Universidade de São Paulo
Instituto de Física e Química de São Carlos

O Efeito das Lesões nas Capacidades de Memorização e Generalização de um Perceptron

Daniela M.L. Barbato

USP / IFQSC / SBI
8-2-001016

Dissertação apresentada ao
Instituto de Física e Química de São Carlos
para obtenção do título de
Mestre em Física Básica.

Orientador: Prof. Dr. José Fernando Fontanari

Departamento de Física e Ciência dos Materiais
São Carlos – 1993

SERVIÇO DE BIBLIOTECA E INFORMAÇÃO - IFQSC
FÍSICA

UNIVERSIDADE DE SÃO PAULO
Instituto de Física e Química de São Carlos

Fone (0162) 72-6222
Fax (0162) 72-2218

Av. Dr. Carlos Botelho, 1465
Caixa Postal 369
CEP 13560.970 - São Carlos - SP
Brasil

MEMBROS DA COMISSAO JULGADORA DA DISSERTACAO DE MESTRADO DE DANIELA MARIA LEMOS BARBATO APRESENTADA AO INSTITUTO DE FISICA E QUIMICA DE SAO CARLOS, DA UNIVERSIDADE DE SAO PAULO, EM 08/09/93

COMISSAO JULGADORA:

Prof.Dr.José Fernando Fontanari

Prof.Dr.Sylvio Goulart Rosa Junior

Prof.Dr.Nestor Felipe Caticha Alfonso

Agradecimentos

Gostaria de agradecer a todos que contribuiram para a realização deste trabalho:

. ao Prof. Fontanari pela orientação, disposição e principalmente pela paciência.

. aos meus pais, Cleusa e Roberto por tudo.

. aos meus irmàos, Pito e Kiza por me aguentarem.

. ao meus avós Líbia e Orlando.

. ao Xuxo pela compreensão.

. ao Mauro e ao Camarada (Gandhi) pela amizade e por resolverem todos os tipos de problemas computacionais.

. ao Prof. Degani pelo empréstimo da workstation

. às minhas grandes amigas Maria Cristina e Marta.

. aos amigos Salviano, Osame, Sérgio, Jean, Ricardo, Cachucho, Rocha, Jeremias, Virgílio, Washington, Miled, Zé Luis, Luis Antônio, Facin, Manzoli, Melquisedech, Zé Camilo, Wilson, Márcio, Wanda, Javier, Mário, Bunny, Miro, Zak e tantos outros.

Conteúdo

Resumo **v**

Abstract **vi**

1 Introdução **1**

2 O Modelo **8**

2.1 Generalização 10

2.2 Lesões 12

2.3 Ruído 14

3 A Mecânica Estatística do Modelo **16**

3.1 Diluição Móvel 20

3.2 Diluição Fixa 25

3.3 Erro de Generalização Médio 27

4 Análise dos Resultados **29**

4.1 Erro de Treinamento 29

4.2 Capacidade de Armazenamento 32

4.3 Erro de Generalização 35

5 Conclusão **43**

A Cálculo da Entropia **48**

A.1 Diluição Móvel 48

A.2 Diluição Fixa 51

B Cálculo de G_0 com a Prescrição Simetria de Réplicas **52**

B.1 Diluição Móvel 52

B.2 Diluição Fixa 53

C Cálculo de G_1 com a Prescrição Simetria de Réplicas **55**

Bibliografia **57**

Lista de Figuras

1.1 Diagrama esquemático de uma célula nervosa ou neurônio. 2

2.1 Estrutura da rede neural com $N = 5$. 9

3.1 Entropia em função do erro de treinamento ϵ_t para $Q = 1$ e $\gamma = 0$. 25

4.1 Erro de treinamento em função de α para a diluição móvel e $\gamma = 1$. 30

4.2 Erro de treinamento em função de α para a diluição móvel e $\gamma = 0.9$. . . . 31

4.3 Erro de treinamento em função de α para a diluição fixa e $\gamma = 1$. 32

4.4 Erro de treinamento em função de α para a diluição fixa e $\gamma = 0.9$. 33

4.5 Erro de treinamento em função de Q para as duas diluições e $\alpha = 1.0$. . . . 33

4.6 Capacidade de armazenamento em função de Q para a diluição móvel. . . . 34

4.7 Capacidade de armazenamento em função de Q para a diluição fixa e $\gamma =$ 1.0, 0.9, 0.5, 0.0, ordenados de cima para baixo. 36

4.8 Erro de generalização em função de α para a diluição móvel e $\gamma = 1$. 37

4.9 Erro de generalização em função de α para a diluição móvel e $\gamma = 0.9$. . . . 39

4.10 Erro de generalização em função de Q para a diluição móvel e $\gamma = 0.9$. . . 39

4.11 Erro de generalização em função de α para a diluição fixa e $\gamma = 1$. 40

4.12 Erro de generalização em função de α para a diluição fixa e $\gamma = 0.9$. 41

4.13 Erro de generalização , $\epsilon'_g = \epsilon_g(Q) - \epsilon_g(Q = 1)$, para as duas diluições e $\alpha = 1$, onde $\epsilon_g(Q = 1) = 0.31$. 42

5.1 Erro de generalização calculado no formalismo canônico em função de α para a diluição fixa e $\gamma = 0.9$. 46

Resumo

Perceptrons são redes neurais sem retroalimentação onde os neurônios estão dispostos em camadas. O perceptron considerado neste trabalho consiste de uma camada de N neurônios sensores $S_i = \pm 1$; $i = 1, \ldots, N$ ligados a um neurônio motor σ através das conexões sinápticas (pesos) W_i; $i = 1, \ldots, N$ cujos valores restringimos a ± 1.

Utilizando o formalismo de Mecânica Estatística desenvolvido por Gardner (1988), estudamos os efeitos de eliminarmos uma fração de conexões sinápticas (diluição) nas capacidades de memorização e generalização da rede neural descrita acima. Consideramos também o efeito de ruído atuando durante o estágio de treinamento do perceptron.

Consideramos dois tipos de diluição : diluição móvel na qual os pesos são cortados de maneira a minimizar o erro de treinamento e diluição fixa na qual os pesos são cortados aleatoriamente. A diluição móvel, que modela lesões em cérebro de pacientes muito jovens, pode melhorar a capacidade de memorização e, no caso da rede ser treinada com ruído, também pode melhorar a capacidade de generalização . Por outro lado, a diluição fixa, que modela lesões em cérebros de pacientes adultos, sempre degrada o desempenho da rede, sendo seu principal efeito introduzir um ruído efetivo nos exemplos de treinamento.

Abstract

Perceptrons are layered, feed-forward neural networks. In this work we consider a perceptron composed of one input layer with N sensor neurons $S_i = \pm 1$; $i = 1, \ldots, N$ which are connected to a single motor neuron σ through the synaptic weights W_i; $i = 1, \ldots, N$, which are constrained to take on the values ± 1 only.

Using the Statistical Mechanics formalism developed by Gardner (1988), we study the effects of eliminating a fraction of synaptic weights on the memorization and generalization capabilities of the neural network described above. We consider also the effects of noise acting during the perceptron training stage.

We consider two types of dilution: annealed dilution, where the weights are cut so as to minimize the training error and quenched dilution, where the weights are cut randomly. The annealed dilution which models brain damage in very young patients can improve the memorization ability and, in the case of training with noise, it can also improve the generalization ability. On the other hand, the quenched dilution wich models lesions on adult brains always degrades the performance of the network, its main effect being to introduce an effective noise in the training examples.

Capítulo 1

Introdução

Uma das principais motivações para o estudo de redes neurais é a compreensão dos mecanismos de funcionamento do cérebro. É sabido que tarefas como memorização , aprendizado, reconhecimento de padrões, percepção e criatividade, entre outras, são por ele executadas automaticamente. Em princípio estas tarefas podem parecer triviais, porém não se sabe claramente como ele as executa.

O cérebro é composto por aproximadamente 10^{10} células especializadas chamadas neurônios, que têm um grau de interconectividade várias ordens de grandeza maior que de um super computador : cada neurônio está ligado em média a outros 10^4 neurônios. Basicamente o neurônio é composto por 3 partes: soma ou corpo celular, axônio e dendritos conforme esquematizado na figura (1.1).

O processamento de informação ou a comunicação entre os neurônios ocorre da seguinte forma: os sinais chegam ao corpo do neurônio pelos dendritos e lá são processados de maneira que se a somatória de todos os sinais for maior que um certo limiar intrínseco do neurônio, ele emite um pulso elétrico que se propaga através do axônio indo para outros

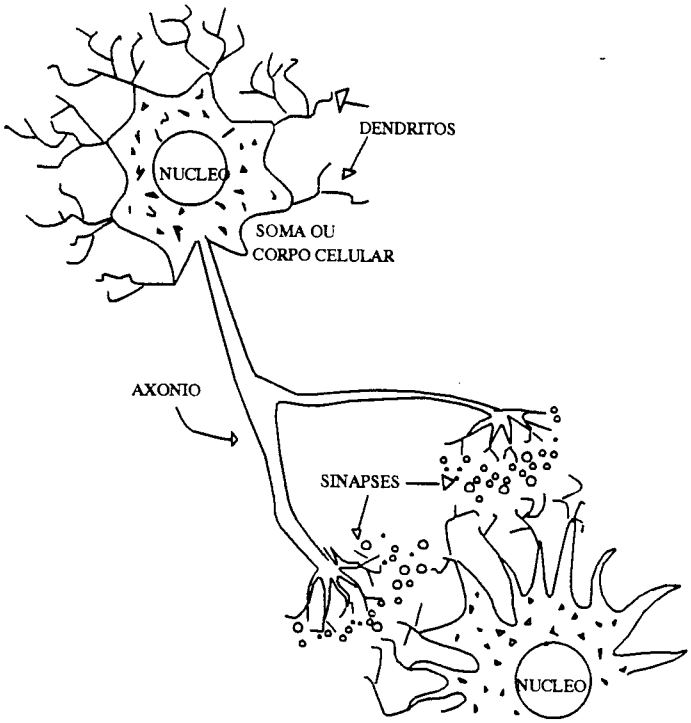

Figura 1.1: Diagrama esquemático de uma célula nervosa ou neurônio.

neurônios, senão ele permanace inativo. A conexão entre os neurônios é feita através das sinapses, que podem ser excitatórias ou inibitórias.

Por volta de 1960 foi estudado por Frank Rosenblatt (1960) um tipo de rede neural denominada *perceptron* que ele considerava ser um modelo simplificado dos mecanismos biológicos para o processamento de informações sensoriais. Apesar dessas redes serem extremamente simples comparadas com a complexidade do cérebro, elas serão usadas no estudo do comportamento deste. Afinal, segundo afirmação de Marvin Minsky, um dos maiores opositores do estudo deste tipo de rede na década de 60, seria surpreendente, dada a simplicidade do perceptron, que a natureza não fizesse uso dele em algum lugar. De fato essa rede nos permite modelar tarefas que o cérebro realmente executa como memorização e generalização . Independentemente do aspecto biológico, o perceptron tem tido grande aplicação em áreas tecnológicas como engenharia de controle (Nguyen e Widrow 1990) e

inteligência artificial (Sejnowski e Rosenberg 1987).

O perceptron é uma rede em camadas sem retroalimentação , onde a informação flui em uma única direção através de camadas distintas de neurônios. Esta rede é constituída por uma camada de entrada (formada por neurônios sensores os quais não estão ligados entre si) que recebe estímulos externos, algumas camadas intermediárias que processam a informação e uma camada de saída (formada por neurônios motores). As camadas estão ligadas pelas sinapses cujas eficiências (pesos) são ajustáveis. Neste trabalho será considerado o perceptron mais simples, composto por uma camada de entrada com N neurônios e uma camada de saída com apenas um neurônio. A fim de modelarmos os elementos biológicos básicos, levamos em conta os fatos de que os neurônios ou emitem pulsos elétricos ou estão inativos e as sinapses podem ser inibitórias ou excitatórias. Assim, tanto os neurônios de entrada como o de saída só podem assumir dois estados, ativo $(+1)$ e inativo (-1). As sinapses responsáveis pela ligação dos N neurônios de entrada com o neurônio de saída assumem valores positivos (excitatória) e negativos (inibitória). Em terminologia matemática podemos definir um perceptron como uma rede que soma todos os sinais da camada de entrada com um determinado peso e fornece o resultado através da função de transferência $g(x)$

$$\sigma = g\left(\sum_{i=1}^{N} W_i\,(S_i + 1) - U\right) \tag{1.1}$$

onde W_i é o peso da conexão entre o neurônio sensor i e o neurônio motor, $S_i = \pm 1$ representa o estado do neurônio sensor i e σ o estado do neurônio motor. Aqui, U é o limiar do neurônio sensor que neste trabalho tomaremos como sendo $U = \sum_{i=1}^{N} W_i$. Ao vetor de N componentes $\vec{S} = (S_1, S_2, \ldots, S_N)$ denominamos padrão de entrada. Aos pares $\left(\vec{S}, \sigma\right)$ denominamos exemplos. As funções de transferência mais usadas são a função sinal $g(x) = \mathrm{sgn}(x)$, que será usada neste trabalho, e a função linear $g(x) = x$. Denominamos

perceptron booleano aquele cuja função de transferência é a função sinal e perceptron linear aquele cuja função de transferência é a linear.

Os perceptrons podem ser treinados para executarem determinadas tarefas, isto é, um conjunto de exemplos $\left(\vec{\xi}^l, t^l\right); l = 1, \ldots, P$, através do ajuste adequado dos pesos sinápticos. Assim, no contexto deste trabalho, aprender significa encontrar os valores corretos dos pesos que implementam a tarefa desejada sem alterar a arquitetura da rede neural. Isso leva em conta o fato biológico de que todas as ligações num cérebro adulto já estão formadas. Foi descoberto por Rosenblatt (1962) um procedimento de treinamento aplicado a um perceptron de uma camada de conexões com a mesma estrutura descrita acima. Esse procedimento é conhecido como Algorítmo do Perceptron e, segundo o Teorema de Convergência do Perceptron (Minsky e Paper 1969), se existir um conjunto de pesos que realize corretamente a tarefa desejada, este procedimento certamente o encontrará. Porém, em certos casos a convergência pode demorar a ser atingida. De fato, no Algorítmo do Perceptron os pesos crescem em módulo arbitrariamente, de maneira que ele só funciona corretamente se seus pesos forem números reais ou inteiros ilimitados. Por exemplo, no caso em que os pesos assumem apenas os valores ± 1, este algorítmo não converge. Portanto, encontrar uma rede neural que execute a tarefa desejada por meio da construção de algorítmos pode não ser a maneira mais adequada, pois o fato de um dado algorítmo não conseguir encontrar uma rede correta não implica em que esta rede não exista. Felizmente este problema pode ser tratado de uma maneira geral sob o ponto de vista da Mecânica Estatística que determina a região onde existem soluções para quaisquer tipos de restrições nos pesos sinápticos. A precursora dessa linha de pesquisa foi Elizabeth Gardner que mostrou que, no caso em que os P exemplos $\left(\vec{\xi}^l, t^l\right)$ são gerados aleatoriamente, existe uma solução com pesos contínuos apenas para $P < 2N$ (Gardner

1988). O mesmo formalismo aplicado para o problema em que os pesos estão restritos aos valores ± 1 mostra que existe solução apenas para $P < 0.83N$ (Krauth e Mézard 1989).

Neste trabalho estudaremos apenas redes neurais com pesos restritos a ± 1, pois este é o modelo mais simples que leva em conta a existência de sinapses excitatórias e inibitórias e, como veremos no decorrer deste trabalho, apresenta um comportamento bastante complexo. Ainda, o estudo de redes com pesos binários pode ser útil para o entendimento dos efeitos da necessária discretização dos pesos reais quando da implementação da rede neural em *hardware*. Finalmente, este modelo é passível de solução exata, o que não ocorre no modelo com pesos contínuos.

Os neurônios são células que não se regeneram, portanto o cérebro está constantemente perdendo um certo número deles. Mesmo assim esta perda não afeta seu bom funcionamento. As lesões também podem acarretar mal funcionamento ou perda de neurônios. O estudo de redes neurais diluídas (com lesões) pode ser útil para o entendimento da robusteza do cérebro contra o mal funcionamento de alguns elementos. Para modelarmos as lesões, devemos cortar uma fração dos pesos sinápticos antes do início do processo de treinamento. Este corte será feito de duas maneiras. A primeira, que denominamos diluição móvel, permite que a rede rearranje suas conexões de forma a minimizar o efeito da lesão, simulando assim a flexibilidade do cérebro infantil. Já a segunda, a qual denominamos diluição fixa, não possui essa flexibilidade modelando assim a rigidez de um cérebro adulto. Nosso principal interesse consiste em entender o comportamento de uma rede neural diluída no que concerne à realização de funções tais como memorização e generalização , ou seja, como ela é capaz de exposta a um certo número de exemplos inferir a regra geradora dos exemplos. A fim de estudarmos as capacidades de memorização e generalização da rede diluída, a qual denominamos estudante, precisamos de

uma outra rede chamada mestra, cuja função é gerar os exemplos. A rede estudante já diluída, é treinada em um determinado número de exemplos gerados pela rede mestra. Durante o treinamento a rede estudante deve encontrar um conjunto de pesos que melhor implemente esses exemplos. Uma vez encontrado esse conjunto, o desempenho da rede é testado em novos exemplos (generalização).

Além da diluição , introduzimos ruído durante o treinamento, de maneira que os padrões de entrada gerados pela rede mestra não são os mesmos apresentados à rede estudante. Assim, planejamos obter a dependência das capacidades de memorização e generalização da rede com o tamanho do conjunto de treinamento, e as intensidades da lesão e do ruído. Essa análise será realizada utilizando-se as técnicas de Mecânica Estatística introduzidas por Elizabeth Gardner (1988). Uma versão resumida do presente trabalho pode ser encontrada em Barbato e Fontanari (1993).

O restante dessa monografia está organizada da seguinte maneira:

(a) No capítulo 2 descrevemos detalhadamente o modelo a ser tratado, definindo os conceitos de generalização , diluição e ruído.

(b) O capítulo 3 está dividido em quatro partes: na primeira fazemos um estudo da Mecânica Estatística do modelo usando o formalismo microcanônico; nas segunda e terceira partes calculamos a entropia do modelo usando a prescrição simetria de réplicas para os casos das diluições móvel e fixa respectivamente; na quarta parte obtemos a expressão do erro de generalização .

(c) No capítulo 4 analisamos numericamente as soluções das equações de ponto de sela derivadas no capítulo 3 para obtermos as capacidades de memorização e generalização da rede diluída.

(d) No capítulo 5 listamos os principais resultados obtidos, fazemos uma breve análise

da aplicação do formalismo canônico para resolução do problema proposto e comentamos sobre nossas perspectivas futuras no estudo de redes neurais.

Capítulo 2

O Modelo

Vamos considerar redes que são compostas de N neurônios sensores binários $S_i = \pm 1$ $(i = 1, \ldots, N)$, N conexões sinápticas $W_i = \pm 1$ $(i = 1, \ldots, N)$ e um neurônio motor, cujo estado é dado por

$$\sigma = \text{sign}\left(N^{-1/2} \sum_{i=1}^{N} W_i S_i\right). \tag{2.1}$$

A arquitetura da rede é mostrada na figura (2.1).

A tarefa da rede neural é realizar um conjunto de P associações entrada/saída (exemplos):

$$\vec{\xi}^l = \left(\xi_1^l, \xi_2^l, \ldots, \xi_N^l\right) \to t^l \quad l = 1, \ldots, P \tag{2.2}$$

onde ξ_i^l e t^l são também variáveis binárias, podendo assumir os valores ± 1. A este conjunto de exemplos chamamos *conjunto de treinamento.* Como veremos no desenvolvimento da Mecânica Estatística do modelo, P deve escalar linearmente com N, isto é, $P = \alpha N$ a fim de que o limite termodinâmico $N \to \infty$ seja bem definido. O desempenho da rede é

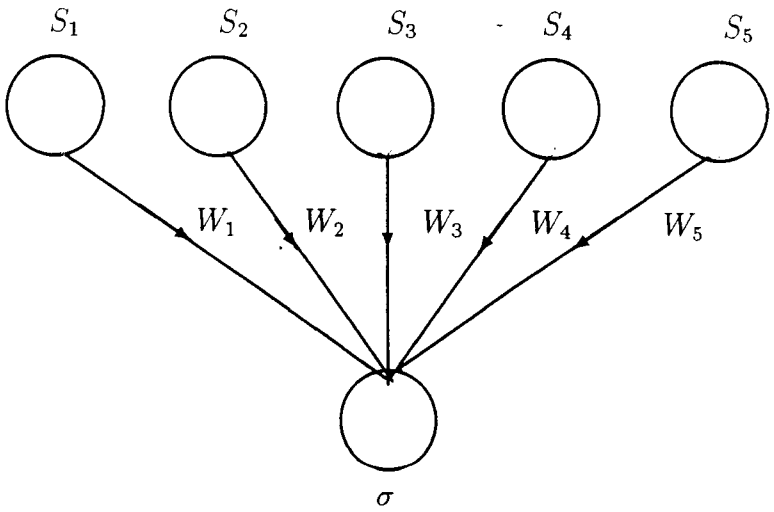

Figura 2.1: Estrutura da rede neural com $N = 5$.

medido pela *energia de treinamento*

$$E(\vec{W}, \vec{\xi}, t) = \sum_{l=1}^{P} \Theta\left(-t^l \, \sigma\left(\vec{W}, \vec{\xi}^l\right)\right) \tag{2.3}$$

onde $\sigma\left(\vec{W}, \vec{\xi}^l\right)$ é a resposta da rede neural com pesos $\vec{W}$ ao estímulo ξ^l. Aqui, $\Theta(x)$ é a função de degrau: $\Theta(x) = 1$ se $x > 0$ e $\Theta(x) = 0$ caso contrário. Assim, dizemos que treinar a rede é encontrar um conjunto de pesos que minimize a energia de treinamento. Note que o valor mínimo da energia de treinamento ($E = 0$) ocorre quando $\sigma^l = t^l$ para $l = 1, \ldots, P$, ou seja, a rede neural realiza com perfeição as associações (2.2). Entretanto, dependendo do número de exemplos, pode não existir um conjunto de pesos $\vec{W}$ para o qual a energia de treinamento se anule. De fato, podemos definir a capacidade de armazenamento α_c como a razão entre o número máximo de exemplos para o qual existe pelo menos uma rede neural com $E = 0$ e o número de neurônios da camada de entrada N.

O modelo acima foi introduzido e estudado por Gardner e Derrida (1988) para o caso

em que não existam correlações entre os padrões de entrada ξ^l e suas respectivas saídas t^l. A este tipo de tarefa denominamos *mapa aleatório*. Naquela análise foram utilizadas técnicas de mecânica estatística desenvolvidas no estudo de sistemas desordenados (vidros de spin), mais especificamente o método das réplicas. No limite termodinâmico, $N \to \infty$, a capacidade de armazenamento encontrada foi $\alpha_c = 4/\pi$. Porém a este resultado está associada uma entropia negativa, que é fisicamente inaceitável, pois a entropia de um sistema com um número finito e discreto de graus de liberdade deve ser sempre não negativa. Outra evidência de que a estimativa de α_c dada acima é incorreta provém de argumentos de teoria de informação : uma rede com N pesos binários não pode armazenar mais que N bits de informação e, tratando-se de um mapa aleatório, são necessários pelo menos P bits para armazenar as P associações . Portanto, $\alpha_c < 1$. A razão por que o cálculo de Gardner e Derrida não é correto está no uso da prescrição simetria de réplicas para descrever a estrutura dos parâmetros de ordem do sistema.

A fim de determinar o valor correto de α_c, ainda para o caso do mapa aleatório, Krauth e Opper (1989) realizaram simulações para $N = 25$ obtendo um valor próximo a 0.82. Esta questão foi finalmente resolvida analiticamente por Krauth e Mézard (1989) que, empregando o prescrição de Parisi para quebrar a simetria de réplicas, obtiveram $\alpha_c = 0.83$. É interessante notar que exatamente para este valor de α, a entropia calculada com a prescrição simetria de réplicas anula-se.

2.1 Generalização

Para o mapa aleatório não existe uma regra que associe os padrões de entrada $\vec{\xi}^l$ ao bit de saída t^l. Naturalmente, o caso mais interessante é quando existe uma correlação entre

entrada e saída. Essa correlação pode ser modelada por uma função Booleana, $t^l = f(\vec{\xi}^l)$, que reconhecemos como sendo a regra geradora do mapa entrada-saída. Vamos considerar a seguinte forma particular para f:

$$t^l = \text{sign}\left(N^{-1/2}\sum_{i=1}^{N} W_i^0 \xi_i^l\right) \qquad l = 1, \ldots, P \tag{2.4}$$

onde $W_i^0 = \pm 1$ podem ser identificados como os pesos de uma outra rede neural que denominamos rede *mestra*. Consequentemente, chamaremos *estudante* á rede que deve realizar o mapa gerado pela rede mestra. Note que uma dada rede mestra gera não apenas as P associações acima, mas todas as 2^N possíveis associações entrada/saída. Uma questão interessante, levantada pela primeira vez por Gardner e Derrida (1989), é o desempenho da rede estudante em responder corretamente a exemplos que não pertençam ao conjunto de treinamento. Com esta intenção definimos o *erro de generalização* E_g como

$$E_g\left(\vec{W}\right) = \int \prod_i d\vec{\xi}_i P\left(\vec{\xi}\right) \Theta\left(-t\,\sigma\left(\vec{W}, \vec{\xi}\right)\right), \tag{2.5}$$

onde $\sigma\left(\vec{W}, \vec{\xi}\right)$ é a resposta da rede estudante ao padrão de entrada $\vec{\xi}$ e $P(\xi_i)$ é a distribuição de probabilidade dos padrões de entrada. Vamos supor que as N componentes ξ_i sejam variáveis aleatórias estatisticamente independentes distribuídas de acordo com a seguinte distribuição de probabilidade

$$P\left(\vec{\xi}\right) = \prod_i p(\xi_i) = \prod_i \left[\frac{1}{2}\delta(\xi_i + 1) + \frac{1}{2}\delta(\xi_i - 1)\right]. \tag{2.6}$$

A contribuição dos $P = \alpha N$ padrões de entrada pertencentes ao conjunto de treinamento para a média na equação (2.6) é desprezível no limite termodinâmico, uma vez que esta média é realizada sobre todos os 2^N possíveis padrões de entrada. Portanto E_g mede de fato o desempenho da rede na classificação de um exemplo novo. Este modelo

foi introduzido e estudado numericamente por Gardner e Derrida (1989). Esses autores observaram que existe um valor máximo $P = 1.35\ (\pm 0.10)N$ a partir do qual o erro de generalização é nulo. Assim, existe um número mínimo de exemplos com os quais a rede estudante deve ser treinada para que realize perfeitamente o mapa completo. A partir deste valor o treinamento se torna redundante. Este tipo de comportamento não ocorre para redes com pesos contínuos, sendo neste caso necessária a apresentação de infinitos exemplos para que o erro de generalização se anule (Györgyi e Tishby 1989). O modelo com pesos binários foi estudado analiticamente por Györgyi (1990) que confirmou os resultados numéricos de Gardner e Derrida (1989), obtendo que o regime de generalização perfeita ocorre para $P > 1.245N$.

2.2 Lesões

Uma lesão cerebral pode ocasionar mal funcionamento de alguns elementos (neurônios, sinapses, etc ...) ou destruição total destes elementos. Vamos nos restringir a lesões que provocam destruição de um certo número de sinapses. Este tipo de processo pode ser modelado estudando-se os efeitos da diluição dos pesos em uma rede neural. Para o tipo particular de arquitetura considerada, figura (2.1), a eliminação do peso W_i implica no desligamento total do neurônio sensor S_i do resto da rede, isto é, para efeitos práticos implica na destruição deste neurônio.

Acredita-se que cada hemisfério do cérebro controle modos diferentes de pensamento: o hemisfério esquerdo realiza pensamentos analíticos e lógicos, sendo em particular responsável pela linguagem; enquanto que o direito é o hemisfério não-científico, responsável pela criação artística . Estas duas metades podem funcionar independentemente. Assim,

quando uma criança sofre algum tipo de lesão em um dos hemisférios do cérebro, o outro hemisfério tenta compensar estas perdas assumindo as funções que foram perdidas. No entanto não podemos saber se esta compensação é completa, pois para isto seria necessário saber como a criança se desenvolveria caso não ocorressem tais lesões, mas podemos afirmar que há uma boa compensação . Isto não ocorre, porém, em pessoas com idades mais avançadas. Assim quanto mais nova for uma pessoa, maior é a capacidade do cérebro de se recuperar (Lindsay e Norman 1977).

As lesões acima podem ser modeladas por dois tipos de diluição , as quais denominamos diluição móvel e diluição fixa[1] (Bouten *et al* 1990a). A primeira é equivalente às lesões sofridas por pessoas com pouca idade, de forma que o sistema neural consegue amenizar o efeito da lesão através de uma redistribuição de funções . A segunda equivale ao caso em que o cérebro já tenha perdido essa flexibilidade.

Numa rede neural em geral a maneira de se executar uma diluição é fazer com que uma fração de seus pesos assumam o valor 0, isto é, deixem de existir. Portanto, numa rede binária diluída os valores que os pesos podem assumir são 0, $+1$ e -1. Após a diluição , a conectividade da rede é medida pelo parâmetro

$$Q = \frac{1}{N}\sum_{i=1}^{N} W_i^2, \tag{2.7}$$

sendo o grau de diluição dado por $1 - Q$. Naturalmente, os NQ pesos restantes podem assumir apenas os valores ± 1.

A diferença entre as diluições móvel e fixa está no modo empregado para cortar os pesos. Na diluição móvel, a escolha das conexões a serem cortadas é feita de forma a minimizar a energia de treinamento. Por outro lado, na diluição fixa os pesos são cortados de maneira aleatória, não levando em conta a energia de treinamento.

[1]Tradução livre dos termos "annealed dilution" e "quenched dilution" respectivamente.

SERVIÇO DE BIBLIOTECA

O estudo dos efeitos de diluição em uma rede neural binária foi realizado por Bouten *et al* (1990b) para o caso do mapa aleatório. Esses autores calcularam a capacidade de armazenamento α_c em função da conectividade Q, encontrando que esta capacidade atinge seu valor máximo ($\alpha_c = 1.17$) para $Q = 0.63$. Note que a capacidade da rede não diluída ($Q = 1$) é $\alpha_c = 0.83$.

2.3 Ruído

Ruído é qualquer tipo de perturbação que deturpa a comunicação de maneira que a mensagem recebida não seja exatamente igual a mensagem emitida. No nosso caso, a informação é gerada e emitida pela rede mestra, sendo recebida e processada pela rede estudante. Vamos supor que cada componente dos padrões de entrada ξ_i^l gerados pela rede mestra seja corrompida de modo que o padrão de entrada que a rede estudante recebe seja $\vec{S}^l$, cujas componentes são extraídas da distribuição de probabilidade condicional

$$P(S_i^l|\xi_i^l) = \frac{1+\gamma}{2}\,\delta(S_i^l - \xi_i^l) + \frac{1-\gamma}{2}\,\delta(S_i^l + \xi_i^l). \tag{2.8}$$

A intensidade do ruído é então medida pelo parâmetro γ que pode variar entre 0 e 1. Quando γ assumir o valor 1, os padrões apresentados à rede estudante são os mesmos gerados pela rede mestra, ou seja, são padrões puros. Porém quando γ assumir o valor 0, recuperamos o problema do mapa aleatório. Assim, o mapa efetivo em que a rede estudante é treinada torna-se

$$\vec{S}^l = \left(S_1^l, S_2^l, \ldots, S_N^l\right) \to t^l \quad l = 1, \ldots, P, \tag{2.9}$$

de forma que a energia de treinamento passa a ser

$$E(\vec{W}, \vec{S}, t) = \sum_{l=1}^{P} \Theta\left(-t^l\,\sigma\left(\vec{W}, \vec{S}^l\right)\right) \tag{2.10}$$

onde $\sigma\left(\vec{W}, \vec{S}^l\right)$ é a resposta da rede neural ao padrão distorcido $\vec{S}^l$. Por outro lado, o erro de generalização permanece sendo medido pela equação (2.5). Assim, no teste da generalização os padrões de entrada são apresentados à rede estudante sem terem sido alterados pelo ruído. Neste estágio já podemos efetuar a média sobre $\vec{\xi}$ na equação (2.5). Para isto devemos lembrar que, de acordo com a equação (2.4), t é uma função de $\vec{\xi}$ e $\vec{W}^0$. Seguindo o apêndice C de Seung *et al* (1992) obtemos

$$E_g\left(\vec{W}\right) = \frac{1}{\pi} \arccos\left(\sqrt{Q}R\right), \tag{2.11}$$

onde Q é a conectividade da rede, equação (2.7), e R mede a correlação entre a rede mestra $\vec{W}^0$ e a rede estudante $\vec{W}$,

$$R = \frac{1}{NQ} \sum_{i=1}^{N} W_i W_i^O. \tag{2.12}$$

Vemos então que o erro de generalização E_g nada mais é que o ângulo formado pelos vetores $\vec{W}$ e $\vec{W}^0$.

Capítulo 3

A Mecânica Estatística do Modelo

Neste capítulo vamos estudar o modelo descrito no capítulo anterior dentro do formalismo do ensemble microcanônico. Nesse formalismo a grandeza fundamental é o número de redes neurais com uma dada energia de treinamento $\mathcal{N}(E)$. Para o caso de redes com pesos binários e conectividade Q temos $\mathcal{N}(E) \in \{0, 1, 2, \ldots, N_{\max}\}$, onde

$$N_{\max} = 2^{NQ}\ C_N^{N(1-Q)} \tag{3.1}$$

com $C_p^k = \frac{p!}{k!(p-k)!}$. Assim, a energia mínima de treinamento, $E_{\min}$, satisfaz a condição

$$\mathcal{N}(E < E_{\min}) = 0. \tag{3.2}$$

Para uma realização específica do conjunto de treinamento, o número de redes neurais $\mathcal{N}(E)$ com energia $E \geq 0$ pode ser calculado com o auxílio da função $Y_{\vec{W}}$ definida como

$$Y_{\vec{W}} = \begin{cases} 1 & \text{se } E(\vec{W}, \vec{S}, t) = E \\ 0 & \text{caso contrário,} \end{cases} \tag{3.3}$$

de modo que podemos escrever

$$\mathcal{N}(E) = \sum_{\{\vec{W}\}} Y_{\vec{W}} \tag{3.4}$$

onde $\sum_{\{\vec{W}\}}$ é a soma sobre todas as $N_{\max}$ redes possíveis.

A fim de eliminarmos a dependência de $\mathcal{N}(E)$ nas realizações dos exemplos de treinamento, devemos realizar médias sobre $\vec{S}^l$ e $\vec{\xi}^l$. Para isso seguiremos a prescrição de tomar médias apenas sobre quantidades extensivas pois estas se tornam automediantes no limite $N \to \infty$ (Binder e Young 1986). A rigor, o fato de grandezas extensivas serem automediantes só pode ser provado para sistemas com interações de curto alcance, no entanto acreditamos que isto também seja válido para sistemas com interações de alcance infinito como o que está sendo tratado neste trabalho. Definimos então a seguinte grandeza extensiva

$$\mathcal{S}(E) = \langle\langle \ln \mathcal{N}(E) \rangle\rangle \tag{3.5}$$

que identificamos com a entropia de Boltzman mediada sobre $\vec{S}^l$ e $\vec{\xi}^l$. O cálculo dessas médias, indicado pela notação $\langle\langle \ldots \rangle\rangle$, pode ser efetuado através do método das réplicas que consiste basicamente em usar a identidade

$$\langle\langle \ln \mathcal{N}(E) \rangle\rangle = \lim_{n \to 0} \frac{1}{n} \ln \langle\langle (\mathcal{N}(E))^n \rangle\rangle \tag{3.6}$$

sendo que calculamos $\langle\langle \mathcal{N}(E)^n \rangle\rangle$ para n *inteiro* e fazemos a continuação analítica para $n \approx 0$. Assim, para calcular a média sobre $\ln \mathcal{N}(E)$ devemos calcular o n-ésimo momento da distribuição de probabilidade de $\mathcal{N}(E)$. Note que a energia de treinamento dada pela equação (2.10) é função das variáveis aleatórias $\vec{S}^l$ e $\vec{\xi}^l$ de modo que tanto E como $\mathcal{N}$ são também variáveis aleatórias. Em particular, a energia de treinamento é, de acordo com a

definição de distribuição de probabilidade de uma função de variável aleatória, distribuída segundo a distribuição

$$P(E_t) = \langle\langle \delta(E_t - E(\vec{W}, \vec{S}, t))\rangle\rangle. \tag{3.7}$$

Para ilustrar o cálculo de $\langle\langle(\mathcal{N}(E))^n\rangle\rangle$ consideremos o caso $n = 1$:

$$\begin{aligned}
\langle\langle\mathcal{N}(E)\rangle\rangle &= \langle\langle \sum_{\{\vec{W}\}} Y_{\vec{W}}\rangle\rangle \\
&= \sum_{\{\vec{W}\}} (0 \times \text{Prob}\,(E_t \neq E) + 1 \times \text{Prob}\,(E_t = E)) \\
&= \langle\langle \sum_{\{\vec{W}\}} \delta(E - E(\vec{W}, \vec{S}, t))\rangle\rangle.
\end{aligned} \tag{3.8}$$

A generalização para o cálculo do n-ésimo momento é imediata:

$$\begin{aligned}
\langle\langle(\mathcal{N}(E))^n\rangle\rangle &= \langle\langle \prod_{a=1}^{n} \sum_{\{\vec{W}^a\}} Y_{\vec{W}^a}\rangle\rangle \\
&= \sum_{\{\vec{W}^1\}} \cdots \sum_{\{\vec{W}^n\}} \text{Prob}\left(E_t^1 = E, \ldots, E_t^n = E\right) \\
&= \langle\langle \prod_{a=1}^{n} \sum_{\{\vec{W}^a\}} \delta(E - E(\vec{W}^a, \vec{S}, t))\rangle\rangle,
\end{aligned} \tag{3.9}$$

onde $\text{Prob}(E_t^1 = E, \ldots, E_t^n = E)$ é a distribuição de probabilidade conjunta das redes $\vec{W}^1, \ldots, \vec{W}^n$ terem energia de treinamento igual a E.

Usando a representação integral da função delta obtemos

$$\langle\langle(\mathcal{N}(E))^n\rangle\rangle = \int_{-i\infty}^{i\infty} \prod_{a=1}^{n} \frac{d\hat{\epsilon}_a}{2\pi i} \exp N \left(\alpha\epsilon_t \sum_{a=1}^{n} \hat{\epsilon}_a + \frac{1}{N} \ln\langle\langle(Z(\hat{\epsilon}, \mathbf{S}, t))^n\rangle\rangle\right), \tag{3.10}$$

onde

$$Z(\hat{\epsilon}, \vec{S}, t) = \sum_{\{\vec{W}\}} \exp(-\hat{\epsilon} E(\vec{W}, \vec{S}, t)) \tag{3.11}$$

é função de partição canônica. O parâmetro $\hat{\epsilon}$ faz então o papel do inverso da temperatura T. Definimos também o *erro de treinamento*

$$\epsilon_t = \frac{E}{\alpha N} \tag{3.12}$$

que assume o valor mínimo $\epsilon_t = 0$ quando a rede estudante consegue realizar perfeitamente o conjunto de treinamento. O pior desempenho ocorre para $\epsilon_t = 1/2$ que corresponde a escolher as saídas de forma aleatória, independentemente das entradas. Note que a uma rede neural $\vec{W}$ com $\epsilon_t = \epsilon^* \geq 1/2$ está associada uma rede simétrica $-\vec{W}$ com $\epsilon_t = 1 - \epsilon^* \leq 1/2$ de forma que o caso $\epsilon_t \geq 1/2$ está intimamente ligado ao caso $\epsilon_t \leq 1/2$. A conexão com o formalismo do ensemble canônico é feita através da relação termodinâmica

$$\frac{\partial \mathcal{S}}{\partial E} = \frac{1}{T}. \tag{3.13}$$

Existem $\mathcal{N}(\epsilon_t)$ redes neurais $\vec{W}$ com erro de treinamento ϵ_t, mas a cada uma delas corresponde um erro de generalização $E_g\left(\vec{W}\right)$ diferente. A grandeza que nos interessa é o erro de generalização *médio* do ensemble de redes neurais com erro de treinamento ϵ_t definido como

$$\epsilon_g = \langle\langle \frac{\sum_{\{\vec{W}\}} E_g\left(\vec{W}\right) \delta\left(\alpha N \epsilon_t - E\left(\vec{W}, \vec{S}^l, t^l\right)\right)}{\sum_{\{\vec{W}\}} \delta\left(\alpha N \epsilon_t - E\left(\vec{W}, \vec{S}^l, t^l\right)\right)} \rangle\rangle. \tag{3.14}$$

Para efetuar as médias na equação acima usamos a identidade

$$\frac{1}{\mathrm{x}} = \lim_{n \to 0} \mathrm{x}^{n-1} \tag{3.15}$$

sendo que calculamos x^{n-1} para n *inteiro* e fazemos a continuação analítica para $n \approx 0$. Deste modo obtemos

$$\epsilon_g = \lim_{n\to 0} \langle\langle \prod_{a=1}^{n} \sum_{\{\vec{W}^a\}} E_g\left(\vec{W}^1\right) \delta\left(\alpha N \epsilon_t - E\left(\vec{W}, \vec{S}^l, t^l\right)\right) \rangle\rangle \tag{3.16}$$

Nas seções seguintes efetuaremos explicitamente os cálculos de $\langle\langle \ln \mathcal{N}(\epsilon_t)\rangle\rangle$, equação (3.6), e ϵ_g, equação (3.16) para os dois tipos de diluição descritos no capítulo anterior.

3.1 Diluição Móvel

Neste caso os pesos podem assumir os valores $0, \pm 1$ de forma que o vínculo (2.7) deve ser imposto através da função delta de Kronecker:

$$\delta\left(Q, \frac{1}{N}\sum_i W_i^2\right) = \int_{-i\pi}^{i\pi} \frac{d\hat{Q}}{2\pi i} \exp\left(-\hat{Q}\left(Q - \frac{1}{N}\sum_i W_i^2\right)\right). \tag{3.17}$$

Para calcularmos $\mathcal{S}(\epsilon_t)$ devemos primeiramente efetuar as médias sobre o conjunto de treinamento, introduzindo os parâmetros de ordem que descrevem o sistema. As integrais sobre estes parâmetros podem ser calculadas exatamente no limite termodinâmico $N \to \infty$ pelo método de ponto do sela. Estes cálculos são mostrados em detalhes no apêndice A. Apresentamos aqui apenas o resultado final para a densidade média de entropia $s = \mathcal{S}/NQ$:

$$\begin{aligned} s(\epsilon_t) = \lim_{n\to 0} \max \frac{1}{n} \Bigg\{ & \alpha' \epsilon_t \sum_a \hat{\epsilon}_a - \sum_{a<b} q_{ab}\hat{q}_{ab} - \sum_a R_a \hat{R}_a \\ & - \sum_a \hat{Q}_a + \frac{1}{Q} G_0\left(\hat{q}_{ab}, \hat{R}_a, \hat{Q}_a\right) + \alpha' G_1(q_{ab}, R_a, \hat{\epsilon}_a) \Bigg\} \end{aligned} \tag{3.18}$$

onde

$$G_0 = \frac{1}{N}\sum_{i=1}^{N} \ln\left\{ \prod_{a=1}^{n} \sum_{\{W^a=0,\pm 1\}} \exp\left(\sum_{a<b} q_{ab} W^a W^b + \sum_a Q^a (W^a)^2 + \sum_a \hat{R}_a W^a W_i^0 \right) \right\} \tag{3.19}$$

e

$$G_1 = \ln \int_{-\infty}^{\infty} Dy \int_{-\infty}^{\infty} \prod_a \frac{dx_a d\tilde{x}_a}{2\pi} \exp\left[-\sum_a \hat{\epsilon}_a \Theta(-y x_a) - \frac{1}{2} \sum_a \tilde{x}_a^2 \left(1 - \gamma'^2 R_a^2\right) \right.$$
$$\left. - \sum_{a<b} \tilde{x}_a \tilde{x}_b \left(q_{ab} - \gamma'^2 R_a R_b\right) + i \sum_a \tilde{x}_a \left(x_a - y\gamma' R_a\right) \right]. \quad (3.20)$$

Introduzimos também as seguintes definições

$$Dy = \frac{dy}{\sqrt{2\pi}} e^{-y^2/2}, \quad (3.21)$$

$$\alpha' = \frac{\alpha}{Q}, \quad (3.22)$$

e

$$\gamma' = \gamma\sqrt{Q}. \quad (3.23)$$

Os parâmetros de ordem $(\hat{\epsilon}^a, \hat{q}_{ab}, \hat{R}^a, \hat{Q}^a, q_{ab}, R^a)$ são determinados impondo-se que a expressão dentro das chaves na equação (3.18) seja máxima. Note que no limite $n \to 0$ a entropia deve ser minimizada pelos parâmetros com dois índices de réplicas e, como usualmente, maximizada pelos parâmetros com apenas um índice de réplicas. Isto se deve ao fato da dimensão da matriz cujas componentes são q_{ab} tornar-se negativa nesse limite (Mézard *et al* 1987). Os parâmetros de ordem

$$q_{ab} = \frac{1}{NQ} \sum_{i=1}^{N} W_i^a W_i^b \qquad a \neq b \quad (3.24)$$

e

$$R_a = \frac{1}{NQ} \sum_{i=1}^{N} W_i^a W_i^0 \quad (3.25)$$

medem, respectivamente, a correlação entre duas redes diferentes ($\vec{W}^a$ e $\vec{W}^b$) com o mesmo erro de treinamento ϵ_t e a correlação de uma rede $\vec{W}^a$ com erro de treinamento ϵ_t com a rede mestra $\vec{W}^0$.

A fim de prosseguirmos os cálculos vamos usar a chamada prescrição simetria de réplicas, ou seja, vamos supor que os valores dos parâmetros de ordem sejam independentes dos seus índices de réplicas:

$$\begin{aligned} q_{ab} &= q \quad \text{e} \quad \hat{q}_{ab} = \hat{q} \quad \forall a < b; \\ R_a &= R \quad \text{e} \quad \hat{R}_a = \hat{R} \quad \forall a; \\ \hat{Q}_a &= \hat{Q} \quad \text{e} \quad \hat{\epsilon}_a = \hat{\epsilon} \quad \forall a. \end{aligned} \tag{3.26}$$

Usamos esta prescrição para calcular G_0 e G_1 e, no limite $n \to 0$, obter a seguinte expressão para densidade média de entropia (apêndices B e C):

$$\begin{aligned} s_{sr}(\epsilon_t) = \alpha'\epsilon_t\hat{\epsilon} + q\hat{q}/2 - R\hat{R} - \hat{Q} + 2\alpha' \int_{-\infty}^{\infty} DzH(\xi_1) \ln\left[e^{-\hat{\epsilon}} + (1 - e^{-\hat{\epsilon}})H(\xi_2)\right] \\ + \frac{1}{Q}\int_{-\infty}^{\infty} Dz \ln\left[1 + 2\exp\left(\hat{Q} - \hat{q}/2\right)\right) \cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)\right] \end{aligned} \tag{3.27}$$

onde

$$\xi_1 = z\sqrt{\frac{\gamma'^2 R^2}{q - \gamma'^2 R^2}}, \tag{3.28}$$

$$\xi_2 = z\sqrt{\frac{q}{1-q}} \tag{3.29}$$

e

$$H(x) = \int_x^{\infty} Dy. \tag{3.30}$$

Os parâmetros de ordem na prescrição simetria de réplicas $(\hat{\epsilon}, \hat{q}, \hat{R}, \hat{Q}, q, R)$ são dados pelas equações de ponto de sela. Para obtê-las devemos igualar a zero as derivadas de s_{rs} com relação a esses parâmetros, resultando nas seguintes equações :

$$Qq = \int Dz \left[\frac{2\exp\left(\hat{Q} - \hat{q}/2\right) \sinh\left(\hat{R} + z\sqrt{\hat{q}}\right)}{1 + 2\exp\left(\hat{Q} - \hat{q}/2\right) \cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)} \right]^2 , \tag{3.31}$$

$$QR = \int Dz \left[\frac{2\exp\left(\hat{Q} - \hat{q}/2\right) \sinh\left(\hat{R} + z\sqrt{\hat{q}}\right)}{1 + 2\exp\left(\hat{Q} - \hat{q}/2\right) \cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)} \right] , \tag{3.32}$$

$$Q = \int Dz \left[\frac{2\exp\left(\hat{Q} - \hat{q}/2\right) \cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)}{1 + 2\exp\left(\hat{Q} - \hat{q}/2\right) \cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)} \right] , \tag{3.33}$$

$$\hat{R} = -\frac{2\alpha' q}{\sqrt{2\pi} R(q - \gamma'^2 R^2)} \int Dz \; \xi_1 \; e^{-\xi_1^2/2} \ln\left[(e^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)\right] , \tag{3.34}$$

$$\hat{q} = \frac{R\hat{R}}{q} + \frac{2\alpha'}{\sqrt{2\pi} q(1-q)} \int Dz \; \xi_2 \; \frac{e^{-\xi_2^2/2} H(\xi_1)}{(e^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)} , \tag{3.35}$$

$$\epsilon_t = 2(e^{\hat{\epsilon}} - 1)^{-1} \int Dz \; H(\xi_1) \; \frac{1 - H(\xi_2)}{(e^{\hat{\epsilon}} - 1)^{-1} + H(\xi_2)} . \tag{3.36}$$

Para comprovar que a prescrição simetria de réplicas corresponde de fato a um máximo da expressão entre chaves na equação (3.18) devemos analisar as matrizes de derivadas segundas de G_0 e G_1 em relação a seus argumentos. A condição para estabilidade local do ponto de sela para a prescrição simetria de réplicas (de Almeida e Thouless 1978) é dada por

$$\alpha \gamma_0 \gamma_1 < 1 \tag{3.37}$$

onde γ_0 e γ_1 são os autovalores das matrizes $\frac{1}{2}n(n+3)$-dimensionais das derivadas segundas de G_0 e G_1 com relação a $\hat{q}_{ab}$ e q_{ab} respectivamente. Seguindo a análise feita por Gardner e Derrida (1988) obtemos os seguintes autovalores

$$\gamma_0 = 4\exp\left(2\hat{Q} - \hat{q}\right) \int Dz \left[\frac{2\exp\left(\hat{Q} - \hat{q}/2\right) + \cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)}{1 + 2\exp\left(\hat{Q} - \hat{q}/2\right)\cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)}\right]^2 \tag{3.38}$$

$$\gamma_1 = 2\int DzH(\xi_1)(\langle x^2\rangle - \langle x\rangle^2)^2 \tag{3.39}$$

onde

$$\langle x\rangle = \frac{i}{\sqrt{2\pi}(1-q)}\frac{e^{-\xi_2^2/2}}{(e^{\hat{\epsilon}}-1)^{-1} + H(\xi_2)} \tag{3.40}$$

e

$$\langle x^2\rangle = -\frac{1}{\sqrt{2\pi}(1-q)}\frac{\xi_2 e^{-\xi_2^2/2}}{(e^{\hat{\epsilon}}-1)^{-1} + H(\xi_2)}. \tag{3.41}$$

O sistema de equações acopladas (3.31)–(3.36) deve ser resolvido numericamente para ϵ_t, α, γ e Q fixos. O comportamento típico da densidade de entropia média em função do erro de treinamento é mostrado na figura (3.1).

A densidade de entropia aumenta com ϵ_t, atingindo seu valor máximo $\ln \mathcal{N}_{\max}$ para $\epsilon_t = 1/2$, onde começa a decrescer a medida que ϵ_t aumenta em direção a 1. Como a energia de treinamento definida na equação (2.3) satisfaz $E(-\vec{W}) = 1 - E(\vec{W})$, a densidade de entropia deve ser invariante com relação á reflexões em torno do ponto $\epsilon_t = 1/2$. Conforme discutimos anteriormente, a densidade de entropia não pode assumir valores negativos, porém a densidade de entropia calculada com a prescrição simetria de réplicas, $s_{sr}(\epsilon_t)$, pode tornar-se negativa. Assim definimos o estado de menor energia, isto é, o estado mais ordenado como sendo o que possui o valor mais baixo de $\epsilon_t \geq 0$ para o qual s_{sr} é positiva. A este valor mínimo do erro de treinamento denotamos $\epsilon_t^{\min}$. Se $\epsilon_t^{\min} = 0$, então $s_{sr}(\epsilon_t) \geq 0$ para todo ϵ_t e, portanto, a prescrição simetria de réplicas, no caso de ser estável, fornece um resultado exato. Como a densidade de entropia é não nula, o número de redes neurais com erro ϵ_t é $\mathcal{O}(e^N)$. Por outro lado, no caso em que $\epsilon_t^{\min} > 0$ temos $s_{sr}(\epsilon_t^{\min}) = 0$, sendo negativa para $\epsilon_t < \epsilon_t^{\min}$. No entanto se a solução com

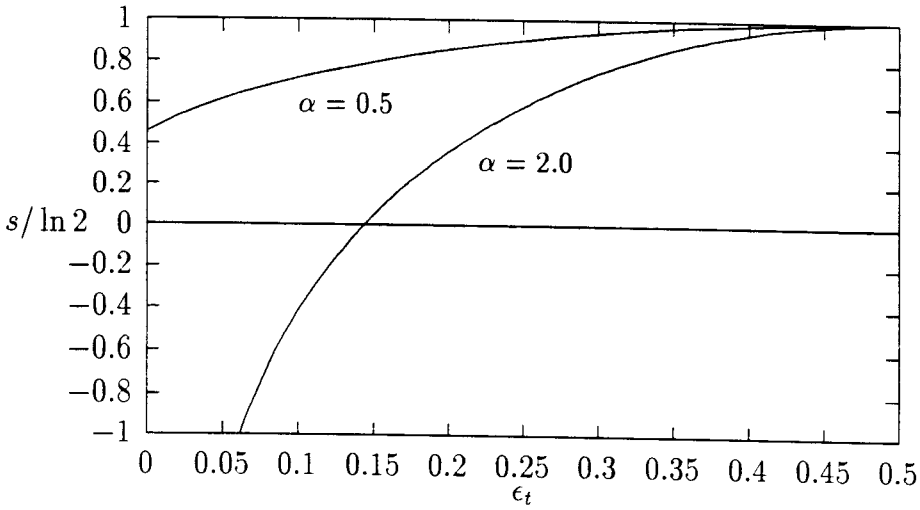

Figura 3.1: Entropia em função do erro de treinamento ϵ_t para $Q = 1$ e $\gamma = 0$.

simetria de réplicas for estável para $\epsilon_t = \epsilon_t^{\min}$, essa prescrição produz um estado ordenado que satisfaz todas as exigências físicas, de modo que acreditamos ser este o estado de mínima energia correto. Uma discussão mais completa deste ponto pode ser encontrada em Fontanari e Meir (1993) onde o método de cálculo utilizado acima foi proposto.

3.2 Diluição Fixa

Neste caso, $N(1-Q)$ pesos escolhidos aleatoriamente são cortados, isto é, assumem o valor zero. No entanto podemos re-indexar os neurônios de forma tal que $W_j = \pm 1$ ($j = 1, \ldots, NQ$) e $W_j = 0$ ($j = NQ+1, \ldots, N$). Assim o vínculo (2.7) é obedecido automaticamente, não sendo necessário o uso da delta de Kronecker. Seguindo o mesmo

procedimento anterior (apêndice A) obtemos a densidade média de entropia

$$s(\epsilon_t) = \lim_{n\to 0} \max \frac{1}{n} \left\{ \alpha' \epsilon \sum_a \hat{\epsilon}_a - \sum_{a<b} q_{ab}\hat{q}_{ab} - \sum_a R_a \hat{R}_a + G_0\left(\hat{q}_{ab}, \hat{R}_a\right) + \alpha' G_1\left(q_{ab}, R_a, \hat{\epsilon}_a\right) \right\} \tag{3.42}$$

onde

$$G_0 = \frac{1}{NQ} \sum_{i=1}^{NQ} \ln \left\{ \prod_{a=1}^{n} \sum_{\{W^a=\pm 1\}} \exp\left(\sum_{a<b} q_{ab} W^a W^b + \sum_a \hat{R}_a W^a W_i^0 \right) \right\} \tag{3.43}$$

e G_1 é dado pela equação (3.20). A equação (3.42) expressa a densidade de entropia de uma rede com $N' = NQ$ neurônios treinada com $P = \alpha' N'$ pares entrada/saída $\left(\vec{S}^l, t^l\right)$ onde os padrões ruidosos originam-se da distribuição de probabilidade dada pela equação (2.8) com γ substituído por $\gamma' = \gamma\sqrt{Q}$. Assim, os principais efeitos da diluição fixa são re-escalar o número de neurônios e introduzir um ruído efetivo no processo de treinamento através da re-escala do parâmetro de ruído γ.

Usando a prescrição simetria de réplicas para calcular G_0 (apêndice B) e tomando o limite $n \to 0$ re-escrevemos a equação (3.42) como

$$\begin{aligned} s_{sr}(\epsilon_t) = \alpha'\epsilon\hat{\epsilon} - \frac{\hat{q}}{2}(1-q) - R\hat{R} + \int_{-\infty}^{\infty} Dz \ln\left[2\cosh\left(\hat{R} + z\sqrt{\hat{q}}\right)\right] \\ +2\alpha' \int_{-\infty}^{\infty} Dz\ H(\xi_1) \ln\left[e^{-\hat{\epsilon}} + \left(1 - e^{-\hat{\epsilon}}\right) H(\xi_2)\right], \end{aligned} \tag{3.44}$$

onde ξ_1 e ξ_2 são dados pelas equações (3.28) e (3.29). Os parâmetros de ordem $\hat{R}, \hat{q}, \hat{\epsilon}$ permanecem inalterados sendo dados pelas equações (3.34), (3.35) e (3.36), enquanto que os outros parâmetros são dados por

$$q = \int Dz \tanh^2\left(\hat{R} + z\sqrt{\hat{q}}\right) \tag{3.45}$$

e

$$R = \int Dz \tanh\left(\hat{R} + z\sqrt{\hat{q}}\right). \tag{3.46}$$

Analogamente à diluição móvel, a solução com simetria de réplicas é estável sempre que a condição (3.37) seja satisfeita, com γ_1 ainda dado pela equação (3.39) e γ_0 dado por

$$\gamma_0 = \int Dz \left[1 - \tanh^2\left(\hat{R} + z\sqrt{\hat{q}}\right)\right]^2 . \tag{3.47}$$

Para se obter o estado de mínima energia aplica-se o mesmo método usado na seção anterior, sendo válidos para esta diluição todos os comentários daquela seção .

3.3 Erro de Generalização Médio

Dando continuidade ao cálculo da equação (3.16) e utilizando as mesmas técnicas das seções anteriores obtemos, para o caso da diluição móvel,

$$\begin{aligned}\epsilon_g &= \lim_{n\to 0} \int \prod_{a=1}^{n} \frac{d\hat{\epsilon}^a}{2\pi i} \int \prod_{a<b} \frac{d\hat{q}_{ab} dq_{ab}}{2\pi/NQi} \int \prod_a \frac{d\hat{R}_a dR_a}{2\pi/NQi} \int \prod_a \frac{d\hat{Q}_a}{2\pi/Ni} \\ &\frac{1}{\pi} \arccos\left(\sqrt{Q}R_1\right) \exp N \left[\alpha\epsilon_t \sum_a \hat{\epsilon}_a - Q \sum_{a<b} q_{ab}\hat{q}_{ab} - Q \sum_a R_a \hat{R}_a \right. \\ &\left. -Q \sum_a \hat{Q}_a + G_0\left(\hat{q_{ab}}, \hat{R}_a, \hat{Q}_a\right) + \alpha G_1\left(q_{ab}, R_a, \hat{\epsilon}_a\right)\right]\end{aligned} \tag{3.48}$$

onde G_0 e G_1 são dados pelas equações (3.19) e (3.20), respectivamente. Para efetuarmos a integral acima devemos aplicar o método da integração por ponto de sela

$$\mathcal{I}(N) = \int_C g(t) e^{NF(t)} dt = \lim_{N\to\infty} \sqrt{\frac{-2\pi}{NF''(t_0)}}\, g(t_0)\, e^{NF(t_0)} \tag{3.49}$$

onde t_0 é determinado pela equação $F'(t_0) = 0$. Obtemos então

$$\begin{aligned}\epsilon_g &= \lim_{n\to 0} \frac{1}{\pi} \arccos\left(\sqrt{Q}R_1\right) \exp N \max \left\{\alpha'\epsilon_t \sum_a \hat{\epsilon}_a - \sum_{a<b} q_{ab}\hat{q}_{ab} - \sum_a R_a \hat{R}_a \right. \\ &\left. -\sum_a \hat{Q}_a + \frac{1}{Q} G_0\left(\hat{q_{ab}}, \hat{R}_a, \hat{Q}_a\right) + \alpha' G_1\left(q_{ab}, R_a, \hat{\epsilon}_a\right)\right\} .\end{aligned} \tag{3.50}$$

Usando a prescrição simetria de réplicas, podemos re-escrever a equação acima como

$$\begin{aligned}\epsilon_g &= \lim_{n\to 0}\frac{1}{\pi}\arccos\left(\sqrt{Q}R_1\right)\exp Nn\max\Big\{\alpha'\epsilon_t\hat{\epsilon}+q\hat{q}/2-R\hat{R}-\hat{Q}\\ &+2\alpha'\int DzH(\xi_1)\ln\left[e^{-\hat{\epsilon}}+(1-e^{-\hat{\epsilon}})H(\xi_2)\right]\\ &+\frac{1}{Q}\int Dz\ln\left[1+2\exp\left(\hat{Q}-\hat{q}/2\right)\cosh\left(\hat{R}+z\sqrt{\hat{q}}\right)\right]\Big\}. \qquad (3.51)\end{aligned}$$

Assim, no limite $n \to 0$, obtemos a forma final do erro de generalização médio

$$\epsilon_g = \frac{1}{\pi}\arccos\left(\sqrt{Q}R\right) \qquad (3.52)$$

sendo que o valor de R é dado pelas equações de ponto de sela (3.31)-(3.36) calculadas para valores fixos de Q,α, ϵ_t e γ.

No caso da diluição fixa, o procedimento para calcular ϵ_g é o mesmo usado aqui, exceto que valor de R é dado pelas equações de ponto de sela (3.45)-(3.46) referentes àquela diluição .

Capítulo 4

Análise dos Resultados

Neste capítulo analisamos a solução numérica das equações de ponto de sela a fim de obtermos as capacidades de memorização e generalização da rede estudante, quando seu treinamento é feito sob condições que variam a intensidade do ruído e a conectividade da rede. Basicamente serão analisados os parâmetros que medem o desempenho da rede: o erro de treinamento mínimo $\epsilon_t^{\min}$, a capacidade de armazenamento α_c e o erro de generalização médio ϵ_g calculado para $\epsilon_t = \epsilon_t^{\min}$. Para simplificar a notação vamos nos referir a $\epsilon_t^{\min}$ como erro de treinamento ϵ_t e ao erro de generalização médio como erro de generalização simplesmente. Todas as análises serão feitas para as diluições fixa e móvel.

4.1 Erro de Treinamento

Vamos analisar primeiramente o erro de treinamento, porque a partir dele é que podemos determinar a capacidade de armazenamento e o erro de generalização da rede estudante.

Para iniciar nossa análise começamos com o caso em que a rede é treinada na ausência

de ruído, isto é, $\gamma = 1$. A figura (4.1) mostra o erro de treinamento em função de α para as conectividades $Q = 0.5$, 0.8 e 0.98, para o caso da diluição móvel. Quando a

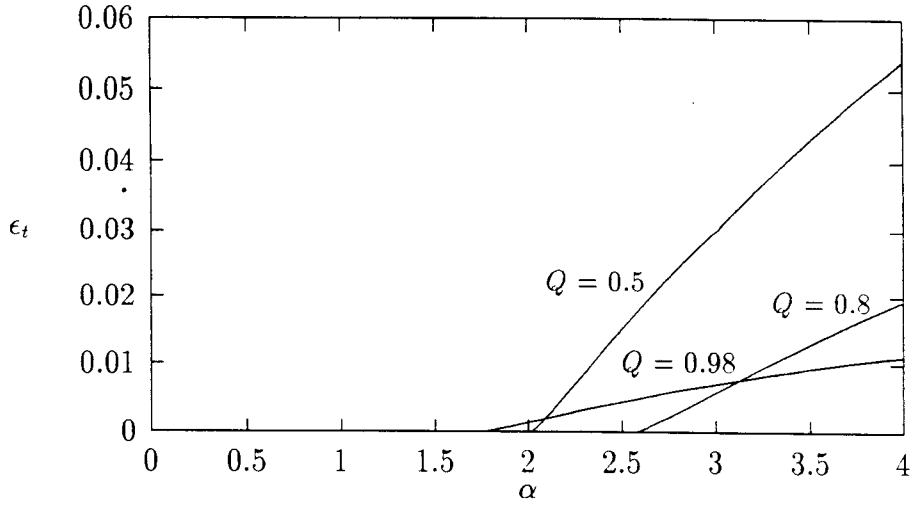

Figura 4.1: Erro de treinamento em função de α para a diluição móvel e $\gamma = 1$.

rede está totalmente conectada ($Q = 1$), ϵ_t é nulo seja qual for o tamanho do conjunto de treinamento. Isto acontece porque sempre existem redes capazes de realizar corretamente um determinado mapa, sendo que a rede mestra é apenas uma dentre as possibilidades. À medida que aumentamos o conjunto de treinamento, há uma diminuição no número dessas redes (a entropia diminui com o aumento de α), porém, mesmo assim, o erro de treinamento continua sendo zero. No caso de $Q < 1$ o erro de treinamento deixa de ser zero a partir de α_c sendo que para cada valor de α existe um valor de Q para o qual ϵ_t é mínimo. A partir de α_c quanto maior for a diluição (menor Q), mais rapidamente o desempenho da rede degrada com o aumento de α. No limite de α muito grande ϵ_t atinge

seu valor máximo (este resultado será explicado na seção (4.3))

$$\epsilon_t \approx \frac{1}{\pi} \arccos\left(\gamma\sqrt{Q}\right) \qquad \alpha \to \infty. \tag{4.1}$$

Assim, neste limite quanto menos diluída estiver a rede (maior Q) melhor o seu desempenho. No caso da rede ser treinada com ruído seu melhor rendimento também ocorre quando ela estiver menos diluída, porém, para valores finitos de α, o erro de treinamento é maior quando a rede está totalmente conectada ($Q = 1$), como podemos ver na figura (4.2). Na diluição fixa o comportamento é análogo ao da diluição móvel: a partir de α_c,

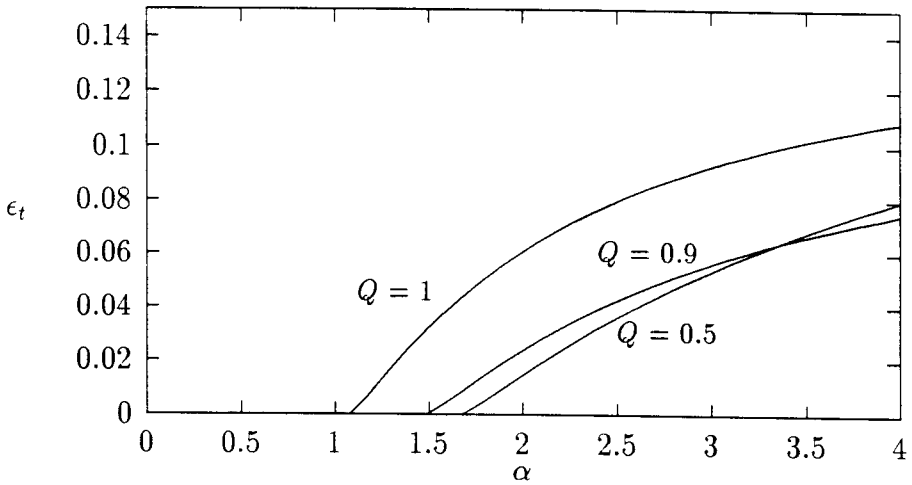

Figura 4.2: Erro de treinamento em função de α para a diluição móvel e $\gamma = 0.9$.

quanto menor for o valor de Q, mais rápida é a degradação do desempenho da rede e, no limite $\alpha \to \infty$, seu comportamento é regido também pela equação (4.1). Note que na diluição fixa ϵ_t se aproxima mais rápido do seu valor máximo que na diluição móvel. A fim de ilustrarmos este comportamento mostramos os gráficos de ϵ_t em função de α para

$\gamma = 1$ e $Q = 0.5,\ 0.8,\ 0.98$ (figura (4.3)) e $\gamma = 0.9$ e $Q = 0.5,\ 0.9,\ 1.0$ (figura (4.4)).

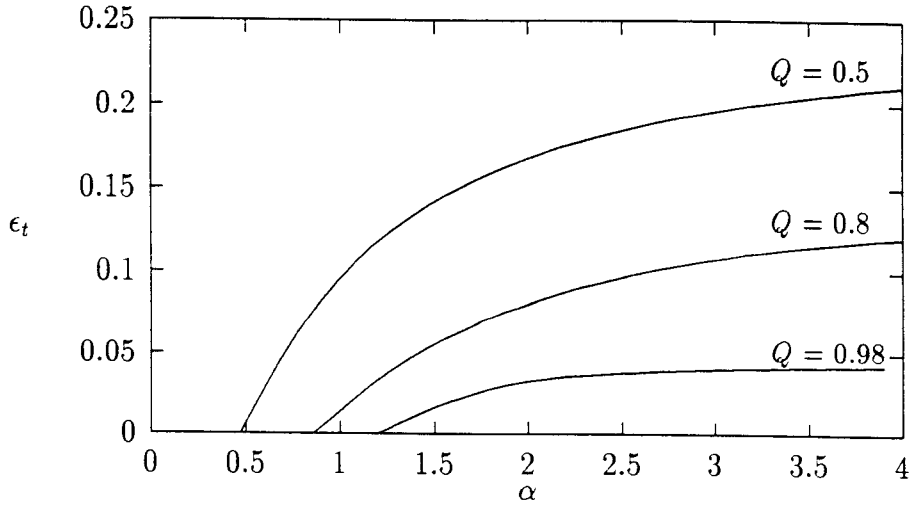

Figura 4.3: Erro de treinamento em função de α para a diluição fixa e $\gamma = 1$.

Podemos comparar o efeito das duas diluições através do gráfico de ϵ_t em função de Q para α e γ fixos (figura (4.5)). Como na diluição móvel o corte dos pesos é feito de maneira a minimizar o erro de treinamento, os valores de ϵ_t neste caso são sempre menores que no caso da diluição fixa.

4.2 Capacidade de Armazenamento

Conforme definido no capítulo 2, α_c é o valor da medida do conjunto de treinamento a partir do qual o erro de treinamento deixa de ser zero. Nesta seção , a determinação de α_c será feita fixando-se o valor de γ e variando-se o valor de Q.

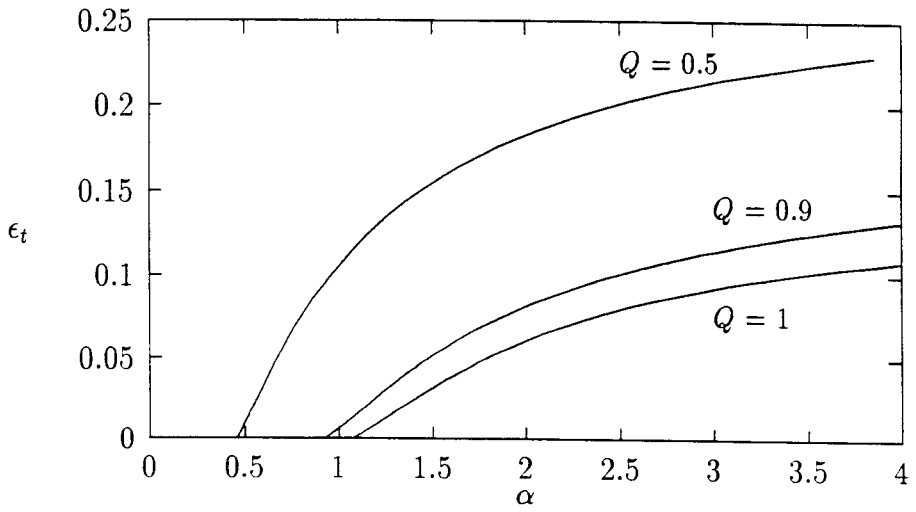

Figura 4.4: Erro de treinamento em função de α para a diluição fixa e $\gamma = 0.9$.

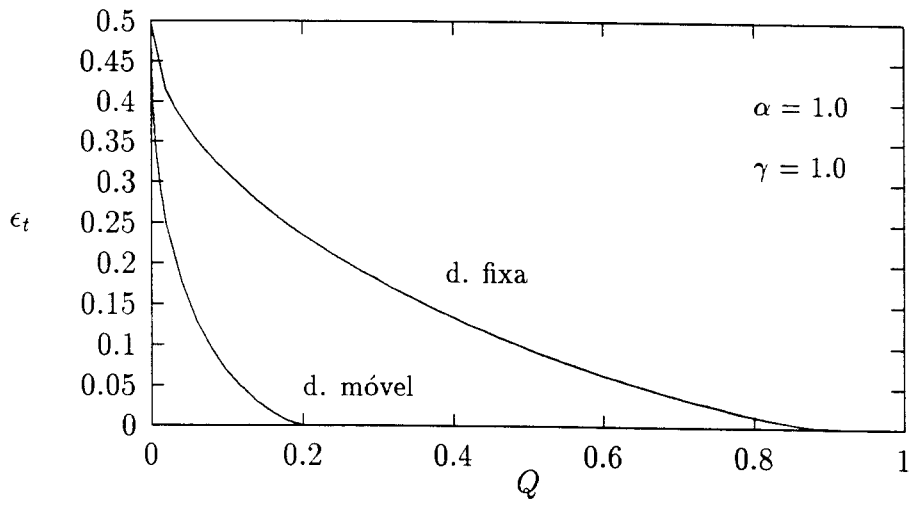

Figura 4.5: Erro de treinamento em função de Q para as duas diluições e $\alpha = 1.0$.

Na figura (4.6), para a diluição móvel, apresentamos o comportamento de α_c em função de Q para $\gamma = 1.0,\ 0.9,\ 0.5,\ 0.0$. Observamos que, partindo de $Q = 1$, o efeito da diluição

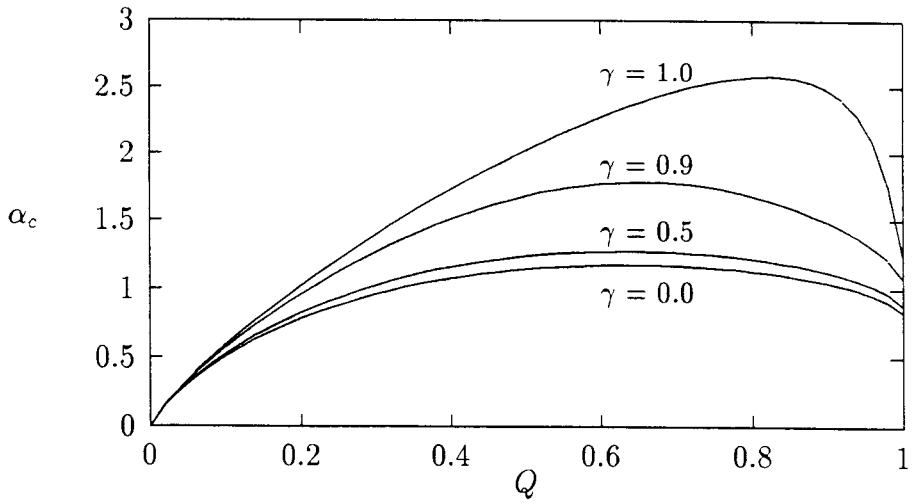

Figura 4.6: Capacidade de armazenamento em função de Q para a diluição móvel.

é aumentar α_c até atingir seu valor máximo, a partir do qual passa a diminuir. O valor de Q correspondente ao máximo de α_c pode ser calculado através da maximização de s_{sr} com relação a Q, ou seja, fazendo Q ser um parâmetro de ordem determinado pela equação $\partial s_{sr}/\partial Q = 0$. Na tabela abaixo relacionamos os valores de Q aos máximos de α_c.

γ	Q	α_c
1.0	0.82	2.58
0.9	0.66	1.78
0.5	0.62	1.27
0.0	0.63	1.17

Os maiores valores de α_c ocorrem para $Q < 1$, ou seja, quando a rede não está totalmente conectada. De fato, permitindo que Q varie aumentamos os graus de liberdade dos pesos de forma que agora possam tomar três valores $W_i = 0, \pm 1$. Portanto, a capacidade computacional da rede aumenta resultando no aumento de α_c.

No caso da diluição fixa o corte dos pesos sempre piora o desempenho da rede, uma vez que quanto maior for sua conectividade (Q) maior será a capacidade de armazenamento. Este comportamento pode ser visto na figura (4.7) onde apresentamos α_c em função de Q para vários valores de γ.

4.3 Erro de Generalização

Conforme discutido na seção 3.3, quando a rede estudante é treinada existem muitas redes neurais $\vec{W}$ com o mesmo erro de treinamento ϵ_t, mas a cada uma delas corresponde um erro de generalização $E_g\left(\vec{W}\right)$ diferente. Assim, devemos considerar o erro de generalização mediado sobre todas as redes neurais com mesmo erro de treinamento ϵ_t. Se a rede estudante for treinada com padrões puros ($\gamma = 1$) e na ausência de diluição ($Q = 1$), o regime de generalização perfeita ocorre para $\alpha \geq 1.245$ (Györgi 1990). Quando α atingir o valor 1.245, além do erro de treinamento, o erro de generalização também será zero.

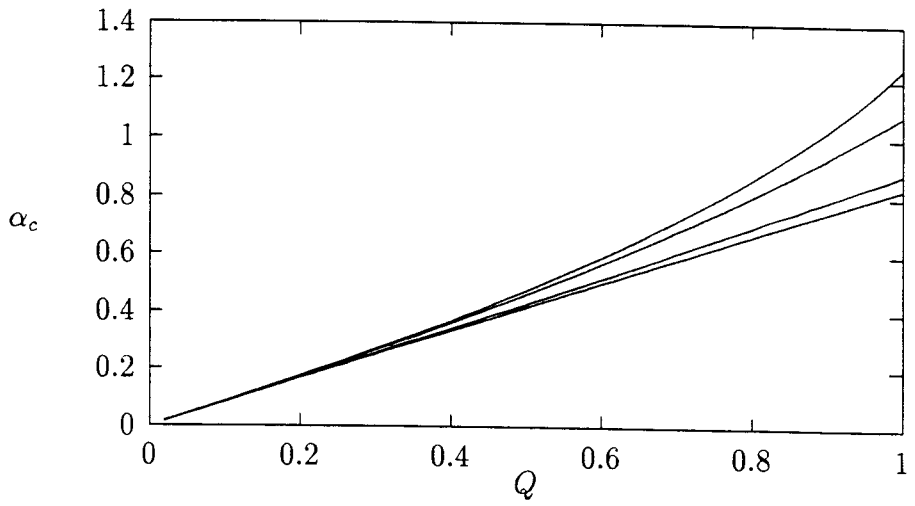

Figura 4.7: Capacidade de armazenamento em função de Q para a diluição fixa e $\gamma = 1.0$, 0.9, 0.5, 0.0, ordenados de cima para baixo.

Neste regime, a única rede (do ponto de vista macroscópico) que realiza perfeitamente o conjunto de treinamento é a rede mestra. Existem muitas outras redes que diferem da rede mestra por um número finito de pesos mas, uma vez que trabalhamos no limite $N \to \infty$, esta diferença não pode ser observada macroscopicamente. Neste caso, para $\alpha < 1.245$ além da rede mestra que tem $E_g\left(\vec{W}^0\right) = 0$, existem outras redes com $\epsilon_t = 0$ mas $E_g\left(\vec{W}\right) > 0$ resultando em um erro de generalização médio diferente de zero. Quando aumentamos o tamanho do conjunto de treinamento, há uma diminuição no valor de ϵ_g, pois menos redes com $E_g\left(\vec{W}\right) > 0$ contribuem para a média.

Podemos ver na figura (4.8), para o caso da diluição móvel com $\gamma = 1$, que há uma

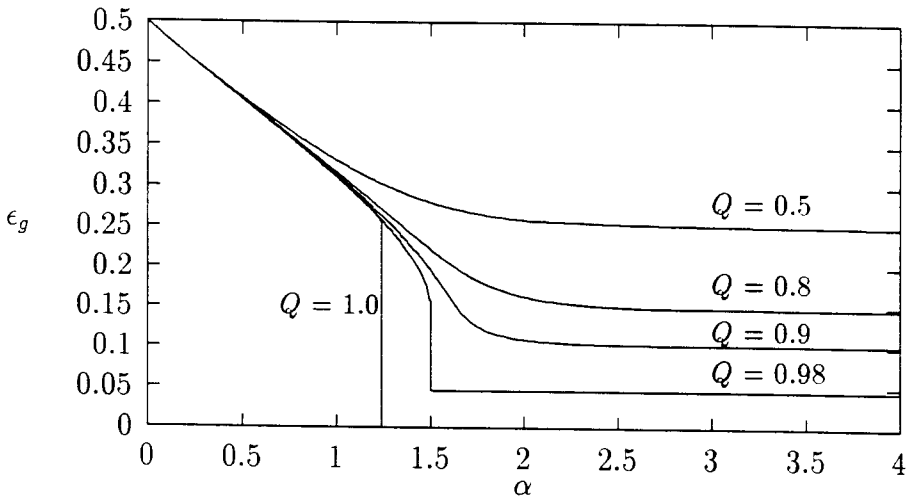

Figura 4.8: Erro de generalização em função de α para a diluição móvel e $\gamma = 1$.

transição descontínua para $Q < 1$ que persiste até $Q \approx 0.94$, a partir de onde as curvas passam a ser suaves. Essas transições ocorrem para $\alpha < \alpha_c$. Quando aumentamos o conjunto de treinamento a rede estudante se aproxima da rede mestra (verificamos nu-

mericamente que $R \to 1$ para qualquer valor de γ e Q), fazendo com que o valor de ϵ_g diminua até atingir seu valor mínimo

$$\epsilon_g \approx \frac{1}{\pi} \arccos\left(\sqrt{Q}\right) \qquad \alpha \to \infty. \tag{4.2}$$

Note que no limite assintótico ϵ_g pode tornar-se menor que ϵ_t para $\gamma < 1$ como consequência da definição (2.5). Se o treinamento da rede e o teste de generalização forem feitos com padrões ruidosos, o argumento do arccos na equação (4.2) torna-se $\gamma\sqrt{Q}$. Por outro lado, nessas condições foi mostrado por Seung *et al* (1992) que, no limite $\alpha \to \infty$, o erro de generalização tende ao de treinamento, $\epsilon_t \to \epsilon_g$. Justificamos assim a equação (4.1). No caso de treinamento na ausência de ruído ($\gamma = 1$) o corte dos pesos degrada a capacidade de generalização da rede pois quanto maior for a diluição , menor será sua semelhança com a rede mestra.

Consideremos agora o caso em que a rede é treinada com padrões ruidosos $\gamma < 1$. De acordo com a figura (4.9), ϵ_g diminui com o aumento de α, pois como o ruído obedece uma determinada distribuição de probabilidade quanto mais exemplos ruidosos forem apresentados à rede, maior é a possibilidade de compensação do efeito do ruído durante o treinamento, de maneira que $R \to 1$ no limite $\alpha \to \infty$. Note que dependendo do tamanho do conjunto de treinamento, a rede tem um desempenho melhor quando não está totalmente conectada, como pode ser observado na figura referida acima, onde o corte dos pesos foi feito usando a diluição móvel com $\gamma = 0.9$.

Para ilustrar melhor o comportamento do erro de generalização , a figura (4.10) mostra ϵ_g em função de Q para $\gamma = 0.9$. Quando α é pequeno o erro de generalização não é sensível ao corte dos pesos e o melhor desempenho ocorre para $Q = 1$. Aumentando-se α, o valor ótimo de Q que fornece o mínimo erro de generalização começa a decrescer. A partir de um determinado α, este valor torna a crescer e tende a 1 quando $\alpha \to \infty$, conforme a

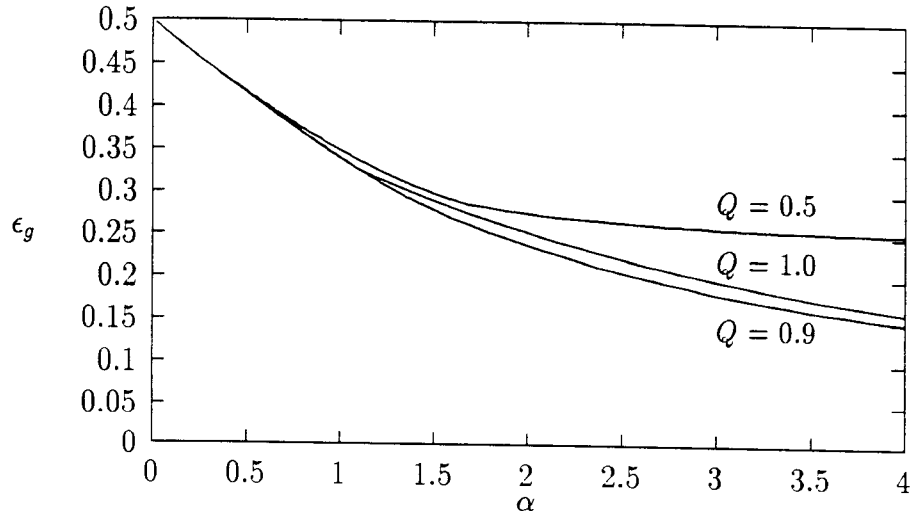

Figura 4.9: Erro de generalização em função de α para a diluição móvel e $\gamma = 0.9$.

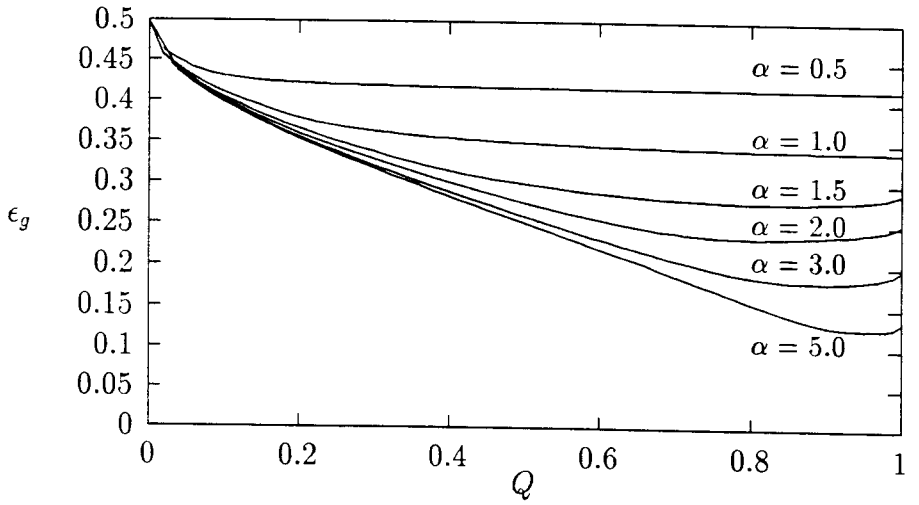

Figura 4.10: Erro de generalização em função de Q para a diluição móvel e $\gamma = 0.9$.

equação (4.2).

Na diluição fixa, para o caso de treinamento sem ruído ($\gamma = 1$) o comportamento de ϵ_g em função de α é análogo ao da diluição móvel, ou seja, quando aumentamos o valor de α há uma diminuição em ϵ_g e no limite $\alpha \to \infty$ o erro de generalização obedece a equação (4.2). Porém, na diluição móvel ϵ_g se aproxima mais rapidamente de seu valor mínimo. A figura (4.11) mostra ϵ_g em função de α para $Q = 0.5,\ 0.8,\ 0.9, 0.99,\ 1$. Neste caso, a

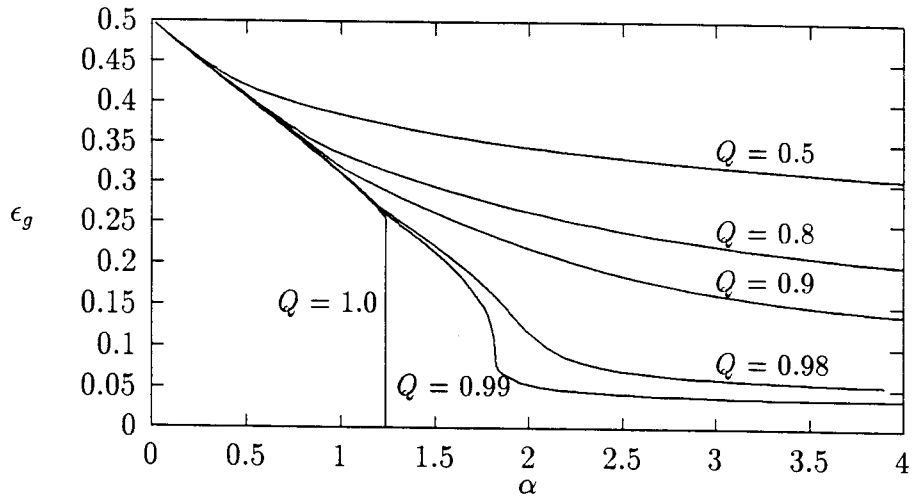

Figura 4.11: Erro de generalização em função de α para a diluição fixa e $\gamma = 1$.

descontinuidade no erro de generalização para $Q < 1$ ocorre para $\alpha > \alpha_c$ e desaparece para valores de Q maiores que no caso da diluição móvel.

Quando a rede é treinada com ruído ($\gamma < 1$), a diluição fixa degrada o erro de generalização para qualquer valor de α: diminuindo-se o valor de Q, aumenta-se o valor de ϵ_g como pode ser observado na figura (4.12) obtida com $\gamma = 0.9$ e $Q = 0.5,\ 0.9,\ 1.0$.

Para melhor observarmos as diferenças entre os efeitos dos dois tipos de diluição na

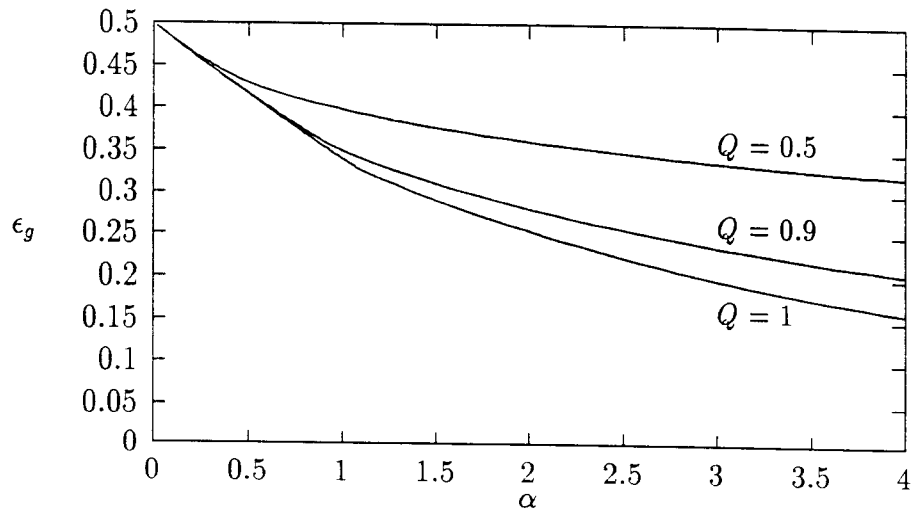

Figura 4.12: Erro de generalização em função de α para a diluição fixa e $\gamma = 0.9$.

capacidade de generalização da rede, apresentamos em uma mesma figura duas curvas de ϵ_g em função de Q. Assim, na figura (4.13) para $\alpha = 1.0$, observamos que o desempenho da rede com a diluição móvel é melhor que na diluição fixa. Ao contrário do resultado obtido para o erro de treinamento (figura (4.5)), este não é um resultado esperado pois, em princípio, não podemos garantir que a rede que tiver menor erro de treinamento terá também o menor erro de generalização .

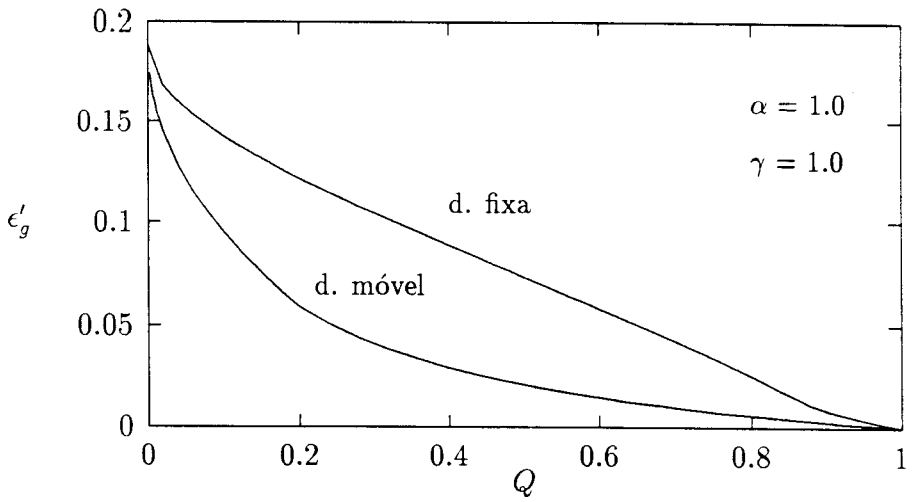

Figura 4.13: Erro de generalização , $\epsilon'_g = \epsilon_g(Q) - \epsilon_g(Q = 1)$, para as duas diluições e $\alpha = 1$, onde $\epsilon_g(Q = 1) = 0.31$.

Capítulo 5

Conclusão

Neste trabalho foram modelados dois tipos de lesões (móvel e fixa) sofridas, em ambos os casos, antes do processo de treinamento. Ainda, durante o treinamento adicionamos ruído na comunicação entre as redes mestra e estudante. Verificamos que o rendimento da rede estudante é menos afetado no caso da diluição móvel. Como essa diluição modela lesões sofridas por pessoas com pouca idade, este resultado era esperado pois neste caso o cérebro é mais flexível e existe uma forma de compensar o que foi lesado. Os principais resultados obtidos estão relacionados abaixo:

(i) O efeito da diluição fixa é re-escalar o número de neurônios $N' = NQ$ e o parâmetro de ruído $\gamma' = \gamma\sqrt{Q}$, fazendo com que o ruído e a diluição sempre degradem o desempenho da rede. Por outro lado, a diluição móvel produz um comportamento extremamente complexo como pode ser ilustrado pelo aumento da capacidade de armazenamento e pela melhora na capacidade de generalização no caso de treinamento com ruído.

(ii) No caso da diluição móvel e treinamento com ruído ($\gamma < 1$), a capacidade de armazenamento α_c atinge seu valor máximo quando a rede não está totalmente conectada.

Para ambas as diluições esta capacidade aumenta com a diminuição do ruído.

(iii) No caso de treinamento sem ruído ($\gamma = 1$), observamos que o erro de generalização diminui com o aumento de α e aumenta com a diminuição de Q, sendo que o menor erro ocorre para $Q = 1$. Isto é válido para ambas as diluições . O erro de generalização apresenta uma descontinuidade para altas conectividades $Q \approx 1$ quando $\alpha < \alpha_c$ para a diluição móvel e quando $\alpha > \alpha_c$ para a diluição fixa.

(iv) No caso de treinamento com ruído ($\gamma < 1$) e diluição móvel, dependendo do tamanho do conjunto de treinamento (α) a melhor generalização ocorre para $Q < 1$. Já na diluição fixa a melhor generalização ocorre para $Q = 1$, independente do valor de α.

(v) Apenas no limite $\alpha \to \infty$ é que as duas diluições produzem os mesmos efeitos: o erro de treinamento tende ao seu valor máximo (equação (4.1)) enquanto o erro de generalização tende ao seu valor mínimo (equação (4.2)). Porém, no caso da diluição móvel o erro de generalização atinge seu limite assintótico mais rapidamente que na diluição fixa. O oposto ocorre para o erro de treinamento.

Acreditamos que a melhora no desempenho da rede observada no caso da diluição móvel seja devida à restrição dos pesos a valores binários. De fato, a análise de perceptrons com pesos contínuos no caso do mapa aleatório ($\gamma = 0$) indica que a capacidade de armazenamento sempre diminui com a diluição (Bouten *et al* 1990a).

O estudo do problema proposto neste trabalho foi feito através da aplicação do formalismo microcanônico, pois este é o formalismo mais adequado para solução de problemas envolvendo perceptrons com pesos discretos (Fontanari e Meir 1993). Este problema foi também analisado sob o ponto de vista do formalismo canônico, onde o estado fundamental é determinado no limite em que a temperatura atinge o valor zero. Nesse formalismo as equações de ponto de sela são idênticas às do formalismo microcanônico, exceto que $\hat{e}$

passa a ser substituído pelo parâmetro fixo $\beta = 1/T$ de maneira que a equação (3.36) deve ser eliminada do sistema de equações de ponto de sela. A aplicação do formalismo canônico não é apropriada para o estudo do problema em questão, pois o valor da capacidade de armazenamento da rede, obtido tomando-se o limite $q \to 1$, é maior que o valor obtido no caso microcanônico, estando então em uma região de entropia negativa. Consequentemente, a dependência do erro de treinamento com α também é diferente da obtida no formalismo microcanônico. O comportamento qualitativo do erro de generalização para o caso da rede ser treinada na ausência de ruído é o mesmo encontrado no formalismo microcanônico. Porém, o formalismo canônico produz valores menores para essa grandeza. Quando $\gamma < 1$ e o valor de α está entre os valores críticos previstos pelos formalismos microcanônico e canônico, o erro de generalização aumenta com α atingindo seu valor máximo para $\alpha = \alpha_c$ calculado no formalismo canônico. A partir daí o erro de generalização passa a diminuir novamente com α como pode ser visto na figura (5.1). Os dois formalismos produzem os mesmos resultados para $\alpha < \alpha_c$, pois neste caso $\epsilon_t = 0$ implicando que $\hat{\epsilon} = 1/T \to \infty$. Acima de α_c o erro de treinamento deixa de ser zero e consequentemente a temperatura obtida através da equação (3.13) também.

As diluições tratadas nesse trabalho foram executadas antes do processo de treinamento, sendo esse um caso análogo a treinar-se um paciente já debilitado. Um caso interessante ocorre quando a diluição é feita após o processo de treinamento. Essa situação é equivalente a algum tipo de perda de memória, como por exemplo amnésia. Pretendemos considerar esta situação num trabalho futuro. Dando prosseguimento a este trabalho planejamos estudar também o efeito de lesões (diluição) nas capacidades de memorização e generalização de um perceptron com pesos reais, uma vez que este parece ser um modelo mais realístico de sistemas biológicos. Como a entropia de um sistema com um contínuo

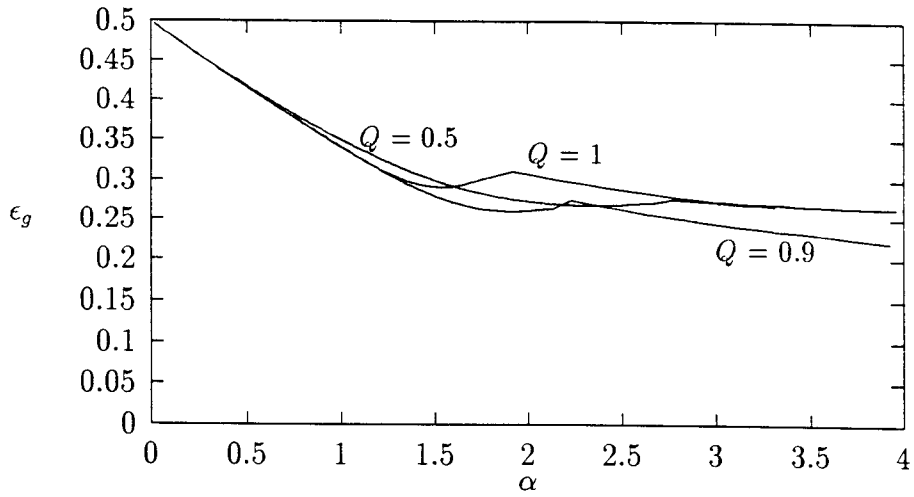

Figura 5.1: Erro de generalização calculado no formalismo canônico em função de α para a diluição fixa e $\gamma = 0.9$.

de graus de liberdade não é necessariamente positiva, o estado de mais baixa energia (erro) é obtido no limite $\hat{\epsilon} \to \infty$ de modo que os dois formalismos, canônico e microcanônico, dão os mesmos resultados. A maior dificuldade nesse estudo encontra-se na análise do regime de erro de treinamento não nulo ($\epsilon_t > 0$), pois neste caso a prescrição simetria de réplicas é instável frente à quebra de simetria de réplicas (Gardner e Derrida 1988). Assim, para estudar o efeito de lesões em um perceptron de pesos reais é necessário utilizar a prescrição de Parisi (Parisi 1980, Mézard *et al* 1987, Binder e Young 1986) para quebrar a simetria de réplicas, o que torna a análise analítica extremamente complicada.

Apêndice A

Cálculo da Entropia

Neste apêndice desenvolveremos o cálculo da entropia para as diluições móvel e fixa partindo da equação (3.10) para obtermos as equações (3.18) e (3.43), respectivamente.

A.1 Diluição Móvel

Dando continuidade ao cálculo da equação (3.10), devemos primeiramente executar as médias sobre o conjunto de treinamento. Para facilitar estes cálculos vamos introduzir as identidades:

$$\int_{-\infty}^{\infty} \prod_{l=1}^{p} dy^l \delta\left(y^l - \sum_{j=1}^{N} W_j^0 \xi_j^l N^{-1/2}\right) = 1 \tag{A.1}$$

e

$$\int_{-\infty}^{\infty} \prod_{l,a} dx_a^l \delta\left(x_a^l - \sum_{j=1}^{N} W_j^a s_j^l N^{-1/2}\right) = 1. \tag{A.2}$$

Usando a representação integral da função delta de Dirac,

$$\delta(x) = \int_{-\infty}^{\infty} \frac{d\tilde{x}}{2\pi} \, e^{i\tilde{x}x} \tag{A.3}$$

escrevemos

$$\begin{aligned}\langle\langle\left(Z\left(\hat{\epsilon},\vec{S},t\right)\right)^n\rangle\rangle &= \langle\langle\prod_{a=1}^{n}\sum_{\{\vec{W}^a\}}\delta\left(Q,\frac{1}{N}\sum_i(W_i^a)^2\right)\int\prod_{l=1}^{P}\frac{d\tilde{y}_l dy_l}{2\pi}\\ &\int\prod_{l,a}\frac{d\tilde{x}_l^a dx_a^l}{2\pi}\exp\left[-\frac{i}{\sqrt{N}}\sum_{l,j}\left(\tilde{y}_l W_j^0\xi_j^l+\sum_a\tilde{x}_a^l W_j^a S_j^l\right)\right.\\ &\left.+i\sum_l\left(\tilde{y}_l y_l+\sum_a\tilde{x}_l^a x_a^l\right)-\sum_{l,a}\hat{\epsilon}^a\Theta\left(-y_l x_l^a\right)\right]\rangle\rangle \qquad (A.4)\end{aligned}$$

onde $W_i^a = 0,\ \pm 1\ (i = 1,\ldots,N);\ (a = 1,\ldots,n)$.

Os termos relevantes para o cálculo das médias são aqueles que dependem de S_j^l e ξ_j^l. Assim,

$$\begin{aligned}\mathcal{M}_{jl} &= \int d\xi_j^l\int dS_j^l P\left(\xi_j^l\right)P\left(S_j^l|\xi_j^l\right)\exp\left[-i\tilde{y}^l W_j^0 N^{-1/2}\xi_j^l-i\sum_a\tilde{x}_l^a W_j^a N^{-1/2}S_j^l\right]\\ &= \left(\frac{1+\gamma}{2}\right)\cos\left(\sum_a\tilde{x}_l^a W_j^a N^{-1/2}+\tilde{y}_l W_j^0 N^{-1/2}\right)\\ &+\left(\frac{1-\gamma}{2}\right)\cos\left(\sum_a\tilde{x}_l^a W_j^a N^{-1/2}-\tilde{y}_l W_j^0 N^{-1/2}\right). \qquad (A.5)\end{aligned}$$

No limite $N\to\infty$ podemos expandir os argumentos dos cossenos mantendo apenas o termo dominante:

$$\prod_{jl}\mathcal{M}_{jl}=\exp\left\{-\frac{1}{2N}\sum_{jl}\left[(\tilde{y}_l)^2+\left(\sum_a\tilde{x}_a^l W_j^a\right)^2+2\gamma\tilde{y}_l W_j^0\sum_a\tilde{x}_a^l W_j^a\right]\right\} \qquad (A.6)$$

onde usamos $\left(W_j^0\right)^2 = 1$. Os parâmetros de ordem são então introduzidos por meio das identidades

$$\int\prod_{a<b}dq_{ab}\,\delta\left(q_{ab}-\frac{1}{NQ}\sum_{j=1}^{N}W_j^a W_j^b\right)=1 \qquad (A.7)$$

e

$$\int\prod_a^n dR_a\,\delta\left(R_a-\frac{1}{NQ}\sum_{j=1}^{N}W_j^0 W_j^a\right)=1, \qquad (A.8)$$

de forma que a equação (3.10) é escrita como

$$\begin{aligned}\langle\langle(\mathcal{N}(E))^n\rangle\rangle &= \int\prod_{a=1}^{n}\frac{d\hat{\epsilon}^a}{2\pi i}\int\prod_{a<b}\frac{d\hat{q}_{ab}dq_{ab}}{2\pi/NQi}\int\prod_{a}\frac{d\hat{R}_a dR_a}{2\pi/NQi}\int\prod_{a}\frac{d\hat{Q}_a}{2\pi/Ni}\\ &\exp N\left[\frac{P}{N}\epsilon_t\sum_a\hat{\epsilon}_a - Q\sum_{a<b}q_{ab}\hat{q}_{ab} - Q\sum_a R_a\hat{R}_a\right.\\ &\left.-Q\sum_a\hat{Q}_a + G_0\left(\hat{q}_{ab},\hat{R}_a,\hat{Q}_a\right) + \frac{P}{N}\,G_1\left(q_{ab},R_a,\hat{\epsilon}_a\right)\right] \qquad (A.9)\end{aligned}$$

onde

$$G_0 = \frac{1}{N}\sum_{i=1}^{N}\ln\sum_{\{W^a=0,\pm1\}}\exp\left\{+\sum_{a<b}\hat{q}_{ab}W^aW^b + \sum_a\hat{Q}^a(W^a)^2 + \sum_a\hat{R}_aW^aW_i^0\right\}, \qquad (A.10)$$

$$\begin{aligned}G_1 &= \ln\int Dy\int\prod_a\frac{dx_a d\tilde{x}_a}{2\pi}\exp\left[-\sum_a\hat{\epsilon}_a\Theta(-yx_a) - \frac{1}{2}\sum_a\tilde{x}_a^2\left(1-\gamma'^2R_a^2\right)\right.\\ &\left.-\sum_{a<b}\tilde{x}_a\tilde{x}_b\left(q_{ab}-\gamma'^2R_aR_b\right) + i\sum_a\tilde{x}_a\left(x_a - y\gamma' R_a\right)\right], \qquad (A.11)\end{aligned}$$

com a notação

$$Dy = \frac{dy}{\sqrt{2\pi}}e^{-y^2/2}. \qquad (A.12)$$

As integrais sobre os parâmetros de ordem são efetuadas através do método de integração por ponto de sela:

$$\mathcal{I}(N) = \int_C e^{NF(t)}dt = \lim_{N\to\infty}\sqrt{\frac{-2\pi}{NF''(t_0)}}\,e^{NF(t_0)} \qquad (A.13)$$

onde t_0 é determinado pela equação $F'(t_0) = 0$. Mas para aplicarmos esse método de integração , a razão P/N deve ser $\mathcal{O}(1)$, o que justifica a escala $P = \alpha N$. Note que no limite termodinâmico apenas o argumento da exponencial contribui para a densidade $\frac{1}{N}\ln\mathcal{I}(N)$. Desse modo obtemos a equação (3.18).

A.2 Diluição Fixa

Dando continuidade ao cálculo da equação (3.10) e seguindo os passos da seção anterior escrevemos

$$\langle\langle\left(Z\left(\hat{\epsilon},\vec{S},t\right)\right)^n\rangle\rangle = \langle\langle\prod_{a=1}^{n}\sum_{\{\vec{W}^a\}}\int\prod_{l=1}^{P}\frac{d\tilde{y}_l dy_l}{2\pi}\int\prod_{l,a}\frac{d\tilde{x}_l^a dx_a^l}{2\pi}\exp\left[i\sum_l\left(\tilde{y}_l y_l+\sum_a\tilde{x}_l^a x_a^l\right)\right.$$
$$\left.-\frac{i}{\sqrt{N}}\sum_{l,j}\left(\tilde{y}_l W_j^0\xi_j^l+\sum_a\tilde{x}_a^l W_j^a S_j^l\right)-\sum_{l,a}\hat{\epsilon}^a\Theta\left(-y_l x_l^a\right)\right]\rangle\rangle, \quad \text{(A.14)}$$

onde $W_i^a = \pm 1 (i = 1, \ldots, NQ); (a = 1, \ldots, n)$. Efetuando as médias sobre S_j^l e ξ_j^l como na seção anterior obtemos a seguinte expressão para o n-ésimo momento de $\mathcal{N}(\epsilon_t)$

$$\langle\langle(\mathcal{N}(E))^n\rangle\rangle = \int\prod_{a=1}^{n}\frac{d\hat{\epsilon}^a}{2\pi i}\int\prod_{a<b}\frac{d\hat{q}_{ab}dq_{ab}}{2\pi/NQi}\int\prod_a\frac{d\hat{R}_a dR_a}{2\pi/NQi}$$
$$\exp N\left[\alpha\epsilon_t\sum_a\hat{\epsilon}_a - Q\sum_{a<b}q_{ab}\hat{q}_{ab} - Q\sum_a R_a\hat{R}_a + QG_0 + \alpha G_1\right] \quad \text{(A.15)}$$

onde

$$G_0 = \frac{1}{NQ}\sum_{i=1}^{NQ}\ln\sum_{\{W^a=\pm 1\}}\exp\left\{+\sum_{a<b}\hat{q}_{ab}W^a W^b + \sum_a\hat{R}_a W^a W_i^0\right\}, \quad \text{(A.16)}$$

e G_1 é dado pela equação (3.20). A integração por ponto de sela no limite $N \to \infty$ nos leva a equação (3.43).

Apêndice B

Cálculo de G_0 com a Prescrição Simetria de Réplicas

Neste apêndice mostraremos o cálculo de G_0 para os dois tipos de diluição , uma vez que o modo utilizado para cortar os pesos afeta apenas a contribuição de G_0 para a densidade de entropia.

B.1 Diluição Móvel

Usando a prescrição simetria de réplicas na equação (3.19) e rearranjando os termos reescrevemos aquela equação como

$$G_0^{sr} = \frac{1}{N}\sum_{i=1}^{N} \ln\left\{\prod_{a=1}^{n}\sum_{\{W^a=0,\pm1\}} \exp\left(\frac{\hat{q}}{2}\left(\sum_a W^a\right)^2 + \left(\hat{Q}-\frac{\hat{q}}{2}\right)\sum_a W_a^2 + \hat{R}W_i^0\sum_a W^a\right)\right\}. \tag{B.1}$$

Aplicando a transformação gaussiana,

$$e^{x^2/2} = \int_{-\infty}^{\infty} Dt\, e^{tx}, \tag{B.2}$$

para linearizar o primeiro termo do argumento da exponencial obtemos

$$\begin{aligned} G_0^{sr} &= \frac{1}{N}\sum_{i=1}^{N} \ln \int_{-\infty}^{\infty} Dt \left\{ \sum_{\{W=0,\pm 1\}} \exp\left[\left(\hat{Q} - \hat{q}/2\right)(W)^2 + \hat{R} W W_i^0 + \sqrt{\hat{q}} W t\right]\right\}^n \\ &\approx \frac{1}{N}\sum_{i=1}^{N} \ln \int_{-\infty}^{\infty} Dt \left\{1 + n \ln\left[1 + 2\exp\left(\hat{Q} - \hat{q}/2\right)\cosh\left(\hat{R} W_i^0 + \sqrt{\hat{q}} t\right)\right]\right\} \\ &\approx \frac{n}{N}\sum_{i=1}^{N} \int_{-\infty}^{\infty} Dt \ln\left[1 + 2\exp\left(\hat{Q} - \hat{q}/2\right)\cosh\left(\hat{R} W_i^0 + \sqrt{\hat{q}} t\right)\right]. \end{aligned} \tag{B.3}$$

Note que $W_i^0 = \pm 1$ e $\cosh(x) = \cosh(-x)$ de modo que efetuando-se a mudança de variáveis $t = t'/W_i^0$ obtemos a expressão final de G_0^{sr}

$$G_0^{sr} \approx n \int_{-\infty}^{\infty} Dt \ln\left[1 + 2\exp\left(\hat{Q} - \frac{\hat{q}}{2}\right)\cosh\left(\hat{R} + t\sqrt{\hat{q}}\right)\right]. \tag{B.4}$$

B.2 Diluição Fixa

Usando a prescrição simetria de réplicas na equação (3.44) obtemos

$$G_0^{sr} = \frac{1}{NQ}\sum_{i=1}^{NQ} \ln\left\{\prod_{a=1}^{n} \sum_{\{W^a=\pm 1\}} \exp\left(\frac{\hat{q}}{2}\left(\sum_a W^a\right)^2 - \frac{\hat{q}}{2}\sum_a W_a^2 + \hat{R} W_i^0 \sum_a W^a\right)\right\}. \tag{B.5}$$

Aplicando a transformação gaussiana, equação (B.2), para linearizar o primeiro termo do argumento da exponencial obtemos, analogamente à seção anterior,

$$
\begin{aligned}
G_0^{sr} &= \frac{1}{NQ}\sum_{i=1}^{NQ} \ln \int_{-\infty}^{\infty} Dt \left\{ \sum_{\{W=\pm 1\}} \exp\left[(-\hat{q}/2)\,(W)^2 + \hat{R}WW_i^0 + \sqrt{\hat{q}}Wt\right]\right\}^n \\
&\approx \frac{1}{NQ}\sum_{i=1}^{NQ} \ln \int_{-\infty}^{\infty} Dt \left\{1 + n \ln\left[2\exp\left(-\hat{q}/2\right)\cosh\left(\hat{R}W_i^0 + \sqrt{\hat{q}}t\right)\right]\right\} \\
&\approx \frac{n}{NQ}\sum_{i=1}^{NQ} \int_{-\infty}^{\infty} Dt \ln\left[2\exp\left(-\hat{q}/2\right)\cosh\left(\hat{R}W_i^0 + \sqrt{\hat{q}}t\right)\right] \\
&\approx -\frac{n\hat{q}}{2} + n\int_{-\infty}^{\infty} Dt \ln 2\cosh\left(\hat{R} + t\sqrt{\hat{q}}\right). \qquad \text{(B.6)}
\end{aligned}
$$

Apêndice C

Cálculo de G_1 com a Prescrição Simetria de Réplicas

Analogamente ao cálculo de G_0, usamos a prescrição simetria de réplicas na equação (3.20), rearranjando seus termos de forma que podemos reescrevê-la como

$$\begin{aligned} G_1^{sr} &= \ln \int Dy \int \prod_a \frac{dx_a d\tilde{x}_a}{2\pi} \exp \left[\frac{R^2\gamma'^2 - q}{2} \left(\sum_a \tilde{x}_a \right)^2 + \frac{q-1}{2} \sum_a (\tilde{x}_a)^2 \right. \\ &\left. -\hat{\epsilon} \sum_a \Theta(-yx_a) + i \sum_a \tilde{x}_a x_a - i\gamma' R \sum_a \tilde{x}_a y \right]. \end{aligned} \tag{C.1}$$

Aplicando a transformação gaussiana, equação (B.2), para linearizar o termo quadrático no argumento da exponencial e tomando o limite $n \to 0$ obtemos,

$$\begin{aligned} G_1^{sr} &= \ln \int_{-\infty}^{\infty} Dt \int_{-\infty}^{\infty} Dy \ \Lambda^n \\ &\approx n \ln \int_{-\infty}^{\infty} Dt \int_{-\infty}^{\infty} Dy \ \ln \Lambda \end{aligned} \tag{C.2}$$

onde

$$\begin{aligned}\Lambda &= \int_{-\infty}^{\infty} \frac{dx d\tilde{x}}{2\pi} \exp\left[-\hat{\epsilon}\Theta(-yx) + \frac{q-1}{2}\tilde{x}^2 + i\left(x - \gamma' R y + t\sqrt{R^2\gamma'^2 - q}\right)\tilde{x}\right] \\ &= \mathrm{e}^{-\hat{\epsilon}\Theta(y)} H\left[\frac{\gamma' R y - t\sqrt{R^2\gamma'^2 - q}}{\sqrt{1-q}}\right] + \mathrm{e}^{-\hat{\epsilon}\Theta(-y)} H\left[\frac{-\gamma' R y + t\sqrt{R^2\gamma'^2 - q}}{\sqrt{1-q}}\right] \qquad \text{(C.3)}\end{aligned}$$

com $H(x)$ dado pela equação (3.20). Realizando uma mudança de variáveis podemos efetuar uma das integrais que aparece na equação (C.2), obtendo a expressão final de G_1

$$G_1 \approx 2n \int_{-\infty}^{\infty} Dz\ H(\xi_1) \ln\left[\mathrm{e}^{-\hat{\epsilon}} + \left(1 - \mathrm{e}^{-\hat{\epsilon}}\right) H(\xi_2)\right] \qquad \text{(C.4)}$$

onde

$$\xi_1 = z\sqrt{\frac{\gamma'^2 R^2}{q - \gamma'^2 R^2}} \qquad \text{(C.5)}$$

e

$$\xi_2 = z\sqrt{\frac{q}{1-q}}. \qquad \text{(C.6)}$$

Bibliografia

de Almeida J R e Thouless DJ 1978 " Stability of the Sherrington-Kirkpatrick solution of a spin model" *J. Phys. A* **11** 983.

Barbato DML e Fontanari JF 1993 " The effects of lesions on the generalization ability of a perceptron"*J. Phys. A* **26** 1862.

Binder K e Young A P 1986 " Spin glasses: Experimental facts, theorical concepts, and open questions " *Rev. Mod. Phys.* **58** 801.

Bouten M, Engel A, Komoda A e Serneels R 1990a " Quenched versus annealed diluition in neural networks " *J. Phys. A* **23** 4643.

Bouten M, Komoda A e Serneels R 1990b " Storage capacity of a diluted neural network with Ising couplings " *J. Phys. A* **23** 2605.

Fontanari JF e Meir R 1993 "The statistical mechanics of the Ising perceptron" *J. Phys. A* **26** 1077.

Gardner E 1988 " The space of interactions in neural network models"*J. Phys. A* **21** 257.

Gardner E e Derrida B 1988 " Optimal storage properties of neural network models " *J. Phys. A* **21** 271.

Gardner E e Derrida B 1989 " Three unfinished works on the optimal storage capacity of networks " *J. Phys. A* **22** 1983.

György G 1990 " First-order transition to perfect generalization in a neural network with binary synapses " *Phys. Rev. A* **41** 7097.

György G e Tishby N 1989 " Statistical theory of learning a rule " *Neural Networks and Spin Glasses* eds. W K Theumann e R Köberle (Singapore: World Scientific).

Krauth W e Mézard M 1989 " Storage capacity of memory networks with binary couplings " *J.Physique* **50** 3057.

Krauth W e Opper M 1989 " Critical storage capacity of the $J = \pm 1$ neural network " *J. Phys A* **22** L519.

Lindsay P H e Norman D A 1977 *Human Information Processing* (New York: Academic Press).

Mézard M, Parisi G e Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific).

Minsky M L e Papert S A 1969 *Perceptrons: An introdution to computational geometry* (Cambridge: MIT Press).

Nguyen D e Widrow B 1990 " The truck backer-upper: an example of self-learning in neural networks " *Neural Networks for Control* eds W T Miller, R S Sutton e P J Werbos (Cambridge: MIT Press).

Parisi G 1980 " The order parameter for Spin Glasses: A function on the interval 0-1" *J. Phys. A* **13** 1101.

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

Sejnowski T S e Rosenberg C R 1987 " Parallel Networks that learn to pronounce English text" Complex Systems **1** 145.

Seung S, Sompolinsky H e Tishby N 1992 " Statistical mechanics of learning from examples " *Phys. Rev.* A **45** 6056.

SERVIÇO DE BIBLIOTECA E INFORMAÇÃO - IFQSC
FISICA