Universidade de São Paulo
Instituto de Física de São Carlos
Departamento de Física e Informática

"Estudo Analítico do Efeito da Diluição em Perceptrons"

Daniela M. L. Barbato

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

Orientador: *Prof. Dr. J.F.Fontanari*

São Carlos

1997

Barbato, Daniela Maria Lemos

Estudo Analítico do Efeito da Diluição em Perceptrons/Daniela Maria Lemos Barbato.–São Carlos, 1998.

129 p.

Tese (Doutorado)–Instituto de Física de São Carlos, 1998.

Orientador: Prof. Dr. José Fernando Fontanari

1. Mecânica Estatística. 2. Redes Neurais. 3. Diluição em Perceptrons. 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 DANIELA MARIA LEMOS BARBATO APRESENTADA AO INSTITUTO DE FÍSICA DE SÃO CARLOS, DA UNIVERSIDADE DE SÃO PAULO, EM 21 DE JANEIRO DE 1998.

COMISSÃO JULGADORA:

Prof. Dr. José Fernando Fontanari/IFSC-USP

Prof. Dr. Roland Koberle/IFSC-USP

Prof. Dr. Osame Kinouchi Filho/IFSC-USP

Prof. Dr. Antonio Carlos Roque da Silva Filho/FFCLRP-USP

Prof. Dr. Nestor Felipe Caticha Alfonso/IF-USP

Índice

Lista de Figuras

Resumo

Perceptrons são redes neurais sem retroalimentação cujos 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 único neurônio motor σ através das conexões sinápticas J_i; $i = 1, \ldots, N$. Utilizando o formalismo da Mecânica Estatística desenvolvido por Gardner e colaboradores, estudamos os efeitos da eliminação de uma fração dos pesos sinápticos (diluição) nas capacidades de aprendizado e generalização de dois tipos de pertceptrons, a saber, o perceptron linear e o perceptron Booleano. No perceptron linear comparamos o desempenho de redes lesadas por diferentes tipos de diluição, que podem ocorrer durante ou após o processo de aprendizado. Essa comparação mostra que a estratégia de minimizar o erro de treinamento não fornece o menor erro de generalização, além do que, dependendo do tamanho do conjunto de treinamento e do nível de ruído, os pesos menores podem se tornar os fatores mais importantes para o bom funcionamento da rede. No perceptron Booleano investigamos apenas o efeito da diluição após o término do aprendizado na capacidade de generalização da rede neural treinada com padrões ruidosos. Neste caso, apresentamos uma comparação entre os desempenhos relativos de cinco regras de aprendizado: regra de Hebb, pseudo-inversa, algoritmo de Gibbs, algoritmo de estabilidade ótima e algoritmo de Bayes. Em particular mostramos que a diluição sempre degrada o desempenho de generalização e o algoritmo de Bayes sempre fornece o menor erro de generalização.

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 J_i; $i = 1, \ldots, N$. Using the Statistical Mechanics formalism developed by Gardner and co-workers, we study the effects of eliminating a fraction of synaptic weights (dilution) on the learning and generalization capabilities of the two types of perceptrons, namely, the linear perceptron and the Boolean perceptron. In the linear perceptron we compare the performances of networks damaged by different types of dilution, which may occur either during or after the learning stage. The comparison between the effects of the different types of dilution, shows that the strategy of minimizing the training error does not yield the best generalization performance. Moreover, this comparison also shows that, depending on the size of the training set and on the level of noise corrupting the training data, the smaller weights may became the determinant factors in the good functioning of the network. In the Boolean perceptron we investigate the effect of dilution after learning on the generalization ability when this network is trained with noisy examples. We present a thorough comparison between the relative performances of five learning rules or algorithms: the Hebb rule, the pseudo-inverse rule, the Gibbs algorithm, the optimal stability algorithm and the Bayes algorithm. In particular, we show that the effect of dilution is always deleterious, and that the Bayes algorithm always gives the best generalization performance.

Capítulo 1

Introdução

O estudo de redes neurais é basicamente centrado em dois aspectos: a modelagem de sistemas cognitivos biológicos e a produção de inteligência artificial.

O cérebro manifesta características muito peculiares que o diferencia de um simples computador. Dentre estas podemos citar a robusteza, que é a capacidade de manter seu bom funcionamento mesmo com a constante perda de neurônios, e a flexibilidade, que é a capacidade de se adaptar a novas situações através do aprendizado. Este sistema realiza tarefas complexas como aprendizado, reconhecimento de padrões, criatividade e percepção. Apesar destas tarefas parecerem triviais, não se sabe realmente como elas são executadas.

Com base nestas características podemos pensar em maneiras mais eficientes de programação para computadores convencionais. O método usual de se programar um computador consiste na elaboração de uma lista de instruções, porém, uma vez que nem todas as tarefas que ele deve desempenhar podem ser decodificadas em tal lista, este método implica em algumas limitações. Além disso, se houver algum tipo de perturbação ou ruído nesta lista de instruções, seu desempenho não será satisfatório.

Quando trabalhamos com redes neurais ao invés de programas convencionais, podemos treiná-las com exemplos que são pares questões/respostas. Desta maneira, estas redes são capazes de inferir regras contidas nestes exemplos e, consequentemente, são capazes de generalizar, ou seja, responder corretamente a novas questões. Esta capacidade de generalização constitui uma das principais motivações para o estudo de redes neurais. É

importante notar que as redes neurais são modelos extremamente simplificados quando comparados a sistemas biológicos.

O sistema nervoso é constituído por células nervosas chamadas neurônios. A parte do sistema nervoso que estamos interessados é o cérebro, que é formado por aproximadamente 10^{10} neurônios. O tamanho e a forma do neurônio variam de acordo com a sua função, mas a estrutura básica é sempre a mesma: soma ou corpo celular, axônio e dendritos. Os dendritos recebem o sinal de entrada de outros neurônios, o soma transforma estes sinais de entrada em sinal de saída e o axônio transmite o sinal de saída para outros neurônios. O ponto de ligação entre uma terminação axônica e outra célula é denominado sinapse. Neste ponto as células estão separadas por uma fenda estreita. Nos mamíferos mais desenvolvidos cada neurônio pode receber até 10^4 entradas sinápticas.

Os sinais no sistema nervoso são transmitidos eletricamente e quimicamente. A transmissão elétrica é baseada em descargas elétricas que começam no corpo celular e se propagam através do axônio para as várias conexões sinápticas. A transmissão do sinal elétrico através da fenda sináptica é efetuada por mecanismos químicos. Quando o sinal chega ao terminal do axônio, os neurotransmissores se difundem na fenda sináptica e atingem o neurônio da célula vizinha induzindo uma mudança de potencial naquela célula. Todos os sinais que chegam a célula vizinha propagam-se para o corpo celular onde são integrados. Se a somatória de todos os sinais exceder um certo limiar, o neurônio emite um pulso que se propaga através do axônio. A contribuição de uma entrada sináptica para o potencial da célula caracteriza o peso ou eficiência sináptica. Como foi postulado por Hebb [21], o peso das conexões sinápticas modifica-se de acordo com seu nível de atividade, assim as sinapses são altamente modificáveis com relação à sua eficiência, desempenhando um papel fundamental no aprendizado.

O estudo de modelos de redes neurais começou com McCulloch e Pitts [35] que introduziram a idéia de um neurônio formal com estados binários baseados no fato de que os neurônios ou emitem pulsos elétricos ou estão inativos. Ainda, aqueles autores mostraram que redes de neurônios formais são capazes de desempenhar qualquer tarefa lógica. Porém, o problema era escolher os pesos sinápticos a fim de que a rede desempenhasse uma determinada tarefa. Foi mostrado por Rosenblatt [44] que redes com conexões modificáveis

chamadas perceptrons podem ser treinadas para classificar certos conjuntos de padrões. Sua proposta original era tentar explicar o funcionamento do cérebro em termos das estruturas cerebrais. Rosenblatt e seus colaboradores desenvolveram um procedimento de treinamento, o algoritmo Perceptron, com a finalidade de encontrar os pesos corretos para uma dada tarefa.

Existem dois tipos básicos de arquitetura de redes neurais: perceptrons e redes atratoras. Nas redes atratoras cada neurônio está conectado a todos os demais e o estado de cada um deles é ajustado segundo uma dinâmica que pode ser síncrona ou assíncrona. As redes atratoras podem ser usadas como memória associativa e são análogas a sistemas de vidros de spin. Perceptrons são redes dispostas em camadas de neurônios, onde a informação flui em uma única direção através destas camadas sem que haja retroalimentação. A estrutura do perceptron é formada por uma camada de entrada constituída por neurônios sensores que não estão ligados entre si, algumas camadas intermediárias e uma camada de saída composta por neurônios motores. As ligações entre os neurônios de diferentes camadas é feita através das sinapses cujas eficiências ou pesos são ajustáveis.

Neste trabalho consideramos apenas os perceptrons mais simples, compostos por uma camada de entrada com N neurônios sensores que recebem estímulos externos e uma camada de saída com apenas um neurônio motor. O modelo matemático de um perceptron leva em conta que os neurônios basicamente têm dois estados de atividade, ou emitem pulsos ou estão inativos, assim os estados dos neurônios sensores são dados por $S_i = \pm 1 \, (i = 1, \ldots . N)$. O perceptron é uma rede que soma todos os sinais da camada de entrada com um determinado peso e fornece o resultado através da função de tranferência $f(x)$, ou seja,

$$\sigma = f\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i S_i\right), \tag{1.1}$$

onde $J_i \, (i = 1. \ldots . , N)$ são os pesos das conexões sinápticas e σ é o estado do neurônio motor. As funções de transferência mais comumente usadas são a função linear, $f(x) = x$, e a função booleana, onde $f(x) = \text{sign}(x)$. Os perceptrons aqui abordados são o perceptron linear e o perceptron Booleano, cujos pesos sinápticos assumem valores reais.

O perceptron de muitas camadas tem grande aplicação prática em diversas áreas como

reconhecimento de caracteres escritos à mão em códigos de endereçamento postal (CEP). predição de estrutura tridimensional de proteínas, reconhecimento de voz e muitas ou tras [8]. Existem algumas tarefas que um perceptron de uma camada não desempenha. Estas tarefas estão relacionadas a propriedade geométrica de separabilidade linear. O perceptron com uma camada é capaz de realizar ou aprender somente as funções que são linearmente separáveis. A seu favor devemos mencionar que além de ser um modelo consistente com fatos biológicos. é suficientemente complexo para produzir um comportamento interessante e suficientemente simples para que possa ser estudado analiticamente.

Programar um computador com uma lista de instruções implica em severas limitações num problema como reconhecimento de voz, onde não se conhece uma regra. Assim é mais fácil ensinar uma rede com exemplos que são pares questões/respostas corretas e deixar que a rede neural descubra a regra por si mesma. Este método é chamado de aprendizado supervisionado pois necessita que um supervisor forneça as respostas corretas para as questões que a rede deve aprender.

Existem muitas maneiras ou algoritmos diferentes de uma rede neural ser treinada. ou seja, dos pesos serem ajustados para desempenharem uma dada tarefa. Além do algoritmo Perceptron que foi desenvolvido para treinar um perceptron Booleano. muitos outros algoritmos foram criados, tais como o algoritmo Adaline [43] para treinar perceptrons lineares e o algoritmo Backpropagation [43], para treinar perceptrons com muitas camadas. O estudo do aprendizado em redes neurais feito através da pesquisa de algoritmos específicos de treinamento nem sempre é eficiente. uma vez que os algoritmos podem apresentar uma convergência muito lenta ou até mesmo nem convergir para a solução correta. Mas não se pode concluir que o fato de algum algoritmo não convergir implique na inexistência de solução para um determinado problema.

Neste trabalho não estamos interessados em investigar algoritmos específicos de treinamento: nosso objetivo é estudar o aprendizado em perceptrons de uma maneira mais geral através da Mecânica Estatística. Nesta abordagem é possível encontrar a região do espaço dos pesos onde existem soluções para uma determinada tarefa. independente da existência de um algoritmo capaz de encontrar estas soluções.

O estudo analítico do aprendizado através da análise do espaço dos pesos foi inici-

ado por Gardner [15]. Nesta abordagem os exemplos de treinamento são pares $\left(\boldsymbol{\xi}^l, t^l\right)$. $l = 1, \ldots, P$ onde o vetor de N componentes $\boldsymbol{\xi}^l = \left(\xi_1^l, \xi_2^l, \ldots, \xi_N^l\right)$ é denominado padrão de entrada e a variável t^l é a saída associada ao padrão $\boldsymbol{\xi}^l$. Aprender consiste em procurar os pesos sinápticos adequados que implementam a regra geradora dos exemplos de treinamento. Assim, os pares de exemplos $\left(\boldsymbol{\xi}^l, t^l\right)$ fazem o papel das variáveis lentas (quenched) e os pesos sinápticos são as variáveis dinâmicas que se ajustam de forma a minimizar uma função energia, construída a partir dos exemplos. A análise das propriedades de equilíbrio deste sistema é feita através do formalismo das réplicas [4]. Em seu trabalho pioneiro Gardner estudou a capacidade de memorização de uma rede neural para o caso do mapa aleatório em que não há correlação entre os padrões de entrada $\boldsymbol{\xi}^l$ e suas respectivas saídas t^l. Naquele trabalho foi mostrado que, quando a rede é treinada com P exemplos $\left(\boldsymbol{\xi}^l, t^l\right)$ $(l = 1, \ldots, P)$ gerados aleatoriamente, existe solução com pesos contínuos apenas para $P < 2N$ no limite termodinâmico $N \to \infty$.

No mapa aleatório não há uma regra que associa os padrões de entrada $\boldsymbol{\xi}^l$ à saída t^l. Um caso mais interessante é o aprendizado supervisionado em que existe uma regra relacionando os pares entrada/saída. Após o aprendizado ter terminado, a rede adquire a capacidade de generalizar, ou seja, a capacidade de predizer a saída t associada a um padrão de entrada $\boldsymbol{\xi}$ que não pertence ao conjunto de treinamento. A grandeza que mede esta capacidade é chamada erro de generalização. O método desenvolvido por Gardner tem sido amplamente aplicado ao estudo analítico do aprendizado por exemplos em perceptrons lineares e Booleanos [45], [20], [36], [38].

Em perceptrons o aprendizado por exemplos foi primeiro estudado analiticamente pelo método de Gardner por Györgyi e Tishby [20]. Estes autores estudaram a capacidade de generalização de um perceptron Booleano com pesos contínuos. Eles também introduziram o problema em que os padrões de treinamento são deturpados por ruído, cuja intensidade é medida por um parâmetro γ que varia entre $\gamma = 0$ (mapa aleatório) e $\gamma = 1$ (padrões puros). Esta situação é bem mais geral e realística, uma vez que é bastante improvável que durante a fase de treinamento não haja nenhum tipo de ruído que interfira no processo de aprendizado da rede neural.

O aprendizado supervisionado em perceptrons Booleanos com pesos contínuos foi pro-

fundamente estudado numérica e analiticamente [45]. Também para esta mesma rede foram estudadas as capacidades de armazenamento e generalização no caso em que o perceptron é treinado com vários algoritmos de treinamento diferentes [36], [38]. Estes trabalhos tinham como objetivo investigar qual é o algoritno de treinamento que produz o menor erro de generalização. Outro estudo interessante em perceptrons Booleanos com pesos contínuos é o que concerne ao algoritmo de Bayes. Era sabido que este algoritmo fornece o melhor desempenho de generalização [12], porém não existia uma energia de treinamento associada a ele que permitisse o estudo analítico de suas propriedades pelo método de Gardner. Através de uma abordagem variacional, foi encontrada por Kinouchi e Caticha [25] a energia de treinamento relacionada com o perceptron de generalização ótima.

O perceptron linear com pesos reais foi primeiro estudado analiticamente com o método das réplicas por Levin e colaboradores [34], e também por Seung e colaboradores [45] para o aprendizado supervisionado. Uma questão interessante relativa ao perceptron linear é a normalização dos pesos. Foi mostrado por Hertz e colaboradores [23] que a escolha da normalização no perceptron linear desempenha um papel fundamental, mas no perceptron Booleano esta escolha é irrelevante. Nesta linha, Fontanari [13] estudou os efeitos sobre os erros de treinamento e generalização de um perceptron linear causados pela escolha da norma dos pesos, como fixa *a priori* ou como um parâmetro ajustável. No caso em que a norma dos pesos é fixa, Kinouchi e Caticha [26] investigaram um algoritmo ótimo de aprendizagem que minimiza o erro de generalização através do método variacional acoplado ao método das réplicas.

É importante notar que o estudo de redes neurais está centrado no ajuste de pesos, tanto no que diz respeito a pesquisar algoritmos de treinamento, quanto na abordagem da Mecânica Estatística. A maioria dos modelos de redes neurais leva em conta a plasticidade sináptica, de maneira que a principal variável do problema de aprendizado é o valor dos pesos sinápticos.

O fato de o cérebro apresentar um comportamento adaptativo, ou seja, ser um orgão dotado de grande flexibilidade, pode ser explicado pelas alterações nos pesos sinápticos [42]. Em estudos recentes tem sido comprovado que o valor das sinapses biológicas é

ajustado de acordo com a experiência [37], fazendo com que as sinapses desempenhem um papel fundamental nas tarefas cognitivas como aprendizado e generalização. Uma vez que o cérebro é um orgão flexível ele é capaz de se recuperar de lesões que provocam a perda de sinapses. As sinapses remanescentes se rearranjam para compensar estas perdas, de maneira que o funcionamento do cérebro seja pouco afetado. É sabido que os mecanismos de compensação que ocorrem no cérebro diminuem sua eficiência a medida que este orgão envelhece. No sistema nervoso em desenvolvimento ocorrem processos de eliminação de sinapses que ainda não são completamente entendidos [9]. O estudo destes processos poderá ajudar na compreenção dos motivos pelos quais a capacidade de recuperação do cérebro contra lesões diminui com o passar do tempo [42].

Motivados por estes mecanismos de compensação que ocorrem no cérebro, nosso principal interesse é investigar o comportamento de redes diluídas, ou seja, redes nas quais parte de suas conexões foi cortada. A diluição pode então ser utilizada para estudar a robusteza de redes neurais, artificiais e biológicas, contra o mal funcionamento de alguns elementos.

A diluição em redes neurais foi investigada primeiramente para o modelo de Hopfield de memória associativa em redes atratoras. Os efeitos da diluição na capacidade de recuperação e na dinâmica destas redes foram amplamente estudados [5],[10],[16],[46].

Existem diferentes tipos de dilução, no que diz respeito ao procedimento utilizado para eliminar os pesos. Foram introduzidos por Bouten, Komoda e Serneels [6] dois tipos de diluição que ocorrem durante o processo de treinamento: a *diluição fixa* em que os pesos são cortados aleatoriamente e a *diluição móvel* em que o corte dos pesos leva em conta os padrões a serem armazenados. Estes autores estudaram a capacidade de armazenamento de um perceptron Booleano com pesos de Ising $J = \pm 1$ no caso do mapa aleatório e consideraram apenas a diluição móvel. Mais tarde, os dois tipos de diluição foram estudados para um perceptron boleano com pesos contínuos também no caso do mapa aleatório [7]. Com relação à capacidade de armazenamento para perceptrons Booleanos, tem sido estudados outros modelos de diluição, como a diluição aleatória que não depende de processos de treinamento [51] e algoritmos de diluição que fornecem capacidades de armazenamento ótimas [14], [31].

No caso do aprendizado por exemplos foi estudado o efeito das diluições no desempenho de generalização de um perceptron Booleano com acoplamentos de Ising [2] e de um perceptron Booleano com pesos contínuos [32]. Porém, neste caso a solução com simetria de réplicas é localmente instável e, consequentemente, os resultados obtidos se tornam pouco confiáveis. A maioria destes trabalhos normalmente aborda diluições em perceptrons Booleanos que ocorrem antes do término do processo de aprendizado.

Neste trabalho além de estudarmos os efeitos dos dois tipos de diluição mencionados acima, vamos introduzir um outro tipo de diluição que ainda não havia sido abordado de forma sistemática na literatura: é a diluição que ocorre após o término do processo de treinamento. Neste caso, a rede é treinada com todas as suas conexões e, uma vez terminado o treinamento, o corte dos pesos é efetuado. Os pesos podem ser cortados de diversas maneiras diferentes, que serão descritas no próximo capítulo. A fim de podermos comparar os efeitos dos vários tipos de diluição, a fração dos pesos deletados deverá ser sempre a mesma em todos os casos. Além da diluição, vamos estudar o efeito do ruído nos exemplos, que atua durante a fase de treinamento da rede. O estudo da diluição após o término do treinamento tem grande interesse prático por duas razões: (1) auxilia na identificação das componentes mais importantes da rede neural, cujo dano possa causar maior prejuízo ao desempenho do sistema; (2) está intimamente ligada a popular técnica da poda que consiste em cortar os pesos menores na esperança de diminuir o erro de generalização e também de diminuir a memória necessária para armazenar os pesos da rede neural [22]. Esta técnica, ligeiramente aperfeiçoada, foi denominada de "dano cerebral ótimo" e aplicada com sucesso em uma rede de muitas camadas treinada com o algoritmo Backpropagation para o reconhecimento de caracteres manuscritos em código de endereçamento postal [22].

Esta tese está organizada da conforme descrito a seguir: No capítulo 2 será explicado o processo de treinamento das redes, bem como a ferramenta matemática empregada nos cálculos dos erros de treinamento e generalização. Serão mostrados também os tipos de diluições e ruídos e a maneira com que são adicionados ao treinamento. No capítulo 3 estudaremos o aprendizado supervisionado do perceptron linear treinado com o algoritmo Adaline. Vamos comparar os efeitos dos vários tipos de diluição sobre os erro de

treinamento e generalização na presença de ruído adicionado aos padrões apresentados ao estudante. No capítulo 4 estudaremos os efeitos da diluição realizada após o término do processo de treinamento em um perceptron Booleano treinado com vários algoritmos diferentes: regra pseudo-inversa, regra de Hebb, algoritmo de Gibbs, algoritmo de estabilidade ótima e algoritmo de Bayes. Finalmente, no capítulo 5 apresentamos nossas principais conclusões e perspectivas de trabalho.

Capítulo 2

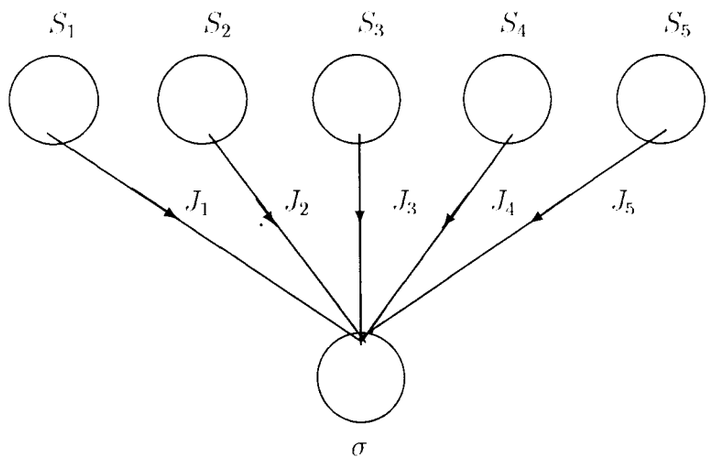
O Modelo

Vamos considerar um perceptron composto por N neurônios sensores binários cujos estados podem assumir os valores $S_i = \pm 1$ $(i = 1, \ldots, N)$. N conexões sinápticas J_i $(i = 1, \ldots, N)$ que assumem valores reais e um neurônio motor cujo estado é dado por

$$\sigma = f\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i S_i\right), \tag{2.1}$$

onde f é a função de transferência.

A arquitetura de um perceptron com $N = 5$ é mostrada na figura abaixo:



2.1 Treinamento

Para que uma rede neural realize uma tarefa, é necessário que ela seja treinada. Uma

tarefa consiste em P associações entrada/saída que denominamos exemplos

$$\boldsymbol{\xi}^l = \left(\xi_1^l, \xi_2^l, \ldots, \xi_N^l\right) \rightarrow t^l \qquad\qquad l = 1, \ldots, P$$

onde ξ_i^l são variáveis binárias que podem assumir os valores ± 1. As saídas t^l tanto podem ser geradas segundo alguma regra, como podem ser aleatórias, ou seja, sem correlação entre os padrões de entrada $\boldsymbol{\xi}^l$ e os dígitos de saída t^l. O conjunto de exemplos é deno minado conjunto de treinamento e, no caso em que as saídas são aleatórias, o conjunto de associações entrada/saída é chamado de mapa aleatório. Como usual, vamos supor que o número de exemplos cresça linearmente com P , isto é, $P = \alpha N$.

A grandeza que mede o êxito da rede na realização de um conjunto de treinamento é a energia de treinamento que normalmente é definida como a soma de uma função monotônica crescente g da diferença entre a saída produzida pela rede σ^l e a saída desejada t^l para todos os exemplos:

$$E\left(\mathbf{J}, \boldsymbol{\xi}, t\right) = \sum_{l=1}^{P} g\left(t^l - \sigma^l\right). \tag{2.2}$$

Existem várias maneiras ou algoritmos diferentes de uma rede ser treinada. À cada algoritmo está associada uma energia de treinamento $E\left(\mathbf{J}, \boldsymbol{\xi}, t\right)$ específica. O objetivo do treinamento é ajustar os pesos através de algum procedimento que vise à minimização desta energia. O procedimento mais comumente usado no ajuste dos pesos é a regra do gradiente descendente

$$\frac{dJ_i}{dt} = -\eta \frac{\partial E\left(\mathbf{J}\right)}{\partial J_i}, \tag{2.3}$$

onde $\eta \ll 1$ é uma constante de proporcionalidade que representa a taxa de aprendizado. Porém, existem algumas funções energia que não são diferenciáveis e, nestes casos, pode-mos usar outras maneiras de ajustar os pesos como o algoritmo de Monte Carlo.

Dependendo do número e da natureza dos exemplos pode não existir um conjunto de pesos $\mathbf{J}$ para o qual a energia de treinamento se anule. A razão entre o número máximo de exemplos para o qual existe pelo menos um conjunto de pesos com energia de treinamento nula e o número de neurônios da camada de entrada N é definida como a capacidade de armazenamento denotada, α_c. É importante salientar que, como neste trabalho o

aprendizado é abordado do ponto de vista da Mecânica Estatística, não vamos encontrar um conjunto de pesos específicos que possibilite o aprendizado de um particular conjunto de treinamento. Ao contrário, nossa ênfase será nas propriedades médias ou típicas de um conjunto de pesos que realiza um conjunto de treinamento típico. Para termos uma idéia quantitativa do desempenho da rede vamos estudar curvas de aprendizado, isto é, os erros de treinamento e generalização como função da razão $\alpha = P/N$.

Para o caso do perceptron linear, vamos estudar as curvas de aprendizado apenas para um algoritmo cuja energia de treinamento é simplesmente dada por

$$E(\mathbf{J}, \boldsymbol{\xi}, t) = \sum_{l=1}^{P} \left(t^l - \sigma^l\right)^2 \tag{2.4}$$

com

$$\sigma^l = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i^l. \tag{2.5}$$

O algoritmo de treinamento que ajusta os pesos de maneira a realizar os P exemplos $\left(\boldsymbol{\xi}^l, t^l\right)$ é dado pela regra delta [43], também conhecida como algoritmo Adaline. Nesta regra a componente J_i é ajustada à cada apresentação de um padrão $\boldsymbol{\xi}^l$, da seguinte maneira

$$\Delta_l J_i = \eta \left(t^l - \sigma^l\right) \xi_i^l. \tag{2.6}$$

onde t^l é a saída desejada, σ^l é a saída produzida pela rede na apresentação do padrão $\boldsymbol{\xi}^l$, ξ_i^l é o valor da i-ésima componente deste padrão, e $\Delta_l J_i$ é a variação da componente J_i do vetor $\mathbf{J}$. O valor final de J_i é encontrado após a apresentação de um ciclo completo, ou seja, após a apresentação dos P padrões. Se a constante de proporcionalidade η for suficientemente pequena esta regra encontra os pesos que minimizam a energia de treinamento [43]. O vetor $\mathbf{J}$ será encontrado após a repetição deste procedimento para todas as N componentes J_i. Foi mostrado que abaixo da capacidade de armazenamento ($\alpha < \alpha_c$), o vetor encontrado com a regra delta coincide com o vetor construído pela regra pseudo-inversa [11]. A pseudo-inversa gera cada componente J_i de acordo com a expressão [29]

$$J_i = \frac{1}{N} \sum_{l,m} t^l \left(\mathcal{C}^{-1}\right)_{lm} \xi_i^m \tag{2.7}$$

onde C é uma matriz simétrica $P \times P$ cujos elementos são dados por

$$C_{lm} = \frac{1}{N} \sum_i \xi_i^l \xi_i^m. \tag{2.8}$$

Para o caso do perceptron Booleano, vamos comparar as curvas de aprendizado obtidas com cinco algoritmos diferentes, dentre eles o algoritmo de Bayes que fornece o menor erro de generalização possível. No caso Booleano é conveniente definir a estabilidade Δ^l,

$$\Delta^l = \frac{t^l}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i^l, \tag{2.9}$$

pois esta grandeza fornece informações úteis sobre o aprendizado, uma vez que se $\Delta^l > 0$ para todo l, a rede encontra os pesos corretos, isto é, possui energia de treinamento nula. A energia de treinamento relativa a todos os algoritmos que vamos estudar no caso Booleano pode ser escrita como

$$E(\mathbf{J}, \boldsymbol{\xi}, t) = \sum_{l=1}^{P} g\left(\Delta^l\right). \tag{2.10}$$

Os algoritmos de treinamento do perceptron Booleano aqui estudados são:

- Regra pseudo-inversa: nesta regra os pesos são obtidos pela minimização da energia [38]

  $$E(\mathbf{J}) = \frac{1}{2} \sum_l \left(1 - \Delta^l\right)^2 \tag{2.11}$$

  e o algoritmo de treinamento é a regra delta dada pela equação (2.6).

- Regra de Hebb: o vetor peso construído a partir desta regra

  $$J_i = \frac{1}{\sqrt{N}} \sum_{l=1}^{P} t^l \xi_i^l, \tag{2.12}$$

  minimiza a seguinte energia de treinamento [19]

  $$E(\mathbf{J}) = -\sum_l \Delta^l. \tag{2.13}$$

- Algoritmo de Gibbs: os pesos são obtidos pela minimização da energia

  $$E(\mathbf{J}) = \sum_l \Theta\left(K - \Delta^l\right) \tag{2.14}$$

  onde $K \geq 0$ é o parâmetro de margem e o algoritmo de treinamento é o método de Monte Carlo.

- Algoritmo de estabilidade ótima: os pesos são obtidos pela minimização da energia

$$E(\mathbf{J}) = \sum_l \Theta\left(K_c - \Delta^l\right) \tag{2.15}$$

onde K_c é o maior valor do parâmetro de margem para o qual o algoritmo de Gibbs encontra um conjunto de pesos que realiza corretamente os P exemplos, ou seja, que produza energia de treinamento nula.

- Algoritmo de Bayes: os pesos minimizam a energia $E(\mathbf{J})$ extraída da abordagem variacional desenvolvida por Kinouchi e Caticha [25]. O erro de generalização de Bayes foi calculado analiticamente através da abordagem da Mecânica Estatística por Opper e Haussler [39]- [40]. Estes autores propuseram também um algoritmo de treinamento que produzia o erro de Bayes. Acreditava-se então que o desempenho de Bayes não poderia ser implementado por um perceptron de uma camada, além do que não existia uma energia de treinamento associada a ele. Mais tarde foi mostrado por Watikin [49] que era possível implementar este algoritmo com um perceptron simples, e na sequência, Kinouchi e Caticha calcularam através de um método variacional a energia de treinamento associada a este algoritmo.

Conforme mencionado, as P associações entrada/saída (exemplos) com as quais a rede é treinada podem ser aleatórias ou pode existir uma regra que associa os padrões de entrada $\boldsymbol{\xi}^l = \left(\xi_1^l, \xi_2^l, \ldots, \xi_N^l\right)$ ao dígito de saída t^l. A regra geradora dos exemplos pode ser modelada por outra rede neural que denominamos rede mestre e, consequentemente, a rede que deverá realizar o mapa gerado por esta rede será denominada rede estudante. Na rede mestre as N conexões sinápticas J_i^0 $(i = 1, \ldots, N)$ são variáveis aleatórias estatisticamente independentes com valor médio zero e variância M. Dado o padrão $\boldsymbol{\xi}^l$ esta rede gera o dígito t^l através da relação

$$t^l = f\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^0 \xi_i^l\right). \tag{2.16}$$

A rede mestra gera um mapa, um conjunto de 2^N associações compostas por todas as possíveis configurações de entrada $\{\boldsymbol{\xi}\}$ e suas respectivas saídas $\{t\}$. O treinamento consiste em apresentar uma parte deste mapa contendo apenas P pares entrada/saída $\left(\boldsymbol{\xi}^l, t^l\right)$ $(l = 1, \ldots, P)$ à rede estudante. Esta rede deverá procurar no espaço dos pesos,

sujeito a algum tipo de vínculo, o conjunto de pesos que reproduz o mapa corretamente. Nesta tese adotamos que as redes mestre e estudante têm a mesma arquitetura, ou seja, são perceptrons de uma única camada com pesos adaptativos.

2.2 Ruído

O treinamento da rede estudante é feito na presença de ruído, ou seja, durante o treinamento ocorre algum tipo de interferência na comunicação entre a rede estudante e a rede mestre. O conjunto de treinamento pode ser corrompido pelo ruído de duas formas: alterando-se os padrões de entrada apresentados à rede estudante ou alterando-se o dígito de saída gerado pela rede mestre.

No caso da alteração dos padrões de entrada, cada componente ξ_i^l é corrompida pelo ruído. Assim sendo, a rede estudante é treinada com $P = \alpha N$ exemplos $\left(\mathbf{S}^l, t^l\right)$ $(l = 1, \ldots, P)$, onde $\mathbf{S}^l$ são os padrões corrompidos cujas componentes S_i^l obedecem a distribuição de probabilidade condicional

$$P\left(S_i^l \mid \xi_i^l\right) = \frac{1+\gamma}{2}\delta\left(S_i^l - \xi_i^l\right) + \frac{1-\gamma}{2}\delta\left(S_i^l + \xi_i^l\right). \tag{2.17}$$

Se $\gamma = 1$ temos o problema de aprendizado com exemplos puros, enquanto que se $\gamma = 0$ temos o problema do mapa aleatório. Este tipo de ruído será abordado tanto no treinamento do perceptron linear como no treinamento do perceptron Booleano.

No caso da alteração do dígito de saída gerado pela rede mestre, a rede estudante é treinada com $P = \alpha N$ exemplos $\left(\boldsymbol{\xi}^l, \zeta^l\right)$ $(l = 1, \ldots, P)$ onde ζ^l é gerado pela distribuição de probabilidade condicional:

$$P\left(\zeta^l \mid t^l\right) = (1-\chi)\,\delta\left(\zeta^l - t^l\right) + \chi\delta\left(\zeta^l + t^l\right). \tag{2.18}$$

O parâmetro $0 \leq \chi \leq 1/2$ mede a intensidade do ruído, quanto maior for o valor de χ maior será a intensidade do ruído. O estudo deste tipo de ruído só faz sentido no caso do treinamento de perceptrons Booleanos, pois as saídas são binárias.

2.3 Generalização

Inferir regras a partir de exemplos é a principal habilidade envolvida nas atividades cognitivas e o principal fator levado em conta no projeto de redes neurais. O treinamento com exemplos tem como objetivo a extração sistemática de regras que estão implícitas nestes exemplos. Sem a capacidade de generalizar todo o conhecimento teria que ser ensinado em itens separados, além de que seria inconcebível e inútil treinar uma rede neural exaustivamente com todo o conjunto de treinamento.

Uma vez que a rede foi treinada, ela deve ser capaz de generalizar, ou seja, de classificar corretamente um padrão de entrada que não pertence ao conjunto de treinamento. Como a energia de treinamento mede apenas o desempenho da rede em classificar um conjunto limitado de exemplos, vamos introduzir a energia de generalização, que deve medir o desempenho da rede em predizer corretamente o dígito de saída associado a um padrão aleatório (padrão teste) a ela apresentado.

Mesmo o treinamento sendo feito na presença de ruído, a energia de generalização pode ser medida de duas maneiras: os padrões de teste apresentados à rede estudante são puros ou estão corrompidos por ruído.

A energia de generalização relativa a apresentação de padrões puros é definida como

$$E_g\left(\{J_i\}\right) = \frac{1}{2}\int \prod_i d\xi_i P\left(\xi_i\right)\left[t - \sigma\left(\{J_i\}, \boldsymbol{\xi}\right)\right]^2 \tag{2.19}$$

onde $\sigma\left(\{J_i\}, \boldsymbol{\xi}\right)$ é a resposta da rede estudante ao padrão de entrada aleatório $\boldsymbol{\xi}$, cujas componentes são geradas pela distribuição de probabilidade

$$P\left(\xi_i\right) = \frac{1}{2}\delta\left(\xi_i - 1\right) + \frac{1}{2}\delta\left(\xi_i + 1\right). \tag{2.20}$$

Esta energia mede o valor médio do quadrado da diferença entre a saída produzida pela rede estudante $\sigma\left(\{J_i\}, \boldsymbol{\xi}\right)$ e a saída desejada t.

Para que a energia de generalização expressa pela equação (2.19) possa ser calculada é necessário que se especifique o tipo de rede que está sendo considerado. Quando estas redes são perceptrons lineares temos

$$t = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i^0 \xi_i \tag{2.21}$$

e

$$\sigma = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i. \tag{2.22}$$

Neste caso, o cálculo explícito de E_g resulta em [45]

$$E_g\left(\{J_i\}\right) = \frac{1}{2}\left(Q + M - 2R\right) \tag{2.23}$$

onde Q é a norma quadrada do perceptron estudante

$$Q = \frac{1}{N} \sum_{i}^{N} J_i^2, \tag{2.24}$$

R é a correlação entre o perceptron mestre e o estudante

$$R = \frac{1}{N} \sum_{i}^{N} J_i^0 J_i, \tag{2.25}$$

e M é a norma quadrada do perceptron mestre

$$M = \frac{1}{N} \sum_{i}^{N} \left(J_i^0\right)^2. \tag{2.26}$$

Quando as redes consideradas são perceptrons Booleanos temos

$$t = \text{sign}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^0 \xi_i\right) \tag{2.27}$$

e

$$\sigma = \text{sign}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i \xi_i\right). \tag{2.28}$$

Como σ e t são variáveis de Ising, podemos reescrever a energia de generalização como

$$E_g\left(\{J_i\}\right) = \int \prod_i d\xi_i P\left(\xi_i\right) \Theta\left[-t\sigma\left(\{J_i\}, \boldsymbol{\xi}\right)\right]. \tag{2.29}$$

O cálculo explícito das integrais sobre ξ_i leva a [45]

$$E_g\left(\{J_i\}\right) = \frac{1}{\pi} \arccos\left(\frac{R}{\sqrt{Q}}\right). \tag{2.30}$$

Quando os padrões de entrada apresentados à rede estudante estão corrompidos com ruído a energia de generalização para perceptrons lineares é dada por [13]

$$E_g\left(\{J_i\}\right) = \frac{1}{2}\left(Q + M - 2\gamma R\right). \tag{2.31}$$

e para perceptrons Booleanos é dada por [20]

$$E_g\left(\{J_i\}\right) = \frac{1}{\pi}\arccos\left(\frac{R\gamma}{\sqrt{Q}}\right). \tag{2.32}$$

No caso Booleano, quando a saída produzida pela rede mestre está corrompida por ruído a energia de generalização é dada por [40]

$$E_g\left(\{J_i\}\right) = \chi + (1-2\chi)\frac{1}{\pi}\arccos\left(\frac{R}{\sqrt{Q}}\right). \tag{2.33}$$

No caso do treinamento com perceptrons lineares sem ruído, existe um número mínimo de exemplos com os quais a rede estudante deve ser treinada para que realize perfeitamente o mapa completo. Isto não ocorre no treinamento com perceptrons Booleanos, pois neste caso são necessários infinitos exemplos para que a energia de generalização se anule.

2.4 Diluição

As lesões no cérebro podem ocorrer por diversas causas que variam desde injúrias sofridas em acidentes até a remoção cirúrgica de um dos hemisférios cerebrais. Curiosamente, é possível recuperar parcial ou totalmente as funções cerebrais perdidas dependendo da idade em que tais lesões tenham ocorrido.

O cérebro é um orgão altamente especializado onde cada um dos dois hemisférios (direito e esquerdo) é responsável por conjuntos de funções diferentes. Por exemplo, o hemisfério esquerdo é responsável pela linguagem e por raciocínios lógicos, enquanto o hemisfério direito é responsável por abstrações e criatividade. Esta especialização se fortalece a medida que o cérebro envelhece. Uma criança que já tenha aprendido a falar e sofre uma lesão nas regiões da linguagem é capaz de recomeçar a falar em poucos anos, pois há uma transferência da dominância da linguagem do hemisfério esquerdo para o direito. Este fato ilustra a plasticidade cerebral, ou seja, a grande capacidade de adaptação da estrutura do cérebro. A transferência da dominância da linguagem não é mais possível a partir do décimo ano, pois as regiões correspondentes do hemisfério não dominante da linguagem já assumiram outras tarefas. Assim, quanto mais jovem for o cérebro, mais plástico ele será, ou seja, a capacidade de recuperação de uma criança que tenha sofrido algum tipo de lesão será muito maior que a de um adulto.

O objetivo do estudo da diluição em redes neurais é modelar e entender os efeitos das lesões em sistemas artificiais dotados da capacidade de aprender. Este estudo tem grande utilidade tanto do ponto de vista biológico quanto do ponto de vista prático. Do ponto de vista biológico, essa investigação pode ajudar no entendimento dos mecanismos de funcionamento do cérebro, pois podemos comparar sintomas apresentados por pessoas que sofreram algum tipo de lesão com o comportamento de uma rede neural diluída [24]. Do ponto de vista prático, esse estudo ajuda a identificar quais são os componentes cuja destruição pode afetar mais drasticamente o desempenho da rede.

A diluição de uma rede neural consiste em anular o valor de uma fração de seus pesos. O corte dos pesos pode ser efetuado basicamente de duas maneiras diferentes, durante ou após o término do processo de aprendizado. No caso do corte dos pesos ser efetuado durante o processo de aprendizado existem duas possibilidades: a diluição móvel em que os pesos são cortados de maneira a minimizar os efeitos da lesão na energia de treinamento e a diluição fixa em que os pesos são cortados aleatoriamente e permanecem fixos durante o estágio de aprendizado não levando em conta a energia de treinamento. Na diluição que ocorre após o término do processo de treinamento, o critério usado para eliminar os pesos é a intensidade dos mesmos. Como queremos comparar o efeito dos vários tipos de diluição é necessário que haja um parâmetro que forneça o grau de diluição. Desta maneira vamos introduzir o parâmetro de diluição κ que expressa a fração dos pesos remanescentes.

Para efetuar a diluição móvel introduzimos as variáveis $c_i = 0, 1$ onde o parâmetro de diluição é dado por

$$\frac{1}{N}\sum_{i=1}^{N} c_i = \kappa \tag{2.34}$$

de modo que podemos definir um novo conjunto de pesos $\{W_i\}$ com $J_i = c_i W_i$ $(i = 1, \ldots, N)$, sendo que tanto c_i como W_i devem ser determinados de forma a minimizar a energia de treinamento. Para efetuar a diluição fixa cortamos $(1 - \kappa) N$ pesos, fazendo $J_i = 0$ para $(i = \kappa N + 1, \ldots, N)$. Conforme mencionado acima, a diluição móvel é apropriada para modelar as lesões sofridas por pessoas com pouca idade, enquanto que a diluição fixa modela as lesões sofridas por adultos.

No caso em que os pesos são cortados após o término do processo de treinamento, vamos considerar três possibilidades: cortar os $(1 - \kappa) N$ pesos menores, cortar os $(1 - \kappa) N$

pesos maiores e cortar $(1-\kappa)N$ pesos aleatoriamente independente de sua intensidade. Para implementar este tipo de diluição fazemos com que a rede seja treinada com toda a sua capacidade sináptica e, uma vez terminado o treinamento, eliminamos os pesos da maneira desejada.

As diluições móvel e fixa basicamente modelam a plsticidade do cérebro, uma vez que ocorrem antes do término do processo de aprendizado. A diluição efetuada após o estágio de treinamento não leva em conta os mecanismos de aprendizado ou auto-reparo, assim ela é apropriada para estudar a robusteza da rede contra danos causados por agentes externos durante o funcionamento desta.

No caso do perceptron linear vamos investigar e comparar os efeitos dos cinco tipos de diluição sobre os erros de treinamento e generalização. No caso do perceptron Booleano, que será treinado com cinco algoritmos diferentes, vamos estudar apenas os efeitos da diluição que ocorre após o término do aprendizado, com o objetivo de investigar qual regra de aprendizado ou algoritmo é mais robusta contra o corte dos pesos.

2.5 Mecânica Estatística do Aprendizado

Nesta seção mostramos como serão calculados nos capítulos 3 e 4 os valores médios das energias de treinamento e generalização através da Mecânica Estatística. Com essas grandezas podemos construir as curvas de aprendizado relativas aos algoritmos já mencionados. Mostraremos também, como será calculada no capítulo 3 a distribuição de probabilidade dos pesos que não foram eliminados no caso da diluição móvel.

2.5.1 Erro de Treinamento

Os processos de aprendizado abordados neste trabalho são tratados sob o ponto de vista da Mecânica Estatística no ensemble canônico. Neste contexto devemos definir uma energia que depende das configurações que o sistema pode assumir. No nosso caso, a energia que descreve o sistema é a energia de treinamento, que depende das configurações dos pesos $\{J_i\}$ e dos exemplos de treinamento $(\boldsymbol{\xi}^l, t^l)$.

Supondo que o sistema considerado esteja em contato com um banho térmico numa

temperatura T, a probabilidade de se encontrar este sistema em uma determinada configuração de pesos é dada pela distribuição de probabilidade de Gibbs

$$P(\{J_i\}) = \frac{e^{-\beta E(\{J_i\})}}{Z}, \tag{2.35}$$

onde $\beta = 1/T$ e Z é a função da partição

$$Z = \mathrm{Tr} \exp\left[-\beta E(\{J_i\})\right]. \tag{2.36}$$

Aqui Tr indica a integração sobre todas as configurações de pesos permitidas que devem satisfazer o vínculo da normalização

$$Q = \frac{1}{N}\sum_{i}^{N} J_i^2 \tag{2.37}$$

e o vínculo da diluição que força o vetor $\mathbf{J}$ a possuir κN componentes diferentes de zero. Nosso principal objetivo é caracterizar a configuração de pesos $\{J_i\}$ que minimiza a energia de treinamento para um dado conjunto de exemplos.

Uma vez que a energia de treinamento depende das configurações dos pesos e dos exemplos, o valor esperado desta energia, ou seja, o erro de treinamento médio é definido como

$$\epsilon_t = \frac{1}{P}\left\langle\left\langle E(\{J_i\})\right\rangle_T\right\rangle \tag{2.38}$$

onde $\langle\ldots\rangle_T$ é a média térmica ou média sobre todas as configurações permitidas de $\{J_i\}$,

$$\langle\ldots\rangle_T = \frac{\mathrm{Tr}\,(\ldots)\, e^{[-\beta E(\{J_i\})]}}{\mathrm{Tr}\, e^{[-\beta E(\{J_i\})]}},$$

e $\langle\langle\ldots\rangle\rangle$ é a média sobre as variáveis lentas, que neste caso são as variáveis aleatórias estatisticamente independentes S_i^l, ξ_i^l e J_i^0. Como é sabido da Mecânica Estatística, os valores esperados das grandezas físicas podem ser obtidos das derivadas da *energia livre* f mediada sobre todas as realizações do conjunto de treinamento,

$$-\beta f = \lim_{N\to\infty} \frac{1}{N}\left\langle\ln Z\right\rangle. \tag{2.39}$$

Assim o erro de treinamento médio, (2.38), associado ao ensemble de configurações que minimizam a energia de treinamento é obtido através da relação

$$\epsilon_t = \frac{1}{\alpha}\lim_{\beta\to\infty}\frac{\partial(\beta f)}{\partial\beta}. \tag{2.40}$$

O cálculo da média sobre as variáveis lentas pode ser efetuado através da aplicação do método das réplicas, tradicionalmente empregado no estudo de vidros de spin [4]. Este método consiste basicamente em usar a identidade

$$\langle\langle \ln Z \rangle\rangle = \lim_{n\to 0} \frac{1}{n} \ln \langle\langle Z^n \rangle\rangle \,. \tag{2.41}$$

onde Z^n é equivalente a uma função de partição de n sistemas idênticos (réplicas) não interagentes. Para obtermos $\langle\langle \ln Z \rangle\rangle$ primeiramente calculamos $\langle\langle Z^n \rangle\rangle$ para n inteiro e então fazemos a continuação analítica para $n \approx 0$. Uma vez que a continuação analítica de n inteiro para n real não pode ser justificada de forma rigorosa, os resultados obtidos com a aplicação deste método em geral são validados pela comparação com os resultados de simulações de Monte Carlo.

2.5.2 Erro de Generalização

A energia de generalização também depende da configuração dos pesos $\{J_i\}$ e dos exemplos de treinamento, assim a grandeza que nos interessa é o valor esperado desta energia ou seja, o erro de generalização médio definido como

$$\epsilon_g = \lim_{\beta\to\infty} \left\langle\left\langle \langle E_g(\{J_i\})\rangle_T \right\rangle\right\rangle . \tag{2.42}$$

O cálculo de ϵ_g é similar ao cálculo de ϵ_t, porém, para facilitar a realização das médias sobre as variáveis lentas, vamos definir uma energia efetiva

$$E^{ef}(\{J_i\}) = E(\{J_i\}) + hE_g(\{J_i\}) \,. \tag{2.43}$$

de maneira que a função de partição agora é dada por

$$Z^{ef} = \mathrm{Tr} \exp\left[-\beta E^{ef}(\{J_i\})\right] . \tag{2.44}$$

A energia livre média pode então ser calculada através da aplicação do método das réplicas de modo que o valor médio do erro de generalização é simplesmente dado por

$$\epsilon_g = \lim_{\beta\to\infty} -\frac{1}{\beta}\frac{\partial}{\partial h} \left\langle\left\langle \ln Z^{ef} \right\rangle\right\rangle |_{h=0} \,. \tag{2.45}$$

2.5.3 Distribuição dos Pesos

No caso das diluições móvel e fixa os pesos são cortados durante o processo de aprendizado. O vetor solução **J** que implementa a tarefa desejada e satisfaz os vínculos é distribuído segundo a distribuição de probabilidade de Gibbs. Uma vez que o vetor **J** foi encontrado, para termos uma idéia quantitativa dos valores que as componentes remanescentes J_i podem assumir vamos calcular a probabilidade de que J_i assuma um valor no intervalo $[J, J + dJ]$. A densidade de probabilidade associada a esta probabilidade é definida como

$$P(J) = \lim_{\beta \to \infty} \langle \langle \delta (J_i - J) \rangle_T \rangle . \tag{2.46}$$

Para calcularmos $P(J)$ vamos introduzir um campo auxiliar h como no cálculo do erro de generalização, conforme será descrito em detalhes no próximo capítulo para o caso da diluição móvel.

Da análise da distribuição dos pesos obtida veremos que o valor absoluto dos pesos remanescentes é sempre maior que um certo valor limite. Este limite inferior depende do grau de diluição: quanto mais diluída for a rede, maior será o limite inferior. Assim, este resultado indica que o corte dos pesos menores após o aprendizado não deverá afetar muito drasticamente o erro de treinamento.

Capítulo 3

Perceptron Linear

Neste capítulo estudaremos as capacidades de aprendizado e de generalização de um perceptron linear com pesos reais. O perceptron linear é um modelo simples e não trivial em que as equações que o descrevem podem ser resolvidas integralmente de forma analítica.

Vamos estudar o comportamento dos erros de treinamento e de generalização quando a rede estudante for sujeita a ruído e diluição durante a fase de treinamento. É importante ressaltar que o algoritmo de treinamento considerado nesta abordagem é o algoritmo Adaline e que apenas para este algoritmo faremos a análise dos erros de treinamento e de generalização.

O estudo analítico do perceptron linear será realizado dentro do formalismo das réplicas. Nesta análise usaremos a prescrição simetria de réplicas para solucionar as equações de ponto de sela, e em seguida, testaremos a estabilidade desta solução. No caso do perceptron Booleano a prescrição simetria de réplicas é instável a partir da capacidade de armazenamento da rede α_c, o que não ocorre com o perceptron linear.

Este capítulo está organizado em quatro seções principais. Na primeira seção o modelo a ser estudado será descrito em detalhes. Na segunda seção será estudada a diluição realizada durante o aprendizado, a qual compreende as diluições móvel e fixa. Para estes dois tipos de diluição calculamos analiticamente os erros de treinamento e de generalização e testamos a estabilidade da prescrição simetria de réplicas. Para o caso da diluição móvel calculamos também a distribuição dos pesos que não foram eliminados no processo de diluição. Na terceira seção calculamos analiticamente os erros de generalização e de

treinamento para o caso da diluição que ocorre após o término do aprendizado, a qual compreende o corte dos pesos menores, o corte dos pesos maiores e o corte aleatório dos pesos. Na última seção comparamos o desempenho da rede para todos os tipos de diluição estudados.

3.1 Modelo

O perceptron estudante consiste de N unidades de entrada binárias $S_i = \pm 1\ (i = 1, \ldots, N)$, N pesos sinápticos reais $J_i\ (i = 1, \ldots, N)$ e uma unidade de saída linear

$$\sigma = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i S_i. \tag{3.1}$$

A tarefa do estudante é realizar o mapa entre as 2^N possíveis configurações de entrada $\{\xi\}$ e suas respectivas saídas $\{t\}$ geradas pelo perceptron mestre

$$t = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^0 \xi_i, \tag{3.2}$$

onde os pesos $J_i^0\ (i = 1, \ldots, N)$ são variáveis estatisticamente independentes extraídas da distribuição de probabilidade Gaussiana

$$P\left(J_i^0\right) = \frac{1}{\sqrt{2\pi M}} \exp\left[-\frac{\left(J_i^0\right)^2}{2M}\right]. \tag{3.3}$$

Embora derivemos todos os nossos resultados para M genérico, todas as figuras deste capítulo são obtidas com $M = 1$. Para realizar essa tarefa o estudante é treinado com $P = \alpha N$ pares entrada/saída $\left(\mathbf{S}^l, t^l\right)\ (l = 1, \ldots, P)$ onde t^l é a saída do mestre associada a entrada $\boldsymbol{\xi}^l$, e cada componente S_i^l é extraída da distribuição de probabilidade condicional

$$P\left(S_i^l \mid \xi_i^l\right) = \frac{1+\gamma}{2}\delta\left(S_i^l - \xi_i^l\right) + \frac{1-\gamma}{2}\delta\left(S_i^l + \xi_i^l\right) \tag{3.4}$$

com

$$P\left(\xi_i^l\right) = \frac{1}{2}\delta\left(\xi_i^l - 1\right) + \frac{1}{2}\delta\left(\xi_i^l + 1\right). \tag{3.5}$$

No caso do perceptron linear a energia de treinamento é definida como

$$E\left(\{J_i\}\right) = \frac{1}{2} \sum_{i=1}^{P} \left(t^l - \sigma^l\right)^2. \tag{3.6}$$

onde $\sigma^l = \sigma\left(\{J_i\}, \mathbf{S}^l\right)$ é a resposta do estudante à entrada com ruído $\mathbf{S}^l$, e t^l é a resposta do mestre ao padrão de entrada puro $\boldsymbol{\xi}^l$.

3.2 Diluição durante o aprendizado

Neste caso o corte dos pesos é feito antes do término do processo de aprendizado. Na diluição móvel o corte dos pesos é feito de maneira a minimizar o efeito da diluição sobre o erro de treinamento e na diluição fixa o corte dos pesos é feito aleatoriamente.

3.2.1 Diluição Móvel

Neste caso o parâmetro κ que mede o grau de diluição, isto é, a fração dos pesos com valores não nulos, é imposto pelo vínculo

$$\kappa = \frac{1}{N}\sum_{i=1}^{N} c_i, \tag{3.7}$$

onde $c_i = 0, 1$ são variáveis binárias que nos permitem redefinir os pesos sinápticos J_i tal que $J_i = c_i W_i$ $(i = 1, \ldots, N)$. Em termos dos novos pesos W_i a energia de treinamento é escrita como

$$E_\kappa\left(\{W_i\}, \{c_i\}\right) = \frac{1}{2}\sum_{l=1}^{P}\left(t^l - \frac{1}{\sqrt{N}}\sum_{i=1}^{N} c_i W_i S_i^l\right)^2. \tag{3.8}$$

A rede estudante deve encontrar as configurações $\{c_i\}$ e $\{W_i\}$ que minimizam esta energia. A fim de evitarmos que as integrais sobre W_i divirjam, será necessário introduzir outros dois vínculos

$$Q = \frac{1}{N}\sum_{i=1}^{N} c_i W_i^2 \tag{3.9}$$

e

$$Q^0 = \frac{1}{N}\sum_{i=1}^{N} (1 - c_i) W_i^2. \tag{3.10}$$

Para obtermos a expressão do erro de treinamento dado pela equação (2.40) devemos calcular primeiramente a energia livre (2.39) usando o formalismo das réplicas. No caso da diluição móvel, a função de partição é dada por

$$Z = \sum_{\{\mathbf{c}\}} \delta_{Kr}\left(\sum_i c_i, \kappa N\right)\int_{-\infty}^{\infty} d\mu\left(\mathbf{W}\right)\exp\left[-\beta E_\kappa\left(\{W_i\}, \{c_i\}\right)\right] \tag{3.11}$$

onde

$$d\mu\left(\mathbf{W}\right) = \prod_i dW_i \delta\left(NQ - \sum_i c_i W_i^2\right)\delta\left(NQ^0 - \sum_i (1 - c_i) W_i^2\right) \tag{3.12}$$

e δ_{Kr} é a delta de Kronecker.

A média sobre o conjunto de treinamento faz com que haja um acoplamento entre as réplicas e introduz de forma natural os parâmetros de ordem que descrevem o sistema. As integrais sobre estes parâmetros são calculadas no limite termodinâmico $N \to \infty$ pelo método do ponto de sela. Estes cálculos estão mostrados detalhadamente no Apêndice A e levam a seguinte expressão para energia livre

$$\beta f^m = \lim_{n\to 0} \text{extr} \frac{1}{n} \left\{ G_0 \left(q_{ab}, \hat{q}_{ab}, R_a, \hat{R}_a, \hat{c}_a, Q^0, \hat{Q}^0_a, \hat{Q}_a \right) \right.$$

$$\left. + \alpha G_1 \left(q_{ab}, R_a \right) + G_2 \left(\hat{q}_{ab}, \hat{c}_a, \hat{Q}_a, \hat{Q}^0_a, \hat{R}_a \right) \right\} \tag{3.13}$$

onde

$$G_0 = -\sum_{a<b}^{n} q_{ab}\hat{q}_{ab} + \sum_{a}^{n} \left(\kappa \hat{c}_a + Q^0 \hat{Q}^0_a - Q\hat{Q}_a - R_a \hat{R}_a \right), \tag{3.14}$$

$$G_1 = \ln \int \prod_{a=1}^{n} \frac{dx_a}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \sum_{a}^{n} x_a^2 \left[1 + \beta \left(Q + M - 2\gamma R_a \right) \right] \right.$$

$$\left. -\beta \sum_{a<b}^{n} x_a x_b \left(q_{ab} + M - 2\gamma R_a \right) \right\} \tag{3.15}$$

e

$$G_2 = \ln \sum_{\{c^a = 0,1\}} \int \prod_{a=1}^{n} dW^a \exp \left\{ -\sum_{a}^{n} \left[\hat{c}_a c^a + \hat{Q}^0_a \left(1 - c^a \right) \left(W^a \right)^2 - \hat{Q}_a c^a \left(W^a \right)^2 \right] \right.$$

$$\left. + \sum_{a}^{n} c^a \hat{R}_a W^a J^0 + \sum_{a<b}^{n} \hat{q}_{ab} c^a c^b W^a W^b \right\} \tag{3.16}$$

O extremo na equação (3.13) é tomado sobre todos os parâmetros de ponto de sela $\left(\hat{c}_a, \hat{q}_{ab}, \hat{Q}_a, \hat{Q}^0_a, \hat{R}_a, q_{ab}, R_a \right)$. Os parâmetros de ordem físicos

$$q_{ab} = \frac{1}{N} \sum_{i=1}^{N} c_i^a c_i^b W_i^a W_i^b \qquad a < b \tag{3.17}$$

e

$$R_a = \frac{1}{N} \sum_{i=1}^{N} c_i^a W_i^a J_i^0 \tag{3.18}$$

medem a correlação entre dois estados de equilíbrio diferentes $\{J_i^a\}$ e $\left\{ J_i^b \right\}$ e a correlação entre o estado $\{J_i^a\}$ e a rede mestre $\{J_i^0\}$, respectivamente.

Para prosseguirmos no cáculo de βf^m seria necessário resolver as equações de ponto de sela para um n genérico e então tomarmos o limite $n \to 0$. Este procedimento seria muito complicado, por isso usaremos a prescrição simetria de réplicas que consiste em supor que os valores dos parâmetros de ordem são independentes dos seus índices de réplicas:

$$\begin{aligned} q_{ab} &= q \quad \text{e} \quad \hat{q}_{ab} = \hat{q} \quad \forall a < b; \\ R_a &= R \quad \text{e} \quad \hat{R}_a = \hat{R} \quad \forall a; \\ \hat{Q}_a &= \hat{Q} \;\; , \;\; \hat{Q}_a^0 = \hat{Q}^0 \;\; \text{e} \;\; \hat{c}_a = \hat{c} \;\; \forall a. \end{aligned} \tag{3.19}$$

Nos Apêndices (B.1) e (C.1) calculamos G_1 e G_2, respectivamente, usando esta prescrição. Assim, no limite $n \to 0$, obtemos a expressão da energia livre média

$$\begin{aligned} -\beta f_{sr}^m &= \frac{1}{2}\hat{q}q - Q\hat{Q} + Q^0\hat{Q}^0 - R\hat{R} - \frac{1-\kappa}{2}\ln\hat{Q}^0 + \kappa\hat{c}' + \frac{\kappa}{2}\ln 2 + \frac{1}{2}\ln\pi \\ &\quad -\frac{\alpha}{2}\ln\left[1+\beta\left(Q-q\right)\right] - \frac{\alpha\beta}{2}\frac{q+M-2\gamma R}{1+\beta\left(Q-q\right)} \\ &\quad + \int Dz \ln\left\{1 + \frac{\exp\left[-\hat{c}' + z^2\left(\hat{q}+M\hat{R}^2\right)/4\left(\frac{\hat{q}}{2}-\hat{Q}\right)\right]}{\sqrt{\hat{q}/2-\hat{Q}}}\right\}, \end{aligned} \tag{3.20}$$

onde definimos o parâmetro $\hat{c}' = \hat{c} - \frac{1}{2}\ln\hat{Q}^0$ a fim de desacoplar $\hat{Q}^0$ dos outros parâmetros relevantes. Ainda, introduzimos a notação para a medida Gaussiana

$$Dz = \frac{dz}{\sqrt{2\pi}} e^{-z^2/2}. \tag{3.21}$$

Os parâmetros de ordem na prescrição simetria de réplicas $\left(\hat{c}, \hat{q}, q, \hat{Q}, \hat{Q}^0, \hat{R}, R\right)$ são dados pelas equações de ponto de sela que são obtidas igualando-se a zero as derivadas de $-\beta f_{sr}^m$ com relação a esses parâmetros. Este procedimento leva às seguintes equações:

$$\hat{q} = \frac{\alpha\beta^2\left(Q+M-2\gamma R\right)}{\left[1+\beta\left(Q-q\right)\right]}, \tag{3.22}$$

$$\hat{R} = \frac{\alpha\beta\gamma}{1+\beta\left(Q-q\right)}, \tag{3.23}$$

$$\hat{Q}^0 = \frac{1-\kappa}{2Q^0}, \tag{3.24}$$

$$\kappa = \frac{1}{\eta^{1/2}} \int Dz \frac{e^{-\hat{c}'} e^{\Xi z^2}}{\Delta}, \tag{3.25}$$

$$q = \frac{1}{2\eta^{5/2}} \int Dz \frac{e^{-\hat{c}'} e^{\Xi z^2}}{\Delta} \left[\eta \left(1 - z^2\right) + z^2 \frac{M\hat{R}^2 + \hat{q}}{2} \right], \tag{3.26}$$

$$Q = \frac{1}{2\eta^{5/2}} \int Dz \frac{e^{-\hat{c}'} e^{\Xi z^2}}{\Delta} \left[\eta + z^2 \frac{M\hat{R}^2 + \hat{q}}{2} \right], \tag{3.27}$$

e

$$R = \frac{M\hat{R}}{2\eta^{3/2}} \int Dz z^2 \frac{e^{-\hat{c}'} e^{\Xi z^2}}{\Delta}. \tag{3.28}$$

Aqui utilizamos as seguintes notações:

$$\Xi = \frac{\left(\hat{q} + M\hat{R}^2\right)}{\left(\frac{\hat{q}}{2} - \hat{Q}\right)}, \tag{3.29}$$

$$\eta = \frac{\hat{q}}{2} - \hat{Q} \tag{3.30}$$

e

$$\Delta = 1 + e^{-\hat{c}'} e^{\Xi z^2} (\eta)^{-1/2}. \tag{3.31}$$

Uma vez que a energia livre média foi calculada, podemos agora determinar a expressão para o erro de treinamento médio, que mede o desempenho da rede estudante em aprender um conjunto de treinamento. Usando a equação (2.40) obtemos

$$\epsilon_t^m = \lim_{\beta \to \infty} \frac{1}{2} \frac{M + Q - 2\gamma R + \beta (Q - q)^2}{\left[1 + \beta (Q - q)\right]^2}. \tag{3.32}$$

A seguir, vamos estudar o comportamento do erro de treinamento ϵ_t^m com relação a norma quadrada Q que deve ser escolhida de maneira a minimizar a energia livre. Para isto devemos tomar o limite $\beta \to \infty$ garantindo que ϵ_t^m seja mínimo e analisar dois regimes diferentes: $q \to Q$ e $q < Q$.

Regime de erro de treinamento não nulo

O primeiro regime a ser analisado é para $q \to Q$ e $\beta \to \infty$ tal que $\beta(Q-q) = x$, onde x assume um valor finito de maneira que o erro de treinamento (3.32) reduz-se a

$$\epsilon_t^m = \lim_{\beta\to\infty} \frac{1}{2}\frac{M+Q-2\gamma R}{(1+x)^2}. \tag{3.33}$$

Neste caso, para encontrar o valor de Q que minimiza a energia livre fazemos

$$\frac{\partial(\beta f_{sr}^m)}{\partial Q} = 0 \tag{3.34}$$

e obtemos outra equação de ponto de sela:

$$\frac{\hat{q}}{2} - \hat{Q} = \frac{\alpha\beta}{2[1+\beta(Q-q)]}. \tag{3.35}$$

Utilizando esta equação para eliminar as variáveis $\hat{c}, \hat{q}, \hat{Q}, \hat{Q}^0$ e $\hat{R}$ podemos obter os valores de x, R e Q :

$$x = \frac{\Lambda_\kappa}{\alpha - \Lambda_\kappa}, \tag{3.36}$$

$$q = Q = \frac{M\Lambda_\kappa}{\alpha - \Lambda_\kappa}\left[1 + \gamma^2(\alpha - 2\Lambda_\kappa)\right] \tag{3.37}$$

e

$$R = M\gamma\Lambda_\kappa, \tag{3.38}$$

onde

$$\Lambda_\kappa = 2\int_{\lambda_\kappa}^{\infty} Dz\, z^2, \tag{3.39}$$

sendo λ_κ é a única solução de

$$\kappa = 2\int_{\lambda_\kappa}^{\infty} Dz. \tag{3.40}$$

Finalmente, substituindo esses resultados em (3.33) obtemos a seguinte expressão para o erro de treinamento médio

$$\epsilon_t^m/M = \frac{1}{2}\left(1 - \gamma^2\Lambda_\kappa\right)\left(1 - \frac{\Lambda_\kappa}{\alpha}\right). \tag{3.41}$$

Das equações (3.32) e (3.36), vemos que o erro de treinamento não nulo ocorre para $0 \leq x < \infty$ e, portanto para $\alpha > \alpha_c^m = \Lambda_\kappa$. Naturalmente esta solução de x finito só faz sentido para o caso $\alpha > \alpha_c^m$. Na figura (3.1) mostramos a dependência de α_c^m com κ.

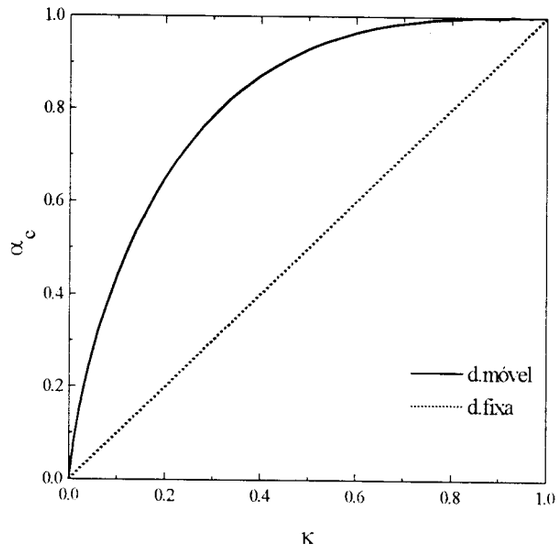


Figura 3.1: Capacidade de armazenamento α_c^m em função da conectividade κ

O primeiro resultado interessante obtido é que o valor de α_c^m não depende do parâmetro de ruído γ, ao contrário do que ocorre com o perceptron Booleano binário [2]. No caso do perceptron linear, a capacidade de armazenamento é determinada apenas pela quebra de independência linear entre as linhas e colunas da matriz $P \times N$ composta pelas variáveis aleatórias S_i^l.

Uma vez que os valores de α_c^m foram determinados podemos analisar o comportamento do erro de treinamento. Quando a rede estudante é treinada na presença de ruído, o erro de treinamento sempre aumenta à medida que o ruído aumenta, ou seja, à medida que γ diminui. Isto pode ser observado na figura (3.2) que mostra o erro de treinamento ϵ_t^m em função do tamanho do conjunto de treinamento α para $\kappa = 1$ e vários valores de γ. Note que todas as curvas iniciam no mesmo ponto, uma vez que α_c^m independe de γ. O efeito da diluição sobre o erro de treinamento é análogo ao efeito do ruído: quanto maior for a diluição, maior será o erro de treinamento. Na figura (3.3) podemos ver o comportamento do erro de treinamento em função de α quando o treinamento é feito na ausência de ruído

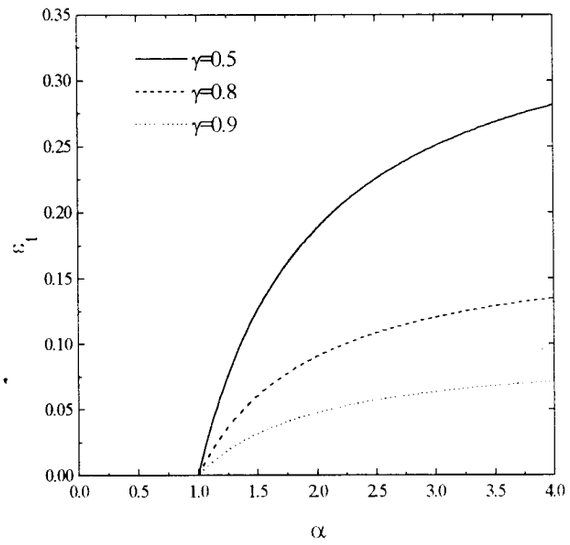


Figura 3.2: Erro de treinamento para a diluição móvel em função de α para $\kappa = 1$

Quando o treinamento for feito na presença de ruído, a figura (3.3) sofre um pequeno deslocamento, ou seja, na presença de ruído o erro de treinamento aumenta em relação ao erro de treinamento obtido com $\gamma = 1$, como pode ser visto na figura (3.4).

Uma vez que a rede estudante encontrou o vetor peso que minimiza o erro de treinamento, vamos investigar a capacidade desta rede de classificar corretamente um exemplo que não pertence ao conjunto de treinamento, ou seja, vamos calcular o erro de generalização. Neste trabalho o erro de generalização será medido para padrões de teste ruidosos. Tomando as equações (2.31) e (2.42), e substituindo os valores dos parâmetros de ordem encontrados resolvendo-se as equações de ponto de sela, temos que para $\alpha > \alpha_c^m$ o erro de generalização é dado por

$$\epsilon_g^m / M = \frac{1}{2} \frac{\alpha \left(1 - \gamma^2 \Lambda_\kappa\right)}{\alpha - \Lambda_\kappa}, \tag{3.42}$$

que, no limite de α grande, pode ser escrito como

$$\epsilon_g^m / M = \frac{1}{2} \left(1 - \gamma^2 \Lambda_\kappa\right) \left(1 + \frac{\Lambda_\kappa}{\alpha}\right) + \mathcal{O}\left(\alpha^{-2}\right). \tag{3.43}$$

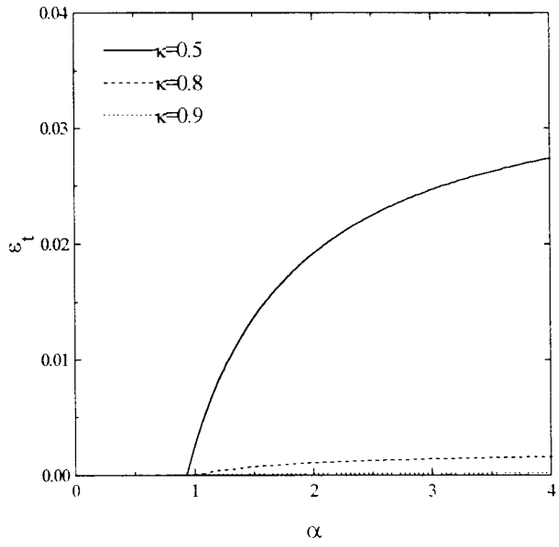


Figura 3.3: Erro de treinamento para a diluição móvel em função de α para $\gamma = 1$

Lembrando que $\alpha_c^m = \Lambda_\kappa$ podemos observar na equação (3.42) que para $\gamma \neq 1$ o erro de generalização diverge em $\alpha = \alpha_c^m$. Ainda, $\epsilon_g^m = 0$ para $\alpha \geq \alpha_c^m = 1$ quando $\gamma = 1$ e $\kappa = 1$.

Regime de erro de treinamento nulo

Dando prosseguimento ao estudo do comportamento do erro de treinamento ϵ_t^m com relação a norma quadrada Q, vamos tomar o limite $\beta \to \infty$ e analisar o regime em que $q < Q$. Neste regime podemos ver pela equação (3.32) que qualquer valor de $Q > q$ produz $\epsilon_t^m = 0$. O sistema de equações de ponto de sela para $\alpha < \alpha_c^m$ não tem solução pois a equação adicional obtida com a minimização de f_{sr}^m com relação a Q torna o sistema de equações inconsistente. Neste caso as equações de ponto de sela são resolvidas supondo que Q seja fixo *a priori* e adquira o valor mínimo possível, correspondente a solução pseudo-inversa. Resolvendo as equações de ponto de sela neste regime obtemos

$$R = M\gamma\alpha \tag{3.44}$$

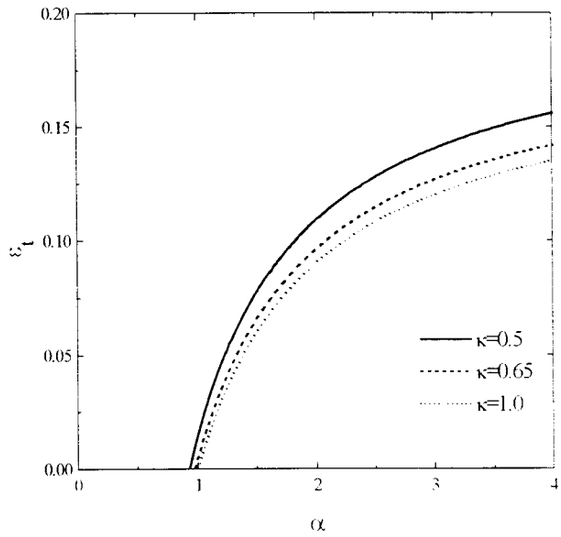


Figura 3.4: Erro de treinamento para a diluição móvel em função de α para $\gamma = 0.8$

e

$$q = \frac{M\alpha(1-\gamma^2\alpha)}{\Lambda_\kappa - \alpha}. \tag{3.45}$$

Para $\alpha < \alpha_c^m$ o menor valor que Q pode assumir é q de maneira que qualquer valor que possamos escolher de $Q > q$, onde q é dado na equação acima, fornece $\alpha_l^m = 0$. A escolha do menor valor de Q é chamada solução pseudo-inversa ou solução de menor norma. Neste regime $\alpha < \alpha_c^m$, encontrar o menor valor de Q corresponde a encontrar a solução de menor norma de um sistema de equações lineares em que o número de incógnitas é maior que o de equações [29]. Como mencionado anteriormente neste capítulo, estamos estudando o desempenho do perceptron linear treinado apenas com o algoritmo Adaline. Neste caso, embora o erro de treinamento seja zero, a energia de generalização depende linearmente de Q, equação (2.31), e portanto quanto menor for o valor de Q menor será o erro de generalização. É importante ressaltar que o algoritmo que fornece o menor erro de generalização para o perceptron linear foi desenvolvido através do método variacional acoplado ao método das réplicas por Kinouchi e Caticha [26]. Tomando o valor de R dado

na equação (3.44) e o valor de Q na solução pseudo-inversa onde $Q = Q^P$, com $Q^P = q$, obtemos o erro de generalização substituindo estes valores na equação (2.45). O resultado é

$$\epsilon_g^m/M = \frac{1}{2}\frac{\Lambda_\kappa - \gamma^2\alpha\,(2\Lambda_\kappa - \alpha)}{\Lambda_\kappa - \alpha}. \tag{3.46}$$

Na equação acima podemos ver que, para $\gamma \neq 1$, o erro de generalização diverge em $\alpha = \alpha_c^m = \Lambda_\kappa$. No caso em que $\kappa = 1$ e $\gamma = 1$ o erro de generalização se anula em $\alpha = 1$, sendo que para $\alpha < 1$ expressão de ϵ_g^m/M é uma reta com coeficiente angular igual a -1.

Vamos analisar o comportamento do erro de generalização para os dois regimes discutidos acima. A figura (3.5) mostra o erro de generalização em função do tamanho do conjunto de treinamento para $\kappa = 1$ e vários valores de γ.

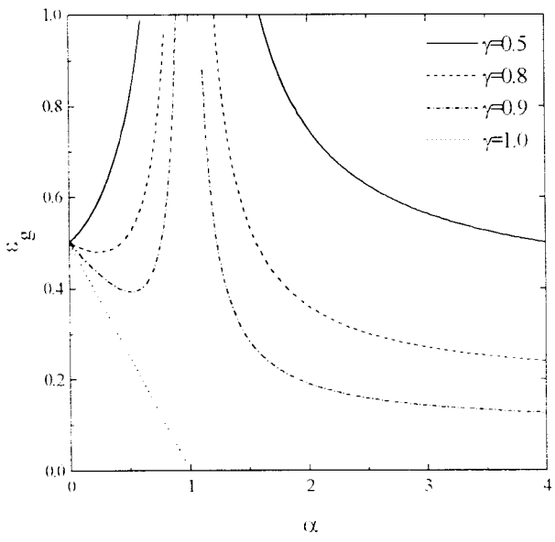


Figura 3.5: Erro de generalização para a diluição móvel em função de α para $\kappa = 1$

Quando a rede é treinada com ruído, o erro de generalização diminui a medida que γ aumenta. No caso do treinamento sem ruído ($\gamma = 1$) e sem diluição ($\kappa = 1$), observamos uma transição contínua em $\alpha = 1$ para o regime de generalização perfeita $\epsilon_g^m = 0$. Assim, para $\alpha > 1$, além do erro de treinamento, o erro de generalização também é nulo. Neste

regime a única rede que realiza perfeitamente o conjunto de treinamento é a rede mestra, ou seja, o único estado fundamental da energia de treinamento é $\{J_i = J_i^0\}$. Para $\alpha < 1$, além da rede mestra que tem $E_g\ (J_i^0) = 0$, existem outras redes com $\epsilon_t = 0$ mas $E_g\ (J_i^0) > 0$, resultando num erro de generalização médio não nulo. À medida que aumentamos o conjunto de treinamento há uma diminuição no valor de ϵ_g, pois cada vez menos redes com $E_g\ (J_i^0) > 0$ passam a contribuir para a média.

A figura (3.6) mostra o efeito da diluição numa rede que foi treinada na ausência de ruído ($\gamma = 1$). Quanto mais diluída for a rede, maior será o erro de generalização.

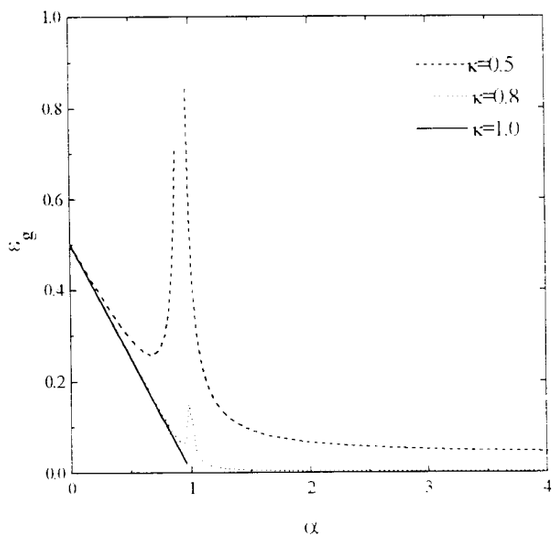


Figura 3.6: Erro de generalização para a diluição móvel em função de α para $\gamma = 1.0$

Quando a rede é treinada com ruído, o efeito da diluição é o mesmo que no caso sem ruído para $\alpha < \alpha_c^m$ e para $\alpha > 1/\gamma^2$. Note que $\alpha = 1/\gamma^2$ é o ponto onde todas as curvas se encontram. Para $\alpha_c^m < \alpha < 1/\gamma^2$, quanto maior for a diluição menor será o erro de generalização. Este comportamento é ilustrado na figura (3.7).

Uma particularidade que ocorre com o perceptron linear é o fato de tanto o ruído quanto a diluição provocarem divergências no erro de generalização quando α se aproxima

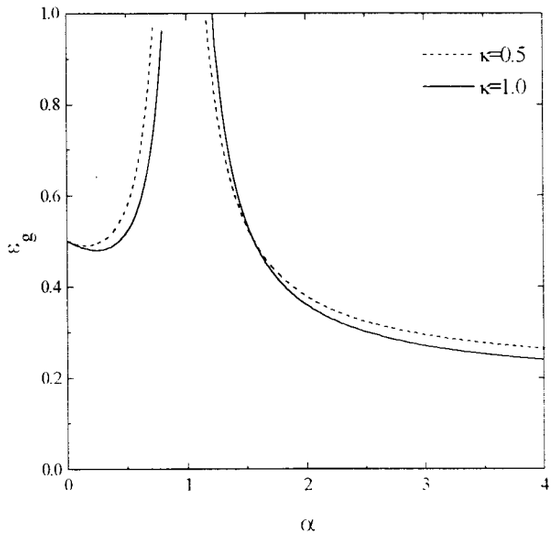


Figura 3.7: Erro de generalização para a diluição móvel em função de α para $\gamma = 0.8$

de α_c^m, pois neste ponto a norma Q diverge.

3.2.2 Teste de estabilidade da solução com simetria de réplicas

Para solucionar as equações de ponto de sela usamos a prescrição simetria de réplicas, ou seja, supomos que os valores dos parâmetros de ordem eram independentes dos seus índices de réplica. A fim de testarmos a validade desta solução devemos verificar se ela satisfaz a condição de estabilidade local dada por [1], [17]

$$\alpha\gamma_1\gamma_2 < 1 \tag{3.47}$$

onde γ_1 e γ_2 são autovalores das matrizes de derivadas segunda de G_1 e G_2 com relação a $\hat{q}_{ab}$ e q_{ab}. Seguindo a análise de Gardner [17] vamos calcular γ_1 e γ_2. O valor de γ_1 é dado por

$$\gamma_1 = \mathcal{P} - 2\mathcal{Q} + \mathcal{R} \tag{3.48}$$

onde

$$\mathcal{P} = \frac{\partial^2 G_1}{\partial q_{ab} q_{ab}}, \quad \mathcal{Q} = \frac{\partial^2 G_1}{\partial q_{ad} q_{ab}}, \quad \mathcal{R} = \frac{\partial^2 G_1}{\partial q_{cd} q_{ab}}. \tag{3.49}$$

Lembrando que

$$G_1 = \ln D$$

com

$$D = \int \prod_{a=1}^{n} \frac{dx_a}{\sqrt{2\pi}} \exp\left[\lambda\left(x_a\right)\right] \tag{3.50}$$

onde

$$\lambda\left(y_a\right) = \left\{ -\frac{1}{2} \sum_{a}^{n} x_a^2 \left[1 + \beta\left(Q + M - 2\gamma R_a\right)\right] \right. \tag{3.51}$$

$$\left. -\beta \sum_{a<b}^{n} x_a x_b \left(q_{ab} + M - 2\gamma R_a\right) \right\}.$$

podemos escrever γ_1 como

$$\gamma_1 = \frac{1}{D} \left\{ \int \prod_{a=1}^{n} \frac{dx_a}{\sqrt{2\pi}} x_a^2 x_b^2 \exp\left[\lambda\left(x_a\right)\right] - 2 \int \prod_{a=1}^{n} \frac{dx_a}{\sqrt{2\pi}} x_a^2 x_b x_d \exp\left[\lambda\left(x_a\right)\right] \right. \tag{3.52}$$

$$\left. \int \prod_{a=1}^{n} \frac{dx_a}{\sqrt{2\pi}} x_a x_b x_c x_d \exp\left[\lambda\left(x_a\right)\right] \right\}.$$

Tomando o limite $n \longrightarrow 0$ temos

$$\gamma_1 = \int_{-\infty}^{\infty} Dt \left[\left\langle x^2 \right\rangle - \left\langle x \right\rangle^2\right]^2 \tag{3.53}$$

onde $\left\langle x^k \right\rangle$ é dado por

$$\left\langle x^k \right\rangle = \frac{\int \frac{dx}{\sqrt{2\beta\pi}} x^k \exp\left\{ -\frac{[1+\beta(Q-q)]x^2}{2\beta} + itx\sqrt{q + M - 2R\gamma} \right\}}{\int \frac{dx}{\sqrt{2\beta\pi}} \exp\left\{ -\frac{[1+\beta(Q-q)]x^2}{2\beta} + itx\sqrt{q + M - 2R\gamma} \right\}}. \tag{3.54}$$

Calculando o primeiro e segundo momento obtemos

$$\gamma_1 = \frac{\beta^2}{\left[1 + \beta\left(Q - q\right)\right]^2}. \tag{3.55}$$

O valor de γ_2 é dado por

$$\gamma_2 = \mathcal{P}' - 2\mathcal{Q}' + \mathcal{R}' \tag{3.56}$$

onde

$$\mathcal{P} = \frac{\partial^2 G_2}{\partial \hat{q}_{ab}\hat{q}_{ab}}, \quad \mathcal{Q} = \frac{\partial^2 G_2}{\partial \hat{q}_{ad}\hat{q}_{ab}}, \quad \mathcal{R} = \frac{\partial^2 G_2}{\partial \hat{q}_{cd}\hat{q}_{ab}}. \tag{3.57}$$

Efetuando estas derivadas e tomando o limite $n \longrightarrow 0$ temos

$$\gamma_2 =< \int_{-\infty}^{\infty} Dt \left[\left\langle W^2 \right\rangle - \langle W \rangle^2\right]^2 >_{J^0} \tag{3.58}$$

onde $\left\langle W^k \right\rangle$ é dado por

$$\left\langle W^k \right\rangle = \frac{\int dW \sum_c W^k \exp\left\{-\hat{c}c - \hat{Q}^0 (1-c) W^2 - cW^2 \left(\frac{\dot{q}}{2} - \dot{Q}\right) + \left(\hat{R}J^0 + t\sqrt{\tilde{q}}\right) cW\right\}}{\int dW \sum_c \exp\left\{-\hat{c}c - \hat{Q}^0 (1-c) W^2 - cW^2 \left(\frac{\dot{q}}{2} - \dot{Q}\right) + \left(\hat{R}J^0 + t\sqrt{\tilde{q}}\right) cW\right\}}. \tag{3.59}$$

Calculando o primeiro e segundo momento encontramos

$$\gamma_2 = \frac{\kappa}{4\eta^2}.$$

onde η é dado pela equação (3.30). A estabilidade da solução com simetria de réplicas deve ser analisada para os limites acima e abaixo de α_c^m. Assim, para $\alpha < \alpha_c^m$ obtemos

$$\alpha\gamma_1\gamma_2 = \frac{\alpha\kappa}{2\Lambda_\kappa^2} < 1, \tag{3.60}$$

e para $\alpha > \alpha_c^m$.

$$\alpha\gamma_1\gamma_2 = \frac{\kappa}{\alpha} < 1. \tag{3.61}$$

Desses resultados podemos concluir que a solução com simetria de réplicas é localmente estável para os dois regimes analisados.

3.2.3 Distribuição dos pesos

O cálculo da distribuição dos pesos tem como finalidade a obtenção da densidade de probabilidade de que uma certa conexão, J_i , assuma o valor $J_i = J$ uma vez que a rede encontrou os pesos que implementam uma determinada tarefa. A densidade de probabilidade que um dado peso assuma o valor $J_i = J$ é dada por

$$\begin{aligned} P(J_i = J) &= \langle\langle\langle \delta(J_i - J)\rangle\rangle\rangle \\ &= \frac{1}{N}\left\langle\left\langle\left\langle \sum_i \delta(J_i - J)\right\rangle\right\rangle\right\rangle. \end{aligned} \tag{3.62}$$

uma vez que esta grandeza é claramente independente da particular conexão escolhida. Para efetuarmos esse cáculo vamos usar um campo auxiliar h, a fim de facilitar a média sobre os exemplos de treinamento como foi feito em seções anteriores. Este cáculo será efetuado apenas para o caso da diluição móvel.

Para calcularmos a expressão acima com o auxílio do campo h, definimos a seguinte função de partição

$$Z = \sum_{\{\mathbf{c}\}} \delta_{Kr}\left(\sum_i c_i, \kappa N\right) \int_{-\infty}^{\infty} d\mu\left(\mathbf{W}\right) \exp\left[-\beta E_\kappa\left(\{W_i\}, \{c_i\}\right) - \beta h \sum_i \delta\left(c_i, 1\right) \delta\left(W_i - J\right)\right] \tag{3.63}$$

onde δ_{Kr} é a delta de Kronecker e $d\mu\left(\mathbf{W}\right)$ é dado pela equação (3.12). Assim, a distribuição dos pesos é dada por

$$\langle\langle\langle \delta\left(J_i - J\right)\rangle\rangle\rangle = -\frac{1}{\beta N}\frac{\partial}{\partial h}\langle\langle \ln Z\rangle\rangle \mid_{h=0} .$$

Novamente usaremos o método das réplicas para calcular $\langle\langle \ln Z\rangle\rangle$. O n -ésimo momento de Z é dado por

$$\begin{aligned}\langle\langle Z^n\rangle\rangle \;=\; & \left\langle\left\langle \prod_a \int_{-\infty}^{\infty} \prod_i dW_i^a \delta\left(QN - \sum_i c_i^a \left(W_i^a\right)^2\right) \delta\left(QN - \sum_i \left(1 - c_i^a\right)\left(W_i^a\right)^2\right)\right.\right. \\ & \sum_{\{c_i^a\}} \delta_{Kr}\left(\sum_{i,a} c_i^a, \kappa N\right) \exp -\frac{\beta}{2}\sum_{l,a}^{l'}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i^0 \xi_i^l - \frac{1}{\sqrt{N}}\sum_{i=1}^{N} c_i W_i S_i^l\right)^2 \\ & \left.\left. \exp\left[-\beta h \sum_i \delta\left(c_i, 1\right)\delta\left(W_i - J\right)\right]\right\rangle\right\rangle . \end{aligned} \tag{3.64}$$

Mediando sobre os exemplos e efetuando a integração por ponto de sela obtemos

$$P\left(J_i = J\right) = \int \frac{Dz}{\sqrt{\pi}} \frac{\exp\left(-\hat{c} + \eta J^2 + zJ\sqrt{\hat{q} + M\hat{R}^2}\right)}{1 + \eta^{-1/2}\exp\left(-\hat{c} + \Xi z^2\right)} \tag{3.65}$$

onde η e Ξ são dados pelas equações (3.30) e (3.29), respectivamente. Vamos analisar o comportamento desta distribuição para os dois regimes de interesse discutidos anteriormente: o regime de erro de treinamento não nulo quando $q \to Q$ e o regime de erro de treinamento nulo quando $Q = Q^{l'}$. Para o regime de erro de treinamento não nulo,

$\alpha > \alpha_c^m$, encontramos

$$P(J_i = J) = \begin{cases} 0 & \text{se} \quad |J| < \lambda_\kappa \sqrt{\frac{M(1+\gamma^2\alpha-2\gamma^2\Lambda_\kappa)}{\alpha-\Lambda_\kappa}} \\ \frac{\sqrt{\alpha-\Lambda_\kappa}}{\sqrt{2\pi M(1+\gamma^2\alpha-2\gamma^2\Lambda_\kappa)}} \exp\left[-\frac{J^2(\alpha-\Lambda_\kappa)}{2M(1+\gamma^2\alpha-2\gamma^2\Lambda_\kappa)}\right] & \text{caso contrário} \end{cases} \tag{3.66}$$

onde Λ_κ e λ_κ são dados pelas equações (3.39) e (3.40). No caso em que não há diluição e na ausência de ruído ($\gamma = 1$) a distribuição dos pesos não depende do tamanho do conjunto de treinamento α e é dada por

$$P(J_i = J) = \frac{1}{\sqrt{2\pi M}} e^{-J^2/2M}.$$

Esta distribuição está ilustrada na figura (3.8). Como as distribuições são simétricas em relação a origem $J = 0$, nesta e nas próximas figuras vamos apresentar apenas a região de J positivo.

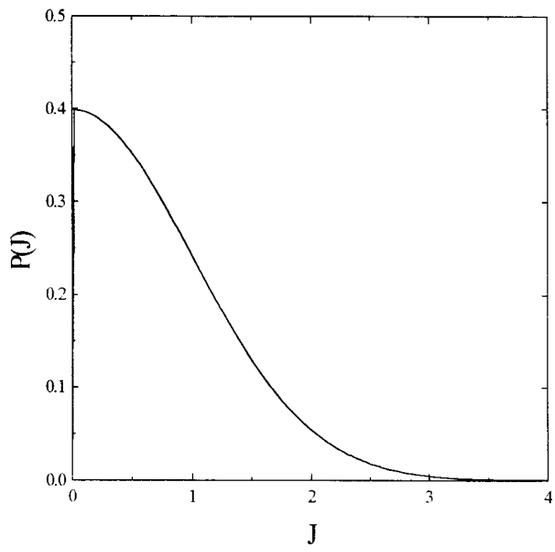


Figura 3.8: Distribuição dos pesos para $\gamma = 1$. $\kappa = 1$. e $\alpha = 3.0$

No caso em que há diluição, a distribuição dos pesos varia com α e. como pode ser visto na figura (3.9). esta distribuição é uma Gaussiana cuja seção central foi removida.

No caso de treinamento com ruído e diluição, quanto maior for α, menor será a seção central removida da Gaussiana como pode ser visto na figura (3.10).

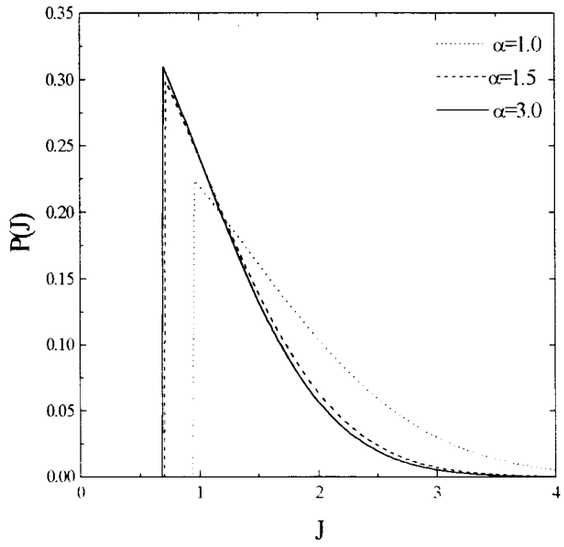


Figura 3.9: Distribuição dos pesos para $\gamma = 1$, e $\kappa = 0.5$

No regime de erro de treinamento nulo, $\alpha \leq \alpha_c^m$ a distribuição dos pesos é dada por

$$P(J_i = J) = \begin{cases} 0 & \text{se} \quad |J| < \lambda_\kappa \sqrt{\frac{M\alpha(1-\gamma^2\alpha)}{\Lambda_\kappa(\Lambda_\kappa - \alpha)}} \\ \frac{\sqrt{\Lambda_\kappa(\Lambda_\kappa-\alpha)}}{\sqrt{2\pi M\alpha(1-\gamma^2\alpha)}} \exp\left[-\frac{J^2\Lambda_\kappa(\Lambda_\kappa-\alpha)}{2M\alpha(1-\gamma^2\alpha)}\right] & \text{caso contrário.} \end{cases} \tag{3.67}$$

No caso de $\kappa = 1$ e $\gamma = 1$ esta distribuição reduz-se a

$$P(J_i = J) = \frac{1}{\sqrt{2\pi M\alpha}} e^{-J^2/2M\alpha}.$$

À medida que α se aproxima de α_c^m o valor dos pesos aumenta. A dependência da distribuição dos pesos com o ruído pode ser vista na figura (3.11).

No caso em que há diluição, a distribuição dos pesos tem o mesmo comportamento como pode ser visto na figura (3.12).

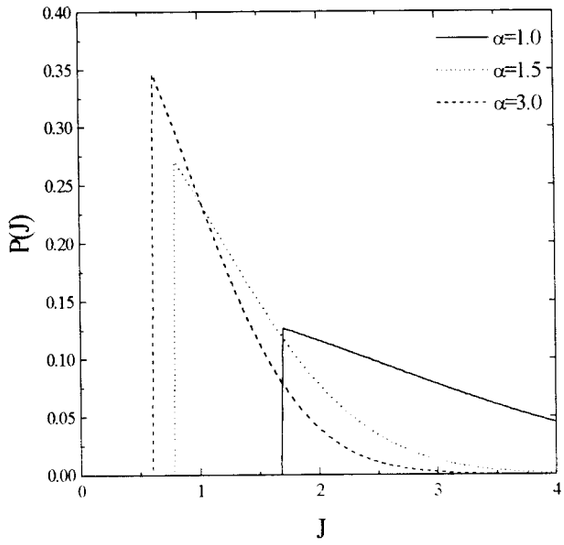


Figura 3.10: Distribuição dos pesos para $\gamma = 0.8$ e $\kappa = 0.5$

Quando o treinamento é feito com ruído o comportamento da distribuição é oposto ao que ocorre no regime de erro de treinamento não nulo: quanto maior for α, maior será a lacuna da Gaussiana como pode ser visto na figura (3.13).Para um determinado valor de α, quanto maior for a diluição maior será a lacuna da Gaussiana. A figura (3.14) mostra o comportamento da distribuição dos pesos para $\alpha = 0.6$ e vários valores de κ.

Em resumo, nesta seção calculamos a distribuição dos pesos remanescentes quando a rede estudante é sujeita a diluição móvel, na qual o corte dos pesos é feito de maneira a minimizar o erro de treinamento. Como será visto no final deste capítulo, este tipo de diluição fornece o menor erro de treinamento. Na figura (3.14) vemos que neste processo de diluição os pesos de menor intensidade é que são eliminados, ou seja, os pesos que não foram cortados assumem valores a partir de um certo limiar, que depende do grau de diluição. Assim, no caso em que a rede é treinada com toda a sua capacidade e uma fração dos pesos é então cortada, é razoável esperar que o corte dos pesos menores forneça os melhores resultados.

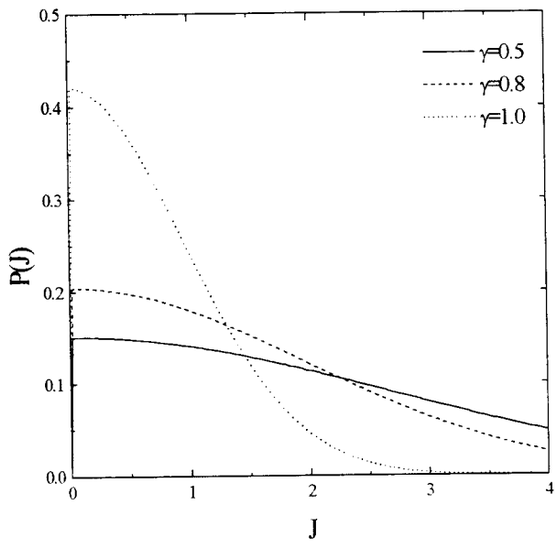


Figura 3.11: Distribuição dos pesos para $\kappa = 1$ e $\alpha = 0.9$

3.2.4 Diluição Fixa

Neste caso $(1 - \kappa)\, N$ pesos escolhidos aleatoriamente são eliminados. Como os exemplos são aleatórios e não há nada que diferencie um determinado sítio i dos outros, podemos sem perda de generalidade eliminar os pesos J_i com $i = N\kappa + 1, \ldots, N$. Assim, não é necessário utilizar explicitamente o vínculo da diluição, bastando efetuar os somatórios sobre os sítios $i = 1, \ldots, N\kappa$. Portanto a energia de treinamento é escrita como

$$E_\kappa\left(\{J_i\}\right) = \frac{1}{2}\sum_{i=1}^{P}\left(t^l - \frac{1}{\sqrt{N}}\sum_{i=1}^{N\kappa} J_i S_i^l\right)^2 . \tag{3.68}$$

A rede estudante deve encontrar as configurações de J_i que minimizam esta energia e satisfazem o vínculo (3.9), que passa a ser escrito como

$$Q = \frac{1}{N}\sum_{i=1}^{N\kappa} J_i^2 . \tag{3.69}$$

O processo de obtenção do erro de treinamento é exatamente o mesmo empregado na seção anterior. Para obtermos a expressão do erro de treinamento dado pela equação

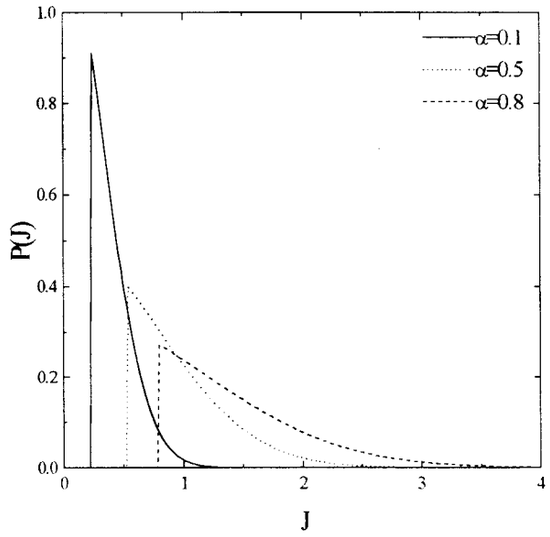


Figura 3.12: Distribuição dos pesos para $\gamma = 1.0$ e $\kappa = 0.5$

(3.32) devemos calcular primeiramente a energia livre (2.39) usando o formalismo das réplicas. No caso da diluição fixa a função de partição torna-se

$$Z = \int_{-\infty}^{\infty} d\mu\left(\mathbf{W}\right) \exp\left[-\beta E\left(\{J_i\}\right)\right] \tag{3.70}$$

onde

$$d\mu\left(\mathbf{W}\right) = \prod_{i}^{N\kappa} dJ_i \delta\left(NQ - \sum_{i}^{N\kappa} J_i^2\right).$$

O cálculo da densidade de energia livre é feito de forma análoga ao caso da diluição móvel (seção A.2). Assim, temos

$$\begin{aligned} \beta f^f &= \lim_{n\to 0} \operatorname{extr} \frac{1}{n} \left\{ G_0\left(q_{ab}, \hat{q}_{ab}, \hat{Q}_a, R_a, \hat{R}_a\right) \right. \\ &\left. + \alpha G_1\left(q_{ab}, R_a\right) + \kappa G_2\left(\hat{q}_{ab}, \hat{Q}_a, \hat{R}_a\right) \right\} \end{aligned} \tag{3.71}$$

onde

$$G_0 = -\sum_{a<b}^{n} q_{ab}\hat{q}_{ab} - \sum_{a}^{n}\left(Q\hat{Q}_a + R_a\hat{R}_a\right).$$

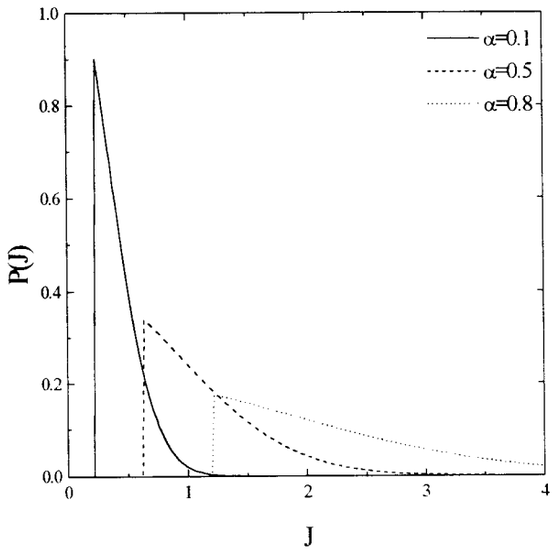


Figura 3.13: Distribuição dos pesos para $\gamma = 0.8$ e $\kappa = 0.5$

$$G_2 = \ln \int \prod_{a=1}^{n} dJ^a \exp\left\{ \sum_a^n \hat{Q}_a \left(J^a\right)^2 + \sum_a^n \hat{R}_a J^a J^0 + \sum_{a<b}^n \hat{q}_{ab} J^a J^b \right\} \tag{3.72}$$

e G_1 é o mesmo que na diluição móvel, dado na equação (3.15). O extremo na equação (3.71) é tomado sobre todos os parâmetros de ponto de sela $\left(\hat{q}_{ab}, \hat{Q}_a, \hat{R}_a, q_{ab}, R_a\right)$. Os parâmetros de ordem físicos são

$$q_{ab} = \frac{1}{N} \sum_{i=1}^{N\kappa} J_i^a J_i^b \qquad a < b \tag{3.73}$$

e

$$R_a = \frac{1}{N} \sum_{i=1}^{N\kappa} J_i^a J_i^0, \tag{3.74}$$

e têm os mesmos significados dos parâmetros dados em (3.17) e (3.18).

Na seção C.2 calculamos G_2 usando a prescrição simetria de réplicas. No limite $n \to 0$, obtemos a seguinte expressão da energia livre média

$$-\beta f_{sr}^f = \frac{1}{2}\hat{q}q - Q\hat{Q} - R\hat{R} + \frac{\kappa}{2}\ln\pi - \frac{\alpha}{2}\ln\left[1 + \beta\left(Q - q\right)\right]$$

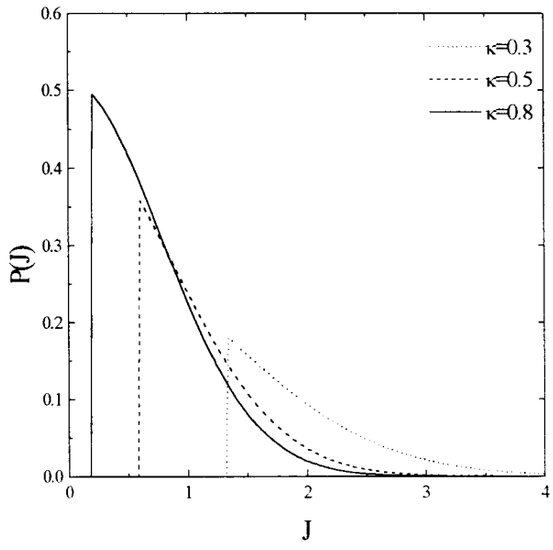


Figura 3.14: Distribuição dos pesos para $\gamma = 1.0$ e $\alpha = 0.6$

$$-\frac{\alpha\beta}{2}\frac{q+M-2\gamma R}{1+\beta(Q-q)} - \frac{\kappa}{2}\ln\left(\frac{\hat{q}}{2}-\hat{Q}\right) + \frac{\kappa}{4}\frac{\hat{q}+M\hat{R}^2}{\left(\frac{\hat{q}}{2}-\hat{Q}\right)}. \tag{3.75}$$

Os parâmetros de ordem na prescrição simetria de réplicas $\left(\hat{q}, q, \hat{Q}, \hat{R}, R\right)$ são dados pelas equações de ponto de sela. Estas equações são obtidas igualando-se a zero as derivadas de $-\beta f_{sr}^{q}$ com relação a estes parâmetros:

$$\hat{q} = \frac{\alpha\beta^2(Q+M-2\gamma R)}{[1+\beta(Q-q)]}, \tag{3.76}$$

$$q = \frac{\kappa}{4}\frac{\hat{q}+M\hat{R}^2}{\left(\frac{\hat{q}}{2}-\hat{Q}\right)^2}, \tag{3.77}$$

$$Q = \frac{\kappa\left[2\left(\hat{q}-\hat{Q}\right)+M\hat{R}^2\right]}{4\left(\frac{\hat{q}}{2}-\hat{Q}\right)^2}, \tag{3.78}$$

$$\hat{R} = \frac{\alpha\beta\gamma}{1+\beta(Q-q)}, \tag{3.79}$$

e

$$R = \frac{\kappa \hat{R}}{2\left(\frac{q}{2} - \hat{Q}\right)}. \tag{3.80}$$

A expressão para o erro de treinamento médio é a mesma encontrada no caso da diluição móvel, equação (3.32). Seguindo o estudo feito com o erro de treinamento naquele tipo de diluição vamos tomar o limite $\beta \to \infty$ e analisar dois regimes diferentes: $q \to Q$ e $q < Q$.

Regime de erro de treinamento não nulo

Neste regime tomamos $q \to Q$ e $\beta \to \infty$ de maneira que $\beta(Q - q) = x$ seja finito. Tomando a derivada de βf_{sr}^{f} com relação a Q, obtemos a equação de ponto de sela

$$\frac{\hat{q}}{2} - \hat{Q} = \frac{\alpha\beta}{2\left[1 + \beta(Q - q)\right]}. \tag{3.81}$$

Utilizando esta equação para eliminar as variáveis $\hat{q}, \hat{Q}$, e $\hat{R}$ calculamos os valores de x, R, e Q :

$$x = \frac{\kappa}{\alpha - \kappa}, \tag{3.82}$$

$$R = M\gamma\kappa, \tag{3.83}$$

e

$$q = Q = \frac{M\kappa}{\alpha - \kappa}\left(1 + \gamma^2(\alpha - 2\kappa)\right). \tag{3.84}$$

Substituindo esses valores na equação

$$\epsilon_t^f = \lim_{\beta \to \infty} \frac{1}{2} \frac{M + Q - 2\gamma R}{(1 + x)^2} \tag{3.85}$$

obtemos o erro de treinamento médio

$$\epsilon_t^f / M = \frac{1}{2}\left(1 - \gamma^2\kappa\right)\left(1 - \frac{\kappa}{\alpha}\right). \tag{3.86}$$

No caso da diluição fixa, o erro de treinamento diferente de zero ocorre para $\alpha > \alpha_c^f = \kappa$ como pode ser visto na equação acima. Neste caso novamente α_c^f não depende do parâmetro de ruído γ, pois α_c^f é determinado apenas pela quebra de independencia linear entre as linhas e colunas da matriz $P \times N$ composta pelas variáveis aleatórias S_i^l.

Na figura (3.1) podemos ver que a capacidade de armazenamento na diluição fixa é bem menor que na diluição móvel.

A seguir vamos analisar o comportamento do erro de treinamento. Na figura (3.15) podemos ver que o efeito da diluição fixa sobre o erro de treinamento é análogo ao efeito da diluição móvel, quanto maior for a diluição maior será o erro de treinamento. Na presença de ruído o efeito da diluição continua sendo o mesmo, como pode ser visto na figura (3.16). que mostra ϵ_t^f em função de α para $\gamma = 0.8$ e vários valores de κ.

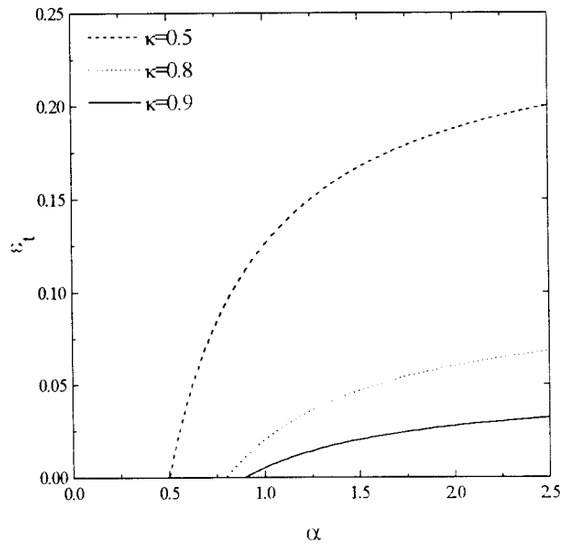


Figura 3.15: Erro de treinamento para a diluição fixa em função de α para $\gamma = 1$

Terminado o cálculo do erro de treinamento o próximo passo é calcular o erro de generalização. Tomando a equação (2.45) e substituindo os valores dos parâmetros de ordem encontrados nas equações de ponto de sela, temos que para $\alpha > \alpha_c^f$ o erro de generalização é dado por

$$\epsilon_g^f / M = \frac{1}{2} \frac{\alpha\,(1 - \gamma^2 \kappa)}{\alpha - \kappa} \tag{3.87}$$

Na equação acima vemos que quando $\kappa \neq 1$ o erro de generalização diverge em $\alpha = \alpha_c^f$

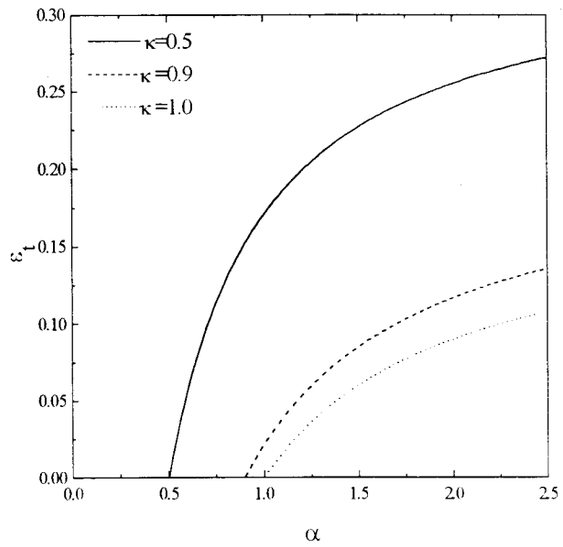


Figura 3.16: Erro de treinamento para a diluição fixa em função de α para $\gamma = 0.8$

sendo sempre nulo para $\alpha \geq \alpha_c^f = 1$ quando $\gamma = 1$ e $\kappa = 1$. Este resultadoé o mesmo encontrado na diluição móvel uma vez que para $\kappa = 1$ não há diluição.

Regime de erro de treinamento nulo

Como no caso da diluição móvel, para $\alpha \leq \kappa$ o erro de treinamento é nulo embora o erro de generalização não o seja. Novamente, o sistema de equações de ponto de sela para $\alpha \leq \kappa$ não pode ser resolvido, uma vez que a equação adicional obtida com a minimização de f_{sr}^f com relação a Q só é válida quando $q \to Q$. Para calcularmos o erro de generalização ϵ_g^f vamos encontrar os valores de q e R supondo Q fixo *a priori* . Assim, resolvendo as equações de ponto de sela com $Q = Q^P = q$ temos

$$R = M\alpha\gamma \tag{3.88}$$

e

$$q = \frac{M\alpha\left(1-\gamma^2\alpha\right)}{\kappa - \alpha}. \tag{3.89}$$

Substituindo os valores de q e R na equação (2.45) encontramos

$$\epsilon_g^f/M = \frac{1}{2}\frac{\kappa - \gamma^2\alpha(2\kappa - \alpha)}{\kappa - \alpha}. \tag{3.90}$$

Podemos observar que para $\kappa \neq 1$ o erro de generalização diverge em $\alpha = \alpha_c^f$ e no caso em que $\gamma = 1$ e $\kappa = 1$ recuperamos o regime de generalização perfeita já obtido com a diluição móvel.

O efeito da diluição fixa sobre o erro de generalização é análogo ao efeito da diluição móvel, quanto maior for a diluição maior será o erro de generalização, como pode ser visto na figura (3.17) que mostra o erro de generalização em função de α para $\gamma = 1$.

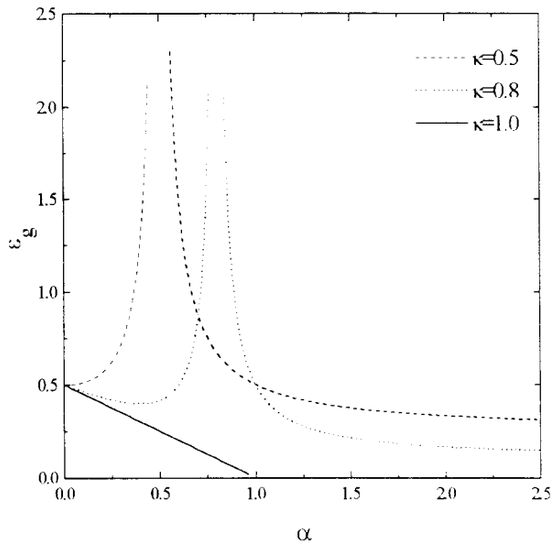


Figura 3.17: Erro de generalização para a diluição fixa em função de α para $\gamma = 1.0$

Quando a rede é treinada na presença de ruído, o efeito da diluição é análogo ao caso acima porém é mais pronunciado como pode ser visto na figura (3.18).

Podemos notar que existe uma similaridade muito grande entre as equações que descrevem a diluição fixa e as equações que descrevem a diluição móvel. Em particular, para obter as equações de ponto de sela e os erros médios de treinamento e generalização basta

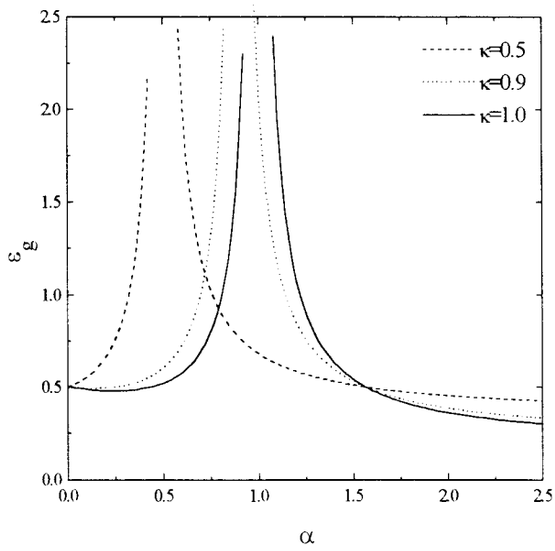


Figura 3.18: Erro de generalização para a diluição fixa em função de α para $\gamma = 0.8$

substituir Λ_κ na diluição móvel por κ. Essa similaridade entre as equações que descrevem as duas diluições é uma característica do perceptron linear, uma vez que nada similar foi observado no estudo do perceptron Booleano binário sujeito a estes dois tipos de diluição [2].

3.2.5 Teste da estabilidade da solução com simetria de réplicas

Como G_1 é o mesmo para ambas as diluições vamos calcular apenas γ_2. Neste caso temos

$$\gamma_2 =< \int_{-\infty}^{\infty} Dt \left[\left\langle W^2 \right\rangle - \langle W \rangle^2\right]^2 >_{J^0} \tag{3.91}$$

onde $\left\langle W^k \right\rangle$ é dado por

$$\left\langle W^k \right\rangle = \frac{\int dW \exp\left\{-\left(\frac{\hat{q}}{2} - \hat{Q}\right) W^2 + \left(\hat{R} J^0 + t\sqrt{\hat{q}}\right) W\right\} W^k}{\int dW \exp\left\{-\left(\frac{\hat{q}}{2} - \hat{Q}\right) W^2 + \left(\hat{R} J^0 + t\sqrt{\hat{q}}\right) W\right\}}. \tag{3.92}$$

Calculando o primeiro e segundo momentos obtemos

$$\gamma_2 = \frac{\kappa}{4\left(\frac{\hat{q}}{2} - \hat{Q}\right)^2}.$$

A estabilidade da solução com simetria de réplicas deve ser analisada para os casos de α acima e abaixo de α_c. Assim, para $\alpha \leq \alpha_c = \kappa$ encontramos

$$\alpha\gamma_1\gamma_2 = \frac{\alpha}{\kappa} < 1, \tag{3.93}$$

e para $\alpha > \alpha_c = \kappa$,

$$\alpha\gamma_1\gamma_2 = \frac{\kappa}{\alpha} < 1. \tag{3.94}$$

De onde concluímos que a solução com simetria de réplicas é localmente estável para este tipo de diluição.

3.3 Diluição após o aprendizado

Uma vez que a rede encontrou as conexões corretas utilizando toda a sua capacidade sináptica, vamos fazer com que parte de suas conexões se anule, de modo que a fração dos pesos remanescentes seja igual a κ.

3.3.1 Corte dos pesos menores

No caso do corte dos pesos menores, queremos que os pesos anulados sejam menores que um certo limiar, ou seja, $|J_i| < \omega$, onde ω é escolhido de maneira a garantir que a fração dos pesos remanescentes iguale a κ.

Para medir o desempenho da rede estudante na realização dos P padrões, definimos o erro de treinamento médio:

$$\epsilon_t^p = \frac{1}{\alpha N} \lim_{\beta\to\infty} \langle\langle E\left(\{J_i\Theta\left(|J_i| - \omega\right)\}\right)\rangle_T\rangle \tag{3.95}$$

com E dada por

$$E\left(\{J_i\Theta\left(|J_i| - \omega\right)\}\right) = \frac{1}{2}\sum_{l=1}^{P}\left(t^l - \frac{1}{\sqrt{N}}\sum_{i=1}^{N} S_i^l J_i\Theta\left(|J_i| - \omega\right)\right)^2 \tag{3.96}$$

e

$$t^l = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^l J_i^0. \tag{3.97}$$

A maior dificuldade encontrada na obtenção de ϵ_t^p é a realização da média sobre os padrões. A fim de facilitarmos este cálculo vamos introduzir a seguinte função de partição

$$Z = \mathrm{Tr} \exp \left\{ -\beta E_{\kappa=1} \left(\{J_i\} \right) + \frac{\beta h}{2} \sum_l \left[t^l - \sum_i \frac{S_i^l J_i \Theta \left(|J_i| - \omega \right)}{\sqrt{N}} \right]^2 \right\} \tag{3.98}$$

onde

$$E_{\kappa=1} \left(\{J_i\} \right) = \frac{1}{2} \sum_l \left[t^l - \sum_i \frac{S_i^l J_i}{\sqrt{N}} \right]^2 . \tag{3.99}$$

Desta maneira para calcularmos o erro de treinamento médio fazemos

$$\epsilon_t^p = -\frac{1}{\beta} \frac{d}{dh} \left\langle \left\langle \ln Z \right\rangle \right\rangle \mid_{h=0} .$$

O cálculo de $\langle\langle \ln Z \rangle\rangle$ é feito através da aplicação do método das réplicas. Assim o n-ésimo momento de Z é dado por

$$\left\langle \left\langle Z^n \right\rangle \right\rangle = \left\langle \left\langle \prod_a \mathrm{Tr} \exp \left\{ -\beta E_{\kappa=1} \left(\{J_i^a\} \right) + \frac{\beta h}{2} \sum_l \left[t^l - \sum_i \frac{S_i^l J_i^a \Theta \left(|J_i^a| - \omega \right)}{\sqrt{N}} \right]^2 \right\} \right\rangle \right\rangle . \tag{3.100}$$

O cálculo detalhado deste momento é apresentado na seção (A.3). Neste caso, é necessária a introdução de novos parâmetros de ordem, além dos já utilizados. São eles:

$$p_{ab} = \frac{1}{N} \sum_{i=1}^{N} J_i^a \Theta \left(|J_i^a| - \omega \right) J_i^b \Theta \left(\left| J_i^b \right| - \omega \right) \tag{3.101}$$

$$P_a = \frac{1}{N} \sum_{i=1}^{N} \left(J_i^a \right)^2 \Theta \left(|J_i| - \omega \right) , \tag{3.102}$$

$$r_{ab} = \frac{1}{N} \sum_{i=1}^{N} J_i^a J_i^b \Theta \left(\left| J_i^b \right| - \omega \right) , \tag{3.103}$$

e

$$S_a = \frac{1}{N} \sum_{i=1}^{N} J_i^0 J_i^a \Theta \left(|J_i^a| - \omega \right) . \tag{3.104}$$

Como mostrado na seção (A.3) apenas G_1 depende de h de modo que podemos escrever

$$\epsilon_t^p = -\frac{1}{\beta} \frac{d}{dh} G_1 \mid_{h=0} .$$

O cálculo explícito de G_1 usando a prescrição simetria das réplicas é apresentado na seção (B.3). Assim, o resultado para o erro de treinamento médio é

$$\epsilon_t^p = \frac{P+M}{2} - \gamma S + \frac{\beta^2 (P-r)^2}{2\left[1+\beta(Q-q)\right]^2}\left[q+M-2\gamma R\right] -$$

$$-\frac{\beta(P-r)}{2\left[1+\beta(Q-q)\right]}\left[(P+r)+2M-2\gamma(R+S)\right] \tag{3.105}$$

Lembramos ao leitor que aqui P e S são parâmetros de ordem simétricos com respeito as réplicas e não devem ser confundidos com o número total de exemplos ($P = \alpha N$) e as componentes dos padrões de entrada corrompidas por ruído. A equação que relaciona ω e κ é

$$\kappa = \left\langle\left\langle\left\langle \frac{1}{N}\sum_i^N \Theta\left(|J_i| - \omega\right)\right\rangle_T\right\rangle\right\rangle \tag{3.106}$$

que fornece

$$\kappa = 2\int_{\omega/\sqrt{Q_1}}^{\infty} Dz \tag{3.107}$$

onde Q_1 é o valor de Q para $\kappa = 1$.

Para calcular ϵ_t^p devemos distinguir os dois regimes mencionados nas diluições já discutidas que ocorrem para $\alpha < 1$ e $\alpha > 1$. As equações de ponto de sela para α qualquer são

$$Q = \frac{2\left(\hat{Q}-\hat{q}\right) + M\hat{R}^2}{4\left(\hat{Q}-\frac{\hat{q}}{2}\right)^2}, \tag{3.108}$$

$$q = \frac{M\hat{R}^2 - \hat{q}}{4\left(\hat{Q}-\frac{\hat{q}}{2}\right)^2}, \tag{3.109}$$

$$\hat{q} = -\frac{\beta^2\alpha\left(q+M-2R\gamma\right)}{\left[1+\beta(Q-q)\right]^2}, \tag{3.110}$$

$$\hat{R} = \frac{-\alpha\beta\gamma}{1+\beta(Q-q)}, \tag{3.111}$$

$$R = \frac{-M\hat{R}}{2\left(\hat{Q}-\frac{\hat{q}}{2}\right)}. \tag{3.112}$$

$$P = Q\Lambda_\kappa, \qquad r = q\Lambda_\kappa, \qquad S = R\Lambda_\kappa \tag{3.113}$$

e

$$\hat{r} = \hat{P} = \hat{p} = \hat{S} = 0. \tag{3.114}$$

Para $\alpha < 1$ ($\beta \to \infty$ e $q < Q$) as equações de ponto de sela reduzem-se a

$$R = M\gamma\alpha \tag{3.115}$$

e

$$q = Q^P = \frac{M\alpha\left(1-\gamma^2\alpha\right)}{1-\alpha}. \tag{3.116}$$

O erro de treinamento médio é então dado por

$$\epsilon_t^p/M = \frac{1-\Lambda_\kappa}{2}\left(1-\Lambda_\kappa+\frac{\alpha\Lambda_\kappa\left(1-\gamma^2\alpha\right)}{1-\alpha}\right). \tag{3.117}$$

Para $\alpha > 1$ ($\beta \to \infty$ e $q \to Q$) temos mais uma equação de ponto sela

$$\frac{\hat{q}}{2} - \hat{Q} = \frac{\alpha\beta}{2\left[1+\beta\left(Q-q\right)\right]}, \tag{3.118}$$

que leva a

$$R = M\gamma \tag{3.119}$$

e

$$q = Q = \frac{M}{\alpha-1}\left(1+\gamma^2\left(\alpha-2\right)\right). \tag{3.120}$$

Substituindo estes valores na equação (3.105) obtemos

$$\epsilon_t^p/M = \frac{\alpha^2\left(1-\gamma^2\Lambda_\kappa\right)-\alpha\left(1+\Lambda_\kappa\left(1-2\gamma^2\right)\right)+\Lambda_\kappa\left(2-\Lambda_\kappa\right)\left(1-\gamma^2\right)}{2\alpha\left(\alpha-1\right)}, \tag{3.121}$$

que, no limite de α grande, pode ser re-escrita como

$$\epsilon_t^p/M = \frac{1-\gamma^2\Lambda_\kappa}{2} - \frac{\Lambda_\kappa\left(1-\gamma^2\right)}{2\alpha} + \mathcal{O}\left(\alpha^{-2}\right). \tag{3.122}$$

Além das curvas resultantes dos cálculos analíticos (indicadas nos gráficos por traços contínuos), a fim de ilustrarmos o comportamento dos erros de treinamento e de genera lização mostramos também as curvas obtidas das simulações. Para realizar as simulações usamos a regra pseudo-inversa (2.7) para $\alpha < 1$ e a regra delta (2.6) para $\alpha > 1$. Em ambos os casos foram feitas 100 médias e o número de neurônios usado foi $N = 100$. Na figura (3.19) podemos ver o erro de treinamento em função de α para $\gamma = 0.8$ e $\kappa = 0.5$.

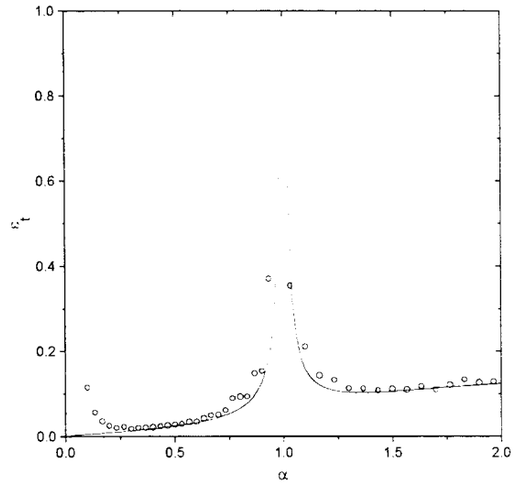


Figura 3.19: Erro de treinamento em função de α para $\gamma = 0.8$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

Concentremo-nos agora no cálculo do erro de generalização. O erro de generalização médio é definido como

$$\epsilon_g^p = \lim_{\beta \to \infty} \left\langle \left\langle \langle E_g \left[\{ J_i \Theta \left(|J_i| - \omega \right) \} \right] \rangle_T \right\rangle \right\rangle \tag{3.123}$$

onde a energia de generalização E_g, calculada como na equação (2.19), é dada por

$$\begin{aligned} E_g &= \frac{1}{2} \left(M + P - 2\gamma S \right) \\ &= \frac{1}{2} \left(M + \Lambda_\kappa - 2\gamma R \Lambda_\kappa \right). \end{aligned} \tag{3.124}$$

O erro de generalização médio é calculado pela mesma maneira descrita na seção (2.3) e deve ser analisado nos casos $\alpha \leq 1$ e $\alpha > 1$. No caso $\alpha \leq 1$ encontramos

$$\epsilon_g^p / M = \frac{\alpha^2 \gamma^2 \Lambda_\kappa - \alpha \left(1 - \Lambda_\kappa \left(1 - 2\gamma^2 \right) \right) + 1}{2 \left(1 - \alpha \right)}. \tag{3.125}$$

e no caso que $\alpha > 1$,

$$\epsilon_g^p / M = \frac{1}{2} + \frac{\Lambda_\kappa \left(1 - \gamma^2 \alpha \right)}{2 \left(\alpha - 1 \right)}. \tag{3.126}$$

Para α grande esta equação pode ser escrita como

$$\epsilon_g^p/M = \frac{1-\gamma^2\Lambda_\kappa}{2} + \frac{\Lambda_\kappa(1-\gamma^2)}{2\alpha} + \mathcal{O}\left(\alpha^{-2}\right). \tag{3.127}$$

A figura (3.20) mostra o comportamento do erro de generalização quando a rede é treinada na ausência de ruído e de diluição, ou seja, quando $\gamma = 1$ e $\kappa = 1$. Na figura (3.21)

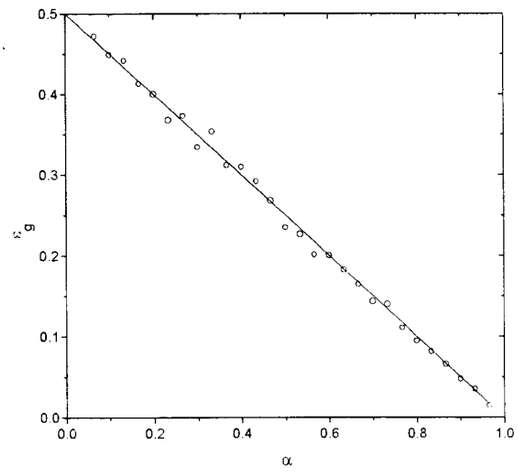


Figura 3.20: Erro de generalização em função de α para $\gamma = 1$ e $\kappa = 1$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

mostramos o efeito da diluição sobre o erro de generalização quando a rede é treinada na ausência de ruído com $\gamma = 1$ e $\kappa = 0.5$. Na figura (3.22) mostramos o efeito do ruído e da diluição sobre o erro de generalização quando a rede é treinada com $\gamma = 0.8$ e $\kappa = 0.5$.

3.3.2 Corte dos pesos maiores

Neste caso queremos que os pesos anulados sejam maiores que um certo limiar, ou seja, $|J_i| > \omega$. Assim o erro de treinamento médio é dado por

$$\epsilon_t^g = \frac{1}{\alpha N} \lim_{\beta\to\infty} \langle\langle E\left[\{J_i\left(1-\Theta\left(|J_i|-\omega\right)\right)\}\right]\rangle_T\rangle\rangle\,, \tag{3.128}$$

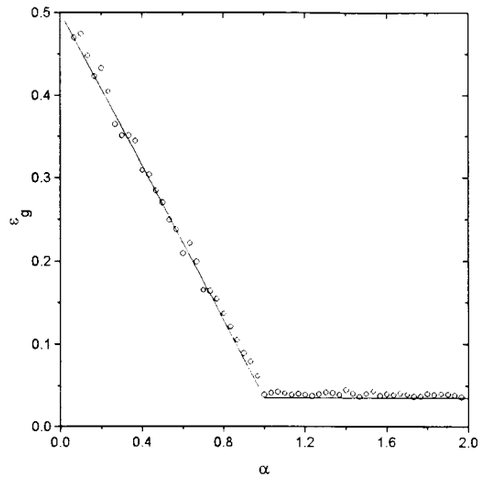


Figura 3.21: Erro de generalização em função de α para $\gamma = 1$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

onde E é dada em (2.4). Já o erro de generalização médio é dado por

$$\epsilon_g^g = \lim_{\beta\to\infty} \left\langle \left\langle \langle E_g \left[\{ J_i \left(1 - \Theta \left(|J_i| - \omega \right) \right) \} \right] \rangle_T \right\rangle \right\rangle , \tag{3.129}$$

onde E_g é dado em (2.19). A relação entre ω e κ é

$$\kappa = \left\langle \left\langle \left\langle \frac{1}{N} \sum_i^N \left(1 - \Theta \left(|J_i| - \omega \right) \right) \right\rangle_T \right\rangle \right\rangle \tag{3.130}$$

que fornece

$$\kappa = 2 \int_0^{\omega/\sqrt{Q_1}} Dz. \tag{3.131}$$

Todas as equações obtidas no cálculo dos erro de treinamento e generalização são idênticas às equações obtidas com o corte dos pesos menores desde que se substitua Λ_κ por $\Lambda_{1-\kappa}$, dado por

$$\Lambda_{1-\kappa} = 2 \int_0^{\omega/\sqrt{Q_1}} Dz\, z^2.$$

As figuras (3.23) e (3.24) mostram os erros de treinamento e de generalização respectivamente, quando a rede é treinada com ruído $\gamma = 0.8$ e com diluição $\kappa = 0.5$

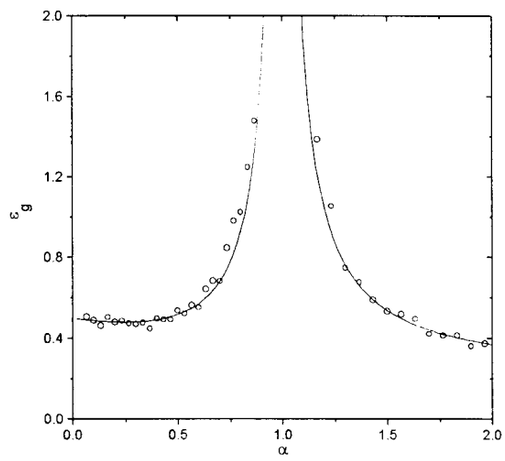


Figura 3.22: Erro de generalização em função de α para $\gamma = 0.8$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

3.3.3 Corte aleatório dos pesos

Neste tipo de diluição os pesos a serem cortados não dependem de seus valores, ou seja, serão cortados $(1 - \kappa) N$ pesos escolhidos aleatoriamente. Neste caso o corte de pesos leva em conta apenas o parâmetro κ

$$\kappa = \frac{1}{N} \sum_{i}^{N} c_i, \tag{3.132}$$

onde c_i são variáveis aleatórias estatisticamente independentes distribuídas de acordo com distribuição de probabilidade

$$P(c_i) = \kappa \delta(c_i - 1) + (1 - \kappa) \delta(c_i). \tag{3.133}$$

O erro de treinamento médio é dado por

$$\epsilon_t^a = \frac{1}{\alpha N} \lim_{\beta \to \infty} \langle\langle\langle E(\{c_i J_i\})\rangle_T\rangle\rangle_c \tag{3.134}$$

onde

$$E(\{c_i J_i\}) = \frac{1}{2} \sum_{i=1}^{P} \left(t^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} S_i^l c_i J_i \right)^2. \tag{3.135}$$

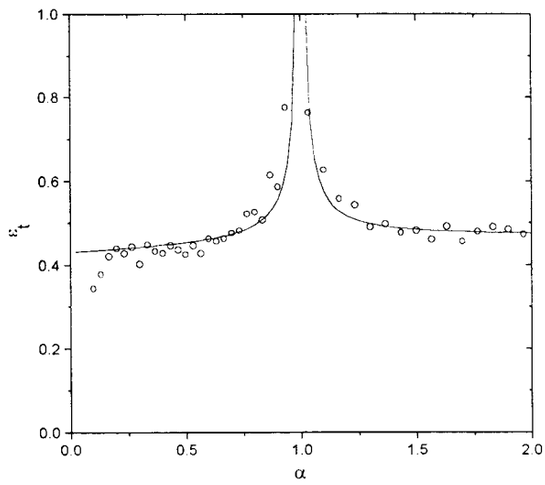


Figura 3.23: Erro de treinamento em função de α para $\gamma = 0.8$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

Para efetuar a média sobre as variáveis c_i vamos escrever ϵ_t^a da seguinte maneira

$$\epsilon_t^a = \frac{1}{\alpha N} \lim_{\beta\to\infty} \left\langle \left\langle \left\langle \frac{\mathrm{Tr}\, e^{-\beta E_{\kappa=1}(\{J_i\})} E\left(\{c_i J_i\}\right)}{\mathrm{Tr}\, e^{-\beta E_{\kappa=1}(\{J_i\})}} \right\rangle \right\rangle_{S_i,\xi_i} \right\rangle_{c_i} . \tag{3.136}$$

Como $E_{\kappa=1}(\{J_i\})$ não depende de c_i, podemos efetuar a média sobre $E(\{c_i J_i\})$ diretamente obtendo

$$\epsilon_t^a = \frac{1}{\alpha N} \lim_{\beta\to\infty} \left\langle \left\langle \frac{\mathrm{Tr}\, e^{-\beta E_{\kappa=1}(\{J_i\})} \left[E\left(\{\kappa J_i\}\right) + \frac{1}{2}\alpha\kappa\left(1-\kappa\right)\sum_i J_i^2\right]}{\mathrm{Tr}\, e^{-\beta E_{\kappa=1}(\{J_i\})}} \right\rangle \right\rangle_{S_i,\xi_i} . \tag{3.137}$$

A expressão acima pode ser escrita como

$$\epsilon_t^a = \frac{1}{\alpha N} \lim_{\beta\to\infty} \left\langle \left\langle \left\langle E\left(\{\kappa J_i\}\right) - \frac{1}{2}\alpha\kappa\left(1-\kappa\right) N Q \right\rangle \right\rangle_T \right\rangle . \tag{3.138}$$

Para efetuarmos a média sobre os exemplos e a média térmica vamos proceder como no caso do corte dos pesos menores, definindo a seguinte função de partição

$$Z = \mathrm{Tr}\exp\left\{ -\beta E_{\kappa=1}(\{J_i\}) - \frac{\beta h}{2}\left[\sum_l \left(t^l - \sum_i \frac{\kappa S_i^l J_i}{\sqrt{N}} \right)^2 + \alpha\kappa\left(1-\kappa\right)\sum_i J_i^2 \right] \right\} \tag{3.139}$$

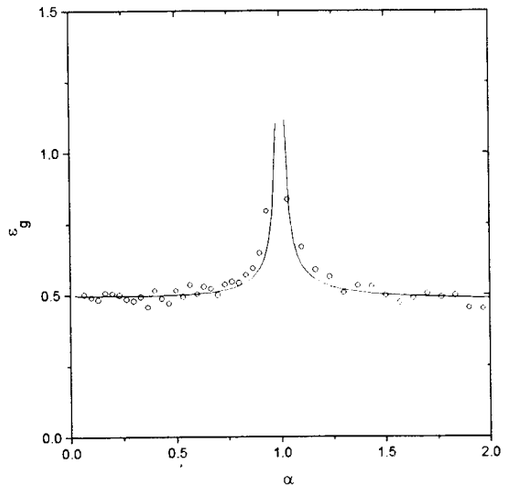


Figura 3.24: Erro de generalização em função de α para $\gamma = 0.8$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

de maneira que o erro de treinamento médio fica dado por

$$\epsilon_t^a = -\frac{1}{\beta}\frac{d}{dh}\langle\langle \ln Z\rangle\rangle \mid_{h=0} . \tag{3.140}$$

Como nos casos anteriores, o cálculo de $\langle\langle \ln Z\rangle\rangle$ é efetuado através do método das réplicas, utilizando-se a prescrição simetria das réplicas conforme mostrado em detalhes nas seções (A.4), (B.4) e (C.4). Neste caso ϵ_t^a é dado por

$$\epsilon_t^a = \frac{M + \kappa(1-\kappa)\alpha Q}{2} + \frac{\kappa(Q-q)(\kappa - 2M\beta)}{2[1+\beta(Q-q)]}$$

$$+\frac{\kappa^2\left[q + M\beta^2(Q-q)^2\right] - 2\kappa\gamma R[1+\beta(Q-q)(1-\kappa)]}{2[1+\beta(Q-q)]^2} \tag{3.141}$$

onde Q e R são dados pela equações (2.24) e (2.25). Devemos considerar novamente os dois limites, α acima e abaixo de $\alpha_c = 1$. O resultado final é

$$\epsilon_t^a/M = \frac{1-\kappa}{2}\left(1 - \kappa + \frac{\alpha\kappa(1-\gamma^2\alpha)}{1-\alpha}\right) \quad \alpha \leq 1 \tag{3.142}$$

e

$$\epsilon_t^a/M = \frac{\alpha^2(1-\gamma^2\kappa) - \alpha(1+\kappa(1-2\gamma^2)) + \kappa(2-\kappa)(1-\gamma^2)}{2\alpha(\alpha-1)} \quad \alpha > 1. \tag{3.143}$$

Para α grande esta equação reduz-se a

$$\epsilon_t^a/M = \frac{1-\gamma^2\kappa}{2} - \frac{\kappa(1-\gamma^2)}{2\alpha} + \mathcal{O}\left(\alpha^{-2}\right). \tag{3.144}$$

Na figura (3.25) mostramos o comportamento do erro de treinamento quando a rede é treinada na presença de ruído com $\gamma = 0.8$, e na presença de diluição com $\kappa = 0.5$.

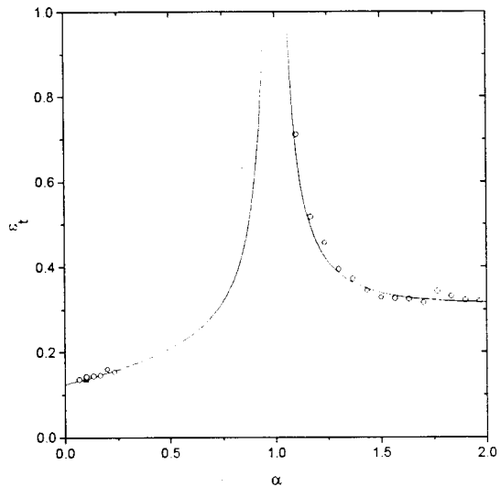


Figura 3.25: Erro de treinamento em função de α para $\gamma = 0.8$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

O erro de generalização médio é definido como

$$\epsilon_g^a = \lim_{\beta\to\infty} \left\langle\left\langle\left\langle E_g\left(\{c_i J_i\}\right)\right\rangle_T\right\rangle\right\rangle_c . \tag{3.145}$$

onde E_g é dado por (2.19). Após efetuar a média sobre os c_i 's como anteriormenteobtemos

$$\epsilon_g^a = \lim_{\beta\to\infty} \left\langle\left\langle\left\langle \frac{1}{2}\left(M + \kappa Q - 2\gamma\kappa R\right)\right\rangle_T\right\rangle\right\rangle \tag{3.146}$$

que reduz-se a

$$\epsilon_g^a/M = \frac{\alpha^2\gamma^2\kappa - \alpha\left(1-\kappa\left(1-2\gamma^2\right)\right)+1}{2\left(1-\alpha\right)} \qquad \alpha \leq 1 \tag{3.147}$$

e

$$\epsilon_g^a/M = \frac{1}{2} + \frac{\kappa\left(1-\gamma^2\alpha\right)}{2\left(\alpha-1\right)} \qquad \alpha > 1. \tag{3.148}$$

Assim, o comportamento assintótico ($\alpha \to \infty$) é dado por

$$\epsilon_g^a/M = \frac{1-\gamma^2\kappa}{2} + \frac{\kappa(1-\gamma^2)}{2\alpha} + \mathcal{O}\left(\alpha^{-2}\right). \tag{3.149}$$

Na figura (3.26) mostramos o efeito da diluição sobre o erro de generalização quando a rede é treinada na ausência de ruído com $\gamma = 1$ e $\kappa = 0.5$. Na figura (3.27) mostramos o efeito do ruído e da diluição sobre o erro de generalização quando a rede é treinada com $\gamma = 0.8$ e $\kappa = 0.5$.

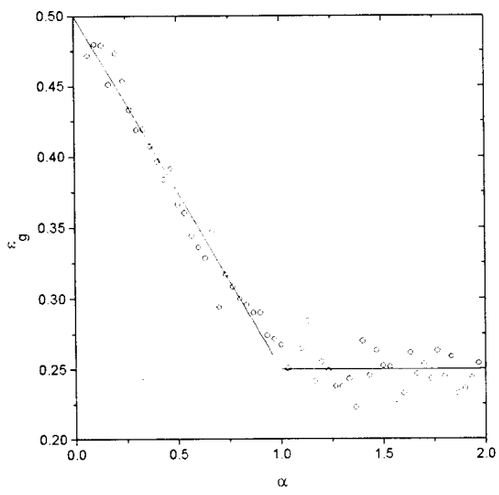


Figura 3.26: Erro de generalização em função de α para $\gamma = 1$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

Como no caso das diluições durante o aprendizado onde havia uma grande similaridade entre as equações, na diluição após o aprendizado podemos notar que para obtermos as equações que descrevem o corte aleatório dos pesos basta trocarmos Λ_κ por κ nas equações que descrevem o corte dos pesos pequenos.

Para todos os tipos de diluição após o aprendizado o erro de treinamento coincide com o erro de generalização, para $\gamma = 1$ e $\alpha > 1$. Neste regime estas grandezas são independentes de α, pois o mínimo global $\{J_i = J_i^0\}$ independe de α .

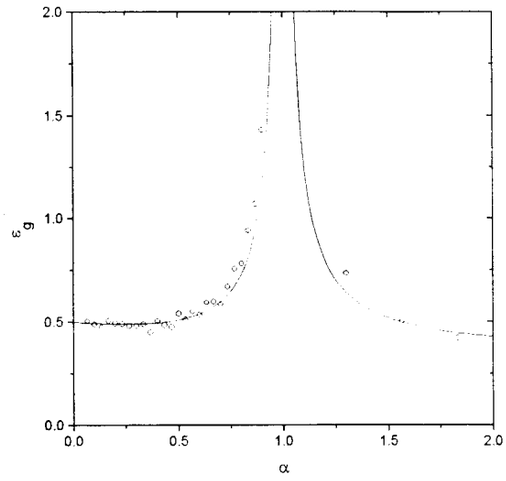


Figura 3.27: Erro de generalização em função de α para $\gamma = 0.8$ e $\kappa = 0.5$. A linha cheia representa os resultados teóricos e os símbolos representam os resultados obtidos das simulações realizadas com $N = 100$.

3.4 Discussão dos resultados

A fim de ilustrarmos os resultados obtidos nas seções anteriores deste capítulo, faremos uma análise das duas principais grandezas que medem o desempenho da rede, a saber, os erros médios de treinamento e de generalização. Nesta seção o principal objetivo é comparar o efeito causado pelos diversos tipos de diluição nas capacidades de treinamento e de generalização da rede estudante.

3.4.1 Erro de treinamento

Vamos começar analisando o comportamento do erro de treinamento no caso em que a rede estudante é treinada na ausência de ruído ($\gamma = 1$). Na figura (3.28) temos o comportamento do erro de treinamento ϵ_t em função do tamanho do conjunto de treinamento α para $\kappa = 0.5$ e para os vários tipos de diluição.

Podemos observar que o menor erro de treinamento ocorre no caso da diluição móvel, uma vez que neste caso o corte dos pesos é feito justamente de maneira a minimizar o

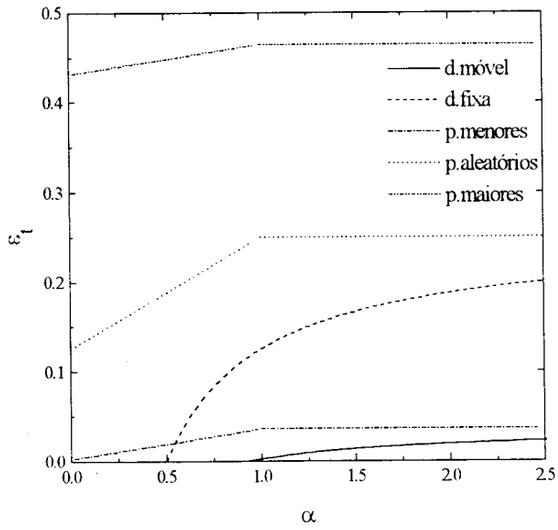


Figura 3.28: Erro de treinamento em função de α para $\gamma = 1$ e $\kappa = 0.5$

efeito da diluição sobre este erro. O corte dos pesos menores é o tipo de diluição que fornece o segundo menor erro de treinamento a partir de $\alpha_c^f = \kappa$, e o corte dos pesos maiores é o tipo de diluição que mais degrada o erro de treinamento. Este fato pode ser melhor ilustrado nas figuras (3.29) e (3.30) em que fixamos os valores do tamanho do conjunto de treinamento em $\alpha = 0.5$ e $\alpha = 2.0$ respectivamente e variamos o valor do grau de diluição κ .

No estudo da distribuição dos pesos feito anteriormente, vimos que a distribuição dos pesos remanescentes é uma Gaussiana sem a parte central, o que significa que estes pesos assumem valores não nulos a partir de um certo limiar, ou seja, sempre os pesos menores são eliminados. Assim era de se esperar que o corte dos pesos pequenos degradasse pouco o erro de treinamento.

O efeito do ruído é degradar ainda mais o erro de treinamento, além do que, neste caso o erro de treinamento diverge em $\alpha = 1$ para as diluições que ocorrem após o aprendizado. Isto pode ser visto na figura (3.31) que mostra o erro de treinamento em função de α para

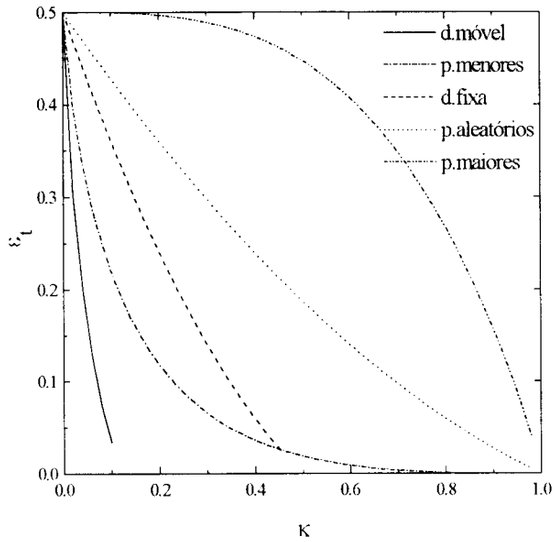


Figura 3.29: Erro de treinamento como função da conectividade κ para $\alpha = 0.5$ e $\gamma = 1.0$

$\gamma = 0.8$ e $\kappa = 0.5$.

3.4.2 Erro de generalização

Vamos começar analisando o erro de generalização ϵ_g no caso em que a rede estudante é treinada na ausência de ruído ($\gamma = 1$). O primeiro resultado interessante é que neste caso o menor erro de generalização ocorre para o corte dos pesos menores como pode ser visto na figura (3.32).Isto pode ser facilmente entendido de acordo com o seguinte argumento: para $\alpha > 1$ sabemos que o corte dos pesos é feito no vetor J_i^0 (o mínimo global $\epsilon_t = 0$ é único), assim no caso do corte dos pesos menores a equação (3.124) é dada por

$$E_g\left(J_i = J_i^0\right) = \frac{1}{2N}\left\{\sum_{i=1}^{N}\left(J_i^0\right)^2 - \sum_{i=1}^{N}\left(J_i^0\right)^2 \Theta\left(|J_i| - \omega\right)\right\}. \tag{3.150}$$

Nesta equação podemos ver que a soma é efetuada apenas sobre os pesos que são cortados. Portanto, quanto menor for o valor destes pesos, menor será o valor de E_g . Para $\alpha < 1$ como as configurações do mínimo global não são conhecidas, não podemos

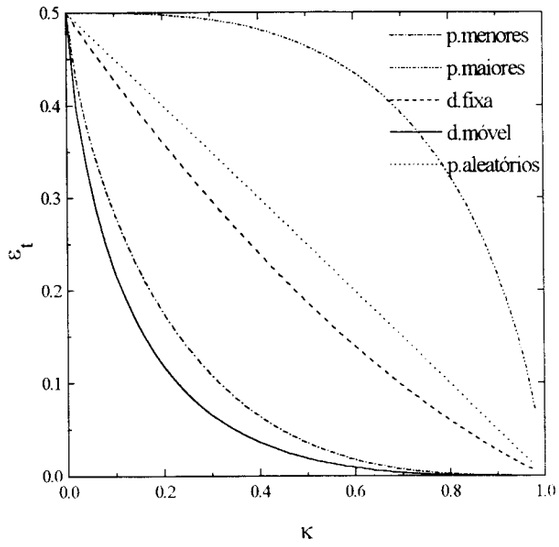


Figura 3.30: Erro de treinamento em função da conectividade κ para $\alpha = 2.0$ e $\gamma = 1.0$

provar a otimização do corte dos pesos menores, que ocorre para qualquer valor de α como pode ser visto nas figuras (3.33) e (3.34) onde fixamos o tamanho do conjunto de treinamento em $\alpha = 0.5$ e $\alpha = 2.0$ e variamos o grau de diluição κ.

No caso em que a rede é treinada na presença de ruído ($\gamma < 1$) o corte dos pesos menores fornece o menor erro de generalização apenas fora do intervalo $2 - 1/\gamma^2 < \alpha < 1/\gamma^2$. A figura (3.35) mostra ϵ_g em função de α para $\kappa = 0.5$ e $\gamma = 0.8$.

No intervalo em que $2-1/\gamma^2 < \alpha < 1/\gamma^2$ temos $\epsilon_g^p > \epsilon_g^a > \epsilon_g^g$. Quanto maior for o ruído, mais importante será a contribuição dos pesos menores ou seja, maior é a sensibilidade da rede ao corte dos pesos menores.

No limite assintótico $\alpha \to \infty$ o erro de generalização deve tender ao erro de treinamento. No caso da diluição móvel $\epsilon_t \to (1 - \gamma^2 \Lambda_\kappa)/2$, que coincide com o valor de ϵ_t no caso do corte dos pesos menores. Assim, o erro de generalização obtido com o corte dos pesos menores tende ao erro de generalização obtido com a diluição móvel.

No caso em que o treinamento é realizado na ausência de ruído, a diluição sempre

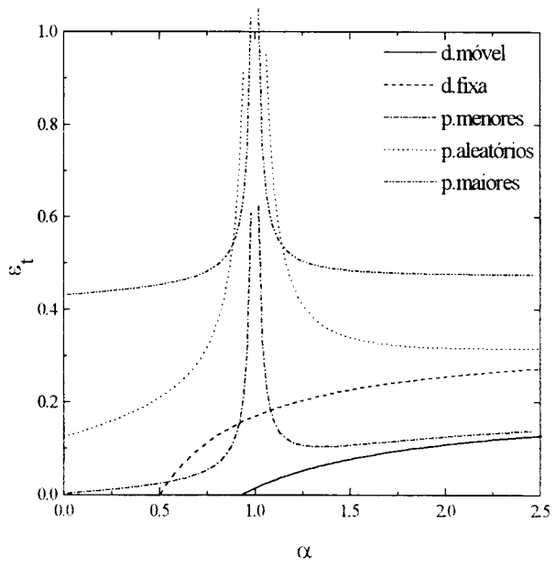


Figura 3.31: Erro de treinamento em função de α para $\gamma = 0.8$ e $\kappa = 0.5$

degrada o erro de generalização como pode ser visto na figura (3.36). Porém, na presença de ruído a rede diluída generaliza melhor que a rede não diluída para $2 - 1/\gamma^2 < \alpha < 1/\gamma^2$, como pode ser visto na figura (3.37). Isto ocorre para qualquer tipo de diluição após o aprendizado.

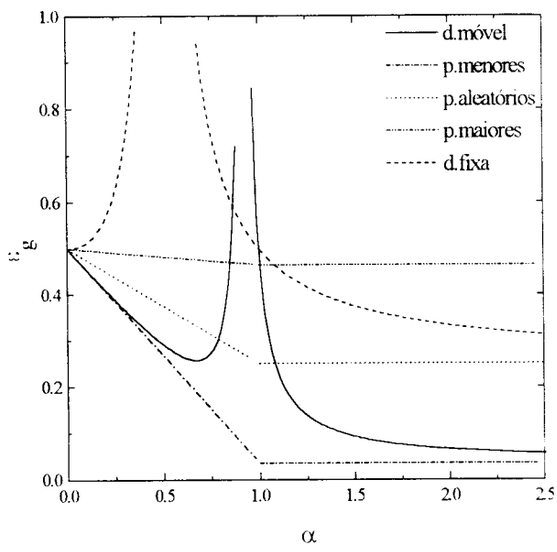


Figura 3.32: Erro de generalização em função de α para $\gamma = 1$ e $\kappa = 0.5$

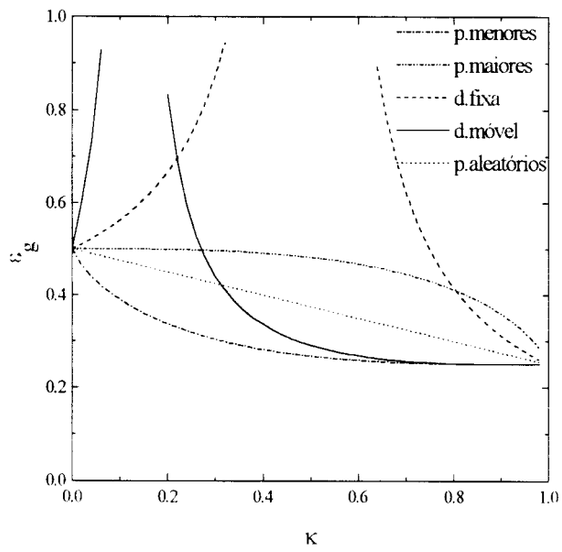


Figura 3.33: Erro de generalização em função de κ para $\gamma = 1$ e $\alpha = 0.5$

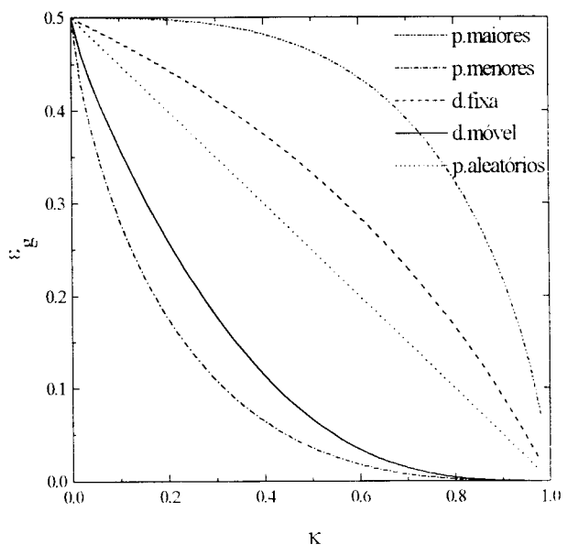


Figura 3.34: Erro de generalização em função de κ para $\gamma = 1$ e $\alpha = 2.0$

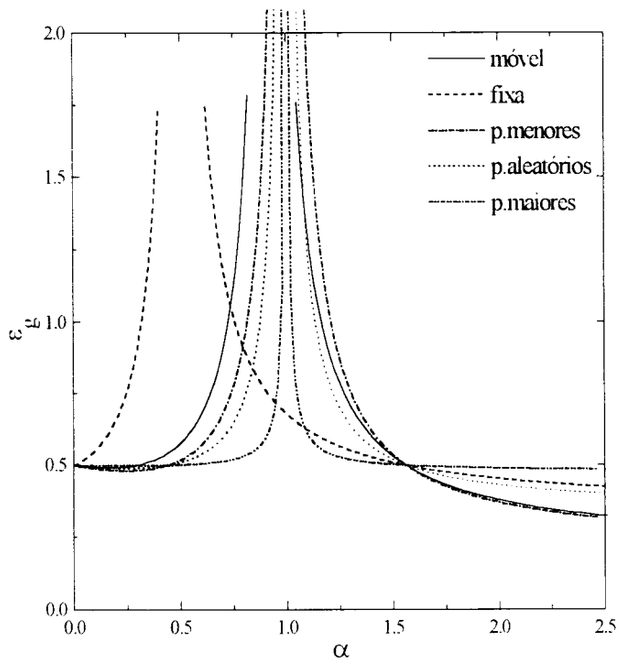


Figura 3.35: Erro de generalização em função de α para $\gamma = 0.8$ e $\kappa = 0.5$

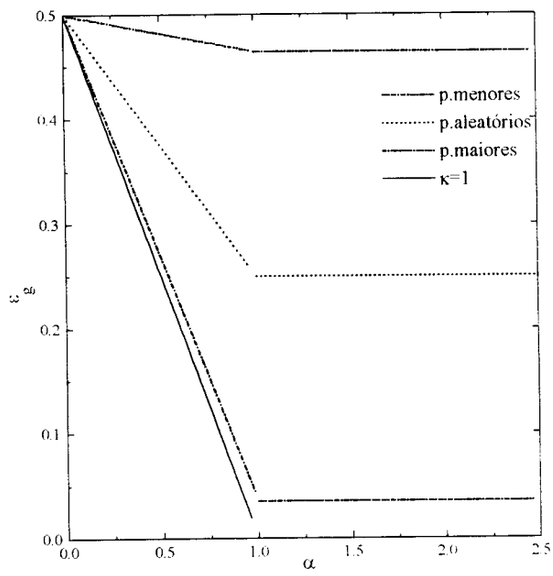


Figura 3.36: Erro de generalização em função de α, para $\gamma = 1$ e $\kappa = 0.5$

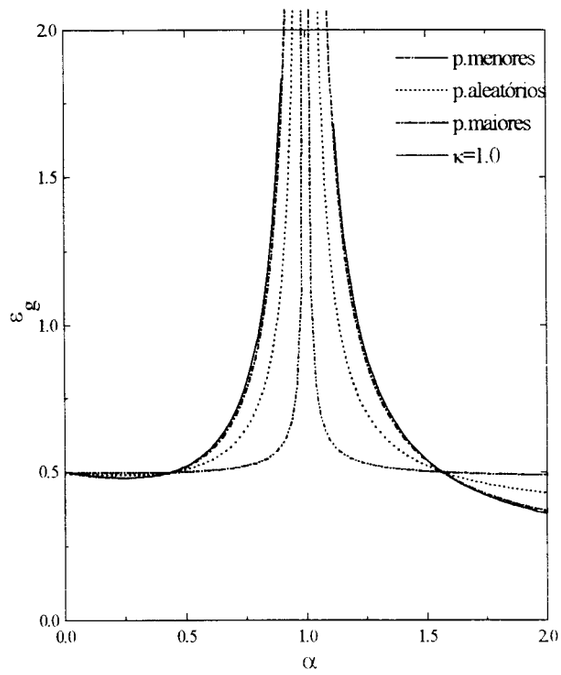


Figura 3.37: Erro de generalização em função de α, para $\gamma = 0.8$ e $\kappa = 0.5$

Capítulo 4

Perceptron Booleano

Neste capítulo vamos considerar apenas o efeito da diluição após o término do aprendizado sobre a capacidade de generalização do perceptron Booleano. O estudo da diluição durante o aprendizado foi considerado por Kuhlmann e Muller [32] no caso do aprendizado por exemplos, e por Bouten e colaboradores [7] no caso do mapa aleatório. Esses trabalhos foram restritos ao algoritmo de Gibbs, definido na seção (2.1). Entretanto, uma vez que nesses casos (diluição móvel) a diluição é correlacionada com o aprendizado, certamente esse processo não é adequado para estudar a robusteza de redes neurais treinadas com diferentes algoritmos frente a eliminação de uma fração de pesos sinápticos.

No que segue vamos mostrar que os parâmetros Q e R sofrem apenas uma reescala sob o efeito da diluição. Este resultado é mostrado para um processo determinístico de diluição no qual o peso J_i passa a $J_i\mathcal{F}(J_i,\kappa)$ e independe da regra de aprendizado considerada. Como antes, κ é o grau de diluição. É claro que $\mathcal{F}$ deve ser uma função theta (ou uma soma de funções thetas) para representar uma diluição. Os resultados serão ilustrados para o corte dos pesos menores e para cinco regras de aprendizado, a saber, Hebb, pseudo-inversa, Gibbs, estabilidade ótima e Bayes.

4.1 Modelo

O perceptron estudante consiste de N unidades de entrada binárias $\xi_i = \pm 1\ (i = 1, \ldots, N)$,

N pesos sinápticos reais $J_i\,(i=1,\ldots,N)$ e uma unidade de saída Booleana

$$\sigma = \text{sign}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i\xi_i\right). \tag{4.1}$$

A tarefa do estudante é realizar o mapa entre as 2^N possíveis configurações de entrada $\{\xi\}$ e suas respectivas saídas $\{t\}$ geradas pelo perceptron mestre

$$t = \text{sign}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i^0\xi_i\right), \tag{4.2}$$

onde os pesos $J_i^0\,(i=1,\ldots,N)$ são variáveis estatisticamente independentes extraídas da distribuição de probabilidade Gaussiana

$$P\left(J_i^0\right) = \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{\left(J_i^0\right)^2}{2}\right]. \tag{4.3}$$

Para realizar essa tarefa o estudante é treinado com $P = \alpha N$ pares entrada/saída $\left(\boldsymbol{\xi}^l,\zeta^l\right)\,(l=1,\ldots,P)$ onde ζ^l é extraída da distribuição de probabilidade condicional

$$P\left(\zeta^l \mid t^l\right) = (1-\chi)\,\delta\left(\zeta^l - t^l\right) + \chi\delta\left(\zeta^l + t^l\right), \tag{4.4}$$

e t^l é a resposta do perceptron mestre para a entrada $\boldsymbol{\xi}^l$. Assim, o parâmetro $\chi \in [0,1/2]$ mede a probabilidade de que o dígito de saída ζ^l apresentado ao estudante esteja errado.

No caso Booleano é conveniente definir a estabilidade Δ^l como

$$\Delta^l = \frac{\zeta^l}{\sqrt{N}}\sum_{i=1}^{N} J_i\xi_i^l, \tag{4.5}$$

de maneira que a energia de treinamento relativa a todos os algoritmos que vamos estudar pode ser escrita como

$$E\left(\mathbf{J},\boldsymbol{\xi},\zeta\right) = \sum_{l=1}^{P} g\left(\Delta^l\right). \tag{4.6}$$

onde a forma funcional da função g depende do algoritmo específico de treinamento.

4.2 Erro de Generalização

Graças a equação (4.6), nesta seção vamos calcular o erro de generalização de maneira genérica para todos os algoritmos de treinamento. O erro de generalização médio para

padrões puros no caso de um perceptron Booleano cujos pesos após a diluição são $J_i\mathcal{F}(J_i,\kappa)$ $(i=1,\ldots,N)$ é escrito como

$$\epsilon_g = \frac{1}{\pi}\arccos\left(\frac{S}{\sqrt{P}}\right) \tag{4.7}$$

onde

$$P = \left\langle \frac{1}{N}\sum_i J_i^2\mathcal{F}^2(J_i,\kappa)\right\rangle, \tag{4.8}$$

$$S = \left\langle \frac{1}{N}\sum_i J_i^0 J_i\mathcal{F}(J_i,\kappa)\right\rangle \tag{4.9}$$

e $\mathcal{F}(J_i,\kappa)$ depende do tipo de diluição considerada. Como exemplo, consideramos o corte dos pesos menores. Neste caso temos

$$\mathcal{F}(J_i,\kappa) = \Theta(|J_i|-\omega) \tag{4.10}$$

onde o limiar de corte ω relaciona-se com o grau de diluição κ através da equação

$$\kappa = \left\langle \frac{1}{N}\sum_i \Theta(|J_i|-\omega)\right\rangle. \tag{4.11}$$

A determinação de ϵ_g depende então apenas do cálculo de P e S. A seguir calcularemos esses parâmetros para um $\mathcal{F}(J_i,\kappa)$.

O parâmetro P é dado por

$$P = -\frac{1}{\beta N}\frac{\partial}{\partial h}\langle \ln Z\rangle\,|_{h=0} \tag{4.12}$$

onde

$$Z = \sum_{\{J_i\}}\exp\left(-\beta E^{ef}\right) \tag{4.13}$$

e

$$E^{ef}(J_i) = E(J_i) + h\sum_i (J_i)^2\mathcal{F}^2(J_i,\kappa). \tag{4.14}$$

Conforme mencionado, a dependência da regra de aprendizado aparece na forma funcional da energia de treinamento (4.6). As médias sobre as variáveis estatisticamente independentes ξ_i^l, ζ^l e J_i^0 podem ser efetuadas através da aplicação do método das réplicas.

Mediando sobre ζ^l e introduzindo as identidades

$$\int_\infty^\infty \prod_{l=1}^{P} dy^l \delta\left(y^l - \frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i^0\xi_i^l\right) = 1 \tag{4.15}$$

e

$$\int_{\infty}^{\infty} \prod_{l,a} dx_a^l \delta\left(x_a^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^a \xi_i^l\right) = 1 \tag{4.16}$$

escrevemos

$$\begin{aligned}\langle\langle [Z(\beta,\xi,t)]^n \rangle\rangle &= \left\langle\left\langle \int \prod_l \frac{dy_l d\hat{y}_l}{2\pi} \int \prod_{l,a} \frac{dx_l^a d\hat{x}_l^a}{2\pi} \exp\left\{ i \sum_l \left(y_l \hat{y}_l + \sum_a x_l^a \hat{x}_l^a \right) \right.\right.\right. \\ &\left.\left.\left. + \sum_l \ln\left[(1-\chi) \exp\left(\beta \sum_a g\left(\operatorname{sign} y^l x_l^a\right)\right) + \chi \exp\left(-\beta \sum_a g\left(\operatorname{sign} y^l x_l^a\right)\right)\right]\right\}\right. \\ &\int \prod_{i,a} dJ_i^a \exp\left[-\beta h \sum_{a,l} \left(J_i^a\right)^2 \mathcal{F}^2\left(J_i, \kappa\right)\right] \\ &\left.\left\langle \exp\left[-\frac{i}{\sqrt{N}} \sum_{l,j} \left(\hat{y}_l J_i^0 \xi_i^l + \sum_a \hat{x}_l^a J_i^a \xi_i^l \right)\right]\right\rangle_{\xi_i^l}\right\rangle\end{aligned} \tag{4.17}$$

Efetuamos as médias sobre a variável ξ_i^l da mesma maneira que foi feita na seção (A.1). Assim, introduzindo os parâmetros de ordem por meio das identidades

$$\int \prod_{a<b} dq_{ab} \delta\left(q_{ab} - \frac{1}{N} \sum_{i=1}^{N} J_i^a J_i^b\right) = 1 \tag{4.18}$$

$$\int \prod_{a}^{n} dR_a \delta\left(R_a - \frac{1}{N} \sum_{i=1}^{N} J_i^0 J_i^a\right) = 1 \tag{4.19}$$

e

$$\int \prod_{a}^{n} dQ_a \delta\left(Q_a - \frac{1}{N} \sum_{i=1}^{N} \left(J_i^a\right)^2\right) = 1$$

podemos reescrever o n-ésimo momento de Z dado na equação (4.17) como

$$\begin{aligned}\langle\langle (Z(\beta,\mathbf{S},t))^n \rangle\rangle &= \int \prod_a \frac{d\hat{Q}_a dQ_a}{2\pi i/N} \int \prod_a \frac{d\hat{R}_a R_a}{2\pi i/N} \int \prod_{a<b} \frac{d\hat{q}_{ab} q_{ab}}{2\pi i/N} \\ &\exp N\left\{ G_0\left(q_{ab}, \hat{q}_{ab}, R_a, \hat{R}_a, Q_a, \hat{Q}_a\right)\right. \\ &\left. + \alpha G_1\left(q_{ab}, R_a\right) + G_2\left(\hat{q}_{ab}, \hat{Q}_a, \hat{R}_a\right)\right\}\end{aligned} \tag{4.20}$$

onde

$$G_0 = i \sum_a Q_a \hat{Q}_a + i \sum_a \hat{R}_a R_a + i \sum_{a<b} \hat{q}_{ab} q_{ab}, \tag{4.21}$$

$$G_1 = \ln \int \frac{dyd\hat{y}}{2\pi} e^{iy\hat{y}-\hat{y}^2/2} \int \prod_a \frac{dx^a d\hat{x}^a}{2\pi} \exp\left\{ \sum_a x^a \hat{x}^a \right.$$

$$\left. -\frac{1}{2}\sum_a^n \hat{x}^a Q_a - \sum_{a<b}^n \hat{x}^a \hat{x}^b q_{ab} - \sum_a^n \hat{x}^a \hat{y} R_a \right\}$$

$$\left[(1-\chi) \exp\left(\beta \sum_a g\left(\operatorname{sign} y x^a\right) \right) + \chi \exp\left(-\beta \sum_a g\left(\operatorname{sign} y x^a\right) \right) \right] \quad (4.22)$$

e

$$G_2 = \ln \int \prod_{a=1}^n dJ^a \exp\left\{ -i \sum_a^n \left[\hat{Q}_a \left(J^a\right)^2 + \hat{R}_a J^a J^0 \right] \right.$$

$$\left. \exp\left[-i \sum_{a<b}^n \hat{q}_{ab} J^a J^b - \beta h \sum_a^n \left(J^a\right)^2 \mathcal{F}^2\left(J_i, \kappa\right) \right] \right\}. \quad (4.23)$$

Aplicando o método de integração por ponto de sela dado pela equação (A.13) para efetuar as integrais sobre os parâmetros de ordem e usando a prescrição simetria das réplicas obtemos

$$\frac{1}{N}\langle \ln Z \rangle = R\hat{R} - \frac{1}{2} q\hat{q} + Q\hat{Q} + \alpha \left[(1-\chi) G_1^+ + \chi G_1^- \right] + G_2 \quad (4.24)$$

onde

$$G_1^{\mp} = \int Dz \int \frac{dyd\hat{y}}{2\pi} e^{-\hat{y}^2/2 + iy\hat{y}} \ln \int \frac{dxd\hat{x}}{2\pi} e^{i x \hat{x}}$$

$$\exp\left[-\frac{1}{2}\left(Q-q\right)\hat{x}^2 + i\sqrt{q} z \hat{x} - R\hat{x}\hat{y} - \beta g\left(\pm x \operatorname{sign} y\right) \right] \quad (4.25)$$

e

$$G_2 = \int Dz \int DJ^0 \ln \int DJ \exp\left[\mathcal{H}\left(J, J^0, z\right) - \beta h J^2 \mathcal{F}^2\left(J, \kappa\right) \right]. \quad (4.26)$$

Aqui, o hamiltoniano efetivo

$$\mathcal{H}\left(J, J^0, z\right) = -\left(\hat{Q} - \hat{q}/2\right) J^2 + i\sqrt{\hat{q}} J z - \hat{R} J J^0$$

independe da forma de $g(\mathbf{J})$ e, consequentemente, independe da regra de aprendizado. De acordo com a convenção anterior as notações Dz e DJ^0 são medidas Gaussianas. As

equações de ponto de sela são obtidas derivando-se a equação (4.24) com relação aos parâmetros $\left(q, \hat{q}, R, \hat{R}, Q, \hat{Q}\right)$e tomando-se $h = 0$. Assim da equação (4.12) obtemos

$$P = \int Dz \int DJ^0 \frac{\int DJ\, J^2 \mathcal{F}^2 (J, \kappa) \exp\left[\mathcal{H}\left(J, J^0, z\right)\right]}{\int DJ \exp\left[\mathcal{H}\left(J, J^0, z\right)\right]} \tag{4.27}$$

que fornece

$$P = Q \int DJ\, J^2 \mathcal{F}^2 \left(\sqrt{Q} J, \kappa\right). \tag{4.28}$$

Para calcular S devemos re-escrever a equação (4.14) como

$$E^{ef}\left(J_i\right) = E\left(J_i\right) + h \sum_i J_i^0 J \mathcal{F}\left(J_i, \kappa\right). \tag{4.29}$$

e seguindo o procedimento acima obtemos facilmente

$$S = R \int DJ\, J^2 \mathcal{F} \left(\sqrt{Q} J, \kappa\right). \tag{4.30}$$

Concluimos então que o erro de generalização (4.7) reduz-se a

$$\epsilon_g = \frac{1}{\pi} \arccos \left[\frac{R}{\sqrt{Q}} \frac{\int DJ\, J^2 \mathcal{F}\left(J_i, \kappa\right)}{\sqrt{\int DJ\, J^2 \mathcal{F}\left(J_i, \kappa\right)}}\right]. \tag{4.31}$$

Assim, uma vez que a regra de aprendizado determina os valores de Q e R o efeito da diluição é apenas uma reescala desses parâmetros. Vamos considerar apenas o corte dos pesos menores no qual $\mathcal{F}$ é dada por (4.10). Como neste caso temos $\mathcal{F}^2 = \mathcal{F}$ definimos a grandeza

$$\Lambda_\kappa = \int DJ\, J^2 \Theta \left(|J| - \omega/\sqrt{Q}\right) \tag{4.32}$$

com ω dado por (4.11). Daí, $P = Q\Lambda_\kappa$ e $S = R\Lambda_\kappa$. O cálculo de Λ_κ é imediato e resulta

$$\Lambda_\kappa = \left[\kappa + \frac{\sqrt{2}}{\pi} \lambda_\kappa e^{-\lambda_\kappa^2/2}\right] \tag{4.33}$$

sendo λ_κ a solução de

$$\kappa = 2 \int_{\lambda_\kappa}^{\infty} Dz. \tag{4.34}$$

Portanto o erro de generalização é dado por

$$\epsilon_g = \frac{1}{\pi} \arccos \left[\sqrt{\Lambda_\kappa} \frac{R}{\sqrt{Q}}\right]. \tag{4.35}$$

Nas seções seguintes vamos calcular Q e R para diversas regras de aprendizado. Esses cálculos são análogos aos desenvolvidos no capítulo anterior e, portanto, não apresentaremos seus detalhes. Além disso, esses cálculos encontram-se na literatura citada, referente a cada seção.

4.2.1 Hebb

Escrever os pesos da regra de Hebb em termos do conjunto de treinamento $\left\{\boldsymbol{\xi},\zeta^l\right\}$ é extremamente simples, porém para que a análise dos efeitos da diluição se torne mais fácil vamos usar a seguinte energia de treinamento [19]

$$E\left(\mathbf{J}\right) = -\sum_l \Delta^l \tag{4.36}$$

com a norma fixa $Q = 1$ que é minimizada pelo vetor peso dado pela regra de Hebb, a saber,

$$J_i = \frac{1}{\sqrt{N}} \sum_l \zeta^l \xi_i^l. \tag{4.37}$$

O cálculo de $G_1^{\pm}$ é muito simples nesse caso, pois $g\left(\Delta\right)$ é uma fução linear. O resultado final para a densidade de energia livre com $h = 0$ é

$$\begin{aligned} -\beta f &= R\hat{R} - \frac{1}{2}q\hat{q} + \hat{Q} + \alpha\frac{\beta^2\left(1-q\right)}{2} + \alpha\left(1-2\chi\right)\beta R\frac{\sqrt{2}}{\sqrt{\pi}} + \int Dz \int DJ^0 \\ &\times \ln \int dJ \exp\left[-\left[\hat{Q} - \hat{q}/2\right]J^2 + \left(i\sqrt{\hat{q}}z - \hat{R}J^0\right)J\right]. \end{aligned} \tag{4.38}$$

onde $\mathcal{F}\left(J_i,\kappa\right)$ é dado pela equação (4.10)

Como usual, para encontrarmos as equações de ponto de sela derivamos a equação (4.38) com relação aos parâmetros de ordem e eliminamos as variáveis auxiliares $\hat{Q},\hat{q},\hat{R}$. Assim,

$$R = \sqrt{\frac{2\alpha\left(1-2\chi\right)^2}{\pi + 2\alpha\left(1-2\chi\right)^2}} \tag{4.39}$$

que coincide com o resultado obtido por Vallet [47] para $\chi = 0$.

O erro de generalização (4.35) no limite de α grande é dado por

$$\epsilon_g^H \approx \frac{1}{\sqrt{2\pi\alpha\left(1-2\chi\right)^2}} \tag{4.40}$$

para $\kappa = 1$ e

$$\epsilon_g^H \approx \frac{1}{\pi}\arccos\sqrt{\Lambda_\kappa} + \left(\frac{\Lambda_\kappa}{1-\Lambda_\kappa}\right)^{1/2}\frac{1}{4\alpha\left(1-2\chi\right)^2} \tag{4.41}$$

para $\kappa < 1$.

4.2.2 Pseudo-Inversa

No caso do treinamento do perceptron Booleano com a regra pseudo-inversa, a energia de treinamento é exatamente a mesma energia usada no perceptron linear. Porém, como $\zeta^l = \pm 1$ podemos ver através de um cálculo simples que a equação (2.4) passa a ser escrita como

$$E\left(\mathbf{J}\right) = \frac{1}{2}\sum_l\left(1 - \frac{\zeta^l}{\sqrt{N}}\sum_i J_i\xi_i^l\right)^2 = \frac{1}{2}\sum_l\left(1-\Delta^l\right)^2. \tag{4.42}$$

No caso da regra pseudo-inversa a equação (4.24) é re-escrita como

$$\begin{aligned} -\beta f \;=\; & R\hat{R} - \frac{1}{2}q\hat{q} + Q\hat{Q} - \frac{\alpha}{2}\ln\left[1+\beta\left(Q-q\right)\right] \\ & -\frac{\alpha\beta}{\left[1+\beta\left(Q-q\right)\right]}\left\{\frac{1+q}{2} + \left(1-2\chi\right)R\frac{\sqrt{2}}{\sqrt{\pi}}\right\} + \int Dz\int DJ^0 \\ & \times\ln\int dJ\exp\left[-\left[\hat{Q}-\hat{q}/2\right]J^2 + \left(i\sqrt{\hat{q}}z - \hat{R}J^0\right)J\right]. \end{aligned} \tag{4.43}$$

As equações de ponto de sela são obtidas derivando-se a equação (4.43) com relação a $\left(R,\hat{R},q,\hat{q},Q,\hat{Q}\right)$. Neste caso novamente vamos analisar os parâmetros de ordem para $\alpha \leq 1$ e para $\alpha > 1$. No caso que $\alpha \leq 1$ existe mais de um vetor peso $\mathbf{J}$ que minimiza a energia de treinamento dada pela equação (4.42), assim devemos escolher a solução de menor norma onde os parâmetros de ordem são dados por

$$Q = \frac{\alpha}{\pi}\frac{\pi - 2\alpha\left(1-2\chi\right)^2}{1-\alpha} \tag{4.44}$$

e

$$R = \sqrt{\frac{2}{\pi}}\alpha\left(1-2\chi\right). \tag{4.45}$$

No caso que $\alpha > 1$ temos

$$Q = \frac{2(\alpha - 2)(1 - 2\chi)^2 + \pi}{\pi(\alpha - 1)} \tag{4.46}$$

$$R = \sqrt{\frac{2}{\pi}}(1 - 2\chi). \tag{4.47}$$

Estes resultados concordam com os resultados obtidos por Opper e colaboradores [38] para o caso sem ruído. Para α grande temos

$$\epsilon_g^P \approx \frac{1}{\sqrt{2\pi\alpha(1 - 2\chi)^2}}\left[1 - \frac{2}{\pi}(1 - 2\chi)^2\right]^{1/2} \tag{4.48}$$

para $\kappa = 1$ e

$$\epsilon_g^P \approx \frac{1}{\pi}\arccos\sqrt{\Lambda_\kappa} + \left(\frac{\Lambda_\kappa}{1 - \Lambda_\kappa}\right)^{1/2}\frac{1}{4\alpha(1 - 2\chi)^2}\left[1 - \frac{2}{\pi}(1 - 2\chi)^2\right] \tag{4.49}$$

para $\kappa < 1$. Neste limite temos $\epsilon_g^H > \epsilon_g^P$ para todo χ .

4.2.3 Algoritmo de Gibbs

O algoritmo de Gibbs escolhe o vetor peso aleatoriamente de acordo com a distribuição de probabilidade de Gibbs

$$\Pr(\mathbf{J}) = \frac{1}{Z}\exp\left[-\beta E(\mathbf{J})\right] \tag{4.50}$$

onde

$$E(\mathbf{J}) = \sum_l \Theta\left(K - \frac{\zeta^l}{\sqrt{N}}\sum_i J_i\xi_i^l\right) \tag{4.51}$$

é a energia de treinamento. β é o inverso da temperatura . $K \geq 0$ é o parâmetro de margem e Z é a função de partição. No caso do algoritmo de Gibbs a normalização dos pesos não é relevante (ela fornece a escala de medida do parâmetro K) de maneira que faremos $Q = 1$. O conjunto de pesos gerado por este algoritmo é caracterizado pelos parâmetros de ordem R e q que são pontos de extremo da seguinte densidade de energia

livre

$$-\beta f = \frac{1}{2}\left[\frac{q-R^2}{1-q}+\ln(1-q)\right]$$

$$+\alpha\int Dt\left[\chi+(1-2\chi)H(\xi_1)\right]\ln\left[e^{-\beta}+\left(1-e^{-\beta}\right)H(\xi_2)\right] \quad (4.52)$$

com

$$\xi_1 = \frac{Rt}{\sqrt{q-R^2}}, \quad (4.53)$$

$$\xi_2 = \frac{K+\sqrt{q}t}{\sqrt{1-q}} \quad (4.54)$$

e

$$H(x) = \int_x^\infty Dt \quad (4.55)$$

onde $Dt = dt/\sqrt{2\pi}e^{-t^2/2}$ é uma medida Gaussiana. As equações de ponto de sela para R e q são dadas por

$$\frac{R^2}{(1-q)} = -\frac{2\alpha u(1-2\chi)}{\sqrt{2\pi(q-R^2)}}\int\frac{e^{-ut^2/2}}{\sqrt{2\pi}}dt\, t\ln v \quad (4.56)$$

e

$$\frac{q^3+R^3-qR^2(1+R)}{2q(1-q)^2} = \frac{\alpha}{\sqrt{2\pi}}\int Dt\left[\chi+(1-2\chi)H(\xi_1)\right]\frac{\left(1-e^{-\beta}\right)e^{\xi_2^2/2}}{v}\xi_3 \quad (4.57)$$

respectivamente, com

$$u = \frac{q}{q-R^2}, \quad (4.58)$$

$$v = \left[e^{-\beta}+\left(1-e^{-\beta}\right)H(\xi_2)\right] \quad (4.59)$$

e

$$\xi_3 = \frac{t+K\sqrt{q}}{\sqrt{q(1-q)}(1-q)}. \quad (4.60)$$

Estabilidade da solução com simetria de réplicas

Para testarmos a validade da solução com simetria de réplicas devemos verificar se esta solução satisfaz a condição de estabilidade local dada por [1]. [17]

$$\alpha\gamma_1\gamma_2 < 1. \quad (4.61)$$

Efetuando o cálculo de γ_1 e γ_2 da mesma maneira que foi descrita no capítulo 3 obtemos

$$\alpha\gamma_1\gamma_2 = 2\alpha\left\{(1-\chi)\int_{K-\sqrt{2x}}^{K} Dt - (1-2\chi)\int_{K-\sqrt{2x}}^{K} DtH(\xi_1)\right\} \tag{4.62}$$

onde $x = \beta(1-q)$. Da equação (4.62) temos que no limite $\beta \to \infty$ a solução só é estável para $K \leq K_c(\alpha, \chi)$. Neste regime o conjunto de treinamento é aprendido perfeitamente produzindo erro de treinamento médio nulo. Na figura (4.1) mostramos K_c em função de α para vários valores de χ. Note que a curva para $\chi = 0$ nunca toca o eixo-α.

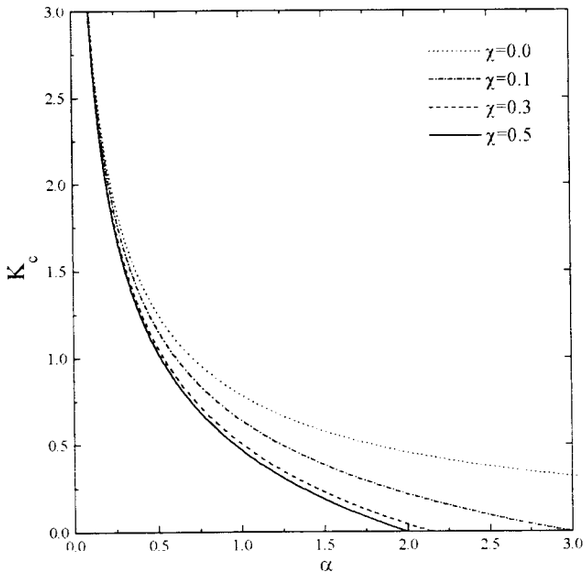


Figura 4.1: Parâmetro de margem K_c em função do tamanho do conjunto de treinamento α

Embora uma escolha cuidadosa do parâmetro de margem K possa melhorar o desempenho do algoritmo de Gibbs [36], neste trabalho consideramos apenas a escolha padrão $K = 0$.

4.2.4 Algoritmo de estabilidade ótima

As propriedades de equilíbrio do único vetor peso gerado pelo algoritmo de estabilidade ótima [30] são obtidas tomando-se o limite $K = K_c(\alpha, \chi)$. Assim, o algoritmo de estabilidade ótima é o algoritmo de Gibbs com o parâmetro de margem K no valor máximo em que este algoritmo é consistente, isto é, apresenta erro de treinamento nulo.

Tomando o limite $q \to 1$ nas equações (4.56) e (4.57) obtemos

$$R = \frac{\alpha(1-2\chi)}{\pi}\left\{\sqrt{1-R^2}\exp -\frac{K^2}{2(1-R^2)} + K\sqrt{2\pi}H\left[-\frac{K}{2\sqrt{1-R^2}}\right]\right\} \tag{4.63}$$

e

$$\frac{1-R^2}{2} = \alpha(1-2\chi)\int_{-K}^{\infty} DtH(\xi_1)(t+K)^2 + \alpha\chi\left\{\frac{K}{\sqrt{2\pi}}e^{-K^2/2} + H(-K)\left(1+K^2\right)\right\}. \tag{4.64}$$

4.2.5 Algoritmo de Bayes

O erro de generalização de Bayes foi calculado analiticamente através da abordagem da Mecânica Estatística por Opper e Haussler [39] [40] com o auxílio do algoritmo de Gibbs. A idéia principal para este cálculo é baseada no algoritmo da maioria. Quando a rede é treinada com $P = \alpha N$ exemplos os vetores peso $\mathbf{J}$ que implementam estes P exemplos são gerados de acordo com a distribuição de Gibbs dada na equação (4.50). O algorítmo de Gibbs não é o melhor algoritmo a ser usado uma vez que se queira maximizar a probabilidade de acerto para a $\boldsymbol{\xi}^{P+1}$ entrada escolhida aleatoriamente. Para que a predição desta entrada tenha a maior probabilidade de estar correta, o procedimento ideal é o seguinte: dada a entrada $\boldsymbol{\xi}^{P+1}$, calcula-se a saída σ^{P+1} para cada vetor $\mathbf{J}$ gerado pela distribuição de Gibbs segundo a regra $\sigma^{P+1} = \text{sgn}\left(1/\sqrt{N}\sum \xi_i^{P+1} J_i\right)$. Este cálculo vai dividir o conjunto dos pesos em dois subconjuntos, tais que o primeiro contenha os pesos $\mathbf{J}$ que classificam $\boldsymbol{\xi}^{P+1}$ como $+1$ e o segundo contenha os pesos $\mathbf{J}$ que classificam $\boldsymbol{\xi}^{P+1}$ como -1. Segundo o algoritmo da maioria, para que a probabilidade de erro seja menor deve-se classificar a entrada $\boldsymbol{\xi}^{P+1}$ de acordo com a classificação do maior subconjunto pois há maior chance do vetor professor estar contido nele. O algoritmo de Bayes é idêntico ao algoritmo da maioria, exceto que a temperatura da distribuição de Gibbs depende do

parâmetro de ruído [40]

$$\beta = \ln \frac{1-\chi}{\chi}. \tag{4.65}$$

Para esta temperatura os extremos da energia livre dada pela equação (4.52) são $q = R$, onde R é solução da equação

$$\frac{R}{\sqrt{1-R}} = \frac{\alpha}{\pi}(1-2\chi)^2 \int Dt \frac{e^{-Rt^2/2}}{\chi + (1-2\chi)\, H\left(\sqrt{R}t\right)}. \tag{4.66}$$

Assim, erro de generalização para o perceptron de Bayes diluído é [39] [40]

$$\epsilon_g^B = \frac{1}{\pi} \arccos\left(\sqrt{\Lambda_\kappa R}\right) \tag{4.67}$$

com R dado pela solução da equação (4.66).

No limite de α grande o erro de generalização é dado por

$$\epsilon_g^B \approx \frac{1}{\pi \alpha \Xi (1-2\chi)^2} \tag{4.68}$$

para $\kappa = 1$ e

$$\epsilon_g^B \approx \frac{1}{\pi} \arccos\sqrt{\Lambda_\kappa} + \left(\frac{\Lambda_\kappa}{1-\Lambda_\kappa}\right)^{1/2} \frac{1}{2\pi\alpha^2\Xi^2(1-2\chi)^4} \tag{4.69}$$

para $\kappa < 1$, onde

$$\Xi = \frac{1}{\pi} \int Dt \frac{e^{-t^2/2}}{\chi + (1-2\chi)\, H(t)}. \tag{4.70}$$

O erro de generalização fornecido pelo algoritmo de Bayes pode agora ser calculado sem o auxílio do algoritmo de Gibbs, encontrando-se os pesos que minimizam a energia $E(\mathbf{J})$ extraída da abordagem variacional desenvolvida por Kinouchi e Caticha [25]. O cálculo efetuado através da minimização desta energia fornece exatamente o mesmo parâmetro dado pela equação (4.66).

4.3 Corte aleatório dos pesos

Neste tipo de diluição são cortados $(1-\kappa)N$ pesos escolhidos aleatoriamente, onde

$$\kappa = \frac{1}{N} \sum_i^N c_i \tag{4.71}$$

e c_i são variáveis aleatórias estatisticamente independentes distribuídas de acordo com a distribuição

$$P(c_i) = \kappa\delta(c_i - 1) + (1 - \kappa)\,\delta(c_i). \tag{4.72}$$

Assim os parâmetros de ordem são dados por

$$P = \overline{\left\langle \frac{1}{N}\sum_i (J_i)^2 (c_i)^2 \right\rangle} \tag{4.73}$$

e

$$S = \overline{\left\langle \frac{1}{N}\sum_i J_i J_i^0 c_i \right\rangle} \tag{4.74}$$

onde a barra indica uma média sobre as variáveis aleatórias c_i. Como a média $\langle \ldots \rangle$ não involve as variáveis c_i (a diluição ocorre após o treinamento) podemos efetuar as médias diretamente nas equações (4.73) e (4.74). Lembrando que $\overline{c_i} = \kappa$ obtemos:

$$P = \kappa Q \qquad \text{e} \qquad S = \kappa R. \tag{4.75}$$

4.4 Discussão dos resultados

Nesta seção vamos ilustrar o desempenho de generalização do perceptron Booleano. Para isto vamos comparar o erro de generalização dos vários algoritmos discutidos acima. Quando o perceptron é treinado na ausência de ruído e de diluição, o menor erro de generalização é obtido pelo algoritmo de Bayes seguido pelo algoritmo de estabilidade ótima. Na figura (4.2) mostramos o erro de generalização como função do tamanho do conjunto de treinamento α para $\kappa = 1$ e $\chi = 0$. Esta figura é exatamente a mesma figura apresentada por Opper e colaboradores [38], exceto pelo fato de termos adicionado as curvas do algoritmo de Gibbs e do algoritmo de Bayes.

O comportamento do erro de generalização torna-se diferente quando o treinamento é realizado na presença de ruído ($\chi > 0$). Como pode ser visto na figura (4.3) que mostra o erro de generalização em função de α para $\kappa = 1.0$ e $\chi = 0.1$, o algoritmo de Bayes continua fornecendo o menor erro de generalização, mas há uma melhora no desempenho das regras pseudo-inversa e de Hebb. As curvas de aprendizado para os algoritmos de

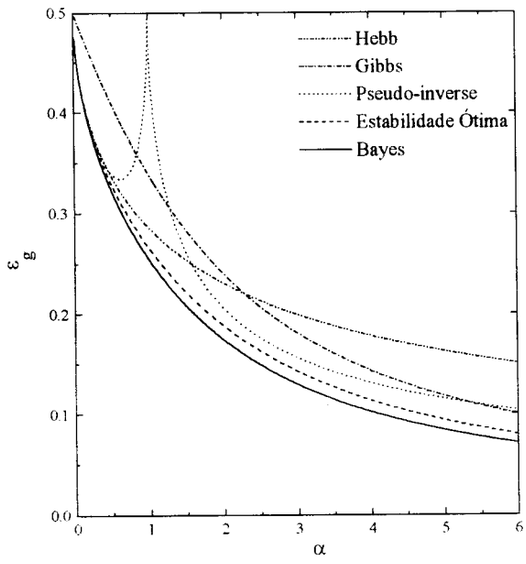


Figura 4.2: Erro de generalização em função de α para $\kappa = 1$ e $\chi = 0$

Gibbs e de estabilidade ótima são mostradas apenas para $\alpha \leq 2.992$, pois além deste valor a solução com simetria de réplicas é localmente instável. Quando o treinamento é realizado na presença de diluição o desempenho relativo dos algoritmos continua o mesmo, porém a diferença entre o desempenho dos algoritmos diminui. À medida que κ diminui o desempenho de todos os algoritmos tende a $\epsilon_g = 0.5$. Este comportamento pode ser visto na figura (4.4) que mostra o erro de generalização em função de α para $\kappa = 0.5$ e $\chi = 0.1$

O efeito da diluição após o aprendizado sobre o erro de generalização de diferentes algoritmos de treinamento não altera a ordem de seus desempenhos. Em particular, o algoritmo de Bayes fornece o menor erro de generalização, mesmo sob o efeito de diluição. Este importante resultado depende de a energia de treinamento para cada algoritmo depender apenas dos pesos através da estabilidade dos padrões dada na equação (4.5), e da estabilidade da solução com simetria de réplicas que foi verificada para todas as curvas de aprendizado apresentadas. Assim o algoritmo de Bayes sempre fornecerá o menor erro

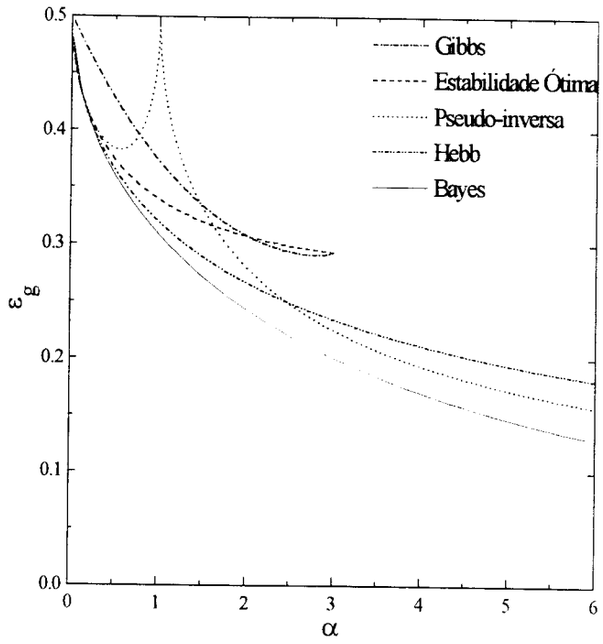


Figura 4.3: Erro de generalização em função de α para $\kappa = 1$ e $\chi = 0.1$

de generalização. Este comportamento se estende ao corte dos pesos maiores e ao corte aleatório dos pesos.

No perceptron Booleano, ao compararmos o desempenho do erro de generalização para uma mesma regra, independente da rede ser treinada com ou sem ruído, vemos que o melhor e o pior desempenho ocorrem para o corte dos pesos menores e para o corte dos pesos maiores respectivamente. Esse comportamento contrasta com o que ocorre no perceptron linear e pode ser visto na figura (4.5) que mostra o erro de generalização em função de α para $\gamma = 0.5$ e $\kappa = 0.8$ para a regra pseudo-imversa

Embora neste capítulo tenhamos considerado apenas o ruído sobre o dígito de saída cuja intensidade é dada pelo parâmetro χ, também realizamos todos os cálculos para o caso do ruído sobre os dígitos do padrão de entrada, medido pelo parâmetro γ. Um resultado interessante é que nos casos de Hebb e da pseudo-inversa existe uma relação direta entre esses dois parâmetros, a saber

$$\gamma = 1 - 2\chi. \tag{4.76}$$

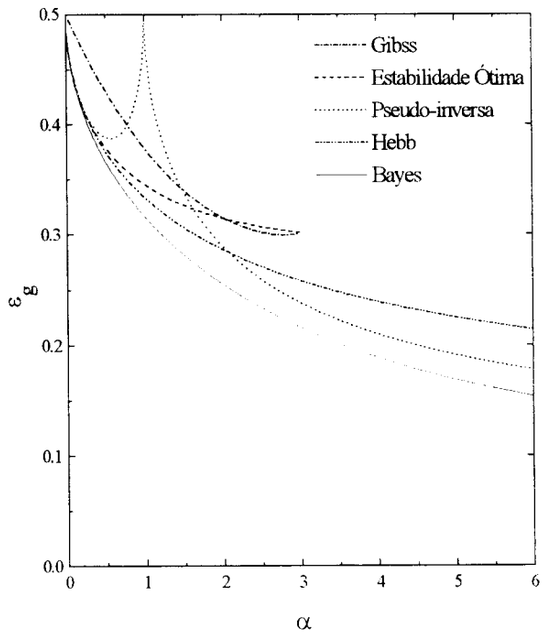


Figura 4.4: Erro de generalização em função de α para $\kappa = 0.5$ e $\chi = 0.1$

Em outras palavras, as equações para as densidades de energia livre calculadas com os dois tipos de ruído tornam-se idênticas utilizando-se a relação acima. Por outro lado não encontramos nenhuma relação entre γ e χ nos outros algoritmos, embora o comportamento qualitativo do erro de generalização sob efeito de γ ou χ seja o mesmo.

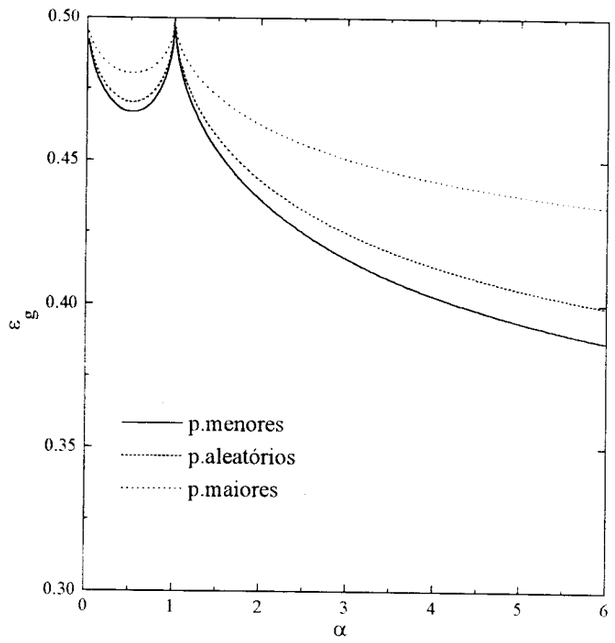


Figura 4.5: Erro de generalização em função de α para $\chi = 0.25$ e $\kappa = 0.8$, para a regra pseudo-inversa

Capítulo 5

Conclusão

O tema principal desta tese foi o estudo dos efeitos da diluição ou lesões na capacidade de aprendizado, medida pelos erros de treinamento e generalização, em redes neurais sem retro-alimentação. As redes neurais escolhidas para essa investigação foram as mais simples possíveis, a saber, o perceptron linear e o perceptron Booleano sem camadas neurais ocultas, a fim de possibilitar uma abordagem completamente analítica a esse projeto. O estudo foi realizado dentro do formalismo da Mecânica Estatística desenvolvido por Elizabeth Gardner para investigar o espaço dos pesos de redes neurais [16],[17].

Há várias motivações para o estudo de diluição ou lesões em redes neurais artificiais. Do ponto de vista biológico, acredita-se que a comparação entre os padrões de comportamento de redes neurais artificiais diluídas e redes neurais biológicas lesionadas venha a esclarecer alguns aspectos do funcionamento do cérebro [24], [41]. Por outro lado, do ponto de vista prático, o estudo de diluição permite-nos determinar a robusteza de uma dada arquitetura ou algoritmo frente a destruição de alguns de seus elementos. Em particular, tal estudo facilitaria a identificação dos elementos mais importantes da rede neural, cuja ausência afetaria mais severamente o desempenho do sistema.

Além dessas motivações, o estudo da diluição após o término do processo de aprendizado é útil para analisar de forma quantitativa o tradicional procedimento de 'poda', ou seja, a eliminação dos pesos de menor intensidade de uma rede neural a fim de melhorar sua capacidade de generalização (e também de reduzir a memória necessária para armazenar os pesos da rede neural). De fato, o resultado principal de nossa investigação

está diretamente relacionado com esse fato de grande interesse prático. Embora na lite ratura 'poda ' seja um termo que designa o corte dos pesos menores, aqui vamos usá-lo num sentido mais amplo, para designar os 3 tipos de diluição após o aprendizado. Mostramos que no caso do perceptron Booleano a diluição sempre deteriora a capacidade de gene ralização da rede, independentemente de haver ou não ruído no processo de aprendizado. Já no caso do perceptron linear, dependendo do valor do parâmetro que mede o tamanho do conjunto de treinamento (α), a 'poda' pode reduzir significativamente o erro de gene ralização no caso de aprendizado com ruído(veja a figura (3.37)). Quanto maior for a intensidade do ruído, maior será a região onde a 'poda' é benéfica. É importante notar que no caso de aprendizado sem ruído, a 'poda' é sempre prejudicial ao desempenho da rede neural (veja a figura (3.36)).

Apesar dessa discrepância aparente entre os padrões de comportamento dos perceptrons linear e Booleano, a diluição após o término do treinamento tem o mesmo efeito sobre os parâmetros de ordem P (norma quadrada dos pesos da rede estudante diluída) e S (correlação entre as redes estudante e mestre) de ambas redes neurais. De fato, esses parâmetros sofrem apenas uma re-escala, $P = \Lambda_\kappa\ Q$ e $S = \Lambda_\kappa\ R$ com $\Lambda_\kappa \leq 1$ (veja equações (4.28) e (4.30)). Esse resultado é bastante geral, dependendo somente do fato da energia de treinamento ser uma função explícita das estabilidades Δ^k dos exemplos de treinamento apenas, o que é satisfeito pelos dois tipos de perceptrons. A origem da diferença entre os perceptrons linear e Booleano está na forma da dependência funcional do erro de generalização com os parâmetros P e S. Em particular,

$$\epsilon_g^B = \arccos\left(\frac{\sqrt{\Lambda_\kappa} R^B}{\sqrt{Q^B}}\right) \tag{5.1}$$

e

$$\epsilon_g^L / M = 1 + \Lambda_\kappa \left(Q^L - 2R^L\right)/M \tag{5.2}$$

para os perceptros Booleano e linear, respectivamente. Uma vez que $\Lambda_\kappa \leq 1$ fica claro então que a diluição após o treinamento é sempre prejudicial no caso Booleano, mas pode ser benéfica no caso linear se $Q^L - 2R^L > 0$. De fato, esta condição pode ser satisfeita no caso do aprendizado com ruído. Concluímos assim que a forma da função de transferência determina o efeito da diluição sobre o erro de generalização.

Bibliografia

[1] de Almeida J R e Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983.

[2] Barbato D M L Fontanari J F 1993 *J. Phys. A: Math. Gen.* **26** 1847.

[3] Barbato D M L Fontanari J F 1995 *Phys. Rev. E* **51** 6219

[4] Binder K. e Young A. P. 1986 *Rev. Mod. Phys.* **58** 801.

[5] Bollé D. e van Mourik J. 1994 *J. Phys. A: Math. Gen.* **27** 1151.

[6] Bouten M. Komoda A e Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 2605.

[7] Bouten M. Engel A. Komoda A e Serneels R 1990 *J. Phys. A: Math. Gen.* **23** 4643.

[8] Brunak S e Lautrup B *Neural Networks :Computers with Intuition* (World Scientific,1990).

[9] Colman H., Nabekura J. e Lichtman J. W. 1997 *Science* **275** 356.

[10] Derrida B., Gardner E., Zippelius A. 1987. *Europhys. Lett.* **4** 167.

[11] Diederich S e Opper M. 1987 *Phys. Rev. Lett.* **58** 949.

[12] Duda R O e Hart P E 1973 *Pattern Classification and Science Analysis* (New York: Wiley).

[13] Fontanari J. F. 1993 *J. Phys. A: Math. Gen.* **26** 6147.

[14] Garcés R., Kuhlmann P. e Eißfeller H. 1992 *J. Phys. A: Math. Gen.* **25** L335.

[15] Gardner E. 1988 *J. Phys. A: Math. Gen.* **21** 257.

[16] Gardner E. 1989 *J. Phys. A: Math. Gen.* **22** 1969.

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

[18] Gardner E. e Derrida B. 1989 *J. Phys. A: Math. Gen.* **22** 1983.

[19] Griniasty M. e Gutfreund H. 1991 *J. Phys. A: Math. Gen.* **24** 715.

[20] Gyorgyi G. e Tishby N., em *Neural Networks and Spin Glasses* eds. W K Theumann e R. Koberle (World Scientific, Singapore. 1990), pp. 3-36.

[21] Hebb D. O. 1949, *The Organization of Behaviour*, Willey, New York.

[22] Hertz J. A., Krogh A e Palmer R.G. *Introduction to the Theory of Neural Computation* Lecture Notes Volume 1, Addison-Wesley Publishing Company, pp139-141.

[23] Hertz J. A., Krogh A. e Thorbergsson 1989 *J. Phys. A: Math. Gen.* **22** 2133.

[24] Hinton G. E., Plaut D. C. e Shallice T. 1993 *Sci. Am.* **269** 76.

[25] Kinouchi O. e Caticha N. 1996 *Phys. Rev. E* **24** R54.

[26] Kinouchi O. e Caticha N. 1995 *Phys. Rev. E* **52** 2878.

[27] Kinouchi O. 1992 Generalização ótima em perceptrons. Dissertação de Mestrado, IFQSC-USP.

[28] Kinouchi O. 1996 Aprendizagem ótima em perceptrons a partir de exemplos com ruído, Tese de Doutorado, IFUSP-USP.

[29] Kohonen T *Self-Organization and Associative Memory* (Springer-Verlag, Berlim,1984)

[30] Krauth W e Mezard M 1987 *J. Phys. A: Math. Gen.* **20** L745.

[31] Kuhlmann P., Garcés R. e Eiβfeller H. 1992 *J. Phys. A: Math. Gen.* **25** L593.

[32] Kuhlmann P. e Muller K. R. 1994 *J. Phys. A: Math. Gen.* **27** 3759.

[33] Le Cun Y., Denker J.S.e Solla S.A. 1990 em *Advances in Neural Information Processing Systems II* (Denver 1989), ed. D.S. Touretzky, (San Mateo: Morgan Kufmann) 598-605.

[34] Levin E. Tishby N. e Solla S. A. 1990 *Proc. IEEE* **78** 1568.

[35] McCulloch W. S. e Pitts W. 1943 *Bull. Math. Biophys.* **5** 115.

[36] Meir R e Fontanari J F 1992 *Phys. Rev. A* **45** 8874.

[37] Nicoll R. A. e Malenka R. C. 1995 *Nature* **377** 115.

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

[39] Opper M e Haussler D 1991 *Phys. Rev. Lett* **66** 2677.

[40] Opper M e Haussler D 1991 em *Proceedings of the IVth Annual Workshop on Computation Learning Theory* (San Mateo: Morgan Kufmann).

[41] Plaut D.C. e Shallice T. 1993 *Cognitive Neuropsych.* **10** 377.

[42] Purves D. e Lichtman J. W. 1980 *Science* **210** 153.

[43] Rumelhart D.E., McClelland J.L. *Parallel Distributed Processing* (MIT Press, 1986).

[44] Rosenblatt F. 1962 *Principles of Neurodynamics.* Spartan. Washington DC.

[45] Seung S. Sompolinsky e Tishby N. 1992 *Phys. Rev. A* **45** 6056.

[46] Sompolinsky H 1986 *Phys. Rev. A* **34** 2571.

[47] Vallet F 1989 *Europhys. Lett.* **8** 747.

[48] Virasoro M.A. 1988 *Europhys. Lett.* **7** 293.

[49] Watikin T L H 1993 *Europhys. Lett.* **21** 871.

[50] Watikin T L H . Rau A e Biehl M 1993 *Rev. Mod. Phys.* **65** 499.

[51] Wong K.Y.M. e Bouten M. 1991 *Europhys. Lett.* **16** 525.

Apêndice A

Cálculo da Energia Livre

Neste apêndice desenvolveremos o cálculo da energia livre para as diluições móvel e fixa e para o corte dos pesos menores e corte aleatório dos pesos.

A.1 Diluição Móvel

Para facilitar as médias sobre o conjunto de treinamento devemos introduzir as identidades

$$\int_{\infty}^{\infty} \prod_{l=1}^{P} dy^l \delta\left(y^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^0 \xi_i^l\right) = 1 \tag{A.1}$$

e

$$\int_{\infty}^{\infty} \prod_{l,a} dx_a^l \delta\left(x_a^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} c_i^a W_i^0 S_i^l\right) = 1. \tag{A.2}$$

Usando a representação integral da função delta de Dirac,

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

escrevemos

$$\langle\langle [Z(\beta, \mathbf{S}, t)]^n \rangle\rangle = \left\langle\left\langle \sum_{\{c_i^a\}} \delta_{Kr}\left(\sum_{i,a} c_i^a, \kappa N\right) \prod_a \delta\left(QN - \sum_i c_i^a (W_i^a)^2\right)\right.\right.$$

$$\left.\left. \delta\left(QN - \sum_i (1 - c_i^a)(W_i^a)^2\right) \int \prod_{i,a} dW_i^a \int \prod_l \frac{dy_l d\hat{y}_l}{2\pi}\right.\right.$$

$$\int \prod_{l,a} \frac{dx_l^a d\hat{x}_l^a}{2\pi} \exp\left\{ -\frac{i}{\sqrt{N}} \sum_{l,j} \left(\hat{y}_l J_i^0 \xi_i^l + \sum_a \hat{x}_l^a c_i^a W_i^a S_i^l \right) \right.$$

$$\left. + i \sum_l \left(y_l \hat{y}_l + \sum_a x_l^a \hat{x}_l^a \right) - \beta \sum_{l,a} (y_l - x_l^a)^2 \right\} \Big\rangle \Big\rangle \tag{A.4}$$

Devemos agora calcular as médias sobre os termos que dependem de ξ_i^l, S_i^l :

$$\mathcal{M}_{j,l} = \int d\xi_i^l \int dS_i^l P\left(\xi_i^l\right) P\left(S_i^l \mid \xi_i^l\right) \exp\left\{ \frac{-i}{\sqrt{N}} \hat{y}_l J_i^0 \xi_i^l - \frac{-i}{\sqrt{N}} \sum_a \hat{x}_l^a c_i^a W_i^a S_i^l \right\}$$

$$= \left(\frac{1+\gamma}{2}\right) \cos\left(\frac{1}{\sqrt{N}} \sum_a \hat{x}_l^a c_i^a W_i^a + \frac{1}{\sqrt{N}} \hat{y}_l J_i^0 \right)$$

$$+ \left(\frac{1-\gamma}{2}\right) \cos\left(\frac{1}{\sqrt{N}} \sum_a \hat{x}_l^a c_i^a W_i^a - \frac{1}{\sqrt{N}} \hat{y}_l J_i^0 \right). \tag{A.5}$$

No limite $N \to \infty$ expandimos os argumentos dos cossenos mantendo apenas os termos dominantes:

$$\prod_{j,l} \mathcal{M}_{j,l} \approx \exp -\frac{1}{2N} \sum_{j,l} \left\{ \left(\hat{y}_l J_i^0 \right)^2 + 2\gamma \hat{y}_l J_i^0 \sum_a \hat{x}_l^a c_i^a W_i^a + \left(\sum_a \hat{x}_l^a c_i^a W_i^a \right)^2 \right\}. \tag{A.6}$$

Os parâmetros de ordem são introduzidos por meio das identidades:

$$\int \prod_{a<b} dq_{ab} \delta\left(q_{ab} - \frac{1}{N} \sum_{i=1}^{N} c_i^a W_i^a c_i^b W_i^b \right) = 1 \tag{A.7}$$

$$\int \prod_a^n dR_a \delta\left(R_a - \frac{1}{N} \sum_{i=1}^{N} c_i^a W_i^a J_i^a \right) = 1. \tag{A.8}$$

Portanto a equação (A.4) é reescrita como

$$\langle\langle (Z(\beta, \mathbf{S}, t))^n \rangle\rangle = \int_{-i\pi}^{i\pi} \prod_a \frac{d\hat{c}_a}{2\pi i/N} \int \prod_a \frac{d\hat{Q}_a}{2\pi i/N} \int \prod_a \frac{d\hat{Q}_a^0}{2\pi i/N} \int \prod_a \frac{d\hat{R}_a R_a}{2\pi i/N}$$

$$\int \prod_{a<b} \frac{d\hat{q}_{ab} q_{ab}}{2\pi i/N} \exp N \left\{ G_0\left(q_{ab}, \hat{q}_{ab}, R_a, \hat{R}_a, \hat{c}_a, Q^0, \hat{Q}_a^0, \hat{Q}_a \right) \right.$$

$$\left. + \alpha G_1\left(q_{ab}, R_a\right) + G_2\left(\hat{q}_{ab}, \hat{c}_a, \hat{Q}_a, \hat{Q}_a^0, \hat{R}_a \right) \right\} \tag{A.9}$$

onde

$$G_0 = \kappa \sum_a \hat{c}_a - Q \sum_a \hat{Q}_a + Q^0 \sum_a \hat{Q}_a^0 - \sum_a \hat{R}_a R_a - \sum_{a<b} \hat{q}_{ab} q_{ab}, \tag{A.10}$$

$$G_1 = \ln \int \prod_{a=1}^{n} \frac{dx_a}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \sum_a^n x_a^2 \left[1 + \beta \left(Q + M - 2\gamma R_a \right) \right] \right.$$

$$\left. -\beta \sum_{a<b}^{n} x_a x_b \left(q_{ab} + M - 2\gamma R_a \right) \right\}. \tag{A.11}$$

e

$$G_2 = \ln \sum_{\{c^a=0,1\}} \int \prod_{a=1}^{n} dW^a \exp \left\{ -\sum_a^n \left[\hat{c}_a c^a + \hat{Q}_a^0 \left(1 - c^a \right) \left(W^a \right)^2 - \hat{Q}_a c^a \left(W^a \right)^2 \right] \right.$$

$$\left. + \sum_a^n c^a \hat{R}_a W^a J^0 + \sum_{a<b}^{n} \hat{q}_{ab} c^a c^b W^a W^b \right\}. \tag{A.12}$$

As integrais sobre os parâmetros de ordem são efetuadas, no limite $N \to \infty$, pelo método de integração por ponto de sela:

$$\mathcal{I}(N) = \int e^{NF(t)} dt = \approx \mathcal{C} e^{NF(t_0)} \tag{A.13}$$

onde t_0 é determinado pela equação $F'(t_0) = 0$ e $\mathcal{C}$ é da ordem $N^{-1/2}$. Como queremos calcular a densidade de energia livre, no limite termodinâmico apenas o argumento da exponencial contribui de forma que podemos desprezar o prefator $\mathcal{C}$. O cálculo explícito de G_1 e G_2 na prescrição simetria de réplicas encontra-se nos Apêndices (B.1) e (C.1), respectivamente.

A.2 Diluição Fixa

As médias sobre o conjunto de treinamento são efetuadas como no caso da diluição móvel. Assim o n-ésimo momento de Z é escrito como

$$\langle\langle \left(Z(\beta, \mathbf{S}, t) \right)^n \rangle\rangle = \int \prod_a \frac{d\hat{Q}_a}{2\pi i/N} \int \prod_a \frac{d\hat{R}_a R_a}{2\pi i/N} \int \prod_{a<b} \frac{d\hat{q}_{ab} q_{ab}}{2\pi i/N}$$

$$\exp N \left\{ G_0 \left(\hat{Q}, \hat{R}_a, R_a, \hat{q}_{ab}, q_{ab} \right) \right.$$

$$+\alpha G_1\left(q_{ab}, R_a\right) + G_2\left(\hat{q}_{ab}, \hat{c}_a, \hat{Q}_a, \hat{Q}_a^0, \hat{R}_a\right)\Big\} \tag{A.14}$$

onde

$$G_0 = -Q\sum_a \hat{Q}_a - \sum_a \hat{R}_a R_a - \sum_{a<b} \hat{q}_{ab} q_{ab}. \tag{A.15}$$

G_1 tem a mesma expressão que no caso da diluição móvel, equação (A.11), e

$$G_2 = \ln \int \prod_{a=1}^{n} dJ^a \exp\left\{\sum_a^n \hat{Q}_a \left(J^a\right)^2 \right.$$

$$\left. + \sum_a^n \hat{R}_a J^a J^0 + \sum_{a<b}^n \hat{q}_{ab} J^a J^b \right\}. \tag{A.16}$$

O cálculo de G_2 está no Apêndice C na seção (C.2).

A.3 Corte dos pesos menores

Neste caso, para realizar as médias sobre o conjunto de treinamento devemos introduzir as identidades

$$\int_{-\infty}^{\infty} \prod_{l=1}^{P} dt^l \delta\left(t^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^0 \xi_i^l\right) = 1. \tag{A.17}$$

$$\int_{-\infty}^{\infty} \prod_{l,a} dy_a^l \delta\left(y_a^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^a S_i^l\right) = 1 \tag{A.18}$$

$$\int_{-\infty}^{\infty} \prod_{l,a} dx_a^l \delta\left(x_a^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} S_i^l J_i^a \Theta\left(|J_i^a| - \omega\right)\right) = 1 \tag{A.19}$$

Usando a representação integral da função delta de Dirac, equação (A.3), escrevemos

$$\langle\langle [Z(\beta, \mathbf{S}.t)]^n \rangle\rangle = \left\langle\left\langle \prod_a \mathrm{Tr} \int \prod_l \frac{dt_l d\hat{t}_l}{2\pi} \int \prod_{l,a} \frac{dy_l^a d\hat{y}_l^a}{2\pi} \int \prod_{l,a} \frac{dx_l^a d\hat{x}_l^a}{2\pi} \right.\right.$$

$$\times \exp\left\{ i \sum_l \left(t_l \hat{t}_l + \sum_a y_l^a \hat{y}_l^a + \sum_a x_l^a \hat{x}_l^a \right) \right.$$

$$- \frac{i}{\sqrt{N}} \sum_{l,j} \left(\sum_a \hat{y}_l^a J_i^a \xi_i^l + \sum_a \hat{x}_l^a W_i^a S_i^l + \hat{t}_l J_i^0 \xi_i^l \right)$$

$$\left.\left. - \frac{\beta}{2} \sum_{l,a} \left(t_l - y_l^a\right)^2 - \frac{\beta h}{2} \sum_{l,a} \left(t_l - x_l^a\right)^2 \right\} \right\rangle\Big\rangle \tag{A.20}$$

com

$$W_i^a = J_i^a \Theta\left(|J_i^a| - \omega\right).$$

Devemos calcular as médias sobre os termos que dependem de ξ_i^l, S_i^l :

$$\mathcal{M}_{j,l} = \int d\xi_i^l \int dS_i^l P\left(\xi_i^l\right) P\left(S_i^l \mid \xi_i^l\right) \exp\left\{-\frac{i}{\sqrt{N}}\hat{t}_l J_i^0 \xi_i^l\right.$$

$$\left. -\frac{i}{\sqrt{N}}\sum_a \hat{y}_l^a J_i^a S_i^l - \frac{i}{\sqrt{N}}\sum_a \hat{x}_l^a W_i^a S_i^l\right\}, \tag{A.21}$$

que leva a

$$\mathcal{M}_{j,l} = \left(\frac{1+\gamma}{2}\right)\cos\left(\frac{\sum_a \hat{y}_l^a J_i^a}{\sqrt{N}} + \frac{\sum_a \hat{x}_l^a W_i^a}{\sqrt{N}} + \frac{\hat{t}_l J_i^0}{\sqrt{N}}\right) +$$

$$\left(\frac{1-\gamma}{2}\right)\cos\left(\frac{\sum_a \hat{y}_l^a J_i^a}{\sqrt{N}} + \frac{\sum_a \hat{x}_l^a W_i^a}{\sqrt{N}} - \frac{\hat{t}_l J_i^0}{\sqrt{N}}\right). \tag{A.22}$$

No limite $N \to \infty$, expandimos os argumentos dos cossenos mantendo apenas os termos dominantes:

$$\prod_{j,l}\mathcal{M}_{j,l} = \exp\left\{-\frac{1}{N}\sum_{j,l}\left[\sum_a \hat{y}_l^a J_i^a \sum_a \hat{x}_l^a W_i^a + \gamma \hat{t}_l J_i^0 \sum_a \left(\hat{y}_l^a J_i^a + \hat{x}_l^a W_i^a\right)\right.\right.$$

$$\left.\left. \frac{\left(\hat{t}_l J_i^0\right)^2}{2} + \frac{\left(\sum_a \hat{y}_l^a J_i^a\right)^2}{2} + \frac{\left(\sum_a \hat{x}_l^a W_i^a\right)^2}{2}\right]\right\}. \tag{A.23}$$

Os parâmetros de ordem $q_{ab}, Q_a, p_{ab}, P_a, r_{ab}, R_a, S_a$ são introduzidos por meio das identidades:

$$\int \prod_{a<b} dq_{ab}\delta\left(q_{ab} - \frac{1}{N}\sum_{i=1}^N J_i^a J_i^b\right) = 1, \tag{A.24}$$

$$\int \prod_a^n dQ_a\delta\left(Q_a - \frac{1}{N}\sum_{i=1}^N (J_i^a)^2\right) = 1, \tag{A.25}$$

$$\int \prod_{a<b} dp_{ab}\delta\left(p_{ab} - \frac{1}{N}\sum_{i=1}^N W_i^a W_i^b\right) = 1, \tag{A.26}$$

$$\int \prod_a^n dP_a\delta\left(P_a - \frac{1}{N}\sum_{i=1}^N (W_i^a)^2\right) = 1, \tag{A.27}$$

$$\int \prod_{a<b} dr_{ab} \delta \left(r_{ab} - \frac{1}{N} \sum_{i=1}^{N} J_i^a W_i^b \right) = 1, \tag{A.28}$$

$$\int \prod_{a}^{n} dR_a \delta \left(R_a - \frac{1}{N} \sum_{i=1}^{N} J_i^0 J_i^a \right) = 1, \tag{A.29}$$

e

$$\int \prod_{a}^{n} dS_a \delta \left(S_a - \frac{1}{N} \sum_{i=1}^{N} J_i^0 W_i^a \right) = 1. \tag{A.30}$$

Usando a representação integral da delta de Dirac, reescrevemos (A.20) como

$$\begin{aligned} \langle\langle [Z(\beta, \mathbf{S}, t)]^n \rangle\rangle &= \int \prod_a \frac{d\hat{Q}_a Q_a}{2\pi i/N} \int \prod_a \frac{d\hat{R}_a R_a}{2\pi i/N} \int \prod_a \frac{d\hat{P}_a dP_a}{2\pi i/N} \int \prod_a \frac{d\hat{S}_a dS_a}{2\pi i/N} \\ & \int \prod_{a<b} \frac{d\hat{q}_{ab} q_{ab}}{2\pi i/N} \int \prod_{a<b} \frac{d\hat{r}_{ab} dr_{ab}}{2\pi i/N} \int \prod_{a<b} \frac{d\hat{p}_{ab} dp_{ab}}{2\pi i/N} \\ & \exp\left[N\left(G_0 + \alpha G_1 + G_2\right)\right] \end{aligned} \tag{A.31}$$

onde

$$\begin{aligned} G_0 &= \sum_a Q^a \hat{Q}_a + \sum_a \hat{R}_a R_a + \sum_a \hat{P}_a P_a + \sum_a \hat{S}_a S_a + \\ & \sum_{a<b} \hat{q}_{ab} q_{ab} + \sum_{a<b} \hat{r}_{ab} r_{ab} + \sum_{a<b} \hat{p}_{ab} p_{ab}, \end{aligned} \tag{A.32}$$

$$\begin{aligned} G_2 &= \ln \prod_a \mathrm{Tr} \exp \left\{ - \sum_a^n \left[\hat{R}_a J^a J^0 + \hat{Q}_a \left(J^a\right)^2 + \hat{S}_a J^0 W^a + \hat{P}_a J^a W^a \right] \right. \\ & \left. - \sum_{a<b}^n \hat{q}_{ab} J^a J^b - \sum_{a<b}^n \hat{p}_{ab} W^a W^b - \sum_{a \neq b}^n \hat{r}_{ab} J^a W^b \right\}, \end{aligned} \tag{A.33}$$

e

$$\begin{aligned} G_1 &= \ln \prod_{a=1}^n \int \frac{dy_a d\hat{y}_a}{2\pi} \int \frac{dx_a d\hat{x}_a}{2\pi} \int \frac{dt d\hat{t}}{2\pi} e^{-\frac{M\hat{t}^2}{2} + it\hat{t}} \exp \left\{ \sum_a^n \left[i y_a \hat{y}_a + \right. \right. \\ & i x_a \hat{x}_a - \frac{\beta}{2} (t - y_a)^2 - \frac{\beta h}{2} (t - x_a)^2 - \gamma \hat{t} \left(\hat{y}_a R_a + \hat{x}_a S_a \right) - \hat{x}_a \hat{y}_a P_a \\ & \left. \left. - \frac{1}{2} \hat{y}_a^2 Q_a - \frac{1}{2} \hat{x}_a^2 P_a \right] - \sum_{a \neq b}^n \left[\hat{y}_a \hat{x}_b r_{ab} + \frac{1}{2} \hat{y}_a \hat{y}_b q_{ab} + \frac{1}{2} \hat{x}_a \hat{x}_b p_{ab} \right] \right\} \end{aligned} \tag{A.34}$$

Os cálculos de G_1 e G_2 estão nos apêndices (B.3) e (C.3) respectivamente.

A.4 Corte aleatório dos pesos

Para facilitar as médias sobre o conjunto de treinamento devemos introduzir as identidades

$$\int_{\infty}^{\infty} \prod_{l=1}^{P} dt^l \delta\left(t^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^0 \xi_i^l\right) = 1, \tag{A.35}$$

$$\int_{\infty}^{\infty} \prod_{l,a} dy_a^l \delta\left(y_a^l - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_i^a S_i^l\right) = 1 \tag{A.36}$$

Usando a representação integral da função delta de Dirac, equação (A.3) escrevemos

$$\begin{aligned}
\langle\langle (Z(\beta, \mathbf{S}, t))^n \rangle\rangle \;=\; & \left\langle \left\langle \prod_a \mathrm{Tr} \int \prod_l \frac{dt_l d\hat{t}_l}{2\pi} \int \prod_{l,a} \frac{dy_l^a d\hat{y}_l^a}{2\pi} \right.\right. \\
& \exp\left\{ +i \sum_l \left(t_l \hat{t}_l + \sum_a y_l^a \hat{y}_l^a \right) - \frac{\beta h \kappa (1-\kappa)\alpha}{2} \sum_{i,a} (J_i^a)^2 \right. \\
& -\frac{\beta}{2} \sum_{l,a} (t_l - y_l^a)^2 - \frac{\beta h \kappa^2}{2} \sum_{l,a} \left(\frac{t_l}{\kappa} - y_l^a \right)^2 \\
& \left.\left. -\frac{i}{\sqrt{N}} \sum_{l,i} \left(\sum_a \hat{y}_l^a J_i^a S_i^l + + \hat{t}_l J_i^0 \xi_i^l \right) \right\} \right\rangle \Big\rangle .
\end{aligned} \tag{A.37}$$

Devemos calcular as médias sobre os termos que dependem de ξ_i^l, S_i^l :

$$\mathcal{M}_{j,l} = \int d\xi_i^l \int dS_i^l P\left(\xi_i^l\right) P\left(S_i^l \mid \xi_i^l\right) \exp\left[-\frac{i}{\sqrt{N}} \left(\hat{t}_l J_i^0 \xi_i^l + \sum_a \hat{y}_l^a J_i^a S_i^l \right) \right]. \tag{A.38}$$

Efetuando estas médias obtemos

$$\begin{aligned}
\mathcal{M}_{j,l} \;=\; & \left(\frac{1+\gamma}{2} \right) \cos\left(\frac{\sum_a \hat{y}_l^a J_i^a}{\sqrt{N}} + \frac{\hat{t}_l J_i^0}{\sqrt{N}} \right) + \\
& \left(\frac{1-\gamma}{2} \right) \cos\left(\frac{\sum_a \hat{y}_l^a J_i^a}{\sqrt{N}} + -\frac{\hat{t}_l J_i^0}{\sqrt{N}} \right).
\end{aligned} \tag{A.39}$$

No limite $N \to \infty$ expandimos os argumentos dos cossenos mantendo apenas os termos dominantes:

$$\prod_{j,l} \mathcal{M}_{j,l} \approx \exp\left\{-\frac{1}{N}\sum_{j,l}\left[\gamma \hat{t}_l J_i^0 \sum_a \left(\hat{y}_l^a J_i^a\right) + \frac{\left(\hat{t}_l J_i^0\right)^2}{2} + \frac{\left(\sum_a \hat{y}_l^a J_i^a\right)^2}{2}\right]\right\}. \tag{A.40}$$

Os parâmetros de ordem, q_{ab}, Q_a, R_a são os mesmos da seção anterior. Assim, a expressão do $n-$ésimo momento de Z é escrita como

$$\begin{aligned}\langle\langle (Z(\beta, \mathbf{S}, t))^n \rangle\rangle &= \int \prod_a \frac{d\hat{Q}_a Q_a}{2\pi i/N} \int \prod_a \frac{d\hat{R}_a R_a}{2\pi i/N} \int \prod_{a<b} \frac{d\hat{q}_{ab} q_{ab}}{2\pi i/N} \\ &\exp\left\{G_0 + \alpha G_1 + G_2\right\}\end{aligned} \tag{A.41}$$

onde

$$G_0 = \sum_a Q^a \hat{Q}_a + \sum_a \hat{R}_a R_a + \sum_{a<b} \hat{q}_{ab} q_{ab}, \tag{A.42}$$

$$\begin{aligned}G_2 &= \ln \prod_a \mathrm{Tr} \exp\left\{-\sum_a^n \left[\hat{R}_a J^a J^0 + \hat{Q}_a \left(J^a\right)^2\right]\right. \\ &\left. -\sum_{a<b}^n \hat{q}_{ab} J^a J^b - \frac{\beta h \kappa (1-\kappa)\alpha}{2} \left(J^a\right)^2\right\}\end{aligned} \tag{A.43}$$

e

$$\begin{aligned}G_1 &= \ln \prod_{a=1}^n \int \frac{dy_a d\hat{y}_a}{2\pi} \int \frac{dt d\hat{t}}{2\pi} e^{-\frac{M\hat{t}^2}{2} + it\hat{t}} \exp\left\{\sum_a^n \left[i y_a \hat{y}_a + \right.\right. \\ &-\frac{\beta}{2}(t - y_a)^2 - \frac{\beta h \kappa^2}{2}\left(\frac{t}{\kappa} - y_a\right)^2 - \gamma \hat{t}\left(\hat{y}_a R_a\right) \\ &\left.\left. -\frac{1}{2}\hat{y}_a^2 Q_a\right] - \sum_{a<b}^n \left[\hat{y}_a \hat{y}_b q_{ab}\right]\right\}.\end{aligned} \tag{A.44}$$

O cálculo explícito de G_1 e de G_2 usando a prescrição simetria de réplicas é apresentado nos apêndices (B.4) e (C.4). respectivamente.

Apêndice B

Cálculo de G_1 com simetria de réplicas

B.1 Diluição móvel

Usando a prescrição simetria de réplicas e rearranjando os termos na expressão (3.15) temos

$$G_1^{sr} = \ln \int \prod_{a=1}^{n} \frac{dx_a}{\sqrt{2\pi}} \exp\left\{ -\frac{\beta}{2} \left(\sum_a^n x_a \right)^2 (q + M - 2\gamma R) - \frac{1}{2} \sum_a^n x_a^2 \left[1 + \beta (Q - q) \right] \right\}. \tag{B.1}$$

Aplicando a transformação gaussiana

$$e^{-x^2/2} = \int_{-\infty}^{\infty} Dz e^{-ixz} \tag{B.2}$$

para desacoplar as réplicas no argumento da exponencial e tomando o limite $n \to 0$ obtemos

$$\begin{aligned} G_1^{sr} &= \ln \int Dz \Lambda^n \\ &\approx n \int Dz \ln \Lambda \end{aligned} \tag{B.3}$$

com

$$\Lambda = \int \frac{dx}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left[1 + \beta (Q - q) \right] x^2 - iz\sqrt{\beta (q + M - 2\gamma R)}x \right\}. \tag{B.4}$$

Efetuando a integral acima obtemos

$$G_1^{sr} \approx n\left\{-\frac{1}{2}\ln\left[1+\beta\left(Q-q\right)\right]-\frac{\beta\left(q+M-2\gamma R\right)}{2\left[1+\beta\left(Q-q\right)\right]}\right\} \tag{B.5}$$

B.2 Corte dos pesos menores

Usando a prescrição simetria de réplicas e rearranjando os termos na expressão (A.34) temos

$$\begin{aligned} G_1^{sr} \;=\; & \ln\prod_{a=1}^{n}\int\frac{dy_a d\hat{y}_a}{2\pi}\int\frac{dx_a d\hat{x}_a}{2\pi}\int\frac{dtd\hat{t}}{2\pi}e^{-\frac{Mt^2}{2}+it\hat{t}} \\ & \exp\sum_a^n\left[iy_a\hat{y}_a+ix_a\hat{x}_a-\frac{\beta}{2}\left(t-y_a\right)^2-\frac{\beta h}{2}\left(t-x_a\right)^2\right] \\ & \exp\left\{-\gamma\hat{t}\left[R\sum_a^n\hat{y}_a+S\sum_a^n\hat{x}_a\right]-\frac{(Q-q)}{2}\sum_a^n\hat{y}_a^2-\frac{q}{2}\left(\sum_a^n\hat{y}_a\right)^2\right. \\ & -\frac{(P-p)}{2}\sum_a^n\hat{x}_a^2-P\sum_a^n\hat{y}_a\hat{x}_a+\frac{p}{2}\left(\sum_a^n\hat{x}_a\right)^2 \\ & \left.-r\sum_a^n\hat{y}_a\sum_b^n\hat{x}_b+r\sum_a^n\hat{y}_a\hat{x}_a\right\}. \end{aligned} \tag{B.6}$$

Para facilitarmos o cálculo desta expressão vamos introduzir, através da função delta de Dirac, as variáveis

$$v=\sum_a^n\hat{y}_a \qquad e \qquad u=\sum_b^n\hat{x}_b. \tag{B.7}$$

Efetuando a integral sobre as variáveis $\hat{x}_a, \hat{y}_a, x_a, y_a, u, v$ temos

$$G_1^{sr}=\ln\int\frac{dtd\hat{t}}{2\pi}e^{-\frac{Mt^2}{2}+it\hat{t}}\int\frac{d\hat{u}d\hat{v}}{2\pi\sqrt{pq-r^2}}e^{f\left(\hat{u},\hat{v},\hat{t}\right)}\left[\eta^{-1/2}e^{g(\hat{u},\hat{v},t)}\right]^n \tag{B.8}$$

onde

$$\eta=1+\beta\left\{(Q-q)+h\left(P-p\right)+\beta h\left[\left(Q-q\right)\left(P-p\right)-\left(P-r\right)^2\right]\right\}, \tag{B.9}$$

$$f\left(\hat{u},\hat{v},\hat{t}\right)=\frac{\left(\gamma\hat{t}S-i\hat{u}\right)^2}{2p}+\frac{\left[\gamma\hat{t}\left(Rp-rS\right)+i\left(r\hat{u}-p\hat{v}\right)\right]^2}{2p\left(pq-r^2\right)} \tag{B.10}$$

e

$$g(\hat{u},\hat{v},t) = -\frac{\beta}{2\eta}\left\{h\left[1+\beta(Q-q)(\hat{u}-t)^2\right]\right.$$

$$\left.+\left[1+\beta h(P-p)(\hat{v}-t)^2\right]-2\beta h(P-r)(\hat{u}-t)(\hat{v}-t)\right\}. \quad \text{(B.11)}$$

Tomando o limite $n \longrightarrow 0$ encontramos

$$\frac{G_1^{sr}}{n} = \ln\eta^{-1/2} + \int \frac{dtd\hat{t}}{2\pi} e^{-\frac{M\hat{t}^2}{2}+it\hat{t}} \int \frac{d\hat{v}d\hat{u}}{2\pi\sqrt{pq-r^2}} e^{f(\hat{u},\hat{v},\hat{t})} g(\hat{u},\hat{v},t). \quad \text{(B.12)}$$

Efetuando as integrais sobre $t, \hat{t}, \hat{u}, \hat{v}$ obtemos

$$\frac{G_1^{sr}}{n} = \ln\eta^{-1/2} + (C_1p + C_2q + C_3r) + \frac{(C_1+C_2+C_3)}{p}\left[pM - \left(\gamma^2S^2 + \frac{B^2}{A}\right)\right] +$$

$$\frac{(Br-\gamma SA)}{pA}\left[2(C_1p+C_2r)+C_3(p+r)\right] - \frac{B}{A}\left[2(C_1r+C_2q)+C_3(q+r)\right]$$

$$\frac{(C_1+C_2+C_3)}{pA^2}\left\{(Br-\gamma SA)\left[(Br-\gamma SA)-2Br\right]+B^2pq\right\} \quad \text{(B.13)}$$

onde

$$C_1 = -\frac{\beta h\left[1+\beta(Q-q)\right]}{2\eta}, \quad \text{(B.14)}$$

$$C_2 = -\frac{\beta\left[1+\beta h(P-p)\right]}{2\eta}, \quad \text{(B.15)}$$

$$C_3 = \frac{\beta^2 h(P-r)}{\eta}, \quad \text{(B.16)}$$

$$A = pq - r^2, \quad \text{e} \quad B = \gamma(Rp - rS). \quad \text{(B.17)}$$

Fazendo $h = 0$ encontramos

$$\frac{G_1^{sr}}{n} \approx -\frac{1}{2}\ln\left[1+\beta(Q-q)\right] - \frac{\beta}{2}\frac{(q+M-2\gamma R)}{\left[1+\beta(Q-q)\right]}. \quad \text{(B.18)}$$

B.3 Corte aleatório dos pesos

Usando a prescrição simetria de réplicas e rearranjando os termos na expressão (A.44) obtemos

$$G_1^{sr} = \ln\prod_{a=1}^{n}\int\frac{dy_a d\hat{y}_a}{2\pi}\int\frac{dtd\hat{t}}{2\pi}e^{-\frac{M\hat{t}^2}{2}+it\hat{t}}$$

$$\exp\left\{\sum_a^n iy_a\hat{y}_a - \sum_a^n \frac{\beta}{2}(t-y_a)^2 - \frac{\beta h\kappa^2}{2}\sum_a^n\left(\frac{t}{\kappa}-y_a\right)^2\right.$$

$$\left.-\gamma\hat{t}R\sum_a^n \hat{y}_a - \frac{(Q-q)}{2}\sum_a^n \hat{y}_a^2 - \frac{q}{2}\left(\sum_a^n \hat{y}_a\right)^2\right\}. \tag{B.19}$$

Fazendo

$$v = \sum_a^n \hat{y}_a \tag{B.20}$$

temos

$$G_1^{sr} = \ln\prod_{a=1}^n \int\frac{dy_a d\hat{y}_a}{2\pi}\int\frac{dtd\hat{t}}{2\pi}e^{-\frac{Mi^2}{2}+it\hat{t}}$$

$$\exp\left\{\sum_a^n iy_a\hat{y}_a - \frac{\beta}{2}\sum_a^n(t-y_a)^2 - \frac{\beta h\kappa^2}{2}\sum_a^n\left(\frac{t}{\kappa}-y_a\right)^2\right.$$

$$\left.\exp -\gamma\hat{t}R\sum_a^n \hat{y}_a - \frac{(Q-q)}{2}\sum_a^n \hat{y}_a^2 - i\hat{v}\sum_a^n \hat{y}_a\right\}. \tag{B.21}$$

Integrando em $\hat{y}_a, y_a$ e v e tomando o limite $n \to 0$ reescrevemos G_1^{sr} como

$$\frac{G_1^{sr}}{n} = -\frac{1}{2}\ln\left[1+\beta(Q-q)\left(1+h\kappa^2\right)\right] + \int\frac{d\hat{v}}{2\pi}e^{-\frac{v^2}{2}}\int\frac{dtd\hat{t}}{2\pi}e^{-\frac{Mi^2}{2}+it\hat{t}}$$

$$\left[\frac{2i\gamma R\hat{t}\hat{v}-\hat{v}^2}{2(Q-q)} - \frac{\beta t^2(1+h)}{2} + \frac{\left[\left(\hat{v}-i\gamma R\hat{t}\right)+\beta(Q-q)t(1+h\kappa)\right]^2}{2(Q-q)\left[1+\beta(Q-q)(1+h\kappa^2)\right]}\right] \tag{B.22}$$

Finalmente, integrando em $\hat{v}, \hat{t}, t$ encontramos

$$\frac{G_1}{n} = -\frac{1}{2}\ln\left[1+\beta(Q-q)\left(1+h\kappa^2\right)\right] - \frac{\beta M(1+h)}{2} - \frac{q}{2(Q-q)}$$

$$\frac{q+\beta(Q-q)(1+h\kappa)\left[2R\gamma+M\beta(Q-q)(1+h\kappa)\right]}{2(Q-q)\left[1+\beta(Q-q)(1+h\kappa^2)\right]}. \tag{B.23}$$

Apêndice C

Cálculo de G_2 com simetria de réplicas

C.1 Diluição móvel

Usando a prescrição simetria de réplicas reescrevemos a equação (3.16) como

$$G_2^{sr} = \ln \sum_{\{c^a=0,1\}} \int \prod_{a=1}^{n} dW^a \exp\left\{\sum_a^n \left[-\hat{c}_a c^a + \left(\hat{Q}^0 + \hat{Q} - \frac{\hat{q}}{2}\right) c^a (W^a)^2 \right.\right.$$

$$\left.\left. -\hat{Q}^0 (W^a)^2 + \hat{R}\sum_a^n c^a W^a J^0 + \frac{\hat{q}}{2}\left(\sum_a^n c^a W^a\right)^2\right]\right\}. \qquad \text{(C.1)}$$

Aplicando a transformação gaussiana (B.2) para para desacoplar as réplicas no argumento da exponencial e tomando o limite $n \to 0$ obtemos

$$\begin{aligned} G_2^{sr} &= \ln \int Dz \Lambda^n \\ &\approx n \int Dz \ln \Lambda \end{aligned} \qquad \text{(C.2)}$$

onde

$$\Lambda = \frac{\sqrt{\pi}}{\sqrt{\hat{Q}^0}} + \frac{\sqrt{\pi} e^{-\hat{c}}}{\sqrt{\frac{\hat{q}}{2} - \hat{Q}}} \exp \frac{\left(\sqrt{\hat{q}} z + \hat{R} J^0\right)^2}{4\left(\frac{\hat{q}}{2} - \hat{Q}\right)}. \qquad \text{(C.3)}$$

Daí, G_2^{sr} é dado por

$$G_2^{sr} = n\int Dz \ln\sqrt{\pi}\left\{\frac{1}{\sqrt{\hat{Q}^0}} + \frac{e^{-\hat{c}}}{\sqrt{\frac{\hat{q}}{2}-\hat{Q}}}\exp\frac{\left(\sqrt{\hat{q}}z+\hat{R}J^0\right)^2}{4\left(\frac{\hat{q}}{2}-\hat{Q}\right)}\right\}. \tag{C.4}$$

Mediando sobre J^0 obtemos a a expressão final para G_2^{sr}

$$G_2^{sr} = n\frac{1}{2}\ln\pi + n\int Dz \ln\left\{\frac{1}{\sqrt{\hat{Q}^0}} + \frac{e^{-\hat{c}}}{\sqrt{\frac{\hat{q}}{2}-\hat{Q}}}\exp\frac{\left(\hat{q}+M\hat{R}^2\right)z^2}{4\left(\frac{\hat{q}}{2}-\hat{Q}\right)}\right\}. \tag{C.5}$$

C.2 Diluição fixa

Usando a prescrição simetria de réplicas reescrevemos a equação (3.72) como

$$\begin{aligned} G_2^{sr} &= \ln\int\prod_{a=1}^{n} dJ^a \exp\left\{\sum_a^n\left(\hat{Q}-\frac{\hat{q}}{2}\right)(J^a)^2\right. \\ &\left. +\hat{R}\sum_a^n J^aJ^0 + \frac{\hat{q}}{2}\left(\sum_a^n J^a\right)^2\right\}. \end{aligned} \tag{C.6}$$

Aplicando a transformação gaussiana dada na equação (B.2) e tomando o limite $n \to 0$ obtemos

$$\begin{aligned} G_2^{sr} &= \ln\int Dz\Lambda^n \\ &\approx n\int Dz\ln\Lambda \end{aligned} \tag{C.7}$$

onde

$$\Lambda = \frac{\sqrt{\pi}}{\sqrt{\frac{\hat{q}}{2}-\hat{Q}}}\exp\frac{\left(\sqrt{\hat{q}}z+\hat{R}J^0\right)^2}{4\left(\frac{\hat{q}}{2}-\hat{Q}\right)}. \tag{C.8}$$

Daí,

$$G_2^{sr} = n\int Dz\ln\left\{\frac{\sqrt{\pi}}{\sqrt{\frac{\hat{q}}{2}-\hat{Q}}}\exp\frac{\left(\sqrt{\hat{q}}z+\hat{R}J^0\right)^2}{4\left(\frac{\hat{q}}{2}-\hat{Q}\right)}\right\}. \tag{C.9}$$

Mediando sobre J^0 obtemos a a expressão final para G_0^{sr}

$$G_2^{sr} = n\frac{1}{2}\ln\pi - n\frac{1}{2}\ln\left(\frac{\hat{q}}{2}-\hat{Q}\right) + n\frac{\left(\hat{q}+M\hat{R}^2\right)}{4\left(\frac{\hat{q}}{2}-\hat{Q}\right)}. \tag{C.10}$$

C.3 Corte dos pesos menores

Usando a prescrição simetria de réplicas na equação (A.33) temos

$$G_2^{sr} = \ln \prod_{a=1}^{n} \mathrm{Tr} \exp \left\{ -\sum_a^n J^a \left[\left(\hat{R} J^0 + \hat{S} J^0 \Theta \left(|J^a| - \omega \right) + \hat{Q} + \hat{P} J^a \Theta \left(|J^a| - \omega \right) \right) \right] \right.$$

$$\left. -\hat{q} \sum_{a<b}^{n} J^a J^b - \hat{r} \sum_{a \neq b}^{n} J^a J^b \Theta \left(\left| J^b \right| - \omega \right) - \hat{p} \sum_{a<b}^{n} J^a \Theta \left(|J^a| - \omega \right) J^b \Theta \left(\left| J^b \right| - \omega \right) \right\} \quad \text{(C.11)}$$

ou

$$G_2^{sr} = \ln \prod_{a=1}^{n} \mathrm{Tr} \exp \left\{ -\hat{R} J^0 \sum_a^n J^a - \hat{S} J^0 \sum_a^n J^a \Theta \left(|J^a| - \omega \right) - \left(\hat{Q} - \frac{\hat{q}}{2} \right) \sum_a^n \left(J^a \right)^2 \right.$$

$$-\frac{\hat{q}}{2} \left(\sum_a^n J^a \right)^2 - \left(\hat{P} - \frac{\hat{p}}{2} - \hat{r} \right) \sum_a^n \left(J^a \right)^2 \Theta \left(|J^a| - \omega \right)$$

$$\left. -\frac{\hat{p}}{2} \left(\sum_a^n J^a \Theta \left(|J^a| - \omega \right) \right)^2 - \hat{r} \sum_a^n J^a \sum_b^n J^b \Theta \left(\left| J^b \right| - \omega \right) \right\}. \quad \text{(C.12)}$$

Fazendo

$$y = \sum_a^n J^a \Theta \left(|J^a| - \omega \right) \qquad \text{e} \qquad x = \sum_a^n J^a \quad \text{(C.13)}$$

temos

$$G_2^{sr} = \ln \int \frac{dx d\hat{x}}{2\pi} \int \frac{dy d\hat{y}}{2\pi} \exp \left\{ i x \hat{x} + i y \hat{y} - \frac{\hat{q}}{2} x^2 - \frac{\hat{p}}{2} y^2 - \hat{r} x y - J^0 \hat{R} x + \hat{S} y J^0 \right\}$$

$$\prod_a \mathrm{Tr} \exp \left\{ -\left(\hat{Q} - \frac{\hat{q}}{2} \right) \sum_a^n \left(J^a \right)^2 - \left(\hat{P} - \frac{\hat{p}}{2} - \hat{r} \right) \sum_a^n \left(J^a \right)^2 \Theta \left(|J^a| - \omega \right) \right.$$

$$\left. -i \hat{y} \sum_a^n J^a \Theta \left(|J^a| - \omega \right) - i \hat{x} \sum_a^n J^a \right\}. \quad \text{(C.14)}$$

Mediando sobre J^0 e fazendo

$$\hat{\eta} = \hat{Q} - \frac{\hat{q}}{2} + \hat{P} - \frac{\hat{p}}{2} - \hat{r} \quad \text{(C.15)}$$

encontramos

$$G_2^{sr} = \ln \int \frac{dx d\hat{x}}{2\pi} \int \frac{dy d\hat{y}}{2\pi} \exp \left\{ i x \hat{x} + i y \hat{y} - \frac{x^2 \left(\hat{q} - M \hat{R}^2 \right)}{2} - \frac{y^2 \left(\hat{p} - M \hat{S}^2 \right)}{2} \right.$$

$$-xy\left(\hat{r}-M\hat{R}\hat{S}\right)\Big\}\prod_a\left\{\int_{-\infty}^{-\omega}dJ\exp\left[-\hat{\eta}J^2-i\left(\hat{x}+\hat{y}\right)J\right]\right.$$

$$\left.\int_{-\omega}^{\omega}dJ\exp\left[\left(\hat{Q}-\frac{\hat{q}}{2}\right)J^2-i\hat{x}J\right]+\int_{\omega}^{\infty}dJ\exp\left[-\hat{\eta}J^2-i\left(\hat{x}+\hat{y}\right)J\right]\right\} \quad \text{(C.16)}$$

Vemos então que G_2^{sr} pode ser escrita como

$$\frac{G_2^{sr}}{n}=\int d\mu\left(\hat{x},\hat{y}\right)\ln F\left(\hat{x},\hat{y}\right), \quad \text{(C.17)}$$

onde

$$d\mu\left(\hat{x},\hat{y}\right) = \frac{dxd\hat{x}}{2\pi}\frac{dyd\hat{y}}{2\pi}\exp\left\{i\left(x\hat{x}+y\hat{y}\right)-\frac{x^2\left(\hat{q}-M\hat{R}^2\right)}{2}\right.$$

$$\left.-\frac{y^2\left(\hat{p}-M\hat{S}^2\right)}{2}-xy\left(\hat{r}-M\hat{R}\hat{S}\right)\right\} \quad \text{(C.18)}$$

e

$$F\left(\hat{x},\hat{y}\right) = \int_{-\infty}^{-\omega}dJ\exp\left[-\hat{\eta}J^2-i\left(\hat{x}+\hat{y}\right)J\right]+\int_{-\omega}^{\omega}dJ\exp\left[\left(\hat{Q}-\frac{\hat{q}}{2}\right)J^2-i\hat{x}J\right]$$

$$\int_{\omega}^{\infty}dJ\exp\left[-\hat{\eta}J^2-i\left(\hat{x}+\hat{y}\right)J\right]. \quad \text{(C.19)}$$

C.4 Corte aleatório dos pesos

Usando a prescrição simetria de réplicas na equação (A.43) temos

$$G_2^{sr} = \ln\prod_{a=1}^{n}\mathrm{Tr}\exp\left\{-\hat{R}\sum_a^n J^aJ^0-\hat{q}\sum_{a<b}^n J^aJ^b\right.$$

$$\left.-\left[\frac{\beta h\kappa\left(1-\kappa\right)\alpha}{2}+\hat{Q}\right]\sum_a^n\left(J^a\right)^2\right\} \quad \text{(C.20)}$$

ou

$$G_2^{sr} = \ln\prod_{a=1}^{n}\mathrm{Tr}\exp\left\{-\hat{R}J^0\sum_a^n J^a-\frac{\hat{q}}{2}\left(\sum_a^n J^a\right)^2\right.$$

$$\left.-\left(\frac{\beta h\kappa\left(1-\kappa\right)\alpha}{2}+\hat{Q}-\frac{\hat{q}}{2}\right)\sum_a^n\left(J^a\right)^2\right\}. \quad \text{(C.21)}$$

Aplicando a transformação gaussiana (B.2) para desacoplar as réplicas e tomando o limite $n \longrightarrow 0$ encontramos

$$\frac{G_2^{sr}}{n} = -\frac{1}{2}\ln\pi - \frac{1}{2}\ln\left(\frac{\beta h\kappa\left(1-\kappa\right)\alpha}{2} + \hat{Q} - \frac{\hat{q}}{2}\right) + \int Dz \frac{\left(iz\sqrt{\hat{q}} + \hat{R}J^0\right)^2}{4\left(\frac{\beta h\kappa(1-\kappa)\alpha}{2} + \hat{Q} - \frac{\hat{q}}{2}\right)}. \tag{C.22}$$

Finalmente efetuando a média sobre J^0 obtemos

$$\frac{G_2^{sr}}{n} = -\frac{1}{2}\ln\pi - \frac{1}{2}\ln\left(\frac{\beta h\kappa\left(1-\kappa\right)\alpha}{2} + \hat{Q} - \frac{\hat{q}}{2}\right) + \frac{\left(M\hat{R}^2 - \hat{q}\right)^2}{4\left(\frac{\beta h\kappa(1-\kappa)\alpha}{2} + \hat{Q} - \frac{\hat{q}}{2}\right)}. \tag{C.23}$$