Propriedades de Recuperação de Memórias em Redes Neurais Atratoras

Camilo Rodrigues Neto

Tese apresentada ao Instituto de Física de São Carlos, da Universidade de São Paulo, para obtenção do título de Doutor em Ciências: Física Básica.

Orientador: José Fernando Fontanari

São Carlos
1997

Rodrigues Neto, Camilo

Propriedades de Recuperação de Memórias em

Redes Neurais Atratoras

113 p.

Tese (Doutorado) – Instituto de Física de São Carlos, 1997.

Orientador: Prof. Dr. José Fernando Fontanari

1. Redes Neurais Atratoras. 2. Propriedades de Recuperação.

I. Título.

UNIVERSIDADE DE SÃO PAULO
Instituto de Física de São Carlos

Av. Dr. Carlos botelho, 1465
CEP 13560-250 - São Carlos - SP
Brasil

Fone (016) 274-3444
Fax (016) 272-2218

MEMBROS DA COMISSÃO JULGADORA DA TESE DE DOUTORADO DE **CAMILO RODRIGUES NETO** APRESENTADA AO INSTITUTO DE FÍSICA DE SÃO CARLOS, UNIVERSIDADE DE SÃO PAULO, EM 05/06/1997.

COMISSÃO JULGADORA:

Prof. Dr. José Fernando Fontanari (IFSC-USP)

Prof. Dr. Luciano da Fontoura Costa (IFSC-USP)

Prof. Dr. José Roberto Drugowich de Felício (FFCLRP-USP)

Prof. Dr. Nestor Felipe Caticha Alfonso (IF-USP)

Prof. Dr. Walter K.Theumann (UFRS)

C:\Meus Documentos\WLA\OFDEFD.DOC

A Kim Hyung Mi, companheira de descobertas e cúmplice de projetos futuros.

Agradecimentos

No longo percurso que separou o impulso inicial da escolha de uma nova área de trabalho e a tremenda energia necessária para concluir esta tese, uma multidão de questões pessoais se impôs com prioridades inadiáveis. Compatibilizar a urgência das questões cotidianas com a necessidade de calma, reflexão e visão de longo prazo da atividade acadêmica foi mais fácil devido ao auxílio de um cem número de pessoas.

Particularmente, a contribuição de meus pais para minha estadia em Rio Claro foi inestimável. Além disso, estarei sempre em dívida para com eles pelos dias e noites que passamos juntos no Pingo d'Água.

Ao Virgílio, pelo apoio nos primeiros tempos de São Carlos, à Dani pelas primeiras conversas profissionais, ao Salviano, Márcio, Mauro, Manzoli e Reginaldo, pelo apoio computacional e ao Domingos, que nos últimos tempos tem sido um contraponto para discussões emocionadas de Física, a todos meus sinceros agradecimentos pela companhia e estímulo.

Devo um agradecimento especial ao Fontanari, que foi, sem dúvida, o principal responsável pelo bom andamento dos trabalhos. Acima de tudo, um grande exemplo de profissionalismo e dedicação. Também de grande importância foram as reuniões semestrais em sua casa com todo o grupo.

Ao professores, funcionários e colegas do IFSC, pelo apoio, a infra-estrutura e estímulo.

Finalmente, devo citar o apoio financeiro do CNPq e da FAPESP para a realização desta tese.

Índice

Agradecimentos **i**

Índice **ii**

Lista de Figuras **iv**

Resumo **1**

Abstract **2**

1 Introdução **3**

2 Mecânica estatística do processo de aprendizado **18**

2.1 Padrões com viés na pseudo-inversa 19

2.1.1 Cálculo da energia livre 19

2.1.2 Cálculo da distribuição de probabilidade das estabilidades . . 26

2.2 Pesos ótimos 30

3 Diagramas de fase para redes extremamente diluídas **32**

3.1 Determinação dos diagramas de fase 34

3.2 Pseudo-inversa 35

3.3 Pesos Ótimos 40

4 Vizinhança dos pontos fixos **46**

4.1 Pseudo-inversa 47

4.1.1 Cálculo da distribuição de probabilidade das estabilidades . . 47
4.2 Pesos ótimos 53

5 Natureza dos atratores **58**
5.1 Determinação dos pesos sinápticos 59
5.2 Determinação dos atratores 61
5.3 Análise dos resultados 64

6 Categorização na pseudo-inversa **77**
6.1 Estabilidade de conceitos categorizados através de exemplos 78
6.1.1 Cálculo da energia livre 78
6.1.2 Cálculo da distribuição de probabilidade das estabilidades dos conceitos 86
6.2 Estabilidade dos exemplos 93
6.2.1 Cálculo da distribuição de probabilidade das estabilidades dos exemplos 94
6.3 Análise termodinâmica para P finito 96

7 Conclusão **101**

Lista de Figuras

3.1 Gráfico da função $g(m)$ para diversos valores de α e T. 35

3.2 Diagrama de fase da rede neural atratora pseudo-inversa. 36

3.3 Zeros da função $g(m)$ em função de α com T constante. 37

3.4 Bacia de atração para a pseudo-inversa à $T = 0$. 38

3.5 Diagrama de fase da rede neural atratora dos pesos ótimos. 40

3.6 Diagrama de fase da rede neural atratora ótima no plano $T = 0$. . . . 41

3.7 Linha tricrítica no espaço de parâmetros (α, κ, T). 42

3.8 Diagrama de fase da rede neural atratora ótima. 44

4.1 Fração de sítios instáveis do padrão de teste na pseudo-inversa. 51

4.2 Fração desítios instáveis do padrão de teste com $\kappa = 0.0$. 53

4.3 O mesmo da figura 4.2, mas com $\kappa = 0.470$. 54

4.4 O mesmo da figura 4.2, mas com $\kappa = 0.5$. 56

4.5 O mesmo da figura 4.2, mas com $\kappa = 0.8$. 57

5.1 Função $g_{pf}(N, \alpha) = \frac{1}{N} \ln \langle \mathcal{N}_{pf} \rangle$ contra $1/N$. 65

5.2 Função $g_c(N, \alpha) = \frac{1}{N} \ln \langle \mathcal{N}_c \rangle$ contra o inverso do número de neurônios. 66

5.3 Y_2 contra o inverso do número de neurônios $1/N$. 67

5.4 Previsão para o segundo momento no caso de bacias uniformes. . . . 68

5.5 A bacia de atração dos estados memorizados. 69

5.6 Função $g_{pf}(N, \alpha)$ contra α. 70

5.7 Função $g_c(N, \alpha)$ contra α. 71

5.8 $\langle Y_2 \rangle$ contra α. 72

5.9 Previsão para o segundo momento no caso de bacias uniformes. . . . 73

5.10 Estimativa da bacia de atração dos estados memorizados. 74
5.11 Função $g_c(N, \alpha)$ contra α para a dinâmica paralela. 76
5.12 $\langle Y_2 \rangle$ contra α para a dinâmica paralela. 77
6.1 Erro de categorização contra α. 92
6.2 Erro de categorização contra d. 93
6.3 Erro de categorização contra s. 94
6.4 Fração de sítios instáveis dos exemplos contra α. 98
6.5 Fração de sítios instáveis dos exemplos contra s. 99

Resumo

Redes neurais atratoras são redes de neurônios artificiais com realimentação e sem estrutura de conexão pré-definida. Estes tipos de redes apresentam uma rica dinâmica dissipativa e são freqüentemente utilizadas como memórias associativas. Tais dispositivos têm a propriedade de recuperar uma memória previamente armazenada, mesmo quando expostos à informação parcial ou degradada daquela memória. Armazenar uma memória significa criar um atrator para ela na dinâmica da rede e isto é feito especificando-se adequadamente os pesos sinápticos. Nesta tese, nos concentramos basicamente em duas maneiras de se definir os pesos sinápticos, que dão origem ao modelo da *pseudo-inversa* e ao modelo dos *pesos ótimos*.

Para redes neurais extremamente diluídas, onde a conectividade C e o número de neurônios N satisfazem à condição $C \ll \ln N$, obtivemos os diagramas de fase no espaço completo de parâmetros dos modelos da pseudo-inversa e dos pesos ótimos através da análise da dinâmica da correlação de recuperação dos padrões armazenados.

Além disso, investigamos as propriedades de recuperação de redes neurais completamente conectadas através de duas abordagens: a investigação analítica da vizinhança dos padrões armazenados e a enumeração exaustiva dos atratores por meio de simulações numéricas.

Finalmente, estudamos analiticamente o problema da categorização no modelo da pseudo-inversa. A categorização em redes neurais atratoras é a capacidade da rede treinada com exemplos de um conceito desenvolver um atrator para este conceito.

Abstract

Attractor neural networks are feedback neural networks with no pre-defined connection structure. These types of neural networks present a rich dissipative dynamics and, in general, are used as associative memory devices. Such devices have the capacity to retrieve a previously stored memory, even when exposed to partial or degraded information. To store a memory means to create an attractor for it in the network dynamics, and this is done by specifying the set of synaptic weighs. In this thesis, we concentrate on two classical ways of specifying the synaptics weighs: the pseudo-inverse and the optimal weighs models.

For extremelly diluted neural networks, for which the connectivity C and the number of neurons N satisfy the condition $C \ll \ln N$, we obtain the phase diagrams in the complete space of the model parameters through the analytical study of the retrieval overlap dynamics.

We also investigate the retrieval properties of fully connected neural networks using two approaches: the analytical study of the neighborhood of the stored patterns, and the exhaustive enumeration of the attractors via numerical simulations.

Finally, we study analytically the problem of categorization in the pseudo-inverse model. Categorization in attractor neural networks is the capacity to create an attractor for a concept to which the network has had access only through a finite number of examples.

Capítulo 1

Introdução

Talvez o objetivo mais audacioso do estudo de redes neurais seja entender o funcionamento do cérebro humano. Nas últimas décadas, fizeram-se grandes progressos na compreensão de como redes de neurônios realizam o processamento sensorial, particularmente em como o cérebro faz o processamento visual. Entretanto, deseja-se não só entender como o cérebro realiza essas funções, mas também as ditas funções superiores, como o aprendizado, a inteligência e a consciência. Além disso, o estudo de redes neurais pretende descobrir como se poderia construir redes de neurônios artificiais com tais propriedades. Apesar desses ambiciosos objetivos, a implementação destas funções através de redes neurais tem encontrado grandes dificuldades e limitações. Mesmo quando utilizadas como memórias associativas ou para fazer reconhecimento de padrões – aplicações onde as redes neurais são mais bem sucedidas – não se pode dizer que o sucesso tenha sido total.

Os modelos de redes neurais têm demonstrado possuir algumas funções cognitivas, como as já citadas *memória associativa* e *aprendizado*. São capazes também de *categorização*, que é a habilidade de criar um atrator para um conceito a partir de exemplos com características comuns. Os modelos de neurônios utilizados são uma simplificação fabulosa do neurônio real. A pergunta que se coloca é saber como um conjunto de elementos simples pode apresentar comportamentos tão complexos como memória, aprendizado e categorização, entre outros. Mesmo se considerar-

mos o neurônio real, com toda a riqueza de detalhes ainda parcialmente conhecidos, como uma rede composta de tais neurônios pode apresentar comportamentos tais como as funções de raciocínio e abstração? Nas últimas décadas, grande atenção vem sendo dada ao estudo do comportamento coletivo de conjuntos de elementos simples. São as propriedades *emergentes* desses sistemas, que aparecem, em geral, através do fenômeno das *transições de fase*, tornando seus comportamentos imprevisíveis e surpreendentes. Exemplo de comportamento coletivo são as propriedades magnéticas de sistemas de spins. Como lá, não seria a inteligência uma propriedade emergente do comportamento coletivo de grandes conjuntos de neurônios?

No esforço de compreender melhor as redes neurais, uma das abordagens possíveis, particularmente à qual os físicos têm-se envolvido mais, é a modelagem teórica e numérica de redes de neurônios artificiais. Antes de descrevermos o neurônio *artificial*, concentremo-nos por um momento no neurônio *real* e em redes desses neurônios, em especial no cérebro de mamíferos superiores.

Muito simplificadamente, o neurônio pode ser dividido em três partes: os *dendritos*, que coletam os sinais de outros neurônios; o *soma*, que transforma os sinais de entrada em um sinal de saída; e o *axônio*, que transmite o sinal de saída para outros neurônios. Numa rede neural os neurônios estão conectados através de *sinápses*, que são pontos de contato entre dendritos e axônios. No cérebro de mamíferos, cada neurônio está conectado em média com 10^4 outros neurônios, tendo o cérebro humano em média 10^{10} neurônios [1]. Os neurônios comunicam-se por meio de pulsos elétricos que se propagam através dos axônios mantendo sua forma e amplitude, até atingirem as sinápses, onde neurotransmissores fazem a ponte com os dendritos de outros neurônios, criando o *potencial pós-sináptico*. Os potenciais pós-sinápticos gerados na árvore dendrítica propagam-se com atenuação para o soma, onde são integrados. Se a somatória dos potenciais pós-sinápticos que chegam ao soma num certo intervalo de tempo é superior a um certo limite, o neurônio dispara um pulso que se propaga por seu axônio, alcançando as conexões sinápticas com os dendritos dos neurônios aos quais está conectado. A contribuição de uma entrada pré-sináptica

ao potencial pós-sináptico caracteriza a *eficiência sináptica.* Uma sinápse pode ser tanto excitatória como inibitória, mas como existe uma grande variedade de fontes de ruído sináptico, a transmissão de um sinal através de uma sinápse acaba tendo um caráter probabilístico [2].

Quanto à arquitetura das conexões, as redes de neurônios artificiais ou simplesmente redes neurais podem ser divididas em duas grandes categorias: uma de redes com estruturas de camadas *sem realimentação*, da qual são exemplos os *perceptrons*; e outra de redes *com realimentação*, sem uma estrutura geométrica definida, chamadas de *redes neurais recorrentes* ou *redes neurais atratoras.* As redes tipo perceptron têm uma dinâmica muito simples: a informação propaga-se num fluxo contínuo de atividade neural da primeira para a última camada. Esse tipo de rede é encontrado nos estágios iniciais do processamento sensorial. Já as redes recorrentes apresentam uma rica dinâmica de atratores. A entrada é assumida como sendo o estado inicial da rede, que evolui segundo sua dinâmica para atratores, pontos fixos ou ciclos da dinâmica. Sob certas condições a dinâmica da rede pode apresentar comportamento caótico [3] [4].

Como o próprio nome indica, o sistema físico denominado *rede neural* recebeu forte influência da analogia com sistemas biológicos reais. Embora muitos dos termos utilizados no estudo desses sistemas físicos preservem os nomes provenientes da motivação original, a abordagem física desse problema ganhou a autonomia de uma disciplina independente.

Dada a complexidade do neurônio real, nos modelos de redes neurais são utilizadas muitas simplificações. Procura-se pelo modelo mais simples de neurônio, e da rede como um todo, que possa abarcar o máximo de propriedades relevantes das redes reais. Com tal propósito em mente, McCulloch e Pitts [5] introduziram, em 1943, a noção de *neurônio formal* como um elemento lógico simples de dois estados e mostraram que redes desses elementos podiam implementar funções lógicas. A unidade básica para os modelos de redes neurais artificiais abordados nesse trabalho são estes neurônios formais, reduzidos a elementos lógicos simples de dois estados,

com funções de ativação bem definidas, eventualmente estocásticas.

Uma rede neural artificial é composta por um conjunto desses neurônios que interagem através dos pesos sinápticos. Hebb propôs em 1949 aquela que viria a ser a regra mais utilizada para especificar os pesos sinápticos [6]. Nesta proposta, a representação de conceitos no cérebro se faria pela taxa de disparo de conjuntos de neurônios e o aprendizado pela modificação dos pesos sinápticos entre os neurônios. Os pesos sinápticos podem assumir valores reais e, como veremos em seção posterior desta tese, dependem fortemente de sua regra definidora.

Nesta tese, nos concentraremos na utilização das redes neurais como memórias associativas. Tal dispositivo tem a propriedade de recuperar uma memória previamente armazenada, mesmo quando exposto à informação parcial ou degradada daquela memória. Na ausência de ruído, o estado de um neurônio no instante $t+1$ é função da soma dos estados dos neurônios com que interage mediados pelos pesos sinápticos J_{ij},

$$S_i(t+1) = \text{sign}\left[\frac{1}{\sqrt{C}}\sum_j J_{ij}S_j(t)\right]. \tag{1.1}$$

Aqui $S_i(t)$ é o estado do $i-ésimo$ neurônio no instante t, que pode assumir os valores -1 e 1 se o neurônio estiver, respectivamente, inativo ou ativo; J_{ij} é o acoplamento sináptico entre os neurônios i e j; os índices i e j correm sobre todos os neurônios da rede, variando de 1 a N; e C é a conectividade da rede neural.

Em redes neurais atratoras, um padrão (ou memória) l é representado pelo vetor $\boldsymbol{\xi}^l = \left(\xi_1^l, \xi_2^l, \ldots, \xi_N^l\right)$, onde ξ_i^l é o estado do neurônio i quando o padrão l é recuperado. Uma rede neural é dita uma memória associativa se para um dado conjunto de padrões armazenados $\left\{\boldsymbol{\xi}^l\right\}$ com $l = 1, \ldots, P$, as respectivas configurações são atratores da dinâmica neural. Dada uma configuração inicial da rede $S_i(t)$, a proximidade do padrão $\boldsymbol{\xi}^l$ a este estado, chamada de *correlação de recuperação* com o padrão armazenado l, é definida como

$$m^l(t) = \frac{1}{N}\sum_i \xi_i^l S_i(t). \tag{1.2}$$

Dado a correlação de recuperação do estado inicial da rede $m^l(0)$ com o padrão $\boldsymbol{\xi}^l$, esta evolui de acordo com sua dinâmica e, eventualmente, alcança o padrão armazenado, tal que $m^l(\infty) \approx 1$. Daí a denominação de memória endereçável por conteúdo para este tipo de sistema. Em geral, as componentes dos padrões ξ_i^l são variáveis aleatórias estatisticamente independentes, geradas pela distribuição de probabilidade

$$p\left(\xi_i^l\right) = \frac{1}{2}\delta\left(\xi_i^l - 1\right) + \frac{1}{2}\delta\left(\xi_i^l + 1\right), \tag{1.3}$$

onde $\delta(x)$ é a função delta de Dirac.

Pode-se dizer que o envolvimento da Física com o estudo de redes neurais ganhou maior impulso a partir 1982 com o trabalho de Hopfield [7], onde foi lançado o conceito de *energia computacional*, que caracteriza o estado da rede. Nesta abordagem, a energia associada a determinado estado $\mathbf{S}$ é dada por

$$E(\mathbf{S}) = -\frac{1}{2}\sum_{ij}^{N} J_{ij} S_i S_j, \tag{1.4}$$

com os pesos sinápticos dados pela regra de Hebb

$$J_{ij} = \frac{1}{N}(1 - \delta_{ij})\sum_{l=1}^{P} \xi_i^l \xi_j^l. \tag{1.5}$$

Aqui, a conectividade é completa $C = N - 1$, isto é, cada neurônio interage com todos os outros neurônios da rede. O que possibilitou descrever a rede através da função energia e o processo de aprendizado, ou memorização dos padrões, como a criação de mínimos dessa energia, foi justamente a regra de Hebb, que leva a redes com acoplamentos simétricos $J_{ij} = J_{ji}$ e autoacoplamentos nulos $J_{ii} = 0$. Se o aprendizado é a criação de mínimos na função energia, a recuperação dos padrões armazenados se faz pela procura desses mínimos.

O modelo proposto por Hopfield foi extensivamente estudado na literatura, apesar da prescrição de Hebb, utilizada para os pesos sinápticos, não garantir que os padrões armazenados sejam pontos fixos da dinâmica no regime em que P cresce linearmente com a conectividade, isto é, $P = \alpha C$. De fato, para $\alpha < \alpha_c \approx 0.144$, os pontos fixos atratores estão muito próximos dos padrões armazenados e o modelo de

Hopfield, mesmo neste regime, continua útil como memória associativa. Este valor máximo da razão P/C é denominado capacidade de armazenamento. Para $\alpha = \alpha_c$, a rede neural sofre uma transição descontínua para o regime de não recuperação, onde a correlação de recuperação $m^l(\infty)$ é da ordem de $1/\sqrt{N}$ [7] [8].

A transmissão de sinal pelas sinápses é um processo repleto de perturbações que introduzem ruído e, conseqüentemente, uma componente estocástica na dinâmica da rede. Na versão estocástica da dinâmica determinística (1.1), proposta por Amit, Gutfreund e Sompolinsky [9] [10], a probabilidade de um neurônio assumir o valor $\sigma = \pm 1$ no instante $t+1$ é

$$P_i[\sigma] = \frac{1}{1 + \exp[-2\sigma\beta h_i(t)]}, \tag{1.6}$$

onde $T = 1/\beta$ é por definição a temperatura que mede a intensidade de ruído e $h_i(t)$ é denominado de campo local no sítio i no instante t, definido por

$$h_i(t) = \frac{1}{\sqrt{C}} \sum_j J_{ij} S_j(t). \tag{1.7}$$

No limite de ruído nulo, quando $\beta \to \infty$, recuperamos a dinâmica determinística (1.1).

Outro momento importante na abordagem Física do problema de redes neurais ocorreu quando o modelo proposto por Hopfield, agora com a introdução da dinâmica estocástica, foi investigado analiticamente no limite $N \to \infty$ por Amit, Gutfreund e Sompolinsky [9] [10], empregando os métodos da mecânica estatística de sistemas desordenados desenvolvidos previamente para o problema de vidros de spin. O estudo anterior do comportamento de sistemas heterogêneos, como os vidros de spin, permitiu estabelecer uma ponte com o estudo das redes neurais. Como o neurônio formal, o spin de Ising é um elemento de dois estados, assim, por analogia uma rede neural pode ser pensada como um sistema de spins de Ising e a similaridade com o modelo de Sherrington-Kirkpatrick (SK) [11] para vidros de spin é imediata. Aqui, o estado dos neurônios são variáveis rápidas[1], enquanto os padrões são variáveis

[1] Tradução livre do inglês *annealed*.

lentas[2]. As propriedades de equilíbrio da rede de Hopfield podem ser caracterizadas a partir da energia livre por sítio

$$f = -\frac{T}{N} \langle\langle \ln Z \rangle\rangle \,, \tag{1.8}$$

onde Z é a função de partição dada por

$$Z = \mathrm{Tr}_{\mathbf{S}} \exp\left[-\beta E\left(\mathbf{S}\right)\right], \tag{1.9}$$

$\mathrm{Tr}_{\mathbf{S}}$ é a soma sobre todos os estados da rede e $\langle\langle \ldots \rangle\rangle$ simboliza a média sobre os padrões armazenados $\boldsymbol{\xi}^l$.

Para realizar a média de (1.8), é necessário recorrer ao método das réplicas, que consiste em primeiro calcular as médias de n funções de partição desacopladas $\langle\langle Z^n \rangle\rangle$ para n inteiro e, então, efetuar a continuação analítica $n \longrightarrow 0$, obtendo $\langle\langle \ln Z \rangle\rangle$ através da identidade

$$\langle\langle \ln Z \rangle\rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle\langle Z^n \rangle\rangle \,. \tag{1.10}$$

Existem alguns problemas com o método das réplicas, como por exemplo, o fato de termos iniciado com n inteiro no cálculo de $\langle\langle Z^n \rangle\rangle$ e passado para n real quando tomamos o limite $n \longrightarrow 0$. Também invertemos a ordem dos limites $n \longrightarrow 0$ e $N \longrightarrow \infty$ [12]. Existem poucos resultados matemáticos rigorosos que justifiquem tais procedimentos. De fato, a maior justificativa tem sido os resultados de simulações, que, em geral, concordam bastante bem com a teoria. Em certos limites e modelos, é necessário, por exemplo, utilizar a quebra de simetria de réplicas para se obter resultados corretos [13].

Em geral, os resultados das análises feitas através da mecânica estatística são apresentados na forma de um diagrama de fases no plano (α, T). No modelo de Hopfield são encontradas três fases diferentes, a saber, a fase de recuperação, a fase de vidro de spin e a fase paramagnética. Para $T = 0$, a estimativa do cálculo de réplicas simétricas para a capacidade de armazenamento crítica é $\alpha_c \approx 0.138$.

[2]Tradução livre do inglês *quenched*.

Outra contribuição importante para o estudo do modelo de Hopfield ocorreu em 1987, quando Derrida, Gardner e Zippelius [14] resolveram exatamente a dinâmica neural no limite de diluição aleatória extrema, ou seja, quando a conectividade C satisfaz a condição $C \ll \ln N$. Neste limite, a fase de vidro de spin não existe e a transição entre a fase de recuperação e a fase paramagnética é contínua, ocorrendo em $\alpha_c = 2/\pi$ à temperatura zero. Outros regimes de diluição foram considerados na literatura, como por exemplo, Sompolinsky [15] [16], que estudou as propriedades de equilíbrio do modelo de Hopfield aleatoriamente diluído no caso em que C é da mesma ordem de N.

Para armazenar padrões, ou seja, criar configurações da rede neural que sejam atratores de sua dinâmica, é necessário especificar adequadamente o conjunto de interações entre os elementos dessa rede, os pesos sinápticos. Um padrão será um atrator se o campo $h_i(t)$ sobre cada neurônio da rede for tal que estabilize seu estado, isto é, se $\xi_i^l h_i(t) > 0$. A quantidade

$$\Delta_i^l \equiv \xi_i^l h_i(t) = \frac{1}{\sqrt{C}} \xi_i^l \sum_j J_{ij} \xi_j^l(t) \tag{1.11}$$

é denominada *estabilidade* do sítio i do padrão l e desempenha um papel fundamental nas análises subseqüentes.

O processo de especificar os pesos sinápticos é chamado de aprendizado e pode ser feito por duas abordagens. A primeira, da qual é exemplo a regra de Hebb, assume uma dependência específica dos J_{ij} com os padrões armazenados e foi historicamente a primeira linha de pesquisa, que deu origem ao modelo de Hopfield. A outra abordagem consiste em tratar os pesos sinápticos como variáveis e ajustá-los de modo a satisfazer certos vínculos. Esta é a proposta de Gardner [17] [18], que, apresentada em 1988, significou uma mudança qualitativa no entendimento e no tratamento teórico do problema. Tratando os estados da rede como variáveis lentas e as interações entre os neurônios como variáveis rápidas, esta abordagem permitiu tratar o problema de interações não simétricas entre os neurônios e obter propriedades de classes de redes neurais, em vez de realizações específicas dessas.

As propriedades de recuperação de padrões em redes neurais são fortemente dependentes da regra utilizada para especificar os pesos sinápticos. Nesta tese, seguindo o espírito da abordagem de Gardner, nos concentramos em dois modelos: o modelo da *pseudo-inversa* e o modelo dos *pesos ótimos*. O modelo da rede neural atratora pseudo-inversa [19], que pode armazenar perfeitamente um conjunto de N padrões linearmente independentes, foi pela primeira vez estudado no contexto da mecânica estatística por Personnaz, Guyon e Dreyfus [20], utilizando uma prescrição explícita para os pesos sinápticos dada por

$$J_{ij} = \frac{1}{N} \sum_{l,k=1}^{P} \xi_i^l \xi_j^k \left(C_{lk}\right)^{-1}, \tag{1.12}$$

onde $C_{lk} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^l \xi_i^k$ são os elementos da matriz de correlação entre os P padrões armazenados. Esta prescrição para os pesos sinápticos tem os termos diagonais diferentes de zero e é denominado no restante desta tese de modelo PGD da pseudo-inversa. As propriedades de equilíbrio desta mesma regra, mas com os termos diagonais zerados, denominado daqui em diante de modelo KS da pseudo-inversa, foram estudadas por Kanter e Sompolinsky [21] através da técnica de réplicas simétricas no regime de temperatura diferente de zero.

O modelo da pseudo-inversa também pode ser analisado através da técnica de Gardner. Neste caso, os pesos J_{ij} são obtidos tomando a solução de menor norma do conjunto de P equações lineares[3]

$$\Delta_i^l \equiv \frac{1}{\sqrt{C}} \xi_i^l \sum_{j \neq i} J_{ij} \xi_j^l = 1, \tag{1.13}$$

com $l = 1, \ldots, P$ para cada sítio i da rede, $i = 1, \ldots, N$ [19]. A capacidade de armazenamento da pseudo-inversa para padrões aleatórios é $\alpha_c = 1$, mas deve ser

[3]Considere a equação vetorial $\mathbf{A}x = b$, onde $\mathbf{A}$ é uma matriz $(P \mathrm{x} N)$, e x e b são vetores coluna, $x \in R^N$ e $b \in R^P$. Se $N > P$, existem infinitas soluções e é tomada a de menor norma. A solução de menor norma para x é dada por $x = \mathbf{A}^T \left(\mathbf{A}\mathbf{A}^T\right)^{-1} b$, onde $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \mathbf{A}^T$ é denominada de pseudo-inversa de $\mathbf{A}$ (pg.49, Kohonen [19]).

Se $N < P$, nem sempre existirá uma solução para x e a prescrição dos pesos obtida pela minimização de (1.14) corresponde ao conjunto de pesos com menor desvio quadratico.

notado que, diferentemente do modelo de Hofield, os padrões armazenados são estritamente estáveis abaixo de α_c. O problema da existência de soluções para a equação (1.13) pode ser visto, do ponto de vista da mecânica estatística, como o processo de aprendizado através da minimização de uma energia adequadamente definida, a saber [22] [23]

$$E_i(\mathbf{J}) = \frac{1}{2}\sum_l \left(1 - \Delta_i^l\right)^2. \tag{1.14}$$

O modelo da pseudo-inversa, entretanto, não é ótimo no sentido de que sua capacidade de armazenamento não é máxima. Gardner [17] mostrou como tratar o ensemble de pesos da rede ótima sob o ponto de vista da mecânica estatística. De fato, os pesos da rede ótima devem satisfazer às desigualdades

$$\Delta_i^l \equiv \frac{1}{\sqrt{C}}\xi_i^l \sum_{j\neq i} J_{ij}\xi_j^l \geq \kappa \tag{1.15}$$

para todo i e l, sujeitas à condição

$$\sum_j (J_{ij})^2 = N. \tag{1.16}$$

O parâmetro de margem $\kappa \geq 0$ é introduzido para garantir que os padrões $\boldsymbol{\xi}^l$ possuam bacias de atração finitas, embora não seja óbvio nem tenha sido mostrado que o tamanho das bacias de atração anule-se para $\kappa = 0$. A capacidade de armazenamento α_c para padrões aleatórios diminui com κ e, em particular, $\alpha_c = 2$ para $\kappa = 0$. O tratamento do modelo dos pesos ótimos pela mecânica estatística foi feito por Gardner e Derrida [18], utilizando a seguinte função energia

$$E_i(\mathbf{J}) = \sum_l \theta\left(\kappa - \Delta_i^l\right). \tag{1.17}$$

A determinação dos pesos sinápticos para a rede ótima, equações (1.15) e (1.16), pode ser feita por um algoritmo originalmente proposto para redes tipo perceptron. Iniciando com uma matriz de pesos arbitrária, com $J_{ii} = 0$, examinam-se as desigualdades (1.15) para cada i e toda vez que uma não for satisfeita, os J_{ij} correspondentes são modificados por

$$J_{ij} \to J_{ij} + \frac{\lambda}{N}\xi_i^l\xi_j^l\left(1 - \delta_{ij}\right)\theta\left(\kappa - \Delta_i^l\right), \tag{1.18}$$

onde λ é um parâmetro positivo que ajusta o tamanho do incremento de J_{ij}. A cada iteração, os pesos são renormalizados para satisfazer (1.16). Gardner [24] mostrou que esse algoritmo pode ser aplicado a redes neurais atratoras e Abbott e Kepler [25] apresentaram uma variante mais eficiente desse algoritmo. Entretanto, garantir que os vínculos (1.13) e (1.15) sejam satisfeitos, ou seja, que os padrões armazenados sejam pontos fixos da dinâmica, não garante um bom desempenho da rede como memória associativa. De fato, seu desempenho depende de propriedades como o tamanho das bacias de atração e a estabilidade dos atratores. Daí a necessidade de se estudar a dinâmica dessas redes. O regime de extrema diluição do modelo dos pesos ótimos à $T = 0$ foi estudada por Gardner [24], enquanto Amit, Evans, Horner e Wong [26] generalizaram esta análise para temperaturas não nulas. É importante notar, entretanto, que ambas as análises foram restritas ao regime de saturação $\alpha = \alpha_c(\kappa)$. Já a versão extremamente diluída da pseudo-inversa foi estudada por Opper, Kleinz, Köhler e Kinzel [27] para temperatura zero.

Dada a importância dos modelos da pseudo-inversa e dos pesos ótimos no contexto das redes neurais atratoras e, particularmente no contexto de memórias associativas, a obtenção dos diagramas de fase no espaço completo dos parâmetros relevantes aos modelos em questão foi um dos objetivos a que nos propusemos nesta tese. Assim, estendemos a análise da pseudo-inversa para o regime de temperatura não nula, obtendo o diagrama de fases do modelo no espaço (α, T). Para o modelo dos pesos ótimos, obtivemos os diagramas de fase fora do regime de saturação, no espaço dos parâmetros (α, κ, T). Nossa contribuição original ao estudo das versões diluídas de redes neurais atratoras [28] está apresentada no capítulo 3.

De acordo com Gardner [24] e Kepler e Abbot [29], o ingrediente fundamental para a investigação analítica do regime de extrema diluição é o conhecimento da distribuição de probabilidade das estabilidades

$$P\left(\Delta_i^l\right) = \left\langle\left\langle\left\langle \delta\left(\Delta_i^l - \frac{1}{\sqrt{C}}\xi_i^l \sum_{j\neq i} J_{ij}\xi_j^l\right)\right\rangle_J\right\rangle\right\rangle, \tag{1.19}$$

onde a notação $\langle\langle\ldots\rangle\rangle$ significa a média sobre os padrões ξ_j^l e $\langle\ldots\rangle_J$ a média sobre

o ensemble de pesos que definem o modelo em questão. Devido à independencia estatística dos padrões ξ_j^l, a distribuição de probabilidade das estabilidades é independente dos índices i e l. Para obter a distribuição das estabilidades (1.19), necessária para levantar os diagramas de fase no regime de extrema diluição, acrescentamos à equação (1.14) o termo auxiliar linear em h

$$E_i(\mathbf{J}, h) = E_i(\mathbf{J}) + h\delta\left(\Delta_i^l - \frac{1}{\sqrt{C}}\xi_i^l \sum_{j\neq i} J_{ij}\xi_j^l\right), \tag{1.20}$$

de modo que

$$P\left(\Delta_i^l\right) = -\lim_{\lambda\to\infty} \frac{1}{\lambda}\frac{\partial \langle\langle \ln Z_i\rangle\rangle}{\partial h}\,|_{h=0}, \tag{1.21}$$

onde Z_i é a função de partição

$$Z_i(h) = \int dJ_{ij}\delta\left(Q_i - \frac{1}{C}\sum_j J_{ij}^2\right)\exp\left[-\lambda E_i(\mathbf{J}, h)\right] \tag{1.22}$$

e λ^{-1} desempenha o papel de um ruído no processo de aprendizado, que não deve ser confundido com o ruído da dinâmica neural $\beta^{-1} = T$. Neste trabalho, nos restringiremos ao limite $\lambda^{-1} \to 0$, ou $\lambda \to \infty$, de forma que não há necessidade de dar uma interpretação física para este parâmetro. O cálculo da distribuição de probabilidade das estabilidades para a pseudo-inversa é apresentada em detalhe no capítulo 2, uma vez que cálculos semelhantes são utilizados nos capítulos 4 e 6.

As propriedades de recuperação de redes neurais atratoras completamente conectadas ($C = N$) foram analisadas utilizando-se duas abordagens. A primeira abordagem é original e consiste na análise da vizinhança dos padrões armazenados nas redes pseudo-inversa e pesos ótimos. Para isso, calculamos a fração ϵ de sítios instáveis num padrão de teste $\boldsymbol{\eta}^l$ à distância de Hamming d do padrão armazenado $\boldsymbol{\xi}^l$. O padrão de teste é gerado pela distribuição condicional de probabilidades

$$p(\eta_i^l \mid \xi_i^l) = \frac{1+b}{2}\,\delta\left(\eta_i^l - \xi_i^l\right) + \frac{1-b}{2}\,\delta\left(\eta_i^l + \xi_i^l\right), \tag{1.23}$$

onde $0 \leq b \leq 1$ mede a correlação entre η_j^l e ξ_j^l. Este parâmetro está relacionado com a distância de Hamming d entre os dois padrões por $d = (1-b)/2$. Nosso

objetivo é obter a fração de sítios instáveis do padrão de teste $\boldsymbol{\eta}^l$, ou seja, a fração entre o número de sítios para os quais a estabilidade

$$\Lambda_i^l = \frac{1}{\sqrt{N}} \eta_i^l \sum_{j \neq i} J_{ij} \eta_j^l \tag{1.24}$$

é negativa e o número total de sítios N. A independência estatística entre η_i^l e ξ_i^l para diferentes sítios permite escrever esta fração como

$$\epsilon = \int_{-\infty}^{0} d\Lambda W(\Lambda), \tag{1.25}$$

onde $W\left(\Lambda^l\right)$ é a distribuição de probabilidade das estabilidades do padrão de teste $\boldsymbol{\eta}^l$,

$$W\left(\Lambda_i^l\right) = \left\langle\left\langle\left\langle \delta\left(\Lambda_i^l - \frac{1}{\sqrt{N}} \eta_i^l \sum_{j \neq i} J_{ij} \eta_j^l\right)\right\rangle_J\right\rangle\right\rangle_{\eta^l, \xi^l}. \tag{1.26}$$

Aqui, a notação $\langle\langle \ldots \rangle\rangle_{\eta^l, \xi^l}$ significa a média sobre os padrões $\boldsymbol{\eta}^l$ e $\boldsymbol{\xi}^l$ enquanto $\langle \ldots \rangle_J$ é a média sobre o ensemble de pesos que satisfaz a equação (1.13) para a pseudo-inversa ou as desigualdades (1.15) para os pesos ótimos.

A dependência de ϵ com d pode dar muita informação sobre as propriedades da vizinhança do padrão armazenado $\boldsymbol{\xi}^l$. Por exemplo, se $\epsilon = d$ de modo que cada sítio invertido torna-se instável, a vizinhança deste padrão é suave. Por outro lado, se um pequeno desvio do ponto fixo $\boldsymbol{\xi}^l$ leva a um aumento abrupto do número de sítios instáveis, sua vizinhança será rugosa. A obtenção da distribuição de estabilidades do padrão de teste (1.26) e da fração de sítios instáveis (1.25) é semelhante à realizada no capítulo 2. A análise da vizinhança dos padrões armazenados para as redes pseudo-inversa e pesos ótimos completamente conectadas, apresentada no capítulo 4, é uma contribuição original ao estudo de redes atratoras [28].

A segunda abordagem que usamos para analisar as propriedades de recuperação de redes neurais atratoras é a enumeração exaustiva dos atratores através de simulações numéricas realizadas por meio de um algoritmo proposto por Gutfreund, Reger e Young [30]. Para redes neurais atratoras com $N \leq 24$, identificamos todos os atratores e estudamos como seu número, suas bacias de atração e, em particular,

as bacias dos padrões armazenados, dependem de α e N. Os resultados dessas simulações, apresentados no capítulo 5, também são originais.

Dentro do contexto de memórias associativas, além da bacia de atração e da estabilidade de um atrator, outra característica importante dos modelos de redes neurais é a capacidade de *categorização*. A categorização é a capacidade de uma rede treinada com exemplos $\boldsymbol{\xi}^{l\nu}$ de um dado conceito $\boldsymbol{\xi}^{l}$, ao qual a rede não tem acesso, criar um atrator para ele. Os exemplos são dados pela distribuição condicional de probabilidades

$$p\left(\xi_i^{l\nu} \mid \xi_i^l\right) = \frac{1+b}{2}\,\delta\left(\xi_i^{l\nu} - \xi_i^l\right) + \frac{1-b}{2}\,\delta\left(\xi_i^{l\nu} + \xi_i^l\right), \tag{1.27}$$

onde $l = 1, \ldots, P$ e $\nu = 1, \ldots, s$, num total de sP exemplos (s exemplos para cada um dos P conceitos).

O erro de categorização dos conceitos é determinado pela fração de sítios instáveis de $\boldsymbol{\xi}^{l}$, dada por

$$\epsilon_c = \int_{-\infty}^{0} d\Lambda W^c(\Lambda), \tag{1.28}$$

onde $W^c\left(\Lambda^l\right)$ é a distribuição de probabilidade das estabilidades do conceito $\boldsymbol{\xi}^{l}$, definida como

$$W^c\left(\Lambda_i^l\right) = \left\langle\left\langle\left\langle \delta\left(\Lambda_i^l - \frac{1}{\sqrt{N}}\xi_i^l \sum_{j\neq i} J_{ij}\xi_j^l\right)\right\rangle_J\right\rangle\right\rangle_{\xi^{l\nu},\xi^l}. \tag{1.29}$$

Aqui, a notação $\langle\langle\ldots\rangle\rangle_{\xi^{l\nu},\xi^l}$ refere-se à média sobre os exemplos $\boldsymbol{\xi}^{l\nu}$ e os conceitos $\boldsymbol{\xi}^{l}$, enquanto $\langle\ldots\rangle_J$ é a média sobre o ensemble de pesos. No capítulo 6, estudamos o problema da categorização em redes neurais atratoras no modelo da pseudo-inversa, como proposto por Fontanari [31]. O que é original aqui é que estudamos a rede pseudo-inversa, enquanto os trabalhos anteriores sobre a categorização em redes neurais atratoras se restringem ao modelo de Hopfield ([31], [32], [33], [34], [35] e [36]).

O propósito deste capítulo introdutório, talvez excessivamente longo, foi o de apresentar um breve histórico da abordagem física do problema de redes neurais e.

também, de colocar claramente as questões e técnicas tratadas no decorrer desta tese. Finalmente, no capítulo 7, resumimos os pontos mais importantes tratados nesta tese.

Capítulo 2

Mecânica estatística do processo de aprendizado

As redes neurais podem ser especificadas por uma energia, ou função custo, adequadamente definida de acordo com a função que se quer que a rede desempenhe. Neste capítulo, utilizaremos o tratamento da mecânica estatística no ensemble canônico proposto por Gardner e Derrida [18] para estudar o modelo da rede pseudo-inversa treinada com padrões estatisticamente independentes com viés, gerados pela distribuição de probabilidade

$$p\left(\xi_j^k\right) = \frac{1+a}{2}\delta\left(\xi_j^k - 1\right) + \frac{1-a}{2}\delta\left(\xi_j^k + 1\right). \tag{2.1}$$

A fim de ilustrar a técnica que utilizaremos novamente nos capítulos 4 e 6, a apresentação dos cálculos será bastante detalhada.

Na seção 2.1, obtemos a energia livre e os parâmetros de ordem para a pseudo-inversa com padrões com viés e, em seguida, obtemos a distribuição de probabilidades das estabilidades, características que utilizaremos nos próximos capítulos. Apresentamos na seção 2.2, para futura referência, a distribuição de probabilidade das estabilidades para o modelo dos pesos ótimos, já obtidos na literatura [24] [29].

2.1 Padrões com viés na pseudo-inversa

A pseudo-inversa pode ser definida como a solução de menor norma do seguinte conjunto de equações lineares [19] [20]

$$\Delta_i^l \equiv \frac{1}{\sqrt{C}} \xi_i^l \sum_{j \neq i} J_{ij} \xi_j^l = 1, \tag{2.2}$$

com $l = 1, \ldots, P$ para cada sítio i da rede, $i = 1, \ldots, N$. Na equação acima, Δ_i^l é denominado estabilidade do sítio i do padrão l. No caso de $P \leq (N-1)$ padrões independentes, existem infinitas soluções para (2.2) e quando $P = N$ a solução é única. Naturalmente, no limite $N \to \infty$, este modelo tem capacidade de armazenamento $\alpha_c = 1$. Em termos de uma energia, a pseudo-inversa pode ser expressa como o cunjunto de pesos $\{J_{ij}\}$ de menor norma $Q_i = \frac{1}{N} \sum_j J_{ij}$ que anula a energia de treinamento

$$E_i(\mathbf{J}) = \frac{1}{2} \sum_k \left(1 - \Delta_i^k\right)^2. \tag{2.3}$$

Escrever a energia dessa forma garante a solução de (2.2), mas o termo entre parênteses poderia ser elevado ao cubo, ou à quarta potência. De fato, a escolha da potência deve ser feita com base nas dificuldades analíticas que uma ou outra opção introduziriam [37]. Seria interessante estudar o efeito dessa potência no número e tipo de atratores da pseudo-inversa.

2.1.1 Cálculo da energia livre

Interessa-nos calcular a energia livre por sítio $f = -\frac{1}{N\lambda} \langle\langle \ln Z \rangle\rangle$, para o que utilizamos o método das réplicas a fim de efetuar a média sobre as variáveis lentas. Esse método consiste em primeiro calcular as médias de n funções de partição desacopladas $\langle\langle Z^n \rangle\rangle$ para n inteiros e então tomar continuação analítica de $n \to 0$, através da identidade

$$\langle\langle \ln Z \rangle\rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle\langle Z^n \rangle\rangle. \tag{2.4}$$

onde $Z^n = \prod_\rho Z^\rho$ é dado por

$$Z_i^n = \int \left[\prod_{j\rho} dJ_{ij}^\rho \delta\left(Q_i - \frac{1}{C}\sum_j \left(J_{ij}^\rho\right)^2\right)\right] \exp\left[-\lambda E_i\left(\mathbf{J}^\rho\right)\right]. \tag{2.5}$$

Substituindo a energia dada, obtemos

$$\begin{aligned} Z_i^n &= \int \left[\prod_{j\rho} dJ_{ij}^\rho \delta\left(Q_i - \frac{1}{C}\sum_j \left(J_{ij}^\rho\right)^2\right)\right] \\ &\quad \prod_{k\rho} \int dx_{ki}^\rho \delta\left(x_{ki}^\rho - \frac{1}{C^{1/2}}\xi_i^k \sum_j J_{ij}^\rho \xi_j^k\right) \exp\left[-\frac{\lambda}{2}\left(1 - x_{ki}^\rho\right)^2\right]. \end{aligned} \tag{2.6}$$

O primeiro passo é escrever a função $\delta(\varphi - x)$ na sua representação integral, $\delta(\varphi - x) = \frac{1}{2\pi}\int d\widetilde{x} \exp\left[-\mathbf{i}\widetilde{x}(\varphi - x)\right]$ e, em seguida, efetuar a média sobre os padrões ξ_j^k. Como a função de partição é independente do sítio i que consideramos, para maior clareza da notação vamos suprimir a referência explícita a este índice na variável de integração $x_k^\rho \equiv x_{ki}^\rho$. Assim,

$$\begin{aligned} \langle\langle\, Z_i^n \rangle\rangle &= \int \left[\prod_{j\rho} dJ_{ij}^\rho \delta\left(Q - \frac{1}{C}\sum_j \left(J_{ij}^\rho\right)^2\right)\right] \\ &\quad \int \left(\prod_{k\rho} \frac{d\widetilde{x}_k^\rho dx_k^\rho}{2\pi}\right) \exp\left[\mathbf{i}\sum_{k\rho} \widetilde{x}_k^\rho x_k^\rho - \frac{\lambda}{2}\sum_{k\rho}\left(1 - x_k^\rho\right)^2\right] \\ &\quad \left\langle\left\langle \exp\left(-\frac{\mathbf{i}}{C^{1/2}}\sum_{k\rho} \widetilde{x}_k^\rho \xi_i^k \sum_j J_{ij}^\rho \xi_j^k\right)\right\rangle\right\rangle. \end{aligned} \tag{2.7}$$

Fazendo a média em ξ_j^k, obtemos

$$\langle\langle \cdots \rangle\rangle = \prod_{kj}\left[\cos\left(\frac{1}{C^{1/2}}\sum_a \widetilde{x}_k^a J_j^a\right) - \mathbf{i}a\xi_i^k \sin\left(\frac{1}{C^{1/2}}\sum_a \widetilde{x}_k^a J_j^a\right)\right]. \tag{2.8}$$

Expandindo em potências de $1/C^{1/2}$, obtemos a seguinte expressão para a média

$$\langle\langle \cdots \rangle\rangle \simeq \prod_{kj} \exp\ln\left[1 - \frac{1}{2C}\left(\sum_\rho \widetilde{x}_k^\rho J_{ij}^\rho\right)^2 - \frac{\mathbf{i}a}{C^{1/2}}\xi_i^k\left(\sum_\rho \widetilde{x}_k^\rho J_{ij}^\rho\right)\right] \tag{2.9}$$

$$\simeq \prod_{kj} \exp\left[-\frac{(1-a^2)}{2C}\left(\sum_\rho \widetilde{x}_k^\rho J_{ij}^\rho\right)^2 - \frac{\mathbf{i}a}{C^{1/2}}\xi_i^k\left(\sum_\rho \widetilde{x}_k^\rho J_{ij}^\rho\right)\right]. \tag{2.10}$$

Escrevendo o quadrado da somatória como

$$\left(\sum_\rho x^\rho\right)^2 = \sum_\rho \left(x^\rho\right)^2 + \sum_{\rho\neq\sigma} x^\rho x^\sigma = \sum_\rho \left(x^\rho\right)^2 + 2\sum_{\rho<\sigma} x^\rho x^\sigma, \tag{2.11}$$

e introduzindo as variáveis auxiliares $q_{\rho\sigma} = \frac{1}{C}\sum_j J^\rho_{ij} J^\sigma_{ij}$ e $M_\rho = \frac{1}{C^{1/2}}\sum_j J^\rho_{ij}$, a função de partição fica escrita como

$$\begin{aligned}
\langle\langle\, Z_i^n \rangle\rangle \;=\; & \int \left[\prod_{j\rho} dJ^\rho_{ij}\delta\left(Q - \frac{1}{C}\sum_j \left(J^\rho_{ij}\right)^2\right)\right] \\
& \int \left[\prod_{\rho<\sigma} dq_{\rho\sigma}\delta\left(q_{\rho\sigma} - \frac{1}{C}\sum_j J^\rho_{ij} J^\sigma_{ij}\right)\right] \\
& \int \left[\prod_{\rho} dM_\rho\delta\left(M_\rho - \frac{1}{C^{1/2}}\sum_j J^\rho_{ij}\right)\right] \\
& \int \left(\prod_{k\rho} \frac{d\widetilde{x}^\rho_k dx^\rho_k}{2\pi}\right) \exp\left[\mathbf{i}\sum_{k\rho} \widetilde{x}^\rho_k x^\rho_k - \frac{\lambda}{2}\sum_{k\rho}(1 - x^\rho_k)^2\right. \\
& \qquad -\frac{(1-a^2)}{2}\sum_{k\rho} Q\left(\widetilde{x}^\rho_k\right)^2 - \left(1 - a^2\right)\sum_k \left(\sum_{\rho<\sigma} q_{\rho\sigma}\widetilde{x}^\rho_k \widetilde{x}^\sigma_k\right) \\
& \qquad \left. -\mathbf{i}a\sum_k \xi^k_i \sum_\rho \widetilde{x}^\rho_k M_\rho\right]. \qquad (2.12)
\end{aligned}$$

Usando novamente a representação integral da delta nas três primeiras integrais e fazendo as mudanças de variáveis $\widetilde{Q}_\rho = \frac{\widehat{Q}_\rho}{\mathbf{i}/C}$, $\widetilde{q}_{\rho\sigma} = \frac{\widehat{q}_{\rho\sigma}}{\mathbf{i}/C}$ e $\widetilde{M}_\rho = \frac{\widehat{M}_\rho}{\mathbf{i}/C^{1/2}}$, a função de partição pode ser posta na forma

$$\begin{aligned}
\langle\langle\, Z^n \rangle\rangle \;\simeq & \int \left(\prod_\rho \frac{d\widetilde{Q}_\rho}{2\pi\mathbf{i}/C}\right) \int \left(\prod_{\rho<\sigma} \frac{dq_{\rho\sigma} d\widetilde{q}_{\rho\sigma}}{2\pi\mathbf{i}/C}\right) \int \left(\prod_\rho \frac{dM_\rho d\widetilde{M}_\rho}{2\pi\mathbf{i}/C^{1/2}}\right) \\
& \exp\left[C\sum_\rho \widetilde{Q}_\rho Q + C\sum_{\rho<\sigma} \widetilde{q}_{\rho\sigma} q_{\rho\sigma}\right. \\
& \qquad \left. + C^{1/2}\sum_\rho \widetilde{M}_\rho M_\rho + CG_0 + \alpha C G_1\right], \qquad (2.13)
\end{aligned}$$

com

$$G_0 = \ln\left\{\int \left(\prod_\rho dJ^\rho_i\right) \exp\left[-\sum_\rho \widetilde{Q}_\rho \left(J^\rho_i\right)^2 - \sum_{\rho<\sigma} \widetilde{q}_{\rho\sigma} J^\rho_i J^\sigma_i - \sum_\rho \widetilde{M}_\rho J^\rho_i\right]\right\} \qquad (2.14)$$

e

$$\begin{aligned}
G_1 = \Bigg\langle \ln\Bigg\{ & \int \left(\prod_\rho \frac{d\widetilde{x}^\rho dx^\rho}{2\pi}\right) \exp\left[\mathbf{i}\sum_\rho \widetilde{x}^\rho x^\rho - \frac{\lambda}{2}\sum_\rho (1 - x^\rho)^2\right. \\
& \left. -\frac{(1-a^2)}{2}\sum_\rho Q\left(\widetilde{x}^\rho\right)^2 - \left(1 - a^2\right)\left(\sum_{\rho<\sigma} q_{\rho\sigma}\widetilde{x}^\rho \widetilde{x}^\sigma\right) - \mathbf{i}a\xi_i \sum_\rho \widetilde{x}^\rho M_\rho\right]\Bigg\}\Bigg\rangle_{\xi_i} \qquad (2.15)
\end{aligned}$$

Para obter a equação acima, foi utilizada a propriedade de automediância $\frac{1}{P}\sum_{k=1}^{P} g\left(\xi_i^k\right) = \langle g\left(\xi_i\right)\rangle_{\xi_i}$ para i fixo e uma função g arbitrária. A média sobre ξ_i, simbolizada por $\langle\ldots\rangle_{\xi_i}$, é efetuada com a distribuição (2.1).

Neste ponto é importante observar que a forma específica da regra de aprendizado (ou energia de treinamento) altera apenas o termo G_1. Os termos restantes são idênticos para qualquer modelo definido pela minimização de uma função das estabilidades Δ_i^l.

No limite $C \longrightarrow \infty$, as integrais em (2.13) podem ser efetuadas pelo método do ponto de sela. A expressão para a energia livre por sítio pode então ser escrita como

$$\begin{aligned} -\lambda f &= \lim_{n\to 0}\frac{1}{nC}\,\mathrm{extr}\left[C\sum_{\rho}\widetilde{Q}_\rho Q + C\sum_{\rho<\sigma}\widetilde{q}_{\rho\sigma}q_{\rho\sigma}\right. \\ &\qquad \left. +C^{1/2}\sum_{\rho}\widetilde{M}_\rho M_\rho + CG_0 + \alpha CG_1\right], \end{aligned} \tag{2.16}$$

onde o extr é tomado em relação aos parâmetros de ponto de sela $\widetilde{Q}_\rho$, $\widetilde{q}_{\rho\sigma}$, $q_{\rho\sigma}$, $\widetilde{M}_\rho$ e M_ρ. O próximo passo é assumir simetria de réplicas nos parâmetros de ponto de sela, $\widetilde{Q}_\rho = \widetilde{Q}$, $\widetilde{q}_{\rho\sigma} = \widetilde{q}$, $q_{\rho\sigma} = q$, $\widetilde{M}_\rho = \widetilde{M}$ e $M_\rho = M$, o que facilita enormemente o cálculo analítico G_1 e de G_0.

Cálculo de G_1

Utilizando a identidade (2.11) e a transformação gaussiana

$$\int Dz \exp\left(\pm bz\right) = \exp\left(b^2/2\right), \tag{2.17}$$

onde $Dz = \sqrt{\frac{1}{2\pi}}\exp\left(-z^2/2\right)$, podemos desacoplar as diferentes réplicas no termo $q\sum_{\rho<\sigma}\widetilde{x}^\rho\widetilde{x}^\sigma$, de modo que a equação (2.15) fica escrita como

$$\begin{aligned} G_1 &= \left\langle \ln\left\{\int Dz \int\left(\prod_\rho \frac{d\widetilde{x}^\rho dx^\rho}{2\pi}\right)\exp\left[\mathbf{i}\sum_\rho \widetilde{x}^\rho x^\rho\right.\right. \\ &\qquad -\frac{\lambda}{2}\sum_\rho\left(1-x^\rho\right)^2 - \frac{\left(1-a^2\right)}{2}\left(Q-q\right)\sum_\rho\left(\widetilde{x}^\rho\right)^2 \\ &\qquad \left.\left. -\mathbf{i}a\xi_i M\sum_\rho \widetilde{x}^\rho - \mathbf{i}\sqrt{\left(1-a^2\right)q}\sum_\rho x^\rho z\right]\right\}\right\rangle_{\xi_i}. \end{aligned} \tag{2.18}$$

O argumento da função logaritmo pode ser reescrito como

$$\int Dz \prod_{\rho} \left\{ \int \frac{d\tilde{x}dx}{2\pi} \exp \left[\mathbf{i}\tilde{x}x - \frac{\lambda}{2}(1-x)^2 - \frac{(1-a^2)}{2}(Q-q)\tilde{x}^2 \right.\right.$$
$$\left.\left. -\mathbf{i}a\xi_i M\tilde{x} - \mathbf{i}\sqrt{(1-a^2)\,q}\tilde{x}z \right] \right\} =$$
$$= \int Dz \{...\}^n$$
$$= \int Dz \exp n \ln \{...\}$$
$$\simeq \int Dz \left[1 + n \ln \{...\}\right]$$
$$\simeq 1 + n \int Dz \ln \{...\}, \tag{2.19}$$

onde tomamos o limite $n \longrightarrow 0$. Daí,

$$\frac{G_1}{n} \simeq \left\langle \int Dz \ln \left\{ \int \frac{d\tilde{x}dx}{2\pi} \exp \left[\mathbf{i}\tilde{x}x - \frac{\lambda}{2}(1-x)^2 \right.\right.\right.$$
$$\left.\left.\left. -\frac{(1-a^2)}{2}(Q-q)\tilde{x}^2 - \mathbf{i}a\xi_i M\tilde{x} - \mathbf{i}\sqrt{(1-a^2)\,q}\tilde{x}z \right] \right\} \right\rangle_{\xi_i}. \tag{2.20}$$

Agora, as integrais sobre $\tilde{x}$ e x são triviais, levando a

$$\frac{G_1}{n} \simeq -\frac{1}{2}\ln\left[1 + \lambda\left(1-a^2\right)(Q-q)\right]$$
$$- \left\langle \frac{\lambda}{2} \frac{(1-a\xi_i M)^2 + (1-a^2)\,q}{1+\lambda(1-a^2)(Q-q)} \right\rangle_{\xi_i}. \tag{2.21}$$

Finalmente, efetuando a média sobre ξ_i, utilizando a distribuição (2.1), obtemos

$$\frac{G_1}{n} \simeq -\frac{1}{2}\ln\left[1 + \lambda\left(1-a^2\right)(Q-q)\right]$$
$$-\frac{\lambda}{2} \frac{(1+a^2M^2 - 2a^2M) + (1-a^2)\,q}{1+\lambda(1-a^2)(Q-q)}. \tag{2.22}$$

Cálculo de G_0

Para calcular G_0, utilizamos novamente a identidade (2.11) e a transformação gaussiana. Colocando o produtório sobre as réplicas em evidência e integrando em J_i, obtemos, após manipular a potência em n,

$$G_0 = \ln \left\{ 1 + \frac{n}{2}\ln\pi - \frac{n}{2}\ln\left(\tilde{Q} - \frac{\tilde{q}}{2}\right) + n\frac{\widetilde{M}^2}{4\left(\tilde{Q} - \frac{\tilde{q}}{2}\right)} \right.$$

$$-n\int Dz\frac{\tilde{q}z^2}{4\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)}-n\int Dz\frac{\mathbf{i}\sqrt{\tilde{q}}\widetilde{M}z}{2\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)}\Bigg\}=$$

$$=\frac{n}{2}\ln\pi-\frac{n}{2}\ln\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)+n\frac{\widetilde{M}^2-\tilde{q}}{4\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)}\ . \tag{2.23}$$

Para finalizar o cálculo da energia livre por sítio, observamos que, devido à simetria de réplicas, o argumento da primeira exponencial da equação (2.13) reduz-se à equação

$$nC\tilde{Q}Q+n\left(n-1\right)C\frac{\tilde{q}q}{2}+nC^{1/2}\widetilde{M}M, \tag{2.24}$$

o que leva à expressão final para a energia livre por sítio

$$\begin{aligned}-\lambda f \;=\; \mathrm{extr}\Bigg\{&\tilde{Q}Q-\frac{\tilde{q}q}{2}+C^{-1/2}\widetilde{M}M+\frac{1}{2}\ln\pi-\frac{1}{2}\ln\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)\\&+\frac{\widetilde{M}^2-\tilde{q}}{4\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)}-\frac{\alpha}{2}\ln\left[1+\lambda\left(1-a^2\right)\left(Q-q\right)\right]\\&-\frac{\lambda\alpha}{2}\frac{\left(1+a^2M^2-2a^2M\right)+\left(1-a^2\right)q}{1+\lambda\left(1-a^2\right)\left(Q-q\right)}\Bigg\}.\end{aligned} \tag{2.25}$$

Cálculo dos parâmetros de ponto de sela

Resta agora determinar os valores dos parâmetros de ponto de sela da pseudo-inversa, $\tilde{Q}$, $\tilde{q}$, q, $\widetilde{M}$ e M, que levam ao extremo da energia livre. Das equações $\partial f/\partial M=0$ e $\partial f/\partial\widetilde{M}=0$, obtemos

$$\widetilde{M}=-2C^{-1/2}\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)M \tag{2.26}$$

e

$$-2C^{-1}\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)M-\frac{\alpha\lambda a^2\left[M-1\right]}{\left[\lambda\left(1-a^2\right)\left(Q-q\right)+1\right]}=0, \tag{2.27}$$

cuja solução no limite $C\longrightarrow\infty$ é $\widetilde{M}=0$ e $M=1$.

As equações $\partial f/\partial\tilde{Q}=0$ e $\partial f/\partial\tilde{q}=0$ permitem escrever $\tilde{Q}$ e $\tilde{q}$ somente em termos de Q e q. A expressão final para a energia livre é

$$\begin{aligned}-\lambda f=\Bigg\{&\frac{1}{2}+\frac{1}{2}\ln\pi-\frac{1}{2}\ln\frac{1}{2}+\frac{1}{2}\ln\left(Q-q\right)+\frac{q}{2\left(Q-q\right)}\\&-\frac{\alpha}{2}\ln\left[\lambda\left(1-a^2\right)\left(Q-q\right)+1\right]-\frac{\alpha}{2}\frac{\lambda\left(1-a^2\right)\left(1+q\right)}{\left[\lambda\left(1-a^2\right)\left(Q-q\right)+1\right]}\Bigg\}.\end{aligned} \tag{2.28}$$

A equação $\partial f/\partial q = 0$ permite-nos obter uma equação geral que relaciona Q e q,

$$\frac{q}{2(Q-q)^2} - \frac{\alpha}{2}\frac{\widehat{\lambda}^2(1+q)}{\left[1+\widehat{\lambda}(Q-q)\right]^2} = 0, \tag{2.29}$$

onde reescalamos o ruído de treinamento $\widehat{\lambda} = \lambda(1-a^2)$.

Devemos tomar o limite $\widehat{\lambda} \to \infty$, pois estamos interessados em caracterizar apenas as configurações $\{J_{ij}\}$ que minimizam a energia de treinamento (1.14). Lembramos que neste limite a energia livre por sítio é igual à energia de treinamento por sítio, $f = \varepsilon_T$. O limite $\lambda \to \infty$ na expressão (2.28) pode ser tomado de duas maneiras. A primeira é quando $Q > q$ e, neste caso, $\varepsilon_T = 0$. Quando uma rede treinada com determinado número de padrões apresenta energia de treinamento nula, isto é, os padrões são os mínimos globais da energia, dizemos que a rede está abaixo da capacidade de armazenamento α_c. Tomando o limite $\widehat{\lambda} \to \infty$ na equação (2.29) obtemos

$$q = \frac{\alpha}{1-\alpha}. \tag{2.30}$$

Assim, enquanto estivermos no regime de operação $\alpha < \alpha_c$, qualquer valor de Q, tal que $Q > q$, será adequado, e escolhemos o que dá a menor norma. A menor norma será dada fazendo Q tomar o menor valor possível, isto é, fazendo $Q = q = \alpha/(1-\alpha)$.

A segunda maneira de tomar o limite $\widehat{\lambda} \to \infty$ na expressão (2.28) é quando $q \to Q$ tal que $\widehat{\lambda}(Q-q) = \widehat{x}$ permanece finito, o que leva a uma energia de treinamento não nula,

$$\varepsilon_T = -\frac{(1-a^2)Q}{2\widehat{x}} + \frac{\alpha(1-a^2)(1+Q)}{2(\widehat{x}+1)}. \tag{2.31}$$

Agora, queremos escolher Q de maneira a minimizar ε_T. A equação $\partial\varepsilon_T/\partial Q = 0$ leva a

$$\widehat{x} = \frac{1}{\alpha-1}. \tag{2.32}$$

Por outro lado, $\widehat{x}$ também deve ser um ponto de extremo de ε_T e a equação $\partial\varepsilon_T/\partial\widehat{x} = 0$ leva a

$$\frac{(1-a^2)Q}{2\widehat{x}^2} = \frac{\alpha(1-a^2)(1+Q)}{2(\widehat{x}+1)^2}. \tag{2.33}$$

As duas equações anteriores levam à solução

$$Q = \widehat{x} = \frac{1}{\alpha - 1}. \tag{2.34}$$

Substituindo esta solução na equação (2.31), obtemos

$$\varepsilon_T = \frac{(1 - a^2)(\alpha - 1)}{2}. \tag{2.35}$$

Neste limite, a energia de treinamento não é nula, isto é, os padrões não são pontos fixos e não são mais recuperados perfeitamente, de onde vemos que o valor da capacidade de armazenamento é $\alpha_c = 1$, pois abaixo desse valor $\varepsilon_T = 0$.

A conclusão a que chegamos pelos resultados anteriores é que o único efeito do viés é reescalar o parâmetro de ruído do processo de aprendizado $\lambda \longrightarrow \widehat{\lambda} = \lambda(1 - a^2)$, de forma que $\alpha_c = 1$, independente do valor de a.

Para finalizar, devemos discutir a validade da prescrição de simetria de réplicas empregada nos cálculos acima. Em geral, a validação desse procedimento é feita através da análise da estabilidade local dos parâmetros de ponto de sela simétricos com relação às réplicas [38] [39]. Esta análise foi realizada para o modelo da pseudo-inversa sem viés por Fontanari [23], onde concluiu-se que os parâmetros simétricos em relação às réplicas são localmente estáveis para quaisquer valores de α e λ. Como o efeito do viés é apenas reescalar λ, a mesma conclusão permanece válida neste caso.

2.1.2 Cálculo da distribuição de probabilidade das estabilidades

A probabilidade de que a estabilidade $\Delta_i^k \equiv \frac{1}{\sqrt{C}}\xi_i^k \sum_{j \neq i} J_{ij}\xi_j^k$ assuma um valor entre γ e $\gamma + d\gamma$ independe dos índices i e k, devido à independência estatística de ξ_i^k, sendo dada por $P(\gamma)\,d\gamma$, onde $P(\gamma) = \left\langle\left\langle\left\langle \delta\left(\gamma - \Delta_i^k\right)\right\rangle_J\right\rangle\right\rangle$, equação (1.19). Para calcular a distribuição de probabilidade das estabilidades $P(\gamma)$, necessária para obter a equação de evolução da correlação de recuperação $m(t)$ no regime de extrema diluição, acrescentamos à energia da pseudo-inversa (1.14) o termo linear em h

$$E_i(\mathbf{J}, h, \gamma) = \frac{1}{2}\sum_k \left(1 - \Delta_i^k\right)^2 + \frac{h}{P}\sum_k \delta\left(\gamma - \Delta_i^k\right), \tag{2.36}$$

de modo que

$$P(\gamma) = -\lim_{\lambda\to\infty} \frac{1}{\lambda} \frac{\partial \langle\langle \ln Z_i \rangle\rangle}{\partial h} |_{h=0}, \tag{2.37}$$

onde Z_i é a função de partição

$$Z_i(h,\gamma) = \int \left[\prod_j dJ_{ij} \delta \left(Q_i - \frac{1}{C} \sum_j J_{ij}^2 \right) \right] \exp\left[-\lambda E_i(\mathbf{J}, h, \gamma) \right]. \tag{2.38}$$

Aqui, h é uma variável auxiliar, cujo significado físico é irrelevante para nossa análise. O primeiro termo da energia garante a solução de (1.13) e o segundo termo é utilizado para obtenção da distribuição de estabilidades. Devido à equação (2.37), o procedimento para cálculo de $P(\gamma)$ é análogo ao empregado no cálculo da energia livre sendo apresentado a seguir para a pseudo-inversa com padrões com viés.

Conforme mencionado na seção anterior, a alteração da energia de treinamento afeta apenas a expressão para G_1, que neste caso é dada por

$$\begin{aligned} G_1 = \Bigg\langle \ln \Bigg\{ \int \left(\prod_\rho \frac{d\tilde{x}^\rho dx^\rho}{2\pi} \right) \exp \Bigg[\mathbf{i} \sum_\rho \tilde{x}^\rho x^\rho - \frac{\lambda}{2} \sum_\rho (1 - x^\rho)^2 \\ -\lambda h \delta(\gamma - x^\rho) - \frac{(1-a^2)}{2} \sum_\rho Q_\rho (\tilde{x}^\rho)^2 \\ - \left(1 - a^2\right) \left(\sum_{\rho<\sigma} q_{\rho\sigma} \tilde{x}^\rho \tilde{x}^\sigma \right) - \mathbf{i} a \xi_i^k \sum_\rho \tilde{x}^\rho M_\rho \Bigg] \Bigg\} \Bigg\rangle_{\xi_i}. \end{aligned} \tag{2.39}$$

Seguindo o procedimento de cálculo detalhado na seção anterior, podemos efetuar as integrais em $\tilde{x}^\rho$ e x^ρ, e colocar G_1 na seguinte forma compacta

$$\frac{G_1}{n} \simeq \langle \int Dz \ln f(z,h) \rangle_{\xi_i}, \tag{2.40}$$

onde

$$\begin{aligned} f(z,h) = & \sqrt{\frac{1}{1 + \lambda(1-a^2)(Q-q)}} \\ & \exp \left[\frac{\lambda^2 (1-a^2)(Q-q) \left[(h-1)^2 - 1 \right] - \lambda \left[1 + 2(h-1)A + A^2 \right]}{2\left[1 + \lambda(1-a^2)(Q-q) \right]} \right] \end{aligned} \tag{2.41}$$

e reescalamos a temperatura com $\widehat{\lambda} = \lambda\left(1-a^2\right)$ e fizemos

$$A = a\xi_i M + \sqrt{\left(1-a^2\right)q}z. \tag{2.42}$$

Para obtermos a distribuição de probabilidade das estabilidades utilizamos (2.37), que leva à seguinte expressão

$$\begin{aligned} P(\gamma) &= -\lim_{\lambda\to\infty}\frac{1}{\lambda n}\frac{\partial G_1}{\partial h}\,|_{h=0} \\ &= -\lim_{\lambda\to\infty}\frac{1}{\lambda}\left\langle\int Dz\frac{f'(z,0)}{f(z,0)}\right\rangle_{\xi_i}, \end{aligned} \tag{2.43}$$

onde $f'(z,0) = \frac{\partial f(z,h)}{\partial h}\,|_{h=0}$. A função do numerador $f'(z,0)$ é facilmente integrada, fornecendo

$$\begin{aligned} f'(z,0) &= \left[2\pi\left(1-a^2\right)(Q-q)\right]^{-1/2} \\ &\quad \exp\left[-\frac{\lambda}{2}(1-\gamma)^2 - \frac{[\gamma-A]^2}{2\left(1-a^2\right)(Q-q)}\right], \end{aligned} \tag{2.44}$$

e para o denominador obtemos

$$\begin{aligned} f(z,0) &= \left[1+\widehat{\lambda}(Q-q)\right]^{-1/2}\exp\left[-\frac{\lambda}{2} - \frac{A^2}{2\left(1-a^2\right)(Q-q)}\right. \\ &\quad \left. + \frac{\left[A+\widehat{\lambda}(Q-q)\right]^2}{2\left(1-a^2\right)(Q-q)\left[1+\widehat{\lambda}(Q-q)\right]}\right]. \end{aligned} \tag{2.45}$$

O cálculo analítico completo para (2.43) fornece

$$\begin{aligned} P(\gamma) &= \left\langle\frac{\left[1+\widehat{\lambda}(Q-q)\right]}{\left[2\pi\left(1-a^2\right)\right]^{1/2}\left[Q+\widehat{\lambda}(Q-q)^2\right]^{1/2}}\right. \\ &\quad \left.\exp\left\{-\frac{\left[\left[1+\widehat{\lambda}(Q-q)\right]\gamma-\widehat{\lambda}(Q-q)-a\xi_i M\right]^2}{2\left[Q+\widehat{\lambda}(Q-q)^2\right]\left(1-a^2\right)}\right\}\right\rangle_{\xi_i} \qquad (2.46) \\ &= \left\langle p_{\xi_i}(\gamma)\right\rangle_{\xi_i}. \qquad (2.47) \end{aligned}$$

Vemos que para dado ξ_i, a expressão para $p_{\xi_i}(\gamma)$ é uma distribuição de probabilidade gaussiana que pode ser escrita na forma geral

$$p_{\xi_i}(\gamma) = \sqrt{\frac{1}{2\pi\delta_{\xi_i}^2}}\exp\left[-\frac{1}{2}\left(\frac{\gamma-\overline{\gamma}_{\xi_i}}{\delta_{\xi_i}}\right)^2\right], \tag{2.48}$$

com média

$$\overline{\gamma}_{\xi_i} = \frac{\widehat{\lambda}(Q-q) + a\xi_i M}{1+\widehat{\lambda}(Q-q)} , \tag{2.49}$$

e variância

$$\delta^2_{\xi_i} = \frac{(1-a^2)\left[Q+\widehat{\lambda}(Q-q)^2\right]}{\left[1+\widehat{\lambda}(Q-q)\right]^2} . \tag{2.50}$$

Finalmente, mediando sobre ξ_i, obtemos a seguinte expressão para a distribuição de probabilidade das estabilidades

$$\begin{aligned} P(\gamma) &= \sqrt{\frac{1}{2\pi\delta^2_{\xi_i}}}\left\{\frac{1+a}{2}\exp\left[-\frac{1}{2}\left(\frac{\gamma-\overline{\gamma}_+}{\delta_{\xi_i}}\right)^2\right]\right. \\ &\qquad \left.\frac{1-a}{2}\exp\left[-\frac{1}{2}\left(\frac{\gamma-\overline{\gamma}_-}{\delta_{\xi_i}}\right)^2\right]\right\} , \end{aligned} \tag{2.51}$$

onde $\overline{\gamma}_\pm = \left[\widehat{\lambda}(Q-q) \pm aM\right] / \left[1+\widehat{\lambda}(Q-q)\right]$.

No regime de $\alpha \leq \alpha_c = 1$, fazendo $\widehat{\lambda} \to \infty$, obtemos $P(\gamma) = \delta(\gamma - 1)$, conforme esperado, pois estamos no regime de recuperação sem erros. No limite $\alpha > \alpha_c = 1$, tomando $\widehat{\lambda}(Q-q) = \widehat{x}$, conseguimos a seguinte expressão para $\overline{\gamma}_\pm$

$$\overline{\gamma}_\pm = \frac{1 \pm a(\alpha-1)}{\alpha} , \tag{2.52}$$

e para a variância

$$\delta^2_{\xi_i} = \frac{(1-a^2)(\alpha-1)}{\alpha^2} , \tag{2.53}$$

onde substituimos $M = 1$ e $Q = \widehat{x} = 1/(\alpha-1)$.

Cálculo da fração de sítios instáveis

Uma grandeza útil para analisar a estabilidade dos padrões armazenados numa rede neural é a fração de sítios instáveis, isto é, a fração de sítios com $\Delta^k_i < 0$, dada por

$$\epsilon = \int_{-\infty}^{0} d\gamma\, P(\gamma) . \tag{2.54}$$

No caso da pseudo-inversa com viés, conhecendo-se sua distribuição de estabilidades, podemos obter a função erro relacionada à recuperação dos padrões armazenados, isto é, a fração de sítios instáveis dessa distribuição.

Para $\alpha \leq \alpha_c$ a função ϵ é dada por $\epsilon = 0$, e para $\alpha > \alpha_c$, expressão para a fração de sítios instáveis é

$$\epsilon = \frac{1}{2}\left\{\frac{1+a}{2}\,\mathrm{erfc}\left[\sqrt{\frac{1}{2}}\frac{\widehat{x}+a}{\sqrt{\widehat{x}\,(1-a^2)}}\right] + \frac{1-a}{2}\,\mathrm{erfc}\left[\sqrt{\frac{1}{2}}\frac{\widehat{x}-a}{\sqrt{\widehat{x}\,(1-a^2)}}\right]\right\} \tag{2.55}$$

Tomando $a = 0$ com $\widehat{x} = \frac{1}{\alpha-1}$, obtemos o caso da pseudo-inversa com padrões independentes

$$\epsilon = \frac{1}{2}\,\mathrm{erfc}\left[\sqrt{\frac{1}{2\,(\alpha-1)}}\right]. \tag{2.56}$$

Já no limite $\alpha \longrightarrow \infty$ ($\widehat{x} \longrightarrow 0$) e $a \neq 0$, obtemos

$$\epsilon = \frac{1-a}{2}, \tag{2.57}$$

indicando que quanto maior o viés (maior a semelhança entre os padrões), menor é a fração de sítios instáveis desses padrões.

2.2 Pesos ótimos

O modelo com máxima capacidade de armazenagem de padrões aleatórios, denominado modelo dos pesos ótimos, foi proposto por Gardner [17], e tem os pesos satisfazendo às desigualdades

$$\Delta_i^k \equiv \frac{1}{\sqrt{C}}\xi_i^k \sum_{j\neq i} J_{ij}\xi_j^k \geq \kappa \tag{2.58}$$

para todo k e i. O parâmetro de margem $\kappa \geq 0$ é introduzido para garantir uma bacia de atração finita para os padrões estáveis $\boldsymbol{\xi}^l$. A capacidade de armazenagem diminui com κ e, em particular, $\alpha_c = 2$ para $\kappa = 0$. Gardner [24] estudou a armazenagem de padrões com e sem viés a $T = 0$ e Gardner e Derrida [18] generalizaram o caso de padrões sem viés para $T \neq 0$. A análise para padrões com viés a $T \neq 0$ foi realizada por Theumann e Erichsen [40].

Os cálculos para se obter a distribuição de probabilidade das estabilidades do modelo dos pesos ótimos *sem* viés são análogos ao da seção anterior para o modelo

da pseudo-inversa, diferindo apenas na expressão de G_1, pois nesse caso a energia de treinamento é dada pela equação (1.17), $E_i = \sum_l \theta\left(\kappa - \Delta_i^l\right)$. A expressão para a distribuição de estabilidades do modelo dos pesos ótimos foi obtida por [24] [29], sendo dada por

$$P(\gamma) = \sqrt{\frac{1}{2\pi(1-q)}}\Theta(\gamma-\kappa)\int_{-\infty}^{\infty} Dy \frac{\exp\left[-\Xi^2(\gamma,y)/2\right]}{H\left[\Xi(\kappa,y)\right]}, \tag{2.59}$$

onde

$$\Xi(x,y) = \frac{x - y\sqrt{q}}{\sqrt{1-q}}. \tag{2.60}$$

Aqui $\Theta(x) = 1$ se $x > 0$ e 0 caso contrário. O parâmetro q mede a correlação entre dois conjuntos distintos de pesos ótimos. Ele é dado pela solução da equação

$$q = \alpha\sqrt{\frac{1-q}{2\pi}}\int_{-\infty}^{\infty} Dy\left(\kappa - y/\sqrt{q}\right)\frac{\exp\left[-\Xi^2(\kappa,y)/2\right]}{H\left[\Xi(\kappa,y)\right]}. \tag{2.61}$$

Estas equações, obtidas com a prescrição de simetria de réplicas, são válidas para $\alpha \leq \alpha_c(\kappa)$. No regime de saturação, $\alpha = \alpha_c(\kappa)$, tem-se $q \to 1$, simplificando bastante a equação (2.59). Para redes neurais booleanas, a escolha da normalização dos pesos é irrelevante, por isso fazemos $Q = 1$ como usual.

O método de cálculo apresentado neste capítulo será utilizado tanto no capítulo 4 para estudar a vizinhança dos padrões armazenados como no capítulo 6 para estudar a categorização na pseudo-inversa. Além disso, as distribuições de estabilidades para a pseudo-inversa e para a rede dos pesos ótimos, obtidas nesse capítulo, serão utilizadas no capítulo 3 para obter os diagramas de fase para essas redes no regime de extrema diluição para o caso de padrões sem viés.

Para concluir, convém mencionar que a solução simétrica com relação às réplicas, equação (2.61), é localmente estável frente à quebra de simetria de réplicas para $\alpha \leq \alpha_c(\kappa)$ [18]. Felizmente, essa é exatamente a região de interesse para as análises apresentadas nos capítulos posteriores, de forma que não precisamos nos preocupar em quebrar a simetria das réplicas, que certamente seria necessária para a investigação do regime $\alpha > \alpha_c(\kappa)$ ([41], [42] e [43]).

Capítulo 3

Diagramas de fase para redes extremamente diluídas

Para estudar os diagramas de fase das redes pseudo-inversa e pesos ótimos, trabalharemos com a versão diluída dessas redes. A vantagem de trabalhar no regime de extrema diluição, isto é, para $C \ll \ln N$, é que a dinâmica pode ser estudada de forma analítica e exata. O ponto crucial, primeiro notado por Derrida, Gardner e Zippelius [14] no contexto do modelo de Hopfield e depois por Gardner [24] em um contexto mais geral, é que, no limite de extrema diluição, a correlação entre os diferentes sítios introduzida durante a evolução da rede pode ser desprezada.

No modelo dos pesos ótimos, Gardner [24] mostrou que a evolução temporal da correlação de recuperação de um estado $S(t)$ que possui correlação não nula apenas com um dos padrões memorizados pode ser estudada analiticamente, obtendo uma solução exata, no caso de redes extremamente diluídas, isto é, $C \ll \ln N$. A idéia básica, desenvolvida independentemente por Keppler e Abbot [29], é escrever o primeiro passo da dinâmica em termos das estabilidades. Isto é feito calculando a distribuição de probabilidade para o primeiro passo da correlação de recuperação $m_1^l \equiv m^l(t=1)$, dada a correlação de recuperação inicial $m_0^l \equiv m^l(t=0)$, através

de

$$P\left(m_1^l|m_0^l\right) = \frac{\mathrm{Tr}_S\left[\delta\left(Nm_1^l, \sum_{i=1}^N \xi_i^l \,\mathrm{sign}\left(\sum_{j=1}^N J_{ij}S_j\right)\right)\delta\left(Nm_0^l, \sum_{j=1}^N S_j\xi_j^l\right)\right]}{\mathrm{Tr}_S\left[\delta\left(Nm_0^l, \sum_{j=1}^N S_j\xi_j^l\right)\right]}, \quad (3.1)$$

onde Tr_S é a soma sobre todos os possíveis estados da rede e $\delta(x,y)$ é a delta de Kronecker. Calculando a distribuição acima, obtém-se para o primeiro passo da dinâmica a seguinte expressão

$$m_1^l = \int_{-\infty}^{\infty} d\Delta^l P\left(\Delta^l\right) \mathrm{erf}\left[\frac{m_0^l\Delta^l}{Q^{1/2}\sqrt{2\left(1-\left(m_0^l\right)^2\right)}}\right], \quad (3.2)$$

onde $Q = Q_i = 1/N\sum_j J_{ij}^2$ é o quadrado da norma dos pesos sinápticos. Aqui, $P\left(\Delta^l\right)$ é a distribuição das estabilidades, como definida em (1.19)

$$P\left(\Delta_i^l\right) = \left\langle\left\langle\left\langle\delta\left(\Delta_i^l - \frac{1}{\sqrt{C}}\xi_i^l\sum_{j\neq i} J_{ij}\xi_j^l\right)\right\rangle_J\right\rangle\right\rangle. \quad (3.3)$$

A notação $\langle\langle\ldots\rangle\rangle$ significa a média sobre os padrões ξ_j^l e $\langle\ldots\rangle_J$ é a média sobre o ensemble de pesos que definem o modelo em questão. Note que $P\left(\Delta_i^l\right) = P(\Delta)$, se supusermos que as variáveis aleatórias ξ_i^l sejam estatisticamente independentes. Daqui em diante, sempre que não houver margem para dúvidas omitiremos o índice dos padrões l da distribuição de probabilidade das estabilidades Δ e da correlação de recuperação m. No limite de extrema diluição a equação (3.2) é válida para qualquer $t \geq 1$ e a equação da dinâmica para a rede diluída é obtida simplesmente substituindo os índices 1 e 0 por $t+1$ e t, respectivamente.

A generalização de (3.2) para o regime de temperatura não nula é simples: após a realização da média térmica, a função $\mathrm{sign}\left(\sum_{j=1}^N J_{ij}S_j\right)$ em (3.1) é substituída pela função $\tanh\left(\beta\sum_{j=1}^N J_{ij}S_j\right)$ [26], o que leva à seguinte equação

$$m_{t+1} = \int_{-\infty}^{\infty} d\Delta P(\Delta)\int_{-\infty}^{\infty} Dy \tanh\left[\frac{1}{T}\left(m_t\Delta + yQ^{1/2}\sqrt{1-m_t^2}\right)\right], \quad (3.4)$$

onde $Dy = dy/\sqrt{2\pi}e^{-y^2/2}$ é a medida gaussiana. As equações para a dinâmica diluídas são válidas para qualquer que seja a prescrição dos pesos sinápticos, bastando conhecer a distribuição de probabilidade das estabilidades do modelo de rede

neural atratora que se está considerando. É importante notar que as equações (3.2) e (3.4) são válidas apenas para padrões sem viés e, mesmo neste caso, apenas para soluções de recuperação, isto é, estados com correlação não nula com somente um dos padrões. De fato, realizamos o cálculo para o caso de padrões com viés, mas a equação dinâmica obtida resultou depender de várias outras grandezas além das estabilidades, de forma que sua análise tornou-se impraticável.

O regime de equilíbrio é descrito pelo parâmetro de ordem m_∞, utilizado para caracterizar as duas possíveis fases do sistema, a saber, a fase de não recuperação do padrão $\boldsymbol{\xi}^l$ ou fase paramagnética ($m_\infty = 0$) e a fase de recuperação ($m_\infty > 0$). A versão diluída do modelo da pseudo-inversa foi estudada por Opper, Kleinz, Köhler e Kinzel [27] no regime de temperatura zero, enquanto Amit, Evans, Horner e Wong [26] trataram do modelo dos pesos ótimos no regime de saturação $\alpha = \alpha_c(\kappa)$. A análise à temperatura zero de Gardner [24] também foi restrita a este regime.

Na seção 3.1, apresentamos o procedimento geral para obter os diagramas de fase para redes extremamente diluídas. Apresentamos em seguida os diagramas de fase para os modelos da pseudo-inversa e dos pesos ótimos, respectivamente, nas seções 3.2 e 3.3.

3.1 Determinação dos diagramas de fase

Uma vez que se conheça a distribuição de estabilidades $P(\Delta)$, a análise da dinâmica de redes neurais diluídas torna-se simples. As diferentes fases no espaço dos parâmetros de controle são determinadas pelos pontos fixos $m_{t+1} = m_t = m^*$ da equação (3.4). Para determinar esses pontos fixos é interessante definir a seguinte função

$$g(m) = m - \int_{-\infty}^{\infty} d\Delta P(\Delta) \int_{-\infty}^{\infty} Dy \tanh\left[\frac{1}{T}\left(m\Delta + yQ^{1/2}\sqrt{1-m^2}\right)\right], \tag{3.5}$$

de modo que os pontos fixos são agora as raízes de

$$g(m) = 0. \tag{3.6}$$

É importante notar que $g(-m) = -g(m)$ e, também, que o ponto fixo paramagnético $m^* = 0$ é sempre uma raiz. Além disso, como $-m^*$ também é raiz, consideraremos na análise que segue somente as raízes não negativas da equação (3.6). A expansão dessa equação em potências de m fornece

$$m^* = 6^{1/2}\sqrt{\frac{g'(m=0)}{g'''(m=0)}} \tag{3.7}$$

para a solução não nula de (3.6). Assim, a equação

$$g'(m=0) = 0 \tag{3.8}$$

determina a linha de transição contínua entre as fases de recuperação ($m^* > 0$) e paramagnética ($m^* = 0$), desde que $g'''(m=0) \neq 0$.

Como usual nesse tipo de análise de campo médio, a determinação do ponto tricrítico (PTC) se faz quando $g'''(m=0)$ e $g'(m=0)$ anulam-se simultaneamente. Além do mais, a equação (3.8) também dá o limite de estabilidade do ponto fixo paramagnético $m^* = 0$: se $g'(m=0) > 0$ trata-se de um ponto fixo atrator e se $g'(m=0) \leq 0$ de um ponto fixo instável.

Para os modelos da pseudo-inversa e dos pesos ótimos, a análise numérica da equação (3.6) mostra que se $m^* = 0$ é estável, ou existe somente uma raiz ou existem duas raízes positivas adicionais, conforme ilustrado na figura 3.1. Se, entretanto, $m^* = 0$ é instável, existe somente uma raiz positiva adicional. Finalmente, para determinar a linha de transição descontínua, é necessário resolver $g(m) = 0$ e $g'(m) = 0$ simultaneamente, uma vez que nessa transição, que ocorre na região onde $m^* = 0$ é estável, as duas raízes positivas coalescem numa raiz dupla antes de desaparecerem.

3.2 Pseudo-inversa

Para obter a equação da dinâmica para a pseudo-inversa no regime em que os padrões são pontos fixos, isto é, para $\alpha < 1$, substituímos a expressão da distribuição das es-

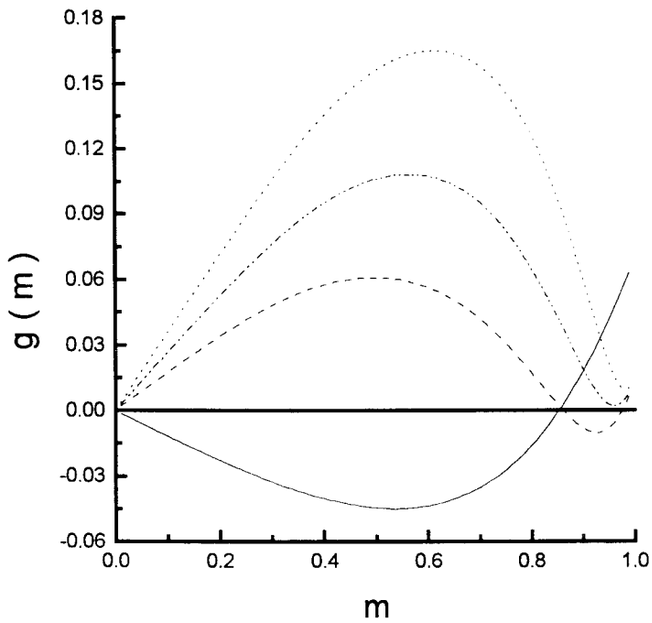

Figura 3.1: Gráfico da função $g(m)$ para diversos valores de α e T, mostrando o número e estabilidade de suas raízes. De baixo para cima (em $m = 0.5$), as curvas são para os seguintes parâmetros (α, T): (0.2, 0.6), (0.45, 0.4), (0.52, 0.4) e (0.6, 0.4). Embora este gráfico seja para o modelo da pseudo-inversa, a análise qualitativa é idêntica para o modelo dos pesos ótimos.

tabilidades (2.51), obtida no capítulo anterior, na equação (3.4), e fazemos a integral em Δ (que é trivial neste caso), o que leva a

$$m_{t+1} = \int_{-\infty}^{\infty} Dy \tanh\left\{\frac{1}{T}\left[m_t + y\sqrt{(1-m_t^2)\,Q}\right]\right\}, \tag{3.9}$$

onde $Q = \alpha/(1-\alpha)$.

O diagrama de fase para $\alpha \leq 1$ é apresentado na figura 3.2. O ponto tricrítico está em $\alpha = 0.249$ e $T = 0.648$. A curva pontilhada intercepta a linha de $T = 0$ em $\alpha = 0.1/(1+\pi/2) \approx 0.389$, em concordância com o resultado de Opper, Kleinz,

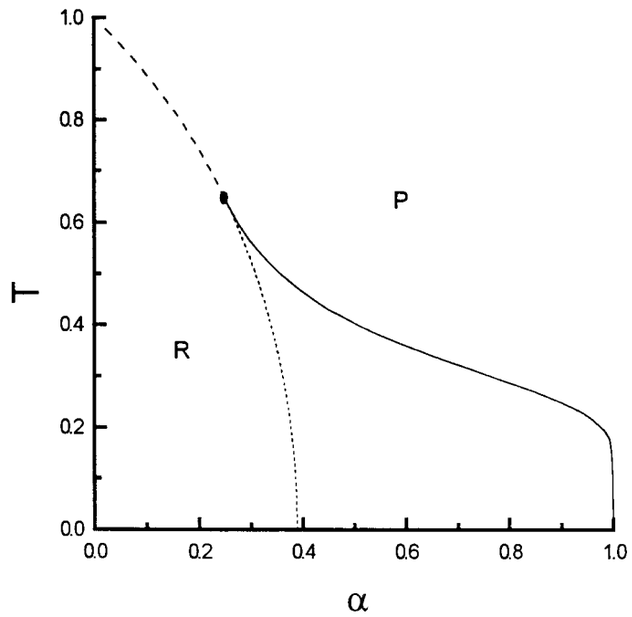

Figura 3.2: Diagrama de fase da rede neural atratora pseudo-inversa, onde se pode ver a fase de recuperação (R) e a fase paramagnética (P). A curva cheia é a transição descontínua , a curva tracejada, a transição contínua e a curva pontilhada, o limite de estabilidade da fase paramagnética. O ponto tricrítico está em $\alpha = 0.249$ e $T = 0.648$.

Köhler e Kinzel [27]. Entre as curvas contínua e pontilhada, ambas as fases coexistem.

A figura 3.3 mostra a dependência da correlação de recuperação m^* com α para alguns valores de temperatura. Para $T = 0.8$ a transição é contínua, o que se percebe pela curva suave que leva m^* a zero. A transição abrupta de m^* para zero pode ser vista para $T = 0.3$, indicando que neste caso a transição é descontínua. Observar que em ambos os casos $m^* \neq 1$ pois $T > 0$. Deve-se notar ainda que o diagrama de fase para a rede não diluída, obtida por Kanter e Sompolinsky [21], apresenta somente a transição descontínua.

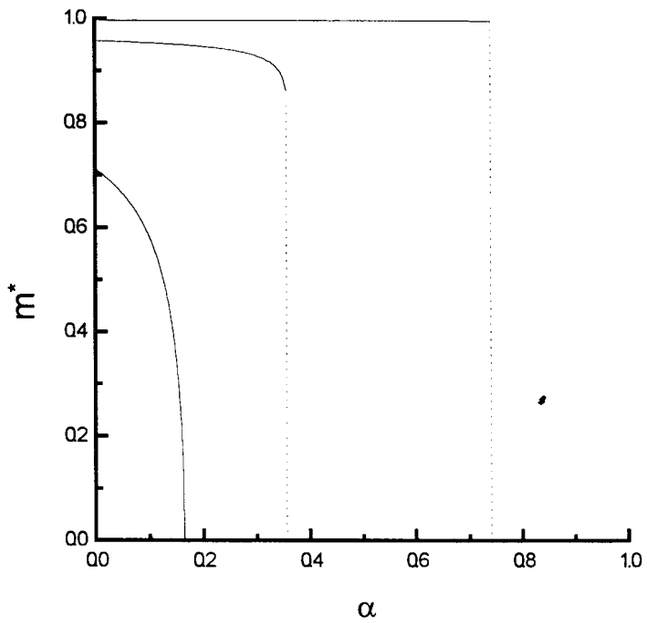

Figura 3.3: Zeros da função $g(m)$ em função de α com T constante para a pseudo-inversa. A primeira curva de baixo para cima é para $T = 0.8$, região em que há uma transição contínua. As duas outras curvas são, de baixo para cima, $T = 0.5$ e $T = 0.3$, região onde ocorre transição descontínua, indicada nesta figura pela linha pontilhada.

Para um dado α, pode-se obter o menor valor da correlação de recuperação inicial tal que haja recuperação. Na figura 3.4, podemos ver claramente a coexistência das fases de recuperação e paramagnética para $T = 0$.

No caso em que os padrões não são armazenados perfeitamente, $\alpha > 1$, a equação (3.4) é escrita da seguinte forma

$$m_{t+1} = \int_{-\infty}^{\infty} Dy \tanh\left\{\frac{1}{T}\left[\overline{\gamma} m_t + y\sqrt{(1 - m_t^2)\, Q + \delta^2 m_t^2}\right]\right\} \tag{3.10}$$

com $Q = 1/(\alpha - 1)$, $\delta^2 = \overline{\gamma}(1 - \overline{\gamma})$ e $\overline{\gamma} = 1/\alpha$. A única raiz desta equação é a

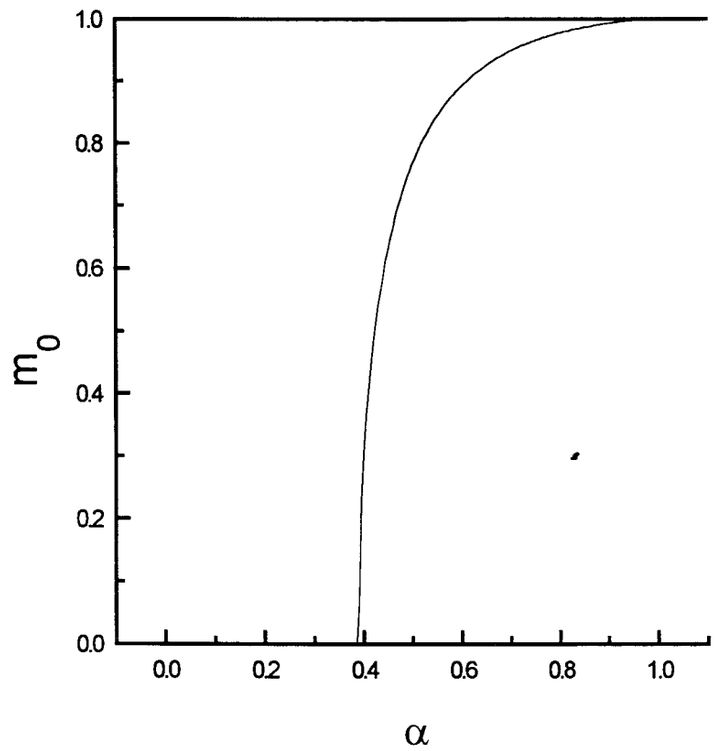

Figura 3.4: Bacia de atração para a pseudo-inversa à $T = 0$. O gráfico mostra o menor valor da correlação de recuperação inicial m_0, tal que seja recuperado o padrão armazenado em função da capacidade de armazenamento α.

paramagnética, $m^* = 0$.

3.3 Pesos Ótimos

Para o modelo dos pesos ótimos, a distribuição de estabilidades foi obtida por Kepler e Abbot [29] e Gardner [24] sendo dada pela equação (2.59). Enquanto os parâmetros de interesse na caracterização dos diagramas de fase da pseudo-inversa eram (α, T), para a rede dos pesos ótimos os parâmetros de interesse são (α, κ, T), onde κ é o parâmetro de margem, introduzido para controlar a bacia de atração dos estados memorizados. Tratemos primeiro do diagrama de fase no plano $\alpha = 0$, mostrado na figura 3.5. Em $T = 0$, o ponto fixo paramagnético torna-se instável para $\kappa > 0.651$. O ponto tricrítico está localizado em $\kappa = 1$ e $T = 0.799$. Para grandes valores de κ, a linha de transição contínua tende à reta $T = \kappa$. No limite de $T = 0$, o ponto fixo de recuperação é $m^* = 1$.

O diagrama de fase no plano $T = 0$, mostrado na figura 3.6, não apresenta a transição contínua entre as fases de recuperação e paramagnética. A curva pontilhada, que delimita a região de estabilidade do ponto fixo $m^* = 0$, intercepta o eixo $\alpha = 0$ em $\kappa = 0.651$. Além disso, ela intercepta a curva de transição descontínua, dada por $\alpha = \alpha_c(\kappa)$, em $\alpha = 0.42$ e $\kappa = 1.2$, em concordância com os resultados de Gardner [24] e Amit, Evans, Horner e Wong [26]. Daqui por diante, nos referiremos à linha $\alpha = \alpha_c(\kappa)$ como linha terminal. Para $\alpha > \alpha_c(\kappa)$, podemos considerar que o conjunto de pesos ótimos é tal que as violações do critério de estabilidade sejam minimizadas, numa maneira similar ao que fizemos para o modelo da pseudo-inversa na região $\alpha > 1$. Entretanto, Amit e outros [26] mostraram que o ponto fixo paramagnético é a única solução de equilíbrio da equação da dinâmica neste caso.

Finalmente, consideremos valores genéricos dos parâmetros de controle α, κ e T. Neste caso, a determinação de $g(m)$, equação (3.5), envolve o cálculo de uma integral tripla que, entretanto, pode ser reduzida a uma integral dupla através de uma mudança apropriada das variáveis de integração, permitindo a resolução analítica da integral sobre Δ. Os diagramas de fase para $\kappa = 0, 0.5, 0.7, 1.0, 1.2, 1.5, 1.7$ e 2.0 são apresentados na figura 3.8. Para $\kappa = 0$, o ponto fixo de recuperação é instável,

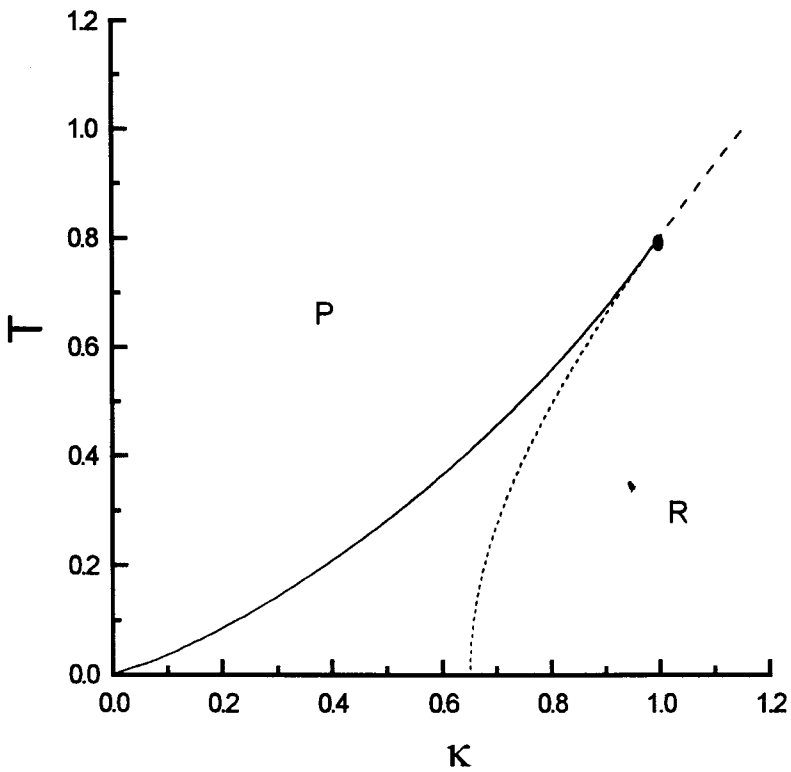

Figura 3.5: Diagrama de fase da rede neural atratora dos pesos ótimos no plano $\alpha = 0$. O ponto tricrítico ocorre em $\kappa = 1$ e $T = 0.799$. Entre as linhas contínua e pontilhada as duas fases coexistem. A convenção é a mesma da figura 3.2.

embora qualquer valor não nulo de κ possa estabilizá-lo. Isto pode ser facilmente visto verificando-se a condição $g'(m = 1) > 0$ para $T = 0$ e $\alpha = 0$, já que se este ponto fixo for instável nestes limites, é muito provável que ele também o seja para $T > 0$ e $\alpha > 0$. Em particular, neste limite, obtemos

$$g'(m) = 1 - \frac{1}{\pi H(\kappa)(1 - m^2)} \exp\left[-\frac{\kappa^2}{2(1 - m^2)}\right], \tag{3.11}$$

onde $H(\kappa) = \frac{1}{2}\operatorname{erfc}\left(\kappa/\sqrt{2}\right) = \frac{1}{\sqrt{2\pi}}\int_\kappa^\infty \exp(-x^2/2)\,dx$. Desta equação vemos que $g'(m^* = 1) \to -\infty$ para $\kappa = 0$, e $g'(m^* = 1) \to 1$ para $\kappa > 0$. À medida que κ aumenta a partir de zero, a região de estabilidade do ponto fixo de recuperação também aumenta. A transição descontínua termina abruptamente na linha terminal

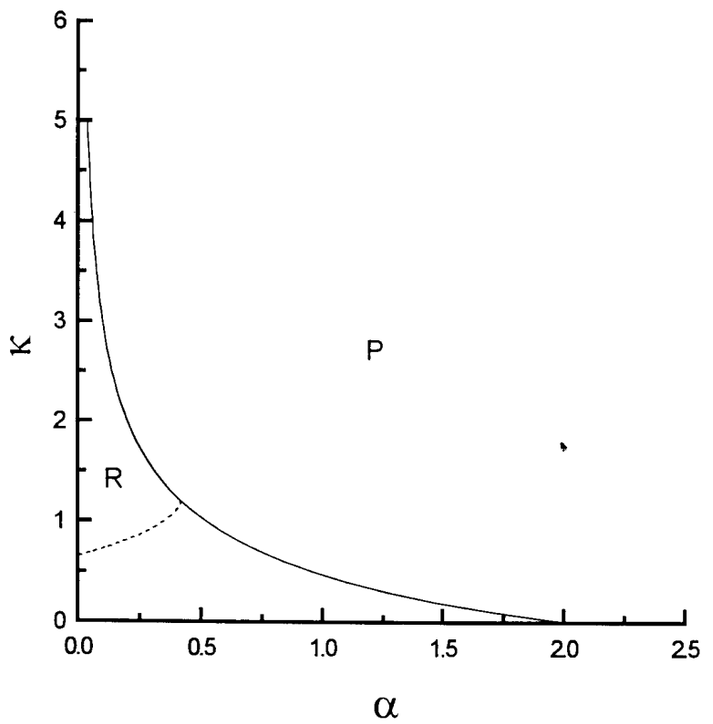

Figura 3.6: Diagrama de fase da rede neural atratora ótima no plano $T = 0$. A convenção é a mesma da figura 3.2.

$\alpha = \alpha_c(\kappa)$. Para $\kappa > 0.651$, aparece uma região próxima à origem $T = 0$ e $\alpha = 0$, onde o ponto fixo paramagnético é instável. Seguindo nossa convenção, a fronteira dessa região é indicada pela curva pontilhada. A primeira intersecção dessa curva com a linha de transição descontínua ocorre em $\alpha = 0$ e $\kappa = 1$; a primeira intersecção com a linha terminal ocorre em $T = 0$ e $\kappa = 1.2$. O ponto tricrítico, gerado em $\kappa = 1$, atinge a linha terminal em $\kappa = 1.7$. Como resultado, a transição descontínua e o PTC desaparecem para $\kappa > 1.7$, após o que o ponto fixo de recuperação se torna o único ponto fixo estável abaixo da curva de transição contínua. Como esperado, aumentando o parâmetro κ aumenta a robustez da rede ao ruído e diminui sua capacidade de armazenamento. A linha tricrítica no espaço tridimensional (α, κ, T) é mostrada na figura 3.7, juntamente com suas projeções nos planos $\alpha = 0$, $\kappa = 0$ e

$T = 0$.

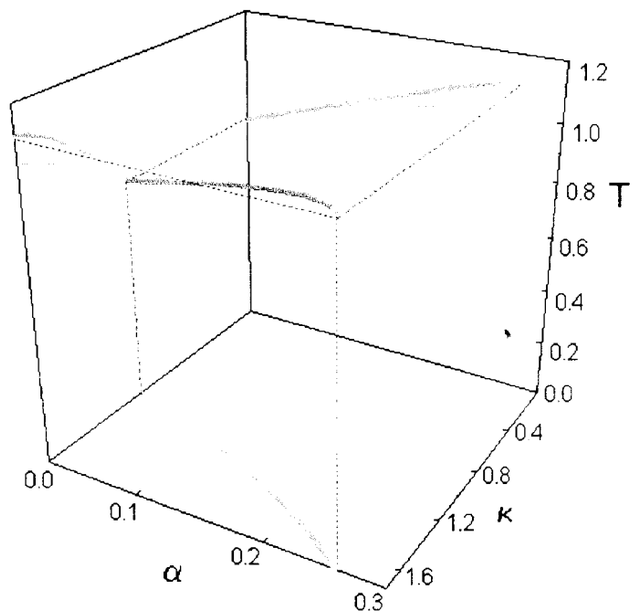

Figura 3.7: Linha tricrítica no espaço de parâmetros (α, κ, T) do modelo dos pesos ótimos. As projeções nos planos $\alpha = 0$, $\kappa = 0$ e $T = 0$ mostram os limites citados no texto.

Comparando o diagrama de fase para a pseudo-inversa, apresentado na figura 3.2, com o diagrama de fase da rede atratora ótima, apresentado na figura 3.8, pode-se concluir que, para a mesma capacidade de armazenamento $\alpha_c = 1$ (atingido com $\kappa \approx 0.470$), a pseudo-inversa tem melhor desempenho. De fato, além de ser mais robusta ao ruído, a rede pseudo-inversa apresenta um regime onde o ponto fixo de recuperação é o único ponto fixo estável, enquanto para a rede atratora ótima esse regime está presente apenas para $\kappa > 0.651$. É importante mencionar que para $\kappa = 0$ e $\alpha \leq 2$ os padrões memorizados, embora sejam pontos fixos, não são atratores. No capítulo 5, verificaremos se essa conclusão também é válida para redes não-diluídas.

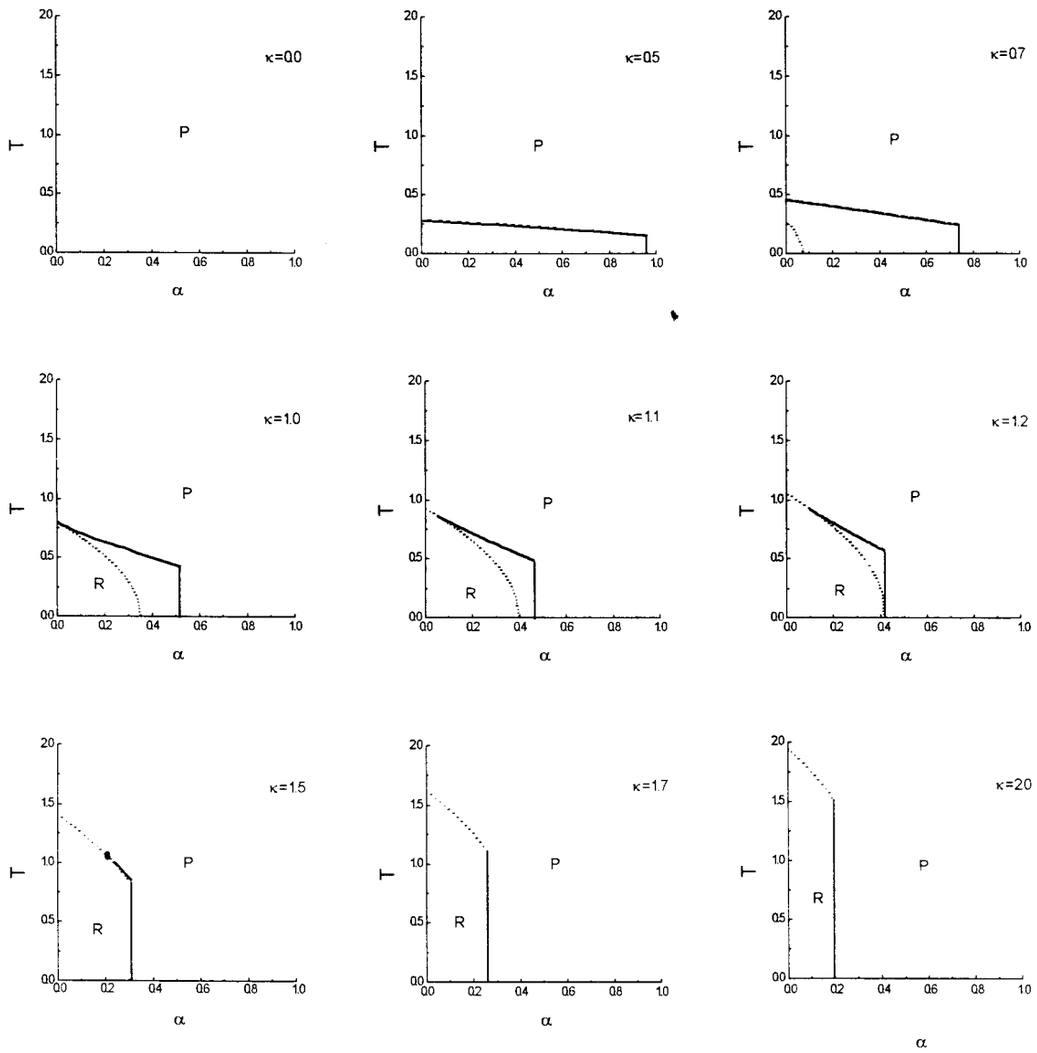

Figura 3.8: Diagrama de fase da rede neural atratora ótima nos planos $\kappa = 0$, 0.5, 0.7, 1.0, 1.2, 1.5, 1.7 e 2.0. A convenção é a mesma da figura 3.2.

Capítulo 4

Vizinhança dos pontos fixos

Neste capítulo, investigaremos a estabilidade dos padrões armazenados em redes neurais atratoras completamente conectadas, através da análise de sua vizinhança. A vizinhança dos estados memorizados será avaliada através da fração de sítios instáveis de um padrão de teste $\boldsymbol{\eta}^k$ à distância de Hamming d do padrão armazenado $\boldsymbol{\xi}^k$. O padrão de teste é gerado pela distribuição condicional de probabilidades

$$p\left(\eta_i^k \mid \xi_i^k\right) = \frac{1+b}{2}\,\delta\left(\eta_i^k - \xi_i^k\right) + \frac{1-b}{2}\,\delta\left(\eta_i^k + \xi_i^k\right), \tag{4.1}$$

onde $0 \leq b \leq 1$ mede a correlação entre $\boldsymbol{\eta}^k$ e $\boldsymbol{\xi}^k$. O parâmetro b está relacionado com a distância de Hamming d entre dois padrões por $d = (1-b)/2$. Lembramos que os padrões memorizados são gerados pela distribuição de probabilidade

$$p\left(\xi_i^k\right) = \frac{1}{2}\,\delta\left(\xi_i^k - 1\right) + \frac{1}{2}\,\delta\left(\xi_i^k + 1\right). \tag{4.2}$$

Na seção 4.1, analisamos o modelo da pseudo-inversa, e na seção 4.2, o modelo dos pesos ótimos.

4.1 Pseudo-inversa

Aqui, como na seção 2.1, a pseudo-inversa é definida pela minimização da energia (1.14)

$$E_i(\mathbf{J}) = \frac{1}{2}\sum_k \left(1 - \Delta_i^k\right)^2, \tag{4.3}$$

com as estabilidades dadas, como antes, por $\Delta_i^k \equiv \frac{1}{\sqrt{N}}\xi_i^l \sum_{j\neq i} J_{ij}\xi_j^l$ e os padrões $\boldsymbol{\xi}^l$ gerados por (4.2). Assim, o cálculo da energia livre e das equações de ponto de sela é idêntico ao do capítulo 2 para o caso sem viés ($a = 0$) e não será repetido aqui.

4.1.1 Cálculo da distribuição de probabilidade das estabilidades

A distribuição de probabilidade das estabilidades do padrão de teste é dada pela equação (1.29),

$$W(\gamma) = \left\langle\left\langle\left\langle\delta\left(\gamma - \Lambda_i^k\right)\right\rangle_J\right\rangle\right\rangle_{\eta^k,\xi^k}, \tag{4.4}$$

onde $\Lambda_i^k \equiv \frac{1}{\sqrt{N}}\eta_i^k \sum_{j\neq i} J_{ij}\eta_j^k$. Para calcular $W(\gamma)$, acrescentamos à energia da pseudo-inversa (1.14) o termo linear em h

$$E_i(\mathbf{J}, h, \gamma) = \frac{1}{2}\sum_k \left(1 - \Delta_i^k\right)^2 + \frac{h}{P}\sum_k \delta\left(\gamma - \Lambda_i^k\right), \tag{4.5}$$

de modo que

$$W(\gamma) = -\lim_{\lambda\to\infty}\frac{1}{\lambda}\frac{\partial\left\langle\left\langle\ln Z_i\right\rangle\right\rangle_{\eta^k,\xi^k}}{\partial h}\Big|_{h=0}, \tag{4.6}$$

onde a função de partição é dada por

$$Z_i(h,\gamma) = \int\left[\prod_j dJ_{ij}\delta\left(Q_i - \frac{1}{N}\sum_j (J_{ij})^2\right)\right]\exp\left[-\lambda E_i(\mathbf{J}, h, \gamma)\right]. \tag{4.7}$$

Aqui, h é uma variável auxiliar, cujo significado físico é irrelevante para nossa análise. O primeiro termo da energia garante a solução de (1.13) e o segundo termo é utilizado para obtenção da distribuição de estabilidades do padrão de teste. Devido à equação (4.6), o procedimento para cálculo de $W(\gamma)$ é análogo ao empregado no cálculo da energia livre do capítulo 2, sendo apresentado a seguir.

Utilizando o método das réplicas, equação (2.4), a função de partição replicada n vezes fica escrita da seguinte forma

$$
\begin{aligned}
Z_i^n &= \int \left[\prod_{j\rho} dJ_{ij}^{\rho} \delta \left(Q_i - \frac{1}{N} \sum_j \left(J_{ij}^{\rho} \right)^2 \right) \right] \\
&\quad \cdot \prod_{k\rho} \int dx_{ki}^{\rho} \delta \left(x_{ki}^{\rho} - \frac{1}{N^{1/2}} \xi_i^k \sum_j J_{ij}^{\rho} \xi_j^k \right) \exp \left[-\frac{\lambda}{2} \left(1 - x_{ki}^{\rho} \right)^2 \right] \\
&\quad \prod_{k\rho} \int dy_{ki}^{\rho} \delta \left(y_{ki}^{\rho} - \frac{1}{N^{1/2}} \eta_i^k \sum_j J_{ij}^{\rho} \eta_j^k \right) \exp \left[-\lambda h \delta \left(\gamma - y_{ki}^{\rho} \right) \right] . \quad (4.8)
\end{aligned}
$$

Utilizando a representação integral da função delta, suprimindo a referência explícita ao sítio i nas variáveis de integração $x_k^{\rho} \equiv x_{ki}^{\rho}$ e $y_k^{\rho} \equiv y_{ki}^{\rho}$ (devido à independência da função de partição com este índice), e explicitando a média sobre os padrões η^k e ξ^k, conseguimos

$$
\begin{aligned}
\langle\langle Z_i^n \rangle\rangle_{\eta^k, \xi^k} &= \int \left[\prod_{j\rho} dJ_{ij}^{\rho} \delta \left(Q - \frac{1}{N} \sum_j \left(J_{ij}^{\rho} \right)^2 \right) \right] \\
&\int \left(\prod_{k\rho} \frac{d\tilde{x}_k^{\rho} dx_k^{\rho}}{2\pi} \right) \exp \left[\mathbf{i} \sum_{k\rho} \tilde{x}_k^{\rho} x_k^{\rho} - \frac{\lambda}{2} \sum_{k\rho} \left(1 - x_k^{\rho} \right)^2 \right] \\
&\int \left(\prod_{k\rho} \frac{d\tilde{y}_k^{\rho} dy_k^{\rho}}{2\pi} \right) \exp \left[\mathbf{i} \sum_{k\rho} \tilde{y}_k^{\rho} y_k^{\rho} - \lambda h \sum_{k\rho} \delta \left(\gamma - y_k^{\rho} \right) \right] \\
&\left\langle\left\langle \exp \left(-\frac{\mathbf{i}}{N^{1/2}} \sum_{k\rho} \tilde{x}_k^{\rho} \xi_i^k \sum_j J_{ij}^{\rho} \xi_j^k - \frac{\mathbf{i}}{N^{1/2}} \sum_{k\rho} \tilde{y}_k^{\rho} \eta_i^k \sum_j J_{ij}^{\rho} \eta_j^k \right) \right\rangle\right\rangle_{\eta^k, \xi^k} . (4.9)
\end{aligned}
$$

Fazendo a média sobre os padrões η_j^k e ξ_j^k, obtemos

$$
\begin{aligned}
\langle\langle \cdots \rangle\rangle_{\eta^k, \xi^k} &\simeq \prod_{kj} \exp \ln \left[1 - \frac{(1-b^2)}{2N} \left(\sum_{\rho} \tilde{x}_k^{\rho} J_{ij}^{\rho} \right)^2 \right. \\
&\qquad \left. - \frac{1}{2N} \left(b \eta_i^k \xi_i^k \sum_{\rho} \tilde{x}_k^{\rho} J_{ij}^{\rho} + \sum_{\rho} \tilde{y}_k^{\rho} J_{ij}^{\rho} \right)^2 \right] \quad (4.10) \\
&\simeq \prod_{kj} \exp \left[-\frac{(1-b^2)}{2N} \left(\sum_{\rho} \tilde{x}_k^{\rho} J_{ij}^{\rho} \right)^2 \right. \\
&\qquad \left. - \frac{1}{2N} \left(b \eta_i^k \xi_i^k \sum_{\rho} \tilde{x}_k^{\rho} J_{ij}^{\rho} + \sum_{\rho} \tilde{y}_k^{\rho} J_{ij}^{\rho} \right)^2 \right] . \quad (4.11)
\end{aligned}
$$

Escrevendo o quadrado da somatória como (2.11) e introduzindo a variável auxiliar $q_{\rho\sigma} = \frac{1}{N}\sum_j J_{ij}^{\rho} J_{ij}^{\sigma}$, a função de partição fica escrita como

$$\begin{aligned}
\langle\langle\, Z_i^n \rangle\rangle_{\eta^k,\xi^k} \;=\; & \int \left[\prod_{j\rho} dJ_{ij}^{\rho}\delta\left(Q - \frac{1}{N}\sum_j \left(J_{ij}^{\rho}\right)^2\right)\right] \int \left[\prod_{\rho<\sigma} dq_{\rho\sigma}\delta\left(q_{\rho\sigma} - \frac{1}{N}\sum_j J_{ij}^{\rho} J_{ij}^{\sigma}\right)\right] \\
& \int \left(\prod_{k\rho} \frac{d\widetilde{x}_k^{\rho} dx_k^{\rho}}{2\pi}\right) \exp\left[\mathbf{i}\sum_{k\rho} \widetilde{x}_k^{\rho} x_k^{\rho} - \frac{\lambda}{2}\sum_{k\rho}\left(1 - x_k^{\rho}\right)^2\right] \\
& \int \left(\prod_{k\rho} \frac{d\widetilde{y}_k^{\rho} dy_k^{\rho}}{2\pi}\right) \exp\left[\mathbf{i}\sum_{k\rho} \widetilde{y}_k^{\rho} y_k^{\rho} - \lambda h \sum_{k\rho} \delta\left(\gamma - y_k^{\rho}\right)\right] \\
& \exp\left[-\frac{\left(1-b^2\right)}{2} Q \sum_{k\rho} \left(\widetilde{x}_k^{\rho}\right)^2 - \left(1-b^2\right) \sum_{k\rho<\sigma} q_{\rho\sigma} \widetilde{x}_k^{\rho} \widetilde{x}_k^{\sigma}\right. \\
& \qquad -\frac{1}{2} Q \sum_{k\rho} \left(b\eta_i^k \xi_i^k \widetilde{x}_k^{\rho} + \widetilde{y}_k^{\rho}\right)^2 \\
& \qquad \left. - \sum_{k\rho<\sigma} q_{\rho\sigma} \left(b\eta_i^k \xi_i^k \widetilde{x}_k^{\rho} + \widetilde{y}_k^{\rho}\right)\left(b\eta_i^k \xi_i^k \widetilde{x}_k^{\sigma} + \widetilde{y}_k^{\sigma}\right)\right]. \qquad (4.12)
\end{aligned}$$

Usando novamente a representação integral da delta nas duas primeiras integrais e fazendo as mudanças de variáveis $\widetilde{Q}_\rho = \frac{\widehat{Q}_\rho}{\mathbf{i}/N}$ e $\widetilde{q}_{\rho\sigma} = \frac{\widehat{q}_{\rho\sigma}}{\mathbf{i}/N}$, a função de partição pode ser posta na seguinte forma

$$\begin{aligned}
\langle\langle\, Z^n \rangle\rangle_{\eta^k,\xi^k} \;\approx\; & \int \left(\prod_{\rho} \frac{d\widetilde{Q}_\rho}{2\pi\mathbf{i}/N}\right) \int \left(\prod_{\rho<\sigma} \frac{dq_{\rho\sigma} d\widetilde{q}_{\rho\sigma}}{2\pi\mathbf{i}/N}\right) \\
& \exp\left[N\sum_{\rho} \widetilde{Q}_\rho Q + N\sum_{\rho<\sigma} \widetilde{q}_{\rho\sigma} q_{\rho\sigma} + N^{1/2}\sum_{\rho} \widetilde{M}_\rho M_\rho\right. \\
& \qquad \left. + NG_0 + \alpha N G_1\right], \qquad (4.13)
\end{aligned}$$

com

$$G_0 = \ln\left\{\int \left(\prod_{\rho} dJ_i^{\rho}\right) \exp\left[-\sum_{\rho} \widetilde{Q}_\rho \left(J_{ij}^{\rho}\right)^2 - \sum_{\rho<\sigma} \widetilde{q}_{\rho\sigma} J_i^{\rho} J_i^{\sigma}\right]\right\} \qquad (4.14)$$

e

$$\begin{aligned}
G_1 \;=\; & \left\langle \ln\left\{\int \left(\prod_{\rho} \frac{d\widetilde{x}^{\rho} dx^{\rho}}{2\pi}\right) \int \left(\prod_{\rho} \frac{d\widetilde{y}^{\rho} dy^{\rho}}{2\pi}\right) \exp\left[\mathbf{i}\sum_{\rho} \widetilde{x}^{\rho} x^{\rho}\right.\right.\right. \\
& -\frac{\lambda}{2}\sum_{\rho}\left(1 - x^{\rho}\right)^2 + \mathbf{i}\sum_{\rho} \widetilde{y}^{\rho} y^{\rho} - \lambda h \sum_{\rho} \delta\left(\gamma - y^{\rho}\right) - \frac{\left(1-b^2\right)}{2} Q \sum_{\rho} \left(\widetilde{x}^{\rho}\right)^2
\end{aligned}$$

$$-\left(1-b^2\right)\sum_{\rho<\sigma} q_{\rho\sigma}\widetilde{x}^{\rho}\widetilde{x}^{\sigma}-\frac{1}{2}Q\sum_{\rho}\left(b\eta_i\xi_i\widetilde{x}^{\rho}+\widetilde{y}^{\rho}\right)^2$$

$$-\sum_{\rho<b}\left(b\eta_i\xi_i\widetilde{x}^{\rho}+\widetilde{y}^{\rho}\right)\left(b\eta_i\xi_i\widetilde{x}^{\sigma}+\widetilde{y}^{\sigma}\right)\Bigg]\Bigg\}\Bigg\rangle_{\eta_i,\xi_i}. \tag{4.15}$$

Na derivação de G_1, foi utilizada a propriedade de automediância, $\sum_k g\left(\eta_i^k\xi_i^k\right) = \alpha N\left\langle\left\langle g\left(\eta_i\xi_i\right)\right\rangle\right\rangle_{\eta_i,\xi_i}$, que resulta na eliminação do índice k. Na equação acima, as médias sobre η_i e ξ_i são indicadas por $\langle\ldots\rangle_{\eta_i,\xi_i}$. Novamente, a forma específica da energia de treinamento altera apenas o termo G_1.

Cálculo de G_1

Para obter a distribuição de estabilidades para o padrão de teste, precisamos somente de G_1. O procedimento para efetuar as integrais em $\widetilde{x}^{\rho}$, x^{ρ}, $\widetilde{y}^{\rho}$ e y^{ρ} é análogo ao apresentado no capítulo 2 e por isso não o reproduzimos aqui. Após calcular estas integrais, a expressão para G_1 pode ser colocada na seguinte forma compacta

$$\frac{G_1}{n} = \langle\int Dt\int Dz\ln f\left(t,\ z,\omega,h\right)\rangle_{\eta_i,\xi_i}, \tag{4.16}$$

onde

$$f\left(t,\ z,\omega,h\right) = \left[1+\lambda\left(1-b^2\right)\left(Q-q\right)\right]^{-1/2}$$
$$\int D\omega\exp\left[-\lambda h\delta\left(\gamma-q^{1/2}z-\left(Q-q\right)^{1/2}\omega\right)\right]$$
$$\exp\left[\frac{\lambda}{2}\frac{\left[1-\left(1-b^2\right)^{1/2}q^{1/2}t-b\eta_i\xi_i\left(q^{1/2}z+\left(Q-q\right)^{1/2}\omega\right)\right]^2}{\left[1+\lambda\left(1-b^2\right)\left(Q-q\right)\right]}\right]. \tag{4.17}$$

E através de (4.4), obtemos a distribuição de estabilidades

$$\begin{aligned} W\left(\gamma\right) &= -\lim_{\lambda\to\infty}\frac{1}{\lambda n}\frac{\partial G_1}{\partial h}\Big|_{h=0} \\ &= -\lim_{\lambda\to\infty}\frac{1}{\lambda}\left\langle\int Dt\int Dz\frac{f'\left(t,\ z,\omega,0\right)}{f\left(t,\ z,\omega,0\right)}\right\rangle_{\eta_i,\xi_i}, \end{aligned} \tag{4.18}$$

onde $f'\left(t,\ z,\omega,0\right) = \frac{\partial f(t,\ z,\omega,h)}{\partial h}\Big|_{h=0}$. Todas as integrais envolvidas são gaussianas e podem ser efetuadas facilmente. A forma final para a distribuição de estabilidades

é

$$W(\gamma) = \langle w_{\eta_i,\xi_i}(\gamma)\rangle_{\eta_i,\xi_i}, \tag{4.19}$$

onde $w_{\eta_i,\xi_i}(\gamma)$ é uma distribuição de probabilidade gaussiana para dado $\eta_i\xi_i$, como em (2.48), dada pela seguinte expressão

$$w_{\eta_i,\xi_i}(\gamma) = \sqrt{\frac{1}{2\pi Q(1-b^2)}}\exp\left[-\frac{(\gamma - b\eta_i\xi_i)^2}{2Q(1-b^2)}\right], \tag{4.20}$$

onde $Q = \alpha/(1-\alpha)$. Realizando a média sobre $\eta_i\xi_i$, obtemos

$$\begin{aligned} W(\gamma) \quad &= \quad = \sqrt{\frac{1}{2\pi Q(1-b^2)}}\left\{\frac{1+b}{2}\exp\left[-\frac{(\gamma-b)^2}{2Q(1-b^2)}\right]\right. \\ &\left. + \frac{1-b}{2}\exp\left[-\frac{(\gamma+b)^2}{2Q(1-b^2)}\right]\right\}, \end{aligned} \tag{4.21}$$

que não é mais gaussiana.

É conveniente enfatizar que as distribuições de probabilidades condicionais $p_{\xi_i}(\gamma)$ (2.48) e $w_{\eta_i,\xi_i}(\gamma)$ (4.20) são gaussianas. Esse resultado será utilizado no capítulo 6 para simplificar os cálculos referentes ao problema da categorização.

Cálculo da fração de sítios instáveis

Desejamos obter a fração de sítios instáveis do padrão de teste $\boldsymbol{\eta}^k$, isto é, a fração ϵ entre o número de sítios para os quais a estabilidade

$$\Lambda_i^k = \frac{1}{\sqrt{N}}\eta_i^k \sum_{j\neq i} J_{ij}\eta_j^k \tag{4.22}$$

é negativa e o número total de sítios N. A independência estatística entre η_i^k e ξ_i^k para diferentes sítios permite escrever esta fração como

$$\epsilon = \int_{-\infty}^{0} d\gamma W(\gamma), \tag{4.23}$$

onde $W(\gamma) = \left\langle\left\langle\left\langle\delta\left(\gamma - \Lambda_i^k\right)\right\rangle_J\right\rangle\right\rangle_{\eta_i,\xi_i}$ é a distribuição de probabilidade das estabilidades do padrão de teste $\boldsymbol{\eta}$, dada pela equação (4.21).

Efetuando a integral sobre γ, obtemos a seguinte expressão para a fração de sítios instáveis em $\boldsymbol{\eta}$

$$\epsilon = \frac{1-b}{2} + bH\left[\frac{b}{\sqrt{Q\left(1-b^2\right)}}\right], \tag{4.24}$$

onde $H(x) = \int_x^\infty Dt$. Esta quantidade é mostrada na figura 4.1 como uma função da distância de Hamming para vários valores de α. Para $\alpha = 0$ obtemos $\epsilon = d$. À medida que α cresce em direção a $\alpha_c = 1$, a vizinhança do padrão armazenado torna-se mais abrupta: para pequenos valores de d a função erro é muito sensível a variações da distância de Hamming, enquanto para grandes valores de d ela se torna praticamente independente da distância dos padrões armazenados. Em particular, para $\alpha = \alpha_c = 1$ tem-se $\epsilon = 0$ se $d = 0$ e $\epsilon = \frac{1}{2}$ caso contrário.

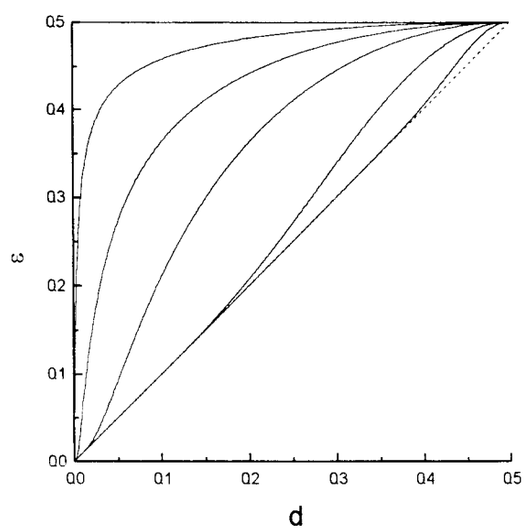

Figura 4.1: Fração de sítios instáveis do padrão de teste η como função da distância de Hamming do padrão armazenado ξ para a rede neural atratora pseudo-inversa. Os parâmetros são (de cima para baixo) $\alpha = 1$, 0.99, 0.9, 0.6, 0.1 e 0.01. A linha pontilhada é $\epsilon = d$, que coincide com a curva de ϵ para $\alpha = 0$.

4.2 Pesos ótimos

Os cálculos para o modelo dos pesos ótimos são análogos aos descritos anteriormente. O fator complicante é que algumas integrais não podem ser efetuadas analiticamente, mas o procedimento de cálculo segue idêntico ao do da pseudo-inversa. Para o modelo dos pesos ótimos, utilizamos a seguinte expressão para a energia

$$E_i\left(\mathbf{J}, h, \gamma\right) = \sum_k \theta\left(1 - \Delta_i^k\right)^2 + \sum_k h\delta\left(\gamma - \Lambda_i^k\right). \tag{4.25}$$

Seguindo os procedimentos indicados na seção anterior, chegamos à seguinte equação para a distribuição de probabilidade das estabilidades do padrão de teste $\boldsymbol{\eta}$

$$W\left(\gamma\right) = \frac{\mathrm{e}^{-\gamma^2/2}}{\sqrt{2\pi}} \left\langle \int_{-\infty}^{\infty} Dy \int_{-\infty}^{\infty} Dx \frac{H\left[\Xi_1\right]}{H\left[\Xi_2\right]} \right\rangle_{\eta_i, \xi_i}, \tag{4.26}$$

onde

$$\Xi_1 = \frac{\kappa - b\gamma\eta_i\xi_i - \sqrt{q\left(1 - b^2\right)}x}{\sqrt{\left(1 - q\right)\left(1 - b^2\right)}} \tag{4.27}$$

e

$$\Xi_2 = \frac{\kappa - b\eta_i\xi_i\left(\gamma q + \sqrt{q\left(1 - q\right)}y\right) - \sqrt{q\left(1 - b^2\right)}x}{\sqrt{1 - q}}. \tag{4.28}$$

O parâmetro de ordem q é dado pela equação (2.61). Como antes, a integral tripla que aparece no cálculo de ϵ, equação (4.23), pode ser reduzida a uma integral dupla através de uma mudança apropriada de variáveis que permite a integração analítica sobre γ, levando a

$$\begin{aligned}\epsilon = \frac{1+b}{2} \int_0^{\infty} Dw \int Dx \frac{H\left[h_1\left(\kappa + w\right) - h_2 x\right]}{H\left[h_3\kappa + h_4 w - h_5 x\right]} \\ + \frac{1-b}{2} \int_0^{\infty} Dw \int Dx \frac{H\left[h_1\left(\kappa - w\right) + h_2 x\right]}{H\left[h_3\kappa - h_4 w + h_5 x\right]},\end{aligned} \tag{4.29}$$

onde $h_1^2 = \frac{\left(1-b^2 q\right)}{(1-q)(1-b^2)}$, $h_2^2 = \frac{q\left(1-b^2\right)}{(1-q)}$, $h_3^2 = b^2 q^2 \frac{\left(1-b^2\right)}{(1-b^2 q)(1-q)}$, $h_4^2 = \frac{1}{(1-q)}$ e $h_5^2 = \frac{q\left(1-b^2 q\right)}{(1-q)}$.

A dependência de ϵ com d para $\kappa = 0$ é apresentada na figura 4.2. Em particular, para $\alpha = 0$, obtemos

$$\epsilon\left(\alpha = 0\right) = \frac{1-b}{2} + \frac{b}{\pi} \arccos b, \tag{4.30}$$

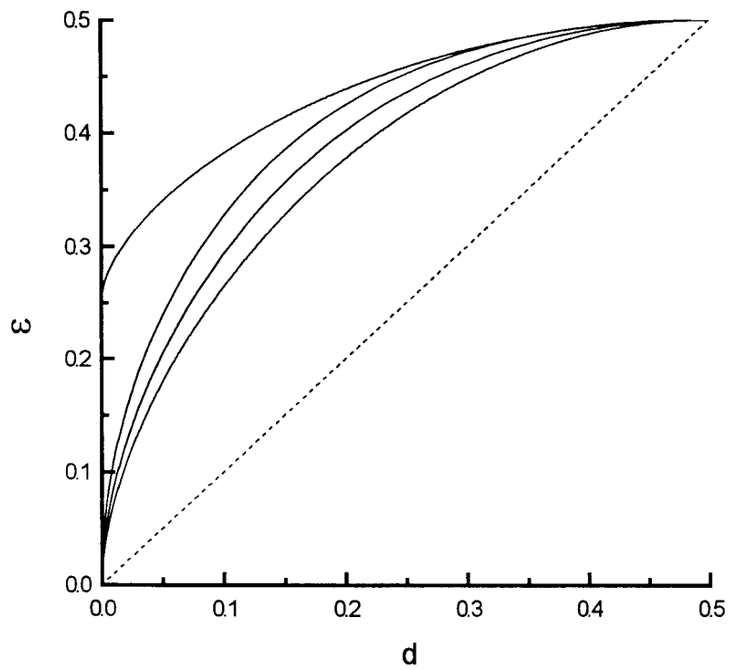

Figura 4.2: Fração de sítios instáveis do padrão de teste η como função da distância de Hamming do padrão armazenado ξ para a rede neural atratora pesos ótimos. Os parâmetros são $\kappa = 0$ e (de cima para baixo) $\alpha = 2$, 1, 0.5 e 0. A linha pontilhada é $\epsilon = d$.

enquanto para $\alpha = \alpha_c = 2$, obtemos $\epsilon(\alpha = 2) = 0$ para $b = 1$ e $\epsilon(\alpha = 2) = 1/4 + \epsilon(\alpha = 0)/2$, caso contrário. É interessante comparar esses resultados com os da pseudo-inversa. Enquanto a dependência de $\epsilon(\alpha = 0)$ com d é muito simples para a pseudo-inversa, $\epsilon(\alpha = 0) = d$, ela é mais complexa para a rede dos pesos ótimos: qualquer desvio do padrão armazenado leva a um aumento abrupto do número de sítios instáveis, uma vez que todas as derivadas de $\epsilon(\alpha = 0)$ divergem em $d = 0$.

Para melhor comparar os modelos da pseudo-inversa e dos pesos ótimos, escolhemos $\kappa \approx 0.470$ de modo que a capacidade de armazenamento de ambos os modelos

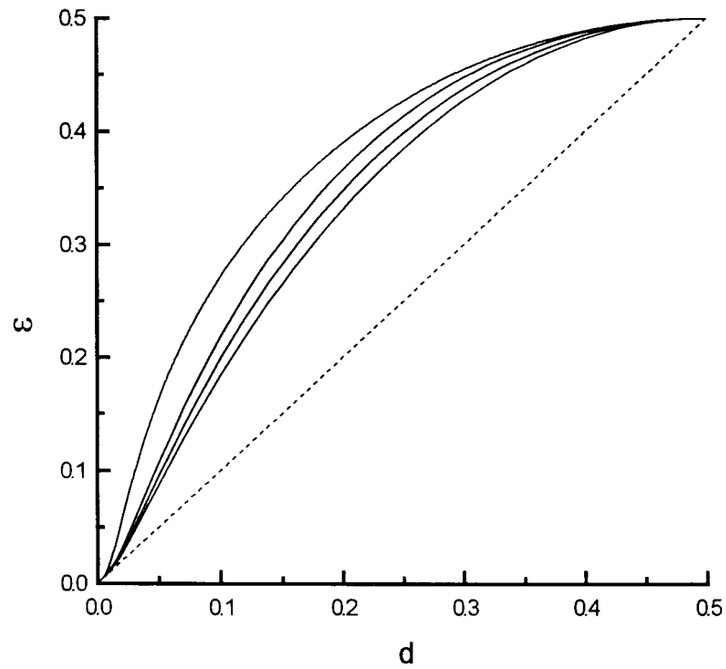

Figura 4.3: O mesmo da figura 4.2, mas com $\kappa = 0.470$ e (de cima para baixo) $\alpha = 1$, 0.5, 0.25 e 0.

sejam iguais ($\alpha_c = 1$). A figura 4.3 mostra a dependência de ϵ com d para este caso. Para pequenos valores de d tem-se $\epsilon \approx d$, o que contrasta com o comportamento não analítico para o caso $\kappa = 0$. É interessante que um valor não zero para κ garanta uma vizinhança suave, isto é, $\epsilon \approx d$, mesmo no regime de saturação $\alpha = \alpha_c$. Grandes valores de κ aumentam a faixa de d para a qual $\epsilon \approx d$. Em particular, para $\kappa \to \infty$ tem-se $\epsilon \to d$ para $\alpha \leq \alpha_c \to 0$. As figuras 4.4 e 4.5, para $\kappa = 0.5$ e 0.8, respectivamente, confirmam a tendência da função erro em suavizar o comportamento em $d = 0$.

Aparentemente, o papel de κ é suavizar a vizinhança dos padrões armazenados. Entretanto, é necessário salientar que embora a função ϵ possa indicar uma vizinhan-

ça suave, isto não diz nada sobre a bacia de atração quando se utiliza a dinâmica (1.1), pois esta depende localmente das estabilidades e não de suas integrais, como em ϵ.

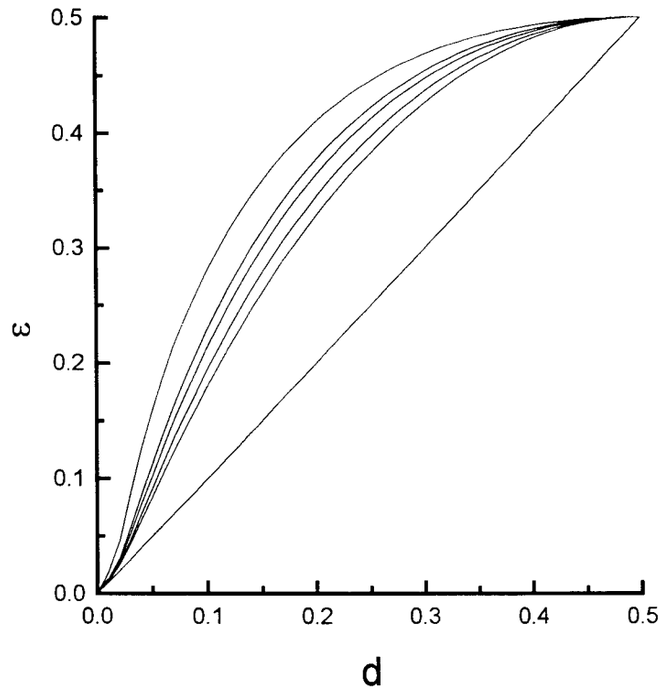

Figura 4.4: O mesmo da figura 4.2, mas com $\kappa = 0.5$ e (de cima para baixo) $\alpha = 0.96$, 0.65, 0.5, 0.25 e 0.

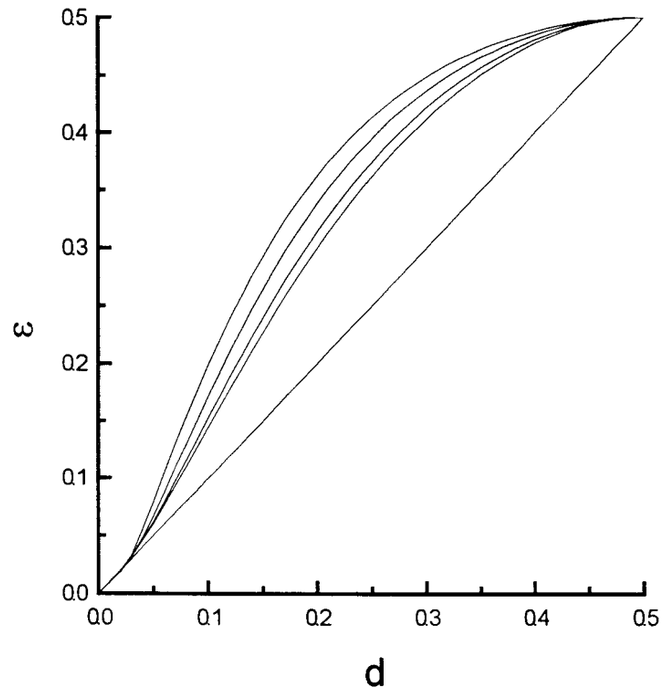

Figura 4.5: O mesmo da figura 4.2, mas com $\kappa = 0.8$ e (de cima para baixo) $\alpha = 0.65$, 0.5, 0.25 e 0.

Capítulo 5

Natureza dos atratores

Como vimos no capítulo 3, para avaliar o desempenho de uma rede neural é interessante conhecer características como o tamanho das bacias de atração e a estabilidade dos atratores, que podem ser avaliadas através do estudo da dinâmica da rede neural. Naquele capítulo, obtivemos os diagramas de fase para redes neurais atratoras no regime de extrema diluição. Somente neste limite é possível obter analiticamente a equação da dinâmica para a correlação de recuperação das redes pseudo-inversa e pesos ótimos. A fim de complementar nosso estudo das propriedades de recuperação de redes neurais atratoras completamente conectadas, é necessário recorrer ao uso de simulação.

Neste capítulo, apresentamos os métodos utilizados para a obtenção explícita dos pesos sinápticos para o modelo dos pesos ótimos e para as três formulações do modelo da pseudo-inversa. Em seguida, descrevemos o algoritmo que realiza a enumeração exaustiva de todas as configurações da rede [30] e apresentamos os resultados obtidos através de simulações numéricas. Devido ao alto custo computacional dessa busca, estudamos redes com $N \leq 24$.

5.1 Determinação dos pesos sinápticos

Pseudo-inversa

O modelo da pseudo-inversa tem sido estudado na literatura em três diferentes formulações, que apresentamos a seguir.

(1) Modelo PGD: A primeira formulação da pseudo-inversa obtém os pesos sinápticos através da equação

$$J_{ij} = \frac{1}{N} \sum_{kl}^{P} \xi_i^k \xi_j^l \left(C^{-1}\right)_{kl}, \tag{5.1}$$

onde os elementos da matriz simétrica $\mathbf{C}$ são dados por $C_{kl} = 1/N \sum_i \xi_i^k \xi_i^l$ [19] e os índices i e j correm sobre os N neurônios da rede. A equação (5.1) produz uma matriz de pesos simétrica com termos diagonais J_{ii} não nulos. A solução do sistema de equações (1.13), dada por (5.1), permite armazenar $P = N$ padrões linearmente independentes. Esse modelo foi estudado numericamente por Personnaz, Guyon e Dreyfus [20].

(2) Modelo KS: Este modelo, dado pela mesma prescrição anterior (5.1), mas com os termos diagonais J_{ii} zerados, foi estudado por Kanter e Sompolinsky [21]. Eliminando os termos diagonais de (5.1), não se pode mais garantir que os padrões sejam atratores da dinâmica. Entretanto, Kanter e Sompolinsky mostraram analiticamente que no limite $N \to \infty$ a capacidade de armazenagem desse modelo é $\alpha_c = 1$. Os atratores da formulação KS não coincidem, em geral, com os da formulação PGD e, além disso, suas bacias de atração diferem enormemente.

(3) Modelo E: É o modelo definido pela solução de menor norma do sistema de equações lineares (1.13). Ele produz uma matriz de pesos assimétrica com termos diagonais nulos, sendo que o número de padrões linearmente independentes que podem ser armazenados nessa formulação é $P = N - 1$.

Pesos ótimos

Os pesos sinápticos do modelo dos pesos ótimos devem satisfazer às desigualdades

$$\Delta_i^l \equiv \frac{1}{\sqrt{N}} \xi_i^l \sum_{j \neq i} J_{ij} \xi_j^l \geq \kappa \tag{5.2}$$

para todo $i = 1, \ldots, N$ e $l = 1, \ldots, P$, sujeitas à condição

$$\sum_j J_{ij}^2 = N, \qquad \forall i. \tag{5.3}$$

A determinação dos pesos sinápticos para a rede ótima, equações (5.2) e (5.3), pode ser feita por um algoritmo originalmente proposto para redes tipo perceptron [44] [45]. Iniciando com uma matriz de pesos arbitrária, com $J_{ii} = 0$, examinam-se as desigualdades (5.2) para cada i e l e, toda vez que uma não for satisfeita (isto é, $\Delta_i^l < \kappa$), os J_{ij} correspondentes são modificados de acordo com a regra

$$J_{ij} \rightarrow J_{ij} + \frac{\lambda}{N} \xi_i^l \xi_j^l \delta \left(1 - \delta_{ij}\right) \theta \left(\kappa - \Delta_i^l\right), \tag{5.4}$$

onde λ é um parâmetro positivo que ajusta o tamanho do passo. A cada iteração os pesos são renormalizados para satisfazer (5.3). Este é o algoritmo padrão que Gardner [24] mostrou convergir também para redes neurais atratoras. Entretanto, sob vários aspectos, este algoritmo não é o mais eficiente. Em primeiro lugar, o tamanho do passo λ de cada iteração de (5.4) é sempre o mesmo, independente do quanto as desigualdades (5.2) ou a normalização (5.3) diferem dos valores corretos. Abbott e Kepler [25] apresentaram uma variante mais eficiente desse algoritmo, aumentando o tamanho do passo se $\Delta_i^l \ll \kappa$ e diminuindo-o se Δ_i^l está próximo de κ. Além disso, o tamanho do passo também é ajustado de acordo com a magnitude de $\sum_j J_{ij}^2$. O algoritmo proposto por aqueles autores é dado pela seguinte regra

$$J_{ij} \rightarrow J_{ij} + \frac{1}{N} \xi_i^l \xi_j^l f\left(\Delta_i^l\right) \|J_{ij}\| \, \delta \left(1 - \delta_{ij}\right) \theta \left(\kappa - \Delta_i^l\right), \tag{5.5}$$

onde $\|J_{ij}\| = \sqrt{\sum_j J_{ij}^2}$ é o fator de normalização e $f\left(\Delta_i^l\right)$ é uma função que deve ser escolhida de forma a otimizar a velocidade de convergência do método. Seguindo as

sugestões de Abbott e Kepler [25], adotamos a função não linear $f\left(\Delta_i^l\right) = \kappa + \delta - \Delta_i^l + \left[\left(\kappa + \delta - \Delta_i^l\right)^2 - \delta^2\right]^{1/2}$ com $\delta = 0.01$. Dado o conjunto de padrões que se quer armazenar na rede neural, nem sempre será possível obter uma matriz de pesos satisfazendo aos vínculos (5.2). De fato, para um dado i, os padrões $\left(\xi_1^l, \ldots, \xi_{i-1}^l, \xi_{i+1}^l, \ldots, \xi_N^l\right)$ devem ser linearmente separáveis para que o algoritmo encontre uma solução. Além disso, a solução a que se chega depende da particular escolha da matriz de pesos inicial.

Ao contrário da matriz de pesos da rede ótima, obtida pelo algoritmo do perceptron e suas variantes, a matriz de pesos da pseudo-inversa para cada uma das três formulações que estudaremos é única. É importante notar que, em todos os casos (pseudo-inversa e pesos ótimos), os padrões $\left\{-\boldsymbol{\xi}^l\right\}$ são também automaticamente memorizados.

5.2 Determinação dos atratores

Para caracterizar os atratores, utilizamos um algoritmo simples e engenhoso, proposto por Gutfreund, Reger e Young [30]. Através deste algoritmo, obtemos o número de atratores, o tipo e sua bacia de atração. Como o número total de estados de uma rede neural binária cresce exponencialmente com N, o algoritmo que realiza a enumeração exaustiva dos atratores deve ser eficiente para não inviabilizar a operação. A seguir, descrevemos simplificadamente o algoritmo utilizado.

O número de estados $\mathbf{S} = (S_1, \ldots, S_i, \ldots, S_N)$ de uma rede neural binária é 2^N, podendo ser enumerados associando-se cada uma das configurações da rede à sua representação decimal. Assim, cada estado da rede é representado pelo inteiro b que pode assumir os valores $b = 0, 1, \ldots, \left(2^N - 1\right)$, num total de 2^N estados[1]. A

[1]Por exemplo, ao estado da rede representado pelo vetor de estado com todos os spins 0 é associado o índice $b = 0$, resultando em $\mathbf{S}^0 \rightarrow (0, \ldots, 0)$. Numa rede de $N = 4$ spins, o estado com todos os spins 1 é associado ao índice $b = 2^4 - 1 = 15$. Assim, $\mathbf{S}^{15} \rightarrow (1, 1, 1, 1)$, ou simplesmente dizemos que a rede está no estado de número $b = 15$.

cada estado b da rede são associadas duas variáveis inteiras: n_b armazena o próximo estado da rede e m_b que identifica a bacia de atração à qual b pertence. Temos ainda o vetor Ω_i que armazena o número de estados na i-ésima bacia de atração. Inicialmente, m_b e Ω_i são zerados. O primeiro passo é obter, para cada estado da rede, o próximo estado, através da equação (1.1), $S_i(t+1) = \text{sign}\left[\frac{1}{\sqrt{C}}\sum_j J_{ij}S_j(t)\right]$, que fica armazenado em n_b. Denota-se o estado atual da rede por a e o número de bacias já identificadas por i (inicialmente $a = i = 0$). O algoritmo, então, segue os seguintes passos.

(a) Se este estado já foi visitado antes ($m_a \neq 0$) passa-se para o próximo estado $a = a + 1$. Se ele não foi visitado ($m_a = 0$), então, identifica-se a bacia de atração à que pertence fazendo $m_a = i + 1$, prosseguindo a dinâmica até se atingir um estado que já tenha sido visitado antes ($m_a \neq 0$). Agora, existem duas possibilidades, (b1) ou (b2), dadas a seguir.

(b1) Se $m_a = i + 1$, então, este estado é um novo atrator e utiliza-se o número de estados visitados desde o estado inicial a, como uma primeira estimativa de sua bacia, Ω_{i+1}.

(b2) Se $m_a < i + 1$, então, foi atingido um estado pertencente a uma bacia de atração já identificada. Volta-se ao estado a e segue-se a dinâmica novamente, mas agora nomeando a bacia de atração com m_a. O número de estados desde a até o estado da bacia m_a é agora somado à antiga estimativa do número de estados nesta bacia de atração, Ω_{m_a}.

(c) Se $a + 1 < 2^N$, então, volta-se para (a), para nova iteração.

Neste algoritmo, utilizamos a regra (1.1) para obter o próximo passo da dinâmica e assim identificar os atratores. De fato, nesta tese, aplicaremos a dinâmica (1.1) de duas maneiras. À primeira denominamos de dinâmica serial, que consiste em aplicar (1.1) a um spin de cada vez, sempre na mesma seqüência. A segunda dinâmica que consideraremos é denominada de dinâmica paralela e consiste em aplicar (1.1) a

todos os spins de uma única vez. Como veremos a seguir, ambas as dinâmicas garantem a recuperação dos estados memorizados, mas com propriedades distintas. Vale salientar que este algoritmo para enumeração exaustiva dos atratores funciona apenas para dinâmicas determinísticas, como é o caso das duas dinâmicas estudadas. Para uma dinâmica serial estocástica, em que o sítio a evoluir é escolhido aleatoriamente a cada passo, não é possível determinar de forma única o sucessor de um dado estado, impossibilitando a contagem do número de estados nas bacias de atração. Da mesma forma, não se aplica este método à dinâmica com temperatura diferente de zero.

Após identificados os atratores e contados o número de estados na bacia de cada um deles, definimos o peso da s-ésima bacia de atração como

$$W_s = \Omega_s / 2^N, \tag{5.6}$$

onde Ω_s é o número de estados que convergem para o s-ésimo atrator. A estrutura das bacias de atração é, então, caracterizada pelos momentos

$$Y_n = \sum_s (W_s)^n, \qquad n = 1,\ 2,\ \ldots \tag{5.7}$$

Quando $n = 1$, a soma dos pesos das bacias de todos os atratores é necessariamente 1 e, de fato, esta propriedade foi utilizada para verificar a correteza do algoritmo. De particular interesse para nossa análise é o caso $n = 2$. O segundo momento Y_2 dá a probabilidade de que dois estados arbitrariamente escolhidos pertençam à mesma bacia de atração. Assim, se $Y_2 \to 0$ no limite $N \to \infty$, então, o espaço de configurações é dominado por muitos atratores cujos pesos W_s se tornam cada vez menores. Por outro lado, se Y_2 permanece finito neste limite, então, uns poucos atratores dominam praticamente todo o espaço de configurações.

Na equação (5.7), o índice s corre sobre todos os atratores. Entretanto, é de interesse também investigar as bacias dos padrões memorizados. Para isso, definimos a quantidade

$$\mathcal{N}_\xi = \frac{1}{P} \sum_{s \in \{\xi\}} \Omega_s, \tag{5.8}$$

onde $s \in \{\xi\}$ significa que este índice corre somente sobre os padrões memorizados. Este é o número médio de estados na bacia de atração de cada padrão.

A seguir, discutimos como as grandezas introduzidas acima dependem dos vários parâmetros que definem os modelos da pseudo-inversa e pesos ótimos. Nossa discussão focalizará principalmente a dinâmica seqüencial. Os resultados da dinâmica paralela serão apresentados brevemente no final do capítulo.

5.3 Análise dos resultados

Nos gráficos mostrados nesta seção, que apresentam os resultados das simulações, cada ponto representa uma média sobre 100 realizações diferentes do conjunto de padrões $\{\boldsymbol{\xi}\}$. Os valores de N serão sempre ímpares para o modelo PGD e pares para o modelo KS da pseudo-inversa e dos pesos ótimos. Essa escolha objetiva evitar que o argumento da função sinal na equação dinâmica (1.1) se anule, o que introduziria uma arbitrariedade desnecessária na dinâmica. Como a matriz de pesos do modelo de pesos ótimos não é simétrica, mesmo no caso da dinâmica seqüencial podem surgir ciclos de período arbitrário (limitado, e claro, pelo número de estados 2^N). Lembramos que para as formulações PGD e KS da pseudo-inversa, cujas matrizes de pesos são simétricas, a dinâmica seqüencial leva apenas a pontos fixos. Assim, deve ficar claro que o número total de pontos fixos, discutido a seguir, não corresponde ao número total de atratores do modelo dos pesos ótimos. A análise do modelo de pesos ótimos ficará restrita a apenas três valores do parâmetro de margem, a saber, $\kappa = 0$, 0.5 e 0.8, para os quais $\alpha_c = 2$, 0.96 e 0.66, respectivamente. Devemos mencionar ainda que, dentro da precisão das figuras que serão apresentadas a seguir, não é possível distinguir entre os resultados da formulação KS e os da formulação E, embora as matrizes de pesos sejam diferentes nos dois casos (simétrica para KS e assimétrica para E). Assim, apresentaremos apenas os resultados para a formulação KS da pseudo-inversa.

Dependência com N para α fixo

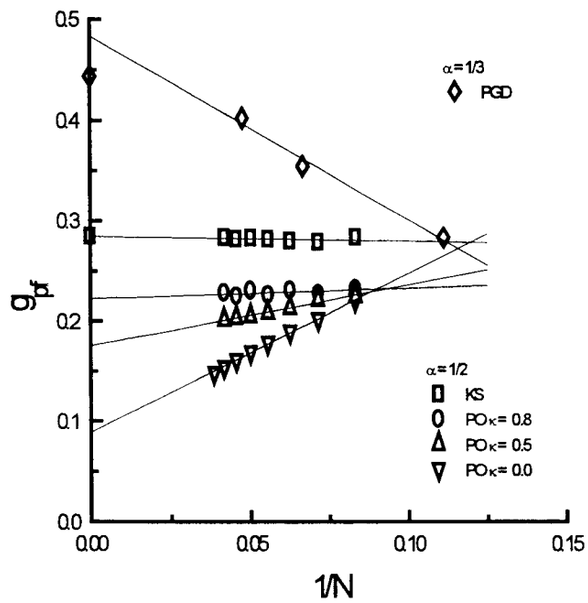

Figura 5.1: Função $g_{pf}(N,\alpha) = \frac{1}{N}\ln\langle\mathcal{N}_{pf}\rangle$ contra o inverso do número de neurônios N para α fixo. Como é claro da figura, o comportamento de $g_{pf}(N,\alpha)$ no limite $1/N \longrightarrow 0$ é claramente exponencial e fornece o valor de $g_\infty(\alpha)$ no limite termodinâmico. Os pontos sobre o eixo $1/N \longrightarrow 0$ são obtidos da previsão teórica. As curvas plotadas dos modelos KS e PO são para $\alpha = 1/2$ e do modelo PGD são para $\alpha = 1/3$.

A fim de determinar como o número médio de pontos fixos $\langle\mathcal{N}_{pf}\rangle$ cresce com N, apresentamos na figura 5.1 a função $g_{pf}(N,\alpha) = \frac{1}{N}\ln\langle\mathcal{N}_{pf}\rangle$ contra $1/N$. Podemos inferir dessa figura a seguinte relação

$$\langle\mathcal{N}_{pf}\rangle = c(\alpha)\exp\left[g_\infty(\alpha)N\right], \tag{5.9}$$

onde $\ln c(\alpha)$ é o coeficiente angular da reta que interpola os dados e $g_\infty(\alpha)$ o ponto onde ela toca o eixo $1/N = 0$. No caso da pseudo-inversa, os valores analíticos,

obtidos para $N \longrightarrow \infty$ por Kuhlmann e Anlauf [46], são também apresentados nesta figura. Embora a concordância entre a extrapolação dos resultados das simulações para o limite $1/N \longrightarrow 0$ e os valores analíticos para o modelo KS seja excelente, essa concordância não é muito boa para o modelo PGD, devido aos pequenos valores de P empregados nas simulações ($P = 3$, 5 e 7 para $N = 9$, 15 e 21, respectivamente). Um resultado importante e original apresentado nesta figura é o de que o número médio de pontos fixos também cresce exponencialmente com N no modelo dos pesos ótimos e, em particular, aumenta com κ.

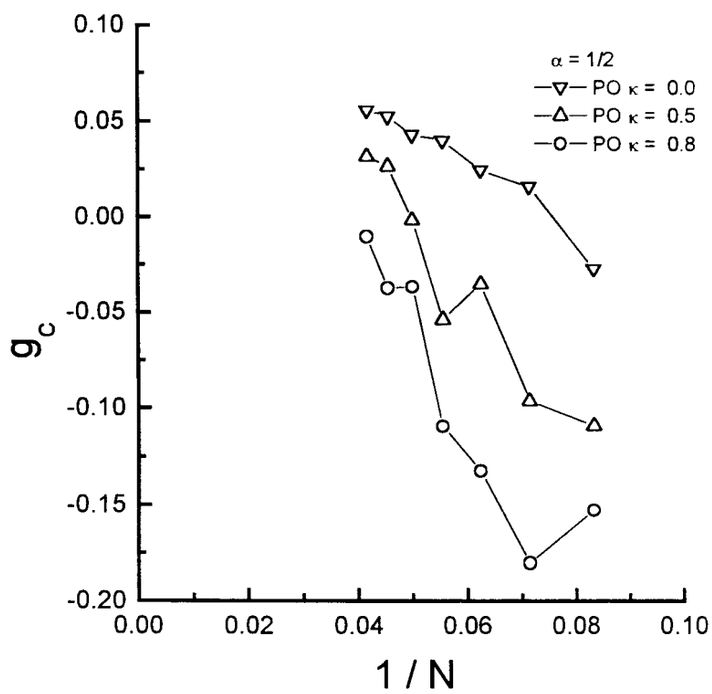

Figura 5.2: Função $g_c(N, \alpha) = \frac{1}{N} \ln \langle \mathcal{N}_c \rangle$ contra o inverso do número de neurônios $1/N$ para α fixo.

Analogamente, introduzimos a função $g_c(N, \alpha) = \frac{1}{N} \ln \langle \mathcal{N}_c \rangle$ para investigar a dependência do número médio de ciclos $\langle \mathcal{N}_c \rangle$ com N e α. Os resultados mostrados

na figura 5.2 para o modelo de pesos ótimos indicam o fato um pouco surpreendente de que o número de ciclos aumenta com N. Note que g_c pode tomar valores negativos se algumas das realizações dos padrões mediadas não produzirem ciclos, uma vez que neste caso teríamos $\langle \mathcal{N}_c \rangle < 1$.

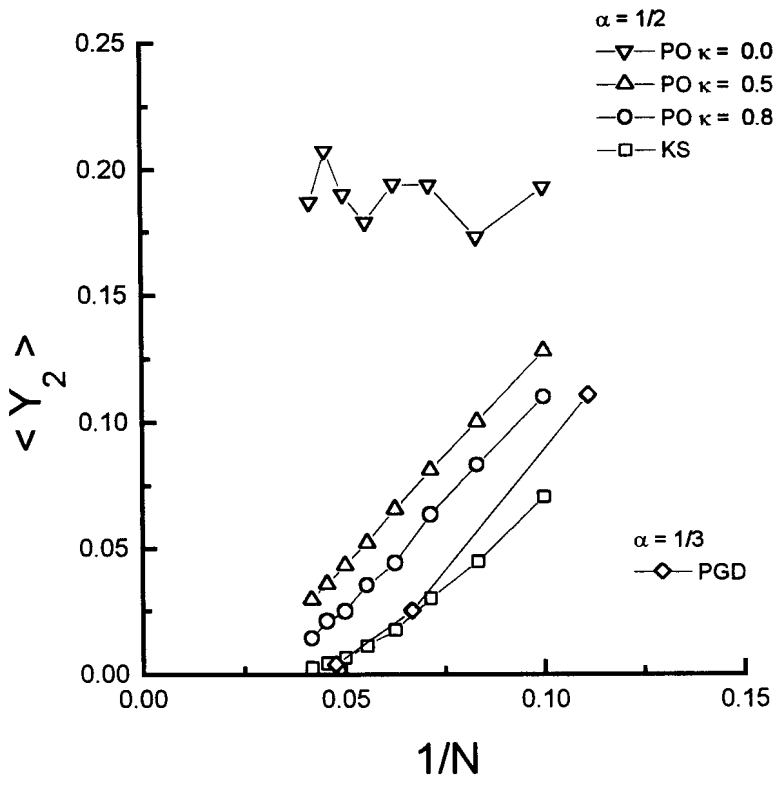

Figura 5.3: Y_2 contra o inverso do número de neurônios $1/N$ para α fixo.

A figura 5.3 apresenta a dependência do valor médio do segundo momento da distribuição dos pesos das bacias de atração, equação (5.7), com N. Note que aqui são considerados todos os atratores: pontos fixos e ciclos. Há uma clara diferença qualitativa entre o modelo de pesos ótimos com $\kappa = 0$ e os outros modelos: o fato de $\langle Y_2 \rangle$ tender a um valor não nulo para $N \to \infty$ implica na existência de atratores com enormes bacias de atração no caso $\kappa = 0$. É instrutivo comparar esses resultados com o de um modelo uniforme, no qual todas as bacias de atração são idênticas, de

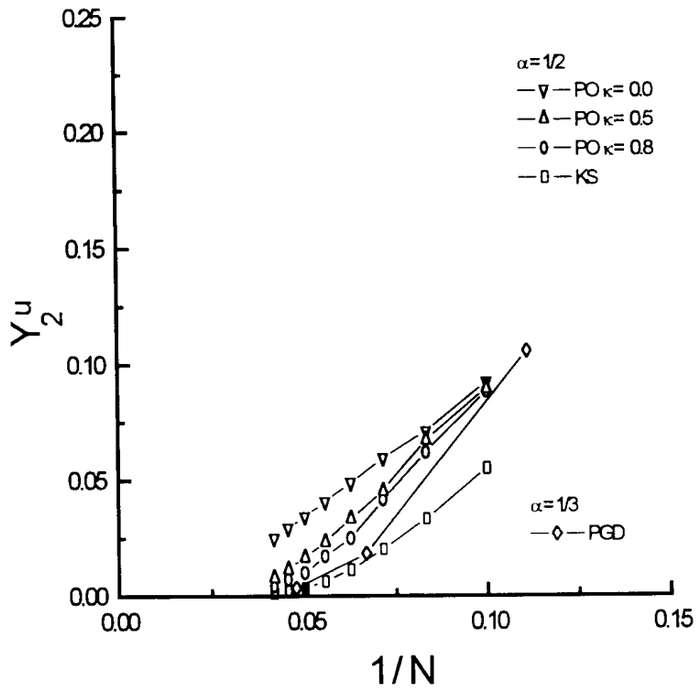

Figura 5.4: Previsão para o segundo momento no caso de bacias uniformes $Y_2^u = \exp(-Ng_t)$ contra o inverso do número de neurônios $1/N$ para α fixo.

modo que

$$\Omega_s = \frac{\exp(N \ln 2) - \exp(Ng_t)}{\exp(Ng_t)} \tag{5.10}$$

para todo s. Aqui, $\exp(Ng_t)$ é o número total de atratores. Segue imediatamente dessa hipótese que $\langle Y_2 \rangle \simeq \exp(-Ng_t)$ e, portanto, vai exponencialmente a zero no limite termodinâmico. Utilizando $g_t = g_{pf}$, obtido da figura 5.1 (que é uma boa aproximação no caso da dinâmica seqüencial, pois o número de ciclos é desprezível frente ao número de pontos fixos), apresentamos na figura 5.4 os dados obtidos para o modelo uniforme. A comparação desses dados com os dos modelos originais indicam que, como esperado, o modelo de pesos ótimos com $\kappa = 0$ tem as bacias de atração mais heterogêneas e, embora não haja concordância quantitativa, o modelo uniforme descreve qualitativamente bem os outros modelos. As discrepâncias, é

claro, são devidas à não homogeneidade das bacias de atração. Aqui vale observar que quando as bacias de atração são idênticas, $\langle Y_2 \rangle$ assume seu menor valor.

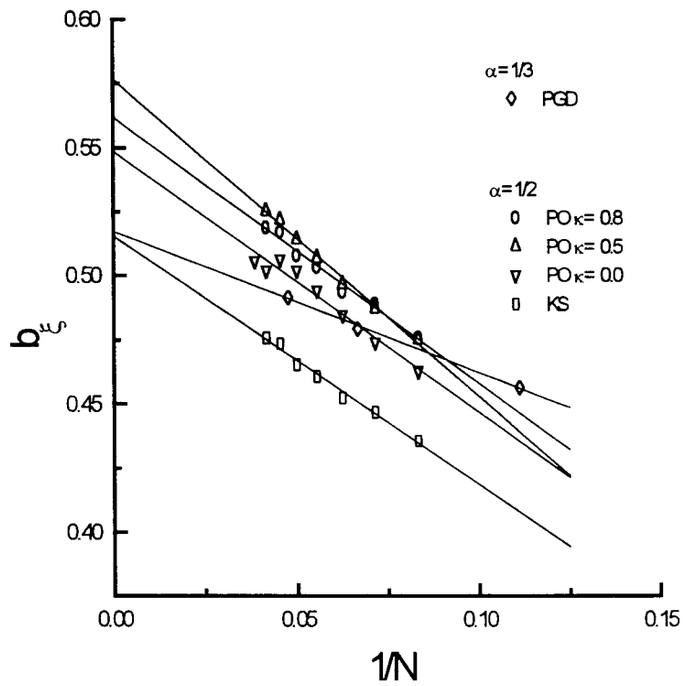

Figura 5.5: A bacia de atração dos estados memorizados pode ser avaliada pela dependência de $b_\xi(\alpha) = \frac{1}{N} \ln \mathcal{N}_\xi$ contra o inverso do número de neurônios $1/N$.

O número médio de estados na bacia de atração de cada um dos padrões memorizados pode ser avaliado pela figura 5.5 onde mostramos $b_\xi = 1/N \ln \langle \mathcal{N}_\xi \rangle$ contra $1/N$. O que surpreende nesta figura é a quase desprezível influência do parâmetro de margem κ sobre as bacias de atração dos padrões memorizados, implicando assim que esse parâmetro não cumpre a função para a qual foi introduzido no modelo. Lembramos que, por outro lado, a análise do modelo extremamente diluído prevê bacias de atração nulas para os padrões com $\kappa = 0$. As possíveis razões dessa discordância serão discutidas no capítulo 7.

Dependência com α para N fixo

Nesta seção, apresentaremos resultados de simulações para $N = 22$ (pesos ótimos e KS) e $N = 21$ (PGD).

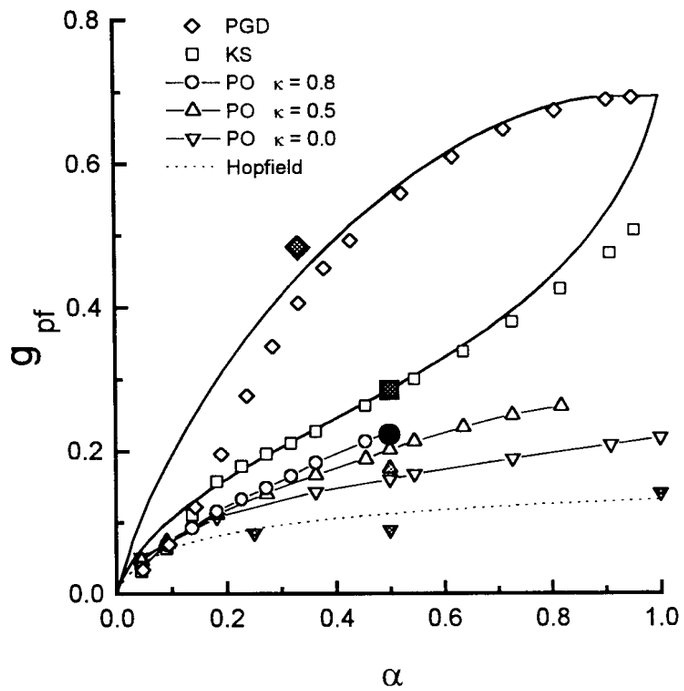

Figura 5.6: Função $g_{pf}(N, \alpha)$ contra α. Aqui, os pontos vazios são os resultados da simulação e os pontos cheios são $g_{\infty}(\alpha)$, a extrapolação de $g_{pf}(N, \alpha)$ no limite $1/N \to 0$, como ilustrado na figura 5.1. Para maior clareza, apenas o maior valor de N simulado é mostrado, a saber $N = 21$, 22 e 22, respectivamente, para os modelos da PGD, KS e PO. A curva tracejada representa a previsão teórica para o modelo de Hopfield [47] [48].

A figura 5.6 mostra $g_{pf}(N, \alpha)$ contra α para os vários modelos considerados. As curvas cheias são os resultados analíticos obtidos para o limite $N \to \infty$ por Kuhlmann e Anlauf [46] para o caso da pseudo-inversa. A discordância entre as

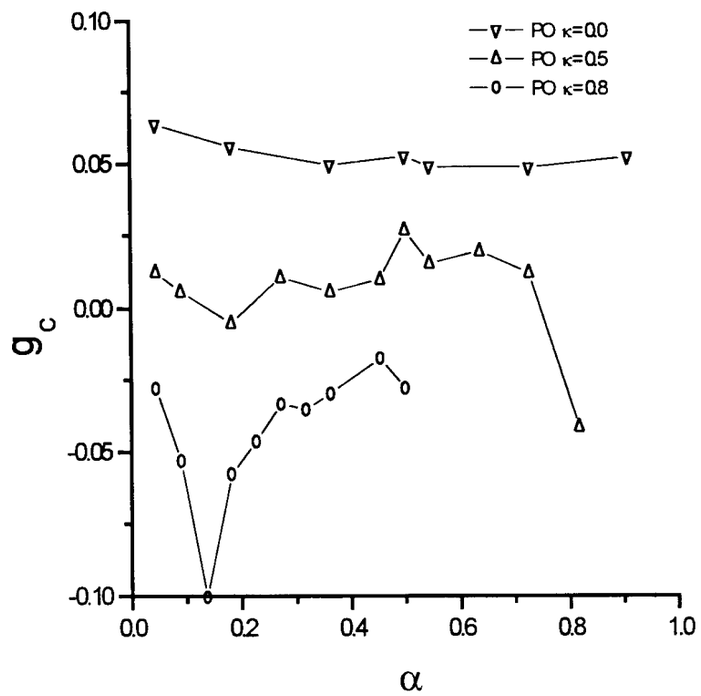

Figura 5.7: Função $g_c(N, \alpha)$ contra α. Lembrar que a dinâmica serial da pseudo-inversa não apresenta ciclos.

simulações e os resultados analíticos para α pequeno já era esperada, uma vez que naqueles cálculos foi utilizado $P \longrightarrow \infty$, enquanto nas simulações o menor valor de α corresponde a $P = 1$. Já a discordância para a formulação KS no caso $\alpha \simeq 1$, deve-se ao fato da capacidade de armazenagem desse modelo ser dada por $\alpha_c = 1 - 1/N$, resultando em $\alpha_c \simeq 0.95$ para $N = 22$, de forma que é esperada alguma discordância próximo a essa região crítica. O resultado $\alpha_c = 1 - 1/N$, embora óbvio para a formulação E da pseudo-inversa, está baseado apenas nos dados das simulações para a formulação KS que indicaram não ser possível encontrar uma matriz de pesos capaz de armazenar $P = N$ padrões aleatórios. Ainda para o modelo da pseudo-inversa, notemos que, no limite $\alpha = 1$, o número médio de pontos fixos tende ao número total de estados, levando a $g_{pf} = \ln 2 \simeq 0.69$. Embora não existam resultados analíticos

para o modelo de pesos ótimos, é claro, pela figura, que o aumento do número médio de pontos fixos com α é muito mais suave nesse modelo. É importante mencionar que medimos também a grandeza $\langle \ln \mathcal{N}_{pf} \rangle$, mas não observamos diferenças significativas em relação aos dados apresentados nas figuras 5.1 e 5.6. A figura 5.7 apresenta $g_c\,(N, \alpha)$ contra α para o modelo de pesos ótimos. No regime em que $g_c > 0$, esta quantidade é praticamente independente de α.

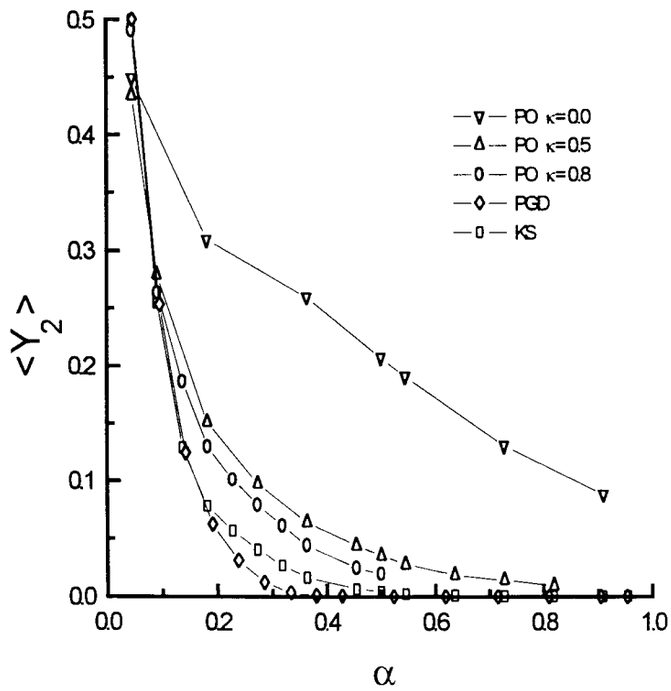

Figura 5.8: $\langle Y_2 \rangle$ contra α.

A figura 5.8, que mostra $\langle Y_2 \rangle$ contra α, ilustra a diminuição do tamanho das bacias de atração dos atratores a medida que α aumenta. Notemos que para $\alpha \longrightarrow 0$ ($P = 1$ nas simulações) temos $\langle Y_2 \rangle = 0.5$ para o modelo da pseudo-inversa, implicando que dinâmica seqüencial possui apenas dois atratores $\boldsymbol{\xi}$ e $-\boldsymbol{\xi}$ dividindo igualitariamente o espaço de configurações. O mesmo não ocorre com os modelos de pesos ótimos, pois, além desses dois pontos fixos, aparecem ciclos como podemos

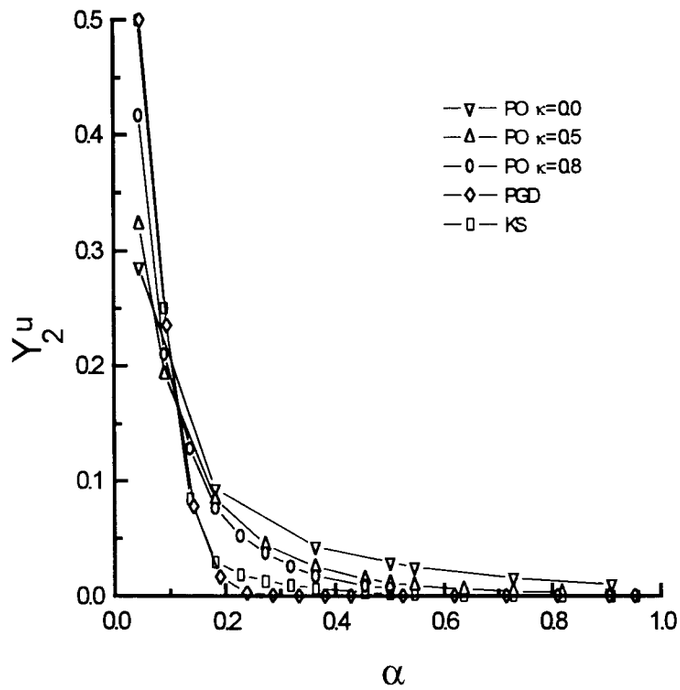

Figura 5.9: Previsão para o segundo momento no caso de bacias uniformes $Y_2^u = \exp(-Ng_t)$ contra α para N fixo.

observar na figura 5.7. Os resultados do modelo uniforme são mostrados na figura 5.9.

Finalmente, a figura 5.10 mostra a estimativa da bacia de atração dos estados memorizados como função de α. Dois aspectos são merecedores de atenção. O primeiro é o fato de no modelo PGD as bacias de atração dos padrões tenderem a zero para $\alpha \geq 0.5$, em concordância com os resultados de Kanter e Sompolinsky [21]. O segundo, que já fora observado na figura 5.5, é a surpreendentemente pequena influência de κ sobre as bacias de atração dos padrões memorizados. Novamente, é interessante observar que, no limite $\alpha \to 0$ ($P = 1$ nas simulações), encontramos $b_\xi = 1/N \ln\left(2^N/2\right) = (1 - 1/N)\ln 2$ para todos os casos, exceto $\kappa = 0$.

Consideraremos, a seguir, a dinâmica paralela. Como demonstrado por Peretto

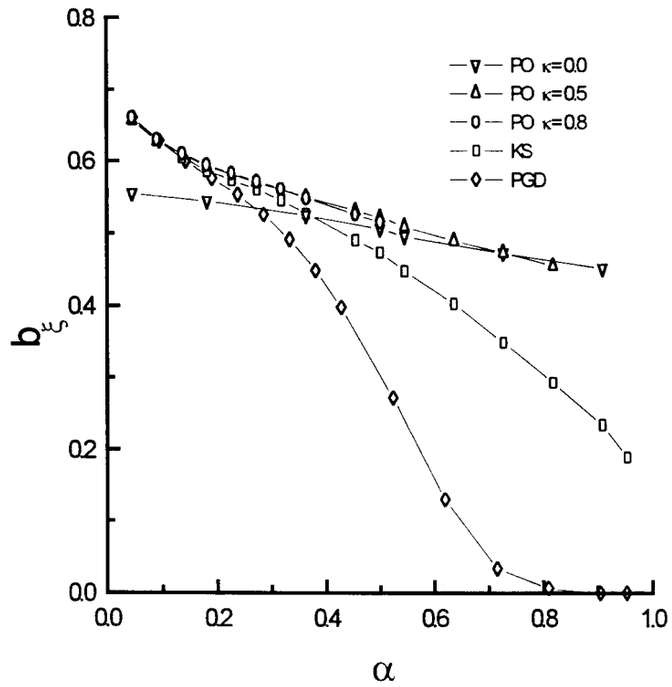

Figura 5.10: Estimativa da bacia de atração dos estados memorizados como função de α.

[49], mesmo com as matrizes de pesos simétricas, a dinâmica paralela pode levar a ciclos de período 2. Não há diferenças relevantes entre o número médio de pontos fixos para as duas dinâmicas, de forma que as figuras 5.1 e 5.6 não são modificadas significativamente. A figura 5.11 ilustra como a dinâmica paralela aumenta enormemente o número de ciclos, exceto para o modelo PGD cujo termo de autointeração positiva inibe o aparecimento de ciclos. Devido a estes novos estados estacionários, a estrutura das bacias de atração é bastante modificada conforme mostra a figura 5.12. O que chama a atenção nesta figura é o comportamento de $\langle Y_2 \rangle$ para o modelo KS na região α próximo a 1: surgem uns poucos atratores, que observamos serem ciclos de período 2, dominando praticamente todo o espaço de configurações. Outro fato curioso com relação ao modelo de pesos ótimos com $\kappa = 0$, mostrado claramente

por estas figuras, é a independência dos ciclos e da estrutura das bacias de atração com α.

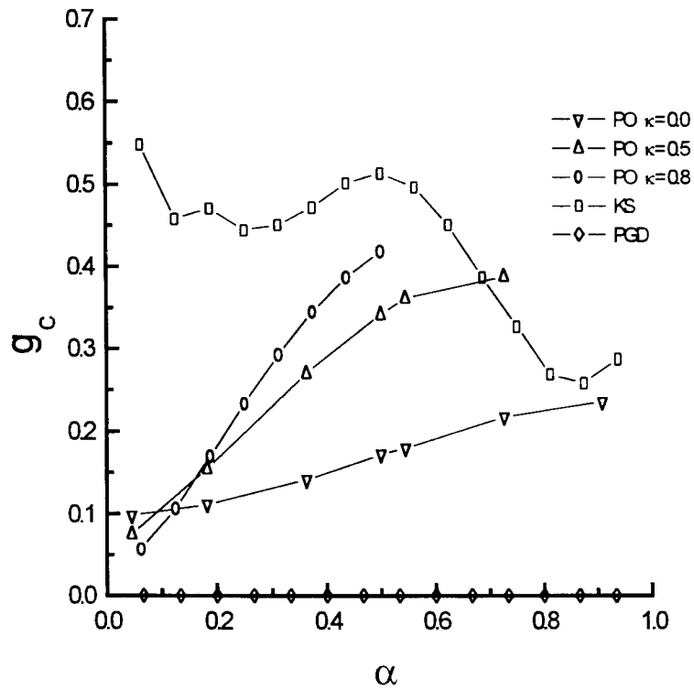

Figura 5.11: Função $g_c(N, \alpha)$ contra α para a dinâmica paralela. $N = 15$ para a PGD e $N = 16$ para os modelos KS e PO.

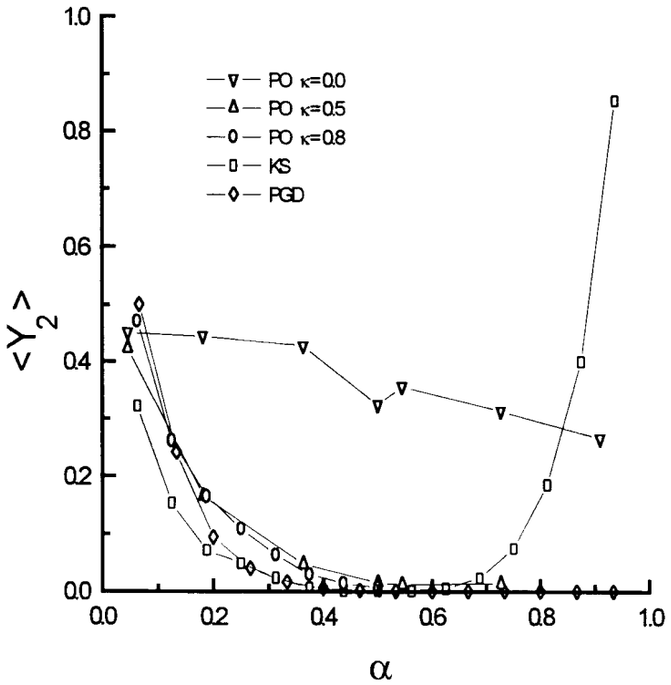

Figura 5.12: $\langle Y_2 \rangle$ contra α para a dinâmica paralela. $N = 15$ para a PGD e $N = 16$ para os modelos KS e PO.

Capítulo 6

Categorização na pseudo-inversa

Neste capítulo, estudaremos o problema da categorização em redes neurais atratoras, como proposto por Fontanari [31]. O problema consiste em treinar uma rede neural com um conjunto de exemplos $\boldsymbol{\xi}^{l\nu}$ dos conceitos $\boldsymbol{\xi}^{l}$, aos quais a rede não tem acesso, com o objetivo de verificar as condições e limites a partir dos quais a rede cria um atrator para estes conceitos. Dados os conceitos, os exemplos são gerados pela distribuição de probabilidade condicional

$$p\left(\xi_i^{l\nu} \mid \xi_i^{l}\right) = \frac{(1+b)}{2}\delta\left(\xi_i^{l\nu} - \xi_i^{l}\right) + \frac{(1-b)}{2}\delta\left(\xi_i^{l\nu} + \xi_i^{l}\right), \tag{6.1}$$

onde $l = 1, \ldots, P$ e $\nu = 1, \ldots, s$, num total de sP exemplos (s exemplos para cada conceito). Os conceitos $\boldsymbol{\xi}^{l}$ são gerados pela mesma distribuição do capítulo 4, equação (4.2),

$$p\left(\xi_i^{k}\right) = \frac{1}{2}\,\delta\left(\xi_i^{k} - 1\right) + \frac{1}{2}\,\delta\left(\xi_i^{k} + 1\right). \tag{6.2}$$

Na seção 6.1, analisamos a estabilidade dos conceitos e na seção 6.2, a estabilidade dos exemplos, em ambos os casos para a formulação E da pseudo-inversa definida pela minimização da função (1.14), como apresentado no capítulo 5. Encerramos o capítulo na seção 6.3, com um estudo da termodinâmica do problema da categorização no modelo KS para o limite P finito. Foi possível estudar a termodinâmica do modelo KS, pois sua matriz de pesos sinápticos é simétrica com os termos de diagonais nulos. Além disso, sua escolha se justifica, pois como vimos no capítulo

5, as formulações KS e E são, dentro dos limites de precisão de nossas simulações, idênticas sob todos os aspéctos considerados.

6.1 Estabilidade de conceitos categorizados através de exemplos

A pseudo-inversa treinada com exemplos é definida pela minimização da energia

$$E_i = \frac{1}{2}\sum_{k\nu}\left(1-\Delta_i^{k\nu}\right)^2, \tag{6.3}$$

onde $\Delta_i^{k\nu} = \frac{1}{N^{1/2}}\xi_i^{k\nu}\sum_{j\neq i} J_{ij}\xi_j^{k\nu}$. A seguir, na seção 6.1.1, obtemos a energia livre e os parâmetros de ordem da pseudo-inversa treinada com exemplos e na seção 6.1.2, a distribuição de probabilidade das estabilidades dos conceitos e sua fração de sítios instáveis.

6.1.1 Cálculo da energia livre

Para obter a energia livre por sítio $f = -\frac{1}{N\lambda}\langle\langle \ln Z\rangle\rangle_{\xi^{k\nu},\xi^k}$, utilizamos, como nos capítulos 2 e 4, o método das réplicas para efetuar as médias sobre as variáveis lentas. Aqui $\langle\langle \ldots \rangle\rangle_{\xi^{k\nu},\xi^k}$ denota a média sobre os exemplos $\xi^{k\nu}$ e sobre os padrões não memorizados ξ^k. A função de partição replicada n vezes $Z^n = \prod_\rho Z^\rho$ se escreve da seguinte maneira

$$Z_i^n = \int\left[\prod_{j\rho} dJ_{ij}^\rho \delta\left(Q_i - \frac{1}{N}\sum_j\left(J_{ij}^\rho\right)^2\right)\right]\exp\left(-\lambda E_i^\rho\right). \tag{6.4}$$

Substituindo a energia (6.3), na expressão para a Z^n, obtemos

$$\begin{aligned} Z_i^n &= \int\left[\prod_{j\rho} dJ_{ij}^\rho \delta\left(Q_i - \frac{1}{N}\sum_j\left(J_{ij}^\rho\right)^2\right)\right] \\ &\prod_{k\nu\rho}\int dx_{k\nu i}^\rho \delta\left(x_{k\nu i}^\rho - \frac{1}{N^{1/2}}\xi_i^{k\nu}\sum_j J_{ij}^\rho \xi_j^{k\nu}\right)\exp\left[-\frac{\lambda}{2}\left(1-x_{k\nu i}^\rho\right)^2\right] \end{aligned} \tag{6.5}$$

Usando a representação integral da função delta, $\delta(\varphi - x) = \frac{1}{2\pi}\int d\widetilde{x}\exp\left(-\mathbf{i}\widetilde{x}\left(\lambda - x\right)\right)$, e escrevendo $x^{\rho}_{k\nu} \equiv x^{\rho}_{k\nu i}$ e $y^{\rho}_{k} \equiv y^{\rho}_{ki}$, pois a função de partição é independente do sítio, obtemos

$$\begin{aligned}
\langle\langle\, Z^n\rangle\rangle_{\xi^{k\nu},\xi^k} &= \int\left[\prod_{j\rho} dJ^{\rho}_j\delta\left(Q_\rho - \frac{1}{N}\sum_j\left(J^{\rho}_j\right)^2\right)\right] \\
&\quad \int\left(\prod_{k\nu\rho}\frac{d\widetilde{x}^{\rho}_{k\nu}dx^{\rho}_{k\nu}}{2\pi}\right)\exp\left[\mathbf{i}\sum_{k\nu\rho}\widetilde{x}^{\rho}_{k\nu}x^{\rho}_{k\nu} - \frac{\lambda}{2}\sum_{k\nu\rho}(1-x^{\rho}_{k\nu})^2\right] \\
&\quad \left\langle\left\langle\exp\left(-\frac{\mathbf{i}}{N^{1/2}}\sum_{k\nu\rho}\widetilde{x}^{\rho}_{k\nu}\xi^{k\nu}_i\sum_j J^{\rho}_j\xi^{k\nu}_j\right)\right\rangle\right\rangle_{\xi^{k\nu},\xi^k}. \qquad (6.6)
\end{aligned}$$

Fazendo a média sobre os exemplos $\xi^{k\nu}_j$ $(j \neq i)$, obtemos

$$\begin{aligned}
\langle\langle\ldots\rangle\rangle_{\xi^{k\nu},\xi^k} &= \prod_{kj\nu}\left\langle\left\langle\exp\left(-\frac{\mathbf{i}}{N^{1/2}}\xi^{k\nu}_i\xi^{k\nu}_j\sum_\rho\widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right)\right\rangle\right\rangle_{\xi^{k\nu},\xi^k} \qquad (6.7) \\
&= \prod_{kj\nu}\left\langle\cos\left(\frac{1}{N^{1/2}}\sum_\rho\widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right)\right. \\
&\qquad \left. -\mathbf{i}b\xi^{k\nu}_i\xi^k_j\sin\left(\frac{1}{N^{1/2}}\sum_\rho\widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right)\right\rangle_{\xi^k}. \qquad (6.8)
\end{aligned}$$

Expandindo em potências de $\frac{1}{N^{1/2}}$, obtemos

$$\begin{aligned}
\langle\langle\ldots\rangle\rangle_{\xi^{k\nu},\xi^k} &\simeq \prod_{kj\nu}\left\langle\exp\ln\left[1 - \frac{1}{2N}\left(\sum_\rho\widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right)^2\right.\right. \\
&\qquad \left.\left. -\frac{\mathbf{i}b}{N^{1/2}}\xi^{k\nu}_i\xi^k_j\left(\sum_\rho\widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right)\right]\right\rangle_{\xi^k} \qquad (6.9) \\
&\simeq \prod_{kj\nu}\left\langle\exp\left[-\frac{(1-b^2)}{2N}\left(\sum_\rho\widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right)^2\right.\right. \\
&\qquad \left.\left. -\frac{\mathbf{i}b}{N^{1/2}}\xi^{k\nu}_i\xi^k_j\left(\sum_\rho\widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right)\right]\right\rangle_{\xi^k}. \qquad (6.10)
\end{aligned}$$

Fazendo a média sobre os conceitos ξ^k_j $(j \neq i)$, obtemos

$$\begin{aligned}
\langle\langle\ldots\rangle\rangle_{\xi^{k\nu},\xi^k} &\simeq \prod_{kj}\left\langle\exp\left[-\frac{\mathbf{i}b}{N^{1/2}}\xi^k_j\sum_{\rho\nu}\xi^{k\nu}_i\ \widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right]\right\rangle_{\xi^k} \qquad (6.11) \\
&\simeq \prod_{kj}\cos\left[\frac{b}{N^{1/2}}\sum_{\rho\nu}\xi^k_i\xi^{k\nu}_i\ \widetilde{x}^{\rho}_{k\nu}J^{\rho}_j\right]. \qquad (6.12)
\end{aligned}$$

Expandindo em potências de $\frac{1}{N^{1/2}}$, obtemos

$$\langle\langle \ldots \rangle\rangle_{\xi^{k\nu},\xi^k} \simeq \prod_{kj} \exp\left\{ -\frac{1}{2N} \left[b \sum_{\rho\nu} J_j^\rho \xi_i^k \xi_i^{k\nu} \; \tilde{x}_{k\nu}^\rho \right]^2 \right\}. \tag{6.13}$$

Finalmente, o termo mediado fica dado por

$$\begin{aligned} \langle\langle \ldots \rangle\rangle_{\xi^{k\nu},\xi^k} \simeq \prod_{kj} \exp\Bigg\{ & -\frac{(1-b^2)}{2N} \sum_\nu \left(\sum_\rho J_j^\rho \tilde{x}_{k\nu}^\rho \right)^2 \\ & -\frac{1}{2N} \left[b \sum_{\rho\nu} J_j^\rho \xi_i^k \xi_i^{k\nu} \; \tilde{x}_{k\nu}^\rho \right]^2 \Bigg\}, \end{aligned} \tag{6.14}$$

onde fica explícita a dependência no sítio i através do termo $\xi_i^k \xi_i^{k\nu}$.

Reescrevendo o quadrado do somatório com o auxílio da identidade $\left(\sum_\rho x\right)^2 = \sum_\rho x^2 + 2\sum_{\rho<\sigma} x_\rho x_\sigma$ e introduzindo a variável auxiliar $q_{\rho\sigma} = \frac{1}{N}\sum_j J_j^\rho J_j^\sigma$, a função de partição fica dada por

$$\begin{aligned} \langle\langle Z^n \rangle\rangle = & \int \left(\prod_{j\rho} dJ_{ij}^\rho \delta \left(Q - \frac{1}{N} \sum_j \left(J_j^\rho \right)^2 \right) \right) \\ & \int \left(\prod_{\rho<\sigma} dq_{\rho\sigma} \delta \left(q_{\rho\sigma} - \frac{1}{N} \sum_j J_j^\rho J_j^\sigma \right) \right) \int \left(\prod_{k\nu\rho} \frac{d\tilde{x}_{k\nu}^\rho dx_{k\nu}^\rho}{2\pi} \right) \\ & \exp \Bigg[\mathbf{i} \sum_{k\nu\rho} \tilde{x}_{k\nu}^\rho x_{k\nu}^\rho - \frac{\lambda}{2} \sum_{k\nu\rho} (1 - x_{k\nu}^\rho)^2 \\ & -\frac{(1-b^2)}{2} \sum_{k\nu\rho} Q_\rho \left(\tilde{x}_{k\nu}^\rho \right)^2 - \left(1 - b^2\right) \sum_{k\nu} \left(\sum_{\rho<\sigma} q_{\rho\sigma} \tilde{x}_{k\nu}^\rho \tilde{x}_{k\nu}^\sigma \right) \\ & -\frac{1}{2} \sum_{k\rho} Q_\rho \left(b \sum_\nu \xi_i^k \xi_i^{k\nu} \; \tilde{x}_{k\nu}^\rho \right)^2 \\ & - \sum_{k\rho<\sigma} q_{\rho\sigma} \left(b \sum_\nu \xi_i^k \xi_i^{k\nu} \; \tilde{x}_{k\nu}^\rho \right) \left(b \sum_\nu \xi_i^k \xi_i^{k\nu} \; \tilde{x}_{k\nu}^\sigma \right) \Bigg]. \end{aligned} \tag{6.15}$$

Usando novamente a representação integral da delta nas duas primeiras integrais e fazendo as mudanças de variáveis $\tilde{Q}_\rho = \frac{\widehat{Q}_\rho}{\mathbf{i}/N}$ e $\tilde{q}_{\rho\sigma} = \frac{\widehat{q}_{\rho\sigma}}{\mathbf{i}/N}$, a função de partição fica dada por

$$\begin{aligned} \langle\langle Z^n \rangle\rangle \simeq & \int \left(\prod_\rho \frac{d\tilde{Q}_\rho}{2\pi\mathbf{i}/N} \right) \int \left(\prod_{\rho<\sigma} \frac{dq_{\rho\sigma} d\tilde{q}_{\rho\sigma}}{2\pi\mathbf{i}/N} \right) \exp \Bigg[N \sum_\rho \tilde{Q}_\rho Q \\ & -N \sum_{\rho<\sigma} q_{\rho\sigma} q_{\rho\sigma} + NG_0 + \alpha N G_1 \Bigg] \end{aligned} \tag{6.16}$$

com

$$G_0 = \ln\left\{\int\left(\prod_\rho dJ_{ij}^\rho\right)\exp\left[-\sum_\rho \tilde{Q}_\rho\left(J_j^\rho\right)^2 - \sum_{\rho<\sigma}\tilde{q}_{\rho\sigma}J_j^\rho J_j^\sigma\right]\right\} \tag{6.17}$$

e

$$\begin{aligned} G_1 \;=\; \Bigg\langle \ln\Bigg\{ \int\left(\prod_{\nu\rho}\frac{d\tilde{x}_\nu^\rho dx_\nu^\rho}{2\pi}\right)\exp\Bigg[\mathbf{i}\sum_{\nu\rho}\tilde{x}_\nu^\rho x_\nu^\rho \\ -\frac{\lambda}{2}\sum_{\nu\rho}(1-x_\nu^\rho)^2 - \frac{(1-b^2)}{2}Q\sum_{\nu\rho}(\tilde{x}_\nu^\rho)^2 \\ -\left(1-b^2\right)\sum_\nu\left(\sum_{\rho<\sigma}q_{\rho\sigma}\tilde{x}_\nu^\rho\tilde{x}_\nu^\sigma\right) - \frac{1}{2}Q\sum_\rho\left(b\sum_\nu\xi_i\xi_i^\nu\;\tilde{x}_\nu^\rho\right)^2 \\ -\sum_{\rho<\sigma}q_{\rho\sigma}\left(b\sum_\nu\xi_i\xi_i^\nu\;\tilde{x}_\nu^\rho\right)\left(b\sum_\nu\xi_i\xi_i^\nu\;\tilde{x}_\nu^\sigma\right)\Bigg]\Bigg\}\Bigg\rangle_{\xi_i^\nu,\xi_i} . \end{aligned} \tag{6.18}$$

Na derivação de G_1, foi utilizada a propriedade de automediância, $\sum_k g\left(\xi_i^k\xi_i^{k\nu}\right) = \alpha N\left\langle\left\langle g\left(\xi_i\xi_i^\nu\right)\right\rangle\right\rangle_{\xi_i^\nu,\xi_i}$, que resulta na eliminação do índice k. No limite $N \to \infty$, as integrais em (6.16) são efetuadas pelo método do ponto de sela e a expressão para a energia livre por sítio pode ser escrita como

$$\begin{aligned} -\lambda f \;=\; \lim_{n\to 0}\frac{1}{nN}\,\mathrm{extr}\Bigg[N\sum_\rho\tilde{Q}_\rho Q - N\sum_{\rho<\sigma}\tilde{q}_{\rho\sigma}q_{\rho\sigma} \\ +NG_0 + \alpha NG_1\Bigg], \end{aligned} \tag{6.19}$$

onde o extr é tomado em relação aos parâmetros de ponto de sela $\tilde{Q}_\rho$, $\tilde{q}_{\rho\sigma}$ e $q_{\rho\sigma}$. O próximo passo é assumir simetria de réplicas nos parâmetros de ponto de sela, $\tilde{Q}_\rho = \tilde{Q}$, $\tilde{q}_{\rho\sigma} = \tilde{q}$ e $q_{\rho\sigma} = q$ $(\rho < \sigma)$, o que facilita enormemente o cálculo analítico de G_1 e G_0.

Cálculo de G_1

O que torna o cálculo de G_1 complicado é o usual acoplamento entre as diferentes réplicas, aliado agora ao acoplamento entre diferentes exemplos de um mesmo conceito. Utilizando a identidade (2.11) e a transformação gaussiana, podemos desacoplar as diferentes réplicas no termo $\sum_{\rho<\sigma} q_{\rho\sigma}\tilde{x}_\nu^\rho\tilde{x}_\nu^\sigma$, de modo que a equação (6.18)

fica escrita como

$$
\begin{aligned}
G_1 \;=\; & \Bigg\langle \ln \Bigg\{ \int \left(\prod_{\nu} Dt_\nu \right) \int Dz \int \left(\prod_{\nu\rho} \frac{d\widetilde{x}_\nu^\rho dx_\nu^\rho}{2\pi} \right) \\
& \exp \Bigg[\mathbf{i} \sum_{\nu\rho} \widetilde{x}_\nu x_\nu - \frac{\lambda}{2} \sum_{\nu\rho} (1 - x_\nu)^2 \\
& - \frac{(1-b^2)}{2} (Q-q) \sum_{\nu\rho} (\widetilde{x}_\nu)^2 \;\; - \frac{(Q-q)}{2} \sum_{\rho} \left(b \sum_{\nu} \xi_i \xi_i^\nu \; \widetilde{x}_\nu \right)^2 \\
& - \mathbf{i} \sqrt{(1-b^2)\, q} \sum_{\nu\rho} \widetilde{x}_\nu t_\nu - \mathbf{i} \sqrt{q} \sum_{\rho} \left(b \sum_{\nu} \xi_i \xi_i^\nu \; \widetilde{x}_\nu \right) z \Bigg] \Bigg\} \Bigg\rangle_{\xi_i^\nu, \xi_i} . \qquad (6.20)
\end{aligned}
$$

O argumento da função logaritmo pode ser reescrito como

$$
\begin{aligned}
& \int \left(\prod_{\nu} Dt_\nu \right) \int Dz \prod_{\rho} \Bigg\{ \int \frac{d\widetilde{x} dx}{2\pi} \exp \Bigg[\mathbf{i} \sum_{\nu} \widetilde{x}_\nu x_\nu - \frac{\lambda}{2} \sum_{\nu} (1 - x_\nu)^2 \\
& - \frac{(1-b^2)}{2} (Q-q) \sum_{\nu} (\widetilde{x}_\nu)^2 - \frac{(Q-q)}{2} \left(b \sum_{\nu} \xi_i \xi_i^\nu \; \widetilde{x}_\nu \right)^2 \\
& \qquad - \mathbf{i} \sqrt{(1-b^2)\, q} \sum_{\nu} \widetilde{x}_\nu t_\nu - \mathbf{i} \sqrt{q} \left(b \sum_{\nu} \xi_i \xi_i^\nu \; \widetilde{x}_\nu \right) z \Bigg] \Bigg\} = \\
= \; & \int \left(\prod_{\nu} Dt_\nu \right) \int Dz \, \{ ... \}^n \\
= \; & \int \left(\prod_{\nu} Dt_\nu \right) \int Dz \exp n \ln \{ ... \} \\
\simeq \; & \int \left(\prod_{\nu} Dt_\nu \right) \int Dz \, [1 + n \ln \{ ... \}] \\
\simeq \; & 1 + n \int \left(\prod_{\nu} Dt_\nu \right) \int Dz \ln \{ ... \} , \qquad (6.21)
\end{aligned}
$$

onde tomamos o limite $n \longrightarrow 0$. Daí,

$$
\begin{aligned}
\frac{G_1}{n} \;\simeq\; & \Bigg\langle \int \left(\prod_{\nu} Dt_\nu \right) \int Dz \ln \int \left(\prod_{\nu} \frac{d\widetilde{x}_\nu dx_\nu}{2\pi} \right) \\
& \exp \Bigg[\mathbf{i} \sum_{\nu} \widetilde{x}_\nu x_\nu - \frac{\lambda}{2} \sum_{\nu} (1 - x_\nu)^2 \\
& - \frac{(1-b^2)}{2} (Q-q) \sum_{\nu} (\widetilde{x}_\nu)^2 - \frac{(Q-q)}{2} \left(b \sum_{\nu} \xi_i \xi_i^\nu \; \widetilde{x}_\nu \right)^2 \\
& - \mathbf{i} \sqrt{(1-b^2)\, q} \sum_{\nu} \widetilde{x}_\nu t_\nu - \mathbf{i} \sqrt{q} \left(b \sum_{\nu} \xi_i \xi_i^\nu \; \widetilde{x}_\nu \right) z \Bigg] \Bigg\rangle_{\xi_i^\nu, \xi_i} . \qquad (6.22)
\end{aligned}
$$

Utilizando a transformação gaussiana para desacoplar os exemplos no termo quadrático $\left(\sum_\nu \xi_i \xi_i^\nu \ \tilde{x}_\nu\ \right)^2$, podemos facilmente efetuar as integrais sobre x_ν e $\tilde{x}_\nu$, resultando

$$\frac{G_1}{n} = \left\langle \int \left(\prod_\nu Dt_\nu\right) \int Dz \ln \int D\eta \prod_\nu \sqrt{\frac{1}{\lambda\left(1-b^2\right)(Q-q)+1}} \right.$$
$$\left. \exp\left(-\frac{\lambda\left(1-A_\nu\ \right)^2}{2\left(1+\lambda\left(1-b^2\right)(Q-q)\right)}\right)\right\rangle_{\xi_i^\nu,\xi_i}, \quad (6.23)$$

onde $A_\nu = \sqrt{(Q-q)\,b^2}\xi_i\xi_i^\nu\ \eta + \ \sqrt{\left(1-b^2\right)q}t_\nu + \sqrt{q}\ b\xi_i\xi_i^\nu\ z$.

A integral em η é gaussiana e pode ser calculada com facilidade, levando a

$$\frac{G_1}{n} = \left\langle \int \left(\prod_\nu Dt_\nu\right) \int Dz \left\{-\frac{(s-1)}{2}\ln\left[1+\lambda(Q-q)\left(1-b^2\right)\right]\right.$$
$$-\frac{1}{2}\ln\left[1+\lambda(Q-q)\left(1-b^2+sb^2\right)\right] - \frac{\lambda\sum_\nu\left(1-\ B_\nu\ \right)^2}{2\left(1+\lambda\left(1-b^2\right)(Q-q)\right)}$$
$$\left.+\frac{\lambda^2(Q-q)\,b^2\left[\sum_\nu\left(1-\ B_\nu\ \right)\xi_i\xi_i^\nu\ \right]^2}{2\left[1+\lambda\left(1-b^2\right)(Q-q)\right]\left[1+\lambda(Q-q)\left(1-b^2+sb^2\right)\right]}\right\}\right\rangle_{\xi_i^\nu,\xi_i}, \quad (6.24)$$

onde $B_\nu = \sqrt{\left(1-b^2\right)q}t_\nu + \sqrt{q}\ b\xi_i\xi_i^\nu\ z$. Novamente, efetuando as integrais gaussianas em z e t_ν e usando $\left\langle \sum_{\nu\neq\mu}\xi_i^\nu\xi_i^\mu\right\rangle_{\xi_i^\nu,\xi_i} = s\,(s-1)\,b^2$, obtemos

$$\frac{G_1}{n} = -\frac{(s-1)}{2}\ln\left[1+\lambda\left(1-b^2\right)(Q-q)\right]$$
$$-\frac{1}{2}\ln\left[1+\lambda(Q-q)\left(1-b^2+sb^2\right)\right] - \frac{\lambda s\,(1+q)}{2\left(1+\lambda\left(1-b^2\right)(Q-q)\right)}$$
$$+\frac{\lambda^2(Q-q)\,b^2 s\left(1-b^2+sb^2\right)(1+q)}{2\left(1+\lambda\left(1-b^2\right)(Q-q)\right)\left(1+\lambda(Q-q)\left(1-b^2+sb^2\right)\right)}. \quad (6.25)$$

Cálculo de G_0

O cálculo de G_0 é idêntico ao do capítulo 2, equação (2.23), e leva à seguinte expressão

$$G_0 = \ln\left\{1+\frac{n}{2}\ln\pi - \frac{n}{2}\ln\left(\tilde{Q}-\frac{\tilde{q}}{2}\right) - n\int Dz\frac{\tilde{q}z^2}{4\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)}\right\} =$$
$$= \frac{n}{2}\ln\pi - \frac{n}{2}\ln\left(\tilde{Q}-\frac{\tilde{q}}{2}\right) - n\frac{\tilde{q}}{4\left(\tilde{Q}-\frac{\tilde{q}}{2}\right)}\ . \quad (6.26)$$

Para finalizar o cálculo da energia livre por sítio, observamos que, devido à simetria de réplicas, o argumento da primeira exponencial da equação (6.16) reduz-se à equação

$$nN\tilde{Q}Q + n(n-1)N\frac{\tilde{q}q}{2}, \tag{6.27}$$

o que leva à seguinte expressão final para a energia livre por sítio

$$\begin{aligned}
-\lambda f \quad = \quad & \tilde{Q}Q - \frac{\tilde{q}q}{2} + \frac{1}{2}\ln\pi - \frac{1}{2}\ln\left(\tilde{Q} - \frac{\tilde{q}}{2}\right) - \frac{\tilde{q}}{4\left(\tilde{Q} - \frac{\tilde{q}}{2}\right)} \\
& -\alpha\frac{(s-1)}{2}\ln\left[1 + \lambda(Q-q)\left(1-b^2\right)\right] \\
& -\alpha\frac{1}{2}\ln\left[1 + \lambda(Q-q)\left(1-b^2+sb^2\right)\right] \\
& -\alpha\frac{\lambda s(1+q)}{2\left(1+\lambda(1-b^2)(Q-q)\right)} \\
& +\alpha\frac{\lambda^2(Q-q)b^2\left[1-b^2+sb^2\right](1+q)s}{2\left(1+\lambda(Q-q)(1-b^2)\right)\left(1+\lambda(Q-q)(1-b^2+sb^2)\right)}.
\end{aligned} \tag{6.28}$$

Cálculo dos parâmetros de ponto de sela

Para obter os valores dos parâmetros de ponto de sela, partimos das equações $\partial f/\partial\tilde{q} = 0$ e $\partial f/\partial\tilde{Q} = 0$, de onde obtemos

$$\tilde{q} = -\frac{q}{(Q-q)^2} \tag{6.29}$$

e

$$\tilde{Q} = \frac{Q-2q}{2(Q-q)^2}. \tag{6.30}$$

Substituindo na expressão para f, obtemos

$$\begin{aligned}
-\lambda f \quad = \quad & \frac{1}{2} + \frac{1}{2}\ln\pi - \frac{1}{2}\ln\left(\frac{1}{2}\right) + \frac{1}{2}\ln(Q-q) + \frac{q}{2(Q-q)} \\
& -\alpha\frac{(s-1)}{2}\ln\left[1 + \lambda(Q-q)\left(1-b^2\right)\right] \\
& -\alpha\frac{1}{2}\ln\left[1 + \lambda(Q-q)\left(1-b^2+sb^2\right)\right] \\
& -\alpha\frac{\lambda s(1+q)}{2\left[1+\lambda(1-b^2)(Q-q)\right]} \\
& +\alpha\frac{\lambda^2(Q-q)b^2\left[1-b^2+sb^2\right](1+q)s}{2\left[1+\lambda(Q-q)(1-b^2)\right]\left[1+\lambda(Q-q)(1-b^2+sb^2)\right]}.
\end{aligned} \tag{6.31}$$

Tomando o limite $\lambda \longrightarrow \infty$, com $Q > q$, na equação $\partial f/\partial q = 0$, obtemos

$$q = \frac{\alpha s}{1 - \alpha s}. \tag{6.32}$$

Neste limite, a rede apresenta energia de treinamento nula, isto é, os padrões são mínimos globais da energia e dizemos que a rede está abaixo da capacidade de armazenamento $\alpha \leq \alpha_c$.

Agora, tomamos o limite $\lambda \longrightarrow \infty$, com $\lambda (Q - q) = x$, na equação (6.31) o que leva à seguinte expressão para a energia de treinamento

$$\begin{aligned} \varepsilon_T &= -\frac{Q}{2x} + \alpha \frac{s(1+Q)}{2[1+x(1-b^2)]} \\ &\quad -\alpha \frac{xb^2(1-b^2+sb^2)(1+Q)s}{2[1+x(1-b^2)][1+x(1-b^2+sb^2)]}. \end{aligned} \tag{6.33}$$

Da equação $\partial \varepsilon_T/\partial Q = 0$, obtemos a expressão para x

$$x^2\left(1-b^2+sb^2\right)\left(1-b^2\right)(\alpha s - 1) + x\left[\alpha s - 2\left(1-b^2\right) - sb^2\right] - 1 = 0\ . \tag{6.34}$$

Podemos isolar αs da equação acima, o que leva à seguinte expressão

$$\alpha s = \frac{[1+x(1-b^2)][1+x(1-b^2+sb^2)]}{x[1+x(1-b^2+sb^2)(1-b^2)]}\ . \tag{6.35}$$

Tomando o limite $x \longrightarrow \infty$, obtemos a capacidade de armazenamento $\alpha_c = 1/s$, como esperado.

Substituindo αs na equação $\partial \varepsilon_T/\partial x = 0$, a expressão para Q fica dada por

$$Q = \frac{xC}{1 - xC}\ , \tag{6.36}$$

onde

$$C = \frac{[1+x(1-b^2)(1-b^2+sb^2)]^2 + b^4(s-1)}{[1+x(1-b^2+sb^2)(1-b^2)][1+x(1-b^2)][1+x(1-b^2+sb^2)]}. \tag{6.37}$$

No limite $b = 0$, obtemos $C = \frac{1}{1+x}$, de forma que $Q = x$.

6.1.2 Cálculo da distribuição de probabilidade das estabilidades dos conceitos

Para calcular a distribuição de probabilidade das estabilidades dos conceitos $W^c(\gamma) = \left\langle\left\langle\left\langle\delta\left(\gamma - \Lambda^l\right)\right\rangle_J\right\rangle\right\rangle_{\xi^{l\nu},\xi^l}$, equação (1.29), acrescentamos à energia (6.3) o termo linear em h,

$$E_i = \frac{1}{2}\sum_{k\nu}\left(1 - \Delta_i^{k\nu}\right)^2 + \frac{h}{P}\sum_k \delta\left(\gamma - \Lambda_i^k\right), \tag{6.38}$$

onde $\Lambda_i^k = \frac{1}{N^{1/2}}\xi_i^k \sum_{j\neq i} J_{ij}\xi_j^k$ e $\Delta_i^{k\nu} = \frac{1}{N^{1/2}}\xi_i^{k\nu}\sum_{j\neq i} J_{ij}\xi_j^{k\nu}$. Daí,

$$W^c(\gamma) = -\lim_{\lambda\to\infty}\frac{1}{\lambda}\frac{\partial\left\langle\left\langle\ln Z_i\right\rangle\right\rangle_{\xi^{l\nu},\xi^l}}{\partial h}|_{h=0}, \tag{6.39}$$

com a função de partição dada por

$$Z_i(h,\gamma) = \int\left[\prod_j dJ_{ij}\delta\left(Q_i - \frac{1}{N}\sum_j (J_{ij})^2\right)\right]\exp\left[-\lambda E_i(\mathbf{J},h,\gamma)\right]. \tag{6.40}$$

Aqui, h é uma variável auxiliar cujo significado físico é irrelevante para nossa análise. O primeiro termo da energia garante a solução de (1.13) e o segundo termo é utilizado para obtenção da distribuição de estabilidades dos conceitos. Devido à equação (6.39), o procedimento para cálculo de $W^c(\gamma)$ é análogo ao empregado no cálculo da distribuição de probabilidade das estabilidades de um padrão de teste, apresentado no capítulo 4. Como o termo G_0 não muda, basta recalcular G_1 a partir da nova energia (6.38), o que leva à seguinte equação (análoga à eq. (6.23))

$$\begin{aligned}\frac{G_1}{n} = \Bigg\langle\int\left(\prod_\nu Dt_\nu\right)\int Dz\ln\int D\eta \\ \int dy\delta\left(y - \sqrt{q}\,z - \sqrt{(Q-q)}\eta\right)\exp\left(-\lambda h\delta(\gamma - y)\right) \\ \prod_\nu\sqrt{\frac{1}{\lambda(1-b^2)(Q-q)+1}}\exp\left(-\frac{\lambda(1-A_\nu)^2}{2(1+\lambda(1-b^2)(Q-q))}\right)\Bigg\rangle_{\xi_i^\nu,\xi_i}\end{aligned} \tag{6.41}$$

onde $A_\nu = \sqrt{(Q-q)\,b^2}\xi_i\xi_i^\nu\,\eta + \sqrt{(1-b^2)\,q}t_\nu + \sqrt{q}\,b\xi_i\xi_i^\nu\,z$. Assim,

$$W^c(\gamma) = -\lim_{\lambda\to\infty}\frac{1}{\lambda n}\frac{\partial G_1}{\partial h}|_{h=0}. \tag{6.42}$$

Notando que as integrais que aparecem em (6.42) são todas gaussianas e lembrando também dos resultados do capítulo 2, equação (2.48), e do capítulo 4, equação (4.20), concluímos que a distribuição de probabilidade das estabilidades dos conceitos pode ser escrita como

$$W^c(\gamma) = \left\langle w^c_{\xi_i^\nu,\xi_i}(\gamma)\right\rangle_{\xi_i^\nu,\xi_i}, \tag{6.43}$$

onde $w^c_{\xi_i^\nu,\xi_i}(\gamma)$ é uma distribuição gaussiana, dada por

$$w^c_{\xi_i^\nu,\xi_i}(\gamma) = \sqrt{\frac{1}{2\pi\delta^2_{\xi_i^\nu,\xi_i}}}\exp\left[-\frac{1}{2}\left(\frac{\gamma-\overline{\gamma}_{\xi_i^\nu,\xi_i}}{\delta_{\xi_i^\nu,\xi_i}}\right)^2\right], \tag{6.44}$$

para ξ_i^ν e ξ_i $(\nu = 1,\ldots,s)$ dados. Com isso em mente, podemos calcular somente os dois primeiros momentos de $w^c_{\xi_i^\nu,\xi_i}(\gamma)$ e reconstituir a distribuição de estabilidades desejada. Para tanto, basta substituirmos $\delta(\gamma - y)$ por y e y^2 na equação (6.41) e eliminarmos as médias sobre ξ_i e ξ_i^ν.

Cálculo do primeiro momento

Substituindo $\delta(\gamma - y)$ por y na equação (6.41) e seguindo o procedimento de cálculo do capítulo 2, obtemos

$$\begin{aligned}\frac{(G_1^p)_{\xi_i^\nu,\xi_i}}{n} &= \int\left(\prod_\nu Dt_\nu\right)\int Dz\left\{-\frac{(s-1)}{2}\ln\left[1+\lambda\left(1-b^2\right)(Q-q)\right]\right.\\
&-\frac{1}{2}\ln\left[1+\lambda(Q-q)\left(1-b^2+sb^2\right)\right]-\lambda h\sqrt{q}\,z\\
&-\frac{s\lambda}{2\left[1+\lambda(1-b^2)(Q-q)\right]}+\frac{2\lambda\sum_\nu B_\nu-\lambda\sum_\nu(B_\nu)^2}{2\left[1+\lambda(1-b^2)(Q-q)\right]}\\
&\left.+\frac{\lambda^2(Q-q)\left[h\left(1+\lambda(1-b^2)(Q-q)\right)-b\sum_\nu\xi_i\xi_i^\nu\,(1-B_\nu)\right]^2}{2\left[1+\lambda(Q-q)(1-b^2+sb^2)\right]\left[1+\lambda(1-b^2)(Q-q)\right]}\right\},\end{aligned} \tag{6.45}$$

onde $B_\nu = \sqrt{(1-b^2)\,q}\,t_\nu + \sqrt{q}\,b\xi_i\xi_i^\nu\,z$.

O primeiro momento, para dados ξ_i^ν e ξ_i, obtém-se de

$$\overline{\gamma}_{\xi_i^\nu,\xi_i} = -\frac{1}{\lambda n}\left[\frac{\partial(G_1^p)_{\xi_i^\nu,\xi_i}}{\partial h}\right]_{h=0}. \tag{6.46}$$

resultando

$$\overline{\gamma}_{\xi_i^\nu,\xi_i} = \int\left(\prod_\nu Dt_\nu\right)\int Dz\,\{\sqrt{q}\,z + \frac{\lambda\,(Q-q)\,b\sum_\nu \xi_i\xi_i^\nu}{[1+\lambda\,(Q-q)\,(1-b^2+sb^2)]} - \frac{\lambda\,(Q-q)\,b\sum_\nu \xi_i\xi_i^\nu\;B_\nu}{[1+\lambda\,(Q-q)\,(1-b^2+sb^2)]}\Bigg\}. \tag{6.47}$$

O primeiro e o terceiro termos são lineares nas variáveis de integração e, portanto, não contribuem. Sobrevive apenas o segundo termo, resultando

$$\overline{\gamma}_{\xi_i^\nu,\xi_i} = \frac{\lambda\,(Q-q)\,b\sum_\nu \xi_i\xi_i^\nu}{1+\lambda\,(Q-q)\,(1-b^2+sb^2)}. \tag{6.48}$$

Agora, devemos tomar o limite $\lambda \to \infty$ das duas maneiras discutidas anteriormente. Para $\alpha \leq \alpha_c = 1/s$, obtemos a seguinte expressão para o primeiro momento

$$\overline{\gamma}_{\xi_i^\nu,\xi_i} = \frac{b\sum_\nu \xi_i\xi_i^\nu}{(1-b^2+sb^2)} \tag{6.49}$$

e para $\alpha > \alpha_c = 1/s$, onde $\lambda\,(Q-q) = x$, obtemos

$$\overline{\gamma}_{\xi_i^\nu,\xi_i} = \frac{xb\sum_\nu \xi_i\xi_i^\nu}{1+x\,(1-b^2+sb^2)}. \tag{6.50}$$

Aqui, x é dado pela solução de (6.34). Note que para $x \to \infty$ ($\alpha = \alpha_c$) os dois limites coincidem.

Cálculo do segundo momento

Substituindo $\delta\,(\gamma - y)$ por y^2 na equação (6.41) e seguindo o procedimento de cálculo do capítulo 2, obtemos

$$\begin{aligned}\frac{(G_1^s)_{\xi_i^\nu,\xi_i}}{n} = \int\left(\prod_\nu Dt_\nu\right)\int Dz\Bigg\{&-\frac{(s-1)}{2}\ln\left(1+\lambda\left(1-b^2\right)(Q-q)\right)\\ &-\frac{1}{2}\ln\left[1+\lambda\,(Q-q)\left(1-b^2+sb^2\right)+2\lambda h\left(1+\lambda\left(1-b^2\right)(Q-q)\right)(Q-q)\right]\\ &+\frac{s\lambda}{2\,(1+\lambda\,(1-b^2)\,(Q-q))} - \lambda hq\;z^2 + \frac{2\lambda\sum_\nu B_\nu - \lambda\sum_\nu (\;B_\nu)^2}{2\,(1+\lambda\,(1-b^2)\,(Q-q))}\\ &+\frac{\left[2\lambda h\sqrt{q}\;z\sqrt{(Q-q)}\,(1+\lambda\,(1-b^2)\,(Q-q)) - \lambda\sqrt{(Q-q)\,b^2}\sum_\nu \xi_i\xi_i^\nu\;\;(1-B_\nu)\right]^2}{2\,[1+\lambda\,(Q-q)\,(1-b^2+sb^2)\\ +2\lambda h\,(1+\lambda\,(1-b^2)\,(Q-q))\,(Q-q)]\,(1+\lambda\,(1-b^2)\,(Q-q))}\Bigg\},\end{aligned} \tag{6.51}$$

onde $B_\nu = \sqrt{(1-b^2)\,q}t_\nu + \sqrt{q}\; b\xi_i\xi_i^\nu\; z$.

O segundo momento, para dado ξ_i^ν e ξ_i, obtém-se de

$$\overline{\gamma^2}_{\xi_i^\nu,\xi_i} = -\frac{1}{\lambda n}\left[\frac{\partial\,(G_1^s)_{\xi_i^\nu,\xi_i}}{\partial h}\right]_{h=0}, \tag{6.52}$$

resultando

$$\begin{aligned}
\overline{\gamma^2}_{\xi_i^\nu,\xi_i} \;=\; & \int\left(\prod_\nu Dt_\nu\right)\int Dz\left\{\frac{(1+\lambda\,(1-b^2)\,(Q-q))\,(Q-q)}{1+\lambda\,(Q-q)\,(1-b^2+sb^2)} + q\;z^2\right.\\
& -\frac{2\lambda\sqrt{q}\,(Q-q)\,b\sum_\nu \xi_i\xi_i^\nu\;B_\nu\;z}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))} + \frac{2\lambda\sqrt{q}\,(Q-q)\,b\sum_\nu \xi_i\xi_i^\nu\;z}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))}\\
& +\frac{\lambda^2\,(Q-q)^2\,b^2\,(\sum_\nu \xi_i\xi_i^\nu\;)^2}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2} + \frac{\lambda^2\,(Q-q)^2\,b^2\,(\;\sum_\nu \xi_i\xi_i^\nu\;B_\nu\;)^2}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}\\
& \left.-\frac{2\lambda^2\,(Q-q)^2\,b^2\,(\sum_\nu \xi_i\xi_i^\nu)\;(\sum_\nu \xi_i\xi_i^\nu\;B_\nu\;)}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}\right\}.
\end{aligned} \tag{6.53}$$

Os termos lineares nas variáveis z e t_ν resultam nulos na integração, enquanto os termos quadraticos fornecem o próprio coeficiente, o que leva à equação

$$\begin{aligned}
\overline{\gamma^2}_{\xi_i^\nu,\xi_i} \;=\; & \frac{(1+\lambda\,(1-b^2)\,(Q-q))\,(Q-q)}{1+\lambda\,(Q-q)\,(1-b^2+sb^2)} + q + \frac{\lambda^2\,(Q-q)^2\,b^2\,(\sum_\nu \xi_i\xi_i^\nu\;)^2}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}\\
& -\frac{2\lambda q\,(Q-q)\,b^2 s}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))} + \frac{\lambda^2\,(Q-q)^2\,qb^4s^2}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}\\
& +\frac{\lambda^2\,(Q-q)^2\,b^2\;\;(1-b^2)\,q}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}\int\left(\prod_\nu Dt_\nu\right)\left(\sum_\nu \xi_i\xi_i^\nu t_\nu\right)^2.
\end{aligned} \tag{6.54}$$

Usando $\int\,(\prod_\nu Dt_\nu)\,(\sum_\nu \xi_i\xi_i^\nu t_\nu)^2 = s$, escrevemos

$$\begin{aligned}
\overline{\gamma^2}_{\xi_i^\nu,\xi_i} \;=\; & \frac{(1+\lambda\,(1-b^2)\,(Q-q))\,(Q-q)}{1+\lambda\,(Q-q)\,(1-b^2+sb^2)} + q + \frac{\lambda^2\,(Q-q)^2\,b^2\,(\sum_\nu \xi_i\xi_i^\nu\;)^2}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}\\
& -\frac{2\lambda q\,(Q-q)\,b^2 s}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))} + \frac{\lambda^2\,(Q-q)^2\,qb^4s^2}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}\\
& +\frac{s\lambda^2\,(Q-q)^2\,b^2\;\;(1-b^2)\,q}{(1+\lambda\,(Q-q)\,(1-b^2+sb^2))^2}.
\end{aligned} \tag{6.55}$$

Para $\alpha \leq \alpha_c = 1/s$, devemos tomar o limite $\lambda \to \infty$, com $q < Q$, o que, com $\delta^2_{\xi_i^\nu,\xi_i} = \overline{\gamma^2}_{\xi_i^\nu,\xi_i} - \overline{\gamma}^2_{\xi_i^\nu,\xi_i}$, leva à seguinte expressão para a variância

$$\delta^2_{\xi_i^\nu,\xi_i} = \frac{Q\,(1-b^2)}{1-b^2+sb^2}, \tag{6.56}$$

com $Q = \alpha s/(1-\alpha s)$, equação (6.32).

Para $\alpha > \alpha_c = 1/s$, tomamos o limite $\lambda(Q-q) \to x$, obtendo

$$\overline{\gamma^2}_{\xi_i^\nu,\xi_i} = \frac{\begin{array}{c} Q\left[1+x\left(1-b^2+sb^2\right)\right]^2 + x^2b^2\left(\sum_\nu \xi_i\xi_i^\nu\right)^2 \\ -2Qxb^2s\ \left[1+x\left(1-b^2+sb^2\right)\right] \\ +Qx^2b^2s\left(1-b^2+sb^2\right) \end{array}}{\left[1+x\left(1-b^2+sb^2\right)\right]^2}, \tag{6.57}$$

onde x e Q são dados por (6.34) e (6.36), respectivamente. A variância é então escrita como

$$\delta^2_{\xi_i^\nu,\xi_i} = \frac{Q\left\{1+x\left(1-b^2\right)\left[2+x\left(1-b^2+sb^2\right)\right]\right\}}{\left[1+x\left(1-b^2+sb^2\right)\right]^2}. \tag{6.58}$$

Agora, a dependência em ξ_i e ξ_i^ν da função $w^c_{\xi_i^\nu,\xi_i}(\gamma)$, equação (6.44), pode ser posta na forma

$$\sum_\nu \xi_i\xi_i^\nu = 2n_+ - s, \tag{6.59}$$

onde n_+ é o número de sítios em que $\xi_i\xi_i^\nu = +1$, levando a $w^c_{\xi_i^\nu,\xi_i}(\gamma) = w^c_{n_+}(\gamma)$.

Finalmente, a equação (6.44) pode ser mediada em ξ_i e ξ_i^ν, o que fornece a seguinte equação

$$\begin{aligned} W^c(\gamma) &= \left\langle w^c_{n_+}(\gamma)\right\rangle_{n_+} \\ &= \sum_{n_+=0}^{s}\binom{s}{n_+}\left(\frac{1+b}{2}\right)^{n_+}\left(\frac{1-b}{2}\right)^{s-n_+} w^c_{n_+}(\gamma). \end{aligned} \tag{6.60}$$

Cálculo da fração de sítios instáveis

A medida da estabilidade do conceito $\boldsymbol{\xi}^l$ é dada pela fração média de sítios instáveis de $\boldsymbol{\xi}^l$, obtida por

$$\epsilon_c = \int_{-\infty}^{0} d\gamma\, W^c(\gamma), \tag{6.61}$$

onde $W^c(\gamma)$ é a distribuição de probabilidade das estabilidades dos conceitos, dada por (6.60).

No regime $\alpha \leq \alpha_c = 1/s$, a fração de sítios instáveis média nos conceitos é escrita como

$$\epsilon_c = \sum_{n_+=0}^{s}\binom{s}{n_+}\left(\frac{1+b}{2}\right)^{n_-}\left(\frac{1-b}{2}\right)^{s-n_-}$$

$$\frac{1}{2}\,\mathrm{erfc}\left[\frac{(2n_+ - s)\,b}{\sqrt{2Q\,(1-b^2)\,(1-b^2+sb^2)}}\right], \tag{6.62}$$

onde $Q = \alpha s/\,(1-\alpha s)$. No regime $\alpha > \alpha_c = 1/s$, ϵ_c é escrita como

$$\begin{aligned}\epsilon_c &= \sum_{n_+=0}^{s}\binom{s}{n_+}\left(\frac{1+b}{2}\right)^{n_+}\left(\frac{1-b}{2}\right)^{s-n_+} \\ &\quad \frac{1}{2}\,\mathrm{erfc}\left[\frac{x\,(2n_+ - s)\,b}{\sqrt{2Q\,\{(1-b^2)\,x\,[x\,(1-b^2+sb^2)+2]+1\}}}\right]\end{aligned} \tag{6.63}$$

onde x e Q são dados por (6.34) e (6.36), respectivamente.

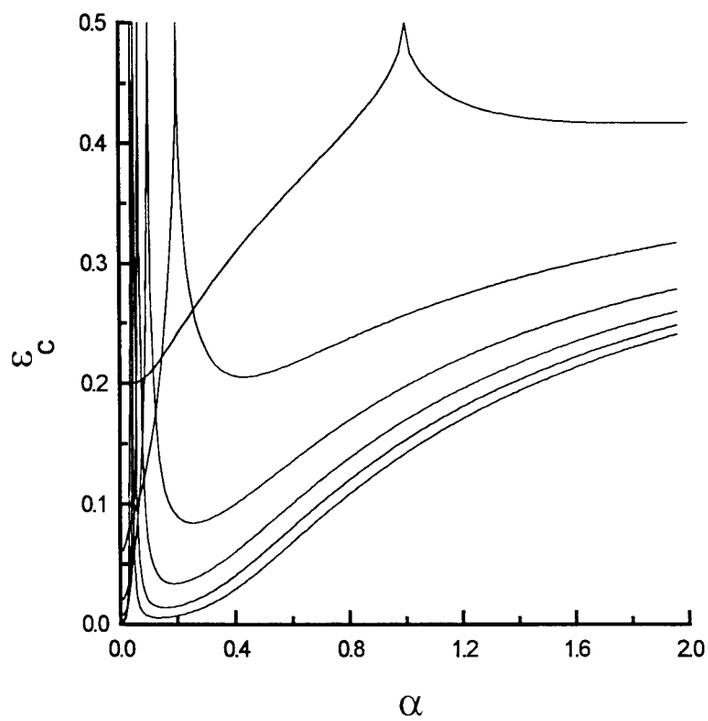

Figura 6.1: Erro de categorização contra α, para $d = 0.2$. As curvas são, de baixo para cima, para $s = 25$, 20, 15, 10, 5 e 1.

A figura 6.1 mostra o erro de categorização contra α para uma rede treinada com s exemplos à distância $d = 0.2$ dos conceitos. Em todas as curvas o pico $\epsilon_c = 0.5$

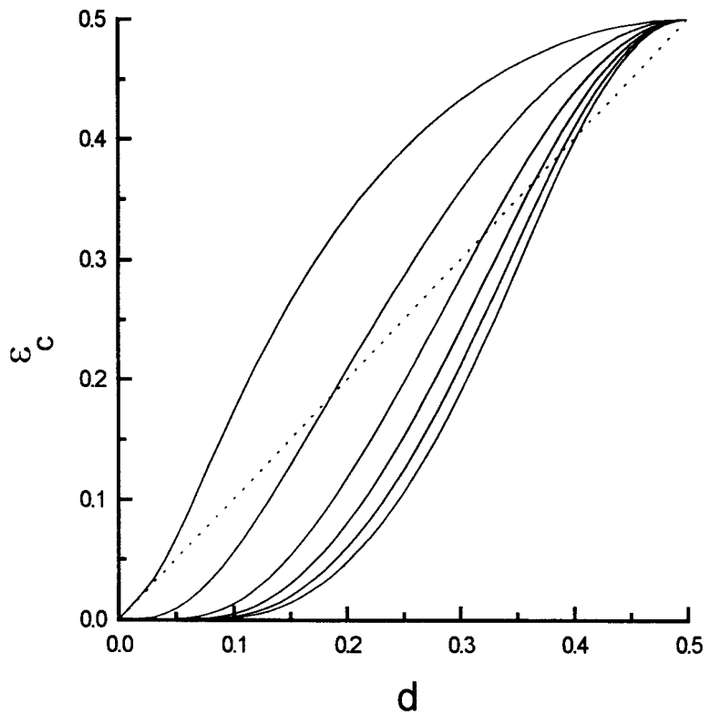

Figura 6.2: Erro de categorização contra d, para $\alpha = 0.6$. As curvas são, de baixo para cima, para $s = 25$, 20, 15, 10, 5 e 1.

ocorre para $\alpha = \alpha_c = 1/s$. Para $\alpha < \alpha_c$ a rede está criando atratores novos para cada exemplo apresentado. Para $\alpha > \alpha_c$ ocorre uma diminuição abrupta no erro de categorização, que logo passa a aumentar novamente. O limite $\alpha \to \infty$ leva a $\epsilon_c = 1/2$.

A figura 6.2 foi obtida para $\alpha = 0.6$. Para $s = 1$, recuperamos o comportamento apresentado na figura 4.1. Apenas esta curva não está na região $\alpha > \alpha_c$. Aqui, novamente, aumentando o número de exemplos s o erro de categorização diminui.

A figura 6.3 permite uma melhor comparação com o conhecido problema da categorização do modelo de Hopfield. Como lá, aqui existe uma região acima da qual o erro de categorização sempre diminui, nesta figura $s_c = 1/\alpha_c = 2$. É interessante notar, então, que a categorização na pseudo-inversa para α não nulo é similar à

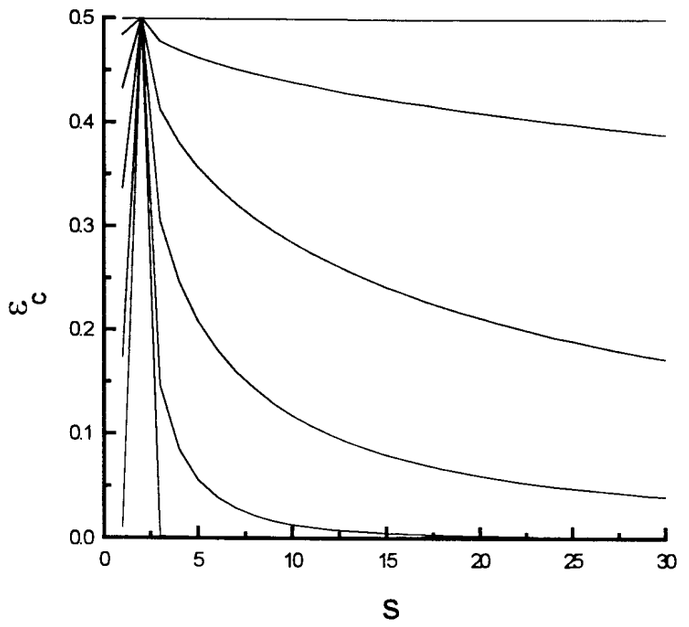

Figura 6.3: Erro de categorização contra s, para $\alpha = 0.5$. As curvas são, de baixo para cima, para $d = 0.01$, 0.1, 0.2, 0.3, 0.4 e 0.49.

categorização no modelo de Hopfield para P finito ($\alpha = 0$).

6.2 Estabilidade dos exemplos

A fim de complementar o estudo da seção anterior, vamos calcular a distribuição de probabilidade das estabilidades dos exemplos e sua fração de sítios instáveis.

$$
\begin{aligned}
&+\frac{\lambda^2 b^2 (Q-q)^2 (1-b^2) \left(\sum_\nu \xi_i \xi_i^\nu\right)^2}{s\left[1+\lambda (Q-q)(1-b^2)\right]^2 \left[1+\lambda (Q-q)\left[1+(s-1)b^2\right]\right]} \\
&+\frac{q\lambda^2 b^2 (Q-q)^2 (1-b^2)\left[1+(s-1)b^2\right]}{\left[1+\lambda (Q-q)(1-b^2)\right]^2 \left[1+\lambda (Q-q)\left[1+(s-1)b^2\right]\right]} \\
&+\frac{\lambda^2 b^2 (Q-q)^2 \left(\sum_\nu \xi_i \xi_i^\nu\right)^2}{s\left[1+\lambda (Q-q)(1-b^2)\right]\left[1+\lambda (Q-q)(1-b^2+sb^2)\right]^2} \\
&+\frac{q\lambda^2 b^2 (Q-q)^2 \left[1+(s-1)b^2\right]^2}{\left[1+\lambda (Q-q)(1-b^2)\right]\left[1+\lambda (Q-q)\left[1+(s-1)b^2\right]\right]^2}. \qquad (6.69)
\end{aligned}
$$

Tomando o limite $\lambda \longrightarrow \infty$ com $\alpha \leq \alpha_c = 1/s$, obtemos

$$
\overline{\gamma^2}_{\xi_i^\nu, \xi_i} = 1, \qquad (6.70)
$$

de forma que $w^e_{\xi_i^\nu, \xi_i}(\gamma) = \delta(\gamma - 1,)$, conforme esperado. A variância fica dada por $\delta^2_{\xi_i^\nu, \xi_i} = 0$, neste regime.

Já o limite $\lambda \longrightarrow \infty$, para $\alpha > \alpha_c = 1/s$, leva a

$$
\begin{aligned}
\overline{\gamma^2}_{\xi_i^\nu, \xi_i} = & \frac{\left[x(1-b^2)\right]^2}{\left[1+x(1-b^2)\right]^2} + \frac{Q}{\left[1+x(1-b^2)\right]^2} - \frac{2b^2 x Q\left[1+b^2(s-1)\right]}{\left[1+x(1-b^2)\right](1+x)} \\
& + \frac{b^2 x^2 (1-b^2)\left(\sum_\nu \xi_i \xi_i^\nu\right)^2}{s\left[1+x(1-b^2)\right]^2\left[1+x\left[1+(s-1)b^2\right]\right]} \\
& + \frac{Q b^2 x^2 (1-b^2)\left[1+b^2(s-1)\right]}{\left[1+x(1-b^2)\right]^2\left[1+x\left[1+(s-1)b^2\right]\right]} \\
& + \frac{b^2 x \left(\sum_\nu \xi_i \xi_i^\nu\right)^2}{s\left[1+x(1-b^2)\right]\left[1+x\left[1+(s-1)b^2\right]\right]^2} \\
& + \frac{Q b^2 x^2 \left[1+(s-1)b^2\right]^2}{\left[1+x(1-b^2)\right]\left[1+x\left[1+(s-1)b^2\right]\right]^2}, \qquad (6.71)
\end{aligned}
$$

onde x e Q são dados por (6.34) e (6.36), respectivamente.

Novamente, escreve-se $\sum_\nu \xi_i \xi_i^\nu = 2n_+ - s$, onde n_+ é o número de sítios em que $\xi_i \xi_i^\nu = +1$, e a equação (6.65) pode ser mediada em n_+, levando à seguinte expressão

$$
\begin{aligned}
W^e(\gamma) &= \left\langle w^e_{2n_+ - s}(\gamma)\right\rangle_{n_+} \\
&= \sum_{n_+=0}^{s} \binom{s}{n_+} \left(\frac{1+b}{2}\right)^{n_+} \left(\frac{1-b}{2}\right)^{s-n_+} w^e_{n_+}(\gamma). \qquad (6.72)
\end{aligned}
$$

Cálculo da fração de sítios instáveis

A medida da estabilidade dos exemplos $\boldsymbol{\xi}^{k\nu}$ é dada pela fração média de sítios instáveis de $\boldsymbol{\xi}^{k\nu}$, obtida por

$$\epsilon_e = \int_{-\infty}^{0} d\gamma W^e(\gamma), \tag{6.73}$$

onde $W^e(\gamma)$ é a distribuição de probabilidade das estabilidades dos exemplos, dada por (6.72).

No regime $\alpha \leq \alpha_c = 1/s$, ϵ_e é nula, pois todos os exemplos são armazenados corretamente. No regime $\alpha > \alpha_c = 1/s$, ϵ_e é escrita como

$$\epsilon_e = \frac{1}{2}\sum_{n_+=0}^{s}\binom{s}{n_+}\left(\frac{1+b}{2}\right)^{n_+}\left(\frac{1-b}{2}\right)^{s-n_+}\operatorname{erfc}\left[\sqrt{\frac{1}{2}}\frac{\overline{\gamma}_{n_+}}{\delta_{n_+}}\right]. \tag{6.74}$$

A figura 6.4 mostra a fração de sítios instáveis dos exemplos ϵ_e contra α para d fixo e vários valores de s. Há um aumento monótono de ϵ_e à medida que α aumenta. Notar que para $\alpha \leq \alpha_c = 1/s$ os exemplos são estáveis, isto é, $\epsilon_e = 0$. Esta região $(\epsilon_e = 0)$, corresponde, na figura 6.1, a α abaixo dos picos nos quais $\epsilon_c = 0.5$.

O fenômeno da categorização fica mais claro na figura 6.5, onde pode-se ver que a partir de $s_c = 1/\alpha_c = 2$, a fração de sítios instáveis dos exemplos ϵ_e contra s para α fixo e vários valores de d aumenta monotonicamente.

6.3 Análise termodinâmica para P finito

O problema da recuperação dos padrões armazenados em redes neurais atratoras pode ser estudado investigando-se a correlação de recuperação no equilíbrio. No caso de redes com matrizes de pesos simétricas e com termos diagonais nulos, é possível realizar o estudo termodinâmico desse problema. Para a formulação KS da pseudo-inversa, esse estudo foi feito por Kanter e Sompolinsk [21] para α geral. Nesta seção, estudaremos a termodinâmica do problema da categorização na formulação KS da pseudo-inversa para P finito (mais precisamente, para $P = 1$).

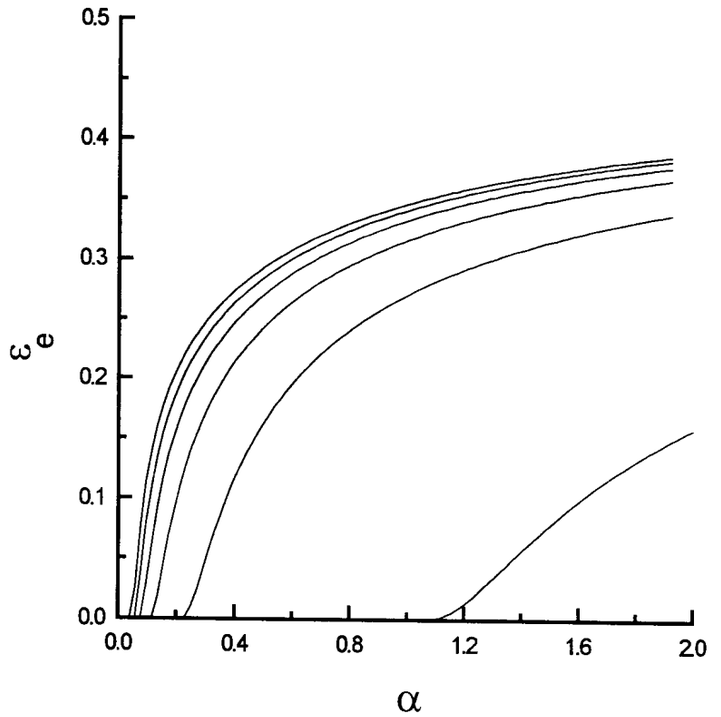

Figura 6.4: Fração de sítios instáveis dos exemplos contra α, para $d = 0.2$. As curvas são, de baixo para cima, para $s = 1$, 5, 10, 15, 20 e 25.

Conforme mencionado, o problema da categorização consiste em estudar a recuperação de um conceito $\boldsymbol{\xi}^1$, ao qual a rede só tem acesso através do conjunto de exemplos $\boldsymbol{\xi}^{1\nu}$, dados pela distribuição de probabilidade condicional (6.1).

Nossa rede é definida pela energia

$$E = -\frac{1}{2}\sum_{ij} J_{ij} S_i S_j, \tag{6.75}$$

onde os pesos sinápticos são dados pela equação (5.1), $J_{ij} = \frac{1}{N}\sum_{\mu\nu} \xi_i^{1\mu}\xi_j^{1\nu}\left(C^{-1}\right)_{\mu\nu}$ com os termos diagonais nulos, $J_{ii} = 0$.

Utilizando a expressão da energia e fazendo a mudança de variável $m_\nu \equiv m_{1\nu} =$

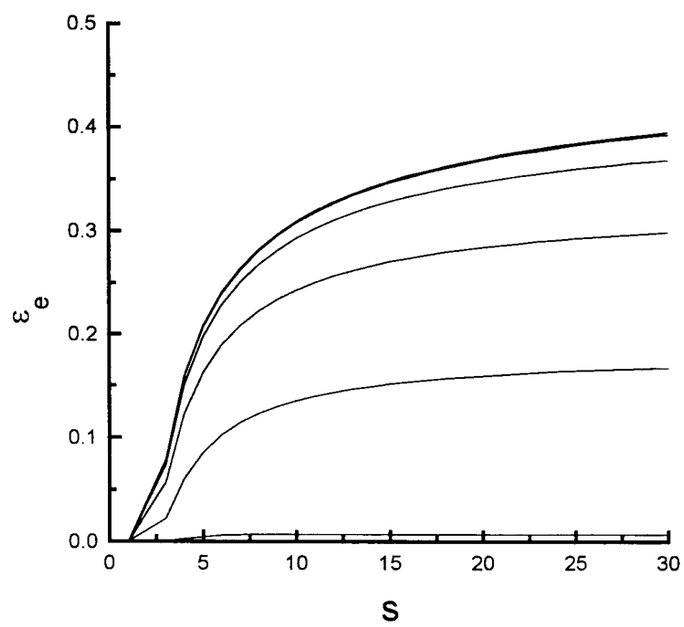

Figura 6.5: Fração de sítios instáveis dos exemplos contra s, para $\alpha = 0.5$. As curvas são, de baixo para cima, para $d = 0.01$, 0.1, 0.2, 0.3, 0.4 e 0.49.

$\frac{1}{N}\sum_i \xi_i^{1\nu} S_i$, obtemos a seguinte expressão para a função de partição

$$\begin{aligned} Z &= \operatorname*{Tr}_S \exp\{-\beta E\} \\ &= \int \left[\prod_\mu \frac{dm_\mu d\widetilde{m}_\mu}{2\pi}\right] \exp\left\{\mathbf{i}\sum_\mu m_\mu \widetilde{m}_\mu + \frac{\beta N}{2}\sum_{\mu\nu}\left(C^{-1}\right)_{\mu\nu} m_\mu m_\nu\right\} \\ &\qquad \operatorname*{Tr}_S \exp\left\{-\frac{\mathbf{i}}{N}\sum_\mu \widetilde{m}_\mu \sum_i \xi_i^{1\mu} S_i\right\}. \end{aligned} \tag{6.76}$$

Fazendo $\widetilde{m}_\mu = \frac{\widehat{m}_\mu}{\mathbf{i}/N}$ e explicitando a média sobre as variáveis $\xi_i^{1\nu}$ e ξ_i^1, obtemos

$$\begin{aligned} \langle\langle Z\rangle\rangle_{\xi^{k\nu},\xi^k} &= \int \left[\prod_\mu \frac{dm_\mu d\widehat{m}_\mu}{2\pi\mathbf{i}/N}\right] \exp\left\{N\sum_\mu m_\mu \widehat{m}_\mu + \frac{\beta N}{2}\sum_{\mu\nu}\left(C^{-1}\right)_{\mu\nu} m_\mu m_\nu\right\} \\ &\qquad \left\langle\left\langle \prod_i \operatorname*{Tr}_S \exp\left\{\left(-\sum_\mu \widehat{m}_\mu \xi_i^{1\mu}\right) S_i\right\}\right\rangle\right\rangle_{\xi^{1\nu},\xi^1}, \end{aligned} \tag{6.77}$$

onde utilizamos o fato dos elementos $C_{\mu\nu}$ serem automediantes no limite de P finito [21]. Isto leva a $C_{\mu\mu} = 1$ e $C_{\mu\nu} = b^2$ para $\mu \neq \nu$. A matriz inversa $(C^{-1})_{\mu\nu}$ pode ser facilmente calculada, resultando

$$C_0 \equiv \left(C^{-1}\right)_{\mu\mu} = \frac{1 + b^2 (s-2)}{(1-b^2)(1-b^2+sb^2)} \tag{6.78}$$

e

$$C_1 \equiv \left(C^{-1}\right)_{\mu\neq\nu} = \frac{b^2}{(1-b^2)(1-b^2+sb^2)}. \tag{6.79}$$

Efetuando a soma sobre S_i, reescrevemos a função de partição como

$$\begin{aligned}\langle\langle Z\rangle\rangle_{\xi^{k\nu},\xi^k} = \int \left[\prod_\mu \frac{dm_\mu d\widehat{m}_\mu}{2\pi \mathbf{i}/N}\right] \exp\Bigg\{-\beta N \Bigg[\sum_\mu m_\mu \widehat{m}_\mu \\ + \frac{\beta}{2}\sum_{\mu\nu}\left(C^{-1}\right)_{\mu\nu} m_\mu m_\nu + \left\langle\left\langle \ln 2\cosh\left(\sum_\mu \widehat{m}_\mu \xi_i^{1\mu}\right)\right\rangle\right\rangle_{\xi^{1\nu},\xi^1}\Bigg]\Bigg\},\end{aligned} \tag{6.80}$$

onde utilizamos a propriedade de automediância $\frac{1}{N}\sum_i g\left(\xi_i^{1\nu}\xi_i^1\right) = \langle g\left(\xi^{1\nu}\xi^1\right)\rangle_{\xi^{1\nu},\xi^1}$.

No limite $N \longrightarrow \infty$, as integrais em dm_μ e $d\widehat{m}_\mu$ podem ser efetuadas pelo método do ponto de sela, levando à seguinte expressão

$$\begin{aligned}-\beta f(\beta, m) \;=\; \mathrm{extr}\Bigg\{\sum_\mu m_\mu \widehat{m}_\mu + \frac{\beta}{2}\sum_{\mu\nu}\left(C^{-1}\right)_{\mu\nu} m_\mu m_\nu \\ + \left\langle\left\langle \ln 2\cosh\left(\sum_\mu \widehat{m}_\mu \xi_i^{1\mu}\right)\right\rangle\right\rangle_{\xi^{1\nu},\xi^1}\Bigg\}.\end{aligned} \tag{6.81}$$

Da equação $\partial f/\partial m_\mu = 0$, obtemos

$$\widehat{m}_\mu = -\beta \sum_\nu \left(C^{-1}\right)_{\mu\nu} m_\nu, \tag{6.82}$$

que, substituindo na expressão para f, leva a

$$f = \frac{1}{2}\sum_{\mu\nu}\left(C^{-1}\right)_{\mu\nu} m_\mu m_\nu - \frac{1}{\beta}\left\langle\left\langle \ln 2\cosh\beta\left(\sum_{\mu\nu}\left(C^{-1}\right)_{\mu\nu} m_\nu \xi_i^{1\mu}\right)\right\rangle\right\rangle_{\xi^{1\nu},\xi^1}. \tag{6.83}$$

Usando as equações (6.78) e (6.79), a equação anterior pode ser escrita como

$$\begin{aligned}f = \frac{1}{2}C_0 \sum_\mu m_\mu^2 + \frac{1}{2}C_1 \sum_{\mu\neq\nu} m_\mu m_\nu \\ -\frac{1}{\beta}\left\langle\left\langle \ln 2\cosh\beta\left(C_0 \sum_\mu m_\mu \xi_i^{1\mu} + C_1 \sum_{\mu\neq\nu} m_\nu \xi_i^{1\mu}\right)\right\rangle\right\rangle_{\xi^{1\nu},\xi^1}.\end{aligned} \tag{6.84}$$

Vamos nos concentrar no caso geral de soluções assimétricas, com $m_\mu = m_1$ para $\mu = 1$ e $m_\mu = m_{s-1}$ para $\mu > 1$. Substituindo $m_\mu = m_1$ para $\mu = 1$ e $m_\mu = m_{s-1}$ para $\mu > 1$ na expressão para a energia livre, além de fazer $X_{s-1} \equiv \sum_{\mu>1}^{s} \xi_i^{1\mu}\xi_i^1$, a equação $\partial f/\partial m_1 = 0$ permite-nos escrever

$$C_0 m_1 + C_1 (s-1) m_{s-1} = \langle\langle [C_0 + C_1 X_{s-1}] \tanh \{\beta \Xi\} \rangle\rangle_{\xi^{1\nu},\xi^1} . \tag{6.85}$$

Já a equação $\partial f/\partial m_{s-1} = 0$, leva a

$$\begin{aligned}(s-1)(C_0 - C_1) m_{s-1} + C_1 (s-1) [m_1 + (s-1) m_{s-1}] = \\ \langle\langle [(s-1) C_1 + X_{s-1} (C_0 + (s-2) C_1)] \tanh \{\beta \Xi\} \rangle\rangle_{\xi^{1\nu},\xi^1} ,\end{aligned} \tag{6.86}$$

onde

$$\begin{aligned}\Xi \;=\; & C_0 m_1 + C_1 (s-1) m_{s-1} + \\ & + X_{s-1} [C_1 m_1 + m_{s-1} (C_0 + (s-2) C_1)] .\end{aligned} \tag{6.87}$$

Apesar de sua aparência complicada, este sistema de equações tem uma solução analítica extremamente simples, a saber,

$$m_1 = \tanh (\beta m_1) \tag{6.88}$$

e

$$m_{s-1} = b^2 m_1 . \tag{6.89}$$

Observe que tanto m_1 como m_{s-1} independem de s. Fica claro, então, que não ocorre categorização no limite P finito, ao contrário do modelo de Hopfield, onde para um dado número crítico de exemplos s_c, ocorre uma transição descontínua para o regime $m_1 = m_{s-1}$, que corresponde à fase de categorização. Note que ocorre uma transição contínua para o regime paramagnético em $\beta = 1$.

Capítulo 7

Conclusão

O objetivo principal desta tese foi estudar as propriedades de recuperação de memórias em dois dos mais populares modelos de redes neurais atratoras: a rede pseudo-inversa e a rede dos pesos ótimos. Além disso, estendemos a análise da categorização, anteriormente restrita ao modelo de Hopfield, para o modelo da pseudo-inversa.

As propriedades de recuperação foram abordadas por três métodos. No primeiro método, nossa análise tratou de redes neurais extremamente diluídas, onde o número médio de conexões C é muito menor que o logaritmo do número de neurônios N. Neste caso, uma vez que a dinâmica do primeiro passo [29] torna-se exata para qualquer tempo [14] [24], pudemos levantar o diagrama de fase do modelo da pseudo-inversa no espaço (α, T). Aqui, α é a razão entre o número de padrões armazenados P e a conectividade C; e T é o parâmetro que mede a intensidade de ruído na dinâmica neural. Assim, generalizamos a análise anterior restrita à $T = 0$ [27]. Também obtivemos o diagrama de fase para a rede dos pesos ótimos no espaço completo de parâmetros (α, κ, T), onde κ é o parâmetro de margem, introduzido originalmente para controlar a bacia de atração das memórias. Até este trabalho [28], a análise da rede dos pesos ótimos estava restrita ao regime de saturação $\alpha = \alpha_c(\kappa)$ [24] [26]. Para os dois modelos, fomos capazes de delimitar uma fase de recuperação, outra de não-recuperação e uma terceira fase em que ambas as soluções coexistem. Um dos principais resultados dessa abordagem foi mostrar que para o modelo dos

pesos ótimos os padrões memorizados, apesar de serem pontos fixos da dinâmica neural, possuem bacia de atração nula para $\kappa = 0$ e $\alpha \leq 2$.

A segunda maneira de estudar as propriedades de recuperação consistiu numa técnica analítica original para investigar a natureza da vizinhança dos padrões armazenados $\boldsymbol{\xi}^l$. Particularmente, calculamos a fração de sítios ϵ que se tornam instáveis quando d sítios do padrão armazenado são alterados. Esta pode ser uma maneira interessante de caracterizar um modelo de memória associativa, pois somos capazes de imaginar uma dinâmica que garanta a recuperação dos padrões armazenados quando $\epsilon \simeq d$. De fato, basta que tal dinâmica apenas altere um sítio se o número de sítios instáveis diminui. Segundo esta dinâmica hipotética, memórias na rede pseudo-inversa (figura 4.1) seriam mais facilmente recuperadas do que memórias na rede dos pesos ótimos (figura 4.2) pois aquela tem a vizinhança mais suave que esta. É importante notar que esses resultados são independentes do grau de diluição da rede. Pretendemos, em breve, estender esses cálculos para o modelo da pseudo-inversa com viés não-nulo ($a \neq 0$).

A terceira maneira pela qual estudamos as propriedades de recuperação foi pela enumeração exaustiva dos atratores das redes pseudo-inversa e pesos ótimos completamente conectadas ($C = N$), através de um algoritmo numérico bastante eficiente [30]. Utilizamos três formulações distintas para obter os pesos sinápticos da rede pseudo-inversa, apresentadas no capítulo 5. Duas dessas formulações, o modelo KS e o modelo E, são, dentro da escala de comparação utilizada, idênticas sob todos os aspectos considerados. O modelo PGD difere do KS apenas por ter os termos diagonais da matriz de conexão não nulos. Conforme já era conhecido da literatura [46], o número de pontos fixos para a pseudo-inversa cresce exponencialmente com o tamanho da rede (figura 5.1). O resultado original dessa figura foi a determinação da dependência, também exponencial, do número de pontos fixos do modelo dos pesos ótimos com o tamanho da rede. Dois aspectos marcantes no que se refere ao modelo dos pesos ótimos são a quase desprezível influência do parâmetro de margem κ no tamanho da bacia de atração dos estados memorizados (figura 5.5)

e a não homogeneidade das bacias de atração de seus atratores quando $\kappa = 0$ (figuras 5.3 e 5.4). Ainda para a rede dos pesos ótimos, a análise numérica mostrou que a bacia de atração dos estados memorizados não é nula para $\kappa = 0$ (figura 5.10). Para a dinâmica paralela, verificamos a predominância dos ciclos de período dois sobre os pontos fixos (figura 5.11 e 5.1). Particularmente, para o modelo KS próximo à saturação ($\alpha \to 1$), esses ciclos dominam praticamente todo o espaço de configurações, pois $\langle Y_2 \rangle \to 1$ (figura 5.12).

Finalmente, estudamos analiticamente o problema da categorização no modelo da pseudo-inversa, que é a capacidade da rede treinada com exemplos de um conceito desenvolver um atrator para este conceito. Fomos capazes de identificar um regime de operação no qual pode-se dizer que o modelo da pseudo-inversa apresenta a capacidade de categorização a partir de um certo número de exemplos $s_c = 1/\alpha$, acima do qual a rede tem o erro de categorização diminuindo monotonicamente com s (figura 6.3). De fato, esses resultados são bastante similares aos obtidos para o modelo de Hopfield no limite de P finito [31] [35], exceto pelo fato de ϵ_c aumentar com s para $s < s_c$ na pseudo-inversa, enquanto ϵ_c praticamente independe de s no modelo de Hopfield. Por último, o estudo da termodinâmica do modelo KS para P finito demonstrou não ocorrer categorização neste limite, contrariamente ao que ocorre no modelo de Hopfield.

Bibliografia

[1] Müller B. e Reinhardt J. *Neural Networks: an introduction.* (Springer: Berlin), 1991.

[2] Gutfreund H. e Toulouse G. *The physics of neural networks*, chapter Cap. 1, page 7. *in* Spin Glasses and Biology Stein D. L., (Word Scientific: London), 1992.

[3] Schiff S. J., Jerger K., Duong D. H., Chang T., Spano M. L. e Ditto W. L. *Nature* **370** 615 (1994).

[4] Hoff A. *Z. Naturforsch* **49a** 589 (1994).

[5] McCulloch W. S. e Pitts W. A. *Bull. Math. Biophys.* **5** 115 (1943).

[6] Hebb D. O. *The Organization of Behavior.* (Wiley: New York), 1949.

[7] Hopfield J. J. *Proc. Natl.Acad. Sci., USA* **79** 2554 (1982).

[8] Kinzel W. *Z. Phys. B - Condensed Matter* **60** 205 (1985).

[9] Amit D. J., Gutfreund H. e Sompolinsky H. *Phys. Rev. A* **32** 1007 (1985).

[10] Amit D. J., Gutfreund H. e Sompolinsky H. *Ann. Phys. (NY)* **173** 30 (1987).

[11] Kirkpatrick S. e Sherrington D. *Phys. Rev.* **B17** 983 (1978).

[12] Van Hemmen J. L. e Palmer R. G. *J. Phys. A: Math. Gen.* **12** 563 (1979).

[13] Mezard M., Parisi G. e Virasoro M. A. *Spin glass theory and beyond.* (Word Scientific Publishing: NJ), 1987.

[14] Derrida B., Gardner E. e Zippelius A. *Europhys. Lett.* **2** 337 (1987).

[15] Sompolinsky H. *Phys. Rev. A* **34** 2571 (1986).

[16] Van Hemmen J. L. e Morgenstern I., editor. *Heidelberg Colloquium on Glassy Dynamics*, volume 275. (Springer: Berlin), 1987.

[17] Gardner E. *J. Phys. A: Math. Gen.* **21** 257 (1988).

[18] Gardner E. e Derrida B. *J. Phys. A: Math. Gen.* **21** 271 (1988).

[19] Kohonen T. *Self-Organisation and Associative Memory.* (Springer: Berlin), 1989.

[20] Personnaz L., Guyon I. e Dreyfus G. *Phys. Rev. A* **34** 4217 (1986).

[21] Kanter I. e Sompolinsky H. *Phys. Rev. A* **35** 380 (1987).

[22] Opper M., Kinzel W., Kleinz J. e Nehl R. *J. Phys. A: Math. Gen.* **23** L581 (1990).

[23] Fontanari J. F. *J. Phys. A: Math. Gen.* **26** 6147 (1993).

[24] Gardner E. *J. Phys. A: Math. Gen.* **22** 1969 (1989).

[25] Abbott L. F. e Kepler T. B. *J. Phys. A: Gen.* **22** L711 (1989).

[26] Amit D. J., Evans M. R., Horner H. e Wong K. Y. M. *J. Phys. A: Math. Gen.* **23** 3361 (1990).

[27] Opper M., Kleinz J., Köhler H. e Kinzel W. *J. Phys. A: Math. Gen.* **22** L407 (1989).

[28] Rodrigues Neto C. e Fontanari J. F. *J. Phys. A: Math. Gen.* **29** 3041 (1996).

[29] Kepler T. B. e Abbot L. F. *J. Physique* **49** 1657 (1988).

[30] Gutfreund H., Reger J. D. e Young A. P. *J. Phys. A: Math. Gen.* **21** 2775 (1988).

[31] Fontanari J. F. *J. Physique* **51** 2421 (1990).

[32] Miranda E. N. *J. Physique I* **1** 999 (1991).

[33] Branchstein M. C. e Arenzon J. J. *J. Physique I* **2** 2019 (1992).

[34] Stariolo D. A. e Tamarit F. A. *Phys. Rev. A* **46** 5249 (1992).

[35] Krebs P. R. e Theumann W. K. *1993* **26** 3983 (1993).

[36] Silva C. R., Tamarit F. A., Lemke N., Arenzon J. J. e Curado E. M. *J. Phys. A* **28** 1593 (1995).

[37] Abbott L. F. *Network* **1** 105 (1990).

[38] De Almeida J. R. e Thouless D. J. *J. Phys. A: Math. Gen.* **11** 983 (1978).

[39] Binder K. e Young A. *Rev. Mod. Phys.* **58** 801 (1986).

[40] Theumann W. K. e Erichsen Jr. R. *J. Phys. A* **24** L565 (1991).

[41] Majer P., Engel A. e Zippelius A. *J. Phys. A: Math. Gen.* **26** 7405 (1993).

[42] Erichsen Jr. R. e Theumann W. K. *J. Phys. A: Math. Gen.* **26** L61 (1993).

[43] Whyte W. e Sherrington D. *J. Phys. A: Math. Gen.* **29** 3063 (1996).

[44] Rosenblatt F. *Principles of neurodynamics.* (Spartan: New York), 1962.

[45] Minsky M. e Papert S. *Perceptrons.* (MIT Press: Cambridge, MA), 1969.

[46] Kuhlmann P. e Anlauf J. K. *J. Phys. A: Math. Gen.* **27** 5871 (1994).

[47] Treves A. e Amit D. J. *J. Phys. A: Math. Gen.* **21** 3155 (1988).

[48] Gardner E. *J. Phys. A* **19** L1047 (1986).

[49] Peretto P. *Biol. Cybernet.* **50** 51 (1984).