Universidade de São Paulo
Instituto de Física

Análise de redes neurais de atratores interagentes por meio de um modelo com solução analítica

Pietro Zanin

Orientador: Prof. Dr. Nestor Felipe Caticha Alfonso

Dissertação de mestrado apresentada ao Instituto de Física da Universidade de São Paulo, como requisito parcial para a obtenção do título de Mestre em Ciências.

Banca Examinadora:
Prof. Dr. Nestor Felipe Caticha Alfonso - Orientador (Universidade de São Paulo)
Prof. Dr. Daniel Adrián Stariolo - UFF
Prof. Dr. Juan Pablo Neirotti - Aston University

São Paulo
2022

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

Zanin, Pietro

Análise de redes neurais de atratores interagentes por meio de um modelo com solução analítica.. São Paulo, 2022.

Dissertação (Mestrado) – Universidade de São Paulo. Instituto de Física. Depto. de Física Geral.

Orientador: Prof. Dr. Nestor Felipe Caticha Alfonso
Área de Concentração: **Mecânica Estatística de Sistemas Complexos**

Unitermos: 1. Mecânica estatística; 2. Método de réplicas; 3. Algoritmos de aprendizado; 4. Modelos de agentes; 5. Redes neurais de atratores.

USP/IF/SBI-038/2022

University of São Paulo
Institute of Physics

Analysis of interacting attractor neural networks with a model with analytic solution

Pietro Zanin

Advisor: Prof. Dr. Nestor Felipe Caticha Alfonso

Dissertation submitted to the Physics Institute of the University of São Paulo in partial fulfillment of the requirements for the degree of Master of Science.

Examining Committee:
Prof. Dr. Nestor Felipe Caticha Alfonso - Advisor (Universidade de São Paulo)
Prof. Dr. Daniel Adrián Stariolo - UFF
Prof. Dr. Juan Pablo Neirotti - Aston University

São Paulo
2022

Agradecimentos

Agradeço à FAPESP e à Cnpq por financiarem este projeto. O projeto da FAPESP foi o processo com o número 2021/07951-7. As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do autor e não necessariamente refletem a visão da FAPESP.

Agradeço a Prof. Dr. Nestor Caticha pela orientação, paciência e confiança. Também agradeço a todo o grupo de estudos do professor pelas sugestões e por terem visto as versões iniciais do projeto, em particular agradeço ao Felippe Alves Pereira pela ajuda com partes do código.

Agradeço à toda minha família e amigos pelo suporte e motivação, principalmente nas fases mais difíceis da pandemia.

Resumo

Neste trabalho construímos e analisamos um modelo com o objetivo de ampliar ideias do algoritmo de *unlearning* para tentar entender redes neurais interagentes. Nos baseamos não só em vários trabalhos de redes neurais que giram em torno desta ideia, mas também em alguns artigos de ciências sociais relacionados.

O modelo é construído introduzindo uma modificação do Hamiltoniano de outros modelos, no qual introduzimos uma interação entre diferentes redes. Mudar a magnitude desta interação leva a resultados diferentes, sendo eles não triviais e ricos. Em geral, discutimos em quais regiões essa interação é benéfica e de que maneira ela pode ser benéfica.

Apesar do modelo ser complexo demais para ser comparado com dados reais, ele apresenta comportamentos qualitativos que mimetizam algumas dinâmicas sociais de maneira interessante. Além disso, o modelo também é de interesse para a área de redes neurais, pois mostra uma maneira em que redes podem ser melhoradas significativamente de maneira eficiente.

Palavras-chave:Mecânica Estatística, Método de réplicas, Algoritmos de Aprendizado,Modelos de agentes,Redes Neurais de Atratores

Abstract

In this work we built and analyzed a model with the goal of enlarging and generalizing ideas from the unlearning algorithm to try to understand interacting neural networks. Besides basing this work in previous neural networks articles related to that subject, we also used some social science articles to further develop the work.

The model is built by introducing a change in the Hamiltonian of other models, in which we introduce an interaction between different networks. Changing the magnitude of this interaction leads to different results, which are interesting and non-trivial. In general, we try to understand the regions where this interaction is beneficial and in which ways it can be beneficial.

Despite the model being too complex to be compared to real data, it shows qualitative behaviors that mimic some social dynamics in an curious way. Additionally, the model is also interesting to the area of neural networks, as it shows a way to train the networks in an efficient way.

Keywords:Statistical Mechanics, replica method, learning algorithms, agent models, attractor neural networks

Contents

Chapter 1

Introdução

O funcionamento do cérebro sempre intrigou pesquisadores, e uma das principais ferramentas clássicas para tentar se entender isso é o uso de redes neurais, que são conjuntos de neurônios conectados em uma maneira específica de tal forma que a rede consegue processar informação de diferentes maneiras.

A inspiração biológica de tais ferramentas é somente uma componente delas, e se tornou comum analisar redes neurais artificiais não mais com um olhar biológico, mas sim de um ponto de vista computacional e estatístico, principalmente motivado pelo enorme sucesso na aplicação de tais ferramentas em processamento de dados, visão computacional e resolução de problemas complexos. Pelo fato de ser significativamente mais simples aplicar redes neurais em um problema e modificá-las de maneira *ad hoc* do que compreendê-las extensivamente, o entendimento da eficácia e funcionamento interno delas ficou atrasado em relação às diversas aplicações na vida real. Notamos porém, que o ainda insuficiente conhecimento teórico já melhora extensivamente o funcionamento das redes neurais em aplicações, o que justifica o forte interesse em tal pesquisa.

Boa parte do esforço e da complexidade nestes sistemas está em entender de que maneira é possível treinar e melhorar as redes de maneira eficiente, a fim de que possam analisar grandes números de dados em um tempo razoável. As técnicas usadas são variadas e ricas, em particular, muitas delas utilizam de novo conceitos biológicos como base.

A particular maneira que a informação é processada depende do tipo da rede, por exemplo, em perceptrons com nenhuma camada intermediária temos que o *input* fornecido a cada neurônio passa por um determinado peso, o que, juntos com os outros inputs, resulta em um único *output*.

Um dos tipos mais relevantes e mais antigos de redes neurais consiste em redes neurais de atratores, também chamados de memórias associativas, onde se espera que a rede neural recupere determinados padrões em base ao estado inicial dela, em particular, se o estado inicial dela é próximo de um padrão, o esperado é que ela o recupere. O exemplo mais intuitivo é o de reconhecimento visual, supondo que temos 10 imagens diferentes de animais, se tivermos uma rede neural de atratores devidamente treinada o esperado é que, dado uma imagem de um gato levemente alterada, a rede consiga retornar a imagem restaurada de um gato.

O oscilador harmônico dessa categoria é a rede de Hopfield [1], que se utiliza do aprendizado Hebbiano, inspirado no famoso psicólogo canadense, para construir sua rede. Apesar de por si só não ser uma rede muito útil na vida real e apresentar algumas inconsistências do ponto de vista biológico, ela serve como um modelo mais simples para entendermos comportamento gerais. Notamos também que redes mais eficazes, como máquinas de Boltzmann restritas

[2], são inspiradas fortemente nelas e que o aprendizado Hebbiano se revelou uma ferramenta extremamente potente e ampla.

Uma aplicação interessante de redes gerais é em entender sistemas sociais, já que estes são sistemas coletivos de alta complexidade que em certas situações exibem características de redes. Os trabalhos estudando isso com redes complexas são antigos e mais conhecidos [3],[4]; mas com redes neurais em particular são mais recentes e menos entendidas [5]. No caso, são de interesse seja *toy models* para nos indicar comportamentos qualitativos, seja modelos aplicados e realistas para tentar fazer previsões e entender estes fenômenos de maneira mais quantitativa.

Boa parte das limitações da rede de atratores de servir como uma memória associativa são devidas ao fato da dinâmica ficar presa em mínimos espúrios, os quais surgem da interação entre os padrões de memória a ser guardados na rede. Algoritmos desenhados para eliminar esses mínimos indesejados recebem o nome de *unlearning*. O objetivo inicial do nosso trabalho foi buscar maneiras de ampliar algoritmos de *unlearning* hebbiano para entender possíveis interações entre redes neurais, em particular focamos na análise de redes neurais de atratores. No caso do nosso mecanismo os mínimos não são eliminados, mas aprendidos; além disso, não são escolhidos aleatoriamente, mas determinados por um agente modelado por uma rede similar. O interesse em entender estas interações vem seja de entender mecanismos alternativos de se conseguir compreender melhor redes neurais em geral, como o *transfer learning*; seja buscar maneiras de se representar comportamentos sociais específicos com redes neurais, como o comportamento de um professor ensinando o aluno.

A literatura sobre *unlearning* hebbiano é rica e parte da criação e discussão do algoritmo nos artigos [6] e [7], as mencionamos e explicamos detalhadamente no capítulo 2, em particular nos baseamos fortemente no trabalho recente [8] para desenvolver nosso projeto e orientar a nossa busca.

O nosso trabalho focou quase totalmente na análise aprofundada de um sistema que apresenta duas redes neurais do tipo de Hopfield interagindo com uma interação quártica, o mesmo foi resolvido por meio do cálculo da sua energia livre por meio de métodos usuais de réplicas e ponto de sela, e conseguimos expressar o comportamento do sistema por meio de diversos diagramas de fase. Um trabalho secundário foi estabelecer um tipo de algoritmo que denominamos *learning* hebbiano coletivo, que é basicamente o algoritmo de *unlearning* hebbiano, porém as mudanças na matriz sináptica vem dos mínimos alcançados pela outra rede, e as mudanças tem o sinal trocado. A explicação dos dois trabalhos se dá no capítulo 3, sendo que as partes mais técnicas se encontram nos apêndices. As duas partes, e principalmente a primeira, apresentaram resultados que consideramos significativamente interessantes, uma discussão e interpretação completa junto deles está exposta nos capítulos 4 e 5. Eles claramente se inserem na categoria de *toy models*, e não é possível tirar conclusões quantitativas, fazer previsões ou interpretar dados reais; porém, como argumentado anteriormente, acreditamos que são modelos com resultados ricos e úteis por meio das interpretações dos seus resultados qualitativos.

Notamos que o nosso trabalho não é exaustivo, existem mais direções a serem buscadas a partir dele. Consideramos interessante observar como a estrutura de vidro de spin se modifica ao introduzirmos o Hamiltoniano de interação, e se é possível relacionar isso a melhoras no processamento de informação das redes ou à sincronização entre as redes. Além disso, generalizações de todo tipo são bem-vindas, seja introduzir interações deste tipo em outros modelos, seja tentar introduzir outras interações em redes desse tipo, notamos em particular que aplicações a modelos mais simples e realistas seriam intrigantes, e seria benéfico buscar maneiras de se comparar com dados da vida real.

Para entender o que está acontecendo no sistema, observamos como os 3 parâmetros de

ordem (m_1, m_2, h) se modificam ao longo do espaço de 5 dimensões dos parâmetros fixos $(\beta, \alpha, t_1, t_2, \epsilon)$. A grande dimensionalidade desse espaço nos levou a fazer 3 tipos de diagramas bi-dimensionais distintos, nos quais podemos verificar as diferentes fases que ocorrem; algumas fases que aparecem são conhecidas de outros modelos, porém também temos fases novas e particulares. β é o inverso da temperatura, $\alpha \equiv \frac{p}{N}$, onde p é o número de padrões e N o número de neurônios, é a medida da carga de padrões no sistema. Os valores t_1 e t_2 se referem às taxas de treinamento dos agentes 1 e 2 respectivamente, elas caracterizam a capacidade dos agentes previamente à interação, esse parâmetro é introduzido em [8]; por fim, o parâmetro ϵ , introduzido por nós, mostra a magnitude da interação entre os agentes. Em relação aos parâmetros de ordem, m_1 e m_2 são os valores esperados dos *overlaps* com o padrão condensado do primeiro e segundo agente respectivamente, e h, o valor esperado do *overlap* entre os agentes, mede a semelhança entre eles.

Dentre as várias fases, enfatizamos três que são interessantes e novas, a de professor-aluno na qual o agente menos capacitado consegue processar mais informação por meio da interação enquanto o agente mais capacitado aprende pouco ou nada, modulando de certa forma uma interação típica professor-aluno. Na fase de mutualismo, as duas redes se ajudam por meio da interação, conseguindo armazenar informação de maneira eficiente em situações onde antes seria impossível. Na fase de ilusão reforçada, os agentes reforçam seus próprios erros, isto é, acreditam estar certos e concordam entre si, porém não conseguem armazenar dados de maneira eficiente. As diferentes fases mostram o caráter rico e não trivial da interação.

Chapter 2

Teoria e Bases necessárias

2.1 Introdução

O nosso trabalho original não foi motivado somente por nossas ideias abstratas, mas ele segue uma linha de pensamentos, modelos e técnicas relacionadas a esse assunto que começou 40 anos atrás, com [6]. Nesse capítulo queremos explicar de maneira clara e relativamente concisa as ideias que permitiram fazermos algo de novo, começando por explicar o algoritmo de *unlearning* hebbiano, como analisá-lo e o que ele consegue modificar em uma rede de Hopfield tradicional.

A grande mudança na capacidade nos faz questionar até quantos padrões é possível uma rede neural com certas restrições armazenar. Essa pergunta foi respondida de maneira elegante e abrangente nos artigos [9] e [10], onde usando técnicas de réplicas se pode entender os limites das redes neurais.

As redes com as matrizes sinápticas introduzidas em [11] e [12] alcançam esse limite, porém de forma diferente e de maneiras qualitativamente distintas. Uma recente melhora [8] dessas redes consegue não só ser mais robusta, mas também controlar essa robustez de maneira mais clara. Em particular, ela se utiliza de ideias de *unlearning* para justificar biologicamente sua construção, e como mostraremos no capítulo 3 é possível generalizá-la ainda mais.

2.2 Sobre sistemas desordenados

Nesta dissertação vamos considerar um sistema cuja desordem é fixa, isto é, ao longo do tempo somente os spins mudam, os quais denotamos por σ. Em geral, a desordem age sobre a matriz sináptica da rede, a qual denotamos J, e a escolha particular da desordem é escolhida seguindo uma determinada probabilidade $P(J)$. Além disso, a desordem é forte o suficiente para mudar significativamente o comportamento do sistema em várias temperaturas, e não somente em algumas regiões específicas.

O fato da desordem ser fixa implica que o resultado particular do sistema depende da escolha particular da desordem, porém calcular o comportamento para uma escolha particular é um processo inútil, já que em geral temos um número da ordem de $\mathcal{O}(\exp(N))$ de possíveis desordens distintas. O procedimento ingênuo seria então calcular a média dos parâmetros de ordem do sistema, porém isto não é o que nos interessa, nos interessa os valores típicos dos parâmetros, já que, ao considerarmos um sistema com um número suficientemente grande de spins, o que vamos observar é o valor típico, não a média.

Para calcularmos isso, é crucial mostrar que a energia livre de interesse não é $f = -\frac{1}{\beta}\log(\langle Z\rangle)$, mas sim $f = -\frac{1}{\beta}\langle\log(Z)\rangle$, onde a média se refere à média sobre a desordem. Provamos isso de duas formas distintas.

Consideremos uma escolha particular da desordem, a energia livre associada a ela é:

$$f(J) = -\frac{1}{\beta}\log(Z_J) \tag{2.1}$$

De forma mais geral, vamos considerar o caso de *partial annealing*, onde as interações também podem mudar ao longo do tempo, sendo que a temperatura em que elas mudam é T' e o seu Hamiltoniano é f_J. A função de partição se torna então

$$Z = \int dJP(J)\exp(\frac{\beta'}{\beta}\log[Z_J)] = \langle Z_J^n\rangle, \tag{2.2}$$

onde denotamos $\frac{\beta'}{\beta} = n$.

A energia livre é então

$$f = -T'\log(\langle Z^n\rangle) \tag{2.3}$$

Para o nosso caso, $n \to 0$, já que com temperatura T' infinita a dinâmica dos spins não influencia a dinâmica das interações, levando efetivamente a uma desordem fixa já que elas não mudam. Assim,

$$f = -T'\log(\langle Z^n\rangle) = -\frac{T}{n}\log(\langle Z^n\rangle) = -\frac{T}{n}\log(\langle 1 + n\log(Z)\rangle) = -T\langle\log(Z)\rangle \tag{2.4}$$

Esse raciocínio vem majoritariamente de [13].

A outra maneira de ver isso é usando o método de máxima entropia.

Consideremos a distribuição $P(s, J)$, as condições impostas nela são a de um ensemble canônico mais o fato de que a probabilidade da desordem é conhecida:

$$\int dsP(s \mid J) = 1; \quad \int dsP(s \mid J)H(s \mid J) = E; \quad \int dsP(s, J) = P(J) \tag{2.5}$$

Além disso, lembremos do conhecido fato de que $f = E - TS$, assim temos que

$$\begin{aligned} S(P(s,J)) &= -\int\int dJdsP(s,J)\log(P(s,J)) \\ &= \int\int dJds\frac{\exp[-\beta H(s \mid J)]}{Z(J)}P_0(J)\{\log[\frac{\exp(-\beta H(s \mid J))}{Z(J)}) + \log(P_0(J)]\} \\ &= -\int\int dJds\frac{\exp[-\beta H(s \mid J)]}{Z(J)}P_0(J)\{-\beta H(s \mid J) - \log[Z(J)] + \log[P_0(J)]\} \\ &= \beta E + \langle\log[Z(J)]\rangle + S[P_0(J)] \end{aligned} \tag{2.6}$$

Ou seja, $f = -\frac{1}{\beta}\langle\log(Z(J))\rangle - TS(P_0(J))$. Note que isto prova o que queríamos demonstrar, já que o último termo é uma constante que depende do sistema em particular que estamos lidando.

2.3 A rede de Hopfield

Falamos de maneira sucinta do modelo de Hopfield e de como pode ser analisado. O sistema consiste em N neurônios totalmente conectados que podem assumir os valores discretos -1 e 1, p padrões fixos descorrelacionados e sem viés, sendo que um padrão é uma configuração particular dos N neurônios. Em geral, não nos focamos na forma particular em que os padrões são obtidos, e os tomamos com uma distribuição de Bernoulli por simplicidade. A interação entre dois neurônios i e j depende dos valores particulares dos padrões, e o Hamiltoniano é da forma de Ising:

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \quad J_{ii} = 0, \quad \mathcal{H} = -\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij}\sigma_i\sigma_j; \tag{2.7}$$

onde J_{ij} é a interação entre os neurônios i e j, o índice superior μ se refere ao índice dos padrões e σ_i representa o valor do neurônio i.

O sentido biológico desse modelo está no fato de que essa forma particular de interação premia a dinâmica a buscar estados que se pareçam com os padrões, assim, representa uma forma de memória associativa. Notamos que existe mais de uma característica que torna essa analogia mais fraca e vaga, como a simetria entre relações de neurônios, neurônios serem totalmente conectados e os padrões serem descorrelacionados, porém mesmo assim o modelo continua sendo uma analogia poderosa e interessante.

Focamos no comportamento do limite termodinâmico, i.e. quando $N \to \infty$. Quando o número de padrões é de ordem $\mathcal{O}(1)$, a energia livre assume uma forma familiar ao do modelo de Ising totalmente conectado:

$$f = \frac{1}{2}\sum_{\mu=1}(m_\mu)^2 - \langle\langle \log\{2\cosh[\beta(\sum_{\mu}\xi^\mu m^\mu)]\}\rangle\rangle, \tag{2.8}$$

onde a média dupla se refere à média sobre a desordem e os possíveis estados e

$$m_\mu = \langle\langle \frac{1}{N}\sum_i \xi_i^\mu \sigma_i\rangle\rangle. \tag{2.9}$$

São possíveis estados onde $m^\mu > 0$ para mais de um valor de μ, mas em geral estamos interessados em estados onde somente um deles é não nulo, já que são mais claros de se analisar. Note que nesse caso a equação de consistência se torna equivalente ao modelo de Ising totalmente conectado. A noção intuitiva de porque a adição de um número de padrões finitos não afetar tanto o comportamento é que o fato deles serem descorrelacionados torna a desordem introduzida pelas correlações não significativas.

O panorama muda quando consideramos números de padrões da ordem de $\mathcal{O}(N)$. Agora, a desordem é significativa, e o comportamento do sistema é alterado significativamente. A energia livre depende do método de réplicas para ser calculada, e ela é

$$\begin{aligned} f = &\frac{1}{2}\sum_{\nu=1}^{p}(m^\nu)^2 + \frac{\alpha}{2\beta}[\log(1-\beta+\beta q) - \frac{\beta q}{1-\beta+\beta q}] \\ &+\frac{\alpha\beta r(1-q)}{2} - \frac{1}{\beta}\int Dz\langle\langle \log\{2\cosh[\beta(\sqrt{\alpha r}z + \sum_{\mu=1}^{p} m^\mu\xi^\mu)]\}\rangle\rangle, \end{aligned} \tag{2.10}$$

onde $Dz \equiv \exp(-\frac{z^2}{2})dz$ e temos um novo parâmetro de ordem

$$q = \langle \frac{1}{N} \sum_{i=1}^{N} \langle S_i \rangle^2 \rangle. \tag{2.11}$$

Ele basicamente nos diz se em uma região os estados típicos tem alguma correlação entre si, independentemente da magnetização. Na fase paramagnética sabemos que isso não acontece, assim $q = 0$. Para que o resultado $m^\mu > 0$ para algum valor de μ faça sentido, é necessário que $q > 0$, assim na fase ferromagnética temos que $q > 0$. Porém, é possível ter $q > 0$ e $m = 0$, esta é a região vidro de spin caracterizada de ter vários ($\mathcal{O}(\exp(N))$) mínimos espúrios, onde $m = 0$.

O rico diagrama de fase pode ser visto na figura a seguir:

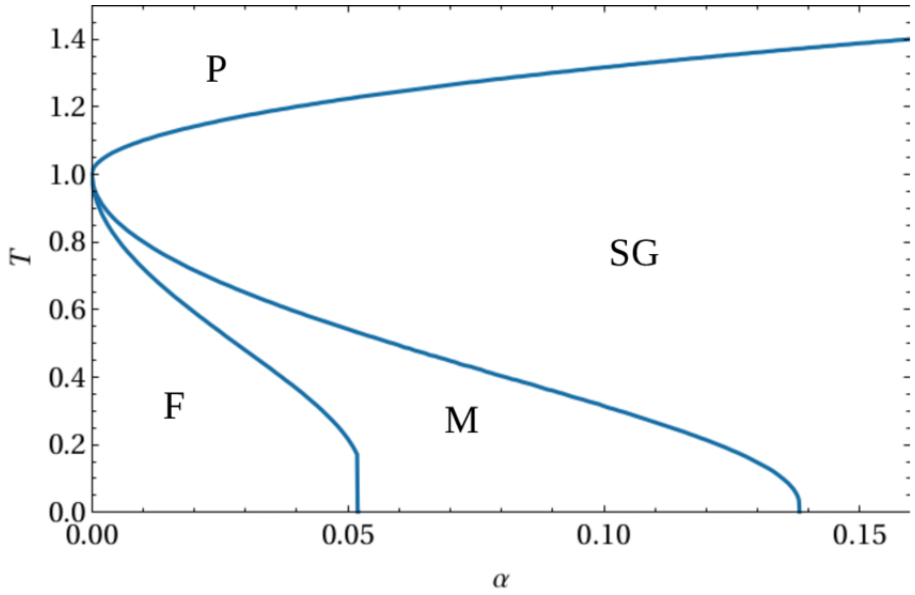


Figure 2.1: Diagrama de fase do modelo de Hopfield. SG se refere à fase de vidro de spin; F e M formam a fase ferromagnética, na parte F temos estabilidade de estados com mais de um padrão condensado enquanto que na parte M não; P se refere à fase paramagnética. Retirado de [8].

2.4 *Unlearning* hebbiano

A rede de Hopfield tradicional consegue armazenar um número de padrões da ordem de $\mathcal{O}(N)$ [1], o que, apesar de ser um resultado muito significativo e importante, leva a pensar se é possível armazenar mais padrões com esse tipo de rede, já que a capacidade máxima é relativamente baixa $\alpha_{max} = \frac{p_{max}}{N} \approx 0.138$ [14]. Notamos que quando falamos de armazenamento de padrões, estamos considerando que eles podem ser recuperados com alguns erros, e não que devem ser recuperados de maneira perfeita. Isso se deve ao fato de que, seja em um contexto biológico, seja em um contexto de máquinas, o aparecimento de poucos erros é pouco significativo e não traz danos para a análise de um problema. Além disso, resultados na literatura sobre recuperação perfeita de padrões, nos indicam que o número de padrões armazenados de maneira

perfeita é no máximo da ordem $\mathcal{O}(\log(N))$[15], assim temos um problema menos interessante e rico.

Em 1983, Crick e Mitchinson[7], ao tentar explicar a presença do sono REM em humanos, criaram a hipótese de que esta fase do sono serve para expelir memórias pouco importantes do cérebro para melhorar o processamento de informação dele. Curiosamente, a poucos meses de distância, o próprio Hopfield criou um algoritmo que pode ser aplicado em simulações de tamanho finito de redes neurais de Hopfield que simula este comportamento [6], posteriormente foi chamado de *unlearning* hebbiano e foi examinado em maior profundidade em [16] e [17], os quais seguiremos majoritariamente. Apesar de não ser o nosso foco, vale a pena citar um artigo posterior de Crick e Mitchinson [18] em que eles expandem os raciocínios biológicos que explicam tal comportamento

Expliquemos agora como o algoritmo funciona. Como de costume, temos N neurônios totalmente conectados, p padrões fixos e um processo dinâmico de Glauber. Em particular, notamos que este processo de *unlearning* deve ser feito com temperatura 0, porém a rede resultante pode ser usada em temperaturas não nulas.

O processo pode ser descrito como a sequência dos seguintes passos:

1. Comece em um estado aleatório usando uma distribuição uniforme.
2. Deixe o sistema relaxar até um ponto fixo. Esse passo justifica $T = 0$, já que não existem pontos fixos com temperatura não nula.
3. Faça a transição $J_{ij} \to J_{ij} - S_i S_j \frac{\epsilon}{N}, \quad \forall i, j$, onde $\epsilon > 0$ geralmente é da ordem de 0.01.

Esse processo é repetido um número d de vezes, biologicamente este número representa a quantidade de tempo empregado no sono REM. Ele aumenta significativamente a capacidade de uma rede neural se é executado no mínimo $d = D_{opt}$ vezes, podendo levar até $\alpha_c = 0.55$. Porém, existe também um número máximo de sonhos $d = D_c$ que pode ser feito sem diminuir a magnetização, assim temos um intervalo onde o algoritmo é realmente útil, notamos que essa propriedade peculiar não tem uma explicação biológica clara dada nos artigos citados anteriormente.

Os resultados a seguir vem de simulações, cujos resultados e detalhes podem ser encontrados em [17]. Apesar de serem feitas há um certo tempo, elas continuam válidas e nos indicam as principais propriedades do algoritmo. Temos que,

$$D_{opt}(\alpha) = \frac{p}{2\epsilon}(1+\alpha), \;\; D_c(\alpha) = \frac{p}{2\epsilon}(3-\alpha), \;\; \alpha_c \approx 0.59; \tag{2.12}$$

sendo que $D_{opt}(\alpha)$ e $D_c(\alpha)$ são respectivamente os números mínimos e máximos de sonhos que tornam possível ter magnetização com uma determinada capacidade α e α_c é a maior capacidade onde encontramos magnetização não-nula.

Podemos estimar o tempo de convergência, que é o tempo que a rede demora para encontrar um estado estável em que não é possível ocorrer qualquer mudança:

$$t_{conv} \sim \sqrt{N}. \tag{2.13}$$

É útil comparar o tempo necessário para construir uma rede de Hopfield tradicional e fazê-la encontrar um estado estável com o tempo necessário para fazer isso adicionando o algoritmo de *unlearning*.

Nos dois casos a construção da matriz sináptica se dá pela mesma maneira, temos que fazer $\frac{N^2}{2}$ somas distintas, tendo as somas p elementos e considerando o caso relevante que $p = \mathcal{O}(N)$, temos que esse passo demora $\mathcal{O}(N^3)$.

Assim, na rede de Hopfield tradicional, depois da construção só é necessário rodar a dinâmica e convergir para um estado estável, o que segundo a equação (2.13) adiciona um tempo da ordem de $\mathcal{O}(N^{\frac{1}{2}})$, assim o tempo total é da ordem de $\mathcal{O}(N^{\frac{7}{2}})$.

Com o algoritmo de *unlearning*, antes de convergir para o estado final temos que rodar o algoritmo $\mathcal{O}(N)$ vezes como indica (2.1), e notamos que temos que fazer $\frac{N^2}{2}$ modificações cada vez que rodamos o algoritmo, o que adiciona um tempo de $\mathcal{O}(N^2)$. Ou seja, acabamos adicionando um tempo da ordem de $\mathcal{O}(N^4)$, assim, no total, temos que o tempo é de $\mathcal{O}(N^{\frac{13}{2}})$, um valor significativamente mais alto do que o tempo necessário para uma rede de Hopfield tradicional.

2.5 Teoria do *unlearning*

O algoritmo não foi construído com a esperança de ter uma explicação analítica clara, o ponto dele é ser um artifício com motivações biológicas para ser usado em simulações de redes de Hopfield. Apesar disso, é possível justificar parcialmente a eficácia do algoritmo, e em particular podemos entendê-lo aproximadamente como uma realização de uma equação diferencial, onde a matriz sináptica varia em função do número de sonhos. Enfatizamos que a explicação a seguir falha em certos regimes e não captura toda a complexidade do algoritmo.

Denotando por t o número de sonhos, podemos escrever que

$$J_{ij}(t+1) = J_{ij}(t) - S_i S_j \frac{\epsilon}{N} \quad \forall i, j. \tag{2.14}$$

Onde temos que $J(0) = J_{hebb}$. Podemos pensar nessa variável discreta t como sendo o tempo.

Considerando a definição de um ponto fixo e um processo de Glauber com $T = 0$:

$$\begin{aligned} J_{ij}(t+1) &= J_{ij}(t) - \frac{\epsilon}{N} \operatorname{sign}(\sum_k^N J_{ik}(t) S_k) \operatorname{sign}(\sum_m^N J_{im}(t) S_m) \qquad (2.15) \\ &= J_{ij}(t) - \frac{\epsilon}{N} \operatorname{sign}(\vec{J_i}(t) \cdot \vec{S}) \operatorname{sign}(\vec{J_j}(t) \cdot \vec{S}), \qquad (2.16) \end{aligned}$$

onde $\vec{J_i}$ é a linha i da matriz sináptica e $\vec{S}$ é o vetor do estado atual de dimensão N.

A partir daqui, usamos uma interpretação geométrica da equação para entendê-la.

O valor de $\operatorname{sign}(\vec{J_i}(t) \cdot \vec{S}) \operatorname{sign}(\vec{J_j}(t) \cdot \vec{S})$ depende somente das relações entre esses 3 vetores, e podemos considerar S como um vetor aleatório com distribuição uniforme, assim somente a relação entre $\vec{J_i}(t)$ e $\vec{J_j}(t)$ determina o sinal desse termo.

Imaginemos 2 hiperplanos, sendo um perpendicular a $\vec{J_i}(t)$ e o outro perpendicular a $\vec{J_j}(t)$. Denominemos ϕ_{ij} o ângulo entre $\vec{J_i}(t)$ e $\vec{J_j}(t)$. $\vec{J_i}(t)$ e $\vec{J_j}(t)$ podem estar em 2 seções, uma é a seção com ângulo ϕ_{ij} e a outra é a seção com ângulo $\pi - \phi_{ij}$. Se $\vec{S}$ está na primeira seção, então o valor do produto dos sinais vai ser negativo, se está na segunda vai ser positivo.

Com esse raciocínio podemos afirmar que, em média

$$\langle J_{ij}(t+1)\rangle = \langle J_{ij}(t)\rangle - \frac{\epsilon}{N}[2\frac{(\pi-\phi_{ij})}{2\pi} - \frac{\phi_{ij}}{\pi}], \tag{2.17}$$

onde a média é sobre várias transições em um mesmo tempo t. Então, temos que

$$\langle J_{ij}(t+1)\rangle - \langle J_{ij}(t)\rangle \equiv \langle \Delta J_{ij}\rangle(t) = \frac{2\epsilon}{N\pi}\arccos(\frac{\vec{J_i}\vec{J_j}}{\mid \vec{J_i} \mid\mid \vec{J_j} \mid}), \tag{2.18}$$

onde a norma é a euclidiana. Se supormos que o argumento do $\arccos$ é pequeno e que a dependência da norma de J em t não é relevante, então temos que

$$\langle \Delta J_{ij}\rangle(t) \equiv \langle \Delta J_{ij}\rangle = \frac{2\epsilon}{\pi N}\frac{J_{ij}^2}{\mid \hat{J} \mid^2}, \tag{2.19}$$

onde $\mid \hat{J} \mid^2 = \sum_{ij} J_{ij}^2$, e $\hat{J}$ representa a matriz sináptica.
Já que consideramos i e j arbitrários, temos que

$$\langle \Delta \hat{J}\rangle = \frac{2\epsilon}{\pi N}\frac{\hat{J}^2}{\mid J \mid^2}, \tag{2.20}$$

Se esta última fórmula fosse exata e se considerarmos ΔJ baixo por causa do limite termodinâmico, teríamos que

$$\dot{\hat{J}} = \frac{2\epsilon}{\pi}\frac{\hat{J}^2}{\mid \hat{J} \mid^2}, \quad \hat{J}(0) = \hat{J}_{hebb}; \tag{2.21}$$

o que constitui uma equação diferencial que explica como a matriz sináptica varia conforme a rede neural sonha. A solução para ela é a matriz da pseudo-inversa, que será explicada com mais detalhes posteriormente, o ponto importante é que ela permite capacidade críticas de até $\alpha = 1$, o que explicaria a eficácia do algoritmo de *unlearning*.

Porém, a existência de D_{opt} nos mostra que não existe convergência de verdade, e o argumento falha pois a suposição de que a dependência da norma de J em t não é relevante não é verdade. Apesar disso, o argumento indica que a melhora ocorre em uma certa região pois a matriz sináptica se aproxima de alguma forma à matriz da pseudo-inversa[12], que vai ser explicada em mais detalhes mais para frente nesse capítulo.

2.6 Limites na capacidade

Anteriormente mostramos que redes de Hopfield podem ter um aumento considerável na capacidade crítica se fizermos mudanças apropriadas na matriz sináptica. Baseado nesse resultados, é relevante se perguntar até quanto é possível a princípio uma rede perfeita armazenar, dadas algumas restrições gerais. Para responder isso, é necessário caracterizar o espaço de fase de

uma matriz sináptica de um rede neural geral, o que nos vai permitir descobrir qual matriz nos leva à capacidade máxima e qual é esta capacidade máxima.

Temos N neurônios totalmente conectados e p padrões fixos. Para assegurar um espaço de fase finito devemos usar alguma restrição, que no caso vai ser a esférica

$$\sum_{j\neq i} J_{ij}^2 = N. \tag{2.22}$$

Isto é, as possíveis eficácias estão confinados em uma superfície de uma N-hiperesfera com raio $\sqrt{N}$. Notamos que nos cálculos e simulações em geral esta restrição particular não vai existir, porém os resultados que apresentaremos a seguir continuam válidos já que essa condição pode ser mudada mudando o raio da esfera.

Estamos interessados em caracterizar a capacidade de recuperar padrões de um região em base nos pontos fixos aqui, assim a temperatura é 0. O campo h_i é a contribuição total que um spin i recebe, isto é,

$$h_i = \frac{1}{\sqrt{N}} \sum_{j\neq i} J_{ij} S_j \tag{2.23}$$

A dinâmica pode ser ser entendida a partir desta quantidade, já que ela tiver o mesmo sinal do que S_i, não teremos mudanças a temperaturas nulas.

Adicionalmente, é possível ter alguns limiares T_i e κ que limitam os pontos fixos, a saber

$$S_i^\mu (h_i - T_i) > \kappa. \tag{2.24}$$

Usualmente colocamos $T_i = \kappa = 0$ e recuperamos a expressão mais familiar

$$S_i^\mu h_i > 0. \tag{2.25}$$

Se tivermos *retrieval*, isto é, todos os padrões são possíveis de serem recuperados, temos que

$$\xi_i^\mu (h_i - T_i) > \kappa. \tag{2.26}$$

A quantidade de interesse é o volume de sinapses que satisfazem a restrição esférica e a equação (2.13), já que com ele conseguimos caracterizar o espaço de fase. O sistema tem desordem fixa como de costume, assim a quantidade de interesse é $\langle \log(V) \rangle$, sendo a média sobre a distribuição fixa de ξ_i^μ. Note que de certa maneira este é um problema inverso: estamos procurando as matrizes sinápticas J dadas certas restrições sobre S, o que é inverso do que é geralmente feito, que é procurar o comportamento usual de S dadas restrições sobre J.

Esse tipo de problema foi resolvido de maneira elegante e completa por Elizabeth Gardner e colaboradores em [9] e [10], os resultados a seguir se referem principalmente a estes dois artigos. Após cálculos cuidadosos, é possível provar que

$$\lim_{N\to\infty} \frac{1}{N} \langle \log(V) \rangle = \alpha \int Dt \log\left(\int_{\frac{\sqrt{q}t+\kappa}{\sqrt{1-q}}}^{\infty} Dz \right) + \frac{\log(1-q)}{2} + \frac{q}{2(1-q)}, \tag{2.27}$$

onde $q = \langle\langle \frac{1}{N} \sum_{j \neq i} J_{ij}^2 \rangle\rangle$ é o análogo ao parâmetro de Edward-Anderson q_{ab} em sistemas usuais de vidros de spin, a e b são os índices das réplicas, este parâmetro enfatiza de novo o fato de estarmos lidando com um problema inverso.

O valor máximo de α é o valor que reduz ao máximo o volume disponível sem torná-lo 0:

$$\alpha_c = \frac{1}{\int_{-\kappa}^{\infty} (t+\kappa)^2 Dt}. \tag{2.28}$$

Se tomarmos $\kappa = 0$, temos que $\alpha_c = 2$. Para matrizes sinápticas simétricas este resultado se reduz a $\alpha_c = 1$, já que só temos $\frac{N(N-1)}{2} \approx \frac{N^2}{2}$ sinapses independentes.

2.7 Obtendo a capacidade máxima: as matrizes sinápticas de Dotsenko e da pseudo-inversa

Dado o resultado provado anteriormente, é de interesse buscar redes neurais que satisfazem a capacidade máxima e entender em quais situações ela é possível de ser alcançada. As duas matrizes que mencionaremos não só alcançam a capacidade máxima em certas situações, mas também são úteis para definir extensões do *unlearning* hebbiano que falaremos posteriormente.

A análise do modelo da pseudo-inversa foi feita em [12], e temos novamente o caso usual de N neurônios totalmente conectados e p padrões descorrelacionados. A diferença é que sua matriz sináptica é

$$J_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu (C^{-1})_{\mu\nu}, \tag{2.29}$$

onde $C_{\mu\nu} = \frac{1}{N} \sum_i \xi_i^\mu \xi_j^\nu$ é a matriz de correlação entre padrões. Notamos que, diferentemente da matriz hebbiana, essa matriz é não-local por meio da introdução da matriz de correlação, isto é, as sinapses entre dois neurônios são afetadas por padrões longe deles e não somente pelos padrões que estão neles, como ocorre na matriz sináptica hebbiana tradicional. Também notamos que o custo computacional de construir essa matriz é $\mathcal{O}(N^4)$, bem maior do que o custo da matriz hebbiana que tem ordem $\mathcal{O}(N^2)$.

O diagrama de fase relevante é

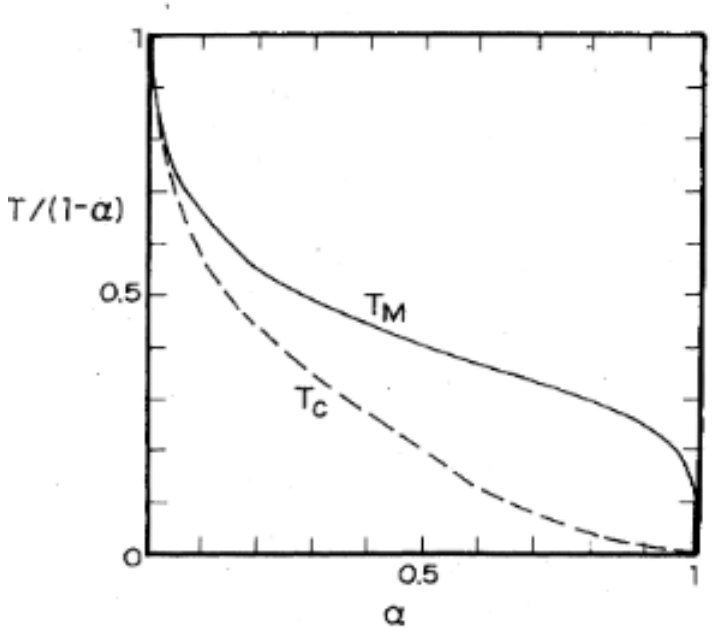


Figure 2.2: Diagrama de fase da matriz pseudo-inversa, obtido de [12]. Aqui, abaixo de T_M temos que os estados onde existe *retrieval* são possíveis, enquanto que abaixo de T_c eles se tornam mínimos globais.

O modelo de Dotsenko, desenvolvido em [11], é similar ao da pseudo-inversa, porém notamos que pode ser entendido como uma via de meio entre a matriz da pseudo inversa e a matriz hebbiana:

$$J_{ij}(t) = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu \xi_j^\nu ((\mathbb{1} + tC)^{-1})_{\mu\nu}, \tag{2.30}$$

onde t é um parâmetro que regula a capacidade do modelo. Para $t \to 0$, recuperamos o modelo de Hopfield, para $t \to \infty$, recuperamos $\alpha_c = 1$. Em particular, pelo fato do espaço de fase ser conectado, no limite de t grande essa matriz é igual à matriz pseudo-inversa vezes uma constante.

Uma característica importante dessa matriz é que o aumento de t diminui a robustez à temperatura, para $t \to \infty$, só temos *retrieval* com $T = 0$.

O diagrama de fase para vários valores de t é

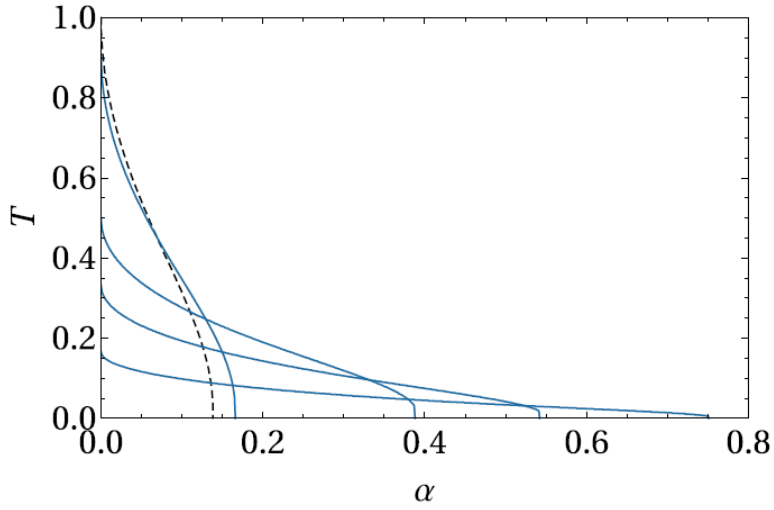


Figure 2.3: Diagrama de fase do modelo de Dotsenko, retirado de [11]. As curvas tem diferentes valores de t: 0,0.1,0.2,0.5 e 1; sendo que a preta representa a $t = 0$ e as de maior número são as que progressivamente tem menor temperatura crítica com $\alpha = 0$. As curvas separam a região ferromagnética da de vidro de spin: à esquerda e embaixo delas, temos ferromagnetismo, no resto do espaço temos vidro de spin. Note como o aumento de t leva simultaneamente ao aumento de valores em α que é possível ter $m \neq 0$ e à diminuição de valores de T tal que $m \neq 0$.

2.8 Algoritmo de *Reinforcement and Removal*

Recentemente em [8], um novo Hamiltoniano[1] para uma rede neural foi definido, levando a algumas propriedades importantes e reunindo várias ideias de artigos mais antigos, em particular juntando a ideia de *unlearning* com uma matriz similar à de Dotsenko com ideias de reforço, que evitam o colapso do sistema ao aumento de temperaturas como o modelo de Dotsenko exibe. Como será mostrado posteriormente, essa matriz vai ser muito útil para o nosso modelo original.

O Hamiltoniano é

$$H = -\frac{1}{2N} \sum_{i,j,i\neq j}^{N} \sum_{\mu,\nu}^{p} \xi_i^\mu \xi_j^\nu (\frac{1+t}{\mathbb{1} + tC})_{\mu,\nu} \sigma_i \sigma_j. \tag{2.31}$$

No limite $t \to 0$ recuperamos a matriz Hebbiana, no limite $t \to \infty$ recuperamos a matriz da pseudo-inversa.

Notamos que a diferença entre esse modelo e o de Dotsenko é sutil, mudamos o numerador de 1 para $1 + t$. Essa mudança faz a parte de reforço do estados puros, mantendo a resistência a mudanças de temperatura, enquanto que o denominador faz a parte de remover mínimos espúrios e aumentar a capacidade máxima.

Um exercício útil é considerar o seguinte Hamiltoniano, que corresponde somente à parte de reforço:

[1]Não confundir o *reinforcement* no nome com *reinforcement learning*, no sentido do algoritmo os autores consideram *reinforcement* como algo que reforça os estados puros, aumentando o espaço de *retrieval*. Citamos a referência [19] para representar o significado usual de reinforcement learning.

$$H' = -\frac{1}{2}\sum_{\mu}\sum_{ij}\xi_i^{\mu}\xi_j^{\mu}(1+t)\sigma_i\sigma_j. \tag{2.32}$$

Comparando ele com o modelo de Hopfield tradicional, vemos que na prática temos somente um reescalonamento da temperatura por meio de $T' = \frac{T}{(1+t)}$ sem um aumento da capacidade.

Assim, os dois mecanismos em conjunto são necessários para termos uma rede neural significativamente mais eficiente em vários aspectos.

É possível calcular a energia livre do modelo com os métodos tradicionais de réplicas e método de ponto de sela. As contas são praticamente iguais às feitas por Dotsenko [11]. Para cada trinca de parâmetros intensivos (α, β, t), a extremização no espaço dos parâmetros de ordem (m, q, Q, Δ, r) leva a cinco equações que podem ser reduzidas a três, pois podemos escrever Δ e r como função das outras três variáveis.

$$\begin{aligned} f(\alpha,\beta,t,m,Q,q,\Delta,r) &= \frac{m^2}{2(1+t)}(1+\frac{t}{\Delta}) + \frac{(\Delta-1)(1+t)}{2t}Q + \frac{\alpha\beta}{2}r(Q-q) \\ &+\frac{\alpha}{2\beta}\{\log[1-\beta(1+t)(Q-q)] - \frac{q\beta(1+t)}{1-\beta(1+t)(Q-q)}\} \\ &+\frac{(1-\Delta)(1+t)}{2t\Delta} + \frac{\log(\Delta)}{2\beta} + \frac{\alpha r t}{2\Delta(1+t)} - \frac{1}{\beta}\int Dx\log\{\cosh[\frac{\beta}{\Delta}(m+\sqrt{\alpha r}x)]\}; \end{aligned} \tag{2.33}$$

$$\begin{aligned} m &= \frac{1+t}{\Delta+t}\int Dx\tanh[\frac{\beta}{\Delta}(m+\sqrt{\alpha r}x)]; \\ r &= \frac{q(1+t)^2}{(1-\beta(1+t)(Q-q))^2};\ \ \Delta = 1+\frac{\alpha t}{1-\beta(1+t)(Q-q)}; \\ q &= Q + \frac{t}{\beta(1+t)\Delta} - \frac{1}{\Delta^2}\int Dx\cosh^{-2}[\frac{\beta}{\Delta}(m+\sqrt{\alpha r}x)]; \\ Q &= \frac{1}{\Delta^2} - \frac{t}{(1+t)\Delta\beta} + \frac{\alpha r t^2}{(1+t)^2\Delta^2} - \frac{t^2m^2}{(1+t)^2\Delta^2} \\ &-\frac{2\beta\alpha r t}{\Delta^3(1+t)}\int Dx\cosh^{-2}[\frac{\beta}{\Delta}(m+\sqrt{\alpha r}x)]. \end{aligned} \tag{2.34}$$

onde $\Delta \equiv 1+\alpha\beta\frac{t}{1+t}(R-r)$.

O significado das variáveis é:

$$\begin{aligned} m &= \langle\langle\frac{1}{N}\sum_i \sigma_i\xi_i^1\rangle\rangle, \quad q_{ab} = \langle\langle\frac{1}{N}\sum_j(\sigma_j^a + i\sqrt{\frac{t}{\beta(1+t)}}\phi_j^a)(\sigma_j^b + i\sqrt{\frac{t}{\beta(1+t)}}\phi_j^b)\rangle\rangle \\ &= \frac{1}{N}\langle\langle\sum_j \sigma_j^a\sigma_j^b\rangle\rangle - \frac{1}{N}\langle\langle\frac{t}{\beta(1+t)}\sum_j\phi_j^a\phi_j^b\rangle\rangle \end{aligned}$$

onde ϕ são variáveis auxiliares que são introduzidas por meio de transformações de Hubbard-Stratonovich usuais, e q e Q são introduzidas por meio da simetria de réplicas $q_{ab} = Q\delta_{ab} + q(1-\delta_{ab})$. A média dupla se refere às médias sobre os padrões fixos e sobre a temperatura.

Para escrever a última igualdade, notamos que $\langle\langle \sigma_j^a \phi^a \rangle\rangle = 0$, assim os termos imaginários vão para 0.

Notamos que m é a magnetização usual, porém q e Q não são exatamente os parâmetros de Edward-Anderson usuais, mas estão relacionados a eles de uma maneira mais complicada.

Além disso, os significados das variáveis r e R não foram entendidos, sabemos somente que estão relacionadas de alguma maneira à semelhança entre réplicas, o que justifica a simetria de réplica usada (i.e. $r_{ab} = R\delta_{ab} + r(1 - \delta_{ab})$, já que réplicas se parecem mais consigo mesmas do que com outras).

Estas equações são suficientes para derivar o seguinte diagrama de fase:

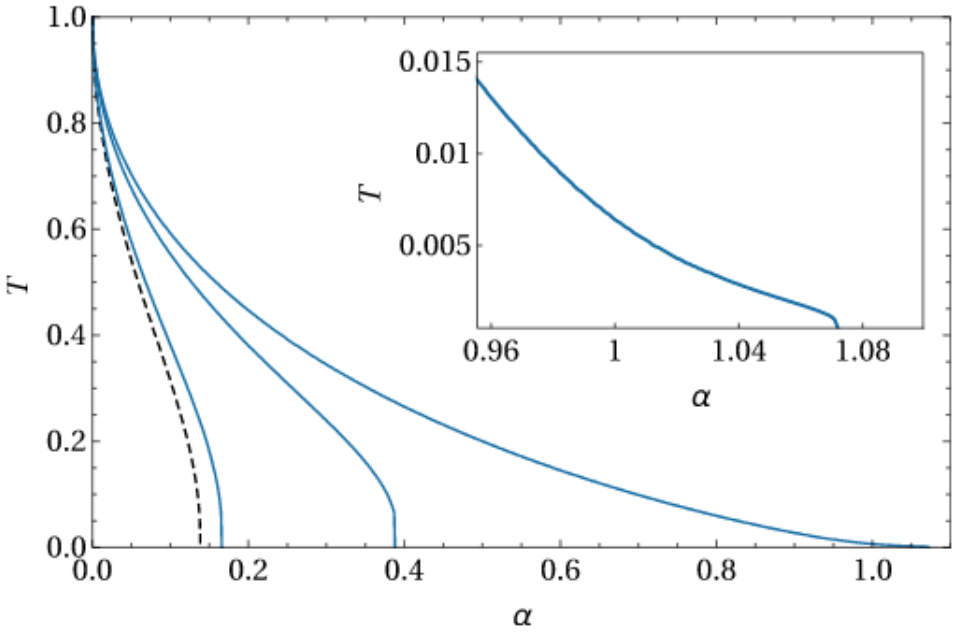


Figure 2.4: Diagrama de fase do algoritmo de *Reinforcement and Removal*, retirado de [8]. Aqui, temos 4 valores diferentes de t, da esquerda para a direita: $t = 0$, $t = 0.1$, $t = 1$, $t = 1000$. O quadro no canto superior direito mostra o comportamento para $t = 1000$ perto de $T = 0$, mostrando a quebra de simetria de réplicas nesta região.

Note como, diferentemente do modelo de Dotsenko, conforme aumentamos t, temos um aumento na capacidade e um aumento na resistência à temperatura, ou seja, só temos benefícios.

Apesar de não ser possível estabelecer um dependência fechada entre α_c e t para todas as temperaturas, a figura da dependência com temperatura 0 a seguir revela um pouco o comportamento

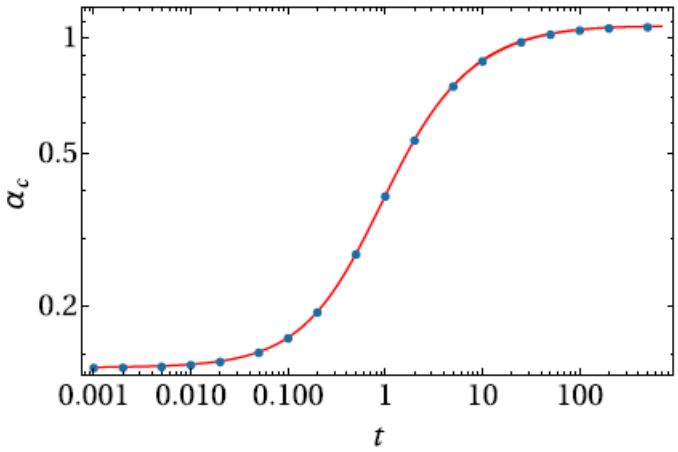


Figure 2.5: Gráfico usando a representação $log - log$ entre t e α com α normalizado e $\beta = \infty$, retirado de [8]. No caso, α normalizado significa $\alpha_{c,norm} = \frac{\alpha_c - \alpha_{min}}{\alpha_{max} - \alpha_{min}}$. O ajuste foi feito manualmente, ele consiste em $\log(\alpha_c) = \frac{\log(t)}{\log(t)+a}$, com $a = 2.84(1)$.

Em geral, a capacidade crítica aumenta significativamente na região aproximadamente entre $t = 0.1$ e $t = 10$, porém para valores mais altos aumenta muito pouco, dando a entender que existe uma convergência para uma certa matriz sináptica, como foi provado rigorosamente em [8].

Baseados nesse resultados, os autores estabeleceram que o parâmetro pode ser entendido como o tempo empregado nos sonhos, já que na teoria original de Crick e Mitchinson o sono REM é relevante até quando consegue remover a grande maioria de memórias espúrias, sendo de pouco uso depois. Diferentemente do *unlearning* hebbiano, esse modelo não exibe uma destruição do processamento de dados quando se sonha muito, o que biologicamente é mais plausível e consistente.

2.9 Trabalhos adicionais

Fazemos menção a trabalhos que, apesar de não serem ligados diretamente ao nosso projeto original, são importantes para o assunto em geral e nos ajudaram a ter uma visão mais ampla do que é possível fazer com estas técnicas e ideias.

O trabalho [20] expandiu a hipótese de Crick e Mitchinson, em particular criando hipóteses sobre os efeitos dos canabinoides nas sinapses neuronais e como eles poderiam evitar a presença de mínimos espúrios aumentando a mobilidade do processamento de informações do cérebro.

Os trabalhos [21] e [22] tratam de como existem algoritmos locais que transformam a matriz sináptica hebbiana na matriz projetora no limite termodinâmico e de quão eficazes eles são com um número N de neurônios finitos. Além disso, também são feitas considerações de que estes algoritmos podem ser transformados exatamente em equações diferenciais ordinárias onde a matriz sináptica evolui com o tempo, e que quando o tempo tendo ao infinito a solução única é ela ser igual à matriz da pseudo-inversa.

Chapter 3

Construção de modelos de redes neurais de atratores interagentes

3.1 Introdução

O ponto chave de nosso trabalho é, assim nos trabalhos anteriores, estender o significado do *unlearning* hebbiano, em particular, estendê-lo de tal modo a permitir que duas redes inicialmente desconectadas se conectem de maneiras particulares, e modifiquem suas sinapses de maneira benéfica ou não.

Atacamos isso de duas maneiras distintas: um algoritmo de *learning* hebbiano coletivo para ser usado em simulações de Monte Carlo e um Hamiltoniano para ter um sistema que possa ser resolvido analiticamente. O sistema a ser resolvido analiticamente usou métodos usuais, como o de réplicas e o de integração de ponto de sela, e posteriormente levou a uma análise numérica demorada das equações de consistência. A primeira parte nos levou a resultados pouco interessantes, mas a segunda parte demandou um tempo significativamente maior para ser completada e seus resultados complexos e ricos, o que justifica nosso foco majoritário nela. Apesar das duas serem distintas, elas apresentaram características similares que tem análogos com situações sociais do mundo real.

3.2 Algoritmo de *learning* hebbiano coletivo

Como vimos no capítulo 2, a base e motivação de algoritmos de *unlearning* é sempre biológica e sempre se baseia em uma rede alterando de maneira adequada sua própria matriz sináptica. O nosso procedimento muda estas duas principais características ao se basear em motivações sociológicas e permitir que outras matrizes sinápticas possam ser mudadas por outras redes, entendendo isso como um processo de comunicação.

O primeiro passo foi construir um algoritmo para ser usado em simulações de tamanho finito que busca mimetizar interações de aprendizado entre diferentes indivíduos, e ele funciona da seguinte maneira:

1. Comece com duas redes no mesmo estado aleatório.
2. Deixe as duas redes relaxar até um ponto fixo.
3. Faça as transições $J_{ij}^{(1)} \to J_{ij}^{(1)} + \frac{\epsilon^{(1)}}{N} S_i^{(2)} S_j^{(2)}$ e $J_{ij}^{(2)} \to J_{ij}^{(2)} + \frac{\epsilon^{(2)}}{N} S_i^{(1)} S_j^{(1)}$,

com $\epsilon^{(1)}, \epsilon^{(2)} \geq 0$ e onde o índice superior denota o número do indivíduo. Diferentemente do *unlearning* hebbiano, cada passo não representa um tempo que se sonhou, mas sim uma interação entre diferentes indivíduos. Nota-se também que se altera o sinal da mudança, pois queremos representar uma interação em que os indivíduos tem confiança uns nos outros, ou seja, pode reforçar determinadas sinapses benéficas, e não somente destruir algumas maléficas.

Após algumas simulações, que podem ser encontradas no apêndice 3, concluímos esse sistema consegue apresentar situações relevantes onde uma rede tem melhora no processamento de informações.

A primeira situação de professor-aluno é a mais relevante. Temos dois indivíduos, um com $\epsilon^{(1)} > 0$ e capacidade inicial baixa e o outro com $\epsilon^{(2)} = 0$ e capacidade alta. Nesse caso, temos que o segundo 'ensina' o primeiro, levando a ter magnetização intensiva ao longo do tempo. É importante frisar que não ocorre um aumento da capacidade crítica aqui, mas sim uma obsessão benéfica que ajuda a rede a processar corretamente uma pequena parte da informação disponível, diferindo do algoritmo de *unlearning* que melhora o *retrieval* de qualquer padrão e consegue efetivamente aumentar a capacidade crítica.

Além disso, temos que uma relação professor-aluno inversa resulta no decréscimo da capacidade, isto é, o aluno com capacidade baixa pode fazer o professor diminuir sua capacidade se trocarmos os valores de ϵ, permitindo assim comportamento mais variados. Por último, não encontramos situações onde interagir demais poderia destruir as magnetizações, assim como ocorre no *unlearning*.

Assim, os resultados obtidos não são muito interessantes, logo não nos ocupamos muito deles e deste algoritmo. Acreditamos que o modelo que introduziremos a seguir é de relevância significativamente maior.

3.3 Solução analítica de um modelo de redes neurais interagentes

A nossa intenção era conseguir criar um modelo baseado nos algoritmos de *unlearning* de tal forma que, ao mesmo tempo que conseguimos obter resultados novos, também conseguimos analisá-lo de maneira analítica, já que assim o entenderíamos de maneira mais clara. Pensando nisso e usando o algoritmo de *reinforcement* e *removal* como base, notamos que o Hamiltoniano a seguir acaba levando a uma mudança no campo sentido pelos spins parecida com as mudanças ocasionadas pelo algoritmo de *unlearning* original:

$$H_{int} \equiv -\frac{\epsilon}{N} \sum_{ij} \sigma_i S_j S_i \sigma_j, \tag{3.1}$$

onde $\epsilon > 0$, σ_i representa o spin i do indivíduo 1 e S_i representa o spin i do indivíduo 2. Os campos sentidos pelo spin i se tornam então

$$h_i^{(1)} = \sum_j \sigma_j (J_{ij}^{(1)} + \frac{\epsilon}{N} S_i S_j), \quad h_i^{(2)} = \sum_j S_j (J_{ij}^{(2)} + \frac{\epsilon}{N} \sigma_i \sigma_j); \tag{3.2}$$

onde $H = -\sum_i \sigma_i h_i^{(1)} - \sum_i S_j h_i^{(2)}$.

Notamos que, apesar da grande similaridade, esse modelo não representa exatamente uma solução analítica dos algoritmos, pois os campos não são exatamente iguais. Também temos

necessariamente uma simetria nas interações, por exemplo, não é possível ter uma situação onde $\epsilon^{(1)} > 0$ e $\epsilon^{(2)} = 0$ como anteriormente. Por fim, é um modelo pouco biológico, que não é diretamente aplicável a alguma situação específica.

Os hamiltonianos individuais das redes são construídos a partir de Hamiltonianos do algoritmo de *reinforcement and removal*, pois não só isso nos permite ter uma gama de possíveis matrizes mais robustas, mas também porque permite analisar interações entre redes diferentes, que individualmente tem comportamentos significativamente distintos.

O hamiltoniano total se torna

$$\begin{aligned} -\beta H = &\frac{\beta}{2N}\sum_{i,j}\sum_{\mu,\nu}\xi_i^{\mu A}\xi_j^{\nu A}(\frac{1+t_1}{\mathbb{1}+t_1 C})_{\mu\nu A}\sigma_i\sigma_j \\ &+\frac{\beta}{2N}\sum_{i,j}\sum_{\mu,\nu}\xi_i^{\mu B}\xi_j^{\nu B}(\frac{1+t_2}{\mathbb{1}+t_2 C})_{\mu\nu B}S_iS_j + \beta\epsilon N(\frac{1}{N}\sum_i \sigma_i S_i)^2. \end{aligned} \tag{3.3}$$

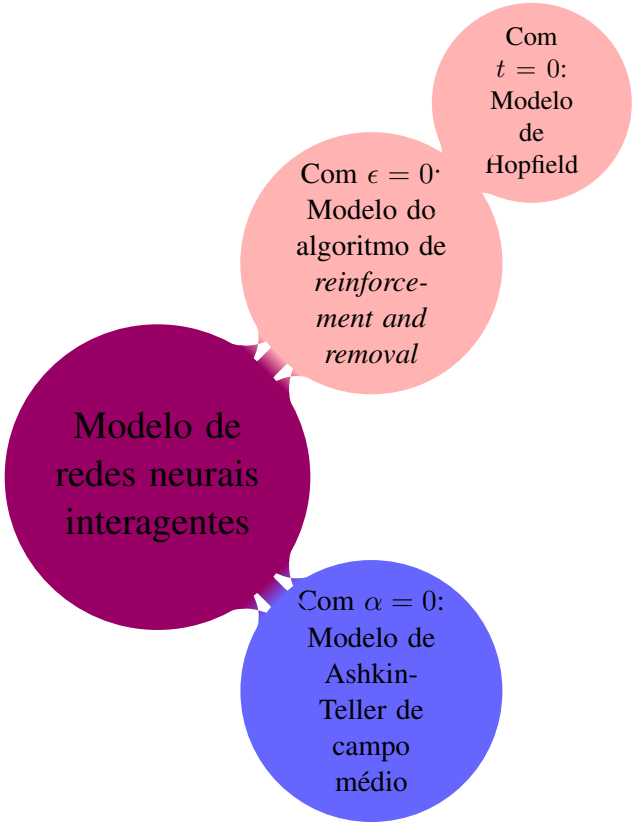
Para simplificar, consideramos que os padrões imbuídos nos indivíduos são os mesmos: $\xi^A = \xi^B$. Assim, ficamos com 5 parâmetros a serem variados: $t_1, t_2, \epsilon, \alpha, \beta$.

Vale a pena fazer um interlúdio para falar sobre a relação deste Hamiltoniano com outros. Com $\epsilon = 0$ temos duas redes com o algoritmo de *reinforcement and removal* separadas, e se adicionalmente fizermos $t_1 = t_2 = 0$, temos duas redes de Hopfield separadas, estas relações não são difíceis de ver pois cada modelo foi criado como uma generalização do anterior. Porém, também existe outra relação não óbvia, que é a que ocorre quando somente há um padrão, i.e. $p = 1$:

$$\begin{aligned} -\beta H &= \frac{\beta}{2N}\sum_{i,j}\xi_i^1\xi_j^1\sigma_i\sigma_j + \frac{\beta}{2N}\sum_{i,j}\xi_i^1\xi_j^1 S_iS_j + \beta\epsilon N(\frac{1}{N}\sum_i\sigma_i S_i)^2 \\ &= \frac{\beta}{2N}\sum_{i,j}\xi_i^1\sigma_i\xi_j^1\sigma_j + \frac{\beta}{2N}\sum_{i,j}\xi_i^1 S_i\xi_j^1 S_j + \beta\epsilon N(\frac{1}{N}\sum_i\xi_i^1\sigma_i\xi_i^1 S_i)^2 \\ &= \frac{\beta}{2N}\sum_{i,j}\sigma_i'\sigma_j' + \frac{\beta}{2N}\sum_{i,j}S_i'S_j' + \beta\epsilon N(\frac{1}{N}\sum_i\sigma_i' S_i')^2, \end{aligned} \tag{3.4}$$

com $\sigma_i' = \xi_i^1\sigma_i$ e $S_i' = \xi_i^1 S_i$, que é simplesmente o modelo de Ashkin-Teller totalmente conectado. Essa informação é útil principalmente ao analisarmos o sistema em números de padrões não extensivos, veremos que esta relação permite ter magnetização com $\beta \leq 1$, o que antes seria impossível. No apêndice 5 se encontram alguns resultados relevantes deste modelo.

A figura a seguir mostra a relação do nosso modelo com outros



Usando o método de réplicas e integração de ponto de sela, conseguimos encontrar uma solução analítica para o modelo encontrando uma expressão para a energia livre, que depende de 9 variáveis independentes, resultando em 9 equações de auto-consistência independentes, sendo que o sistema pode ser caracterizado de maneira satisfatória usando somente 3 variáveis:

$$m_1 = \langle\langle \frac{1}{N}\sum_i \xi_i^1 \sigma_i \rangle\rangle, \quad m_2 = \langle\langle \frac{1}{N}\sum_i \xi_i^1 S_i \rangle\rangle, \quad h = \langle\langle \frac{1}{N}\sum_i \sigma_i S_i \rangle\rangle, \tag{3.5}$$

sendo que h é uma variável nova que geralmente não aparece em modelos de vidros de spin, e que nos indica a semelhança entre as duas redes.

Como anteriormente, as outras variáveis não tem significado físico claro, sendo elas $q^\sigma, q^S, q^{\sigma S}, Q^\sigma, Q^S, Q^{\sigma S}$, em particular

$$q_{ab}^\rho = \langle\langle \frac{1}{N}\sum_j (\sigma_j^{a,\rho} + i\sqrt{\frac{t}{\beta(1+t)}}\phi^{a,\rho})(\sigma_j^{b,\rho} + i\sqrt{\frac{t}{\beta(1+t)}}\phi^{b,\rho})\rangle\rangle, \tag{3.6}$$

onde ρ pode ser σ, S ou σS. A partir daqui ρ tem esse significado. A justificativa para os termos imaginários irem a zero é a mesma que enunciamos na parte do algoritmo de *reinforcement and removal*. Além disso, como antes, o significado deste termo não é claro, e não será usado para diferenciar fases ou comportamentos.

Notamos que precisamos introduzir uma simetria de réplicas particular para esse sistema, que, apesar de ser basicamente a mesma usada em [8], parte de uma justificativa diferente. Ela tem a seguinte forma:

$$
\begin{aligned}
&m_1^a = m_1, m_2^a = m_2, h^a = h \quad \forall a, \\
&q_{ab}^\rho = Q^\rho \delta_{ab} + q^\rho (1 - \delta_{ab}) \quad \forall a, b, \rho, \\
&r_{ab}^\rho = R^\rho \delta_{ab} + r^\rho (1 - \delta_{ab}) \quad \forall a, b, \rho.
\end{aligned} \tag{3.7}
$$

As 3 primeiras formas não são surpreendentes: é usual pensar que a magnetização de cada réplica tem que ser igual, e com o mesmo raciocínio também deve ser a similaridade entre os indivíduos com o mesmo número de réplicas. As 4 formas relacionadas a $q^\sigma, q^S, r^\sigma, r^S$ seguem [8], mas as últimas 2 relacionadas a $q^{\sigma S}$ e $r^{\sigma S}$ não são óbvias, já que somente sabemos que $q_{ab}^{\sigma S}$ e $r_{ab}^{\sigma S}$ estão relacionados de alguma maneira vaga entre réplicas de número a e b. A justificativa em usar esse Ansatz consiste no fato de que o Hamiltoniano de interação somente introduz interação entre réplicas de mesmo número, assim elas devem ter uma similaridade especial que deve ser distinguida.

Vale a pena mencionar que tentamos inicialmente usar um Ansatz $q_{ab}^{\sigma S} = Q^{\sigma S}, r_{ab}^{\sigma S} = R^{\sigma S} \forall a, b$, porém o descartamos pois leva a resultados contraditórios. Adicionalmente, verificamos que o Ansatz (3.7) não acaba degenerando no Ansatz acima, isto é, em situações com $\epsilon \neq 0$ temos que $Q^{\sigma S} \neq q^{\sigma S}$ e $R^{\sigma S} \neq r^{\sigma S}$.

Com exceção da necessidade de fazer uma transformação particular de Hubbard-Stratonovich de maneira cuidadosa, o resto dos cálculos não tem nenhuma grande novidade em relação a [8], e eles estão detalhados com mais cuidado no apêndice.

A energia livre é o resultado da extremização dados os parâmetros intensivos $(\alpha, \beta, t_1, t_2, \epsilon)$ no espaço dos parâmetros de ordem $(m_1, m_2, h, q^\sigma, q^S, q^{\sigma S}, Q^\sigma, Q^S$
$, Q^{\sigma S}, \Delta^\sigma, \Delta^S, \Delta^{\sigma S}, r^\sigma, r^S, r^{\sigma S})$ leva a 15 equações que podem ser reduzidas a 9, pois podemos escrever $(\Delta^\sigma, \Delta^S, \Delta^{\sigma S}, r^\sigma, r^S, r^{\sigma S})$ como função das outras 9 variáveis.

$$
\begin{aligned}
&f(m_1, m_2, h, Q^\sigma, Q^S, Q^{\sigma S}, q^\sigma, q^S, q^{\sigma S}, t_1, t_2, \beta, \alpha, \epsilon) = \\
&\frac{m_1^2}{2 + 2t_1} + \frac{m_2^2}{2 + 2t_2} + \epsilon h^2 + \frac{(\Delta^\sigma - 1)(1 + t_1)}{2t_1} Q^\sigma + \frac{\log[\Delta^\sigma \Delta^S - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2]}{2\beta} \\
&+ \frac{\alpha\beta}{2} r^\sigma (Q^\sigma - q^\sigma) + \frac{(\Delta^S - 1)(1 + t_2)}{2t_2} Q^S + \frac{\alpha\beta}{2} r^S (Q^S - q^S) \\
&+ \Delta^{\sigma S} Q^{\sigma S} + \frac{\alpha\beta r^{\sigma S}}{2} (Q^{\sigma S} - q^{\sigma S}) + \frac{\alpha}{2\beta} \log\{\frac{1}{2}[2 + \beta(1 + t_1)(q^\sigma - Q^\sigma) + \beta(1 + t_2)(q^S - Q^S) \\
&- \beta\sqrt{((1 + t_1)(q^\sigma - Q^\sigma) - (1 + t_2)(q^S - Q^S))^2 + 4(1 + t_1)(1 + t_2)(-Q^{\sigma S} + q^{\sigma S})^2}]\} \\
&+ \frac{\alpha}{2\beta} \log\{\frac{1}{2}[2 + \beta(1 + t_1)(q^\sigma - Q^\sigma) + \beta(1 + t_2)(q^S - Q^S) \\
&+ \beta\sqrt{((1 + t_1)(q^\sigma - Q^\sigma) - (1 + t_2)(q^S - Q^S))^2 + 4(1 + t_1)(1 + t_2)(-Q^{\sigma S} + q^{\sigma S})^2}]\} \\
&- \alpha\{q^\sigma (1 + t_1)[1 - \beta(Q^S - q^S)(1 + t_2)] \\
&+ q^S (1 + t_2)[1 - \beta(Q^\sigma - q^\sigma)(1 + t_1)] + \beta(1 + t_1)(1 + t_2) q^{\sigma S}(-q^{\sigma S} + Q^{\sigma S})\} \\
&\times\{2[1 - \beta(1 + t_1)(Q^\sigma - q^\sigma)][1 - \beta(1 + t_2)(Q^S - q^S)] - 2\beta^2 (1 + t_1)(1 + t_2)(-Q^{\sigma S} + q^{\sigma S})^2\}^{-1} \\
&- \frac{1}{\beta} \int Dx Dy \log\{\cosh[\beta(\upsilon + \Upsilon)] \exp(\beta\eta) + \cosh[\beta(\upsilon - \Upsilon)] \exp(-\beta\eta)\} \\
&- \frac{\Delta^S t_1'^2 \beta(\alpha r^\sigma + m^2 + \frac{1}{\beta^2 t_1'^4}) + \Delta^\sigma t_2'^2 \beta(\alpha r^S + m_2^2 + \frac{1}{\beta^2 t_2'^4}) + 2t_1'^2 t_2'^2 \beta^2 \Delta^{\sigma S}(\frac{r^{\sigma S}\alpha}{2} + m_1 m_2)}{2\Delta^\sigma \Delta^S - 2t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2}.
\end{aligned} \tag{3.8}
$$

onde

$$\eta = 2\epsilon h + \frac{\Delta^{\sigma S}}{\Delta^{\sigma}\Delta^{S} - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2}, \tag{3.9}$$

$$t_1' = i\sqrt{\frac{t_1}{\beta(t_1+1)}}, t_2' = i\sqrt{\frac{t_2}{\beta(t_2+1)}}, r_{\pm} = \sqrt{1 \pm \frac{r_{\sigma S}}{2\sqrt{r^{\sigma} r^{S}}}},$$

$$\upsilon = \frac{\Delta^{S}[m_1 + \sqrt{\frac{\alpha r^{\sigma}}{2}}(r_+ x + r_- y)] + t_2'^2 \beta \Delta^{\sigma S}[m_2 + \sqrt{\frac{\alpha r^{S}}{2}}(r_+ x - r_- y)]}{\Delta^{\sigma}\Delta^{S} - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2},$$

$$\Upsilon = \frac{\Delta^{\sigma}[m_2 + \sqrt{\frac{\alpha r^{S}}{2}}(r_+ x - r_- y)] + t_1'^2 \beta \Delta^{\sigma S}[m_1 + \sqrt{\frac{\alpha r^{\sigma}}{2}}(r_+ x + r_- y)]}{\Delta^{\sigma}\Delta^{S} - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2}.$$

As equações de consistência se encontram no apêndice 1.

Mencionamos sucintamente o problema de estabelecermos se a solução é estável, e se sim, onde que é. Como foi provado por Dotsenko [11], o modelo com a matriz de Dotsenko tem uma quebra da simetria de réplicas parecida com a do modelo de Hopfield, ou seja, ela somente é quebrada em uma região pequena, trazendo poucas consequências para a solução. Se baseando nisso, é razoável assumir que o modelo de *reinforcement and removal* também tem uma quebra de estabilidade pouca significativa, como foi enunciado explicitamente no próprio trabalho de Fachechi *et al* [8].

Para o nosso modelo, com baixas interações não nos afastamos muito do modelo anterior, assim não deveríamos esperar imprecisões grandes na solução nessa região. Porém, a situação fica menos clara com maiores interações, já que nosso modelo é significativamente diferente de outro modelos com a estabilidade conhecida. Calcular a estabilidade do jeito normal, como foi feito em Almeida e Thouless [23], demoraria um tempo de ordem de grandeza maior do trabalho apresentado, assim decidimos não fazer esta parte, e simplesmente assumimos que a simetria usada é suficientemente estável para conseguirmos resultados representativos.

3.4 Trabalhos adicionais

Nos trabalhos de Sakata e Hukushima[24] e Takayama [25] foram criados modelos cujos Hamiltonianos são parecidos com o nosso, eles consideraram dois sistemas distintos, um deles é um ferromagneto puro e o outro é um vidro de spin, e existe uma interação na forma de vidro de spin entre os dois. Apesar de serem modelos mais gerais e não representarem redes neurais, existe uma relação de manipulação entre os dois sistemas, onde o sistema de vidro de spin pode desestabilizar ou estabilizar o ferromagneto; além disso, o sistema apresenta uma reentrância peculiar, assim como em algumas parte particulares o nosso aparenta apresentar. Por último, notamos que seria interessante considerar *partial annealing* no contexto do nosso modelo da mesma maneira que foi feito em [24].

O trabalho Kuva *et al.*[26] considera o problema inverso de tentar encontrar um Hamiltoniano dados mínimos da energia, e podemos relacioná-lo com o fato de que, ao interagir, as redes no nosso sistema fazem um processo similar, de tentar entender qual a melhor matriz sináptica dadas as respostas da outra rede.

Por fim, citamos o trabalho [5], que apesar de considerar um modelo mais aplicado e usar

outros métodos, serviu como uma das principais motivações a tentar usar redes neurais interagentes para entender fenômenos sociais.

Chapter 4

Resultados

4.1 Introdução

Enunciamos e explicamos detalhadamente os resultados encontrados que avaliamos como relevantes. A complexidade do sistema e o número grande de parâmetros a serem mudados levou a 3 formas distintas de se representar o espaço de fase, cada uma complementando a visão da outra. Elas são coerentes entre si, e acreditamos são suficientes para entender as principais propriedades, embora não sejam totalmente exaustivas.

Além disso, explicamos com detalhes como e com quais técnicas chegamos nesses resultados, e de que maneira eles podem ser conectados a resultado interessantes. Em particular, notamos que nosso modelo em uma certa região se torna outro modelo muito mais bem compreendido, e que resultados clássicos deste explicam uma propriedade nova daquele.

4.2 Técnicas usadas

Antes de apresentar os resultados, é útil falar sobre como foi feita a resolução numérica das equações. Todos os programas foram desenvolvidos e executados em Python 3.8.

Supomos que temos M equações para as M variáveis independentes, sendo elas dadas por $x_i = g_i(\vec{x})$, onde $\vec{x} = \{x_1, x_2, ..., x_M\}$, no caso queremos encontrar $\vec{x}$ que satisfaz todas essas equações simultaneamente. O método de ponto fixo consiste então nos seguintes passos:

- Começar com valores iniciais relativamente próximos do esperado, os chamamos de $\vec{x}^{(0)}$.
- Atualizar a primeira componente de $\vec{x}^{(0)}$ por meio de $x_1^{(1)} = \eta x_1^{(0)} + (1-\eta) g_1(\vec{x}^{(0)})$, onde $0 \leq \eta < 1$.
- Fazer o mesmo com as outras componentes, notando que o argumento de g_i já tem as componentes atualizadas até $i-1$.
- Comparar a diferença absoluta máxima entre as componentes do vetor novo $\vec{x}^{(1)}$ e do vetor velho $\vec{x}^{(0)}$ com um certo valor ε. Se for maior, voltar ao passo 1 usando o vetor novo como velho, se for menor, parar o programa e considerar esta solução como a solução definitiva. Para evitar tempos de computação demasiadamente grandes, se o número de interações passa de certo valor alto, o programa para.

É possível provar, e.g. ver [27], que para todo sistema de equações com pelo menos uma solução esse algoritmo leva a uma solução arbitrariamente próxima da solução real se $\eta \geq \eta_c$, se as condições iniciais forem bem calibradas e as derivadas $\frac{\partial g_i}{\partial x_j}$ forem limitadas. Infelizmente, a menos de problemas bem específicos, não é possível determinar η_c, pode ser difícil dar valores iniciais razoáveis e não é clara a relação entre o erro da solução aproximada e o limite. Além disso, o valor adotado para η e o valor adotado para o limite do erro são relacionados, já que se aumentarmos o primeiro temos uma mudança menor a cada passo e precisamos diminuir o limite, também aumentando significativamente o tempo da resolução numérica.

No nosso caso, o segundo problema é de menor importância, já que tirando as variáveis m_1, m_2 e h, cujos valores iniciais adequados estão por volta de 0.99, os valores iniciais das outras 6 variáveis não mudam significativamente os resultados finais se estiverem nos intervalos $[-1, 1]$, isso foi concluído após testar o programa em diferentes regiões com diferentes condições iniciais.

Em relação ao primeiro problema, chegamos à conclusão de que $\eta_c = 0.9$ é um valor razoável para esse parâmetro, já que rende resultados consistentes com o caso conhecido $\epsilon = 0$ em diferentes situações e permite tempos razoáveis para a resolução numérica; similarmente determinamos $\varepsilon = 3 \cdot 10^{-5}$ adequado comparando com o caso $\epsilon = 0$, em particular com esse valor nunca passamos o número máximo de interações de 2000, indicando que sempre houve uma boa convergência.

4.3 Diagramas mostrando a relação entre T e α_c

Nota: por convenção, t_1 se refere ao menor valor de t e t_2 ao maior em todos os gráficos e diagramas.

Para caracterizar as fases introduzimos os parâmetros

$$\Delta m_1 = m_1(\epsilon) - m_1(0), \quad \Delta m_2 = m_2(\epsilon) - m_2(0), \quad \Delta h = h(\epsilon) - h(0), \tag{4.1}$$

que mostram as mudanças nos parâmetros de ordem devido à interação entre as redes. Com essa representação obtemos visualizações que são mais similares a modelos do tipo de Ashkin-Teller, e nesse caso dividimos os diagramas em 7 fases:

Fases	m_1	m_2	h	$\Delta m_1(\epsilon)$	$\Delta m_2(\epsilon)$	$\Delta h(\epsilon)$
Insuficiente	≈ 0	$\neq 0$	≈ 0	≈ 0	≈ 0	≈ 0
Professor-estudante	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	≈ 0	$\neq 0$
Mutualismo	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$
Desordenada	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
Ilusão reforçada	≈ 0	≈ 0	$\neq 0$	≈ 0	≈ 0	$\neq 0$
Indiferente	$\neq 0$	$\neq 0$	$\neq 0$	≈ 0	≈ 0	≈ 0
Amensalismo	≈ 0	≈ 0	≈ 0	≈ 0	$\neq 0$	≈ 0

Lembramos que está implícito que nos casos em que o valor na tabela é $\neq 0$, temos que o

valor é positivo também.

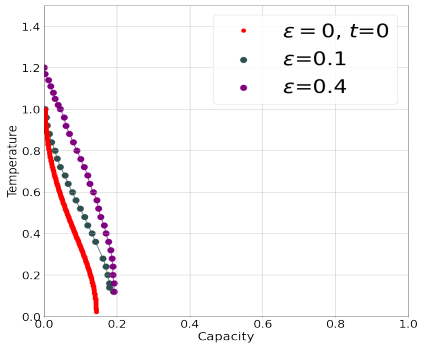
As fases indiferente, de professor-estudante e de mutualismo podem ser englobadas em uma categoria mais geral, que chamamos de Pseudo-Baxter pois nessa região temos que $m_1 \neq m_2$ em geral, diferente da igualdade $m_1 = m_2$ que geralmente ocorre em fases Baxter.

Nos casos em que analisamos, a fase insuficiente acaba tendo uma pequena área que desaparece conforme aumentamos ϵ e as fases pseudo-Baxter e de ilusão reforçada vão cobrindo a fase desordenada e para ϵ grande acabamos sempre tendo uma área significativa das últimas 3 fases.

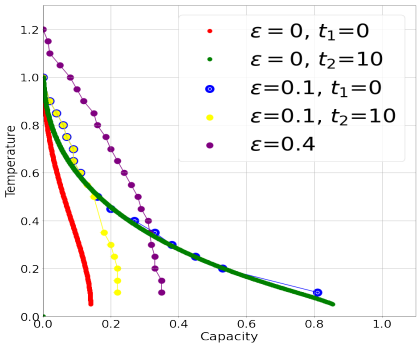
O aumento do valor ϵ leva em geral a um aumento da capacidade crítica, porém em alguns casos onde a diferença entre t_1 e t_2 é significativa em temperaturas baixas o inverso acontece. Os 6 diagramas de fase a seguir mostram esse fenômeno. No caso, neles fixamos ϵ, t_1 e t_2, variando β para encontrar a capacidade máxima.

À esquerda das linhas dos gráficos, temos magnetização intensiva, e à direita não temos. Assim, as linhas nos dizem como a capacidade crítica varia conforme a temperatura muda, ou como a temperatura crítica muda conforme a capacidade varia.

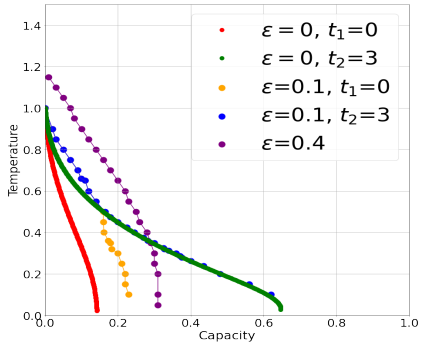
Em casos onde a diferença entre t_1 e t_2 é pequena, como no diagrama 1, as redes tem uma capacidade crítica conjunta seja com $\epsilon = 0.1$ seja com $\epsilon = 0.4$, no sentido de que não existe espaço onde uma rede tem magnetização e a outra não, assim temos que só são necessárias duas linhas para denotar as capacidades críticas com interação. Quando a diferença entre t_1 e t_2 é grande, como no diagrama 2, as capacidades críticas são diferentes, assim devemos usar 3 linhas distintas para as capacidades críticas. As linhas com $\epsilon = 0$ foram colocadas a fim de comparação, para entender o que muda com a introdução da interação.



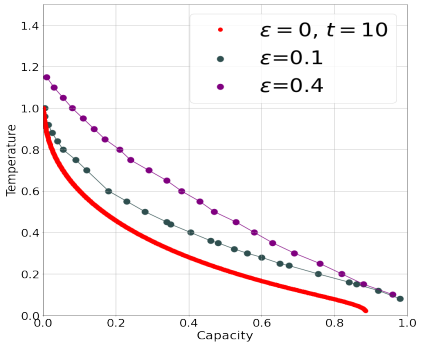
(a) Diagrama 1. Parâmetros: $(t_1, t_2, \epsilon) = (0, 0, 0/0.1/0.4)$. Duas redes com capacidades iniciais baixas interagindo. Notamos que as redes sempre tem a mesma capacidade crítica, e independente da interação conseguem ter uma melhora em relação ao caso sem interação.



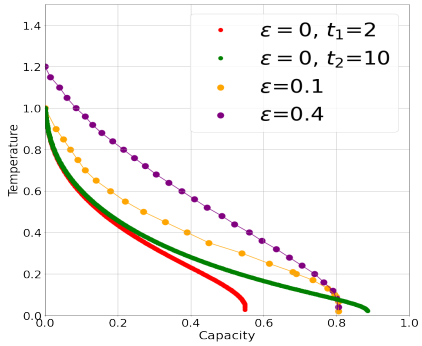
(b) Diagrama 2. Parâmetros: $(t_1, t_2, \epsilon) = (0, 10, 0/0.1/0.4)$. Aqui temos uma situação de uma rede com capacidade inicial baixa interagindo com uma rede com capacidade inicial alta. Notamos que as redes tem capacidade crítica diferente em geral, somente com interações maiores é que se juntam. Além disso, com interações maiores e temperaturas baixas temos diminuição da capacidade crítica para as redes com maior valor de t.



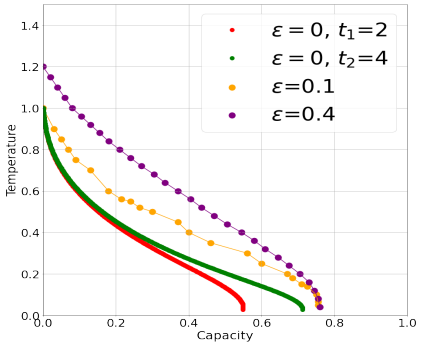
(a) Diagrama 3. Parâmetros: $(t_1, t_2, \epsilon) = (0, 3, 0/0.1/0.4)$. Uma rede com capacidade inicial baixa interagindo com uma rede com capacidade inicial alta. Este diagrama tem propriedades similares ao diagrama 2, já que a diferença entre um indivíduo com $t = 3$ e $t = 10$ não é tão grande, em geral as curvas apresentam uma menor capacidade crítica, mas não tem grande diferença qualitativa.



(b) Diagrama 4. Parâmetros: $(t_1, t_2, \epsilon) = (10, 10, 0/0.1/0.4)$. Duas redes com capacidades iniciais altas interagindo. Vemos que somente temos melhora na capacidade conforme aumentamos a interação, em particular temos um grande aumento em temperaturas altas e um aumento mais modesto para temperaturas baixas, que parece respeitar aproximadamente o limite $\alpha_c = 1$ que temos para modelos de redes neurais com Hamiltonianos da forma de Ising.



(a) Diagrama 5. Parâmetros: $(t_1, t_2, \epsilon) = (2, 10, 0/0.1/0.4)$. Uma rede com capacidade inicial média interagindo com uma rede com capacidade inicial alta. Para temperaturas altas, não temos diferenças significativas com o diagrama 4, já que temos de novo um aumento significativo da capacidade crítica conforme aumentamos a interação. Para temperaturas baixas temos a diferença que a capacidade crítica diminui em relação ao indivíduo com maior capacidade inicial, fenômeno que não era visualizado anteriormente.

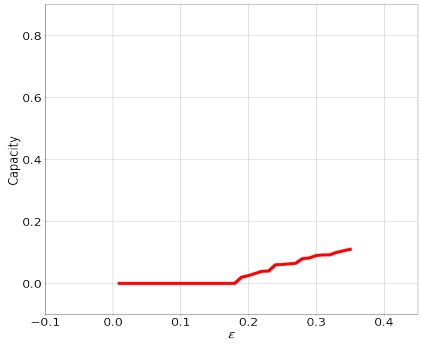


(b) Diagrama 6. Parâmetros: $(t_1, t_2, \epsilon) = (2, 4, 0/0.1/0.4)$. Duas redes com capacidades iniciais médias interagindo. Não temos diferenças qualitativas em relação ao diagrama 4, de novo a interação aumenta significativamente a capacidade crítica, em particular em temperaturas mais altas.

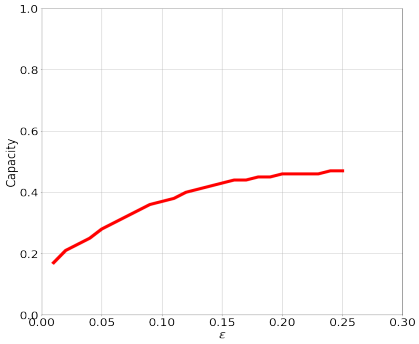
4.4 Diagramas mostrando a relação entre ϵ e α_c

Outra representação que achamos interessante analisar foi fixar β, t_1 e t_2, variando ϵ para encontrar a capacidade crítica. Nesses diagramas, é mais nítida a mudança no comportamento do sistema quando aumentamos ϵ. Observamos que, curiosamente, o indivíduo com maior t somente aumenta sua capacidade crítica quando o indivíduo com menor t o alcança, veja diagramas 13 e 14. Além disso, todos os diagramas exibem um certo platô a partir de certo ponto, indicando que aumentar ϵ só traz diferenças significativas até alguma magnitude. Por fim, vemos novamente que para temperaturas baixas e diferenças de t significativas temos uma diminuição da capacidade crítica no indivíduo t_2, e em particular essa transição aparenta ser de primeira ordem.

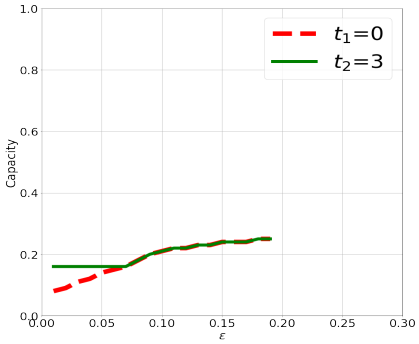
Nesses diagramas, abaixo da linha temos magnetização, e acima da linha não, quando a linha está sobre o eixo das abscissas temos que não existe capacidade que deixe possível ter magnetização nessa região. Assim como os diagramas anteriores, temos diagramas que temos capacidade crítica conjunta como o diagrama 1, assim só precisamos de uma linha para representá-los; já quando esta diferença é grande, como no diagrama 13, precisamos de duas linhas para representá-los.



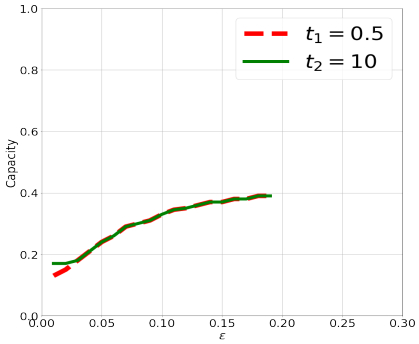
(a) Diagrama 7. Parâmetros: $(t_1, t_2, \beta) = (10, 10, 0.9)$. Este diagrama representa a situação onde consideramos uma região ($\beta \leq 1$) em que não é possível armazenar padrões sem interação, mas com interação é permitido ter armazenamento significativo, pelo fato de considerarmos redes iguais, temos que a capacidade é conjunta. Notem como os resultados batem com os conhecidos resultados do modelo de Ashkin-Teller.



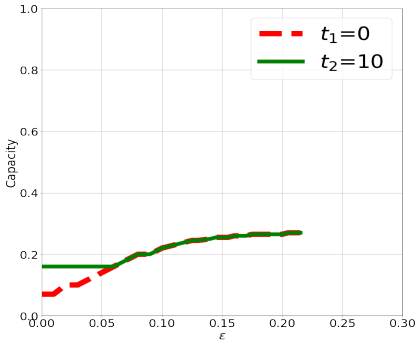
(b) Diagrama 8.Parâmetros: $(t_1, t_2, \beta) = (2, 4, 2)$. Neste aqui, vemos como duas redes similares conseguem se ajudar em altas temperaturas. É necessária somente uma linha, já que t_1 e t_2 são muito próximos.



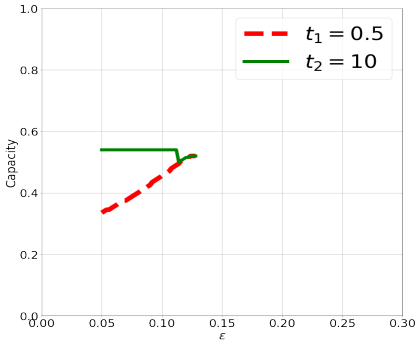
(a) Diagrama 9. Parâmetros: $(t_1, t_2, \beta) = (0, 3, 2)$. Para diferenças significativas entre t_1 e t_2, é necessária uma interação significativa para juntar as redes, sendo nesse caso $\epsilon \approx 0.06$. Para valores de interações mais fortes, a capacidade máxima do segundo agente também aumenta, assim passamos de uma fase professor-aluno para uma fase de mutualismo.



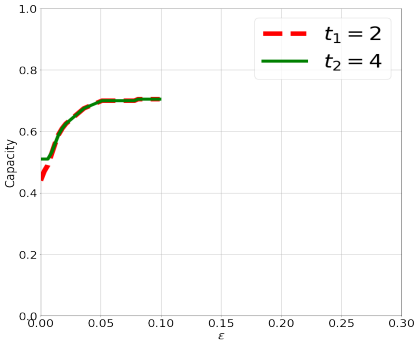
(b) Diagrama 10. Parâmetros: $(t_1, t_2, \beta) = (0.5, 10, 5)$. Esse diagrama é parecido com o diagrama 9, porém vemos que pela diferença entre t_1 e t_2 ser significativamente menor, eles se encontram com uma interação menor, e conseguem alcançar capacidade críticas significativamente maiores.



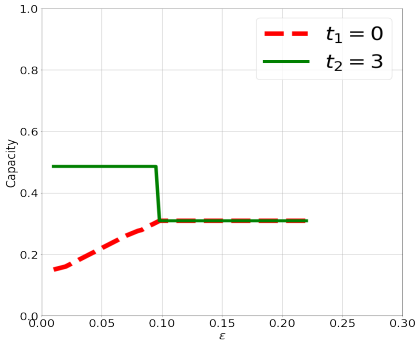
(a) Diagrama 11. Parâmetros: $(t_1, t_2, \beta) = (0, 10, 2)$. Este diagrama não apresenta muitas diferenças qualitativas em relação ao diagrama 9, temos somente que a capacidade crítica conjunta é ligeiramente maior.



(b) Diagrama 12. Parâmetros: $(t_1, t_2, \beta) = (0, 10, 5)$. Este diagrama apresenta um caso onde a rede 2 basicamente não se altera com a interação, mesmo com interações fortes. Apesar disso, a rede 1 aumenta bastante sua capacidade crítica.



(a) Diagrama 13. Parâmetros: $(t_1, t_2, \beta) = (2, 4, 5)$. Diferentemente do diagrama 8, temos que existe uma região significativa onde as duas redes tem capacidades criticas distintas, mas não é necessária uma forte interação para juntá-las. De novo, a interação aumenta significativamente a capacidade crítica entre eles, eventualmente chegando a um platô.



(b) Diagrama 14. Parâmetros: $(t_1, t_2, \beta) = (0, 10, 5)$. Este diagrama apresenta uma situação peculiar, que só encontramos em baixas temperaturas, altas capacidade e fortes interações: a diminuição da capacidade crítica com aumento da interação. Vemos que inicialmente as capacidade críticas estão muito separadas, e a rede 1 não consegue alcançar a rede 2, mas sim uma via de meio conjunta que acaba abaixando a capacidade da rede 2. Perceba também que essa transição parece ser de primeira ordem para a rede 2.

4.5 Diagramas mostrando a relação entre ϵ e T_c

Por fim, fixamos α, t_1 e t_2 e variamos ϵ para ver a mudança no β crítico.

A seguir se encontram 8 diagramas que dão uma ideia de quando e como estas fases aparecem.

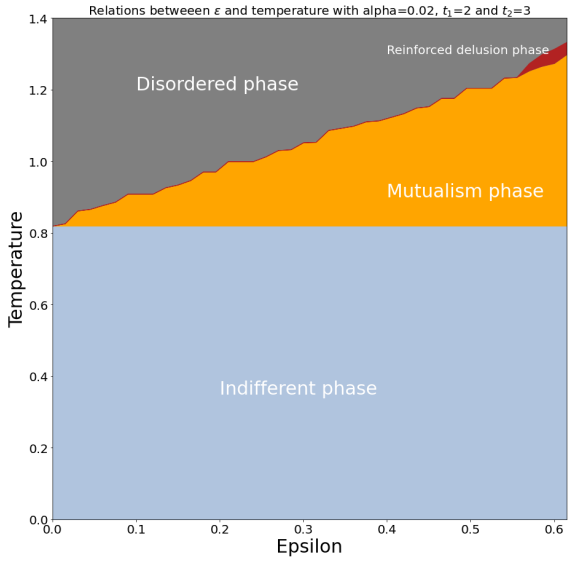


Diagrama 15. Neste caso, o fato de t_1 e t_2 serem muito próximos tem como consequência a fase insuficiente não ter tamanho apreciável, assim não a representamos. Vemos como a interação aumenta substancialmente as fases pseudo-Baxter, e como é necessária uma interação muito forte para aparecer a fase de ilusão reforçada.

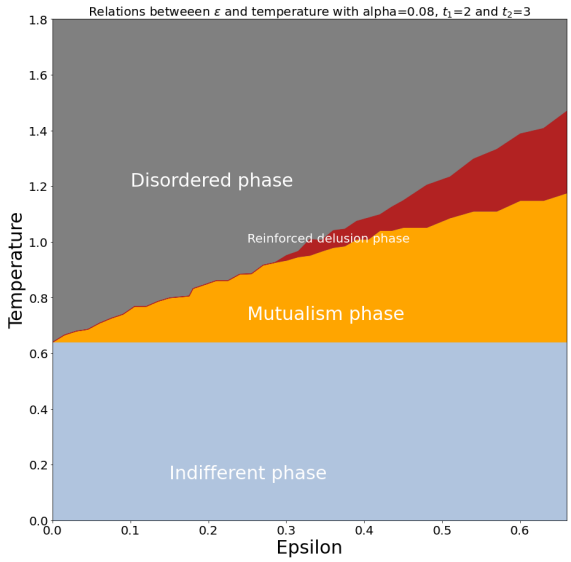


Diagrama 16. Um aumento médio da capacidade em relação ao diagrama 15 leva a um aumento do espaço da fase desordenada e principalmente um aumento na fase de ilusão reforçada, diminuindo as fases de mutualismo e indiferente.

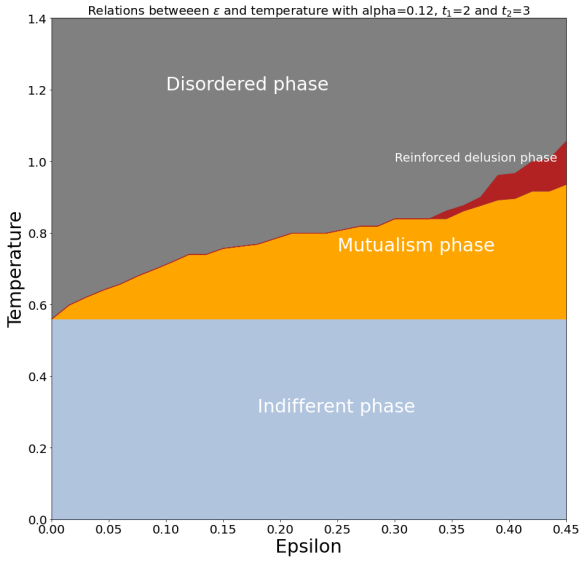


Diagrama 17. Um aumento grande da capacidade em relação ao diagrama 15 reforça as consequências mencionadas no diagrama anterior, agora o aumento mais significativo está na fase desordenada.

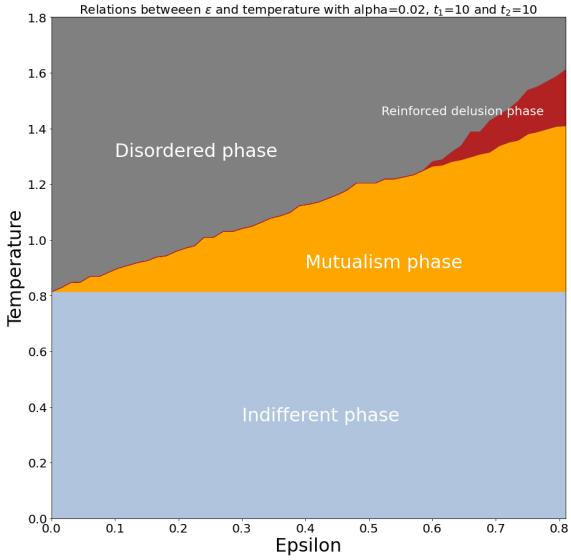


Diagrama 18. Temos $t_1 = t_2$, assim de novo não temos fase insuficiente. Notamos que o esquema qualitativo não difere muito dos 3 diagramas anteriores, o aumento da capacidade aumenta significativamente a fase de mutualismo e indiferente até um certo ponto onde começa a aparecer a fase de ilusão reforçada, e ela aumenta mais conforme aumentamos a interação.

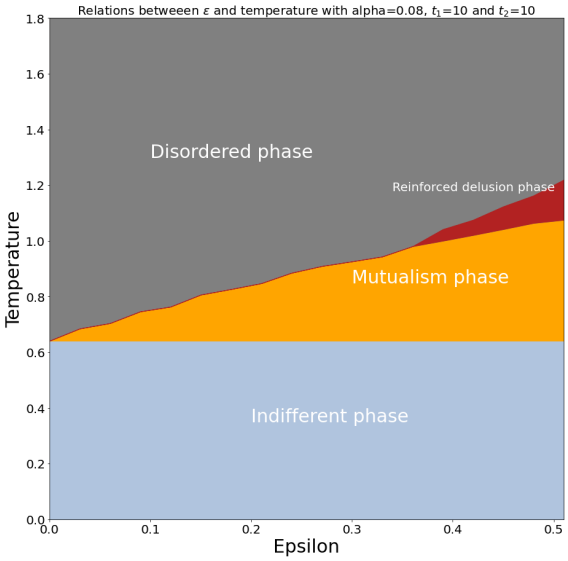


Diagrama 19. Os efeitos de aumentar a capacidade que podem ser vistos nesse diagrama e no seguinte são de novo similares aos casos com $t_1 = 2$ e $t_2 = 3$: aumento da fase desordenada e da fase de ilusão reforçada.

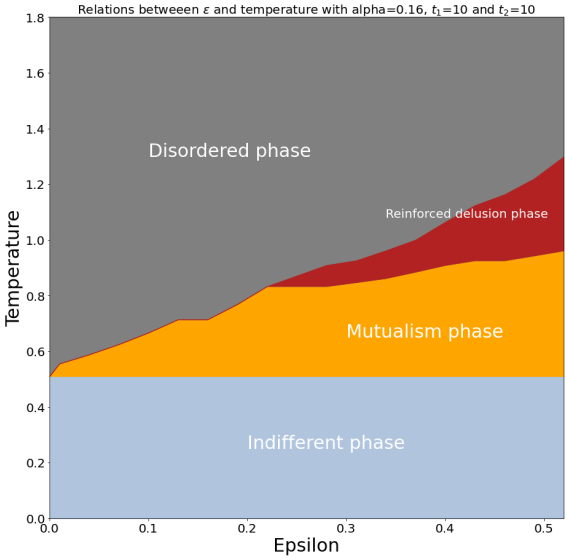


Diagrama 20

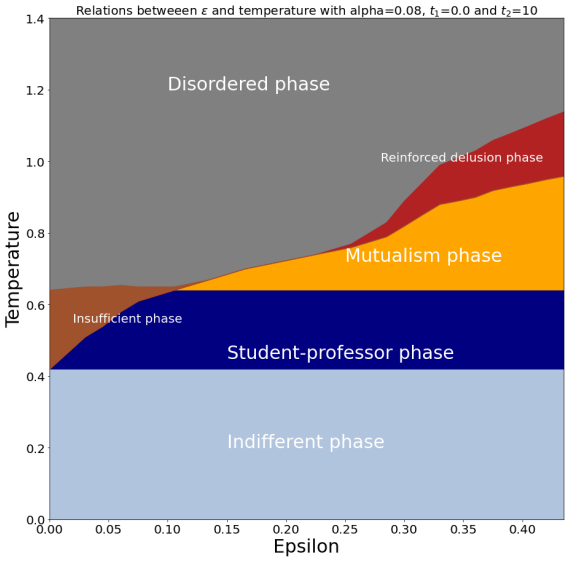


Diagrama 21.Agora temos fase insuficiente, já que t_1 e t_2 são significativamente distintos. Conforme a interação aumenta, vemos que ela rapidamente diminui, e ocupa um espaço pequeno nesse caso. Temos também o aparecimento da fase professor aluno, que a partir de $\epsilon \approx 0.12$ substitui totalmente a fase insuficiente.

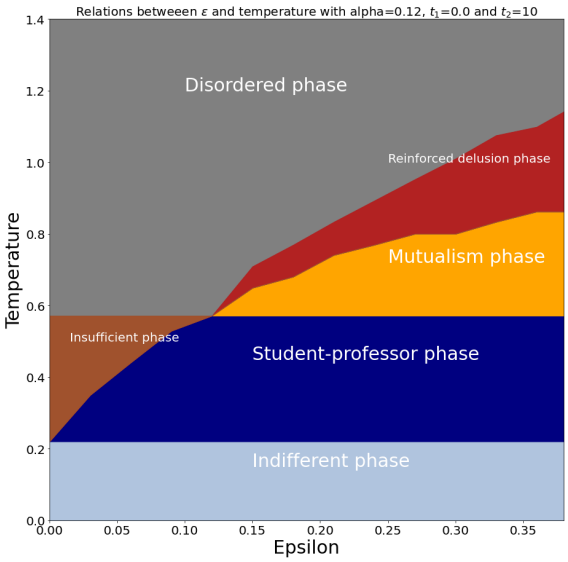


Diagrama 22.O aumento da capacidade leva a um aumento da fase insuficiente também, nesse caso permitindo que ela e a fase de ilusão reforçada se encostem.

Apresentamos quatro cortes verticais dos diagramas anteriores que ilustram um pouco como Δm_1, Δm_2 e Δh variam com a temperatura para um ϵ fixo.

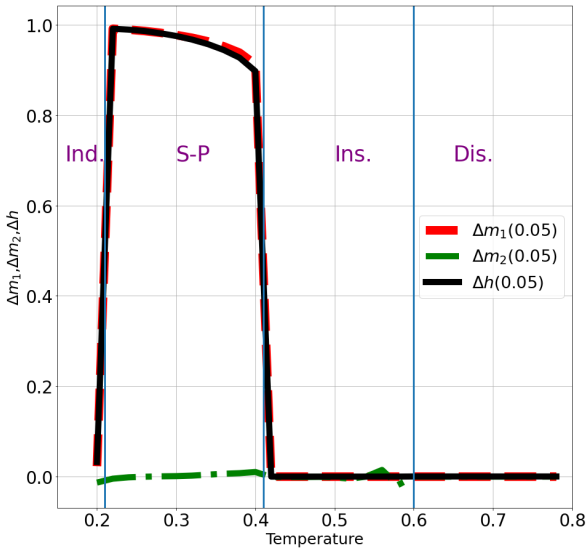


Diagrama 23. Parâmetros: $(t_1, t_2, \alpha, \epsilon) = (0, 10, 0.12, 0.05)$. Esta imagem deve ser comparada com o diagrama 22. Vemos que com interação fraca e capacidade alta, a fase de mutualismo acaba sendo pequena, e em $T \approx 0.4$ já estamos na fase desordenada.

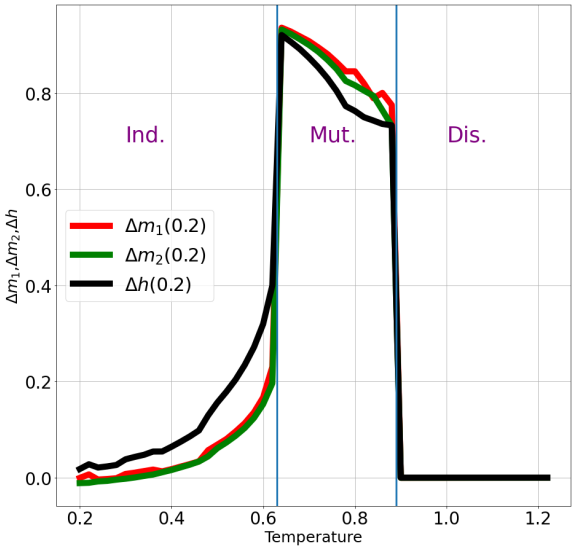


Diagrama 24. Parâmetros: $(t_1, t_2, \alpha, \epsilon) = (2, 3, 0.08, 0.2)$. Esta imagem deve ser comparada com o diagrama 16. Vemos que, pelo fato de t_1 ter valor próximo a t_2, temos que qualitativamente Δm_1 e Δm_2 são similares, eles só alcançam valores significativos a partir de $T \approx 0.62$, onde entram na fase de mutualismo, e depois de $T \approx 0.88$ eles não conseguem mais se ajudar, entrando na fase desordenada.

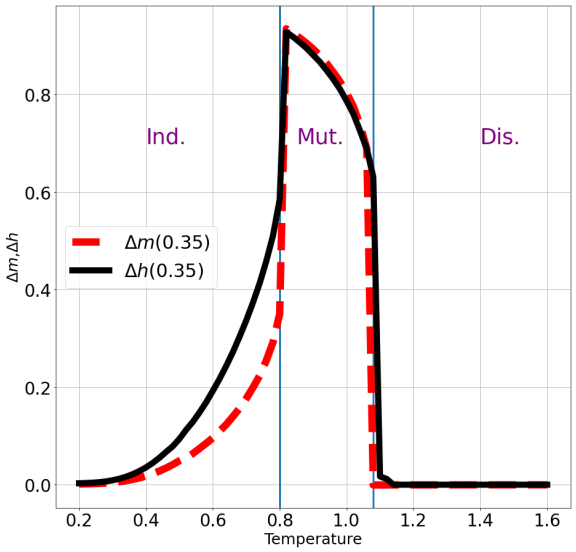


Diagrama 25. Parâmetros: $(t_1, t_2, \alpha, \epsilon) = (10, 10, 0.02, 0.35)$. No caso, consideramos somente $\Delta m \equiv \Delta m_1 = \Delta m_2$, já que $t_1 = t_2$, esta imagem é para ser visualizada junto com o diagrama 18. Vemos que até a temperatura 0.8, Δm e Δh não tem valores muito altos, somente depois dessa temperatura e até $T \approx 1.05$ temos valores significativos. Depois, temos que a interação não consegue mais ajudar os agentes, e eles entram na fase desordenada.

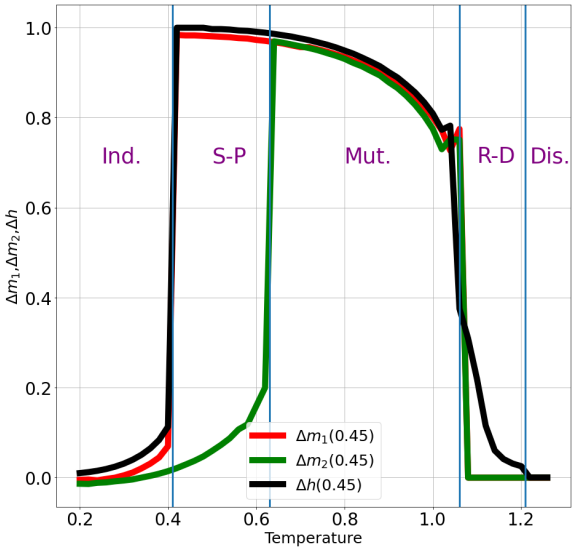


Diagrama 26, esta imagem é para ser comparada com o diagrama 21. Parâmetros: $(t_1, t_2, \alpha, \epsilon) = (0, 10, 0.08, 0.45)$. Entre $T \approx 0.4$ e $T \approx 0.62$ temos a fase de professor-estudante, já que somente m_1 tem melhoras significativas. Conforme aumentamos mais a interação. m_2 também mehora significativamente e entramos na fase de mutualismo. Por último, por volta de $T \approx 1.05$, somente Δh tem valores diferente de 0, o que corresponde à fase de ilusão reforçada.

É importante também fazer uma breve discussão sobre a questão de diferenciar fase paramagnética de fase de vidro de spin na região desordenada. Em sistemas mais usuais, isto pode ser feito de maneira relativamente simples ao olhar para o parâmetro de Edwards-Anderson q_{EA}, se ele é não-nulo, estamos em uma fase de vidro de spin, senão, estamos na paramagnética.

Porém, no modelo de *Reinforcement and Removal*, isto é um problema, pois se $t \neq 0$, temos que $Q, q \neq q_{EA}$, e não existe maneira clara de se proceder. No nosso sistema, temos ainda duas dificuldades extras. A primeira é que a estrutura energética tem uma forma curiosa e pouco clara, pois não temos só uma parte usual que forma um sistema de vidro de spin, mas também existe uma componente que reforça estados similares nas duas redes, assim não é claro como a ergodicidade do sistema se modifica; a fim de exemplo, considere a estrutura do sistema com $\epsilon >> 1$, são mínimos igualmente espaçados em relação aos spins com ligeiras variações. A outra dificuldade é que, considerando o sistema como uma rede de $2N$ neurônios, só temos fase paramagnética quando os dois estão na fase paramagnética, já que se um estiver na fase de vidro de spin não há mais ergodicidade, o que complica ainda mais a análise.

Considerando isso junto com o fato que estamos mais interessados na variações das magnetizações e parâmetro h, chamamos a fase onde $m_1 = m_2 = h = 0$ simplesmente de desordenada, e notamos que é interessante tentar entender melhor essa questão, principalmente em relação à estrutura energética.

Chapter 5

Interpretação dos Resultados e verificação

5.1 Introdução

Os resultados por si só não dizem a história do sistema. É necessária uma interpretação aprofundada e comparações com situações externas para realmente entendermos os possíveis significados do sistema e toda sua complexidade. Em geral, chegamos à conclusão de que, apesar de ser claramente um *toy model* que não pode ser aplicado diretamente em dados do mundo real, ele indica vários comportamentos e fenômenos particulares que realmente ocorrem.

Também interpretamos as comparações que fizemos com simulações de Monte Carlo do nosso modelo novo. Apesar de elas indicarem que o modelo funciona significativamente bem em quase todo o espaço de fase, em certas regiões parece que ele não captura todos os acontecimentos. Fazemos hipóteses sobre esse problema, e infelizmente não chegamos a uma solução clara. Mesmo assim, a nossa solução parece ser muito boa e certamente indica novas propriedade e características que geralmente são ausentes de modelos de redes neurais. No capítulo 6 exibimos as conclusões no contexto mais geral, relacionando os resultados apresentados nesse capítulo às outras partes.

5.2 Interpretação dos resultados

Repetimos para evitar confusão que t_1 se refere ao menor valor de t e t_2 ao maior em todos os gráficos e diagramas.

Em geral, os diagramas da seção 4.3 podem ser divididos em dois grupos: situações onde a diferença entre os t_i é grande o suficiente para que existam regiões com interação ϵ não desprezível onde os dois indivíduos tem diferenças no comportamento; e situações onde ela não é, e os dois indivíduos tem o mesmo comportamento seja com $\epsilon = 0.1$ seja com $\epsilon = 0.4$. Clarificamos que quando dizemos que os t tem valores com diferença significativa, dizemos no sentido que indivíduos não interagentes tem comportamento muito distinto com esses valores de t, e não a diferença no sentido comum $t_2 - t_1$. Como exemplo, notamos que o comportamento para $t = 4$ é mais parecido com o comportamento para $t = 10$ do que com o comportamento para $t = 0$.

Feita essa categorização, alguns comportamentos são notáveis:

1) Quando a diferença entre t_1 e t_2 é grande o suficiente, a introdução de uma interação com temperaturas não baixas beneficia de maneira significativa os dois indivíduos, porém para temperaturas menores o indivíduo com maior valor de t acaba tendo uma menor capacidade do

que tem sem interação. Isso ocorre nos diagramas 2, 3 e 5 por exemplo, e é mais prominente quando a diferença é maior.

É possível entender isso como uma diferença da flexibilidade dos indivíduos: em temperaturas altas o sistema com maior t tem menor flexibilidade, assim quem muda para encontrar uma concordância é o indivíduo com menor t, mas em temperaturas baixa isso se inverte. De um ponto de vista social, podemos entender que em situações onde é mais difícil armazenar informação (T alto) a autoridade do indivíduo com maior valor de t é pouco desafiada, enquanto que em situações onde é mais fácil armazenar informação (T baixo) o indivíduo com menor t tem mais liberdade e acaba tendo mais dominância na interação.

2) Relacionado a esse fenômeno, vemos que nas situações da primeira categoria os indivíduos com t_2 mudam basicamente nada com interação fraca ($\epsilon = 0.1$), diferentemente dos indivíduos t_1. Isso nos lembra uma situação de professor-aluno, onde com poucas interações somente o aluno aprende, o professor precisa de uma interação mais aprofundada para aprender.

3) As situações na segunda categoria nos lembram mais uma situação de colegas, pois já com interação fraca os dois aprendem juntos, e com exceção de temperaturas muito baixas no diagrama 5 não temos situações onde o indivíduo t_2 perde capacidade de interpretação com o aumento de ϵ.

Olhando para os diagramas da seção 4.4, vemos em geral que os comportamentos vistos no diagramas anteriores são confirmados. De novo, os dividimos em duas categorias: diagramas que não tem regiões onde a capacidade crítica dos indivíduos é diferente (13 a 18) e os que tem (o resto). A segunda categoria confirma o comportamento de colegas de indivíduos com baixa diferença. A capacidade somente aumenta conforme a interação se intensifica, até convergir em um certo valor que vai depender de t_1 e t_2. Aqui notamos que a capacidade não aumenta indefinidamente conforme aumentamos a interação, o que não é óbvio já que os limites comumente estabelecidos em redes neurais são válidos somente para Hamiltonianos de Ising, o que não é nosso caso. Também verificamos que lugares onde era impossível ter magnetização antes com interação suficiente é possível ter, em particular nos referimos à região com $\beta \leq 1$ (diagramas 7,8 e 9). Porém, não encontramos lugares onde o limite geral $\alpha_c = 1$ fosse ultrapassado de maneira mais significativa do que a quebra de réplicas do modelo sem interação passa. Para os casos da segunda categoria, vemos mais claramente como a interação pode ser benéfica: no diagrama 13, a capacidade do indivíduo menor mais que triplica, e a capacidade total alcançada para ϵ grande é maior do que a soma das capacidade sem interação dos dois indivíduos, mostrando que juntos eles armazenam melhor a informação do que separados. Porém, as situações de interação maléfica vistas anteriormente também são confirmadas (diagrama 18). Nele vemos que, apesar do indivíduo t_1 aumentar a capacidade crítica, em aproximadamente $\epsilon = 0.1$ o indivíduo t_2 sofre uma queda brusca, que lembra uma transição de primeira ordem, para então estabilizar em um certo valor. Nesse caso, o indivíduo t_2 teve flexibilidade suficiente para perder capacidade crítica.

Por último,discutimos os diagramas da seção 4.5. Neles é mais claro perceber as diferentes fases e como elas se modificam conforme variamos a magnitude da interação. Como visto no outros diagramas, a fase de insuficiência é relativamente pequena e ocorre somente quando a diferença entre t_1 e t_2 é significativa, desaparecendo geralmente por volta de $\epsilon \approx 0.06$. Enquanto isso, as fases indiferente, professor-estudante e de mutualismo ocupam boa parte do espaço, sendo que ocupam mais se a capacidade é fixada em um valor menor. A fase de ilusão reforçada é uma fase que não aparece de forma tão clara nos outros tipos de diagrama, e é interessante ver que ela ocupa um espaço significativo, fazendo a analogia que existem várias regiões em que os agentes podem reforçar as crenças erradas dos outros, principalmente quando temos muitos

padrões, ou seja, muita informação a ser processada.

Esse conjunto de características são interessantes pois são não triviais e plurais, em particular não temos somente relações benéficas, mas sim relações que podem atrapalhar, sendo mais similares com relações de aprendizado no dia-a-dia. A questão de dependência epistêmica entra nisso nos sentido que o aumento de aprendizado conjunto depende de algumas características, em particular que no mínimo uma das partes confie suficientemente na outra para ocorrer alguma evolução. Observamos que a existência de lugares onde α_c decresce com ϵ nos aproxima de volta ao *unlearning* hebbiano original, onde, diferentemente do modelo do algoritmo de *reinforcement and removal*, temos valores ótimos de ϵ, e aumentar ϵ pode ser muito deletério para a capacidade de um sistema.

5.3 Comparação com simulações de Monte Carlo

Reproduzir todos os resultados com precisão por Monte Carlo é uma tarefa que requer um tempo muito maior do tempo de obtenção dos resultados, pelo fato de que é necessário usar redes de tamanhos muito grandes, utilizar algoritmos sofisticados e fazer muitos passos de Monte Carlo para realmente ver todos os detalhes. Também notamos que temos que fazer a média sobre a desordem, o que demanda mais tempo e dificulta ficar mudando muito α como fizemos nas soluções numéricas. As simulações foram feitas com o algoritmo de Metropolis, e em particular fixamos os valores de $(\alpha, \beta, t_1, t_2)$ e vemos como a magnetização varia conforme mudamos ϵ.

Fases	
Passo de MC para a termalização	4000
Passos por amostra por configuração particular de desordem	4000
Amostras por configuração de desordem	30
Configurações de desordem	10
Número de spins	625-1225

Apesar de vermos poucos casos, acreditamos que eles conseguem nos dar uma ideia dos fenômenos que o nosso modelo consegue capturar ou não, e encontramos uma boa concordância entre as simulações e as soluções numéricas para vários valores dos parâmetros, como pode ser constatado nas imagens a seguir:

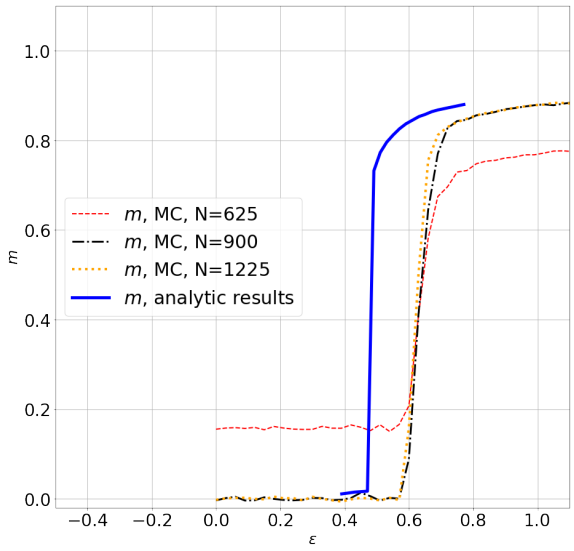


Diagrama 27. Parâmetros: $(t_1, t_2, \beta, \alpha) = (10, 10, 0.95, 0.01)$.Este diagrama e o próximo não só mostram boa concordância com os resultados analíticos e numéricos, como também indicam que aumentar o número de spins leva a uma melhor precisão, que é o esperado. Notamos que mesmo assim existe uma pequena diferença entre os pontos em que a solução analítica e numérica começam a ter magnetização macroscópica.

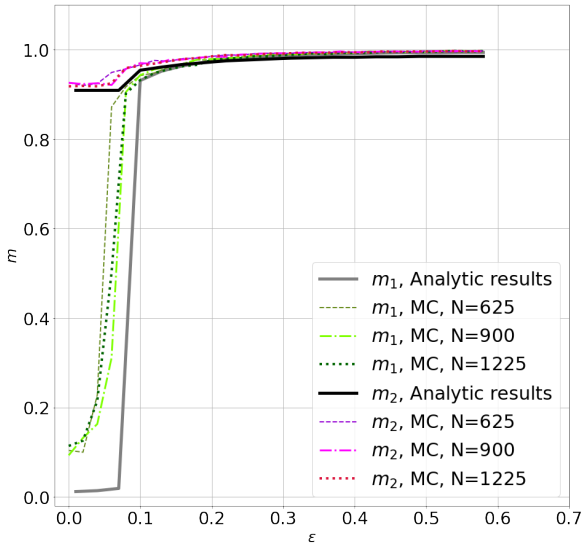


Diagrama 28. Parâmetros: $(t_1, t_2, \beta, \alpha) = (0, 10, 2, 0.1)$.Este diagrama e o próximo mostram uma concordância ainda maior, no caso, estamos considerando um caso onde os valores de t_1 e t_2 são significativamente diferentes.

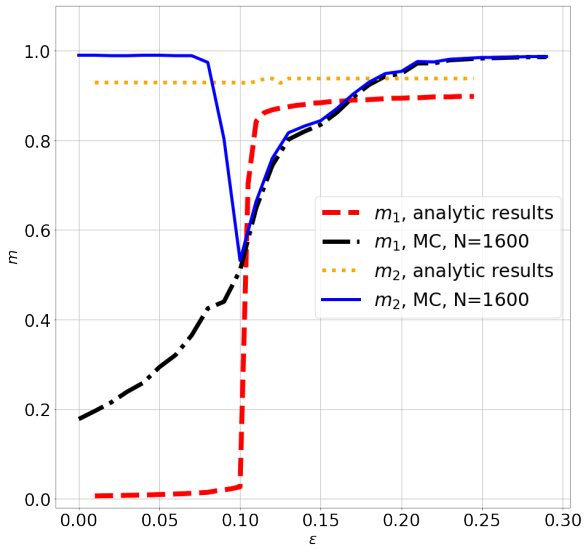


Diagrama 29. Parâmetros: $(t_1, t_2, \beta, \alpha) = (0.5, 10, 5, 0.45)$. Embora em boa parte desse diagrama e do próximo parece ter uma certa concordância, a solução analítica parece perder um comportamento qualitativo relevante: o vale observado em m_2 entre 0.1 e 0.15 ϵ. Apesar desse fenômenos não ser ausente em todos as situações analíticas, em geral o Monte Carlo prevê que ele ocorra muito mais frequentemente do que a solução analítica indica.

Encontramos uma situação estranha em temperaturas baixas e em lugares onde a diferença entre t é significativa: parece que, depois da magnetização diminuir nos lugares esperados, ela aumenta se aumentarmos ainda mais o valor de ϵ, criando a figura a seguir

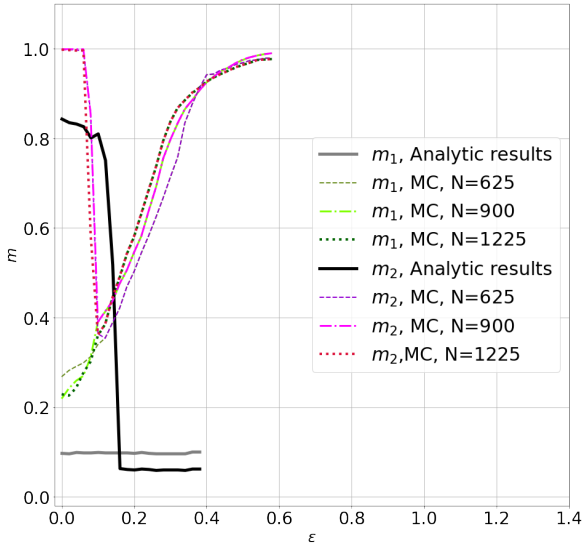


Diagrama 30. Parâmetros: $(t_1, t_2, \beta, \alpha) = (0, 3, 10, 0.5)$.Vemos uma discordância significativa entre as soluções, já que a queda das magnetizações conforme aumentamos ϵ observada na solução analítica simplesmente não bate com o vale observado na solução de Monte Carlo.

A primeira explicação para essa diferença entre os resultados foi que nessa região o Ansatz usado para a réplicas não é válido, e teríamos que quebrar a simetria entre as réplicas para obter algo mais preciso, já que esta é uma região onde se esperaria o fenômeno de quebra de réplicas se ele acontecesse.

A segunda explicação parte de que notamos que o número de passos de Monte Carlo para termalizar o sistema aumenta de maneira exponencial quando aumentamos o ϵ além de certo valor, e em particular esse fenômeno é mais nítido em temperatura baixas. A intuição para esse fenômeno é clara: um valor de ϵ alto aprofunda os mínimos relacionados aos padrões, e sair dos padrões acaba levando muito mais tempo. Um exemplo desse aumento é a seguinte figura:

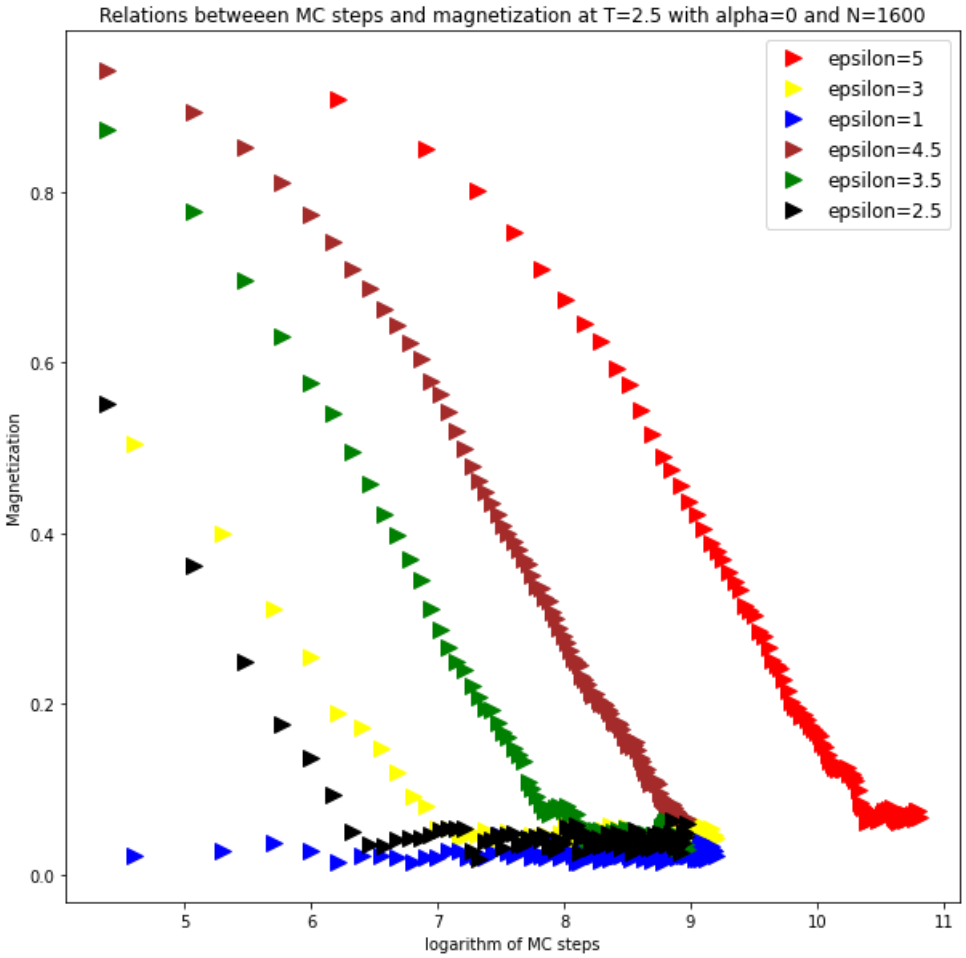


Diagrama 30

Mesmo assim, aumentar os passos de termalização em alguns casos não foi o suficiente. Em relação a isso, pode ser possível também que o algoritmo de Metropolis não seja adequado nesta região, já que em baixas temperaturas ele tem uma convergência lenta.

No capítulo 6, relacionado às conclusões, discutimos esse problema no contexto mais geral e porque acreditamos que ele não invalida nosso trabalho.

Chapter 6

Conclusão

O nosso modelo tenta sintetizar várias ideias em torno de *unlearning*, ao mesmo tempo que tentamos entender maneiras em que as redes podem ser melhoradas ou modificadas ao interagir com outras redes e tentamos entender estes fenômenos de um ponto de vista social. Diferentemente de outros modelos ou algoritmos, pudemos encontrar uma solução analítica e não depender somente de simulações, levando à possibilidade de analisar vários comportamentos possíveis com certa precisão que com simulações não seria possível. Notamos porém que verificar a validade da solução requer um esforço analítico muito forte, sendo uma ordem de grandeza além das contas necessárias para encontrar a solução; assim, notando que o modelo que serviu de base exibe uma solução estável em quase todo o espaço, supomos que pelo menos para baixas interações o nosso modelo consegue capturar os fenômenos existentes de maneira razoável. Ao comparar com simulações de Monte Carlo, vimos que em quase todo o espaço encontramos boa concordância, indicando que nessas regiões a nossa solução é adequada. Em algumas regiões infelizmente encontramos discordância significativas, porém sendo essa região relativamente pequena, acreditamos que esse fato não invalida nossa solução e os fenômenos que conseguimos encontrar dela.

Evidenciamos algumas conclusões que acreditamos serem relevantes: na grande maioria das situações, a introdução da interação aumenta significativamente a capacidade de pelo menos uma das redes de armazenar informação de maneira adequada, não só podendo transformar uma rede inicialmente pouca capacitada em uma altamente capacitada, mas indo além dos limites que existem para redes individuais, o que significa que treinar redes por meio de interações pode ser mais potente do que treinar redes individualmente de maneiras usuais. O fato de que interações podem levar a quedas na capacidade crítica em algumas regiões é inesperado e interessante, e pode dar uma direção em como agentes inicialmente capacitados podem chegar a uma situação piorada, apesar de ter muita informação disponível.

Notamos que as nossas interpretações dos fenômenos do sistema como analogias a interações sociais de aprendizado são vagas e não quantitativas, porém servem para dar sentido à história do sistema e enfatizar alguns pontos; é de se pensar que sistemas mais simples com estas ideias talvez consigam obter resultados quantitativos significativos.

Há algumas modificações, generalizações e análises adicionais que podem ser feitas no nosso modelo, e o principal passo é tentar entender em quais sistemas isso é possível. Além de redes de tipo Hopfield, uma interação com esta forma pode ser interessante, e seria interessante testar isso em redes que são mais realistas biologicamente; além disso, nossa interação conecta todos os neurônios da rede 1 com todos da rede 2, é válido perguntar quais mudanças ocorrem se introduzirmos uma interação de primeiros vizinhos.

É possível introduzir um terceiro agente e mais duas interações no nosso modelo e calcular analiticamente a energia livre. A análise das equações se torna muito mais complicada do que a análise aqui apresentada, porém os resultados seriam provavelmente interessantes e não triviais e talvez possam sugerir extensões para mais agentes. Também é possível avançar na análise da estrutura energética, separando os estados de vidro de spin dos estados paramagnéticos.

Chapter 7

Apêndice

7.1 Equações de consistência

Minimizando a energia a livre em relação a 15 variáveis, conseguimos 15 equações, das quais 9 são independentes.

As equações auxiliares (i.e., de variáveis não independentes) são

$$r^{\sigma} = \{q^{\sigma}(1+t_1)^2[1-\beta(Q^S-q^S)(1+t_2)]^2 \tag{7.1a}$$
$$+\beta(Q^{\sigma S}-q^{\sigma S})(1+t_1)^2(1+t_2)[q^{\sigma S}(1-\beta(Q^S-q^S)(1+t_2))+q^S(1+t_2)\beta(Q^{\sigma S}-q^{\sigma S})]\}$$
$$\times\{[1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1)][1-\beta(Q^S-q^S)(1+t_2)]-(Q^{\sigma S}-q^{\sigma S})^2(1+t_1)(1+t_2)\beta^2\}^{-2},$$

$$\Delta^{\sigma} = 1+\alpha t_1\{((1-\beta(Q^S-q^S)(1+t_2))^2(1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1)) \tag{7.1b}$$
$$-(1-\beta(Q^S-q^S)(1+t_2))(Q^{\sigma S}-q^{\sigma S})^2\beta^2(1+t_1)(1+t_2))\}$$
$$\times\{[1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1)][1-\beta(Q^S-q^S)(1+t_2)]-(Q^{\sigma S}-q^{\sigma S})^2(1+t_1)(1+t_2)\beta^2\}^{-2},$$

$$r^{S} = (q^S(1+t_2)^2(1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1))^2 \tag{7.1c}$$
$$+\beta(Q^{\sigma S}-q^{\sigma S})(1+t_1)(1+t_2)^2(q^{\sigma S}(1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1))+q^{\sigma}(1+t_1)\beta(Q^{\sigma S}-q^{\sigma S})))$$
$$\times\{[1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1)][1-\beta(Q^S-q^S)(1+t_2)]-(Q^{\sigma S}-q^{\sigma S})^2(1+t_1)(1+t_2)\beta^2\}^{-2},$$

$$\Delta^{S} = 1+(\alpha t_2((1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1))^2(1-\beta(Q^S-q^S)(1+t_2)) \tag{7.1d}$$
$$-(1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1))(Q^{\sigma S}-q^{\sigma S})^2\beta^2(1+t_1)(1+t_2)))$$
$$\times\{[1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1)][1-\beta(Q^S-q^S)(1+t_2)]-(Q^{\sigma S}-q^{\sigma S})^2(1+t_1)(1+t_2)\beta^2\}^{-2},$$

$$\Delta^{\sigma S} = \alpha\beta(-q^{\sigma S}+Q^{\sigma S})(1+t_1)(1+t_2) \tag{7.1e}$$
$$\times\{2[1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1)][1-\beta(Q^S-q^S)(1+t_2)]-2\beta^2(-Q^{\sigma S}+q^{\sigma S})^2(1+t_1)(1+t_2)\}^{-1},$$

$$r^{\sigma S} = (1+t_1)(1+t_2)\{q^{\sigma S} \tag{7.1f}$$
$$\times[(1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1))(1-\beta(Q^{S}-q^{S})(1+t_2))-\beta^2(Q^{\sigma S}-q^{\sigma S})^2(1+t_1)(1+t_2)]^{-1}$$
$$+2[(-q^{\sigma S}+Q^{\sigma S})\beta(q^{\sigma}(1+t_1)(1-\beta(Q^{S}-q^{S})(1+t_2))$$
$$+q^{S}(1+t_2)(1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1))+q^{\sigma S}(1+t_1)(1+t_2)\beta(-q^{\sigma S}+Q^{\sigma S}))]$$
$$\times[(1-\beta(Q^{\sigma}-q^{\sigma})(1+t_1))(1-\beta(Q^{S}-q^{S})(1+t_2))-\beta^2(Q^{\sigma S}-q^{\sigma S})^2(1+t_1)(1+t_2)^2]^{-1}\}.$$

As equações de auto-consistência independentes são

$$h=\int DxDy\frac{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta)-\cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta)+\cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}, \tag{7.1g}$$

$$m_1(\frac{1}{1+t_1}-\frac{\Delta^S t_1'^2\beta}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2})=\frac{m_2t_1'^2t_2'^2\beta^2\Delta^{\sigma S}}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2} \tag{7.1h}$$
$$+\int DxDy\{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta)+\cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)\}^{-1}$$
$$\times\{\frac{\Delta^S}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2}[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)+\sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$+\frac{t_1'^2\beta\Delta^{\sigma S}}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2}[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)-\sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]\},$$

$$m_2(\frac{1}{1+t_2}-\frac{\Delta^{\sigma} t_2'^2\beta}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2})=\frac{m_1t_1'^2t_2'^2\beta^2\Delta^{\sigma S}}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2} \tag{7.1i}$$
$$+\int DxDy\{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta)+\cosh[\beta(\upsilon-\Upsilon)]\exp(-\eta)\}^{-1}$$
$$\times\{\frac{\Delta^{\sigma S}t_2'^2\beta}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2}[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)+\sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$+\frac{\Delta^{\sigma}}{\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2}[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)-\sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]\},$$

$$\frac{Q^{\sigma}(1+t_1)}{2t_1}=-\frac{\Delta^S}{2\beta\Delta^{\sigma}\Delta^S-2t_1'^2t_2'^2\beta^3(\Delta^{\sigma S})^2}-[2(\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2)^2]^{-1} \tag{7.1j}$$
$$\times\{(\Delta^S)^2t_1'^2\beta[\alpha r^{\sigma}+m_1^2+\frac{1}{\beta^2t_1'^4}]+t_1'^2t_2'^4\beta^3(\Delta^{\sigma S})^2[\alpha r^S+m_2^2+\frac{1}{\beta^2t_2'^4}]$$
$$+2t_1'^2t_2'^2\beta^2\Delta^{\sigma S}\Delta^S[\frac{r^{\sigma S}\alpha}{2}+mM]\}+\int DxDy$$
$$\times\{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta)+\cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta))[\Delta^{\sigma}\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2]^2\}^{-1}$$
$$\times\{[-(\Delta^S)^2(m_1+\sqrt{\frac{\alpha r^{\sigma}}{2}}(r_+x+r_-y))-t_2'^2\beta\Delta^S\Delta^{\sigma S}(m_2+\sqrt{\frac{\alpha r^S}{2}}(r_+x-r_-y))]$$
$$\times[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)+\sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$-[t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2(m_2+\sqrt{\frac{\alpha r^S}{2}}(r_+x-r_-y))+t_1'^2\beta\Delta^S\Delta^{\sigma S}(m_1+\sqrt{\frac{\alpha r^{\sigma}}{2}}(r_+x+r_-y))]$$
$$\times[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)-\sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$-\Delta^S\Delta^{\sigma S}[\cosh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)-\cosh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]\},$$

$$\frac{Q^S(1+t_2)}{2t_2} = -\frac{\Delta^\sigma}{2\beta\Delta^\sigma\Delta^S - 2t_1'^2t_2'^2\beta^3(\Delta^{\sigma S})^2} - \{2[\Delta^\sigma\Delta^S - t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2]^2\}^{-1} \quad (7.1k)$$
$$\times\{(\Delta^\sigma)^2t_2'^2\beta[\alpha r^S + m_2^2 + \frac{1}{\beta^2t_2'^4}] + t_2'^2t_1'^4\beta^3(\Delta^{\sigma S})^2[\alpha r^\sigma + m_1^2 + \frac{1}{\beta^2t'^4}]$$
$$+2t_1'^2t_2'^2\beta^2\Delta^{\sigma S}\Delta^\sigma[\frac{r^{\sigma S}\alpha}{2} + m_1m_2]\} + \int DxDy$$
$$\times\{(\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta))[\Delta^\sigma\Delta^S - t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2]^2\}^{-1}$$
$$\times\{[-(\Delta^\sigma)^2(m_2 + \sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y)) - t_1'^2\beta\Delta^\sigma\Delta^{\sigma S}(m_1 + \sqrt{\frac{\alpha r^S}{2}}(r_+x + r_-y))]$$
$$\times[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) - \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)] - [t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2$$
$$\times(m_1 + \sqrt{\frac{\alpha r^\sigma}{2}}(r_+x + r_-y)) + t_2'^2\beta\Delta^\sigma\Delta^{\sigma S}(m_2 + \sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y))]$$
$$\times[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) + \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$-\Delta^\sigma\Delta^{\sigma S}[\cosh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) - \cosh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]\},$$

$$Q^{\sigma S} = \frac{t_1'^2t_2'^2\beta\Delta^{\sigma S}}{\Delta^\sigma\Delta^S - t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2} + \frac{t_1'^2t_2'^2\beta^2(\frac{r^{\sigma S}}{2} + mM)}{\Delta^\sigma\Delta^S - t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2} \quad (7.1l)$$
$$+\{t_1'^2t_2'^2\beta^2\Delta^{\sigma S}[\Delta^St_1'^2\beta(\alpha r^\sigma + m_1^2 + \frac{1}{\beta^2t_1'^4})$$
$$+\Delta^\sigma t_2'^2\beta(\alpha r^S + m_2^2 + \frac{1}{\beta^2t_2'^4}) + 2t_1'^2t_2'^2\beta^2\Delta^{\sigma S}(\frac{r^{\sigma S}\alpha}{2} + mM)]\}$$
$$\times(\Delta^\sigma\Delta^S - t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2)^{-2} + \int DxDy\{[\cosh(\beta(\upsilon+\Upsilon))\exp(\beta\eta)$$
$$+\cosh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)][\Delta^\sigma\Delta^S - t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2]^2\}^{-1}$$
$$\times\{[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) + \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$\times[2t_1'^2t_2'^2\beta^2\Delta^{\sigma S}\Delta^S(m_1 + \sqrt{\frac{\alpha r^\sigma}{2}}(r_+x + r_-y))$$
$$+(t_1'^2t_2'^4\beta^3(\Delta^{\sigma S})^2 + t_2'^2\beta\Delta^\sigma\Delta^S)(m_2 + \sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y))]$$
$$+[2t_1'^2t_2'^2\beta^2\Delta^{\sigma S}\Delta^\sigma(m_2 + \sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y))$$
$$+(t_1'^2\beta\Delta^\sigma\Delta^S + t_1'^4t_2'^2\beta^3(\Delta^{\sigma S})^2)(m_1 + \sqrt{\frac{\alpha r^\sigma}{2}}(r_+x + r_-y)]$$
$$\times[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) - \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)] + (\Delta^\sigma\Delta^S + t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2)$$
$$\times[\cosh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) - \cosh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]\},$$

$$q^{\sigma} = Q^{\sigma} - \frac{\Delta^S t_1'^2}{\Delta^{\sigma}\Delta^S - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2} - \frac{1}{\alpha\beta}\int DxDy \tag{7.1m}$$
$$\times\{[\cosh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) + \cosh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)](4\Delta^{\sigma}\Delta^S - 4t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2)\}^{-1}$$
$$\times\{(\sinh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \sinh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)\}\{\Delta^S[4\sqrt{\frac{\alpha}{2r^{\sigma}}}(r_+x + r_-y)$$
$$+\frac{r^{\sigma S}}{r^{\sigma}}\sqrt{\frac{\alpha}{2r^S}}(-\frac{x}{r_+} + \frac{y}{r_-})] - r^{\sigma S}t_2'^2\beta\Delta^{\sigma S}\sqrt{\frac{\alpha}{2(r^{\sigma})^3}}[\frac{x}{r_+} + \frac{y}{r_-}]\}$$
$$+[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) - \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$\times[t_1'^2\beta\Delta^{\sigma S}(4\sqrt{\frac{\alpha}{2r^{\sigma}}}(r_+x + r_-y) + \frac{r^{\sigma S}}{r^{\sigma}}\sqrt{\frac{\alpha}{2r^S}}(-\frac{x}{r_+} + \frac{y}{r_-})) - r^{\sigma S}\Delta^{\sigma}\sqrt{\frac{\alpha}{2(r^{\sigma})^3}}(\frac{x}{r_+} + \frac{y}{r_-})],$$

$$q^S = Q^S - \frac{\Delta^{\sigma} t_2'^2}{\Delta^{\sigma}\Delta^S - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2} - \frac{1}{\alpha\beta}\int DxDy \tag{7.1n}$$
$$\times\{[\cosh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) + \cosh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)](4\Delta^{\sigma}\Delta^S - 4t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2)\}^{-1}$$
$$\times\{[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) - \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$\times[\Delta^{\sigma}(4\sqrt{\frac{\alpha}{2r^S}}(r_+x - r_-y) - \frac{r^{\sigma S}}{r^S}\sqrt{\frac{\alpha}{2r^{\sigma}}}(\frac{x}{r_+} + \frac{y}{r_-})) + r^{\sigma S}t_1'^2\beta\Delta^{\sigma S}\sqrt{\frac{\alpha}{2(r^S)^3}}(-\frac{x}{r_+} + \frac{y}{r_-})]$$
$$+[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) + \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$\times[t_2'^2\beta\Delta^{\sigma S}(4\sqrt{\frac{\alpha}{2r^S}}(r_+x - r_-y) - \frac{r^{\sigma S}}{r^S}\sqrt{\frac{\alpha}{2r^{\sigma}}}(\frac{x}{r_+} + \frac{y}{r_-})) + r^{\sigma S}\Delta^S\sqrt{\frac{\alpha}{2(r^S)^3}}(-\frac{x}{r_+} + \frac{y}{r_-})]\},$$

$$q^{\sigma S} = Q^{\sigma S} - \frac{t_1'^2 t_2'^2\beta\Delta^{\sigma S}}{\Delta^{\sigma}\Delta^S - t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2} - \frac{1}{\alpha\beta}\int DxDy \tag{7.1o}$$
$$\times\{2[\cosh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) + \cosh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)][\Delta^{\sigma}\Delta^S - t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2]\}^{-1}$$
$$+\{[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) + \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$\times[\Delta^S\sqrt{\frac{\alpha}{2r^S}}(\frac{x}{r_+} - \frac{y}{r_-}) + t_2'^2\beta\Delta^{\sigma S}\sqrt{\frac{\alpha}{2r^{\sigma}}}(\frac{x}{r_+} + \frac{y}{r_-})]$$
$$+[\sinh(\beta(\upsilon+\Upsilon))\exp(\beta\eta) - \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)]$$
$$\times[\Delta^{\sigma}\sqrt{\frac{\alpha}{2r^{\sigma}}}(\frac{x}{r_+} + \frac{y}{r_-}) + t_1'^2\beta\Delta^{\sigma S}\sqrt{\frac{\alpha}{2r^S}}(\frac{x}{r_+} - \frac{y}{r_-})]\}.$$

Para $\beta \to \infty$, 2 modificações são necessárias: trocar q^{ρ} por $c^{\rho} \equiv \beta(Q^{\rho} - q^{\rho})$, de maneira similar ao que foi feito em [8], e trocar as combinações de funções trigonométricas hiperbólicas por combinações de funções sinal. Este segundo processo é feito explicitamente no apêndice 2.

Temos 3 combinações distintas:

$$I_1 = \frac{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) - \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}, \tag{7.2}$$
$$I_2 = \frac{\sinh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \sinh(\beta(\upsilon-\Upsilon))\exp(-\beta\eta)}{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)},$$
$$I_3 = \frac{\sinh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) - \sinh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}.$$

A primeira combinação se torna

$$
\begin{aligned}
I_1 = \frac{1}{4}\{&[\mathrm{sign}(\upsilon+\Upsilon)+1][\mathrm{sign}(\upsilon-\Upsilon)+1]\,\mathrm{sign}(\eta+\Upsilon) \\
&+[\mathrm{sign}(\upsilon+\Upsilon)+1][-\mathrm{sign}(\upsilon-\Upsilon)+1]\,\mathrm{sign}(\eta+\upsilon) \\
&+[-\mathrm{sign}(\upsilon+\Upsilon)+1][\mathrm{sign}(\upsilon-\Upsilon)+1]\,\mathrm{sign}(\eta-\upsilon) \\
&+[-\mathrm{sign}(\upsilon+\Upsilon)+1][-\mathrm{sign}(\upsilon-\Upsilon)+1]\,\mathrm{sign}(\eta-\Upsilon)\}.
\end{aligned} \tag{7.3}
$$

A segunda combinação se torna

$$
\begin{aligned}
I_2 = \frac{1}{4}\{&(\mathrm{sign}(\upsilon+\Upsilon)+1)(\mathrm{sign}(\upsilon-\Upsilon)+1) \\
&+[\mathrm{sign}(\upsilon+\Upsilon)+1][-\mathrm{sign}(\upsilon-\Upsilon)+1]\,\mathrm{sign}(\eta+\upsilon) \\
&+(-\mathrm{sign}(\upsilon+\Upsilon)+1][\mathrm{sign}(\upsilon-\Upsilon)+1]\,\mathrm{sign}(\upsilon-\eta) \\
&-[-\mathrm{sign}(\upsilon+\Upsilon)+1][-\mathrm{sign}(\upsilon-\Upsilon)+1]).
\end{aligned} \tag{7.4}
$$

A terceira combinação se torna

$$
\begin{aligned}
I_3 = \frac{1}{4}\big(&(\mathrm{sign}(\upsilon+\Upsilon)+1)(\mathrm{sign}(\upsilon-\Upsilon)+1)\,\mathrm{sign}(\eta+\Upsilon) \\
&+(-\mathrm{sign}(\upsilon+\Upsilon)+1)(-\mathrm{sign}(\upsilon-\Upsilon)+1)\,\mathrm{sign}(\Upsilon-\eta) \\
&+(\mathrm{sign}(\upsilon+\Upsilon)+1)(-\mathrm{sign}(\upsilon-\Upsilon)+1)-(-\mathrm{sign}(\upsilon+\Upsilon)+1)(\mathrm{sign}(\upsilon-\Upsilon)+1)\big).
\end{aligned} \tag{7.5}
$$

É possível eliminar uma das integrais com estas funções sinais, mas não vale o esforço e não diminui o tempo computacional para resolvê-las de maneira significativa. Não encontramos nenhum fenômeno particularmente interessante nessa região, por isso focamos no comportamento geral.

7.2 A energia livre e as equações de consistência resultantes

O nosso hamiltoniano é

$$
H_{tot} = H^A + H^B + H_{int}, \tag{7.6}
$$

sendo H^A e H^B os hamiltonianos usuais dos indivíduos e H_{int} um Hamiltoniano de interação,

$$
H_{int} \equiv -\frac{\epsilon}{N}\sum_{ij}\sigma_i S_j S_i \sigma_j. \tag{7.7}
$$

$\Longrightarrow$

$$
\begin{aligned}
Z = \sum_{\sigma,S}\exp[&\frac{\beta}{2N}\sum_{i,j}\sum_{\mu,\nu}\xi_i^{\mu A}\xi_j^{\nu A}(\frac{1+t_1}{1+t_1 C})_{\mu\nu A}\sigma_i\sigma_j \\
&+\frac{\beta}{2N}\sum_{i,j}\sum_{\mu,\nu}\xi_i^{\mu B}\xi_j^{\nu B}(\frac{1+t_2}{1+t_2 C})_{\mu\nu B}S_i S_j + \beta\epsilon N(\frac{1}{N}\sum_i \sigma_i S_i)^2].
\end{aligned} \tag{7.8}
$$

Assim, temos que

$$\langle Z^n \rangle = \langle \sum_{\sigma^n, S^n} \exp[\frac{\beta}{2N} \sum_a \sum_{i,j} \sum_{\mu,\nu} \xi_i^{\mu A} \xi_i^{\nu A} (\frac{1+t_1}{\mathbb{1}+t_1 C})^A_{\mu\nu} \sigma_i^a \sigma_j^a \tag{7.9}$$
$$+ \frac{\beta}{2N} \sum_a \sum_{i,j} \sum_{\mu,\nu} \xi_i^{\mu B} \xi_j^{\nu B} (\frac{1+t_2}{\mathbb{1}+t_2 C})^A_{\mu\nu} S_i^a S_j^a + \sum_a \beta\epsilon N (\frac{1}{N} \sum_i \sigma_i^a S_i^a)^2] \rangle.$$

Por simplicidade, consideramos que os padrões de A e B são os mesmos, então

$$\langle Z^n \rangle = \langle \sum_{\sigma^n, S^n} \exp[\frac{\beta}{2N} \sum_a \sum_{i,j} \sum_{\mu,\nu} \xi_i^{\mu} \xi_i^{\nu} (\frac{1+t_1}{\mathbb{1}+t_1 C})_{\mu\nu} \sigma_i^a \sigma_j^a \tag{7.10}$$
$$+ \frac{\beta}{2N} \sum_a \sum_{i,j} \sum_{\mu,\nu} \xi_i^{\mu} \xi_j^{\nu} (\frac{1+t_2}{\mathbb{1}+t_2 C})_{\mu\nu} S_i^a S_j^a + \beta\epsilon N (\frac{1}{N} \sum_i \sigma_i^a S_i^a)^2] \rangle.$$

Comecemos mexendo na função de partição

$$Z = \sum_{\sigma,S} \exp[\frac{\beta}{2N} \sum_{i,j} \sum_{\mu,\nu} \xi_i^{\mu} \xi_j^{\nu} ((\frac{1+t_1}{\mathbb{1}+t_1 C})_{\mu\nu} \sigma_i \sigma_j + (\frac{1+t_2}{\mathbb{1}+t_2 C})_{\mu\nu} S_i S_j) + \beta\epsilon N (\frac{1}{N} \sum_i \sigma_i S_i)^2] \tag{7.11}$$

Para tratar apropriadamente estes termos, precisamos examinar de alguma maneira as matrizes inversas e linearizar os termos quadráticos. A primeira parte pode ser feita da seguinte maneira:

$$\exp(\frac{1}{2} B^T A^{-1} B) = \sqrt{\frac{\det(A)}{(2\pi)^n}} \int \exp(-\frac{1}{2} x^T A x + B^T x) d^n x; \quad \exp(\frac{\beta(1+t)}{2N} \sigma^T \xi^T (\frac{1}{\mathbb{1}+tC}) \xi\sigma)$$
$$= \int d^p z \exp(-\frac{1}{2} \sum_{\mu\nu} (\mathbb{1}+tC)_{\mu\nu} z_\mu z_\nu + \sqrt{\frac{\beta(1+t)}{N}} \sum_\mu \sum_i \xi_i^\mu \sigma_i z_\mu). \tag{7.12}$$

A segunda parte pode ser feita organizando as variáveis e aplicando uma transformação de Hubbard-Stratonovich:

$$\sum_{\mu\nu} (\mathbb{1}_{\mu\nu} + tC)_{\mu\nu} z_\mu z_\nu = \sum_\mu (z_\mu)^2 + \frac{t}{N} (\xi z)^2, \tag{7.13}$$
$$\int d^p z \exp(-\frac{1}{2} \sum_{\mu\nu} (\mathbb{1}+tC)_{\mu\nu} z_\mu z_\nu + \sqrt{\frac{\beta(1+t)}{N}} \sum_\mu \sum_i \xi_i^\mu \sigma_i z_\mu)$$
$$= \int d^p z \exp(-\frac{1}{2} \sum_\mu (z_\mu)^2 - \frac{t}{2N} (\xi z)^2 + \sqrt{\frac{\beta(1+t)}{N}} \sum_i \sum_\mu \xi_i^\mu \sigma_i z_\mu)$$
$$= \int D^p z D^N \phi \exp(i\sqrt{\frac{t}{N}} \sum_{i,\mu} \phi_i \xi_i^\mu z_\mu + \sqrt{\frac{\beta(1+t)}{N}} \sum_{i,\mu} \xi_i^\mu \sigma_i z_\mu),$$

onde usamos a medida gaussiana $Dx \equiv dx \exp(-\frac{x^2}{2})$ para simplificar a notação. Notamos que esse procedimento é o mesmo para os indivíduos 1 e 2.

Para linearizar o termo ligado à interação, fazemos também uma transformação de Hubbard-Stratonovich:

$$\exp(\beta\epsilon N(\frac{1}{N}\sum_i \sigma_i S_i)^2) = \int Dh \exp(\sqrt{2\beta\epsilon N}h(\frac{1}{N}\sum_i \sigma_i S_i)). \tag{7.14}$$

Observação: como o leitor deve ter percebido, as igualdades estão certas a menos de constantes multiplicativas. Ao longo desse apêndice as ignoramos, já que na energia livre aparecem como constantes aditivas que não tem importância para as equações que regem o sistema. Além disso, todas as integrais são de $-\infty$ a ∞, exceto quando notado.

As variáveis resultantes da inversão e linearização ligadas ao indivíduo 1 são x e ϕ e ao indivíduo 2 são y e Φ. A variável ligada à linearização da parte da interação é h. Obtemos que

$$Z = \sum_{\{\sigma,S\}} \int D^p x D^N \phi \int D^p y D^N \Phi \int Dh \exp[i\sqrt{\frac{t_1}{N}}\sum_{i,\mu}\phi_i\xi_i^\mu x_\mu + i\sqrt{\frac{t_2}{N}}\sum_{i,\mu}\Phi_i\xi_i^\mu y_\mu$$
$$+\sqrt{\frac{\beta}{N}}\sum_{i,\mu}\xi_i^\mu(\sigma_i x_\mu\sqrt{1+t_1} + S_i y_\mu\sqrt{1+t_2}) + h\sqrt{\frac{2\beta\epsilon}{N}}\sum_i \sigma_i S_i]. \tag{7.15}$$

Voltando para a média da função de partição replicada, temos que

$$\langle Z^n\rangle = \langle\prod_a \sum_{\sigma^n,S^n}\int \prod_{\mu,a} Dx_\mu^a Dy_\mu^a \prod_{i,a} D\Phi_i^a D\phi_i^a \prod_a Dh^a \tag{7.16}$$
$$\times \exp[\sqrt{\frac{\beta}{N}}\sum_{\mu,i,a}\xi_i^\mu(\sqrt{1+t_1}x_\mu^a\sigma_i^a + \sqrt{1+t_2}y_\mu^a S_i^a) + i\sqrt{\frac{1}{N}}\sum_{\mu,i,a}(x_\mu^a\phi_i^a\sqrt{t_1} + y_\mu^a\Phi_i^a\sqrt{t_2})$$
$$+\sqrt{\frac{2\beta\epsilon}{N}}\sum_a\sum_i \sigma_i^a S_i^a h^a]\rangle.$$

Vamos considerar o caso onde só temos um padrão condensado, que SPG podemos considerar ser o padrão $\mu = 1$. Este passo simplifica as contas e nos permite ter expressões fechadas, além de que esse é o caso mais interessante, já que ele que nos diz as regiões onde temos retrieval e onde não temos.

$$\langle Z^n\rangle = \sum_{\sigma^n,S^n}\int \prod_a Dx_1^a Dy_1^a \prod_{i,a} D\Phi_i^a D\phi_i^a \prod_a Dh^a \tag{7.17}$$
$$\times \exp[\sqrt{\frac{\beta(1+t_1)}{N}}\sum_{i,a} x_1^a\xi_i^1(\sigma_i^a + i\sqrt{\frac{t_1}{\beta(t_1+1)}}\phi_i^a) + \sqrt{\frac{\beta(1+t_2)}{N}}y_1^a\xi_i^1(S_i^a + i\sqrt{\frac{t_2}{\beta(t_2+1)}}\Phi_i^a)$$
$$+\sqrt{\frac{2\beta\epsilon}{N}}\sum_a\sum_i \sigma_i^a S_i^a h^a]\langle\prod_a\prod_{\mu\geq 2} Dx_\mu^a Dy_\mu^a \exp\{\sum_{i,a}\sum_{\mu\geq 2}$$
$$\times\xi_i^\mu[\sqrt{\frac{\beta(t_1+1)}{N}}x_\mu^a(\sigma_i^a + i\sqrt{\frac{t_1}{\beta(t_1+1)}}\phi_i^a) + \sqrt{\frac{\beta(t_2+1)}{N}}y_\mu^a(S_i^a + i\sqrt{\frac{t_2}{\beta(t_2+1)}}\Phi_i^a)]\}\rangle.$$

A partir de agora, usaremos a notação $T' = i\sqrt{\frac{T}{\beta(T+1)}}$ e $t' = i\sqrt{\frac{t}{\beta(t+1)}}$. Podemos fazer a média explicitamente

$$\langle \exp\{\sum_{ia}\sum_{\mu\geq 2}\xi_i^\mu[\sqrt{\frac{\beta(t_1+1)}{N}}x_\mu^a(\sigma_i^a+t_1'\phi_i^a)+\sqrt{\frac{\beta(t_2+1)}{N}}y_\mu^a(S_i^a+t_2'\Phi_i^a)]\}\rangle \tag{7.18}$$
$$= \{\cosh[\sum_{ia}\sum_{\mu\geq 2}(\sqrt{\frac{\beta(t_1+1)}{N}}x_\mu^a(\sigma_i^a+t_1'\phi_i^a)+\sqrt{\frac{\beta(t_2+1)}{N}}y_\mu^a(S_i^a+t_2'\Phi_i^a))]\}^N$$
$$= \exp\{\frac{\beta}{2N}\sum_{iab}\sum_{\mu\geq 2}(1+t_1)(\sigma_i^a+t_1'\phi_i^a)(\sigma_i^b+t_1'\phi_i^b)x_\mu^a x_\mu^b+(1+t_2)y_\mu^a y_\mu^b(S_i^a+t_2'\Phi_i^a)(S_i^b+t_2'\Phi_i^b)$$
$$+2\sqrt{(1+t_1)(1+t_2)}[(\sigma_i^a+t_1'\phi_i^a)(S_i^b+t_2'\Phi_i^b)x_\mu^a y_\mu^b]\}.$$

Na primeira igualdade fatorizamos sobre o índice i, na segunda fizemos a média explicitamente e na terceira fizemos uma expansão em Taylor, que está justificada pelo limite termodinâmico.

Usando funções delta e introduzindo as variáveis de Edward-Anderson, esta expressão se torna

$$\int\prod_{ab}dq_{ab}^\sigma dq_{ab}^S dq_{ab}^{\sigma S}\delta[q_{ab}^\sigma-\frac{1}{N}\sum_i(\sigma_i^a+t_1'\phi_i^a)(\sigma_i^b+t_1'\phi_i^b)] \tag{7.19}$$
$$\times\delta[q_{ab}^S-\frac{1}{N}\sum_i(S_i^a+t_2'\Phi_i^a)(S_i^b+t_2'\Phi_i^b)]\delta[q_{ab}^{\sigma S}-\frac{1}{N}\sum_i(\sigma_i^a+t_1'\phi_i^a)(S_i^b+t_2'\Phi_i^b)]$$
$$\times\exp[\frac{\beta}{2}\sum_{ab}\sum_{\mu\geq 2}(1+t_1)q_{ab}^\sigma x_\mu^a x_\mu^b+(1+t_2)y_\mu^a y_\mu^b q_{ab}^S+\sqrt{1+t_1}\sqrt{1+t_2}(q_{ab}^{\sigma S}x_\mu^a y_\mu^b+x_\mu^b y_\mu^a q_{ba}^{\sigma S})].$$

Para podermos mexer com facilidade nesses deltas, usamos a representação integral das distribuições delta, as denominamos de r_{ab}^σ, r_{ab}^S, $r_{ab}^{\sigma S}$, sendo elas relacionadas respetivamente a q_{ab}^σ, q_{ab}^S, $q_{ab}^{\sigma S}$.

$$\int\prod_{ab}dq_{ab}^\sigma dq_{ab}^S dq_{ab}^{\sigma S}dr_{ab}^\sigma dr_{ab}^\sigma dr_{ab}^S dr_{ab}^{\sigma S}\exp\{N\sum ir_{ab}^\sigma[q_{ab}^\sigma-(\sigma^a+t_1'\phi^a)(\sigma^b+t_1'\phi^b)] \tag{7.20}$$
$$+N\sum_{ab}ir_{ab}^S[q_{ab}^S-(S^a+t_2'\Phi^a)(S^b+t_2'\Phi^b)]+N\sum_{ab}ir_{ab}^{\sigma S}[q_{ab}^{\sigma S}-(\sigma^a+t_1'\phi^a)(S^b+t_2'\Phi^b)]$$
$$+\frac{\beta}{2}\sum_{ab}[\sum_{\mu\geq 2}(1+t_1)q_{ab}^\sigma x_\mu^a x_\mu^b+(1+t_2)y_\mu^a y_\mu^b q_{ab}^S+\sqrt{1+t_1}\sqrt{1+t_2}(q_{ab}^{\sigma S}x_\mu^a y_\mu^b+x_\mu^b y_\mu^a q_{ba}^{\sigma S})]\}.$$

Assim, temos que

$$\langle Z^n\rangle = \sum_{\sigma^n,S^n}\int\prod_a Dx_1^a Dy_1^a \prod_{i,a} D\Phi_i^a D\phi_i^a \prod_a Dh^a \tag{7.21}$$
$$\times \exp[\sqrt{\frac{\beta(1+t_1)}{N}}\sum_{i,a} x_1^a\xi_i^1(\sigma_i^a + t_1'\phi_i^a) + \sqrt{\frac{\beta(1+t_2)}{N}}y_1^a\xi_i^1(S_i^a + t_2'\Phi_i^a)$$
$$+\sqrt{2\beta\epsilon N}\sum_a S_i^a\sigma_i^a h^a]\int\prod_{a,b}\prod_{\mu\geq 2}Dx_\mu^a Dy_\mu^b\int\prod_{ab}dq_{ab}^\sigma dq_{ab}^S dq_{ab}^{\sigma S} dr_{ab}^\sigma dr_{ab}^\sigma dr_{ab}^S dr_{ab}^{\sigma S}$$
$$\times\exp\{N\sum ir_{ab}^\sigma[]q_{ab}^\sigma - (\sigma^a + t_1'\phi^a)(\sigma^b + t_1'\phi^b)]$$
$$+N\sum_{ab} ir_{ab}^S[q_{ab}^S - (S^a + t_2'\Phi^a)(S^b + t_2'\Phi^b)] + N\sum_{ab} ir_{ab}^{\sigma S}[q_{ab}^{\sigma S} - (\sigma^a + t_1'\phi^a)(S^b + t_2'\Phi^b)]$$
$$+\frac{\beta}{2}\sum_{ab}\sum_{\mu\geq 2}(1+t_1)q_{ab}^\sigma x_\mu^a x_\mu^b + (1+t_2)y_\mu^a y_\mu^b q_{ab}^S + \sqrt{1+t_1}\sqrt{1+t_2}(q_{ab}^{\sigma S}x_\mu^a y_\mu^b + x_\mu^b y_\mu^a q_{ba}^{\sigma S})\}.$$

Examinemos as integrais em x e y na parte não condensada. A soma sobre os padrões é separada, então podemos fatorar as integrais em p-1 integrais com a seguinte fórmula

$$\int\prod_{a,b}Dx^a Dy^b\exp\{\frac{\beta}{2}\sum_{ab}[(1+t_1)q_{ab}^\sigma x^a x^b + (1+t_2)y^a y^b q_{ab}^S \tag{7.22}$$
$$+\sqrt{(1+t_1)(1+t_2)}(q_{ab}^{\sigma S}x^a y^b + x^b y^a q_{ab}^{\sigma S})]\}$$
$$=\int\prod_{a,b}dx^a dy^b\exp\{\frac{1}{2}\sum_{ab} -\delta_{ab}x^a x^b - \delta_{ab}y^a y^b + \beta[(1+t_1)q_{ab}^\sigma x^a x^b + (1+t_2)y^a y^b q_{ab}^S$$
$$+\sqrt{1+t_1}\sqrt{1+t_2}(q_{ab}^{\sigma S}x_\mu^a y_\mu^b + x_\mu^b y_\mu^a q_{ba}^{\sigma S})]\} = \int\prod_{a,b}dx^a dy^b\exp(-\frac{1}{2}z^T[\mathbb{1} - \beta\hat{q})z],$$

onde definimos a matriz $\hat{q}$ quadrada de tamanho 2n e o vetor de dimensão 2n z:

$$\hat{q}\equiv\begin{pmatrix}(1+t_1)\hat{q}^\sigma & \sqrt{(1+t_1)(1+t_2)}\hat{q}^{\sigma S}\\ \sqrt{(1+t_1)(1+t_2)}(\hat{q}^{\sigma S})^T & (1+t_2)\hat{q}^S\end{pmatrix},\quad z\equiv\begin{pmatrix}x_1\\ \dots\\ x_n\\ y_1\\ \dots\\ y_n\end{pmatrix},$$

onde $\hat{q}^\rho$ são matrizes nxn tal que $\hat{q}^\rho = q_{ab}^\rho$.
Com essas definições, também podemos dizer que $\prod_{a,b}^n dx^a dy^b = \prod_a^{2n} dz^a$
Assim, temos

$$\int\prod_a dz^a\exp[-\frac{1}{2}z(\mathbb{1}-\beta q)z] = [\det(\mathbb{1}-\beta q)]^{\frac{1}{2}}. \tag{7.23}$$

Também fazemos as seguintes mudanças de variáveis para recuperar o significado desejado das variáveis:

$$x_1 \to \sqrt{\frac{\beta N}{1+t_1}} m_1^a, y_1 \to \sqrt{\frac{\beta N}{1+t_2}} m_2^a, \tag{7.24}$$
$$r_{ab}^\rho \to i\frac{\alpha\beta^2}{2} r_{ab}^\rho \forall \rho, h^a \to \sqrt{2\epsilon\beta N} h^a,$$

onde ρ pode ser σ, S ou σS. A partir daqui ρ terá esse sentido.
A expressão completa se torna então

$$\langle Z^n \rangle = \int \prod_a dm_1^a dm_2^a dh^a \prod_{ab} dq_{ab}^\sigma dq_{ab}^S dq_{ab}^{\sigma S} dr_{ab}^\sigma dr_{ab}^S dr_{ab}^{\sigma S} \tag{7.25}$$
$$\times \exp\{-\frac{\beta N}{2} \sum_a [\frac{(m_1^a)^2}{1+t_1} + \frac{(M_1^a)^2}{1+t_2} + 2\epsilon (h^a)^2] - \log[\det(\mathbb{1} - \beta\hat{q})]^{\frac{p}{2}}$$
$$-\frac{N\alpha\beta^2}{2} \sum_{ab} (r_{ab}^\sigma q_{ab}^\sigma + r_{ab}^S q_{ab}^S + r_{ab}^{\sigma S} q_{ab}^{\sigma S})$$
$$+\log[\sum_{\sigma^n, S^n} \int \prod_i \prod_a D\Phi_i^a D\phi_i^b \exp(\frac{\alpha\beta^2}{2} \sum_{ab} (r_{ab}^\sigma (\sigma_i^a + t'\phi_i^a)(\sigma_i^b + t_1'\phi_i^b)$$
$$+r_{ab}^S (S_i^a + t_2'\Phi_i^a)(S_i^b + t_2'\Phi_i^b) + r_{ab}^{\sigma S}(\sigma_i^a + t'\phi_i^a)(S_i^b + t_2'\Phi_i^b))$$
$$+\beta \sum_i \sum_a \xi_i^1 (m_1^a(\sigma_i^a + t_1'\phi_i^a) + m_2^a(S_i^a + t_2'\Phi_i^a)) + 2\beta\epsilon \sum_a S_i \sigma_i^a h^a)]\},$$

onde aproximamos $p - 1 \approx p$, já que estamos interessados na região onde p é de ordem $\mathcal{O}(N)$.

A simetria de réplicas usada é

$$m_1^a = m_1, m_2^a = m_2 \quad \forall a, \tag{7.26}$$
$$q_{ab}^\rho = Q^\rho \delta_{ab} + q^\rho (1 - \delta_{ab}) \quad \forall a, b, \rho,$$
$$r_{ab}^\rho = R^\rho \delta_{ab} + r^\rho (1 - \delta_{ab}) \quad \forall a, b, \rho,$$
$$h^a = h \quad \forall a.$$

Os primeiros termos se tornam

$$\frac{1}{2n} \sum_a (\frac{(m_1^a)^2}{1+t_1} + \frac{(M_1^a)^2}{1+t_2} + \epsilon (h^a)^2) = \frac{m_1^2}{2+2t_1} + \frac{m_2^2}{2+2t_2} + \frac{\epsilon h^2}{2} \tag{7.27}$$
$$\frac{\alpha\beta}{2n} \sum_{ab} r_{ab}^\sigma q_{ab}^\sigma = \frac{(\Delta^\sigma - 1)(1+t_2)}{2t_2} Q^\sigma + \frac{\alpha\beta}{2} r^\sigma (Q^\sigma - q^\sigma)$$
$$\frac{\alpha\beta}{2n} \sum_{ab} r_{ab}^S q_{ab}^S = \frac{(\Delta^S - 1)(1+t_1)}{2t_1} Q^S + \frac{\alpha\beta}{2} r^S (Q^S - q^S)$$
$$\frac{\alpha\beta}{2n} \sum_{ab} r_{ab}^{\sigma S} q_{ab}^{\sigma S} = \frac{\alpha\beta}{2} (R^{\sigma S} Q^{\sigma S} - q^{\sigma S} r^{\sigma S})$$
$$= \frac{\alpha\beta}{2} (r^{\sigma S} + \frac{2}{\alpha\beta} \Delta^{\sigma S}) Q^{\sigma S} - \frac{\alpha\beta}{2} q^{\sigma S} r^{\sigma S} = \Delta^{\sigma S} Q^{\sigma S} + \frac{\alpha\beta r^{\sigma S}}{2} (Q^{\sigma S} - q^{\sigma S}),$$

onde trocamos de variáveis $R^\rho \to \Delta^\rho$ por meio de

$$\Delta^\sigma \equiv 1 + \alpha\beta\frac{t_1}{1+t_1}(R^\sigma - r^\sigma), \Delta^S \equiv 1 + \alpha\beta\frac{t_2}{1+t_2}(R^S - r^S), \Delta^{\sigma S} \equiv \frac{\alpha\beta}{2}(R^{\sigma S} - r^{\sigma S}),$$

para calcular o logaritmo determinante de $\mathbb{1} - \beta\hat{q}$, precisamos achar seus autovalores e suas respectivas degenerescências, já que usamos a identidade $\log(\det(M)) = \sum_i \log(\lambda_i)$, onde λ_i são os autovalores e $\sum_i$ percorre todos os autovalores.

Para facilitar as contas, denominamos

$$\begin{aligned} &a = 1 - \beta Q^\sigma(1+t_1),\ \ b = -\beta q^\sigma(1+t_1),\ \ c = 1 - \beta Q^S(1+t_2),\ \ d = -\beta q^S(1+t_2) \\ &e = -\beta Q^{\sigma S}\sqrt{(1+t_1)(1+t_2)},\ \ f = -\beta q^{\sigma S}\sqrt{(1+t_1)(1+t_2)}. \end{aligned} \tag{7.28}$$

A matriz tem 4 autovalores distintos, suas degenerescências e seus valores são:

$$\begin{aligned} &g_1 = 1, \lambda_1 = \frac{1}{2}\{a + (n-1)b + c + (n-1)d \\ &-\sqrt{[a+(n-1)b+c+(n-1)d]^2 - 4[(a+(n-1)b)(c+(n-1)d) - (e+(n-1)f)^2]}\}, \\ &g_2 = 1, \lambda_2 = \frac{1}{2}\{a + (n-1)b + c + (n-1)d \\ &+\sqrt{[a+(n-1)b+c+(n-1)d]^2 - 4[(a+(n-1)b)(c+(n-1)d) - (e+(n-1)f)^2]}\}, \\ &g_3 = n-1, \lambda_3 = \frac{1}{2}\{a - b + c - d - \sqrt{(a-b+c-d)^2 + 4[(-a+b)(c-d) + (e-f)^2]}\}, \\ &g_4 = n-1, \lambda_4 = \frac{1}{2}\{a - b + c - d + \sqrt{(a-b+c-d)^2 + 4[(-a+b)(c-d) + (e-f)^2]}\}. \end{aligned} \tag{7.29}$$

Assim, temos que

$$\begin{aligned} &\sum_i \log(\lambda_i) = \log(\lambda_1) + \log(\lambda_2) + \log(\lambda_3)^{n-1} + \log(\lambda_4)^{n-1} \\ &= n[\log(\lambda_3) + \log(\lambda_4)] + \log(\frac{\lambda_1}{\lambda_3}) + \log(\frac{\lambda_2}{\lambda_4}). \end{aligned} \tag{7.30}$$

Temos que $\lambda_3 = \lambda_1 \mid_{n=0}$ e que $\lambda_4 = \lambda_2 \mid_{n=0}$, assim

$$\begin{aligned} &\sum_i \frac{\log(\lambda_i)}{n} = \log(\lambda_3) + \log(\lambda_4) + \frac{1}{\lambda_3}\frac{\partial\lambda_1}{\partial n} \mid_{n=0} + \frac{1}{\lambda_4}\frac{\partial\lambda_2}{\partial n} \mid_{n=0} \\ &= \log\{\frac{1}{2}[a - b + c - d - \sqrt{(a-b-c+d)^2 + 4(e-f)^2}]\} \\ &+ \log\{\frac{1}{2}[a - b + c - d + \sqrt{(a-b-c+d)^2 + 4(e-f)^2}]\} \\ &+ \frac{b + d - \frac{(b-d)(a-b-c+d)+2f(e-f)}{\sqrt{(a-b-c+d)^2+4(e-f)^2)}})}{a - b + c - d - \sqrt{(a-b-c+d)^2 + 4(e-f)^2}} \\ &+ \frac{b + d + \frac{(b-d)(a-b-c+d)+2f(e-f)}{\sqrt{(a-b-c+d)^2+4(e-f)^2)}})}{a - b + c - d + \sqrt{(a-b-c+d)^2 + 4(e-f)^2}}. \end{aligned} \tag{7.31}$$

Manipulando um pouco, obtemos que

$$\sum_i \frac{\log(\lambda_i)}{n} = \log\{\frac{1}{2}[a-b+c-d-\sqrt{(a-b-c+d)^2+4(e-f)^2}]\} \tag{7.32}$$
$$+\log\{\frac{1}{2}[a-b+c-d+\sqrt{(a-b-c+d)^2+4(e-f)^2}]\} + \frac{b(c-d)+d(a-b)-f(e-f)}{(a-b)(c-d)-(e-f)^2}.$$

Multiplicando pelo fator $\frac{\alpha}{2\beta}$ que aparece na energia livre e usando as variáveis originais, temos que

$$\begin{aligned}
&\frac{\alpha}{2\beta}\log\{\frac{1}{2}[2+\beta(1+t_1)(q^\sigma-Q^\sigma)+\beta(1+t_2)(q^S-Q^S) \\
&-\beta\sqrt{((1+t_1)(q^\sigma-Q^\sigma)-(1+t_2)(q^S-Q^S))^2+4(1+t_1)(1+t_2)(-Q^{\sigma S}+q^{\sigma S})^2}]\} \\
&+\frac{\alpha}{2\beta}\log\{\frac{1}{2}(2+\beta(1+t_1)(q^\sigma-Q^\sigma)+\beta(1+t_2)(q^S-Q^S) \\
&+\beta\sqrt{((1+t_1)(q^\sigma-Q^\sigma)-(1+t_2)(q^S-Q^S))^2+4(1+t_1)(1+t_2)(-Q^{\sigma S}+q^{\sigma S})^2}]\} \\
&-\alpha\{q^\sigma(1+t_1)[1-\beta(Q^S-q^S)(1+t_2)] \\
&+q^S(1+t_2)[1-\beta(Q^\sigma-q^\sigma)(1+t_1)]+\beta(1+t_1)(1+t_2)q^{\sigma S}(-q^{\sigma S}+Q^{\sigma S})\} \\
&\times\{2[1-\beta(1+t_1)(Q^\sigma-q^\sigma)][1-\beta(1+t_2)(Q^S-q^S)]-2\beta^2(1+t_1)(1+t_2)(-Q^{\sigma S}+q^{\sigma S})^2\}^{-1}.
\end{aligned} \tag{7.33}$$

Agora focamos no último termo, que é a parte em que a soma sobre os estados possíveis é feita.

$$\begin{aligned}
&\frac{1}{n\beta N}\log\{\sum_{\sigma_N^n,S_N^n}\int\prod_{i,a}D\Phi_i^aD\phi_i^a\exp[\frac{\alpha\beta^2}{2}\sum_i\sum_{ab}(r_{ab}^\sigma(\sigma_i^a+t_1'\phi_i^a)(\sigma_i^b+t_1'\phi_i^b) \\
&+r_{ab}^S(S_i^a+t_2'\Phi_i^a)(S_i^b+t_2'\Phi_i^b)+r_{ab}^{\sigma S}(\sigma_i^a+t_1'\phi_i^a)(S_i^b+t_2'\Phi_i^b)) \\
&+\beta\sum_i\sum_a(m_1^a(\sigma_i^a+t_1'\phi_i^a)+m_2^a(S_i^a+t_2'\Phi_i^a))+\beta\epsilon\sum_i\sum_a S_i^a\sigma_i^ah^a]\}.
\end{aligned} \tag{7.34}$$

Fatoramos as N integrais

$$\begin{aligned}
&\frac{1}{n\beta}\log\{\sum_{\sigma^n,S^n}\int\prod_a D\Phi^aD\phi^a\exp[\frac{\alpha\beta^2}{2}\sum_{ab}(r_{ab}^\sigma(\sigma^a+t_1'\phi^a)(\sigma^b+t_1'\phi^b) \\
&+r_{ab}^S(S^a+t_2'\Phi^a)(S^b+t_2'\Phi^b)+r_{ab}^{\sigma S}(\sigma^a+t_1'\phi^a)(S^b+t_2'\Phi^b)) \\
&+\beta\sum_a(m_1^a(\sigma^a+t_1'\phi^a)+m_2^a(S^a+t_2'\Phi^a))+\beta\epsilon\sum_a S^a\sigma^ah^a]\}.
\end{aligned} \tag{7.35}$$

Aplicamos a simetria de réplicas e fazemos a seguinte manipulação:

$$\sum_{ab} r^{\sigma}(\sigma^a + t_1'\phi^a)(\sigma^b + t_1'\phi^b) + \sum_{ab} r^{S}(S^a + t_2'\Phi^a)(S^b + t_2')$$
$$+ \sum_{ab} r^{\sigma S}(\sigma^a + t_1'\phi^a)(S^b + t_2') \tag{7.36}$$
$$= \frac{1}{2}[\sqrt{r^{\sigma}}\sum_a(\sigma^a + t_1'\phi^a) + \sqrt{r^{S}}\sum_a(S^a + t_2'\Phi^a)]^2(1 + \frac{r^{\sigma S}}{2\sqrt{r^{\sigma}r^{S}}})$$
$$+\frac{1}{2}[\sqrt{r^{\sigma}}\sum_a(\sigma^a + t_1'\phi^a) - \sqrt{r^{S}}\sum_a(S^a + t_2'\Phi^a)]^2(1 - \frac{r^{\sigma S}}{2\sqrt{r^{\sigma}r^{S}}}),$$

e denotamos $r_+ \equiv \sqrt{1 + \frac{r^{\sigma S}}{2\sqrt{r^{\sigma}r^{S}}}}$ e $r_- \equiv \sqrt{1 - \frac{r^{\sigma S}}{2\sqrt{r^{\sigma}r^{S}}}}$.
Assim, obtemos que

$$\frac{1}{n\beta}\log\{\sum_{\sigma^n, S^n}\int\prod_a D\Phi^a D\phi^a \exp[\frac{\alpha\beta^2}{2}\sum_a((R^{\sigma} - r^{\sigma})(\sigma^a + t_1'\phi^a)^2 + (R^{S} - r^{S})(S^a + t_2'\Phi^a)^2$$
$$+(R^{\sigma S} - r^{\sigma S})(\sigma^a + t_1'\phi^a)(S^a + t_2'\Phi^a) + \frac{1}{2}(\sqrt{r^{\sigma}}(\sigma^a + t_1'\phi^a) + \sqrt{r^{S}}(S^a + t_2'\Phi^a))^2 r_+^2 \tag{7.37}$$
$$+\frac{1}{2}(\sqrt{r^{\sigma}}(\sigma^a + t_1'\phi^a) - \sqrt{r^{S}}(S^a + t_2'\Phi^a))^2 r_-^2)$$
$$+\beta\sum_a(m_1(\sigma^a + t_1'\phi^a) + m_2(S^a + t_2'\Phi^a)) + \beta\epsilon\sum_a S^a\sigma^a h^a]\}.$$

Para simplificar significativamente as contas, fazemos a seguinte mudança de variável:

$$\psi = \phi + \frac{\sigma}{t_1'}, \quad \Psi = \Phi + \frac{S}{t_2'}, \tag{7.38}$$

obtendo

$$\frac{1}{n\beta}\log\{\int DxDy[\sum_{\sigma,S}\int D\Psi D\psi \exp(\frac{\alpha\beta^2}{2}((R^{\sigma} - r^{\sigma})(t_1'\psi)^2 + (R^{S} - r^{S})(t_2'\Psi)^2$$
$$+(R^{\sigma S} - r^{\sigma S})t_1't_2'\psi\Psi) + \frac{\psi\sigma}{t_1'} + \frac{\Psi S}{t_2'} + r_+x\beta\sqrt{\frac{\alpha}{2}}(\sqrt{r^{\sigma}}t_1'\psi + \sqrt{r^{S}}t_2'\Psi) \tag{7.39}$$
$$+r_-y\beta\sqrt{\frac{\alpha}{2}}(\sqrt{r^{\sigma}}t_1'\psi - \sqrt{r^{S}}t_2'\Psi) + 2\beta\epsilon\sigma Sh + \beta(m_1t_1'\psi + m_2t_2'\Psi))]^n\}.$$

Fazemos as duas integrais

$$\frac{1}{n\beta}\log\{\int DxDy[\sum_{\sigma,S}\exp((\Delta^S t'^2\beta^2(\sqrt{\frac{\alpha r^\sigma}{2}}(r_+x+r_-y)+m_1+\frac{\sigma}{\beta t_1'^2})^2$$
$$+\Delta^\sigma t_2'^2\beta^2(\sqrt{\frac{\alpha r^S}{2}}(r_+x-r_-y)+m_2+\frac{S}{\beta t_2'^2})^2+2t_1'^2t_2'^2\beta^3\Delta^{\sigma S} \tag{7.40}$$
$$*(\sqrt{\frac{\alpha r^\sigma}{2}}(r_+x+r_-y)+m_1+\frac{\sigma}{\beta t_1'^2})(\sqrt{\frac{\alpha r^S}{2}}(r_+x-r_-y)+m_2+\frac{S}{\beta t_2'^2}))$$
$$*(2(\Delta^\sigma\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2))^{-1}+2\beta\epsilon\sigma Sh)(\Delta^\sigma\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2)^{-\frac{1}{2}}]^n\}.$$

Abrindo os quadrados, o termo se torna

$$\frac{1}{n\beta}\log\{\int DxDy[\sum_{\sigma,S}(\Delta^\sigma\Delta^S-t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2)^{-\frac{1}{2}} \tag{7.41}$$
$$*\exp((\Delta^S t_1'^2\beta^2(\frac{\alpha r^\sigma}{2}(r_+x+r_-y)^2+m_1^2+\frac{1}{\beta^2t_1'^4}+m_1\sqrt{2\alpha r^\sigma}(r_+x+r_-y)+\frac{2m_1\sigma}{\beta t_1'^2}$$
$$+\sqrt{2\alpha r^\sigma}\frac{\sigma(r_+x+r_-y)}{\beta t_1'^2})+\Delta^\sigma t_2'^2\beta^2(\frac{\alpha r^S}{2}(r_+x-r_-y)^2+m_2^2+\frac{1}{\beta^2t_2'^4}$$
$$+m_2\sqrt{2\alpha r^S}(r_+x-r_-y)+\sqrt{2\alpha r^S}\frac{S(r_+x-r_-y)}{\beta t_2'^2}+\frac{2m_2S}{\beta t_2'^2})$$
$$+2t_1'^2t_2'^2\beta^3\Delta^{\sigma S}(\frac{\sqrt{r^Sr^\sigma}\alpha(r_+^2x^2-r_-^2y^2)}{2}+m_2\sqrt{\frac{\alpha r^\sigma}{2}}(r_+x+r_-y)$$
$$+\frac{S}{\beta t_2'^2}\sqrt{\frac{\alpha r^\sigma}{2}}(r_+x+r_-y)+m_1\sqrt{\frac{\alpha r^S}{2}}(r_+x-r_-y)+m_1m_2+\frac{m_1S}{\beta t_2'^2}$$
$$+\frac{\sigma}{\beta t_1'^2}\sqrt{\frac{\alpha r^S}{2}}(r_+x-r_-y)+\frac{m_2\sigma}{\beta t_1'^2}+\frac{\sigma S}{\beta^2t_1'^2t_2'^2}))(2\Delta^\sigma\Delta^S-2t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2)^{-1}+2\beta\epsilon\sigma Sh)]^n\}.$$

Para facilitar a visualização, dividimos os termos independentes dos estados de σ e S do dependentes.

Os independentes podem ser integrados sobre x e y imediatamente, eles são

$$\{\Delta^S t_1'^2\beta^2[\frac{\alpha r^\sigma}{2}(r_+x+r_-y)^2+m_1^2+\frac{1}{\beta^2t_1'^4}+m_1\sqrt{2\alpha r^\sigma}(r_+x+r_-y)]$$
$$+\Delta^\sigma t_2'^2\beta^2[\frac{\alpha r^S}{2}(r_+x-r_-y)^2+m_1^2+\frac{1}{\beta^2t_2'^4}+m_2\sqrt{2\alpha r^S}(r_+x-r_-y)]$$
$$+2t_1'^2t_2'^2\beta^3\Delta^{\sigma S}[\sqrt{r^\sigma r^S}\frac{\alpha(r_+^2x^2-r_-^2y^2)}{2}+m_2\sqrt{\frac{\alpha r^\sigma}{2}}(r_+x+r_-y) \tag{7.42}$$
$$+m_1\sqrt{\frac{\alpha r^S}{2}}(r_+x-r_-y)+m_1m_2]\}[2\Delta^S\Delta^\sigma-2t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2]^{-1}.$$

Quando integrados, eles se tornam

$$\frac{\Delta^S t_1'^2\beta^2(\alpha r^\sigma+m_1^2+\frac{1}{\beta^2t_1'^4})+\Delta^\sigma t_2'^2\beta^2(\alpha r^S+m_2^2+\frac{1}{\beta^2t_2'^4})+2t_1'^2t_2'^2\beta^3\Delta^{\sigma S}(\frac{r^{\sigma S}\alpha}{2}+m_1m_2)}{2\Delta^\sigma\Delta^S-2t_1'^2t_2'^2\beta^2(\Delta^{\sigma S})^2} \tag{7.43}$$

Notamos que para essa integral convergir, é necessário que $\Delta^\sigma \Delta^S > t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2$, já que senão temos uma exponencial cujo argumento é uma expressão quadrática positiva, e nenhum outro termos da integral conseguiria reduzi-lo a zero no infinito. Embora não seja claro se e quando ocorre essa inversão de sinal, hipotetizamos que se for ocorrer, ocorreria para valores muito altos de ϵ, onde teríamos valores elevados de $\Delta^{\sigma S}$. Testamos essa condição, e ela sempre foi satisfeita sem nenhum problema, o que indica certa consistência, embora não constitua uma prova definitiva já que os próprios valores de Δ^ρ são calculados assumindo a desigualdade.

Os dependentes são

$$\sum_{\sigma,S} \exp\{[\Delta^S t_1'^2 \beta^2 (\frac{2m_1\sigma}{\beta t'^2} + \sqrt{2\alpha r^\sigma}\frac{\sigma(r_+x + r_-y)}{\beta t_1'^2}) \tag{7.44}$$
$$+\Delta^\sigma t_2'^2 \beta^2 (\sqrt{2\alpha r^S}\frac{S(r_+x - r_-y)}{\beta t_2'^2} + \frac{2m_2 S}{\beta t_2'^2}) + 2t_1'^2 t_2'^2 \beta^3 \Delta^{\sigma S}(\frac{S}{\beta t_2'^2}\sqrt{\frac{\alpha r^\sigma}{2}}(r_+x + r_-y) + \frac{m_1 S}{\beta t_2'^2}$$
$$+\frac{\sigma}{\beta t_1'^2}\sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y) + \frac{m_2\sigma}{\beta t_1'^2} + \frac{\sigma S}{\beta^2 t_1'^2 t_2'^2})](2\Delta^\sigma\Delta^S - 2t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2)^{-1} + 2\beta\epsilon\sigma S h\}$$
$$= \sum_{\sigma,S} \exp\{[\Delta^S\beta(2m_1\sigma + \sqrt{2\alpha r^\sigma}\sigma(r_+x + r_-y)) + \Delta^\sigma\beta(\sqrt{2\alpha r^S}S(r_+x - r_-y) + 2m_2 S)$$
$$+2\beta\Delta^{\sigma S}(S t_1'^2\beta\sqrt{\frac{\alpha r^\sigma}{2}}(r_+x + r_-y) + m_1 S t_1'^2\beta$$
$$+\beta t_2'^2\sigma\sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y) + \beta t_2'^2 m_2\sigma + \sigma S)][2\Delta^\sigma\Delta^S - 2t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2]^{-1} + 2\beta\epsilon\sigma S h\}$$
$$= \sum_{\sigma,S} \exp\{\beta\sigma[\frac{\Delta^S(m_1 + \sqrt{\frac{\alpha r^\sigma}{2}}(r_+x + r_-y)) + t_2'^2\beta\Delta^{\sigma S}(m_2 + \sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y))}{\Delta^\sigma\Delta^S - t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2}]$$
$$+\beta S[\frac{\Delta^\sigma(m_2 + \sqrt{\frac{\alpha r^S}{2}}(r_+x - r_-y)) + t_1'^2\beta\Delta^{\sigma S}(m_1 + \sqrt{\frac{\alpha r^\sigma}{2}}(r_+x + r_-y))}{\Delta^\sigma\Delta^S - t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2}]$$
$$+\beta\sigma S[2\epsilon h + \frac{\Delta^{\sigma S}}{\Delta^\sigma\Delta^S - t_1'^2 t_2'^2\beta^2(\Delta^{\sigma S})^2}]\}.$$

Notemos que

$$\sum_{\sigma,S} \exp(A\sigma + BS + C\sigma S) = \exp(A + B + C) + \exp(A - B - C)$$
$$+ \exp(-A + B - C) + \exp(-A - B + C) = \exp(C)\cosh(A + B) + \exp(-C)\cosh(A - B).$$

Assim, os termos dependentes dos estados são

$$\frac{1}{\beta}\int DxDy \log\{\cosh[\beta(\upsilon + \Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon - \Upsilon)]\exp(-\beta\eta)\}, \tag{7.45}$$

onde denotamos

$$\eta = 2\epsilon h + \frac{\Delta^{\sigma S}}{\Delta^{\sigma}\Delta^{S} - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2} \tag{7.46}$$

$$\upsilon = \frac{\Delta^{S}[m_1 + \sqrt{\frac{\alpha r^{\sigma}}{2}}(r_+ x + r_- y)] + t_2'^2 \beta \Delta^{\sigma S}[m_2 + \sqrt{\frac{\alpha r^{S}}{2}}(r_+ x - r_- y)]}{\Delta^{\sigma}\Delta^{S} - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2}$$

$$\Upsilon = \frac{\Delta^{\sigma}[m_2 + \sqrt{\frac{\alpha r^{S}}{2}}(r_+ x - r_- y)] + t_1'^2 \beta \Delta^{\sigma S}[m_1 + \sqrt{\frac{\alpha r^{\sigma}}{2}}(r_+ x + r_- y)]}{\Delta^{\sigma}\Delta^{S} - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2}.$$

Juntando as equações (5.41), (5.39), (5.29) e (5.23), obtemos que a energia livre é, a menos de constantes aditivas,

$$\begin{aligned}
&f(m_1, m_2, h, Q^{\sigma}, Q^{S}, Q^{\sigma S}, q^{\sigma}, q^{S}, q^{\sigma S}, t_1, t_2, \beta, \alpha, \epsilon) = \\
&\frac{m_1^2}{2+2t_1} + \frac{m_2^2}{2+2t_2} + \epsilon h^2 + \frac{(\Delta^{\sigma}-1)(1+t_1)}{2t_1} Q^{\sigma} + \frac{\log[\Delta^{\sigma}\Delta^{S} - t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2]}{2\beta} \\
&+\frac{\alpha\beta}{2} r^{\sigma}(Q^{\sigma} - q^{\sigma}) + \frac{(\Delta^{S}-1)(1+t_2)}{2t_2} Q^{S} + \frac{\alpha\beta}{2} r^{S}(Q^{S} - q^{S}) \\
&+\Delta^{\sigma S} Q^{\sigma S} + \frac{\alpha\beta r^{\sigma S}}{2}(Q^{\sigma S} - q^{\sigma S}) + \frac{\alpha}{2\beta}\log\{\frac{1}{2}[2 + \beta(1+t_1)(q^{\sigma} - Q^{\sigma}) + \beta(1+t_2)(q^{S} - Q^{S}) \\
&-\beta\sqrt{((1+t_1)(q^{\sigma} - Q^{\sigma}) - (1+t_2)(q^{S} - Q^{S}))^2 + 4(1+t_1)(1+t_2)(-Q^{\sigma S} + q^{\sigma S})^2}]\} \\
&+\frac{\alpha}{2\beta}\log\{\frac{1}{2}[2 + \beta(1+t_1)(q^{\sigma} - Q^{\sigma}) + \beta(1+t_2)(q^{S} - Q^{S}) \\
&+\beta\sqrt{((1+t_1)(q^{\sigma} - Q^{\sigma}) - (1+t_2)(q^{S} - Q^{S}))^2 + 4(1+t_1)(1+t_2)(-Q^{\sigma S} + q^{\sigma S})^2}]\} \\
&-\alpha\{q^{\sigma}(1+t_1)[1 - \beta(Q^{S} - q^{S})(1+t_2)] \\
&+q^{S}(1+t_2)[1 - \beta(Q^{\sigma} - q^{\sigma})(1+t_1)] + \beta(1+t_1)(1+t_2) q^{\sigma S}(-q^{\sigma S} + Q^{\sigma S})\} \\
&\times\{2[1 - \beta(1+t_1)(Q^{\sigma} - q^{\sigma})][1 - \beta(1+t_2)(Q^{S} - q^{S})] - 2\beta^2(1+t_1)(1+t_2)(-Q^{\sigma S} + q^{\sigma S})^2\}^{-1} \\
&-\frac{1}{\beta}\int Dx Dy \log\{\cosh[\beta(\upsilon + \Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon - \Upsilon)]\exp(-\beta\eta)\} \\
&-\frac{\Delta^{S} t_1'^2 \beta(\alpha r^{\sigma} + m_1^2 + \frac{1}{\beta^2 t_1'^4}) + \Delta^{\sigma} t_2'^2 \beta(\alpha r^{S} + m_2^2 + \frac{1}{\beta^2 t_2'^4}) + 2t_1'^2 t_2'^2 \beta^2 \Delta^{\sigma S}(\frac{r^{\sigma S}\alpha}{2} + m_1 m_2)}{2\Delta^{\sigma}\Delta^{S} - 2t_1'^2 t_2'^2 \beta^2 (\Delta^{\sigma S})^2}.
\end{aligned} \tag{7.47}$$

Para obter as equações de consistência só é necessário derivar este expressão pelas variáveis e igualar a 0. Diferentemente de [8], não é necessária nenhuma outra modificação ou simplificação para chegar nelas, já que apesar de termos tentado ter facilitado a análise dessas expressões acreditamos que não é possível simplificá-las.

7.3 Explicação das mudanças nas equações de consistência com temperatura 0

Este apêndice menor serve para explicar como chegamos às equações de temperatura 0. Peguemos a primeira combinação.

$$
\begin{aligned}
I_1 &= \frac{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) - \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}, \\
I_2 &= \frac{\sinh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \sinh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}, \\
I_3 &= \frac{\sinh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) - \sinh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}{\cosh[\beta(\upsilon+\Upsilon)]\exp(\beta\eta) + \cosh[\beta(\upsilon-\Upsilon)]\exp(-\beta\eta)}.
\end{aligned} \tag{7.48}
$$

Podemos escrever que

$$
\begin{aligned}
I_1 &= 1 - \frac{2}{\frac{\cosh[\beta(\upsilon+\Upsilon)]}{\cosh[\beta(\upsilon-\Upsilon)]}\exp(2\beta\eta) + 1} \\
&= \begin{cases}
1 - \frac{2}{\exp[2\beta(\eta+\Upsilon)]+1} = \operatorname{sign}(\eta+\Upsilon), & \text{se } \upsilon+\Upsilon>0, \upsilon-\Upsilon>0 \\
1 - \frac{2}{\exp[2\beta(\eta+\upsilon)]+1} = \operatorname{sign}(\eta+\upsilon), & \text{se } \upsilon+\Upsilon>0, \upsilon-\Upsilon<0 \\
1 - \frac{2}{\exp[2\beta(\eta-\upsilon)]+1} = \operatorname{sign}(\eta-\upsilon), & \text{se } \upsilon+\Upsilon<0, \upsilon-\Upsilon>0 \\
1 - \frac{2}{\exp[2\beta(\eta-\Upsilon)]+1} = \operatorname{sign}(\eta-\Upsilon), & \text{se } \upsilon+\Upsilon<0, \upsilon-\Upsilon<0
\end{cases}.
\end{aligned} \tag{7.49}
$$

onde usamos o fato de que $\lim_{\beta\to\infty}\cosh(\beta x) = \frac{\exp(\operatorname{sign}(x)\beta x)}{2}$.

Notemos que restrições do tipo $f(x,y) > 0$ podem ser incorporados à equação como $\frac{\operatorname{sign}(f(x,y))+1}{2}$. Assim, temos que

$$
\begin{aligned}
I_1 = \frac{1}{4}\{&[\operatorname{sign}(\upsilon+\Upsilon)+1][\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\eta+\Upsilon) \\
&+[\operatorname{sign}(\upsilon+\Upsilon)+1][-\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\eta+\upsilon) \\
&+[-\operatorname{sign}(\upsilon+\Upsilon)+1][\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\eta-\upsilon) \\
&+[-\operatorname{sign}(\upsilon+\Upsilon)+1][-\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\eta-\Upsilon)\}.
\end{aligned} \tag{7.50}
$$

Para a segunda combinação, podemos escrever que

$$
I_2 = \begin{cases}
\frac{\exp[\beta(\upsilon+\Upsilon+\eta)]+\exp[\beta(\upsilon-\Upsilon-\eta)]}{\exp[\beta(\upsilon+\Upsilon+\eta)]+\exp[\beta(\upsilon-\Upsilon-\eta)]} = 1, & \text{se } \upsilon+\Upsilon>0, \upsilon-\Upsilon>0 \\
\frac{\exp[\beta(\upsilon+\Upsilon+\eta)]-\exp[\beta(-\upsilon+\Upsilon-\eta)]}{\exp[\beta(\upsilon+\Upsilon+\eta)]+\exp[\beta(-\upsilon+\Upsilon-\eta)]} = 1 - \frac{2}{\exp[2(\upsilon+\eta)]+1} = \operatorname{sign}(\upsilon+\eta), & \text{se } \upsilon-\Upsilon<0, \upsilon+\Upsilon>0 \\
\frac{-\exp[\beta(-\upsilon-\Upsilon+\eta)]-\exp[\beta(\upsilon-\Upsilon-\eta)]}{\exp[\beta(\upsilon+\Upsilon+\eta)]+\exp[\beta(-\upsilon+\Upsilon-\eta)]} = 1 - \frac{2}{\exp[2(\upsilon+\eta)]+1} = \operatorname{sign}(\upsilon-\eta), & \text{se } \upsilon-\Upsilon>0, \upsilon+\Upsilon<0 \\
\frac{-\exp[\beta(\upsilon+\Upsilon+\eta)]-\exp[\beta(\upsilon-\Upsilon-\eta)]}{\exp[\beta(\upsilon+\Upsilon+\eta)]+\exp[\beta(\upsilon-\Upsilon-\eta)]} = -1, & \text{se } \upsilon+\Upsilon<0, \upsilon-\Upsilon<0
\end{cases},
$$

$$
\begin{aligned}
I_2 = \frac{1}{4}\{&[\operatorname{sign}(\upsilon+\Upsilon)+1][\operatorname{sign}(\upsilon-\Upsilon)+1] \\
&+[\operatorname{sign}(\upsilon+\Upsilon)+1][-\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\eta+\upsilon) \\
&+[-\operatorname{sign}(\upsilon+\Upsilon)+1][\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\upsilon-\eta) - [-\operatorname{sign}(\upsilon+\Upsilon)+1][-\operatorname{sign}(\upsilon-\Upsilon)+1]\}.
\end{aligned} \tag{7.51}
$$

Temos que $I_3(\upsilon,\Upsilon) = I_2(\Upsilon,\upsilon)$, assim

$$
\begin{aligned}
I_3 = \frac{1}{4}\{&[\operatorname{sign}(\upsilon+\Upsilon)+1][\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\eta+\Upsilon) \\
&+[-\operatorname{sign}(\upsilon+\Upsilon)+1][-\operatorname{sign}(\upsilon-\Upsilon)+1]\operatorname{sign}(\Upsilon-\eta) \\
&+[\operatorname{sign}(\upsilon+\Upsilon)+1][-\operatorname{sign}(\upsilon-\Upsilon)+1] - [-\operatorname{sign}(\upsilon+\Upsilon)+1][\operatorname{sign}(\upsilon-\Upsilon)+1]\}.
\end{aligned} \tag{7.52}
$$

7.4 Simulações de Monte Carlo testando o algoritmo de *learning* hebbiano coletivo

Fizemos 4 simulações de Monte Carlo distintas testando o algoritmo de *learning* hebbiano coletivo. Usamos o algoritmo de Metropolis clássico e escolhemos os padrões com uma distribuição de Bernoulli para garantir que sejam descorrelacionados. Para atualizar as duas redes, escolhemos aleatoriamente os neurônios a serem atualizados.

Nessas simulações focamos em ver como o algoritmo modifica a magnetização e como ele aproxima as matrizes sinápticas da matriz da pseudo-inversa, testando as ideias desenvolvidas no capítulo 2.

	Primeira simulação	Segunda simulação	Terceira simulação	Quarta simulação
N	100	100	100	100
p	40	40	50	10
t_1	0	0,8	0	0
t_2	10	0,8	0,8	0
$\epsilon^{(1)}$	0,01	0,01	0,01	0,01
SI	3500	1000	2000	3000
m_1	0,98	0,32	0,14	0,46
m_2	1	0,92	0,4	1
OPI_1	0,897	0,802	0,617	0,67
OPI_2	1	0,894	0,7	0,95
MC	100	100	100	100

t_1 é o valor de t do primeiro indivíduo, t_2 o valor de t do segundo indivíduo. D é o valor da desconfiança. SI é o número de interações sociais. m_i é o valor final da magnetização do indivíduo i, OPI_i é o valor final do overlap do matriz sináptica do indivíduo i com a matriz da pseudo-inversa. MC é o número de passos de Monte Carlo para a iteração das redes individuais.

A primeira simulação mostra uma rede com $t_2 = 10$ ensinando uma rede com $t_1 = 0$, a permitindo ter magnetização com $\alpha = 0.4$, o que antes não era possível. Além disso, a matriz se aproxima significativamente da pseudo-inversa, chegando a ter aproximadamente 90% de semelhança. Nesse caso, a interação é benéfica.

A terceira simulação é parecida, porém nesse caso temos mais padrões e o professor é menos habilitado, assim não temos magnetização intensiva e as matrizes estão significativamente longe da pseudo inversa. O fato de que se usou menos interações sociais é significativo também, talvez com mais interações tenhamos uma melhora nas redes. Nesse caso, a interação não é benéfica e nem maléfica.

A segunda e quarta simulação mostram situações onde a interação acaba sendo maléfica. Dois indivíduos com a mesma capacidade tenta ensinar, porém o resultado não é benéfico, vejam que a rede aluno acaba tendo um deterioração da sua magnetização e semelhança com a pseudo-inversa.

Em geral, concluímos que interações benéficas ocorrem principalmente quando a diferença de aprendizado prévio entre professor e aluno é mais significativa e o professor já consegue processar informação satisfatoriamente na região. Caso contrário, a interação pode não ter efeito ou até ser maléfica, mostrando que o algoritmo consegue de certa forma mimetizar o aprendizado com 2 pessoas.

7.5 Resultados relevantes do modelo de Ashkin-Teller

As soluções possíveis são $m_1 = m_2 = h = 0$; $m_1 = m_2 \equiv m \neq 0, h \neq 0$; $m_1 = m_2 = 0, h \neq 0$, assim, podemos simplesmente caracterizar o modelo analisando as equações de m e h. Simplificando ao máximo as equações, temos que

$$m = \frac{\sinh(2\beta m)\exp(4\beta\epsilon h)}{\cosh(2\beta m)\exp(4\beta\epsilon h)+1}, \quad h = \frac{\cosh(2\beta m)\exp(4\beta\epsilon h)-1}{\cosh(2\beta m)\exp(4\beta\epsilon h)+1}, \tag{7.53}$$

sendo o Hamiltoniano

$$-\beta H = \frac{\beta}{2N}\sum_{i,j}\sigma_i\sigma_j + \frac{\beta}{2N}\sum_{i,j}S_iS_j + \beta\epsilon N(\frac{1}{N}\sum_i \sigma_i S_i)^2. \tag{7.54}$$

Temos 3 casos distintos:

- h e m vão a 0 em $\beta_c = 1$ com uma transição de segunda ordem, como acontece no modelo individual. Isso ocorre na região $0 \leq \epsilon < 0.135$;
- h e m vão para 0 em um $0.5 < \beta_c < 1$ com uma transição de primeira ordem. Isso ocorre na região $0.135 \leq \epsilon < 0.7$;
- m vai a 0 em uma temperatura $0.5 < \beta_{c1} < 1$ em uma transição de primeira ordem, nesse ponto h sofre uma descontinuidade, depois indo a 0 em um ponto $0.5 < \beta_{c2} < \beta_{c1}$ em uma transição de segunda ordem. Isso ocorre para $0.7 < \epsilon < 1$
- m vai a 0 em uma temperatura $0.5 < \beta_{c3} < 1$ e nesse ponto a primeira derivada de h em relação à temperatura sofre uma descontinuidade, depois indo a 0 em um ponto $0.5 < \beta_{c4} < \beta_{c3}$ em uma transição de segunda ordem. Isso ocorre para $\epsilon \geq 1$.

Nos casos 3 e 4, onde a magnetização vai a 0 antes de h, podemos calcular a temperatura crítica de h explicitamente:

$$h = \frac{\cosh(2\beta m)\exp(4\beta\epsilon h)-1}{\cosh(2\beta m)\exp(4\beta\epsilon h)+1} = \frac{\exp(4\beta\epsilon h)-1}{\exp(4\beta\epsilon h)+1} = \tanh(2\beta\epsilon), \tag{7.55}$$

assim $\beta_{ch} = \frac{1}{2\epsilon}$ nestas regiões.
No caso 4, é possível encontrar um expressão para a temperatura crítica de m:

$$m = \frac{(2\beta m + \frac{4\beta^3 m^3}{3})\exp(4\beta\epsilon h)}{1+\exp(4\beta\epsilon h)+2\beta^2 m^2\exp(4\beta\epsilon h)} = (\beta m + \frac{2\beta^3 m^3}{3}) * (\frac{1+\exp(-4\beta\epsilon h)}{2} - \beta^2 m^2)$$
$$m(\frac{1+\exp(-4\beta\epsilon h)}{2} - \beta) = -\frac{\beta^3 m^3}{3}m^3, \quad \beta_{c1} = \frac{1+\exp(-4\beta_{c1}\epsilon h)}{2}. \tag{7.56}$$

Juntando as duas expressões, temos que no caso 4

$$\beta_{cm} = \exp[4\epsilon(\beta_{cm}-1)]\cosh[]4\epsilon(1-\beta_{cm})]. \tag{7.57}$$

A figura a seguir mostra a relação das temperatura críticas com ϵ:

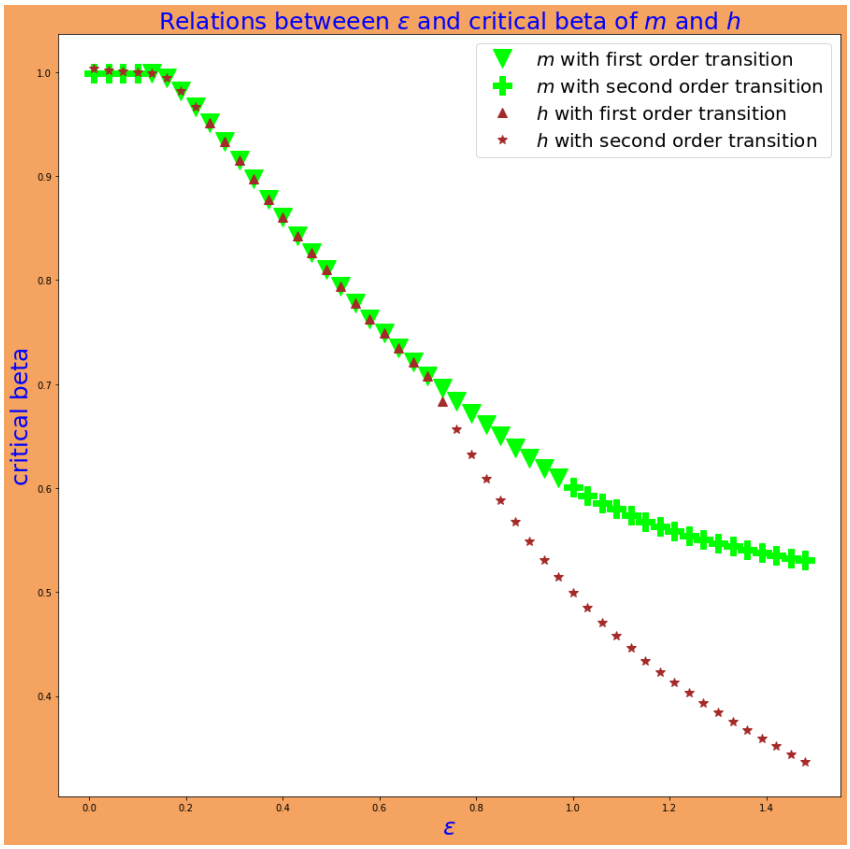


Figure 7.1: Comportamento de β_{c1} e β_{c2} em função de ϵ.

Bibliography

[1] J J Hopfield. "Neural networks and physical systems with emergent collective computational abilities." In: *Proceedings of the National Academy of Sciences* 79.8 (1982), pp. 2554–2558. DOI: 10.1073/pnas.79.8.2554. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.79.8.2554. URL: https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

[2] Rumelhart, James Mcclelland, and James L. *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations*. Jan. 1986.

[3] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Rev. Mod. Phys.* 74 (1 Jan. 2002), pp. 47–97. DOI: 10.1103/RevModPhys.74.47. URL: https://link.aps.org/doi/10.1103/RevModPhys.74.47.

[4] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. "Statistical physics of social dynamics". In: *Rev. Mod. Phys.* 81 (2 May 2009), pp. 591–646. DOI: 10.1103/RevModPhys.81.591. URL: https://link.aps.org/doi/10.1103/RevModPhys.81.591.

[5] Nestor Caticha and Felippe Alves. "Trust, law and ideology in a Neural Network agent model of the US-Appellate Courts". In: *ESANN 2019 - Proceedings 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (2019).

[6] John Hopfield, D Feinstein, and R Palmer. "'Unlearning' has a stabilizing effect in collective memories". In: *Nature* 304 (July 1983), pp. 158–9. DOI: 10.1038/304158a0.

[7] Francis H. C. Crick and Graeme J. Mitchison. "The function of dream sleep". In: *Nature* 304 (1983), pp. 111–114.

[8] Alberto Fachechi, Elena Agliari, and Adriano Barra. "Dreaming neural networks: Forgetting spurious memories and reinforcing pure ones". In: *Neural Networks* 112 (2019), pp. 24–40. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2019.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0893608019300176.

[9] E Gardner. "The space of interactions in neural network models". In: *Journal of Physics A: Mathematical and General* 21.1 (1988), pp. 257–270. DOI: 10.1088/0305-4470/21/1/030. URL: https://doi.org/10.1088/0305-4470/21/1/030.

[10] E Gardner and B Derrida. "Optimal storage properties of neural network models". In: *Journal of Physics A: Mathematical and General* 21.1 (Jan. 1988), pp. 271–284. DOI: 10.1088/0305-4470/21/1/031. URL: https://doi.org/10.1088/0305-4470/21/1/031.

[11] V S Dotsenko, N D Yarunin, and E A Dorotheyev. “Statistical mechanics of Hopfield-like neural networks with modified interactions”. In: *Journal of Physics A: Mathematical and General* 24.10 (May 1991), pp. 2419–2429. DOI: 10.1088/0305-4470/24/10/026. URL: https://doi.org/10.1088/0305-4470/24/10/026.

[12] I. Kanter and H. Sompolinsky. “Associative recall of memory without errors”. In: *Phys. Rev. A* 35 (1 Jan. 1987), pp. 380–392. DOI: 10.1103/PhysRevA.35.380. URL: https://link.aps.org/doi/10.1103/PhysRevA.35.380.

[13] Viktor Dotsenko. *Introduction to the Replica Theory of Disordered Statistical Systems*. Collection Alea-Saclay: Monographs and Texts in Statistical Physics. Cambridge University Press, 2000. DOI: 10.1017/CBO9780511524592.

[14] Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. “Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks”. In: *Phys. Rev. Lett.* 55 (14 Sept. 1985), pp. 1530–1533. DOI: 10.1103/PhysRevLett.55.1530. URL: https://link.aps.org/doi/10.1103/PhysRevLett.55.1530.

[15] Daniel J. Amit. *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press, 1989. DOI: 10.1017/CBO9780511623257.

[16] Stefan Wimbauer, Nikolaus Klemmer, and J. Leo van Hemmen. “Universality of unlearning”. In: *Neural Networks* 7.2 (1994), pp. 261–270. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(94)90020-5. URL: https://www.sciencedirect.com/science/article/pii/0893608094900205.

[17] J. L. van Hemmen and N. Klemmer. “Unlearning and Its Relevance to REM Sleep: Decorrelating Correlated Data”. In: *Neural Network Dynamics*. Ed. by J. G. Taylor et al. London: Springer London, 1992, pp. 30–43. ISBN: 978-1-4471-2001-8.

[18] Francis Crick and Graeme Mitchison. “REM sleep and neural nets”. In: *Behavioural Brain Research* 69.1 (1995). The Function of Sleep, pp. 147–155. ISSN: 0166-4328. DOI: https://doi.org/10.1016/0166-4328(95)00006-F. URL: https://www.sciencedirect.com/science/article/pii/016643289500006F.

[19] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[20] Osame Kinouchi and Renato Kinouchi. “Dreams, endocannabinoids and itinerant dynamics in neural networks: re elaborating Crick-Mitchison unlearning hypothesis”. In: (Aug. 2002).

[21] Alexander Plakhov and S. Semenov. “Neural networks : iterative unlearning algorithm converging to the projector rule matrix”. In: *http://dx.doi.org/10.1051/jp1:1994105* 4 (Feb. 1994). DOI: 10.1051/jp1:1994105.

[22] Serguei Semenov and Irina Shuvalova. “Some results on convergent unlearning algorithm”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky, M.C. Mozer, and M. Hasselmo. Vol. 8. MIT Press, 1995. URL: https://proceedings.neurips.cc/paper/1995/file/4a213d37242bdcad8e7300e202e7caa4-Paper.pdf.

[23] J R L de Almeida and D J Thouless. “Stability of the Sherrington-Kirkpatrick solution of a spin glass model”. In: *Journal of Physics A: Mathematical and General* 11.5 (May 1978), pp. 983–990. DOI: 10.1088/0305-4470/11/5/028. URL: https://doi.org/10.1088/0305-4470/11/5/028.

[24] Ayaka Sakata and Koji Hukushima. “Partial annealing of a coupled mean-field spin-glass model with an embedded pattern”. In: *Phys. Rev. E* 83 (2 Feb. 2011), p. 021105. DOI: 10.1103/PhysRevE.83.021105. URL: https://link.aps.org/doi/10.1103/PhysRevE.83.021105.

[25] Hajime Takayama. “On Phase Diagrams of the Extended Mean Field Model for Spin Glasses –Mean Field Description of Random Network Model for Spin Glass-Ferromagnetic Systems–”. In: *Journal of the Physical Society of Japan* 61.7 (1992), pp. 2512–2521. DOI: 10.1143/JPSJ.61.2512. eprint: https://doi.org/10.1143/JPSJ.61.2512. URL: https://doi.org/10.1143/JPSJ.61.2512.

[26] S.M. Kuva, O. Kinouchi, and N. Caticha. “Learning a spin glass: Determining Hamiltonians from metastable states”. In: *Physica A: Statistical Mechanics and its Applications* 257.1 (1998), pp. 28–35. ISSN: 0378-4371. DOI: https://doi.org/10.1016/S0378-4371(98)00126-5. URL: https://www.sciencedirect.com/science/article/pii/S0378437198001265.

[27] MdShafiful Alam. “Iterative Methods to Solve Systems of Nonlinear ALgebraic Equations”. In: *Master thesis of Western Kentucky University* (2018).