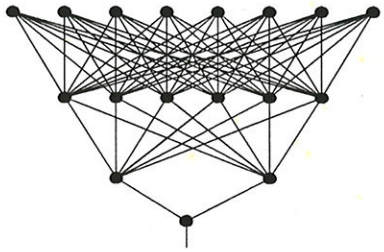
UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FISICA

Generalização e Robustez: Aprendizagem em Redes Neurais na Presença de Ruído



Roberta Simonetti

SBI-IFUSP
305M810T2626

Tese apresentada ao Instituto de Física da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Banca Examinadora :

Prof. Dr. Nestor Felipe Caticha Alfonso (orientador)

Profa. Dra. Alba Theumann (IF -UFRGS)

Prof. Dr. Carlos Eugenio Imbassahy Carneiro (IF -USP)

Profa. Dra. Rita Maria Zorzenon dos Santos (IF - UFF)

Prof. Dr. Silvio Roberto Salinas (IF - USP)

Prof. Iberê Luiz Caldas
Presidente da Comissão de Pós-Graduação

SÃO PAULO
1997

UNIVERSIDADE DE SÃO PAULO
ex. 1
SERVIÇO DE BIBLIOTECA E INFORMAÇÃO
R 626
INSTITUTO DE FÍSICA

09/05/97

530.13
S598g
D
ex.1

FICHA CATALOGRÁFICA
Preparada pelo Serviço de Biblioteca e Informação do Instituto de Física Universidade de São Paulo

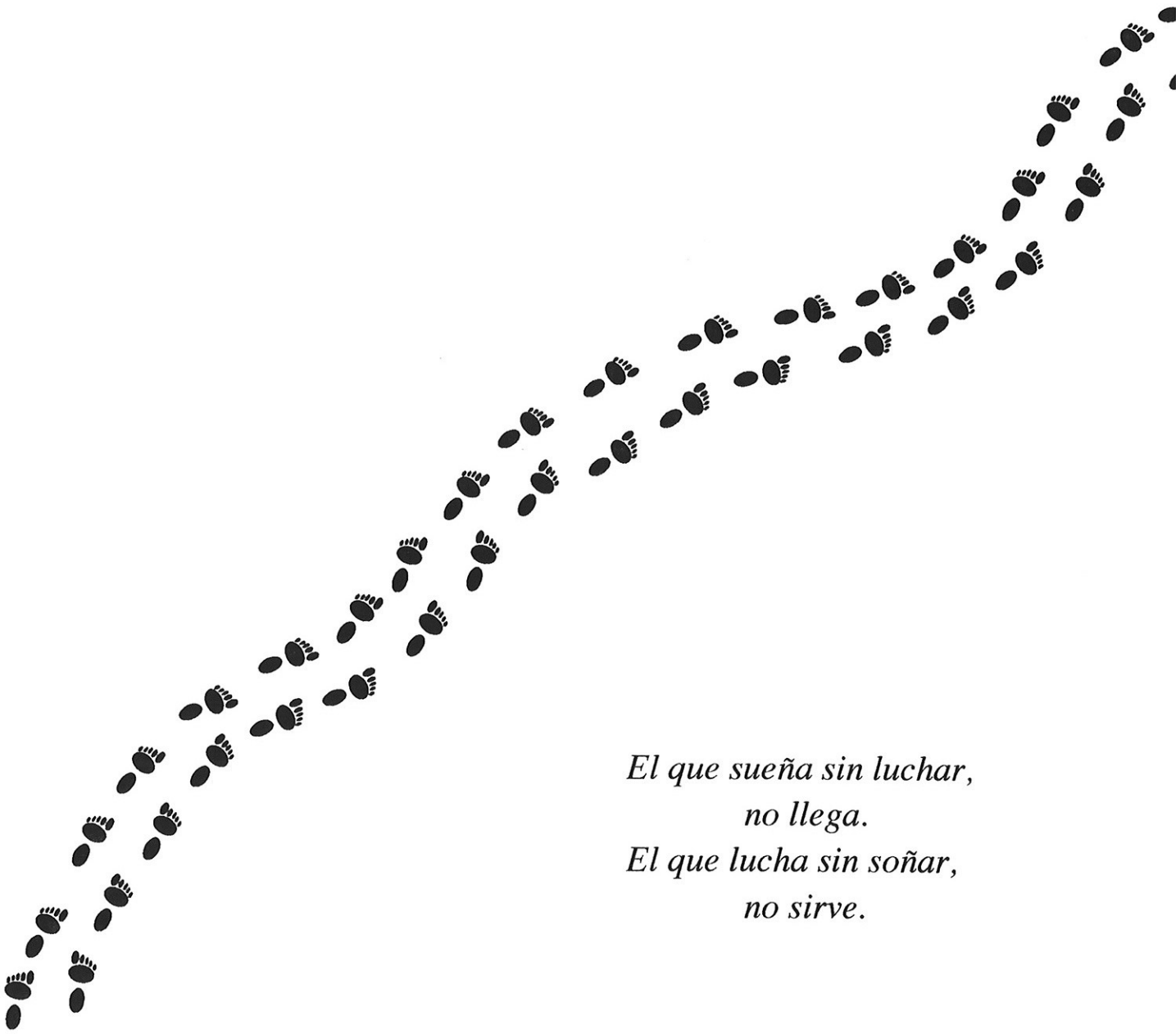
Simonetti, Roberta
Generalização e robustez: aprendizagem em redes neurais na presença de ruído. São Paulo, 1997.

Tese (Doutorado) - Universidade de São Paulo. Instituto de Física - Departamento de Física Geral
Área de Concentração: Física do Estado Sólido
Orientador: Prof. Dr. Nestor Felipe Caticha Alfonso

Unitermos: 1. Mecânica estatística; 2. Redes neurais artificiais; 3. Dinâmicas de aprendizagem; 4. Generalização; 5. Processos estocásticos.

USP/IF/SBI - 13/97

Ao meu querido
Cristiano



El que sueña sin luchar,
no llega.
El que lucha sin soñar,
no sirve.

AGRADECIMENTOS

Esta série de agradecimentos é dedicada às pessoas que contribuíram, direta ou indiretamente para a realização deste trabalho, e sem ele não teria razão de ser. Então, nada mais natural do que iniciar esta lista agradecendo ao Nestor Caticha. Quero agradecê-lo pela oportunidade de realizar este trabalho, pelo seu estímulo, e desejar que esta orientação, que aqui se encerra, possa se transformar numa frutífera colaboração futura.

Ao Cristiano, meu grande companheiro, amigo, esposo... que adora spaguetti a la Simonetti e que sabe ler a minha sobrancelha... pelo seu apoio quase incondicional aos meus "empreendimentos" (quase todos!!!) e principalmente pela oportunidade de crescermos juntos, de lutarmos juntos e de sonharmos juntos...

Aos meus pais, Maria e Gianfranco, por todas as oportunidades que me permitiram chegar até aqui. Por todo amor e carinho que me dedicaram, mesmo diante das mais diversas dificuldades da vida. A Maria, em especial, pelo seu exemplo de luta e de vida. Ao Gianfranco, pelo apoio e compreensão, e por me manter tão ligada às nossas origens storeses...

A Cristina, minha grande amiga, e irmã, por me ajudar a compreender melhor a beleza da natureza, e por saber me ouvir.

À Jô, pela sua constante torcida, amizade, interesse e, especialmente, pela sua vida: um verdadeiro exemplo de luta e de coragem.

À Nonna Mimi, pelos seus minestrones, ragús, torte di mele... e por ter me ensinado muitas outras coisas...

À Vera Maria Rodrigues de Mattos, e a todos os Rodrigues de Mattos, que sempre me recebem de braços abertos...

Aos meus tios Adriana, Anna Lina e Filippo Pardini, pelos ótimos momentos que temos passado juntos todos esses anos!

Ao Carlos Escobar, pela amizade, pelo apoio, por me ouvir e me ajudar, tantas e tantas vezes...

À Inés Prieto Schmidt e ao Ricardo Sauerwein, pela nossa "loooonga" amizade, apesar dos tantos quilômetros que estão nos separando! (e pelas "30 razões"...)

Ao mui caro Javier Satulowsky (seguro!) que, "apesar de ser argentino", é um grande sujeito! E, como dizem por aí: talvez o último cara bom que existe na Terra!

À Kaline Coutinho e ao Sylvio Canuto pela amizade, que começou de repente e cresce a cada dia... e pelos momentos agradáveis que passamos juntos, sempre acompanhados de maravilhosos petiscos!

Ao Nelson Fiedler-Ferrara, pela amizade e pelas maravilhosas pizzas, que enchem muito mais do que o estômago...

Ao Mário Tamashiro, meu amigo e companheiro do chá das cinco, por compartilhar comigo a sua música e pelas incontáveis ajudas...

Ao Mauro Copelli, pela mão e tantas discussões... e aos "catichetes": Silvia Kuva, Renato Vicente e Osame Kinouchi.

A todos os professores do grupo de Mecânica Estatística, que contribuem para a formação de um ambiente científico agradável, em especial ao professor Silvio Salinas, à professora Amélia Império Hamburger, e aos professores Carlos Eugênio I. Carneiro e Mário José de Oliveira.

A todos os amigos do departamento, em especial ao Marcos Yamaguti, pela sua boa vontade em manter os computadores funcionando, permitindo que nosso trabalho possa ser realizado. Aqui, preciso incluir novamente a Kaline Coutinho!

Às secretárias, Miriam e Dirce, sempre muito simpáticas, atenciosas e prontas a ajudar.

Ao Instituto de Física e à Universidade de São Paulo, pela acolhida durante todos esses anos...

À CAPES e ao CNPq, pelo apoio financeiro.

Em especial, à FAPESP, pela prorrogação concedida, sem a qual meu trabalho não teria sido concluído, e esta tese não teria sido escrita.

RESUMO

Neste trabalho investigamos o aprendizado supervisionado on-line, com ênfase nas habilidades de generalização, de redes neurais feedforward. O estudo de algoritmos de aprendizagem ótimos, no sentido da generalização, é estendido para duas diferentes classes de arquiteturas: a máquina paridade com estrutura de árvore e K unidades escondidas, e o perceptron reversed wedge, uma máquina de uma camada com função de transferência não monotônica. O papel do ruído é de fundamental importância na teoria de aprendizagem. Neste trabalho estudamos os processos com ruído que podem ser parametrizados por uma única quantidade, o nível de ruído. No caso da máquina paridade analisamos o aprendizado na presença de ruído multiplicativo (na saída). O algoritmo ótimo é muito superior aos algoritmos de aprendizagem até então apresentados, como o algoritmo de mínima ação (LAA), como podemos ver, por exemplo, através do comportamento do erro de generalização que decai, após a apresentação de p exemplos, com $1/p$ ao invés de $1/p^{1/3}$ como no caso do LAA. Além deste fato, observa-se que não existe um nível de ruído crítico a partir do qual a rede não é capaz de generalizar, como ocorre no LAA. Além do ruído multiplicativo, no caso do perceptron reversed wedge consideramos também o ruído aditivo. Analisamos a função de modulação fornecida pelo algoritmo ótimo e as curvas de aprendizagem. A aprendizagem ótima requer o uso de parâmetros que usualmente não estão disponíveis. Neste caso estudamos a influência da utilização de uma estimativa do nível de ruído sobre as curvas de aprendizado. Estes resultados são apresentados na forma do que chamamos de diagrama de robustez, no espaço de nível de ruído real versus nível de ruído estimado. As linhas de transição deste diagrama definem regiões com comportamentos dinâmicos diferentes. Entre as propriedades mais interessantes encontradas, destacamos a universalidade do diagrama de robustez para ruído multiplicativo, uma vez que é exatamente o mesmo para a máquina paridade e comitê com estrutura de árvore, e para o perceptron reversed-wedge. Entretanto, esta universalidade não se estende para o caso de ruído aditivo, uma vez que, neste caso, os diagramas dependem da arquitetura em questão.

ABSTRACT

In this work online supervised learning is investigated with emphasis on the generalization abilities of feedforward neural networks. The study of optimal learning algorithms, in the sense of generalization, is extended to two different classes of architectures; the tree parity machine (PM) with K hidden units and the reverse wedge perceptron (RWP), a single layer machine with a non monotonic transfer function. The role of noise is of fundamental importance in learning theory, and we study noise processes which can be parametrized by a single quantity, the noise level. For the PM we analize learning in the presence of multiplicative or output noise. The optimal algorithm is far superior than previous learning algorithms, such as the Least Action Algorithm (LAA), since for example, the generalization error's decay is proportional to $1/p$ instead of $1/p^{1/3}$ for the LAA, after p examples have been used for training. Furthermore there is no critical noise level, beyond which no generalization ability is attainable, as is the case for the LAA. For the RW perceptron in addition to multiplicative noise we also consider additive noise. The optimal algorithm modulation function and the learning curves are analized. Optimal learning requires using certain usually unavailable parameters. In this case, we study the influence that misevaluation of the noise levels has on the learning curves. The results are presented in terms of what we have called Robustness Phase Diagrams (RPD), in a space of real noise level against assumed noise level. The RPD boundary lines separate between different dynamical behaviours. Among the most interesting properties, we have found the universality of the RPD for multiplicative noise, since it is exactly the same for the PM, RWP and the tree committee machine. However this universality does not hold for the additive noise case, since RPD's are shown to be architecture dependent.

Índice

1 Introdução **1**

2 Introdução às Redes Neurais **4**

2.1 Motivação e Inspiração 5

2.2 Conceitos básicos 6

2.2.1 Redes Neurais de Atratores - RNA 7

2.2.2 Redes Neurais em Camadas - RNC 9

3 Dinâmicas de aprendizagem **18**

3.1 Dinâmica de aprendizagem on-line 19

3.1.1 O algoritmo de Hebb 22

3.1.2 O algoritmo de Mínima Ação 23

3.1.3 O algoritmo ótimo 26

4 Máquina paridade: caso K=2 **28**

4.1 Características gerais da máquina paridade $K = 2$ 29

4.2 O algoritmo de mínima ação e a máquina paridade 32

4.3 O algoritmo ótimo e a máquina paridade 34

4.3.1 A função modulação ótima: interpretação 36

4.3.2 As equações de evolução 40

4.4 Simulações 43

5 Máquina paridade: caso K geral **47**

5.1 Extensão dos resultados 48

5.2 As equações dinâmicas e o peso ótimo 49

5.2.1 *Solução próxima à origem: o ponto fixo* $\rho_k = 0$ 51
5.2.2 *Comportamento assintótico:* $\rho \to 1$ 52

6 Aprendizagem na presença de ruído **54**
6.1 *Ruído multiplicativo* 55
6.1.1 *O LAA na aprendizagem com ruído* 55
6.1.2 *O algoritmo ótimo na aprendizagem com ruído* 56
6.2 *Limite assintótico de* e_G 58
6.3 *Robustez da aprendizagem na ausência de conhecimento do nível de ruído* . . 59

7 Perceptron com função de transferência não monotônica **66**
7.1 *Introdução* 67
7.2 *Erro de generalização* 70
7.3 *Aprendizagem na presença de ruído multiplicativo* 71
7.3.1 *Função modulação* 71
7.3.2 *A equação de evolução do overlap* 74
7.3.3 *Diagrama de robustez* 75
7.4 *Características da rede especial* $K = K^*$ 79
7.5 *Aprendizagem na presença de ruído aditivo* 80
7.5.1 *Função modulação e equação de evolução do overlap* 80
7.6 *Erro de generalização* 85
7.6.1 *Comportamento no limite* $\alpha \to \infty$ 86

8 Considerações Finais **88**

A Derivação da probabilidade condicional **91**
A.1 *Máquina Paridade* 91
A.1.1 *Derivação de* $P(b_1|\Sigma_B, h_1, h_2)$ 91
A.1.2 *Derivação de* $P^K(b_k|\Sigma_B, \{h_j\})$ 92
A.2 *Perceptron Reversed-Wedge* 94
A.2.1 *Derivação de* $P(b|\xi, h)$ 94
A.2.2 *Derivação de* $P(\tilde{\sigma}, b)$ 96

Referências *98*

Lista de Figuras

2.1 *Representação de redes neurais totalmente conectadas* 8

2.2 *Representação de um perceptron simples* 9

2.3 *Representação de uma rede neural multicamada* 10

2.4 *Arquitetura das redes neurais multicamada* 12

2.5 *Representação espacial dos vetores* $\mathbf{B}_k$ *e* $\mathbf{J}_k$ 13

2.6 *Representação gráfica do overlap* 14

2.7 *Representação vetorial do professor real e seu correspondente corrompido por ruído* 17

3.1 *Procedimento esquemático do algoritmo de mínima ação* 25

4.1 *Representação de uma máquina paridade com 2 ramos* 29

4.2 *Erro de generalização versus overlap para a MP* $K = 2$ 31

4.3 *Função modulação nos ramos 1 e 2 em função do campo local* h_1 *para LAA* . 33

4.4 *Função modulação ótima nos ramos 1 e 2 versus campo local* h_1 37

4.5 *Função modulação reescalada versus campo local reescalado* $\tilde{h}_1$ 38

4.6 *Função modulação ótima reescalada versus campo local reescalado* $\tilde{h}_1$ 39

4.7 $\dot{\rho} \times \rho$*: perceptron e máquina paridade* $K = 2$ 41

4.8 *Erro de generalização (teórico) versus* α 42

4.9 *Erro de generalização versus* α*: teoria e simulação* 45

5.1 *MP K geral:* $e_G \times \alpha$ 49

6.1 *Diagrama de robustez I* 62

6.2 *Diagrama de robustez II* 63

6.3 *Função modulação sem ruído para diferentes valores do overlap* 64

6.4 *Função modulação reescalada com ruído* 65

7.1 *Função* $g(h)$ *no PRW* 67

7.2 *Interpretação geométrica do PRW* 68
7.3 *Função de transferência do PRW* 69
7.4 *Região de integração do erro de generalização do PRW* 70
7.5 *Erro de generalização PRW* 71
7.6 *Função modulação ótima no PRW: sem ruído* 73
7.7 *Função modulação ótima no PRW: com ruído multiplicativo* 74
7.8 *Diagrama de robustez PRW: ruído multiplicativo* 77
7.9 $\dot{\rho} \times \rho$ *para* K *fixo e diferentes pares* (η, ε) 78
7.10 *Função modulação ótima no PRW:* K^* *sem ruído* 80
7.11 $\dot{\rho} \times \rho$ *para* K^* *e diferentes pares* (η, ε) 81
7.12 *Função modulação ótima no PRW* 82
7.13 *Diagrama de robustez PRW: ruído aditivo* 84
7.14 *PRW:* $e_G \times \alpha$ *para diferentes valores de* K 85
7.15 *PRW:* $e_G \times 1/\alpha$ 86

1 Introdução

Uma das áreas de mais intensa atividade no estudo de Redes Neurais Artificiais é a da busca de algoritmos de aprendizado mais eficientes. Uma das redes neurais mais estudadas é o perceptron, pois constitui um bloco fundamental na construção de redes mais complexas e apresenta uma riqueza de comportamento aliada a uma relativa simplicidade. Entretanto esta rede apresenta capacidades computacionais limitadas, por exemplo o perceptron simples booleano só pode ser usado nos problemas linearmente separáveis. Outras redes, mais complexas, implementam problemas que o perceptron simples não é capaz: são as redes que apresentam camadas internas, também chamadas de unidades escondidas pois não se tem acesso ao valor de suas atividades. A possibilidade de construir máquinas com propriedades computacionais universais levou ao estudo das redes multicamadas.

É conhecido que as redes neurais artificiais apresentam propriedades que podem simular, ainda que de forma rudimentar, habilidades cognitivas básicas de grande interesse. Dentre estas habilidades, aparentemente simples, está a capacidade de memorização, que permite o reconhecimento ou a classificação de padrões, e de generalização, que é a capacidade de inferência de regras a partir de um conjunto dado de exemplos. Estes e outros conceitos centrais usados no estudo das redes neurais serão apresentados no capítulo 2.

Quando se quer utilizar uma rede neural para executar uma determinada tarefa, o primeiro passo é escolher a rede mais adequada. Isso justifica a importância do estudo de diferentes tipos de redes neurais: o conhecimento de suas propriedades permite fazer tal escolha e o conhecimento das características de sua aprendizagem permite otimizar sua utilização. Feita a escolha da rede ideal para executar a tarefa desejada, resta a questão da escolha do algoritmo de aprendizagem mais eficiente. As redes neurais podem ser ensinadas, isto é, elas podem aprender a partir das informações contidas nos exemplos apresentados, e não a partir de regras pré-estabelecidas. Dessa forma, um algoritmo eficiente é aquele que permite extrair a máximo de informações que estão contidas nesses exemplos.

Muitos esforços têm sido feitos no sentido de se descobrir a melhor maneira segundo a qual uma rede neural pode aprender. Existem vários algoritmos de aprendizagem para redes neurais, que podem ser implementados basicamente de duas maneiras. Uma delas é conhecida como off-line, onde o aprendizado se dá através da apresentação exaustiva dos exemplos pertencentes ao conjunto de treinamento. Uma outra forma possível é conhecida como on-line, em cujo caso os exemplos são apresentados uma única vez. No capítulo 3 apresentamos a dinâmica de alguns algoritmos on-line.

Recentemente têm sido determinados os algoritmos que, sob certas condições, levam ao desempenho ótimo (no sentido da generalização), tanto off-line quanto on-line, para redes com diferentes arquiteturas. Estes casos incluem o perceptron simples booleano com pesos contínuos e binários, o perceptron linear e a máquina comitê.

Nossa proposta é investigar as propriedades de generalização, quando o aprendizado é feito de maneira on-line, em outros dois tipos de redes. Uma delas é uma rede multicamada com estrutura de árvore, conhecida como máquina paridade (MP), a outra é um perceptron conhecido na literatura como reversed-wedge (PRW).

No capítulo 4 apresentamos a máquina paridade com duas unidades escondidas ($K = 2$), e deduzimos, através do método variacional, a função modulação ótima. Apresentamos também alguns resultados numéricos e as simulações para esta rede. Ainda neste capítulo discutimos os principais resultados existentes na literatura (Kabashima, 1994) sobre a aprendizagem on-line na máquina paridade $K = 2$ com estrutura de árvore, e comparamos com os nossos resultados.

A extensão destes resultados é apresentada no capítulo 5, onde estudamos o aprendizado de uma máquina paridade com um número K qualquer de unidades escondidas. Calculamos a função modulação ótima, as equações de evolução que descrevem o aprendizado e discutimos o comportamento assintótico do erro de generalização.

No capítulo 6 estudamos a situação na qual os exemplos não são apresentados corretamente à rede aluno, mas estão sujeitos a um certo nível de ruído, responsável pela modificação da saída do professor. Investigamos as propriedades de tolerância ao ruído, quando a máquina paridade é treinada usando o algoritmo ótimo. Nosso interesse é saber se esse algoritmo é robusto, isto é, se a rede é capaz de generalizar mesmo na presença deste tipo de ruído. Para tal, analisamos o comportamento assintótico do erro de generalização. Tratamos duas situações distintas: inicialmente assumimos conhecido o nível de ruído presente na saída do professor (ruído real) e a seguir, a situação na qual não se tem acesso a essa grandeza, em cujo caso utiliza-se uma estimativa deste valor (ruído estimado).

No capítulo 7 apresentamos os resultados do estudo da aprendizagem no perceptron

reversed-wedge, que é um perceptron com função de transferência não monotônica. Novamente investigamos o aprendizado através da função modulação e do comportamento do erro de generalização. Estudamos o caso mais geral, isto é, quando o aprendizado se dá na presença de ruído. Consideramos duas fontes distintas de ruído: o ruído na saída e o ruído nas sinapses. Levamos em conta a possibilidade do desconhecimento do nível de ruído pela rede aluno e analisamos as diferentes situações de aprendizagem.

Finalmente, no capítulo 8 apresentamos algumas considerações finais. Alguns cálculos mais detalhados podem ser acompanhados no apêndice, que se encontra no final deste trabalho.

2 Introdução às Redes Neurais

Neste capítulo fazemos uma breve introdução às Redes Neurais Artificiais e discutimos alguns aspectos que inspiram e motivam o seu estudo. Apresentamos os conceitos básicos, os tipos de redes mais estudados, as arquiteturas, as dinâmicas de aprendizagem e as grandezas utilizadas para avaliar seu desempenho.

2.1 Motivação e Inspiração

Nas últimas duas décadas o interesse pelas redes neurais artificiais (RNA's) cresceu consideravelmente. Se por um lado este fenômeno se justifica pela essência das questões em debate, relacionadas à natureza do cérebro, por outro, justifica-se pelo grande potencial tecnológico que elas podem proporcionar.

O estudo das RNA's não tem como objetivo modelar o cérebro, visto a complexidade e sofisticação de sua estrutura. Ao invés disso, busca-se inspiração nas redes neurais naturais para elaborar modelos cognitivos artificiais.

O cérebro é uma das estruturas mais complexas que se conhece, e seu conhecimento vem sendo perseguido pelo homem há centenas de anos. Na segunda metade do século XIX haviam duas escolas: os chamados reticularistas, que acreditavam que o sistema nervoso fosse uma rede contínua e ininterrupta de fibras nervosas, e os neuronistas, que acreditavam que ele era composto por um vasto número de unidades celulares interconectadas.

Por volta de 1880 Camillo Golgi, inventou uma técnica que possibilitou o tingimento das fibras nervosas através de uma reação utilizando de bicromato de prata. Alguns anos mais tarde, a utilização desta técnica permitiu ao espanhol Santiago Ramon y Cajal provar experimentalmente a existência dos neurônios.

Hoje se sabe que existem vários tipos de neurônios, que variam de forma, tamanho e função. Apesar disso, todos eles apresentam uma estrutura muito similar: possuem uma parte central, chamada corpo celular ou soma, cuja dimensão típica é da ordem de $10 - 80\mu m$*; do corpo celular partem várias (até* 10^5*) ramificações: os chamados dendrites, que são receptores de sinais e cujo diâmetro é da ordem de alguns* μm*. Além dos dentrites, do corpo celular também parte uma fibra tubular chamada axônio, cuja função é transmitir os sinais a outras células nervosas ou a fibras musculares. Seu comprimento pode variar de frações de milímetro a um metro.*

Os sinais nervosos podem ser transmitidos de duas formas: elétrica ou química. A transmissão elétrica ocorre principalmente no interior do neurônio, enquanto que a química ocorre entre diferentes neurônios, isto é, nas sinapses, que podem ser de dois tipos: excitatórias ou inibitórias.

De maneira bem simplificada, podemos pensar no cérebro como um conjunto variado, e muito numeroso, de células nervosas "razoavelmente simples" e altamente interconectadas. Por esta intricada teia caminham, a altíssimas velocidades, impulsos elétricos, fruto dos mais variados estímulos, que dão origem às diferentes habilidades cognitivas que o cérebro é capaz de executar.

As principais idéias por trás das redes neurais artificiais surgiram inspiradas neste tipo de estrutura. As redes neurais artificiais funcionam basicamente com um conjunto, em geral numeroso, de processadores simples, usualmente chamados de nodos ou neurônios, conectados entre si. Espera-se que estas estruturas compartilhem com o cérebro algumas propriedades importantes, tais como:

- **capacidade de realizar operações altamente paralelas** ⇒ *existe algo em torno de cem bilhões de neurônios no cérebro, muitos dos quais recebem e enviam impulsos elétricos simultaneamente;*
- **flexibilidade** ⇒ *o cérebro é capaz de se adaptar facilmente a novas circunstâncias;*
- **robustez** ⇒ *o cérebro tem a capacidade de tratar informações afetadas por ruído.*

As redes neurais artificiais foram inicialmente propostas como uma alternativa aos computadores convencionais para a realização de tarefas simples, como o processamento de dados de entrada e saída (McCulloch & Pitts, 1943). Computadores convencionais são, via de regra, constituídos por um único processador que executa funções complicadas, onde cada passo desta computação deve ser explicitamente programado, isto é, define-se uma lista de instruções que deve ser rigorosamente seguida. Entretanto, este tipo de programação sofre algumas limitações, como nos casos em que as regras não são exatamente conhecidas. Além disso, ela não é robusta a erros: basta às vezes a presença de um caractere errado, por exemplo, ou na posição incorreta, para que o programa não funcione!

2.2 Conceitos básicos

O cérebro apresenta uma notável habilidade de aprender. O processo de aprendizagem está intimamente ligado às mudanças nas conexões entre diferentes neurônios, chamadas de conexões sinápticas ou simplesmente sinapses (Hebb, 1949). Tais mudanças ocorrem ao longo da vida, provocadas por estímulos de diversas naturezas. Não se sabe ao certo quantos neurônios existem no cérebro. Algumas estimativas, que se baseiam na densidade de neurônios medida em pequenas regiões, apontam limites inferiores da ordem de 3×10^{10}, *mas este número pode até chegar a* 10^{11}. *Cada um destes neurônios se conecta com aproximadamente* $10^3 - 10^4$ *neurônios. Isso implica que no cérebro existem algo em torno de* $10^{13} - 10^{15}$ *conexões sinápticas. Esses números são extraordinariamente grandes e, de certa forma, evidenciam a ocorrência do aprendizado, pois de outra forma todas essas informações deveriam estar armazenadas na molécula de DNA. Junte-se a isso uma grande massa de*

evidências experimentais confirmando o fato que as conexões sinápticas são ajustadas ao longo da vida, à medida que se aprende.

Uma rede neural pode ser ensinada. Este é um ponto crucial para uma RNA, pois nele se fundamenta seu aprendizado. Isto é muito atraente do ponto de vista tecnológico pois dessa forma é possível eliminar a necessidade da programação explícita e detalhada.

Todos esses aspectos favorecem o estudo das redes neurais artificiais. No entanto, durante muitos anos houve pouco interesse e poucos avanços ocorreram. Uma das causas desta estagnação foi a dificuldade de formalização destes modelos.

2.2.1 Redes Neurais de Atratores - RNA

O modelo básico de uma rede neural de atratores consiste num conjunto de N unidades (McCulloch & Pitts, 1943) totalmente conectadas, que representam os neurônios biológicos. O estado de cada neurônio, representado pela variável binária S_i[1], é influenciado pelo estado de todos os outros neurônios S_j $(i \neq j)$. Tal dependência se dá através do campo local h_i, definido como

$$h_i = \sum_j^N J_{ij} S_j \tag{2.1}$$

A variável J_{ij} é a conexão sináptica entre os neurônios i e j, e mede a contribuição do neurônio j ao estado do neurônio i. A probabilidade de um neurônio i se encontrar no estado S_i no instante $t + \Delta t$ é dada por

$$P(S_i) = \frac{1}{1 + \exp\{-2\beta h_i(t) S_i\}} \tag{2.2}$$

onde a "temperatura" $T \equiv 1/\beta$ representa o ruído[2] existente nas sinapses. Na ausência de ruído, a dinâmica de um neurônio se reduz simplesmente a $S_i(t + \Delta t) = sinal\,(h_i(t))$.

O estabelecimento deste modelo teve a participação de muitos indivíduos, além dos mencionados acima, contribuíram também Hebb (1949), sugerindo que o aprendizado ocorre através da modificação das conexões sinápticas, e Cragg e Temperley (1954) pela analogia entre a memória e histerese em sistemas ferromagnéticos.

No início da década de 80, o estudo das redes neurais teve um grande impulso entre os físicos. A razão disso foi o trabalho de Hopfield (1982) que abriu a possibilidade de se tratar analiticamente as redes neurais, usando para esse fim ferramentas matemáticas já bem

[1] S_i pode assumir os valores $+1$ e -1, que correspondem aos estados biológicos ativo e inativo, respectivamente

[2] o responsável por esta relação foi W.A.Little (1974), estabelecendo desta forma uma analogia com a Termodinâmica

desenvolvidas. A analogia com os sistemas de spins desordenados (vidros de spin) contribuiu fortemente para o desenvolvimento das redes neurais, principalmente devido aos avanços na compreensão de suas propriedades termodinâmicas.

Hopfield percebeu que se o estado de cada neurônio for atualizado um a um (dinâmica assíncrona) através de um processo estocástico e impondo algumas restrições sobre as características das conexões ($J_{ij} = J_{ji}$ e $J_{ii} = 0$) é possível escrever uma função energia E para descrever a dinâmica da rede. Na ausência de ruído ($T = 0$), o sistema caminha no espaço de configurações para o mínimo local de E. Se $T \neq 0$, o sistema relaxa para um estado de equilíbrio descrito pela distribuição de Gibbs ($\exp\{-\beta E\}$). Dessa forma os atratores da dinâmica constituem mínimos da energia livre e podem ser estudados através de métodos da mecânica estatística de equilíbrio. A figura 2.1 mostra a estrutura deste tipo de rede: cada neurônio da rede se conecta com todos os demais, além disso as setas nos dois sentidos indicam que ambos os neurônios se influenciam mutuamente e da mesma forma (isso equivale à simetria $J_{ij} = J_{ji}$).

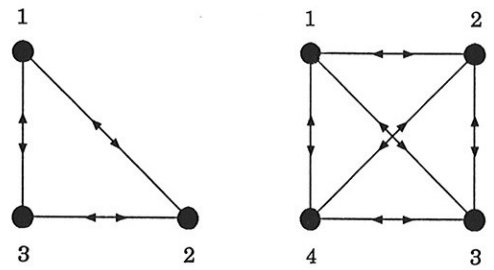


Figura 2.1: Representação esquemática de redes neurais totalmente conectadas.

Hopfield observou ainda que configurações escolhidas a priori podem constituir pontos fixos da dinâmica. Do ponto de vista biológico esta propriedade é muito interessante pois estes pontos fixos podem ser interpretados como memórias, análogas às memórias de sistemas biológicos. Estas redes são conhecidas como Redes Neurais de Atratores, ou redes recorrentes, e o problema de armazenamento de memórias constitui, ainda hoje, área de interesse e atividade (Amit, 1989; Domany et al., 1991: Hertz et al., 1991).

O processo de armazenamento de memória é também chamado de aprendizado não supervisionado, pois não existe um mapeamento de pares entrada/saída previamente determinado. Neste caso, a rede apresenta uma dinâmica intrínseca e, dada uma entrada, evolui naturalmente para um estado, que representa um dos pontos fixos da dinâmica, ao qual associa-se um dos padrões de entrada.

A primeira pergunta que se faz neste tipo de estudo é: dado um conjunto de p

padrões, existe uma rede neural que tem estes padrões como pontos fixos da dinâmica? Considerando a existência desta rede, a pergunta seguinte é: quantos padrões podem ser armazenados, em função do tamanho N da rede? Este número é conhecido como capacidade de armazenamento ou capacidade de memorização da rede (Gardner, 1988).

Além das redes neurais de atratores, existe um outro tipo de rede, do qual tratamos neste trabalho, que são as redes neurais em camadas, que passamos a descrever.

2.2.2 Redes Neurais em Camadas - RNC

A estrutura das redes neurais em camadas difere essencialmente da estrutura das redes de atratores. Nas RNC os neurônios não estão totalmente conectados, em vez disso eles se dispõem em camadas e as conexões sinápticas comunicam apenas os neurônios de diferentes camadas. Estas redes são também conhecidas como redes não recorrentes, ou feedforward, pois a informação corre apenas em um sentido. Esta característica implica que não existe retro-alimentação, isto é, a saída de um neurônio da camada i serve de entrada para o neurônio da camada seguinte, $i+1$, não ocorrendo a situação inversa, de forma que os neurônios não se influencia mutuamente como ocorre nas redes de atratores. Nas redes em camadas o estado dos neurônios de entrada configura a pergunta, e a resposta corresponde à configuração dos neurônios de saída.

O perceptron é o representante mais simples deste tipo de rede, sendo formado por um conjunto de N entradas, representadas na figura 2.2 por S_i, N conexões sinápticas, representadas pelas componentes do vetor N-dimensional $\mathbf{J}$, e uma saída, σ [3].

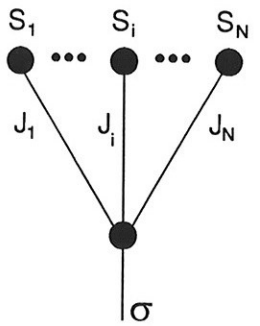


Figura 2.2: Representação esquemática de um perceptron simples.

[3] Vamos adotar a seguinte notação: σ indica tanto a saída de um perceptron simples, como a atividade (estado) de um neurônio escondido. O símbolo Σ é utilizado para indicar a saída de uma rede multicamada.

Embora apresente uma estrutura simples, o perceptron tem sido largamente estudado (Seung et al., 1992; Watkin et al., 1993) pois constitui um elemento básico na construção de redes mais complexas e, apesar de sua aparente simplicidade, apresenta comportamentos interessantes. Além do perceptron, existem redes mais sofisticadas: são as redes multicamadas, que apresentam camadas intermediárias, além das camadas de entrada e saída (figura 2.3). O número de unidades escondidas está relacionado com o número de representações internas possíveis da rede, que por sua vez determina a complexidade do mapeamento implementado e os tipos de problemas para os quais a rede pode ser utilizada.

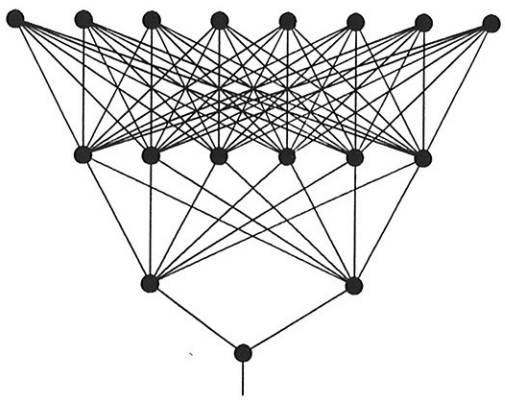


Figura 2.3: Representação de uma rede neural multicamada.

As RNC são caracterizadas pelos vetores sinápticos $\mathbf{J}_{m,k}$ *onde o índice* m *representa a* m*-ésima camada, sendo* M *o número de camadas intermediárias e* k *representa a* k*-ésima unidade escondida da camada* m*. A dimensão dos vetores sinápticos varia de acordo com a arquitetura da rede, que discutiremos adiante. O estado dos neurônios da camada* m *é representado pelo vetor* $\mathbf{S}_m$*, cuja dimensão é igual ao número de unidades existentes na referida camada. Como vimos, o perceptron simples não possui camadas intermediárias, sendo representado apenas por um vetor sináptico de dimensão* N*.*

Em cada unidade escondida, define-se o campo local como:

$$h^{\mu}_{m,k} \equiv \frac{\mathbf{S}^{\mu}_{m} \cdot \mathbf{J}_{m,k}}{|\mathbf{J}_{m,k}|} \ , \tag{2.3}$$

onde, para uma dada entrada $\mathbf{S}^{\mu}_{m}$ *a rede neural associa uma saída,* Σ^{μ}_{J}*, que é função dos vetores* $\mathbf{S}^{\mu}_{m}$ *e* $\mathbf{J}_{m,k}$*. O índice* μ *indica o* μ*-ésimo exemplo de um conjunto qualquer de exemplos* ($\mu = 1...p$)*. Esta dependência pode ser escrita em termos dos campos locais* $h^{\mu}_{m,k}$ *como:*

$$\Sigma^{\mu}_{J} = \mathcal{F}\left(\{h^{\mu}_{m,k}\}\right) \ . \tag{2.4}$$

A função $\mathcal{F}$ é conhecida como função ganho ou função de transferência, que pode ser, entre outras, uma sigmóide, uma função linear, uma função sinal. No caso da função de transferência ser a função sinal, as redes são denominadas booleanas e a saída pode assumir apenas dois valores: $+1$ e -1.

A máquina paridade e a máquina comitê são redes neurais multicamadas que apresentam uma camada escondida, e portanto necessitam de dois vetores sinápticos para serem definidas. O primeiro vetor conecta a camada de entrada à camada escondida, o segundo conecta as unidades da camada escondida à saída. As componentes do segundo vetor sináptico são fixas e iguais a um. Dessa forma podemos suprimir o índice m. A máquina comitê implementa a regra da maioria, isto é, a saída é dada pelo sinal da soma dos estados dos neurônios da camada escondida (σ_k^μ),

$$\Sigma_J^\mu = sinal\left(\sum_{k=1}^{K} \sigma_k^\mu\right) , \tag{2.5}$$

onde $\sigma_k^\mu = sinal(h_k^\mu)$, com h_k^μ sendo definido pela equação (2.3). A saída na máquina paridade é dada pelo produto destas grandezas, ao invés da soma, isto é,

$$\Sigma_J^\mu = sinal\left(\prod_{k=1}^{K} \sigma_k^\mu\right) . \tag{2.6}$$

Daí vem o termo paridade: a saída da MP será positiva sempre que $\sigma_k^\mu = -1$ para um número par de unidades escondidas, e será negativa sempre que $\sigma_k^\mu = -1$ para um número ímpar.

Arquiteturas

As RNC podem ter basicamente duas arquiteturas, como mostra a figura 2.4.

O esquema à direita (b) mostra uma rede com arquitetura "overlapping", isto é, cada unidade interna está conectada a todas as N unidades de entrada. Isto implica que os exemplos apresentados são os mesmos para cada uma das unidades escondidas. O esquema à esquerda (a) representa uma rede com uma arquitetura um pouco mais simples: são redes com "arquitetura não overlapping" ou "estrutura de árvore". Elas podem ser divididas em K ramos, onde K é o número de unidades escondidas. Cada unidade escondida se conecta apenas à uma fração (N/K) das unidades de entrada. Dessa forma, é interessante "decompor" o vetor de entrada $\mathbf{S}^\mu$ em k vetores $\{\mathbf{S}_k^\mu\}$, com $k = 1...K$, bem como o vetor sináptico $\mathbf{J} \equiv \{\mathbf{J}_k\}$.

A máquina paridade, que é a RNC estudada neste trabalho, possui estrutura de árvore e as discussões que se seguem são feitas levando em conta essa característica. É

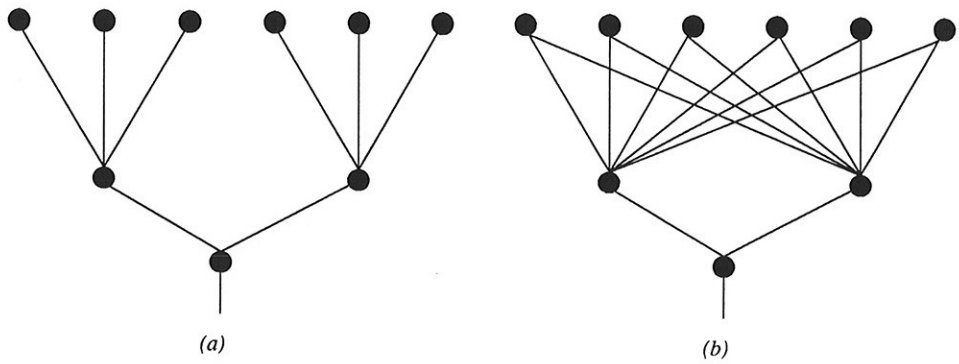


Figura 2.4: Redes neurais multicamada com diferentes arquiteturas: não overlapping (a) e overlapping (b).

bom lembrar que as componentes do segundo vetor sináptico são fixas e iguais a um e que portanto basta um conjunto de K vetores sinápticos $\{\mathbf{J}_k\}$ para definir a rede neural.

Objetivo e essência do aprendizado supervisionado

As RNC podem aprender a partir de exemplos, que são pares de entrada ($\mathbf{S}^\mu$) e saída (Σ^μ). Tais exemplos podem pertencer a um conjunto conhecido como conjunto de aprendizagem, que será representado por $\mathcal{L}$. A existência deste conjunto é a característica básica do aprendizado supervisionado. O processo de aprendizagem permitirá o reconhecimento de padrões vistos anteriormente, isto é, que pertencem ao conjunto $\mathcal{L}$ ou a reconstituição de uma regra, habilidade conhecida como generalização ou extração de regras, que é medida através da apresentação de exemplos que não pertencem necessariamente ao conjunto $\mathcal{L}$. O conjunto de aprendizagem $\mathcal{L}$ é um mapeamento $\mathbf{S}^\mu \longrightarrow \Sigma^\mu$, que pode ser definido de maneira aleatória ou determinado por uma outra rede, denominada rede professor, representada pelo vetor $\mathbf{B}$, que pode, ou não, ter a mesma arquitetura da rede aluno.

No estudo de redes neurais em camadas, a pergunta a ser feita muda um pouco em relação às redes neurais de atratores. A questão agora é encontrar a rede neural, isto é, o vetor sináptico, que implementa corretamente um mapeamento definido a priori pelo conjunto de aprendizagem. No caso da existência do professor, o objetivo do processo de aprendizagem pode ser entendido como a busca de uma rede que se aproxime da rede professor. Para isso parte-se de uma rede aluno cujas componentes do vetor sináptico deverão ser escolhidas a priori. As condições iniciais podem variar: o rede aluno pode partir de tabula rasa, isto é, $\mathbf{J} = 0$, ou $\mathbf{J}$ pode ser escolhido aleatoriamente. Esta rede será treinada, isto é, suas conexões sinápticas serão modificadas, de acordo com alguma dinâmica, visando aproximá-la da rede

professor.

É usual representar graficamente uma rede neural por vetores no espaço N dimensional. Vamos representar o k-ésimo vetor sináptico da rede aluno e da rede professor por dois vetores no espaço bidimensional, e outros dois vetores $\mathbf{S}_k$ e $\mathbf{S}_k^*$, que representam duas entradas, de acordo com a figura 2.5.

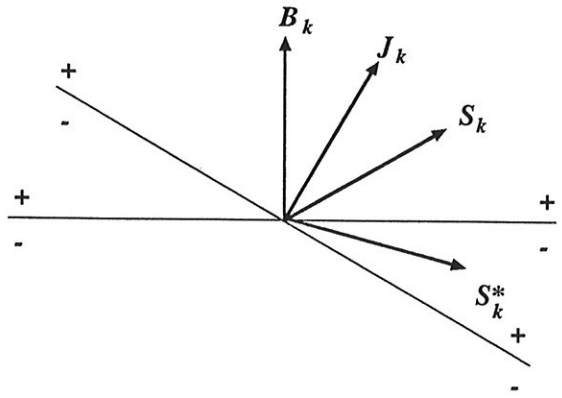


Figura 2.5: Representação espacial dos vetores $\mathbf{B}_k$ e $\mathbf{J}_k$.

Cada plano perpendicular à $\mathbf{B}_k$ e $\mathbf{J}_k$ define duas regiões. Numa delas os exemplos são classificados positivamente e na outra, negativamente. Sendo redes não coincidentes, a rede aluno e a rede professor vão definir regiões distintas, e portanto um exemplo pode ser classificado de maneira diferente pelas duas. Por exemplo, o vetor $\mathbf{S}_k$ é classificado como $+$ por ambas as redes, já o vetor $\mathbf{S}_k^*$ é classificado pela rede professor como $-$ e pela rede aluno como $+$. É fácil observar que isso acontecerá sempre que o exemplo estiver contido na região hachurada da figura 2.6.

É interessante definir uma grandeza, que representa a proximidade entre os vetores $\mathbf{B}_k$ e $\mathbf{J}_k$. Esta grandeza é chamada *overlap* professor-aluno e é dada pelo produto escalar normalizado dos vetores $\mathbf{B}_k$ e $\mathbf{J}_k$:

$$\rho_k = \frac{\mathbf{B}_k \cdot \mathbf{J}_k}{|\mathbf{J}_k||\mathbf{B}_k|} \,. \tag{2.7}$$

Podemos também escrever o overlap em função do ângulo definido pelos vetores $\mathbf{B}_k$ e $\mathbf{J}_k$. É fácil verificar, pela figura 2.6, que $\rho_k = \cos\,\theta$.

Dinâmica de aprendizagem: *on-line versus off-line*

Dissemos que o processo de aprendizagem consiste em fazer modificações no vetor sináptico. A questão é: como modificar $\mathbf{J}_k$ de modo a obter uma rede que implemente o

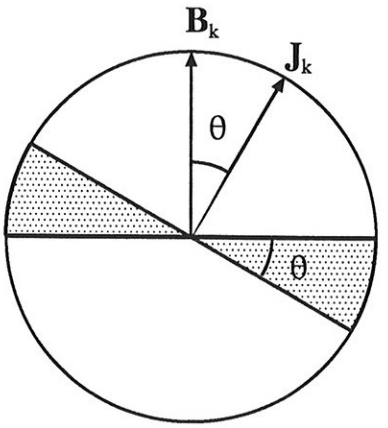


Figura 2.6: Representação gráfica do overlap.

mapeamento desejado? O ponto de partida é que tais mudanças devem ser baseadas nas informações disponíveis, que são os vetores de entrada e suas respectivas saídas, ou seja, nas informações contidas no conjunto de aprendizagem.

Entre outras, é possível implementar o aprendizado de duas formas: off-line e on-line. Na primeira, o aluno pode acessar livremente o conjunto de treinamento, portanto os exemplos podem ser apresentados exaustivamente, e as modificações no vetor $\mathbf{J}_k$ são feitas após a apresentação dos αN exemplos. Esta situação de aprendizagem é estudada usando métodos de mecânica estatística de equilíbrio: existe uma função energia que deve ser minimizada a cada passo, onde um passo corresponde à apresentação de todos os exemplos (Hertz et al., 1991; Seung et al., 1992; Watkin et al., 1993). O mínimo desta energia corresponde ao estado de equilíbrio do sistema no limite termodinâmico. Uma outra forma possível é modificar as conexões sempre que um novo exemplo for apresentado. Esse é um processo de natureza dinâmica e não podemos assumir que o sistema está em equilíbrio. O objetivo é minimizar a energia a cada passo, isto é, à medida que um exemplo é apresentado. Feita a apresentação, o exemplo é descartado, de modo que o aluno tem acesso restrito aos exemplos. Essa característica é interessante do ponto de vista computacional, já que não requer o armazenamento de um grande número de padrões.

Uma importante diferença entre estas situações de aprendizagem é que no caso off-line pode existir uma fase de treinamento relacionada a um certo valor de α, isto é, para uma dada quantidade de padrões αN, a rede é capaz de memorizar os padrões sem que o erro de generalização diminua, quer dizer, sem que a rede esteja necessariamente mais próxima da rede professor. Só a partir de um valor de α a rede entra na fase de generalização, quando

o erro de generalização começa a decrescer.

O desempenho da rede

Para medir o desempenho de uma rede neural no aprendizado supervisionado, podem ser utilizados diferentes parâmetros: o erro de treinamento, o erro de generalização e o erro de predição, descritos a seguir.

- **Erro de treinamento ou de memorização** $(e_T) \Rightarrow$ *esta grandeza mede o desempenho da rede aluno quando lhe são apresentados exemplos* $(\mathbf{S}^\mu)$ *já vistos, e que portanto pertencem ao conjunto de aprendizagem* $(\mathbf{S}^\mu \in \mathcal{L})$. *É a probabilidade do aluno fornecer uma resposta* Σ_J^μ *igual à do professor* Σ_B^μ:

$$e_T = \frac{1}{p} \sum_{\mu=1}^{p} \frac{1 - \Sigma_B^\mu \cdot \Sigma_J^\mu}{2} \,. \tag{2.8}$$

- **Erro de generalização** $(e_G) \Rightarrow$ *esta grandeza mede o desempenho da rede aluno quando lhe são apresentados exemplos* $(\mathbf{S}^\mu)$ *que não fazem parte necessariamente do conjunto de aprendizagem. É uma medida da probabilidade da rede aluno fornecer a resposta correta:*

$$e_G = \left\langle \frac{1 - \Sigma_B^\mu \cdot \Sigma_J^\mu}{2} \right\rangle_{\{\mathbf{S}^\mu\}} \,. \tag{2.9}$$

- **Erro de predição** $(e_P) \Rightarrow$ *esta grandeza mede a probabilidade da rede aluno fornecer uma resposta correta quando os novos exemplos apresentados são corrompidos por um ruído* ε. *Neste caso a resposta do professor é* ξ^μ. *Pode ser escrito como:*

$$e_P = \left\langle \frac{1 - \xi_B^\mu \cdot \Sigma_J^\mu}{2} \right\rangle_{\{\mathbf{S}^\mu\}} \,. \tag{2.10}$$

Aprendizagem na presença de ruído

A aprendizagem supervisionada é realizada a partir da apresentação de exemplos, constituídos por pares de entrada e saída, podendo esta última ser determinada de forma aleatória ou por uma rede professor. A situação ideal é aquela na qual os exemplos são transmitidos corretamente à rede aluno, isto é, cada entrada é associada à verdadeira saída Σ. *Entretanto, esse processo pode ocorrer na presença de ruído, o que implica que a rede aluno não recebe a informação corretamente. As fontes de ruído podem ter diferentes origens, vamos discutir duas delas: ruído multiplicativo e ruído aditivo.*

Ruído Multiplicativo

Esse tipo de ruído é também conhecido como ruído na saída (output noise), pois sua presença afeta diretamente seu valor. A alteração da saída ocorre com uma certa probabilidade, que está relacionada ao nível de ruído ε.

No caso de uma rede booleana, modificar a saída significa inverter o seu sinal, multiplicando-o por -1. Esta operação dá origem ao termo multiplicativo. Sendo Σ a saída correta e ξ a saída corrompida por um certo nível de ruído ε podemos escrever:

$$\begin{aligned} \xi = +\Sigma_B &\implies \text{com probabilidade } 1-\varepsilon \\ \xi = -\Sigma_B &\implies \text{com probabilidade } \varepsilon \end{aligned}$$

e a probabilidade condicional destas duas grandezas é:

$$P(\xi|\Sigma_B) = \varepsilon\,\delta(\xi + \Sigma_B) + (1-\varepsilon)\,\delta(\xi - \Sigma_B)\,. \tag{2.11}$$

Ruído Aditivo

Uma outra possibilidade é alterar a saída indiretamente, corrompendo os exemplos $\mathbf{S}$, *os vetores sinápticos* $\mathbf{J}$ *ou o professor* $\mathbf{B}$. *Vamos considerar o último caso e imaginar a existência de um vetor professor corrompido* $\tilde{\mathbf{B}}$, *que difere de* $\mathbf{B}$ *de acordo com o nível de ruído representado por ω. A grandeza ω pode ser interpretada como overlap entre* $\mathbf{B}$ *e* $\tilde{\mathbf{B}}$:

$$\omega = \mathbf{B}\cdot\tilde{\mathbf{B}}\,. \tag{2.12}$$

Na ausência de ruído os vetores $\mathbf{B}$ *e* $\tilde{\mathbf{B}}$ *coincidem e corresponde à situação na qual $\omega = 1$. A figura 2.7 mostra a representação gráfica destes dois vetores:*

Usando a definição do campo local (2.3), podemos escrever:

$$b \equiv \mathbf{S}\cdot\mathbf{B}\,, \tag{2.13}$$

$$\tilde{b} \equiv \mathbf{S}\cdot\tilde{\mathbf{B}}\,. \tag{2.14}$$

Sendo o vetor $\mathbf{S}$ *aleatório, as grandezas b e $\tilde{b}$ são também variáveis aleatórias, que no limite termodinâmico se distribuem segundo:*

$$P(b,\tilde{b}) = \frac{1}{2\pi\sqrt{1-w^2}}\,\exp\left\{\frac{-b^2-\tilde{b}^2+2b\tilde{b}w}{2(1-w^2)}\right\}\,. \tag{2.15}$$

O aprendizado supervisionado na presença de ruído já foi analisado em vários tipos de rede (Györgyi & Tishby, 1990; Opper & Haussler, 1991; Kabashima, 1994; Biehl et al., 1995; Copelli et al., 1996; Opper & Kinzel, 1996; Kim & Sompolinsky, 1996; Copelli et al., 1997). Nos capítulos 6 e 7 discutiremos mais esse assunto.

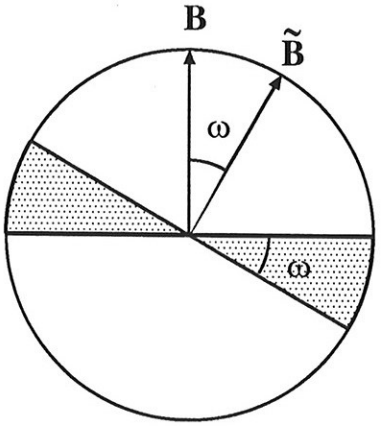


Figura 2.7: Representação esquemática do professor real e seu correspondente corrompido por ruído.

Neste trabalho estudamos a aprendizagem supervisionada on-line na máquina paridade e num perceptron com função de transferência não monotônica, e para medir o desempenho deste aprendizado verificamos o comportamento do overlap e do erro de generalização em função dos exemplos apresentados. No próximo capítulo discutiremos com mais detalhes a dinâmica de aprendizado on-line.

3 Dinâmicas de aprendizagem

Neste capítulo discutimos a dinâmica de aprendizagem on-line e a estrutura de alguns algoritmos que a implementam, entre eles o algoritmo ótimo. Apresentamos a função modulação em cada um dos casos e discutimos o seu papel.

3.1 Dinâmica de aprendizagem on-line

Vimos no capítulo anterior que o objetivo do aprendizado é obter uma rede que seja capaz de generalizar, isto é, responder corretamente a novos exemplos. Isto equivale a obter o menor erro de generalização possível, preferencialmente nulo, ainda que assintoticamente, isto é, para $\alpha \to \infty$, *com* $\alpha = p/N$. *A questão central pode ser colocada da seguinte forma: como modificar as conexões sinápticas de modo a obter a rede desejada? As discussões que se seguem são bastante gerais e servem para qualquer tipo de rede com arquitetura de árvore, como o perceptron simples (obtido tomando-se* $K = 1$*) ou redes multicamadas (casos em que* $K > 1$*), como a máquina comitê ou a máquina paridade. Como mencionamos no capítulo anterior, a máquina paridade possui apenas uma camada interna com* K *unidades e o vetor que conecta esta camada à saída tem componentes fixas e iguais a um. Dessa forma a dinâmica de aprendizado consiste em modificar as componentes do vetor da primeira camada, podendo o índice* m *ser suprimido.*

Na arquitetura de árvore, a entrada $\mathbf{S}^\mu$ *é um vetor* N*-dimensional que pode ser dividido em* K *vetores de dimensão* N/K *representados por* $\mathbf{S}_k^\mu$, *cujas componentes são denotadas por* S_{ki}^μ *e podem ser escolhidas, sem perda de generalidade, como* ± 1. *Esta arquitetura provoca o desacoplamento dos campos locais, já que cada unidade escondida recebe uma informação diferente, pois "enxerga" uma porção diferente do vetor de entrada. Isso faz com que cada um dos ramos possa aprender o ramo correspondente no professor e ser interpretado como um perceptron simples, representado pelo vetor sináptico* N/K*-dimensional* $\mathbf{J}_k$.

A essência do aprendizado on-line consiste em modificar as conexões sinápticas cada vez que um novo exemplo é apresentado. O aprendizado é, portanto, um processo estocástico com dinâmica discreta, descrita pela regra de aprendizagem que pode ser representada genericamente pela equação:

$$J_{ki}^{\mu+1} = J_{ki}^{\mu} + \Delta J_{ki}^{\mu} \,, \tag{3.1}$$

onde o índice μ *representa o* μ*-ésimo exemplo e faz o papel do tempo na dinâmica. Isso significa que, a cada passo, as componentes do vetor sináptico,* J_{ik}^{μ}, *são modificadas pelo incremento* ΔJ_{ki}^{μ}, *que depende do exemplo apresentado e pode ser escrito de maneira geral como:*

$$\Delta J_{ki}^{\mu} = \gamma F_k^{\mu} S_{ki}^{\mu} \,. \tag{3.2}$$

Esta forma se justifica pelo fato que $\mathbf{\Delta J}_k^{\mu}$ *deve ser um vetor, e os únicos vetores disponíveis, além do vetor sináptico, são os exemplos* $\mathbf{S}^\mu$. *O fator* γ *foi introduzido com o objetivo de controlar o módulo de* $\mathbf{\Delta J}_k^{\mu}$ *e será escolhido como* $\gamma = 1/N$. *Podemos então*

reescrever a regra de aprendizagem:

$$J_{ki}^{\mu+1} = J_{ki}^{\mu} + \frac{1}{N} F_k^{\mu} S_{ki}^{\mu} \,. \tag{3.3}$$

F_k^{μ} *é chamada função modulação e define o algoritmo de aprendizagem. Pela equação (3.3) fica claro o seu papel: ela representa a importância dada a cada exemplo* $\mathbf{S}^{\mu}$, *em outras palavras ela é uma medida da quantidade de informação relevante que cada exemplo introduz durante o aprendizado. Por isso, a função* F_k^{μ} *é também chamada de função peso.*

Uma maneira direta de verificar o desempenho de um algoritmo de aprendizagem on-line é estudar a evolução temporal do overlap e do módulo do vetor sináptico, isto é, estudar como estas grandezas variam à medida que novos exemplos são apresentados. Esta evolução é descrita por um conjunto de $2K$ *equações acopladas, que podem ser facilmente obtidas a partir da equação (3.3): a primeira equação é obtida tomando-se o módulo do vetor* $\mathbf{J}_k^{\mu}$ *na equação (3.3) e a segunda, tomando-se o produto escalar normalizado entre (3.3) e* $\mathbf{B}_k$.

Vamos inicialmente deduzir a equação de evolução do módulo do vetor $\mathbf{J}_k$, *que é dado por* $J_k = \sqrt{(\mathbf{J}_k)^2}$. *Usando o fato que* $S_{ki} = \pm 1$ *e a definição de* h_k^{μ}, *dada pela equação (2.3), chega-se ao resultado:*

$$(J_k^{\mu+1})^2 = (J_k^{\mu})^2 \left[1 + \frac{2F_k^{\mu} h_k^{\mu}}{N J_k^{\mu}} + \frac{(F_k^{\mu})^2}{KN (J_k^{\mu})^2} \right] \,. \tag{3.4}$$

Como estamos interessados no limite termodinâmico ($N \to \infty$) *podemos usar a aproximação* $\sqrt{1+x} \approx 1 + x/2$ *para argumentos pequenos* ($x \to 0$) *e extrair a raiz quadrada de (3.4). Assim teremos:*

$$J_k^{\mu+1} = J_k^{\mu} + \frac{1}{N} F_k^{\mu} h_k^{\mu} + \frac{1}{2KN} \frac{(F_k^{\mu})^2}{J_k^{\mu}} \,. \tag{3.5}$$

Para encontrar a equação de evolução dinâmica do overlap, procedemos de maneira análoga. Partindo da definição do overlap dada pela equação (2.7) e do resultado obtido na equação (3.5), mantendo os termos até ordem N^{-1} *chegamos à:*

$$\rho_k^{\mu+1} = \rho_k^{\mu} + \frac{F_k^{\mu} b_k^{\mu}}{N J_k^{\mu}} - \frac{(F_k^{\mu})^2 \rho_k^{\mu}}{2KN (J_k^{\mu})^2} - \frac{F_k^{\mu} h_k^{\mu} \rho_k^{\mu}}{N J_k^{\mu}} \,. \tag{3.6}$$

Definindo $\Delta J_k^{\mu} = J_k^{\mu+1} - J_k^{\mu}$ *e* $\Delta \rho_k^{\mu} = \rho_k^{\mu+1} - \rho_k^{\mu}$, *podemos reescrever as equações (3.5) e (3.6) como:*

$$\Delta J_k^{\mu} = \frac{F_k^{\mu} h_k^{\mu}}{N} + \frac{(F_k^{\mu})^2}{2KN J_k^{\mu}} \tag{3.7}$$

$$\Delta\rho_k^{\mu+1} = \frac{\rho_k^\mu F_k^\mu}{N J_k^\mu}\left(\frac{b_k^\mu}{\rho_k^\mu} - h_k^\mu - \frac{F_k^\mu}{2K J_k^\mu}\right) . \tag{3.8}$$

Tomando $\alpha = p/N$, onde p é o número de exemplos apresentados e N é o tamanho da rede, estas equações se tornam:

$$\frac{\Delta J_k^\mu}{\Delta\alpha} = F_k^\mu h_k^\mu + \frac{(F_k^\mu)^2}{2K J_k^\mu} \tag{3.9}$$

$$\frac{\Delta\rho_k^{\mu+1}}{\Delta\alpha} = \frac{\rho_k^\mu F_k^\mu}{J_k^\mu}\left(\frac{b_k^\mu}{\rho_k^\mu} - h_k^\mu - \frac{F_k^\mu}{2K J_k^\mu}\right) . \tag{3.10}$$

e no limite $N \to \infty$ elas podem ser escritas na forma diferencial, uma vez que as mudanças em J_k^μ e ρ_k^μ serão infinitesimais. Assim teremos:

$$\frac{dJ_k^\mu}{d\alpha} = J_k^\mu\left(\frac{F_k^\mu h_k^\mu}{J_k^\mu} + \frac{(F_k^\mu)^2}{2K(J_k^\mu)^2}\right) \tag{3.11}$$

$$\frac{d\rho_k^\mu}{d\alpha} = \frac{\rho_k^\mu}{J_k^\mu} F_k^\mu\left(\frac{b_k^\mu}{\rho_k^\mu} - h_k^\mu - \frac{F_k^\mu}{2K J_k^\mu}\right) . \tag{3.12}$$

Estamos interessados nas propriedades médias do aprendizado e, portanto, a evolução do parâmetro ρ_k e de $|\mathbf{J}_k|$, não deve depender de um conjunto específico de exemplos, por isso é feita uma média sobre todos os possíveis exemplos, $\mathbf{S}^\mu$. Uma vez que a dependência nos exemplos se dá através dos campos locais, b_k^μ e h_k^μ, fazer a média sobre os exemplos corresponde a fazer a média sobre os valores dos campos, utilizando para isso a probabilidade conjunta $P^K(b_k^\mu, \Sigma_B^\mu, \{h_j^\mu\})$.

No limite termodinâmico ρ_k^μ e $\mathbf{J}_k^\mu$ se tornam quantidades auto-mediantes, isto é, a largura da distribuição vai à zero com $N \to \infty$ e, portanto, suas flutuações podem ser desprezadas. Dessa forma podemos considerá-las independentes do conjunto aleatório de exemplos e tomar seus valores médios, que serão representados por ρ_k e $\mathbf{J}_k$. Entretanto suas derivadas não são auto-mediantes e por isso se torna necessário fazer a média:

$$\frac{dJ_k}{d\alpha} = J_k\left\langle\frac{(F_k^\mu)^2}{2K J_k^2} + \frac{F_k^\mu h_k^\mu}{J_k}\right\rangle_{b_k^\mu, \Sigma_B^\mu, \{h_j^\mu\}} \tag{3.13}$$

$$\frac{d\rho_k}{d\alpha} = \dot{\rho}_k = \frac{\rho_k}{J_k}\left\langle F_k^\mu\left(\frac{b_k^\mu}{\rho_k} - h_k^\mu - \frac{F_k^\mu}{2K J_k}\right)\right\rangle_{b_k^\mu, \Sigma_B^\mu, \{h_j^\mu\}} \tag{3.14}$$

onde

$$\langle(\cdots)\rangle = \sum_{\Sigma_B^\mu=\pm1}\int P^K(b_k^\mu, \Sigma_B^\mu, \{h_j^\mu\})\, db_k^\mu \left[\prod_{j=1}^K dh_j^\mu\right](\cdots) . \tag{3.15}$$

Neste ponto é importante enfatizar que o processo de aprendizagem só pode contar com informações conhecidas e, já que o professor não é conhecido, não temos acesso ao valor dos seus campos locais, b_k^μ. *O que se pode fazer nesse caso, é integrar o lado direito da equação (3.14) sobre todos os possíveis valores de* b_k^μ. *Usando o fato que a probabilidade conjunta* $P^K(b_k^\mu, \Sigma_B^\mu, \{h_j^\mu\}) = P^K(\Sigma_B^\mu, \{h_j^\mu\}) \cdot P^K(b_k^\mu | \Sigma_B^\mu, \{h_j^\mu\})$, *obtemos:*

$$\frac{d\rho_k}{d\alpha} = \frac{\rho_k}{J_k} \sum_{\Sigma_B=\pm 1} \int P^K(\Sigma_B^\mu, \{h_j^\mu\}) \left[\prod_{j=1}^{K} dh_j^\mu\right] F_k^\mu \left(\frac{1}{\rho_k} < b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}} - h_k^\mu - \frac{F_k^\mu}{2KJ_k}\right) , \quad (3.16)$$

onde

$$< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}} = \int b_k^\mu \, P^K(b_k^\mu|\Sigma_B^\mu, \{h_j^\mu\}) \, db_k^\mu \, . \quad (3.17)$$

A equação (3.16) expressa a taxa com que a regra é extraída, isto é, ela determina a rapidez da variação do overlap com apresentação de um novo exemplo. A questão é como escolher a função F_k^μ, *levando em conta o objetivo de obter um algoritmo que implemente uma rede capaz de generalizar de maneira eficiente, isto é, como obter uma rede cujo erro de generalização seja minimizado com o menor número possível de exemplos, ou cujo overlap atinja mais rapidamente o seu valor máximo* ($\rho = 1$). *Antes de responder a essa questão, vamos discutir outros dois algoritmos e analisar o desempenho de uma máquina paridade quando se usam estes algoritmos.*

3.1.1 O algoritmo de Hebb

A única restrição na escolha da função modulação é que ela só dependa de grandezas conhecidas, que podem ser, por exemplo, os elementos do conjunto de aprendizagem, os campos locais do aluno, a dimensão dos vetores de entrada, o número de unidades escondidas e a saída do professor.

A escolha mais simples possível é tomar $F_k^\mu = \Sigma_B^\mu$. *Esta escolha equivale ao algoritmo de Hebb, já utilizado com sucesso no caso do perceptron simples e da máquina comitê. Em ambos os casos o erro de generalização vai assintoticamente à zero com* $\alpha^{-\frac{1}{2}}$. *Vejamos o que ocorre no caso da máquina paridade quando se utiliza o algoritmo de Hebb. Vamos investigar primeiro de maneira "intuitiva". Tomemos, por exemplo, uma máquina paridade com 3 unidades escondidas, ou 3 ramos. Seja* σ_B^i *a saída do i-ésimo ramo e* Σ_B^μ *a saída global do professor para um dado exemplo. As configurações possíveis desta rede são:*

	1	2	3	4	5	6	7	8
σ_B^1	−	+	+	−	+	+	−	−
σ_B^2	−	+	−	+	+	−	+	−
σ_B^3	−	−	+	+	+	−	−	+
Σ_B^μ	−	−	−	−	+	+	+	+

Tomemos, por exemplo, o aprendizado no ramo 1. Nas situações 1,4,5 e 6 σ_B^1 *coincide com* Σ_B^μ*, nas demais (2,3,7,8) ocorre o inverso. Portanto, em 50% das vezes toma-se o sinal correto para o aprendizado, e nas outras 50% o sinal adotado é errado. Isso significa que não existe correlação entre estas duas grandezas e que este problema equivale ao aprendizado com ruído máximo (1/2) no qual a rede é incapaz de generalizar.*

É importante notar que o algoritmo de Hebb não utiliza a informação contida nos outros ramos para implementar modificações no vetor sináptico. Algoritmos que apresentam esta característica são ditos algoritmos locais. A única informação utilizada é a saída global da rede.

Formalmente, podemos verificar o desempenho do algoritmo de Hebb analisando a evolução do overlap em função dos exemplos apresentados. Partindo da equação (3.14) e tomando $F_k^\mu = \Sigma_B^\mu$*, para um* K *geral, chega-se à equação:*

$$\frac{d\rho_k}{d\alpha} = \dot{\rho}_k = -\frac{2}{K\pi^2}\frac{\rho_k}{J_k^2}\prod_{i=1}^{K}\arccos(\rho_i) \ . \tag{3.18}$$

A partir deste resultado pode-se concluir que o overlap nunca atinge o seu valor máximo ($\rho_k = 1$)*. Mas, mais do que isso, ele mostra que o overlap sempre se anula, independentemente do seu valor inicial. Se* $\rho_k(0) > 0$*, seu valor tende a diminuir, até chegar à zero, pois nesse caso* $\dot{\rho}_k < 0$*. Caso seu valor seja inicialmente negativo, ou se no decorrer do processo ele atinge um valor negativo,* $\dot{\rho}_k$ *assume um valor positivo e faz com que o overlap tenda a zero novamente. Portanto o valor* $\rho_k = 0$ *é um ponto fixo estável deste sistema dinâmico e mostra que, usando o algoritmo de Hebb, uma máquina paridade não será capaz de generalizar.*

3.1.2 O algoritmo de Mínima Ação

Originalmente chamado de "Least Action Algorithm" (LAA), este algoritmo foi proposto por Mitchison e Durbin (Mitchison & Durbin, 1989), quando investigavam, empiricamente, os limites para a capacidade de armazenamento de padrões em redes multicamada. O objetivo deste algoritmo não era encontrar todas as soluções possíveis, mas superar a eficiência do algoritmo de retro-propagação na tarefa de memorizar vetores aleatórios.

O algoritmo de "Mínima Ação" foi proposto no contexto da aprendizagem off-line, isto é quando os exemplos são apresentados exaustivamente. A idéia básica do LAA é modificar as sinapses da rede responsáveis pelas respostas erradas. Seu princípio de funcionamento é o seguinte: quando a rede fornece uma resposta errada, é produzida uma lista com o valor do campo em cada uma das unidades escondidas. Então, escolhe-se para ser modificada

aquela cujo valor do campo local é menor em módulo. A modificação é feita usando o algoritmo perceptron (Hertz et al., 1991), onde apenas os exemplos cuja estabilidade[1] *é menor do que um certo valor produzem modificação no vetor sináptico. Tal modificação, entretanto, independe do valor da estabilidade, sendo a mesma para todos os exemplos que satisfazem a primeira condição.*

A filosofia por trás desta regra é selecionar a unidade que requer a mínima mudança, desde que tal mudança provoque a inversão da saída. Isso garante que ao longo deste processo serão feitas as mínimas modificações que levam ao objetivo final, que é o armazenamento de vetores aleatórios. Esta é a origem do nome: Least Action Algorithm, ou "algoritmo de mínima ação".

Em seu trabalho, Mitchison e Durbin não obtém uma expressão exata para a capacidade de armazenamento, apenas estabelecem seus limites para uma rede de duas camadas, onde cada peso sináptico é variável. As principais conclusões deste trabalho são que o limite superior depende linearmente do número de entradas e que o número de pares entrada-saída por sinapse que podem ser armazenados é maior do que para o caso de uma rede simples com saída linear. Entretanto existem algumas restrições para este algoritmo, uma delas é que ele só se aplica ao treinamento de redes com unidades escondidas, e portanto não é tão geral como o algoritmo de retro-propagação. Por outro lado serve para tratar o caso de máquinas Booleanas, para as quais retro-propagação não funciona.

Em 1994, Kabashima (Kabashima, 1994) estudou uma máquina paridade com estrutura de árvore e duas unidades escondidas. Seu interesse era investigar, não mais a capacidade da máquina paridade memorizar corretamente um dado conjunto de exemplos, mas a sua habilidade de inferir uma regra a partir de um dado conjunto de exemplos, em outras palavras, a capacidade de generalização da rede. Neste novo enfoque, Kabashima propõe a utilização do LAA de maneira on-line, isto é, os exemplos não são apresentados exaustivamente ao aluno até serem memorizados, ao invés disso eles são apresentados uma única vez e na medida em que isso ocorre são feitas mudanças no vetor sináptico.

As condições de aprendizado são:

- *considera-se a existência de uma rede professor que determina a priori a regra de aprendizagem;*
- *a rede aluno e a rede professor apresentam a mesma arquitetura de árvore, o que implica que as unidades escondidas recebem estímulos diferentes;*
- *supõe-se que as componentes de cada exemplo são escolhidas aleatoriamente a partir*

[1] que é definida pelo produto entre a saída do professor e campo local

de uma distribuição uniforme.

O procedimento básico do LAA (figura 3.1) é o seguinte:

1. *Um novo exemplo é apresentado ao aluno.*

2. *Se a resposta do aluno estiver correta suas conexões permanecem inalteradas e um novo exemplo é apresentado. Volta ao passo* **1.**

3. *Se a resposta do aluno estiver errada, calculam-se os campos locais em ambas as unidades escondidas, e modifica-se a conexão cujo valor absoluto é mínimo, seguindo a prescrição do algoritmo perceptron.*

4. *Volta ao passo* **1.**

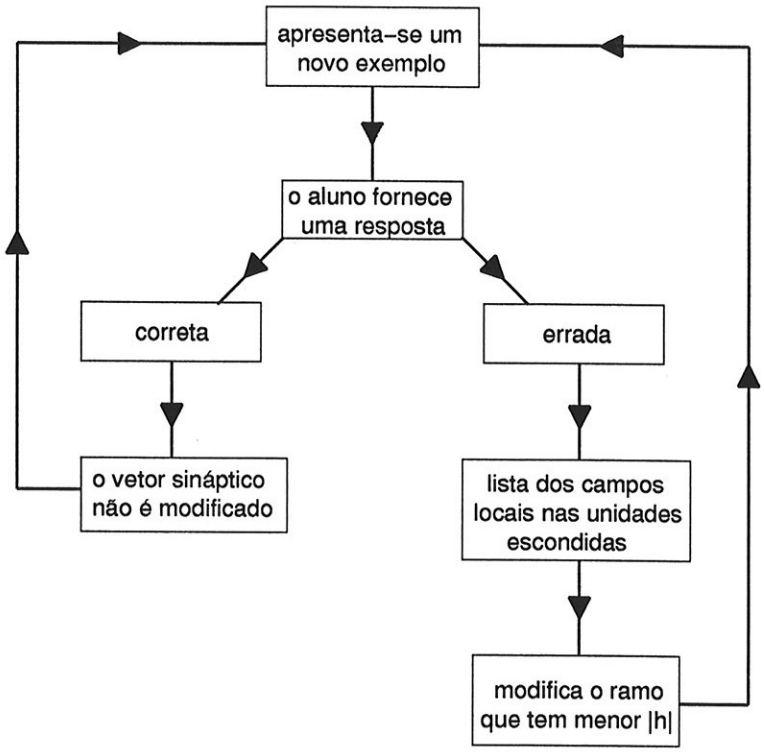


Figura 3.1: Procedimento esquemático do algoritmo de mínima ação

Na versão original este procedimento é iterado ciclicamente até que a rede aluno tenha memorizado todo o conjunto de exemplos, ou quando o número de iterações tenha atingido um limite superior pré-estabelecido. No próximo capítulo discutiremos os principais resultados da aplicação deste algoritmo a uma máquina paridade ($K = 2$).

3.1.3 O algoritmo ótimo

Vimos que a escolha do algoritmo de Hebb não fornece à rede a capacidade de generalizar. É preciso escolher a função modulação de maneira mais inteligente, tentando extrair o máximo de informação possível sobre o professor contida em cada um dos exemplos do conjunto de aprendizagem. Essa é a essência da idéia da otimização.

Maximizar o ganho com cada exemplo implica na maximização de $\dot{\rho}$. Assim a determinação da função modulação ótima F_k^μ torna-se um problema de cálculo variacional. Vamos reescrever o problema de maneira que seja possível a utilização de técnicas do cálculo variacional para encontrar os pontos estacionários de uma função. Assumimos a existência de um conjunto de funções:

$$F_k^\mu = \tilde{F}_k^\mu + \epsilon\, \eta(b_k^\mu, h_k^\mu)\,, \tag{3.19}$$

onde função $\tilde{F}_k^\mu$ é aquela que maximiza $\dot{\rho}$. A função $\eta(b_k^\mu, h_k^\mu)$ não precisa ser definida a priori, basta assumir inicialmente que seja uma função integrável. Substituindo (3.19) na equação (3.16), temos

$$\dot{\rho}_k = \frac{\rho_k}{J_k} \int D\Gamma \left[\frac{< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}}}{\rho_k} (\tilde{F}_k^\mu + \epsilon\,\eta) - (\tilde{F}_k^\mu + \epsilon\,\eta)\, h_k^\mu - \frac{(\tilde{F}_k^\mu + \epsilon\,\eta)^2}{2KJ_k} \right]\,, \tag{3.20}$$

onde $D\Gamma$ é definido por

$$D\Gamma = \sum_{\Sigma_B^\mu = \pm 1} \int P^K(\Sigma_B^\mu, \{h_j^\mu\}) \left[\prod_{j=1}^{K} dh_j^\mu \right]\,. \tag{3.21}$$

Dessa forma $\dot{\rho}$ se torna uma função do parâmetro infinitesimal ϵ. O nosso objetivo é encontrar a função $\tilde{F}_k^\mu$ tal que a variação infinitesimal $\delta\dot{\rho}$, definida como

$$\delta\dot{\rho} = \left(\frac{d\dot{\rho}}{d\epsilon} \right)_{\epsilon=0} d\epsilon\,, \tag{3.22}$$

seja nula. Derivando (3.20) em relação à ϵ, e impondo $\epsilon = 0$, obtemos:

$$\frac{d\dot{\rho}_k}{d\epsilon} = \frac{\rho_k}{J_k} \int \eta \left[\frac{< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}}}{\rho_k} - h_k^\mu - \frac{\tilde{F}_k^\mu}{KJ_k} \right] D\Gamma\,. \tag{3.23}$$

Para que $\delta\dot{\rho}$ seja nulo, dada a arbitrariedade de η, basta impor que o termo entre colchetes no integrando seja nulo, ou seja

$$\left[\frac{< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}}}{\rho_k} - h_k^\mu - \frac{\tilde{F}_k^\mu}{KJ_k} \right] = 0\,. \tag{3.24}$$

Esse procedimento nos leva à função modulação ótima:

$$\tilde{F}_k^\mu = KJ_k \left(\frac{< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}}}{\rho_k} - h_k^\mu \right) . \tag{3.25}$$

Substituindo $< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}}$ *dado pela equação (3.17), podemos escrever:*

$$\tilde{F}_k^\mu = KJ_k \int P^K(b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}) \left(\frac{b_k^\mu}{\rho_k} - h_k^\mu \right) db_k^\mu , \tag{3.26}$$

onde

$$P^K(b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}) = \frac{P^K(b_k^\mu,\Sigma_B^\mu,\{h_j^\mu\})}{P^K(\Sigma_B^\mu,\{h_j^\mu\})} \tag{3.27}$$

é a probabilidade de obter um certo valor de b_k^μ*, conhecendo a saída do professor* (Σ_B^μ) *para um dado exemplo* μ *e os campos locais do aluno* $(\{h_j^\mu\})$*, e* $P^K(\Sigma_B^\mu,\{h_j^\mu\})$ *a probabilidade conjunta, que pode ser escrita como:*

$$P^K(\Sigma_B^\mu,\{h_j^\mu\}) = P^K(\Sigma_B^\mu|\{h_j^\mu\})\, P^K(\{h_j^\mu\}) . \tag{3.28}$$

A expressão dada pela equação (3.25) é válida para qualquer escolha da distribuição de exemplos, embora seja a função modulação ótima no caso desta distribuição levar a um erro de generalização que decresce monotonicamente com ρ*. Ela é bastante geral também porque as variáveis conhecidas podem ser outras, apesar de terem sido definidas a priori nesta dedução. Isso implica que esta mesma função pode ser utilizada no caso do aprendizado na presença de ruído, só que neste caso uma das grandezas conhecidas é a saída corrompida por ruído,* ξ*, no lugar da saída correta,* Σ_B*.*

No próximo capítulo descreveremos com mais detalhes a máquina paridade e usaremos estes resultados para obter a forma da função modulação a ser implementada no seu aprendizado.

4 Máquina paridade caso K=2

Neste capítulo tratamos o caso específico da máquina paridade com duas unidades escondidas. Apresentamos a função modulação ótima e discutimos sua interpretação, deduzimos as equações de evolução do overlap com o professor e do módulo do vetor sináptico. Discutimos o comportamento do erro de generalização e apresentamos as simulações desta rede, comparando estes resultados com os existentes na literatura.

4.1 Características gerais da máquina paridade $K = 2$

A máquina paridade é uma rede neural multicamada simples com: N unidades de entrada S_{ki} ($k = 1, 2$ e $i = 1, ..., N/2$); 2 unidades escondidas e uma unidade de saída Σ.

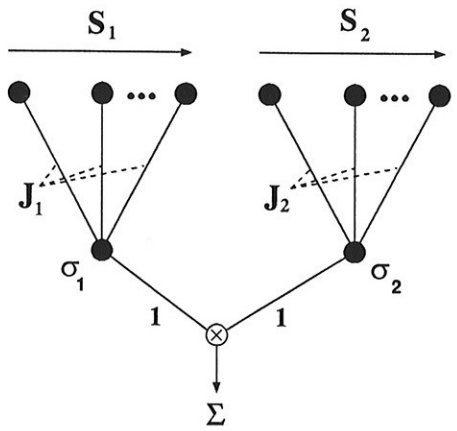


Figura 4.1: Representação esquemática de uma MP.

Esta rede possui arquitetura de árvore, isto significa que ela pode ser dividida em 2 ramos que não compartilham os exemplos. Cada uma das 2 unidades escondidas está ligada a uma porção do vetor de entrada através do vetor $N/2$-dimensional denotado por $\mathbf{J}_k$, com $k = 1, 2$, que é modificado ao longo do processo de aprendizagem. As conexões entre as unidades escondidas e a saída são mantidas fixas e iguais a um. O que caracteriza esta máquina é a maneira com que ela obtém a saída, que é função do produto dos campos locais em cada um dos ramos:

$$\Sigma_J = \mathcal{F}\left(\prod_{k=1}^{2} h_k\right) = \mathcal{F}\left(h_1 \cdot h_2\right) , \tag{4.1}$$

sendo o campo local na k-ésima unidade escondida (h_k) é definido por:

$$h_k = \sum_{i=1}^{N/2} \frac{J_{ki}\, S_{ki}}{|\mathbf{J}_k|}. \tag{4.2}$$

A função $\mathcal{F}$ que aparece na equação da saída (4.1) é a função de transferência, que nesse caso é a função sinal. J_{ki} são as $N/2$ componentes do vetor sináptico $\mathbf{J}_k$ e podem assumir qualquer valor real; $S_{ki} \in \{+1, -1\}$ são as componentes do vetor de entrada $\mathbf{S}_k$.

O objetivo de um algoritmo de aprendizagem é obter o vetor $\mathbf{J}$ que implementa as saídas corretamente para o maior número de exemplos possível. Esta performance, no caso

do aprendizado supervisionado com professor, é medida comparando-se a saída da rede com a saída dada pelo professor:

$$\Sigma_B = sinal\left(\prod_{k=1}^{2} b_k\right) = sinal\,(b_1 \cdot b_2) \tag{4.3}$$

onde b_k é o campo local no professor, dado por:

$$b_k = \sum_{i=1}^{N/2} \frac{B_{ki}\, S_{ki}}{|\mathbf{B}_k|} \qquad com \qquad |\mathbf{B}_k| = 1. \tag{4.4}$$

A escolha da função sinal faz com que as saídas $\Sigma_{B(J)}$ assumam apenas dois valores: $+1$ e -1, caracterizando as redes booleanas.

Para verificar o aprendizado, medimos o erro de generalização, que é o valor esperado do número de respostas incorretas, isto é, quando $\Sigma_J \neq \Sigma_B$, calculado sobre todos os exemplos possíveis, para que esta medida independa de um dado conjunto de exemplos. Podemos escrever o erro de generalização como:

$$e_G = \langle e_G\,(\mathbf{S}^\mu)\rangle_{\{\mathbf{S}^\mu\}}\ . \tag{4.5}$$

Computar erros na máquina paridade implica em contar quantas vezes o campo local num dos ramos da rede aluno difere do campo local num dos ramos da rede professor. Isto pode ser escrito através da função de Heaviside:

$$e_G^{MP} = \left\langle \Theta\left(-\prod_{k=1}^{2} b_k^\mu h_k^\mu\right)\right\rangle_{\{\mathbf{S}^\mu\}}\ . \tag{4.6}$$

A dependência nos exemplos está embutida nos campos h_k^μ e b_k^μ, de modo que podemos tomar a média sobre os exemplos como uma média sobre a distribuição destes campos:

$$e_G^{MP} = \int_{\Sigma_J \neq \Sigma_B} \prod_{k=1}^{2} db_k^\mu\, dh_k^\mu\, P(b_k^\mu, h_k^\mu)\ , \tag{4.7}$$

onde $P(b_k^\mu, h_k^\mu)$ é a probabilidade de distribuição conjunta dos campos b_k^μ e h_k^μ. De acordo com o teorema do limite central, quando $N \to \infty$, b_k^μ e h_k^μ são variáveis gaussianas com média zero e correlação ρ_k^μ . Portanto:

$$P(b_k^\mu, h_k^\mu) = \frac{1}{2\pi\sqrt{1-(\rho_k^\mu)^2}}\ \exp\left\{\frac{-(b_k^\mu)^2 - (h_k^\mu)^2 + 2h_k^\mu b_k^\mu \rho_k^\mu}{2(1-(\rho_k^\mu)^2)}\right\}\ , \tag{4.8}$$

sendo ρ_k^μ o overlap entre o aluno e o professor definido por:

$$\rho_k^\mu = \frac{\mathbf{B}_k \cdot \mathbf{J}_k^\mu}{|\mathbf{J}_k^\mu|}\ . \tag{4.9}$$

Usando a equação (4.7), é fácil verificar que o erro de generalização é dado por:

$$e_G^{MP} = e_G^1 + e_G^2 - 2e_G^1 e_G^2 \,, \tag{4.10}$$

onde $e_G^{1(2)}$ é o erro de generalização no ramo 1(2), dado por $e_G = arccos(\rho_{1(2)})/\pi$. Este resultado pode ser interpretado de maneira simples e intuitiva. O erro de generalização mede a probabilidade da saída da rede estar incorreta. Nessa rede, a saída Σ será incorreta sempre que um dos ramos fornecer a resposta errada e o outro, a correta, e vice-versa. Matematicamente, esta afirmação pode ser escrita da seguinte maneira:

$$e_G^{MP} = e_G^1(1 - e_G^2) + e_G^2(1 - e_G^1) \,. \tag{4.11}$$

O primeiro termo corresponde à probabilidade do ramo 1 errar e o ramo 2 acertar, o segundo termo representa a situação inversa. É fácil verificar que as equações (4.10) e (4.11) são idênticas. Considerando o caso simétrico, isto é, quando o overlap coincide nos dois ramos ($\rho_1 = \rho_2$) a expressão em (4.10) se reduz a:

$$e_G^{MP} = 2e_G - 2(e_G)^2 \,. \tag{4.12}$$

A figura 4.2 mostra o comportamento do erro de generalização em função do overlap.

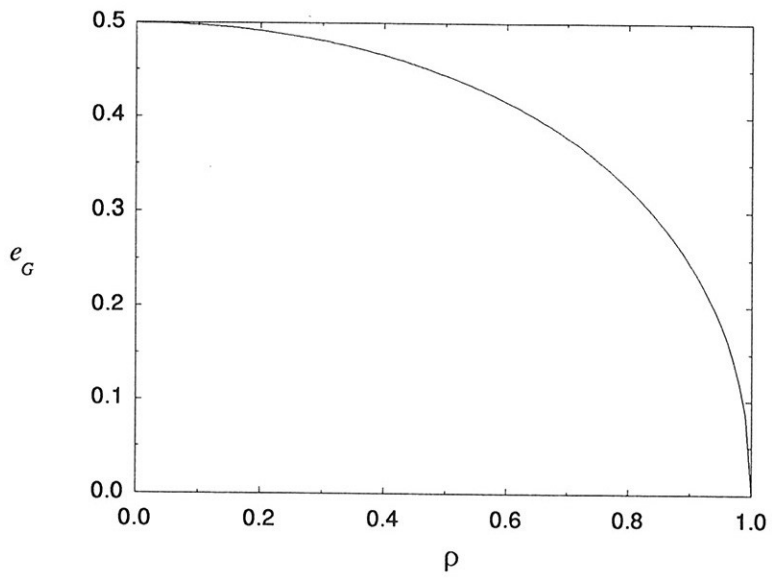


Figura 4.2: Erro de generalização *versus* overlap para a MP $K = 2$ (caso simétrico)

Observamos que o erro de generalização é uma função monotônica decrescente do overlap. Isso significa que a função modulação obtida através do método variacional será

a função ótima, porque cada exemplo apresentado provoca a máxima mudança no overlap, aproximando-o do seu valor máximo, que corresponde ao valor mínimo de e_G. Na próxima seção apresentamos o algoritmo de mínima ação e os resultados de sua aplicação no aprendizado da máquina paridade. Em seguida investigamos a performance desta rede quando se utiliza o algoritmo ótimo.

4.2 O algoritmo de mínima ação e a máquina paridade

Vamos retomar a discussão iniciada na seção 3.1.2, e verificar o desempenho do algoritmo de mínima ação na aprendizado da máquina paridade. Para Kabashima as componentes do vetor sináptico são modificadas à medida que um novo exemplo é apresentado, de acordo com a regra de aprendizagem dada pela equação (3.1), cujo incremento $\Delta \mathbf{J}_k^\mu$ é:

$$\Delta \mathbf{J}_k^\mu = -\Theta(-\Sigma_B \prod_{i=1}^{2} \mathbf{J}_i^\mu \cdot \mathbf{S}_i^\mu) \cdot \Theta(|\mathbf{J}_j^\mu \cdot \mathbf{S}_j^\mu| - |\mathbf{J}_k^\mu \cdot \mathbf{S}_k^\mu|) \cdot sinal(\mathbf{J}_k^\mu \cdot \mathbf{S}_k^\mu) \cdot \mathbf{S}_k^\mu \tag{4.13}$$

onde $k, j = 1, 2$ $(k \neq j)$ e $\Theta(x)$ é a função de Heaviside. Comparando esta expressão com (3.3), vemos que a função modulação no ramo 1 é dada pelo produto de três termos:

$$F_1^{LAA} = \Theta(-\Sigma_B h_1^\mu h_2^\mu) \cdot \Theta(|h_1^\mu| - |h_2^\mu|) \cdot [-sinal(h_1^\mu)] \ , \tag{4.14}$$

com $h_i^\mu = \mathbf{J}_i^\mu \cdot \mathbf{S}_i^\mu$. O primeiro termo é responsável pela manutenção do vetor sináptico, já que ele se anula para $\Sigma_B = h_1 h_2$, isto é, quando as respostas do professor e aluno coincidem. O segundo termo corresponde à mínima ação: ele garante que o ramo a ser modificado é aquele cujo módulo do campo local é menor. Finalmente o terceiro termo garante a inversão do sinal do campo local no ramo modificado, de modo que a saída possa ser invertida.

A figura 4.3 mostra o comportamento típico da função modulação F_1^{LAA} e de F_2^{LAA} em função do campo h_1^μ. Para exemplificar o que ocorre, escolhemos o caso específico onde $h_2^\mu = -2$ e $\Sigma_B = -1$. Neste caso, a saída dada pela rede aluno será positiva se o campo local no ramo 1 for negativo, e neste caso a rede estará errando. Portanto espera-se que nesta situação ocorram mudanças no vetor sináptico. Pelo gráfico observamos que na região onde $h_1^\mu > 0$, a função modulação é nula em ambos os ramos, pois nesse caso $\Sigma_B = \Sigma_J$, a rede aluno acerta e portanto não "precisa" ser modificada. Já na região em que $h_1^\mu < 0$ observam-se mudanças, isto é, a função modulação num dos ramos é diferente de zero. A questão é determinar qual ramo será modificado. Pela prescrição do LAA, será modificado o ramo cujo valor absoluto do campo local for menor, portanto enquanto $|h_1^\mu| < 2$ este será o vetor modificado e quando esta situação se inverter será modificado o vetor sináptico do ramo 2.

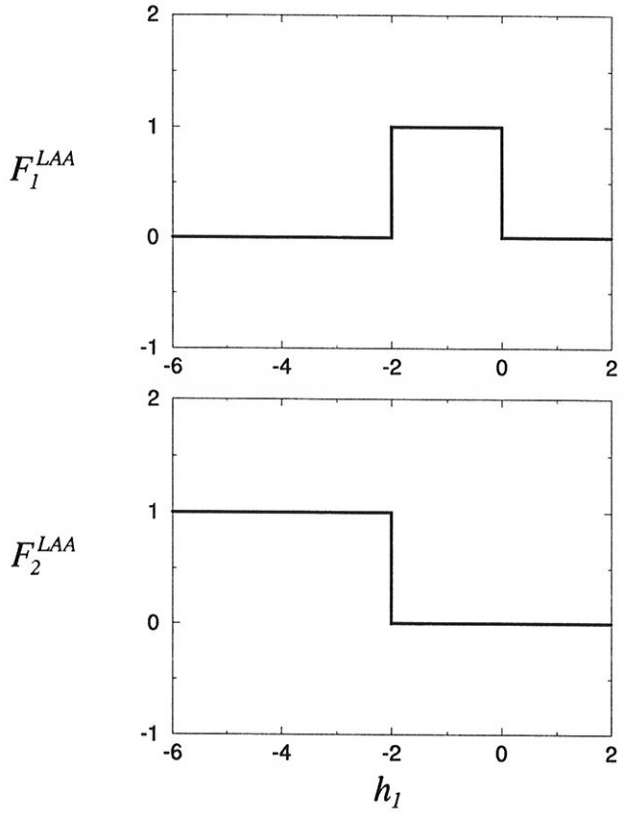


Figura 4.3: A figura mostra a função modulação nos ramos 1 e 2 em função do campo local h_1 para $h_2 = -2$ e $\Sigma_B = -1$.

As equações diferenciais que descrevem a evolução do módulo do vetor sináptico e do overlap podem ser obtidas de maneira análoga à utilizada na seção 3.1. Kabashima observou que ambas dependem do balanço entre dois fatores. Para valores pequenos do vetor sináptico $|\mathbf{J}| \approx 0$, $\dot{\rho}$ é negativo, indicando que nestas condições o overlap necessariamente diminui. Essa característica é explicada por Kabashima da seguinte forma: "quando um vetor é modificado numa direção aleatória, perde-se informação numa outra direção específica". Entretanto, esse balanço só implica na diminuição do overlap para valores pequenos de $|\mathbf{J}|$, existindo um valor crítico $|\mathbf{J}_c|$ acima do qual $\dot{\rho}$ torna-se positivo e portanto a rede passa a generalizar. Nos estágios iniciais do aprendizado ($\rho \sim 0$), as equações de evolução podem ser aproximadas por:

$$\frac{dJ}{d\alpha} \sim \frac{1}{J}\left[\frac{1}{4} - \frac{J(\sqrt{2}-1)}{\sqrt{\pi}}\right] \tag{4.15}$$

$$\dot{\rho} = \frac{d\rho}{d\alpha} \sim -\frac{\rho}{J^2}\left[\frac{1}{4} - \frac{2J}{\pi\sqrt{\pi}}\right] . \tag{4.16}$$

Impondo a condição $\dot{\rho} > 0$, encontra-se o valor crítico $J_c \approx 0.696$ a partir do qual o aluno começa a adquirir a capacidade de generalizar. Assim, enquanto $J < J_c$ a rede passa por uma fase de simples memorização, na qual o valor do overlap não varia.

Uma maneira de verificar o desempenho do algoritmo é estudar o comportamento assintótico do erro de generalização, verificando como, em função do número de exemplos, seu valor mínimo é atingido. No limite $\alpha \to \infty$, espera-se que o overlap atinja seu valor máximo (1) e que, portanto, possa ser escrito como $\rho = 1 - \delta$, com $\delta << 1$. Neste limite as equações de evolução podem ser aproximadas por:

$$\frac{dJ}{d\alpha} \sim \frac{1}{J}\left[\frac{\sqrt{2\delta}}{\pi} - \frac{\sqrt{2}}{\pi}\delta J\right] \tag{4.17}$$

$$\frac{d\delta}{d\alpha} \sim -\frac{1}{J^2}\left[\frac{\sqrt{2\delta}}{\pi} - \frac{4}{\sqrt{2\pi}}\delta J\right] \tag{4.18}$$

cuja solução assintótica é $\delta \sim 0.24\alpha^{-2/3}$.

Como vimos na seção anterior, o erro de generalização para uma máquina paridade é, no caso simétrico, dado por $e_G^{MP} = 2e - 2e^2$, onde $e = \frac{1}{\pi}arccos(\rho)$ é o erro de generalização em um dos ramos. Quando o erro se aproxima de zero, o termo linear domina, e podemos escrever a aproximação:

$$e_G^{MP} \sim \frac{2}{\pi}arccos(\rho)\ . \tag{4.19}$$

Invertendo esta expressão, usando a expansão do cosseno para argumentos pequenos e as equações acima, obtemos o erro de generalização assintótico:

$$e_G^{MP} \sim 0.441\alpha^{-1/3}. \tag{4.20}$$

Kabashima também investiga o aprendizado na presença de ruído, cujos resultados serão apresentados no capítulo 6. Vamos estudar o aprendizado numa máquina paridade quando se utiliza o algoritmo ótimo e confrontar o resultado com o obtido nesta seção.

4.3 O algoritmo ótimo e a máquina paridade

No capítulo anterior obtivemos a expressão que determina a função modulação a ser implementada no k-ésimo ramo de uma rede neural. O resultado obtido é bastante geral e válido para qualquer rede neural com estrutura de árvore e K ramos:

$$\tilde{F}_k^\mu = KJ_k^\mu\left(\frac{< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}}}{\rho_k} - h_k^\mu\right)\ . \tag{4.21}$$

Como observado anteriormente, uma condição necessária para que esta função modulação seja ótima é que o erro de generalização decresça monotonicamente com o overlap. Como mostramos (seção 4.1) esta condição é satisfeita no caso da máquina paridade $K = 2$. O valor esperado do campo b_k^μ é calculado levando-se em conta o conhecimento que se tem sobre as diferentes variáveis do problema, que são os campos locais da rede aluno (h_1^μ, h_2^μ) e a saída do professor Σ_B^μ.

Neste ponto podemos considerar duas situações. Na primeira, para modificar o ramo k, o aluno só tem acesso ao valor do campo no k-ésimo ramo, isto é, no próprio ramo, e não nos demais. Nesta situação o algoritmo tem caráter local e o valor esperado do campo local no professor é calculado a partir de:

$$< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu, h_k^\mu} = \int b_k^\mu \, P(b_k^\mu|\Sigma_B^\mu, h_k^\mu) \, db_k^\mu \, . \tag{4.22}$$

Realizando este cálculo chegamos ao resultado trivial $F_k^\mu = 0$, que leva a: $\dot{\rho}_k/\rho_k = \dot{J}_k/J_k = 0$. Este resultado indica que a máquina paridade não é capaz de aprender quando se utiliza um algoritmo local de maneira on-line. Isso explica porque o algoritmo de Hebb (seção 3.1.1) não funciona na máquina paridade: para que o aprendizado ocorra é preciso que se conheçam os campos locais nos outros ramos, isto é, o algoritmo precisa ser não local. Esta é a segunda situação a ser considerada.

A obtenção da função modulação[1] *requer que seja especificada a distribuição de probabilidade dos campos b_k, dados os campos h_k e a saída Σ_B, deve-se portanto especificar a probabilidade condicional $P(b_k|\Sigma_B, h_1, h_2)$. Esta probabilidade, obtida no apêndice A, é dada por:*

$$P(b_1|\Sigma_B, h_1, h_2) = \frac{P(b_1|h_1)}{P(\Sigma_B|h_1, h_2)} H\left(-\Sigma_B \frac{\sigma_{b_1} h_2}{\lambda_2}\right) \tag{4.23}$$

onde

$$P(\Sigma_B|h_1, h_2) = H\left(\Sigma_B \frac{h_2}{\lambda_2}\right) H\left(\frac{h_1}{\lambda_1}\right) + H\left(-\Sigma_B \frac{h_2}{\lambda_2}\right) H\left(-\frac{h_1}{\lambda_1}\right) \, . \tag{4.24}$$

e $\sigma_{b_k} = sinal(b_k)$. Nestas expressões foram também usadas as definições de λ_k:

$$\lambda_k = \rho_k^{-1}\sqrt{1 - \rho_k^2} \, , \tag{4.25}$$

e da função $H(x)$:

$$H(x) = \frac{1}{2} \, erfc\left(\frac{x}{\sqrt{2}}\right) = \int_{-\infty}^{x} Dx \, . \tag{4.26}$$

[1]aproveitando a simetria do problema, vamos calcular a função modulação no ramo 1. Para o ramo 2 basta trocar os índices $1 \leftrightarrow 2$.

A combinação de todos estes resultados fornece a melhor função modulação para modificar o vetor sináptico no ramo 1:

$$\tilde{F}_1 = \frac{2J_1\lambda_1}{\sqrt{2\pi}}\Sigma_B e^{-\frac{h_1^2}{2\lambda_1^2}} \frac{H(-\frac{h_2}{\lambda_2}) - H(\frac{h_2}{\lambda_2})}{H\left(\Sigma_B\frac{h_2}{\lambda_2}\right)H\left(\frac{h_1}{\lambda_1}\right) + H\left(-\Sigma_B\frac{h_2}{\lambda_2}\right)H\left(-\frac{h_1}{\lambda_1}\right)} . \tag{4.27}$$

4.3.1 A função modulação ótima: interpretação

A primeira característica importante a ser notada na forma estrutural da função modulação ótima é que há uma dependência direta no valor dos campos locais. Isso significa que esta função modulação depende do exemplo apresentado, sendo dado um peso diferente a cada exemplo. Como vimos na seção 4.2, esta característica já estava presente no LAA, entretanto naquele caso a função modulação se restringe aos valores 0 ou 1, e no caso do algoritmo ótimo isso não ocorre.

A segunda característica importante, e que não está presente no LAA, é que a função modulação depende explicitamente do valor de ρ. Isto implica que ela evolui com o tempo, isto é, à medida que novos exemplos são apresentados e o overlap muda de valor. De outra maneira podemos dizer que a função modulação leva em conta o estado de aprendizagem da rede. Este fato pode ser facilmente observado na figura 4.4, que mostra a função modulação nos ramos 1 e 2 em função do campo local no ramo 1, para um valor fixo do campo local no ramo 2 ($h_2 = -3$), da saída ($\Sigma_B = -1$) e diferentes valores do overlap.

O interessante da dependência de $\tilde{F}$ com o overlap é que nos estágios iniciais da aprendizagem, isto é, para valores pequenos de ρ, por exemplo para $\rho = 0.1$ (linhas sólidas), o peso dado aos exemplos varia muito pouco, bem diferente do que ocorre quando o overlap se aproxima de seu valor máximo, por exemplo para $\rho = 0.7$ (linhas tracejadas). Essa característica pode ser interpretada da seguinte forma: a "atenção" dispensada pelo aluno a cada exemplo depende do desempenho. No início ele "presta a mesma atenção" a todos os exemplos, mas à medida que ele aprende torna-se "mais crítico", selecionando aqueles exemplos que lhe fornecem informações mais relevantes.

Outra característica importante a se perceber através destes gráficos é a diferença entre as formas de $\tilde{F}_1$ e $\tilde{F}_2$. Vemos que os ramos sofrem diferentes mudanças para um mesmo exemplo apresentado. Vamos discutir com mais detalhes esta questão.

Confiança e Surpresa

O objetivo do aprendizado é encontrar uma rede que imite o professor, isto é, que forneça as respostas corretas. Dessa forma, é de se esperar que as mudanças mais signifi-

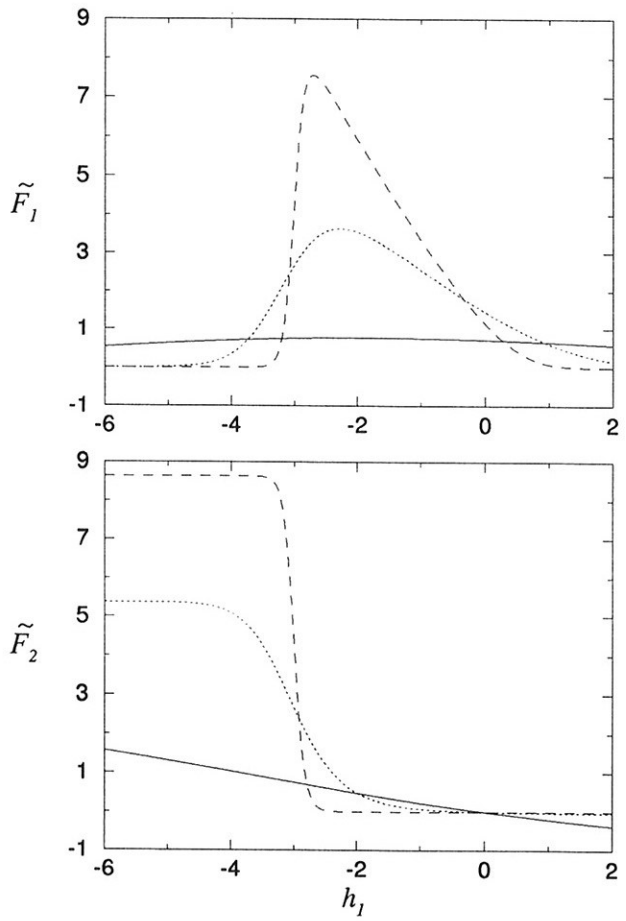


Figura 4.4: Função modulação ótima nos ramos 1 e 2 em função do campo local h_1 para h_2 e Σ_B fixos, e diferentes valores do overlap: 0.1(——), 0.4(- - - - -), 0.7(· · · · · ·).

cativas sejam efetuadas quando a resposta dada pelo aluno estiver errada. Isso ocorre nas seguintes situações:

$$\begin{matrix}\Sigma_B = +1 \\ \Sigma_J = -1\end{matrix} \Rightarrow \begin{cases} h_1 > 0 & h_2 < 0 \\ h_1 < 0 & h_2 > 0 \end{cases} \qquad \begin{matrix}\Sigma_B = -1 \\ \Sigma_J = +1\end{matrix} \Rightarrow \begin{cases} h_1 > 0 & h_2 > 0 \\ h_1 < 0 & h_2 < 0 \end{cases}$$

Na máquina paridade, que é uma rede booleana, as modificações nas conexões sinápticas devem ser feitas com o objetivo de inverter a saída da rede. A questão é: qual dos dois ramos é mais conveniente modificar?

Para responder a essa pergunta, e facilitar a análise do papel desempenhado pela função modulação, é conveniente definir a função modulação reescalada $\tilde{F}_1/J_1\lambda_1$ e os campos reescalados $\tilde{h}_1 \equiv h_1/\lambda_1$ e $\tilde{h}_2 \equiv h_2/\lambda_2$. Dessa forma elimina-se a dependência no parâmetro

de ordem ρ.

As figuras 4.5 e 4.6 mostram o comportamento das funções modulação reescaladas de acordo com o valor do campo reescalado $\tilde{h}_1$, para dois valores do campo reescalado no ramo 2 ($\tilde{h}_2 = -5$ e $\tilde{h}_2 = +4$, respectivamente) e $\Sigma_B = -1$. Na figura 4.5 observamos que as mudanças mais significativas ocorrem na região em que $\tilde{h}_1 < 0$, já que neste caso $\tilde{h}_1 \cdot \tilde{h}_2 > 0$. Na figura 4.6 ocorre o contrário, isto é, as maiores mudanças ocorrem para $\tilde{h}_1 > 0$, pois nesse caso também $\tilde{h}_1 \cdot \tilde{h}_2 > 0$, diferindo portanto da saída do professor.

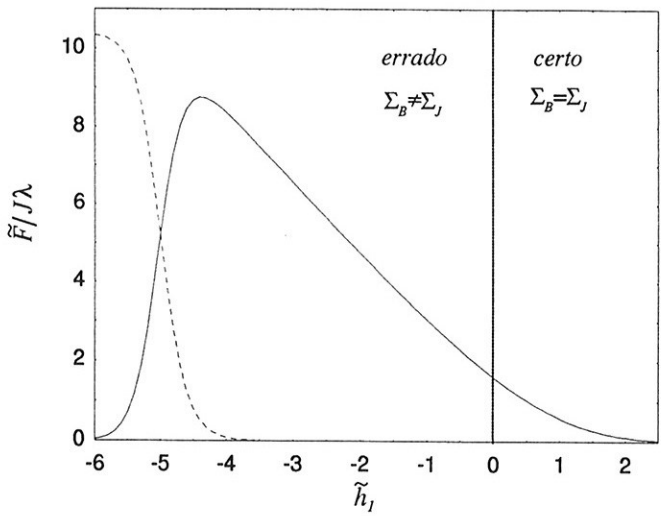


Figura 4.5: Função modulação ótima reescalada nos ramos 1 (linha cheia) e 2 (linha tracejada) em função do campo local reescalado $\tilde{h}_1$ para $\tilde{h}_2 = -5$ e $\Sigma_B = -1$

Podemos relacionar o valor do módulo do campo local com a *confiança* que o aluno tem num dos seus ramos. Quanto maior for esse valor, mais confiança ele tem de que aquele ramo esteja correto. A confiança está também relacionada com a dificuldade em classificar um dado exemplo. Se $|h|$ for grande, o exemplo apresentado é "fácil" de classificar. Por outro lado um valor pequeno do campo $|h|$ representa um exemplo "difícil". Esta interpretação vem do fato que quando $|h|$ é pequeno qualquer pequena variação pode alterar sua resposta. Tal comportamento pode ser entendido como uma "falta de confiança" do aluno na sua resposta.

É nesse contexto que surge a idéia de *surpresa*: o aluno confia demais na sua resposta mas, surpreendentemente, ela está errada. Este é um tipo bom de experiência, na medida em que produz as modificações mais significativas no estado da rede aluno.

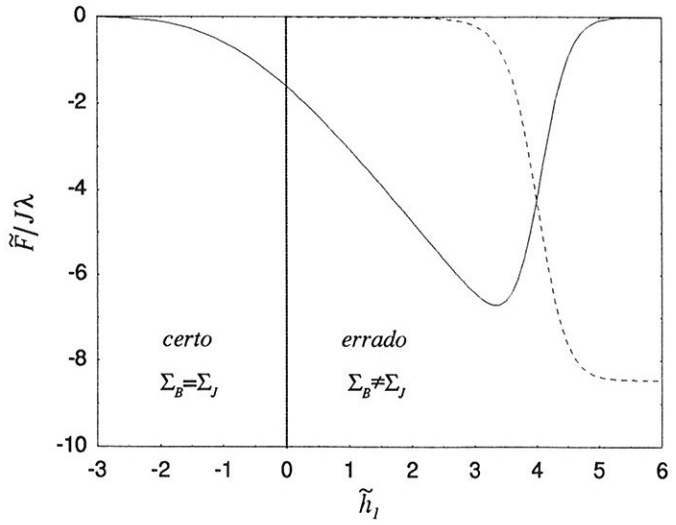


Figura 4.6: Função modulação reescalada nos ramos 1 (linha cheia) e 2 (linha tracejada) em função do campo local reescalado $\tilde{h}_1$ para $\tilde{h}_2 = +4$ e $\Sigma_B = -1$

A questão é: que valor de $|h|$ determina o "estado de confiança" do aluno na sua resposta? Observamos que este valor depende diretamente da relação entre os campos nos dois ramos, h_1 e h_2. Existem três situações possíveis:

- $|h_1| > |h_2| \Rightarrow$ *nesse caso o campo no ramo 1 é maior em módulo do que em 2, e portanto o ramo 1 é "mais confiável" e deve estar correto. Então, a maneira mais fácil de alterar a resposta do aluno, tornando-a igual à do professor, é modificar as conexões do ramo 2, no qual o aluno tem "menos confiança".*

- $|h_1| < |h_2| \Rightarrow$ *esse caso é o oposto do primeiro. Agora o aluno confia mais no ramo 2 e portanto as mudanças mais significativas devem ocorrer no ramo 1.*

- $|h_1| = |h_2| \Rightarrow$ *nesse caso o aluno confia igualmente nos dois ramos. Esta é a situação de aprendizagem mais difícil e delicada: ambos os ramos são igualmente modificados.*

É fácil observar pelas figuras 4.5 e 4.6 que enquanto $|h_1| < |h_2|$ o ramo 1 é o que sofre maiores modificações. Na região em que $|h_1| > |h_2|$ a situação se inverte. Finalmente, quando $|h_1| \sim |h_2|$ ocorrem modificações de mesma magnitude em ambos os ramos. É interessante notar a semelhança que existe entre essas funções e aquelas obtidas através do LAA (figura 4.3). Como já mencionamos, algumas características da função modulação ótima estão também presentes na F^{LAA}.

4.3.2 As equações de evolução

O nosso objetivo agora é calcular as equações de evolução de J_k e do overlap ρ_k usando a função modulação obtida. Para o ramo[2] 1, substituindo o resultado da equação (4.27) nas equações (3.13) and (3.14), e após fazer a soma Σ_B e a integração em b_1 chega-se ao resultado :

$$\frac{d\rho_1}{d\alpha} = \frac{\rho_1\lambda_1^2}{\pi}\int Dh_1 Dh_2\, e^{-h_1^2/\lambda_1^2}\frac{\left[H(-\frac{h_2}{\lambda_2}) - H(\frac{h_2}{\lambda_2})\right]^2}{H\left(\frac{h_2}{\lambda_2}\right)H\left(\frac{h_1}{\lambda_1}\right) + H\left(-\frac{h_2}{\lambda_2}\right)H\left(-\frac{h_1}{\lambda_1}\right)} \tag{4.28}$$

$$\frac{dJ_1}{d\alpha} = \frac{J_1\lambda_1^2}{\pi}\int Dh_1 Dh_2\, e^{-h_1^2/\lambda_1^2}\frac{\left[H(-\frac{h_2}{\lambda_2}) - H(\frac{h_2}{\lambda_2})\right]^2}{H\left(\frac{h_2}{\lambda_2}\right)H\left(\frac{h_1}{\lambda_1}\right) + H\left(-\frac{h_2}{\lambda_2}\right)H\left(-\frac{h_1}{\lambda_1}\right)}\,, \tag{4.29}$$

onde $Dh_k = dh_k\,(2\pi)^{-1}\,\exp\{-h_k^2/2\}$.

Existem dois pontos importantes neste resultado a serem enfatizados. O primeiro é que o valor inicial do overlap $\rho_k(0)$ não pode ser zero, pois neste caso obtém-se a solução trivial: $d\rho_k/d\alpha = 0$. Isto significa que a rede necessita de um conhecimento prévio ($\rho_k(0) \neq 0$) sem o qual ela não será capaz de generalizar. Esse comportamento é típico desta rede não sendo observado no caso do perceptron simples. Sua origem está relacionada à simetria interna da máquina paridade. Para efeito de comparação, na figura 4.7 mostramos o gráfico $\dot{\rho} \times \rho$ para o perceptron simples e para a máquina paridade $K = 2$.

Observamos que no caso do perceptron simples a variação do overlap nos estágios iniciais de aprendizagem tende a infinito. A condição $\rho(0) \neq 0$ não torna menos atraente o algoritmo ótimo, uma vez que este valor pode ser tão pequeno quanto se queira[3]. A partir da integração numérica da equação (4.28) obtém-se o comportamento do overlap em função de α, e usando a equação (4.12) é possível conhecer como o erro de generalização varia em função de α. Esse comportamento está indicado na figura 4.8, no caso simétrico, para diferentes valores de $\rho_k(0)$. Por esse gráfico vê-se que quanto mais conhecimento a priori a rede possuir, mais rápido será o decaimento do erro de generalização, embora que assintoticamente a taxa de decaimento independa das condições iniciais. Esse assunto será discutido adiante.

O segundo ponto importante é que vale a relação $\dot{J}_k/J_k = \dot{\rho}_k/\rho_k$. Essa relação, igualmente obtida do caso de outras dinâmicas otimizadas (perceptron e máquina comitê), é satisfeita pela solução $J_k(\alpha) = c\rho_k(\alpha)$, cuja constante c pode ser escolhida como sendo igual à unidade. Esta equação nos mostra que, se for satisfeita a condição inicial $J_k(0) = \rho_k(0)$, esta igualdade será preservada pela dinâmica. Também indica que o overlap, cujo valor é desconhecido uma vez que não se conhece o professor, é em média igual ao módulo do vetor

[2] também neste caso, para obter as equações de evolução no ramo 2 basta trocar os índices $1 \leftrightarrow 2$.

[3] Vamos discutir mais sobre esse assunto na seção 4.4

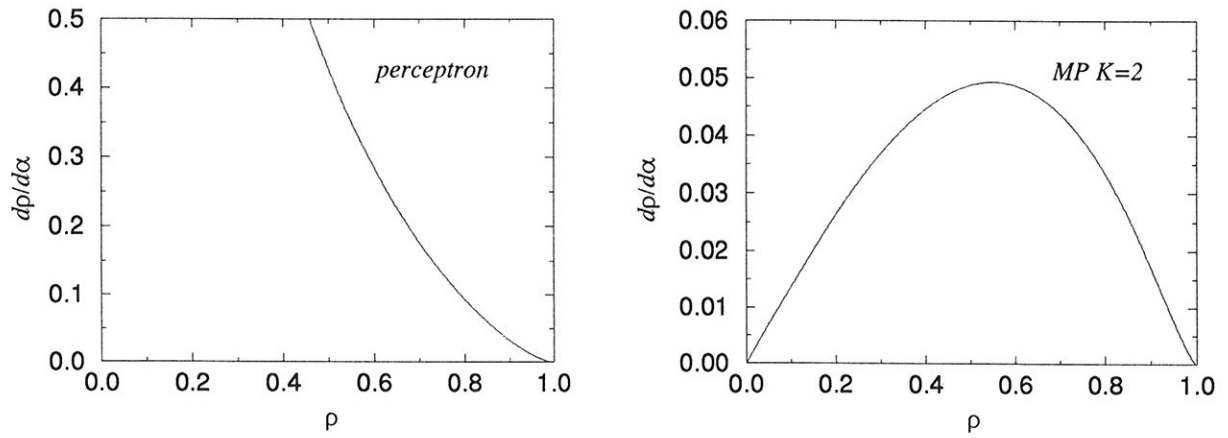


Figura 4.7: Os gráficos mostram o comportamento $\dot{\rho}$ em função de ρ. O gráfico à esquerda indica o comportamento típico de um perceptron simples: nos estágios iniciais do aprendizado $\dot{\rho} \to \infty$; o gráfico à direita, o diferente comportamento da máquina paridade: um conhecimento a priori ($\rho(0) \neq 0$) é necessário para dar início ao processo de aprendizagem.

sináptico J_k, e pode ser substituído pelo seu valor durante o processo de aprendizagem. Essa possibilidade é fundamental para efeitos de implementação do algoritmo ótimo.

Comportamento assintótico de $\dot{\rho}$ e de e_G para $\alpha \to \infty$

Nosso interesse agora é investigar o comportamento de $d\rho/d\alpha$ quando o número de exemplos apresentados é muito grande, isto é, para $\alpha \to \infty$. Nesta situação espera-se que o valor de ρ seja próximo de um. Podemos considerar que o overlap evolui de maneira uniforme nos dois ramos, de modo que a seguinte análise serve tanto para $k = 1$ como para $k = 2$. Vamos portanto assumir o caso simétrico: $\rho_1 \approx \rho_2 = \rho$ e $\lambda_1 \approx \lambda_2 = \lambda$. No limite $\rho_{1,2} \to 1$ temos $\lambda_{1,2} \to 0$, como se pode ver pela equação (4.25). Partimos do resultado fornecido pela equação (4.28), isto é:

$$\frac{d\rho}{d\alpha} = \frac{\rho\lambda^2}{\pi} \int Dh_1 Dh_2 \, e^{-h_1^2/\lambda^2} \frac{\left[H\left(-\frac{h_2}{\lambda}\right) - H\left(\frac{h_2}{\lambda}\right)\right]^2}{H\left(\frac{h_2}{\lambda}\right) H\left(\frac{h_1}{\lambda}\right) + H\left(-\frac{h_2}{\lambda}\right) H\left(-\frac{h_1}{\lambda}\right)} \,. \tag{4.30}$$

Se tomarmos diretamente nesta equação o limite $\lambda \to 0$ obtemos a solução trivial: $d\rho/d\alpha = 0$. Para contornar este problema, e investigar o que ocorre no limite $\rho \sim 1$, fazemos a mudança de variável $x = h_1/\lambda$. Dessa forma podemos tomar o limite $\lambda \to 0$ mais adiante.

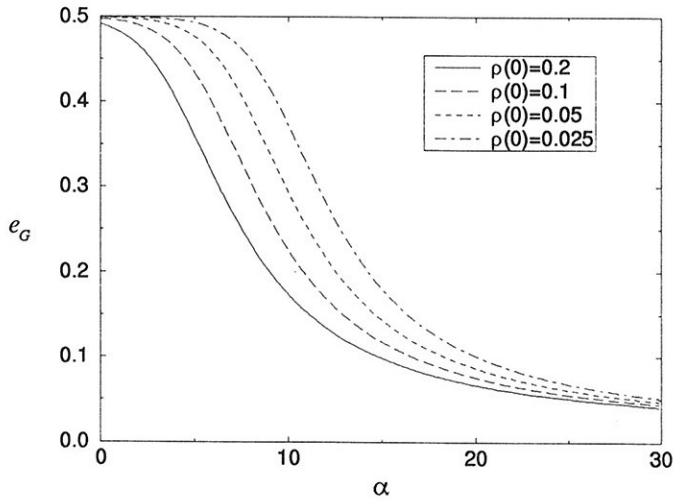


Figura 4.8: Erro de generalização versus α para diferentes valores iniciais ρ_0.

Reescrevendo a equação (4.30) em termos da nova variável e da relação entre λ e ρ, obtemos:

$$\frac{d\rho}{d\alpha} = \frac{(1-\rho)^{3/2}}{\pi\rho^2} \int DxDh_2\, e^{-x^2/2} \frac{\left[H\left(-\frac{h_2}{\lambda}\right) - H\left(\frac{h_2}{\lambda}\right)\right]^2}{H\left(\frac{h_2}{\lambda}\right) H(x) + H\left(-\frac{h_2}{\lambda}\right) H(-x)} \,. \tag{4.31}$$

No limite $\lambda \to 0$ podemos utilizar as aproximações: $H(-\infty) = 1$ e $H(+\infty) = 0$, e chegamos ao limite assintótico da equação de evolução do overlap:

$$\frac{d\rho}{d\alpha} \approx \frac{(1-\rho^2)^{3/2}}{\rho^2} \frac{I_0}{\pi} \,, \tag{4.32}$$

onde a integral I_0 foi calculada numericamente, tendo sido encontrado o resultado:

$$I_0 = \int \frac{dx}{\sqrt{2\pi}} \frac{e^{-x^2}}{H(x)} \approx 2.27 \,. \tag{4.33}$$

Nosso interesse é saber qual a dependência do overlap com α, para isso podemos supor que, quando $\alpha \to \infty$, o overlap vai para 1 com alguma potência (γ) de α:

$$\rho \approx 1 - A\alpha^{-\gamma} \,. \tag{4.34}$$

Neste limite podemos desprezar os termos de ordens superiores e escrever $\rho^2 \approx 1 - 2A\alpha^{-\gamma}$. Derivando a equação (4.34) em relação à α obtemos:

$$\frac{d\rho}{d\alpha} \approx A\gamma\alpha^{-\gamma-1} \,. \tag{4.35}$$

Igualando esta equação à (4.32), obtemos os valores $A = \pi^2/2I_0^2 \approx 0.96$ *e* $\gamma = 2$. *Substituindo estes resultados na equação (4.34) obtemos uma forma para o comportamento assintótico do overlap:*

$$\rho \approx 1 - \frac{\pi^2}{2I_0^2}\alpha^{-2} . \tag{4.36}$$

Com esse resultado podemos obter o comportamento assintótico do erro de generalização em função de α. *Lembrando que para cada ramo* $e_G(\rho) = \pi^{-1}\arccos(\rho)$, *invertendo essa relação e expandindo o cosseno para argumentos pequenos obtemos:*

$$e_G(\alpha) \sim \frac{\sqrt{2A}}{\pi\alpha} \approx \frac{0.44}{\alpha} . \tag{4.37}$$

Inserindo este valor na equação (4.10) e desprezando termos de ordem $\mathcal{O}(\alpha^{-2})$, *obtemos finalmente o limite assintótico do erro de generalização da máquina paridade* $K = 2$ *em função de* α:

$$e_G^{MP}(\alpha) \approx \frac{0.88}{\alpha} . \tag{4.38}$$

Esse resultado é interessante por duas razões. Em primeiro lugar ele reflete o mesmo comportamento assintótico obtido a partir da otimização em outras redes (Kinouchi & Caticha, 1992; Copelli & Caticha 1995), e é duas vezes o erro previsto para o algoritmo de Bayes para um perceptron (Opper & Haussler, 1991). Por outro lado ele deve ser confrontado com o resultado obtido através da aplicação do LAA. Vê-se claramente que no limite de $\alpha \to \infty$ *o algoritmo ótimo tem uma performance muito superior do que o LAA, para o qual o erro de generalização vai para zero com* $\alpha^{-1/3}$. *Isto significa que se para obter um dado valor de* e_G *forem necessários* α_{ot} *exemplos usando o algoritmo ótimo, se usarmos o LAA serão necessários* $\alpha_{LAA} = (\alpha_{ot}/2)^3$ *exemplos. Por exemplo, uma máquina treinada de forma ótima com 20 exemplos terá o mesmo desempenho de uma máquina treinada com LAA após 1000 exemplos!*

4.4 Simulações

As simulações foram feitas em linguagem FORTRAN usando uma estação de trabalho SPARC5 da SUN. O tempo de CPU para estas simulações é bem reduzido, tendo sido os dados obtidos praticamente de forma interativa.

Como mencionamos na seção 4.3.2, a rede aluno necessita de um conhecimento prévio para ser capaz de generalizar. Isto pode ser visto através da equação de evolução que se anula para $\rho = 0$, *e implica que o overlap permanecerá sempre com o mesmo valor, portanto* $\rho = 0$ *é um ponto fixo da dinâmica.*

A pergunta que podemos fazer agora é: quanto conhecimento prévio é necessário? Em outras palavras: qual o mínimo valor inicial do overlap para que a rede seja capaz de generalizar?

Para responder a essa questão, vamos verificar a ordem de grandeza do overlap quando o professor e o aluno são escolhidos de maneira aleatória e independente. Seja o professor dado pelo vetor normalizado $\mathbf{B}$, *cujas componentes são* $B_i = \pm 1/\sqrt{N}$; *e uma rede aluno dada pelo vetor* $\mathbf{J}$, *cujas componentes são* $J_i = \pm a$ *e cujo módulo vale* $a\sqrt{N}$, *de forma que* $< J_i >= 0$ *e* $< B_i >= 0$. *O valor médio de* ρ *será zero com variância dada por* $\sigma^2(\rho)$:

$$\sigma^2(\rho) = \frac{1}{|\mathbf{B}|^2|\mathbf{J}|^2} \sum_{i=1}^{N} \sigma^2(B_i J_i) \,. \tag{4.39}$$

Como B_i *e* J_i *são independentes vale* $\sigma(B_i J_i) = \sigma(B_i)\,\sigma(J_i)$, *e portanto:*

$$\sigma^2(\rho) = \frac{1}{B^2 J^2} \sum_{i=1}^{N} \sigma^2(B_i)\sigma^2(J_i) \,. \tag{4.40}$$

A variância de B_i *e de* J_i *são dadas, respectivamente, por:*

$$\sigma^2 =< B_i^2 > - < B_i >^2 = \frac{1}{N}\sum_{i=1}^{N} B_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} B_i\right)^2 \approx \mathcal{O}\left(\frac{1}{N}\right) \,, \tag{4.41}$$

$$\sigma^2 =< J_i^2 > - < J_i >^2 = \frac{1}{N}\sum_{i=1}^{N} J_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} J_i\right)^2 \approx a^2 = \mathcal{O}(1) \,. \tag{4.42}$$

Portanto

$$\sigma(\rho) = \sqrt{\sigma^2(\rho)} \approx \frac{1}{\sqrt{N}} \,. \tag{4.43}$$

Assim, ao sortear uma rede aluno, seu overlap com o professor será aleatório (ρ_{aleat}) *e terá ordem de magnitude* $\mathcal{O}(1/\sqrt{N})$. *Entretanto, a equação de evolução do overlap é correta em ordem* $\mathcal{O}(1)$. *Assim, a ordem de grandeza da precisão do overlap na equação é maior do que o overlap inicial da rede aluno usada na simulação, isto é,* $\rho \gg \rho_{aleat}$. *Para manter a precisão das equações é necessário que o conhecimento prévio* ρ_0 *seja* $\mathcal{O}(1)$. *Por outro lado é interessante que a rede não possua muito conhecimento, pois esta é uma situação comum nos problemas reais, portanto o overlap inicial deve satisfazer:*

$$\frac{1}{\sqrt{N}} \ll \rho_0 \ll 1. \tag{4.44}$$

Isso mostra que ρ_0 *depende do tamanho* N *da rede. Para investigar o desempenho do algoritmo através de sua simulação, precisamos compará-la com o resultado numérico*

das equações. A integração numérica da equação (4.28) nos fornece o valor de $\rho(\alpha)$ com o qual podemos calcular o erro de generalização também em função de α. Esta integração requer o estabelecimento de uma condição inicial, que serve como base para a escolha do aluno na simulação. É claro que ao sortear a rede aluno não é possível encontrar um que tenha exatamente o valor desejado do overlap, assim escolhemos dois valores que delimitam os valores possíveis do overlap da rede aluno sorteada.

O intervalo desejado de ρ é determinado pela equação (4.44) e depende do tamanho N da rede. Para selecionar alunos com overlap menores, precisamos trabalhar com redes muito grandes, o que traz dificuldades computacionais. Portanto, a escolha do tamanho da rede e do intervalo de overlap devem ser feitas levando-se em consideração este problema.

O gráfico da figura 4.9 mostra o resultado da simulação (pontos) para duas redes de diferentes tamanhos (N), que devem partir de diferentes escolhas de ρ_0, e o erro de generalização teórico.

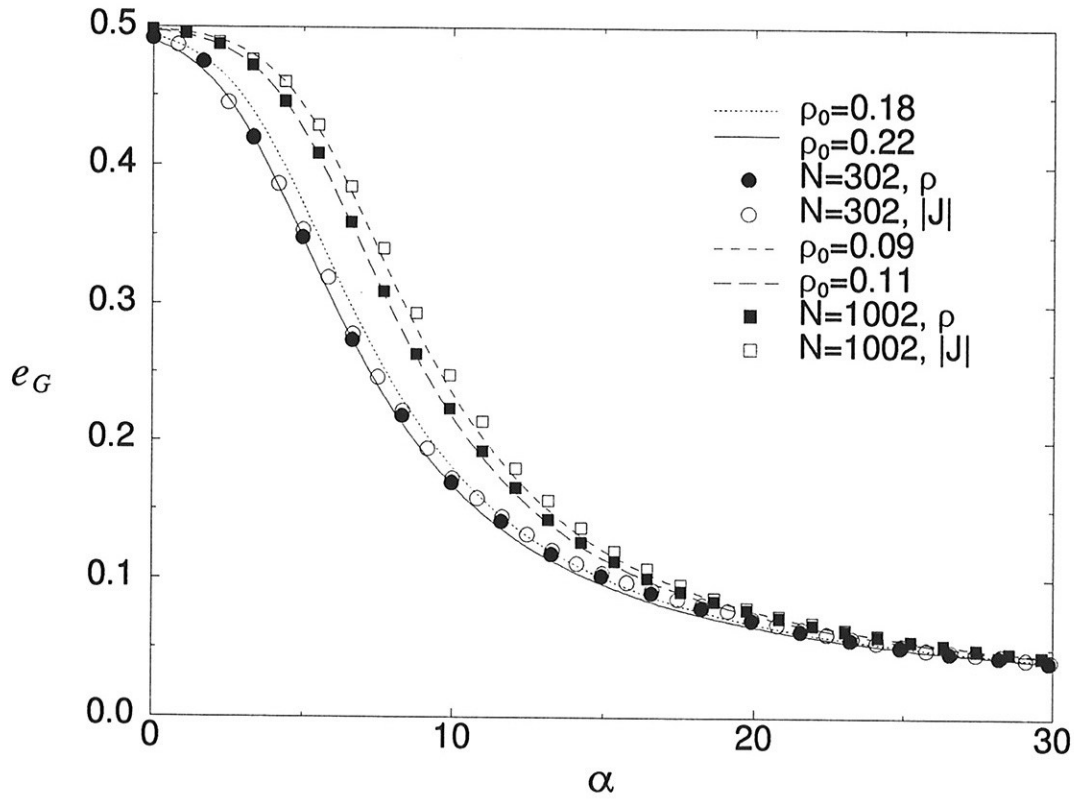


Figura 4.9: Erro de generalização obtido através da integração numérica de $d\rho/d\alpha$ (linhas), simulação usando ρ (símbolos cheios) e J (símbolos vazios) para N=302 e N=1002.

Para os tamanhos de rede selecionados $N = 302$ *(*$\rho_{aleat} \approx 0.06$*) e* $N = 1002$ *(*$\rho_{aleat} \approx 0.03$*), foram escolhidos alunos cujo overlap inicial pertencia aos intervalos* $[0.18 - 0.22]$ *e* $[0.09 - 0.11]$*, respectivamente. As curvas foram obtidas a partir da integração numérica da equação de evolução do overlap utilizando como valor inicial os limites inferior e superior dos intervalos citados.*

É importante reforçar o seguinte ponto: na expressão da função modulação aparece explicitamente o valor do overlap ρ*. Entretanto, este overlap não é conhecido, uma vez que não se tem conhecimento do professor, e portanto, esta grandeza não pode ser usada na simulação. Este não é exatamente um problema pois, como podemos observar a partir das equações de evolução, o overlap e o módulo do vetor* $\mathbf{J}$ *diferem apenas por uma constante, de tal forma que podemos usar o módulo de* $\mathbf{J}$*, ao qual temos acesso, no lugar do overlap. Para efeito de verificação, realizamos simulações utilizando ambas as grandezas. Os pontos indicados pelos símbolos cheios foram calculados usando-se o overlap, e coincidem com os pontos indicados pelos símbolos vazios, que foram obtidos usando-se o módulo de* $\mathbf{J}$*.*

Pelo gráfico da figura (4.9) observamos também que quanto menor for o valor inicial do overlap, mais lento será o decaimento do erro de generalização. A origem deste retardamento, também observada na dinâmica off-line, deve-se à simetria presente neste tipo de rede (Hansel et al., 1992): se em um número par de ramos as conexões $\mathbf{J}$ *forem transformadas em* $-\mathbf{J}$ *a saída não se altera, e portanto não se altera* e_G*. Estes platôs também ocorrem em máquinas comitê totalmente conectadas e nestes casos sua origem também está relacionada às simetrias entre os diferentes ramos. Devido à existência deste ponto fixo pode-se questionar a utilidade de descrever a dinâmica através de um conjunto de equações diferenciais, entretanto deve-se levar em conta que afastando-se da origem este é um excelente método para descrever e compreender a dinâmica de aprendizagem destas redes.*

5 Máquina paridade caso K geral

Neste capítulo estendemos os resultados obtidos no capítulo anterior. Introduzimos um conjunto de relações de recorrência para as distribuições de probabilidade relevantes, o que permite estudar o caso de máquinas paridade com um K geral. Determinamos a função modulação ótima, discutimos o comportamento do erro de generalização, seu limite assintótico e seu comportamento próximo à origem.

5.1 Extensão dos resultados

Os resultados do capítulo anterior podem ser estendidos para uma máquina paridade com um número K qualquer de unidades escondidas. O erro de generalização pode ser calculado a partir da expressão:

$$e_G^{(K)}(\rho_1 ... \rho_K) = \int \prod_{k=1}^{K} db_k \, dh_k \, P(b_k, h_k) \theta \left(- \prod_{i=1}^{K} b_i h_i \right) . \tag{5.1}$$

Uma máquina paridade com K ramos pode ser entendida como uma máquina de 2 ramos: um perceptron, que é uma máquina paridade com K=1 e uma outra máquina paridade com $K-1$ ramos. Dessa forma, usando o resultado do erro de generalização para uma máquina paridade $K = 2$, dado pela equação (4.10), podemos escrever o erro de generalização da máquina paridade K geral através da relação de recorrência:

$$e_G^{(K)}(\rho_1 ... \rho_K) = e_G^{(K-1)}(\rho_1 ... \rho_{K-1}) + e_G^{(1)}(\rho_K) - 2 e_G^{(1)}(\rho_K) e_G^{(K-1)}(\rho_1 ... \rho_{K-1}) , \tag{5.2}$$

cuja solução é:

$$e_G^{(K)}(\rho_1 ... \rho_K) = \frac{1}{2} \left[1 - \prod_{k=1}^{K} \left(1 - 2 e_G^{(1)}(\rho_k) \right) \right] , \tag{5.3}$$

onde $e_G^{(1)}(\rho_k) = \pi^{-1} \arccos(\rho_k)$ é o erro de generalização do k-ésimo ramo, que é um perceptron simples. Isto reforça a idéia que cada um dos ramos da máquina paridade corresponde a um perceptron simples. No caso simétrico, que significa tomar $\rho_k = \rho$ para todo k na equação (5.1), obtemos os seguintes valores para e_G no limite $K \to \infty$

$$\begin{aligned} e_G^{\infty}(\rho) &= \frac{1}{2} \quad \text{para} \quad \forall \ \rho < 1 \\ e_G^{\infty}(\rho) &= 0 \quad \text{para} \quad \rho = 1 . \end{aligned}$$

É fácil verificar estes resultados: para os valores de $\rho < 1$, o erro de generalização pertence ao intervalo $]0, \frac{1}{2}]$ e portanto $(1 - 2e_G) < 1$. Dessa forma o termo $(1 - 2e_G)^K$ se anula no limite $K \to \infty$ e obtemos o resultado $e_G^{\infty}(\rho) = \frac{1}{2}$. Quando o overlap atinge seu valor máximo, $\rho = 1$, o erro de generalização é nulo e portanto $(1 - 2e_G)^K = 1$, assim obtemos o resultado $e_G^{\infty}(1) = 0$. Esse comportamento também pode ser observado pelo gráfico da figura 5.1 que mostra o erro de generalização em função do overlap, no caso simétrico, para redes com diferentes valores de K: 1, 2, 3, 4, 5, 6, 8, 10, 12, 14, 16, 18 e 20.

Pelo gráfico podemos ainda observar que à medida que aumenta o número de camadas internas, a rede demora mais para generalizar, isto é, o erro de generalização só começa a diminuir para valores cada vez maiores de ρ. O ponto importante para as análises que se seguem é o fato que ainda assim o erro de generalização é monotônico em ρ.

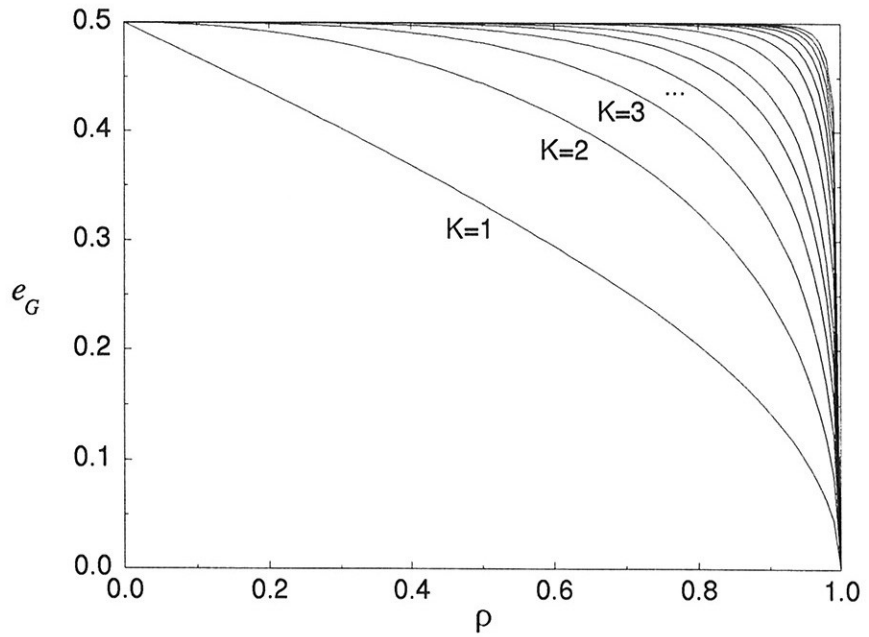


Figura 5.1: Erro de generalização da máquina paridade em função do overlap para um número K de unidades escondidas: o erro de generalização, apesar de decair mais lentamente, é monotônico em ρ, independente do valor de K.

5.2 As equações dinâmicas e o peso ótimo

As equações de evolução obtidas no capítulo 3 são gerais e servem para qualquer rede multicamada com estrutura de árvore. Dessa forma, a expressão geral da função modulação ótima no k-ésimo ramo é a mesma, e portanto podemos partir do resultado:

$$\tilde{F}_k^\mu = KJ_k \left(\frac{< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}}}{\rho_k} - h_k^\mu \right) . \tag{5.4}$$

O valor esperado dos campos internos b_k^μ, são obtidos levando-se em conta a informação contida nos campos locais da rede aluno, e portanto depende de K. Esta dependência se dá através da probabilidade condicional $P^K(b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\})$, pois

$$< b_k^\mu >_{b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}} = \int b_k^\mu \, P^K(b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}) \, db_k^\mu . \tag{5.5}$$

Dessa forma, o primeiro passo para calcular a função modulação ótima é obter as probabilidades conjuntas $P^K(b_k^\mu,\Sigma_B^\mu,\{h_j^\mu\})$ e $P^K(\Sigma_B^\mu,\{h_j^\mu\})$, já que:

$$P^K(b_k^\mu|\Sigma_B^\mu,\{h_j^\mu\}) = \frac{P^K(b_k^\mu,\Sigma_B^\mu,\{h_j^\mu\})}{P^K(\Sigma_B^\mu,\{h_j^\mu\})} . \tag{5.6}$$

Detalhes da obtenção destas probabilidades podem ser encontrados no apêndice A. Abandonando o índice μ, por simplicidade, obtemos o resultado:

$$P^K(b_k, \Sigma_B, \{h_j\}) = P(h_k, b_k) \{ \delta_{\sigma_{b_k},1} P^{K-1}(\Sigma_B, \{h_j\}_{j\neq k}) + \delta_{\sigma_{b_k},-1} P^{K-1}(-\Sigma_B, \{h_j\}_{j\neq k}) \} \tag{5.7}$$

onde

$$P^{K-1}(\pm\Sigma_B, \{h_j\}_{j\neq k}) = \int \prod_{j=1}^{K-1} db_j \, P(h_j, b_j)\theta\left(\pm\Sigma_B \prod_{j=1}^{K-1} \sigma_{b_j}\right) \tag{5.8}$$

A probabilidade $P^K(\Sigma_B, \{h_j\})$ pode ser obtida integrando $P^K(b_k, \Sigma_B, \{h_j\})$ em b_k. Dessa forma obtemos as seguintes relações de recorrência para $P^K(\Sigma_B, \{h_j\})$ e $P^K(\Sigma_B|\{h_j\})$:

$$P^K(\Sigma_B, \{h_j\}) = H\left(-\frac{h_k}{\lambda_k}\right) P(h_k) P^{K-1}(\Sigma_B, \{h_j\}_{j\neq k}) + H\left(\frac{h_k}{\lambda_k}\right) P(h_k) P^{K-1}(-\Sigma_B, \{h_j\}_{j\neq k}) \tag{5.9}$$

$$P^K(\Sigma_B|\{h_j\}) = H\left(-\frac{h_k}{\lambda_k}\right) P^{K-1}(\Sigma_B|\{h_j\}_{j\neq k}) + H\left(\frac{h_k}{\lambda_k}\right) P^{K-1}(-\Sigma_B|\{h_j\}_{j\neq k}) \,, \tag{5.10}$$

onde λ_k é definido por $\lambda_k = \rho_k^{-1}\sqrt{1-\rho_k^2}$. Finalmente obtemos a probabilidade condicional procurada:

$$P^K(b_k|\Sigma_B, \{h_j\}) = \frac{P(b_k|h_k)\,[\,\delta_{\sigma_{b_k},1} P^{K-1}(\Sigma_B|\{h_j\}_{j\neq k}) + \delta_{\sigma_{b_k},-1} P^{K-1}(-\Sigma_B|\{h_j\}_{j\neq k})\,]}{H\left(\frac{h_k}{\lambda_k}\right) P^{K-1}(-\Sigma_B|\{h_j\}_{j\neq k}) + H\left(-\frac{h_k}{\lambda_k}\right) P^{K-1}(\Sigma_B|\{h_j\}_{j\neq k})} \,. \tag{5.11}$$

Usando este resultado na equação (5.4) chegamos à função modulação ótima no ramo k:

$$\tilde{F}_k = \frac{K}{\sqrt{2\pi}} J_k \lambda_k e^{-h_k^2/2\lambda_k^2} \frac{\left[P^{K-1}(\Sigma_B|\{h_j\}_{j\neq k}) - P^{K-1}(-\Sigma_B|\{h_j\}_{j\neq k})\right]}{P^K(\Sigma_B|\{h_j\})} \,, \tag{5.12}$$

com $P^K(\Sigma_B|\{h_j\})$ dado pela equação (5.10).

Através deste resultado e da equação (3.14), após fazer a soma em $\Sigma_B = \pm 1$, chega-se às equações de evolução procuradas:

$$\frac{d\rho_k}{d\alpha} = \frac{K}{2\pi} \rho_k \lambda_k^2 \int \left[\prod_{j=1}^{K} Dh_j\right] e^{-h_k^2/\lambda_k^2} \frac{\left[P^{K-1}(1|\{h_j\}_{j\neq k}) - P^{K-1}(-1|\{h_j\}_{j\neq k})\right]^2}{P^K(1|\{h_j\})} \tag{5.13}$$

e do módulo de $\mathbf{J}_k$:

$$\frac{dJ_k}{d\alpha} = \frac{K}{2\pi} J_k \lambda_k^2 \int \left[\prod_{j=1}^{K} Dh_j\right] e^{-h_k^2/\lambda_k^2} \frac{\left[P^{K-1}(1|\{h_j\}_{j\neq k}) - P^{K-1}(-1|\{h_j\}_{j\neq k})\right]^2}{P^K(1|\{h_j\})} \,. \tag{5.14}$$

Antes de prosseguir a discussão, é importante notar que a integral em (5.13) e (5.14) pode ser escrita como:

$$\int Dh_k \, e^{-h_k^2/\lambda_k^2} \frac{\prod_{j\neq k} Dh_j \left[H\left(-\frac{h_j}{\lambda_j}\right) - H\left(\frac{h_j}{\lambda_j}\right)\right]^2}{P^K(1|\{h_j\})} \,. \tag{5.15}$$

5.2.1 Solução próxima à origem: o ponto fixo $\rho_k = 0$

Os fatores $\left[H\left(-\frac{h_j}{\lambda_j}\right) - H\left(\frac{h_j}{\lambda_j}\right)\right]$, *em (5.15) existem sempre que* $K > 1$. *Notamos que eles se anulam para* $\rho_j = 0$, *pois nesse caso temos* $[H(0) - H(0)]$, *uma vez que* $\lambda_j = \infty$. *Isso indica que, se o processo de aprendizagem parte de uma situação onde* $\rho_j = 0$, *ele não funcionará de maneira on-line, pois* $\dot{\rho}_j = 0$.

Para investigar o comportamento nos primeiros estágios do aprendizado, isto é, para pequenos valores de α, *tomamos o limite* $\rho \approx 0$. *Usando a expansão da função* $H(x)$ *para argumentos pequenos* $H(x) \approx 1/2 - x/\sqrt{2\pi}$, *é possível escrever o numerador do integrando como:*

$$\prod_{l \neq k} Dh_l \left[\frac{2}{\pi} h_l^2 \rho_l^2\right] \tag{5.16}$$

e seu denominador assume o valor $1/2$. *Dessa forma temos*

$$\frac{d\rho_k}{d\alpha} \approx \frac{K}{\pi \rho_k} \int Dh_k \left(\frac{2}{\pi}\right)^{K-1} \left[\prod_{j \neq k} Dh_j \, h_j^2\right] \left[\prod_{j \neq k} \rho_j^2\right] . \tag{5.17}$$

Fazendo as integrais, podemos finalmente escrever a equação diferencial (5.13) para pequenos valores de α *como:*

$$\frac{d\rho_k}{d\alpha} \approx \frac{K}{2} \left(\frac{2}{\pi}\right)^K \frac{1}{\rho_k} \prod_{l \neq k} \rho_l^2 . \tag{5.18}$$

Através desta equação observamos que um perceptron $(K = 1)$ *que parte de* $\rho_k = 0$ *é capaz de aprender rapidamente, pois a derivada do overlap é infinita. Portanto, neste caso,* $\rho_k = 0$ *não constitui um ponto fixo atrativo da dinâmica. Para redes com* $K > 1$, *é fácil observar que* $\rho_k = 0$, *pode ser um ponto fixo, dependendo das condições iniciais nos outros* $K - 1$ *ramos.*

Se a rede não possui nenhum conhecimento inicial, isto é, $\rho_k(0) = 0$ *para todo* k, *ela não será capaz de aprender de maneira on-line, pois o overlap não evolui no tempo, já que sua derivada é nula. Entretanto isso não ocorre se* $\rho_k(0) \neq 0$, *em pelo menos* $K - 1$ *ramos, ainda que seja muito pequeno. Vamos considerar* $\rho_k^2(0) = \epsilon_k \neq 0$, *isto significa que existe um conhecimento inicial, e definir* $E_k = \prod_{i \neq k} \epsilon_i$. *Nestas condições, a evolução inicial do overlap pode ser escrita como uma série de potências:*

$$\begin{aligned} \rho_k &= \left[\epsilon_k + E_k \Lambda_K \alpha + \frac{(\Lambda_K \alpha)^2}{2} \epsilon_k E_k^2 \sum_{i \neq k} \frac{1}{\epsilon_i^2}\right. \\ &+ \left.\frac{E_k^3}{6} (\Lambda_K \alpha)^3 \left(2\epsilon_k^2 \sum_{i \neq j \neq k} \frac{1}{\epsilon_i^2 \epsilon_j^2} + \sum_i \frac{1}{\epsilon_i^2}\right) + \mathcal{O}(\Lambda_K^4 \alpha^4)\right]^{\frac{1}{2}} , \end{aligned} \tag{5.19}$$

com $\Lambda_K = K(2/\pi)^K$.

O ponto fixo $\rho(0) = 0$ do aprendizado on-line está relacionado ao fenômeno de retardamento que ocorre no aprendizado off-line (Opper, 1994): o aprendizado só se inicia depois que um número crítico de exemplos ($\alpha_c N$) for apresentado. Nos casos em que N é finito, uma escolha inicial aleatória das conexões é suficiente para fornecer o conhecimento prévio necessário, já que neste caso $\rho(0) \approx \mathcal{O}(\sqrt{N})$, como discutido no capítulo 4.

5.2.2 Comportamento assintótico: $\rho \to 1$

Podemos, sem perda de generalidades, considerar que o overlap evolui de forma semelhante nos K ramos, de modo que é razoável a aproximação simétrica, isto é, $\rho_k = \rho$. No limite $\alpha \to \infty$ o overlap tende a um, e o erro de generalização pode ser escrito como[1]*:*

$$e_G^{(K)}(\rho) \approx \frac{K}{\pi} \arccos \rho \,. \tag{5.20}$$

Procedendo de maneira análoga à utilizada na seção 4.3.2, encontramos o comportamento da equação de evolução do overlap para $\alpha \to \infty$:

$$\frac{d\rho}{d\alpha} \approx \frac{K}{2\pi} \frac{(1-\rho^2)^{\frac{3}{2}}}{\rho^2} I_0 \,, \tag{5.21}$$

com I_0 definida em (4.33). Para obter o comportamento assintótico $e_G(\alpha)$, consideramos que $\rho \approx 1 - B\alpha^{-\theta}$, e portanto a derivada do overlap em relação a α é

$$\frac{d\rho}{d\alpha} \approx B\theta\alpha^{-\theta-1} \,. \tag{5.22}$$

Comparando as equações (5.21) e (5.22), encontramos os valores $\theta = 2$ e $B = 2(\pi/I_0K)^2$. Invertendo a equação (5.20) e utilizando a expansão do cosseno para argumentos pequenos, podemos escrever o erro de generalização em função de α como:

$$e_G^{(K)}(\alpha) \approx \sqrt{2B}\frac{K}{\alpha\pi} \approx \frac{C}{\alpha} \,, \tag{5.23}$$

com $C = (2/I_0) \approx 0.88$.

Essa equação nos mostra que no limite assintótico, o erro de generalização independe de K, e estende o resultado obtido na seção 4.3.2. Nos mostra também que a máquina

[1] *É fácil obter este resultado, partindo da equação (5.3) e expandindo o binômio $(1-2e)^K$ em série. Temos $(1-2e)^K = 1 - 2eC_{K,1} + 4e^2C_{K,2} - \ldots$. Como estamos interessados no limite $\rho \to 1$, e sabemos que neste limite $e \to 0$, podemos desprezar os termos de ordens superiores, mantendo só os dois primeiros, de modo que $e_G^{(K)} \approx \frac{1}{2}[1 - 1 + 2Ke] \approx Ke$.*

paridade apresenta um erro de generalização muito maior do que cada ramo individualmente, de fato neste limite seu erro é a soma dos erros de cada um dos ramos. É fácil observar este resultado a partir da relação de recorrência dada pela equação (5.2): no limite $\alpha \to \infty$, os termos dominantes são aqueles proporcionais a α^{-1}, isto é, só sobrevivem os termos lineares em e_G:

$$e_G^{(K)}(\rho_1 ... \rho_K) \approx e_G^{(1)}(\rho_1) + e_G^{(1)}(\rho_2) + e_G^{(1)}(\rho_3) + ... \tag{5.24}$$

que corresponde, no caso simétrico, a:

$$e_G^{(K)}(\rho) \approx K e_G^{(1)}(\rho) \, . \tag{5.25}$$

Da mesma forma, podemos escrever:

$$e_G^{(K)}(\alpha) \approx K e_G^{(1)}(\alpha) \approx K \frac{C}{\alpha_p} \, , \tag{5.26}$$

onde $\alpha_p = K\alpha$ é uma medida do número de exemplos dividido pelo número de acoplamentos de cada ramo. Nesta escala, cada um dos ramos aprende com a mesma taxa, independentemente do valor de K. Entretanto, a escala correta para se medir o decaimento do erro da máquina paridade é α e não α_p. Isto leva a um cancelamento dos $K's$, e fornece um erro nominal que independe de K. Esta característica é compartilhada também pela máquina comitê e vai ocorrer para qualquer rede com estrutura de árvore, desde que o aprendizado seja otimizado (Copelli, em preparação).

6 Aprendizagem na presença de ruído

Neste capítulo apresentamos os resultados da aplicação do LAA à máquina paridade $K = 2$. Analisamos a tolerância à ruído da máquina paridade K geral, quando o aprendizado é realizado de maneira on-line utilizando o algoritmo ótimo. Estudamos o comportamento do erro de generalização quando os exemplos são apresentados à rede aluno com um nível de ruído supostamente conhecido. Investigamos também situações mais realistas onde nem sempre é possível conhecer o nível de ruído presente na rede professor, nesses casos as características da dinâmica são descritas através dos chamados diagramas de robustez.

6.1 Ruído multiplicativo

Até este ponto tratamos a situação na qual o aprendizado ocorre a partir da apresentação correta dos exemplos. Esta é a situação ideal, entretanto nem sempre possível de ser obtida. Vamos supor agora que o aprendizado se dá na presença de ruído, de forma que os exemplos transmitidos ao aluno estão corrompidos por um certo nível de ruído. Como vimos no capítulo 2, o ruído pode afetar o aprendizado de várias formas: pode estar presente nas sinapses, nos exemplos ou provocar mudanças diretamente na saída do professor. Esta última será considerada neste capítulo.

Vamos assumir que a saída do professor seja invertida com uma certa probabilidade ε, **supostamente conhecida**. *Chamaremos de* ξ *a nova saída do professor, corrompida por ruído. Levando em conta estas informações podemos escrever:*

$$\begin{aligned} \xi = +\Sigma_B &\implies \text{com probabilidade } 1-\varepsilon \\ \xi = -\Sigma_B &\implies \text{com probabilidade } \varepsilon \end{aligned}$$

e a probabilidade de obter a saída ξ, *dada a saída correta do professor* Σ_B *é:*

$$P(\xi|\Sigma_B) = \varepsilon\,\delta(\xi + \Sigma_B) + (1-\varepsilon)\delta(\xi - \Sigma_B)\,. \tag{6.1}$$

Obviamente se $\varepsilon = 0$ *temos* $P(\xi|\Sigma_B) = \delta(\xi - \Sigma_B)$, *isto é,* $\xi = +\Sigma_B$, *com probabilidade um e recuperamos todos os resultados obtidos no capítulo anterior. Vamos inicialmente apresentar alguns resultados, obtidos por Kabashima, da utilização do LAA na presença de ruído, em seguida analisaremos o desempenho do algoritmo ótimo.*

6.1.1 O LAA na aprendizagem com ruído

Em seu trabalho, Kabashima não leva em conta a existência de ruído quando propõe o algoritmo de aprendizagem, considerando-o apenas na evolução dinâmica do overlap. Não vamos apresentar os detalhes das contas, mas sim discutir os comportamentos assintóticos, isto é, nos estágios iniciais de aprendizagem e após a apresentação de $\alpha \to$ *exemplos . No limite* $\rho \approx 0$ *as equações de evolução são:*

$$\frac{dJ}{d\alpha} \approx \frac{1}{J}\left[\frac{1}{4} - \frac{J(\sqrt{2}-1)}{\sqrt{\pi}}\right] \tag{6.2}$$

$$\frac{d\rho}{d\alpha} \approx -\frac{\rho}{J^2}\left[\frac{1}{4} - (1-2\varepsilon)\frac{2J}{\pi\sqrt{\pi}}\right]\,. \tag{6.3}$$

Este conjunto de equações mostra que o estado $\rho \approx 0$ *se torna instável quando* J *supera o valor crítico* $J_c = 0.696(1-2\varepsilon)^{-1}$, *que pode ser obtido impondo-se a condição*

$d\rho/d\alpha > 0$, ao mesmo tempo em que J converge para um ponto atrativo $J'_c \approx 1.069$, obtido a partir da condição $dJ/d\alpha < 0$. Quando $J_c < J'_c$, J é finalmente atraído para a região $J > J_c$, para qualquer condição inicial. Neste momento ρ começa a crescer. Estas equações descrevem corretamente a dinâmica apenas para valores pequenos de ρ, entretanto, quando ρ atinge um valor finito, o ponto atrativo da dinâmica de J é deslocado e atinge um valor superior, isso faz com que seja mantida a condição para o crescimento de J, garantindo que ρ atinja um valor finito. Isso ocorre para valores pequenos de ε.

Para valores maiores de ε a situação é diferente. Kabashima observou que existe um nível de ruído crítico $\varepsilon_c \approx 0.175$, determinado pela condição $J'_c < J_c$, acima do qual J converge para J'_c antes que ρ atinja um valor finito. Isso mostra que o estado $\rho = 0$ é um ponto fixo estável da dinâmica e implica que, nestas condições, o aluno perde totalmente a habilidade de generalização, mesmo no limite $\alpha \to \infty$.

6.1.2 O algoritmo ótimo na aprendizagem com ruído

Queremos investigar se o algoritmo ótimo é robusto a ruído multiplicativo, isto é, se nesta situação a máquina paridade continua sendo capaz de generalizar e com que eficiência ela o faz. Para isso, vamos estudar a evolução do overlap e o comportamento assintótico do erro de generalização. O primeiro passo é conhecer qual a função modulação mais adequada neste caso, isto é, a função modulação ótima. Vamos proceder da mesma forma que fizemos para o caso sem ruído. Procedendo de forma análoga àquela descrita no capítulo 3, podemos partir da função modulação ótima:

$$\tilde{F}_k = KJ_k \left(\frac{< b_k >_{b_k|\xi,\{h_j\}}}{\rho_k} - h_k \right) \tag{6.4}$$

onde agora as grandezas conhecidas sobre as quais é feita a média em b_k são: a saída corrompida ξ e os campos do aluno $\{h_j\}$, portanto $< b_k >_{\xi,\{h_j\}}$ é o valor esperado do campo no k-ésimo ramo do professor, calculado usando a probabilidade condicional $P^K(b_k|\xi,\{h_j\})$ que pode ser escrita como:

$$P^K(b_k|\xi,\{h_j\}) = \frac{P^K(b_k,\xi,\{h_j\})}{P^K(\xi,\{h_j\})} \tag{6.5}$$

e a probabilidade conjunta $P^K(\xi,\{h_j\})$ é:

$$P^K(\xi,\{h_j\}) = P^K(\{h_j\})P^K(\xi|\{h_j\}) \,. \tag{6.6}$$

O primeiro passo é saber qual a probabilidade de obter uma saída ξ dados os campos $\{h_j\}$, que pode ser escrita da seguinte forma:

$$P^K(\xi|\{h_j\}) = \sum_{\Sigma_B} P^K(\xi,\Sigma_B|\{h_j\}) = \sum_{\Sigma_B} P^K(\xi|\Sigma_B,\{h_j\})P^K(\Sigma_B|\{h_j\}) \,. \tag{6.7}$$

Neste ponto notamos o seguinte: a probabilidade de obter um dado ξ depende unicamente da saída correta do professor Σ_B e não dos campos locais do aluno, portanto podemos escrever $P^K(\xi|\Sigma_B, \{h_j\}) = P(\xi|\Sigma_B)$. Voltando à $P^K(\xi|\{h_j\})$, temos :

$$P^K(\xi|\{h_j\}) = \sum_{\Sigma_B} P^K(\xi|\Sigma_B) P^K(\Sigma_B|\{h_j\}) \,, \tag{6.8}$$

sendo a probabilidade $P^K(\xi|\Sigma_B)$ dada pela equação (6.1). De maneira análoga à feita nos casos anteriores (capítulos 4, 5 e apêndice A), calculamos a probabilidade $P^K(b_k|\xi, \{h_j\})$ e obtivemos a seguinte expressão:

$$P^K(b_k|\xi, \{h_j\}) = \frac{P(b_k, h_k)\left[\varepsilon P^{K-1}(-\xi\sigma_{bk}, \{h_j\}_{j\neq k}) + (1-\varepsilon)P^{K-1}(\xi\sigma_{bk}, \{h_j\}_{j\neq k})\right]}{P^K(\xi, \{h_j\})} \tag{6.9}$$

onde a probabilidade conjunta $P^K(\xi, \{h_j\})$ é:

$$\begin{aligned} P^K(\xi, \{h_j\}) &= H\left(\frac{h_k}{\lambda_k}\right) P(h_k)\left[\varepsilon\, P^{K-1}(\xi, \{h_j\}_{j\neq k}) + (1-\varepsilon)\, P^{K-1}(-\xi, \{h_j\}_{j\neq k})\right] \\ &+ H\left(-\frac{h_k}{\lambda_k}\right) P(h_k)\left[\varepsilon\, P^{K-1}(-\xi, \{h_j\}_{j\neq k}) + (1-\varepsilon)\, P^{K-1}(\xi, \{h_j\}_{j\neq k})\right] \,, \end{aligned}$$

a probabilidade condicional $P^K(\xi|\{h_j\})$ é:

$$\begin{aligned} P^K(\xi|\{h_j\}) &= H\left(\frac{h_k}{\lambda_k}\right)\left[\varepsilon\, P^{K-1}(\xi|\{h_j\}_{j\neq k}) + (1-\varepsilon)\, P^{K-1}(-\xi|\{h_j\}_{j\neq k})\right] \\ &+ H\left(-\frac{h_k}{\lambda_k}\right)\left[\varepsilon\, P^{K-1}(-\xi|\{h_j\}_{j\neq k}) + (1-\varepsilon)\, P^{K-1}(\xi|\{h_j\}_{j\neq k})\right] \,, \end{aligned} \tag{6.10}$$

e

$$P^{K-1}(\pm\xi\sigma_{bk}, \{h_j\}_{j\neq k}) = \int \prod_{j\neq k} db_j\, P(b_j, h_j)\theta\left(\pm\xi\sigma_{bk}\prod_{j\neq k} b_j\right) \,. \tag{6.11}$$

Introduzindo a notação:

$$\sum_{\{+\}} \quad e \quad \sum_{\{-\}} \tag{6.12}$$

que representa a soma sobre todas as possíveis configurações do conjunto $\{\epsilon_j = \pm 1\}$ sujeitas, respectivamente, às condições:

$$\prod_j \epsilon_j = \xi \qquad e \qquad \prod_j \epsilon_j = -\xi \,,$$

podemos reescrever a probabilidade condicional $P^K(\xi|\{h_j\})$ como:

$$P^K(\xi|\{h_j\}) = \varepsilon \sum_{\{-\}}\left[\prod_{j=1}^{K} H\left(-\epsilon_j\frac{h_j}{\lambda_j}\right)\right] + (1-\varepsilon)\sum_{\{+\}}\left[\prod_{j=1}^{K} H\left(-\epsilon_j\frac{h_j}{\lambda_j}\right)\right] \,. \tag{6.13}$$

Substituindo este resultado na equação (6.4), após a integração em b_k obtemos finalmente a função modulação para o ramo k:

$$\tilde{F}_k = \frac{K}{\sqrt{2\pi}} \frac{J_k \lambda_k}{G_\xi^\varepsilon} (1-2\varepsilon) e^{-\frac{h_k^2}{2\lambda_k^2}} \left[P^{K-1}(\xi|\{h_j\}_{j\neq k}) - P^{K-1}(-\xi|\{h_j\}_{j\neq k}) \right] \tag{6.14}$$

com $G_\xi^\varepsilon = P^K(\xi|\{h_j\})$ dado pela equação (6.13) e

$$P^{K-1}(\pm\xi|\{h_j\}_{j\neq k}) = \int \prod_{j\neq k} db_j \, P(b_j|h_j) \theta \left(\pm\xi \prod_{j\neq k} b_j \right) . \tag{6.15}$$

O termo que aparece entre colchetes na equação (6.14) pode ser reescrito como:

$$P^{K-1}(\xi|\{h_j\}_{j\neq k}) - P^{K-1}(-\xi|\{h_j\}_{j\neq k}) = \xi \prod_{j\neq k} \left[H\left(-\frac{h_j}{\lambda_j}\right) - H\left(\frac{h_j}{\lambda_j}\right) \right] , \tag{6.16}$$

e dessa forma reescrevemos a função modulação ótima:

$$\tilde{F}_k = \frac{K}{\sqrt{2\pi}} \frac{J_k \lambda_k}{G_\xi^\varepsilon} \xi (1-2\varepsilon) e^{-\frac{h_k^2}{2\lambda_k^2}} \prod_{j\neq k} \left[H\left(-\frac{h_j}{\lambda_j}\right) - H\left(\frac{h_j}{\lambda_j}\right) \right] . \tag{6.17}$$

A partir de $\tilde{F}_k$ encontramos as equações de evolução de ρ_k e J_k, que são respectivamente:

$$\frac{d\rho_k}{d\alpha} = \frac{K}{\pi} \rho_k \lambda_k^2 (1-2\varepsilon)^2 \int \frac{Dh_k}{G_+^\varepsilon} e^{-\frac{h_k^2}{\lambda_k^2}} \prod_{j\neq k} \left[H\left(-\frac{h_j}{\lambda_j}\right) - H\left(\frac{h_j}{\lambda_j}\right) \right]^2 \tag{6.18}$$

e

$$\frac{dJ_k}{d\alpha} = \frac{K}{\pi} J_k \lambda_k^2 (1-2\varepsilon)^2 \int \frac{Dh_k}{G_+^\varepsilon} e^{-\frac{h_k^2}{\lambda_k^2}} \prod_{j\neq k} \left[H\left(-\frac{h_j}{\lambda_j}\right) - H\left(\frac{h_j}{\lambda_j}\right) \right]^2 , \tag{6.19}$$

onde a probabilidade G_+^ε é obtida tomando $\xi = +1$ em (6.10). Quando $\xi = \Sigma_B$ e $\varepsilon = 0$, isto é, na ausência de ruído, recuperam-se os resultados obtidos no capítulo 5.

6.2 Limite assintótico de e_G

Estamos interessados no limite assintótico do erro de generalização, e portanto, no limite em que o número de exemplos é muito grande ($\alpha \to \infty$). Fazendo uma mudança de variável $h_k/\lambda_k = x$ e usando as aproximações para a função $H(x)$ definidas na seção 4.3.2, chegamos ao resultado

$$\frac{d\rho_k}{d\alpha} \approx \frac{K}{2\pi} \frac{(1-\rho_k^2)^{\frac{3}{2}}}{\rho_k^2} I(\varepsilon) \tag{6.20}$$

sendo $I(\varepsilon)$ definido pela expressão:

$$I(\varepsilon) = (1-2\varepsilon)^2 \int Dx \, \frac{e^{-x^2/2}}{\varepsilon + (1-2\varepsilon)H(x)} \, . \tag{6.21}$$

Na ausência de ruído, $I(\varepsilon) = I_0$, definida na equação (4.33). Considerando o caso simétrico, vemos que a equação (6.20) tem a mesma forma de (5.21), diferindo apenas por $I(\varepsilon)$. Portanto podemos escrever

$$e_G^K(\alpha) \approx \frac{C(\varepsilon)}{\alpha} \, , \tag{6.22}$$

com $C(\varepsilon) = 2/I(\varepsilon)$. Quando o ruído é grande ($\varepsilon \approx 0.5$), $I(\varepsilon)$ pode ser aproximada por $\sqrt{2}(1-2\varepsilon)^2$ e o erro assintótico é dado por:

$$e_G^K(\alpha) \approx \frac{\sqrt{2}}{(1-2\varepsilon)^2} \, \alpha^{-1} \, . \tag{6.23}$$

Com estes resultados mostramos que uma máquina paridade pode generalizar, mesmo quando os exemplos são transmitidos ao aluno com um certo nível de ruído, não existindo um nível crítico de ruído, como aparece quando se usa o LAA. É claro que se $\varepsilon = 0.5$ não existe aprendizagem. Nesta situação, não há sentido falar em generalização, já que não há mais uma regra a ser aprendida, pois neste caso a saída do professor é aleatória. A equação (6.22) também nos mostra que mesmo em situações nas quais o nível real ruído é desconhecido é possível obter o comportamento assintótico α^{-1}.

6.3 Robustez da aprendizagem na ausência de conhecimento do nível de ruído

Vamos considerar uma situação, mais realista, na qual não se tem conhecimento do nível de ruído ε na saída do professor. Neste caso o que se pode fazer é estimar o nível de ruído presente no professor, que chamamos de ruído estimado e representamos por η.

Vamos portanto levar em conta a existência de dois níveis de ruído: o ruído real ε, e o ruído estimado η, e investigar a capacidade de generalização da máquina paridade. Claramente a melhor situação ocorre quando o ruído é conhecido, isto é, $\eta = \varepsilon$, que já foi considerada na seção anterior.

A função modulação ótima só pode ser calculada a partir de grandezas conhecidas, assim deve ser uma função de η, que é o nível de ruído estimado. Entretanto, o overlap evolui de acordo com o nível real de ruído (ε), portanto para o cálculo da evolução temporal do overlap este deve ser o nível de ruído utilizado.

Levando em conta as considerações acima, a função modulação ótima para um ramo k é dada pela equação (6.17), onde no lugar do ruído real ε utiliza-se sua estimativa η:

$$\tilde{F}_k = \frac{K\xi}{\sqrt{2\pi}} \frac{J_k \lambda_k}{G_\xi^\eta} (1-2\eta)\, e^{-h_k^2/2\lambda_k^2} \prod_{j\neq k} \left[H\left(-\frac{h_j}{\lambda_j}\right) - H\left(\frac{h_j}{\lambda_j}\right) \right] \tag{6.24}$$

onde, usando a notação definida em (6.12) escrevemos:

$$G_\xi^\eta = \eta \sum_{\{-\}} \left[\prod_{i=1}^{K} H\left(-\epsilon_i \frac{h_i}{\lambda_i}\right) \right] + (1-\eta) \sum_{\{+\}} \left[\prod_{i=1}^{K} H\left(-\epsilon_i \frac{h_i}{\lambda_i}\right) \right] . \tag{6.25}$$

A evolução do overlap no k-ésimo ramo, é descrita pela equação:

$$\frac{d\rho_k}{d\alpha} = \frac{K}{2\pi} \rho_k \lambda_k^2 (1-2\eta) \sum_{\xi=\pm 1} \int \left[\prod_{j=1}^{K} dh_j P(h_j) \right] e^{-h_k^2/\lambda_k^2} \prod_{i\neq k} \left[H_i^- - H_i^+ \right]^2 \frac{N(\rho,\varepsilon,\eta,h)}{(G_\xi^\eta)^2} \tag{6.26}$$

onde usamos a notação

$$H_i^\pm = H\left(\pm \frac{h_i}{\lambda_i}\right) , \tag{6.27}$$

G_ξ^η e G_ξ^ε foram definidos em (6.25) e (6.13) respectivamente, e definimos $N(\rho,\varepsilon,\eta,h)$ como:

$$N(\rho,\varepsilon,\eta,h) = (1-2\varepsilon) G_\xi^\eta - (1-2\eta) \frac{G_\xi^\varepsilon}{2} . \tag{6.28}$$

Estamos interessados na capacidade de generalização da máquina paridade nestas condições, isso implica na investigação da evolução do overlap e do comportamento assintótico do erro de generalização.

Observando a equação (6.26) notamos que o sinal de $\dot{\rho}_k$ é determinado pelo sinal da integral, que é função do nível real de ruído (ε) e de sua estimativa (η). Isso nos mostra que a capacidade de generalizar depende diretamente da relação entre estas duas grandezas.

Para tornar completa a análise do comportamento de $\dot{\rho}$, vamos reescrever a integral na equação (6.26). Substituindo G_ξ^η e G_ξ^ε, fazendo a soma sobre ξ e rearranjando a equação de forma a fatorar ε, temos:

$$\frac{d\rho_k}{d\alpha} = \frac{K}{2\pi} \rho_k \lambda_k^2 (1-2\eta) \left\{ \mathcal{A}(\rho,\eta,K)\,(2\eta - \varepsilon - 2\varepsilon\eta) + \mathcal{B}(\rho,\eta,K)\,(1 - 3\varepsilon + 2\varepsilon\eta) \right\} \tag{6.29}$$

com $\mathcal{A}(\rho,\eta,K)$ e $\mathcal{B}(\rho,\eta,K)$ definidos pelas integrais:

$$\mathcal{A}(\rho,\eta,K) = \int \left[\prod_{j=1}^{K} dh_j\, P(h_j) \right] e^{-h_k^2/\lambda_k^2} \frac{\prod_{i\neq k} \left[H_i^- - H_i^+ \right]^2}{(G_\xi^\eta)^2} \sum_{\{-\}} \left[\prod_{i=1}^{K} H\left(-\epsilon_i \frac{h_i}{\lambda_i}\right) \right] \tag{6.30}$$

$$\mathcal{B}(\rho,\eta,K) = \int \left[\prod_{j=1}^{K} dh_j\, P(h_j)\right] e^{-h_k^2/\lambda_k^2} \frac{\prod_{i\neq k}\left[H_i^- - H_i^+\right]^2}{(G_\xi^\eta)^2} \sum_{\{+\}} \left[\prod_{i=1}^{K} H\left(-\epsilon_i \frac{h_i}{\lambda_i}\right)\right] , \quad (6.31)$$

que são calculadas numericamente para um valor de ρ fixo.

Impondo a condição $\dot{\rho} = 0$, encontramos a relação entre η e ε que define as diferentes regiões no diagrama $\eta \times \varepsilon$, relacionadas à capacidade de generalização da rede. Sob tal condição temos:

$$\varepsilon = \frac{\eta\mathcal{A}(\rho,\eta,K) + \frac{1}{2}\mathcal{B}(\rho,\eta,K)}{\eta\left[\mathcal{A}(\rho,\eta,K) - \mathcal{B}(\rho,\eta,K)\right] + \frac{1}{2}\left[\mathcal{A}(\rho,\eta,K) + 3\mathcal{B}(\rho,\eta,K)\right]} . \quad (6.32)$$

Sendo $\mathcal{A}(\rho,\eta,K)$ e $\mathcal{B}(\rho,\eta,K)$ funções do overlap, vamos estudar o que ocorre nos limites $\rho \approx 0$ e $\rho \to 1$. Nos estágios iniciais do aprendizado ($\rho \approx 0$), temos:

$$\mathcal{A}(\rho,K) \approx \mathcal{B}(\rho,K) \approx \pi\left(\frac{2}{\pi}\right)^K \prod_{l\neq k} \rho_l^2 . \quad (6.33)$$

Notamos dois fatos importantes neste resultado. Em primeiro lugar deve-se notar que a dependência em η desaparece de $\mathcal{A}$ e $\mathcal{B}$. Em segundo lugar a equivalência entre estas duas grandezas provoca o desaparecimento de K na equação (6.32), devido à linearidade em $\mathcal{A}$ e $\mathcal{B}$, de modo que nesse limite a equação (6.32) fica reduzida a:

$$\varepsilon = \frac{\eta}{2} + \frac{1}{4} . \quad (6.34)$$

Esta equação divide o diagrama de robustez $\eta \times \varepsilon$ em duas regiões, como pode ser observado através da figura 6.1.

Na região à direita da linha tem-se $\dot{\rho} < 0$ para qualquer valor $\rho < 1$ e $\rho_\infty = 0$. Isto implica que para estes pares de valores (η,ε) a rede não é capaz de inferir a regra, não possuindo capacidade de generalizar. O nível de ruído real é elevado e foi subestimado pela rede aluno, dessa forma não é possível encontrar o professor. A solução $\rho = 0$ é um ponto fixo atrativo da dinâmica.

Na região à esquerda tem-se $\dot{\rho} > 0$, indicando que a rede tem a capacidade de generalizar, isto é, o overlap aumenta à medida que são apresentados novos exemplos. Entretanto, a condição $\dot{\rho} > 0$ não é suficiente para determinar de que maneira a rede irá generalizar. Precisamos investigar o comportamento assintótico de ρ, isto é, verificar se ele chega ao seu valor máximo ($\rho_\infty = 1$) ou a um valor menor ($\rho_\infty < 1$) para $\alpha \to \infty$.

Para distinguir estas duas situações, analisamos o comportamento da equação (6.32) no limite $\rho \to 1$. Isso nos permite saber em que região $\dot{\rho}$ se anula enquanto o overlap atinge

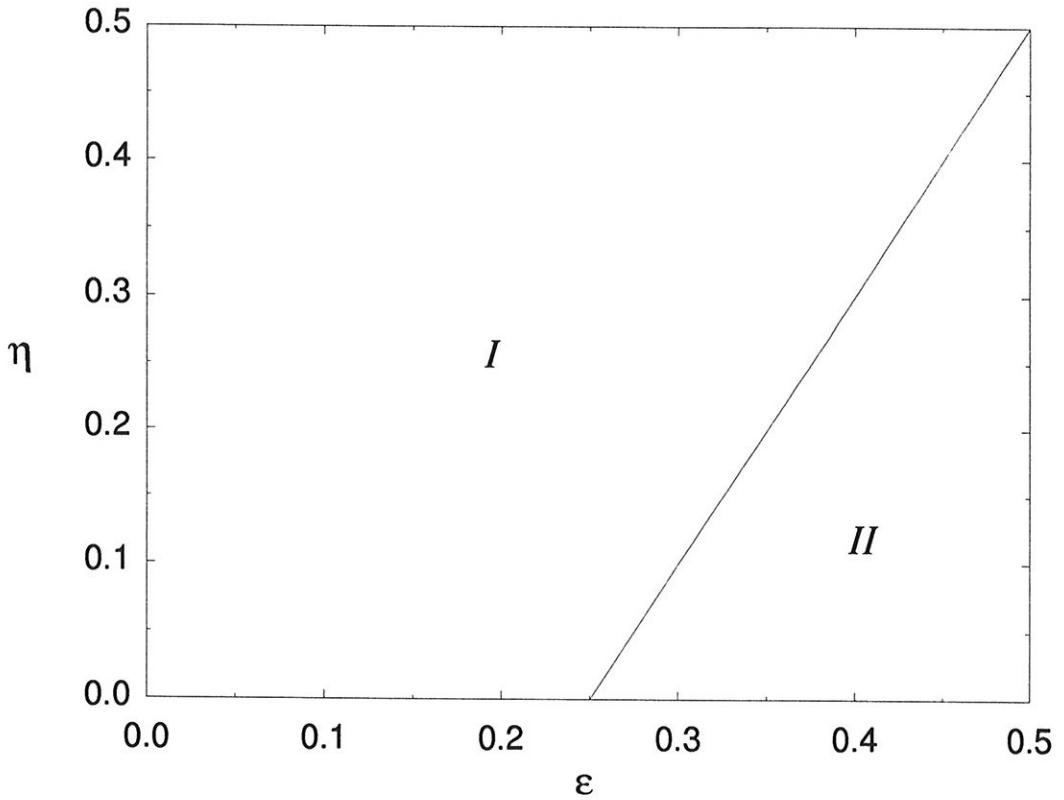


Figura 6.1: Diagrama de robustez I.

o valor máximo, e em que região $\dot{\rho}$ se anula antes que este valor seja atingido. Para isso calculamos $\mathcal{A}(\rho,\eta,K)$ e $\mathcal{B}(\rho,\eta,K)$ em $\rho \sim 1$:

$$\mathcal{A}(\rho,\eta,K) = \mathcal{A}(\rho,\eta) \approx \sqrt{1-\rho^2}\int e^{-x^2/2}\,Dx\,\frac{H(x)}{[\eta H(x)+(1-\eta)H(-x)]^2} \tag{6.35}$$

$$\mathcal{B}(\rho,\eta,K) = \mathcal{B}(\rho,\eta) \approx \sqrt{1-\rho^2}\int e^{-x^2/2}\,Dx\,\frac{H(-x)}{[\eta H(x)+(1-\eta)H(-x)]^2}\,. \tag{6.36}$$

Notamos que também neste limite desaparece a dependência em K, o que mostra que as características da capacidade de generalização descritas no diagrama de robustez são válidas para um valor arbitrário de unidades escondidas.

A solução da equação (6.32) foi obtida numericamente, dando origem à curva que se encontra no diagrama da figura 6.2, que divide a região I em duas: Ia e Ib.

As fases Ia e Ib são caracterizadas por valores diferentes do parâmetro ρ_∞, isto é, diferem quanto ao comportamento assintótico do overlap.

Na região Ia o overlap atinge assintoticamente o valor máximo ($\rho_\infty = 1$). Isto significa que para estes valores de η e ε a rede é capaz de generalizar perfeitamente. Na

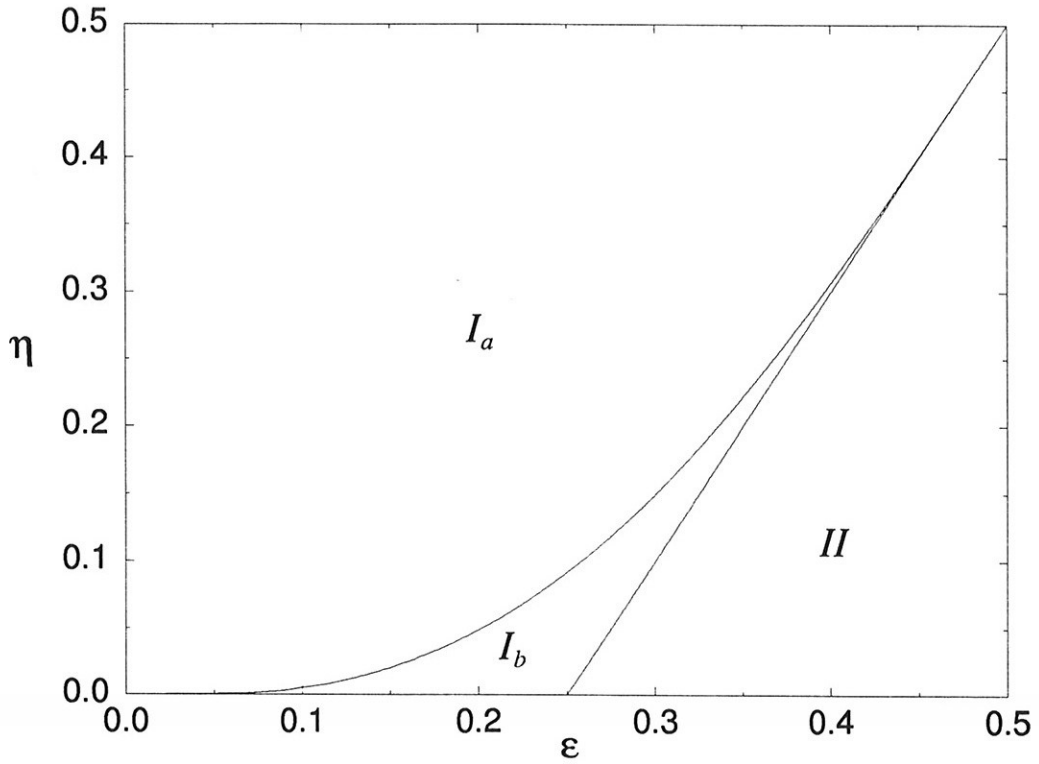


Figura 6.2: Diagrama de robustez II.

região Ib o ruído é subestimado, sendo menor do que o ruído real. Apesar deste fato, no limite de $\alpha \to \infty$ *a rede ainda é capaz de generalizar, embora o valor* $\rho = 1$ *não seja atingido* $(0 < \rho_\infty < 1)$.

Esse diagrama apresenta outra característica muito interessante: sua universalidade, sendo obtido exatamente o mesmo para a máquina comitê, perceptron, enfim para qualquer rede com estrutura de árvore, nas mesmas condições de aprendizagem, o que indica que as propriedades de generalização destas redes descritas pelo diagrama são as mesmas (Copelli et al., 1997).

Vimos que na fase II, caso em que o aluno desconhece e subestima o nível de ruído no professor, ele não é capaz de generalizar.

A perda da capacidade de generalização na máquina paridade foi apontada por Kabashima, que entretanto não leva em consideração a existência de ruído no algoritmo de aprendizado, isto é, na função modulação, aparecendo apenas na evolução do overlap.

Nessas condições, Kabashima encontra um nível de ruído crítico ≈ 0.175 *acima do qual a rede não é capaz de generalizar. Nosso estudo mostra que este resultado surge em*

conseqüência do algoritmo utilizado, não sendo uma característica da máquina paridade.

Uma possível forma de interpretar a situação descrita por Kabashima é a seguinte: para um dado estágio de aprendizagem, ou seja, para um valor fixo de ρ, o algoritmo LAA aproxima de forma precária, a função modulação ótima que utiliza um dado valor da estimativa do ruído, que chamamos de $\eta_{LAA}(\rho)$.

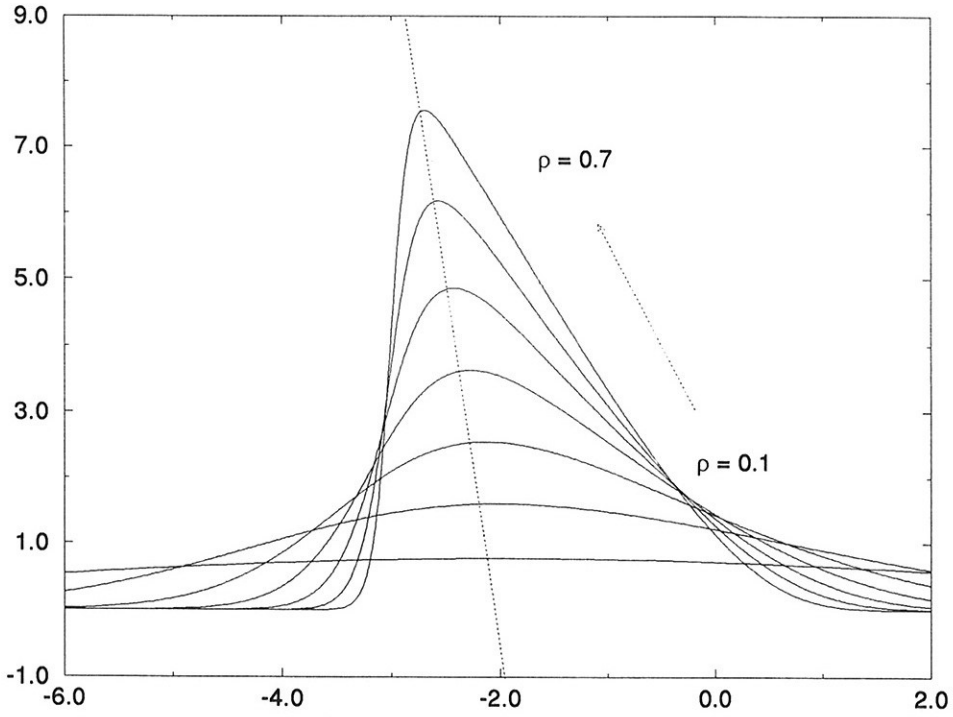


Figura 6.3: Função modulação sem ruído para o ramo 1 em função do seu campo local para alguns valores do overlap. O campo local no ramo 2 é fixo ($h_2 = -3.0$), bem como é fixa a saída do professor ($\Sigma_B = -1$). Vemos que o crossover para a região de desconfiança ocorre para valores menores de h_1, à medida que o overlap aumenta, correspondendo a valores menores do ruído (gráfico 6.4).

Pelo gráfico da figura 6.3 observamos que à medida que o overlap aumenta, o crossover ótimo para a região de desconfiança ocorre para valores menores da estabilidade. Este comportamento está relacionado ao ganho de experiência do aluno.

Pelo gráfico da figura 6.4, verificamos a ocorrência do mesmo tipo de comportamento: para valores menores de $\eta_{LAA}(\rho)$ o crossover ótimo para a região de desconfiança ocorre para valores menores de $\tilde{h}_1$. Quando o nível de ruído ε supera o valor crítico e a melhor aproximação ótima se encontra na fase II, certamente a máquina paridade que utilize o LAA será incapaz de aprender completamente.

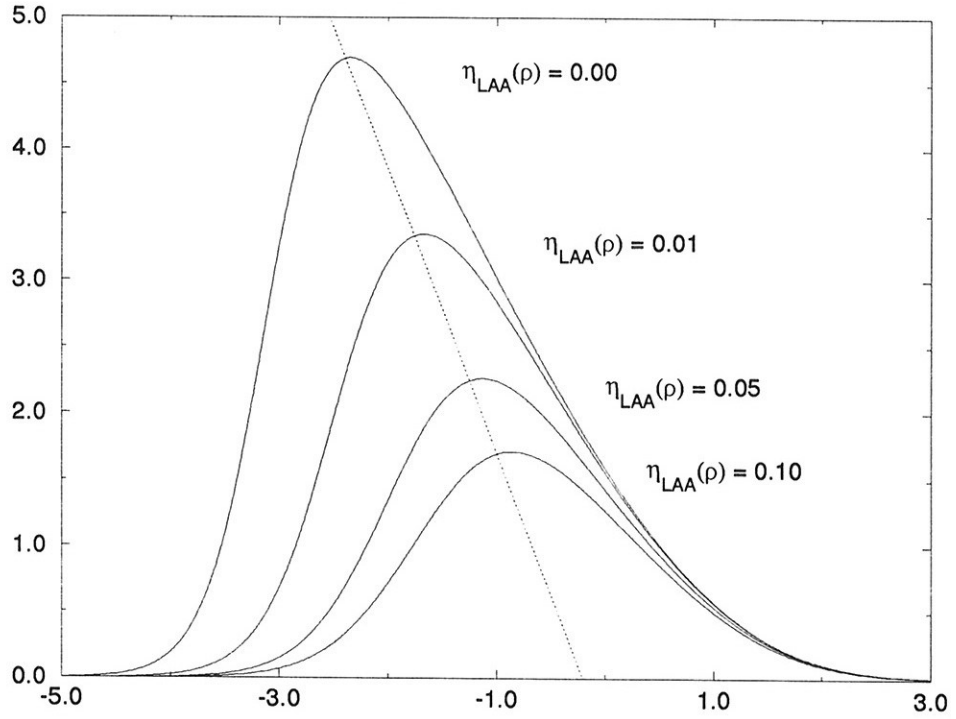


Figura 6.4: Função modulação reescalada com ruído para o ramo 1 em função do seu campo local reescalado. O campo local no ramo 2 é fixo ($\tilde{h}_2 = -3.0$), bem como é fixa a saída do professor ($\Sigma_B = -1$). Vemos que o crossover para a região de desconfiança ocorre para valores menores de $\tilde{h}_1$ à medida que diminui o nível de ruído estimado.

Uma questão que pode ser colocada é: como estimar o nível de ruído? É claro que não interessa superestimá-lo demais e ao mesmo tempo seu valor não deve ser subestimado, já que podemos cair na fase Ib ou mesmo II. Recentemente esse problema foi investigado para o caso do perceptron booleano (Riegler, 1997). A idéia proposta é uma maneira de se estimar o verdadeiro valor de ε através da adaptação on-line do parâmetro η. Para isso, define-se uma dinâmica para η tal que, assintoticamente, tenha-se $\eta \to \varepsilon$. Uma maneira de se fazer isso é diminuir o valor de η sempre que a resposta do aluno estiver correta e aumentar no caso contrário.

7 Perceptron com função de transferência não monotônica

Neste capítulo estudamos um percetron cuja função de transferência não é monotônica no campo, conhecido na literatura como perceptron reversed-wedge. Estudamos duas situações, na primeira o aprendizado ocorre na presença de ruído multiplicativo e na segunda, ruído aditivo. Analisamos diretamente o caso geral no qual é desconhecido o nível real de ruído. As outras situações, isto é, nível de ruído conhecido e ausência de ruído, podem ser facilmente derivadas a partir destes casos gerais.

7.1 Introdução

No capítulo 2 definimos o perceptron e algumas de suas características. Vimos que uma rede neural como esta é definida a partir de um vetor sináptico e do mapeamento implementado pela função de transferência, $\mathcal{F}$. Em alguns casos, $\mathcal{F}$ é uma função monotônica do campo local (linear ou não) e pode ser invertida, o que significa que dada a saída é possível conhecer o valor do campo local correspondente. Como exemplo citamos o *perceptron linear*, onde $\mathcal{F}_{PL} = h$ e o *perceptron analógico*, onde $\mathcal{F}_{PA} = tgh(h)$. Em outros casos, $\mathcal{F}$ não pode ser invertida, o que implica que conhecer a saída não é suficiente para saber qual o valor do campo local correspondente. É o caso dos *perceptrons booleanos*, cuja função de transferência é $\mathcal{F}_{PB} = sinal[g(h)]$. A principal característica do perceptron booleano é a sua saída binária, isto é, que pode assumir apenas dois valores: $+1$ ou -1, devido à presença da função sinal, independente da forma de $g(h)$. A função $g(h)$ pode ser monotônica no campo h, por exemplo se $g(h) = h$, ou não monotônica. Vamos estudar neste capítulo um perceptron booleano em que $g(h)$ é uma função não monotônica do campo local dada por:

$$g(h) = (h + K)\, h\, (h - K)\ . \tag{7.1}$$

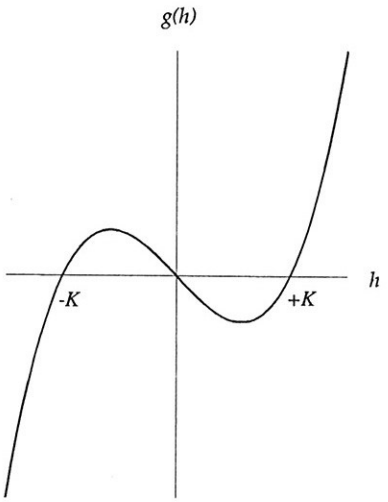


Figura 7.1: Função $g(h)$ no PRW: a escolha $K \neq 0$ significa introduzir uma "cunha" de largura $2K$ em torno da origem $h = 0$.

O papel de K pode ser interpretado em termos geométricos: a presença de $K \neq 0$ equivale à inserção de dois hiperplanos paralelos, um de cada lado do hiperplano ortogonal à $\mathbf{J}$ passando pela origem. Estes planos dão origem a duas novas regiões de classificação, que não aparecem no caso do perceptron simples, como podemos ver na figura 7.2.

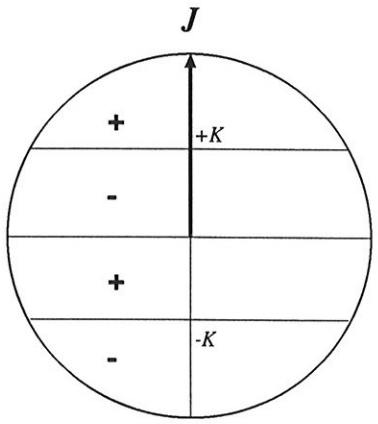


Figura 7.2: Interpretação geométrica do PRW: a presença de $K \neq 0$ dá origem a duas novas regiões de classificação.

O grande interesse em estudar esse tipo de rede vem do fato que um perceptron com tal estrutura é capaz de realizar classificações não linearmente separáveis. A saída do professor e do aluno, supondo que ambos têm a mesma arquitetura, são dadas respectivamente por:

$$\sigma_B = sinal\left[g\left(b\right)\right] \qquad e \qquad \sigma_J = sinal\left[g\left(h\right)\right] \tag{7.2}$$

onde h e b são os campos locais definidos no capítulo 2. Considerando a forma de $g(h)$ definida em (7.1), a saída das redes professor e aluno serão:

$$\begin{array}{llll} \sigma_{B,J} = +1 & se & -K < b, h < 0 & ou \quad b, h > +K \\ \sigma_{B,J} = -1 & se & 0 < b, h < +K & ou \quad b, h < -K \end{array}$$

o que também pode ser observado através da figura 7.3. É fácil ver que para $K = 0$ recupera-se a função de transferência do perceptron simples.

É interessante notar que esse tipo de perceptron implementa um mapeamento igual ao de uma máquina paridade totalmente conectada com três unidades escondidas, sendo que em duas destas unidades existe um threshold de $+K$ e $-K$, respectivamente. Entretanto, o número de representações internas de um perceptron reversed-wedge é 4, e não 8, como no caso da máquina paridade $K = 3$, dessa forma as habilidades deste perceptron devem ser comparadas às de uma máquina paridade $K = 2$, pois ambos dividem o espaço em 4 regiões.

No caso binário ($J_i = \pm 1$) esta rede também apresenta outras características interessantes. Estudando o aprendizado supervisionado off-line usando o algoritmo de contagem de erros, Bex et al. (1995) observaram que para um valor específico[1] de K este perceptron

[1] que será discutido na seção 7.4.

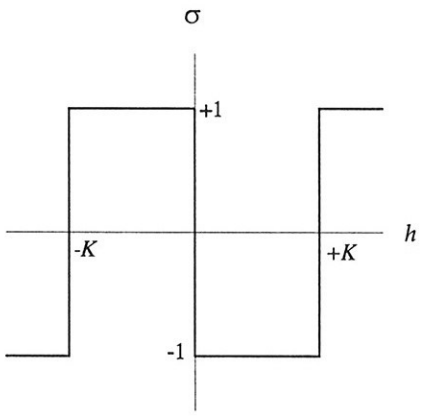


Figura 7.3: Função de transferência do PRW

satura o limite teórico para capacidade de armazenamento $\alpha = 1$ e também de generalização, já que observam uma transição em $\alpha = 1$ para generalização perfeita. No caso da capacidade de armazenamento, $\alpha = 1$ constitui um limite superior, uma vez que cada componente do vetor sináptico pode armazenar no máximo 1 bit de informação, e no caso da generalização constitui um limite inferior, já que cada exemplo pode carregar no máximo 1 bit de informação sobre o professor. Esses valores devem ser comparados com os resultados obtidos a partir do estudo do perceptron binário, também para o algoritmo de contagem de erros, no limite em que p e N tendem a infinito, sendo mantida finita a razão p/N. Para a capacidade de armazenamento foi encontrado o valor $\alpha_c \approx 0.83$ (Krauth & Mézard, 1989), mais tarde confirmado por simulações (Krauth & Opper, 1989) e no caso da generalização, observa-se a existência de uma transição de fase descontínua para perfeita generalização ($e_G = 0$) em $\alpha \approx 1.245$ (Györky, 1990).

Nosso interesse é estudar as propriedades de generalização do perceptron reversed-wedge com componentes reais para o aprendizado supervisionado on-line. Este capítulo está organizado da seguinte forma: na seção 7.2 discutimos o comportamento do erro de generalização, que está associado à aplicabilidade do método variacional para obtenção do algoritmo ótimo. Na seção 7.3 estudamos o aprendizado na presença de ruído multiplicativo, na seção 7.4, o caso do perceptron especial K^* e na 7.5 o caso da aprendizagem com ruído aditivo. Em ambas consideramos desconhecido o nível de ruído e obtemos a função modulação ótima, estudamos a evolução do overlap com a apresentação de novos exemplos e construímos o diagrama de robustez, identificando os diferentes regimes de generalização existentes.

7.2 Erro de generalização

Para que a função modulação obtida através do método variacional seja ótima, é suficiente que o erro de generalização seja uma função monotônica do overlap. Portanto o primeiro passo é verificar como se comporta o erro de generalização em função de ρ. Como vimos no capítulo 2, o erro de generalização pode ser obtido a partir da integral:

$$e_G = \int_{-\infty}^{+\infty} dh \int_{-\infty}^{+\infty} db\, P(h,b) \frac{1-\sigma_B\sigma_J}{2}\,, \tag{7.3}$$

cujo integrando é diferente de zero quando professor e aluno discordarem da resposta, isto é, se $\sigma_B \neq \sigma_J$. Dessa forma a integral pode ser reescrita como:

$$e_G = \int\int_{\sigma_B \neq \sigma_J} P(h,b)\, db\, dh\,. \tag{7.4}$$

A figura 7.4 mostra as 8 regiões onde é feita a integração da equação (7.4).

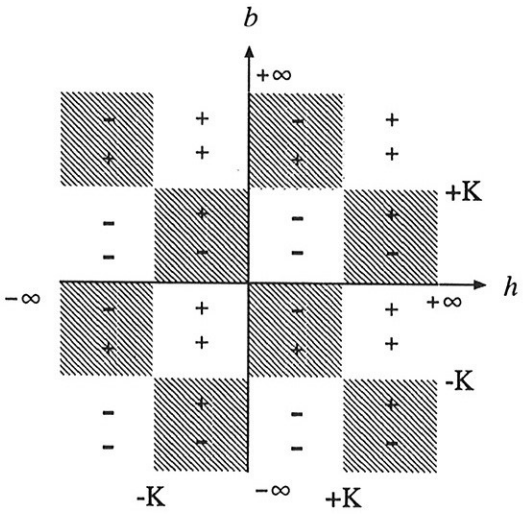


Figura 7.4: A integral é calculada nas regiões hachuradas, onde $\sigma_B \neq \sigma_J$. O sinal superior indica o valor da saída do professor e o inferior, do aluno.

Trocando o sinal de h, de b e/ou de ambos nas diferentes regiões, algumas destas integrais se tornam idênticas. Aproveitando esta simetria podemos escrever:

$$e_G = 4\int_K^{\infty} db \int_0^K dh\, P(b,h) + 2\int_0^K db \int_0^K dh\, P(b,-h) + 2\int_K^{\infty} db \int_K^{\infty} dh\, P(b,-h)\,. \tag{7.5}$$

Para verificar o comportamento de e_G precisamos fazer estas integrais numericamente. Este cálculo foi feito para diferentes valores de K, tendo sido obtido o gráfico da figura 7.5.

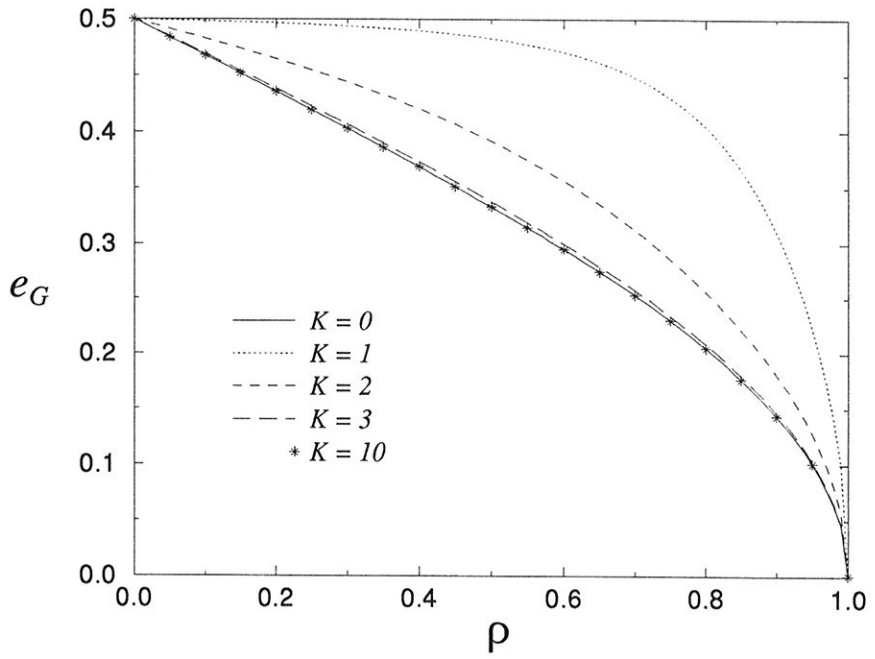


Figura 7.5: Erro de generalização do perceptron com função de transferência não monotônica nos campos locais. Note-se que no limite $K \to \infty$ recupera-se o caso $K = 0$.

Vemos, portanto, que o erro de generalização é uma função monotônica do overlap, o que nos permite implementar o algoritmo ótimo, como fizemos no caso da máquina paridade.

7.3 Aprendizagem na presença de ruído multiplicativo

Nesta seção, as condições de aprendizagem são muito semelhantes àquelas do capítulo 6. Vamos apenas recordar a notação utilizada. Seja σ_B a saída correta do professor, que é invertida com uma certa probabilidade. A saída fornecida ao aluno será ξ, tal que:

$$\begin{aligned} \xi = +\sigma_B &\implies \text{com probabilidade } 1-\varepsilon \\ \xi = -\sigma_B &\implies \text{com probabilidade } \varepsilon \end{aligned}$$

e a probabilidade de obter a saída ξ, dada a saída correta do professor σ_B é:

$$P(\xi|\sigma_B) = \varepsilon\, \delta(\xi + \sigma_B) + (1-\varepsilon)\, \delta(\xi - \sigma_B)\,. \tag{7.6}$$

7.3.1 Função modulação

No caso aqui estudado ε não está disponível à rede aluno, dessa forma ele utilizará uma estimativa fixa deste valor, denotada por η. A função modulação ótima deve ser calcu-

lada a partir de grandezas conhecidas, e por isso deve ser uma função de η. Para calculá-la, partimos da equação:

$$\tilde{F} = J\left(\frac{< b >_{b|\xi,h}}{\rho} - h\right) , \tag{7.7}$$

onde

$$< b >_{b|\xi,h} = \int db\, b\, P(b|\xi, h) . \tag{7.8}$$

Detalhes do cálculo da probabilidade $P(b|\xi, h)$ podem ser encontrados no apêndice A. Este cálculo nos leva à expressão:

$$P(b|\xi, h) = \frac{P(b|h)}{D(\xi, \eta, \rho, h)} \left[\eta\, \delta_{-\xi,\sigma} + (1-\eta)\, \delta_{+\xi,\sigma}\right] , \tag{7.9}$$

onde usamos a notação $P(\xi|h) \equiv D(\xi, \eta, \rho, h)$ definido como:

$$\begin{aligned} D(\xi,\eta,\rho,h) &= \eta\left\{H\left(-\xi\frac{\rho h}{\sqrt{1-\rho^2}}\right) - \xi H\left(\frac{K-\rho h}{\sqrt{1-\rho^2}}\right) + \xi H\left(\frac{K+\rho h}{\sqrt{1-\rho^2}}\right)\right\} \\ &+(1 - \eta)\left\{H\left(\xi\frac{\rho h}{\sqrt{1-\rho^2}}\right) + \xi H\left(\frac{K-\rho h}{\sqrt{1-\rho^2}}\right) - \xi H\left(\frac{K+\rho h}{\sqrt{1-\rho^2}}\right)\right\} , \end{aligned} \tag{7.10}$$

a probabilidade condicional $P(b|h)$

$$P(b|h) = \frac{exp\left\{-\frac{(b-\rho h)^2}{2(1-\rho^2)}\right\}}{\sqrt{2\pi(1-\rho^2)}} , \tag{7.11}$$

e

$$\sigma = sinal\left[b\left(b^2 - K^2\right)\right] . \tag{7.12}$$

Substituindo estas expressões na equação (7.7), e fazendo a mudança de variável $b \to (b - \rho h)/\sqrt{1-\rho^2}$, podemos escrever a função modulação ótima como[2]:

$$\tilde{F} = \xi\frac{J\lambda}{D(\xi,\eta,\rho,h)}(2\eta - 1)\left\{\int_{-\frac{\rho h}{\sqrt{1-\rho^2}}}^{\infty} b\, Db - \int_{\frac{K+\rho h}{\sqrt{1-\rho^2}}}^{\infty} b\, Db - \int_{\frac{K-\rho h}{\sqrt{1-\rho^2}}}^{\infty} b\, Db\right\} . \tag{7.13}$$

Realizando as integrais em b obtemos o resultado:

$$\tilde{F} = \xi\frac{J\lambda}{\sqrt{2\pi}}\frac{(1-2\eta)}{D(\xi,\eta,\rho,h)}e^{-h^2/2\lambda^2}\left[2e^{-K^2/2\lambda^2\rho^2}\cosh(Kh/\lambda^2\rho) - 1\right] . \tag{7.14}$$

Esta expressão nos mostra que a função modulação ótima depende explicitamente do valor do campo local h, do overlap ρ e, certamente, da largura da cunha K. A presença

[2]esta forma de escrever $\tilde{F}$ será útil quando discutirmos o K especial (seção 7.4).

de h indica a dependência nos exemplos apresentados. Como no caso da máquina paridade, a dependência em ρ indica que o peso dado a um mesmo exemplo varia de acordo com o estágio de aprendizagem no qual a rede se encontra. Para $\rho = 0$ (linha sólida) todos os exemplos são igualmente importantes e recebem o mesmo peso: $\tilde{F}$ é uma constante. As figuras mostram a função modulação ótima para alguns valores de K, entre eles $K = 0$, que equivale a um perceptron simples. Para efeitos de comparação apresentamos o caso em que $\eta = 0$ (figura 7.6) e $\eta = 0.2$ (figura 7.7).

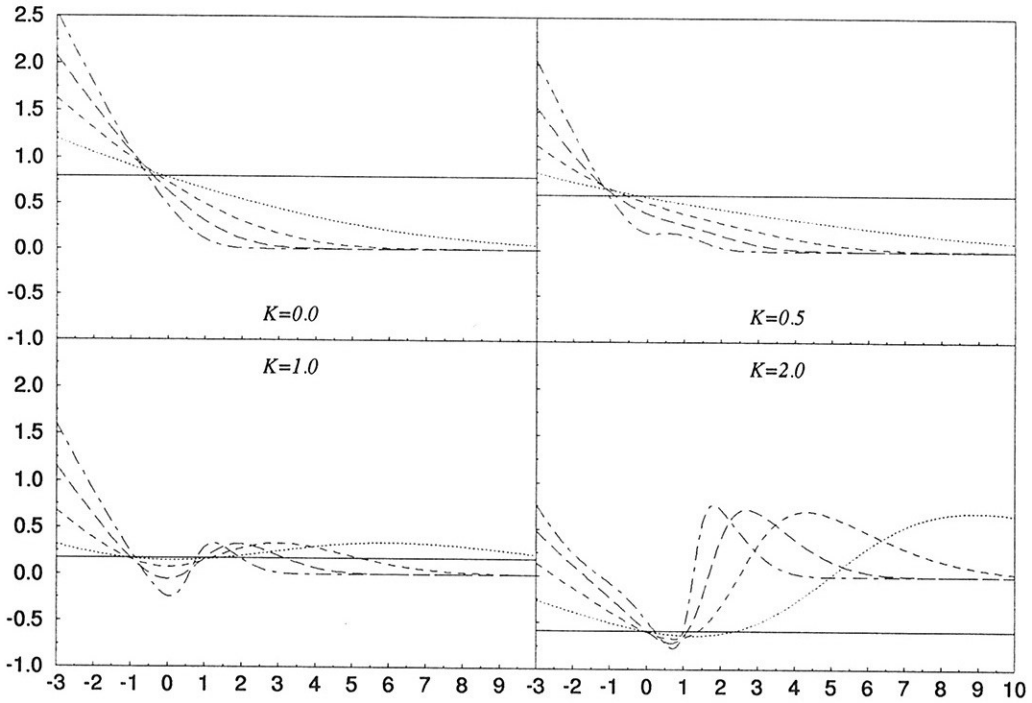


Figura 7.6: Função modulação ótima no PRW em função do valor da estabilidade $h\xi$, para $\eta = 0$ e diferentes valores de K e do overlap: $\rho = 0$ (——), $\rho = 0.2$ (·····), $\rho = 0.4$ (- - - -), $\rho = 0.6$ (– – –) e $\rho = 0.8$ (– – – -). É interessante notar a dependência com o valor do overlap: nos estágios inicias de aprendizagem a função é uma constante. Para $K \to \infty$ recupera-se a forma de $\tilde{F}$ refletida em torno de $h = 0$ e de $\tilde{F} = 0$: $\tilde{F}(K = 0, -h) = -\tilde{F}(K \to \infty, h)$.

É interessante notar que a presença de ruído multiplicativo altera o módulo da função modulação, mas principalmente a sua forma. Na ausência de ruído as mudanças mais significativas ocorrem para valores negativos da estabilidade, e tornam-se cada vez maiores à medida que a estabilidade diminui. Esse comportamento aponta uma "falta de surpresa" por parte do aluno, aliada a uma "total confiança" na resposta dada pelo professor. Em oposição, quando a resposta do professor está contaminada pela presença de ruído multiplicativo,

surgem regiões onde o aluno passa a não ter tanta confiança na resposta do professor e a função modulação diminui, até se anular.

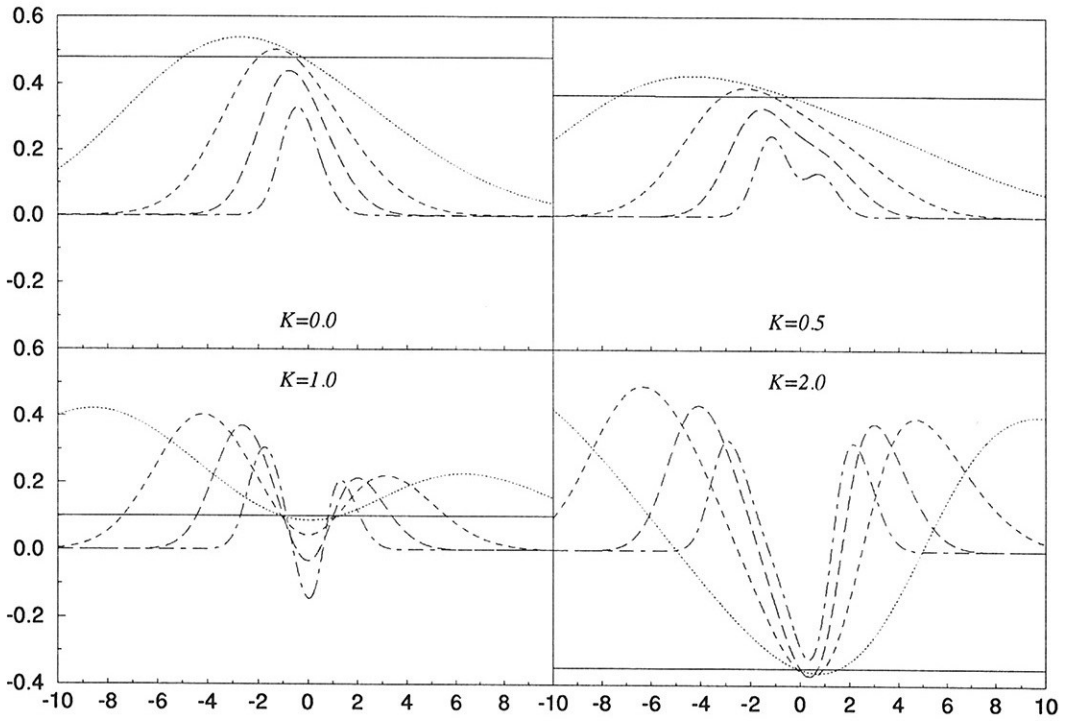


Figura 7.7: Função modulação ótima no PRW em função do valor da estabilidade $h\xi$, na presença de ruído multiplicativo. Consideramos a situação para $\eta = 0.2$ e diferentes valores de K e do overlap: $\rho = 0$ (——), $\rho = 0.2$ (·····), $\rho = 0.4$ (- - - -), $\rho = 0.6$ (– – –) e $\rho = 0.8$ (– - – -). É interessante notar a dependência com o valor do overlap: nos estágios inicias de aprendizagem a função é uma constante.

7.3.2 A equação de evolução do overlap

Vamos agora obter a equação de evolução do overlap. Vimos que para calcular a função modulação, a ser implementada pelo algoritmo, só podem ser usadas grandezas conhecidas[3]*, e portanto deve-se utilizar a estimativa do ruído. Entretanto, deve-se notar que o overlap evolui de acordo com o nível real de ruído que está presente na probabilidade*

[3] o overlap não é conhecido, entretanto existe uma maneira de solucionar este problema através da utilização do módulo do vetor J em seu lugar, como discutido em Kinouchi & Caticha, 1993.

conjunta $P(b,h,\xi)$. Para encontrar sua expressão, partimos do resultado:

$$\frac{d\rho}{d\alpha} = \frac{\rho}{J}\sum_{\xi=\pm 1}\int P(b,h,\xi)\, db\, dh\, \tilde{F}\left(\frac{b}{\rho} - h - \frac{\tilde{F}}{2J}\right) , \tag{7.15}$$

onde $P(b,h,\xi)$ é dado por:

$$P(b,h,\xi) = P(h)\, P(b|h)\, \left[\eta\, \delta_{-\xi,\sigma} + (1-\eta)\, \delta_{+\xi,\sigma}\right] . \tag{7.16}$$

Com o resultado de $\tilde{F}$ substituído na equação acima, realizando a soma sobre ξ e a integral em b, obtemos finalmente a equação de evolução do overlap:

$$\begin{aligned}\frac{d\rho}{d\alpha} &= \frac{\rho\lambda^2}{\pi}(1-2\eta)\int dh\, P(h)\frac{e^{-\frac{h^2}{\lambda^2}}}{D_\eta^2}\left[2e^{-\frac{K^2}{2\lambda^2\rho^2}}\cosh\left(\frac{Kh}{\lambda^2\rho}\right) - 1\right]^2 \\ &\times \left\{D_\eta(1-2\varepsilon) - \frac{D_\varepsilon}{2}(1-2\eta)\right\} .\end{aligned} \tag{7.17}$$

A notação abreviada D_η representa a função definida em (7.10) quando se toma $\xi = +1$ e D_ε é dada pela mesma expressão, substituindo η por ε. A evolução do overlap, e portanto a capacidade de generalização da rede, é determinada pela equação (7.17). Para que o overlap possa crescer é preciso que $\dot{\rho}$ seja positivo. Observando esta equação, fica claro que o sinal de $\dot{\rho}$ depende da relação entre η e ε definida pelo termo entre chaves.

7.3.3 Diagrama de robustez

O procedimento utilizado nesta seção é semelhante ao adotado na seção 6.3 para obter o diagrama de robustez da máquina paridade. O interesse em analisar a evolução de ρ é determinar as diferentes regiões de aprendizado em função da estimativa do ruído e de seu valor real. Tal análise é feita considerando o valor de $\dot{\rho}$ para diferentes pares de valores (η,ε). A partir desses valores construimos o diagrama de robustez, isto é, o diagrama $\eta \times \varepsilon$. Para analisar o comportamento de $\dot{\rho}$ vamos reescrever a integral em (7.17). Substituindo D_η e D_ε pelas suas respectivas definições e rearranjando os termos obtemos:

$$\frac{d\rho}{d\alpha} = \frac{\rho\lambda^2}{2\pi}(1-2\eta)\left\{(2\eta - 2\varepsilon\eta - \varepsilon)\ \mathcal{A}(\eta,K,\rho) + (1 + 2\varepsilon\eta - 3\varepsilon)\ \mathcal{B}(\eta,K,\rho)\right\} . \tag{7.18}$$

As funções $\mathcal{A}(\eta,K,\rho)$ e $\mathcal{B}(\eta,K,\rho)$ são definidas pelas integrais:

$$\mathcal{A}(\eta,K,\rho) = \int dh\, \frac{P(h)}{D_\eta^2}e^{-\frac{h^2}{\lambda^2}}\left[2e^{-\frac{K^2}{2\lambda^2\rho^2}}\cosh\left(\frac{Kh}{\lambda^2\rho}\right) - 1\right]^2\left[H\left(-\frac{h}{\lambda}\right) + H^+ - H^-\right] \tag{7.19}$$

$$\mathcal{B}(\eta, K, \rho) = \int dh \, \frac{P(h)}{D_\eta^2} e^{-\frac{h^2}{\lambda^2}} \left[2e^{-\frac{K^2}{2\lambda^2\rho^2}} \cosh\left(\frac{Kh}{\lambda^2\rho}\right) - 1 \right]^2 \left[H\left(\frac{h}{\lambda}\right) - H^+ + H^- \right] \tag{7.20}$$

onde usamos a notação

$$H^\pm = H\left(\frac{K}{\sqrt{1-\rho^2}} \pm \frac{h}{\lambda}\right) .$$

O termo entre chaves na equação (7.18) define as regiões onde $\dot{\rho}$ é positivo, negativo ou nulo, e depende, além do ruído, de K e do overlap. Para investigar os diferentes regimes de aprendizado precisamos analisar o comportamento assintótico de ρ, isso nos permite saber se a rede é capaz de generalizar ou não, e de que maneira ela o faz. Vamos considerar inicialmente os estágios iniciais da aprendizagem ($\rho \to 0$) e em seguida analisaremos o limite $\rho \to 1$.

Comportamento no limite $\rho \to 0$

Este limite corresponde a $\lambda \to \infty$. Neste caso podemos usar a aproximação da função $H(x)$ para argumentos pequenos [4]. *Usando este resultado verificamos que D_η e D_ε tornam-se independentes do valor de K, η, ε e h: ambos se reduzem ao valor 1/2. Com isso fica fácil calcular as integrais $\mathcal{A}$ e $\mathcal{B}$, qua agora são funções exclusivamente de K:*

$$\mathcal{A}(K) = \mathcal{B}(K) \approx 2\left[2e^{-\frac{K^2}{2}} - 1\right]^2 . \tag{7.21}$$

Usando estes resultados, podemos reescrever a equação (7.18) como:

$$\dot{\rho} \approx \frac{\rho\lambda^2}{\pi}(1-2\eta)\left[2e^{-K^2/2} - 1\right]^2 \{1 - 4\varepsilon + 2\eta\} . \tag{7.22}$$

O sinal de $\dot{\rho}$ é determinado pelo termo entre chaves, isto é, $\dot{\rho}$ muda de sinal quando:

$$\eta = 2\varepsilon - \frac{1}{2} . \tag{7.23}$$

Esta equação dá origem à primeira linha do diagrama de robustez, que o divide em duas regiões I e II. Na região II, isto é, à direita desta linha, $\dot{\rho}$ é negativo para $\rho < 1$ e se anula quando $\rho = 1$. Isto significa que ela delimita a região na qual a rede não é absolutamente capaz de generalizar. Para valores de (η, ε) que se encontram na região à esquerda desta linha a rede é capaz de generalizar. Entretanto, ser capaz de generalizar significa que o overlap pode crescer, mas não garante que seu valor máximo seja atingido. Para conhecer as características da capacidade de generalização nesta fase precisamos investigar o valor assintótico de ρ para $\alpha \to \infty$.

[4] $H(x) \approx 1/2$ para $x \to 0$

Comportamento no limite $\rho \to 1$

Não é possível tomar este limite diretamente na equação (7.17) por causa das divergências que surgem devido à presença de λ no denominador de alguns termos. Assim, este caso é estudado a partir da análise numérica de $\dot{\rho}$. Nosso interesse é saber para que valores (η, ε) $\dot{\rho}$ se anula. Nesta condição, podemos escrever:

$$\varepsilon = \frac{\eta\mathcal{A} + \mathcal{B}/2}{\eta(\mathcal{A} - \mathcal{B}) + 1/2(\mathcal{A} + 3\mathcal{B})} \,. \tag{7.24}$$

A solução numérica desta equação no limite $\rho \to 1$ junto com o resultado obtido para o limite $\rho \to 0$ fornece o diagrama da figura 7.8.

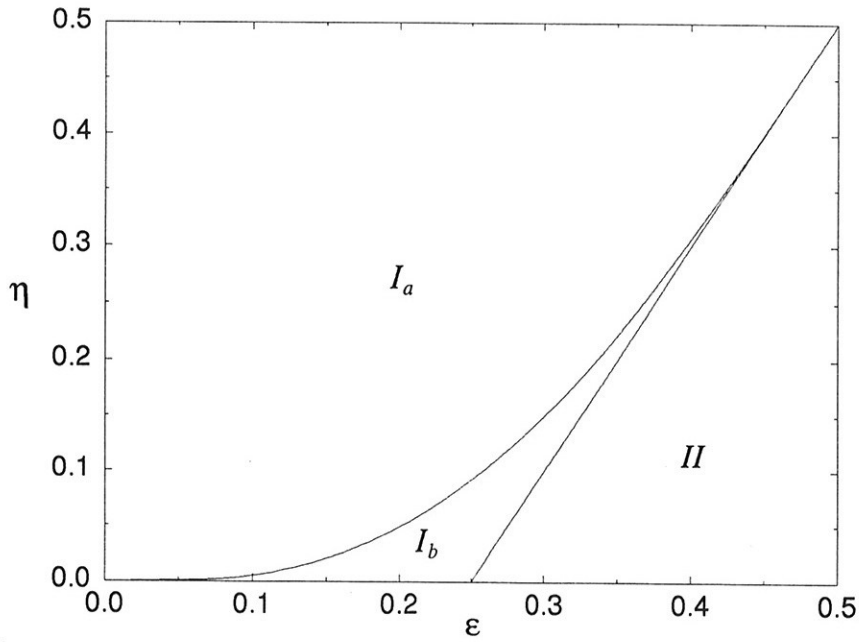


Figura 7.8: Diagrama de robustez PRW para o caso de ruído multiplicativo.

Como dissemos, na região à direita do diagrama (II), a rede é incapaz de generalizar, pois neste caso $\dot{\rho}(0) \leq 0$. Isso significa que $\rho = 0$ é um ponto fixo atrativo da dinâmica. A linha à esquerda divide a região de generalização em duas: na região Ia a rede se encontra num regime de perfeita generalização, sendo $\dot{\rho}$ sempre positivo, até se anular, quando simultaneamente ρ atinge seu valor máximo. Em Ib a capacidade de generalização é parcial, sendo que no limite $\alpha \to \infty$, o overlap atinge um valor inferior a 1, quando $\dot{\rho}$ se anula e passa a ser negativo. Para ilustrar melhor as características dos diferentes regimes de generalização

presentes no diagrama de robustez, mostramos na figura 7.9 como $\dot{\rho}$ varia em função de ρ para diferentes valores de (η, ε) e $K = 1$.

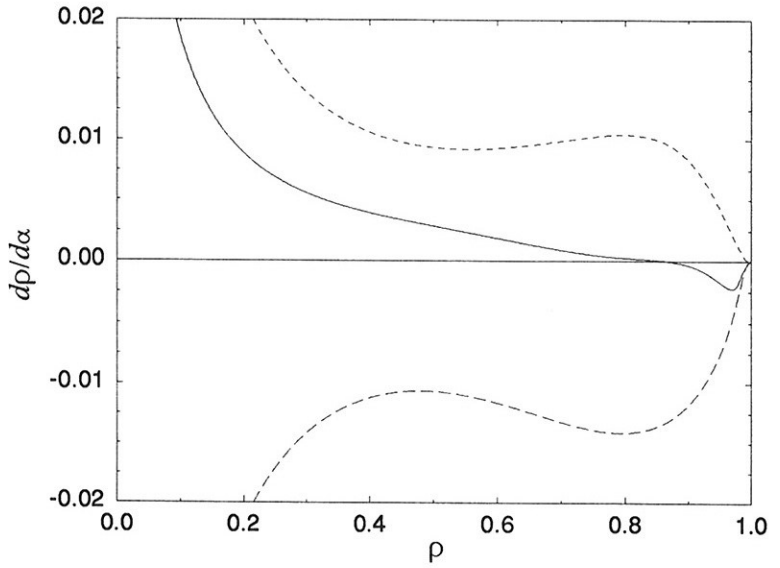


Figura 7.9: O gráfico mostra o comportamento de $\dot{\rho}$ em função de ρ para $K = 1$ e diferentes pares (η, ε) correspondendo às diferentes fases. Fase Ia (0.1,0.2)(- - - -); fase Ib (0.03,0.23)(——); fase II (0.1,0.4)(– – –).

Representando a fase *Ia* escolhemos os valores $\eta = 0.1$ e $\varepsilon = 0.2$ (traços), neste caso $\dot{\rho}$ é sempre positivo e o overlap atinge assintoticamente seu valor máximo. Para a fase *Ib* escolhemos os valores $\eta = 0.03$ e $\varepsilon = 0.23$ (linha cheia), neste caso $\dot{\rho}$ se anula para $\rho \approx 0.87$, inferior ao valor máximo. A terceira curva (traços longos) mostra o comportamento típico de $\dot{\rho}$ na fase *II*: $\dot{\rho}$ é sempre negativo e portanto não ocorre generalização. Neste caso os valores escolhidos foram $\eta = 0.1$ e $\varepsilon = 0.4$ A primeira característica importante deste resultado é sua validade para qualquer valor de K. Ele descreve o comportamento assintótico de ρ, independentemente das condições iniciais, exceto para $K = K^* = \sqrt{2 \ln 2}$. A segunda característica importante a ser notada é que este diagrama é idêntico ao diagrama obtido para a máquina paridade (capítulo 6), isso significa que estas duas redes apresentam as mesmas características e regimes de aprendizagem. A universalidade destes diagramas foi discutida no caso da máquina comitê e paridade (Copelli et al., 1997).

7.4 Características da rede especial $K = K^*$

A rede K^* já foi apontada como sendo um caso especial (seção 7.1) pois satura o limite teórico para capacidade de armazenamento ($\alpha_c = 1$) e apresenta uma transição de fase de primeira ordem descontínua para perfeita generalização também em $\alpha_c = 1$, para o aprendizado off-line. Partindo da função modulação ótima obtida na seção 7.3.1 (equação 7.14) nos estágios iniciais de aprendizagem, e tomando $\rho = 0$ a função modulação ótima se reduz a

$$\tilde{F} = \xi \frac{2(1-2\eta)}{\sqrt{2\pi}} \left[2e^{-\frac{K^2}{2}} - 1 \right] , \tag{7.25}$$

e dessa forma perde a dependência no campo h. Além disso, podemos ver que $\tilde{F}$ se anula quando $K = K^* = \sqrt{2 \ln 2}$. Isso justifica a afirmação que o diagrama de robustez é válido para qualquer valor de K, e independe das condições iniciais, exceto para K^*. Este perceptron especial, diferentemente de todos os outros, necessita de um conhecimento a priori para poder generalizar. Para entender qual a origem deste comportamento, vamos partir da equação (7.13) e reescrever a função modulação, para o caso $\rho = 0$, da seguinte forma:

$$\tilde{F} = 2\xi(2\eta - 1) \left\{ \int_0^\infty b \, Db - 2 \int_K^\infty b \, Db \right\} . \tag{7.26}$$

Notamos que o termo entre chaves pode ser escrito como:

$$\int_0^K b \, Db - \int_K^\infty b \, Db , \tag{7.27}$$

e se anula quando estas duas integrais forem iguais. Mas qual seu significado? A primeira integral calcula o valor esperado de b entre 0 e K (que é numericamente igual ao valor esperado calculado entre $-K$ e 0), e a segunda, entre K e ∞. Quando essas duas grandezas são iguais não existe informação disponível sobre o professor, já que rede aluno não consegue saber se a saída do professor é, por exemplo, positiva porque $-K < b < 0$ ou porque $b > +K$, e vice-versa. Neste caso $\rho = 0$ é um ponto fixo repulsivo da dinâmica. Entretanto, basta um pequeno valor inicial de ρ não nulo para que o sistema evolua. No gráfico da figura 7.10 mostramos o comportamento da função modulação, que tem basicamente a mesma forma da função modulação para os outros valores de K. Para $\rho = 0$, $\tilde{F}$ é uma constante, como nos outros casos, só que neste caso ela é nula.

Para verificar a validade do diagrama de robustez no caso de $K = K^*$, mostramos no gráfico da figura 7.11 o valor de $\dot{\rho}$ em função de ρ, para pares de valores (η, ε) que pertencem às regiões Ia (traços), Ib (linha cheia) e II (traços longos). Na região Ia, $\dot{\rho}$ é sempre positivo e a rede é capaz de generalizar perfeitamente ($\rho_\infty = 1$), embora necessite de um conhecimento prévio. Em Ib o valor assintótico do overlap pertence ao intervalo $]0, 1[$ e em II, $\rho_\infty = 0$.

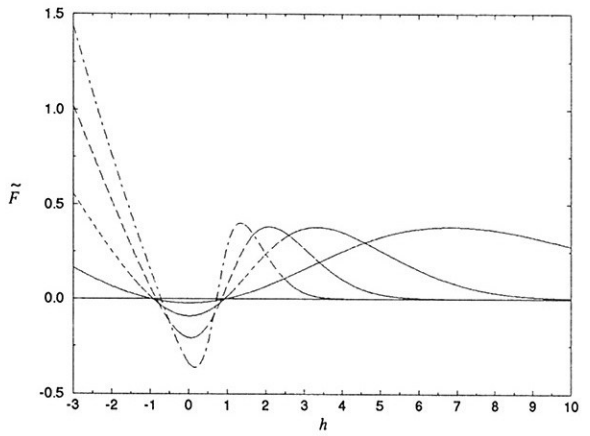


Figura 7.10: O gráfico mostra o comportamento de $\tilde{F}$ em função do campo h (caso $\xi = +1$) para $K = K^*$ e diferentes estágios de aprendizagem: $\rho = 0$ (——), $\rho = 0.2$ (·····), $\rho = 0.4$ (- - - -), $\rho = 0.6$ (– – –) e $\rho = 0.8$ (– - – -). Para $\rho = 0$ temos $\tilde{F} = 0$.

7.5 Aprendizagem na presença de ruído aditivo

Vamos investigar as propriedades de generalização no caso da aprendizagem na presença de ruído aditivo, definido no capítulo 2. Para facilitar a notação, vamos definir:

$$R = \sqrt{1 - \omega^2 \rho^2} \qquad e \qquad \bar{R} = \sqrt{1 - \bar{\omega}^2 \rho^2}$$

e para facilitar a comparação entre os dois tipos de ruído, vamos utilizar as grandezas Ω e $\bar{\Omega}$, que são as taxas efetivas de inversão da resposta do professor e definidas como:

$$\Omega = \frac{1}{\pi} \arccos(\omega) \qquad e \qquad \bar{\Omega} = \frac{1}{\pi} \arccos(\bar{\omega}) \,, \tag{7.28}$$

de modo que quando $\mathbf{B} = \tilde{\mathbf{B}}$, temos $\omega = 1$ e $\Omega = 0$, que corresponde ao caso sem ruído ($\varepsilon = 0$); já quando $\mathbf{B}$ e $\tilde{\mathbf{B}}$ forem ortogonais, $\omega = 0$ e $\Omega = 1/2$, teremos a situação de máximo ruído, que corresponde a $\varepsilon = 1/2$. Dessa forma o diagrama de robustez a ser estudado será $\bar{\Omega} \times \Omega$.

7.5.1 Função modulação e equação de evolução do overlap

Podemos partir da expressão geral definida em (7.7), onde neste caso as variáveis disponíveis são: a saída do professor corrompida $\tilde{\sigma}$ e o campo local do aluno, portanto vamos calcular:

$$\tilde{F} = \frac{J}{\rho} \int db \; (b - \rho h) \; P(b|\tilde{\sigma}, h) \,. \tag{7.29}$$

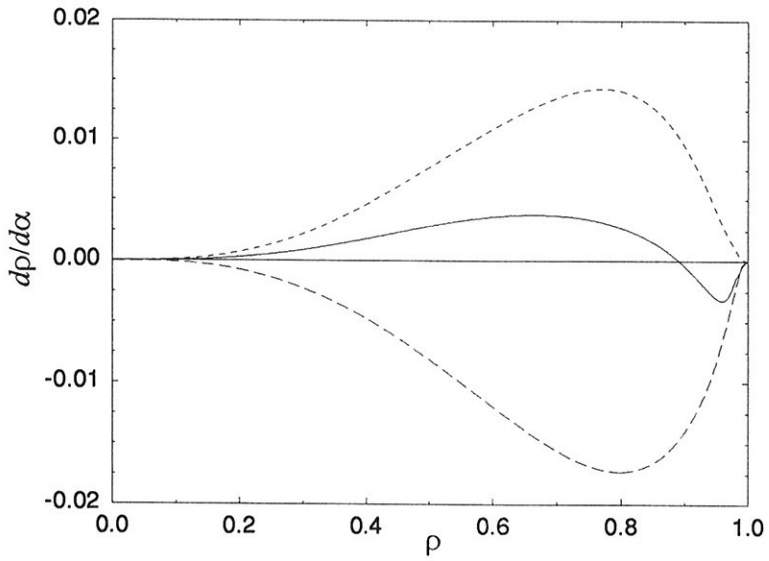


Figura 7.11: O gráfico mostra o comportamento de $\dot{\rho}$ em função de ρ para $K = K^*$ e diferentes pares (η, ε): (0.1,0.2)(- - - -); (0.03,0.23)(—); (0.1,0.4)(– – –).

É fácil verificar que probabilidade condicional $P(b|\tilde{\sigma}, h)$ pode ser escrita em termos de outras probabilidades:

$$P(b|\tilde{\sigma}, h) = \frac{P(\tilde{\sigma}|b)P(b|h)}{P(\tilde{\sigma}|h)}, \tag{7.30}$$

onde a probabilidade $P(b|h)$ é

$$P(b|h) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{(b-\rho h)^2}{2(1-\rho^2)}\right\} \tag{7.31}$$

e as outras duas são dadas respectivamente por[5]:

$$P(\tilde{\sigma}|b) = H\left(\frac{\tilde{\sigma}\omega b}{\sqrt{1-\omega^2}}\right) + \tilde{\sigma}H\left(\frac{K-\omega b}{\sqrt{1-\omega^2}}\right) - \tilde{\sigma}H\left(\frac{K+\omega b}{\sqrt{1-\omega^2}}\right) \tag{7.32}$$

$$P(\tilde{\sigma}|h) = H\left(\frac{\tilde{\sigma}\omega\rho h}{\sqrt{1-\omega^2\rho^2}}\right) + \tilde{\sigma}H\left(\frac{K-\omega\rho h}{\sqrt{1-\omega^2\rho^2}}\right) - \tilde{\sigma}H\left(\frac{K+\omega\rho h}{\sqrt{1-\omega^2\rho^2}}\right). \tag{7.33}$$

Substituindo estas expressões na equação (7.29), obtemos integrais cujos integrandos são dados pelo produto de uma gaussiana, um termo linear e uma função $H(x)$, que podem ser realizadas por partes, utilizando o resultado:

$$\frac{\partial H\left[f(x)\right]}{\partial x} = -\frac{1}{\sqrt{2\pi}} e^{-f^2(x)/2} \frac{\partial f(x)}{\partial x}. \tag{7.34}$$

[5]detalhes da obtenção destas probabilidades podem ser encontrados no apêndice A.

Após estes cálculos, chega-se à seguinte expressão para a função modulação:

$$\tilde{F} = \tilde{\sigma}\frac{J\lambda^2\rho\,\bar{\omega}}{\sqrt{2\pi}\,\bar{R}}\frac{e^{-\frac{\bar{\omega}^2\rho^2h^2}{2\bar{R}^2}}}{D(\tilde{\sigma},\bar{\omega},\rho,h)}\left[2e^{-\frac{K^2}{2\bar{R}^2}}\cosh\left(\frac{K\bar{\omega}\rho h}{2\bar{R}}\right)-1\right], \tag{7.35}$$

onde usamos a notação $D(\tilde{\sigma},\omega,\rho,h) \equiv P(\tilde{\sigma}|h)$[6]. *Note-se que* $\bar{\omega}$ *que aparece na função modulação é o valor estimado do ruído. O gráfico da figura 7.12 mostra o comportamento de* $\tilde{F}$ *para um nível de ruído* $\Omega = 0.2$.

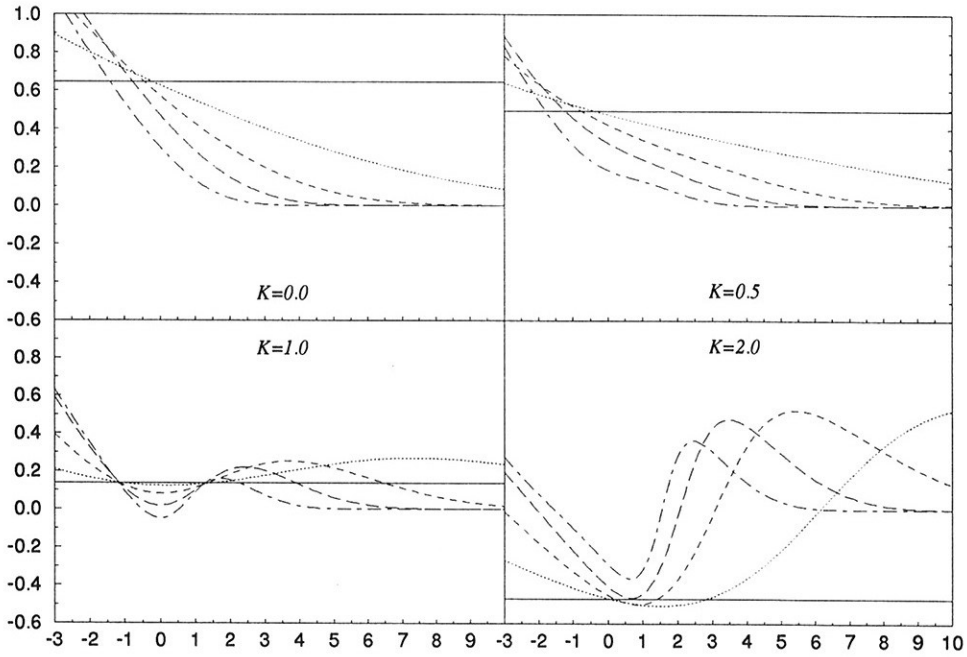


Figura 7.12: Função modulação ótima no PRW em função do valor da estabilidade $h\xi$, *para* $\Omega = 0.2$ *e diferentes valores de* K *e do overlap:* $\rho = 0$ (——), $\rho = 0.2$ (·····), $\rho = 0.4$ (- - - -), $\rho = 0.6$ (– – –) *e* $\rho = 0.8$ (– – – -).

Por esse gráfico, observamos que a presença de ruído aditivo altera o módulo da função modulação, mas praticamente não altera a sua forma. Nesse caso, assim como na ausência de ruído (figura 7.6), não se observam regiões onde o aluno perde a confiança no professor, que se caracterizam pela diminuição de $\tilde{F}$ *para valores grandes e negativos da estabilidade, como ocorre no caso do ruído multiplicativo (figura 7.7).*

Procedendo de maneira análoga à utilizada nos outros casos aqui estudados, dedu-

[6]para $D(\tilde{\sigma},\bar{\omega},\rho,h)$ basta substituir ω por $\bar{\omega}$.

zimos a equação de evolução do overlap, que neste caso é dada por:

$$\begin{aligned}\frac{d\rho}{d\alpha} &= \frac{\lambda^4\rho^3}{\pi}\int P(h)\,dh\,\frac{\bar{\omega}}{\bar{R}}\frac{D(\omega,\rho,h)}{D(\bar{\omega},\rho,h)}e^{-\frac{\bar{\omega}^2\rho^2h^2}{2\bar{R}^2}}\left[2e^{-\frac{K^2}{2\bar{R}^2}}-1\right]\\ &\times\left\{\frac{\omega}{R}\frac{e^{-\frac{\omega^2\rho^2h^2}{2R^2}}}{D(\omega,\rho,h)}\left[2e^{-\frac{K^2}{2R^2}}-1\right]-\frac{1}{2}\frac{\bar{\omega}}{\bar{R}}\frac{e^{-\frac{\bar{\omega}^2\rho^2h^2}{2\bar{R}^2}}}{D(\bar{\omega},\rho,h)}\left[2e^{-\frac{K^2}{2\bar{R}^2}}-1\right]\right\}\,. \qquad (7.36)\end{aligned}$$

$D(\omega,\rho,h)$ pode ser obtido tomando-se $\tilde{\sigma}=+1$ na equação (7.33). Definindo as funções $\mathcal{A}(\rho,\omega,\bar{\omega},K)$ e $\mathcal{B}(\rho,\omega,\bar{\omega},K)$

$$\mathcal{A}(\rho,\omega,\bar{\omega},K)=\int P(h)\,dh\,\frac{\omega}{R}\frac{\bar{\omega}}{\bar{R}}\frac{e^{-\frac{\omega^2\rho^2h^2}{2R^2}}e^{-\frac{\bar{\omega}^2\rho^2h^2}{2\bar{R}^2}}}{D(\bar{\omega},\rho,h)}\left[2e^{-\frac{K^2}{2R^2}}-1\right]\left[2e^{-\frac{K^2}{2\bar{R}^2}}-1\right] \qquad (7.37)$$

$$\mathcal{B}(\rho,\omega,\bar{\omega},K)=\int P(h)\,dh\,\frac{\bar{\omega}^2}{\bar{R}^2}\frac{D(\omega,\rho,h)}{D^2(\bar{\omega},\rho,h)}e^{-\frac{\bar{\omega}^2\rho^2h^2}{\bar{R}^2}}\left[2e^{-\frac{K^2}{2\bar{R}^2}}-1\right]^2\,, \qquad (7.38)$$

é possível reescrever $\dot{\rho}$ como:

$$\frac{d\rho}{d\alpha}=\frac{\lambda^4\rho^3}{\pi}\left\{\mathcal{A}(\rho,\omega,\bar{\omega},K)-\frac{1}{2}\mathcal{B}(\rho,\omega,\bar{\omega},K)\right\}\,. \qquad (7.39)$$

O termo que aparece entre chaves na equação (7.39) define o sinal de $\dot{\rho}$. Analisando seu valor nos limites assintóticos podemos construir o diagrama de robustez, como foi feito para o caso do ruído multiplicativo.

Diagrama de robustez

Para construir o diagrama $\bar{\Omega}\times\Omega$ precisamos investigar o comportamento assintótico de $\dot{\rho}$. Vamos inicialmente olhar o limite $\rho=0$. Neste limite é possível obter analiticamente a linha do diagrama que separa a fase onde $\dot{\rho}$ é sempre negativo, exceto para $\rho=1$, implicando na existência de um regime onde a rede não é capaz de generalizar. Tomando $\rho=0$, $\bar{R}=R=1$ e é fácil verificar que equações (7.37) e (7.38) se reduzem a:

$$\begin{aligned}\mathcal{A}(\omega,\bar{\omega},K) &= 2\omega\bar{\omega}\left[2e^{-K^2/2}-1\right]^2 \qquad (7.40)\\ \mathcal{B}(\bar{\omega},K) &= 2\bar{\omega}^2\left[2e^{-K^2/2}-1\right]^2\,, \qquad (7.41)\end{aligned}$$

e de acordo com a equação (7.39), $\dot{\rho}$ muda de sinal quando $\bar{\omega}=2\omega$. É importante enfatizar que este resultado independe de K. Utilizando a relação entre ω e Ω ($\bar{\omega}$ e $\bar{\Omega}$), podemos escrever:

$$\Omega=\frac{1}{\pi}\arccos\left[\frac{1}{2}\cos\left(\bar{\Omega}\pi\right)\right]\,. \qquad (7.42)$$

Esta equação determina analiticamente a curva do diagrama de robustez que separa as regiões Ib e II, que são respectivamente as regiões onde a rede generaliza imperfeitamente e onde ela não é capaz de generalizar. Para obter a outra curva, que separa as regiões de generalização perfeita e imperfeita, olhamos para o limite $\rho = 1$, *que nesse caso pode ser tomado diretamente na equação (7.36). Olhando para as funções* $\mathcal{A}$ *e* $\mathcal{B}$, *observamos que é preservada a dependência em* K, *exceto para os limites assintóticos de* $\bar{\Omega}$ *e* Ω. *A análise do sinal de* $\dot{\rho}$ *foi feita numericamente: a cada valor de* $\bar{\Omega}$ *procurou-se o valor de* Ω *para o qual* $\dot{\rho}$ *muda de sinal.*

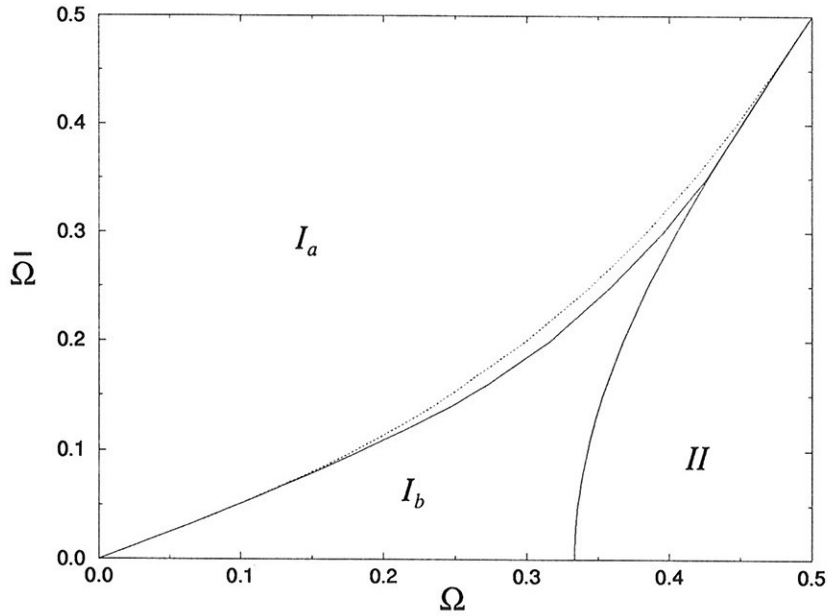


Figura 7.13: Diagrama de robustez PRW para o caso de ruído aditivo. Observa-se a existência de três fases: generalização perfeita (Ia), generalização imperfeita (Ib) e incapacidade de generalização (II). A transição entre as fases Ia e Ib depende do valor de K: *a linha tracejada tem* $K = 0$ *e a linha cheia,* $K = 2$.

As características do diagrama $\bar{\Omega} \times \Omega$ *e, portanto, as propriedades de generalização por ele descritas, são semelhantes ao caso do ruído multiplicativo, no que diz respeito à existência de três regimes de generalização. As principais diferenças são: a forma da linha que separa Ib e II, que no caso de ruído multiplicativo é uma reta, a forma da linha que separa as fases de generalização perfeita e imperfeita, que no limite* $\bar{\Omega} = \Omega = 0$ *apresenta uma inclinação não nula e finalmente a dependência em* K *para* $\rho = 1$. *Estas características já foram discutidas por Copelli et al., 1997.*

7.6 Erro de generalização

Uma maneira de analisar o desempenho de um algoritmo de aprendizagem numa dada rede se dá através da investigação do comportamento assintótico do erro de generalização. Vamos considerar o caso sem ruído. Isso pode ser feito sem perda de generalidades pois a potência de decaimento de e_G com α não é afetada pela presença de ruído, o que muda sim é a constante multiplicativa de α, o que, portanto, não influencia a análise aqui apresentada. O gráfico da figura 7.14 mostra como se comporta o erro de generalização com α para alguns valores de K.

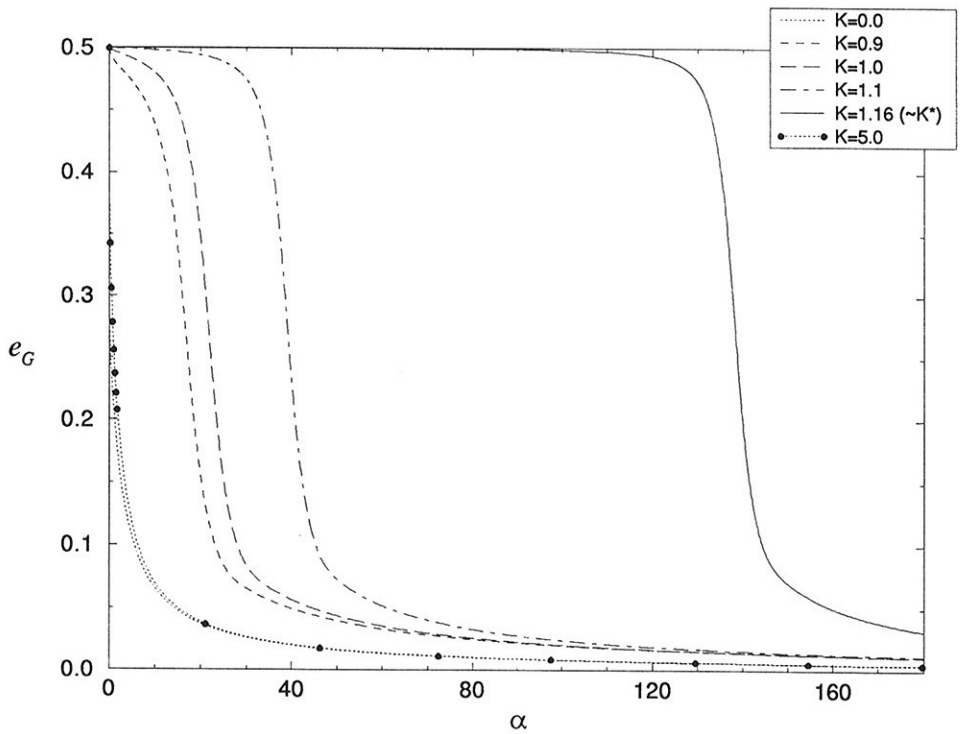


Figura 7.14: Este gráfico mostra o comportamento do erro de generalização em função de α para diferentes valores de K. Vemos que à medida que K se aproxima de K^*, o decaimento fica mais lento, ocorrendo o inverso à medida que K se afasta de K^*.

É interessante notar a presença de platôs para valores de K próximos a K^*. Essa lentidão no decaimento do erro de generalização está relacionada à simetria existente para $K = K^*$, isto é, não há informação sobre o valor do campo local do professor, como mencionamos na seção 7.4. Notamos também que a largura do platô diverge com $|K - K^*|^{-\gamma}$, onde $\gamma \approx 1$.

7.6.1 Comportamento no limite $\alpha \longrightarrow \infty$

Partindo da suposição que o erro de generalização assintótico é descrito por uma lei de potência tipo C/α^{γ}, construimos os gráficos da figura 7.15. Através destes gráficos, é possível determinar os valores de C e γ.

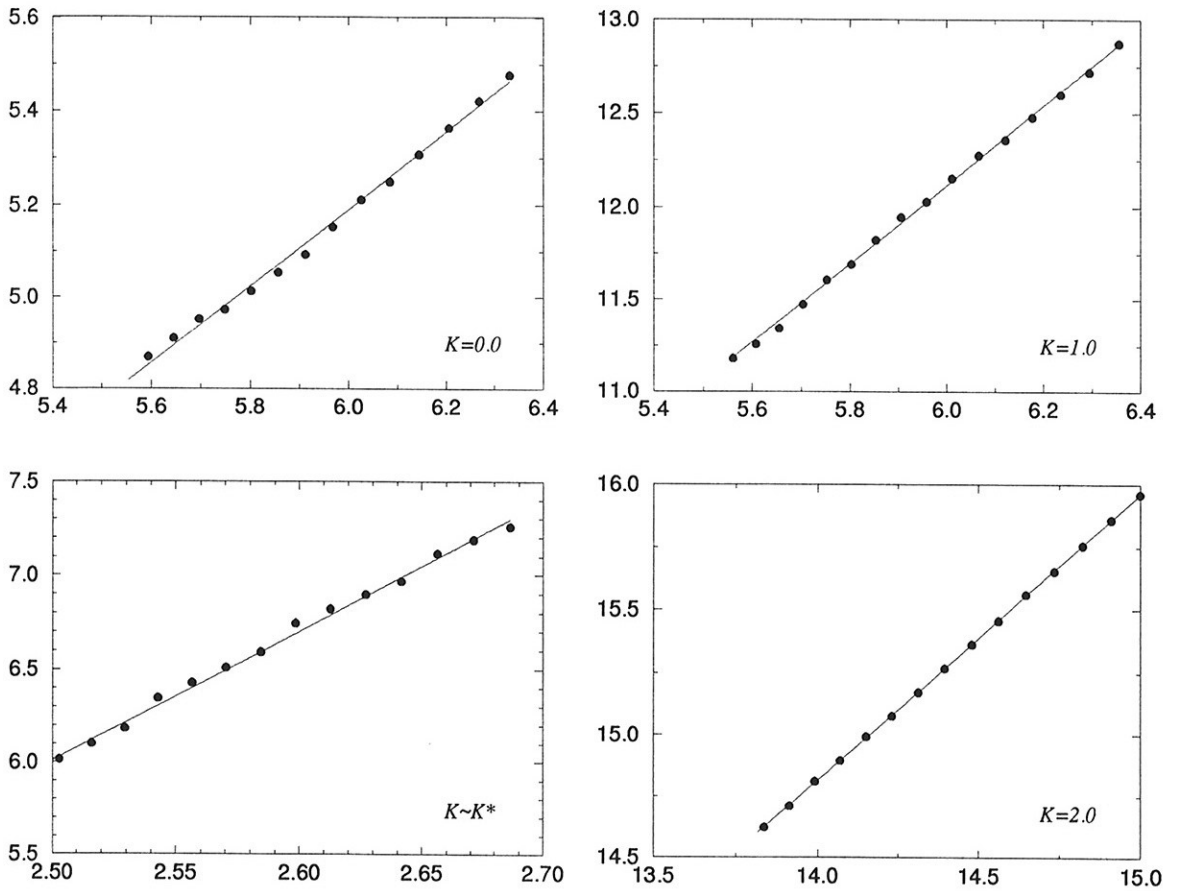


Figura 7.15: A partir deste gráfico é possível determinar a potência do decaimento de e_G com α. As curvas obtidas são lineares, o que indica que assintoticamente o erro decai com α^{-1}. A inclinação varia com K, indicando que a constante multiplicativa é função desta grandeza.

Estes gráficos mostram o comportamento de e_G com $1/\alpha$, e mostram que a potência de decaimento é de fato 1, independentemente do valor de K.

Através destes gráficos também verificamos que a constante C varia em função do parâmetro K, pois varia a inclinação da reta, e portanto que $C = C(K)$. A tabela abaixo mostra alguns valores aproximados de C em função do valor de K. Observamos que à medida que $K \longrightarrow \infty$ temos $C(K) \longrightarrow C(0)$.

K	$C(K)$
0.0	0.88
1.0	2.11
1.16	6.90
2.0	1.15
5.0	0.88

Com esses resultados, verificamos que o caso K^* é a rede mais difícil de ser ensinada, apesar de ter sido apontada (Bex et al., 1995) como a melhor rede para se construir redes mais complexas.

8 Considerações Finais

Neste trabalho estudamos a habilidade de generalização no aprendizado supervisionado on-line de duas redes neurais com arquitetura de árvore: uma máquina paridade e um perceptron com função de transferência não monotônica.

Inicialmente enfocamos o aprendizado supervisionado on-line usando o algoritmo ótimo no caso específico de uma máquina paridade com duas unidades escondidas, que havia sido estudada anteriormente por Kabashima utilizando um algoritmo conhecido como algoritmo de mínima ação (LAA). Embora algumas características do algoritmo ótimo lembrem o LAA, as diferenças são responsáveis pela obtenção de resultados muito superiores. Em seu estudo, Kabashima observa a existência de um valor crítico J_c a partir do qual emerge a capacidade de generalização; obtém um erro de generalização que decai a zero com $\alpha^{-1/3}$ e aponta a existência de um valor crítico de ruído multiplicativo $\varepsilon \approx 0.175$, acima do qual a rede perde totalmente a capacidade de generalização.

Em oposição a estes resultados, verificamos que o processo de otimização fornece um comportamento qualitativa e quantitativamente diferente para as três características descritas. Em primeiro lugar, quando a rede é ensinada de maneira ótima, não existe um comprimento crítico J_c a partir do qual ela passa a generalizar. O que se observa é que, caso as condições iniciais forem tais que $\rho_0 = \mathcal{O}(1)$, a rede se torna capaz de generalizar logo que tem início o processo de aprendizagem, isto é, quando são apresentados os primeiros exemplos, basta que o valor inicial do overlap não seja nulo. A condição $\rho_0 \neq 0$ aponta a necessidade de um conhecimento prévio, sem o qual o processo de aprendizagem não se inicia. Também indica que $\rho_0 = 0$ constitui um ponto fixo repulsivo da dinâmica on-line, entretanto deve-se enfatizar que uma pequena flutuação deste valor é suficiente para que o overlap evolua, de forma que seu valor inicial pode ser tão pequeno quanto se queira. Observamos que quanto menos conhecimento a rede possuir a priori, mais lenta será a aquisição da capacidade de generalização, isto é, mais lento será o decaimento do erro de generalização nos estágios iniciais da aprendizagem. A origem deste retardamento está relacionada à simetria da máquina paridade e também foi detectado no caso do aprendizado supervisionado off-line

(Hansel et al., 1992). A relação entre o efeito de retardamento na dinâmica off-line e o ponto fixo repulsivo da dinâmica on-line parece ser geral e também foi detectada no caso do aprendizado não supervisionado off-line de um perceptron a partir de uma distribuição estruturada de exemplos (Biehl & Mietzner, 1993; Biehl, 1994).

Em segundo lugar, quando a rede aprende de maneira ótima, o erro de generalização decai assintoticamente mais rápido, sendo que para valores grandes de α este decaimento é $0.88\alpha^{-1}$ na ausência de ruído, e mesmo que o aprendizado ocorra na presença de ruído multiplicativo o erro decai assintoticamente a zero com α^{-1} para qualquer nível de ruído. Esses resultados mostram que algumas das características encontradas por Kabashima têm origem na escolha do algoritmo utilizado e desaparecem quando se utiliza o algoritmo ótimo.

Os resultados para a máquina paridade $K = 2$ foram generalizados para valores arbitrários de K. Neste caso novamente encontramos o valor $0.88\alpha^{-1}$ para o erro de generalização na ausência de ruído, independente do número K de unidades escondidas. Este parece ser um resultado universal para o aprendizado otimizado de redes multicamadas com estrutura de árvore, já que foi obtido também no caso da máquina comitê com estrutura de árvore nas mesmas condições de aprendizagem.

Em situações de aprendizagem mais realistas o nível de ruído (ε) não é conhecido, e portanto deve ser estimado (η). Utilizando o algoritmo ótimo, verificamos que a máquina paridade é capaz de generalizar perfeitamente na presença de ruído multiplicativo sempre que for conhecido o nível de ruído. No caso do desconhecimento desta quantidade a rede é capaz de generalizar, ainda que este valor seja subestimado. O estudo da aprendizagem com ruído leva à construção de diagramas de fase no espaço dos ruídos (ruído estimado versus ruído real) chamados de diagramas de robustez. Nestes diagramas observa-se a existência de três regimes de aprendizagem. Quando o ruído é superestimado a rede é capaz de generalizar perfeitamente, isto é, seu overlap atinge assintoticamente o valor máximo. Esta constitui uma das fases. Ainda que o ruído seja subestimado, a rede é capaz de generalizar em duas situações. Se o ruído for levemente subestimado, assintoticamente, a rede aprende de maneira perfeita. Entretanto, determinadas escolhas de η levam à capacidade de generalização imperfeita: o valor máximo de ρ é inferior a um, quando $\dot{\rho}$ se torna negativo. Esta constitui a segunda fase. A terceira fase ocorre quando a variação do overlap com os exemplos apresentados é negativa, o que implica na diminuição do overlap até atingir o valor zero, que neste caso constitui um ponto fixo atrativo da dinâmica.

Neste trabalho também estudamos a aprendizagem supervisionada on-line na presença de ruído multiplicativo e aditivo para o perceptron com função de transferência não monotônica. No primeiro caso obtivemos exatamente o mesmo diagrama de robustez, e portanto mostramos que esta rede possui as mesmas propriedades de generalização de uma

máquina paridade. Isso reforça a idéia de uma universalidade de comportamentos que ainda precisa ser entendida a fundo. Observamos também que esses comportamentos independem do valor de K e das condições iniciais, exceto para $K = K^*$, que constitui um caso especial. Algumas características desta rede especial foram apontadas por Bex et al. (1995). Segundo estes autores, a rede K^* é considerada como o melhor perceptron para construir redes mais complexas, pois satura os limites de memorização e generalização, como mencionamos no capítulo 7. Entretanto, nossos resultados mostram que essa rede é a mais difícil de ser ensinada, devido à simetria entre estabilidades positivas e negativas. Também em função desta simetria, esta rede especial necessita de um conhecimento a priori para escapar do ponto fixo repulsivo da dinâmica $\rho = 0$. Esta simetria é do mesmo tipo que leva ao aparecimento de platôs na máquina paridade. No caso de ruído aditivo o diagrama também apresenta os três regimes de generalização, entretanto depende da arquitetura da rede (depende de K). Essa mesma característica foi observada no caso da máquina paridade e da máquina comitê (Copelli et al., 1997).

A Derivação das probabilidades condicionais

A.1 Máquina Paridade

A.1.1 Derivação de $P(b_1|\Sigma_B, h_1, h_2)$

Queremos calcular a probabilidade de b_1 dados: a saída do professor e os campos locais do aluno . Esta probabilidade pode ser escrita em termos das probabilidades conjuntas $P(b_1, \Sigma_B, h_1, h_2)$ e de $P(\Sigma_B, h_1, h_2)$:

$$P(b_1|\Sigma_B, h_1, h_2) = \frac{P(b_1, \Sigma_B, h_1, h_2)}{P(\Sigma_B, h_1, h_2)} \,. \tag{A.1}$$

A probabilidade conjunta $P(b_1, \Sigma_B, h_1, h_2)$ pode ser calculada a partir de:

$$P(b_1, \Sigma_B, h_1, h_2) = \int db_2 \; P(h_1, h_2, b_1, b_2) \; \delta \left(\Sigma_B - \prod_{k=1,2} \sigma_{b_k} \right) \,, \tag{A.2}$$

com $\sigma_{bk} = sinal(b_k)$. Devido à estrutura de árvore desta rede, a probabilidade $P(h_1, h_2, b_1, b_2)$ se torna um produto de 2 termos $P(h_j, b_j)$, dados pela equação (4.8). Então, podemos escrever:

$$P(h_1, h_2, b_1, b_2) = P(h_1, b_1) \; P(h_2, b_2) \,. \tag{A.3}$$

Usando este resultado, obtemos:

$$P(b_1, \Sigma_B, h_1, h_2) = \int db_2 P(h_1, b_1) P(h_2, b_2) \delta(\Sigma_B - \sigma_{b_1} \sigma_{b_2}). \tag{A.4}$$

Escrevendo:

$$P(h_2, b_2) = P(h_2) P(h_2|b_2) \tag{A.5}$$

chegamos à:

$$P(b_1, \Sigma_B, h_1, h_2) = P(h_1, b_1)P(h_2) \int db_2 P(h_2|b_2)\delta(\Sigma_B - \sigma_{b_1}\sigma_{b_2}). \tag{A.6}$$

Utilizando a representação de Fourier para a função delta:

$$P(b_1, \Sigma_B, h_1, h_2) = P(h_2)P(h_1, b_1) \int db_2 P(h_2|b_2) \int \frac{dx}{2\pi} e^{(ix\Sigma_B - ix\sigma_{b_1}\sigma_{b_2})}, \tag{A.7}$$

a integral em b_2 resulta em:

$$\int db_2 P(h_2|b_2) e^{-ix\sigma_{b_1}\sigma_{b_2}} = e^{(ix\sigma_{b_1})} \cdot H\left(\frac{h_2}{\lambda_2}\right) + e^{(-ix\sigma_{b_1})} \cdot H\left(-\frac{h_2}{\lambda_2}\right). \tag{A.8}$$

Substituindo este resultado em (A.7) e fazendo a integral em x, obtemos o resultado:

$$P(b_1, \Sigma_B, h_1, h_2) = P(h_1, b_1)P(h_2)H\left(-\Sigma_B \frac{\sigma_{b_1} h_2}{\lambda_2}\right). \tag{A.9}$$

O passo seguinte é calcular $P(\Sigma_B, h_1, h_2)$, o que pode ser feito integrando em b_1 a probabilidade $P(b_1, \Sigma_B, h_1, h_2)$. Usando a relação $P(h_1, b_1) = P(h_1)P(h_1|b_1)$, precisamos calcular:

$$P(\Sigma_B, h_1, h_2) = P(h_1)P(h_2) \int db_1 P(h_1|b_1) H\left(-\Sigma_B \frac{\sigma_{b_1} h_2}{\lambda_2}\right). \tag{A.10}$$

Esta integral resulta em:

$$\int db_1 P(h_1|b_1) H\left(-\Sigma_B \sigma_{b_1} \frac{h_2}{\lambda_2}\right) = H\left(\Sigma_B \frac{h_2}{\lambda_2}\right) H\left(\frac{h_1}{\lambda_1}\right) + H\left(-\Sigma_B \frac{h_2}{\lambda_2}\right) H\left(-\frac{h_1}{\lambda_1}\right), \tag{A.11}$$

e finalmente obtemos a probabilidade condicional procurada:

$$P(b_1|\Sigma_B, h_1, h_2) = \frac{P(b_1|h_1) H\left(-\Sigma_B \sigma_{b_1} \frac{h_2}{\lambda_2}\right)}{H\left(\Sigma_B \frac{h_2}{\lambda_2}\right) H\left(\frac{h_1}{\lambda_1}\right) + H\left(-\Sigma_B \frac{h_2}{\lambda_2}\right) H\left(-\frac{h_1}{\lambda_1}\right)}. \tag{A.12}$$

A.1.2 Derivação de $P^K(b_k|\Sigma_B, \{h_j\})$

Queremos calcular a probabilidade de b_k, dada a saída do professor (Σ_B), e os campos locais ($\{h_j\}$), isto é, a probabilidade condicional do campo no k-ésimo ramo do professor. Sabemos que esta probabilidade pode ser escrita em termos das probabilidades conjuntas $P^K(b_k, \Sigma_B, \{h_j\})$ e de $P^K(\Sigma_B, \{h_j\})$:

$$P^K(b_k|\Sigma_B, \{h_j\}) = \frac{P^K(b_k, \Sigma_B, \{h_j\})}{P^K(\Sigma_B, \{h_j\})}, \tag{A.13}$$

onde

$$P^K(\Sigma_B, \{h_j\}) = P^K(\Sigma_B|\{h_j\})P(\{h_j\}) \,. \tag{A.14}$$

O primeiro passo é calcular $P^K(b_k, \Sigma_B, \{h_j\})$, *que pode ser feito integrando-se a probabilidade* $P^K(\{b_j\}, \{h_j\})$ *em* $b_{j\neq k}$ *e impondo que:*

$$\Sigma_B = \prod_{j=1}^{K} \sigma_{b_j} \,.$$

Podemos então escrever:

$$P^K(b_k, \Sigma_B, \{h_j\}) = \int \left[\prod_{j\neq k}^{K-1} db_j\right] \prod_{j=1}^{K} P(h_j, b_j)\, \theta\left(\Sigma_B \prod_{j=1}^{K} \sigma_{b_j}\right) \,, \tag{A.15}$$

e separando o termo $j = k$ *temos:*

$$P^K(b_k, \Sigma_B, \{h_j\}) = P(h_k, b_k) \int \prod_{j\neq k}^{K-1} db_j\, P(h_j, b_j)\, \theta\left(\sigma_{b_k}\Sigma_B \prod_{j\neq k}^{K-1} \sigma_{b_j}\right) \,. \tag{A.16}$$

O sinal do campo no ramo k *pode ser positivo ou negativo, e como não é possível conhecê-lo, precisamos levar em conta as duas situações, portanto:*

$$\begin{aligned} P^K(b_k, \Sigma_B, \{h_j\}) &= P(h_k, b_k)\left\{\delta_{\sigma_{b_k},+1} \int \prod_{j\neq k}^{K-1} P(h_j, b_j)\, \theta\left(\Sigma_B \prod_{j\neq k}^{K-1} \sigma_{b_j}\right)\right. \\ &+ \left.\delta_{\sigma_{b_k},-1} \int \prod_{j\neq k}^{K-1} P(h_j, b_j)\, \theta\left(-\Sigma_B \prod_{j\neq k}^{K-1} \sigma_{b_j}\right)\right\} \,. \end{aligned} \tag{A.17}$$

Usando a notação:

$$P^{K-1}(\pm\Sigma_B, \{h_j\}_{j\neq k}) = \int \prod_{j\neq k}^{K-1} P(h_j, b_j)\, \theta\left(\pm\Sigma_B \prod_{j\neq k}^{K-1} \sigma_{b_j}\right) \,, \tag{A.18}$$

finalmente obtemos:

$$\begin{aligned} P^K(b_k, \Sigma_B, \{h_j\}) &= P(h_k, b_k)\left\{\delta_{\sigma_{b_k},+1} P^{K-1}(\Sigma_B, \{h_j\}_{j\neq k})\right. \\ &+ \left.\delta_{\sigma_{b_k},-1} P^{K-1}(-\Sigma_B, \{h_j\}_{j\neq k})\right\} \,. \end{aligned} \tag{A.19}$$

O passo seguinte é calcular $P^K(\Sigma_B, \{h_j\})$, *que pode ser feito integrando a probabilidade* $P^K(b_k, \Sigma_B, \{h_j\})$ *na variável* b_k.

$$\begin{aligned} P^K(\Sigma_B, \{h_j\}) &= \int db_k\, P^K(b_k, \Sigma_B, \{h_j\}) \\ &= P^{K-1}(-\Sigma_B, \{h_j\}_{j\neq k}) \int_{-\infty}^{0} P(b_k, h_k)\, db_k + P^{K-1}(\Sigma_B, \{h_j\}_{j\neq k}) \int_{0}^{\infty} P(b_k, h_k)\, db_k \\ &= P(h_k)\left\{P^{K-1}(\Sigma_B, \{h_j\}_{j\neq k}) H\left(-\frac{h_k}{\lambda_k}\right) + P^{K-1}(-\Sigma_B, \{h_j\}_{j\neq k}) H\left(\frac{h_k}{\lambda_k}\right)\right\} \end{aligned} \tag{A.20}$$

onde utilizamos os resultados

$$\begin{aligned}\int_{-\infty}^{0} P(b_k|h_k)\, db_k &= H\left(+\frac{h_k}{\lambda_k}\right)\\ \int_{0}^{-\infty} P(b_k|h_k)\, db_k &= H\left(-\frac{h_k}{\lambda_k}\right).\end{aligned}$$

Finalmente, usando as equações (A.19) e (A.20), obtemos a probabilidade condicional desejada:

$$P^K(b_k|\Sigma_B,\{h_j\}) = \frac{P(b_k|h_k)\left\{\delta_{\sigma_{b_k},+1}P^{K-1}(\Sigma_B|\{h_j\}_{j\neq k}) + \delta_{\sigma_{b_k},-1}P^{K-1}(-\Sigma_B|\{h_j\}_{j\neq k})\right\}}{H\left(-\frac{h_k}{\lambda_k}\right)P^{K-1}(\Sigma_B|\{h_j\}_{j\neq k}) + H\left(\frac{h_k}{\lambda_k}\right)P^{K-1}(-\Sigma_B|\{h_j\}_{j\neq k})} \tag{A.21}$$

A.2 Perceptron Reversed-Wedge

A.2.1 Derivação de $P(b|\xi,h)$

Para obter $P(b|\xi,h)$, isto é a distribuição de probabilidade do campo local b no professor dada a saída corrompida por ruído ξ e o campo local do aluno h, precisamos conhecer as distribuições $P(b,\xi,h)$ e $P(\xi,h)$. Vamos inicialmente deduzir $P(\xi,h)$. Sabemos que:

$$P(\xi,h) = P(\xi|h)\,P(h)\,, \tag{A.22}$$

onde $P(\xi|h)$ pode ser escrita em termos da saída correta do professor σ[1]*:*

$$P(\xi|h) = \sum_{\sigma} P(\xi|h,\sigma)\,P(\sigma|h)\,. \tag{A.23}$$

Entretanto, o valor de ξ só depende de σ e não do valor de h. Dessa forma podemos escrever $P(\xi|h,\sigma) = P(\xi|\sigma)$, ou seja:

$$\begin{aligned}P(\xi|h) &= \sum_{\sigma} P(\xi|\sigma)\,P(\sigma|h)\\ P(\xi|h) &= \sum_{\sigma} P(\xi|\sigma)\,\frac{P(\sigma,h)}{P(h)}\,.\end{aligned} \tag{A.24}$$

A distribuição de probabilidade $P(h)$ já é conhecida, assim como $P(\xi|\sigma)$ (equação 7.6). Falta calcular $P(\sigma,h)$, o que pode ser feito integrando em b, a probabilidade $P(b,\sigma,h)$ dada por:

$$P(b,\sigma,h) = P(b,h)\,\delta_{\sigma,\sigma'}\,, \tag{A.25}$$

[1] vamos omitir o índice B para não sobrecarregar a notação

onde $\sigma' = sinal[b^2(b^2 - K^2)]$ e $P(b,h)$ é

$$P(b,h) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-h^2 - b^2 + 2bh\rho}{2(1-\rho^2)}\right\} . \tag{A.26}$$

Assim, precisamos fazer a integral:

$$P(\sigma,h) = \int_{-\infty}^{+\infty} P(b,\sigma,h)\, db = P(h) \int_{-\infty}^{+\infty} dbP(b|h)\, \delta_{\sigma,\sigma'} . \tag{A.27}$$

onde usamos o fato que $P(b,h) = P(b|h)\, P(h)$. Para resolver esta integral, dividimos o intervalo de integração em 4 partes onde σ' assume os valores ± 1.

$$\begin{aligned} P(\sigma,h) &= P(h)\left\{\delta_{\sigma',-1}\left[\int_{-\infty}^{-K} db\, P(b|h) + \int_{0}^{K} db\, P(b|h)\right]\right. \\ &+ \left.\delta_{\sigma',+1}\left[\int_{-K}^{0} db\, P(b|h) + \int_{K}^{+\infty} db\, P(b|h)\right]\right\} . \end{aligned} \tag{A.28}$$

Essas integrais são muito simples e resultam em funções $H(x)$:

$$\begin{aligned} P(\sigma,h) &= P(h)\left\{\delta_{\sigma,-1}\left[H\left(-\frac{\rho h}{\sqrt{1-\rho^2}}\right) + H\left(\frac{K+\rho h}{\sqrt{1-\rho^2}}\right) - H\left(\frac{K-\rho h}{\sqrt{1-\rho^2}}\right)\right]\right. \\ &+ \left.\delta_{\sigma,+1}\left[H\left(+\frac{\rho h}{\sqrt{1-\rho^2}}\right) - H\left(\frac{K+\rho h}{\sqrt{1-\rho^2}}\right) + H\left(\frac{K-\rho h}{\sqrt{1-\rho^2}}\right)\right]\right\} . \end{aligned} \tag{A.29}$$

Vemos que dependendo do valor assumido por σ, sobrevive apenas um dos termos da equação acima. Podemos incorporar esse resultado, escrevendo:

$$P(\sigma,h) = P(h)\left\{H\left(\frac{\sigma\rho h}{\sqrt{1-\rho^2}}\right) - \sigma H\left(\frac{K+\rho h}{\sqrt{1-\rho^2}}\right) + \sigma H\left(\frac{K-\rho h}{\sqrt{1-\rho^2}}\right)\right\} . \tag{A.30}$$

Substituindo esse resultado na expressão de $P(\xi|h)$, fazendo a soma sobre $\sigma = \pm 1$ e rearranjando os termos de modo a fatorar o ruído ε obtemos:

$$\begin{aligned} P(\xi|h) &= \varepsilon\left\{H\left(\frac{-\xi\rho h}{\sqrt{1-\rho^2}}\right) - \xi H\left(\frac{K-\rho h}{\sqrt{1-\rho^2}}\right) + \xi H\left(\frac{K+\rho h}{\sqrt{1-\rho^2}}\right)\right\} + \\ (1 &- \varepsilon)\left\{H\left(\frac{+\xi\rho h}{\sqrt{1-\rho^2}}\right) + \xi H\left(\frac{K-\rho h}{\sqrt{1-\rho^2}}\right) - \xi H\left(\frac{K+\rho h}{\sqrt{1-\rho^2}}\right)\right\} . \end{aligned} \tag{A.31}$$

Já temos $P(\xi|h)$ e $P(h)$, resta encontrar $P(b,h,\xi)$.

$$\begin{aligned} P(b,h,\xi) &= P(\xi|b,h)\, P(b,h) \\ &= \sum_{\sigma} P(\sigma,\xi|b,h)\, P(b,h) \\ &= \sum_{\sigma} P(\xi|\sigma,b,h)\, P(\sigma|b,h)\, P(b,h) \\ &= \sum_{\sigma} P(\xi|\sigma,b,h)\, P(\sigma,b,h) . \end{aligned} \tag{A.32}$$

Como já mencionamos, o valor de ξ só depende de σ, dessa forma a probabilidade condicional $P(\xi|\sigma, b, h)$ se reduz a $P(\xi|\sigma)$. Utilizando a equação (A.25) e fazendo a soma sobre σ, podemos escrever:

$$P(b, h, \xi) = P(b, h)\ \{\varepsilon\, \delta_{-\xi,\sigma'} + (1-\varepsilon)\, \delta_{+\xi,\sigma'}\} \tag{A.33}$$

e finalmente

$$P(b|h, \xi) = \frac{P(b|h)}{P(\xi|h)}\ \{\varepsilon\, \delta_{-\xi,\sigma'} + (1-\varepsilon)\, \delta_{+\xi,\sigma'}\}\ , \tag{A.34}$$

com $\sigma' = sinal[b^2(b^2 - K^2)]$. Observação: o nível de ruído que aparece aqui é ε. Quando calculamos a função modulação ótima considerando a ausência de conhecimento sobre o nível de ruído, a grandeza que aparece em $\tilde{F}$ é o ruído estimado η.

A.2.2 Derivação de $P(\tilde{\sigma}, b)$

Esta probabilidade pode ser calculada a partir da expressão

$$P(\tilde{\sigma}, b) = \int d\tilde{b} P(\tilde{\sigma}|\tilde{b}) P(\tilde{b}|b)\ . \tag{A.35}$$

onde

$$P(\tilde{b}|b) = \frac{P(\tilde{b}, b)}{P(b)} = \frac{1}{\sqrt{2\pi(1-\omega^2)}} \exp\left\{-\frac{(\tilde{b} - \omega b)^2}{2(1-\omega^2)}\right\}\ , \tag{A.36}$$

que pode ser obtida facilmente a partir da probabilidade $P(\tilde{b}, b)$ definida em (2.15) e

$$P(\tilde{\sigma}|\tilde{b}) = \delta\left(\tilde{\sigma} - sinal[\tilde{b}(\tilde{b}^2 - K^2)]\right)\ . \tag{A.37}$$

Para resolver a integral em (A.35) dividimos o intervalo de integração em 4 partes, devido à presença da função sinal:

$$P(\tilde{\sigma}|b) = \delta(\tilde{\sigma}+1)\left\{\int_{-\infty}^{-K} P(\tilde{b}|b)\, d\tilde{b} + \int_0^K P(\tilde{b}|b)\, d\tilde{b}\right\} + \delta(\tilde{\sigma}-1)\left\{\int_{-K}^{0} P(\tilde{b}|b)\, d\tilde{b} + \int_K^{\infty} P(\tilde{b}|b)\, d\tilde{b}\right\} \tag{A.38}$$

e obtemos 4 integrais cujas soluções são:

$$\int_{-\infty}^{-K} P(\tilde{b}|b)\, d\tilde{b} = H\left(\frac{K + \omega b}{\sqrt{1-\omega^2}}\right) \tag{A.39}$$

$$\int_0^K P(\tilde{b}|b)\, d\tilde{b} = H\left(\frac{-\omega b}{\sqrt{1-\omega^2}}\right) - H\left(\frac{K - \omega b}{\sqrt{1-\omega^2}}\right) \tag{A.40}$$

$$\int_{-K}^{0} P(\tilde{b}|b)\, d\tilde{b} = H\left(\frac{\omega b}{\sqrt{1-\omega^2}}\right) - H\left(\frac{K + \omega b}{\sqrt{1-\omega^2}}\right) \tag{A.41}$$

$$\int_K^\infty P(\tilde{b}|b)\, d\tilde{b} = H\left(\frac{K-\omega b}{\sqrt{1-\omega^2}}\right) \tag{A.42}$$

Com estes resultados, e incorporando o $\tilde{\sigma}$, obtemos finalmente a probabilidade desejada:

$$P(\tilde{\sigma}|b) = H\left(\frac{\tilde{\sigma}\omega b}{\sqrt{1-\omega^2}}\right) + \tilde{\sigma}H\left(\frac{K-\omega b}{\sqrt{1-\omega^2}}\right) - \tilde{\sigma}H\left(\frac{K+\omega b}{\sqrt{1-\omega^2}}\right) . \tag{A.43}$$

A derivação de $P(\tilde{\sigma}|h)$ pode ser obtida a partir deste resultado em:

$$P(\tilde{\sigma}|h) = \int db\, P(\tilde{\sigma}|b)P(b|h) . \tag{A.44}$$

O cálculo é totalmente análogo, sendo muito semelhantes os passos, as integrais obtidas e suas soluções. Dessa forma, não iremos apresentá-lo explicitamente. A idéia é escrever a probabilidade $P(\tilde{\sigma}|b)$ como

$$P(\tilde{\sigma}|b) = \int d\tilde{b}\, P(\tilde{\sigma}|\tilde{b})P(\tilde{b}|b) . \tag{A.45}$$

e resolver primeiro a integral em b, cujo resultado é a probabilidade $P(\tilde{b}|h)$ dada por:

$$P(\tilde{b}|h) = \frac{1}{\sqrt{2\pi(1-\omega^2\rho^2)}} \exp\left\{-\frac{(\tilde{b}-\omega\rho h)^2}{2(1-\omega^2\rho^2)}\right\} . \tag{A.46}$$

Novamente, o intervalo de integração é dividido em 4 partes, dando origem as funções $H(x)$. O resultado encontrado é:

$$P(\tilde{\sigma}|h) = H\left(\frac{\tilde{\sigma}\omega\rho h}{\sqrt{1-\omega^2\rho^2}}\right) + \tilde{\sigma}H\left(\frac{K-\omega\rho h}{\sqrt{1-\omega^2\rho^2}}\right) - \tilde{\sigma}H\left(\frac{K-\omega\rho h}{\sqrt{1-\omega^2\rho^2}}\right) . \tag{A.47}$$

Referências

AMIT, D.J.

1989 *Modeling Brain Function*
Cambridge, Cambridge University Press.

BEX, G.J., SERNEELS, R. & VAN DEN BROECK, C.

1995 *Storage capacity and generalization error for the reversed-wedge Ising perceptron*
Phys. Rev. E **51** *(6309).*

BIEHL, M., RIEGLER, P. & STECHERT, M.

1995 *Learning from noisy data: an exactly solvable model*
Phys. Rev. E **52** *(R4624).*

BOFFETTA G., MONASSON R. & ZECCHINA, R.

1993 *Symmetry breaking in non-monotonic neural networks*
J.Phys. A **26** *(L507).*

COPELLI, M. & CATICHA, N.

1995 *On-line learning in the committee machine*
J. Phys. A: Math. Gen. **28** *(1615).*

COPELLI, M., EICHHORN, R., KINOUCHI, O., BIEHL, M., SIMONETTI, R., RIEGLER, P. & CATICHA, N.

1997 *Noise Robustness in Multilayer Neural Networks*
aceito para publicação no Europhys. Lett.

COPELLI, M., KINOUCHI, O. & CATICHA, N.

1996 *Equivalence between learning in noisy perceptrons and tree committee machines*
Phys. Rev. E **53** *(6341).*

CRAGG, B.G. & TEMPERLEY, H.V.

1954 *Organization of Neurones: A Cooperative Analogy*
EEG and Clinical Neurophysiology **6** *(85).*

DOMANY, E., VAN HEMMEN, J.L. & SCHULTEN, K.

1991 *Models of Neural Networks*
Berlin, Springer.

GYÖRGYI, G.

1990 *First-order transition to perfect generalization in a neural network with binary synapses*
J. Phys. A: Math. Gen. **41** *(7097).*

GYÖRGYI, G. & TISHBY, N.

1990 *Neural Networks and Spin Glasses*
Editores: W.K.Theumann e R. Köberle. World Scientific, Singapore.

HANSEL, D., MATO, G. & MEUNIER, C.

1992 *Memorization without Generalization in a Multilayer Neural Network*
Europhys. Lett. **20** *(471).*

HEBB, D.O.

1949 *The Organization of Behavior*
Wiley, NY.

HERTZ, J.A., KROGH, A. & PALMER, R.G.

1991 *Introduction to the Theory of Neural Computation*
Redwood City, CA, Addinson-Wesley.

HOPFIELD, J.J.

1992 *Neural networks and physical systems with emergent collective computational abili-*

ties
Proc. Natl. Acad. Sci. USA **79** *(2554).*

KABASHIMA, Y.

1994 *Perfect loss of generalization in K=2 parity machine*
J. Phys. A: Math. Gen. **27** *(1917).*

KIM, J.W. & SOMPOLINSKY, H.

1996 *On-line Gibbs Learning*
Phys. Rev. Lett. **76** *(3021).*

UNIVERSIDADE DE SÃO PAULO
ex. 1
SERVIÇO DE BIBLIOTECA E INFORMAÇÃO
2626
INSTITUTO DE FÍSICA

KINOUCHI, O. & CATICHA, N.

1992 *Optimal generalization in perceptrons*
J. Phys. A: Math. Gen. **25** *(6243).*

KINOUCHI, O. & CATICHA, N.

1993 *Lower bounds on generalization errors for drifting rules*
J. Phys. A: Math. Gen. **26** *(6161).*

KRAUTH, W. & MÉZARD, M.

1989 *Storage capacity of memory networks with binary couplings*
J. Phys. (France) **50** *(3057).*

KRAUTH, W. & OPPER, M.

1989 *Critical storage capacity of the $J = \pm 1$ neural network*
J. Phys. A: Math. Gen. **22** *(L519).*

LITTLE, W. A.

1974 *The Existence of Persistent States in the Brain*
Mathematical Biosciences **19** *(101).*

MCCULLOCH, W.S. & PITTS, W.

1943 *A logical calculus of the ideas immanent in nervous activity*
Bull. Math. Biophys. **5** *(115).*

MITCHISON, G.J. & DURBIN, R.M.

1989 *Bounds on the learning capacity of some multy-layer networks*
Biol. Cybern. **60** *(345).*

MÜLLER, B. & REINHARDT, J.

1990 *Neural Networks - An Introduction*
Berlin, Springer.

OPPER, M.

1994 *Learning and Generalization in a Two-Layer Neural Network: The Role of the Vapnik-Chervonenkis Dimension*
*Phys. Rev. Lett.***72** *(2113).*

OPPER, M. & HAUSSLER, D.

1991 *Generalization Performance of Bayes Optimal Classification Algorithm for Learning a Perceptron*
*Phys. Rev. Lett.***66** *(2677).*

OPPER, M. & KINZEL, D.

1996 *Models of Neural Networks III*
Editores: E. Domany, J.L. van Hemmen e K. Schulten
Springer, Berlin.

REIMERS, L. & ENGEL, A.

1996 *Weight space structure and generalization in the reversed-wedge perceptron*
J. Phys. A: Math. Gen. **29** *(3923).*

RIEGLER, P.

1997 *Dynamics of On-line Learning in Neural Networks*
Tese de Doutoramento, Institut für Theoretische Physik, Universität Würzburg.

SEUNG, H.S., SOMPOLINSKY, H. & TISHBY, N.

1992 *Statistical mechanics of learning from examples*
J. Phys. A: Math. Gen. **45** *(6065).*

SIMONETTI, R. & CATICHA, N.

1996 *On-line learning in parity machines*
J. Phys. A: Math. Gen. **29** *(4859).*

WATKIN, T.L.H., RAU, A. & BIEHL, M.

1993 *The statistical mechanics of a learning rule*
Rev. Mod. Phys. **65** *(499).*