquad (6.45)$$

4. Dann gilt für die Hilfsvariablen

$$(1-q)\hat{R} = -\frac{f_s R + I_2\sqrt{\frac{f_l}{\alpha}}D}{I_1 + I_2\sqrt{\frac{f_l}{\alpha}}C} \qquad (6.46)$$

$$(1-q)\hat{K} = C(1-q)\hat{R} + D \qquad (6.47)$$

Man bildet die partiellen Ableitungen und löst das Gleichungssystem

$$\left(\frac{\partial s}{\partial q}, \frac{\partial s}{\partial k}, \frac{\partial s}{\partial R}\right) = (0,0,0) \tag{6.48}$$

Das beim Quersummenalgorithmus im zweiten Schritt gelernte Perzeptron erreicht im Limes $q \to 1$ seine optimale Stabilität. In den Sattelpunktgleichungen vereinfachen sich die z–Integrale in diesem Limes. Die zugehörigen Formeln und die drei Sattelpunktgleichungen sind in Anhang D aufgeführt. Bei gegebenen Parametern α, f_s, f_l und B_0 erhalten wir aus der Sattelpunktgleichung für q im Limes $q \to 1$ die kritische Stabilität κ. Aus den beiden anderen Sattelpunktgleichungen gewinnen wir k und die Überlappung R. Schließlich berechnen wir die Verallgemeinerungsrate

$$G(\alpha) = 1 - \frac{1}{\pi} \arccos R \tag{6.49}$$

6.5 Ergebnisse

Die Sattelpunktgleichungen aus Anhang D wurden bei gegebenen Parametern α, f_l und f_s gelöst. Um die auftretenden Effekte möglichst stark auszuprägen, wurde stets $B_0 = 1$ gesetzt, d.h. der Lehrer hat die einfache Form

$$\underline{B} = (1, \ldots, 1, 0, \ldots, 0)^T \tag{6.50}$$

Dabei ist $B_j = 1$ für $j = 1, \ldots, Nf_l$ und $B_j = 0$ für $j = Nf_l + 1, \ldots, N$.

Berechnet wurden die optimale Stabilität $\kappa(\alpha)$ und die Verallgemeinerungsrate $G(\alpha)$ des mit dem Quersummenalgorithmus verdünnten Schülers. Wir vergleichen die Ergebnisse mit denen des (bezüglich der Stabilität) optimal verdünnten[3] Perzeptrons [Mu92], [KuMu93]. Wie erwartet sehen wir in Abbildung 6.3, daß die Stabilität des Quersummenalgorithmus niedriger als die optimale Stabilität ist. Die $\kappa(\alpha)$–Kurve der optimalen Verdünnung ist streng monoton fallend, während beim Quersummenalgorithmus ein lokales Minimum bei niedrigen α auftreten kann. Betrachten wir die Verallgemeinerungsraten $G(\alpha)$ in Abbildung 6.4, so beobachten wir ein weiteres überraschendes Ergebnis: Der Quersummenalgorithmus verallgemeinert (in gewissen Parameterbereichen) besser als der (bezüglich der Stabilität) optimale Verdünnungsalgorithmus. Die $G(\alpha)$–Kurven steigen natürlich auch streng monoton, da mit jedem hinzugekommenen Muster $\underline{\xi}^\mu$ und der zugehörigen, vom Lehrer vorgelegten Ausgabe

$$S^\mu = \text{sign}\left(\frac{1}{\sqrt{Nf_l}} \sum_{j=1}^{Nf_l} B_j \xi_j^\mu\right) \tag{6.51}$$

das Volumen abnimmt, in dem der Lehrer liegen kann.

Um das Auftreten der Minima in den $\kappa(\alpha)$–Kurven beim Quersummenalgorithmus besser zu verstehen, betrachten wir Abbildung 6.5. Hier wurden Verdünnungsparameter f_s des Schülers in der Umgebung des Verdünnungsparameters $f_l = 0.1$ des Lehrers gewählt.

[3] Wir benutzen im folgenden den Ausdruck „optimale Verdünnung“ stets in Zusammenhang mit der Stabilität

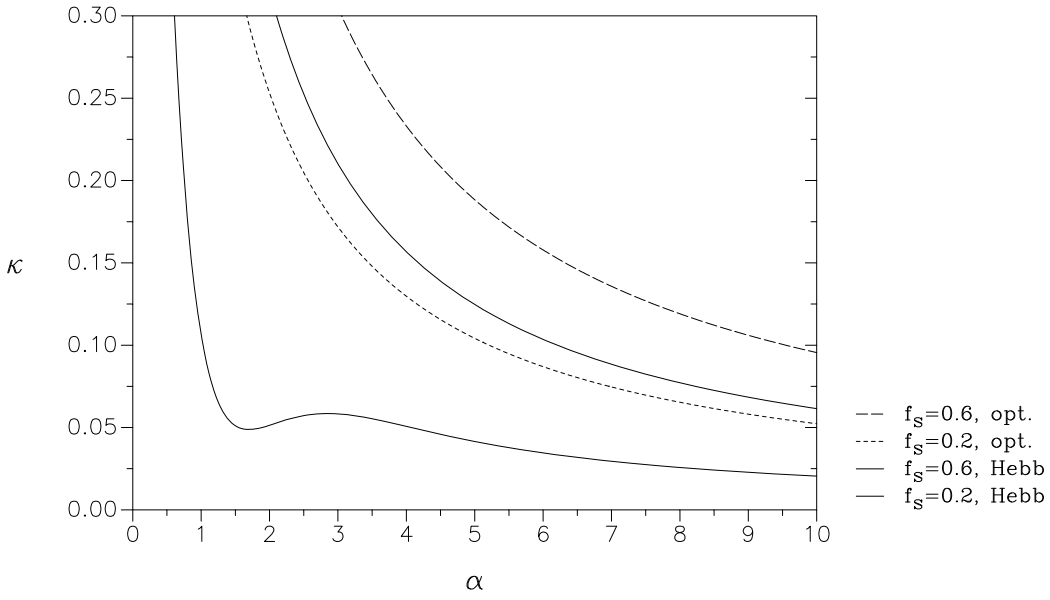


Abbildung 6.3: Vergleich der Stabilitäten $\kappa(\alpha)$ beim Verallgemeinerungsproblem für $f_l = 0.1$ und B_0=1. Die gestrichelte Linie steht für die optimale Verdünnung (opt.), die durchgezogene für den Quersummenalgorithmus (Hebb).

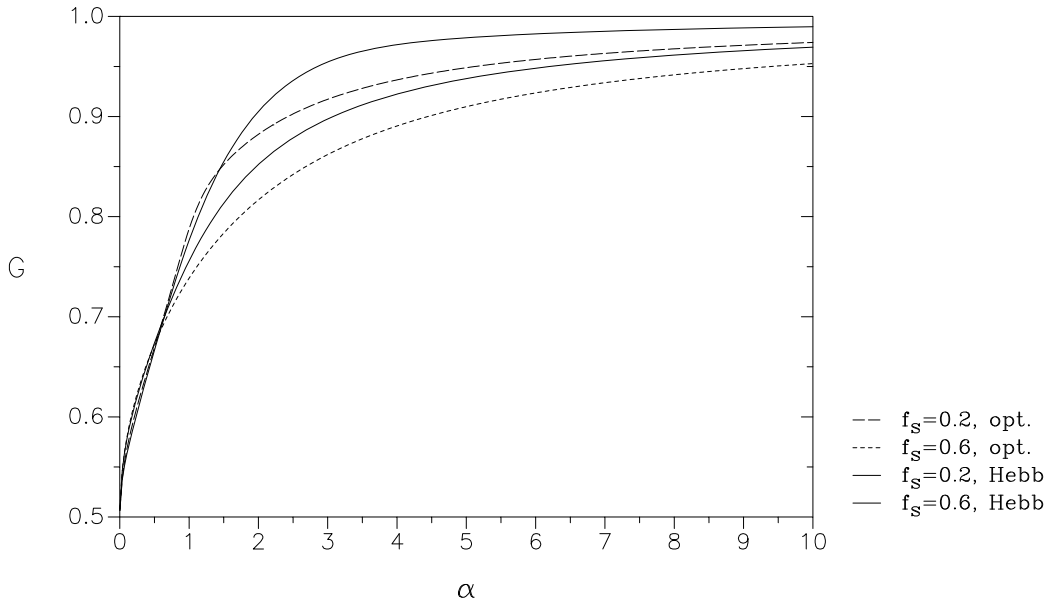


Abbildung 6.4: Vergleich der Verallgemeinerungsraten $G(\alpha)$ für den Parametersatz der Abb. 6.3.

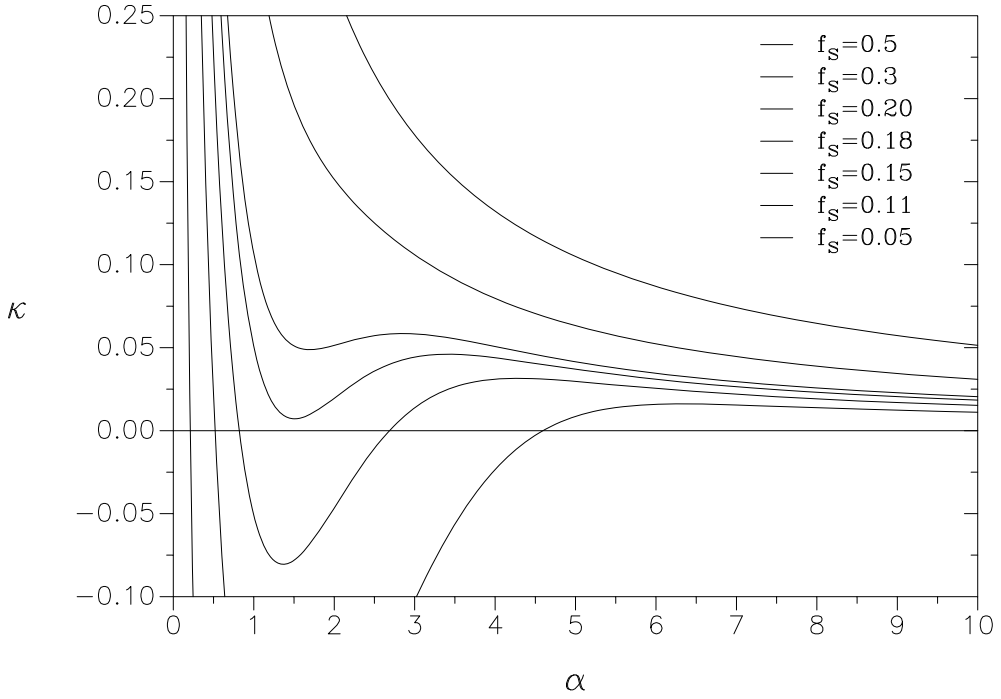


Abbildung 6.5: Die Stabilitäten $\kappa(\alpha)$ beim Quersummenalgorithmus für f_s–Werte in der Nähe des vorgegebenen Verdünnungsparameters $f_l = 0.1$ des Lehrers.

Für $f_s < f_l$ ist das vom Lehrer gestellte Problem von einem gewissen α an nicht mehr lernbar, da der Schüler nicht genug Plätze zur Verfügung hat, um alle Muster korrekt zu klassifizieren. Die zugehörige $\kappa(\alpha)$–Kurve fällt deshalb streng monoton und erreicht schon bei kleinen α negative κ–Werte. Ein ähnliches Verhalten liegt auch bei der optimalen Verdünnung vor [Mu92].

Im Falle $f_s \geq f_l$ liefert der optimale Verdünnungsalgorithmus Stabilitäten $\kappa \geq 0$. Der praktikable Quersummenalgorithmus hingegen hat bei mittleren Werten von α und f_s–Werten, die nahe bei f_l liegen ($f_s \geq f_l$), noch nicht genügend Plätze des Lehrers herausgefunden, um das vom Lehrer gestellte Problem zu lernen. Es treten negative Stabilitäten auf. Bei größeren α ist das Problem zwar schwerer zu lernen, doch holt der Schüler das Wissen über den Lehrer noch schneller nach. Bei größeren α hat der Quersummenalgorithmus also die relevanten Plätze des Lehrers so gut herausgefunden, daß wieder positive Stabilitäten erreicht werden. Ist f_s deutlich größer als f_l, so fällt κ wieder streng monoton in α. Der Schüler hat genug Plätze zur Verfügung, um bei mittleren Werten von α noch positive κ sicherzustellen. Außerdem findet er die relevanten Plätze des Lehrers schneller heraus. In Abbildung 6.6 wird dies nocheinmal anhand des Ergebnisses für f_c deutlich (siehe Abschnitt 6.3). Der Quersummenalgorithmus startet bei $\alpha = 0$ bei dem Wert

$$f_c = f_s \cdot f_l \tag{6.52}$$

des zufälligen Verdünnens.

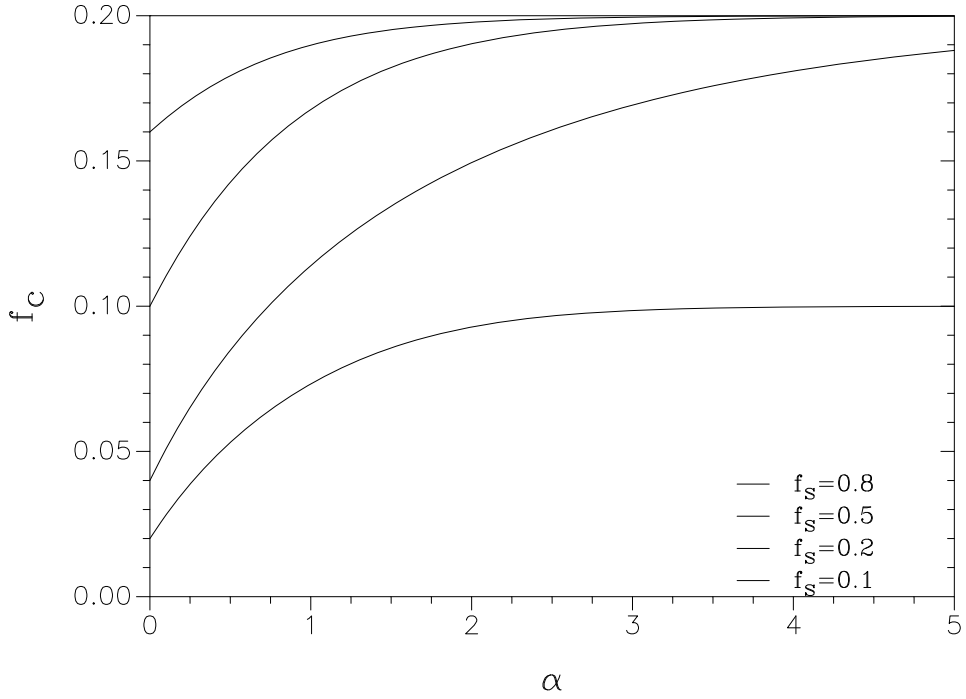


Abbildung 6.6: Die normierte Zahl f_c der übereinstimmenden Plätze von Lehrer und Schüler bei vorgegebenem $f_l = 0.2$. Die Kurven für $f_s = 0.1, 0.2, 0.5$ und 0.8 sind in aufsteigender Reihenfolge abgebildet.

Für $f_s \geq f_l$ gilt

$$f_c \to f_l \quad (\alpha \to \infty) \tag{6.53}$$

Im Falle $f_s < f_l$ strebt f_c ebenfalls sein Maximum an:

$$f_c \to f_s \quad (\alpha \to \infty) \tag{6.54}$$

Wie oben beschrieben, erreichen wir hohe Stabilitäten, wenn wir den Schüler möglichst schwach verdünnen. Hohe Werte von f_s wirken sich aber nachteilig auf die Verallgemeinerungsrate aus, wenn α hinreichend groß ist, siehe Abbildungen 6.7 und 6.8. Wir beobachten einen „overfitting“ – Effekt, der auch bei Funktionsappproximationen auftritt. Führt man die Approximation mit zu vielen freien Parametern durch, so kann man zwar die gegebenen Punkte des Graphen bequem lernen, doch besteht die Gefahr, daß die approximierende Funktion zwischen den Stützstellen zu sehr schwankt, daß sie also die Ausgangsfunktion schlecht verallgemeinert.

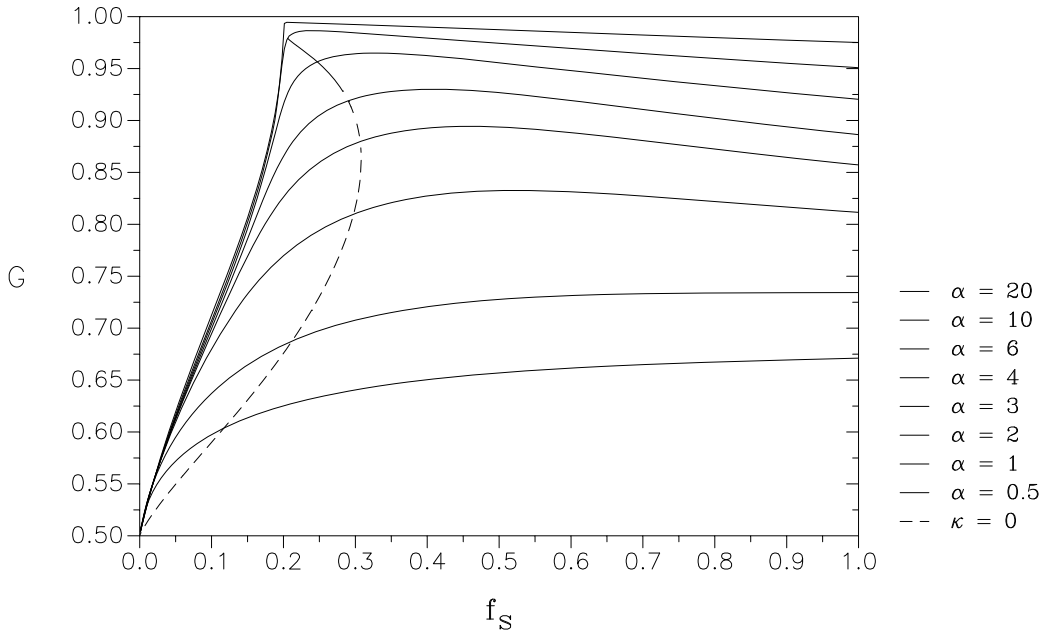


Abbildung 6.7: Der „overfitting“ – Effekt in der Verallgemeinerungsrate $G(f_s)$ des Quersummenalgorithmus bei festem $f_l = 0.2$. Links der gestrichelten $\kappa = 0$– Linie sind die Stabilitäten negativ, rechts davon positiv.

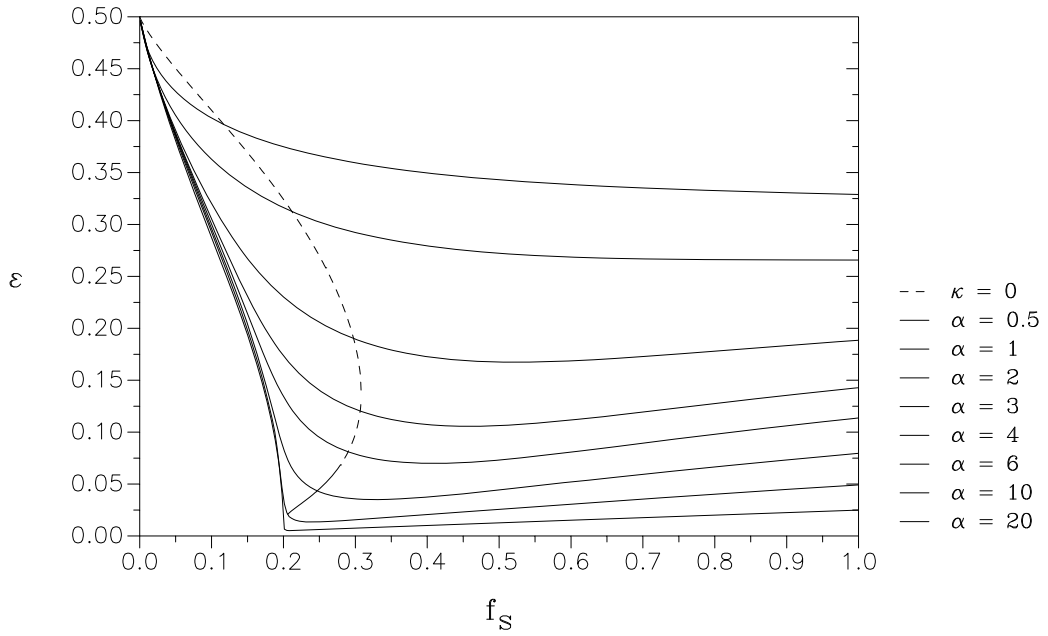


Abbildung 6.8: Der Verallgemeinerungfehler $\varepsilon(f_s) = 1 - G(f_s)$ des Quersummenalgorithmus. G entstammt der obigen Abbildung 6.7.

Bei unserem Verdünnungsproblem tritt das „overfitting“ erst bei größeren α auf. In diesem Fall agiert (für $f_s \geq f_l$) der Schüler auf den aktiven Plätzen des Lehrers. Je größer f_s ist, desto mehr überflüssige Neuronen des Schülers sind aktiv, so daß in der Überlappung

$$R = \frac{1}{N\sqrt{f_l f_s}} \sum_{j=1}^{Nf_l} c_j T_j B_j \tag{6.55}$$

im Zähler zu viele Terme fehlen.

Da $\underline{T}$ nur aus der analytischen Replika–Rechnung bekannt ist, vergegenwärtigen wir uns die obige Aussage nocheinmal an dem zu R analogen Ausdruck für die normierte Überlappung R_c der Verdünnungsvektoren des Lehrers und des Schülers:
Da $B_0 = 1$ ist, gilt

$$\underline{c}^{(B)} = \underline{B} = (1, \ldots, 1, 0, \ldots, 0)^T \tag{6.56}$$

Der Schüler hat den Verdünnungsvektor

$$\underline{c}^{(J)} = (c_1, \ldots, c_N)^T \tag{6.57}$$

mit

$$c_j^{(J)} = \theta\left(\left| \frac{1}{\sqrt{N\alpha}} \sum_{\mu=1}^{p} S^\mu \xi_j^\mu \right| - w \right) \quad \forall j \tag{6.58}$$

Dann ist R_c der cos des Winkels zwischen den beiden Verdünnungsvektoren

$$R_c = \cos \measuredangle \left(\underline{c}^{(B)}, \underline{c}^{(J)} \right) = \frac{\sum_{j=1}^{Nf_l} c_j}{N\sqrt{f_l f_s}} = \frac{f_c}{\sqrt{f_l f_s}} \tag{6.59}$$

Im Limes $\alpha \to \infty$ gilt nach Gl.(6.22)

$$R_c \to \frac{\min(f_l, f_s)}{\sqrt{f_l f_s}} = \min\left(\sqrt{\frac{f_l}{f_s}}, \sqrt{\frac{f_s}{f_l}} \right) =: R_c^\infty \tag{6.60}$$

Dieses exakte Ergebnis verdeutlicht den „overfitting“ – Effekt. Im Fall $f_s < f_l$ stellt der Grenzwert von R_c eine obere Schranke für die Überlappung R der Kopplungen dar:

Aus der Cauchy–Schwarzschen Ungleichung und der Normierungsbedingung der Kopplungen des Schülers

$$\sum_{j=1}^{N} c_j {T_j}^2 = N f_s$$

folgt zunächst allgemein [4]

$$R(\alpha) \leq \frac{\sqrt{\sum_{j=1}^{Nf_l} c_j {T_j}^2} \cdot \sqrt{\sum_{j=1}^{Nf_l} c_j {B_j}^2}}{\sqrt{N f_s} \cdot \sqrt{N f_l}} \leq \frac{\sqrt{\sum_{j=1}^{Nf_l} c_j {B_j}^2}}{\sqrt{N f_l}} \tag{6.61}$$

[4] Man beachte, daß c_j eine Komponente des Verdünnungsvektors des Schülers ist (Gl.(6.55)).

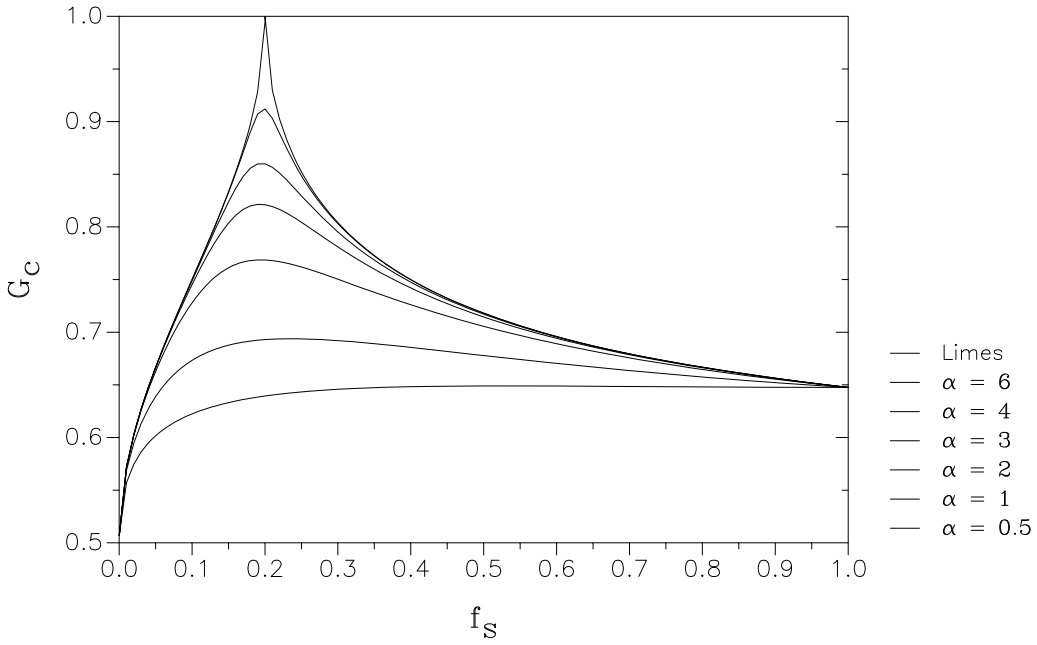


Abbildung 6.9: Der „overfitting“ – Effekt in der Verallgemeinerungsrate G_c der Verdünnungsvektoren (siehe Gl.(6.63) bei festem $f_l = 0.2$.

Im vorliegenden Fall $B_0 = 1$ und für $f_s < f_l$ gilt dann

$$R(\alpha) \leq \sqrt{\frac{f_c}{f_l}} \leq \sqrt{\frac{f_s}{f_l}} = R_c^{\infty} \tag{6.62}$$

In der Replika–Rechnung wird darüber hinaus stets $R(\alpha) < R_c(\alpha)$ im Falle $f_s < f_l$ beobachtet. Wir bilden also zum Vergleich die zu R_c gehörende „Verallgemeinerungsrate“

$$G_c(\alpha) := 1 - \frac{1}{\pi} \arccos R_c \tag{6.63}$$

und stellen sie in Abb. 6.9 für verschiedene α–Werte dar.

Für $f_s < f_l$ verhält sich G_c wie G aus Abbildung 6.7. Im Fall $f_s > f_l$ sieht man jedoch eine Diskrepanz der beiden Größen G_c und G. Die Abbildungen 6.7 und 6.9 unterscheiden sich qualitativ im Limes $\alpha \to \infty$. Die Kopplungen des Schülers werden bei hohen α auf den nicht relevanten Plätzen des Lehrers immer kleiner. Für $f_s > f_l$ gilt also

$$R \to 1 \quad (\alpha \to \infty) \tag{6.64}$$

im Gegensatz zu

$$R_c < 1 \quad (\alpha \to \infty) \tag{6.65}$$

Der „overfitting“ – Effekt wird auch beim optimal verdünnten Perzeptron beobachtet. Er ist hier stärker ausgeprägt als beim Quersummenalgorithmus. Außerdem setzt er früher ein, siehe Abbildung 6.10. Offenbar ist das optimal verdünnte Perzeptron für $f_s > f_l$ zu sehr damit beschäftigt, mit den

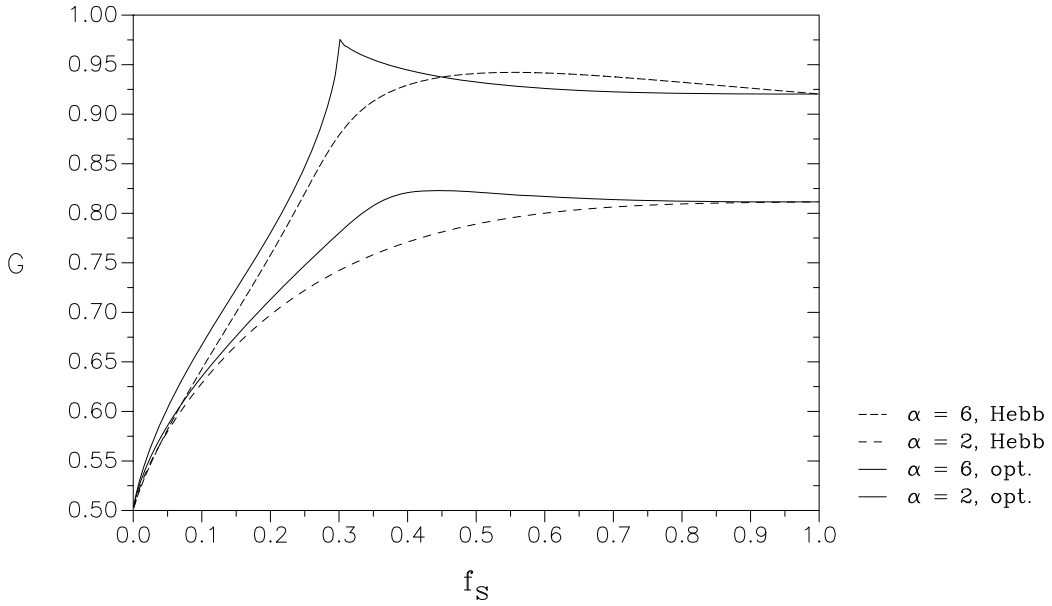


Abbildung 6.10: Der „overfitting“ – Effekt beim Quersummenalgorithmus im Vergleich mit der optimalen Verdünnung. Der Verdünnungsparameter des Lehrers ist $f_l = 0.3$. Wir beobachten einen starken „overfitting“ – Effekt bei $\alpha = 6$, wo die $G(f_s)$–Kurve des Quersummenalgorithmus (gestrichelte Linie) die Kurve der optimalen Verdünnung (durchgezogene Linie) schneidet.

überschüssigen Kopplungen die Stabilität zu optimieren. Es büßt dabei an Verallgemeinerungsfähigkeit ein. Am Punkt $f_s = f_l$ bleibt dem optimal verdünnten Perzeptron jedoch schon bei mittleren Werten von α keine andere Wahl, als zum Lehrer ähnliche Kopplungen zu wählen.

Die Stabilität des Lehrers[5] $\kappa_l = 0$ ist im Fall $f_s = f_l$ schon bei mittleren α kaum zu verbessern, so daß schon sehr früh $R \sim 1$ erreicht wird.

Für $f_s > f_l$ kann es hingegen bezüglich der Verallgemeinerungsrate günstiger sein, den Quersummenalgorithmus zu verwenden, da dieser einen schwächeren „overfitting“ – Effekt aufweist. Damit ist das obige überraschende Ergebnis aus Abbildung 6.4 geklärt.

Um einen verdünnten Lehrer gut zu verallgemeinern, können wir also den Quersummenalgorithmus benutzen. Von einem gewissen α an ist es wegen des „overfitting“ – Effektes günstig, auch den Schüler zu verdünnen. Man muß jedoch darauf achten, daß man im Bereich positiver Stabilitäten, also rechts der

[5] Es gilt $\kappa_l = 0$, weil die lokalen Felder

$$h_\mu = \frac{1}{\sqrt{N f_l}} \sum_{j=1}^{N f_l} B_j \xi_j^\mu$$

gaußverteilt sind und somit die lokalen Energien $S^\mu h_\mu = |h_\mu|$ beliebig nahe an den Nullpunkt heranreichen.

gestrichelten $\kappa = 0 -$ Linie in Abbildung 6.7 bleibt. Andernfalls ist die Aufgabe nicht mehr lernbar, und es treten Konvergenzprobleme bei den Perzeptron–Algorithmen auf. Auch der MinOver–Algorithmus [KrMe87], der bei negativen κ formell angewendet werden kann, ist noch nicht hinreichend in dieser Hinsicht überprüft.

Bei negativen κ treten auch Probleme in der analytischen Rechnung auf. Dort ist nämlich bereits beim unverdünnten Perzeptron die Replika–Symmetrie gebrochen [GaDe88]. Die berechneten Verallgemeinerungsraten müßten dann ebenfalls korrigiert werden. Für $\kappa \geq 0$ gehen wir beim Quersummenalgorithmus wieder davon aus, daß die replikasymmetrische Lösung stabil ist, da ein einfaches Perzeptron–Problem auf den verbleibenden Plätzen gelöst wurde. Im Fall der optimalen Verdünnung wurde die Verallgemeinerungsrechnung ebenfalls unter der Annahme der Replika–Symmetrie durchgeführt [Mu92]. Die gewonnenen Ergebnisse stimmen qualitativ mit denen des Quersummenalgorithmus überein. Wir erwarten jedoch im Fall der optimalen Verdünnung, daß eine RSB1–Rechnung analog Kapitel 3 auch im Fall $\kappa \geq 0$ eine quantitative Verbesserung bringt.

Kapitel 7

Zusammenfassung und Ausblick

7.1 Zusammenfassung

In der vorliegenden Arbeit wurde die lineare Verdünnung des optimalen Perzeptrons untersucht. Die Speicherkapazitäten und die Verallgemeinerungsraten linear verdünnter Perzeptrone wurden analytisch berechnet. Die wichtigsten Ergebnisse sind im folgenden aufgeführt:

1. Ein neues Rechenverfahren wurde vorgestellt, mit dem Speicherkapazitäten und Verallgemeinerungsraten von neuronalen Einschichtnetzwerken gewonnen werden können. Das Verfahren kommt ohne die Anwendung der Replika–Methode aus.

2. Für die Speicherkapazität α des optimal verdünnten Perzeptrons konnte eine obere Schranke gefunden werden. Damit wurde bewiesen, daß die replikasymmmetrische Rechnung in [Bo+90] zumindest im Bereich kleiner f falsch ist.

3. Die Näherung für die Speicherkapazität des optimal verdünnten Perzeptrons wurde in Replika–Symmetriebrechung erster Stufe berechnet. Sie liegt unterhalb der oberen Schranke. Die Speicherkapazität pro besetztem Platz, die effektive Speicherkapazität

$$\alpha_{eff} = \frac{p}{Nf} = \frac{\alpha}{f} \tag{7.1}$$

genannt wird, wird größer als 2 und divergiert für $f \to 0$ in diesem Limes

$$\alpha_{eff} = -\frac{1}{\ln 2} \ln f \tag{7.2}$$

4. Mit dem Quersummenalgorithmus wurde ein einfacher, praktikabler Verdünnungsalgorithmus vorgestellt. In einer replikasymmetrischen Gardner – Rechnung wurde eine effektive Speicherkapazität

$$\alpha_{eff} > 2 \tag{7.3}$$

für den Algorithmus berechnet. Die Rechnung wurde mit Hilfe von Computersimulationen überprüft. In der Diskussion wurde geklärt, daß das Ergebnis $\alpha_{eff} > 2$ nicht gegen den Satz von Cover verstößt. Das α_{eff} des Quersummenalgorithmus ist kleiner als das der optimalen Verdünnung. Im Limes $f \to 0$ divergiert α_{eff} schwächer als im optimalen Fall.

5. Um α_{eff} weiter zu verbessern, wurde das Schneideverfahren untersucht. Die analytische Rechnung läßt sich für die Einschritt – Version des Schneideverfahrens durchführen. Es handelt sich um eine doppelte, replikasymmetrische Gardner–Rechnung, die zum Quersummenalgorithmus analoge Ergebnisse liefert. Effektive Speicherkapazitäten α_{eff}, die weit größer als 2 sind, wurden auch in der zugehörigen Computersimulation beobachtet. Es wurde gezeigt, daß das Nachlernen des Perzeptrons optimaler Stabilität im zweiten Schritt besonders bei kleinen Werten des Verdünnungsparameters f und bei großen Speicherkapazitäten α nötig ist.

   Die Mehrschritt – Version des Schneideverfahrens konnte nur mit Hilfe von Computersimulationen untersucht werden [Gc92]. Das Mehrschrittverfahren liefert Speicherkapazitäten in der Nähe der RSB1–Näherung für den optimalen Fall. Eine weitere Ähnlichkeit zur RSB1–Näherung zeigt sich in der Wahrscheinlichkeitsverteilung der Kopplungen nach dem letzten Lernschritt. In beiden Fällen tritt eine Höckerstruktur auf, in der kleine Beträge der Kopplungen unterdrückt sind. Da das Mehrschrittverfahren viel Rechenzeit braucht, ist für praktische Anwendungen ein Ein– oder Zweischrittverfahren vorzuziehen.

6. Die Verallgemeinerungsrate G des Quersummenalgorithmus wurde in einer replikasymmetrischen Gardner–Rechnung ermittelt. Die Ergebnisse stimmen qualitativ mit dem G des optimal verdünnten Perzeptrons [Mu92] überein. Es zeigt sich, daß der Quersummenalgorithmus dazu benutzt werden kann, die relevanten Plätze eines verdünnten Lehrers herauszufinden. Man muß dazu den Schüler soweit ausdünnen, daß gerade noch positive Stabilitäten erreicht werden. Verallgemeinerungsrate und Stabilität sind dabei Kontrahenten: Überschüssige Neuronen werden im Fall $f_s > f_l$ fast ausschließlich zur Optimierung der Stabilität benutzt, so daß ein „overfitting“ – Effekt auftritt, der sich ungünstig auf die Verallgemeinerungsrate auswirkt.

7.2 Ausblick

7.2.1 Ergänzungen technischer Natur

Einige der in der Arbeit vorgestellten Rechnungen können mit anderen Methoden nachgerechnet werden bzw. mit weiteren Untersuchungen abgerundet werden. Dazu mache ich folgende Vorschläge

1. Das in Kapitel 2 vorgestellte Verfahren kann auf den Quersummenalgorithmus und das Einschrittverfahren angewendet werden. Dabei ist auch

eine Anwendung auf die zugehörigen Verallgemeinerungsrechnungen möglich. Die in diesen Rechnungen auftretende Korrelationsmatrix enthält dann die Verdünnungsvektoren $\underline{c}$. Die Determinante der Korrelationsmatrix sollte[1] deshalb mit Grassmann–Variablen dargestellt werden [FuSh90]. Im weiteren Verlauf der Rechnung entstehen Ordnungsparameter aus Grassmann – Variablen, die mathematisch schwer zu behandeln sind. Ein Vergleich mit den Rechnungen der vorliegenden Arbeit könnte bei der Interpretation der Ordnungsparameter hilfreich sein.

2. Die analytische Rechnung zum Einschrittschneideverfahren kann um endlich viele Schritte erweitert werden. Die Ordnungsparameter vorhergehender Schritte sind im aktuellen Schritt jeweils bekannt, so daß ein Iterationsverfahren möglich ist.

3. Das Schneideverfahren der Verdünnung kann so abgeändert werden, daß man alte, bereits entfernte Plätze wieder aktiviert. Beispielsweise kann man in einem Zweischrittverfahren kleine Beträge der Kopplungen des zweiten (und damit vorletzten) Lernschritts durch größere Beträge entfernter Kopplungen des ersten Lernschritts ersetzen und die zugehörigen Neuronen ebenfalls austauschen. Die Güte solcher Algorithmen wäre in einer Computersimulation herauszufinden.

7.2.2 Verdünnte Mehrschichtnetzwerke

Nachdem in der vorliegenden Arbeit verdünnte Einschichtnetzwerke behandelt wurden und sowohl bezüglich der Speicherkapazität als auch bezüglich der Verallgemeinerungsrate effektive Verdünnungsalgorithmen vorgestellt worden sind, stellt sich die Frage nach der Konzeption von verdünnten Mehrschichtnetzwerken. Dazu könnte man in folgenden Schritten vorgehen.

1. Als Einzelbausteine des zu konstruierenden verdünnten Mehrschichtnetzes verwenden wir grundsätzlich ein verdünntes, mit einem Schneideverfahren erzeugtes Perzeptron. Mit einer direkten Anwendung auf die bekannten unverdünnten Mehrschichtnetzwerke [MeNa89], [Bi91], [BiOp91], [Ru90], [He+91] lassen sich erste Ergebnisse gewinnen.

2. Im nächsten Schritt könnte man die Verdünnungen der Einzelbausteine variabel gestalten, um die Effizienz zu erhöhen.

3. Die Zwischenschichtneuronen des Mehrschichtnetzes könnten anschließend variabel gestaltet werden. Als Grundlage hierfür könnte man eine Indexmaschine [Sch+90] bzw. einen Tabellenautomaten [Mi92] benutzen. Mit den freien, „nicht überwachten“ Zwischenschichtneuronen lassen sich höhere Stabilitäten erzeugen. Jedoch ist in einem Zwischenschritt das Studium verdünnter nicht überwachter Einschichtnetze anzuraten. Hier ist

[1] Eine andere Möglichkeit der Darstellung der Determinante ist in [BrMo80] aufgezeigt. Es handelt sich um einen Replika–Limes $m \to -2$. Da die angesprochene Rechenmethode aber ohne Replika auskommen sollte, wäre die Anwendung der Darstellung von Bray und Moore hier inkonsequent.

zu beachten, daß aufgrund der freien Ausgaben bereits bei der analytischen Behandlung vollvernetzter Netze eine Replika–Symmetriebrechung auftritt [Mi92].

4. Da jedes nicht linear separable Klassifikationsproblem mit Mehrschichtnetzwerken erfaßt werden kann, ist in einer abschließenden Untersuchung der günstigste Verwendungszweck des jeweiligen Modells des verdünnten Mehrschichtnetzwerkes herauszufinden. Man orientiert sich dabei an zwei Zielen. Zum einen möchte man das Klassifikationsproblem mit einem möglichst stark verdünnten Netzwerk lösen, um den Speicherbedarf gering zu halten. Zum anderen sollen die relevanten Komponenten der Eingangsmuster erkannt werden. Bei der Bildverarbeitung ginge dies mit dem Herausfinden der für das Problem wichtigen Vorverarbeitungsfunktionen einher.

Anhang A

Häufig verwendete Formeln

A.1 δ–Funktion, θ–Funktion und Kronecker–δ

Die Fourierdarstellung der δ–Funktion ist

$$\delta(x) = \int_{-\infty}^{+\infty} \frac{dk}{2\pi}\, e^{ikx} \tag{A.1}$$

Eine Testfunktion g erfüllt die Gleichung

$$g(a) = \int_{-\infty}^{+\infty} g(x)\delta(x-a)dx \tag{A.2}$$

Diese Gleichung wird oft verwandt, um Entkopplungen vorzunehmen. Insbesondere schreibt man für Ausdrücke der Form $a = \frac{1}{N}\sum_{j=1}^{N} {J_j}^2$

$$g\left(\frac{1}{N}\sum_{j=1}^{N} {J_j}^2\right) = \int_{-\infty}^{+\infty} dx\; g(x)\cdot N \int_{-\infty}^{+\infty} \frac{dk}{2\pi}\; \exp\left(ik\left(Nx - \sum_{j=1}^{N} {J_j}^2\right)\right) \tag{A.3}$$

Dabei heißt k die zu x konjugierte Variable.

Die θ–Funktion ist definiert als

$$\theta(x) = \begin{cases} 1 & \text{für} \quad x \geq 0 \\ 0 & \text{für} \quad x < 0 \end{cases} \tag{A.4}$$

Es gilt

$$\theta(x) = \int_0^{\infty} d\lambda\; \delta(\lambda - x) = \int_0^{\infty} d\lambda \int_{-\infty}^{+\infty} \frac{dk}{2\pi} \exp\left(ik(\lambda - x)\right) \tag{A.5}$$

Die Fourierdarstellung des Kronecker-δ

$$\delta_{Kr}(x-y) = \begin{cases} 1 & \text{für} \quad x = y \\ 0 & \text{für} \quad x \neq y \end{cases} \tag{A.6}$$

ist analog zu Gl.(A.1)

$$\delta_{Kr}(x-y) = \int_{-i\pi}^{i\pi} \frac{d\psi}{2\pi i}\, e^{\psi(x-y)} \tag{A.7}$$

A.2 Gauß–Integrale

Es werden zwei häufig verwendete Gauß–Integral–Formeln aufgeführt:

1.
$$\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}bz^2 + cz\right) \frac{dz}{\sqrt{2\pi}} = \frac{1}{\sqrt{b}} \exp\left(\frac{c^2}{2b}\right) , \tag{A.8}$$

dabei muß $b > 0$ sein. c kann komplex sein. Man nennt die Formel dann auch oft „Hubbard–Stratonovich–Identität“ [Hu59]. Sind b und c rein imaginär, so handelt es sich um das sogenannte „Fresnel–Integral“. Die Konvergenz ist dann ebenfalls gesichert.

2. Sei $\mathbf{C}$ eine reelle symmetrische positiv definite Matrix und $\vec{J}$ irgendein Vektor, der auch komplexe Komponenten haben kann. Dann gilt (siehe [NeOr88]):

$$\left(\prod_{\nu=1}^{p} \int_{-\infty}^{+\infty} \frac{dx_\nu}{\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\sum_{\mu,\nu} x_\mu C_{\mu\nu} x_\nu + \sum_{\nu=1}^{p} x_\nu J_\nu\right)$$
$$= (\det \mathbf{C})^{-\frac{1}{2}} \exp\left(\frac{1}{2}\sum_{\mu,\nu} J_\mu (C_{\mu\nu})^{-1} J_\nu\right) .$$

In Vektorschreibweise lautet die Formel

$$\left(\prod_{\nu=1}^{p} \int_{-\infty}^{+\infty} \frac{dx_\nu}{\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\vec{x}^T \mathbf{C}\vec{x} + \vec{x}^T\vec{J}\right)$$
$$= (\det \mathbf{C})^{-\frac{1}{2}} \exp\left(\frac{1}{2}\vec{J}^T \mathbf{C}^{-1}\vec{J}\right) . \tag{A.9}$$

Generell schreiben wir

$$Dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \tag{A.10}$$

A.3 Die Φ–Funktion

Die Definition der Φ–Funktion ist

$$\Phi(t) := \int_{-\infty}^{t} D\lambda \tag{A.11}$$

Dabei ist $D\lambda$ gemäß Gl.(A.10) erklärt.

Es gelten folgende Entwicklungen [AbSt65]

$$\Phi(t) = \frac{1}{2}\left(1 + \sqrt{\frac{2}{\pi}}\, t - \frac{1}{3\sqrt{2\pi}}\, t^3 + \mathcal{O}\left(t^5\right)\right) \quad (t \to 0) \tag{A.12}$$

$$\Phi(t) = 1 - \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}t^2}\left(1 - \frac{1}{t^2} + \mathcal{O}\left(\frac{1}{t^4}\right)\right) \quad (t \to \infty) \tag{A.13}$$

$$\Phi(t) = -\frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}t^2}\left(1 - \frac{1}{t^2} + \mathcal{O}\left(\frac{1}{t^4}\right)\right) \quad (t \to -\infty) \tag{A.14}$$

Die asymptotische Reihe der Φ–Funktion ist

$$\Phi(t) = 1 - \frac{1}{\sqrt{2\pi}t} \cdot e^{-\frac{1}{2}t^2} \cdot \sum_{m=0}^{\infty} (-1)^m \cdot \frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2m-1)}{t^{2m}} \tag{A.15}$$

Dies gilt im Limes $t \to \infty$.

Uneigentliche Integrale über die Φ–Funktion [GrRy65]:

$$\int_0^\infty Dx \ \Phi(-ax) = \frac{1}{2\pi} \arctan \frac{1}{a} \quad (\text{für } a > 0) \tag{A.16}$$

$$\int_0^\infty Dx \ x \cdot \Phi(ax) = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \cdot \left(1 + \frac{a}{\sqrt{1+a^2}}\right) \tag{A.17}$$

Bei der Berechnung der Verallgemeinerungsrate G des Perzeptrons tritt die folgende Formel auf [Sw91].

$$G = 2 \cdot \int_0^\infty DH \ \Phi\left(H \cdot \frac{R}{\sqrt{1-R^2}}\right) = 1 - \frac{1}{\pi} \arccos R \tag{A.18}$$

Ein uneigentliches Integral über die gesamte reelle Achse

$$\int_{-\infty}^{+\infty} Dz \ \Phi(az+b) = \Phi\left(\frac{b}{\sqrt{1+a^2}}\right) \tag{A.19}$$

A.4 Die allgemeine Gauß–Formel

Für reelle Argumente t, a, b, c, u, v, w und für $ca^2 < 1$ ist das folgende Gauß–Integral definiert.

$$\mathrm{gf}(t,a,b,c,u,v,w) := \int_{-t}^{\infty} Dy \exp\left(\frac{1}{2}c(ay+b)^2\right) \cdot \left(uy^2 + vy + w\right) \tag{A.20}$$

Dy ist dabei der Gauß–Faktor aus Gl.(A.10).

Mit Hilfe der Regel der partiellen Integration läßt sich die Funktion gf durch Φ–Funktionen ausdrücken. Man definiert zunächst die Hilfsvariablen d, α, β:

$$d = 1 - ca^2$$

Damit ist

$$\begin{aligned} \alpha &= t\sqrt{d} + \frac{abc}{\sqrt{d}} \\ \beta &= \frac{abc}{\sqrt{d}} \end{aligned}$$

Es gilt die Formel

$$\begin{aligned} &\mathrm{gf}(t,a,b,c,u,v,w) = \\ &\quad \frac{1}{\sqrt{d}} \exp\left(\frac{cb^2}{2d}\right) \Phi(\alpha) \cdot \\ &\quad \cdot \left[\frac{1+\beta^2}{d}u + \frac{\beta}{\sqrt{d}}v + w + \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\alpha^2\right)}{\Phi(\alpha)} \cdot \left(\frac{2\beta - \alpha}{d}u + \frac{1}{\sqrt{d}}v\right)\right] \end{aligned} \tag{A.21}$$

Diese Formel läßt sich sehr leicht als Unterprogramm–Routine programmieren. Die Routine wird numerisch besonders stabil, wenn man eine innere Routine für den Ausdruck [...] in der obigen Formel programmiert. Der Grund dafür ist, daß sich der Fall $\alpha \to -\infty$ dann mit der asymptotischen Entwicklung (A.14) erfassen läßt.

Anhang B

Eine Formel zur Replika–Symmetriebrechung erster Stufe

Die in Abschnitt 3.4.2 verwendete Formel (3.43) wird hergeleitet, indem man von der Parisi–Parametrisierung der Sattelpunktmatrix $Q_{\rho\sigma}$ ausgeht (siehe Gl.(3.38)). Dabei setzen wir allgemein $Q_{\rho\rho} = Q$ auf der Hauptdiagonale.

Formel zur RSB1:
Sei $\mathrm{Tr}_{\{S^\rho\}}$ eine Spur über Größen S^ρ. Sie faktorisiere gemäß

$$\mathrm{Tr}_{\{S^\rho\}} = \prod_{\rho=1}^{n} \mathrm{Tr}_{S^\rho}$$

Dann gilt in RSB1:

$$\begin{aligned} A &= \mathrm{Tr}_{\{S^\rho\}} \exp\left(\gamma \sum_{\rho\leq\sigma} Q_{\rho\sigma} S^\rho S^\sigma\right) && \text{(B.1)} \\ &= 1 + n \cdot \int_{-\infty}^{+\infty} Dz \cdot \frac{1}{m} \cdot \\ &\quad \ln\left[\int_{-\infty}^{+\infty} Dy \cdot \left(\mathrm{Tr}_S \exp\left(S\left(y\sqrt{\gamma(q_1-q_0)} + z\sqrt{\gamma q_0}\right) + \gamma\left(Q - \frac{1}{2}q_1\right)S^2\right)\right)^m\right] + \\ &\quad + \mathcal{O}\left(n^2\right) && \text{(B.2)} \end{aligned}$$

Wir rechnen dies nach, indem wir die Replika im Exponenten aufteilen

$$\begin{aligned} A &= \mathrm{Tr}_{\{S^\rho\}} \exp\left(\frac{\gamma}{2}\sum_{k=1}^{n/m} \sum_{\rho(k),\sigma(k)} q_1 S^\rho S^\sigma + \frac{\gamma}{2}{\sum_{\rho,\sigma}}' q_0 S^\rho S^\sigma + \right. \\ &\quad \left. + \gamma\left(Q - \frac{1}{2}q_1\right)\sum_{\rho=1}^{n}(S^\rho)^2\right) && \text{(B.3)} \end{aligned}$$

Dabei numeriert der Index k alle Blöcke. Die zugehörigen Indizes $\rho(k), \sigma(k)$ numerieren die m Replika innerhalb der Blöcke. Die Summe $\sum_{\rho,\sigma}{}'$ hingegen geht über alle ρ, σ, wobei ρ und σ nicht dem gleichen Block angehören. Man addiert im Exponenten

$$0 = \frac{\gamma}{2} \sum_{k=1}^{n/m} \sum_{\rho(k),\sigma(k)} q_0 S^\rho S^\sigma - \frac{\gamma}{2} \sum_{k=1}^{n/m} \sum_{\rho(k),\sigma(k)} q_0 S^\rho S^\sigma$$

Damit gilt

$$\begin{aligned} A &= \mathrm{Tr}_{\{S^\rho\}} \exp\left(\frac{\gamma}{2}(q_1 - q_0) \sum_{k=1}^{n/m} \sum_{\rho(k),\sigma(k)} S^\rho S^\sigma + \frac{\gamma}{2} q_0 \sum_{\rho,\sigma} S^\rho S^\sigma + \right. \\ &\quad \left. + \gamma \left(Q - \frac{1}{2} q_1 \right) \sum_{\rho=1}^{n} (S^\rho)^2 \right) \end{aligned} \tag{B.4}$$

Die quadratischen Terme im Exponenten werden mit Hilfe der Hubbard–Stratonovich–Identität (A.8) entkoppelt, und wir erhalten

$$\exp\left(\frac{\gamma}{2} q_0 \left(\sum_{\rho=1}^{n} S^\rho \right)^2 \right) = \int_{-\infty}^{+\infty} Dz \exp\left(\sqrt{\gamma q_0} z \sum_{\rho=1}^{n} S^\rho \right) \tag{B.5}$$

Dabei gilt einfach

$$\sum_{\rho=1}^{n} S^\rho = \sum_{k=1}^{n/m} \sum_{\rho=m(k-1)+1}^{km} S^\rho \tag{B.6}$$

Innerhalb der quadratischen Blöcke liefert die Entkopplung

$$\begin{aligned} &\exp\left(\frac{\gamma}{2}(q_1 - q_0) \sum_{k=1}^{n/m} \left(\sum_{\rho=(k-1)m+1}^{mk} S^\rho \right)^2 \right) = \\ &\left(\prod_{k=1}^{n/m} \int_{-\infty}^{+\infty} Dy_k \right) \exp\left(\sqrt{\gamma(q_1 - q_0)} \sum_{k=1}^{n/m} y_k \sum_{\rho=(k-1)m+1}^{mk} S^\rho \right) \end{aligned} \tag{B.7}$$

Da die Spur über die S^ρ faktorisiert, erhalten wir

$$\begin{aligned} A = & \\ &\int_{-\infty}^{+\infty} Dz \left[\int_{-\infty}^{+\infty} Dy \cdot \left(\mathrm{Tr}_S \exp\left(S\left(y\sqrt{\gamma(q_1 - q_0)} + z\sqrt{\gamma q_0} \right) + \gamma\left(Q - \frac{1}{2} q_1 \right) S^2 \right) \right)^m \right]^{\frac{n}{m}} \end{aligned} \tag{B.8}$$

Die Entwicklung nach kleinen n liefert schließlich das obige Ergebnis (B.2).

Anhang C

Die Sattelpunktgleichungen und die Kopplungsverteilung zur RSB1 des optimal verdünnten Perzeptrons

C.1 Die Durchführung des Limes $q_1 \to 1$ in den Sattelpunktgleichungen

Bei der Herleitung der Sattelpunktgleichungen zur Entropie s aus Gleichung (3.51) entstehen zwei Arten von Gleichungen, deren Asymptotik ähnlich zu behandeln ist. Es sind dies die Sattelpunktgleichungen bezüglich q_0, q_1 auf der einen und bezüglich s_0, s_1, ψ auf der anderen Seite. Die Sattelpunktgleichung für m hingegen ist eine Mischung.

In die Ableitungen nach q_0 und q_1 gehen das folgende Integral und seine Ableitungen ein.

$$I_1(z) = \int_{-\infty}^{+\infty} Dy \left(\Phi\left(-\frac{\kappa + y\sqrt{q_1 - q_0} + z\sqrt{q_0}}{\sqrt{1-q_1}} \right) \right)^m \tag{C.1}$$

Mit dem in Abschnitt 3.4.3 angegebenen Skalierungsansatz für $q_1 \to 1$ erhält man unter Verwendung von Formel (A.14)

$$\left(\Phi\left(-\frac{\kappa + y\sqrt{q_1 - q_0} + z\sqrt{q_0}}{\sqrt{1-q_1}} \right) \right)^m \to \tag{C.2}$$
$$\theta(y+g) \exp\left(-\frac{1}{2} c \Big(\kappa + y\sqrt{1-q_0} + z\sqrt{q_0} \Big)^2 \right) + \theta(-g-y) \quad (q_1 \to 1)$$

Dabei habe ich die Abkürzung

$$g = \frac{\kappa + z\sqrt{q_0}}{\sqrt{1-q_0}} \tag{C.3}$$

eingeführt.

In den Ableitungen des ursprünglichen Ausdrucks [1] für I_1 nach q_0 und q_1 entsteht unter Beachtung der Formel (A.14) der folgende Ausdruck

$$(\Phi(\ldots))^m \cdot \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\ldots)^2\right)}{\Phi(\ldots)} \rightarrow \tag{C.4}$$
$$\theta(y+g) \cdot \exp\left(-\frac{1}{2}c\Big(\kappa + y\sqrt{1-q_0} + z\sqrt{q_0}\Big)^2\right) \cdot$$
$$\cdot\sqrt{\frac{c}{m}}\Big(\kappa + y\sqrt{1-q_0} + z\sqrt{q_0}\Big) \quad (q_1 \rightarrow 1)$$

Dabei ist

$$(\ldots) = \left(-\frac{\kappa + y\sqrt{q_1-q_0} + z\sqrt{q_0}}{\sqrt{1-q_1}}\right)$$

In die Ableitungen der Entropie nach den Hilfsvariablen s_0, s_1 und ψ geht das folgende Integral I_2 ein

$$I_2(z) = \int_{-\infty}^{+\infty} Dy \ \left(1 + \exp\left(\frac{1}{2} \cdot (y\sqrt{s_1-s_0} + z\sqrt{s_0})^2 + \frac{\eta}{2}\right)\right)^m \tag{C.5}$$

Hier ist im Limes $q_1 \rightarrow 1$ eine etwas kompliziertere Fallunterscheidung notwendig. Führt man

$$g_\pm = \frac{\pm\sqrt{-h} - z\sqrt{t_0}}{\sqrt{t_1-t_0}}, \quad g_+ > g_- \tag{C.6}$$

ein, so gilt für den Integranden in I_2 im Limes $q_1 \rightarrow 1$

$$(1+\exp(\ldots))^m \rightarrow$$
$$\exp\left(\frac{1}{2}\Big(y\sqrt{t_1-t_0} + z\sqrt{t_0}\Big)^2 + \frac{h}{2}\right) \cdot (\theta(y-g_+) + \theta(g_- - y)) +$$
$$+1 \cdot \theta(g_+ - y) \cdot \theta(y - g_-) \tag{C.7}$$

Leitet man den ursprünglichen Ausdrucks I_2 nach s_0, s_1, ψ ab, so entsteht der asymptotische Ausdruck

$$(1+\exp(\ldots))^m \cdot \frac{\exp(\ldots)}{1+\exp(\ldots)} \rightarrow$$
$$\exp\left(\frac{1}{2}\Big(y\sqrt{t_1-t_0} + z\sqrt{t_0}\Big)^2 + \frac{h}{2}\right) \cdot (\theta(y-g_+) + \theta(g_- - y)) + 0 \tag{C.8}$$

In die Sattelpunktgleichung für m gehen ebenfalls die obigen asymptotischen Ausdrücke ein. Zusätzlich benötigt man noch zwei Ausdrücke. Mit Formel (A.14) erhält man im Limes $q_1 \rightarrow 1$, d.h. $m \rightarrow 0$

$$\ln\Phi\left(-\frac{\kappa + y\sqrt{q_1-q_0} + z\sqrt{q_0}}{\sqrt{1-q_1}}\right) \rightarrow -\frac{1}{2}\theta(y+g)\frac{c}{m} \cdot \Big(\kappa + y\sqrt{1-q_0} + z\sqrt{q_0}\Big)^2 \tag{C.9}$$

[1] Die hergeleiteten asymptotischen Ausdrücke werden nur in den Sattelpunktgleichungen verwendet. Die Ausdrücke werden nicht in die Entropie s vor dem Ableiten eingesetzt, da dies auf falsche Ergebnisse führen kann.

Der dazu analoge Ausdruck beim Integral I_2 ist

$$\ln\left(1+\exp\left(\frac{1}{2}\cdot\left(y\sqrt{s_1-s_0}+z\sqrt{s_0}\right)^2+\frac{\eta}{2}\right)\right)\rightarrow$$
$$\frac{1}{2m}\cdot\left(\left(y\sqrt{t_1-t_0}+z\sqrt{t_0}\right)^2+h\right)\cdot\left(\theta(y-g_+)+\theta(g_--y)\right) \quad \text{(C.10)}$$

Durch die θ–Funktionen in den obigen Ausdrücken werden die inneren y–Integrale in allen sechs Sattelpunktgleichungen zu Gauß–Integralen über quadratische Funktionen. Diese lassen sich mit der in Anhang A.4 angegebenen Formel in Φ–Funktionen überführen.

C.2 Das Endergebnis für die RSB1 der optimalen Verdünnung des Perzeptrons

Legt man den in den Gln.(3.54) und (3.55) angegebenen Skalierungsansatz zugrunde, so hat das Gleichungssystem (3.53) die Lösungen q_0, c und h. Die Sattelpunktgleichung für m liefert dabei das kritische α. Die Variablen t_0 und t_1 lassen sich durch die anderen vier Variablen ausdrücken. Im folgenden gebe ich eine Anleitung zur Berechnung des nichtlinearen Gleichungssystems in den vier Variablen q_0, c, h, α, die sich eng an ein Computerprogramm anlehnt.

Gegeben seien also q_0, c, h, α. Das oben angegebene Integral I_1 ist dann

$$I_1(z)=\mathrm{gf}\Big(g,\sqrt{1-q_0},\kappa+z\sqrt{q_0},-c,0,0,1\Big)+\Phi(-g) \quad \text{(C.11)}$$

Dabei bezeichnet gf das in Anhang A.4 eingeführte allgemeine Gauß–Integral, welches sich mit Φ–Funktionen berechnen läßt. Wie oben ist

$$g=\frac{\kappa+z\sqrt{q_0}}{\sqrt{1-q_0}} \quad \text{(C.12)}$$

Für I_2 gilt

$$I_2(z) = \exp\left(\frac{h}{2}\right)\Big(\mathrm{gf}\Big(-g_+,\sqrt{t_1-t_0},z\sqrt{t_0},1,0,0,1\Big)+$$
$$+\mathrm{gf}\Big(g_-,-\sqrt{t_1-t_0},z\sqrt{t_0},1,0,0,1\Big)\Big)+\Phi(g_+)-\Phi(g_-) \quad \text{(C.13)}$$

Wie oben ist

$$g_\pm=\frac{\pm\sqrt{-h}-z\sqrt{t_0}}{\sqrt{t_1-t_0}},\quad g_+>g_- \quad \text{(C.14)}$$

Des weiteren werden die Konstanten D_1 und D_2 eingeführt:

$$D_1 = \frac{\alpha}{2}c\int_{-\infty}^{+\infty}Dz\cdot\frac{1}{I_1}\cdot \quad \text{(C.15)}$$
$$\cdot\mathrm{gf}\left(g,\sqrt{1-q_0},\kappa+z\sqrt{q_0},-c,-1,\left(z\sqrt{\frac{1-q_0}{q_0}}-g\right),(\kappa+z\sqrt{q_0})\cdot\frac{z}{\sqrt{q_0}}\right)$$

$$D_2 = \frac{\alpha}{2}c^2 \int_{-\infty}^{+\infty} Dz \cdot \frac{1}{I_1} \cdot \tag{C.16}$$
$$\cdot \mathrm{gf}\Big(g, \sqrt{1-q_0}, \kappa + z\sqrt{q_0}, -c, 1-q_0, 2\sqrt{1-q_0}(\kappa + z\sqrt{q_0}), (\kappa + z\sqrt{q_0})^2\Big)$$

Aus D_1 und D_2 bilde ich

$$B = \frac{D_1}{D_2} \tag{C.17}$$

Dann sind t_0 und t_1 einfach gegeben durch

$$t_1 = \frac{1}{1 + \frac{1}{c} - q_0 B}, \quad t_0 = B t_1 \tag{C.18}$$

Vor der Formulierung der Sattelpunktgleichungen wird noch der konstante Term

$$\text{Term} = 1 - s_1 + (1-m)q_1 s_1 + m q_0 s_0 \tag{C.19}$$

in der Entropie in Gl.(3.51) betrachtet. Er läßt sich mit Hilfe der Sattelpunktgleichung zu q_0 vereinfachen. Man erhält

$$\text{Term} = m s_0 \cdot \frac{fm}{2D_1} = t_0 \cdot \frac{fm}{2D_1} = \mathcal{O}(m) \tag{C.20}$$

Damit lautet das nichtlineare Gleichungssystem:

1. $\frac{\partial s}{\partial s_0} = 0$:

$$0 = \frac{D_1}{t_0} \cdot q_0 + \frac{1}{2} \int_{-\infty}^{+\infty} Dz \; \frac{1}{I_2} \; \exp\left(\frac{h}{2}\right) \cdot \tag{C.21}$$
$$\cdot \left(\mathrm{gf}\left(-g_+, \sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, -1, z\left(\sqrt{\frac{t_1 - t_0}{t_0}} - \sqrt{\frac{t_0}{t_1 - t_0}} \right), z^2 \right) + \right.$$
$$\left. + \mathrm{gf}\left(g_-, -\sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, -1, -z\left(\sqrt{\frac{t_1 - t_0}{t_0}} - \sqrt{\frac{t_0}{t_1 - t_0}} \right), z^2 \right) \right)$$

2. $\frac{\partial s}{\partial s_1} = 0$:

$$0 = -\frac{D_1}{t_0}\left(1 + \frac{1}{c}\right) + \frac{1}{2} \int_{-\infty}^{+\infty} Dz \; \frac{1}{I_2} \; \exp\left(\frac{h}{2}\right) \cdot \tag{C.22}$$
$$\cdot \left(\mathrm{gf}\left(-g_+, \sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, 1, z\sqrt{\frac{t_0}{t_1 - t_0}}, 0 \right) + \right.$$
$$\left. + \mathrm{gf}\left(g_-, -\sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, 1, -z\sqrt{\frac{t_0}{t_1 - t_0}}, 0 \right) \right)$$

3. $\frac{\partial s}{\partial \eta} = 0$:

$$\begin{aligned} 0 \quad = \quad & -f + \int_{-\infty}^{+\infty} Dz \; \frac{1}{I_2} \; \exp\left(\frac{h}{2}\right) \cdot \\ & \cdot \Big(\mathrm{gf}\Big(-g_+, \sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, 0, 0, 1\Big) \; + \\ & + \mathrm{gf}\Big(g_-, -\sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, 0, 0, 1\Big)\Big) \end{aligned} \tag{C.23}$$

4. $\frac{\partial s}{\partial m} = 0$:

$$\begin{aligned} 0 = \; & \\ & D_1 q_0 - D_2\left(1 + \frac{1}{c}\right) - \int_{-\infty}^{+\infty} Dz(\alpha \ln I_1 + \ln I_2) + \\ & + \frac{1}{2} \int_{-\infty}^{+\infty} Dz \; \frac{1}{I_2} \; \exp\left(\frac{h}{2}\right) \cdot \\ & \cdot \Big(\mathrm{gf}\Big(-g_+, \sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, t_1 - t_0, 2z\sqrt{t_1 - t_0}\sqrt{t_0}, h + z^2 t_0\Big) \; + \\ & + \mathrm{gf}\Big(g_-, -\sqrt{t_1 - t_0}, z\sqrt{t_0}, 1, t_1 - t_0, -2z\sqrt{t_1 - t_0}\sqrt{t_0}, h + z^2 t_0\Big)\Big) \end{aligned} \tag{C.24}$$

Die vier Gleichungen wurden mit der Powell–Hybrid–Methode gelöst [Po70].

C.3 Die Wahrscheinlichkeitsdichte der Kopplungen in RSB1–Näherung

Für das in Kapitel 3 behandelte optimal verdünnte Perzeptron berechnen wir die Wahrscheinlichkeitsdichte der Kopplungen in der RSB1–Näherung.

Da $N(1-f)$ Kopplungen zu Null gesetzt werden, gilt für die Wahrscheinlichkeitsdichte einer Kopplung (ich wähle die N – te Kopplung)

$$W(J) = (1-f)\delta(J) + P_r(J) \tag{C.25}$$

$P_r(J)$ ist die Wahrscheinlichkeitsdichte für den Fall, daß die N – te Kopplung nicht weggeschnitten wird, der Verdünnungsvektor

$$\underline{c} = (c_1, \ldots, c_N)^T \tag{C.26}$$

also die Komponente

$$c_N = 1 \tag{C.27}$$

aufweist.

Dann lautet der Ansatz zur Berechnung von $P_r(J)$ mit dem Phasenraumvolumen V_{ges} aus Gl.(3.16)

$$\begin{aligned} P_r(J) = & \\ & \frac{1}{V_{ges}} \cdot \sum_{\{\underline{c}\}} \int_{-i\pi}^{i\pi} \frac{d\psi}{2\pi i} \exp\left(\psi\left(Nf - \sum_{j=1}^{N} c_j\right)\right) \cdot \\ & \cdot \frac{1}{C_{norm}} \cdot \left(\prod_{j=1}^{N} \int_{-\infty}^{+\infty} \frac{dT_j}{\sqrt{2\pi}}\right) \delta\left(\sum_{j=1}^{N} c_j {T_j}^2 - Nf\right) \exp\left(-\frac{1}{2}\sum_{j=1}^{N}(1-c_j){T_j}^2\right) \cdot \\ & \cdot \left(\prod_{\mu=1}^{p} \theta\left(\frac{1}{\sqrt{Nf}} \sum_{j=1}^{N} c_j T_j \sigma_j^\mu - \kappa\right)\right) \cdot \delta_{Kr}(c_N - 1)\, \delta(J - T_N) \end{aligned} \tag{C.28}$$

Man berechnet $P_r(J)$ im Limes $N \to \infty$ mit Hilfe einer Replika–Mittelung (siehe Abschnitt 5.2.1). Die Rechnung verläuft analog zur Rechnung in [Bo+90] (siehe auch [Gc92]). Analog zu Abschnitt 3.4.2 wird dann die Replika–Symmetriebrechung erster Stufe durchgeführt. Man erhält mit den Lösungsvariablen t_0, t_1, D_1 aus dem obigen Abschnitt C.2 das folgende Endergebis am kritischen Punkt $q_1 = 1$:

$$\begin{aligned} P_r(J) = & \\ & \int_{-\infty}^{+\infty} Dz \cdot \frac{1}{I_2} \cdot \sqrt{\frac{2D_1}{2\pi(t_1 - t_0)t_0 f}} \cdot \\ & \cdot \exp\left[-\frac{1}{2}J^2 \cdot \frac{2D_1}{t_0 f} \cdot \frac{1-(t_1-t_0)}{t_1 - t_0} + z\sqrt{\frac{2D_1}{f}} \cdot \frac{J}{t_1 - t_0} - \frac{1}{2}z^2 \cdot \frac{t_0}{t_1 - t_0} + \frac{h}{2}\right] \cdot \\ & \cdot (\theta(J - J_0) + \theta(-J_0 - J)) \end{aligned} \tag{C.29}$$

I_2 ist aus Gl.(C.13) bekannt. Die Schranke J_0 für die symmetrische Verteilung $P_r(J)$ ist durch die Lösungsvariablen gegeben:

$$J_0 = \sqrt{\frac{-h \cdot t_0 f}{2D_1}} \tag{C.30}$$

Da wegen Gl.(C.25) die Normierung

$$\int_{-\infty}^{+\infty} dJ \; P_r(J) = f \tag{C.31}$$

gilt, ist in Abbildung 5.7 die Dichte

$$P(J) = \frac{1}{f} \cdot P_r(J) \tag{C.32}$$

dargestellt.

Anhang D

Die Sattelpunktgleichungen zur Bestimmung der Verallgemeinerungsrate des Quersummenalgorithmus

In den Sattelpunktgleichungen (6.48) vereinfachen sich die z–Integrale mit Hilfe des Limes

$$L = \lim_{q\to 1} \frac{\Phi'\left(-\frac{\kappa-k-uR+z\sqrt{q-R^2}}{\sqrt{1-q}}\right)}{\Phi\left(-\frac{\kappa-k-uR+z\sqrt{q-R^2}}{\sqrt{1-q}}\right)}\sqrt{1-q} = \left(\kappa-k-uR+z\sqrt{1-R^2}\right)\cdot\theta(z-(-t)) \tag{D.1}$$

wobei die Abkürzung

$$t = \frac{\kappa-k-uR}{\sqrt{1-R^2}} \tag{D.2}$$

eingeführt wurde. Für die Limites der Hilfsvariablen (siehe Gln.(6.44) – (6.47)) führen wir die Abkürzungen

$$\hat{S} = \hat{R}(1-q), \quad \hat{G} = \hat{K}(1-q) \tag{D.3}$$

ein.

Wir führen dann die vereinfachte Gauß–Formel ein

$$\mathcal{I}(s,u,v,w) = \int_{-s}^{\infty} Dz\ (uz^2+vz+w) = \mathrm{gf}(s,0,0,0,u,v,w) \tag{D.4}$$

Dabei ist gf die allgemeine Gauß–Formel aus Anhang A.4.

Damit lauten die Sattelpunktgleichungen im Limes $q \to 1$

$$-\hat{S} = \sqrt{\frac{2}{\pi}}\hat{G} + 2\alpha\int_0^{\infty} Du\ \mathcal{I}\left(t,R,(Rt+u\sqrt{1-R^2}),ut\sqrt{1-R^2}\right) \tag{D.5}$$

$$-\hat{G} = 2\alpha\int_0^{\infty} Du\ \mathcal{I}\left(t,0,1,t\right)\sqrt{1-R^2} \tag{D.6}$$

$$
\begin{aligned}
-f_s &= \frac{1}{\alpha}\hat{G}^2 - \frac{\hat{S}^2}{f_s}\, I_1 - 2\hat{S}\hat{G}\frac{I_2}{f_s}\sqrt{\frac{f_l}{\alpha}} - \frac{f_l\hat{G}^2}{\alpha f_s}\, I_3 + \\
&\quad -(1-f_l)\frac{I_4}{\alpha f_s}\,\hat{G}^2 - 2\alpha(1-R^2)\int_0^\infty Du\, \mathcal{I}\left(t,1,2t,t^2\right)
\end{aligned}
\tag{D.7}
$$

Die Konstanten I_1 – I_4 sind bereits in den Gleichungen (6.40) – (6.43) eingeführt worden. Für den in Abschnitt 6.5 behandelten, einfachen Lehrer mit B_0=1 lauten sie

$$
\begin{aligned}
I_1 &= \Phi(w_1) + \Phi(-w_2) \\
I_2 &= \sqrt{\frac{2\alpha}{\pi f_l}}\,(\Phi(w_1) + \Phi(-w_2)) + \\
&\quad + \frac{1}{\sqrt{2\pi}}\left(\exp(-\frac{1}{2}w_2^2) - \exp(-\frac{1}{2}w_1^2)\right) \\
I_3 &= \left(1 + \frac{2\alpha}{\pi f_l}\right) I_1 + \sqrt{\frac{4\alpha}{\pi^2 f_l}}\left(\exp(-\frac{1}{2}w_2^2) - \exp(-\frac{1}{2}w_1^2)\right) + \\
&\quad + \frac{1}{\sqrt{2\pi}}\left(w_2\exp(-\frac{1}{2}w_2^2) - w_1\exp(-\frac{1}{2}w_1^2)\right) \\
I_4 &= 2\left(\Phi(-w) + \frac{w}{\sqrt{2\pi}}\exp(-\frac{1}{2}w^2)\right)
\end{aligned}
$$

wobei w_1 und w_2 durch die Schranke w gegeben sind

$$
\begin{aligned}
w_1 &= -w - \sqrt{\frac{2\alpha}{\pi f_l}} \\
w_2 &= w - \sqrt{\frac{2\alpha}{\pi f_l}}
\end{aligned}
$$

Literaturverzeichnis

[AbSt65] M. Abramovitz und I. Stegun: *Handbook of Mathematical Functions,* National Bureau of Standards, Washington 1965.

[Am+87] D.J. Amit, H. Gutfreund und H. Sompolinsky: *Statistical Mechanics of Neural Networks near Saturation,* Annals of Physics,**173** (1987), 30.

[Am89a] D.J. Amit: *Field Theory, the Renormalization Group, and Critical Phenomena,* McGraw–Hill, Singapur 1989

[Am89b] D.J. Amit: *Modelling brain function: the world of attractor neural networks,* Cambridge University Press, Cambridge 1989

[AnBi90] J.K. Anlauf und M. Biehl: *The AdaTron: An Adaptive Perceptron Algorithm,* Europhys.Lett.,**10** (1990), 687.

[AT78] J.R.L. de Almeida und D.J. Thouless: *Stability of the Sherrington–Kirkpatrick Solution of a Spin Glass Model,* J.Phys.,**A11** (1978), 983.

[Bi91] M. Biehl: *Lernverfahren für Neuronale Netzwerke mit vorwärtsgerichteter Informationsverarbeitung,* Dissertation, Justus–Liebig–Universität Gießen 1992

[BiOp91] M. Biehl und M. Opper: *Tilinglike learning in the parity machine,* Phys.Rev.,**A44** (1991), 6888.

[BiYo86] K. Binder und A.P. Young: *Spin glasses: Experimental facts, theoretical concepts and open questions,* Rev.Mod.Phys.,**58** (1986), 801.

[Bo+90] M. Bouten et al.: *Quenched versus Annealed Dilution in Neural Networks,* J.Phys.,**A23** (1990), 4643.

[Bo92] S. Bös: *Neuronale Netzwerke mit Mehrzustandspins,* Dissertation, Justus–Liebig–Universität Gießen 1992

[Br61] N.G. de Brujn: *Asymptotic Methods in Analysis,* North–Holland, Amsterdam 1961

[BrMo80] A.J. Bray und M.A. Moore: *Metastable states in spin glasses,* J.Phys.,**A22** (1980), L469.

[Br+92] N. Brunel, J.–P. Nadal und G. Toulouse: *Information capacity of a perceptron,* J.Phys.,**A25** (1992), 5017.

[Co65] T. Cover: *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,* IEEE Trans Electron. Comp.,**14** (1965), 326.

[Ct92] K. Sarnow: *Gemeinsam statt einsam (Neuronales Netz auf Transputerbasis),* c't,**9** (1991), 30.

[DiOp89] S. Diederich und M. Opper: *Replicators with Random Interactions– a Solvable Model,* Phys.Rev.(Rapid Comm.),**A39** (1989), 4333.

[Do+89] E. Domany, W. Kinzel und R. Meir: *Layered Neural Networks,* J.Phys.,**A22** (1989), 2081.

[En+92] A. Engel et al.: *Storage capacity and learning algorithms for two–layer neural networks,* Phys.Rev.,**A45** (1992), 7590.

[EnHo+92] A.C.D. van Enter, A. Hof und J. Miekisz: *Overlap distributions for deterministic systems with many pure states,* J.Phys.,**A25** (1992), L1133.

[EnHe84] A.C.D. van Enter und J.L. van Hemmen: *Statistical Mechanics Formalism for Spin–Glasses,* Phys.Rev.,**A29** (1984), 355.

[Fe68] W. Feller: *An Introduction to Probability Theory and Its Applications,* Wiley & Sons, New York 1968

[FI91] Deutsche Bundespost Telekom, Forschungsinstitut beim FTZ (Hrsg.): *Jahresbericht 1991,* Darmstadt 1992

[FiHe91] K.H. Fischer und J Hertz: *Spin Glasses,* Cambridge University Press, Cambridge 1991

[Fl87] R. Fletcher: *Practical Methods of Optimization,* Wiley & Sons, New York 1987

[FuSh90] T. Fukai und M. Shiino: *Large suppression of spurious states in neural networks of nonlinear analog neurons,* Phys.Rev.,**A42** (1990), 7459.

[Ga86] E. Gardner: *Structure of metastable states in the Hopfield model,* J.Phys.,**A19** (1986), L1047.

[Ga88] E. Gardner: *The Space of Interactions in Neural Network Models,* J.Phys.,**A21** (1988), 257.

[GaDe88] E. Gardner und B. Derrida: *Optimal storage properties of neural network models,* J.Phys.,**A21** (1988), 271.

[Gc92] R. Garcés: *Learning Algorithms for diluted Neuronal Networks,* Diplomarbeit, Justus–Liebig–Universität Gießen 1992

[Gc+92] R. Garcés, P. Kuhlmann und H. Eißfeller: *In search of an optimal dilution algorithm for feedforward networks,* J.Phys.,**A25** (1992), L1335.

[Gl63] R.J. Glauber: *Time–Dependent Statistics of the Ising Model,* J.Math.Phys.,**4** (1963), 294.

[GoId83] D. Goldfarb und A. Idnani: *A numerically stable dual method for solving strictly convex quadratic programs,* Mathematical Programming,**27** (1983), 1.

[GrRy65] I.S. Gradshteyn und I.M. Ryzhik: *Table of Integrals, Series, and Products,* New York 1965

[GyTi90] G. Györgi und N. Tishby in W.K. Theumann und R. Köberle (Hrsg.): *Statistical Theory of Learning a Rule,* World Scientific, Singapur 1990

[He49] D.O. Hebb: *The Organisation of Behavior,* Wiley, New York, 1949.

[He+91] J. Hertz, A. Krogh und R. Palmer: *Introduction to the Theory of Neural Computation,* Addison–Wesley, New York 1991

[HeKu91] J.L. van Hemmen und R. Kühn: *Collective Phenomena in Neural Networks in „Physics of Neural Networks", Hrsg. J.L. van Hemmen, E. Domany und K. Schulten,* Springer, Berlin 1991

[HeNi91] R. Hecht–Nielsen: *Neurocomputing,* Addison–Wesley, Reading 1991

[HePa79] J.L. van Hemmen und R.G. Palmer: *The replica method and a solvable spin glass model,* J.Phys.,**A12** (1979), 563.

[HM92] Deutsche Messe AG, Hannover: *Katalog zur Hannover–Messe 1992,* Hannover 1992

[Ho82] J.J. Hopfield: *Neural networks and physical systems with emergent collective computational abilities,* Proc. Natl. Acad. Sci. USA,**79** (1982), 2554.

[Hu59] J. Hubbard: *Calculation of Partition Functions,* Phys.Rev.Lett.,**3** (1959), 77.

[Huang] K. Huang: *Statistical Mechanics,* Wiley & Sons, New York 1963 und 1987

[IEEE92] IEEE (Hrsg.): *IJCNN–91–Seattle: International Joint Conference on Neural Networks (Cat. No. 91CH3049–4),* IEEE, New York 1991

[IMSL] IMSL, Inc.: *Manual zur Fortran–Bibliothek der mathematischen und statistischen Unterprogramme, Version 2.0,* IMSL, Inc., Houston 1991

[Is25] E. Ising: *Beitrag zur Theorie des Ferromagnetismus*,Z.Phys.,**31** (1925), 253.

[Ja89] S.A. Janowsky: *Pruning versus clipping in neural networks*, Phys.Rev.,**A39** (1989), 6600.

[KaSo87] I. Kanter und H. Sompolinsky: *Associative recall of memory without errors*, Phys.Rev.,**A35** (1987), 380.

[Ki90] W. Kinzel: *Statistical Mechanics of the Perceptron with Maximal Stability in „Statistical Mechanics of Neural Networks", Hrsg. Luis Garrido*, Springer, Berlin 1990

[KiOp89] W. Kinzel und M. Opper: *Dynamics of Learning in „Physics of Neural Networks", Hrsg. J.L. van Hemmen, E. Domany und K. Schulten*, Springer, Berlin 1991

[Ko83] I. Kondor: *Parisi's mean–field solution for spin glasses as an analytic continuation in the replica number*, J.Phys.,**A16** (1983), L127.

[Ko88] T. Kohonen: *Self–Organisation and Associative Memory*, Springer, Berlin, 1984 und 1988

[KrMe87] W. Krauth und M. Mézard: *Learning Algorithms with Optimal Stability in Neural Networks*, J.Phys.,**A20** (1987), L745.

[Ku90] P. Kuhlmann: *Zahl der metastabilen Zustände bei einem Neuronalen Netzwerk mit Projektorkopplungen*, Diplomarbeit, Justus–Liebig–Universität Gießen 1990

[Ku+92] P. Kuhlmann, R. Garcés und H. Eißfeller: *A dilution algorithm for neural networks*, J.Phys.,**A25** (1992), L593.

[KuMu93] P. Kuhlmann und K.–R. Müller: *On the Generalisation Ability of Diluted Perceptrons*, in Vorbereitung

[Me+87] M. Mézard, G. Parisi und M.A. Virasoro: *Spin Glass Theory and beyond*, World Scientific, Singapur, 1987.

[MeNa89] M. Mézard und J.–P. Nadal: *Learning in feedforward layered neural networks: the tiling algorithm*, J.Phys.,**A22** (1989), 2191.

[Mi92] A. Mietzner: *Lernen bei frei wählbaren Ausgaben*, Diplomarbeit, Justus–Liebig–Universität Gießen 1992

[MiDu89] G. J. Mitchison und R. Durbin: *Bounds on the Learning Capacity of Some Multi–Layer Networks*, Biol. Cybern.,**60** (1989), 345.

[MiPa69] M.L. Minsky und S.A. Papert: *Perceptrons*, MIT Press, Cambridge 1969 und 1988

[Mu92] K.–R. Müller: *Spärlich verbundene neuronale Netze und ihre Anwendung*, Dissertation, Universität Karlsruhe 1992

[Ne91] R. Nehl: *Lernen und Verallgemeinern in Neuronalen Netzen,* Diplomarbeit, Justus–Liebig–Universität Gießen 1991

[NeOr88] J.W. Negele und H. Orland: *Quantum Many–Particle–Systems,* Addison–Wesley, Redwood City, 1988.

[NR88] W.H. Press et al.: *Numerical Recipes in C,* Cambridge University Press, Cambridge 1988

[Op88] M. Opper: *Learning times of neural networks: Exact solution for a Perceptron algorithm,* Phys.Rev.,**A38** (1988), 3824.

[Op+90] M. Opper, W. Kinzel, J. Kleinz und R. Nehl: *On the ability of the optimal perceptron to generalise,* J.Phys.,**A23** (1990), L581.

[Pa80a] G. Parisi: *The order parameter for spin glasses: A function on the interval 0–1,* J.Phys.,**A13** (1980), 1101.

[Pa80b] G. Parisi: *A sequence of approximated solutions to the S–K model for spin glasses,* J.Phys.,**A13** (1980), L115.

[Pe84] P. Peretto: *Collective properties of neural networks: A statistical physics approach,* Biol. Cybern., **50** (1984), 51

[Pe+85] L. Personnaz, I. Guyon, G. Dreyfus: *Information storage and retrieval in spin–glass neural networks,* J. Phys. (Paris) **46** (1985), L359

[Pe+86] L. Personnaz, I. Guyon und G. Dreyfus: *Collective computational properties of neural networks: New learning mechanisms,* Phys.Rev.,**A34** (1986), 4217.

[Po70] M.J.D. Powell in „Numerical Methods for Nonlinear Algebraic Equations“, Hrsg.: P. Rabinowitz: *G,*ordon and Breach, London 1970

[Re+82] D.R. Reilly, L.N. Cooper und C. Elbaum: *A Neural Model for Category Learning,* Biol. Cybern.,**45** (1982), 35.

[Reif] F. Reif: *Fundamentals of statistical and thermal physics,* McGraw–Hill, Singapur 1965

[Ri+91] H. Ritter, T. Martinetz, K. Schulten: *Neuronale Netze,* Addison–Wesley, Bonn 1991

[Ro58] F. Rosenblatt: *The perceptron: a probabilistic model for information storage and organization in the brain,* Psych. Rev.,**65** (1958), 386.

[Ru86] D.E. Rumelhart und J.L. McClelland: *Parallel distributed processing,* Bradford Books, Cambridge und London, 1986.

[Ru90] P. Ruján: *Learning in Multilayer Networks: A Geometric Computational Approach in „Statistical Mechanics of Neural Networks“, Hrsg. Luis Garrido,* Springer, Berlin, 1990

[Ru91] P. Rujàn: *A Fast Method for Calculating the Perceptron with Maximal Stability,* Preprint, Oldenburg 1991

[Sch+90] H.J. Schmitz et al.: *Fast Recognition of Real Objects by an Optimized Hetero–Associative Neural Network,* J. Phys (Paris),**51** (1990), 167.

[SK75] D. Sherrington und S. Kirkpatrick: *Solvable Model of a spin glass,*Phys.Rev.Lett.,**35** (1975), 1792.

[SK78] D. Sherrington und S. Kirkpatrick: *Infinite–ranged models of spin–glasses,* Phys.Rev.,**B17** (1978), 4384.

[Sn92] A. Scharnagl: *Verallgemeinern bei einem nicht lernbaren Problem: Perzeptron lernt Komitee–Maschine,* Diplomarbeit, Justus–Liebig–Universität Gießen 1992

[So86] H. Sompolinsky: *Neural Networks with Nonlinear Synapses and Static Noise,* Phys.Rev.,**A34** (1986), 2571.

[Sw91] H. Schwarze: *Verallgemeinern in einem Mehrschichtnetzwerk,* Diplomarbeit, Justus–Liebig–Universität Gießen 1991

[Wa+92] T.L.H. Watkin, A. Rau, M. Biehl: *The statistical mechanics of learning a rule,* Preprint, Oxford 1992

[WaLe90] A. Waibel und K.–F. Lee (Eds.): *Readings in speech recognition,* Morgan Kaufmann Pub., San Mateo 1990

[Wen91] A. Wendemuth: *Lernen magnetisierter Muster in Neuronalen Netzwerken,* Diplomarbeit, Justus–Liebig–Universität Gießen 1991

[Wi63] R.O. Winder: *Bounds on Threshold Gate Realizability,* IEEE Trans. Electron. Comp.,**EC–12** (1963), 561.

[Wo91] K.Y.M. Wong: , Private Mitteilung, 1991

Danksagung

An dieser Stelle möchte ich mich bei Prof. Dr. W. Kinzel für die Bereitstellung des Themas und die Unterstützung bei der Durchführung bedanken. Herrn Priv. Doz. Dr. M. Opper danke ich für seine Diskussionsbereitschaft und seine Geduld.

Holger Eißfeller und Rodrigo Garcés danke ich herzlich für viele interessante Diskussionen. Die gemeinsam erlebte Arbeitsatmosphäre wird mir stets in guter Erinnerung bleiben. Ebenfalls danke ich Andrea Scharnagl, Andreas Mietzner und Michael Biehl für ihre sowohl mentale als auch tatkräftige Unterstützung.

Für die freundliche und hilfsbereite Aufnahme in der neuen Umgebung an der Universität Würzburg danke ich Herrn Priv. Doz. Dr. G. Reents und Michael Vogel. „Last but not least“ gilt mein Dank Frau C. Kneipp und Frau U. Eitelwein für ihre Unterstützung bei allen bürokratischen Problemen.

Peter Kuhlmann

6332 Ehringshausen

Hiermit versichere ich, die vorliegende Arbeit selbständig und unter ausschließlicher Verwendung der angeführten Hilfsmittel und Quellen angefertigt zu haben.

Gießen, im Mai 1993