Diplomarbeit

Modellierung adulter Neurogenese im Hippokampus

von

Malte J. Rasch

Diplomstudiengang Biophysik
Humboldt Universität zu Berlin

betreut von Dr. Laurenz Wiskott

29. September 2003

Inhaltsverzeichnis

Kapitel 1

Einleitung

Erfahrungen können sich in plastischen Modifikationen des zentralen Nervensystems manifestieren [17]. Dazu gehören Änderungen der Stärke synaptischer Verbindungen, Reorganisation der Konnektivität des neuronalen Netzes durch Axon- und Dendritenwachstum, sowie die Bildung neuer Neurone [17]. Neurogenese ist jedoch auf einige Bereiche des Gehirns beschränkt [4], u.a. entstehen in einer Substruktur des Hippokampus vieler Spezies, einschließlich Primaten [17], zeitlebens neue Neurone [3]. Verhaltensexperimente an Mäusen legen die Vermutung nahe, dass Neurogenese im Hippokampus mit Lernen und der Aufnahme von Erfahrungen im Zusammenhang steht [4, 5, 14, 22, 23]. Beispielsweise scheint ein Käfig, der Mäusen mehr Möglichkeiten zur Interaktion mit der Umwelt bietet als ein Vergleichskäfig, Neurogenese im Hippokampus zu stimulieren [23]. Die kognitive Funktion adulter Neurogenese im Hippokampus ist jedoch unbekannt [4, 17].

Ein intakter Hippokampus ist notwendig für die Aneignung von Gedächtnisinhalten [21, 30]. Er scheint dafür zuständig zu sein, die Integration multimodaler Information aus nahezu allen assoziativen Hirnarealen zu einem gemeinsamen Produkt neuronaler Aktivität zu vollziehen. Die Funktion einer solchen Kompression hochdimensionaler Information zu einem niedrigdimensionalen Abbild hat zur Hypothese geführt, dass das Entstehen neuer Neurone ein Phänomen der Adaptation dieser Kodierung an Veränderungen der Statistik neuronaler Aktivität ist [22, 24].

Da nämlich angenommen wird, dass der Hippokampus die komprimierte Information speichert [39, 42], und die Gedächtnisinhalte während des Erinnerns in die neocorticalen Areale zurück projiziert werden, trifft auf dieses System eine Form des Plastizitäts-Stabilitäts-Dilemmas zu [24]: Einerseits fordert die Effizienz der Kompression eine Anpassung an veränderliche Daten, andererseits muss die Dekodierung der Gedächtnisinhalte trotz Adaptation gewahrt werden [30].

Anhand eines mathematischen Modells wollen wir den Beitrag der Neurogenese zur Lösung des Dilemmas analysieren. Wir untersuchen deshalb den Adaptationsprozess eines Modellsystems des Hippokampus für Änderungen der (Statistik der) verarbeiteten Information. Dazu vergleichen wir verschiedene Weisen, wie die Kodierung sich verhalten könnte. Wir werden feststellen, dass in bestimmten Situationen die Bildung neuer Neurone zur Unterstützung der Adaptation vorteilhaft ist, während sie in anderen Situationen keinen Erfolg verspricht. Wir charakterisieren beide Fälle mathematisch exakt. Insbesondere zeigt sich, dass Neurogenese tatsächlich den Kode der Gedächtnisinhalte des Hippokampus stabilisieren kann. Darüberhinaus befürwortet das Szenario des "Sammelns von Erfahrungen" in unserem Modell eine Adaptationsstrategie mittels Neurogenese, was die erwähnten experimentellen Untersuchungen zu bestätigen scheinen.

1.1 Übersicht

Nachdem im nächsten Kapitel auf die biologischen Grundlagen des Hippokampus und der Neurogenese eingegangen wird, formulieren wir im dritten Kapitel das verwendete mathematische Modell und leiten einige Aussagen über das Modell her. In Kapitel 4 vergleichen wir verschiedene Adaptationsstrategien in einem numerischen "Experiment", das den Adaptationsprozess simuliert. Kapitel 5 beschäftigt sich mit der mathematischen Untersuchung der durch Neurogenese gestützten Adaptation. Abschließend werden wir die Ergebnisse im Kapitel 6 diskutieren.

Kapitel 2

Hippokampus

2.1 Kurzer Überblick über die Anatomie

Der Hippokampus ist eine Hirnstruktur des medialen Temporallappens [21]. Er ist histologisch in mehrere Substrukturen gegliedert (vergl. Abb. 2.1). Er beinhaltet den Gyrus dentatus (GD), die Zellfelder CA1-4 und weitere Strukturen (z.B. Subiculum, Presubiculum) [21]. Der entorhinale Cortex (EC) bildet die Schnittstelle zwischen dem Hippokampus und neocorticalen assoziativen Arealen [42]. Information aus dem Neocortex mündet in die Schichten I-III des EC [7]. Durch das Faserbündel Tractus perforans wird Schicht II des EC mit den granulären Zellen im Hilus des Gyrus dentatus, sowie mit den Pyramidenzellen in CA3 [7, 44, 21] verbunden; dies sind die vorherrschenden Zellformen der jeweiligen Substruktur. In einer Reihe vorwärtsgerichteter Fasern bewegt sich der Hauptinformationsfluss vom Gyrus dentatus durch die Moosfasern zu den Pyramidenzellen des CA3-Feldes, von dort über die Schaffer-Kollateralen zur CA1-Region und via Subiculum/Presubiculum zurück zum entorhinalen Cortex, jetzt endend in Schicht V und VI (siehe auch Abb. 2.1) [2, 7, 44, 38, 42].

Neben dem beschriebenen circulären Hauptinformationsfluss existieren Querverbindungen [42]. Aus der dritten und fünften Schicht des entorhinalen Cortex entspringen Fasern zum CA1-Zellfeld. Die Zuordnung der Schichten im EC variieren z.T. je nach Spezies (genaueres in [44]). Desweiteren gibt es komissurale Fasern, die die Seiten des bilateralen Hippokampus verknüpfen und Verbindungen zu sublimbischen Strukturen, welche die allgemeine Aktivität des Hippokampus regeln, jedoch nichts mit der verarbeiteten Information als solche zu tun haben scheinen [42]. Schließlich sind die rekurrenten Verbindungen innerhalb des CA3-Feldes hervorzuheben, denn durch sie besitzen die dortigen Pyramidenzellen eine Interkonnektivität von ca. 5% [42].

Der Hippokampus empfängt Information aus praktisch allen assoziativen Areal des Neocortex, z.B. inferior temporaler visueller Cortex, superior temporaler auditorischer Cortex und parietaler Cortex [37, 42]. Die multimodale Informationen erreicht via parahippokampalen und perirhinalen Cortex, sowie dem entorhinalen Cortex den Hippokampus [29, 37, 42]. Nach Verarbeitung kann Information den Hippokampus auf zwei

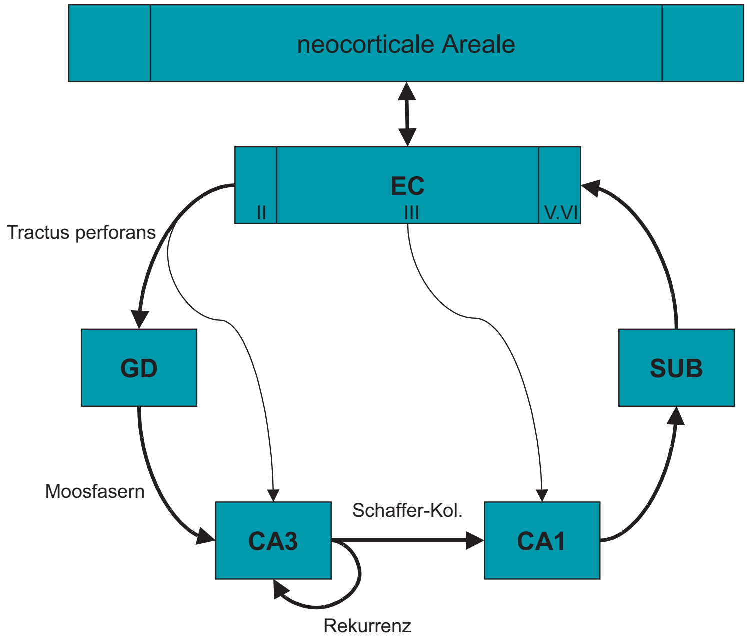

Abbildung 2.1: Schematische Darstellung der wesentlichen Projektionen zwischen den Substrukturen des Hippokampus (Gyrus dentatus (GD), CA3 und CA1-Region, Subiculum (SUB)) und corticalen Arealen (entorhinaler Cortex (EC) und neocorticale Areale). Die Schichten des EC sind bezeichnet. Fette Pfeile markieren den Hauptinformationsfluss.

Wegen verlassen. Einerseits via Fimbria und Fornix zum Thalamus anterior [37]. Andererseits projiziert Schicht V/VI des EC zurück in die erste und zweite Schicht der Cortizes, aus denen Information den Hippokampus erreicht [37].

Im Menschen stehen im GD ca. 11 ± 3 Millionen den 6 ± 1.5 und 3 ± 1 Millionen Neurone in CA1[1] bzw. CA2-4[2] gegenüber. Subiculum und Presubiculum umfassen etwa 5 ± 1 bzw. 10 ± 2.5 Millionen Zellen [15]. Die ca. 8 ± 1 Millionen Neurone des entorhinalen Cortex teilen sich in folgender Weise auf seine Schichten auf (in Millionen): 0.7 (II), 3.7 (III), 1.2 (V) und 2.6 (VI) [49]. Die Aktivitätsmuster des Hippokampus sind spärlich (0.4% gleichzeitig aktive Zellen in GD und 2.4% in CA1 und CA3), während EC weniger spärliche (7%) und auch weniger selektive Aktivität zeigt [6, 29, 42].

Im Allgemeinen gibt es eine relativ kleine Population von Interneuronen, die, entgegen dem Großteil der Neurone, inhibitorische Synapsen bilden [2, 11]. An Synapsen des Tractus perforans, der Moosfasern und der Schaffer-Kollateralen ist Langzeitpotenzierung (LTP) nachgewiesen worden. LTP bedeutet, dass die Effizienz der synaptischen Signalübertragung plastisch moduliert wird, und zwar erfolgt eine Stärkung bei gekoppelter Aktivität von post- und präsynaptischem Neuron; man spricht von Hebb'scher Lernregel der synaptischen Gewichte [21]. Die Anzahl und Art der ausgebildeten Syn-

[1] Andere Autoren erwähnen 16 statt 6 Millionen CA1-Zellen [49]. Da die kleiner Zahl an alt gestorbenen Menschen erfasst wurde, könnte die Diskrepanz auf den Befund altersbedingter Abnahme der CA1-Neurone zurückzuführen sein [15, 40].

[2] genauer: 2.25 Millionen in CA2-3. 1 Millionen Neurone in CA4.

apsen der beschriebenen Fasern ist sehr verschieden. Eine CA1-Pyramidenzellen erhält ca. $5 \cdot 10^5$ synaptische Eingängen in Ratten. Ein Drittel der Eingängen stammt vom Tractus perforans, die übrigen werden von den Schaffer-Kollateralen gebildet [42]. Auffallend ist die Spärlichkeit der Verbindung der Moosfasern auf Pyramidenzellen in CA3: In der Ratte münden höchstens 50 Moosfasern auf eine gegebene CA3-Zelle; in Anbetracht der Zellzahlenverhältnisse beider Regionen bildet eine granuläre Zelle des GD also durchschnittlich nur 15 CA3-Synapsen aus [42]. Moosfasersynapsen scheinen den Pyramidenzellen besonders starke excitatorische Signale zu vermitteln, münden aber auch auf inhibitorische Interneurone [43]. Desweiteren erhält eine CA3-Pyramidenzelle sehr viel mehr, jedoch schwächere Synapsen des Tractus perforans aus EC (ca. $4 * 10^3$ in Ratten) und Eingänge von rekurrenten Fasern ($1.2 * 10^4$ in Ratten). Welcher der drei Eingänge die mittlere Spikeerzeugung der CA3-Region am stärksten beeinflusst, ist unklar [43].

In der vorliegenden Arbeit werden wir den Schwerpunkt auf den Hauptinformationsfluss von EC über GD, CA3, CA1, und von hier zurück zu EC, legen und in der Modellierung von vielen biologischen Details abstrahieren.

2.2 Funktionelle Bedeutung des Hippokampus

Der Hippokampus ist für das deklarative Gedächtnis notwendig [21]. Deklarative Gedächtnisinhalte sind Formen von Erinnerung, die explizit abgerufen werden können; dazu zählt die Erinnerung an erlebte Episoden und semantisches Wissen [38]. Das explizite Gedächtnis steht im Gegensatz zu unbewussten, motorischen Fertigkeiten, dem impliziten Gedächtnis [21].

Studien an Patienten mit Hirnläsionen lassen vermuten, dass der Hippokampus für die Aneignung von Gedächtnisinhalten notwendig ist, jedoch möglicherweise nur als ihr intermediärer Speicher dient [30, 29]. Denn Patienten behalten nach der Schädigung des Hippokampus weit zurückliegende Erinnerungen, während kürzliche erworbene Erinnerungen verloren gehen (retrograde Amnesie), desweiteren ist der Bildung neuer Erinnerungen gestört (anterograde Amnesie) [30, 21]. Daher vermutet man, dass Erinnerungen aus dem Hippokampus als Zwischenspeicher in ein stabileres Langzeitgedächtnis anderer Hirnarealen transferiert werden (Konsolidierung) [1, 30]; dies könnte während des Schlafs geschehen [28]. Der Prozess der Konsolidierung dauert möglicherweise mehrere Jahre für bestimmte Gedächtnisformen in Menschen [30, 32].

Verhaltensexperimente an Mäusen und Ratten haben hippokampusabhängige Lernprobleme identifiziert (z.B. "Morris-Wasserlabyrinth" [31]). Insbesondere räumliche Aufgaben scheinen einen intakten Hippokampus zu benötigen, so dass der Hippokampus zwar von den Einen als eine Art räumliches Gedächtnis bezeichnet wird [8, 33, 35], andere jedoch allgemeiner die Fähigkeit zur Integration von Informationen als herausragende Eigenschaft des Hippokampus nennen [42]. Letzteres wird damit unterstützt, dass der Hippokampus am Endpunkt der Reizverarbeitung steht; die Eingangsschicht, der entorhinale Cortex, erhält Projektionen aus verschiedensten Hirnarealen. Diese Information wird durch eine Zone größter Konvergenz, der CA3-Region [40], geleitet und kann so zu einem zusammenfassenden Muster neuronaler Aktivität integriert werden. Diese integrative Rolle des Hippokampus könnte zur Erfassung einer gesamten Szenerie als solche notwendig sein und so die Grundlage zur Speicherung von Episoden oder räumlichen Begebenheiten darstellen ("Schnappschuss"-Gedächtnis) [38].

In theoretischen Arbeiten wird die CA3-Region wegen ihrer rekurrenten Konnektivität häufig als der eigentliche Informationsspeicher angesehen [26, 42]; CA3 ähnelt einem spärlich verbundenen Autoassoziativspeicher (Hopfield-Netz [18, 19]). Diese Region scheint daher verantwortlich für das Assoziieren gespeicherter Inhalte [25]. Dem vorgelagertem GD wird die Aufgabe zugesprochen die Redundanz neocorticaler Information zu vermindern [30] und sie durch Orthogonalisierung mittels einer spärliche Darstellung dem Speicher zuzuführen [13, 25, 42]. Eine derartige Kodierung des GD erscheint sowohl für das Konzept der Integration von Information, als auch aus Gründen der Speicherkapazität von CA3 notwendig [42]. Die Moosfasern sind aufgrund ihrer starken postsynaptischen Potenziale geeignet, Muster in die CA3-Region "einzuschreiben" [42]. Tatsächlich bestimmen sie in Phasen gewisser GD Aktivität den Einfluss auf die CA3-Region [43].

Für die CA1-Region ist vorgeschlagen worden, dass sie die Invertierung der durch den GD bewirkten Kodierung vollzieht [42, 29]. Es wurde jedoch auch ein Beitrag dieser Region zur zeitlichen Trennung von Information festgestellt [13, 25].

Obwohl es viele konkurrierende Theorien über die hauptsächliche Funktion des hippokampalen Systems gibt, werden wir im wesentlichen der Argumentation Treves und Rolls [42] folgen, und den Hippokampus vereinfacht als mittelfristigen Speicher ansehen, der "Schnappschüsse" aktueller Hirnaktivität aufnimmt. Der Speicher wird von einem Kodierer (GD) und einem Dekodierer (CA1) attestiert (genaueres im Abschnitt 3.1).

2.3 Adulte Neurogenese im Gyrus dentatus

Im Hippokampus vieler Spezies, darunter Primaten, werden zeitlebens neue Neurone gebildet [14, 17]. Die postnatale Entstehung neuer Neurone beschränkt sich auf Zellen des Gyrus dentatus [17]. Vorläuferzellen befinden sich in der subgranulären Zone des Gyrus dentatus; durch Proliferation gebildete Tochterzellen wandern von dort in die granuläre Zellschicht [3]. Reife Neurone erhalten synaptische Eingänge und senden, gleich den "alten" granulären Zellen, axonale Verbindungen zur CA3-Region aus [3, 17]. Sie scheinen nicht alte Neurone zu ersetzen [24], sondern sich in die Schicht der bestehender Neurone gezielt einzugliedern [17].

Die Vorläuferzellen weisen eine basale Teilungsaktivität auf [14, 23, 45], welche durch verschiedene neuroendokrine Substanzen positiv wie negativ reguliert wird [17]. Regulation der Proliferation könnte möglicherweise über den EC als Mediator indirekt durch affarente Hirnareale geschehen [17]. Es wird darüber hinaus eine Regulierung der

Überlebenswahrscheinlichkeit neu geborener Zellen nachgewiesen [45]. Dies ist möglich, da viele angehende Neurone vor ihrer Reife durch Apoptose sterben [3]. Man schätzt, dass das postnatale Wachstum die granulären Zellen des GD in Mäusen insgesamt um etwa 10% vermehrt [24].

In Experimenten an Mäusen wurde beobachtet, dass sich Umweltbedingungen und Verhaltensweisen auf die Neurogenese auswirkt. Es scheinen sowohl das freiwillige Benutzen von Laufrädern [45], das Leben in "reizreicher" Umgebung [23] (mit Stimuli angereicherter Käfig, wie etwa Laufrad, Röhren, Möglichkeit für soziale Kontakte, etc. [46]) als auch nachweislich hippokampusabhängige Experimente [14] die Neubildung von granulären Zellen im Gyrus dentatus zu stimulieren. An bestimmten frei lebenden Vögeln wurde eine saisonale Schwankung neuer Neurone im Hippokampus nachgewiesen, die auf die Speicherung von räumlicher Information (Futterplätze, Gefahrstellen, etc.) zurückgeführt wurden [5].

Erhöhte Bildung funktionell aktiver Neurone wird mit komplexeren Anforderungen an das Gedächtnis in Verbindung gebracht [17], obwohl es hier auch leicht abweichende Meinungen gibt [45]. Man vermutet, dass Neurogenese mit der gesammelten Erfahrung eines Tieres über seine Umwelt einhergeht und die Gedächtnisleistung besonders für räumliche Aufgaben steigert [22, 23].

2.3.1 Hypothetische Funktion

Über die ausgeübte Funktion der Neurogenese in der Informationsverarbeitung des Hippokampus ist wenig bekannt [4]. Das liegt vorallendingen daran, dass die zellulären Mechanismen der kognitiven Funktion des Hippokampus selbst noch schlecht verstanden sind [22]. Stellt man sich jedoch auf den Standpunkt, dass der Hippokampus der Integration von Informationen am Endpunkt der Reizverarbeitung dient, bemerkt man, dass Neurogenese an einer strategisch günstigen Position geschieht [22]: die Informationen aus vielen Bereichen des Neocortex werden über EC auf nur ca. 300000 Neurone (in Mäusen) des Gyrus dentatus projiziert [22]; dies ist eine dramatische Dimensionsreduktion. Da die Information vom Hippokampus zum Ausgangspunkt im Neocortex zurück projiziert wird [42], werden also Neurone im Bereich der engsten Stelle des Informationsflusses platziert [22]; die nachfolgende CA3-Region gilt als Punkt größter Konvergenz (Abschnitt 2.1, vergl. aber auch Diskussion).

Der Gyrus dentatus übt eine Kodierung der Information für die Speicherung in der CA3-Region aus (Abschnitt 2.2). Neurogenese geschieht also möglicherweise zur Anpassung des Kodes des Kodierers. Letzteres ist funktionell sinnvoll, denn eine Adaptation des Kodierers ist notwendig, um Redundanz der den Hippokampus erreichenden Informationen effizient entfernen zu können. Anpassung müsste in einem statistischen Sinne passieren; wären beispielsweise die Aktivität der Neurone im Neocortex gruppenweise miteinander korreliert, könnte dies im GD mit wenigen ("Gruppen"-) Neuronen kodiert werden. Um ein bestimmtes Verhältnis der Aktivität der Gruppen zueinander effizient speichern zu können, ist es nötig, die Gruppen herauszusuchen, die besonders charakteristische, i.e. starke, Variation zueinander aufweisen. Dann ist jedes einzelne gespeicherte Muster besonders aussagekräftig, weil sie sich voneinander stark unterscheiden. Das Auffinden solcher Charakteristika wäre Aufgabe einer adaptiven Kodierung.

Würde sich die Gruppenzuordnung der Neurone im Neocortex wesentlich verändern oder, etwa nach Wechsel in eine vielfältigere Umwelt, komplexere Korrelationen aufweisen, wäre eine Anpassung des Kodierers notwendig. Wir wollen in der vorliegenden Arbeit analysieren, inwieweit zusätzliche Neurone diesem Zweck behilflich sein könnten.

Kapitel 3

Modell

3.1 Einführung und Motivation des Modellsystems

3.1.1 Biologische Basis

Als Grundlage der Modellierung der Neurogenese dient ein einfaches funktionelles Schema des Hippokampus [42] (Abb. 3.1). Es basiert auf der These, dass der Hippokampus die Aufgabe hat, Muster – wenn auch nur vorübergehend – zu speichern. Unter "Muster" verstehen wir die Aktivität (Feuerraten) einer Ansammlung von Neuronen, welche beispielsweise durch ein vom Organismus erlebtes Ereignis ausgelöst wird. Da die CA3-Region durch starke Vernetzung ihrer Pyramidenzellen einem Autoassoziativspeicher ähnelt (Abschnitt 2.1), wird angenommen, dass hier Muster gespeichert werden.

Obwohl der Hippokampus in mehr Substrukturen gegliedert ist (Abschnitt 2.1), legen wir in bezug auf den Speicher zwei funktionelle Einheiten fest. Der Kodierer bestehe aus dem Gyrus dentatus (GD); er erhält Muster vom entorhinalen Cortex (EC) und gibt diese transformiert weiter an den Speicher CA3. Ein Ziel der Kodierung ist die Kompression und Integration neocorticaler Information (Abschnitt 2.2). Kompression wird durch Redundanzreduktion erreicht. Desweiteren wird die Information in einen spärlichen Kode umgewandelt [42], ein Aspekt den wir in dieser Arbeit nur diskutieren können (Abschnitt 6).

Die zweite logische Einheit ist ein Dekodierer, der die kodierten Speicherinhalte der CA3-Region entziffert, welche dann über den EC in neocorticale Bereiche zurück transferiert werden. Dieser Prozess wird als Erinnerung an ein Erlebnis interpretiert. Wir identifizieren Treves und Rolls folgend [42] die CA1-Region als den Dekodierer.

Die Gliederung des Hippokampus in drei logische Einheiten, i.e. Kodierer, Speicher und Dekodierer, führt zunächst auf ein einfaches Modell des Hippokampus, welches in Abb. 3.1 schematisch dargestellt ist. Neocorticale Areale projizieren Information via EC auf den Gyrus dentatus. GD "schreibt" die kodierten Muster in den Speicher CA3. Über CA1 werden die Speicherinhalten dekodiert, und somit erinnert oder ins stabilere (corticale) Langzeitgedächtnis konsolidiert. Kodierung und Dekodierung wird durch spezielle "Wahl" der synaptischen Verbindungsstärken der Fasern zwischen dem

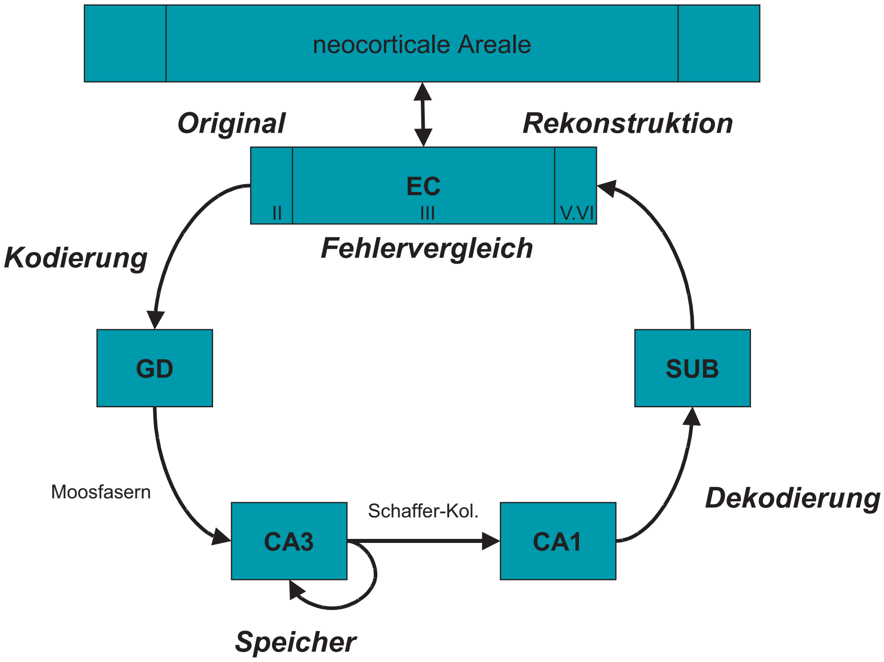

Abbildung 3.1: Biologische Grundlage des Modells. Neocorticale Information (Original), die im Hippokampus gespeichert werden soll, erreicht via entorhinalem Cortex (EC) den Hippokampus. Auf dem Weg zum Gyrus dentatus wird die Information kodiert und so in den Speicher, die CA3-Region, aufgenommen. Auf dem Rückweg (beim Erinnern oder im "Adaptationsmodus", siehe Abschnitt 3.1.2) wird sie dekodiert, d.h. das Original wird rekonstruiert. Zur Adaptation von Kodierung und Dekodierung wird in den Schichten der Cortizes ein Vergleich der Rekonstruktion mit dem Original zur Bestimmung des Rekonstruktionsfehlers vorgenommen.

Speicher CA3 und EC erreicht. Wir nehmen an, dass die Moosfasern vorallendingen für das Einschreiben der Muster in den Speicher verantwortlich sind (vergl. Abschnitt 2.1), während die Synapsen des Tractus perforans auf GD-Neurone die eigentliche Kodierung, d.h. die Redundanz- und Dimensionsreduktion neocorticaler Muster, vornehmen. Die Dekodierung passiere sowohl an den Synapsen der Schaffer-Kollateralen auf CA1 als auch an den Synapsen, die die CA1-Pyramidenzellen via Subiculum mit EC-Neuronen verbinden. Da die Information schließlich Bereiche des Neocortex erreicht, könnte man auch jene Synapsen zur Dekodierung zählen.

3.1.2 Modell zur Untersuchung der Adaptation

In Abschnitt 2.3 wurde die Hypothese geäußert, dass Neurogenese eine Form der Adaptation des Kodierers sei. Anpassung der Kodierung an die charakteristische Aktivität neocorticaler Areale ist prinzipiell auch ohne Neurogenese möglich. Die synaptischen Gewichte müssen sich dafür plastisch derart arrangieren, dass diejenigen Komponenten der neuronalen Muster, welche für die Aktivität der neocorticalen Bereiche charakteristisch sind, möglichst effizient kodiert und dekodiert, i.e. rekonstruiert, werden können. Ändern sich die Charakteristika, muss der Kodierung und Dekodierung mit geeigneter Modifikation der Gewichte reagieren. Wir wollen zunächst klären wie der Anpassungsprozess der Kodierung und Dekodierung im vorgestellten Modell vonstatten gehen könnte. Danach wird das Modell auf die Untersuchung des Adaptationsprozess spezialisiert.

Für eine Anpassung des Kodes ist ein Optimierungskriterium nötig; es muss die Möglichkeit bestehen, die Güte eines Kodes zu überprüfen. Da der Hippokampus unseren Annahmen zufolge die Aufgabe hat, Muster zu speichern, um sie zu einem späteren Zeitpunkt zu reaktivieren, ist die Abweichung der rekonstruierten von der ursprünglichen Information, i.e. der Rekonstruktionsfehler, ein geeignetes Maß für die Güte von Kodierung und Dekodierung. Berechnung der Abweichung zwischen Rekonstruktion eines Musters und seinem Original könnte in den corticalen Schichten (im EC [7] und möglicherweise auch in den neocorticalen Arealen) geschehen (vergl. Abb. 3.1). Dazu müsste ein Muster aus den neocorticalen Arealen im Hippokampus kodiert, dekodiert und zurück projiziert werden, während gleichzeitig das originäre Muster aktiv bleibt, um einen Vergleich zu ermöglichen [38]. Dieser postulierte Adaptationsmodus wäre nur ein Modus operandi unter mehreren (Konsolidierung, Speicherung, Erinnerung) und wird auch in der Literatur diskutiert [7].

Um Effekte der Neurogenese im Adaptationsprozess des Kodierers besser verstehen und untersuchen zu können, werden wir im Modell des Hippokampus (Abb. 3.1) Kodierung und Dekodierung in den Mittelpunkt der Betrachtungen setzen. Wir beschränken es deshalb auf die minimalen Anforderungen des Adaptationsmodus. Alle Neurone der corticalen Areale (EC und Neocortex), die auf den Hippokampus projizieren, seien zu einer logischen Neuronenschicht zusammengefasst; wir werden diese Neuronenschicht im Folgenden einfach als EC bezeichnen. Diese Schicht sei mit den Neuronen des Hippokampus verbunden, welche ihrerseits eine logische Neuronenschicht bilden; wir nennen sie die verborgene Schicht (s.u.). Der Hippokampus leitet dann die Information zurück zum EC. An den Synapsen auf die Hippokampusneurone werde die Kodierung, auf dem Weg zurück zum EC die Dekodierung vollzogen. Gespeichert werde die kodierte Information, d.h. die Aktivität der Neurone der verborgenen Schicht. In dieser Vereinfachung weisen die Strukturen GD, CA3 und CA1 identische Aktivität auf, die Kodierungs- und Dekodierungsvorgänge wurden vollständig den Synapsen zwischen den corticale Arealen und dem Hippokampus zugesprochen. Da der Speicher CA3 nicht explizit modelliert wird, befindet sich das Modell stets im Adaptationsmodus.

Abb. 3.2 zeigt ein Schema des vereinfachten Modells. Die Struktur der Vernetzung gleicht einem bekannten neuronalen Netzwerk, dem linearen Autokodierer [24, 18, 41]. In unserem Fall besteht er aus drei neuronalen Schichten, wobei EC gleichzeitig als Eingabe und- Ausgabeschicht dient, während dazwischen die verborgene Schicht, i.e. der Hippokampus, liegt.

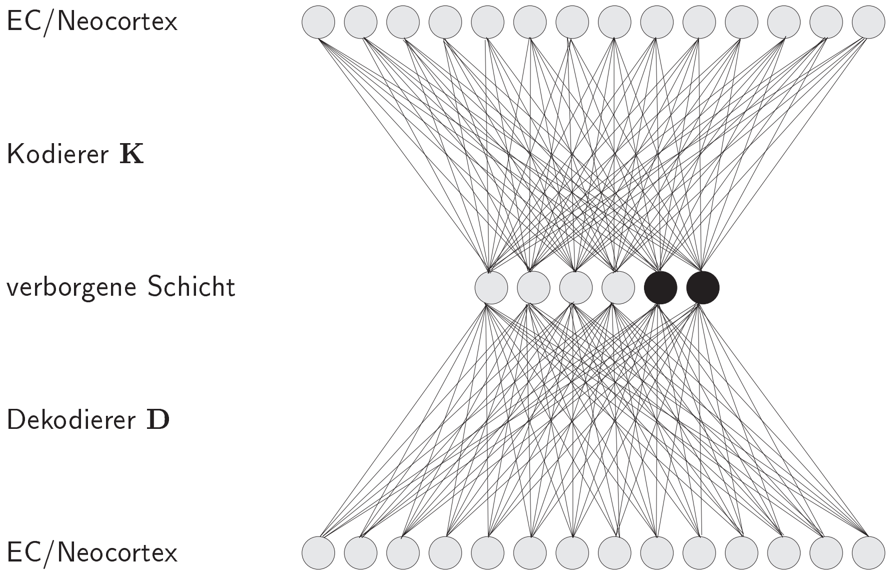

Abbildung 3.2: Darstellung des Modells. Neurone sind durch Kreise, synaptische Verbindungen durch Linien angedeutet. Jedes Neuron aus der EC-Schicht ist mit jedem der verborgenen Schicht verbunden. Die EC-Schicht ist doppelt gezeichnet (oben und unten). Neurogenese geschieht durch das Vergrößern der verborgenen Schicht (schwarze Kreise).

3.1.3 Adaptation und Neurogenese

Neurogenese scheint mit der Aufnahme von Erfahrung im Zusammenhang zu stehen (siehe Abs. 2.3). Um die Bildung neuer Erfahrungen in der Modellierung zu konkretisieren, bemühen wir folgendes Konzept eines Umgebungswechsels.

Stellen wir uns vor, dass ein Tier in einer begrenzten Umwelt lebt, in der es einer bestimmten Sorte von Eindrücken ausgesetzt ist. Die Eindrücke rufen bestimmte neuronale Aktivitätsverteilungen, i.e. neuronale Muster, im Neocortex hervor, welche vom Hippokampus mittels der Kodierung abstrahiert werden. Wird das Tier unbekannten Eindrücken ausgesetzt, etwa durch einen Wechsel in eine fremde Umgebung, sollte sich auch die Aktivitätsverteilung des Neocortex beeinflusst zeigen, so dass sich die charakteristischen Komponenten der neuronalen Muster verändern. Adaptation der Kodierung an die neue Situation würde die Bedeutung der Abstraktion früherer Eindrücke verschieben, so dass in der Konsequenz Eindrücke der "alten" Sorte vergessen oder missverstanden werden würden. Sollte das Tier in die fremde Umgebung übersiedeln, scheint das Vergessen nebensächlich. Erweitert das Tier seine Umwelt jedoch um die neue Umgebung, müssen alle Eindrücke effizient kodiert werden können; das Tier hat zusätzliche Erfahrungen gesammelt und muss damit, im Sinne eines angepassten Kodes, umgehen können. Insgesamt ist dadurch die Umwelt des Tieres, und damit die Aktivitätsverteilung des Neocortex, komplexer geworden. Genau bei dem Zuwachs an Erfahrung vermuten wir die Notwendigkeit neuer Neurone im Kodierer.

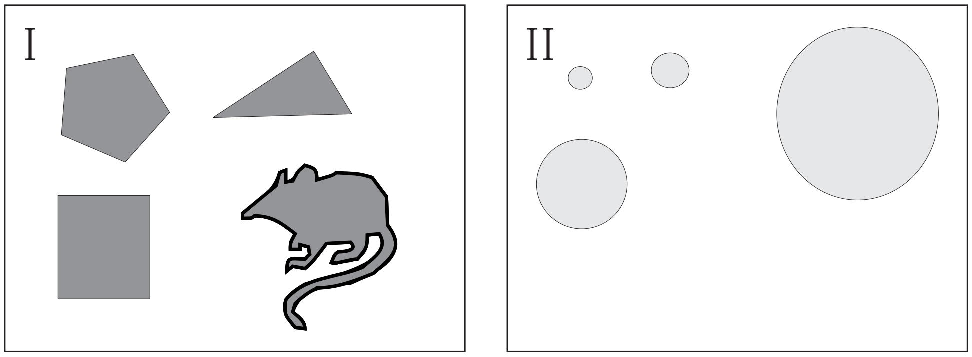

Abbildung 3.3: Umgebungswechsel. Zwei Umgebungen, in denen unterschiedliche Sorten Eindrücke herrschen, symbolisiert durch Vielecke bzw. Kreise. Umgebung I ist das bekannte Umfeld der Maus, während Umgebung II fremde Eindrücke beherbergt (siehe Text).

Wir wollen die angesprochenen Adaptationsvorgänge an einer Maus und zwei Umgebungen karikieren. In Abb. 3.3 sind die Eindrücke pointiert durch Vielecke bzw. Kreise in den Umgebungen angedeutet. Die Eindrücke innerhalb einer Umgebung ähneln sich, eine Kodierung würde somit einen Redundanz befreiten Kode bilden können. Um die Eindrücke von Umgebung I unterscheiden zu können, reicht es beispielsweise ein "Eckenneuron" zu benutzen, dass die Ecken der Vielecke zählt. Kommt die Maus jedoch aus ihrer gewohnten Umgebung I in die fremde Umgebung II, ist das "Eckenneuron" offensichtlich eine ungeeignete Kodierung; die Maus könnte größere und kleinere Kreise nicht voneinander unterscheiden (Abb. 3.3). Kodierer und Dekodierer müssen sich auf die neue Situation einstellen, weil sich die charakteristischen Merkmale der Eindrücke geändert haben. Das "Eckenneuron" könnte zum Beispiel in ein "Größenneuron" umgewandelt werden, welches proportional zur Größe der Kreise feuert. Es ist klar, dass das

"Größenneuron" seinerseits in Umgebung I keine effiziente Kodierung darstellt. Plastizität alleine genügt also nicht, sich in beiden Umgebungen zurecht zu finden; es sei denn es gäbe ein gemeinsames Merkmal zur Unterscheidung aller Eindrücke. Die Stabilität des Kodes früherer Eindrücke ist außerdem wichtig für das Abrufen von Gedächtnisinhalten des Hippokampus; diese müssen auch nach dem Adaptationsprozess vom Dekodierer korrekt dekodiert werden können. Die Lösung des Problems scheint auf der Hand zu liegen: Wenn zum "Eckenneuron" ein neues Neuron im Kodierer hinzu käme, dass als "Größenneuron" fungiert, könnten Eindrücke beider Umgebungen korrekt klassifiziert werden.

Aus dieser Überlegung heraus postulieren wir, dass die synaptischen Gewichte der Neurone im Kodierer nur in der Frühphase ihrer funktionellen Existenz plastisch sind, später dagegen stabil [24]. Experimentell ist dies nicht eindeutig bestätigt, es ist aber durchaus möglich [46]. Die Plastizität der Neurone nehme also im Laufe ihrer Existenz ab. Dann sind neu gebildete Neurone eher in der Lage, sich an fremde Umgebungen anzupassen, als bestehende, ältere Neurone. Letztere haben sich, nach unserer Überlegung, bereits an frühere Umgebungen angepasst und versuchen ihre Gewichte im Umfeld von neuen Eindrücken beizubehalten. Neurogenese wird in unserem Modell durch Erhöhung der Anzahl der Neurone der verborgenen Schicht beschrieben (symbolisiert durch schwarze Kreise in Abb. 3.2). Tatsächlich vergrößern sich dadurch auch die kodierte Darstellung der Muster, d.h. die Speicherinhalte, und der Dekodierer, obwohl physiologisch Neurogenese nur im GD gefunden wird (siehe Abs. 2.3); dies ist ein Tribut an die Einfachheit des Modells (vergl. Diskussion).

Wir gehen davon aus, das sich (die Gewichte der) Neurone des Dekodierers, im Gegensatz zu denen des Kodierers, stets plastisch anpassen, und zwar derart, dass der Rekonstruktionsfehler der verarbeiteten Muster im Mittel minimiert wird. Mit der Minimierung des Rekonstruktionsfehlers werden sich weite Teile des folgenden Abschnitts 3.2 beschäftigen.

3.2 Mathematische Formulierung

3.2.1 Definition des Modells

Jedes der n EC-Neurone hat eine Verbindung zu jedem der m Neurone der verborgenen Schicht (vergl. Abb. 3.2). Kodierer[1] $\mathbf{K} \in \mathbb{R}^{(m,n)}$ und Dekodierer $\mathbf{D} \in \mathbb{R}^{(n,m)}$ sind Matrizen, die die synaptischen Gewichte der EC-Neurone auf die verborgene Schicht bzw. der Neurone der verborgenen Schicht auf die EC-Neurone repräsentieren (Abb. 3.2).

Wir gehen davon aus, dass die Frequenz der Spikes der Neurone, also ihre Feuerrate, linear mit der Summe der synaptischen Eingänge steigt. Ein Zeilenvektor $\mathbf{k}_i^T$ von $\mathbf{K}$ entspricht den Gewichten der synaptischen Eingänge aller EC-Neurone auf das i-

[1] Wir verwenden folgende Konvention zur Schreibweise von Variablen. Fett gedruckte Großbuchstaben bezeichnen Matrizen, fett gedruckte Kleinbuchstaben (Spalten-) Vektoren und kursiv gedruckte Buchstaben Skalare oder sonstiges. Bsp.: $\mathbf{K}$, $\mathbf{k}$, K, k.

te Neuron der verborgenen Schicht, dessen Feuerrate also durch die gewichtete Summe $y_i = \mathbf{k}_i^T \mathbf{x} = \sum_{j=1}^{n} k_{ij} x_j$ gegeben ist, wenn $\mathbf{x}$ die Feuerraten der EC-Neurone bezeichnet. Analog sind die Einträge des Spaltenvektors $\mathbf{d}_i$ von $\mathbf{D}$ die Gewichte der Rückverbindungen des i-ten Neurons der verborgenen Schicht zu jedem EC-Neuron.

Das Modell des Hippokampus (Abb. 3.2) führt zur Rekonstruktion eines EC-Musters, i.e. eine Kombination von Feuerraten $\mathbf{x} \in \mathbb{R}^n$ (vergl. auch Abschnitt 3.2.2), insgesamt zwei lineare Operationen aus:

$$\mathbf{z} = \mathbf{DKx} \tag{3.1}$$

Die Funktionsweise des Modells ist also wie folgt (Abb. 3.4). Eine Kombination von Feuerraten der EC-Neurone $\mathbf{x} \in \mathbb{R}^n$, wird durch den Kodierer $\mathbf{K}$ linear transformiert, i.e. $\mathbf{y} = \mathbf{Kx}$. Vektor $\mathbf{y} \in \mathbb{R}^m$ bezeichnet die Feuerraten der Neurone der verborgenen Schicht. Das kodierte Muster $\mathbf{y}$ wird durch $\mathbf{D}$ erneut linear transformiert, i.e. $\mathbf{z} = \mathbf{Dy}$. Die Rekonstruktion $\mathbf{z}$ des Originals $\mathbf{x}$ erreicht so wieder die EC-Neurone. Die Aktivität der verborgenen Schicht $\mathbf{y}$ wird in unserem einfachen Modell als gespeichertes Muster angesehen.

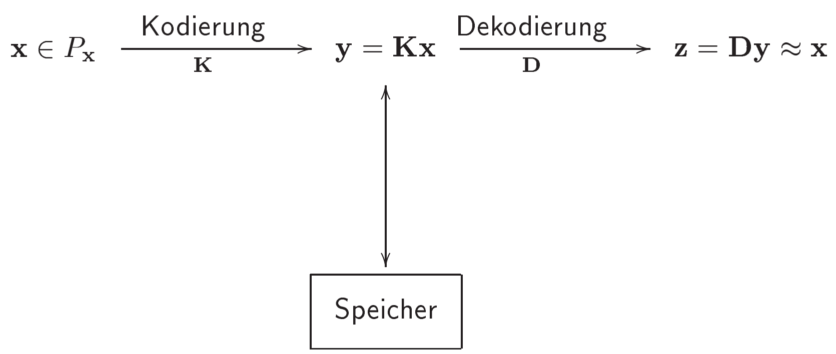

Abbildung 3.4: Hippokampus als Speicher mit Kodierer und Dekodierer. EC-Muster $\mathbf{x}$, nach $P_{\mathbf{x}}$ verteilt, wird mittels $\mathbf{K}$ kodiert (lineare Transformation). Das kodierte Zwischenformat $\mathbf{y} = \mathbf{Kx}$ wird im Speicher abgelegt. Der Dekodierer $\mathbf{D}$ rekonstruiert aus dem kodierten Signal eine Approximation $\mathbf{z} = \mathbf{DKx}$ an das Eingangsdatum $\mathbf{x}$.

Eindrücke einer Umgebung (siehe Abs. 3.1.3) manifestieren sich in einer bestimmten Aktivitätsverteilung im Neocortex; sie lösen also ein bestimmtes EC-Muster $\mathbf{x}$ aus. Die Gesamtheit der Eindrücke einer Umgebung wird folglich durch die Angabe einer Wahrscheinlichkeitsdichte $P_{\mathbf{x}}$ der EC-Muster $\mathbf{x}$ modelliert. Bei Wechsel der Umgebung verändert sich die Verteilung[2] $P_{\mathbf{x}}$ der EC-Muster $\mathbf{x}$, denn die Charakteristik der Eindrücke werden sich von Umgebung zu Umgebung unterscheiden.

Wir verwenden die mittlere, quadratische Abweichung zwischen Originalmuster $\mathbf{x}$ und Rekonstruktion $\mathbf{z}$ als ein Kriterium zur Adaptation des Kodierers und des Deko-

[2] Den Begriff "Verteilung" verwenden wir synonym zu "Wahrscheinlichkeitsdichte".

dierers an eine Verteilung $P_{\mathbf{x}}$, d.h. [3]

$$\varepsilon := \left\langle |\mathbf{x} - \mathbf{z}|^2 \right\rangle \tag{3.2}$$

$$\text{[Gl. 3.1]} \qquad = \left\langle |\mathbf{x} - \mathbf{DKx}|^2 \right\rangle \tag{3.3}$$

Während die Adaptation des Kodierers nicht nur durch den Rekonstruktionsfehler bedingt ist, nehmen wir an, dass die Gewichte des Dekodierer stets bestrebt sind, den minimalen Rekonstruktionsfehler zu erreichen (siehe Abs. 3.1.3). Der optimale Dekodierer $\mathbf{D}^{\mathrm{opt}}$ ist also diejenige Matrix $\mathbf{D} \in \mathbb{R}^{(n,m)}$, die die kodierte Darstellung $\mathbf{y} = \mathbf{Kx}$ im Mittel über die $\mathbf{x}$ einer Umgebung mit geringstem Fehler rekonstruiert:

$$\mathbf{D}^{\mathrm{opt}} = \min_{\mathbf{D} \in \mathbb{R}^{(n,m)}} \left\langle |\mathbf{x} - \mathbf{z}|^2 \right\rangle \tag{3.4}$$

Nachdem im nächsten Abschnitt zunächst Voraussetzungen und Bezeichnungen für alle weiteren Rechnungen genannt werden, werden wir den optimalen Dekodierer im Sinne von Gl. 3.4 näher bestimmen. Damit wird sich Abschnitt 3.2.4 beschäftigen. Danach (Abschnitt 3.2.5) werden wir Kodierer $\mathbf{K}$ herleiten, welche die charakteristischen Aspekte der Muster einer Umgebung besonders gut erfassen. Das Kapitel wird mit einem Beispiel abgeschlossen.

3.2.2 Voraussetzungen und Bezeichnungen

Wir möchten an dieser Stelle Voraussetzungen angeben, die in allen weiteren Rechnungen der vorliegenden Arbeit verwendet werden, sofern nichts Gegenteiliges gesagt wird.

Verteilung Die Aktivität der Neurone ("Feuerraten") können in der mathematischen Formulierung auch negativ sein. Wir stellen uns dabei vor, dass Abweichungen von einer mittleren Feuerrate modelliert werden. Dadurch ist die Verteilung $P_{\mathbf{x}}$ der $\mathbf{x}$ ebenfalls mittelwertfrei. Die Zweitemomentematrix $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ der Verteilung $P_{\mathbf{x}}$ ist daher gleich der Kovarianzmatrix der Verteilung. Kovarianzmatrizen sind positiv semidefinite, symmetrische Matrizen. Wir werden fordern, dass sie zusätzlich invertierbar sind. Damit hat $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ nur positive Eigenwerte und ist deshalb (und aufgrund ihrer Symmetrie) sogar positiv definit. Die Eigenwerte der Kovarianzmatrix entsprechen der Varianz in den Richtungen ihrer Eigenvektoren und sind daher anschauliche Größen. Wir schreiben auch "$\mathbf{x} \in P_{\mathbf{x}}$" für die Tatsache, dass $\mathbf{x}$ gemäß Wahrscheinlichkeitsdichte $P_{\mathbf{x}}$ verteilt ist.

Kodierer und Dekodierer Wir werden stets annehmen, dass jeder Kodierer $\mathbf{K}$ Höchstrang hat. Damit sind seine Zeilenvektoren linear unabhängig, der Rang des Kodierers entspricht also seiner Spaltenzahl, i.e. der Anzahl der Neurone der verborgenen

[3] Links neben den Gleichungen in eckigen Klammern sind zum leichteren Verständnis Anmerkungen zu Umformungen aufgelistet. Diese Notationsweise wird beibehalten.

Schicht, $\operatorname{Rg}\mathbf{K} = m$. Da wir die Spaltenzahl von $\mathbf{K}$ später variieren werden, stellt die Annahme keine Beschränkung der Allgemeinheit dar. Außerdem sei $\mathbf{K}$ (und auch $\mathbf{D}$) unabhängig von einzelnen Mustern $\mathbf{x} \in P_{\mathbf{x}}$. Der Hintergrund dieser Unabhängigkeit ist die Annahme, dass die Anpassung der Kodierung und Dekodierung auf einer sehr viel längeren Zeitskala verläuft als die Rekonstruktion einzelner Muster.

Bild und Kern Im Folgenden wird von Bild und Kern einer Matrix gesprochen, deshalb wollen wir die Begriffe rekapitulieren und die Bezeichner klären (für Nachweise siehe z.B. [34]).

Eine Matrix $\mathbf{M} \in \mathbb{R}^{(n,m)}$ ist eine lineare Abbildung (Homomorphismus) der Art $\mathbb{R}^m \mapsto \mathbb{R}^n$. Das Bild dieser Matrix, in Zeichen $\operatorname{Bild}\mathbf{M}$, ist der Unterraum des Vektorraums $\mathbb{R}^n$, der durch die Spaltenvektoren von $\mathbf{M}$ aufgespannt wird. Sind $\mathbf{m}_i$, $i = 1,\ldots,m$, die Spaltenvektoren von $\mathbf{M}$, wird das Bild von $\mathbf{M}$ also durch alle Linearkombinationen der Spaltenvektoren erzeugt, wir schreiben (in spitzen Klammern) $\operatorname{Bild}\mathbf{M} = \langle \mathbf{m}_i | i \in \mathbb{N}_m \rangle$, wobei $\mathbb{N}_m$ für die Menge der natürlichen Zahlen $\{1,\ldots,m\}$ steht.

Der Kern von $\mathbf{M}$ ist ein Unterraum von $\mathbb{R}^m$ (und nicht von $\mathbb{R}^n$), wir schreiben $\operatorname{Kern}\mathbf{M}$. Für alle Vektoren $\mathbf{v} \in \operatorname{Kern}\mathbf{M}$ gilt $\mathbf{M}\mathbf{v} = \mathbf{0}$.

Es gilt $\operatorname{Bild}\mathbf{M} \cup \operatorname{Kern}\mathbf{M}^T = \mathbb{R}^n$ und $\operatorname{Bild}\mathbf{M} \cap \operatorname{Kern}\mathbf{M}^T = \emptyset$. Die Summe der Dimensionen von $\operatorname{Bild}\mathbf{M}$ und $\operatorname{Kern}\mathbf{M}^T$ ist daher n.

Spur Wir wollen an dieser Stelle darauf hinweisen, dass die Spur $\operatorname{Sp}\mathbf{C}$ einer symmetrischen Matrix $\mathbf{C}$ gleich der Summe der Eigenwerte von $\mathbf{C}$ ist. Eine symmetrische Matrix besitzt stets eine Eigenwertdarstellung $\mathbf{C} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$, wobei $\mathbf{W} = (\mathbf{w}_1,\ldots,\mathbf{w}_n)$ die Matrix der Eigenvektoren $\mathbf{w}_i$ und $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$ die Diagonalmatrix der Eigenwerte ist. Damit folgt nun

$$\begin{aligned}
& & \operatorname{Sp}\mathbf{C} &= \operatorname{Sp}\left(\mathbf{W}\mathbf{\Lambda}\mathbf{W}^T\right) \\
&[\operatorname{Sp}(\mathbf{AB}) = \operatorname{Sp}(\mathbf{BA})] & &= \operatorname{Sp}\left(\mathbf{W}^T\mathbf{W}\mathbf{\Lambda}\right) \\
&[\mathbf{W}^T\mathbf{W} = \mathbf{I}] & &= \operatorname{Sp}(\mathbf{\Lambda}) \\
& & &= \sum_{i=1}^{\dim\mathbf{C}} \lambda_i.
\end{aligned} \tag{3.5}$$

Die Eigenvektoren einer symmetrischen Matrix $\mathbf{C} \in \mathbb{R}^{(n,n)}$ sind stets so wählbar, dass sie eine Orthonormalbasis des $\mathbb{R}^n$ bilden.

3.2.3 Rekonstruktionsfehler

In diesem Abschnitt wird der Rekonstruktionsfehler (Gl. 3.3) genauer betrachten. Wir schreiben um die Abhängigkeiten deutlich werden zu lassen

$$\varepsilon(\mathbf{K},\mathbf{D},P_{\mathbf{x}}) = \left\langle |\mathbf{x} - \mathbf{DKx}|^2 \right\rangle_{\mathbf{x}\in P_{\mathbf{x}}} \tag{3.6}$$

Diese Gleichung lässt sich umformulieren:

$$\varepsilon = \left\langle |\mathbf{x} - \mathbf{DKx}|^2 \right\rangle \tag{3.7}$$

$$= \left\langle (\mathbf{x} - \mathbf{DKx})^T (\mathbf{x} - \mathbf{DKx}) \right\rangle \tag{3.8}$$

$$= \left\langle \mathbf{x}^T (\mathbf{I} - \mathbf{DK})^T (\mathbf{I} - \mathbf{DK})\, \mathbf{x} \right\rangle \tag{3.9}$$

$$= \left\langle \mathbf{x}^T \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} \mathbf{x} \right\rangle \tag{3.10}$$

Die genaue Bedeutung von $\hat{\mathbf{Q}} := (\mathbf{I} - \mathbf{DK})^T$ wird später erläutert. Man kann aber jetzt schon erkennen, dass $\hat{\mathbf{x}} := \hat{\mathbf{Q}}\mathbf{x}$ der Anteil des Vektors $\mathbf{x}$ ist, der zum Fehler ε beiträgt; wir nennen $\hat{\mathbf{x}}$ daher Residuenvektor. Der Rekonstruktionsfehler gleicht nach Gl. 3.10 der Varianz der Residuenvektoren

$$\varepsilon = \left\langle |\hat{\mathbf{x}}|^2 \right\rangle . \tag{3.11}$$

Ist ein Vektor $\mathbf{x}^\perp$ orthogonal zu den Zeilenvektoren in $\mathbf{K}$, i.e. $\mathbf{Kx}^\perp = \mathbf{0}$, folgt, dass $\mathbf{x}^\perp$ gleich seinem Residuenvektor ist, $\hat{\mathbf{x}} = \mathbf{x}^\perp - \mathbf{DKx}^\perp = \mathbf{x}^\perp$. Zu $\mathbf{K}$ orthogonale Vektoren werden nicht kodiert und können demnach auch nicht gespeichert werden. Denn der Kodierer bildet die Vektoren auf den Nullvektor ab, $\mathbf{y} = \mathbf{Kx}^\perp = \mathbf{0}$, so dass keine Aktivität an der verborgenen Schicht ausgelöst wird, die gespeichert werden könnte. Die Zeilenvektoren von $\mathbf{K}$ bestimmen daher den Unterraum der Vektoren, welche nicht in den Speicher gelangen können.

Benutzt man nun neben der Linearität Eigenschaften der Spurabbildung $\mathrm{Sp}\,(\cdot)$, lässt sich Gl. 3.10 weiter umformen:

$$\varepsilon = \left\langle \mathbf{x}^T \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} \mathbf{x} \right\rangle \tag{3.12}$$

$$[\gamma = \mathrm{Sp}\,(\gamma)\,,\ \gamma \in \mathbb{R}] \qquad = \left\langle \mathrm{Sp}\left(\mathbf{x}^T \hat{\mathbf{Q}}^T \hat{\mathbf{Q}} \mathbf{x} \right) \right\rangle \tag{3.13}$$

$$[\mathrm{Sp}\,(\mathbf{AB}) = \mathrm{Sp}\,(\mathbf{BA})] \qquad = \left\langle \mathrm{Sp}\left(\hat{\mathbf{Q}} \mathbf{x}\mathbf{x}^T \hat{\mathbf{Q}}^T \right) \right\rangle \tag{3.14}$$

$$[\text{Linearität von } \langle\cdot\rangle] \qquad = \mathrm{Sp}\left(\hat{\mathbf{Q}} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \hat{\mathbf{Q}}^T \right) \tag{3.15}$$

Gl. 3.15 zeigt, dass der Rekonstruktionsfehler ε nur über die Kovarianzmatrix $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ der Verteilung $P_\mathbf{x}$ abhängt.

3.2.4 Optimaler Dekodierer

Wir betrachten nun das Problem, eine Matrix $\mathbf{D}^{\mathrm{opt}} \in \mathbb{R}^{(n,m)}$ zu finden, die den Rekonstruktionsfehler minimiert:

$$\mathbf{D}^{\mathrm{opt}}(\mathbf{K}, P_\mathbf{x}) := \min_{\mathbf{D} \in \mathbb{R}^{(n,m)}} \varepsilon(\mathbf{K}, \mathbf{D}, P_\mathbf{x}) \tag{3.16}$$

$$= \min_{\mathbf{D} \in \mathbb{R}^{(n,m)}} \left\langle |\mathbf{x} - \mathbf{DKx}|^2 \right\rangle_{\mathbf{x} \in P_\mathbf{x}}$$

Wir interessieren uns für den Fall $n > m$. Dann sind $\mathbf{D} \in \mathbb{R}^{(n,m)}$ und $\mathbf{K} \in \mathbb{R}^{(m,n)}$ nicht quadratisch und es gibt kein Inverses zu $\mathbf{K}$. Gesucht ist die Matrix $\mathbf{D}$, die am ehesten die Kodierung $\mathbf{K}$ invertiert, d.h. den Rekonstruktionsfehler minimiert.

Da die offene Menge $\mathbb{R}^{(n,m)}$ betrachtet wird, muss an der Stelle eines Minimums $\mathbf{D} = \mathbf{D}^{\mathrm{opt}}$ die Ableitung des Rekonstruktionsfehler für jede Komponente von $\mathbf{D}$ verschwinden. Um dies auszunutzen, werden zwei Aussagen über Matrixableitungen benötigt. Seien eine beliebige Matrix $\mathbf{M}$, eine symmetrische Matrix $\mathbf{C}$, und Vektoren $\mathbf{u}$ und $\mathbf{v}$ derart gewählt, dass die Matrixmultiplikationen in den folgenden Gleichungen definiert sind (äußeres oder inneres Produkt). Dann gilt

$$\frac{\partial}{\partial \mathbf{M}} \left(\mathbf{u}^T \mathbf{M} \mathbf{v}\right) = \mathbf{u}\mathbf{v}^T \tag{3.17}$$

und

$$\frac{\partial}{\partial \mathbf{M}} \left(\mathbf{u}^T \mathbf{M}^T \mathbf{M} \mathbf{u}\right) = 2\mathbf{M}\mathbf{u}\mathbf{u}^T \tag{3.18}$$

Damit lässt sich ausrechnen:

$$\begin{aligned}
&[\text{Gl. 3.8}] && \left.\frac{\partial \varepsilon}{\partial \mathbf{D}}\right|_{\mathbf{D}^{\mathrm{opt}}} = \left\langle \frac{\partial}{\partial \mathbf{D}} \left((\mathbf{x} - \mathbf{DKx})^T (\mathbf{x} - \mathbf{DKx})\right)\Big|_{\mathbf{D}^{\mathrm{opt}}} \right\rangle \\
&[\gamma^T = \gamma, \gamma \in \mathbb{R}] && = \left\langle \frac{\partial}{\partial \mathbf{D}} \left(\mathbf{x}^T\mathbf{x} - 2\mathbf{x}^T\mathbf{DKx} + \mathbf{x}^T\mathbf{K}^T\mathbf{D}^T\mathbf{DKx}\right)\Big|_{\mathbf{D}^{\mathrm{opt}}} \right\rangle \\
&[\text{Gl. 3.17, Gl. 3.18}] && = -2\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T + 2\mathbf{D}^{\mathrm{opt}}\mathbf{K}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T \\
&[\text{Min. ges.}] && = \mathbf{\Theta}
\end{aligned} \tag{3.19}$$

Die Matrix $\mathbf{\Theta}$ sei die (passende) Nullmatrix, also eine Matrix mit lauter Nullen als Einträgen. In Gl. 3.19 wurde Gl. 3.17 mit $\mathbf{u} := \mathbf{x}$ und $\mathbf{v} := \mathbf{Kx}$, sowie Gl. 3.18 mit $\mathbf{u} := \mathbf{Kx}$ und jeweils mit $\mathbf{M} := \mathbf{D}$ verwendet. Es ergibt sich schließlich, weil $\mathbf{K}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T$ wegen der Invertierbarkeit von $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ und dem Höchstrang von $\mathbf{K}$ (siehe Abs. 3.2.2) invertierbar ist:

$$\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T = \mathbf{D}^{\mathrm{opt}}\mathbf{K}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T \tag{3.20}$$

$$\mathbf{D}^{\mathrm{opt}} = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T \left(\mathbf{K}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T\right)^{-1} \tag{3.21}$$

Im Anhang wird gezeigt, dass in Gl. 3.21 tatsächlich ein Minimum vorliegt und nicht etwa ein Maximum (siehe Abs. A.1.1). Wir nennen $\mathbf{D}^{\mathrm{opt}}$ den optimalen Dekodierer (bei fester Dimensionalität).

Gleichung 3.21 gibt den Dekodierer an, welcher eine Kodierung am ehesten invertiert. Es fällt auf, dass die Kovarianz $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ der Muster aus $P_{\mathbf{x}}$ genutzt wird, um $\mathbf{D}^{\mathrm{opt}}$ zu bestimmen. Der Dekodierer nutzt die Kovarianz, um die Varianz der Rekonstruktion in relevanten, d.h. sehr variablen, Richtungen der Verteilung $P_{\mathbf{x}}$ zu verstärken und die Varianz in irrelevanten Richtungen zu vermindern. Diese Verzerrung der Rekonstruktion zugunsten der vorherrschenden “Form” der Muster, ergibt eine verbesserte Rekonstruktion. Nach der Skizze in Abb. 3.3, würde etwa ein abgerundetes Rechteck in Umgebung I, in der Vielecke vorherrschen, in der Rekonstruktion “eckiger” gemacht werden, um den Rekonstruktionsfehler, i.e. die Abweichung der Rekonstruktion von der vorherrschenden “Form” der Originale (Vielecke), zu minimieren. Dieser Effekt wird uns später noch beschäftigen (Kapitel 5).

Folgerungen

Der optimaler Dekodierer projiziert die niedrigdimensionale, kodierte Darstellung der Muster $\mathbf{Kx}$ i.A. nicht in den gleichen Unterraum des $\mathbb{R}^n$ zurück, wo der Kodierer sie “her” hatte, d.h. $\mathrm{Bild}\,\mathbf{D}^{\mathrm{opt}} \neq \mathrm{Bild}\,\mathbf{K}^T$. Vektoren $\mathbf{x}'$ aus dem von den (Zeilen-)Vektoren des Kodierers aufgespannten Raum, i.e. $\mathbf{x}' \in \mathrm{Bild}\,\mathbf{K}^T$, können also durch die Rekonstruktion modifiziert werden, wir haben es oben bereits angedeutet. An einem Beispiel wollen wir dies nachweisen, indem wir $\mathbf{D}^{\mathrm{opt}}\mathbf{Kx}' \neq \mathbf{x}'$ zeigen. Sei $\langle \mathbf{xx}^T \rangle = \mathrm{diag}(\lambda_1, \lambda_2)$ und $\mathbf{K} = \begin{pmatrix} 1 & 1 \end{pmatrix}$. Der zugehörigen optimale Dekodierer lautet dann nach Gl. 3.21 $\mathbf{D}^{\mathrm{opt}} = \frac{1}{\lambda_1+\lambda_2} \begin{pmatrix} \lambda_1 & \lambda_2 \end{pmatrix}^T$. Die Rekonstruktion des Vektors $\mathbf{x}' = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$, der offensichtlich im Bild von $\mathbf{K}^T$ liegt, ist gegeben durch

$$\mathbf{D}^{\mathrm{opt}}\mathbf{Kx}' = \frac{2}{\lambda_1 + \lambda_2} \begin{pmatrix} \lambda_1 & \lambda_2 \end{pmatrix}^T \neq \begin{pmatrix} 1 & 1 \end{pmatrix}^T = \mathbf{x}'. \tag{3.22}$$

Die Rekonstruktion von $\mathbf{x}'$ ist also nicht mit der Richtung des Kodierers identisch. Dagegen gilt allgemein die Aussage

$$\mathbf{D}^{\mathrm{opt}}\mathbf{Kx}'' = \mathbf{x}'', \quad \mathbf{x}'' \in \mathrm{Bild}\,\mathbf{D}^{\mathrm{opt}}, \tag{3.23}$$

d.h. für ein Vektor $\mathbf{x}'' \in \mathrm{Bild}\,\mathbf{D}^{\mathrm{opt}}$ ist die Rekonstruktion tatsächlich mit dem ursprünglichen Vektor identisch. Denn es gilt zunächst

$$[\text{Gl. } 3.21] \qquad \mathbf{KD}^{\mathrm{opt}} = \mathbf{I}, \tag{3.24}$$

wie man leicht aus der Gleichung für den optimalen Dekodierer (Gl. 3.21) erkennt. Da nun $\mathbf{x}''$ wegen $\mathbf{x}'' \in \mathrm{Bild}\,\mathbf{D}^{\mathrm{opt}}$ als Linearkombination der Vektoren in $\mathbf{D}^{\mathrm{opt}}$ darstellbar ist, i.e. $\mathbf{x}'' = \mathbf{D}^{\mathrm{opt}}\mathbf{m}$, folgt die Aussage Gl. 3.23 dann aus Gl. 3.24: $\mathbf{D}^{\mathrm{opt}}\mathbf{Kx}'' = \mathbf{D}^{\mathrm{opt}}\mathbf{KD}^{\mathrm{opt}}\mathbf{m} = \mathbf{D}^{\mathrm{opt}}\mathbf{m} = \mathbf{x}''$.

Wir haben diese Aussagen gebracht, um darauf hinzuweisen, dass $\mathbf{P} := \mathbf{D}^{\mathrm{opt}}\mathbf{K}$ keine “echte”, i.e. orthogonale, Projektion sein muss, sondern nur eine “nicht orthogonale” Projektion ist. Während Projektionen stets idempotent sind, ist eine orthogonale Projektion, i.e. eine Projektion im engeren Sinne, zusätzlich symmetrisch. Die Idempotenz von $\mathbf{P}$ folgt aus $\mathbf{KD}^{\mathrm{opt}} = \mathbf{I}$ (Gl. 3.24):

$$\mathbf{PP} = \mathbf{D}^{\mathrm{opt}}\mathbf{KD}^{\mathrm{opt}}\mathbf{K} = \mathbf{K}\,\mathbf{I}\,\mathbf{D}^{\mathrm{opt}} = \mathbf{P}. \tag{3.25}$$

Da jedoch Bild $\mathbf{K}^T \neq$ Bild $\mathbf{D}^{\text{opt}}$ ist[4] (siehe oben), ist $\mathbf{P}$ im Allgemeinen nicht symmetrisch. Das besagt nämlich die Äquivalenz (Beweis siehe Abschnitt A.1.2)

$$\text{Bild}\,\mathbf{K}^T = \text{Bild}\,\mathbf{D}^{\text{opt}} \quad \Longleftrightarrow \quad \mathbf{P} = \mathbf{P}^T. \tag{3.26}$$

Aus dem Nachweis im Anhang können wir genau angeben, wann ein "echte" Projektion vorliegt (Abschnitt A.1.2). Denn dort wird gezeigt, dass $\mathbf{P} = \mathbf{P}^T$ genau dann gilt, wenn die Zeilenvektoren aus $\mathbf{K}$ durch Linearkombinationen von genau m Eigenvektoren $\mathbf{w}_i$ von $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ darstellbar sind; die Zeilenvektoren von $\mathbf{K}$ spannen dann einen "Eigenraum" von $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ auf[5]. Wir haben also für eine Indexmenge $\mathcal{I} \subset \mathbb{N}_n$ mit $|\mathcal{I}| = m$

$$\mathbf{P} = \mathbf{P}^T \quad \Longleftrightarrow \quad \text{Bild}\,\mathbf{K}^T = \langle \mathbf{w}_i | i \in \mathcal{I} \subset \mathbb{N}_n \rangle \tag{3.28}$$

Wir schreiben $\langle \mathbf{w}_i | i \in \mathcal{I} \rangle$ für den von den Eigenvektoren $\mathbf{w}_i$, $i \in \mathcal{I}$ von $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$, aufgespannten Unterraum in $\mathbb{R}^n$ (Eigenraum).

Wenn der Kodierer $\mathbf{K}^T = (\mathbf{k}_1, \ldots, \mathbf{k}_m)$ einen Eigenraum von $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ aufspannt und die Gewichtsvektoren zusätzlich orthogonale zueinander sind, hängt der zugehörige optimale Dekodierer $\mathbf{D}^{\text{opt}} = (\mathbf{d}_1, \ldots, \mathbf{d}_m)$ nicht mehr von der Kovarianzmatrix $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ ab. Denn es gilt (Beweis siehe Abschnitt A.1.3):

$$\text{Bild}\,\mathbf{K}^T = \text{Bild}\,\mathbf{D}^{\text{opt}} \text{ und } \mathbf{k}_i^T \mathbf{k}_j = 0,\ i \neq j \quad \Longrightarrow \quad \mathbf{d}_j = \frac{1}{|\mathbf{k}_j|^2} \mathbf{k}_j,\ j \in \mathbb{N}_m \tag{3.29}$$

Der optimale Dekodierer zu einem solchen $\mathbf{K}$ wird nicht mehr von der Kovarianzmatrix $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ beeinflusst. Jetzt sind nicht nur die Bilder der Kodierungs- und Dekodierungsmatrix identisch, die einzelnen Gewichtsvektoren von Kodierer und Dekodierer zeigen sogar in die gleiche Richtung.

[4] Der kodierte Raum stimmt also i.A. nicht mit dem dekodierten Raum überein. Man kann sich das veranschaulichen. Es ist etwa so, als ob ein Schatten eines Objekts auf ein zu den (parallelen) Lichtstrahlen schiefes Blatt Papier geworfen wird. Malt man den Umriss des Schattens auf das Papier, stimmt die gemalte Linie nicht mit dem Umriss des Objekts überein (nicht orthogonale Projektion). Würde man jedoch das Papier senkrecht zum Licht halten, sind Schatten und Umriss des Objekts identisch (orthogonale Projektion).

[5] Der mathematische Grund, warum der Dekodierer das Bild nicht verändert, wenn der Kodierer einen Eigenraum aufspannt, liegt am folgenden allgemeineren Sachverhalt: Vektoren $\mathbf{x}^\perp$, die senkrecht zu Eigenvektoren $\mathbf{w}_j$, $j \in \mathcal{J} \subset \mathbb{N}_n$, einer Matrix $\mathbf{C}$ stehen, sind auch nach der Matrixmultiplikation, $\mathbf{C}\mathbf{x}^\perp$, orthogonal zu diesen Eigenvektoren $\mathbf{w}_j$. Die Vektoren $\mathbf{x}^\perp$ können also durch $\mathbf{C}$ nicht aus dem Unterraum $\mathbb{R}^n \setminus \langle \mathbf{w}_j | j \in \mathcal{J} \rangle$ "herausgedreht" werden. Auf Kovarianzmatrizen bezogen bedeutet diese Aussage, dass Vektoren $\mathbf{x}^\perp$ keine Kovarianz mit den Richtungen $\mathbf{w}_j$ besitzen. Ein Dekodierer würde im Mittel einen größeren Fehler machen, wenn die Rekonstruktion von $\mathbf{x}^\perp$ Varianz in Richtungen $\mathbf{w}_j$ aufweisen würde, als ein Dekodierer, der die Orthogonalität zu diesen Richtungen beibehält (vergl. auch mit Formel für optimalen Dekodierer Gl. 3.21). Man kann diesen Sachverhalt zusammenfassen mit der Aussage ($\mathcal{I} \subset \mathbb{N}_n$)

$$\text{Bild}\,\mathbf{K}^T \subset \langle \mathbf{w}_i | i \in \mathcal{I} \rangle \quad \Longleftrightarrow \quad \text{Bild} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T \subset \langle \mathbf{w}_i | i \in \mathcal{I} \rangle \tag{3.27}$$

Rekonstruktionsfehler für optimalen Dekodierer

Wir wollen jetzt den Rekonstruktionsfehler an der Stelle des optimalen Dekodierers bestimmen. Es sei

$$\begin{aligned} \varepsilon_{\mathbf{x}}(\mathbf{K}) &:= \varepsilon(\mathbf{K}, \mathbf{D}^{\text{opt}}(\mathbf{K}, P_{\mathbf{x}}), P_{\mathbf{x}}) && (3.30)\\ \text{[Gl. 3.15]} \qquad &= \operatorname{Sp}\left(\hat{\mathbf{P}} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \hat{\mathbf{P}}^T\right), && (3.31) \end{aligned}$$

wobei $\hat{\mathbf{P}} := \hat{\mathbf{Q}}^{\text{opt}} = \left(\mathbf{I} - \mathbf{D}^{\text{opt}}\mathbf{K}\right)$. Dies lässt sich weiter umformen. Wegen der Symmetrie der Matrix $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \hat{\mathbf{P}}^T$,

$$\begin{aligned} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \hat{\mathbf{P}}^T &= \left\langle \mathbf{x}\mathbf{x}^T \right\rangle - \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T \mathbf{D}^{\text{opt}\,T} \\ \text{[Gl. 3.21]} \qquad &= \left\langle \mathbf{x}\mathbf{x}^T \right\rangle - \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T \left(\mathbf{K} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T\right)^{-1} \mathbf{K} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \\ \text{[Gl. 3.21]} \qquad &= \left\langle \mathbf{x}\mathbf{x}^T \right\rangle - \mathbf{D}^{\text{opt}}\mathbf{K} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \\ &= \hat{\mathbf{P}} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle, && (3.32) \end{aligned}$$

wird aus Gl. 3.31

$$\varepsilon_{\mathbf{x}} = \operatorname{Sp}\left(\hat{\mathbf{P}}\hat{\mathbf{P}} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle\right). \tag{3.33}$$

$\hat{\mathbf{P}}$ ist die zu $\mathbf{P}$ komplementäre (i.A. nicht orthogonale) Projektion und ist daher ebenfalls idempotent:

$$\hat{\mathbf{P}}\hat{\mathbf{P}} = \mathbf{I} - 2\mathbf{P} + \mathbf{P}\mathbf{P} = \hat{\mathbf{P}}. \tag{3.34}$$

Einsetzen in Gl. 3.33 ergibt schließlich

$$\begin{aligned} \varepsilon_{\mathbf{x}} &= \operatorname{Sp}\left(\hat{\mathbf{P}} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle\right) && (3.35)\\ &= \operatorname{Sp}\left(\left\langle \mathbf{x}\mathbf{x}^T \right\rangle - \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T \left(\mathbf{K} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T\right)^{-1} \mathbf{K} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle\right), && (3.36) \end{aligned}$$

Das ist der optimale Rekonstruktionsfehler, d.h. der Rekonstruktionsfehler bei Verwendung des optimalen Dekodierers $\mathbf{D}^{\text{opt}}(\mathbf{K}, P_{\mathbf{x}})$ für einen vorgegebenen Kodierer $\mathbf{K}$.

Ist $\mathbf{P}_* = \mathbf{D}^{\text{opt}}(\mathbf{K}_*, P_{\mathbf{x}})\mathbf{K}_*$ eine "echte" Projektion, d.h. $\mathbf{P}_*^T = \mathbf{P}_*$, lässt sich der Rekonstruktionsfehler auf eine einfache Form bringen. Nach Gl. 3.28 ist dies nämlich gleichbedeutend damit, dass die Zeilenvektoren des Kodierers einen Eigenraum von $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle = \sum_{j=1}^n \lambda_j \mathbf{w}_j \mathbf{w}_j^T$ aufspannen, i.e. $\operatorname{Bild} \mathbf{K}_*^T = \langle \mathbf{w}_i | i \in \mathcal{I} \subset \mathbb{N}_n \rangle$. Für $\mathbf{P}_*$ folgt[6]

$$\mathbf{P}_*\mathbf{w}_i = \begin{cases} \mathbf{w}_i, & i \in \mathcal{I} \\ \mathbf{0}, & \text{sonst} \end{cases} \tag{3.37}$$

[6] Denn wegen $\operatorname{Bild} \mathbf{K}_*^T = \operatorname{Bild} \mathbf{D}_*^{\text{opt}}$ (siehe Abs. A.1.2) ist $\operatorname{Bild} \mathbf{P}_* = \operatorname{Bild} \mathbf{K}_*^T$, d.h. es gilt $\mathbf{P}_*\mathbf{w}_i = \mathbf{w}_i$, $i \in \mathcal{I}$. Aus $\operatorname{Kern} \mathbf{P}_* = \mathbb{R}^n \setminus \operatorname{Bild} \mathbf{P}_*$ folgt dann $\mathbf{P}_*\mathbf{w}_j = \mathbf{0}$, $j \in \mathcal{I} \setminus \mathbb{N}_n$

Dadurch wird der Rekonstruktionsfehler zu

$$
\begin{aligned}
&\text{[Gl. 3.35]} && \varepsilon_{\mathbf{x}}^{*} = \operatorname{Sp}\left(\left(\mathbf{I}-\mathbf{P}_{*}\right)\left\langle \mathbf{x}\mathbf{x}^{T}\right\rangle\right) && (3.38)\\
& && = \operatorname{Sp}\left\langle \mathbf{x}\mathbf{x}^{T}\right\rangle - \operatorname{Sp}\left(\sum_{j=1}^{n}\lambda_j \mathbf{P}_{*}\mathbf{w}_j\mathbf{w}_j^{T}\right) && (3.39)\\
&\text{[Gl. 3.37]} && = \operatorname{Sp}\left\langle \mathbf{x}\mathbf{x}^{T}\right\rangle - \operatorname{Sp}\left(\sum_{i\in\mathcal{I}}\lambda_i \mathbf{w}_i\mathbf{w}_i^{T}\right) && (3.40)\\
&[\operatorname{Sp}\left(\mathbf{w}_i\mathbf{w}_i^{T}\right)=1] && = \operatorname{Sp}\left\langle \mathbf{x}\mathbf{x}^{T}\right\rangle - \sum_{i\in\mathcal{I}}\lambda_i && (3.41)\\
&\text{[Gl. 3.5]} && = \sum_{j\in\mathbb{N}_n\setminus\mathcal{I}}\lambda_j && (3.42)
\end{aligned}
$$

Das ist die Summe der Eigenwerte, die zu den Eigenvektoren gehören, welche nicht im Bildraum des Kodierers $\mathbf{K}_{*}$ liegen. Die Herleitung von Gl. 3.42 zeigt außerdem, dass der optimale Rekonstruktionsfehler für zwei Kodierer, welche den gleichen Eigenraum aufspannen, identisch ist.

3.2.5 Gewichtsvektoren des Kodierers

Wir wollen nun überlegen, wie ein Kodierer $\mathbf{K}^{\text{opt}}(P_{\mathbf{x}})$ gestaltet sein muss, um

$$\mathbf{K}^{\text{opt}} = \min_{\mathbf{K}} \varepsilon_{\mathbf{x}}(\mathbf{K}) \tag{3.43}$$

unter den Bedingung $\mathbf{K}^{\text{opt}} \in \mathbb{R}^{(m,n)}$ und $m = \operatorname{Rg}\mathbf{K}^{\text{opt}}$ zu erfüllen. Wir gehen also davon aus, dass die Minimierung bei Verwendung des optimalen Dekodierer (Gl. 3.21) geschieht.

Wir benutzen zur Minimierung die Gleichung 3.31 des Rekonstruktionsfehlers für optimale Dekodierer $\varepsilon_{\mathbf{x}}$. Ist $\left\langle \mathbf{x}\mathbf{x}^{T}\right\rangle = \sum_{i=1}^{n}\lambda_i\mathbf{w}_i\mathbf{w}_i^{T}$ die Spektraldarstellung der Kovarianzmatrix , erhält man mit $\hat{\mathbf{P}} = \mathbf{I} - \mathbf{D}^{\text{opt}}\mathbf{K}$

$$
\begin{aligned}
&\text{[Gl. 3.31]} && \varepsilon_{\mathbf{x}} = \operatorname{Sp}\left(\hat{\mathbf{P}}\left\langle \mathbf{x}\mathbf{x}^{T}\right\rangle\hat{\mathbf{P}}^{T}\right) && (3.44)\\
&[\operatorname{Sp}(\mathbf{AB}) = \operatorname{Sp}(\mathbf{BA})] && = \sum_{i=1}^{n}\lambda_i\left(\hat{\mathbf{P}}\mathbf{w}_i\right)^{T}\hat{\mathbf{P}}\mathbf{w}_i && (3.45)
\end{aligned}
$$

Wir wollen zunächst $\mathbf{w}_i^{T}\hat{\mathbf{P}}^{T}\hat{\mathbf{P}}\mathbf{w}_i = |\hat{\mathbf{P}}\mathbf{w}_i|^2$ näher bestimmen. Sei dazu $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{n-m})$ eine orthonormale Basis von Kern $\mathbf{K}$, $\mathbf{K} = (\mathbf{k}_1, \ldots, \mathbf{k}_m)^{T}$. Da die Zeilenvektoren von $\mathbf{H}$ und $\mathbf{K}^{T}$ zusammen genommen eine Basis von $\mathbb{R}^{n}$ sind, lassen sich die Eigenvektoren $\mathbf{w}_i$ von $\left\langle \mathbf{x}\mathbf{x}^{T}\right\rangle$ als deren Linearkombination darstellen:

$$\mathbf{w}_i = \sum_{j=1}^{m}\mu_{ji}\mathbf{k}_j + \sum_{j=1}^{n-m}\nu_{ji}\mathbf{h}_j \tag{3.46}$$

Wegen der Orthonormalität der Vektoren in $\mathbf{H}$ ist $\nu_{ji} = \mathbf{h}_j^T \mathbf{w}_i$. Da die Vektoren $\mathbf{h}_j$ nach Definition im Kern von $\mathbf{K}$ liegen, also $\mathbf{K}\mathbf{h}_j = \mathbf{0}$, ist

$$\hat{\mathbf{P}}\mathbf{h}_j = (\mathbf{I} - \mathbf{D}^{\mathrm{opt}}\mathbf{K})\mathbf{h}_j = \mathbf{h}_j \tag{3.47}$$

und es lässt sich mit Gl. 3.46 ausrechnen

$$[\mathbf{k}_j^T \mathbf{h}_i = \mathbf{0}] \qquad \mathbf{w}_i^T \hat{\mathbf{P}}^T \hat{\mathbf{P}} \mathbf{w}_i = \left| \sum_{j=1}^{m} \mu_{ji} \hat{\mathbf{P}} \mathbf{k}_j \right|^2 + \left| \sum_{j=1}^{n-m} (\mathbf{h}_j^T \mathbf{w}_i) \mathbf{h}_j \right|^2 \tag{3.48}$$

$$=: \quad \rho_i \quad + \quad \sigma_i \tag{3.49}$$

Wir können zunächst zusammenfassen, indem wir Gl. 3.49 in Gl. 3.45 einsetzen:

$$\varepsilon_{\mathbf{x}} = \sum_{i=1}^{n} \lambda_i \; (\rho_i + \sigma_i) \tag{3.50}$$

Wir wollen jetzt zeigen, dass an die Zahlen ρ_i und σ_i Nebenbedingungen geknüpft sind. Wenn wir diese kennen, können wir den prinzipiell möglichen, minimalen Wert des Rekonstruktionsfehlers angeben. Danach werden wir zeigen, dass dieser Wert für bestimmte Kodierer $\mathbf{K}$ tatsächlich angenommen werden kann.

Man sieht zunächst aus der Definition (Gl. 3.49), dass $\rho_i \geq 0$ und $\sigma_i \geq 0$ sind. Die Zahlen σ_i lassen sich umformen zu

$$\sigma_i = \left| \sum_{j=1}^{n-m} (\mathbf{h}_j^T \mathbf{w}_i) \mathbf{h}_j \right|^2 \tag{3.51}$$

$$[\mathbf{h}_i^T \mathbf{h}_j = \delta_{ij}] \qquad = \sum_{j=1}^{n-m} (\mathbf{h}_j^T \mathbf{w}_i)^2. \tag{3.52}$$

Wir bemerken, dass die Summe $\sum_{i=1}^{n} \sigma_i$ konstant ist, d.h. nicht von der Lage der Gewichtsvektoren des Kodierers $\mathbf{K}$ abhängt:

$$\begin{aligned} \sum_{i=1}^{n} \sigma_i &= \sum_{j=1}^{n-m} \sum_{i=1}^{n} (\mathbf{h}_j^T \mathbf{w}_i)^2 \\ [\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)] \qquad &= \sum_{j=1}^{n-m} \mathbf{h}_j^T \mathbf{W} \mathbf{W}^T \mathbf{h}_j \\ [\mathbf{W}\mathbf{W}^T = \mathbf{I},\ \mathbf{h}_j^T \mathbf{h}_j = 1] \qquad &= n - m \end{aligned} \tag{3.53}$$

Außerdem folgt aus ähnlichen Gründen

$$\sigma_i = \sum_{j=1}^{n-m} (\mathbf{h}_j^T \mathbf{w}_i)^2 \tag{3.54}$$

$$\leq \sum_{j=1}^{n} (\mathbf{h}_j^T \mathbf{w}_i)^2 \tag{3.55}$$

$$= \mathbf{w}_i^T \mathbf{H}' \mathbf{H}'^T \mathbf{w}_i \tag{3.56}$$

$$= 1 \tag{3.57}$$

$\mathbf{H}' = (\mathbf{h}_1, \ldots, \mathbf{h}_n)$ sei eine Erweiterung der Basis $\mathbf{H}$ des Kerns von $\mathbf{K}$ zu einer orthonormalen Basis des $\mathbb{R}^n$.

Könnte man nun $\rho_i \geq 0$ und $0 \leq \sigma_i \leq 1$ unter Beachtung von $\sum_{i=1}^n \sigma_i = n - m$ beliebig wählen, würde der Rekonstruktionsfehler (Gl. 3.50) offensichtlich den kleinsten Wert annehmen bei

$$\sigma_i = \begin{cases} 0, & i \leq m \\ 1, & \text{sonst.} \end{cases} \tag{3.58}$$

$$\rho_i = 0, \quad i \in \mathbb{N}_n. \tag{3.59}$$

Hier haben wir geordnete Eigenwerte vorausgesetzt, $\lambda_i \geq \lambda_j$, $i > j$, und $\lambda_i > 0$. Wenn nun diese Werte für ein spezielles $\mathbf{K}^* \in \mathbb{R}^{(m,n)}$ angenommen werden, ist gezeigt, dass ein solches $\mathbf{K}^*$ den Rekonstruktionsfehler tatsächlich minimiert.

Nach Definition (Gl. 3.49) gibt die Zahl σ_i den Anteil des Eigenvektors $\mathbf{w}_i$ an, der durch die Basis des Kerns von $\mathbf{K}$ dargestellt wird. Ist also $\sigma_i = 1$, liegt der Eigenvektor $\mathbf{w}_i$ vollständig im Kern von $\mathbf{K}$, und andersherum bedeutet $\sigma_i = 0$, dass der Eigenvektor $\mathbf{w}_i$ vollständig durch eine Linearkombination der Zeilenvektoren aus $\mathbf{K}$ dargestellt werden kann.

Die Bedingung an die Zahlen σ_i für ein Minimum (Gl. 3.58) besagt daher, dass ein optimaler Kodierer $\mathbf{K}^{\text{opt}}$ den größten Eigenraum von $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ aufspannt, $\langle \mathbf{w}_j | j \in \mathbb{N}_m \rangle = \text{Bild}\,\mathbf{K}^T$. Wir wollen überprüfen, ob ein solcher Kodierer auch die Bedingung an die ρ_i (Gl. 3.59) erfüllt.

Im Anhang (Abschnitt A.1.2) zeigen wir die Äquivalenz

$$\langle \mathbf{w}_i | i \in \mathcal{I} \subset \mathbb{N}_n \rangle = \text{Bild}\,\mathbf{K}^T \quad \Longleftrightarrow \quad \text{Bild}\,\mathbf{D}^{\text{opt}} = \text{Bild}\,\mathbf{K}^T. \tag{3.60}$$

Da ein optimaler Kodierer $\mathbf{K}^{\text{opt}}$ laut Gl. 3.58 einen Eigenraum aufspannt, können wir die Äquivalenz anwenden und wir haben $\text{Bild}\,\mathbf{D}^{\text{opt}} = \text{Bild}\,\mathbf{K}^{\text{opt}T}$. Damit ist $\mathbf{k}_j \in \text{Bild}\,\mathbf{D}^{\text{opt}}$ und aus ρ_i wird wegen $\hat{\mathbf{P}}\mathbf{x} = \mathbf{0}$ für ein $\mathbf{x} \in \text{Bild}\,\mathbf{D}^{\text{opt}}$ (Gl. 3.23):

$$\begin{aligned} &[\text{Gl. } 3.49] & \rho_i &= \left| \sum_{j=1}^m \mu_{ji} \hat{\mathbf{P}} \mathbf{k}_j \right|^2 \\ &[\text{wg. } \mathbf{k}_j \in \text{Bild}\,\mathbf{K}^T = \text{Bild}\,\mathbf{D}^{\text{opt}}] & &= 0. \end{aligned} \tag{3.61}$$

Insgesamt sind die Forderungen für σ_i (Gl. 3.58) und und ρ_i (Gl. 3.59) erfüllt; ein optimales $\mathbf{K}$ ist bereits durch die Wahl der σ_i vorgegeben. Wir erhalten schließlich als Bedingung für Gl. 3.43 ein $\mathbf{K}^{\text{opt}}$ (bei fester Dimensionalität)

$$\text{Bild}\,\mathbf{K}^{\text{opt}T} = \langle \mathbf{w}_j | j \in \mathbb{N}_m \rangle \tag{3.62}$$

Ein optimaler Kodierer spannt den größten Eigenraum der Kovarianzmatrix $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ der Verteilung $P_\mathbf{x}$ auf. Der optimale Kodierer ist durch diese Bedingung offensichtlich nicht eindeutig bestimmt (im Gegensatz zum optimalen Dekodierer, vergl. Gl. 3.21).

Insbesondere sind die Zeilenvektoren eines optimalen Kodierers nicht notwendigerweise normiert oder orthogonal zueinander.

Die Menge der Kodierer $\mathbf{K} \in \mathbb{R}^{(m,n)}$, die zu einer gegebenen Verteilung $P_\mathbf{x}$ optimal sind, für die also Gl. 3.62 gilt, bezeichnen wir $\mathcal{K}^{\text{opt}}_{(m,n)}(P_\mathbf{x})$. Ist eine konkrete Kovarianzmatrix $\mathbf{C}$ vorgegeben, werden wir die Tatsache, dass die m Zeilenvektoren des Kodierers $\mathbf{K} \in \mathbb{R}^{(m,n)}$ den größten m-dimensionalen Eigenraum von $\mathbf{C}$ erzeugen, durch die übersichtlichere Notation

$$\mathbf{K} \sim \mathrm{C}_m \tag{3.63}$$

zum Ausdruck bringen. Man beachte, dass kein Fettdruck verwendet wird, um Verwechslungen zu vermeiden.

Folgerungen

Sind Kodierer $\mathbf{K}^{\text{opt}} \sim \mathrm{C}_m$ und zugehöriger Dekodierer $\mathbf{D}^{\text{opt}} = \mathbf{D}^{\text{opt}}(\mathbf{K}^{\text{opt}}, \mathbf{C})$ optimal für eine Verteilung $P_\mathbf{x}$ mit Kovarianzmatrix $\mathbf{C}$, ist die Projektion $\mathbf{P} = \mathbf{D}^{\text{opt}}\mathbf{K}^{\text{opt}}$ symmetrisch (Gl. 3.28 und Gl. 3.25). Der Rekonstruktionsfehler für $\mathbf{K}^{\text{opt}}$ und $\mathbf{D}^{\text{opt}}$ ist gegeben durch (siehe Gl. 3.50)

$$\varepsilon_\mathbf{x}(\mathbf{K}^{\text{opt}}) = \sum_{i=m+1}^{n} \lambda_i, \tag{3.64}$$

das ist die Summe der $n - m$ kleinsten Eigenwerte der Kovarianzmatrix $\mathbf{C}$. Die Eigenwerte sind hier der Größe nach geordnet.

Sind zwei optimale Kodierer mit unterschiedlicher Dimension der verborgenen Schicht $\mathbf{y}$ vorgegeben, also $\mathbf{K}^{\text{opt}}_m \sim \mathrm{C}_m$ und $\mathbf{K}^{\text{opt}}_{m'} \sim \mathrm{C}_{m'}$, folgt, da alle Eigenwert als positiv vorausgesetzt sind, aus Gl. 3.64 sofort

$$\varepsilon_\mathbf{x}(\mathbf{K}^{\text{opt}}_m) < \varepsilon_\mathbf{x}(\mathbf{K}^{\text{opt}}_{m'}), \quad \text{falls } m > m' \tag{3.65}$$

Der minimale Rekonstruktionsfehler ist also stets kleiner, wenn man mehr Neuronen der verborgenen Schicht benutzt. Man kann die Verminderung des Fehlerwerts genau angeben. Ist nämlich $m > m'$, folgt wegen der Optimalität beider Kodierer aus Gl. 3.64

$$\varepsilon_\mathbf{x}(\mathbf{K}^{\text{opt}}_m) - \varepsilon_\mathbf{x}(\mathbf{K}^{\text{opt}}_{m'}) = \sum_{j=m'+1}^{m} \lambda_j \tag{3.66}$$

Wie stark $\varepsilon_\mathbf{x}$ sinkt, hängt also von den Eigenwerten, i.e. der Varianz, der zusätzlich kodierten Richtungen ab. Im Grenzfall $m = 0$ findet keine Kodierung statt und es ist $\varepsilon_\mathbf{x} = \mathrm{Sp}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle = \left\langle |\mathbf{x}|^2 \right\rangle$. Im anderen Grenzfall, wenn $m = n$, ist $\mathrm{Bild}\,\mathbf{K}^T = \mathbb{R}^n$ und die Kodierung verlustfrei, $\varepsilon_\mathbf{x} = 0$.

3.2.6 Zusammenfassung

Wir haben in den vorangegangenen Abschnitten untersucht, wie die Kodierung und Dekodierung gestalten sein müssen, damit Eindrücke einer Umgebung optimal aufbereitet werden können. Konkret haben wir bei vorgegebener Verteilung $P_{\mathbf{x}}$, mit $\mathbf{x} \in \mathbb{R}^n$, und der Dimension der verborgenen Schicht $\mathbf{y} \in \mathbb{R}^m$ versucht den Rekonstruktionsfehler $\varepsilon(\mathbf{K}, \mathbf{D}, P_{\mathbf{x}}) = \left\langle |\mathbf{x} - \mathbf{DKx}|^2 \right\rangle$ zu minimieren. Wir haben dazu $\mathbf{K} \in \mathbb{R}^{(m,n)}$, und $\mathbf{D} \in \mathbb{R}^{(n,m)}$ betrachtet. Zunächst ließ sich ε umformulieren in

$$\varepsilon(\mathbf{K}, \mathbf{D}, P_{\mathbf{x}}) = \mathrm{Sp}\left(\hat{\mathbf{Q}} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \hat{\mathbf{Q}}^T\right), \tag{3.67}$$

mit $\hat{\mathbf{Q}} := (\mathbf{I} - \mathbf{DK})$. Somit hängt ε nur über die Kovarianzmatrix $\mathbf{C} := \left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ von Verteilung $P_{\mathbf{x}}$ ab, d.h. es ist $\varepsilon = \varepsilon(\mathbf{K}, \mathbf{D}, \mathbf{C})$.

Ist $\mathbf{C}$ invertierbar, existiert genau eine Dekodierungsmatrix $\mathbf{D}^{\mathrm{opt}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}, \mathbf{C})$, mit $\mathbf{D}^{\mathrm{opt}} \in \mathbb{R}^{(n,m)}$, welche den Rekonstruktionsfehler ε (für gegebenes $\mathbf{K}$) minimiert, nämlich

$$\mathbf{D}^{\mathrm{opt}}(\mathbf{K}, \mathbf{C}) = \mathbf{C}\mathbf{K}^T \left(\mathbf{K}\mathbf{C}\mathbf{K}^T\right)^{-1} \tag{3.68}$$

Hat der Kodierer $\mathbf{K}$ zueinander orthogonale Gewichtsvektoren und spannt zusätzlich einen Eigenraum von $\mathbf{C}$ auf, hängt der optimale Dekodierer nicht mehr von der Kovarianzmatrix $\mathbf{C}$ ab.

Der Rekonstruktionsfehler an der Stelle des optimalen Dekodierers ist

$$\varepsilon_{\mathbf{C}}(\mathbf{K}) = \mathrm{Sp}\left(\hat{\mathbf{P}}\mathbf{C}\right). \tag{3.69}$$

Die idempotente Matrix $\hat{\mathbf{P}} := \mathbf{I} - \mathbf{D}^{\mathrm{opt}}\mathbf{K}$ ist i.A. nicht symmetrisch.

Der minimale Fehlerwert der Rekonstruktionsfehler $\varepsilon_{\mathbf{C}}(\mathbf{K}^{\mathrm{opt}})$ (bei gegebener Dimensionalität der verborgenen Schicht m) wird durch Nutzung eines optimalen Kodierer $\mathbf{K}^{\mathrm{opt}} \sim \mathrm{C}_m$ erreicht. Dieser spannt den größten m-dimensionalen Eigenraum der Kovarianzmatrix $\mathbf{C}$ auf. Das Minimum des Fehlers ist die Summe der $n - m$ kleinsten Eigenwerte von $\mathbf{C}$.

3.2.7 Beispiel

Es ist illustrativ, die Ergebnisse der letzten Abschnitte an einem niedrigdimensionalem Beispiel anzuwenden. Es sei $n = 2$ und die Kovarianzmatrix der Verteilung $P_{\mathbf{x}}$

$$\mathbf{C} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{3.70}$$

mit $\lambda_1 \geq \lambda_2 > 0$. Dadurch ist eine mittelwertfreie multivariate Normalverteilung $P_{\mathbf{x}}$ definiert, die Abbildung 3.5 A veranschaulicht. Der Kodierer sei ein beliebiger normierter Vektor, habe also die Gestalt

$$\mathbf{K}(\varphi) = \begin{pmatrix} \sin\varphi & \cos\varphi \end{pmatrix}. \tag{3.71}$$

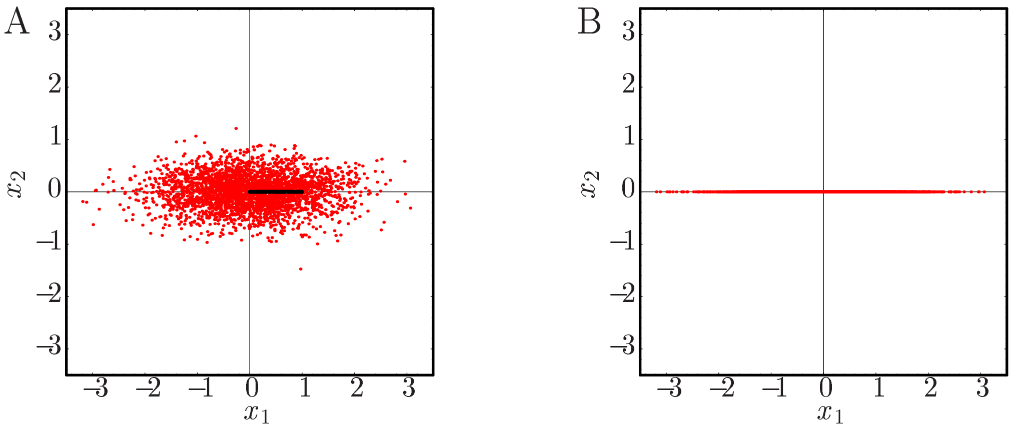

Abbildung 3.5: *A*: Multivariate Normalverteilung mit Kovarianzmatrix $\mathbf{C} = \mathrm{diag}\,(0.9, 0.1)$. Eingezeichnet sind optimaler Kodierer und Dekodierer (sie liegen übereinander, durchgezogene Linie). *B*: Rekonstruktion der Muster aus Abb. A. Die Datenpunkte werden zunächst auf den durch den Kodierer vorgegebenen Raum (x_1-Achse) in den $\mathbb{R}^1$ projiziert und mittels Dekodierer im Ursprungsraum $\mathbb{R}^2$ dargestellt. Am Optimum liegen Dekodierer und Kodierer übereinander, die rekonstruierten Muster befinden sich daher entlang der x_1-Achse.

Damit ist $m = 1$. Der optimale Dekodierer lässt sich nach Gl. 3.21 berechnen

$$\begin{aligned} \mathbf{D}^{\mathrm{opt}}(\varphi) &= \mathbf{CK}(\varphi)^T \left(\mathbf{K}(\varphi)\mathbf{CK}(\varphi)^T\right)^{-1} \\ &= \frac{1}{\lambda_1 \sin^2\varphi + \lambda_2 \cos^2\varphi} \begin{pmatrix} \lambda_1 \sin\varphi \\ \lambda_2 \cos\varphi \end{pmatrix} \end{aligned} \tag{3.72}$$

und der Rekonstruktionsfehler unter Anwendung des optimalen Dekodierers schreibt sich nach Gl. 3.35

$$\varepsilon_{\mathbf{C}}(\varphi) = \lambda_1 + \lambda_2 - \frac{\lambda_1^2 \sin^2\varphi + \lambda_2^2 \cos^2\varphi}{\lambda_1 \sin^2\varphi + \lambda_2 \cos^2\varphi}. \tag{3.73}$$

Eine kurze Rechnung zeigt, dass die Extrema von Gl. 3.73 an den Stellen $\varphi \in \{z\frac{\pi}{2} | z \in \mathbb{Z}\}$ liegen. Es ist $\varepsilon_{\mathbf{C}}(z\pi) = \lambda_1$ und $\varepsilon_{\mathbf{C}}(z\pi + \frac{\pi}{2}) = \lambda_2$. Am Minimum des Fehlers $z\pi + \frac{\pi}{2}$ gilt $\mathbf{K}^{\mathrm{opt}} = \pm \begin{pmatrix} 1 & 0 \end{pmatrix}$; das ist der Eigenvektor zum größten Eigenwert λ_1 der Matrix $\mathbf{C}$, also spannt $\mathbf{K}^{\mathrm{opt}}$ tatsächlich den größten Eigenraum von $\mathbf{C}$ auf, $\mathbf{K}^{\mathrm{opt}} \sim \mathrm{C}_1$. Da $\mathbf{K}^{\mathrm{opt}}$ optimal ist, zeigt der zugehörige optimale Dekodierer in die gleiche Richtung. Bei Variation des Kodierers pendelt der Wert des Rekonstruktionsfehlers zwischen den beiden Eigenwerten als Extrema hin und her (siehe Abb. 3.6); gilt $\lambda_1 = \lambda_2$, so ist der Fehler konstant. Abb. 3.5 B zeigt die Rekonstruktion von Mustern der Verteilung $P_{\mathbf{x}}$ bei Verwendung der optimalen Kodierung.

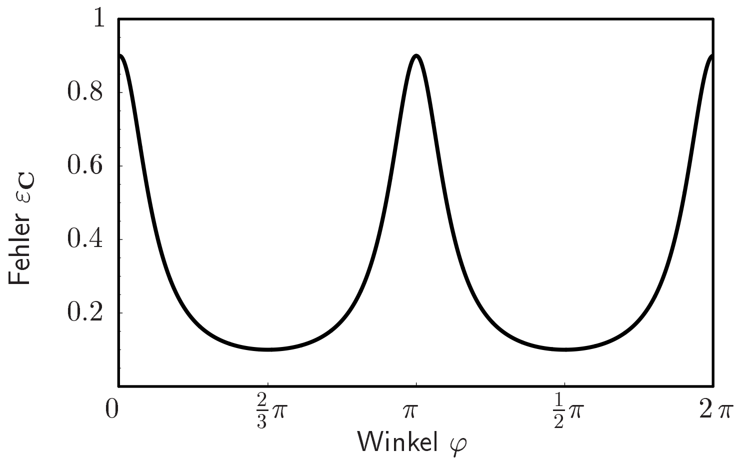

Abbildung 3.6: Der Fehlerwert $\varepsilon_{\mathbf{C}}(\mathbf{K}(\varphi))$ (Gl. 3.73) für eine Drehung des Kodierer in der Ebene (festgelegt durch φ). Er pendelt zwischen den Eigenwerten $\lambda_1 = 0.9$ und $\lambda_2 = 0.1$ als Extrema.

Kapitel 4

Experiment

4.1 Einführung

Nachdem wir uns im vorangegangenen Kapitel mit der Optimierung von Kodierer und Dekodierer an Eindrücke mit gegebener Statistik befasst haben, wollen wir nun den Adaptationsprozess des Kodes an Änderungen der mittleren Aktivitätsverteilung neocorticaler Muster untersuchen. Eine solche Änderung könnte aus dem Zuwachs an Erfahrungen oder aus dem Leben in einer komplexeren Umwelt resultieren. Um eine solche Änderungen sprachlich kürzer benennen zu können, sagen wir symbolisch, dass ein Tier in eine andere Umgebung wechselt, in der es neuen Eindrücken, d.h. einer veränderten Statistik der Muster, ausgesetzt ist (vergl. auch Abschnitt 3.1.3).

Kommt ein Tier nun in eine unbekannte Umgebung, muss die Kodierung reagieren, um der neuen Situation gerecht zu werden. Wie haben in Abschnitt 3.1.3 motiviert, dass es, in Anbetracht der Funktion des Hippokampus als Speicher, sinnvoll sein könnte, nur neugebildete Neurone an die unbekannten Eindrücke adaptieren zu lassen, während die übrigen Neurone den Kode bekannter Eindrücke konservieren, i.e. ihre Gewichte stabilisieren. Um zu testen, inwiefern eine derartige Adaptationsstrategie Erfolg verspricht, entwerfen wir ein "Experiment", das den Adaptationsprozess schematisiert und vergleichen sie mit alternativen Möglichkeiten den Kodierer an die neue Umgebung anzupassen. Insbesondere sind wir natürlich daran interessiert, ob sich die Strategie, neue Neurone zu verwenden, gegenüber einer "gewöhnlichen" plastischen Anpassung des Kodes ohne Neurogenese behauptet.

Das Experiment besteht im wesentlichen aus einem Umgebungswechsel eines Tieres aus bekannter in eine unbekannte Umgebung. Dort adaptiert es an die Statistik neuer Eindrücke. Nach dem Adaptationsprozess wird überprüft, ob die Rekonstruktion der Eindrücke der ersten Umgebung "vergessen" wurde. Diese Vorgänge werden in drei Teile gegliedert (Abb. 4.1):

(a) Zunächst befinde sich das Tier in bekannter Umgebung I, in der bestimmte Eindrücke, i.e. Muster einer Verteilung $\mathbf{a} \in P_\mathbf{a}$ (symbolisiert durch Vielecke in Abb. 4.1), vorherrschen.

(b) Das Tier kommt in eine unbekannte Umgebung II, in der sich die Eindrücke verändert haben, d.h. die Muster sind nach einer anderen Statistik verteilt, $\mathbf{b} \in P_{\mathbf{b}}$ (Kreise in Abb. 4.1). Das Tier adaptiere an die neue Situation.

(c) Abschließend werde die Maus mit Eindrücken aus Umgebung I konfrontiert,i.e. $\mathbf{a} \in P_{\mathbf{a}}$, jedoch ohne dass eine erneute Anpassung der Kodierung oder Dekodierung geschieht.

Eine Adaptationsstrategie ist im Experiment durch die Angabe von Kodierermatrizen aus Umgebung I und Umgebung II definiert, denn der Dekodierer strebe stets zum Optimum (bei gegebener Kodierung) (Gl. 3.21).

Während sich also der Dekodierer an die herrschenden Eindrücke der Umgebungen $\mathbf{a} \in P_{\mathbf{a}}$ bzw. $\mathbf{b} \in P_{\mathbf{b}}$ einstellt, gibt eine Adaptationsstrategie die Wahl der Kodierer $\mathbf{K}_{\mathrm{I}}$ und $\mathbf{K}_{\mathrm{II}}$ in Umgebung I bzw. Umgebung II vor. Wir schreiben deshalb für eine Adaptationsstrategie S zusammenfassend $\{\mathbf{K}_{\mathrm{I}}; \mathbf{K}_{\mathrm{II}}\}$. Da die Dekodierung sich in Teil (a) und (b) an die Verteilungen und die jeweiligen Kodierer der Strategien anpasst, ist sie gegeben durch die zugehörigen Optima (Gl. 3.21), d.h. $\mathbf{D}_{\mathrm{I}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{I}}, P_{\mathbf{a}})$ und $\mathbf{D}_{\mathrm{II}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{II}}, P_{\mathbf{b}})$ in Umgebung I bzw. Umgebung II. Teil (c) des Experimentes bewertet die Fähigkeit einer Strategie, frühere Muster trotz Adaptation effizient zu rekonstruieren. Die Kriterien zur Auswertung des Experimentes werden im Abschnitt 4.3 ausgearbeitet. Wir wollen jedoch zunächst Adaptationsstrategien mathematisch festlegen.

4.2 Mathematische Formulierung der Strategien

4.2.1 Neue Neurone im Adaptationsprozess

Die einfachste Möglichkeit Neurogenese im Modell zu realisieren, ist die Erhöhung der Dimensionalität der verborgenen Schicht $\mathbf{y}$ (vergl. Abb. 3.2). Obwohl sich dadurch sowohl die Zeilenanzahl des Kodierers $\mathbf{K}$ als auch die Spaltenanzahl des Dekodierers $\mathbf{D}$ erhöht, werden wir diese vereinfachende Sichtweise der Neurogenese verwenden. Seien dann $\mathbf{K}_{\mathrm{I}} = (\mathbf{k}_1, \dots, \mathbf{k}_{n_{\mathrm{I}}})^T$ und $\mathbf{D}_{\mathrm{I}} = (\mathbf{d}_1, \dots, \mathbf{d}_{n_{\mathrm{I}}})$ die Kodierungs- und Dekodierungsmatrix der ersten Umgebung I (Abb. 4.1). Beide Matrizen werden an $\mathbf{a}$-Muster angepasst. Der Kodierer sei also auf Verteilung $P_{\mathbf{a}}$ optimiert; er spannt daher einen Eigenraum der Kovarianzmatrix $\mathbf{A}$ auf, d.h. es gilt $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_{n_{\mathrm{I}}}$ (zur Notation vergl. Gl. 3.63).

Wenn die Maus in eine bis dato fremde Umgebung II wechselt, "passiere" Neurogenese, d.h. nun gelte $\mathbf{K}_{\mathrm{II}} = (\mathbf{k}_1, \dots, \mathbf{k}_{n_{\mathrm{I}}}, \mathbf{k}_{n_{\mathrm{I}}+1}, \dots, \mathbf{k}_{n_{\mathrm{II}}})^T$ und für den Dekodierer analog $\mathbf{D}_{\mathrm{II}} = (\mathbf{d}_1, \dots, \mathbf{d}_{n_{\mathrm{I}}}, \mathbf{d}_{n_{\mathrm{I}}+1}, \dots, \mathbf{d}_{n_{\mathrm{II}}})$. Wir können $\mathbf{K}_{\mathrm{II}}$ als Blockmatrix schreiben, indem wir die bereits in Umgebung I existenten, "alten" Neurone und die in Umgebung II durch Neurogenese hinzukommenden, "neuen" Neurone jeweils zusammenfassen, $\mathbf{K}_{\mathrm{II}}^T = \left(\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}\,T} \quad \mathbf{K}_{\mathrm{II}}^{\mathrm{neu}\,T}\right)$; analog sei $\mathbf{D}_{\mathrm{II}} = \left(\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} \quad \mathbf{D}_{\mathrm{II}}^{\mathrm{neu}}\right)$. Laut unserer Hypothese (Abschnitt 3.1.3) passen sich neu gebildete Neurone des Kodierers besser an unbekannte Eindrücke neuer Umgebungen an als ältere Neurone. Da Neurogenese hier abrupt

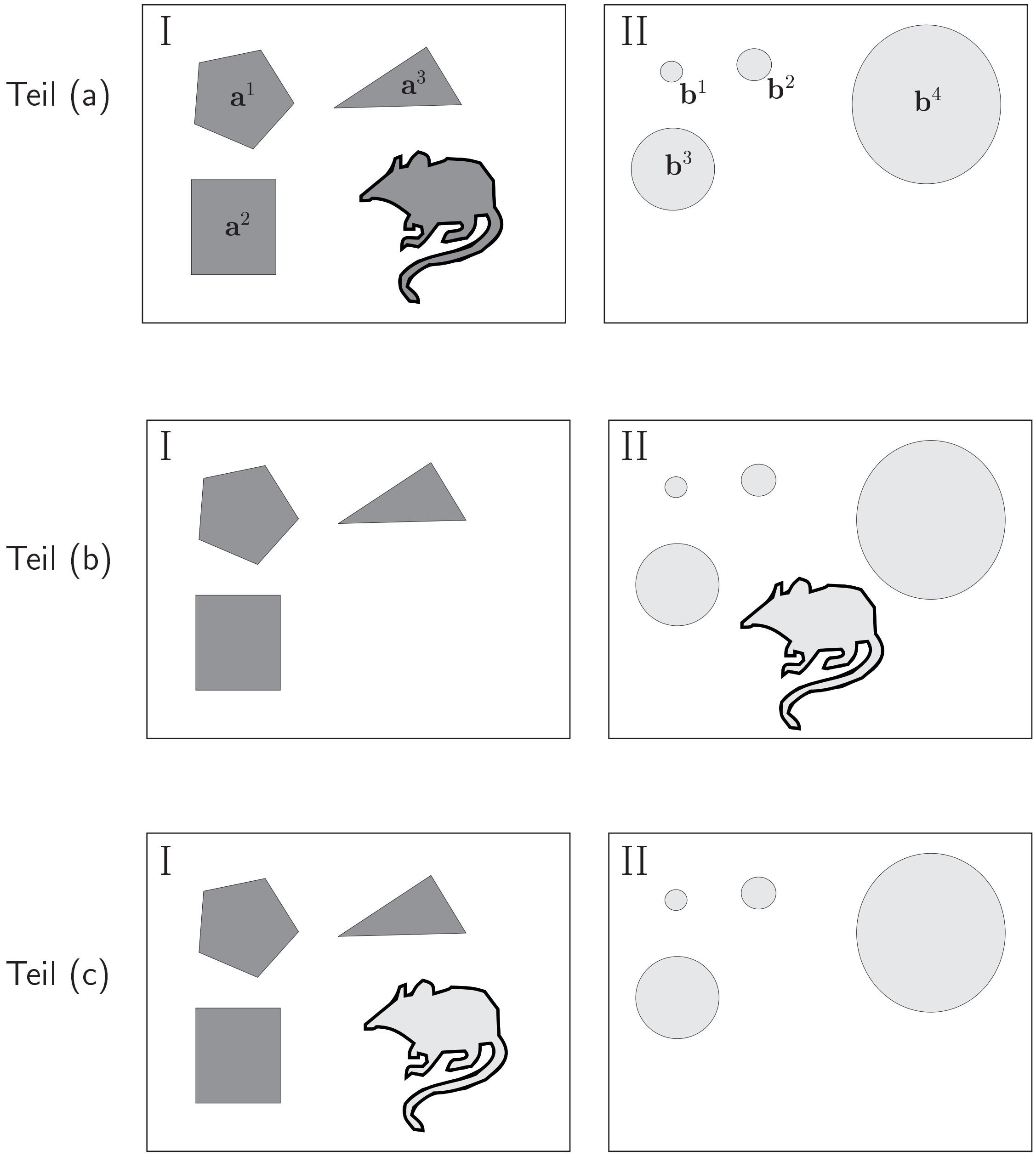

Abbildung 4.1: Experiment. Auf der linken Seite ist Umgebung I, auf der rechten Seite Umgebung II schematisch dargestellt. *Teil (a)*: Die Maus befindet sich in bekannter Umgebung I. *Teil (b)*: Die Maus ist in Umgebung II, die Statistik der Muster hat sich geändert, der Hippokampus ist mit Mustern $\mathbf{b} \in P_\mathbf{b}$ konfrontiert (Kreise). Der Hippokampus adaptiert an die veränderte Situation, Adaptation ist durch die Einfärbung der Maus symbolisiert. Eine Adaptationsstrategie gibt die Kodierung vor, während der Dekodierer optimal ist. *Teil (c)*: Aus Umgebung II wird die Maus nun in die alte Umgebung I versetzt. Sie hat jetzt keine Zeit Kodierer oder Dekodierer, die noch an Muster $\mathbf{b} \in P_\mathbf{b}$ aus Umgebung II adaptiert sind, erneut auf Muster $\mathbf{a} \in P_\mathbf{a}$ aus Umgebung I zu adaptieren. Hier wird getestet, ob der Adaptationsprozess die Rekonstruktion früher bekannter Muster beeinflusst.

geschieht, sollen sich vereinfacht die Gewichte der neuen Neurone $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}}$ an die fremden $\mathbf{b}$-Muster der Umgebung II anpassen, während die übrigen Gewichtsvektoren $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ unverändert bleiben. Die alten Neurone kodieren also weiterhin optimal für $\mathbf{a}$-Muster.

Wir nehmen an, dass neue Neurone den bestehenden Kodierer um zusätzliche orthogonale Dimensionen ergänzen. Damit die Varianz der $\mathbf{b}$-Muster möglichst gut kodiert wird, müssen die Gewichte der neuen Neurone daher die größten Hauptkomponenten von dem Teil der Verteilung $P_{\mathbf{b}}$ lernen, der senkrecht zu den Gewichten der alten Neurone $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ steht. Mathematisch heißt das, dass die Gewichte der neuen Neurone $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}}$ die größte Varianz der Projektion der $\mathbf{b}$-Muster in den Kern von $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ kodieren, i.e. $\hat{\mathbf{P}}_{\mathrm{I}}\mathbf{b} = (\mathbf{I} - \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}, \mathbf{A})\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}})\mathbf{b}$, wegen $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}} \sim \mathrm{A}_{n_{\mathrm{I}}}$. Die Gewichte $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}}$ spannen also den größten Eigenraum der Matrix $\mathbf{B}^{\perp}$ auf, $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}} \sim \mathrm{B}_{n_{\mathrm{II}}-n_{\mathrm{I}}}^{\perp}$, wobei

$$\mathbf{B}^{\perp} := \hat{\mathbf{P}}_{\mathrm{I}}\mathbf{B}\hat{\mathbf{P}}_{\mathrm{I}}. \tag{4.1}$$

Wir führen eine symbolische Bezeichnung für die vorgestellte Adaptationsstrategie mittels Neurogenese $S_{\perp}^{\mathrm{NG}}$ ein:

$$S_{\perp}^{\mathrm{NG}}: \quad \{\mathbf{K}_{\mathrm{I}}; \mathbf{K}_{\mathrm{II}}\} \sim \left\{\mathrm{A}_{n_{\mathrm{I}}}; \left(\mathrm{A}_{n_{\mathrm{I}}}, \mathrm{B}_{g}^{\perp}\right)\right\}, \tag{4.2}$$

mit $g := n_{\mathrm{II}} - n_{\mathrm{I}}$. Das Symbol $\perp$ in $S_{\perp}^{\mathrm{NG}}$ verdeutlicht die Tatsache, dass die Gewichte neuer und alter Neurone im Kodierer stets orthogonal zueinander sind.

Gl. 4.2 bedeutet zusammengefasst, dass der Kodierer $\mathbf{K}_{\mathrm{I}}$ in Umgebung I n_{I} Neurone hat und auf $\mathbf{a}$-Muster optimiert ist, $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_{n_{\mathrm{I}}}$, und zum Kodierer in Umgebung II g Neurone hinzukommen, die auf $\mathbf{B}^{\perp}$ optimiert sind, $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}} \sim \mathrm{B}_{g}^{\perp}$, während die alten Neurone stabil bleiben.

4.2.2 Alternative Adaptationsstrategien

Die Adaptationsstrategie $S_{\perp}^{\mathrm{NG}}$ (Gl. 4.2) besitzt sowohl plastische als auch stabile Gewichte im Kodierer. Die neuen Neurone adaptieren jedoch nicht direkt an die $\mathbf{b}$-Muster, sondern kodieren eine zu den bestehenden Neuronen senkrechten Teil von $\mathbf{b}$ (siehe Abschnitt 4.2.1). Eine Alternative ist daher eine Strategie, die mit $S_{\perp}^{\mathrm{NG}}$ identisch ist, mit Ausnahme davon, dass die neuen Neurone direkt den größten Eigenraum von $\mathbf{B}$ (anstatt von $\mathbf{B}^{\perp}$) aufspannen, also $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}} \sim \mathrm{B}_{g}$. Das hat zur Folge, dass neue und alte Gewichte nicht senkrecht zueinander stehen müssen, also einen beliebigen Winkel einschließen (symbolisiert durch: $\angle$). Wie nennen diese Strategie deshalb $S_{\angle}^{\mathrm{NG}}$:

$$S_{\angle}^{\mathrm{NG}}: \quad \{\mathbf{K}_{\mathrm{I}}; \mathbf{K}_{\mathrm{II}}\} \sim \{\mathrm{A}_{n_{\mathrm{I}}}; (\mathrm{A}_{n_{\mathrm{I}}}, \mathrm{B}_{g})\}. \tag{4.3}$$

Die Konstanz der Gewichte der alten Neurone im Kodierer hat zur Folge, dass in den Strategien $S_{\perp}^{\mathrm{NG}}$ und $S_{\angle}^{\mathrm{NG}}$ (Gl. 4.2 und Gl. 4.3) der Kodierer $\mathbf{K}_{\mathrm{II}}$ i.A. nicht auf $\mathbf{B}$ optimal ist. Das liegt an der Zielsetzung der Strategien, denn der Kode der $\mathbf{a}$-Muster soll möglichst erhalten bleiben. Natürlich ist es notwendig, dem eine Nullhypothese in Form einer Adaptationsstrategie gegenüber zu stellen, die keine Stabilität im Kodierer aufweist. Erst im Vergleich der Strategien kann man ersehen, wie erfolgreich durch zusätzliche Neurone der Kode früherer Eindrücke tatsächlich "gemerkt" werden kann. Wir nennen diese, sich plastisch verhaltene Strategie S^{P}

$$S^{\mathrm{P}}: \quad \{\mathbf{K}_{\mathrm{I}}; \mathbf{K}_{\mathrm{II}}\} \sim \{\mathrm{A}_{n_{\mathrm{I}}}; \mathrm{B}_{n_{\mathrm{II}}}\}. \tag{4.4}$$

Der Kodierer in Umgebung I ist auf **a**-Muster optimiert, $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_{n_{\mathrm{I}}}$, während der Kodierer in Umgebung II für **b**-Muster optimal ist, $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_{n_{\mathrm{II}}}$. Um die Vergleichbarkeit der plastischen Strategie S^{P} mit $S^{\mathrm{NG}}_{\angle}$ und $S^{\mathrm{NG}}_{\perp}$ zu garantieren und Effekte der Neurogenese zu kompensieren, werden wir in S^{P} verschiedene Dimensionalitäten der Kodierer untersuchen (vergl. Tabelle 4.1).

Den plastischen Strategien stellen wir einen Grenzfall gegenüber, in dem sich die Gewichte des Kodierers in Umgebung II überhaupt nicht verändern; sie bleiben stabil, d.h. auf **a**-Muster optimiert:

$$S^{\mathrm{S}}: \quad \{\mathbf{K}_{\mathrm{I}}; \mathbf{K}_{\mathrm{II}}\} \sim \{\mathrm{A}_{n_{\mathrm{I}}}; \mathrm{A}_{n_{\mathrm{I}}}\} \tag{4.5}$$

Man erkennt, dass $S^{\mathrm{NG}}_{\perp}$ und $S^{\mathrm{NG}}_{\angle}$ für $g = 0$ in die Strategie S^{S} übergehen (vergl. Gl. 4.2 und Gl. 4.3).

Als letztes nehmen wir zufällige Kodierungen zum Vergleich mit den übrigen Strategien auf. Sie seien mit

$$S^{\mathrm{R}}: \quad \{\mathbf{K}_{\mathrm{I}}; \mathbf{K}_{\mathrm{II}}\} \sim \{\mathrm{Z}_{n_{\mathrm{I}}}; \mathrm{Z}'_{n_{\mathrm{II}}}\} \tag{4.6}$$

bezeichnet. Hier wird Kodierung durch zufällig gewählten Gewichtsvektoren ausgeführt. In Analogie zur Notation für optimale Kodierer (Gl. 3.63) schreiben wir[1] $\mathbf{K}_{\mathrm{I}} \sim \mathrm{Z}_{n_{\mathrm{I}}}$ und $\mathbf{K}_{\mathrm{II}} \sim \mathrm{Z}_{n_{\mathrm{II}}}$ für eine Kodierung bestehend aus n_{I} bzw. n_{II} zufälligen Vektoren.

Tabelle 4.1 fasst alle in der Durchführung des Experiments verwendeten Adaptationsstrategie zusammen.

4.3 Bewertungskriterien

Um den Ausgang des Experiments für eine Adaptationsstrategie zu bewerten, stützen wir uns auf die Evaluation des Rekonstruktionsfehlers. Durch eine Adaptationsstrategie werden die Kodierer $\mathbf{K}_{\mathrm{I}}$ und $\mathbf{K}_{\mathrm{II}}$ vorgegeben (siehe Abs. 4.2). Die Dekodierungsmatrizen sind dann durch die jeweiligen Optima $\mathbf{D}_{\mathrm{I}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{I}}, \mathbf{A})$ bzw. $\mathbf{D}_{\mathrm{II}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{II}}, \mathbf{B})$ gegeben (siehe Abschnitt 4.1). Wir werden im Experiments die Plastizität, d.h. die Fähigkeit zur Adaptation, und die Stabilität des Kodes nach Adaptation bewerten.

4.3.1 Plastizität

Das Tier ist in den Abschnitten (a) und (b) des Experiments genau einer Sorte Eindrücke ausgesetzt. Die Güte der Adaptation wird daher vom Rekonstruktionsfehler für

[1] Sie spannen sozusagen den größten Eigenraum einer zufälligen Kovarianzmatrix $\mathbf{Z}$ auf.

Art der Strategie	$\mathbf{K}_{\mathrm{I}}$	$\mathbf{K}_{\mathrm{II}}$	Erklärung
S^{R}	Z_ℓ	Z'_ℓ	Zufällige Wahl der Kodierung in Umgebung I und Umgebung II mit ℓ Neuronen. Die Vektoren werden in jeder Umgebung neu generiert.
S^{R}	Z_ℓ	$\mathrm{Z}'_{\ell+g}$	Zufällige Wahl der Kodierung in Umgebung I und Umgebung II. Die Dimensionalität des Kodierers in Umgebung II wird um g auf $n_{\mathrm{I}}+g$ Neurone erhöht, jedoch erneut zufällig gewählt.
S^{R}	$\mathrm{Z}_{\ell+g}$	$\mathrm{Z}'_{\ell+g}$	zufällige Kodierung in beiden Umgebungen, $\ell+g$ Vektoren.
S^{P}	A_ℓ	B_ℓ	Plastizität ohne Stabilität. Kodierer (mit ℓ Neuronen) adaptiert vollständig an die Eindrücke **b** aus der neuen Umgebung II.
S^{P}	A_ℓ	$\mathrm{B}_{\ell+g}$	Plastizität ohne Stabilität. Der Kodierer (mit ℓ Neuronen) wird in Umgebung II um weitere g Neurone ergänzt, es adaptieren alle Neurone **b**-Muster.
S^{P}	$\mathrm{A}_{\ell+g}$	$\mathrm{B}_{\ell+g}$	Plastizität ohne Stabilität. Kodierer (mit $\ell+g$ Neuronen) adaptiert vollständig an die Eindrücke **b** aus Umgebung II.
S^{S}	A_ℓ	A_ℓ	Stabilität ohne Plastizität. Kodierer (mit ℓ Neuronen) bleibt als ganzes in neuer Umgebung II unverändert.
$S^{\mathrm{NG}}_{\angle}$	A_ℓ	$(\mathrm{A}_\ell, \mathrm{B}_g)$	Stabilität und Plastizität. Während die Gewichte der ℓ Neurone des Kodierers aus Umgebung I in Umgebung II beibehalten werden, adaptieren g neu hinzukommende Neurone an die Eindrücke **b** aus Umgebung II.
$S^{\mathrm{NG}}_{\perp}$	A_ℓ	$\left(\mathrm{A}_\ell, \mathrm{B}^{\perp}_g\right)$	Stabilität und Plastizität. Während die Gewichte der ℓ Neurone des Kodierers aus Umgebung I in Umgebung II beibehalten werden, adaptieren g neu hinzukommende Neurone an den Teil der **b**-Muster, welcher orthogonal zu den ℓ stabilen Gewichtsvektoren ist.

Tabelle 4.1: Adaptationsstrategien zum Experiment. In Umgebung I gibt es Muster $\mathbf{a} \in P_{\mathbf{a}}$ (mit Kovarianzmatrix **A**) in Umgebung II ändert sich die Statistik der Eindrücke zu $\mathbf{b} \in P_{\mathbf{b}}$ (mit Kovarianzmatrix **B**) (vergl. Abb. 4.1). Zur Notationsweise vergl. Gl. 3.63 und Abschnitt 4.2.

Muster der jeweiligen Verteilung $P_\mathbf{a}$ bzw. $P_\mathbf{b}$ erfasst (mit Kovarianzmatrix $\mathbf{A}$ bzw. $\mathbf{B}$). Wir verwenden zur Bewertung in Teil (a) bzw. (b) deshalb den optimalen Rekonstruktionsfehler unter Nutzung der jeweiligen Kodierung der Strategie (Gl. 3.35):

(a) $\varepsilon_\mathbf{A}(\mathbf{K}_\mathrm{I}) = \varepsilon(\mathbf{K}_\mathrm{I}, \mathbf{D}^\mathrm{opt}(\mathbf{K}_\mathrm{I}, \mathbf{A}), \mathbf{A})$

(b) $\varepsilon_\mathbf{B}(\mathbf{K}_\mathrm{II}) = \varepsilon(\mathbf{K}_\mathrm{II}, \mathbf{D}^\mathrm{opt}(\mathbf{K}_\mathrm{II}, \mathbf{B}), \mathbf{B})$

4.3.2 Stabilität

In Teil (c) des Experiments fragen wir nach der Stabilität des Kodes, d.h. nach dem Bekanntheitsgrad früherer **a**-Muster nach Adaptation an **b**-Muster. Dies misst folgender Rekonstruktionsfehler, den wir zunächst definieren

$$\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}) := \varepsilon(\mathbf{K}, \mathbf{D}^\mathrm{opt}(\mathbf{K}, \mathbf{B}), \mathbf{A}) \tag{4.7}$$

Die Bezeichnung $\varepsilon_{\mathbf{A}|\mathbf{B}}$ soll andeuten, dass nach der Adaptation an **b**-Muster, der Rekonstruktionsfehler für **a** Muster ausgewertet wird (mit gleichem $\mathbf{K}$). Oder angelehnt an die Sprechweise für bedingte Wahrscheinlichkeiten: "Rekonstruktionsfehler der **a**-Muster gegeben den für **b**-Muster optimalen Dekodierer".

Den Rekonstruktionsfehler $\varepsilon_{\mathbf{A}|\mathbf{B}}$ werden wir zur Auswertung des Teils (c) benutzen (es ist $\mathbf{D}_\mathrm{II} = \mathbf{D}^\mathrm{opt}(\mathbf{K}_\mathrm{II}, \mathbf{B})$)

(c) $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathrm{II}) = \varepsilon\left(\mathbf{K}_\mathrm{II}, \mathbf{D}_\mathrm{II}, \mathbf{A}\right)$

Das Fehlermaß misst im gewissen Sinne den Bekanntheitsgrad der Muster, mit denen das Modell konfrontiert wird. Sein Wert würde am kleinsten sein, nämlich gleich dem optimalen Rekonstruktionsfehler $\varepsilon_\mathbf{A}(\mathbf{K}_\mathrm{II})$, wenn die Dekodierung auf $\mathbf{A}$ anstatt auf $\mathbf{B}$ optimiert wäre.

Abrufen von Speicherinhalten

Explizite Speicherung ist in unserem einfachen Modell nicht integriert, vielmehr nimmt es an, dass Muster im kodierter Form $\mathbf{y} = \mathbf{Kx}$ gespeichert werden (vergl. Abb. 3.4). Trotzdem wollen wir untersuchen, ob die vorgestellten Adaptationsstrategien prinzipiell in der Lage wären, die Anforderungen der Speicherung gerecht zu werden.

Nehmen wir an, das aus Umgebung I stammende Tier befindet sich in Umgebung II (Teil (b)). Da die Aktivität der verborgenen Schicht in unserem Modell gleich der im Hippokampus gespeicherten Information ist, würde ein **a**-Muster aus Umgebung I in der Form $\mathbf{y} = \mathbf{K}_\mathrm{I}\mathbf{a}$ gespeichert sein. In Umgebung II werden die Gedächtnisinhalte aus Umgebung I mit der adaptierten Dekodierung $\mathbf{D}_\mathrm{II}$ abgerufen; die Rekonstruktion eines gespeicherte **a**-Musters lautet daher $\mathbf{D}_\mathrm{II}\mathbf{K}_\mathrm{I}\mathbf{a}$.

Wenn wir annehmen, dass alle Eindrücke aus Umgebung I gespeichert werden, können wir die Güte der Wiederherstellung gespeicherter Information (aus Umgebung I) als den mittleren Rekonstruktionsfehler von $\mathbf{a} \in P_{\mathbf{a}}$ mit Nutzung von $\mathbf{K}_{\mathrm{I}}$ als Kodierer und $\mathbf{D}_{\mathrm{II}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{II}}, \mathbf{B})$ als Dekodierer benennen:

$$\eta_{\,\mathbf{A}|\mathbf{B}} := \varepsilon\,(\mathbf{K}_{\mathrm{I}}, \mathbf{D}_{\mathrm{II}}, \mathbf{A}) \tag{4.8}$$

Es ist klar, dass die Wiederherstellungsqualität dann optimal ist, also $\eta_{\,\mathbf{A}|\mathbf{B}}$ minimal, wenn der Dekodierer aus der Umgebung I benutzt wird, um die gespeicherte $\mathbf{a}$-Muster zu rekonstruieren ($\mathbf{D}_{\mathrm{I}}$ anstatt $\mathbf{D}_{\mathrm{II}}$ in Gl. 4.8). Dann entspricht $\eta_{\,\mathbf{A}|\mathbf{B}}$ dem optimalen Rekonstruktionsfehler $\varepsilon_{\mathbf{A}}(\mathbf{K}_{\mathrm{I}})$ (siehe Gl. 3.35); die Rekonstruktion der Speicherinhalte wäre optimal. Eine gute Wiederherstellung der gespeicherten Muster erfordert also möglichst unveränderte Gewichte im Dekodierer.

Sollte sich die Dimensionalität der verborgenen Schicht durch Neurogenese vergrößern, ändert die Dekodierungsmatrix ihre Gestalt. Das hat zur Folge, dass die Matrixmultiplikation $\mathbf{D}_{\mathrm{II}}\mathbf{K}_{\mathrm{I}}$ (Gl. 4.8) dann nicht definiert ist. Physiologisch beschränkt sich Neurogenese jedoch auf die Gyrus dentatus. Um die Neurogenese auf die Kodierung zu einzugrenzen, wäre es notwendig, eine zusätzliche Verbindung zwischen Kodierung und Dekodierung zu etablieren. Um jedoch nicht von der wesentliche Anforderung der Speicherung, nämlich der Stabilität des Dekodierers, durch ein komplizierteres Modell abzulenken, beschränken wir uns darauf, die Stabilität der "alten" Neurone im Dekodierer zu testen (siehe Abs. 4.2.1), wenn sich die Dimensionalität des Kodierers bei einer Strategie in Umgebung II ändern sollte:

$$\eta_{\,\mathbf{A}|\mathbf{B}} = \varepsilon\left(\mathbf{K}_{\mathrm{I}}, \mathbf{D}_{\mathrm{II}}^{\mathrm{alt}}, \mathbf{A}\right). \tag{4.9}$$

4.4 Definition der Verteilungen

Um das Experiment für verschiedene Strategien numerisch berechnen zu können, müssen wir zunächst die durch die Umwelt ausgelösten Wahrscheinlichkeitsverteilungen neocorticaler Aktivität $P_{\mathbf{a}}$ und $P_{\mathbf{b}}$ konkretisieren. Da über "wahre" Statistik der Feuerraten so gut wie nichts bekannt ist, müssen Annahmen gemacht werden. Dies wird dadurch erleichtert, dass die zur Auswertung des Experimentes verwendeten Fehlermaße nur von der Kovarianzmatrix der Verteilungen abhängen (Gl. 3.15). Die Verteilungen $P_{\mathbf{a}}$ und $P_{\mathbf{b}}$ können daher o.B.d.A. als n-dimensionale Normalverteilungen angesehen werden. Wir haben schon weiter oben (Abschnitt 3.2.2) darauf hingewiesen, dass die Verteilungen zusätzlich mittelwertfrei sind: die Neurone verarbeiten nur die Fluktuation um einen Mittelwert. Nach diesen Vorgaben reicht es aus, die Kovarianzmatrizen $\mathbf{A}$ und $\mathbf{B}$ der Verteilungen $P_{\mathbf{a}}$ und $P_{\mathbf{b}}$ zu spezifizieren. Es ist konzeptionell wichtig einzusehen, dass ein Eigenvektor der Kovarianzmatrix die Varianz der Feuerraten einer Kombination von Neuronen darstellt; dasselbe gilt für den Begriff "Richtung", i.e. ein Vektor im Raum der Feuerraten der EC-Neurone $\mathbb{R}^n$. Da die Wahl des Koordinatensystems jedoch willkürlich ist, seien die Hauptkomponenten der Verteilung $P_{\mathbf{a}}$ aus Gründen der Übersichtlichkeit gerade die Koordinatenachsen,

$$\mathbf{A} := \mathrm{diag}(\mu_1, \ldots, \mu_n). \tag{4.10}$$

Wir nehmen an, dass nur ein kleiner Teil der Varianz der EC-Neurone zum Speichern relevante Informationen mit sich führt. Entscheidend über die Relevanz ist die Stärke der Fluktuation dieser Neurone, i.e. die Größe des zur Hauptkomponenten gehörigen Eigenwertes. Abgesehen von den wenigen großen Eigenwerten, welche die relevante Information tragen, sei die Varianz in den übrigen Richtungen gleichverteiltes, irrelevantes "Rauschen", d.h. diese Richtungen weisen gleichmäßig kleine Eigenwerte auf. Dabei verstehen wir "Rauschen" nicht im eigentlichen Sinne als statistische Fluktuationen um "wahre" Muster $\mathbf{x}$, welche aufgrund eines gestörten Signalweges entstehen. Vielmehr sind es die feinen Details, die die einzelnen Muster unterscheiden, welche sich statistisch in einem "Rauschen" äußern. Das "Rauschen" ist die Ähnlichkeit oder Redundanz der Muster, wogegen die relevanten Informationen die wesentlichen Unterscheidungsmerkmale der Eindrücke einer Umgebung sind. Diese Annahmen über die Art des Spektrums stehen im Einklang mit unserer Modellvorstellung. Denn um eine Redundanzreduktion (siehe Abs. 3.1.3) durchführen zu können, muss die Information auch redundant sein.

Wir wählen zunächst wie beschrieben das Spektrum der Kovarianzmatrix $\mathbf{A}$ der Umgebung I:

$$\mu_i = \begin{cases} \frac{\alpha}{\gamma} \exp(-\tau\,(i-1)), & i \leq n_{\text{info}} \\ \frac{1-\alpha}{n-n_{\text{info}}}, & \text{sonst.} \end{cases} \tag{4.11}$$

Wobei $\gamma := \sum_{j=0}^{n_{\text{info}}-1} \exp(-\tau\, j)$ eine Normierungskonstante und $0 \leq \alpha \leq 1$, $\tau \in \mathbb{R}$ und $n_{\text{info}} \in \mathbb{N}_n$ Parameter sind. Das Spektrum ist so gewählt, dass der Anteil α der Varianz exponentiell auf die ersten n_{info} Dimensionen und der restliche Teil $1-\alpha$ gleichmäßig auf die übrigen Eigenrichtungen aufgeteilt ist. Ersteres ist die relevante Information letzteres "Rauschen". Es gilt außerdem

$$\operatorname{Sp} \mathbf{A} = \sum_{i=1}^{n} \mu_i = 1. \tag{4.12}$$

Abb. 4.2 zeigt das Spektrum mit Parametern, die bei Durchführung des Experiments verwendet werden (Tab. 4.2).

Muster $\mathbf{b} \in P_{\mathbf{b}}$ aus Umgebung II seien in anderen Aspekten redundant und relevant als $\mathbf{a} \in P_{\mathbf{a}}$. Wir nehmen jedoch an, dass das Spektrum beider Kovarianzmatrizen identisch ist. Mit diesen Annahmen können wir $\mathbf{B}$ als räumliche Drehung von $\mathbf{A}$ interpretieren. Es sei also:

$$\mathbf{B}(\mathbf{R}) = \mathbf{R}^T \mathbf{A} \mathbf{R}, \tag{4.13}$$

wobei $\mathbf{R}$ eine Rotationsmatrix (orthogonale Matrix mit $\det \mathbf{R} = 1$) ist, $\mathbf{R} \in \mathcal{R}_n := \{\mathbf{R} | \mathbf{R}^T \mathbf{R} = \mathbf{I}_n \text{ und } \det \mathbf{R} = 1\}$. Nach Gl. 4.13 besitzen beide Kovarianzmatrizen also identisches Eigenwertspektrum; die Eigenvektoren von $\mathbf{B}$ sind gegenüber denen von $\mathbf{A}$ lediglich im Raum gedreht. Um die Unabhängigkeit von speziellen Drehungen zu gewährleisten, werden wir im Experiment über verschiedene Drehungen $\mathbf{R} \in \mathcal{R}_n$ mitteln.

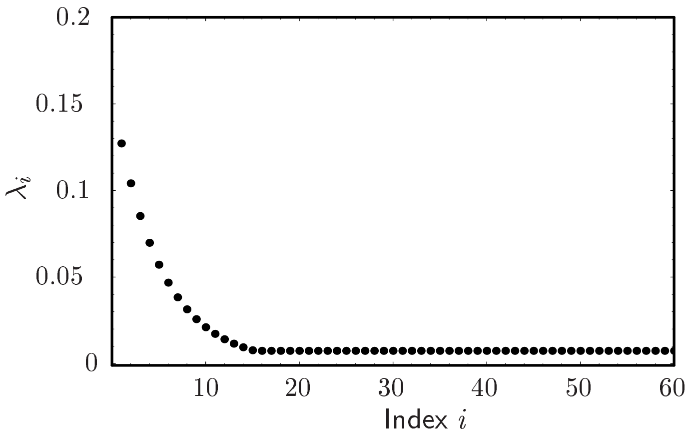

Abbildung 4.2: Geordnete Eigenwerte der Kovarianzmatrizen **A** und **B**, welche für die Ergebnisse in den Tabellen 4.2 und 4.3 verwendet wurden. $\tau = 0.2$, $n = 60$, $\alpha = 2/3$, $n_{\text{info}} = 15$. Die $n_{\text{info}} = 15$ ersten (und größten) Eigenwerte tragen die relevante Information, zusammengenommen $\alpha = 2/3$ des Spektrums, während die übrigen als "Rauschen" interpretiert werden; das sind die restlichen $1/3$ des Spektrums.

		Rekonstruktion Teil (a)	Rekonstruktion Teil (b)	Rekonstruktion Teil (c)	Wiederherstellung gesp. Muster	Mittelwert
Strat.	$\{\mathbf{K}_{\mathrm{I}};\mathbf{K}_{\mathrm{II}}\}$	$\varepsilon_{\mathbf{A}}$	$\langle\varepsilon_{\mathbf{B}}\rangle$	$\langle\varepsilon_{\mathbf{A}\mid\mathbf{B}}\rangle$	$\langle\eta_{\mathbf{A}\mid\mathbf{B}}\rangle$	Σ
S^{R}	$\{\mathrm{Z}_\ell;\mathrm{Z}'_\ell\}$	0.53 (0.01)	0.53 (0.02)	0.91 (0.03)	1.51 (0.07)	0.87 (0.03)
S^{R}	$\left\{\mathrm{Z}_\ell;\mathrm{Z}'_{\ell+g}\right\}$	0.53 (0.01)	0.44 (0.01)	0.81 (0.03)	1.91 (0.14)	0.92 (0.05)
S^{R}	$\left\{\mathrm{Z}_{\ell+g};\mathrm{Z}'_{\ell+g}\right\}$	0.44 (0.01)	0.44 (0.01)	0.81 (0.03)	1.67 (0.09)	0.84 (0.03)
S^{P}	$\{\mathrm{A}_\ell;\mathrm{B}_\ell\}$	0.33 (0.00)	0.33 (0.00)	0.75 (0.01)	1.65 (0.05)	0.77 (0.02)
S^{P}	$\{\mathrm{A}_\ell;\mathrm{B}_{\ell+g}\}$	0.33 (0.00)	0.30 (0.00)	0.67 (0.02)	1.65 (0.05)	0.74 (0.02)
S^{P}	$\{\mathrm{A}_{\ell+g};\mathrm{B}_{\ell+g}\}$	0.30 (0.00)	0.30 (0.00)	0.67 (0.02)	1.69 (0.05)	0.74 (0.02)
S^{S}	$\{\mathrm{A}_\ell;\mathrm{A}_\ell\}$	0.33 (0.00)	0.53 (0.02)	0.77 (0.05)	0.77 (0.05)	0.60 (0.03)
$S^{\mathrm{NG}}_{\angle}$	$\{\mathrm{A}_\ell;(\mathrm{A}_\ell,\mathrm{B}_g)\}$	0.33 (0.00)	0.36 (0.01)	0.41 (0.01)	0.56 (0.03)	0.42 (0.01)
$S^{\mathrm{NG}}_{\perp}$	$\left\{\mathrm{A}_\ell;\left(\mathrm{A}_\ell,\mathrm{B}_g^{\perp\mathrm{A}_\ell}\right)\right\}$	0.33 (0.00)	0.36 (0.01)	0.41 (0.01)	0.45 (0.01)	0.39 (0.01)

Tabelle 4.2: Vergleichstest ($n = 60$, $\ell = 15, g = 5$, $\tau = 0.2$, $\alpha = 2/3$). Zeilenweise sind die Ergebnisse der Adaptationsstrategien aufgelistet. Spaltenweise sind verschiedene Fehlermaße angeordnet (genaueres im Text). Es wurde über 5000 zufällige Rotationen $\mathbf{R} \in \mathcal{R}_{60}$ gemittelt (und über $\mathbf{K}$ in S^{R}), Standardabweichungen sind angegeben. Die Auswahl der Rotationsmatrizen $\mathbf{R}$ ist für jede Strategie identisch. Spannt ein Dekodierer einen Eigenraum einer Verteilung auf, z.B. $\mathbf{K} \sim \mathrm{A}_N$, werden die größten Eigenvektoren als Spaltenvektoren des Kodierers verwendet. Damit sind diese Kodierer orthonormiert (vergl. dazu Tab. 4.3). In Strategien S^{R} wird $\mathbf{K}$ stets neu generiert (in $[0, 1)$ gleichverteilte Matrixeinträge werden normiert). Alle Zahlen wurden numerisch berechnet, auch wenn ein analytischer Wert existiert (vergl. Text). Die Grauwertkästchen verdeutlichen das Abschneiden einer Strategie, indem der Grauwert proportional zum Fehlerwert steigt, i.e. dunkler wird. Das Kästchen ist weiß, wenn der Fehlerwert bei 1/3 (oder darunter) liegt und schwarz, wenn er über 1 steigt. Ist gerade die relevante Information kodiert, wird der Wert 1/3 angenommen werden (wegen $\alpha = 2/3$). Dagegen gleicht ein Wert von 1 dem Fehlerwert, der im Grenzfall $\mathrm{Rg}\,\mathbf{K} = 0$ erzielt wird, also dann, wenn überhaupt keine Kodierung existiert (wg. $\mathrm{Sp}\,\mathbf{A} = \mathrm{Sp}\,\mathbf{B} = 1$).

		Rekonstruktion Teil (a)	Rekonstruktion Teil (b)	Rekonstruktion Teil (c)	Wiederherstellung gesp. Muster	Mittelwert
Strategie	$\{\mathbf{K}_{\mathrm{I}};\mathbf{K}_{\mathrm{II}}\}$	$\varepsilon_{\mathbf{A}}$	$\langle\varepsilon_{\mathbf{B}}\rangle$	$\left\langle\varepsilon_{\mathbf{A}\mid\mathbf{B}}\right\rangle$	$\left\langle\eta_{\mathbf{A}\mid\mathbf{B}}\right\rangle$	Σ
S^{R} $\mid\angle\mid$	$\left\{\mathrm{Z}_{\ell+g};\mathrm{Z}'_{\ell+g}\right\}$	0.44 (0.01)	0.44 (0.01)	0.81 (0.03)	1.67 (0.09)	0.84 (0.03)
S^{R} $\mid\perp\mid$	$\left\{\mathrm{Z}^{\perp}_{\ell+g};\mathrm{Z}'^{\perp}_{\ell+g}\right\}$	0.44 (0.01)	0.44 (0.01)	0.81 (0.03)	1.48 (0.05)	0.79 (0.03)
S^{P} $\angle$	$\{\mathrm{A}_{\ell};\mathrm{B}_{\ell+g}\}$	0.33 (0.00)	0.30 (0.00)	0.67 (0.02)	$5\cdot10^{5}$ $(2\cdot10^{7})$	—
S^{P} $\mid\angle\mid$	$\{\mathrm{A}_{\ell};\mathrm{B}_{\ell+g}\}$	0.33 (0.00)	0.30 (0.00)	0.67 (0.02)	$3\cdot10^{5}$ $(7\cdot10^{6})$	—
S^{P} $\mid\perp\mid$	$\{\mathrm{A}_{\ell};\mathrm{B}_{\ell+g}\}$	0.33 (0.00)	0.30 (0.00)	0.67 (0.02)	1.67 (0.07)	0.74 (0.02)
S^{P} $\mid\perp\mid$E	$\{\mathrm{A}_{\ell};\mathrm{B}_{\ell+g}\}$	0.33 (0.00)	0.30 (0.00)	0.67 (0.02)	1.65 (0.05)	0.74 (0.02)

Tabelle 4.3: Vergleichstest ($n = 60$, $\ell = 15, g = 5$, $\tau = 0.2$, $\alpha = 2/3$). Identische Versuchsdurchführung wie in Tab. 4.2, nur der Übersicht halber in gesonderter Tabelle aufgeführt. An die Gewichtsvektoren der Kodierer sind zusätzliche Bedingungen geknüpft. $\angle$: beliebig ausgerichtete Vektoren, $\perp$: orthogonale Vektoren, $|\cdot|$: normierte Vektoren, E Eigenvektoren. Man erkennt, dass Teil (a)–(c) des Experiments, in Übereistimmung mit der Theorie (Gl. 3.62) nicht von der Art der Basis des durch den Kodierer aufgespannten Raums abhängt. Für orthogonale Kodierer ist die Wiederherstellungsqualität gespeicherter Muster sehr viel besser als für nicht orthogonale. Der optimale Dekodierer wird für S^{P} $\angle$ durch den Anpassungsprozess extrem verzerrt. (S^{P} $\angle$: in [0,1) gleichverteilte Linearkombinationen der Eigenvektoren des Eigenraums. S^{R}: in [0,1) gleichverteilte Einträge).

4.5 Auswertung

Die Bewertungskriterien sind in Tabelle 4.4 zusammengefasst. Alle verwendeten Adaptationsstrategien sind in Tabelle 4.1 aufgelistet und kurz erläutert. Tabelle 4.2 schließlich zeigt die Ergebnisse des Experimentes für die Parameter der Verteilungen $n = 60$ (Dimension EC), $\tau = 0.2$ (Form des Spektrums), $\alpha = 2/3$ (Aufteilung relevanter / redundanter Varianz) und $n_{\text{info}} = 15$ (Anzahl relevanter Dimensionen) (vergl. Abb. 4.2), und der Strategien $\ell = 15$ (Neurone in Umgebung I) und $g = 5$ (Anzahl neuer Neurone in Strategien mit Neurogenese) (vergl. Tab. 4.1).

Teil	Umg.	Kov.	Kodierer	Dekodierer	Auswertung
(a)	I	$\mathbf{A}$	$\mathbf{K}_{\text{I}}$	$\mathbf{D}_{\text{I}} = \mathbf{D}^{\text{opt}}(\mathbf{K}_{\text{I}}, \mathbf{A})$	$\varepsilon_{\mathbf{A}} = \varepsilon\left(\mathbf{K}_{\text{I}}, \mathbf{D}^{\text{opt}}(\mathbf{K}_{\text{I}}, \mathbf{A}), \mathbf{A}\right)$
(b)	II	$\mathbf{B}(\mathbf{R})$	$\mathbf{K}_{\text{II}}$	$\mathbf{D}_{\text{II}} = \mathbf{D}^{\text{opt}}(\mathbf{K}_{\text{II}}, \mathbf{B}(\mathbf{R}))$	$\langle\varepsilon_{\mathbf{B}}\rangle = \langle\varepsilon(\mathbf{K}_{\text{II}}, \mathbf{D}_{\text{II}}, \mathbf{B}(\mathbf{R}))\rangle_{\mathbf{R}\in\mathcal{R}_n}$
(c)	I	$\mathbf{A}$	$\mathbf{K}_{\text{II}}$	$\mathbf{D}_{\text{II}} = \mathbf{D}^{\text{opt}}(\mathbf{K}_{\text{II}}, \mathbf{B}(\mathbf{R}))$	$\left\langle\varepsilon_{\mathbf{A}\vert\mathbf{B}}\right\rangle = \langle\varepsilon(\mathbf{K}_{\text{II}}, \mathbf{D}_{\text{II}}, \mathbf{A})\rangle_{\mathbf{R}\in\mathcal{R}_n}$
Wdh.	II	$\mathbf{A}$	$\mathbf{K}_{\text{I}}$	$\mathbf{D}_{\text{II}} = \mathbf{D}^{\text{opt}}(\mathbf{K}_{\text{II}}, \mathbf{B}(\mathbf{R}))$	$\left\langle\eta_{\ \mathbf{A}\vert\mathbf{B}}\right\rangle = \langle\varepsilon(\mathbf{K}_{\text{I}}, \mathbf{D}_{\text{II}}, \mathbf{A})\rangle_{\mathbf{R}\in\mathcal{R}_n}$

Tabelle 4.4: Zusammenfassung des Experiments einer Strategie $S = \{\mathbf{K}_{\text{I}}; \mathbf{K}_{\text{II}}\}$. Die Kovarianzmatrizen $\mathbf{A}$ und $\mathbf{B}(\mathbf{R})$ sind in Abschnitt 4.4 spezifiziert.

Bevor wir auf die Ergebnisse im Detail eingehen, wollen wir einige allgemeine Feststellungen formulieren. Wenn nichts Gegenteiliges gesagt wird, beziehen wir uns im Folgenden auf die Tabelle 4.2.

- *Zufällige Kodierer verursachen einen größeren Rekonstruktionsfehler als optimierte Kodierer.* Dies ist ersichtlich aus Vergleich zwischen Strategien S^{R} mit S^{P} in Teil (a) und (b). Obwohl die Aussage trivial erscheint, ist die Konsequenz aus ihr bedeutend. Denn damit lohnt sich eine Adaptation des Kodierers an die Statistik der Muster im Vergleich zu einer zufälligen Kodierung.

- *Je mehr Vektoren zur Kodierung genutzt werden, desto besser die Rekonstruktion.* In der Ableitung des Rekonstruktionsfehlers für ein optimales $\mathbf{K}$ wurde dies bereits demonstriert (Gl. 3.65), ein Blick auf Tab. 4.2 bestätigt die theoretischen Resultate. Man sieht den Effekt z.B. im Vergleich von Strategie $\{\text{A}_\ell; \text{B}_\ell\}$ mit $\{\text{A}_{\ell+g}; \text{B}_{\ell+g}\}$ (1. und 2. Spalte). In der einen werden ℓ Dimensionen zur Kodierung genutzt, in der anderen $\ell + g$ Neurone.

  Ob neue Neurone tatsächlich relevante Information kodieren, entscheidet die Struktur der Verteilung. Nach unserer Wahl der Parameter ist dies für die Strategien mit $\mathbf{K}_{\text{I}} \sim \text{A}_{\ell+g}$ oder $\mathbf{K}_{\text{II}} \sim \text{B}_{\ell+g}$ nicht der Fall, denn es ist $n_{\text{info}} = \ell$. Da also ℓ Neurone bereits die relevante Varianz auffangen, kodieren hinzukommende Neurone zwangsläufig "Rauschen" (siehe Abs. 4.4). Es werden dann zwar mehr Details der Muster kodiert, die komprimierte Darstellung $\mathbf{y}$ an der verborgenen Schicht wird dadurch jedoch möglicherweise unnötig aufgebläht. Wäre dagegen $n_{\text{info}} > \ell$

gewählt, würden ℓ Neurone im Kodierer relevante Information vernachlässigen, die durch zusätzliche Neurone kodiert werden könnte.

- *Durch den Anpassungsprozess an die* $\mathbf{b}$*-Muster wird der Kode der* $\mathbf{a}$*-Muster in allen Strategien verändert* Der "Bekanntheitsgrad" der $\mathbf{a}$-Muster (Teil (c), 3. Spalte), sowie die Wiederherstellung gespeicherter Muster (4. Spalte) hat sich nach Adaptation in Umgebung II gegenüber Umgebung I (Teil (a), 1. Spalte) in allen Strategien verschlechtert.

- *Orthogonalität der Gewichtsvektoren des Kodierers ist für die Wiederherstellungsqualität gespeicherter Muster entscheidend.* Obwohl in den Teilen (a) bis (c) die Orthogonalität des Kodierers keine Rolle spielt (Tab. 4.3), sind die Fehlerwerte der Wiederherstellungsqualität $\eta_{\mathbf{A}|\mathbf{B}}$ für nicht orthogonale Kodierer durchgehend höher (siehe Tab. 4.3, aber auch 0.56 gegenüber 0.45 für $S^{\mathrm{NG}}_{\angle}$ bzw. $S^{\mathrm{NG}}_{\perp}$ in Tab. 4.2, 4. Spalte).

- *Die Adaptationsstrategie mittels Neurogenese* $S^{NG}_{\perp}$ *wird am ehesten den Anforderungen des Experiments gerecht.* Im Mittel über die vier Bewertungskriterien hat die Strategie $S^{\mathrm{NG}}_{\perp}$ den kleinsten Wert (0.39). Sie erfüllt demnach am besten die Erfordernisse nach Anpassung an $\mathbf{b}$-Muster und Konservierung des Kodes bekannter $\mathbf{a}$-Muster. Plastische Strategien S^{P} sind zwar innerhalb einer Umgebung vorteilhaft, "vergessen" bei der Anpassung jedoch die Kodierung früherer Muster (vergl. 2. Spalte mit 3. und 4. Spalte in Tab. 4.2).

4.5.1 Teil (a)

In Teil (a) des Experiments (Abb. 4.1) wurde der mittlere Rekonstruktionsfehler von $\mathbf{a}$-Mustern ausgewertet, hierbei wurden $\mathbf{K}_{\mathrm{I}}$ und $\mathbf{D}_{\mathrm{I}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{I}}, \mathbf{A})$ benutzt.

Weil im Experiment $\ell = n_{\mathrm{info}}$ gewählt wurde (Abb. 4.2), erfassen die ℓ Gewichtsvektoren der Strategien mit $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_\ell$ oder $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_\ell$ die relevante Information in Umgebung I bzw. Umgebung II. Denn diese Kodierer spannen den größten Eigenraum der jeweiligen Kovarianzmatrix auf, und kodieren damit alle n_{info} relevanten Dimensionen, i.e. die größten Hauptkomponenten (vergl. Definition der Kovarianzmatrizen , Abschnitt 4.4). Da $\alpha = 2/3$ der Varianz durch die $n_{\mathrm{info}} = 15$ größten Hauptkomponenten der Verteilungen übertragen wird, ist der theoretische Wert des Rekonstruktionsfehler dieser Strategien 1/3 (Gl. 3.64). Die numerischen Fehlerwerte stimmen damit überein (vergl. Tab. 4.2).

Spannt der Kodierer einen höher dimensionalen Eigenraum auf als durch die Zahl der relevanten Dimensionen $n_{\mathrm{info}} = 15$ vorgegeben, $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_{\ell+g}$, kodieren die $g = 5$ zusätzlichen Neurone "Rauschen", d.h. Eigenrichtungen mit Eigenwert $\frac{1-\alpha}{n-n_{\mathrm{info}}} = 1/135$ (Gl. 4.11). Dadurch ist der Rekonstruktionsfehler um $g/135 \approx 0.037$ geringer als 1/3. So erklärt sich der Wert von 0.30 (1. Spalte).

Bei zufälliger Wahl des Kodierers in den Strategien S^{R} ist der Rekonstruktionsfehler geringer als man auf den ersten Blick erwarten würde, denn der Wert 0.53 für $\mathbf{K}_{\mathrm{I}} \sim \mathrm{Z}_\ell$

steht im Gegensatz zur folgender Abschätzung. Auf jede Dimension fallen durchschnittlich $1/n = 1/60$ des Spektrums ab, so dass $\ell = 15$ Neurone $15/60 = 0.25$ Prozentpunkte kodieren und somit ein Rekonstruktionsfehler von 0.75 übrig bleiben sollte (wg. $\mathrm{Sp}\,\mathbf{A} = \mathrm{Sp}\,\mathbf{B} = 1$). Diese Überschlagsrechnung geht jedoch fälschlicherweise davon aus, dass der kodierte Teil der Verteilung durch den Dekodierer aus der niedrigdimensionalen Darstellung (der verborgenen Schicht) in den Ursprungsraum $\mathbb{R}^n$ unverändert zurück projiziert wird (Annahme $\mathrm{Bild}\,\mathbf{D}^{\mathrm{opt}} = \mathrm{Bild}\,\mathbf{K}^T$, siehe Abschnitt 3.2.4). Dies trifft i.A. nicht zu. Vielmehr resultiert der bessere Wert 0.53 aus der Nutzung der Korrelation des kodierten Raumes mit anderen Richtungen starker Varianz (siehe Abschnitt 3.2.4 und Abschnitt 5.3.1).

4.5.2 Teil (b)

Die Kovarianzmatrix $\mathbf{B}$ der Verteilung $P_{\mathbf{b}}$ hat das gleiche Eigenwertspektrum wie $\mathbf{A}$ aus Umgebung I. Der Rekonstruktionsfehler ist für die plastischen Strategien S^{P} und den Strategien S^{R} daher in Teil (b) identisch zu Teil (a), sofern die gleiche Dimensionalität des Kodierers verwendet wird. Man sieht, dass sich die Fehlerwerte entsprechen, z.B. 0.33 für $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_\ell$ und $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_\ell$ oder 0.44 für $\mathbf{K}_{\mathrm{I}} \sim \mathrm{Z}_{\ell+g}$ und $\mathbf{K}_{\mathrm{II}} \sim \mathrm{Z}_{\ell+g}$.

Die Verminderung des Rekonstruktionsfehlers bei $g = 5$ zusätzlichen Neuronen ist in $\{\mathrm{Z}_\ell; \mathrm{Z}_{\ell+g}\}$ stärker ausgeprägt, als es in der plastischen Strategie $\{\mathrm{A}_\ell; \mathrm{B}_{\ell+g}\}$ der Fall ist (0.09 gegenüber 0.03). Dies hängt damit zusammen, dass mit $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_\ell$ bereits alle relevanten Dimensionen, und damit die größten Eigenvektoren, kodiert werden, in $\mathbf{K}_{\mathrm{I}} \sim \mathrm{Z}_\ell$ dagegen i.A. nicht. Im letztgenannten Fall ist im Mittel mehr zusätzliche Information da, welche ergänzend kodiert werden kann.

Interessanter ist das Verhalten der Strategien, welche Gewichtsvektoren aus Umgebung I konservieren und höchstens ein Teil ihrer Neurone an Umgebung II anpassen; das sind S^{S}, $S^{\mathrm{NG}}_{\angle}$ und $S^{\mathrm{NG}}_{\perp}$. Wird keine Adaptation an **b**-Muster durchgeführt, haben die Gewichtsvektoren des Kodierers keinen Bezug zu den relevanten Dimensionen in $\mathbf{B}$, weil $\mathbf{B}$ gegenüber $\mathbf{A}$ beliebig gedreht ist. Anstatt $\mathbf{K}_{\mathrm{I}}$ stabil zu halten, lassen sich deshalb genauso gut zufällige Vektoren zur Kodierung benutzen, was am Vergleich $\mathbf{K}_{\mathrm{II}} \sim \mathrm{Z}_\ell$ (S^{R}) und $\mathbf{K}_{\mathrm{II}} \sim \mathrm{A}_\ell$ (S^{S}) deutlich wird: beide haben den Fehlerwert 0.53 (2. Spalte).

Der Rekonstruktionsfehler der **b**-Muster liegt in der Strategie $S^{\mathrm{NG}}_{\angle}$, mit $\mathbf{K}_{\mathrm{II}} \sim (\mathrm{A}_\ell, \mathrm{B}_g)$, bei 0.36, sie adaptiert in Umgebung II damit nur etwas schlechter als die plastischen Strategien S^{P} mit $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_\ell$, obwohl statt $\ell = 15$ (in S^{P}) nur $g = 5$ (in $S^{\mathrm{NG}}_{\angle}$) Neurone zur Adaptation an Umgebung II genutzt werden. Dies hat seinen Grund an der exponentiellen Form des Spektrums für die n_{info} relevanten Dimensionen (Abb. 4.2, Gl. 4.11). Daher kodieren $g = 5$ plastische Neurone, $\mathbf{K}^{\mathrm{neu}}_{\mathrm{II}} \sim \mathrm{B}_g$ einen Anteil von $\sum_{i=1}^{5} \frac{\alpha}{\gamma} \exp(-\tau(i-1)) \approx 0.443$ der Summe des Spektrums. Für die übrigen $\ell = 15$ Neurone in $S^{\mathrm{NG}}_{\angle}$ gilt $\mathbf{K}^{\mathrm{alt}}_{\mathrm{II}} \sim \mathrm{A}_g$, sie kodieren daher durchschnittlich etwa $(1 - 0.443)/(n - n_{\mathrm{info}})\ell \approx 0.187$. Der Rekonstruktionsfehler sollte demnach ungefähr bei 0.37 liegen. Auch hier können die alten Neurone die Kovarianz ausnutzen, da sie i.A. keinen Eigenraum von $\mathbf{B}$ aufspannen (Abschnitt 3.2.4), um den Rekonstruktionsfehler weiter zu verbessern. Da das Spektrum der Eigenwerte ohne die größten $g = 5$ relativ gleichförmig sind (vergl. Abb. 4.2), ist die erzielte Verbesserung auf 0.36 allerdings gering.

Die Rechnung zeigt, dass Strategien mitunter andere Details der Muster zur Kodierung auswählen, als es unserer Einteilung in irrelevante und relevante Information, gemessen an der Größe der Eigenwerte, entspricht (vergl. Abschnitt 4.4). Das dies nur eine Verschlechterung um 0.03 (S^{P} gegenüber $S_{\angle}^{\mathrm{NG}}$) nach sich zieht, liegt wiederum an dem relativ gleichverteilten Spektrum, wenn man die ersten 5 Eigenwerte nicht berücksichtigt: zwischen den letzten relevanten Dimensionen (z.B. 10 – 15 in Abb. 4.2) und den irrelevanten Dimensionen (15 – 60 in Abb. 4.2) besteht nur ein geringer Unterschied in der Größe des Eigenwerte. Bei den verwendeten Parametern des Spektrums (insbesondere $\tau = 0.2$, Gl. 4.11), könnte man deshalb näherungsweise die ersten 5 Hauptkomponenten als alleinig relevant kategorisieren.

Dass $S_{\perp}^{\mathrm{NG}}$ und $S_{\angle}^{\mathrm{NG}}$ den gleichen Rekonstruktionsfehler aufweisen, obwohl die neuen Neurone einerseits orthogonal zu den alten Neuronen stehen, also auf $\mathbf{B}^{\perp}$ optimiert sind ($S_{\perp}^{\mathrm{NG}}$, Gl. 4.2), und andererseits die größten Hauptkomponenten von $\mathbf{B}$ lernen ($S_{\angle}^{\mathrm{NG}}$, Gl. 4.3), kann man durch die geringe Anzahl relevanter Dimensionen erklären ($n_{\mathrm{info}} = 15$ gegenüber $n = 60$). Deshalb ist es sehr wahrscheinlich, dass durch Drehung der Kovarianzmatrix $\mathbf{B}$ ihre $g = 5$ größten Hauptkomponenten orthogonal zu $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ sind. Da $\mathbf{B}^{\perp}$ die Projektion von $\mathbf{B}$ auf den Kern von $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ beschreibt (Gl. 4.1), sind dann die Gewichtsvektoren $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}}$ beider Strategien nahezu identisch.

4.5.3 Teil (c)

Im Teil (c) wurde auf Veränderungen des Kodes früherer Eindrücke getestet, indem der "Bekanntheitsgrad" der **a**-Muster nach Adaptation an **b**-Muster gemessen wurde. Hier rekonstruieren also Kodierer $\mathbf{K}_{\mathrm{II}}$ und der auf $\mathbf{B}$ optimale Dekodierer **a**-Muster. Im Unterschied zu den ersten beiden Teilen des Experiments ist der Dekodierer damit nicht auf die aktuelle Umgebung optimiert, denn der optimale Dekodierer ist i.A. abhängig von der Verteilung (Gl. 3.21).

Man erkennt, dass die Strategien $S_{\perp}^{\mathrm{NG}}$ und $S_{\angle}^{\mathrm{NG}}$ den Kode der **a**-Muster zumindest teilweise bewahren. Dennoch ist die Rekonstruktion der **a**-Muster nach dem Adaptationsprozess im Mittel deutlich schlechter als sie es vorher war (0.41 gegenüber 0.33). Obwohl auch die Strategie S^{S} mit $\mathbf{K}_{\mathrm{II}} \sim \mathrm{A}_{\ell}$ ebenso viele Neurone stabil belässt wie $S_{\perp}^{\mathrm{NG}}$ und $S_{\angle}^{\mathrm{NG}}$, ist die Rekonstruktion in (c) sehr viel schlechter (0.77 gegenüber 0.41). Die Stabilität des Kodierers alleine erreicht offensichtlich nicht, dass frühere Muster nach Adaptation des Dekodierers an neue Eindrücke bekannt bleiben; der Einbau zusätzlicher, plastischer Neurone scheint dagegen die Bewahrung ihres Kodes zu fördern.

Mit Ausnahme von $S_{\perp}^{\mathrm{NG}}$ und $S_{\angle}^{\mathrm{NG}}$ wirkt sich der suboptimale Dekodierer fatal auf die Rekonstruktion aus. Interessanterweise ist der Fehlerwert für $\mathbf{K}_{\mathrm{II}} \sim \mathrm{A}_{\ell}$ (S^{S}) mit 0.77 sogar größer als für $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_{\ell}$ (S^{P}), welcher 0.67 beträgt. Immerhin ist die Rekonstruktion in der stabilen Strategie S^{S} etwas besser als der Wert bei für eine zufälliger

Kodierung S^{R} (0.77 gegenüber 0.91).

Bei Strategien mit $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_{n_{\mathrm{II}}}$, das sind die plastischen Strategien S^{P} (vergl. Tabelle 4.1), lässt sich der "Bekanntheitsgrad" aus Teil (c) analytisch berechnen. Dann ist nämlich $\hat{\mathbf{P}}_{\mathrm{II}} = (\mathbf{I} - \mathbf{D}_{\mathrm{II}}\mathbf{K}_{\mathrm{II}})$ symmetrisch, weil Kodierer $\mathbf{K}_{\mathrm{II}}$ und Dekodierer $\mathbf{D}_{\mathrm{II}}$ gleichermaßen auf Verteilung $P_{\mathbf{b}}$ optimiert sind (Gl. 3.26), und es folgt (vergl. Abschnitt 4.4)

$$\begin{aligned}
&[\text{Gl. } 4.7] && \left\langle \varepsilon_{\mathbf{A}|\mathbf{B}} \right\rangle_{\mathbf{R}\in\mathcal{R}_n} = \left\langle \mathrm{Sp}\left(\left(\hat{\mathbf{P}}_{\mathrm{II}}\right)^T \hat{\mathbf{P}}_{\mathrm{II}} \mathbf{A} \right) \right\rangle_{\mathbf{R}\in\mathcal{R}_n} \\
&[\hat{\mathbf{P}}_{\mathrm{II}}\hat{\mathbf{P}}_{\mathrm{II}} = \hat{\mathbf{P}}_{\mathrm{II}}] && \qquad\qquad = \mathrm{Sp}\left(\left\langle \hat{\mathbf{P}}_{\mathrm{II}} \right\rangle_{\mathbf{R}\in\mathcal{R}_n} \mathbf{A} \right)
\end{aligned} \tag{4.14}$$

Es ist leicht einzusehen, dass durch Drehung der Eigenvektoren von $\mathbf{B}(\mathbf{R})$ in alle möglichen Richtungen $\hat{\mathbf{P}}_{\mathrm{II}}$ im Mittel einer Matrix $\mathbf{W}$ mit geweißtem Spektrum entspricht. Denn $\hat{\mathbf{P}}_{\mathrm{II}}$ ist eine Projektionsmatrix auf den Kern von $\mathbf{K}_{\mathrm{II}}$ und besitzt deshalb genau $\dim \mathrm{Kern}\, \mathbf{K}_{\mathrm{II}} = n - n_{\mathrm{II}}$ Eigenwerte mit Wert 1, während die übrigen gleich Null sind. Dadurch sind alle Eigenwerte der geweißte Matrix $\mathbf{W}$ identisch $\lambda_i = \frac{n-n_{\mathrm{II}}}{n}$. Es ergibt sich somit aus Gl. 4.14 ein einfacher Ausdruck:

$$\begin{aligned}
&[\mathbf{W} = \tfrac{n-n_{\mathrm{II}}}{n}\,\mathbf{I}] && \left\langle \varepsilon_{\mathbf{A}|\mathbf{B}} \right\rangle_{P_{\mathbf{b}}\in\mathcal{P}} = \frac{n - n_{\mathrm{II}}}{n}\,\mathrm{Sp}\,\mathbf{A} \\
&[\mathrm{Sp}\,\mathbf{A} = 1] && \qquad\qquad = 1 - \frac{n_{\mathrm{II}}}{n}
\end{aligned} \tag{4.15}$$

Rechnet man die Werte für $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_{n_{\mathrm{II}}}$ mit $n_{\mathrm{II}} = \ell = 15$ und $n_{\mathrm{II}} = \ell + g = 20$ nach ($n = 60$), erhält man 3/4 bzw. 2/3 in Übereinstimmung mit den numerischen Resultaten in Tabelle 4.2.

4.5.4 Wiederherstellung gespeicherter Muster

Das Fehlermaß $\eta_{\,\mathbf{A}|\mathbf{B}}$ bewertet die Güte der Dekodierung gespeicherter Muster nach dem Adaptationsprozess.

Es zeigt sich ein ähnliches Bild wie in Teil (c): mit Ausnahme der Strategien $S^{\mathrm{NG}}_{\angle}$ und $S^{\mathrm{NG}}_{\perp}$ ist die Rekonstruktion gespeicherter Muster fatal, man sieht es an den hohen Fehlerwerten (4. Spalte in Tab. 4.2). Die Werte liegen sogar deutlich über 1; würde keine Kodierung verwendet werden, also $\mathrm{Rg}\,\mathbf{K} = 0$, wäre der der Wert des Rekonstruktionsfehler genau 1 (wg. $\mathrm{Sp}\,\mathbf{A} = \mathrm{Sp}\,\mathbf{B} = 1$).

Man kann die Beobachtung machen, dass, im Gegensatz zur Situation im Kodierer, mehr Neurone im Dekodierer die Rekonstruktion sogar verschlechtern, zu sehen an $\{\mathrm{Z}_\ell; \mathrm{Z}'_\ell\}$ gegenüber $\{\mathrm{Z}_{\ell+g}; \mathrm{Z}'_{\ell+g}\}$ (1.51 bzw. 1.67). Dass die Werte sich für die Strategien $\{\mathrm{A}_\ell; \mathrm{B}_{\ell+g}\}$ und $\{\mathrm{A}_{\ell+g}; \mathrm{B}_{\ell+g}\}$ unterscheiden (könnten) (1.65 gegenüber 1.69), obwohl beides Mal $\mathbf{K}_{\mathrm{II}} \sim \mathrm{B}_{\ell+g}$ gilt, liegt daran, dass $\eta_{\,\mathbf{A}|\mathbf{B}}$ nur die Abweichung der alten Neurone im Dekodierer berücksichtigt (Gl. 4.9); die Anzahl der alten Neurone ist wegen $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_\ell$ bzw. $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_{\ell+g}$ in beiden Strategien unterschiedlich. Dass diese Werte sich nur sehr wenig (wenn überhaupt signifikant) unterscheiden, rechtfertigt die Motivation,

$\eta_{\mathbf{A}|\mathbf{B}}$ auf die alten Neurone zu beschränken (siehe Abs. 4.3). Der größte Teil des Rekonstruktionsfehlers scheint durch die alten Neurone provoziert zu sein, die zusätzlichen Neurone bieten dafür offenbar keine Abhilfe.

Mit Ausnahme der stabilen Strategie S^{S} sind die Fehlerwerte $\eta_{\mathbf{A}|\mathbf{B}}$ für alle Strategien größer als $\varepsilon_{\mathbf{A}|\mathbf{B}}$ (Teil (c)). Dies resultiert aus einer Anpassung des Dekodierers $\mathbf{D}_{\mathrm{II}}$ an den jeweiligen Kodierer in Umgebung II. Denn zur Berechnung der Fehlermaße werden unterschiedliche Kodierer, nämlich $\mathbf{K}_{\mathrm{I}}$ in $\eta_{\mathbf{A}|\mathbf{B}}$ und $\mathbf{K}_{\mathrm{II}}$ in $\varepsilon_{\mathbf{A}|\mathbf{B}}$ (Gl. 4.7 bzw. Gl. 4.9), verwendet, die Dekodierer dagegen entsprechen sich. Beide Maße nutzen $\mathbf{D}_{\mathrm{II}}$, wenn man von der Einschränkung auf die alten Neurone bei $\eta_{\mathbf{A}|\mathbf{B}}$ einmal absieht. Ist der Kodierer stabil, so ist die Differenz zwischen den Fehlerwerten gering; so zu sehen an S^{S}, $S^{\mathrm{NG}}_{\angle}$ und $S^{\mathrm{NG}}_{\perp}$ (0–0.15 Punkte Differenz). Den Grenzfall zeigt die stabile Strategie S^{S}, in der der Kodierer konstant ist $\mathbf{K}_{\mathrm{I}} = \mathbf{K}_{\mathrm{II}} \sim \mathrm{A}_{\ell}$, so dass beide Maße den gleichen Wert liefern (0.67). Verändert sich der Kodierer dagegen stark, wie es bei zufälliger Wahl der Fall ist (S^{R}), ist auch die Differenz der Maße groß (z.B. 0.60 für $\{\mathrm{Z}_{\ell}; \mathrm{Z}'_{\ell}\}$).

Es ist bemerkenswert, dass die Rekonstruktion gespeicherter Muster mit der Strategie $S^{\mathrm{NG}}_{\perp}$ einen deutlich geringeren mittleren Fehler aufweist als $S^{\mathrm{NG}}_{\angle}$ (0.45 zu 0.56), insbesondere deswegen, weil beide Strategien in den Teilen (a) bis (c) identisch bewertet worden sind. Diesem Phänomen liegt die erzwungene Orthogonalität des Kodierers $\mathbf{K}_{\mathrm{II}}$ in $S^{\mathrm{NG}}_{\perp}$ zugrunde. In Strategie $S^{\mathrm{NG}}_{\angle}$ ist die kodierte Darstellung der alten Neurone $\mathbf{K}^{\mathrm{alt}}_{\mathrm{II}}\mathbf{b}$ möglicherweise mit derjenigen der neuen Neurone $\mathbf{K}^{\mathrm{neu}}_{\mathrm{II}}\mathbf{b}$ korreliert. Deshalb werden die alten Gewichtsvektoren des Dekodierers $\mathbf{D}^{\mathrm{alt}}_{\mathrm{II}}$ durch den neuen Teil der Kodierung $\mathbf{K}^{\mathrm{neu}}_{\mathrm{II}}$ beeinflusst (bei Adaptation in Umgebung II). Da das Maß $\eta_{\mathbf{A}|\mathbf{B}}$ auf die alten Neurone im Dekodierer beschränkt ist, wird diese Veränderung durch einen erhöhten Fehlerwert in $S^{\mathrm{NG}}_{\angle}$ gegenüber $S^{\mathrm{NG}}_{\perp}$ wiedergegeben (Differenz 0.11) (vergl. auch Kapitel 5).

4.6 Zusammenfassung

Wir haben Anpassung an unbekannte Eindrücke für verschieden Adaptationsstrategien des Kodierers untersucht. Anhand eines numerischen Experiments haben wir (i) die Fähigkeit zur Anpassung an die Statistik neuer Eindrücke (Plastizität) und (ii) das Vergessen der Rekonstruktion von Erinnerungen durch den Adaptationsprozess bewertet.

Wir haben festgestellt, dass die Adaptationsstrategien, die die Gewichte bestehender Neurone im Kodierer konservieren und die Plastizität neu gebildeter Neurone zur Anpassung nutzen ($S^{\mathrm{NG}}_{\perp}$, $S^{\mathrm{NG}}_{\angle}$), den Anforderung des Experiments am ehesten genügen – zumindest für die in Abschnitt 4.4 motivierte Wahl der Parameter.

Zusätzliche Neurone verbessern zwar generell die Adaptationsfähigkeit, der geringere Rekonstruktionsfehler ist aber möglicherweise erkauft durch Kodierung unwesentlicher Details; die Folge ist eine ineffiziente Kodierung. Darüber hinaus wird durch Neurogenese die Bewahrung des Kodes früherer Eindrücke gefördert, sofern bestehende Gewichtsvektoren des Kodierers sich der Anpassung entziehen. Stabilität der Neurone im Kodierer alleine ist nicht ausreichend, um den Kode früherer Erinnerungen zu behalten oder die Dekodierung von Gedächtnisinhalten zu garantieren. Orthogonale Gewichtsvektoren verbessern dagegen die Wiederherstellungsqualität gespeicherter Eindrücke wesentlich.

Kapitel 5

Untersuchung des Adaptationsprozesses

Die Durchführung des Experiments (Kapitel 4) hat gezeigt, dass Neurogenese das Modell des Hippokampus befähigt, den Anforderungen nach Stabilität und Plastizität gerecht zu werden. Im Experiment haben wir eine spezielle Statistik neocorticaler Aktivitäten angenommen. Zwar erscheint unsere Wahl des Spektrums plausibel (vergl. Motivation im Abschnitt 4.4), aber natürlich bedeutet das nicht, dass diese Wahl biologisch korrekt ist. Deshalb wollen wir in den folgenden Abschnitten mathematisch die Änderungen der Verteilungen charakterisieren, bei denen neu gebildete Neurone den Adaptationsprozess unseres Modells sinnvoll unterstützen. Da $S_{\perp}^{\mathrm{NG}}$ (Gl. 4.2) die erfolgreichste Strategie des Experiments ist (siehe Tab. 4.2), betrachten wir diese Form der Adaptation hauptsächlich und gehen nur am Rande auf die Strategie $S_{\angle}^{\mathrm{NG}}$ (Gl. 4.3).

Es wird uns die Stabilität des Kodes für alte Erinnerungen in Abhängigkeit von der Statistik der Eindrücke in Umgebung II beschäftigen. Dafür wollen wir zunächst eine vereinfachte Notation einführen. Voraussetzungen und Bezeichnungen der Strategie $S_{\perp}^{\mathrm{NG}}$ sind in der Tabelle 5.1 zusammengefasst.

Umgebung I ($\mathbf{A}$)	Umgebung II ($\mathbf{B}$)
$\mathbf{K_a}$	$\mathbf{K}_{\mathrm{II}}^T = \begin{pmatrix} \mathbf{K_a}^T & \mathbf{K}_{\mathbf{b}^\perp}^T \end{pmatrix}$
$\mathbf{D_a^A}$	$\mathbf{D}_{\mathrm{II}} = \begin{pmatrix} \mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} & \mathbf{D}_{\mathrm{II}}^{\mathrm{neu}} \end{pmatrix} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathrm{II}}, \mathbf{B})$

Tabelle 5.1: Voraussetzungen der Strategie $S_{\perp}^{\mathrm{NG}}$. In Umgebung I sind Kodierer und zugehöriger Dekodierer auf $\mathbf{a}$-Muster optimiert. In Umgebung II kommen neue Gewichtsvektoren hinzu, die alten Gewichte des Kodierers, $\mathbf{K_a}$, verändern sich nicht. Die neuen Vektoren des Kodierers, $\mathbf{K}_{\mathbf{b}^\perp}$, sind auf $\mathbf{B}^\perp$ optimiert, und der Dekodierer $\mathbf{D}_{\mathrm{II}}$ ist als Ganzes auf die Kovarianzmatrix $\mathbf{B}$, unter Nutzung des Kodierers $\mathbf{K}_{\mathrm{II}}$, optimiert. Wie üblich seien die Kovarianzmatrizen $\mathbf{A}$ und $\mathbf{B}$ invertierbar und alle Kodierer haben Höchstrang, $n_{\mathrm{I}} := \mathbf{K_a}$, $n_{\mathrm{II}} := \mathrm{Rg}\,\mathbf{K}_{\mathrm{II}}$ und $g := \mathrm{Rg}\,\mathbf{K}_{\mathbf{b}^\perp}$, wobei $g = n_{\mathrm{II}} - n_{\mathrm{I}}$ ist. Zur Notationsweise siehe Abschnitt 5.1.

5.1 Notation

Der Kodierer $\mathbf{K}_{\mathrm{II}}$ der Strategien mit Neurogenese gliedert sich in zwei Teile: die Gewichtsvektoren der "alten", schon in Umgebung I existenten Neurone und diejenigen der hinzukommenden, "neuen" Neurone. Deshalb können wir $\mathbf{K}_{\mathrm{II}}^T = \begin{pmatrix} \mathbf{K}_{\mathrm{II}}^{\mathrm{alt}\,T} & \mathbf{K}_{\mathrm{II}}^{\mathrm{neu}\,T} \end{pmatrix}$ schreiben (siehe Abs. 4.2.1). Die Strategie $S_{\perp}^{\mathrm{NG}} = \left\{ \mathrm{A}_{n_{\mathrm{I}}}; \left(\mathrm{A}_{n_{\mathrm{I}}}, \mathrm{B}_g^{\perp} \right) \right\}$ sieht vor, dass die Gewichte der alten Neurone sich nicht verändern, nachdem sie in Umgebung I für die Kodierung von $\mathbf{a}$-Mustern optimiert worden sind; es ist also $\mathbf{K}_{\mathrm{I}} \sim \mathrm{A}_{n_{\mathrm{I}}}$ und $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}} = \mathbf{K}_{\mathrm{I}}$. Die Gewichte der neuen Neurone $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}}$ adaptieren hingegen auf den zu $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ senkrechten Teil der $\mathbf{b}$-Muster (Gl. 4.2), es ist also $\mathbf{K}_{\mathrm{II}}^{\mathrm{neu}} \sim \mathrm{B}_g^{\perp \mathrm{A}_{n_{\mathrm{I}}}}$. Da wir in diesem Kapitel lediglich die Strategie $S_{\perp}^{\mathrm{NG}}$ untersuchen wollen, werden wir eine vereinfachte Notation einführen, welche die Voraussetzungen reflektiert: wir schreiben $\mathbf{K}_{\mathbf{a}} := \mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ und $\mathbf{K}_{\mathbf{b}^{\perp}} := \mathbf{K}_{\mathrm{II}}^{\mathrm{neu}}$. Die Indizes deuten an, dass beide Blockmatrizen des Kodierers $\mathbf{K}_{\mathrm{II}}$ auf $\mathbf{a}$-Muster bzw. auf den zu $\mathbf{K}_{\mathrm{II}}^{\mathrm{alt}}$ senkrechten Teil der $\mathbf{b}$-Muster optimiert sind. Zusammengenommen schreiben wir den Kodierer der Umgebung II der Strategie $S_{\perp}^{\mathrm{NG}}$ im Folgenden $\mathbf{K}_{\mathrm{II}}^T = \begin{pmatrix} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{b}^{\perp}}^T \end{pmatrix}$. In Umgebung I wird der Kodierer $\mathbf{K}_{\mathbf{a}} = \mathbf{K}_{\mathrm{I}}$ verwendet.

Der Dekodierer der Umgebung II ist analog zum Kodierer in "alte" und "neue" Neurone gegliedert, $\mathbf{D}_{\mathrm{II}} = \begin{pmatrix} \mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} & \mathbf{D}_{\mathrm{II}}^{\mathrm{neu}} \end{pmatrix}$. Da der Dekodierer als Ganzes an die $\mathbf{b}$-Muster adaptiert, sind die Blockmatrizen $\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}}$ und $\mathbf{D}_{\mathrm{II}}^{\mathrm{neu}}$ i.A. vom gesamten, durch $\mathbf{K}_{\mathrm{II}}$ kodierten Raum abhängig. Wir können also nicht ohne weiteres z.B. $\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}}$ als Dekodierer zu $\mathbf{K}_{\mathbf{a}}$ identifizieren. Letzteres ist jedoch möglich, falls die von $\mathbf{K}_{\mathbf{a}}$ und $\mathbf{K}_{\mathbf{b}^{\perp}}$ kodierten Aspekte der $\mathbf{b}$-Muster miteinander unkorreliert sind, d.h. $\left\langle \mathbf{K}_{\mathbf{a}} \mathbf{b}\mathbf{b}^T \mathbf{K}_{\mathbf{b}^{\perp}}^T \right\rangle = \mathbf{\Theta}$ oder gleichbedeutend $\mathbf{K}_{\mathbf{a}} \mathbf{B} \mathbf{K}_{\mathbf{b}^{\perp}}^T = \mathbf{\Theta}$. Dies zeigt folgende Rechnung.

$$
\begin{aligned}
\mathbf{D}_{\mathrm{II}} &= \mathbf{B}\mathbf{K}_{\mathrm{II}}^T \left(\mathbf{K}_{\mathrm{II}} \mathbf{B} \mathbf{K}_{\mathrm{II}}^T \right)^{-1} \\
&= \mathbf{B} \begin{pmatrix} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{b}^{\perp}}^T \end{pmatrix} \left(\begin{pmatrix} \mathbf{K}_{\mathbf{a}} \\ \mathbf{K}_{\mathbf{b}^{\perp}} \end{pmatrix} \mathbf{B} \begin{pmatrix} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{b}^{\perp}}^T \end{pmatrix} \right)^{-1} \\
&= \mathbf{B} \begin{pmatrix} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{b}^{\perp}}^T \end{pmatrix} \begin{pmatrix} \mathbf{K}_{\mathbf{a}} \mathbf{B} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{a}} \mathbf{B} \mathbf{K}_{\mathbf{b}^{\perp}}^T \\ \mathbf{K}_{\mathbf{b}^{\perp}} \mathbf{B} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{b}^{\perp}} \mathbf{B} \mathbf{K}_{\mathbf{b}^{\perp}}^T \end{pmatrix}^{-1} \\
[\mathbf{K}_{\mathbf{a}} \mathbf{B} \mathbf{K}_{\mathbf{b}^{\perp}}^T = \mathbf{\Theta}] \qquad &= \mathbf{B} \begin{pmatrix} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{b}^{\perp}}^T \end{pmatrix} \begin{pmatrix} \left(\mathbf{K}_{\mathbf{a}} \mathbf{B} \mathbf{K}_{\mathbf{a}}^T \right)^{-1} & \mathbf{\Theta} \\ \mathbf{\Theta} & \left(\mathbf{K}_{\mathbf{b}^{\perp}} \mathbf{B} \mathbf{K}_{\mathbf{b}^{\perp}}^T \right)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{B}\mathbf{K}_{\mathbf{a}}^T \left(\mathbf{K}_{\mathbf{a}} \mathbf{B} \mathbf{K}_{\mathbf{a}}^T \right)^{-1} & \mathbf{B}\mathbf{K}_{\mathbf{b}^{\perp}}^T \left(\mathbf{K}_{\mathbf{b}^{\perp}} \mathbf{B} \mathbf{K}_{\mathbf{b}^{\perp}}^T \right)^{-1} \end{pmatrix} \\
[\text{Gl. } 3.21] \qquad &= \begin{pmatrix} \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathbf{a}}, \mathbf{B}) & \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathbf{b}^{\perp}}, \mathbf{B}) \end{pmatrix}
\end{aligned}
\tag{5.1}
$$

Sind also die Vektoren der kodierten Teilräume alter und neuer Neurone miteinander unkorreliert, können die Blockmatrizen des Dekodierers $\mathbf{D}_{\mathrm{II}}$ getrennt voneinander betrachtet werden und es ist $\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathbf{a}}, \mathbf{B})$ und $\mathbf{D}_{\mathrm{II}}^{\mathrm{neu}} = \mathbf{D}^{\mathrm{opt}}(\mathbf{K}_{\mathbf{b}^{\perp}}, \mathbf{B})$. Da wir uns zunächst auf diesen Fall beschränken wollen, können wir, analog zum Kodierer $\mathbf{K}_{\mathrm{II}}$, eine einfachere Notation einführen, die die Optimalität der Dekodierer wiedergibt. Wir schreiben $\mathbf{D}_{\mathbf{a}}^{\mathbf{B}} := \mathbf{D}^{\text{opt}}(\mathbf{K}_{\mathbf{a}}, \mathbf{B})$ und $\mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}} := \mathbf{D}^{\text{opt}}(\mathbf{K}_{\mathbf{b}^\perp}, \mathbf{B})$. Die Indizes drücken die Optimalität für Muster der Kovarianzmatrix $\mathbf{B}$ unter Verwendung der Kodierer $\mathbf{K}_{\mathbf{a}}$ bzw. $\mathbf{K}_{\mathbf{b}^\perp}$ aus. Gilt also die Unkorreliertheit $\mathbf{K}_{\mathbf{a}}\mathbf{B}\mathbf{K}_{\mathbf{b}^\perp}^T = \mathbf{\Theta}$, folgt $\mathbf{D}_{\mathrm{II}} = \begin{pmatrix} \mathbf{D}_{\mathbf{a}}^{\mathbf{B}} & \mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}} \end{pmatrix}$.

Da Kodierer und Dekodierer in Umgebung I für $\mathbf{a}$-Muster optimal sind, schreiben wir auch hier einfach $\mathbf{K}_{\mathrm{I}} = \mathbf{K}_{\mathbf{a}}$ und $\mathbf{D}_{\mathrm{I}} = \mathbf{D}_{\mathbf{a}}^{\mathbf{A}} := \mathbf{D}^{\text{opt}}(\mathbf{K}_{\mathbf{a}}, \mathbf{A})$. Man beachte, dass die Dekodierer $\mathbf{D}_{\mathbf{a}}^{\mathbf{A}}$ und $\mathbf{D}_{\mathbf{a}}^{\mathbf{B}}$ im Allgemeinen nicht identisch sind, obwohl gleiche Kodierergewichte (nämlich $\mathbf{K}_{\mathbf{a}}$) zur Optimierung verwendet wurden (vergl. Gl. 3.21).

5.2 Stabilitätsmaße

Das Fehlermaß $\varepsilon_{\mathbf{A}|\mathbf{B}}$ (Gl. 4.7) quantifiziert den Bekanntheitsgrad der $\mathbf{a}$-Muster nach Adaptation an $\mathbf{b}$-Muster. Wir wollen jetzt mathematisch zeigen, dass $\varepsilon_{\mathbf{A}|\mathbf{B}}$ (in Strategie $S_\perp^{\mathrm{NG}}$) durch die Veränderung der alten Gewichte des Dekodierers hauptsächlich beeinflusst wird. Dabei setzen wir die Unkorreliertheit $\mathbf{K}_{\mathbf{a}}\mathbf{B}\mathbf{K}_{\mathbf{b}^\perp}^T = \mathbf{\Theta}$ voraus und benutzen deshalb im Folgenden die Bezeichnungen aus Abschnitt 5.1.

Laut Definition von $\varepsilon_{\mathbf{A}|\mathbf{B}}$ (Gl. 4.7) gilt bei Nutzung des Kodierers $\mathbf{K}_{\mathrm{II}} = \begin{pmatrix} \mathbf{K}_{\mathbf{a}}^T & \mathbf{K}_{\mathbf{b}^\perp}^T \end{pmatrix}$

$$\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_{\mathrm{II}}) = \mathrm{Sp}\left(\hat{\mathbf{P}}_{\mathrm{II}}^T \hat{\mathbf{P}}_{\mathrm{II}} \mathbf{A}\right), \tag{5.2}$$

wobei $\hat{\mathbf{P}}_{\mathrm{II}} := (\mathbf{I} - \mathbf{D}_{\mathrm{II}}\mathbf{K}_{\mathrm{II}})$ die zugehörige komplementäre Projektion ist (siehe Abs. 3.2.4). Man beachte, dass $\hat{\mathbf{P}}_{\mathrm{II}}$ zwar idempotent, aber i.A. nicht symmetrisch ist, da $\mathbf{K}_{\mathrm{II}}$ als Ganzes i.A. keinen Eigenraum von $\mathbf{B}$ aufspannt (Gl. 3.28); es verhält sich ja nur ein Teil der Neurone plastisch. $\hat{\mathbf{P}}_{\mathrm{II}}$ lässt sich mit den eingeführten Blockmatrizen schreiben (siehe Abs. 5.1):

$$\hat{\mathbf{P}}_{\mathrm{II}} = \mathbf{I} - \begin{pmatrix} \mathbf{D}_{\mathbf{a}}^{\mathbf{B}} & \mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}} \end{pmatrix} \begin{pmatrix} \mathbf{K}_{\mathbf{a}} \\ \mathbf{K}_{\mathbf{b}^\perp} \end{pmatrix} \tag{5.3}$$

$$= \mathbf{I} - \mathbf{D}_{\mathbf{a}}^{\mathbf{B}}\mathbf{K}_{\mathbf{a}} - \mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}}\mathbf{K}_{\mathbf{b}^\perp}, \tag{5.4}$$

Damit kann man in Gl. 5.2 aus multiplizieren

$$\begin{aligned} \varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_{\mathrm{II}}) &= \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_{\mathbf{a}}^{\mathbf{B}}\mathbf{K}_{\mathbf{a}} - \mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}}\mathbf{K}_{\mathbf{b}^\perp}\right)^T \right. \\ &\qquad \left. \left(\mathbf{I} - \mathbf{D}_{\mathbf{a}}^{\mathbf{B}}\mathbf{K}_{\mathbf{a}} - \mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}}\mathbf{K}_{\mathbf{b}^\perp}\right)\mathbf{A}\right) \\ [\mathrm{Sp}(\cdot),\ \text{s.u.}] \qquad &= \mathrm{Sp}\left(\left(\mathbf{I} - 2\,\mathbf{D}_{\mathbf{a}}^{\mathbf{B}}\mathbf{K}_{\mathbf{a}} + \mathbf{K}_{\mathbf{a}}^T\mathbf{D}_{\mathbf{a}}^{\mathbf{B}\,T}\mathbf{D}_{\mathbf{a}}^{\mathbf{B}}\mathbf{K}_{\mathbf{a}} \right.\right. \\ &\qquad \left.\left. -\, 2\,\mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}}\mathbf{K}_{\mathbf{b}^\perp} + \mathbf{K}_{\mathbf{b}^\perp}^T\mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}\ T}\mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}}\mathbf{K}_{\mathbf{b}^\perp}\right)\mathbf{A}\right) \end{aligned} \tag{5.5}$$

Neben Eigenschaften der Spur (angedeutet durch $\mathrm{Sp}(\cdot)$) wurde zur Ableitung von Gl. 5.5 der Sachverhalt $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{A}\mathbf{K}_{\mathbf{a}}^T = \mathbf{\Theta}$ benutzt. Letzteres gilt erstens deswegen, weil, wegen der Optimalität von $\mathbf{K}_{\mathbf{a}}$ auf $\mathbf{a}$-Muster, die Zeilenvektoren von $\mathbf{K}_{\mathbf{a}}$ einen Eigenraum von $\mathbf{A}$ aufspannen (Gl. 3.62); damit lässt sich eine Matrix $\mathbf{M}$ finden, so dass $\mathbf{A}\mathbf{K}_\mathbf{a}^T = \mathbf{K}_\mathbf{a}^T\mathbf{M}$ gilt (vergl. “3. $\Longleftrightarrow$ 4.”, Abschnitt A.1.2). Da zweitens die Gewichtsvektoren neuer und alter Neurone des Kodierers senkrecht zueinander stehen, $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{K}_\mathbf{a}^T = \mathbf{\Theta}$, hat man zusammengenommen $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{A}\mathbf{K}_\mathbf{a}^T = \mathbf{K}_{\mathbf{b}^\perp}\mathbf{K}_\mathbf{a}^T\mathbf{M} = \mathbf{\Theta}$. Bearbeiten alte und neue Neurone **a**-Muster, sind die kodierten Formate bei Verwendung der Strategie $S_\perp^{\mathrm{NG}}$ also stets unkorreliert (im Gegensatz zur Situation in $S_\angle^{\mathrm{NG}}$).

Wenn man davon ausgeht, dass die alten Neurone $\mathbf{K}_\mathbf{a}$ bereits die wesentliche Variabilität der **a**-Muster erfassen, werden die neuen Neurone nur noch unwesentliche Details der **a**-Muster zusätzlich kodieren. Man kann deshalb näherungsweise $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{A} \approx \mathbf{\Theta}$ annehmen. Mit dieser Näherung wird aus Gl. 5.5

$$
\begin{aligned}
&[\mathbf{K}_{\mathbf{b}^\perp}\mathbf{A} \approx \mathbf{\Theta}] && \varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_{\mathrm{II}}) \approx \mathrm{Sp}\left(\left(\mathbf{I} - 2\,\mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a} + \mathbf{K}_\mathbf{a}^T\mathbf{D}_\mathbf{a}^{\mathbf{B}^T}\mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right) \\
& && = \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)^T\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right) && (5.6)\\
&[\text{Gl. } 4.7] && = \varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a}) && (5.7)\\
&[\text{Gl. } 4.8] && = \eta_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a}) && (5.8)
\end{aligned}
$$

Die Wiederherstellungsqualität gespeicherter Muster $\eta_{\mathbf{A}|\mathbf{B}}$ ist wegen der Konstanz der Gewichte alte Neurone im Kodierer für $S_\perp^{\mathrm{NG}}$ identisch mit dem auf die alten Neurone beschränktem Fehlermaß $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a})$. Nun sehen wir auch, dass die Näherung $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{A} \approx \mathbf{\Theta}$ im Experiment zutrifft (0.04 Differenz zwischen den Fehlermaßen, Tab. 4.2). Dies liegt daran, dass die relevante Information von $\mathbf{K}_\mathbf{a}$ bereits kodiert wurden (wegen $\ell = n_{\mathrm{info}}$, Abschnitt 4.5). Der Rekonstruktionsfehler der **a**-Muster mithilfe des Dekodierers $\mathbf{D}_\mathbf{a}^\mathbf{B}$, i.e. $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a})$, ist offensichtlich dann minimal, wenn $\mathbf{D}_\mathbf{a}^\mathbf{B}$ für **A** (anstatt für **B**) optimal wäre, d.h. wenn $\mathbf{D}_\mathbf{a}^\mathbf{B}$ und $\mathbf{D}_\mathbf{a}^\mathbf{A}$ identisch sind. Das Minimum von $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a})$ ist der optimale Rekonstruktionsfehler $\varepsilon_\mathbf{A}(\mathbf{K}_\mathbf{a})$ (Gl. 3.35). Im Allgemeinen gilt jedoch aufgrund der Abhängigkeit des optimalen Dekodierers von der Verteilung $\mathbf{D}_\mathbf{a}^\mathbf{B} \neq \mathbf{D}_\mathbf{a}^\mathbf{A}$ (Gl. 3.21). Das bedeutet insbesondere

$$
\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a}) \geq \varepsilon_\mathbf{A}(\mathbf{K}_\mathbf{a}). \tag{5.9}
$$

Die Rekonstruktion von **a**-Mustern nach der Adaptation an **b**-Muster in Umgebung II kann, für gleiche Anzahl Neurone, nicht besser werden, als sie vor der Adaptation gewesen ist (vergl. Teil (a) und Teil (c) im Experiment, Tab. 4.2). Würde man alle Neurone gemeinsam betrachten, wäre dies zwar prinzipiell möglich (Gl. 3.65), nur ist es unwahrscheinlich, da die neuen Neurone $\mathbf{K}_{\mathbf{b}^\perp}$ ja nicht an **a**-Muster adaptieren, um die Rekonstruktion zu verbessern.

Wir wollen $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a})$ nun weiter umformen

$$
\begin{aligned}
&[\text{Gl. } 4.7] && \varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a}) = \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)^T\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right) \\
&[\mathrm{Sp}\,(\cdot)] && \qquad = \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right) - \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}^T\mathbf{D}_\mathbf{a}^{\mathbf{B}^T}\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right)
\end{aligned} \tag{5.10}
$$

Der erste Summand in Gl. 5.10 ähnelt der Form nach dem optimalen Rekonstruktionsfehler $\varepsilon_\mathbf{A} = \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{A}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right)$. Tatsächlich ist $\mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{A}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right) = \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right)$, obwohl i.A. $\mathbf{D}_\mathbf{a}^\mathbf{B} \neq \mathbf{D}_\mathbf{a}^\mathbf{A}$ gilt. Das liegt daran, dass einerseits $\mathbf{D}_\mathbf{a}^\mathbf{B}$ und $\mathbf{D}_\mathbf{a}^\mathbf{A}$ optimale Dekodierer bzgl. der gleichen Kodierung sind und daher $\mathbf{K}_\mathbf{a}\mathbf{D}_\mathbf{a}^\mathbf{B} = \mathbf{K}_\mathbf{a}\mathbf{D}_\mathbf{a}^\mathbf{A} = \mathbf{I}$ folgt (Gl. 3.24). Andererseits spannt $\mathbf{K}_\mathbf{a}$ einen Eigenraum von $\mathbf{A}$ auf, so dass eine Matrix $\mathbf{M}$ mit der Eigenschaft $\mathbf{K}_\mathbf{a}\mathbf{A} = \mathbf{M}\mathbf{K}_\mathbf{a}$ gefunden werden kann (vergl. "3. $\Longleftrightarrow$ 4.", Abschnitt A.1.2). Man sieht es wie folgt

$$
\begin{aligned}
&[\mathbf{K}_\mathbf{a}\mathbf{A} = \mathbf{M}\mathbf{K}_\mathbf{a}] && \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right) = \mathrm{Sp}\,\mathbf{A} - \mathrm{Sp}\left(\mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{M}\mathbf{K}_\mathbf{a}\right) \\
&[\mathrm{Sp}\,(\cdot)] && = \mathrm{Sp}\,\mathbf{A} - \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}\mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{M}\right) \\
&[\mathbf{K}_\mathbf{a}\mathbf{D}_\mathbf{a}^\mathbf{B} = \mathbf{K}_\mathbf{a}\mathbf{D}_\mathbf{a}^\mathbf{A} = \mathbf{I}] && = \mathrm{Sp}\,\mathbf{A} - \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}\mathbf{D}_\mathbf{a}^\mathbf{A}\mathbf{M}\right) \\
& && = \mathrm{Sp}\left(\left(\mathbf{I} - \mathbf{D}_\mathbf{a}^\mathbf{A}\mathbf{K}_\mathbf{a}\right)\mathbf{A}\right) \\
&[\text{Gl. } 3.35] && = \varepsilon_\mathbf{A}(\mathbf{K}_\mathbf{a})
\end{aligned} \tag{5.11}
$$

In der Herleitung von Gl. 5.11 haben wir die Identität $\mathrm{Sp}\left(\mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{B}}^T\mathbf{A}\right) = \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{A}}^T\mathbf{A}\right)$ gezeigt. Es wurde schon früher bewiesen (Gl. 3.28), dass $\mathbf{D}_\mathbf{a}^\mathbf{A}\mathbf{K}_\mathbf{a}$, wegen der Optimalität beider Matrizen für die gleiche Verteilung, eine symmetrische Projektion ist; deshalb gilt $\mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{A}}^T = \mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{A}}^T\mathbf{D}_\mathbf{a}^\mathbf{A}\mathbf{K}_\mathbf{a}$. Insgesamt lässt sich Gl. 5.10 jetzt zu folgendem Ergebnis auflösen.

$$
\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a}) - \varepsilon_\mathbf{A}(\mathbf{K}_\mathbf{a}) = \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{B}}^T\mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\mathbf{A}\right) - \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{B}}^T\mathbf{A}\right) \tag{5.12}
$$

$$
\begin{aligned}
&= \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{B}}^T\mathbf{D}_\mathbf{a}^\mathbf{B}\mathbf{K}_\mathbf{a}\mathbf{A}\right) - \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}^T{\mathbf{D}_\mathbf{a}^\mathbf{A}}^T\mathbf{D}_\mathbf{a}^\mathbf{A}\mathbf{K}_\mathbf{a}\mathbf{A}\right) \\
&= \mathrm{Sp}\left(\mathbf{K}_\mathbf{a}\mathbf{A}\mathbf{K}_\mathbf{a}^T\left({\mathbf{D}_\mathbf{a}^\mathbf{B}}^T\mathbf{D}_\mathbf{a}^\mathbf{B} - {\mathbf{D}_\mathbf{a}^\mathbf{A}}^T\mathbf{D}_\mathbf{a}^\mathbf{A}\right)\right)
\end{aligned} \tag{5.13}
$$

Da der Dekodierer in Umgebung I durch $\mathbf{D}_\mathrm{I} = \mathbf{D}_\mathbf{a}^\mathbf{A}$ gegeben ist und die Gewichte dieser alten Neurone in Umgebung II laut $S_\perp^\mathrm{NG}$ gerade $\mathbf{D}_\mathrm{II}^\mathrm{alt} = \mathbf{D}_\mathbf{a}^\mathbf{B}$ war – sofern die Unkorreliertheit der Kodierungen $\mathbf{K}_\mathbf{a}\mathbf{B}\mathbf{K}_{\mathbf{b}\perp}^T = \mathbf{\Theta}$ gilt (siehe oben) –, lässt Gl. 5.13 den Schluss zu, dass die Stabilität der alten Gewichte des Dekodierers für die Güte der Wiederherstellung gespeicherter Muster und der Rekonstruktion vermeintlich "bekannter" Eindrücke hauptsächlich verantwortlich ist. Die Veränderung der Dekodierung wird zusätzlich von der Kovarianzmatrix der kodierten Darstellung der $\mathbf{a}$-Muster, i.e. $\mathbf{K}_\mathbf{a}\mathbf{A}\mathbf{K}_\mathbf{a}^T$, gewichtet, um den Zuwachs des Rekonstruktionsfehlers $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K}_\mathbf{a})$ vom Minimum $\varepsilon_\mathbf{A}(\mathbf{K}_\mathbf{a})$ festzulegen.

Da die Veränderung der Gewichte des Dekodierers durch den Adaptationsprozess ausschlaggebend für die Güte der Rekonstruktion gespeicherter Muster ist, können wir uns nun auf die Untersuchung des Dekodierers beschränken.

5.3 Stabilität des Kodes alter Neurone

Wir wollen jetzt analysieren, an welche Art neuer Eindrücke $\mathbf{b}$ das Modellsystem adaptieren kann, ohne $\mathbf{a}$-Muster zu "vergessen". Dazu untersuchen wir, für welche Kovarianzmatrizen $\mathbf{B}$ die Konstanz der alten Gewichte des Dekodierers gewährleistet ist.

In diesem Abschnitt gelte weiterhin $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{B}\mathbf{K}_{\mathbf{a}}^T = \mathbf{\Theta}$; es ist also $\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} = \mathbf{D}_{\mathbf{a}}^{\mathbf{B}}$ und auch $\mathbf{D}_{\mathrm{I}} = \mathbf{D}_{\mathbf{a}}^{\mathbf{A}}$ (vergl. Abschnitt 5.1).

Es lässt sich nachweisen (siehe Abs. A.1.4), dass die Bilder beider Dekodierer, $\mathbf{D}_{\mathbf{a}}^{\mathbf{A}}$ und $\mathbf{D}_{\mathbf{a}}^{\mathbf{B}}$, genau dann identisch sind, wenn $\mathbf{K}_{\mathbf{a}}$, neben dem größten Eigenraum von $\mathbf{A}$, auch einen Eigenraum der Kovarianzmatrix $\mathbf{B}$ aufspannt. Seien nämlich $\mathbf{v}_i$ die Eigenvektoren von $\mathbf{B}$, dann gilt

$$\operatorname{Bild}\mathbf{D}_{\mathbf{a}}^{\mathbf{B}} = \operatorname{Bild}\mathbf{D}_{\mathbf{a}}^{\mathbf{A}} \quad \Longleftrightarrow \quad \operatorname{Bild}\mathbf{K}_{\mathbf{a}}^T = \langle \mathbf{v}_j | j \in \mathcal{J} \subset \mathbb{N}_n \rangle \tag{5.14}$$

Es ist klar, dass die Dimension des Eigenraums $\langle \mathbf{v}_j | j \in \mathcal{J} \rangle$ mit der des Bildes von $\mathbf{K}_{\mathbf{a}}^T$ übereinstimmt, i.e. $|\mathcal{J}| = n_{\mathrm{I}}$. Gleichheit der Bilder der Dekodierermatrizen ist zwar notwendig, aber nicht hinreichend für die Gleichheit der Gewichte der Dekodierer.

Sind die Gewichtsvektoren (Zeilenvektoren) des Kodierers $\mathbf{K}_{\mathbf{a}}$ jedoch orthogonal, ist der zugehörige optimale Dekodierer nicht von der Verteilung abhängig (Gl. 3.29). Dann ist die Gleichheit der Bilder beider Dekodierer in Gl. 5.14 gleichbedeutend mit der Identität der Matrizen, $\mathbf{D}_{\mathbf{a}}^{\mathbf{B}} = \mathbf{D}_{\mathbf{a}}^{\mathbf{A}}$. Orthogonalität der Gewichtsvektoren im Kodierer erweist sich daher als besonders förderlich für die Stabilität des Dekodierers. Wir haben diesen Effekt bereits den Ergebnissen des Experiments entnommen (siehe Abs. 4.5); man vergleiche z.B. die extremen Fehlerwerte für nicht orthogonale Gewichtsvektoren in Tab. 4.3.

Die Auswahl derjenigen Kovarianzmatrizen $\mathbf{B}$, welche die Gewichte des Dekodierers nach Adaptation nicht beeinflussen, ist durch die Bedingung, dass $\mathbf{K}_{\mathbf{a}}$ einen Eigenraum von $\mathbf{B}$ aufspannt, stark eingeschränkt (vergl. Diskussion). Allerdings ist in Gl. 5.14 nicht die Forderung gestellt, dass $\mathbf{K}_{\mathbf{a}}$ den größten Eigenraum aufspannen muss; es kann ein beliebiger sein. Deshalb können wir Verteilungen $P_{\mathbf{b}}$ der $\mathbf{b}$-Muster, die auf die Stabilität des Kodes günstig wirken, in Abschnitt 5.3.2 näher charakterisieren.

Die Äquivalenz Gl. 5.14 beinhaltetet gleichfalls eine Aussage über Verteilungen $P_{\mathbf{b}}$, in denen der Adaptationsprozess die Gewichte des "alten" Dekodierers trotz der Konservierung der Gewichte des Kodierers $\mathbf{K}_{\mathbf{a}}$ verändert. Gilt die nämlich Unkorreliertheit der kodierten Darstellungen $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{B}\mathbf{K}_{\mathbf{a}}^T = \mathbf{\Theta}$, sind es all jene Verteilungen $P_{\mathbf{b}}$, in denen $\mathbf{K}_{\mathbf{a}}$ keinen Eigenraum von $\mathbf{B}$ aufspannt. Im folgenden Unterabschnitt geben wir ein Beispiel für eine derartige Verteilung.

5.3.1 "Ungünstigste" Verteilung

Anhand eines dreidimensionalen Beispiels soll eine mögliche Modifikation der Gewichte des Dekodierers nach dem Adaptationsprozess veranschaulicht werden. Es wird also $\mathbf{D}_{\mathbf{a}}^{\mathbf{A}} \neq \mathbf{D}_{\mathbf{a}}^{\mathbf{B}}$ nachgewiesen. Zugleich wird die zweidimensionale Verteilung angegeben, welche die Gewichte des Dekodierers, bei vorgegebenen Spektrum der Verteilung, am stärksten beeinflusst, d.h. seine Norm in Umgebung II maximiert.

Es wird "rückwärts" vorgegangen, indem zuerst $\mathbf{B}$ und $\mathbf{K}_{\mathbf{b}^\perp}$ und danach $\mathbf{K}_{\mathbf{a}}$ und $\mathbf{A}$ festgelegt werden. Dies hat allein den Vorteil, dass die Gleichungen übersichtlicher werden, weil die Koordinatenachsen durch die Hauptkomponenten von $\mathbf{B}$ vorgegeben werden können. Es sei also

$$\mathbf{B} = \mathrm{diag}\,(\lambda_1, \lambda_2, \lambda_3) \tag{5.15}$$

die dreidimensionale Kovarianzmatrix der $\mathbf{b}$-Muster, mit $\lambda_1 \geq \lambda_2 \geq \lambda_3 > 0$. Der Kodierer $\mathbf{K_a}$ in Umgebung I sei eindimensional, in Umgebung II komme eine weitere Dimension hinzu, d.h. $\mathbf{K}_{\mathrm{II}} = \begin{pmatrix}\mathbf{K_a} & \mathbf{K_{b^\perp}}\end{pmatrix}$. Um mit diesen Voraussetzungen die Bedingung $\mathbf{K_{b^\perp}BK_a}^T = \mathbf{\Theta}$, die Unkorreliertheit beider Teilräume des Kodierers $\mathbf{K}_{\mathrm{II}}$, zu erfüllen, müssen $\mathbf{K_a}$ und $\mathbf{K_{b^\perp}}$ Linearkombination unterschiedlicher Eigenvektoren von $\mathbf{B}$ sein (siehe auch Gl. 3.27). Es gibt also zunächst zwei qualitativ verschiedene Möglichkeiten, durch Festlegung von $\mathbf{K_a}$ die Unkorreliertheit zu erreichen: (i) $\mathbf{K_a}$ liegt in einem eindimensionalen Eigenraum von $\mathbf{B}$, also parallel zu einer Koordinatenachse, oder (ii) $\mathbf{K_a}$ liegt in einer Ebene die von zwei Eigenvektoren von $\mathbf{B}$ aufgespannt wird (und es gilt nicht (i)). Laut Aussage Gl. 5.14 würde im Fall (i) der Dekodierer stabil bleiben, also $\mathbf{D_a^A} = \mathbf{D_a^B}$. Wir interessieren uns daher für (ii) und wählen für den Kodierer $\mathbf{K_a}$ die Ebene, die durch $\mathbf{e}_2$ und $\mathbf{e}_3$ aufgespannt wird; damit steht $\mathbf{K_a}$ senkrecht zu $\mathbf{e}_1$. Das neue Neuron im Kodierer, $\mathbf{K_{b^\perp}}$, adaptiert an den Teil eines $\mathbf{b}$-Musters, der orthogonal zu $\mathbf{K_a}$ ist. Es lernt also den größten Eigenvektor von $\mathbf{B}^\perp$ (und nicht von $\mathbf{B}$). Da nach Konstruktion die größte Hauptkomponente $\mathbf{e}_1$ von $\mathbf{B}$, stets senkrecht zu $\mathbf{K_a}$ ist, ist sie auch (größter) Eigenvektor der Matrix $\mathbf{B}^\perp$, so dass $\mathbf{K_{b^\perp}}$ diese Richtung kodiert; es ist also $\mathbf{K_{b^\perp}} = \gamma\, \mathbf{e}_1{}^T$, $\gamma \in \mathbb{R}$.

Nach Konstruktion haben wir damit $\mathbf{K_a B K}_{\mathbf{b}^\perp}^T = 0$ erreicht und wissen nun, dass die Blockmatrizen des Dekodierers $\mathbf{D}_{\mathrm{II}}$ unabhängig voneinander zu betrachten sind (Gl. 5.1). Wir können daher $\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} = \mathbf{D_a^B}$ und $\mathbf{D}_{\mathrm{II}}^{\mathrm{neu}} = \mathbf{D}_{\mathbf{b}^\perp}^{\mathbf{B}}$ schreiben (vergl. Abschnitt 5.1). Wir interessieren uns für die Stabilität der alten Neurone im Dekodierer, müssen daher $\mathbf{D_a^B}$ mit $\mathbf{D_a^A}$ vergleichen.

Da sich wegen der Unkorreliertheit des alten und neuen Kodierers, $\mathbf{K_a B K}_{\mathbf{b}^\perp}^T = 0$, alle interessierenden Größen in der Ebene senkrecht zu $\mathbf{K}_{\mathbf{b}^\perp}^T = \gamma \mathbf{e}_1$ befinden, können wir uns auf die Betrachtung dieser Ebene beschränken. Damit sind die $\mathbf{b}$-Muster effektiv zweidimensional:

$$\tilde{\mathbf{B}} = \mathrm{diag}\,(\lambda_2, \lambda_3) \tag{5.16}$$

Man beachte, dass dadurch $\mathbf{K_a}$, wie auch $\mathbf{D_a^B}$ und $\mathbf{D_a^A}$, zwei- statt dreidimensionale Vektoren sind, wir belassen es aber o.B.d.A. bei der gleichen Notation.

Im Abschnitt 5.2 haben wir gesehen, dass Veränderung der alten Dekodierergewichte direkten Einfluss auf die Wiederherstellungsqualität gespeicherter Muster, $\eta_{\mathbf{A}|\mathbf{B}}(\mathbf{K_a})$, und gleichermaßen auf den "Bekanntheitsgrad" alter Eindrücke, $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K_a})$, ausübt (Gl. 5.13). In der Tat ist für einen eindimensionalen Kodierer die Maximierung der Norm von $\mathbf{D_a^B}$ unter Variation der Kovarianzmatrix $\mathbf{B}$ gleichbedeutend mit der Maximierung von $\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K_a})$. Denn nehmen wir übersichtshalber $\mathbf{K_a}$ als normiert an, $\mathbf{K_a K_a}^T = 1$, folgt zunächst $\mathbf{D}_\mathbf{a}^\mathbf{A} = \mathbf{K}_\mathbf{a}^T$ (Gl. 3.29) und mit Gl. 5.13:

$$\begin{aligned} \varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K_a}) - \varepsilon_{\mathbf{A}}(\mathbf{K_a}) &= \mathrm{Sp}\left(\mathbf{K_a}\mathbf{A}\mathbf{K}_\mathbf{a}^T\left({\mathbf{D}_\mathbf{a}^\mathbf{B}}^T\mathbf{D}_\mathbf{a}^\mathbf{B} - {\mathbf{D}_\mathbf{a}^\mathbf{A}}^T\mathbf{D}_\mathbf{a}^\mathbf{A}\right)\right) \\ \text{[s.o. , } \mathbf{K_a} \sim \mathrm{A}_1\text{]} \qquad &= \mu_1\left(|\mathbf{D}_\mathbf{a}^\mathbf{B}|^2 - 1\right), \end{aligned} \tag{5.17}$$

wobei μ_1 der größte Eigenwert von $\mathbf{A}$ sei, der ja zum Eigenvektor $\mathbf{K_a}$ (wg. der Optimalität $\mathbf{K_a} \sim \mathrm{A}_1$) gehört.

Wir werden nun die Norm des Dekodierers $\mathbf{D}_\mathbf{a}^\mathbf{B}$ für ein beliebig vorgegebenes Spektrum maximieren (Gl. 5.15), d.h. die Drehung der Kovarianzmatrix $\mathbf{B}$ zu $\mathbf{A}$ variieren. Um eine Matrix $\mathbf{B}$ gemäß den gemachten Voraussetzungen zu wählen, legen wir jedoch die Lage des Kodierers $\mathbf{K_a}$ (anstatt von $\mathbf{B}$) in der Ebene fest. Da $\mathbf{K_a}$ stets der größte Eigenvektor von $\mathbf{A}$ ist, definieren wir dadurch verschiedene $\mathbf{B}$, nur dass wir stets die Hauptkomponenten von $\mathbf{B}$ als Koordinatenachsen benutzen (siehe oben). Es sei also $\mathbf{K_a}$ eine beliebiger, übersichtshalber normierter Vektor im $\mathbb{R}^2$:

$$\mathbf{K_a}(\varphi) := \begin{pmatrix} \sin\varphi & \cos\varphi \end{pmatrix}. \tag{5.18}$$

Damit ist der Dekodierer in Umgebung I gerade durch den transponierten Gewichtsvektor des Kodierers $\mathbf{D}_\mathbf{a}^\mathbf{A}(\varphi) = \mathbf{K}_\mathbf{a}^T(\varphi)$ (wg. Gl. 3.29) gegeben und nach den Ergebnissen des vorhergehenden Beispiels folgt (Abschnitt 3.2.7, Gl. 3.72):

$$\begin{aligned} \mathbf{D}_\mathbf{a}^\mathbf{B}(\varphi) &= \tilde{\mathbf{B}}\mathbf{K_a}(\varphi)^T\left(\mathbf{K_a}(\varphi)\tilde{\mathbf{B}}\mathbf{K_a}(\varphi)^T\right)^{-1} \\ &= \frac{1}{\lambda_2\sin^2\varphi + \lambda_3\cos^2\varphi}\begin{pmatrix} \lambda_2\sin\varphi \\ \lambda_3\cos\varphi \end{pmatrix} \end{aligned} \tag{5.19}$$

Die Norm $|\mathbf{D}_\mathbf{a}^\mathbf{B}(\varphi)|$ schreibt sich

$$|\mathbf{D}_\mathbf{a}^\mathbf{B}(\varphi)| = \frac{\sqrt{\lambda_2^2\sin^2\varphi + \lambda_3^2\cos^2\varphi}}{\lambda_2\sin^2\varphi + \lambda_3\cos^2\varphi} \tag{5.20}$$

Nun können wir die Extrema ermitteln. Neben den eher ersichtlichen Extrema $\varphi \in \{z\frac{\pi}{2}|z \in \mathbb{Z}\}$, das sind die Nullstellen von $\cos^2\varphi \cdot \sin^2\varphi$, findet man nach etwas Algebra weitere bei

$$\varphi^* = z\pi + \arccos\left(\pm\sqrt{\frac{\lambda_2}{\lambda_2+\lambda_3}}\right). \tag{5.21}$$

Während $|\mathbf{D}_\mathbf{a}^\mathbf{B}(z\frac{\pi}{2})| = 1$ ist, gilt

$$|\mathbf{D}_\mathbf{a}^\mathbf{B}(\varphi^*)| = \frac{1}{2}\frac{\lambda_2+\lambda_3}{\sqrt{\lambda_2\lambda_3}}. \tag{5.22}$$

Da für positive Zahlen das arithmetische Mittel größer gleich dem geometrischen ist, sind die Stellen φ^* stets Maxima. Sind die Hauptkomponenten der Verteilungen $\mathbf{A}$ und $\mathbf{B}$ um den Winkel φ^* zueinander gedreht, ergibt sich nach dem Adaptationsprozess der alten Neurone des Dekodierer an $\mathbf{b}$-Muster die maximale Norm des Gewichtsvektors (bei vorgegebenen Eigenwerten).

Als Abschluss des Beispiels berechnen wir den Rekonstruktionsfehler vermeintlich "bekannter" $\mathbf{a}$-Muster an der Stelle φ^*, der ja (bei vorgegebenen Eigenwerten) am größten ist (Gl. 5.17):

$$\begin{aligned}\varepsilon_{\mathbf{A}|\mathbf{B}}(\mathbf{K_a}(\varphi^*)) - \varepsilon_{\mathbf{A}}(\mathbf{K_a}(\varphi^*)) &= \mu_1 \left(\frac{1}{4} \frac{(\lambda_2 + \lambda_3)^2}{\lambda_2 \lambda_3} - 1 \right) \\ &= \mu_1 \frac{1}{4} \frac{(\lambda_2 - \lambda_3)^2}{\lambda_2 \lambda_3} \end{aligned} \tag{5.23}$$

Man erkennt die Abhängigkeit von den Eigenwerten gut. Je wichtiger die kodierte Hauptkomponente $\mathbf{K_a}(\varphi^*)$ ist, d.h. je größer der zugehörige Eigenwert μ_1 ist, desto schlechter wirkt sich natürlich eine verzerrte Rekonstruktion aus.

Interessanterweise hängt die Größe der Norm (und damit auch die Größe des Rekonstruktionsfehlers $\varepsilon_{\mathbf{A}|\mathbf{B}}$) kritisch von dem Verhältnis der durch die neuen Neurone $\mathbf{K}_{\mathbf{b}^\perp}$ nicht kodierten Eigenwerte von $\mathbf{B}$ ab, i.e. λ_2 und λ_3. Falls nämlich $\lambda_2 \gg \lambda_3$ gilt, die Hierarchie dieser Eigenwerte also stark ausgeprägt ist, wird die Norm des Dekodierers $|\mathbf{D}_{\mathbf{a}}^{\mathbf{B}}(\varphi^*)|$ sehr groß; im Grenzfall $\lambda_3 \to 0$ bei konstantem λ_2 divergiert sie sogar:

$$|\mathbf{D}_{\mathbf{a}}^{\mathbf{B}}(\varphi^*)| = \frac{1}{2} \frac{\lambda_2 + \lambda_3}{\sqrt{\lambda_2 \lambda_3}} \approx \frac{1}{2} \sqrt{\frac{\lambda_2}{\lambda_3}} \xrightarrow{\lambda_3 \to 0} \infty \tag{5.24}$$

Sind die Eigenwerte λ_i, $i = 2, 3$ dagegen identisch, bleibt der Dekodierer unverändert; die Richtung des Kodierers und Dekodierers ist identisch.

Es sei bemerkt, dass eine Vergrößerung der Norm des optimalen Dekodierers immer auch eine Zunahme des Winkels zwischen Kodierungsvektor und zugehörigem optimalem Dekodierungsvektor zur Folge hat (und andersherum). Denn durch die Bedingung $\mathbf{K}\mathbf{D}^{\text{opt}} = \mathbf{I}$ (Gl. 3.24), i.e. $\mathbf{k}_i^T \mathbf{d}_j = \delta_{ij}$ in Vektorschreibweise, ist das Skalarprodukt zwischen Gewichtsvektoren des Kodierers und des optimalen Dekodierers festgelegt. Verlängert sich der Vektor $\mathbf{d}_j$, muss er gleichzeitig von $\mathbf{k}_j$ "weg gedreht" werden, um $\mathbf{k}_j^T \mathbf{d}_j = 1$ aufrecht zu erhalten.

Die Veränderung der Norm und Richtung des Dekodierers verbessern den Rekonstruktionsfehler der $\mathbf{b}$-Muster, weil die Rekonstruktion dadurch den größten Hauptkomponenten der Verteilung angepasst wird. Da das kodierte $\mathbf{b}$-Muster, $\mathbf{K_a b}$, mit Richtungen der Verteilung korreliert ist, deren Varianz stärker ist als in Richtung des Kodierers $\mathbf{K_a}$, kann der Dekodierer die Rekonstruktion verbessern, indem er die kodierte, niedrigdimensionale Form der Daten in Richtungen stärkerer Varianz in den Ursprungsraum $\mathbb{R}^n$ zurück projiziert (vergl. Abb. 5.1). Der Effekt ist, dass der Dekodierer $\mathbf{D}_{\mathbf{a}}^{\mathbf{B}}$ nicht in die gleiche Richtung zeigt wie der Kodierer $\mathbf{K_a}$. Da $\mathbf{D}_{\mathbf{a}}^{\mathbf{A}}$ dagegen, wegen der Optimalität von Kodierer und Dekodierer auf $\mathbf{a}$-Muster, die kodierte Darstellung "korrekt" zurück projiziert, also in den Unterraum des Ursprungsraums, aus dem der Kodierer $\mathbf{K_a}$ die kodierte Darstellung gewonnen hat (Bild $\mathbf{K}_{\mathbf{a}}^T$ = Bild $\mathbf{D}_{\mathbf{a}}^{\mathbf{A}}$, siehe Gl. 3.28), verändern sich also die Gewichtsvektoren des Dekodierers nach Adaptation, d.h. $\mathbf{D}_{\mathbf{a}}^{\mathbf{B}} \neq \mathbf{D}_{\mathbf{a}}^{\mathbf{A}}$. In der Konsequenz dieser Veränderung können **a**-Muster trotz Stabilität des Kodierers nicht mehr korrekt dekodiert werden (Abb. 5.1 F).

Der Kodierer $\mathbf{K_a}(\varphi^*)$ ist derart, dass die Hauptkomponenten von $\mathbf{B}$, die mit dem Vektor $\mathbf{K_a}(\varphi^*)$ korreliert sind, in der kodierten Darstellung identische Standardabweichung besitzen, denn es ist[1] $\mathbf{K_a}(\varphi^*)\mathbf{e}_i\sqrt{\lambda_i} = \text{konst.}$, $i = 2, 3$. War die Datenverteilung ursprünglich z.B. zigarrenförmig, liegen alle "Randpunkte" in der kodierten Darstellung übereinander. Es wird dann die Rekonstruktion dadurch optimiert, dass der Dekodierer in die Längsachse der Zigarre gedreht und gestreckt wird (siehe Abb. 5.1).

5.3.2 Folgerungen

Wir haben uns bis jetzt auf den Fall der Strategie $S_\perp^{\mathrm{NG}}$ beschränkt, in dem die alten und neuen Kodiererneurone, $\mathbf{K_a}$ bzw. $\mathbf{K_{b^\perp}}$, in orthogonalen Eigenräumen der **b**-Verteilung "leben"; mit den Eigenvektoren[2] $\mathbf{v}_i$, $i \in \mathbb{N}_n$, von $\mathbf{B}$ galt nämlich stets

$$\mathrm{Bild}\,\mathbf{K_a}^T \subset \langle \mathbf{v}_i | i \in \mathcal{I}_{\mathrm{alt}} \subset \mathbb{N}_n \rangle \quad \text{und} \quad \mathrm{Bild}\,\mathbf{K}_{\mathbf{b}^\perp}^T \subset \langle \mathbf{v}_j | j \in \mathbb{N}_n \backslash \mathcal{I}_{\mathrm{alt}} \rangle \tag{5.25}$$

Denn die Unkorreliertheit der Vektoren beider kodierten Unterräume ist notwendig für die Unkorreliertheit ihrer kodierten Darstellungen, i.e. $\mathbf{K_{b^\perp} B K_a}^T = \mathbf{\Theta}$ (folgt aus der Linearität der Kodierung). Letzteres haben wir in den vorangehenden Abschnitten vorausgesetzt und dazu zwei qualitativ unterschiedliche Situationen vorgefunden (mit $\mathcal{I}_{\mathrm{alt}}$ aus Gl. 5.25):

(i) Die Gewichte der alten Neurone im Kodierer, i.e. $\mathbf{K_a}$, spannen einen Eigenraum von $\mathbf{B}$ auf, d.h. $|\mathcal{I}_{\mathrm{alt}}| = n_{\mathrm{I}}$.

(ii) $\mathrm{Bild}\,\mathbf{K_a}^T$ ist ein echter Teilraum eines Eigenraums von $\mathbf{B}$, d.h. $|\mathcal{I}_{\mathrm{alt}}| > n_{\mathrm{I}}$, und es gilt nicht (i).

[1] Anmerkung (ohne Beweis): Man kann diesen ungünstigen Kodierungsvektor $\mathbf{K_a}(\varphi^*)$ auch im $\mathbb{R}^n$ finden, dann ist er nämlich durch folgende Gleichung definiert

$$\mathbf{k} := \frac{1}{c} \sum_{i=n_{\mathrm{II}}+1}^{n} \frac{1}{\sqrt{\lambda_i}} \mathbf{v}_i$$

Dabei ist $c^2 := \sum_{j=n_{\mathrm{II}}+1}^{n} \frac{1}{\lambda_j}$ eine Normierungskonstante, so dass $\mathbf{k}^T\mathbf{k} = 1$. Die Eigenvektoren $\mathbf{v}_i$ von $\mathbf{B}$ mit den Indizes $n_{\mathrm{II}}+1, \ldots, n$ sind diejenigen, die zwar mit dem Kodierer $\mathbf{K}_{\mathrm{II}}$ korreliert, aber nicht vollständig kodiert sind. Ist die Hierarchie der zugehörigen Eigenwerte stark ausgeprägt, wird der optimale Dekodierer $\mathbf{D_a^B}$ gegenüber $\mathbf{D_a^A}$ stark verzerrt (vergl. Gl. 5.24). Die zu Gl. 5.22 analoge Gleichung der Norm des Dekodierers ist dann

$$\mathbf{D_a^B}^T \mathbf{D_a^B} = \frac{1}{(n - n_{\mathrm{II}})^2} \left(\sum_{i=n_{\mathrm{II}}+1}^{n} \lambda_i \right) \left(\sum_{i=n_{\mathrm{II}}+1}^{n} \frac{1}{\lambda_i} \right)$$

[2] Die n Eigenvektoren $\mathbf{v}_i$ seien stets so gewählt, dass sie zusammengenommen eine Orthonormalbasis des $\mathbb{R}^n$ bilden.

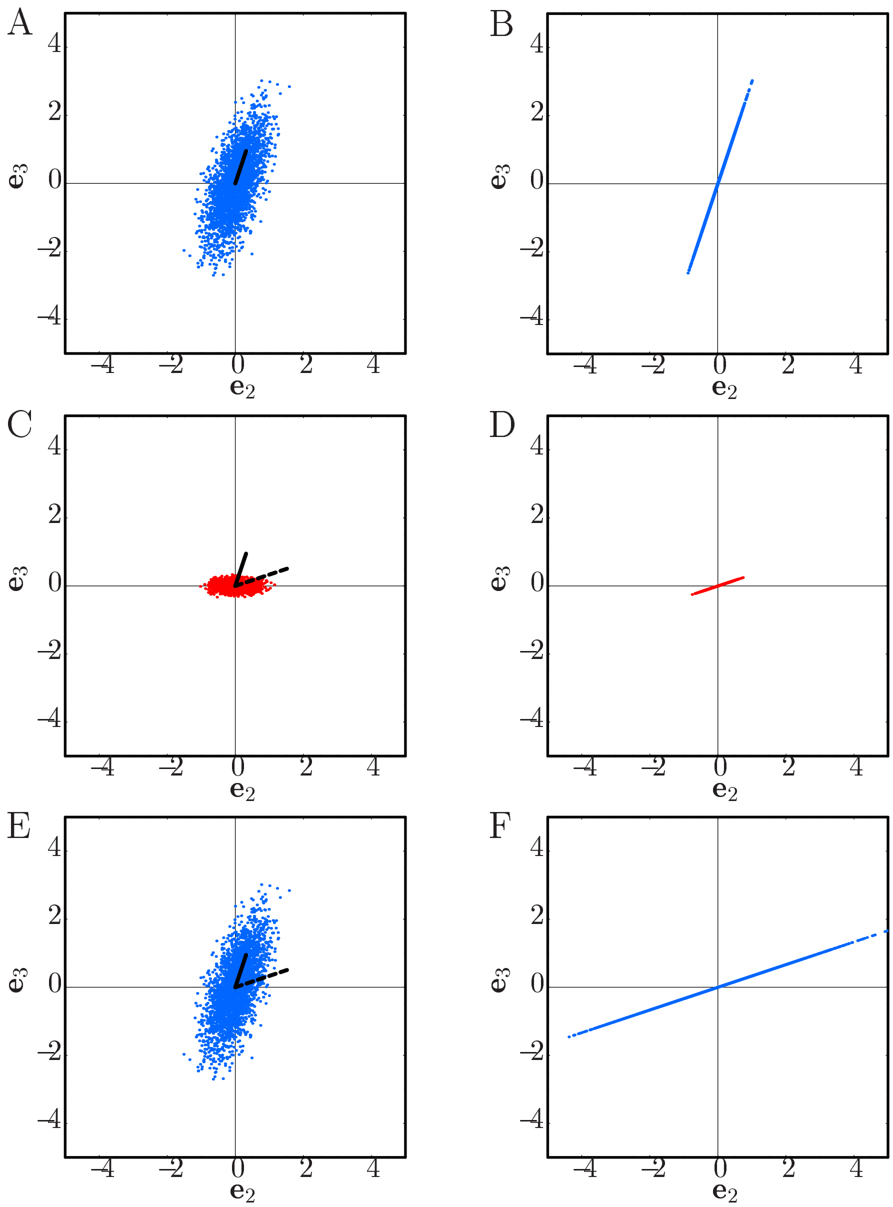

Abbildung 5.1: Beispiel für die Veränderung des Dekodierers trotz stabilem Kodierer. Abb. A und B entsprechen Teil (a), Abb. C und D Teil (b) und Abb. E und F Teil (c) des Experiments (Abb. 4.1). *Linke Seite*: Multivariate Normalverteilungen mit Kodierer und Dekodierer. *Rechte Seite*: Rekonstruktion der Muster der gegenüberliegenden Verteilung. *Abbildung A*: Optimaler Kodierer $\mathbf{K_a}(\varphi^*)$ (durchgezogene Linie) und optimaler Dekodierer $\mathbf{K_a}(\varphi^*)^T$ (sie liegen übereinander). Verteilung: Eigenwerte der auf $\mathbf{K}_{\mathbf{b}^\perp} = \gamma \mathbf{e}_1{}^T$ projizierten Kovarianzmatrix $\mathbf{A}$, i.e. $\tilde{\mathbf{A}}$, sind $\mu_1 = 0.9$ und $\mu_2 = 0.09$. Die dritte Eigenrichtung von $\mathbf{A}$, mit Eigenwert $\mu_3 = 0.01$, liegt in Richtung $\mathbf{e}_1$ und ist daher nicht zu sehen. *Abbildung B*: Rekonstruktion mit optimalem Kodierer und Dekodierer aus Abb. A. *Abbildung C*: Verteilung $\tilde{\mathbf{B}}$, die Projektion von $\mathbf{B}$ in den Kern von $\mathbf{K}_{\mathbf{b}^\perp} = \gamma \mathbf{e}_1{}^T$. $\mathbf{B}$ hat die gleichen Eigenwerte wie $\mathbf{A}$, $\lambda_1 = 0.9, \lambda_2 = 0.09$ und $\lambda_3 = 0.01$, liegt aber anders im $\mathbb{R}^3$: die Koordinatenachsen entsprechen den Eigenvektoren von $\mathbf{B}$. Der größte Eigenvektor zeigt in Richtung $\mathbf{e}_1$ und wird durch $\mathbf{K}_{\mathbf{b}^\perp}$ kodiert. Kodierer $\mathbf{K_a}(\varphi^*)$ (durchgezogene Linie) und auf $\mathbf{B}$ optimierter Dekodierer $\mathbf{D}^{\mathbf{B}}_{\mathbf{a}}(\varphi^*)$ (gestrichelte Linie) sind eingezeichnet. Der Dekodierer ist offensichtlich nach Umgebungswechsel verzerrt, also $\mathbf{D}^{\mathbf{A}}_{\mathbf{a}} = \mathbf{D}_{\mathrm{I}} \neq \mathbf{D}^{\mathrm{alt}}_{\mathrm{II}} = \mathbf{D}^{\mathbf{B}}_{\mathbf{a}}$ (vergl. Abb. A). *Abbildung D*: Rekonstruktion der Muster aus Abb. C. *Abbildung E*: Um die Rekonstruktion von $\mathbf{a}$-Mustern nach dem Adaptationsprozess an $\mathbf{b}$-Muster zu testen ist erneut $\tilde{\mathbf{A}}$ aus Abb. A zu sehen. Nun wird Kodierer $\mathbf{K_a}(\varphi^*)$ (durchgezogene Linie) und Dekodierer $\mathbf{D}^{\mathbf{B}}_{\mathbf{a}}$ (gestrichelte Linie) anstatt $\mathbf{D}^{\mathbf{A}}_{\mathbf{a}}$ verwendet. *Abbildung F*: Rekonstruktion der Verteilung $\tilde{\mathbf{A}}$. Man erkennt deutlich den Unterschied zu Abb. B; die Rekonstruktion der vermeintlich bekannten $\mathbf{a}$-Muster ist durch den Dekodierer wegen $\mathbf{D}^{\mathbf{A}}_{\mathbf{a}} \neq \mathbf{D}^{\mathbf{B}}_{\mathbf{a}}$ gedreht und gestreckt.

Fall (i)

Tritt Fall (i) ein (vergl. Abb. 5.2 A), wird der kodierte Raum der **b**-Verteilung auch zur Dekodierung benutzt, d.h. Bild $\mathbf{K}_\mathbf{a}^T = \text{Bild}\,\mathbf{D}_\mathbf{a}^\mathbf{B}$. Wird beispielsweise eine Ebene von den Gewichtsvektoren des Kodierers $\mathbf{K}_\mathbf{a}$ im Raum der Feuerraten der EC-Neurone, $\mathbb{R}^n$, definiert, projiziert der Dekodierer im Fall (i) auf diese Ebene zurück. Die Rekonstruktion der Daten besteht dann nur aus Varianz in der durch den Kodierer festgelegten Ebene. Dies ist auch vor dem Umgebungswechsel so, weil Kodierer $\mathbf{K}_\mathbf{a}$ und Dekodierer $\mathbf{D}_\mathbf{a}^\mathbf{A}$ in Umgebung I auf die **a**-Muster optimiert sind und deshalb Bild $\mathbf{K}_\mathbf{a}^T = \text{Bild}\,\mathbf{D}_\mathbf{a}^\mathbf{A}$ folgt (Gl. 3.28). Sind die Gewichtsvektoren des Kodierers $\mathbf{K}_\mathbf{a}$ nun zusätzlich orthogonal, kann sogar die Konstanz der Gewichte des Dekodierers nach der Adaptation an **b**-Muster, i.e. $\mathbf{D}_\mathbf{a}^\mathbf{A} = \mathbf{D}_\mathbf{a}^\mathbf{B}$ vorausgesagt werden. Durch die Stabilität der alten Dekodierergewichte wird eine getreue Rekonstruktion von **a**-Muster gewahrt.

Fall (ii)

Liegt der Fall (ii) vor (vergl. Abb. 5.2 D), wird der Dekodierer durch den Adaptationsprozess verändert, i.e. $\mathbf{D}_\mathbf{a}^\mathbf{A} \neq \mathbf{D}_\mathbf{a}^\mathbf{B}$. Das liegt daran, dass der kodierte Raum, Bild $\mathbf{K}_\mathbf{a}^T$, mit mehr als n_I Hauptkomponenten der **b**-Verteilung korreliert ist. Damit projiziert der Dekodierer in Umgebung II nicht mehr in den durch den Kodierer $\mathbf{K}_\mathbf{a}$ definierten Raum, denn es ist Bild $\mathbf{K}_\mathbf{a}^T \neq \text{Bild}\,\mathbf{D}_\mathbf{a}^\mathbf{B}$ (Gl. 3.28). Würde Kodierer $\mathbf{K}_\mathbf{a}$ in einem solchen Fall eine Ebene im $\mathbb{R}^n$ definieren, bestünde die Rekonstruktion der Daten nach der Adaptation aus Varianz in einer anderen Ebene (nämlich in der durch die Gewichtsvektoren des Dekodierers $\mathbf{D}_\mathbf{a}^\mathbf{B}$ definierten Ebene). Diese Verzerrung des Dekodierers, löst eine Verschlechterung der Rekonstruktion vermeintlich bekannter **a**-Muster aus, denn die bestmögliche Dekodierung dieser Muster wäre $\mathbf{D}_\mathbf{a}^\mathbf{A}$.

Ist die Hierarchie der Eigenwerte $\lambda_i,\ i \in \mathcal{I}_\mathrm{alt}$, des Eigenraums $\langle \mathbf{v}_i | i \in \mathcal{I}_\mathrm{alt} \rangle$ kleinster Dimension, in dem das Bild des Kodierers $\mathbf{K}_\mathbf{a}$ echt enthalten ist (siehe Gl. 5.25), ausgeprägt, ist eine starke Abweichung des Dekodierers $\mathbf{D}_\mathbf{a}^\mathbf{B}$ gegenüber $\mathbf{D}_\mathbf{a}^\mathbf{A}$ möglich. Wir haben dies im Beispiel des Abschnitts 5.3.1 verdeutlicht.

Grenzfall

Sind im Fall (ii) die Eigenwerte $\lambda_i,\ i \in \mathcal{I}_\mathrm{alt}$ alle gleich, $\lambda_i = \lambda_j,\ i, j \in \mathcal{I}_\mathrm{alt}$, erfolgt keine Verzerrung, sondern die Gewichte des Dekodierers bleiben erhalten, $\mathbf{D}_\mathbf{a}^\mathbf{A} = \mathbf{D}_\mathbf{a}^\mathbf{B}$ (siehe auch Gl. 5.23). In diesem Grenzfall ist nämlich ein beliebiger Vektor des Kodierers $\mathbf{K}_\mathbf{a}$ bereits ein (degenerierter) Eigenvektor von $\mathbf{B}$. Es ist also genau genommen ein Spezialfall von (i) (Abb. 5.2 B).

Die Gleichheit der Eigenwerte $\lambda_i,\ i \in \mathcal{I}_\mathrm{alt}$, bedeutet, dass die Variabilität der **b**-Muster in jeder Richtung des Unterraums $\langle \mathbf{v}_i | i \in \mathcal{I}_\mathrm{alt} \rangle$ gleichverteilt ist. Diese Struktur der Verteilung nannten wir im Experiment "Rauschen" und identifizierten es mit "unwesentlichen Details" der Muster (Abschnitt 4.4). Im Experiment wurde die Kovarianzmatrix $\mathbf{A}$ in beliebiger Weise gedreht, um die Kovarianzmatrix $\mathbf{B}$ zu erzeugen.

Durch den "Rauschanteil" von 3 : 1 (45 : 15) wurde annährend die Gleichheit der Eigenwerte $\lambda_i, i \in \mathcal{I}_{\text{alt}}$, erreicht; die alten Kodiererneurone $\mathbf{K_a}$ haben nur das "Rauschen" der **b**-Muster "gesehen" und die wesentliche Variabilität von **b** liegt orthogonal zu $\mathbf{K_a}$ (Voraussetzung Gl. 5.25). Unter anderem deshalb ist die Rekonstruktion der **a**-Muster nach dem Anpassungsprozess in $S_\perp^{\text{NG}}$ verglichen mit den anderen Strategien recht gut (vergl. Tab. 4.2, Teil (c)).

5.3.3 Neurogenese induziert Stabilität

Wir haben bis jetzt die Unkorreliertheit der Kodierung neuer und alter Neurone, $\mathbf{K_a}$ und $\mathbf{K_{b^\perp}}$, für **b**-Muster vorausgesetzt, i.e. $\mathbf{K_a B K_{b^\perp}^T} = \mathbf{\Theta}$. Dies ist in der Strategie $S_\perp^{\text{NG}}$ nicht notwendigerweise erfüllt. Zwar sind neue und alte Kodierer zueinander orthogonal, $\mathbf{K_a K_{b^\perp}^T} = \mathbf{\Theta}$, aber $\mathbf{K_{b^\perp}}$ spannt keinen Eigenraum von $\mathbf{B}$ auf, sondern einen Eigenraum von $\mathbf{B}^\perp$. Die Eigenvektoren beider Matrizen sind i.A. nicht identisch. Denn während alle (nicht degenerierten) Eigenvektoren von $\mathbf{B}^\perp = \hat{\mathbf{P}}_\text{I}\mathbf{B}\hat{\mathbf{P}}_\text{I}$ (Gl. 4.1) senkrecht zu $\mathbf{K_a}$ stehen ($\hat{\mathbf{P}}_\text{I} = (\mathbf{I} - \mathbf{D_a^A}^T\mathbf{K_a})$ ist eine symmetrische Projektion auf den Kern des Kodierers $\mathbf{K_a}$, siehe Gl. 3.28), können die Eigenrichtungen von $\mathbf{B}$ beliebig ausgerichtet sein.

Wenn die Kodierungen beider Blockmatrizen $\mathbf{K_a}$ und $\mathbf{K_{b^\perp}}$ in Umgebung II korreliert sind, d.h. $\mathbf{K_a B K_{b^\perp}^T} \neq \mathbf{\Theta}$, kann man nur eine Aussage über den gesamten Kodierer $\mathbf{K}_{\text{II}}^T = \begin{pmatrix} \mathbf{K_a}^T & \mathbf{K_{b^\perp}^T} \end{pmatrix}$ und Dekodierer $\mathbf{D}_{\text{II}} = \mathbf{D}^{\text{opt}}(\mathbf{K}_{\text{II}}, \mathbf{B})$ machen. Nun hilft aber folgende allgemeine Aussage, die wir schon häufiger verwendet haben: Spannt ein Kodierer einen Eigenraum einer Kovarianzmatrix auf, projiziert der zur Kovarianzmatrix und Kodierer gehörende optimale Dekodierer wieder in denselben Raum zurück, und andersherum (wegen Gl. 3.26 $\Longleftrightarrow$ Gl. 3.28). Auf die Situation in Umgebung II mit $\mathbf{K}_{\text{II}}$, $\mathbf{D}_{\text{II}}$ und $\mathbf{B}$ bezogen, lautet die Aussage:

$$\text{Bild}\,\mathbf{K}_{\text{II}}^T = \langle \mathbf{v}_j | j \in \mathcal{J}_{\text{II}} \subset \mathbb{N}_n \rangle \quad \Longleftrightarrow \quad \text{Bild}\,\mathbf{K}_{\text{II}}^T = \text{Bild}\,\mathbf{D}_{\text{II}} \tag{5.26}$$

Setzt man die Orthogonalität des gesamten Kodierers $\mathbf{K}_{\text{II}}$ voraus, folgt aus der Gleichheit der Bilder von Dekodierer und Kodierer (Gl. 5.26), dass jeder Gewichtsvektor $\mathbf{d}_i$ des Dekodierers $\mathbf{D}_{\text{II}}$ durch den zugehörigen Spaltenvektor $\mathbf{k}_i$ des Kodierer $\mathbf{K}_{\text{II}}$ festgelegt ist (Gl. 3.29): $\mathbf{d}_i = \frac{1}{|\mathbf{k}_i|^2}\mathbf{k}_i$. Da nun die alten Neurone in Umgebung I die gleiche Kodierung benutzten wie in Umgebung II, kann man auch jetzt die Stabilität der alten Gewichte im Dekodierer bei Umgebungswechsel erwarten, $\mathbf{D}_{\text{II}}^{\text{alt}} = \mathbf{D}_\text{I}$. Wir bemerken, dass Orthogonalität zwischen neuem und altem Kodierer durch die Strategie $S_\perp^{\text{NG}}$ stets gewährleistet wird, denn es ist $\mathbf{K_{b^\perp} K_a^T} = \mathbf{\Theta}$.

Die Äquivalenz (Gl. 5.26) ist eine allgemeinere Aussage als die weiter oben hergeleitete (Gl. 5.14), weil in hier (Gl. 5.26) nicht gefordert ist, dass die alten Gewichte des Kodierers, $\mathbf{K_a}$, für sich einen Eigenraum von $\mathbf{B}$ aufspannen müssen. Zur Verdeutlichung konstruieren wir ein Beispiel, das durch die frühere Aussage (Gl. 5.14) nicht abgedeckt ist. Es sei angenommen, dass der Kodierer $\mathbf{K_a}$ keinen Eigenraum von $\mathbf{B}$ aufspannt. Bild $\mathbf{K_a}^T$ ist also ein echter Unterraum eines Eigenraums $\mathcal{V}$. Der Eigenraum $\mathcal{V}$ habe nur eine Dimension mehr als das Bild des Kodierers, d.h. Bild $\mathbf{K_a}^T \subset \mathcal{V} = \langle \mathbf{v}_i | i \in \mathcal{I} \subset \mathbb{N}_n \rangle$, mit $|\mathcal{I}| = n_{\mathrm{I}} + 1$. Damit nun $\mathbf{K}_{\mathrm{II}}$ insgesamt einen Eigenraum aufspannt, muss die zur Vervollständigung fehlende Dimension mittels einem neuen Neuron, i.e. ein Gewichtsvektor aus $\mathbf{K}_{\mathbf{b}^\perp}$, kodiert werden. Offensichtlich ist dies prinzipiell möglich. Es wird jedoch nur geschehen, wenn diese fehlende Richtung zu den g größten Hauptkomponenten von $\mathbf{B}^\perp$ gehört oder eine Linearkombination dieser Hauptkomponenten ist. Denn $\mathbf{K}_{\mathbf{b}^\perp}$ kodiert ja laut Strategie eben diese g größten Hauptkomponenten von $\mathbf{B}^\perp$ (Gl. 4.2). Ist die Richtung aber "relevant genug", wird sie von den neuen Neuronen kodiert werden. Man sieht, dass die Aussage (Gl. 5.26) Verteilungen $P_{\mathbf{b}}$ beinhaltet, die wir noch nicht zu denen gezählt haben, in denen trotz Adaptation der Kode der alten Neurone stabil bleibt.

Interessanterweise würde der alte Dekodierer ohne Hinzufügen neuer Neurone nicht stabil bleiben, $\mathbf{D}_{\mathrm{I}} \neq \mathbf{D}_{\mathrm{II}}^{\mathrm{alt}}$, falls Bild $\mathbf{K}_{\mathbf{a}}^T$ ein echter Teilraum eines Eigenraums von $\mathbf{B}$ wäre und keine neuen Neurone diesen zu einem Eigenraum von $\mathbf{B}$ vervollständigen (wg. Gl. 3.26 $\Longleftrightarrow$ Gl. 3.28). Das Hinzufügen neuer Neurone im Kodierer kann also die Stabilität alter Neurone im Dekodierer vermitteln.

Entscheidend bei dem Effekt, dass durch neue Gewichtsvektoren die Stabilität der alten Dekodierergewichte gewahrt wird, ist die Orthogonalität der Gewichtsvektoren des Kodierers $\mathbf{K}_{\mathrm{II}}$ (siehe oben); also insbesondere die Orthogonalität alter und neuer Gewichtsvektoren zueinander, $\mathbf{K}_{\mathbf{a}}\mathbf{K}_{\mathbf{b}^\perp}^T = \mathbf{\Theta}$. In der alternativen Strategie $S_\angle^{\mathrm{NG}}$ (Gl. 4.3), in der die neuen Neurone an $\mathbf{B}$ anstatt an $\mathbf{B}^\perp$ adaptieren, gilt dies i.A. nicht. Wir führen daher die im Experiment beobachtete (deutlich) schlechtere Wiederherstellungsqualität gespeicherter Muster der Strategie $S_\angle^{\mathrm{NG}}$ gegenüber $S_\perp^{\mathrm{NG}}$ auf diesen Effekt zurück (0.56 bzw. 0.44, vergl. Tab. 4.2, 4. Spalte).

5.4 Zusammenfassung

Wir haben den Adaptationsprozess für die Strategie $S_\perp^{\mathrm{NG}}$, die neue Neurone zur Anpassung der Kodierung nutzt, mathematisch untersucht. Da die Anpassungsfähigkeit an fremde Eindrücke im wesentlichen durch die Zahl neuer (plastischer) Neurone determiniert ist, haben wir uns in der Analyse der Stabilität des Kodes gewidmet. Wir haben drei qualitativ unterschiedliche "Arten" von Umgebungswechseln, d.h. verschiedene Statistiken neuer Eindrücke $P_{\mathbf{b}}$, gefunden, in denen der Kode der Erinnerungen durch den Adaptationsprozess nicht verändert wird (vergl. Abb. 5.2).

(A) Die Hauptkomponenten von $\mathbf{B}$ sind gegenüber $\mathbf{A}$ nicht, oder nur um rechte Winkel, gedreht. Das Spektrum der Eigenwerte von $\mathbf{B}$ ist beliebig (Abb. 5.2 A).

(B) Alle Eigenvektoren $\mathbf{v}_i$ von $\mathbf{B}$, die mit dem von $\mathbf{K}_{\mathbf{a}}$ kodierten Raum korreliert sind, für die also $\mathbf{K}_{\mathbf{a}}\mathbf{v}_i \neq \mathbf{0}$ gilt, besitzen identische Eigenwerte. Der auf $\mathbf{a}$-Muster spezialisierte Kodierer $\mathbf{K}_{\mathbf{a}}$ kodiert demnach nur das "Rauschen" der $\mathbf{b}$-Verteilung. Die relevanten Informationen der $\mathbf{b}$-Muster, i.e. ihre größte Variabilität, befinden sich in Richtungen, die orthogonal zu den Gewichtsvektoren des Kodierers $\mathbf{K}_{\mathbf{a}}$ sind (Abb. 5.2 B).

(C) Der Kodierer der alten Neurone $\mathbf{K_a}$ korreliert schwach mit wenigen nicht vollständig erfassten Richtungen der $\mathbf{b}$-Verteilung, welche relativ große Varianz aufweisen. Die neuen Neurone $\mathbf{K}_{\mathbf{b}^\perp}$ sind in der Lage diese Richtungen zu kodieren, so dass der gesamte Kodierer $\mathbf{K}_{\mathrm{II}}$ einen Eigenraum von $\mathbf{B}$ aufspannt (Abb. 5.2 C).

Tritt keiner der genannten Fälle ein, wird der Kode durch den Adaptationsprozess modifiziert (Abb. 5.2 D). In diesem Fall ist eine ausgeprägte Hierarchie der Eigenwerte, die zu Eigenvektoren von $\mathbf{B}$ gehören, welche nicht orthogonal zu den Gewichtsvektoren des alten Kodierers $\mathbf{K_a}$ sind, besonders ungünstig. Dann wird der Kode besonders stark durch den Adaptationsprozess beeinflusst.

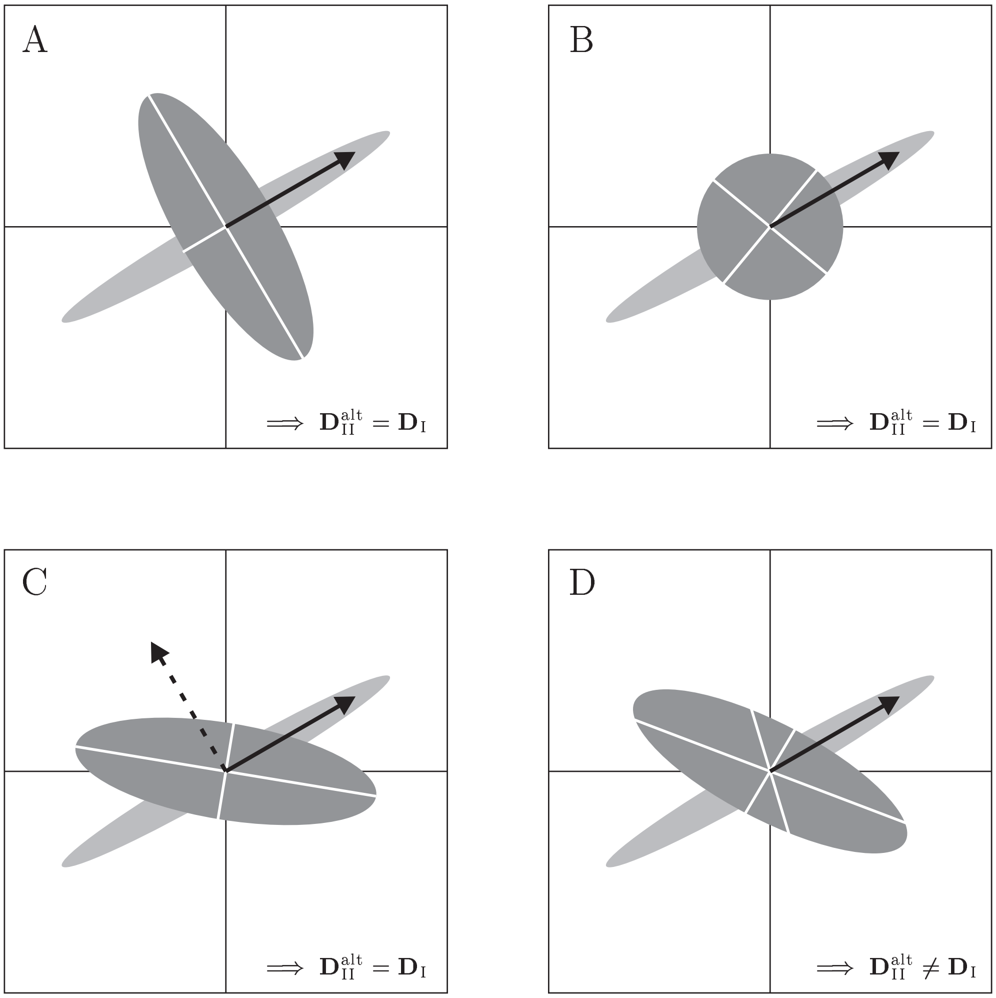

Abbildung 5.2: Stabilität des Dekodierers für die Strategie $S_{\perp}^{\mathrm{NG}}$. Es sind vier unterschiedliche Kombinationen von $P_{\mathbf{a}}$ und $P_{\mathbf{b}}$ Verteilungen schematisch in der Projektion auf die Papierebene gezeigt. $P_{\mathbf{a}}$ ist in allen Abbildungen identisch (helleres Oval im Hintergrund). Der Kodierer der alten Neurone $\mathbf{K_a}$ (durchgezogener Pfeil) liegt in der Papierebene und kodiert stets die größte Hauptkomponente von $P_{\mathbf{a}}$. Im Vordergrund ist die Projektion von $P_{\mathbf{b}}$ auf die Papierebene gezeigt (dunkleres Oval), weiße Achsen markieren die Lage ihrer Hauptkomponenten, i.e. die auf die Papierebene projizierten Eigenvektoren der Kovarianzmatrix $\mathbf{B}$. In Abb. A–C bleiben die Gewichte des Dekodierers aus Umgebung I nach Adaptation an $P_{\mathbf{b}}$ in Umgebung II unverändert, $\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} = \mathbf{D}_{\mathrm{I}}$; in Abb. D dagegen nicht, i.e. $\mathbf{D}_{\mathrm{II}}^{\mathrm{alt}} \neq \mathbf{D}_{\mathrm{I}}$. *Abbildung A*: $\mathbf{K_a}$ liegt in einem (eindimensionalen) Eigenraum von $\mathbf{B}$, der Dekodierer ist also stabil. *Abbildung B*: Das projizierte Spektrum von $\mathbf{B}$ ist weiß (Verteilung ist kreisförmig), d.h. alle Richtungen auf der Papierebene sind Eigenvektoren, also auch $\mathbf{K_a}$; der Dekodierer ist stabil. *Abbildung C*: Zwei der Eigenvektoren von $\mathbf{B}$ liegen in der Papierebene, die übrigen sind dazu orthogonal und deshalb auf den Nullpunkt projiziert. Kodierer $\mathbf{K_a}$ kann nicht die Stabilität des Dekodierers gewährleisten. Wird der Kodierer jedoch um ein zusätzliches Neuron (Gewichtsvektor, gestrichelter Pfeil) erweitert, spannen beide Vektoren gemeinsam einen Eigenraum von $\mathbf{B}$ auf, nämlich die Papierebene. Damit ist der Dekodierer insgesamt stabil, $\mathbf{D}_{\mathrm{I}} = \mathbf{D}_{\mathrm{II}}^{\mathrm{alt}}$. *Abbildung D*: Die Richtung des Kodierers $\mathbf{K_a}$ ist mit drei Eigenvektoren von $\mathbf{B}$ (weiße Linien) korreliert, denn sie liegt nicht senkrecht zu ihnen. Da $\mathbf{K_a}$ also kein Eigenraum von $\mathbf{B}$ aufspannt, wird der Dekodierer verändert, $\mathbf{D}_{\mathrm{I}} \neq \mathbf{D}_{\mathrm{II}}^{\mathrm{alt}}$. Ein zusätzliches Neuron würde daran nichts ändern (wohl aber zwei).

Kapitel 6

Diskussion

Im numerischen Experiment (Kapitel 4) haben wir den Adaptationsprozess des Modellsystems untersucht. Es wurde die Anpassung an eine neue Umgebung simuliert und gefragt, inwiefern trotz Adaptation die Stabilität des Kodes der Gedächtnisinhalte aufrechterhalten werden kann. Wie haben gezeigt, dass mithilfe von Neurogenese das Modell des Hippokampus (Abb. 3.2) Anforderungen an Stabilität und Plastizität des Kodes bei bestimmten Umgebungswechseln gerecht wird ($S_{\perp}^{\mathrm{NG}}$ in Tab. 4.2), während uneingeschränkt plastische Anpassung dazu führt, dass der Kode früherer Erinnerungen "vergessen" wird (S^{P} in Tab. 4.2). Letzteres ist als Stabilität- und Plastizitätsdilemma bekannt [24] (auch "catastrophic inference" genannt, siehe [30]). Entscheidend am Erfolg einer Adaptationsstrategie mittels Neurogenese ist jedoch die Art der Eindrücke der neuen Umgebung, d.h. ihre Wahrscheinlichkeitsverteilung $P_{\mathbf{b}}$.

In einer mathematischen Analyse (Kapitel 5) wurden jegliche Änderungen der Wahrscheinlichkeitsverteilung neocorticaler Aktivität (für beliebige Dimensionen) beschrieben, in denen die Anpassungsstrategie durch Neurogenese die Fähigkeit zur Dekodierung der Gedächtnisinhalte bewahrt. Andererseits wurden Merkmale der Verteilungen angegeben, in denen der Adaptationsprozess dazu führt, dass trotz Neurogenese und der Stabilität des Kodierers der Kode früherer Erinnerungen vom Dekodierer "vergessen" wird.

Da die Situationen des Erfolgs der Adaptation mittels Neurogenese im Modell mathematisch exakt charakterisiert wurden, können Rückschlüsse auf den Erfolg der Strategie in biologisch plausiblen Situationen gemacht werden. Dazu wollen wir zunächst die verschiedenen Fälle des Umgebungswechsels diskutieren, bevor wir dann auf das Modell selbst zu sprechen kommen. Abschließend folgt ein Vergleich mit anderen Arbeiten und ein Ausblick über mögliche Richtung zukünftiger Studien.

6.1 Adaptationsprozess

6.1.1 Ähnlich komplexe Umgebungen

Wir sind davon ausgegangen, dass neocorticale Aktivitätsmuster einer Umgebung zueinander ähnlich sind und sich in relativ wenigen, wesentlichen Merkmalen unterscheiden (siehe Abs. 4.4). Beim Umgebungswechsel des Experiments wurden die informativen Komponenten beliebig anderen Dimensionen zugesprochen. Die Komplexität der Umgebung, gemessen an der Anzahl relevanter Dimensionen, blieb jedoch erhalten. Je weniger relevante Dimensionen bei konstanter Neuronenzahl nun tatsächlich existieren, desto höher ist die Wahrscheinlichkeit, dass die bestehenden Neurone des Kodierers keine wesentliche Information der neuen Eindrücke kodiert, also nur noch "Rauschen" auffängt. In einer solchen Situation ist nicht die Stabilität des Kodes, sondern die Adaptation an neue Eindrücke kritisch für den Erfolg der Adaptationsstrategie mithilfe von Neurogenese ($S_{\perp}^{\mathrm{NG}}$, Gl. 4.2), denn der Kode der Gedächtnisinhalte bleibt in dieser Situation erhalten (vergl. Abb. 5.2 B). Die durch Neurogenese hinzukommenden Gewichte des Kodierers müssen jedoch die gesamte Information der neuen Umgebung aufnehmen. Ist diese tatsächlich von gleicher Komplexität, müsste sich die Anzahl der Neurone daher ungefähr verdoppeln, um neue Eindrücke akkurat rekonstruieren zu können. Dies ist nicht realistisch, denn erstens würde die Redundanzreduktion für beide Sorten Eindrücke ineffektiv werden (ihre kodierte Repräsentation ist doppelt so groß), und zweitens widerspricht eine Verdopplung biologischen Tatsachen, da die Zahl neuer Neurone klein gegenüber der Anzahl bestehender ist [22]. Zwar könnte man die Strategie insofern ergänzen, dass auch ein Teil der bestehenden Neurone plastisch adaptiert, dies würde aber zur Lasten der Stabilität des Kodes gehen.

Wir können also zusammenfassen, dass eine Adaptation mithilfe der Bildung neuer Neurone in Situationen eines vollständigen Wechsels der Merkmale neocorticaler Muster ungeeignet erscheint.

6.1.2 Wachsende Komplexität: "Erfahrung sammeln"

Man könnte das Experiment auch dahingehend interpretieren, dass die Maus in der zweiten Umgebung neue Erfahrungen sammelt, später jedoch mit dem Gesamtschatz an Erfahrungen lebt; am Ende des Experiments befindet sich die Maus sozusagen in beiden Umgebungen gleichzeitig, also insgesamt in einer komplexeren Umwelt.

Tatsächlich stellt das Protokoll des Experiments das Modellsystem vor höheren Anforderungen an Stabilität und Plastizität, als es für eine Interpretation als das Sammeln von Erfahrung nötig wäre. Denn in der neuen Umgebung wird die Maus ausschließlich mit neuen "Erfahrungen" konfrontiert; wir haben dies im vorhergehenden Abschnitt besprochen. Realistischer wäre es, wenn die Maus zusätzliche Erfahrungen in ihrer bekannten Umwelt aufnimmt. Zwar wissen wir nicht, inwiefern sich Erfahrungen tatsächlich in der Statistik der Aktivität des Neokortex äußern, wir können aber aufgrund der mathematischen Analyse (siehe Abs. 5.4) die Arten von Erfahrung genau definieren, für welche sich eine Adaptationsstrategie mithilfe von Neurogenese in unserem Modell lohnt. Wir können die Arten von Erfahrung wie folgt charakterisieren.

1. Die Aktivität assoziativer Areale wird komplexer, d.h. es kommen charakteristische Merkmale hinzu.

2. Die charakteristischen Merkmale neuer Eindrücke unterscheiden sich nur durch die Aktivität weniger neocorticaler Neurone von den bekannten Eindrücken.

Im ersten Fall kodieren neu gebildete Neurone die hinzukommenden relevanten Dimensionen, im zweiten Fall ergänzen sie die Kodierung, um den niedrigdimensionalen Teilraum der Änderungen zu vervollständigen (siehe Abs. 5.3.3). In beiden Fällen wird durch neue Neurone die Stabilität des Kodes gewahrt. Dazu werden nur wenige Neurone benötigt, was mit experimentellen Beobachtungen übereinstimmt [22]. Insgesamt scheint das Sammeln von Erfahrung im oben genannten Sinne im Einklang mit einer Neurogenese gestützten Adaptation zu stehen.

6.1.3 Unabhängige Umgebungen: Mögliche Aufgabe des Speichers

Eine beliebige Drehung der Hauptkomponenten ist, im Gegensatz zur wachsenden Komplexität, nur eingeschränkt für eine Adaptationsstrategie mit Neurogenese ($S_\perp^{\mathrm{NG}}$) geeignet (siehe Abs. 6.1.1). Mathematisch betrachtet ist jedoch der Kodierer der alten Neurone $\mathbf{K_a}$ "fast immer" mit allen Hauptkomponenten der (beliebigen) **b**-Verteilung korreliert. Dies ist im zweidimensionalen anschaulich. Ist nämlich der Kodierer ein beliebiger 2D-Vektor, liegt er praktisch nie orthogonal zu einem der beiden Eigenvektoren von $\mathbf{B}$, die wir uns auf den Koordinatenachsen vorstellen. Einerseits könnte dieser mathematisch häufige Fall biologisch eher selten sein; der Zusammenhang von Neurogenese mit "Erfahrungen sammeln" deutet dies an (siehe oben). Andererseits wäre der Speicher CA3 in der Lage den Dekodierer zu stabilisieren. Wir wollen die Idee näher erläutern.

Der Speicher hat **a**-Muster gespeichert, wenn das Modellsystem mit **b**-Mustern konfrontiert wird. Die Gedächtnisinhalte könnten in Umgebung II dem Dekodierer zusätzlich zu den aufgenommenen Eindrücken präsentiert werden, etwa durch spontane Aktivierung der Inhalte [7]. Der Dekodierer würde so gleichzeitig Muster beider Verteilungen dekodieren, er "sieht" daher effektiv Muster einer Kovarianzmatrix $\mathbf{C} = \alpha\,\mathbf{A} + (1-\alpha)\,\mathbf{B}$, wobei α das Verhältnis von aktivierten Gedächtnisinhalten zu verarbeiteten neuen Eindrücken ist. Durch die Summierung beider Kovarianzmatrizen werden, mit Erhöhung des Anteils α, die Hauptkomponenten von $\mathbf{C}$ denen von $\mathbf{A}$ angeglichen, so dass der Dekodierer stabilisiert werden kann (siehe Abs. 5.3). Abb. 6.1 stellt die Wirkung einer Variation von α auf die Stabilität des Dekodierers dar. Man sieht, dass schon durch kleine Verhältnisse α die Wiederherstellung gespeicherter Muster nach einer "gemischten" Adaptation deutlich verbessert ist; ab $\alpha \approx 1/3$ scheint der Kode der **a**-Muster ausreichend gut bewahrt.

Spontane Aktivierung von Gedächtnisspuren im Hippokampus ist ein bekanntes Phänomen und wird häufig mit dem Transferieren der Inhalten in ein (corticales) Lang-

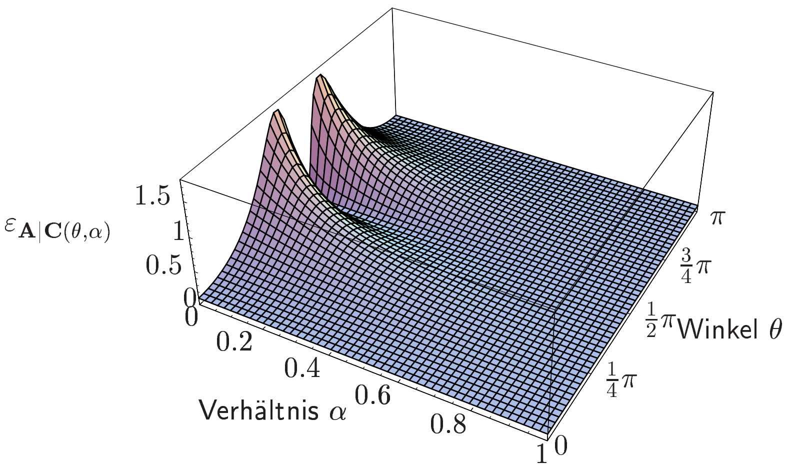

Abbildung 6.1: Aktivierung gespeicherter Inhalte. Der Rekonstruktionsfehler $\varepsilon_{\mathbf{A}|\mathbf{C}(\theta,\alpha)}$ ("Bekanntheitsgrad"), mit $\mathbf{C}(\theta,\alpha) = \alpha\,\mathbf{A} + (1-\alpha)\,\mathbf{B}(\theta)$, ist ausgewertet. Bei dem Adaptationsprozess an **b**-Muster hat der Dekodierer im Verhältnis α gespeicherte **a**-Muster und unbekannte **b**-Muster gemischt verarbeitet. θ variiert Kovarianzmatrizen $\mathbf{B}(\theta)$ (siehe unten). Der Fehlerwert $\varepsilon_{\mathbf{A}|\mathbf{C}(\theta,\alpha)}$ ist groß, wenn die Rekonstruktion von **a**-Mustern nach dem Adaptationsprozess des Dekodierers schlecht ist. Die Kodierung ist optimal für **a**-Muster und wird durch die Adaptation nicht verändert, $\mathbf{K_a} \sim \mathrm{A}_1$ (Es wird also analog zu Abb. 5.1 nur der Raum senkrecht zu $\mathbf{K}_{\mathbf{b}^\perp}$ betrachtet, vergl. Abschnitt 5.3.1). Die Kovarianzmatrizen sind gegeben durch $\mathbf{A} = \mathrm{diag}\,(0.9, 0.1)$ und $\mathbf{B}(\theta) = \mathbf{R}(\theta)\mathbf{A}\mathbf{R}(\theta)^T$, die um den Winkel θ gedrehte Kovarianzmatrix $\mathbf{A}$ ($\mathbf{R}(\theta)$ sind Rotationsmatrizen). Das Verhältnis $\alpha = 0$ entspricht der Situation, in der im Adaptationsprozess an $\mathbf{B}$ keine Speicherinhalte zur Hilfe genommen werden, um den Kode der **a**-Muster zu stabilisieren. Für $\alpha = 1$ findet keine Adaptation an die **b**-Muster statt, die Rekonstruktion der **a**-Muster ist also optimal geblieben (Fehlerwert 0.1, i.e. der kleinste Eigenwert von $\mathbf{A}$). Ist $\mathbf{B}$ im rechten Winkel zu $\mathbf{A}$ gedreht, verändert der Adaptationsprozess den Dekodierer nicht, unabhängig vom Mischverhältnis α. Denn dann spannt $\mathbf{K_a}$ einen Eigenraum von $\mathbf{B}$ auf (vergl. Abb. 5.2 A). Bei Winkeln $\theta = \varphi^*$ (siehe Abs. 5.3.1) ist der Rekonstruktionsfehler der **a**-Muster dagegen am größten (vergl. Abb. 5.1).

zeitgedächtnis in Verbindung gebracht, (Konsolidierung) [51, 28]. In unserer Überlegung würde der Speicher durch spontane Aktivierung zusätzlich für die korrekte Dekodierung seiner Inhalte sorgen. Dazu muss die Dauer des Zwischenspeicherns von Information im Hippokampus länger sein, als das Intervall, für das eine Adaptation der Kodierung notwendig wird. Diese Annahme scheint gerechtfertigt, da zumindest im Menschen Gedächtnisinhalte z.T. mehrere Jahre hippokampusabhängig sind [30, 32].

6.2 Biologische Plausibilität

Unser Modell beschränkt die tatsächliche Konnektivität und Komplexität des Hippokampus stark (vergl. Abschnitt 2.1), und kann daher nur als erster Versuch gelten, die Funktion der Neurogenese zu eruieren. Das Ziel dieser Arbeit ist es, die prinzipielle Eignung einer Adaptationsstrategie mithilfe von Neurogenese zu untersuchen, und nicht ein Detail getreues Abbild des Hippokampus zu entwerfen. Deswegen ist es notwendig, Ergebnisse und Erfordernisse des Modells auf ihre biologische Plausibilität zu überprüfen.

6.2.1 Modell

Wir haben gezeigt, dass der optimale Kodierer einen "Eigenraum" aufspannt. Derartige Gewichtsverteilungen sind mit der Hebb'schen Lernregel zu erreichen [18]. Hebb'sche Plastizität ist an Synapsen des Tractus perforans und den Moosfasern (Kodierer) und auch an den Schaffer-Kollateralen (Dekodierer) nachgewiesen worden (siehe Abs. 2.1). Die Orthogonalität der Gewichte des Kodierers könnte durch inhibitorische Neurone gewährleistet werden [11]. Diese Ergebnisse sind daher mit der Biologie vereinbar.

Die Linearität der Kodierung in unserem Modell ist eine starke Vereinfachung, da Neurone nichtlineare Systeme sind [18]. Die Aktivität der Neurone zeigt Phänomene der Sättigung, so dass z.B. stark verzerrende Vektoren des Dekodierers (vergl. Tab. 4.3) biologisch nicht plausibel sind. Für Aktivitäten unterhalb der Sättigungsschwelle ist der lineare Ansatz als erste Näherung gültig, eine lineare Kodierung ist deshalb aufgrund ihrer analytischen Einfachheit gewählt worden.

Das Modell als solches wurde bereits im Abschnitt 3.1.2 mit der biologischen Grundlage verglichen. Dort haben wir jedoch den Punkt zurückgestellt, inwiefern das Zusammenfassen aller corticalen und aller hippokampalen Neurone zu jeweils einer logischen Neuronenschicht gerechtfertigt ist (siehe Abb. 3.2). Während wir ersteres im Abschnitt 6.4 diskutieren, kommen wir jetzt auf das Zusammenfassen der Hippokampusneurone zu sprechen. Es stellt sich die Frage, ob das Vergrößern der verborgenen Schicht mit dem physiologisch auf den GD beschränktem Wachstum der Neurone vergleichbar ist. Da auf dem Weg zum GD unserer Vorstellung zufolge die Kodierung erfolgt, ist die Frage für den Kodierer positiv zu beantworten. Das Wachstum des Dekodierers ist dagegen unrealistisch. Würde der Dekodierer im Modell nicht mitwachsen, würden die "alten" Gewichte des Dekodierers die Aufgabe übernehmen müssen, die

"Ausgabedaten" neuer Neurone im Kodierer mitzubearbeiten. Denn neue und alte Neurone des Kodierers projizieren dann ihre Darstellung neocorticaler Information auf eine Neuronenschicht konstanter Dimensionalität, i.e. auf die Neurone des Dekodierers. Das führt dazu, dass die kodierten Darstellungen von neuen und alten Neuronen (aus Sicht des Dekodierers) miteinander korreliert sind, sofern sie gleichzeitig aktiv sind. Korrelation der kodierten Darstellungen haben wir als besonders ungünstig auf die Stabilität des Dekodierers wirkend identifiziert (vergl. $S_{\perp}^{\mathrm{NG}}$ gegenüber $S_{\angle}^{\mathrm{NG}}$ in Tab. 4.2, 4. Spalte). Der Schlüssel für dieses Problem könnte die Erzeugung einer kodierten Darstellung sein, die durch ihre Spärlichkeit Korrelationen zwischen Mustern vermindert [18].

Die lineare Kodierung unseres Modells, welche die Komponenten größter Varianz zur Darstellung benutzt (PCA), garantiert keine spärliche Aktivität der verborgenen Schicht. Zwar könnten spärliche neocorticale Muster sich auf die Spärlichkeit des Kodes der verborgenen Schicht auswirken (neuronale Verteilungen werden u.a. als hypergeometrisch modelliert [29]), dies würde aber wahrscheinlich nicht ohne weiteres den hohen Grad der Spärlichkeit im GD erklären, der experimentell gefunden wird (nur 0.4% gleichzeitig aktive Zellen, siehe Abschnitt 2.3).

Veränderte Kodierungsschemen könnten einen spärlicheren Kode im GD erzeugen. Würden zum Beispiel nicht die Komponenten größter Varianz, sondern diejenigen, welche statistisch am "unabhängigsten" sind, zur Darstellung neocorticaler Muster verwendet, könnte die Korrelation zwischen Mustern im GD vermindert werden (Analyse unabhängiger Komponenten, ICA). ICA ist mit der Maximierung der Spärlichkeit verwandt und wurde deshalb in der Literatur als Schema für die Kodierung des Hippokampus bereits motiviert [24]. Inwiefern sich das veränderte Kodierungsschema dann tatsächlich auf die Stabilität des Dekodierers auswirkt, müsste in einer Erweiterung des Modells analysiert werden.

6.2.2 Neurogenese

Neurogenese ist ein komplizierter Prozess, der von der Proliferation von Stammzellen über die Differentation bis zur Reife mehrere Wochen dauert [47]. Unser Analogon der Neurogenese, das plötzliche Entstehen neuer neuronaler Einheiten der verborgenen Schicht, ist offensichtlich eine grobe Abstraktion der natürlichen Vorgänge. Die genannte Zeit von vier Wochen zur Bildung neuer Neurone ist jedoch nicht direkt mit der Neurogenese unseres Modells zu vergleichen. Wir betrachten nämlich die funktionelle Aktivierung neuer Zellen, ein Prozess, der wesentlich kürzer dauern dürfte als die Entstehung neuer Zellen im physiologischen Sinn [16].

Die Plastizität der synaptischen Gewichte "neuer" Neurone gegenüber der Stabilität der Gewichte "alter" Neurone, wie wir es für unsere Adaptationsstrategie ($S_{\perp}^{\mathrm{NG}}$) fordern, ist experimentell weder nachgewiesen noch widerlegt [24]. Es müssen Experimente zeigen, ob diese Hypothese so oder ähnlich (etwa durch ortsbedingtes, plastisches Verhalten der Neurone [24]) zutrifft.

Bildung neuer Neurone ist in unserem Modell über die Auswertung des Rekonstruktionsfehlers regulierbar. Ist er groß, müssen Kodiererneurone gebildet werden, um die Kodierung der relevanten Information zu gewährleisten. Wir haben bereits zur Motivierung des Modells darauf hingewiesen (siehe Abs. 3.1.2), dass der Fehler möglicherweise in Schichten des EC ausgewertet wird [7]. Tatsächlich wird die Fähigkeit des EC zur Regulierung der Neurogenese in der Literatur genannt [17]. Aus Sicht des Modells könnte auch die Modifikation der synaptischen Gewichte des Dekodierers Neurogenese induzieren, die Ausrichtung neuer Neurone im Kodierer würden dann die sich anbahnende Veränderung des Dekodierers Einhalt gebieten können.

6.3 Vergleich mit anderen Arbeiten und Ausblick

Die Funktion neuer Neurone im Hippokampus ist nicht bekannt [24, 4] und theoretische Arbeiten über den Hippokampus und seine Funktionen (z.B. [20, 42, 30, 26, 10, 29]) lassen den Aspekt der Neurogenese außen vor oder weisen ihr nachträglich eine Funktion zu [50].

Die Architektur unseres Modells (Autokodierer Abb. 3.2) ist ein in der Theorie bekanntes künstliches Netzwerk [18], Effekte des Wachstums der verborgenen Schicht wurden hier jedoch nicht untersucht. Interessanterweise wird gerade ein solches Autokodierernetzwerk als Konzept eines "Neuheitsdetektors" für Informationverarbeitungen vorgeschlagen [41], eine Eigenschaft die auch dem Hippokampus nachgesagt wird [25, 48]. Es wird auf Plastizitäts-Stabilitäts-Dilemma in einem ähnlichen Netzwerken hingewiesen (siehe [30]), ein Wachstum von Neuronen wurde aber hierbei nicht als Lösung diskutiert.

Es gibt jedoch Arbeiten, die sich mit wachsenden, künstlichen neuronalen Netzwerken auseinandersetzen (z.B. [9, 12, 36]), welche jedoch mehr den Nutzen derartiger Systeme in der Technik als ihre biologische Relevanz in den Mittelpunkt stellen. Im "Cascade-Correlation" Netzwerk [9] wird, ähnlich wie in unserem Modell, von neuronalen Einheiten zwar der quadratische Fehler minimiert, neue Neurone bearbeiten jedoch auch das Ausgangssignal der bestehenden Neurone, so dass ein mehrschichtiges Netzwerk aufgebaut wird. Da granuläre Zellen im Hippokampus Axone in die CA3-Region aussenden, ist eine solche Architektur hier nicht anwendbar.

6.4 Grenzen und Ausblick

Autoassoziativspeicher (wie CA3 [42]) können, in ihrer Implementierung als (graduelles) Hopfieldnetz [19] pro Dimension, i.e. Neuron, höchstens zwei (reele) Werte speichern. In der Kodierung unseres Modells sind jedoch die Hauptkomponenten der Verteilung neocorticaler Muster erfasst worden, d.h. die Komponenten mit der größten Varianz. Ein Neuron der verborgenen Schicht verändert seine Feuerrate daher über einen möglichst großen Bereich. Es erscheint unsinnig, diese informative Schwankung der Feuerrate durch nur zwei Werte im Speicher abzulegen, zumal dann gerade die charakterisieren-den Schwankungen der Muster wieder aufgehoben werden würden; von der Ausgabe des “Eckenneurons” (siehe Abs. 3.1.3) könnte der Speicher beispielsweise nur “wenige” und “viele” Ecken unterscheiden. Es muss also eine Umwandlung des (angenommenen) Ratenkodes in einer Art binären Kode auf dem Weg zum Speicher existieren (siehe unten).

Im Modell sind wir davon ausgegangen, dass die Anzahl der Neurone, die direkt auf den GD (die verborgene Schicht) projizieren, größer als die Anzahl der Neurone im GD ist. Natürlich gibt es im Neokortex sehr viel mehr Neurone als im Hippokampus [21]. Es wird jedoch in der Literatur angenommen, dass die Projektionen des Neokortex zunächst auf Schichten des entorhinalen Cortex münden und nicht direkt auf GD [29, 42]. Da die Anzahl der Neurone in Schicht II des EC, aus der nachweislich Axone zum GD gesendet werden (siehe Abs. 2.1), kleiner ist als im GD (Verhältnis 1:12 im Menschen, siehe Abschnitt 2.1), würde die Kompression neocorticaler Muster an Neuronen des EC stattfinden und nicht im GD, wo jedoch die Neurogenese stattfindet. In der Tat schreiben manche Theorien die Dimensionsreduktion neocorticaler Muster in ähnlicher Form, wie wir sie beschrieben haben (jedoch ohne Neurogenese), den Verbindungen zwischen Neokortex und EC zu [29].

Die vorgeschlagene Adaptationsstrategie mittels Neurogenese würde mit einer größeren Anzahl EC-Neurone als Neurone der verborgenen Schicht nicht funktionieren, denn sie basiert auf der Annahme eines “Flaschenhalses”, d.h. einer Dimensionsreduktion (siehe Abs. 3.1.3). Die Idee der Strategie, dass zusätzliche, charakteristische Komponenten der Information durch hinzukommende Neurone kodiert werden, während bestehende den Kode konservieren, könnte mit einer abgewandelten Art der Kodierung jedoch trotzdem zutreffen. Zum Beispiel könnte eine erste Umwandlung von Ratenkode in einen zum Speichern notwendigen binären Kode von den Neuronen des Gyrus dentatus bewerkstelligt werden. Würde etwa die Frequenz (oder ein anderes kontinuierliches Merkmal) der EC-Neurone in, sagen wir 10, Stufen aufgeteilt, entspräche dies einer effektiven Projektion der EC-Frequenzmuster in einen (10-mal) höher dimensionalen Raum. Dazu müssten GD-Neurone für verschiedene Frequenzbereiche der EC-Neurone sensitiv sein. Nach dieser Überlegung würden die GD-Neurone die in der vorliegenden Arbeit beschriebene Redundanzreduktion ausführen, nur dass die Eingangsdaten zur verbogenen Schicht aus diesem (effektiv) hochdimensionalen Raum kämen. Ein derartiges Schema wäre aus weiteren Gründen plausibel. Zum einen würde eine solche Kodierung einen spärlichen Kode im GD hervorrufen, denn ein EC-Neuron kann offensichtlich nur mit einer Rate zur Zeit feuern, so dass in unserem Beispiel höchstens 1/10 der GD-Neurone gleichzeitig aktiv sind. Eine spärliche Verteilung der Feuerraten ist charakteristisch für den GD [29, 42]; dieser Aspekt wurde in der bisherigen Formulierung des Modells nicht berücksichtigt (siehe oben). Zum anderen ist der GD Untersuchungen zufolge in der Trennung von Mustern involviert [25]. Eine Diskretisierung der Feuerraten würde eine Trennung von Mustern in natürlicherweise ermöglichen; man erkennt die Effektivität der Trennung auch an der Spärlichkeit der erzielten Kodierung, denn spärliche Repräsentation ist ein Weg Muster zu orthogonalisieren [18].

Ob eine derartige Kodierung stattfindet oder ob andere Kodierungsmechanismen vielversprechender sind (z.B. "overcomplete ICA" [27]), müssen zukünftige Studien zeigen.

6.5 Schlusswort

In dieser Arbeit wurde die Idee untersucht, ob die funktionelle Bedeutung der Neurogenese in der Lösung des Konflikts zwischen Stabilität und Plastizität liegen könnte. Es wurden numerische und analytische Resultate präsentiert, welche die Idee unter bestimmten Bedingungen unterstützen. Auf der anderen Seite wurden jedoch auch Situationen beschrieben, die – zumindest im einfachen Modell – problematisch für diese Theorie sind. Das Modell kann aufgrund seiner Einfachheit die Antwort, ob die Hauptaufgabe der Neurogenese in der Adaptation einer Kodierung zu suchen ist, nicht endgültig entscheiden. Es müssen daher weiterführende, detailliertere Studien folgen, um die kognitive Funktion adulter Neurogenese in einem Bereich des Gehirns zu verstehen, der uns (möglicherweise) in diesem Prozess des Lernens unterstützen wird.

Anhang A

Mathematischer Anhang

A.1 Beweise

A.1.1 Extremum aus Gl. 3.21 ist Minimum.

Die Matrix $\mathbf{\Gamma} = (\gamma_{ij}) := \mathbf{K}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T$ ist eine invertierbare Matrix, da wir $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle$ als invertierbar und den Höchstrang von $\mathbf{K}$ voraussetzen. Dann blieb zu zeigen, ob das einzige gefundene Extremum $\mathbf{D}^{\text{opt}}$ tatsächlich ein Minimum ist. Entwickelt man ε um $\mathbf{D}^{\text{opt}}$ in den Variablen $\mathbf{D} = (d_{ij})$ in eine Taylorreihe und bricht nach der zweiten Ordnung ab, so erhält man, mit $\xi_{ij} := d_{ij} - d_{ij}^{\text{opt}}$, folgendes Ergebnis.

$$
\begin{aligned}
\varepsilon - \varepsilon_{\mathbf{x}} &= \frac{1}{2}\sum_{ijkl}\left(\frac{\partial}{\partial d_{ij}}\frac{\partial}{\partial d_{kl}}\mathbf{D}\mathbf{\Gamma}\mathbf{D}^T\right)\xi_{ij}\xi_{kl} + \mathcal{O}(\xi^3) \\
&= \sum_{ijkl}\left(\frac{\partial}{\partial d_{ij}}\sum_s d_{ks}\gamma_{sl}\right)\xi_{ij}\xi_{kl} + \mathcal{O}(\xi^3) \\
&= \sum_{ijl}\gamma_{jl}\xi_{ij}\xi_{il} + \mathcal{O}(\xi^3) \\
&= \sum_i \boldsymbol{\xi}_i^T\mathbf{\Gamma}\boldsymbol{\xi}_i + \mathcal{O}(\xi^3) \qquad \text{(A.1)}
\end{aligned}
$$

Dabei sei $\boldsymbol{\xi}_i \in \mathbb{R}^m$ der i-te Zeilenvektor der Matrix $\mathbf{D} - \mathbf{D}^{\text{opt}}$.

Die Matrix $\mathbf{\Gamma}$ ist aufgrund ihrer Struktur eine positiv semidefinite Matrix. Denn es gilt folgende Äquivalenz [34]: eine reele, quadratische Matrix $\mathbf{X}$ ist positiv semidefinit und symmetrisch, g.d.w. eine Matrix $\mathbf{Y}$ existiert, mit $\mathbf{X} = \mathbf{Y}^T\mathbf{Y}$. Außerdem ändert Mittelwertbildung nichts an der Definitheit der Matrix. Da $\mathbf{\Gamma}$ als invertierbar vorausgesetzt ist, folgt aus der Symmetrie von $\mathbf{\Gamma}$ sogar die positive Definitheit ($\det\mathbf{\Gamma} = \prod_i \lambda_i \neq 0$, also insgesamt $\lambda_i > 0$. λ_i seien die Eigenwerte von $\mathbf{\Gamma}$). Damit gilt für $\boldsymbol{\xi}_i \neq 0$

$$\boldsymbol{\xi}_i^T\mathbf{\Gamma}\boldsymbol{\xi}_i > 0. \qquad \text{(A.2)}$$

In der Nähe des Optimums können Terme höherer Ordnung in ξ vernachlässigt werden, so dass in einer Umgebung von $\varepsilon_{\mathbf{x}}$ gilt:

$$\varepsilon - \varepsilon_{\mathbf{x}} \geq 0 \qquad \text{(A.3)}$$

Damit ist gezeigt, dass $\mathbf{D}^{\mathrm{opt}}$ mit den gemachten Voraussetzungen (konstante Dimensionen aller Schichten, Regularität von $\mathbf{\Gamma}$) tatsächlich das Minimum ist.

A.1.2 Beweis von Gl. 3.60

Sei $\hat{\mathbf{P}} = \left(\mathbf{I} - \mathbf{D}^{\mathrm{opt}}\left(\mathbf{K}, P_{\mathbf{x}}\right)\mathbf{K}\right)$, wobei $\mathbf{x} \in \mathbb{R}^n$ und $\mathbf{K} \in \mathbb{R}^{(m,n)}$, mit $m = \operatorname{Rg}\mathbf{K}$ Höchstrang. Die Spektraldarstellung der Kovarianzmatrix sei $\left\langle \mathbf{x}\mathbf{x}^T \right\rangle = \sum_{i=1}^n \lambda_i \mathbf{w}_i \mathbf{w}_i^T$ und $\lambda_i > 0$. Dann sind folgende Aussagen äquivalent

1. $\hat{\mathbf{P}}^T = \hat{\mathbf{P}}$
2. $\operatorname{Bild}\mathbf{K}^T = \operatorname{Bild}\mathbf{D}^{\mathrm{opt}}$
3. $\operatorname{Bild}\mathbf{K}^T = \operatorname{Bild}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T$
4. $\exists\, \mathcal{I} \subset \mathbb{N}_n : \quad \langle \mathbf{w}_i | i \in \mathcal{I} \rangle = \operatorname{Bild}\left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T$

Beweis "1. $\Longleftrightarrow$ 2."

Da offensichtlich $\hat{\mathbf{P}} = (\mathbf{I} - \mathbf{P})$ genau dann symmetrisch ist, wenn $\mathbf{P}$ symmetrisch ist, reicht es, die Aussage für $\mathbf{P} = \mathbf{D}^{\mathrm{opt}}\mathbf{K}$ nachzuweisen. Zunächst gilt

$$\mathbf{P}\mathbf{x} = \mathbf{x}, \quad \Longleftrightarrow \quad \mathbf{x} \in \operatorname{Bild}\mathbf{D}^{\mathrm{opt}} \tag{A.4}$$

Denn einerseits gilt mit$\mathbf{m} := \mathbf{K}\mathbf{x}$ die Aussage $\mathbf{x} = \mathbf{P}\mathbf{x} = \mathbf{D}^{\mathrm{opt}}\mathbf{K}\mathbf{x} = \mathbf{D}^{\mathrm{opt}}\mathbf{m}$, d.h. $\mathbf{x}$ ist Linearkombination der Spaltenvektoren von $\mathbf{D}^{\mathrm{opt}}$ und damit $\mathbf{x} \in \operatorname{Bild}\mathbf{D}^{\mathrm{opt}}$. Andererseits gilt ist $\mathbf{y}$ als Linearkombination von $\mathbf{D}^{\mathrm{opt}}$ darstellbar, folgt $\operatorname{Bild}\mathbf{D}^{\mathrm{opt}} \ni \mathbf{y} = \mathbf{D}^{\mathrm{opt}}\mathbf{m}' = \mathbf{D}^{\mathrm{opt}}\mathbf{K}\mathbf{D}^{\mathrm{opt}}\mathbf{m}' = \mathbf{P}\mathbf{y}$ wegen $\mathbf{K}\mathbf{D}^{\mathrm{opt}} = \mathbf{I}$ (Gl. 3.24).

Weiter ist wegen $\mathbf{K}\mathbf{x} = 0 \Longleftrightarrow \mathbf{x} \in \operatorname{Kern}\mathbf{K}$ und $\mathbf{P} = \mathbf{D}^{\mathrm{opt}}\mathbf{K}$ offensichtlich

$$\mathbf{P}\mathbf{x} = \mathbf{0}, \quad \Longleftrightarrow \quad \mathbf{x} \in \operatorname{Kern}\mathbf{K}. \tag{A.5}$$

Die entsprechenden Aussagen über $\mathbf{P}^T = \mathbf{K}^T {\mathbf{D}^{\mathrm{opt}}}^T$ folgert man analog. Diese sind

$$\mathbf{P}^T\mathbf{x} = \mathbf{x} \quad \Longleftrightarrow \quad \mathbf{x} \in \operatorname{Bild}\mathbf{K}^T \tag{A.6}$$

und

$$\mathbf{P}^T\mathbf{x} = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{x} \in \operatorname{Kern}{\mathbf{D}^{\mathrm{opt}}}^T. \tag{A.7}$$

Ist nun $\mathbf{P} = \mathbf{P}^T$ gilt natürlich auch $\mathbf{P}\mathbf{x} = \mathbf{P}^T\mathbf{x}$ und mit Gl. A.4 und Gl. A.6 ergibt sich sofort $\operatorname{Bild}\mathbf{K}^T = \operatorname{Bild}\mathbf{D}^{\mathrm{opt}}$.

Setzt man $\operatorname{Bild}\mathbf{K}^T = \operatorname{Bild}\mathbf{D}^{\mathrm{opt}}$ voraus, folgt auch $\operatorname{Kern}\mathbf{K} = \operatorname{Kern}{\mathbf{D}^{\mathrm{opt}}}^T$. Da sich jeder Vektor $\mathbf{x} \in \mathbb{R}^n$ aus zwei Vektoren $\mathbf{x}_B \in \operatorname{Bild}\mathbf{K}^T$ und $\mathbf{x}_K \in \operatorname{Kern}\mathbf{K}$ zusammensetzen lässt,$\mathbf{x} = \mathbf{x}_B + \mathbf{x}_K$, weil Basen von $\operatorname{Kern}\mathbf{K}$ und $\operatorname{Bild}\mathbf{K}^T$ zusammen eine Basis des

$\mathbb{R}^n$ ausmachen, können wir ableiten

$$
\begin{aligned}
& & \mathbf{Px} &= \mathbf{Px}_B + \mathbf{Px}_K \\
&[\text{Gl. A.4 bzw. Gl. A.5 }] & &= \mathbf{x}_B + \mathbf{0} \\
&[\text{Gl. A.6 bzw. Gl. A.7 }] & &= \mathbf{P}^T\mathbf{x}_B + \mathbf{P}^T\mathbf{x}_K \\
& & &= \mathbf{P}^T\mathbf{x} \qquad \text{(A.8)}
\end{aligned}
$$

Da $\mathbf{x} \in \mathbb{R}^n$ beliebig war ist damit $\mathbf{P} = \mathbf{P}^T$ gezeigt.

Beweis "2. $\Longleftrightarrow$ 3."

Zu zeigen ist, dass die Vektoren der Matrix $\mathbf{D}^{\text{opt}} = (\mathbf{d}_1, \dots, \mathbf{d}_m)$ als Linearkombinationen der Vektoren der Matrix $\mathbf{M} = \langle \mathbf{xx}^T \rangle \mathbf{K}^T = (\mathbf{m}_1, \dots, \mathbf{m}_m)$ darstellbar sind. Nach der Gleichung für $\mathbf{D}^{\text{opt}}$ (Gl. 3.21) gilt

$$
\begin{aligned}
\mathbf{D}^{\text{opt}} &= \mathbf{M}\left(\mathbf{K}\langle \mathbf{xx}^T \rangle \mathbf{K}^T\right)^{-1} && \text{(A.9)} \\
&= \mathbf{MN}, && \text{(A.10)}
\end{aligned}
$$

nennt man $\mathbf{N} = (\nu_{ij}) := \left(\mathbf{K}\langle \mathbf{xx}^T \rangle \mathbf{K}^T\right)^{-1}$. Man hat daher $\forall\, j$

$$\mathbf{d}_j = \sum_{i=1}^{m} \nu_{ij}\mathbf{m}_i. \tag{A.11}$$

Dies war zu beweisen.

Beweis "3. $\Longleftrightarrow$ 4."

Sei $\langle \mathbf{xx}^T \rangle = \sum_{i=1}^n \lambda_i \mathbf{w}_i \mathbf{w}_i^T$, wie es oben vorausgesetzt wurde. Da die Eigenvektoren $\mathbf{w}_i$, $i \in \mathbb{N}_n$, eine Basis des $\mathbb{R}^n$ sind, lässt sich $\mathbf{K}^T = (\mathbf{k}_1, \dots, \mathbf{k}_m)$ durch sie darstellen

$$\mathbf{k}_j = \sum_{i=1}^{n} \mu_{ji}\mathbf{w}_i. \tag{A.12}$$

Damit folgt

$$
\begin{aligned}
&[\mathbf{M} = \langle \mathbf{xx}^T \rangle \mathbf{K}^T] & \mathbf{m}_j &= \langle \mathbf{xx}^T \rangle \mathbf{k}_j && \text{(A.13)} \\
&[\text{Gl. A.12}] & &= \sum_{i=1}^{n} \lambda_i \mathbf{w}_i \mathbf{w}_i^T \sum_{s=1}^{n} \mu_{js}\mathbf{w}_s && \text{(A.14)} \\
&[\mathbf{w}_i^T\mathbf{w}_s = \delta_{is}] & &= \sum_{i=1}^{n} \lambda_i \mu_{ji} \mathbf{w}_i && \text{(A.15)}
\end{aligned}
$$

Ist nun Bild $\mathbf{K}^T$ = Bild $\langle \mathbf{xx}^T \rangle \mathbf{K}^T$ vorgegeben, ist jede Linearkombination $\sum_{j=1}^m \rho_j \mathbf{m}_j$ durch eine Linearkombination der Gewichtsvektoren $\sum_{j=1}^m \nu_j \mathbf{k}_j$ darstellbar. Wählen wir eine Linearkombination, folgt mit Gl. A.15 zunächst

$$\sum_{j=1}^{m} \rho_j \mathbf{m}_j = \sum_{i=1}^{n} \left(\lambda_i \sum_{j=1}^{m} \rho_j \mu_{ji}\right) \mathbf{w}_i. \tag{A.16}$$

und somit

$$\sum_{i=1}^{n}\left(\lambda_i \sum_{j=1}^{m} \rho_j \mu_{ji}\right) \mathbf{w}_i = \sum_{j=1}^{m} \nu_j \mathbf{k}_j \tag{A.17}$$

Da die ρ_j beliebig sind und $\lambda_i > 0$ gilt, fordert die Gleichung, dass die Vektoren $\mathbf{w}_i$, $i \in \mathbb{N}_n$ den Raum $\operatorname{Bild} \mathbf{K}^T$ aufspannen. Da die Eigenvektoren eine Basis der $\mathbb{R}^n$ sind, dürfen nur für $i \in \mathcal{I} \subset \mathbb{N}_n$ mit $|\mathcal{I}| = \operatorname{Rg} \mathbf{K} = m$ die Vorfaktoren von $\mathbf{w}_i$ in Gl. A.17 ungleich Null sein (dies wird erreicht durch die μ_{ji}). Das bedeutet insgesamt $\langle \mathbf{w}_i | i \in \mathcal{I} \rangle = \operatorname{Bild} \mathbf{K}^T$, was zu zeigen war.

Setzt man $\langle \mathbf{w}_i | i \in \mathcal{I} \rangle = \operatorname{Bild} \mathbf{K}^T$ voraus, bedeutet dies $\mathbf{k}_j = \sum_{i \in \mathcal{I}} \sigma_{ji} \mathbf{w}_i$ und eingesetzt in Gl. A.13

$$\mathbf{m}_j = \sum_{i=1}^{n} \lambda_i \mathbf{w}_i \mathbf{w}_i^T \sum_{s \in \mathcal{I}} \sigma_{js} \mathbf{w}_s \tag{A.18}$$

$$[\mathbf{w}_i^T \mathbf{w}_s = \delta_{is}] \qquad = \sum_{i \in \mathcal{I}} \lambda_i \sigma_{ji} \mathbf{w}_i \tag{A.19}$$

Wegen $\mathbf{m}_j = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{k}_j$ ist damit $\operatorname{Bild} \mathbf{K}^T = \operatorname{Bild} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{K}^T$ gezeigt.

A.1.3 Beweis von Gl. 3.29

Wir wollen unter den gleichen Voraussetzungen wie in Abschnitt A.1.2 mit $\mathbf{D}^{\text{opt}} = (\mathbf{d}_1, \ldots, \mathbf{d}_m)$ und $\mathbf{K}^T = (\mathbf{k}_1, \ldots, \mathbf{k}_m)$ folgende Aussage nachweisen:

$$\operatorname{Bild} \mathbf{K}^T = \operatorname{Bild} \mathbf{D}^{\text{opt}} \text{ und } \mathbf{k}_i^T \mathbf{k}_j = 0,\ i \neq j \quad \Longrightarrow \quad \mathbf{d}_i = \frac{1}{|\mathbf{k}_i|^2} \mathbf{k}_i,\ i \in \mathbb{N}_m \tag{A.20}$$

Sei $\operatorname{Bild} \mathbf{K}^T = \operatorname{Bild} \mathbf{D}^{\text{opt}}$ und die Orthogonalität von $\mathbf{K}$ vorgegeben. $\mathbf{K}\mathbf{D}^{\text{opt}} = \mathbf{I}$ (Gl. 3.24) bedeutet in Vektorschreibweise $\mathbf{k}_i^T \mathbf{d}_j = \delta_{ij}$. Daher folgt

$$[\operatorname{Bild} \mathbf{K}^T = \operatorname{Bild} \mathbf{D}^{\text{opt}}] \qquad \mathbf{d}_i = \sum_{j=1}^{m} \mu_{ij} \mathbf{k}_j \tag{A.21}$$

$$[\mathbf{k}_i^T \mathbf{d}_j = \mathbf{k}_i^T \mathbf{k}_j = 0,\ i \neq j] \qquad = \mu_{ii} \mathbf{k}_i \tag{A.22}$$

Daraus ergibt sich wegen $\mathbf{k}_i^T \mathbf{d}_i = 1$ sofort $\mu_{ii} = \frac{1}{|\mathbf{k}_i|^2}$, was zu zeigen war.

A.1.4 Beweis von Gl. 5.14

Mit Voraussetzungen wie in Tab. 5.1 wollen wir die Aussage

$$\operatorname{Bild} \mathbf{D}_{\mathbf{a}}^{\mathbf{A}} = \operatorname{Bild} \mathbf{D}_{\mathbf{a}}^{\mathbf{B}} \text{ und } \mathbf{K}_{\mathbf{b}^\perp} \mathbf{B} \mathbf{K}_{\mathbf{a}}^T = \mathbf{\Theta} \quad \Longleftrightarrow \quad \operatorname{Bild} \mathbf{K}_{\mathbf{a}}^T = \langle \mathbf{v}_j | j \in \mathcal{J} \subset \mathbb{N}_n \rangle \tag{A.23}$$

beweisen. $\mathbf{v}_i$ seien (orthonormale) Eigenvektoren von $\mathbf{B}$.

Richtung “$\Longrightarrow$”

Wir setzen $\operatorname{Bild}\mathbf{D}_\mathbf{a}^\mathbf{A} = \operatorname{Bild}\mathbf{D}_\mathbf{a}^\mathbf{B}$ voraus. Da $\mathbf{K_a}$ optimal für **a**-Muster ist, spannt $\mathbf{K_a}$ einen Eigenraum von $\mathbf{A}$ auf. Damit gilt $\operatorname{Bild}\mathbf{K}_\mathbf{a}^T = \operatorname{Bild}\mathbf{D}_\mathbf{a}^\mathbf{A}$ (Gl. 3.28 und Gl. 3.26). Nach der Voraussetzung gilt dann auch $\operatorname{Bild}\mathbf{K}_\mathbf{a}^T = \operatorname{Bild}\mathbf{D}_\mathbf{a}^\mathbf{B}$. Da $\mathbf{D}_\mathbf{a}^\mathbf{B}$ ebenfalls zu $\mathbf{K_a}$ optimal ist (wegen $\mathbf{K}_{\mathbf{b}^\perp}\mathbf{B}\mathbf{K}_\mathbf{a}^T = \mathbf{\Theta}$, siehe Gl. 5.1), können wir analog schlussfolgern (Gl. 3.28 und Gl. 3.26), dass $\mathbf{K_a}$ auch einen Eigenraum von $\mathbf{B}$ aufspannt, was zu zeigen war.

Richtung “$\Longleftarrow$”

Sei nun $\operatorname{Bild}\mathbf{K}_\mathbf{a}^T = \langle \mathbf{v}_j | j \in \mathcal{J} \subset \mathbb{N}_n \rangle$ vorausgesetzt, der Kodierer $\mathbf{K_a}$ spannt also einen Eigenraum von $\mathbf{B}$ auf. Daraus ergibt sich sofort (“$4 \Longrightarrow 3$” aus Abschnitt A.1.2)

$$\operatorname{Bild}\mathbf{K}_\mathbf{a}^T = \operatorname{Bild}\mathbf{B}\mathbf{K}_\mathbf{a}^T \tag{A.24}$$

Dadurch folgt wegen $\mathbf{K_a}\mathbf{K}_{\mathbf{b}^\perp}^T = \mathbf{\Theta}$ schon $\mathbf{K_a}\mathbf{B}\mathbf{K}_{\mathbf{b}^\perp}^T = \mathbf{\Theta}$. Der Rest ergibt sich wie oben: Da $\mathbf{K_a}$ Eigenräume beider Kovarianzmatrizen aufspannt, sind auch die Bilder der optimalen Dekodierer gleich, $\operatorname{Bild}\mathbf{D}_\mathbf{a}^\mathbf{A} = \operatorname{Bild}\mathbf{D}_\mathbf{a}^\mathbf{B}$.

Literaturverzeichnis

[1] P. Alvarez and L. R. Squire. Memory consolidation and medial temporal lobe: a simple network model. *Proceedings of the National Academy of Sciences of the United States of America*, 91:7041–7045, 1994.

[2] D. G. Amaral and M. P. Witter. The three-dimensional organization of the hippocampal formation: A review of anatomical data. *Neuroscience*, 31(3):571–591, 1989.

[3] P. Ambrogini, R. Cuppini, C. Cuppini, S. Ciaroni, T. Cecchini, P. Ferri, S. Sartini, and P. Del Grande. Spatial learning effects immature granule cell survival in adult rat dentate gyrus. *Neuroscience Letters*, 286:21–24, 2000.

[4] M. Barinaga. Newborn neurons search for meaning. *Science*, 299:32–34, 2003.

[5] A. Barnea and F. Nottebohm. Recruitment and replacement of hippocampal neurons in young and adult chickadees: An addition to the theory of hippocampal learning. *Proceedings of the National Academy of Sciences of the United States of America*, 93:714–718, Jan. 1996.

[6] C. A. Barnes, B. L. McNaughton, S. J. Mizumori, B. W. Leonard, and L. H. Lin. Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progress in Brain Research*, 83:287–300, 1990.

[7] J. J. Chrobak, A. Lörincz, and G. Buzsáki. Physiological patterns in the hippocampo-entorhinal cortex system. *Hippocampus*, 10(4):457–465, 2000.

[8] H. Eichenbaum. The topography of memory. *Nature*, 402(6762):597–599, Dec. 1999.

[9] S. E. Fahlmann and C. Lebiere. The cascade-correlation learning architecture. In D. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, NIPS 1989, pages 524–532. Morgan-Kaufmann, 1990.

[10] M. J. Frank, J. W. Rudy, and R. C. OŔeilly. Transitivity, flexibility, conjunctive representations, and the hippocampus. ii. a computational analysis. *Hippocampus*, 13:299–312, 2003.

[11] T. F. Freund and G. Buzsáki. Interneurons of the hippocampus. *Hippocampus*, 6:347–470, 1996.

[12] B. Fritzke. Growing cell structures – a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460, 1994.

[13] P. E. Gilbert, R. P. Kesner, and I. Lee. Dissociating hippocampal subregions: A double dissociation between dentate gyrus and CA1. *Hippocampus*, 11:626–636, 2001.

[14] E. Gould, A. Beylin, P. Tanapat, A. Reeves, and T. J. Shors. Learning enhances adult neurogenesis in the hippocampal formation. *Nature Neuroscience*, 2(3):260–265, Mar. 1999.

[15] A. J. Harding, G. M. Halliday, and J. J. Kril. Variation in hippocampal neuron number with age an brain volume. *Cerebral Cortex*, 8:710–718, Dec. 1998.

[16] N. B. Hastings and E. Gould. Rapid extension of axons into the ca3 region by adult-generated granule cells. *Journal of Comparative Neurology*, 413:146–154, 1999.

[17] N. B. Hastings, P. Tanapat, and E. Gould. Comparative views of adult neurogenesis. *The Neuroscientist*, 6(5):315–325, 2000.

[18] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA., 1991.

[19] J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 81:3088–3092, May 1984.

[20] S. Káli and P. S. Dayan. The involvement of recurrent connections in area CA3 in establishing the properties of place fields: a model. *The Journal of Neuroscience*, 20(19):7463–7477, Oct. 2000.

[21] E. R. Kandel, J. H. Schwartz, and T. M. Jessel, editors. *Neurowissenschaften. Eine Einführung*. Spektrum Akademischer Verlag, 1995.

[22] G. Kempermann. Why new neurons? possible functions for adult hippocampal neurogenesis. *The Journal of Neuroscience*, 22(3):635–638, Feb. 2002.

[23] G. Kempermann, H. G. Kuhn, and F. H. Gage. More hippocampal neurons in adult mice living in an enriched environment. *Nature*, 386:493–495, Apr. 1997.

[24] G. Kempermann and L. Wiskott. What is the functional role of new neurons in the adult dentate gyrus. In *Stem cells in the nervous system: Function and clinical implications*, Paris, Jan. 2003. Fondation Ipsen. to be published.

[25] R. P. Kesner, P. E. Gilbert, and G. V. Wallenstein. Testing neural network models of memory with behavioral experiments. *Current Opinion in Neurobiology*, 10(2):260–265, Apr. 2000.

[26] W. B. Levy. A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, 6(6):579–590, 1996.

[27] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

[28] D. Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London, Series B*, 262(841):23–81, July 1971.

[29] J. L. McClelland and N. H. Goddard. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6(6):654–665, 1996.

[30] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, July 1995.

[31] R. G. M. Morris, P. Garrud, J. Rawlings, and J. O'Keefe. Place navigation is impaired in rats with hippocampal lesions. *Nature*, 297:681–683, 1982.

[32] L. Nadel and V. Bohbot. Consolidation of memory. *Hippocampus*, 11(1):56–60, 2001.

[33] L. Nadel and H. Eichenbaum. Introduction to the special issue on place cells. *Hippocampus*, 9(4):341–345, 1999.

[34] E. Oeljeklaus and R. Remmert. *Lineare Algebra I.* Heidelberger Taschenbücher Bd. 150, Berlin – Heidelberg – New York, 1974.

[35] J. O'Keefe. A review of the hippocampal place cells. *Progress in Neurobiology*, 13(4):419–439, 1979.

[36] S. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6:270–284, 1994.

[37] E. T. Rolls. Hippocampo-cortical and cortico-cortical backprojections. *Hippocampus*, 10:380–388, 2000.

[38] E. T. Rolls. Memory systems in the brain. *Annual Review of Psychology*, 51:599–630, 2000.

[39] E. T. Rolls and A. Treves. *Neural Networks and Brain Function.* Oxford University Press, 1998.

[40] S. R. Schulz and E. T. Rolls. Analysis of information transmission in the schaffer collaterals. *Hippocampus*, 9:582–598, 1999.

[41] B. B. Thompson, R. J. Marks II, J. j. Choi, A. M. El-Sharkawi, M.-Y. Huang, and C. Bunje. Implicit learning in autoencoder novelty assessment. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, pages 2878–2883. 2002 IEEE World Congress on Computational Intelligence, Honolulu, May12-17, 2002.

[42] A. Treves and E. T. Rolls. Computational analysis of the role of the hippocampus in memory. *Hippocampus*, 4(3):374–391, June 1994.

[43] N. N. Urban, D. A. Henze, and G. Barrionuevo. Revisiting the role of the hippocampal mossy fiber synapse. *Hippocampus*, 11:408–417, 2001.

[44] T. van Groen, P. Miettinen, and I. Kadish. The entorhinal cortex of the mouse: Organization of the projection to the hippocampal formation. *Hippocampus*, 13:133–149, 2003.

[45] H. van Praag, G. Kempermann, and F. H. Gage. Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nature Neuroscience*, 2:266–270, 1999.

[46] H. van Praag, G. Kempermann, and F. H. Gage. Neural consequences of enviromental enrichment. *Nature Reviews Neuroscience*, 2000.

[47] H. van Praag, A. F. Schinder, B. R. Christie, N. Toni, T. D. Palmer, and F. H. Gage. Functional neurogenesis in the adult hippocampus. *Nature*, 415:1030–1034, 2002.

[48] O. S. Vinogradova. Hippocampus as comperator: Role of the two input and two output systems of the hippocampus in selection and registration of information. *Hippocampus*, 11:578–598, 2001.

[49] M. J. West and L. Slomianka. Corrigendum: Total number of neurons in the layers of the human enthorinal cortex. hippocampus 1998;8:69–82. *Hippocampus*, 8:426, 1998.

[50] G. M. Wittenberg, M. R. Sullivan, and J. Z. Tsien. Synaptic reentry reinforcement based network model for long-term memory consolidation. *Hippocampus*, 12:637–647, 2002.

[51] S. Zola-Morgan and L. R. Squire. The primate hippocampal formation: evidence for a time-limited role in memory storage. *Science*, 250(4978):288–290, Oct. 1990.

Hiermit erkläre ich, Malte Rasch, dass zur Anfertigung dieser Diplomarbeit keine unerlaubten Hilfeleistungen in Anspruch genommen wurden.