Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

IL MODELLO DI HOPFIELD: UN'APPLICAZIONE DEL MODELLO DI ISING ALLE RETI NEURALI

Relatore:
Prof. Elisa Ercolessi

Correlatore:
Dott.ssa Barbara Bravi

Presentata da:
Silvia Ferri

Anno Accademico 2015/2016

A Luca

Abstract

Il Modello di Hopfield è un tentativo di modellizzare il comportamento di una memoria associativa come proprietà emergente di un network costituito da unità a due stati interagenti tra loro, e costituisce un esempio di come gli strumenti della meccanica statistica possano essere applicati anche al campo delle reti neurali. Nel presente elaborato viene esposta l'analogia tra il Modello di Hopfield e il Modello di Ising nel contesto delle transizioni di fase, applicando a entrambi i modelli la teoria di campo medio. Viene esposta la dinamica a temperatura finita e ricavata e risolta l'equazione di punto a sella per il limite di non saturazione del Modello di Hopfield. Vengono inoltre accennate le principali estensioni del Modello di Hopfield.

Indice

Introduzione

Gli ultimi decenni hanno visto la nascita e lo sviluppo di un'ampia attività di ricerca che va spesso sotto il nome di biofisica: un numero crescente di fisici hanno iniziato a occuparsi di problematiche che un tempo erano dominio di biologi, biochimici o fisiologi.

Ma che cos'è la biofisica? Si pone questa domanda William Bialek [1], che inscrive la fisica tra le discipline definite dai mezzi di indagine e non dall'oggetto di studio: dunque biofisica è l'utilizzo della fisica come stile conoscitivo, e in particolare come indagine teorica che si serve di strumenti formali in cerca di principi esplicativi, applicato a fenomeni legati al mondo vivente.

Uno degli strumenti formali principali è quello dell'analogia come potente mezzo di modellizzazione: tentare di trovare analogie con sistemi già ampiamente spiegati e compresi può far luce su meccanismi nuovi anche in campi ancora poco conosciuti e aiutare a costruire modelli trattabili matematicamente. Nel presente lavoro di tesi si cercherà di evidenziare un esempio paradigmatico di questo meccanismo come mezzo d'indagine: l'analogia tra sistemi magnetici, oggetto di studio della meccanica statistica largamente esplorato nell'ambito delle transizioni di fase, e i tentativi di modellizzare il sistema nervoso per comprenderne i principi di funzionamento. Dunque questo elaborato si inserisce nel contesto del trasferimento di una sempre più vasta applicazione di metodologie propriamente fisiche e matematiche al campo della biologia e, nello specifico, delle neuroscienze. Questo campo, sviluppatosi in particolare a partire dagli anni 40 del Novecento e con una grande accelerazione a partire dagli anni 80, ha visto grandi sforzi di modellizzare i meccanismi cerebrali a partire da modelli basati sulla fisiologia dei neuroni, ma benché varie branche continuino a focalizzarsi sugli aspetti fisiologici e anche psicologici, si tratta di uno dei casi, non rari nel panorama scientifico, in cui il formalismo matematico ha poi preso vita propria aprendo un campo di indagine a parte.

Tra le applicazioni del formalismo fisico e matematico all'ambito biologico la comprensione del funzionamento del sistema nervoso rappresenta un problema di particolare interesse. Si tratta di un oggetto di studio indubbiamente affascinante e ancora in gran parte misterioso; il nostro cervello è il modo in cui la biologia ha risolto in 600 milioni di anni il problema di processare una grande mole di informazioni, caratterizzate da rumore e ridondanza, in un ambiente in costante evoluzione, mediante network costituiti da un grandissimo numero di cellule nervose altamente interconnesse. Il tentativo degli scienziati, che coinvolge una vasta gamma di ruoli tra matematici, fisici, biologi, informa-

tici e persino psicologi, è capire i principi che sottostanno al meccanismo di information processing in queste strutture complesse. Ne emerge un campo di ricerca decisamente interdisciplinare che si interessa allo studio dei meccanismi di elaborazione dell'informazione in network complessi costituiti da unità elementari molto semplici interagenti tra loro, che possono essere i neuroni del sistema nervoso così come unità elettroniche in implementazioni hardware o software ispirate ai sistemi neurali.

Ci si riferisce spesso a questo campo con il termine neural network, tuttavia non bisogna dimenticare che, pur sottolineando con l'aggettivo "neural" il fatto che gran parte dell'ispirazione viene dalle neuroscienze, non c'è nessuna ambizione di descrizione fisiologica fedele del sistema nervoso reale. La modellizzazione del sistema nervoso è un campo a sé, e anche se trae ispirazione da analogie biologiche l'obiettivo principale dei neural network sono il comportamento e le proprietà computazionali di network artificiali. Fare chiarezza sui meccanismi di funzionamento del sistema nervoso si riflette in passi avanti in svariati campi di ricerca legati all'intelligenza artificiale quali computer vision, pattern recognition, speech understanding.

Ma non è questo l'unico scopo: accanto al piano pratico è imprescindibile un piano di riflessione teorica, stimolato dai problemi posti dalla modellizzazione dei neural network che esibiscono un comportamento ricco e non banale, e si può sperare con la riflessione sulla modellizzazione del sistema nervoso di giungere a tracciare linee guida per nuovi modelli in neuroscienze e psicologia cognitiva. Lungi dal fornire una descrizione fedele, la modellizzazione permette proprio in virtù dell'astrazione e della rappresentazione formale di identificare meccanismi esplicativi non evidenti con un semplice approccio descrittivo e spunti che possano fungere da input concettuali per una riflessione successiva. Si tratta quindi di modelli interpretativi e non descrittivi, che si servono di principi teorici e computazionali per esplorare il significato di vari aspetti del funzionamento del sistema nervoso, tentando di rispondere alla domanda sul perché esso opera nel modo che osserviamo; in questa operazione, i modelli interpretativi fungono da ponte di collegamento tra livelli di conoscenza diversi, nonché tra discipline diverse.

Poiché l'ispirazione proviene principalmente dalle neuroscienze, dobbiamo innanzitutto comprendere sul piano biologico le proprietà essenziali dei neuroni e del sistema nervoso nel processare informazioni; questo consente poi di astrarre e realizzare modelli per neural network artificiali che possano essere simulati, indagati e analizzati.

Il sistema nervoso è costituito da un grande numero di cellule interagenti (le stime sono dell'ordine di 10^{11} cellule nervose, ciascuna interagente con all'incirca altre 10^4). Le cellule nervose, o neuroni, sono cellule eccitabili e sono peculiari tra le cellule del nostro organismo per la loro abilità nel propagare segnali rapidamente e su lunghe distanze. Questo avviene tramite la generazione di impulsi elettrici chiamati potenziali d'azione, o più semplicemente *spike*, che possono viaggiare attraverso le fibre nervose; i neuroni rappresentano e trasmettono informazioni inviando sequenze di *spike*. Un neurone è costituito da tre parti: un corpo cellulare, i dendriti (che costituiscono l'input del segnale) e un assone (che costituisce l'output); le estremità dell'assone (che connettono il neurone

cosiddetto pre-sinaptico con più neuroni detti post-sinaptici) si chiamano terminazioni sinaptiche, o semplicemente sinapsi, e la loro forza è variabile secondo un meccanismo detto plasticità sinaptica che è ritenuto alla base del fenomeno dell'apprendimento. Il potenziale d'azione altro non è che una variazione della differenza di potenziale dovuta alla diversa concentrazione di ioni tra interno ed esterno della cellula, separati da una membrana. Esso si propaga lungo l'assone agendo come segnale elettrico, e all'estremità induce il rilascio di neurotrasmettitori, che causano l'apertura selettiva di canali ionici nella membrana del neurone post-sinaptico; a seconda del segno degli ioni che possono attraversare tali canali la differenza di potenziale aumenta o diminuisce, con un effetto rispettivamente eccitatorio o inibitorio. Il potenziale risultante per la cellula post-sinaptica determina, se supera un certo valore di soglia, l'invio da parte di essa di un potenziale d'azione.

Come si può modellizzare dunque questo meccanismo? Sono stati seguiti principalmente due approcci differenti. Da una parte, i tentativi di descrivere il flusso dei segnali elettrici in analogia alla propagazione di segnali in un circuito, assimilando gli elementi costitutivi del sistema nervoso a resistenze, capacità e condensatori; questo approccio muove i passi dal lavoro di Hodgkin e Huxley [2], cui valse il premio Nobel nel 1963. Dall'altra parte, un approccio che vede come precursori all'inizio degli anni 40 McCulloch e Pitts [3], e consiste nel modellizzare il sistema nervoso come un network di elementi interconnessi, che possono assumere solo due stati e funzionano secondo un meccanismo di soglia; tali network sono in grado di eseguire operazioni logiche arbitrarie.

Il modello di McCulloch-Pitts è stato riformulato negli anni 50 da Cragg e Temperley [4] in analogia a un sistema magnetico formato da spin, e qualche anno dopo Caianiello [5] propose una teoria che tentava di definire una "*thinking machine*" che riproducesse alcune caratteristiche del sistema nervoso, sfruttando le regole della meccanica statistica e dando forma al meccanismo anticipato da Hebb [6] per le modificazioni sinaptiche. Perché questo campo ritorni in auge con un vasto impiego delle tecniche della meccanica statistica bisogna aspettare lo straordinario impulso dato dal lavoro di Hopfield [7] negli anni 80, che ha riportato l'attenzione sui neural network con connessioni simmetriche precedentemente abbandonati perché ritenuti poco plausibili sul piano biologico e quindi poco degni di nota.

La prima caratteristica dei sistemi magnetici che ha portato a ritenerli un'utile analogia per il comportamento del sistema nervoso è la dipendenza dello stato del sistema dalla sua storia passata, in quanto è evidente che la storicità è una caratteristica imprescindibile dei sistemi biologici. Questa idea è presente nei sistemi magnetici mediante l'isteresi, ovvero la dipendenza della magnetizzazione in un certo istante dalla magnetizzazione negli istanti precedenti. Altre caratteristiche comuni al sistema nevoso e ai sistemi magnetici sono i concetti di disordine e rumore: il disordine è un concetto tipico dei sistemi magnetici in fenomeni noti come transizioni di fase, e il rumore può essere modellizzato in maniera analoga alla temperatura.

I materiali magnetici consistono in una serie di spin, ovvero unità con un momento di

dipolo magnetico intrinseco di natura quantistica, disposti in un reticolo che rappresenta la struttura cristallina del materiale, che può essere messo in analogia con il network costituito dai neuroni.

Questo tipo di sistemi sono descritti dalla meccanica statistica, e ci sono caratteristiche del sistema nervoso quali il grande numero di unità costituenti e la natura stocastica dei processi neurali che portano a supporre di poter estendere il formalismo della meccanica statistica anche alla descrizione del sistema nervoso. Essa si occupa infatti di descrivere il comportamento collettivo di un numero molto grande di unità microscopiche semplici, interagenti tra loro. Lo scopo della meccanica statistica è studiare proprietà emergenti a livello macroscopico dall'interazione tra le unità microscopiche, e il mezzo utilizzato è la descrizione probabilistica. Essa si rivela uno strumento potente non tanto per colmare la mancata conoscenza dei dettagli microscopici, quanto per far emergere proprietà macroscopiche che non sono ottenibili né indagabili se non con il cambio di prospettiva che capovolge l'interpretazione meccanicistica e deterministica in una probabilistica.

Un'area di ricerca relativamente recente della meccanica statistica è quella delle transizioni di fase, fenomeni indubbiamente interessanti e onnipresenti in natura che dal punto di vista sperimentale danno vita a una fenomenologia estremamente vasta, e dal punto di vista teorico possono essere sfruttati per spiegare una grande varietà di comportamenti.

È in questo ampio quadro, al crocevia tra neuroscienze e meccanica statistica, che si inserisce il presente elaborato, fornendo un esempio dell'analogia precedentemente descritta tra due modelli entrambi paradigmatici per gli ambiti cui appartengono. Il primo membro dell'analogia, cui sarà dedicato il primo capitolo, è il Modello di Ising, che si colloca all'interno di una disciplina ormai ampiamente approfondita e consolidata e che dispone di mezzi di indagine, come quelli della meccanica statistica all'equilibrio, molto potenti. All'interno di questa disciplina il Modello di Ising ricopre una posizione particolarmente importante sia per il ruolo fondativo che ha avuto nello sviluppo della meccanica statistica dei sistemi magnetici, sia per le innumerevoli applicazioni che ne sono seguite e modelli che si riconducono ad esso come caso semplice e risolvibile in maniera esatta sotto opportune condizioni. Il Modello di Ising verrà qui presentato nel contesto delle transizioni di fase, discutendo il ruolo della magnetizzazione e in particolare la teoria di campo medio, che fornisce una notevole semplificazione del problema. Verrà inoltre presentato, accanto alla descrizione della meccanica statistica all'equilibrio, un approccio dinamico che studia come tale equilibrio si forma, permettendo di tenere conto di elementi stocastici.

Dall'altra parte, il secondo membro dell'analogia, cui verrà dedicato il secondo capitolo, è il Modello di Hopfield, caratterizzato anch'esso da un ruolo paradigmatico, in quanto è stato il precursore di innumerevoli studi sui neural network che sono fioriti dagli anni 80 in poi. Rispetto alla meccanica statistica, questo è un campo più recente e meno consolidato, ma che sta ricevendo recentemente un grande impulso e nuova attenzione anche grazie agli strumenti oggi disponibili sul piano computazionale. Nel secondo capi-

tolo il Modello di Hopfield verrà presentato nella sua forma più semplice, seguendo sia un approccio di tipo dinamico che un approccio di tipo energetico, descrivendo come si può modellizzare un meccanismo di memoria associativa in maniera formalmente analoga al Modello di Ising. La teoria di campo medio discussa per il Modello di Ising verrà applicata a questo sistema ottenendo una transizione di fase che descrive il passaggio da una memoria funzionante secondo i meccanismi della memoria associativa a una memoria non più funzionante secondo tali meccanismi. Una memoria funzionante è in grado di immagazzinare pattern come attrattori dinamici e di recuperarli, e verrà descritto come ciò avviene al variare dei parametri che regolano il comportamento del sistema, ovvero la temperatura (opportunamente definita) e il numero di pattern immagazzinati.

Infine, nel terzo capitolo verranno presentate alcune estensioni di questo modello, che mirano da un lato a renderlo più plausibile sul piano biologico e dall'altro a facilitare implementazioni hardware. Verrà sottolineato come le caratteristiche principali del Modello di Hopfield, nonostante emergano da un modello estremamente semplificato, siano resistenti a modifiche di vario tipo e possano quindi configurarsi come interessanti direzioni verso cui indirizzare ragionamenti e modelli successivi.

Capitolo 1

Il Modello di Ising

Il Modello di Ising costituisce un esempio paradigmatico di un modello in cui è possibile studiare le transizioni di fase, interessanti fenomeni della meccanica statistica generalmente complicati da descrivere dal punto di vista matematico, che risultano in questo caso trattabili in maniera esatta in due dimensioni in assenza di campo magnetico esterno e in una dimensione anche in presenza di un campo magnetico esterno.

Verrà di seguito presentata una introduzione al Modello di Ising, dopo aver delineato il quadro delle transizioni di fase in cui esso si inserisce. Verrà posta l'attenzione in particolare sulla magnetizzazione spontanea, che dipende dalla dimensionalità del modello considerato, e che si dimostrerà essere assente in una dimensione e presente in due dimensioni.

Il Modello di Ising verrà trattato secondo la teoria di campo medio, un metodo approssimato che fornisce soluzioni esatte nel limite in cui ogni unità interagisce con tutte le altre, mentre nel caso di interazioni locali l'accuratezza di tale approssimazione dipende dalle dimensioni d del modello, che diventa esatto per $d \geq 4$.

Infine, verrà affrontato un approccio differente, che invece di studiare la meccanica statistica dei sistemi all'equilibrio studia come tale equilibrio si forma dinamicamente: il vantaggio questo tipo di trattazione è che permette di tenere conto della stocasticità della dinamica. Verrà presentata la dinamica di Glauber come esempio di regola dinamica stocastica che soddisfa la cosiddetta condizione di *detailed balance*, necessaria affinché all'equilibrio si recuperi la distribuzione di Boltzmann.

1.1 Transizioni di fase

Le transizioni di fase costituiscono uno dei fenomeni di principale interesse della meccanica statistica e della termodinamica, e possono essere definite come il passaggio di un sistema termodinamico da una fase a un'altra in seguito alla variazione di parametri esterni, caratterizzato dal brusco cambiamento di una o più proprietà fisiche. Esse sono onnipresenti in natura, dal tipico esempio del passaggio tra le varie fasi di aggregazione

della materia (solido, liquido, gas), ai fenomeni della superfluidità e superconduttività, alle transizioni nei materiali ferromagnetici, che sono descritte dal Modello di Ising e saranno oggetto della presente trattazione. Verrà qui seguita la trattazione di Huang [8].

Il punto critico in cui avviene la transizione è definito da una temperatura critica T_c corrispondente a una discontinuità, che spesso individua due fasi con diversa simmetria spaziale. In questo caso è evidente che il punto a cui avviene la transizione deve costituire una discontinuità, in quanto a causa del fatto che differiscono per la presenza o assenza di simmetria le due fasi dovranno essere descritte da funzioni diverse delle variabili termodinamiche, che non potranno essere continuate analiticamente nel punto critico. Per studiare una transizione di fase con cambio di simmetria è utile definire un **parametro d'ordine**, ovvero una grandezza fisica che rappresenta la principale differenza qualitativa tra le due fasi. Nel punto critico le due fasi coesistono, e dunque il parametro d'ordine sarà nullo; in prossimità del punto critico esso assume valori molto piccoli e può essere usato come parametro di espansione perturbativa nella descrizione dei fenomeni critici.

Nel caso delle transizioni di fase nei ferromagneti, il parametro d'ordine è la magnetizzazione $\mathcal{M}$. Per $T > T_c$ non c'è magnetizzazione spontanea, e dunque $\mathcal{M} = 0$: non c'è una direzione preferenziale e il sistema presenta simmetria rotazionale. Per $T < T_c$ si ha una rottura dell'invarianza rotazionale, e il sistema presenta una direzione privilegiata: proprio a causa di questa rottura di simmetria abbiamo bisogno di un parametro aggiuntivo per descrivere questa fase, dato appunto dalla magnetizzazione $\mathcal{M}$ che non è più nulla.

Quando il parametro d'ordine, che ci permette di descrivere compiutamente il sistema in prossimità del punto critico, varia di $d\mathcal{M}$, il lavoro compiuto sul sistema è

$$dL = H d\mathcal{M} \,. \tag{1.1}$$

Questa equazione definisce il campo coniugato H, che nel caso del ferromagnetismo è il campo magnetico esterno. Possiamo usare H e T come variabili termodinamiche indipendenti per descrivere il nostro sistema. Le funzioni termodinamiche possono essere derivate da un potenziale termodinamico definito in generale da $\phi(H,T) = e^{-\beta\phi(H,T)}$; trattando sistemi magnetici risulta comodo usare come potenziale termodinamico l'libera di Gibbs $G(H,T)$. In funzione di $G(H,T)$ si possono quindi ottenere:

$$M = -\frac{\partial G}{\partial H} \qquad \text{[Magnetizzazione]} \tag{1.2}$$

$$\chi = \frac{1}{V}\frac{\partial M}{\partial H} \qquad \text{[Suscettività]} \tag{1.3}$$

$$E = G - T\frac{\partial G}{\partial T} \qquad \text{[Energia interna]} \tag{1.4}$$

$$C = -T\frac{\partial^2 G}{\partial T^2} \qquad \text{[Calore specifico]} \tag{1.5}$$

Altre informazioni utili sulle transizioni di fase sono contenute nella **funzione di correlazione**, che esprime la correlazione spaziale tra variabili microscopiche (gli spin nel caso di un sistema ferromagnetico). Assumendo che ci sia una densità spaziale di parametro d'ordine $m(\mathbf{r})$, dove $\mathbf{r}$ è il vettore posizione, si può ottenere la magnetizzazione come media sul volume e media termodinamica (indicata dalle parentesi angolate):

$$\mathcal{M} = \left\langle \int d^3 r m(\mathbf{r}) \right\rangle . \tag{1.6}$$

La funzione di correlazione è definita come

$$\Gamma(\mathbf{r}) = \langle m(\mathbf{r}) m(0) \rangle - \langle m(\mathbf{r}) \rangle \langle m(0) \rangle . \tag{1.7}$$

Punti vicini nel sistema tendono ad essere correlati; lontano dal punto critico la correlazione si estende a una certa distanza ξ, detta lunghezza di correlazione. Essa può essere definita mediante il comportamento asintotico della funzione di correlazione

$$\Gamma(\mathbf{r}) \sim e^{-\frac{r}{\xi}} \qquad T \neq T_c . \tag{1.8}$$

Molte quantità termodinamiche divergono al punto critico. Introducendo il parametro

$$t = \frac{T - T_c}{T_c} \tag{1.9}$$

possiamo pensare le grandezze termodinamiche, per $t \to 0$, come scomponibili in una parte regolare, che rimane finita, e una parte singolare, che può essere divergente o avere derivate divergenti; studiando la dipendenza da t della parte singolare, si ottengono delle leggi di potenza parametrizzate da un insieme di **esponenti critici**, che risultano gli stessi per vaste classi di fenomeni dette classi di universalità. Infatti le proprietà termodinamiche di un sistema vicino a una transizione di fase dipendono solo da un piccolo numero di caratteristiche, come la dimensionalità e la simmetria, e non dalle proprietà microscopiche del sistema. Si riportano di seguito le leggi che definiscono i sei esponenti critici comunemente riconosciuti, indicati con le lettere greche α, β, γ, δ, η, ν

$$C \sim |t|^{-\alpha} \tag{1.10}$$

$$M \sim |t|^{\beta} \tag{1.11}$$

$$\chi \sim |t|^{-\gamma} \tag{1.12}$$

$$M \sim H^{-\frac{1}{\delta}} \tag{1.13}$$

$$\Gamma(r) \sim r^{-(d-2+\eta)} e^{-\frac{r}{\xi}} \tag{1.14}$$

$$\xi \sim |t|^{-\nu} \tag{1.15}$$

dove d nella (1.14) è la dimensionalità dello spazio.

Le transizioni di fase possono essere classificate in base al tipo di discontinuità che si presenta al punto critico. Una prima classificazione si deve a Ehrenfest, che individua come transizioni di fase di primo ordine quelle in cui si ha una discontinuità finita in una delle derivate prime dell'energia libera, e come transizioni di secondo ordine quelle in cui la discontinuità si presenta in una delle derivate seconde. Tuttavia si tratta di un metodo inesatto perché non considera il caso in cui una delle derivate tenda a infinito.

La classificazione moderna descrive invece come transizioni di fase di primo ordine quelle che coinvolgono un calore latente (al punto della transizione il sistema assorbe o rilascia calore pur mantenendo costante la temperatura) e che sono legate a una discontinuità dell'entropia. Queste transizioni sono caratterizzate da un valore finito della lunghezza di correlazione. Alcuni esempi sono le transizioni solido-liquido e liquido-gas, e la condensazione di Bose-Einstein. Le transizioni di secondo ordine sono dette anche transizioni continue in quanto l'entropia non è discontinua, ma è non analitica per $T = T_c$. Questo tipo di transizioni non coinvolgono un calore latente, e la lunghezza di correlazione diverge nel punto critico. Sono transizioni di secondo ordine la superconduttività, la superfluidità e il ferromagnetismo.

1.2 Introduzione al Modello di Ising

La trattazione delle transizioni di fase non è semplice dal punto di vista matematico, e solo pochi modelli possono essere risolti in maniera esatta senza pesanti sforzi di calcolo numerico. Tra questi, il Modello di Ising, che è diventato un modello base per la meccanica statistica proprio per la possibilità di essere risolto in maniera esatta in due dimensioni in assenza di campo esterno e in una dimensione anche in presenza di campo esterno, cui si aggiunge una grandissima versatilità che lo rende applicabile a vari tipi di problemi. Esso fu sviluppato inizialmente per descrivere le transizioni di fase nei ferromagneti, ma è stato poi applicato anche per descrivere altri fenomeni, dalle transizioni ordine-disordine nelle leghe binarie, alla descrizione degli spin glasses (metalli che presentano una struttura amorfa invece che cristallina), fino ad applicazioni alle reti neurali, come verrà esposto nel Capitolo 2 in cui si analizzerà il Modello di Hopfield.

Il Modello di Ising costituisce un tentativo di simulare la struttura fisica di un materiale ferromagnetico, che viene modellizzato come un reticolo di N punti, detti siti. A ogni sito è associata una variabile di spin S_i $i = 1, ..., N$ che può assumere solo due valori: $S_i = \pm 1$, detti rispettivamente spin up e spin down. La configurazione del sistema è data dall'insieme dei valori di tutti gli spin, indicato con il vettore $\mathbf{S} = (S_1, S_2, ..., S_N)$, e l'energia del sistema nella configurazione $\mathbf{S}$ è descritta dalla seguente Hamiltoniana

$$\mathcal{H}(\mathbf{S}) = -\sum_{\langle ij \rangle} J_{ij} S_i S_j - \sum_{i=1}^{N} H_i S_i, \tag{1.16}$$

dove H_i è il campo magnetico esterno di cui risente l'i-esimo spin, e con $\langle ij \rangle$ si indicano le coppie di primi vicini, contate ciascuna una volta sola. Poiché $\langle ij \rangle = \langle ji \rangle$, la somma è estesa a $\gamma \frac{N}{2}$ termini, dove γ è il numero di primi vicini di un sito, detto numero di coordinazione del reticolo.

J_{ij} è l'integrale di scambio che esprime l'energia dovuta all'interazione quantomeccanica tra una coppia di spin degli elettroni, e che è una conseguenza dalla natura antisimmetrica della funzione d'onda che descrive una coppia di elettroni. Il valore dell'integrale di scambio, che è quindi il responsabile delle interazioni ferromagnetiche, è

$$J_{ij} = \int \psi_j^*(1)\psi_i^*(2)U_{ij}\psi_j(2)\psi_i(1)\, d^3\vec{r}_1 d^3\vec{r}_2, \tag{1.17}$$

dove ψ_i è la funzione d'onda che rappresenta l'i-esimo elettrone, e U_{ij} è il potenziale coulombiano che agisce tra due elettroni. L'interazione di scambio decresce rapidamente con la distanza, dunque possiamo limitarci a considerarla solo tra primi vicini. Nel caso di interazioni isotrope J_{ij} è una costante, e se inoltre il campo magnetico esterno è costante l'Hamiltoniana diventa

$$\mathcal{H}(\mathbf{S}) = -J \sum_{\langle ij \rangle} S_i S_j - H \sum_{i=1}^{N} S_i. \tag{1.18}$$

Si può notare che se $J > 0$ (materiali ferromagnetici) una coppia di spin allineati contribuisce a far diminuire l'energia, mentre una coppia di spin anti-allineati la fa aumentare. Dunque il principio di minimizzazione dell'energia tende a far allineare gli spin, mentre la presenza della temperatura tende a causare una configurazione disordinata e quindi a massimizzare l'entropia: si tratta dei due principi in competizione nelle transizioni di fase. Nel caso dei materiali anti-ferromagnetici ($J < 0$) la configurazione favorita dal principio di minimizzazione dell'energia presenta invece ogni spin anti-allineato rispetto ai vicini.

La funzione di partizione del sistema è data da

$$Z(H,T) = \sum_{S_1=\pm 1} \sum_{S_2=\pm 1} \cdots \sum_{S_N=\pm 1} e^{-\beta \mathcal{H}(\mathbf{S})}. \tag{1.19}$$

Dalla funzione di partizione possiamo ricavare l'energia libera di Gibbs

$$G(H,T) = -k_B T \ln Z(H,T). \tag{1.20}$$

A partire da queste due funzioni si possono calcolare le funzioni termodinamiche che

descrivono il sistema

$$E(H,T) = -\frac{\partial \ln Z(H,T)}{\partial \beta} \qquad \text{[Energia interna]} \tag{1.21}$$

$$C(H,T) = \frac{\partial E}{\partial T} \qquad \text{[Calore specifico a volume costante]} \tag{1.22}$$

$$M(H,T) = -\frac{\partial G}{\partial H} = \left\langle \sum_{i=1}^{N} S_i \right\rangle \qquad \text{[Magnetizzazione]} \tag{1.23}$$

dove le parentesi angolate nell'ultima uguaglianza indicano la media termodinamica. Queste grandezze sono estensive, ovvero scalano linearmente con il numero N di unità che costituiscono il sistema (ricordando che nel limite termodinamico occorre mandare a ∞ simultaneamente N e il volume V in modo che il rapporto $\frac{N}{V}$ rimanga costante). Può essere comodo definire corrispondenti grandezze intensive dividendo per N, e risulta in particolare utile definire la magnetizzazione per spin

$$m(H,T) = \frac{M(H,T)}{N} \,. \tag{1.24}$$

1.3 Magnetizzazione spontanea

Facendo riferimento alla magnetizzazione come definita nell'equazione (1.23), la quantità $M(0,T)$ è detta magnetizzazione spontanea e se è non nulla il materiale è ferromagnetico, ovvero presenta una magnetizzazione anche in assenza di campo esterno. Come già menzionato, si tratta del parametro d'ordine delle transizioni di fase nei materiali magnetici. Per quanto riguarda le proprietà di simmetria, al di sopra di una certa temperatura critica T_c, detta temperatura di Curie, il sistema non ha una direzione privilegiata e non c'è magnetizzazione spontanea. A $T \leq T_c$ avviene un fenomeno noto come rottura di simmetria: il sistema acquista una direzione privilegiata e la magnetizzazione spontanea non è più nulla. Siamo quindi interessati ad analizzare la magnetizzazione in assenza di un campo esterno.

Notiamo che se $H = 0$ l'Hamiltoniana

$$\mathcal{H}(\mathbf{S}) = -J \sum_{\langle ij \rangle} S_i S_j \tag{1.25}$$

è invariante per trasformazioni $S_i \to -S_i \quad \forall i$ in cui vengono invertiti simultaneamente tutti gli spin, dunque la magnetizzazione

$$M(H,T) = \left\langle \sum_{i=1}^{N} S_i \right\rangle = \frac{1}{Z} \sum_{\mathbf{S}} e^{-\beta \mathcal{H}(\mathbf{S})} \sum_{i=1}^{N} S_i \tag{1.26}$$

risulta nulla, in quanto il contributo di ogni configurazione $\mathbf{S} = (S_1, ..., S_N)$ viene cancellato da quello della configurazione $-\mathbf{S} = (-S_1, ..., -S_N)$. Tuttavia il modo corretto di calcolare la magnetizzazione spontanea è calcolare il limite termodinamico in presenza di un H arbitrariamente piccolo, e solo successivamente considerare il limite $H \to 0^{\pm}$.

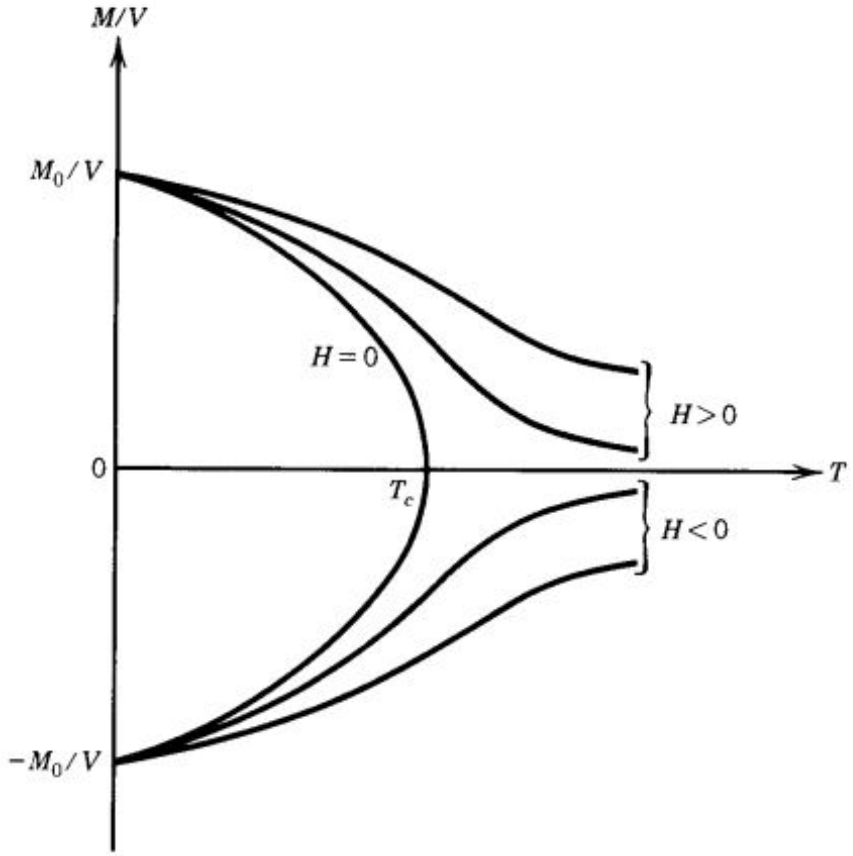

Figura 1.1: Magnetizzazione spontanea per unità di volume in funzione della temperatura.

In Figura 1.1 è rappresentato l'andamento della magnetizzazione spontanea per unità di volume $\frac{M}{V}$ in funzione della temperatura: per $H \to 0^+$ e $H \to 0^-$, $\frac{M}{V}$ si avvicina all'uno o all'altro dei due rami della curva per $H = 0$, e risulta quindi evidente che non è corretto fare la media dei due rami, che darebbe zero.

1.3.1 Assenza di magnetizzazione spontanea in 1 dimensione

Nel Modello di Ising la presenza di magnetizzazione spontanea dipende dalla dimensione del reticolo che si sta considerando.

Consideriamo un Modello di Ising 1-dimensionale e analizziamo la configurazione in cui gli spin sono tutti allineati: l'energia descritta dall'Hamiltoniana (1.25) si trova in un minimo assoluto, mentre l'entropia è nulla. Immaginiamo ora di creare una parete di dominio, ovvero di invertire tutti gli spin alla destra di un certo sito, come mostrato in Figura 1.2.

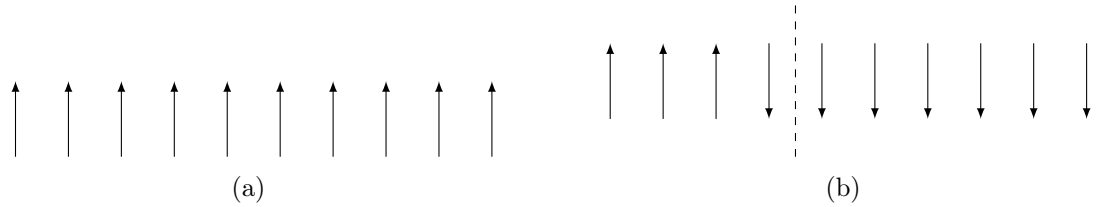

Figura 1.2: Creazione di una parete di dominio in 1 dimensione.

L'energia aumenta di $2J$, mentre l'entropia aumenta di $k_B \ln(N-1)$ poiché abbiamo $N-1$ modi di scegliere il punto in cui situare la parete di dominio. La variazione di energia libera è

$$\Delta F = \Delta E - T\Delta S = 2J - k_B T \ln(N-1). \tag{1.27}$$

Dunque, per $T > 0$ e $N \to \infty$ la creazione di domini abbassa l'energia libera, e sarà quindi favorita: avverrà spontaneamente la creazione di nuovi domini, finché gli spin non saranno tutti disposti in maniera casuale, processo esattamente opposto alla magnetizzazione spontanea.

1.3.2 Esistenza di magnetizzazione spontanea in 2 dimensioni

Consideriamo una configurazione arbitraria degli spin in un reticolo 2-dimensionale, e seguiamo la trattazione di Griffith [9]. In questo caso la definizione di una parete di dominio non è immediata come nel caso 1-dimensionale; consideriamo come parete una linea continua tracciata tra spin up e spin down e indichiamo con b la sua lunghezza, ovvero il numero di spazi reticolari che attraversa. Per rendere la definizione unica occorre definire anche un verso, che scegliamo in maniera tale che gli spin down stiano a sinistra e gli spin up a destra; nel caso in cui ci sia ancora ambiguità scegliamo la parete che si piega verso destra. In questo modo otteniamo pareti che non si incrociano mai; domini con la stessa forma ma situati in posizioni diverse sono considerati diversi, domini con stessa forma e posizione ma diverso verso di percorrenza sono considerati diversi. Un esempio di pareti di dominio ottenute in questo modo è riportato in Figura 1.3.

Immaginiamo ora di applicare un campo esterno con segno positivo, la cui influenza diventa arbitrariamente piccola nel limite di un reticolo infinito: imponiamo la condizione al contorno che tutti gli spin sul bordo siano up, allineati col campo esterno, e si ha quindi una rottura di simmetria. Vogliamo dimostrare che c'è magnetizzazione spontanea (a

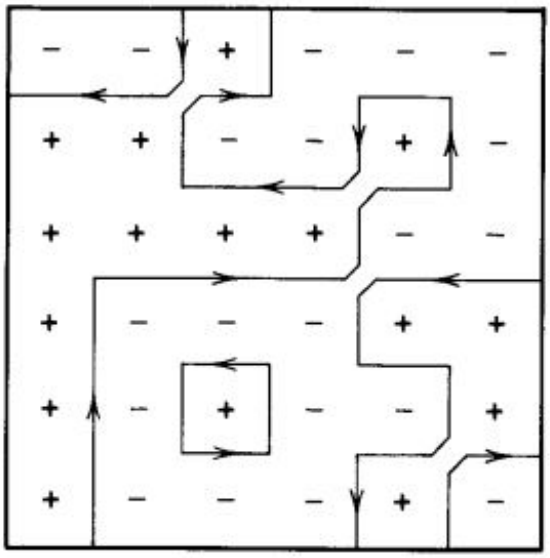

Figura 1.3: Esempio di pareti di dominio in 2 dimensioni.

temperatura sufficientemente bassa ma finita), ovvero che il numero medio di spin down è minore della metà:

$$\frac{\langle N_- \rangle}{N} < \frac{1}{2}\,. \tag{1.28}$$

Grazie alla condizione al contorno abbiamo solo domini chiusi; classifichiamo tutti i possibili domini con l'etichetta (b, i), dove b è la lunghezza del dominio, e i indicizza le varie curve all'interno della classe delle curve di lunghezza b. Una parete di lunghezza b è in generale un poligono formato da b segmenti di lunghezza unitaria, disposti in orizzontale oppure in verticale; notiamo che per avere una curva poligonale chiusa b non deve essere minore di 4 e deve essere pari. Se ricollochiamo i segmenti in modo da ottenere una figura il più possibile vicina a un quadrato, sarà sicuramente contenuta nel quadrato di lato $\frac{b}{4}$: dunque l'area contenuta in un dominio di lunghezza b sarà necessariamente $\leq \frac{b^2}{16}$.

Indichiamo con $m(b)$ il numero di pareti di lunghezza b (tenendo conto della condizione al contorno), e cerchiamo una condizione che limiti superiormente $m(b)$. Poiché, nel creare una parete di lunghezza b senza tenere conto della condizione al contorno, abbiamo N modi di posizionare il primo segmento e 3 modi (verso l'alto, verso il basso o orizzontale) di posizionare ciascuno dei restanti $(b-1)$ segmenti, avremo sicuramente

$$m(b) < 3^{b-1} N\,. \tag{1.29}$$

Per una configurazione fissata, definiamo

$$X(b,i) = \begin{cases} 1 & \text{se la parete } (b,i) \text{ è presente nella configurazione} \\ 0 & \text{altrimenti} \end{cases} . \tag{1.30}$$

Nella configurazione in esame, il numero N_- di spin down (che coincide con l'area di un dominio) soddisfa la relazione

$$N_- \leq \sum_{\substack{b\geq 4 \\ b\, pari}} \frac{b^2}{16} \sum_{i=1}^{m(b)} X(b,i) \, . \tag{1.31}$$

Calcoliamo ora la media termodinamica di $X(b,i)$:

$$\langle X(b,i)\rangle = \frac{\sum'_{\mathbf{S}} e^{-\beta\mathcal{H}(\mathbf{S})}}{\sum_{\mathbf{S}} e^{-\beta\mathcal{H}(\mathbf{S})}}, \tag{1.32}$$

dove l'apice nella prima sommatoria indica che è ristretta alle configurazioni in cui compare $X(b,i)$. Consideriamo una configurazione $\mathbf{C}$ fissata in cui compare $X(b,i)$, e una configurazione $\widetilde{\mathbf{C}}$ ottenuta a partire da $\mathbf{C}$ invertendo tutti gli spin all'interno del dominio (b,i). Le energie di queste due configurazioni sono legate dalla relazione

$$\mathcal{H}(\mathbf{C}) = \mathcal{H}(\widetilde{\mathbf{C}}) + 2Jb \, , \tag{1.33}$$

dunque

$$\langle X(b,i)\rangle = \frac{\sum'_{\mathbf{S}} e^{-\beta\mathcal{H}(\mathbf{S})}}{\sum_{\mathbf{S}} e^{-\beta\mathcal{H}(\mathbf{S})}} \leq e^{-2\beta Jb} \leq e^{-\beta Jb}. \tag{1.34}$$

Torniamo ora alla (1.31) e ne calcoliamo la media termodinamica, sfruttando la disuguaglianza appena trovata e ricordando la (1.29)

$$\langle N_-\rangle \leq \sum_{\substack{b\geq 4 \\ b\, pari}} \frac{b^2}{16} \sum_{i=1}^{m(b)} \langle X(b,i)\rangle \leq \sum_{\substack{b\geq 4 \\ b\, pari}} \frac{b^2}{16} m(b) e^{-\beta Jb} \leq \frac{N}{48} \sum_{\substack{b\geq 4 \\ b\, pari}} b^2 3^b e^{-\beta Jb} \, . \tag{1.35}$$

Effettuando il cambio di variabile $b = 2b' + 4$ per ottenere un indice che assuma tutti i valori interi positivi, questa espressione diventa

$$\frac{\langle N_-\rangle}{N} = \frac{1}{12} \sum_{b=0}^{+\infty} (b+2)^2 (9e^{-2\beta J})^2 (9e^{-2\beta J})^b \, . \tag{1.36}$$

Ponendo per semplicità

$$x = 9e^{-2\beta J} \tag{1.37}$$

ed esprimendo i termini in x^b come derivate di opportune potenze di x, si ha:

$$\begin{aligned}\frac{\langle N_- \rangle}{N} &\leq \frac{1}{12}x^2 \left[\frac{d^2}{dx^2} \left(\sum_{b=0}^{+\infty} x^{b+2} \right) + \frac{d}{dx} \left(\sum_{b=0}^{+\infty} x^{b+1} \right) + \sum_{b=0}^{+\infty} x^b \right] \\ &= \frac{1}{12}x^2 \left[\frac{d^2}{dx^2} \left(\frac{x^2}{1-x} \right) + \frac{d}{dx} \left(\frac{x}{1-x} \right) + \frac{1}{1-x} \right] \\ &= \frac{x^2}{3(1-x)^3} \left(1 - \frac{3}{4}x + \frac{1}{4}x^2 \right) , \end{aligned} \tag{1.38}$$

dove nel penultimo passaggio si è sfruttata la serie geometrica $\sum_{b=0}^{\infty} x^b = 1/(1-x)$. La quantità così ottenuta tende a 0 per $x \to 0$ (ovvero per $\beta \to \infty$). Dunque, per definizione di limite, scegliendo in maniera opportuna δ e ϵ, abbiamo che per $\epsilon = \frac{1}{2} \quad \exists\, \delta > 0$ t.c., per $|x - x_0| < \delta$ (ovvero, ricordando la sostituzione (1.37), per β sufficientemente grande), $\frac{\langle N_- \rangle}{N} < \frac{1}{2}$. Ovvero, la condizione (1.28) è soddisfatta e abbiamo dimostrato che esiste magnetizzazione spontanea, poiché il numero medio di spin down è minore della metà, il che significa che si ha un allineamento preferenziale degli spin con il campo esterno.

1.4 Teoria di campo medio

L'Hamiltoniana del Modello di Ising in assenza di un campo esterno può essere riscritta come

$$\mathcal{H} = -\sum_i h_i S_i \tag{1.39}$$

con

$$h_i = \frac{J}{2} \sum_{j \in \langle i \rangle} S_j \, , \tag{1.40}$$

dove $j \in \langle i \rangle$ indica i primi vicini di un dato spin i. Ovvero, ogni spin i interagisce con il campo locale h_i dovuto ai suoi primi vicini; si tratta di un problema in generale difficile da trattare, perché tale campo locale è disomogeneo, e ognuno dei primi vicini dello spin i-esimo ha un valore variabile, a sua volta influenzato dai propri primi vicini. Per semplificare la trattazione si ricorre all'approssimazione di campo medio, che presenteremo seguendo la trattazione di Greiner, Neise, Stöcker [10].

Nell'approssimazione di campo medio ogni spin è pensato come singolarmente interagente con un campo medio omogeneo dovuto alla magnetizzazione media dei suoi γ primi vicini

$$h = \langle h_i \rangle = J\gamma \langle S \rangle \, . \tag{1.41}$$

Tale approssimazione è accurata nel limite in cui ogni spin interagisce con tutti gli altri, mentre per sistemi con interazioni solo tra primi vicini l'accuratezza della teoria di campo

medio dipende dalla dimensionalità d del reticolo: l'approssimazione non è una buona per $d = 1, 2$, migliora per $d = 3$, diventa accurata per $d \geq 4$.

Sfruttando la seguente identità

$$(S_i - \langle S_i \rangle)(S_j - \langle S_j \rangle) = S_i S_j - S_i \langle S_j \rangle - S_j \langle S_i \rangle + \langle S_i \rangle \langle S_j \rangle \tag{1.42}$$

da cui

$$S_i S_j = S_i \langle S_j \rangle + S_j \langle S_i \rangle - \langle S_i \rangle \langle S_j \rangle + (S_i - \langle S_i \rangle)(S_j - \langle S_j \rangle) \ , \tag{1.43}$$

e tenendo conto che $\langle S_i \rangle = \langle S_j \rangle = \langle S \rangle$ poiché la media termodinamica non dipende dal sito che stiamo considerando, l'Hamiltoniana diventa

$$\begin{aligned} \mathcal{H}(\mathbf{S}) &= -J \sum_{\langle ij \rangle} [S_i \langle S \rangle + S_j \langle S \rangle - \langle S \rangle^2 + (S_i - \langle S \rangle)(S_j - \langle S \rangle)] \\ &= -J \langle S \rangle \gamma \sum_{i=1}^{N} S_i + J \frac{\gamma}{2} N \langle S \rangle^2 - J \sum_{\langle ij \rangle} (S_i - \langle S \rangle)(S_j - \langle S \rangle) \ . \end{aligned} \tag{1.44}$$

Nel primo termine riconosciamo h, il campo medio generato dai vicini, mentre il secondo termine è una costante che non dipende da una orientazione specifica, e l'ultimo termine tiene conto delle deviazioni degli spin dal proprio valor medio: l'approssimazione di campo medio consiste nel trascurare quest'ultimo termine. L'Hamiltoniana di campo medio risulta quindi

$$\mathcal{H}^{cm}(\mathbf{S}) = -J \langle S \rangle \gamma \sum_{i=1}^{N} S_i + J \frac{\gamma}{2} N \langle S \rangle^2 \ . \tag{1.45}$$

Se consideriamo il valore di aspettazione otteniamo

$$\langle \mathcal{H}^{cm}(\mathbf{S}) \rangle = -J \langle S \rangle \gamma N \langle S \rangle \gamma + J \frac{\gamma}{2} N \langle S \rangle^2 = -J \frac{\gamma}{2} N \langle S \rangle^2 \ . \tag{1.46}$$

In presenza di un campo esterno H^{ext}, l'Hamiltoniana completa di campo medio sarà data dalla somma di due Hamiltoniane di singola particella, e dunque risulterà notevolmente semplificata rispetto alla trattazione esatta:

$$\mathcal{H}^{cm} = J \frac{\gamma}{2} N \langle S \rangle^2 - (h + H^{ext}) \sum_{i=1}^{N} S_i \ . \tag{1.47}$$

La magnetizzazione risulta

$$M = \left\langle \sum_{i=1}^{N} S_i \right\rangle = N \langle S \rangle \ , \tag{1.48}$$

oppure può essere calcolata a partire dall'energia libera

$$M = -\frac{\partial}{\partial H} G(N, H, T, \langle S \rangle)\Big|_{N,H,T,\langle S \rangle} . \tag{1.49}$$

Da queste due espressioni si ottiene una equazione implicita per $\langle S \rangle$, detta equazione di autoconsistenza

$$N\langle S \rangle = -\frac{\partial}{\partial H} G(N, H, T, \langle S \rangle)\Big|_{N,H,T,\langle S \rangle} , \tag{1.50}$$

per esplicitare la quale dovremo calcolare G, a partire dalla funzione di partizione Z. Calcoliamo quindi la funzione di partizione, a partire dall'Hamiltoniana di campo medio (1.47)

$$\begin{aligned} Z(N, H, T, \langle S \rangle) &= \sum_{S_1=\pm 1} \cdots \sum_{S_N=\pm 1} e^{-\beta J \frac{\gamma}{2} N \langle S \rangle^2} e^{\beta(h+H^{ext}) \sum_{i=1}^{N} S_i} \\ &= e^{-\beta J \frac{\gamma}{2} N \langle S \rangle^2} \left[\sum_{S=\pm 1} e^{\beta(h+H^{ext}) S_1} \right]^N \\ &= e^{-\beta J \frac{\gamma}{2} N \langle S \rangle^2} \left\{ 2 \cosh[\beta(h + H^{ext})] \right\}^N . \end{aligned} \tag{1.51}$$

L'energia libera risulta

$$G = -k_B T \ln Z = \frac{N}{2} J \langle S \rangle^2 - N k_B T \ln\{2 \cosh[\beta(h + H^{ext})]\} . \tag{1.52}$$

Possiamo quindi calcolare la magnetizzazione

$$M = -\frac{\partial G}{\partial H} = N \tanh[\beta(h + H^{ext})] \tag{1.53}$$

da cui l'equazione di autoconsistenza diventa

$$\langle S \rangle = \tanh[\beta(h + H^{ext})] = \tanh[\beta(J\gamma \langle S \rangle + H^{ext})] . \tag{1.54}$$

Ricordando la definizione della magnetizzazione per spin (1.24) e la (1.48), risulta evidente che $\langle S \rangle$ altro non è che la magnetizzazione per spin, e possiamo quindi scrivere, ponendo il $H^{ext} = 0$ per cercare la magnetizzazione spontanea,

$$m = \tanh(\beta J m) . \tag{1.55}$$

Questa equazione è nota anche come equazione di Curie-Weiss; per semplicità possiamo porre $x = \beta J \gamma m$, ottenendo

$$\frac{1}{\beta J \gamma} x = \tanh x , \tag{1.56}$$

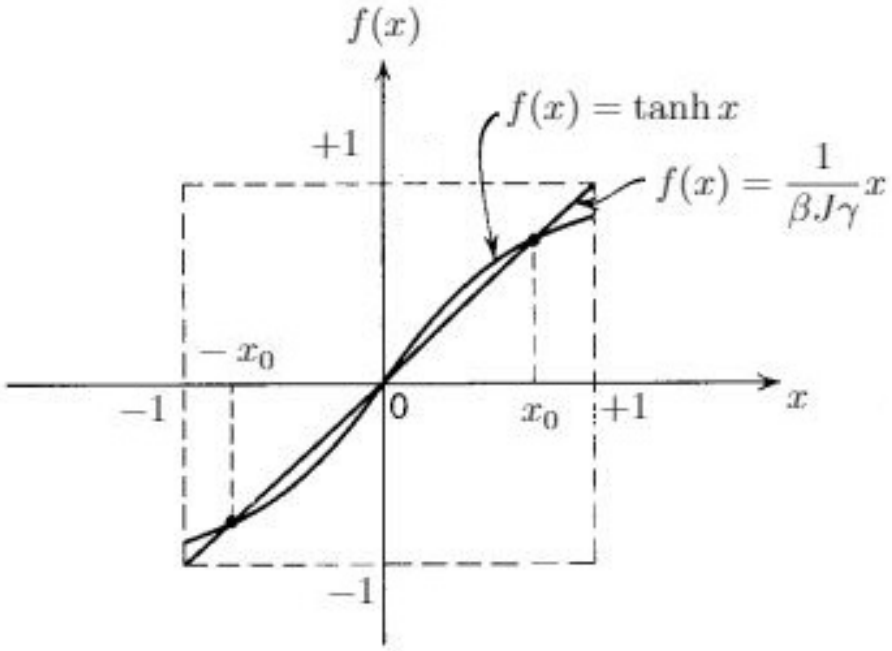

Figura 1.4: Soluzione grafica dell'equazione (1.56).

che può essere risolta per via grafica cercando l'intersezione tra la tangente iperbolica e la retta di pendenza $\frac{1}{\beta J\gamma}$, come riportato in Figura 1.4.

Risulta evidente che in base alla pendenza della retta si hanno due tipi di soluzioni diverse. Definendo

$$T_c = \frac{J\gamma}{k_B}\,, \tag{1.57}$$

per $T < T_c$ (se cioè la pendenza della retta è maggiore di 1) si ha solamente la soluzione banale $x = 0$, da cui $m = 0$: ovvero l'orientazione degli spin è randomica, e non c'è magnetizzazione spontanea. Per $T < T_c$ (cioè, viceversa, la pendenza della retta è minore di 1) si hanno anche due soluzioni non banali $x = \pm x_0$, da cui $m = \pm m_0$, ovvero è presente una magnetizzazione spontanea. In poche parole, il sistema presenta una transizione di fase in $T = T_c$.

Per essere precisi, per ottenere la (1.55) dovremmo considerare il limite $H^{ext} \to 0^{\pm}$, come specificato nella sezione precedente: questo spiega il doppio segno di m; quando il campo esterno è non nullo infatti gli spin sono orientati in maniera concorde ad esso, e daranno quindi origine a una magnetizzazione spontanea $+m_0$ per $H^{ext} \to 0^+$, mentre daranno origine a una magnetizzazione spontanea $-m_0$ per $H^{ext} \to 0^-$. Nella teoria

di campo medio abbiamo quindi ritrovato che la magnetizzazione spontanea riproduce esattamente il comportamento descritto nel grafico in Figura 1.1.

Possiamo vedere la transizione tra una fase simmetrica e una fase asimmetrica anche come una rottura dell'ergodicità. L'ipotesi ergodica afferma che aspettando un tempo sufficientemente lungo la traiettoria del sistema passa arbitrariamente vicino a un qualsiasi punto nello spazio delle fasi, ovvero tutte le configurazioni del sistema sono ugualmente probabili. A $T < T_c$ l'ergodicità è rotta nel senso che le configurazioni del sistema non sono tutte ugualmente probabili, bensì all'equilibrio il sistema potrà trovarsi soltanto nelle configurazioni che danno origine ai valori della magnetizzazione spontanea $m = \pm m_0$, dunque aspettare un tempo lungo non sarà sufficiente affinché la traiettoria del sistema copra tutto lo spazio delle fasi.

Possiamo sfruttare la teoria di campo medio anche per calcolare i principali esponenti critici. A titolo di esempio, cerchiamo il valore per l'esponente critico β definito nella (1.11). Vicino al punto critico possiamo espandere la tangente iperbolica ricordando che $\tanh(x) \simeq x - \frac{1}{3}x^3 + \dots$, e otteniamo

$$\begin{aligned} \frac{1}{\beta J \gamma} &= \frac{T}{T_c} x \simeq x - \frac{1}{3}x^3 + \dots \\ x &\left[\left(1 - \frac{T}{T_c} \right) - \frac{1}{3}x^2 \right] = 0 \,, \end{aligned} \tag{1.58}$$

da cui

$$x_0 = \frac{T}{T_c} m = \left[3 \left(1 - \frac{T}{T_c} \right) \right]^{\frac{1}{2}} \,. \tag{1.59}$$

L'esponente critico per la magnetizzazione risulta quindi $\beta = \frac{1}{2}$, vicino al valore sperimentale $\beta = 0.33$ ottenuto per il Modello di Ising con $d = 3$ [8].

Per basse temperature ($T \to 0$) la pendenza della retta è molto piccola, e dunque l'intersezione avviene per grandi valori di x. Possiamo espandere $\tanh x$ per grandi x, e la (1.56) diventa

$$\frac{T}{T_c} x = 1 - 2e^{-2x} + \dots \,, \tag{1.60}$$

che può essere risolta iterativamente. L'approssimazione di ordine 0 è $x_0 = \frac{T}{T_c}$, e inserendola nel secondo membro si ottiene l'approssimazione al primo ordine

$$m = \frac{T}{T_c} x_0 \simeq 1 - 2e^{-2\frac{T_c}{T}} \,; \tag{1.61}$$

ovvero, per basse T, la teoria di campo medio prevede deviazioni esponenzialmente piccole dal valore di saturazione $m = 1$.

L'andamento della magnetizzazione in funzione della temperatura può essere graficato, ottenendo la curva tipica delle transizioni di fase nei ferromagneti riportata in Figura 1.5, dove è evidente la transizione che avviene a $T = T_c$ tra la fase con $m = 0$ e la fase con $m \neq 0$.

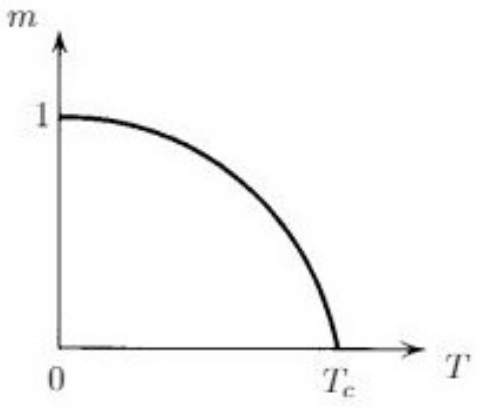

Figura 1.5: Andamento della magnetizzazione spontanea in funzione della temperatura.

1.5 Dinamica di Glauber

L'oggetto principale della trattazione svolta fin'ora sono state le transizioni di fase all'equilibrio, le cui proprietà seguono dalla funzione di partizione Z, e che sono governate dai principi generali di minimizzazione dell'energia e massimizzazione dell'entropia. Tuttavia, in alcuni casi può essere utile un approccio diverso, come quello cinetico seguito da Krapivsky, Redner e Ben-Naim [11], ovvero studiare come l'ordine si forma dinamicamente: questo approccio permette di tenere conto di fluttuazioni stocastiche, introducendo una regola dinamica stocastica, in cui ogni valore a un certo istante è fissato in base a una probabilità invece che con certezza.

La regola dinamica deve essere tale da far convergere il sistema a uno stato di equilibrio in cui la probabilità di trovare il sistema in un certo stato α sia indipendente dal tempo e sia data dalla distribuzione di Boltzmann

$$P_{eq}(\alpha) = \frac{e^{-\beta\mathcal{H}(\alpha)}}{Z}. \tag{1.62}$$

In questa descrizione dinamica non ci sono principi regolatori generali di portata paragonabile ai principi di minimizzazione dell'energia e massimizzazione dell'entropia, dunque la descrizione dinamica non sarà unica. Tuttavia, abbiamo un vincolo che deriva dalla richiesta che il sistema converga a uno stato di equilibrio governato dalla distribuzione di Boltzmann.

Se indichiamo con α e α' due generici stati del sistema, e con $W(\alpha \to \alpha')$ e $W(\alpha' \to \alpha)$ i rate di transizione rispettivamente dallo stato α allo stato α' e viceversa, la condizione di equilibrio assume la seguente forma, detta **condizione di *detailed balance***

$$P_{eq}(\alpha)W(\alpha \to \alpha') = P_{eq}(\alpha')W(\alpha' \to \alpha), \tag{1.63}$$

ovvero il numero medio di transizioni dallo stato α allo stato α' deve essere uguale al numero medio di transizioni dallo stato α' allo stato α. Questa condizione può essere riscritta come

$$\frac{W(\alpha \to \alpha')}{W(\alpha' \to \alpha)} = \frac{P_{eq}(\alpha')}{P_{eq}(\alpha)} = e^{-\beta(\mathcal{H}(\alpha')-\mathcal{H}(\alpha))} = e^{-\beta\Delta\mathcal{H}}, \tag{1.64}$$

dove $\Delta\mathcal{H}$ è la variazione di energia dovuta alla transizione.

Prendiamo in considerazione in particolare solo transizioni in cui viene selezionato casualmente uno spin, che viene invertito, mentre gli altri spin del sistema rimangono invariati. Questo concetto può essere formalizzato introducendo l'operatore di inversione di spin F_i, la cui azione sullo stato $\mathbf{S} = (S_1, ..., S_i, ..., S_N)$ è data da

$$F_i\mathbf{S} = (S_1, ..., -S_i, ..., S_N). \tag{1.65}$$

Calcoliamo la variazione di energia del sistema dovuta a questa transizione riscrivendo l'Hamiltoniana del modello di Ising come

$$\mathcal{H}(\mathbf{S}) = -\frac{1}{2}\sum_k S_k \widetilde{h}_k(\mathbf{S}) - \sum_k H_k S_k \tag{1.66}$$

con

$$\widetilde{h}_i(\mathbf{S}) = \sum_{k\neq i} J_{ik} S_k \tag{1.67}$$

$$\widetilde{h}_i(F_i\mathbf{S}) = \widetilde{h}_i(\mathbf{S}) \tag{1.68}$$

$$\widetilde{h}_k(F_i\mathbf{S}) = \sum_{l\neq k} J_{kl} S_l = \sum_{l\neq k,i} J_{kl} S_l - J_{ki} S_i = \widetilde{h}_k(\mathbf{S}) - 2J_{ki}S_i \quad \forall k \neq i. \tag{1.69}$$

Nella (1.67) è escluso dalla sommatoria il caso $k = i$ poiché si è posto $J_{ii} = 0$ per escludere i casi di auto-accoppiamento e la (1.68) segue dal fatto che gli stati $\mathbf{S}$ e $F_i\mathbf{S}$ differiscono solo per lo spin i-esimo, che è escluso dalla sommatoria.

Dunque la variazione di energia dovuta all'inversione dello spin i-esimo, servendosi di (1.68) e (1.69) risulta

$$\begin{aligned}\mathcal{H}(F_i\mathbf{S}) - \mathcal{H}(\mathbf{S}) &= -\frac{1}{2}\sum_{k\neq i} S_k \left(\widetilde{h}_k(F_i\mathbf{S}) - \widetilde{h}_k(\mathbf{S})\right) + \frac{S_i}{2}\left(\widetilde{h}_k(F_i\mathbf{S}) - \widetilde{h}_k(\mathbf{S})\right) + 2H_iS_i \\ &= 2S_i h_i(\mathbf{S})\end{aligned} \tag{1.70}$$

con

$$h_i(\mathbf{S}) = \sum_{k\neq i} J_{ik} S_k - H_i S_i. \tag{1.71}$$

Il rapporto tra i rate di transizione dato dalla condizione di *detailed balance*, come risulta da (1.64), è quindi

$$\frac{w_i(\mathbf{S})}{w_i(F_i\mathbf{S})} = \frac{P_{eq}(F_i\mathbf{S})}{P_{eq}(\mathbf{S})} = \frac{1 - S_i \tanh(\beta h_i(\mathbf{S}))}{1 - S_i \tanh(\beta h_i(\mathbf{S}))}. \tag{1.72}$$

dove nell'ultima uguaglianza si è fatto uso della seguente identità, valida dal momento che $S = \pm 1$

$$e^{AS} = \cosh A + S \sinh A = \cosh A(1 + S \tanh A). \tag{1.73}$$

La forma più semplice per i rate di transizione affinché soddisfino la condizione di *detailed balance* è quella scelta da Glauber [12]

$$w_i(\mathbf{S}) = \frac{1}{2}\left[1 - S_i \tanh(\beta \sum_j J_{ij} S_j)\right] . \tag{1.74}$$

Un sistema che evolve secondo questa regola dinamica stocastica è quindi descritto dalla dinamica di Glauber, tuttavia, come già sottolineato, la scelta della regola dinamica non è unica. Un'altra scelta comune è, per esempio, l'algoritmo di Metropolis

$$W(\alpha \to \alpha') = \begin{cases} 1 & \Delta\mathcal{H} < 0 \\ e^{-\beta\Delta\mathcal{H}} & \text{altrimenti.} \end{cases} \tag{1.75}$$

Capitolo 2

Il Modello di Hopfield

Il Modello di Hopfield [7] modellizza neuroni e sistema nervoso come una rete neurale, in modo formalmente analogo a come il Modello di Ising modellizza un materiale ferromagnetico come reticolo di spin. La base da cui parte è quindi un insieme di unità molto semplici, in grado di assumere soltanto due stati, detti *firing* e *not firing*, che corrispondono ai casi in cui il neurone emette o meno un potenziale d'azione. In un network di N neuroni, ogni unità sarà quindi un neurone, indicato con $S_i = \pm 1 \; i = 1, ..., N$, dove $+1$ indica lo stato *firing* e -1 lo stato *not firing*.

Il problema di base è studiare proprietà emergenti del sistema, ovvero proprietà non dovute singolarmente alle unità che costituiscono il sistema stesso, bensì alla loro interazione. Siamo interessati in particolare al meccanismo che descrive la memoria associativa come la possibilità di immagazzinare e recuperare memorie, intese come pattern, ovvero configurazioni del sistema in cui ognuna delle N unità ha uno stato fissato; ognuna di queste configurazioni può essere vista come una stringa di N bit, in cui l'i-esimo bit specifica lo stato dell'i-esima unità. Date p memorie, lo stato che assume ogni neurone S_i nelle diverse memorie si indica con ξ_i^μ, dove $i = 1, ..., N$ indicizza i neuroni, mentre $\mu = 1, ..., p$ indicizza le memorie. Un certo pattern ν può essere descritto con un vettore $\boldsymbol{\xi}^\nu$ in cui l'indice ν è fissato, mentre l'indice i assume tutti i valori corrispondenti alle varie unità: $\boldsymbol{\xi}^\nu = (\xi_1^\nu, \ldots, \xi_N^\nu)$.

Poiché ognuna delle N unità può assumere solo 2 stati lo spazio delle configurazioni è costituito da 2^N configurazioni; l'evoluzione del sistema può essere vista come una traiettoria in questo spazio, e risulta naturale descrivere il recupero di una memoria come la persistenza di un certo pattern, in opposizione al passaggio del sistema da una configurazione transiente all'altra. Le memorie immagazzinate devono quindi risultare degli attrattori, ovvero degli stati stabili verso i quali il sistema rilassa; vedremo tuttavia che si formano anche degli attrattori indesiderati, ovvero degli stati stabili non corrispondenti a una delle memorie immagazzinate.

Si possono seguire due approcci differenti ma analoghi: l'evoluzione del sistema può essere trattata mediante una descrizione dinamica, in cui lo stato delle unità evolve secondo una determinata regola dinamica, oppure può essere descritta in termini di una

funzione energia, che risulterà analoga all'Hamiltoniana del Modello di Ising, e il principio che regola l'evoluzione sarà quindi la minimizzazione dell'energia.

Nelle prime due sezioni del presente capitolo verranno presentati entrambi gli approcci, e verrà applicata la teoria di campo medio al Modello di Hopfield per descrivere la dinamica a temperatura finita, fino a ottenere una descrizione della transizione di fase che caratterizza il passaggio da una memoria utile a una memoria inutile. Una memoria è ritenuta utile se esibisce il comportamento tipico della memoria associativa, ovvero, partendo da una configurazione vicina nello spazio delle configurazioni a uno dei pattern immagazzinati come attrattori, è in grado di recuperare tale pattern, con una percentuale di bit errati non superiore a un certo limite stabilito come accettabile. Viceversa, la memoria è considerata inutile se gli stati verso cui evolve sono stati randomici, ovvero in cui in media solo la metà dei bit sono corretti rispetto ai pattern da recuperare.

Nell'ultima sezione verrà ricavata l'equazione di punto a sella, equivalente al limite termodinamico e da cui si deriva la soluzione di campo medio. Un parametro utile per descrivere il sistema è il rapporto tra pattern immagazzinati e unità costituenti il network $\alpha = \frac{p}{N}$. Verrà preso in analisi inizialmente il limite di non saturazione ($\alpha = 0$), per il quale si cercheranno le soluzioni dell'equazione di punto a sella, di cui verrà studiata la stabilità, distinguendo tra attrattori desiderati e indesiderati. Verrà poi affrontato il caso generale, illustrando il diagramma di fase del Modello di Hopfield che descrive la presenza di stati stabili al variare del livello di rumore presente nel network e del parametro α.

2.1 Descrizione dinamica

2.1.1 Regola dinamica e condizione di stabilità

Ogni neurone, o meglio ogni unità (si usa preferibilmente questo termine per sottolineare che la trattazione presente rimane pur sempre una modellizzazione e non una descrizione fedele della fisiologia del sistema nervoso) viene trattato nel Modello di Hopfield secondo la descrizione di McCulloch-Pitts [3].

Il neurone di McCulloch-Pitts è un'unità che riceve stimoli (i potenziali d'azione) dai neuroni vicini, pesati secondo certi pesi J_{ij}, detti coefficienti sinaptici, che indicano la forza della connessione sinaptica tra ogni coppia di neuroni i e j. Se la somma di tutti gli stimoli che raggiungono l'unità S_i supera un certo valore di soglia θ_i, essa invia a sua volta un potenziale d'azione, ovvero assume lo stato *firing* $S_i = 1$; altrimenti, l'unità assume lo stato *not firing* $S_i = -1$. Questo comportamento è riassunto nella **regola dinamica** che descrive il passaggio dell'unita i-esima dallo stato che assume in un certo step temporale t allo stato aggiornato nello step temporale successivo $t + 1$:

$$S_i(t+1) = \operatorname{sgn}\left(\sum_j J_{ij} S_j(t) - \theta_i\right). \tag{2.1}$$

Il modello formulato da Hopfield [7] prevede una regola dinamica di questo tipo in cui la soglia di *firing* θ_i è nulla. Tale modello considera infatti una dinamica random in cui tutte le unità assumono con uguale probabilità i valori $+1$ e -1. Ciò avviene se $\theta_i = 0$ considerando pattern random, poiché verrà mostrato a breve che i J_{ij} possono essere scritti come funzione dei pattern $\boldsymbol{\xi}^\mu$ e se questi sono random anche i coefficienti sinaptici lo sono, mentre se θ_i non fosse nullo i valori $+1$ e -1 non sarebbero equiprobabili per $S_i(t+1)$. Porre $\theta_i = 0$ è ragionevole tenendo conto che i coefficienti sinaptici per ogni coppia di unità non sono fissati bensì dipendono dai pattern che vengono presentati al sistema, dunque non è necessario nemmeno definire per ogni unità una soglia di *firing* fissata bensì si può considerare un valore medio, che possiamo porre per semplicità uguale a zero.

Nella trattazione seguente ometteremo inoltre il riferimento allo step temporale indicando con S' lo stato dell'unità aggiornato secondo la regola dinamica; l'evoluzione dello stato di ogni unità avverrà dunque secondo

$$S'_i = \operatorname{sgn}\left(\sum_j J_{ij} S_j\right). \tag{2.2}$$

L'aggiornamento della configurazione del sistema secondo la regola dinamica appena descritta può essere eseguito in due modi: in maniera sincrona oppure asincrona. Nel primo caso tutte le unità vengono aggiornate contemporaneamente, tuttavia questo comporterebbe la necessità di un central clock che sincronizzi il processo. Nel secondo caso, invece, viene aggiornata un'unità alla volta, e si tratta di una modalità più naturale per i neural network. L'unità da aggiornare può essere scelta randomicamente ad ogni step temporale, oppure possiamo lasciare che ogni unità indipendentemente si aggiorni o meno secondo una certa probabilità costante per unità di tempo.

Una memoria $\boldsymbol{\xi}^\nu$ viene immagazzinata quando la regola dinamica porta ogni neurone S_i ad assumere lo stato ξ_i^ν che ha nella memoria ν, ovvero

$$S_i = \operatorname{sgn}\left(\sum_j J_{ij}\xi_j^\nu\right) = \xi_i^\nu \quad \forall i. \tag{2.3}$$

In poche parole, continuando ad applicare la regola dinamica al neurone i-esimo, questo non modifica il suo stato, che continua a rimanere quello assunto nella memoria ν. L'equazione (2.3) rappresenta quindi la **condizione di stabilità** della memoria ν.

2.1.2 Coefficienti sinaptici

L'informazione sull'immagazzinamento della memoria è contenuta nei coefficienti J_{ij}, detti nel Modello di Hopfield coefficienti sinaptici; talvolta ci si riferisce ad essi direttamente come connessioni sinaptiche, o come matrice sinaptica, in quanto è immediato

che considerando tutti i valori che possono assumere gli indici i e j si ottiene una matrice. Una sinapsi è detta eccitatoria se $J_{ij} > 0$, mentre è detta inibitoria se $J_{ij} < 0$. I coefficienti sinaptici corrispondono all'interazione di scambio tra due spin del modello di Ising, e modificano il proprio valore in base agli stati ricorrenti dei neuroni pre-sinaptico e post-sinaptico: la funzione principale di queste modificazioni sinaptiche è creare attrattori per l'evoluzione dinamica del sistema, il che permette il processo di recupero di una memoria immagazzinata.

Dovremo quindi trovare un'espressione per i J_{ij} che consenta al sistema, evolvendo secondo la regola dinamica (2.2), di arrivare a una configurazione stabile (secondo la condizione di stabilità (2.3)) corrispondente a una delle memorie immagazzinate. Per testare se una forma dei J_{ij} è accettabile occorrerà innanzitutto verificare se i pattern da immagazzinare risultino a propria volta stabili, e controllare quindi se piccole deviazioni dai pattern vengono corrette mentre il sistema evolve (ovvero se, partendo da una configurazione vicina a una memoria immagazzinata, il sistema evolve spontaneamente verso quella memoria).

Il caso più semplice è quello in cui si ha un solo pattern da immagazzinare $\boldsymbol{\xi}^\nu$, che indichiamo con l'indice ν fissato. Risulta immediato che la condizione di stabilità

$$S_i = \mathrm{sgn}\left(\sum_j J_{ij}\xi_j^\nu\right) = \xi_i^\nu \quad \forall i \tag{2.4}$$

è soddisfatta da $J_{ij} \propto \xi_i^\nu \xi_j^\nu$, in quanto risulta

$$S_i = \mathrm{sgn}\left(\sum_j cost. \times \xi_i^\nu \xi_j^\nu \xi_j^\nu\right) = \xi_i^\nu \ , \tag{2.5}$$

dove si è sfruttato il fatto che $\xi_j^\nu \xi_j^\nu = 1$ sia che ξ_j^ν valga $+1$ sia che valga -1. Conviene scrivere la costante di proporzionalità in modo tale che

$$J_{ij} = \frac{1}{N}\xi_i^\nu \xi_j^\nu \ , \tag{2.6}$$

dove il fattore $\frac{1}{N}$ viene inserito per assicurarsi che nel limite termodinamico in cui $N \to \infty$ il contributo dei J_{ij} rimanga finito. Possiamo verificare immediatamente che piccole deviazioni dai pattern vengono corrette mentre il sistema evolve secondo la regola dinamica (2.2): infatti se nella sommatoria un certo numero di bit (minore della metà) è sbagliato, ovvero si discosta dal pattern considerato, prevarranno comunque i bit giusti e il segno risultante sarà quello di ξ_i^ν, come richiesto dalla condizione di stabilità. Possiamo quindi affermare che il pattern $\boldsymbol{\xi}^\nu$ è un **attrattore** per la dinamica del sistema.

Si nota che anche il pattern $-\boldsymbol{\xi}^\nu$, detto *reversed state* in quanto tutte le unità hanno stato invertito rispetto al pattern $\boldsymbol{\xi}^\nu$, è a propria volta un attrattore: lo spazio delle configurazioni risulterà quindi diviso in maniera simmetrica in due bacini di attrazione,

come riportato in Figura 2.1, e a seconda che nello stato iniziale il sistema contenga meno o più della metà dei bit sbagliati rispetto al pattern $\boldsymbol{\xi}^\nu$, convergerà rispettivamente ad esso oppure al suo *reversed state*.

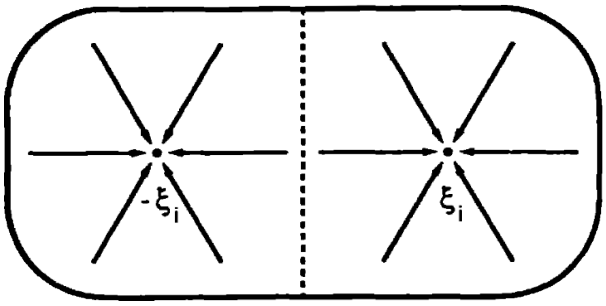

Figura 2.1: Rappresentazione schematica dei bacini di attrazione che si formano nel caso di un solo pattern immagazzinato nel network.

Il convergere del sistema a uno di questi due stati ricorda molto da vicino la rottura dell'ergodicità che accompagna una transizione tra fasi con diversa simmetria descritta nella sezione 1.4 per un sistema ferromagnetico: in quetso caso, se lasciamo evolvere il sistema e aspettiamo un tempo sufficientemente lungo, le configurazioni non saranno tutte ugualmente probabili, bensì il sistema si troverà con certezza o nella configurazione corrispondente a $\boldsymbol{\xi}^\nu$ o in quella corrispondente a $-\boldsymbol{\xi}^\nu$. Richiedere che si formino degli attrattori per la dinamica del sistema equivale quindi in un certo senso a imporre la non-ergodicità, e vedremo più avanti che effettivamente ciò può essere descritto come una transizione tra una fase (non ergodica) in cui sono presenti attrattori corrispondenti ai pattern immagazzinati e una fase in cui tali attrattori non sono più presenti.

L'espressione per i coefficienti sinaptici appena esposta per il caso con un solo pattern da immagazzinare può essere estesa al caso di p pattern, secondo quella che viene detta **Regola di Hebb generalizzata**

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \,, \tag{2.7}$$

dove l'indice μ è ora variabile e assume tutti i valori compresi tra 1 e p. Come formulata originariamente [6], la Regola di Hebb affermava che il cambiamento nella forza sinaptica che agisce tra due neuroni è proporzionale alla correlazione tra l'attività del neurone pre-sinaptico e quella del neurone post-sinaptico.

Dall'espressione generalizzata (2.7) si può infatti notare che il coefficiente sinaptico J_{ij} è positivo per i neuroni i e j se, sommando su tutte le p memorie, sono di più le memorie in cui essi assumono lo stesso stato (in quanto $\xi_i^\mu \xi_j^\mu = +1$ se valgono entrambi $+1$ oppure entrambi -1 nella memoria μ) rispetto a quelle in cui assumono stati opposti. Ovvero, ogni memoria contribuisce con un termine positivo al coefficiente sinaptico J_{ij}

se l'attività dei neuroni i e j è correlata, mentre contribuisce con un termine negativo se è anti-correlata. Tuttavia, questa espressione si discosta leggermente dalla formulazione originale di Hebb, in quanto il peso J_{ij} incrementa positivamente anche se i neuroni sono entrambi *not firing* ($\xi_i^\mu = \xi_j^\mu = -1$), fatto che fisiologicamente non ha senso. Inoltre la regola (2.7) può far cambiare una sinapsi da eccitatoria ($J_{ij} > 0$) a inibitoria ($J_{ij} < 0$) aggiungendo nuovi pattern, cosa che si pensa non accada nella realtà.

2.1.3 Capacità di immagazzinamento

Poiché aggiungendo nuovi pattern da memorizzare i coefficienti sinaptici cambiano, possiamo aspettarci che cambi anche la stabilità dei pattern stessi e che ci sia un limite al numero di pattern che possono essere immagazzinati come attrattori. Oltre tale limite si dice che si ha un breakdown, ovvero il sistema non funziona più in maniera efficace come memoria associativa. Vogliamo quindi determinare il numero massimo p_{max} di memorie immagazzinabili in un neural network, ovvero la capacità di immagazzinamento; con tale termine ci si riferisce spesso anche al rapporto tra memorie immagazzinate e unità costituenti il network $\alpha = \frac{p_{max}}{N}$.

Consideriamo innanzitutto un pattern ν fissato, e vogliamo esprimere la condizione di stabilità in funzione della presenza di altri pattern immagazzinati. La condizione di stabilità data da (2.3) può essere scritta in forma più compatta come

$$S_i = \mathrm{sgn}(h_i^\nu) = \xi_i^\nu \quad \forall i, \tag{2.8}$$

dove

$$\begin{aligned} h_i^\nu &= \sum_j J_{ij}\xi_j^\nu \\ &= \frac{1}{N}\sum_{j=1}^{N}\xi_i^\nu\xi_j^\nu\xi_j^\nu + \frac{1}{N}\sum_{j=1}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p}\xi_i^\mu\xi_j^\mu\xi_j^\nu \\ &= \xi_i^\nu + \frac{1}{N}\sum_{j=1}^{N}\sum_{\substack{\mu=1\\ \mu\neq\nu}}^{p}\xi_i^\mu\xi_j^\mu\xi_j^\nu. \end{aligned} \tag{2.9}$$

Il secondo termine in questa espressione è detto *crosstalk term*, in quanto contiene la sovrapposizione tra lo stato delle unità nel pattern ν che stiamo considerando e in tutti gli altri pattern immagazzinati.

Definiamo per comodità la quantità

$$C_i^\nu = -\xi_i^\nu\frac{1}{N}\sum_{j=1}^{n}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu\xi_j^\nu. \tag{2.10}$$

Risulterà $C_i^\nu < 0$ se il *crosstalk term* ha lo stesso segno di ξ_i^ν, e in questo caso il *crosstalk term* non inficia la stabilità del pattern ν perché non altera il segno di ξ_i^ν, che sarà quindi stabile. Se invece il *crosstalk term* e ξ_i^ν hanno segni diversi, e dunque $C_i^\nu > 0$, possono presentarsi due casi, in base al valore di C_i^ν. Se $C_i^\nu < 1$, esso non può cambiare il segno di h_i^ν, e avremo quindi $S_i = \mathrm{sgn}(h_i^\nu) = \xi_i^\nu$. In questo caso i pattern immagazzinati sono stabili (ovvero se il sistema parte da uno di essi ci rimane), e risultano anche degli attrattori, perché se una piccola frazione di bit sono errati, verranno corretti nel corso dell'evoluzione. Se invece $C_i^\nu > 1$, esso altera il segno di h_i^ν rispetto alla memoria immagazzinata ξ_i^ν, e la regola dinamica porterà il sistema ad allontanarsi da tale memoria, che risulterà instabile.

Vogliamo quindi valutare la probabilità che un certo bit sia instabile:

$$P_{error} = P(C_i^\nu > 1), \tag{2.11}$$

che come risulta evidente da (2.10) cresce all'aumentare di p.

Se scegliamo un criterio di performance accettabile fissando la probabilità di errore ammessa, possiamo determinare la capacità massima di immagazzinamento, ovvero il massimo valore di p che soddisfa il criterio scelto.

Assumiamo per semplicità $N, p \gg 1$ (si tratta di una condizione verosimile e che semplifica i conti): C_i^ν è la somma di circa Np termini che possono assumere solo i valori $+1$ e -1, e possiamo quindi vederlo come una distribuzione binomiale con media 0 e varianza $\sigma^2 = \frac{p}{N}$, che per $N, p \gg 1$ tende a una gaussiana, con media 0 e varianza $\sigma^2 = \frac{p}{N}$. Dunque risulta

$$P_{error} = \frac{1}{\sqrt{2\pi}\sigma} \int_1^\infty e^{-\frac{x^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty e^{-\frac{x^2}{2\sigma^2}}\, dx - \frac{1}{\sqrt{2\pi}\sigma} \int_0^1 e^{-\frac{x^2}{2\sigma^2}}\, dx \tag{2.12}$$

che con la sostituzione $u = \frac{x}{\sqrt{2\sigma^2}}$ diventa

$$P_{error} = \frac{1}{2} - \int_0^{\frac{1}{\sqrt{2\sigma^2}}} e^{-u^2}\, du = \frac{1}{2}\left(1 - erf\left(\frac{1}{\sqrt{2\sigma^2}}\right)\right) = \frac{1}{2}\left(1 - erf\left(\sqrt{\frac{N}{2p}}\right)\right), \tag{2.13}$$

dove si è posto

$$erf\left(\frac{1}{\sqrt{2\sigma^2}}\right) \coloneqq \frac{2}{\sqrt{\pi}} \int_0^{\frac{1}{\sqrt{2\sigma^2}}} e^{-u^2}\, du\,. \tag{2.14}$$

Fissando quindi una probabilità di errore che riteniamo accettabile, per esempio $P_{error} < 0.001$, otteniamo il numero massimo di memorie p_{max} che possono essere immagazzinate compatibilmente con il criterio scelto.

P_{error}	$\frac{p_{max}}{N}$
0.001	0.105
0.0036	0.138
0.01	0.185
0.05	0.37
0.1	0.61

Tabella 2.1: Esempi dei valori di $\frac{p_{max}}{N}$ necessari per ottenere i un fissato valore di P_{error}.

Queste stime riguardano la stabilità iniziale dei pattern, tuttavia una volta che una certa frazione dei bit (compatibile con il criterio di performance scelto) si inverte il sistema continua ad evolvere, e altri bit potrebbero a loro volta invertirsi, causando nella peggiore delle ipotesi un fenomeno a valanga, che porta la configurazione del sistema molto lontana dalla memoria da immagazzinare. Perciò la stima di p_{max} appena riportata rappresenta semplicemente un limite superiore.

2.2 Descrizione mediante una funzione energia

2.2.1 Hamiltoniana del Modello di Hopfield

L'Hamiltoniana che descrive il Modello di Hopfield è analoga a quella del Modello di Ising definita in (1.16), sostituendo gli spin del reticolo con le unità S_i della rete neurale. Trascurando il campo esterno, l'Hamiltoniana ha la forma

$$\mathcal{H}(\mathbf{S}) = -\frac{1}{2}\sum_{i,j} J_{ij} S_i S_j. \tag{2.15}$$

Come per il Modello di Ising, si tratta di una funzione della configurazione del sistema $\mathbf{S}$ (dove con configurazione si intende l'insieme degli stati assunti da tutte le unità, che come già menzionato può essere vista come una stringa di N bit).

L'Hamiltoniana definisce una superficie equi-energetica nello spazio delle configurazioni, e l'evoluzione del sistema avviene secondo il principio di minimizzazione dell'energia: il punto che rappresenta il sistema nello spazio delle configurazioni evolve tendendo verso i minimi della superficie individuata dall'Hamiltoniana. Pertanto questi minimi fungono da attrattori, e non sono altro che i pattern memorizzati. I minimi sono circondati da "valli", o bacini di attrazione delle memorie: ovvero, se il sistema si trova nello stato iniziale in un punto nello spazio delle configurazioni compreso nel bacino di attrazione di una certa memoria, evolverà fino alla configurazione corrispondente a quella memoria, e diremo quindi che è stato in grado di recuperarla.

Tenendo conto della simmetria delle connessioni sinaptiche ($J_{ij} = J_{ji}$), si può riscrivere l'Hamiltoniana come:

$$\mathcal{H}(\mathbf{S}) = C - \sum_{(i,j)} J_{ij} S_i S_j \tag{2.16}$$

dove la costante C racchiude i casi con $i = j$, mentre la sommatoria è estesa a tutte le coppie distinte (i, j). Conviene porre $J_{ii} = 0$ (e dunque $C = 0$), ovvero escludere i casi di auto-accoppiamento, in quanto questi potrebbero dare origine ad attrattori non desiderati, detti stati spuri, in aggiunta a quelli dati dalle memorie da immagazzinare.

Infatti isolando il termine di auto-accoppiamento nella regola dinamica si ottiene:

$$S_i' = \operatorname{sgn}\left(\sum_j J_{ij} S_j\right) = \operatorname{sgn}\left(J_{ii} S_i + \sum_{j \neq i} J_{ij} S_j\right). \tag{2.17}$$

Se domina il primo termine, ovvero

$$J_{ii} > \sum_{j \neq i} J_{ij} S_j, \tag{2.18}$$

si ha che $S_i' = \operatorname{sgn}(S_i) = S_i$, ovvero ogni S_i che soddisfa la condizione (2.18) non viene alterato dalla regola dinamica e risulta uno stato stabile indesiderato.

Con questa accortezza, l'Hamiltoniana del modello di Hopfield diventa:

$$\mathcal{H}(\mathbf{S}) = -\sum_{(i,j)} J_{ij} S_i S_j. \tag{2.19}$$

2.2.2 Regola dinamica e minimizzazione dell'energia

La descrizione dell'evoluzione del sistema mediante l'Hamiltoniana è compatibile con la descrizione dinamica seguita nella sezione precedente. Si può infatti dimostrare che la regola dinamica (2.2) effettivamente minimizza l'energia descritta dall'Hamiltoniana (2.19).

Poiché ogni unità può assumere solo due valori, si hanno due casi. Il caso in cui lo stato dell'unità in questione non viene modificato dalla regola dinamica ($S_k' = S_k$) è banale, in quanto in tal caso l'energia non cambia. Altrimenti ($S_k' = -S_k$) si ha che

$$\begin{aligned}
\mathcal{H}' - \mathcal{H} &= \left(-\sum_i \sum_{j<i} J_{ij} S_i' S_j\right) - \left(-\sum_i \sum_{j<i} J_{ij} S_i S_j\right) \\
&= -\sum_{j \neq k} J_{kj} S_k' S_j - \sum_{i \neq k} \sum_{j<i} J_{ij} S_i' S_j + \sum_{j \neq k} J_{kj} S_k S_j + \sum_{i \neq k} \sum_{j<i} J_{ij} S_i S_j \\
&= 2 S_k \sum_{j \neq k} J_{kj} S_j
\end{aligned} \tag{2.20}$$

Si può vedere che questa espressione è negativa, da cui $\mathcal{H}' < \mathcal{H}$, ovvero la regola dinamica minimizza l'energia. Infatti

$$\operatorname{sgn}\left(S_k \sum_j J_{kj} S_j\right) = \operatorname{sgn}(S_k) \operatorname{sgn}\left(\sum_j J_{kj} S_j\right) = S_k S_k' = -S_k^2 < 0. \tag{2.21}$$

2.2.3 Regola di Hebb e minimizzazione dell'energia

La regola di Hebb può essere ricavata direttamente a partire dal principio di minimizzazione dell'energia. Nel caso più semplice, in cui si ha un solo pattern da memorizzare, dobbiamo imporre che l'energia sia minima quando è massima la sovrapposizione tra il pattern $\boldsymbol{\xi}^\nu$ che vogliamo memorizzare e la configurazione del sistema $\mathbf{S}$, in modo che tale pattern diventi un attrattore. Questa condizione è soddisfatta se scegliamo un'Hamiltoniana della forma

$$\mathcal{H} = -\frac{1}{2N}\left(\sum_i S_i \xi_i^\nu\right)^2, \tag{2.22}$$

dove il fattore $\frac{1}{2N}$ è stato scelto per fare in modo che l'energia sia estensiva e scali secondo il numero di unità del sistema $\mathcal{O}(N)$, mentre l'elevamento al quadrato è stato inserito affinché anche i *reversed state* siano attrattori, come emerge dalla descrizione dinamica.

Per generalizzare al caso di molti pattern è sufficiente sommare su tutti i pattern, in modo da rendere ognuno di essi stabile

$$\mathcal{H} = -\frac{1}{2N}\sum_{\mu=1}^{p}\left(\sum_i S_i \xi_i^\mu\right)^2. \tag{2.23}$$

Esplicitando il quadrato in questa espressione si ottiene

$$\begin{aligned}
\mathcal{H} &= -\frac{1}{2N}\sum_{\mu=1}^{p}\left(\sum_i S_i\xi_i^\mu\right)\left(\sum_j S_j\xi_j^\mu\right) \\
&= -\frac{1}{2}\sum_{i,j}\underbrace{\left(\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu\right)}_{J_{ij}} S_iS_j \\
&= -\frac{1}{2}\sum_{i,j} J_{ij}S_iS_j.
\end{aligned} \tag{2.24}$$

Dunque, partendo da un'espressione che soddisfacesse la minimizzazione dell'energia, abbiamo ottenuto esattamente l'Hamiltoniana del Modello di Hopfield, con i coefficienti sinaptici dati dalla regola di Hebb.

Quello appena seguito è un procedimento del tutto generale: per identificare i coefficienti sinaptici si scrive una Hamiltoniana i cui minimi soddisfino il problema di interesse e si sviluppa il quadrato, ottenendo in generale termini del tipo S_iS_j, termini lineari in S_i, e termini costanti. Il coefficiente dei primi ci dà J_{ij}, nei secondi possiamo riconoscere dei termini di soglia, e le costanti non sono influenti in quanto l'energia è sempre definita a meno di una costante.

2.3 Dinamica a temperatura finita

Nel Modello di Hopfield viene introdotto un concetto analogo a quello di temperatura, intesa come disordine termico, per tenere conto delle fluttuazioni dovute a varie cause non controllabili (i neuroni potrebbero emettere potenziali d'azione con intensità variabile, ci sono ritardi nelle sinapsi, fluttuazioni variabili nel rilascio dei neurotrasmettitori, e così via). Si tratta più propriamente di una pseudo-temperatura, in quanto non ha nulla a che fare con le dimensioni fisiche né con il significato fisico di una temperatura. Spesso si lavora con la temperatura inversa β definita da

$$\beta = \frac{1}{T}, \tag{2.25}$$

ponendo $k_B = 1$.

Questo disordine termico, detto *thermal disorder* o *annealed disorder* (dal termine *anneal* che significa "ricuocere" riferito a un vetro o a un metallo), dovuto alla stocasticità che coinvolge le unità microscopiche, è da distinguere da un altro tipo di disordine, detto *quenched disorder*, riferito ai pattern immagazzinati. Essi possono essere infatti visti in un certo senso come disordine, in quanto vengono estratti randomicamente da una distribuzione, ovvero avranno una distribuzione random; è da notare quindi che i pattern da immagazzinare non si differenziano da tutte le altre possibili configurazioni perché siano in sé configurazioni particolari, ma per il semplice fatto di essere attrattori. Il disordine associato alle configurazioni corrispondenti ai pattern immagazzinati è tuttavia un disordine imposto dall'esterno (imponendo che il pattern sia attrattore) e che persiste all'evolvere del sistema (la definizione viene dal termine inglese *quench*, che riferito a un metallo fuso significa raffreddarlo istantaneamente e in un certo senso "congelarne", ovvero fissarne, la struttura, nella quale permane il disordine tipico dello stato liquido nonostante sia allo stato solido). Le nozioni di *annealed disorder* e *quenched disorder* si differenziano quindi per le scale temporali coinvolte: il primo riguarda configurazioni che evolvono sulla stessa scala di tempo delle variabili dinamiche microscopiche, mentre il secondo riguarda configurazioni che evolvono su scale di tempo molto più lunghe, e che quindi rispetto alla scala di tempo delle variabili microscopiche risultano "congelate".

Come nel Modello di Ising, anche nel Modello di Hopfield l'evoluzione del sistema è quindi governata dalla competizione tra la presenza di un campo esterno, che tende ad allineare gli stati delle unità e prevale per basse temperature, e l'effetto del disordine termico che prevale ad alte temperature. Per tenere conto delle fluttuazioni, occorre introdurre una dinamica stocastica, valida per temperature finite, che si riduce alla dinamica deterministica esposta fin'ora nel limite $T \to 0$. Ciò può essere fatto mediante la Dinamica di Glauber [12] [11], sostituendo la regola dinamica deterministica (2.2) con la regola dinamica stocastica di Glauber (1.74).

A ogni step temporale verrà dunque selezionata randomicamente una unità, che verrà aggiornata secondo

$$S_i' = \begin{cases} +1 & \text{con probabilità } f_\beta(h_i) \\ -1 & \text{con probabilità } 1 - f_\beta(h_i) \end{cases} \tag{2.26}$$

con

$$f_\beta(h_i) = w_i(\mathbf{S}) = \frac{1}{2}\left[1 - S_i \tanh\left(\beta \sum_j J_{ij} S_j\right)\right] \tag{2.27}$$

che può essere riscritta come

$$f_\beta(h_i) = \frac{e^{\beta h_i}}{e^{\beta h_i} + e^{-\beta h_i}} = \frac{1}{1 + e^{-2\beta h_i}}. \tag{2.28}$$

Per $\beta \to \infty$, ovvero $T \to 0$, f_β tende alla step function che descrive la regola dinamica deterministica, mentre per $\beta \to 0$, ovvero $T \to \infty$, $f_\beta \to \frac{1}{2}$, ovvero gli stati delle unità sono completamente randomici, e ogni unità ha la stessa probabilità di essere *firing* o *not firing*. Notando che $1 - f_\beta(h_i) = f_\beta(-h_i)$, la regola dinamica stocastica può essere scritta come

$$P(S_i' = \pm 1) = f_\beta(\pm h_i) = \frac{1}{1 + e^{\mp 2\beta h_i}}. \tag{2.29}$$

Si può mostrare che la (2.28) è la probabilità condizionata che, data una configurazione $\mathbf{S}$ in cui selezioniamo randomicamente un'unità S_i da aggiornare, S_i' valga $+1$, sapendo che tutte le altre unità sono rimaste invariate rispetto a $\mathbf{S}$. Tale probabilità va calcolata secondo la misura di probabilità data dalla distribuzione di Boltzmann

$$\mu(\mathbf{S}) = \frac{e^{-\beta \mathcal{H}(\mathbf{S})}}{Z}. \tag{2.30}$$

Indicando $\widetilde{\mathbf{S}} = \{S_k\}_{k \neq i}$ si ha

$$\begin{aligned} P(S_i' = +1 | \widetilde{\mathbf{S}}) &= \frac{P(S_i' = +1 \cap \widetilde{\mathbf{S}})}{P(\widetilde{\mathbf{S}})} = \frac{\mu(\widetilde{\mathbf{S}}, 1)}{\mu(\widetilde{\mathbf{S}}, 1) + \mu(\widetilde{\mathbf{S}}, -1)} \\ &= \frac{e^{-\beta \mathcal{H}(\widetilde{\mathbf{S}}, 1)}}{e^{-\beta \mathcal{H}(\widetilde{\mathbf{S}}, 1)} + e^{-\beta \mathcal{H}(\widetilde{\mathbf{S}}, -1)}} = \frac{e^{\beta h_i(\mathbf{S})}}{e^{\beta h_i(\mathbf{S})} + e^{-\beta h_i(\mathbf{S})}} \end{aligned} \tag{2.31}$$

L'ultima uguaglianza segue da

$$\begin{aligned} \mathcal{H}(\widetilde{\mathbf{S}}, 1) &= -\sum_{j,k \neq i} J_{jk} S_j S_k - S_i \sum_k J_{ik} S_k = \widetilde{\mathcal{H}}(\mathbf{S}) - h_i(\mathbf{S}), \\ \mathcal{H}(\widetilde{\mathbf{S}}, -1) &= -\sum_{j,k \neq i} J_{jk} S_j S_k - S_i \sum_k J_{ik} S_k = \widetilde{\mathcal{H}}(\mathbf{S}) + h_i(\mathbf{S}). \end{aligned}$$

Possiamo quindi applicare la teoria di campo medio, già vista per il Modello di Ising nella sezione 1.4, al Modello di Hopfield, considerando unità stocastiche che evolvono secondo la regola (2.26). Consideriamo per ora il caso in cui abbiamo pochi pattern p da immagazzinare rispetto al numero di unità che costituiscono il network, ovvero $p \ll N$. La trattazione seguente sarà quindi sempre applicabile nel limite termodinamico in cui $N \to \infty$.

Calcoliamo innanzitutto il valore medio della magnetizzazione per un singola unità immersa in un campo esterno h_i:

$$\begin{aligned}\langle S_i \rangle &= (+1)P(S_i = +1) + (-1)P(S_i = -1) \\ &= \frac{1}{1 + e^{-2\beta h_i}} + \frac{1}{1 + e^{2\beta h_i}} \\ &= \tanh(\beta h_i)\end{aligned} \tag{2.32}$$

Il campo locale con cui ogni unità interagisce è quello dovuto alla magnetizzazione delle altre unità: $h_i = \sum_j J_{ij} S_j$; applichiamo la teoria di campo medio sostituendo questo campo con il suo valore medio

$$\langle h_i \rangle = \sum_j J_{ij} \langle S_j \rangle. \tag{2.33}$$

Otteniamo quindi la seguente equazione di autoconsistenza

$$\langle S_i \rangle = \tanh(\beta \langle h_i \rangle) = \tanh\left(\beta \sum_j J_{ij} \langle S_j \rangle\right) = \tanh\left(\frac{\beta}{N} \sum_{j,\mu} \xi_i^\mu \xi_j^\mu \langle S_j \rangle\right). \tag{2.34}$$

dove nell'ultima uguaglianza si è fatto uso della regola di Hebb (2.7) per i coefficienti sinaptici.

In questo modo ci siamo ricondotti a N equazioni non lineari in N incognite, con il vantaggio che non sono più presenti variabili stocastiche. Siamo interessati a cercare stati stabili secondo la teoria di campo medio, ovvero stati che soddisfano l'equazione di autoconsistenza. Facciamo quindi un Ansatz: ipotizziamo che siano stabili stati del tipo

$$\langle S_i \rangle = m \xi_i^\nu \quad \forall i, \tag{2.35}$$

ovvero il cui valor medio sia proporzionale a un certo pattern immagazzinato. Intuitivamente, questo significa che ogni unità è in media nello stesso stato in cui si trova nel pattern ν.

L'equazione di autoconsistenza per pattern di questo tipo diventa:

$$\begin{aligned} m\xi_i^\nu &= \tanh\left(\frac{\beta}{N}\sum_{j,\mu}\xi_i^\mu\xi_j^\mu m\xi_j^\nu\right) \\ &= \tanh\left(\frac{\beta}{N}\sum_j \xi_i^\nu\xi_j^\nu m\xi_j^\nu + \frac{\beta}{N}\sum_j\sum_{\substack{\mu=1\\\mu\neq\nu}}^{p}\xi_i^\mu\xi_j^\mu m\xi_j^\nu\right) \\ &= \tanh\left(m\xi_i^\nu + \frac{\beta}{N}\sum_j\sum_{\substack{\mu=1\\\mu\neq\nu}}^{p}\xi_i^\mu\xi_j^\mu m\xi_j^\nu\right). \end{aligned} \tag{2.36}$$

Il secondo termine nell'argomento della tangente iperbolica (detto *crosstalk term*) può essere trascurato nel caso $p \ll N$. L'equazione di autoconsistenza risultante

$$m\xi_i^\nu = \tanh(\beta m\xi_i^\nu), \tag{2.37}$$

ovvero

$$\langle S_i\rangle = \tanh(\beta\langle S_i\rangle), \tag{2.38}$$

è analoga a quella per la magnetizzazione nei ferromagneti, e si avrà quindi una transizione di fase come descritto nella sezione 1.4 per il Modello di Ising.

Si può definire una temperatura critica $T_c = 1$ al di sotto della quale ci sono due soluzioni non banali, ovvero gli stati di memoria definiti come nell'Ansatz (2.35) sono stabili.

Possiamo vedere la (2.35) anche come

$$m = \frac{\langle S_i\rangle}{\xi_i^\nu} = \xi_i^\nu\langle S_i\rangle = P(\text{bit i corretto}) - P(\text{bit i non corretto}) \tag{2.39}$$

da cui

$$P(\text{bit i corretto}) = \frac{m+1}{2}. \tag{2.40}$$

Dunque il numero medio di bit corretti nella memoria recuperata sarà

$$\langle N_{correct}\rangle = NP(\text{bit i corretto}) = \frac{N}{2}m + \frac{N}{2} \tag{2.41}$$

Si può osservare chiaramente una transizione di fase, che individua nella temperatura critica T_c il confine tra una memoria utile e una memoria inutile. Per $T > T_c$ abbiamo $m = 0$, e dunque $\langle N_{correct}\rangle = \frac{N}{2}$: solo la metà dei bit in media sono corretti, come ci aspetteremmo in un pattern random e non nella configurazione corrispondente al recupero di un pattern immagazzinato, perciò la memoria è inutile. In $T = T_c$ c'è una

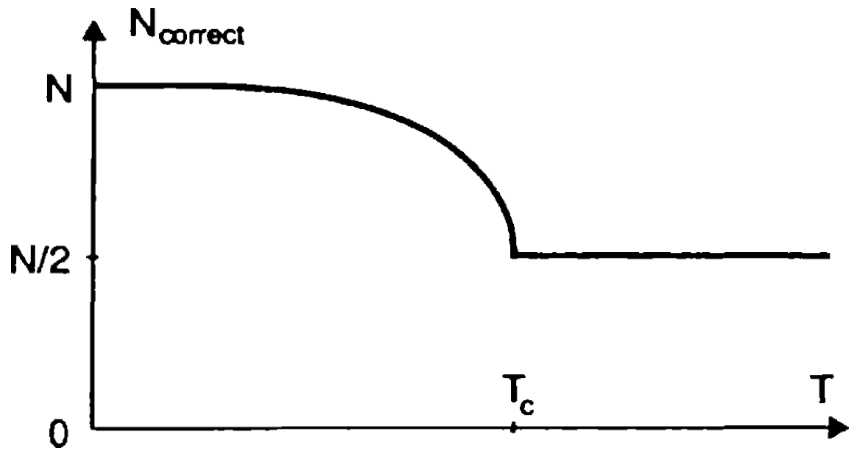

Figura 2.2: Andamento di $\langle N_{correct} \rangle$ in funzione della temperatura, da cui emerge la transizione di fase tra una memoria utile e una memoria inutile.

discontinuità brusca, e per $T < T_c$ il numero di bit corretti comincia a diventare superiore alla metà, aumentando al diminuire di T, fino a $\langle N_{correct} \rangle \to N$ per $T \to 0$. Questo andamento è riassunto in Figura 2.2. Possiamo considerare la memoria utile quando la configurazione stabile si differenzia significativamente dal caso puramente random in cui $\langle N_{correct} \rangle = \frac{N}{2}$; il limite $T \to 0$ rappresenta il caso ideale in cui tutti i bit sono corretti, ovvero la configurazione del sistema corrisponde esattamente a un pattern immagazzinato, ma possiamo ammettere anche una certa percentuale di errore, in base al criterio di performance accettabile che intendiamo stabilire.

Si può notare che questo andamento è analogo, a meno di un offset, a quello dalla curva riportata in Figura 1.5, che descrive la magnetizzazione nella transizione di fase che caratterizza il comportamento dei ferromagneti.

2.4 Stati stabili nella teoria di campo medio

2.4.1 Stati spuri

Abbiamo visto che sia seguendo la descrizione dinamica sia seguendo la descrizione energetica le memorie immagazzinate $\boldsymbol{\xi}^\mu$ sono attrattori; esse sono dette *retrieval states*. Tuttavia ci possono essere anche altri attrattori indesiderati, detti stati spuri. Lo scopo della sezione presente è analizzare i tipi di stati spuri che si formano in relazione al numero di pattern immagazzinati nel network e alla temperatura.

Il primo esempio di stati stabili diversi dai *retrieval states*, già menzionato, è quello dei *reversed states* $-\boldsymbol{\xi}^\mu$, che presentano tutti i bit invertiti rispetto alla memoria da immagazzinare, e sono attrattori poiché sia la dinamica che l'Hamiltoniana presentano una simmetria per la trasformazione $S_i \to -S_i \; \forall i$. Non si tratta esattamente di stati

spuri in quanto sono riconducibili in maniera immediata agli attrattori desiderati: è sufficiente definire un bit di segno che indichi se è necessario invertire o meno tutti i bit rimanenti. Il formarsi di questi stati stabili non compromette quindi l'efficacia della memoria nel recuperare i pattern immagazzinati. I *retrieval states* e i *reversed states* sono detti **stati ferromagnetici**, e costituiscono in tutto $2p$ stati dinamicamente stabili che corrispondono agli attrattori che vogliamo si formino nel network.

Come abbiamo visto il formarsi di stati stabili avviene solamente al di sotto della temperatura critica $T_c = 1$, e si configura quindi come il passaggio da una fase disordinata a una fase ordinata tipico di una transizione di fase. Il parametro d'ordine naturale per descrivere questa transizione di fase, analogo alla magnetizzazione media nei ferromagneti, è la sovrapposizione degli stati del sistema con i pattern immagazzinati, detta overlap, data per una certa memoria ν da

$$m_\nu(\mathbf{S}) = \frac{1}{N}\sum_{i=1}^{N}\langle S_i\rangle\xi_i^\nu \,. \tag{2.42}$$

Si tratta di un'estensione del concetto espresso nella sezione precedente dalla (2.39). Usando una notazione vettoriale, se indichiamo con $\mathbf{m} = (m_1, \ldots, m_p)$ il vettore che contiene gli overlap della configurazione del sistema con ciascuno dei pattern, e con $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^p)$ il vettore che contiene lo stato dell'i-esima unità in ciascuna delle memorie, possiamo scrivere

$$\mathbf{m}(\mathbf{S}) = \frac{1}{N}\sum_{i=1}^{N}\langle S_i\rangle\boldsymbol{\xi}_i \,. \tag{2.43}$$

Poiché negli stati ferromagnetici lo stato del sistema coincide esattamente con uno dei pattern immagazzinati (o con il suo inverso), la sovrapposizione sarà massima con questo stato e nulla con tutti gli altri. Se il sistema, per esempio, si trova nella configurazione stabile corrispondente al pattern recuperato ν, il parametro d'ordine soddisferà le condizioni

$$m_\nu = m\delta_{\nu\mu} \,, \qquad \langle S_i\rangle = \xi_i^\nu m \quad \forall i \,, \tag{2.44}$$

che nel limite di assenza di rumore $T \to 0$ diventano

$$m = 1 \,, \qquad \langle S_i\rangle = \xi_i^\nu \quad \forall i \,. \tag{2.45}$$

Gli stati ferromagnetici saranno caratterizzati da un vettore di overlap del tipo

$$\mathbf{m} = (m, 0, \ldots, 0) \,, \tag{2.46}$$

dove è sempre possibile rinominare gli indici in modo che il pattern recuperato sia il primo. Le equazioni (2.44) corrispondono all'Ansatz (2.35) fatto nella sezione precedente.

Stati spuri veri e propri, meno semplici da trattare rispetto ai *reversed states*, sono gli **stati misti** che hanno una sovrapposizione non nulla con più di uno tra i pattern da immagazzinare. In questo caso il vettore $\mathbf{m}$ sarà della forma

$$\mathbf{m} = (\underbrace{m, \dots, m}_{n volte}, \underbrace{0, \dots, 0}_{(p-n) volte}) \,. \tag{2.47}$$

Vedremo che ognuno degli stati misti esiste solamente al di sotto una certa temperatura critica $T_c < 1$, dunque basterà lavorare a una temperatura maggiore della più alta tra le temperature critiche degli stati misti presenti per eliminarne l'influenza. In questo caso il rumore (se della giusta entità) ricopre un ruolo fondamentale, eliminando gli stati misti come stati stabili e garantendo quindi un buon funzionamento del network come memoria associativa. Stati del tipo (2.47) sono detti stati misti simmetrici, in quanto la sovrapposizione con i diversi pattern, se non nulla, ha lo stesso valore per tutti i pattern. Possono esserci anche stati misti asimmetrici, che hanno cioè sovrapposizioni di entità diverse con diversi pattern, descritti da Amit, Gutfreund e Sompolinsky [13]. Questo tipo di stati, presenti solo a basse temperature, non verranno qui analizzati nel dettaglio.

Un ultimo tipo di stati spuri sono i cosiddetti **stati di spin glass**, che non saranno oggetto di questa trattazione. Si tratta di stati corrispondenti a minimi locali non correlati con nessuno dei pattern immagazzinati, che solitamente hanno bacini di attrazione molto piccoli e quindi inficiano in maniera solo marginale il funzionamento del network. Per una trattazione approfondita degli stati di spin glass si veda un secondo paper di Amit, Gutfreund e Sompolinsky [14].

2.4.2 Equazione di punto a sella nel caso $\alpha = 0$

Consideriamo inizialmente il cosiddetto limite di non saturazione, ovvero il caso in cui il numero p di pattern immagazzinati rimanga finito nel limite termodinamico in cui $N \to \infty$, dunque

$$\alpha = \frac{p}{N} \to 0 \,. \tag{2.48}$$

Seguendo la trattazione di Coolen, Kühn e Sollich [15], possiamo riscrivere l'Hamiltoniana del Modello di Hopfield (2.15), servendoci della regola di Hebb (2.7), come

$$\begin{aligned}\mathcal{H}(\mathbf{S}) &= -\frac{1}{2}\sum_{i,j}\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu S_iS_j \\ &= -\frac{N}{2}\sum_{\mu=1}^{p}\left(\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu S_i\right)\left(\frac{1}{N}\sum_{j=1}^{N}\xi_j^\mu S_j\right)+\frac{1}{2N}\sum_{i=1}^{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_i^\mu S_iS_i \\ &= -\frac{N}{2}\sum_{\mu=1}^{p}(m_\mu(\mathbf{S}))^2+\frac{1}{2}p \\ &= -\frac{N}{2}\mathbf{m}^2(\mathbf{S})+\frac{1}{2}p\,, \end{aligned} \tag{2.49}$$

dove nel penultimo passaggio si è sfruttata la definizione di overlap data in (2.42), e il secondo termine della somma è stato aggiunto per cancellare i termini diagonali $i = j$ escludendo quindi i casi di auto-accoppiamento, in modo che risulti $J_{ii} = 0$. Notiamo che l'Hamiltoniana dipende dalla configurazione del sistema $\mathbf{S}$ solo mediante il vettore di overlap $\mathbf{m}(\mathbf{S})$, che scegliamo come parametro d'ordine, e in funzione del quale svilupperemo la trattazione successiva.

La funzione di partizione risulta, trascurando il termine costante che non è rilevante nel limite termodinamico:

$$Z = \sum_{\mathbf{S}} e^{-\beta\mathcal{H}(\mathbf{S})} = \sum_{\mathbf{S}} e^{-\frac{N\beta}{2}\mathbf{m}^2(\mathbf{S})}\,. \tag{2.50}$$

Il tentativo di riformulare il problema in funzione dell'overlap $\mathbf{m}(\mathbf{S})$ invece che della configurazione del sistema $\mathbf{S}$ ci porta a introdurre la funzione di partizione vincolata e l'energia libera vincolata come quantità che diano informazioni sulla distribuzione dei valori di $\mathbf{m}(\mathbf{S})$ all'equilibrio. Per definire tali quantità osserviamo che la probabilità che $\mathbf{m}(\mathbf{S})$ assuma un certo valore $\overline{\mathbf{m}}$ fissato è data da

$$P(\overline{\mathbf{m}}) = \langle\delta(\overline{\mathbf{m}} - \mathbf{m}(\mathbf{S}))\rangle = \sum_{\mathbf{S}} p_{eq}(\mathbf{S})\delta(\overline{\mathbf{m}} - \mathbf{m}(\mathbf{S}))\,, \tag{2.51}$$

che, imponendo che all'equilibrio la distribuzione delle configurazioni del sistema $\mathbf{S}$ sia quella di Boltzmann, diventa

$$P(\overline{\mathbf{m}}) = \frac{1}{Z}\sum_{\mathbf{S}} e^{-\beta\mathcal{H}(\mathbf{S})}\delta(\overline{\mathbf{m}}) = \frac{1}{Z}\sum_{\mathbf{S}} e^{-\beta\mathcal{H}(\overline{\mathbf{m}})} = \frac{Z(\overline{\mathbf{m}})}{Z} = e^{-\beta(F(\overline{\mathbf{m}})-F)}\,. \tag{2.52}$$

Questa espressione ci porta quindi a definire la funzione di partizione vincolata e l'energia libera vincolata rispettivamente come

$$Z(\overline{\mathbf{m}}) = \frac{1}{Z}\sum_{\mathbf{S}}\delta(\overline{\mathbf{m}} - \mathbf{m}(\mathbf{S}))e^{-\beta\mathcal{H}(\mathbf{S})} \tag{2.53}$$

$$F(\overline{\mathbf{m}}) = -T \ln Z(\overline{\mathbf{m}}) \,. \tag{2.54}$$

L'aggettivo "vincolata" sottolinea il fatto che la somma nella funzione di partizione è ristretta alle configurazioni $\mathbf{S}$ che danno il corretto valore $\overline{m}$. Nel seguito indicheremo $\overline{m}$ semplicemente con m, lasciando inteso che si tratta di un valore fissato.

La funzione di partizione non vincolata può essere quindi scritta a partire della funzione di partizione vincolata come

$$Z = \int d\mathbf{m}\, Z(\mathbf{m}) \,, \tag{2.55}$$

con

$$Z(\mathbf{m}) = e^{-\beta\mathcal{H}}\mathcal{D}(\mathbf{m}) = e^{\frac{N\beta\mathbf{m}^2}{2}}\mathcal{D}(\mathbf{m}) \,, \tag{2.56}$$

dove la densità degli stati è data da

$$\mathcal{D}(\mathbf{m}) = \sum_{\mathbf{S}} \delta(\mathbf{m} - \mathbf{m}(\mathbf{S})) \,. \tag{2.57}$$

La funzione δ nella (2.57) è una δ p-dimensionale, data dal prodotto di p funzioni δ 1-dimensionali

$$\delta(\mathbf{m} - \mathbf{m}(\mathbf{S})) = \prod_{\mu=1}^{p} \delta(m_\mu - m_\mu(\mathbf{S})) \,, \tag{2.58}$$

e ci assicura che gli unici stati possibili siano quelli in cui il parametro $\mathbf{m}$ assume i valori dati dalla sovrapposizione tra la configurazione effettiva del sistema e i pattern $\mathbf{m}(\mathbf{S})$.

La densità degli stati può essere riscritta sfruttando la rappresentazione integrale della funzione δ mediante trasformata di Fourier, introducendo il momento coniugato $\mathbf{x} = (x_1, \ldots, x_p)$ come variabile di integrazione, con l'accortezza di riscalare questa variabile in modo da avere una quantità estensiva nell'esponente così che anche l'energia risulti estensiva. Abbiamo quindi

$$\mathcal{D}(\mathbf{m}) = \left(\frac{N}{2\pi}\right)^p \int d\mathbf{x}\, e^{iN\mathbf{x}\cdot\mathbf{m}} \sum_{\mathbf{S}} e^{-i\frac{1}{N}\sum_{i=1}^{N} S_i \boldsymbol{\xi}_i \cdot \mathbf{x}} \,,$$

che fattorizzando la sommatoria diventa

$$\begin{aligned}
\mathcal{D}(\mathbf{m}) &= \left(\frac{N}{2\pi}\right)^p \int d\mathbf{x}\, e^{iN\mathbf{x}\cdot\mathbf{m}} \prod_{i=1}^{N} \sum_{S_i=\pm 1} e^{-iS_i\boldsymbol{\xi}_i\cdot\mathbf{x}} \\
&= \left(\frac{N}{2\pi}\right)^p \int d\mathbf{x}\, e^{iN\mathbf{x}\cdot\mathbf{m}} \prod_{i=1}^{N} 2\cos(\boldsymbol{\xi}_i \cdot \mathbf{x}) \\
&= \left(\frac{N}{2\pi}\right)^p \int d\mathbf{x}\, e^{N\left[i\mathbf{x}\cdot\mathbf{m} + \frac{1}{N}\sum_{i=1}^{N} \ln 2\cos(\boldsymbol{\xi}_i\cdot\mathbf{x})\right]} \,.
\end{aligned} \tag{2.59}$$

Questa espressione può essere semplificata ponendo

$$g(\mathbf{m}, \mathbf{x}) = -i\mathbf{x} \cdot \mathbf{m} - \langle \ln 2 \cos(\boldsymbol{\xi} \cdot \mathbf{x}) \rangle_{\boldsymbol{\xi}} \tag{2.60}$$

e facendo uso della notazione

$$\langle \ell(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} = \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{\xi}_i) \,, \tag{2.61}$$

per indicare la media sulle unità, in modo da ottenere

$$\mathcal{D}(\mathbf{m}) = \left(\frac{N}{2\pi} \right)^p \int d\mathbf{x} \, e^{-Ng(\mathbf{m},\mathbf{x})} \,. \tag{2.62}$$

Possiamo ora usare l'approssimazione del punto a sella sviluppando $g(\mathbf{m}, \mathbf{x})$ intorno al suo punto di minimo $g(\mathbf{m}) = \min_{\mathbf{x}} g(\mathbf{m}, \mathbf{x})$

$$g(\mathbf{m}, \mathbf{x}) = g(\mathbf{m}) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \mathbf{H} (\mathbf{x} - \mathbf{x}^*) + \mathcal{O}(|\mathbf{x} - \mathbf{x}^*|^3) \,, \tag{2.63}$$

dove H_{ij} è la matrice Hessiana calcolata nel punto di minimo

$$H_{ij} = \left. \frac{\partial^2 g}{\partial x_i \partial x_j} \right|_{\mathbf{x}^*} \,. \tag{2.64}$$

La densità degli stati, mantenendo i termini fino $\mathcal{O}(|\mathbf{x} - \mathbf{x}^*|^2)$, diventa

$$\mathcal{D}(\mathbf{m}) = \left(\frac{N}{2\pi} \right)^p e^{-Ng(\mathbf{m})} \int d\mathbf{x} \, e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T \mathbf{H} (\mathbf{x}-\mathbf{x}^*)} \quad = \left(\frac{N}{2\pi} \right)^{\frac{p}{2}} \frac{1}{\sqrt{\det(H)}} e^{-Ng(\mathbf{m})} \,, \tag{2.65}$$

dove si è sfruttata la formula per l'integrale gaussiano in p dimensioni

$$\int d^p\mathbf{x} \, e^{-\frac{1}{2}\mathbf{x}^T \mathbf{A} \mathbf{x}} = \frac{(2\pi)^{\frac{p}{2}}}{\sqrt{\det \mathbf{A}}} \,. \tag{2.66}$$

L'energia libera per unità di neurone nel limite termodinamico è

$$\begin{aligned} f &= \lim_{N\to\infty} \frac{F}{N} = -\lim_{N\to\infty} \frac{T}{N} \ln \int d\mathbf{m} \, e^{\frac{N\beta\mathbf{m}^2}{2}} \mathcal{D}(\mathbf{m}) \\ &= -\lim_{N\to\infty} \left\{ \frac{T}{N} \frac{p}{2} \ln \frac{N}{2\pi} + \frac{T}{N} \ln \frac{1}{\sqrt{\det(H)}} + \frac{T}{N} \ln \int d\mathbf{m} \, e^{-\beta N \left[-\frac{\mathbf{m}^2}{2} + T g(\mathbf{m}) \right]} \right\} \\ &= -\lim_{N\to\infty} \frac{T}{N} \ln \int d\mathbf{m} \, e^{-\beta N f(\mathbf{m})} \\ &= \min_{\mathbf{m}} f(\mathbf{m}) \,, \end{aligned} \tag{2.67}$$

con

$$f(\mathbf{m}) = -\frac{\mathbf{m}^2}{2} + Tg(\mathbf{m}) \,, \tag{2.68}$$

dove nell'ultimo passaggio si è sfruttato il fatto che per $N \to \infty$ l'integrale è dominato dai contributi in cui $f(\mathbf{m})$ è piccolo, e si può quindi approssimare l'integrale con il valore massimo di $e^{-\beta N f(\mathbf{m})}$, dato dal minimo di $f(\mathbf{m})$.

Soffermandoci un attimo sull'espressione dell'energia libera vincolata (2.68) notiamo che il primo termine non è altro che l'Hamiltoniana (2.49), a meno del termine costante e divisa per il numero di unità N, e anche al secondo termine può essere data un'interpretazione fisica. Ricordando le note formule di meccanica statistica per l'energia interna

$$E = \langle \mathcal{H}(\mathbf{S}) \rangle \tag{2.69}$$

e per l'energia libera di Helmholtz

$$F = E - TS \,, \tag{2.70}$$

che divisa per N diventa

$$f = e - Ts \,, \tag{2.71}$$

dove le lettere minuscole indicano grandezze intensive, risulta evidente che si può definire una funzione $s(\mathbf{m}) = -g(\mathbf{m})$ che corrisponde all'entropia (vincolata) per neurone. Dunque la funzione $g(\mathbf{m})$, a meno del segno, ha il significato fisico di entropia per neurone.

Torniamo al calcolo dell'energia libera non vincolata e cerchiamo quindi il valore di $\mathbf{m}$ che minimizza $f(\mathbf{m})$:

$$\frac{\partial f(\mathbf{m})}{\partial m_\mu} = -m_\mu + T\frac{\partial g(\mathbf{m})}{\partial m_\mu} = 0 \qquad \mu = 1, \ldots, p \,, \tag{2.72}$$

per trovare il quale sarà prima necessario determinare $g(\mathbf{m}) = \min_{\mathbf{x}} g(\mathbf{m}, \mathbf{x})$:

$$\frac{\partial g(\mathbf{m}, \mathbf{x})}{\partial x_\mu} = 0 \,, \tag{2.73}$$

da cui, ricordando che

$$g(\mathbf{m}, \mathbf{x}) = -i\mathbf{x} \cdot \mathbf{m} - \frac{1}{N} \sum_{i=1}^{N} \ln 2 \cos(\boldsymbol{\xi}_i \cdot \mathbf{x}) \,, \tag{2.74}$$

si ha

$$i m_\mu = \langle \xi^\mu \tan(\boldsymbol{\xi} \cdot \mathbf{x}) \rangle_{\boldsymbol{\xi}} \qquad \mu = 1, \ldots, p \,, \tag{2.75}$$

o in notazione vettoriale

$$i\mathbf{m} = \langle \boldsymbol{\xi} \tan(\boldsymbol{\xi} \cdot \mathbf{x}) \rangle_{\boldsymbol{\xi}} \,. \tag{2.76}$$

Questa equazione non può essere risolta esplicitamente per $\mathbf{x}$; lasciando indicato con $\mathbf{x}(\mathbf{m})$ il valore di $\mathbf{x}$ che soddisfa la (2.76), g calcolato nel punto di minimo sarà $g(\mathbf{m}) = g(\mathbf{m}, \mathbf{x}(\mathbf{m}))$. Possiamo ora calcolare:

$$\frac{\partial g(\mathbf{m})}{\partial m_\mu} = \sum_{\nu=1}^{p} \frac{\partial g(\mathbf{m}, \mathbf{x}(\mathbf{m}))}{\partial x_\nu} \frac{\partial x_\nu(\mathbf{m})}{\partial m_\mu} + \frac{\partial g(\mathbf{m}, \mathbf{x}(\mathbf{m}))}{\partial m_\mu} = \frac{\partial g(\mathbf{m}, \mathbf{x}(\mathbf{m}))}{\partial m_\mu} = -i x_\mu \,. \quad (2.77)$$

Notiamo che per svolgere questo calcolo è stato sufficiente sapere che $\mathbf{x}(\mathbf{m})$ minimizza g rispetto a $\mathbf{x}$ (e dunque le derivate prime rispetto agli x_ν saranno nulle), senza conoscere l'espressione esatta di $\mathbf{x}(\mathbf{m})$. Inserendo la (2.77) nella (2.72), otteniamo

$$m_\mu = T \frac{\partial g(\mathbf{m})}{\partial m_\mu} = -iT x_\mu \qquad \mu = 1, \ldots, p \,, \quad (2.78)$$

da cui

$$x_\mu = i\beta m_\mu \qquad \mu = 1, \ldots, p \,, \quad (2.79)$$

o in notazione vettoriale

$$\mathbf{x} = i\beta \mathbf{m} \,. \quad (2.80)$$

Le condizioni (2.76) e (2.80) sono due equazioni accoppiate che insieme ci garantiscono di avere un punto a sella, e possono essere riscritte equivalentemente sostituendo la (2.80) nella (2.76) per eliminare il momento coniugato $\mathbf{x}$, ottenendo

$$i\mathbf{m} = \langle \boldsymbol{\xi} \tan(i\beta \boldsymbol{\xi} \cdot \mathbf{m}) \rangle_{\boldsymbol{\xi}} \quad (2.81)$$

da cui, sfruttando il fatto che $\tan(ix) = i \tanh(x)$,

$$\mathbf{m} = \langle \boldsymbol{\xi} \tanh(\beta \boldsymbol{\xi} \cdot \mathbf{m}) \rangle_{\boldsymbol{\xi}} \,. \quad (2.82)$$

La (2.82) è nota come **equazione del punto a sella** per il parametro d'ordine $\mathbf{m}$.

L'equazione di punto a sella può essere ricavata in maniera analoga sostituendo la (2.80) in (2.60) per eliminare la dipendenza di g da $\mathbf{x}$, ottenendo

$$g(\mathbf{m}) = \beta \mathbf{m}^2 - \langle \ln 2 \cosh(\beta \boldsymbol{\xi} \cdot \mathbf{m}) \rangle_{\boldsymbol{\xi}} \,, \quad (2.83)$$

da cui si può calcolare direttamente l'energia libera

$$\widetilde{f}(\mathbf{m}) = -\frac{\mathbf{m}^2}{2} + T g(\mathbf{m}) = -\frac{\mathbf{m}^2}{2} + T \langle \ln 2 \cosh(\beta \boldsymbol{\xi} \cdot \mathbf{m}) \rangle_{\boldsymbol{\xi}} \,. \quad (2.84)$$

La tilde indica che si tratta in realtà di una pseudo-energia libera, perché abbiamo utilizzato impropriamente la condizione di punto a sella per $\mathbf{m}$ invece che per $\mathbf{x}$. La pseudo-energia libera $\widetilde{f}(\mathbf{m})$ ha gli stessi punti a sella di $f(\mathbf{m})$, perciò bisogna ricordare

che è corretto utilizzarla solo se si lavora in corrispondenza dei punti a sella, espediente che tornerà utile in seguito.

Vale la pena soffermarsi su una proprietà dell'energia libera detta *self-averaging*: nel limite termodinamico essa dipende solo dalla distribuzione dei pattern, e non dal campione specifico dei pattern in esame. L'energia libera dipende infatti dalla statistica degli $\boldsymbol{\xi}_i = (\xi_i^1, \dots, \xi_i^N)$, ognuno dei quali può assumere solo 2^p valori possibili. Dunque la media sulle unità è semplicemente la media su questi 2^p valori, pesati ciascuno con la propria frequenza relativa. Per un campione fissato di p pattern basterà scorrere la lista degli $\boldsymbol{\xi}_i$ e contare le frequenze relative dei diversi valori che possono assumere; se i pattern sono estratti randomicamente e indipendentemente da una distribuzione uniforme, poiché la probabilità è per definizione il limite per $N \to \infty$ delle frequenze di conteggio, nel limite termodinamico tutte le frequenze relative diventeranno esattamente uguali con probabilità 1 indipendentemente dal particolare campione di pattern in esame. Possiamo quindi vedere le medie sulle unità anche come medie sugli $\{\boldsymbol{\xi_i}\}$ secondo

$$\langle \ell(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} = \frac{1}{N} \sum_{i=1}^{N} \ell(\boldsymbol{\xi}_i) = \frac{1}{2^p} \sum_{\boldsymbol{\xi} \in \{-1,1\}^p} \ell(\boldsymbol{\xi}_i) \, . \tag{2.85}$$

Notiamo che, se ipotizziamo che la distribuzione da cui vengono estratti i pattern sia uniforme, varranno le seguenti proprietà, che seguono dal fatto che un bit fissato mediato su tutti i pattern ha la stessa probabilità di assumere i valori ± 1:

$$\langle \xi^\mu \rangle_{\boldsymbol{\xi}} = 0 \tag{2.86}$$

$$\langle \xi^\mu \xi^\nu \rangle_{\boldsymbol{\xi}} = \delta_{\mu\nu} \tag{2.87}$$

$$\langle \xi^\mu \xi^\nu \xi^\rho \xi^\lambda \rangle_{\boldsymbol{\xi}} = \delta_{\mu\nu}\delta_{\rho\lambda} + \delta_{\mu\rho}\delta_{\nu\lambda} + \delta_{\mu\lambda}\delta_{\nu\rho} - 2\delta_{\mu\nu}\delta_{\nu\rho}\delta_{\rho\lambda} \, . \tag{2.88}$$

2.4.3 Stati stabili nel caso $\alpha = 0$

Procediamo ora a risolvere l'equazione di punto a sella. Le soluzioni più importanti sono costituite dagli stati ferromagnetici: ipotizziamo quindi ci siano soluzioni della forma

$$\mathbf{m} = (m, 0, \dots, 0) \, . \tag{2.89}$$

L'equazione di punto a sella diventa in questo caso

$$m_\mu = \langle \xi^\mu \tanh(\beta m \xi^1) \rangle_{\boldsymbol{\xi}} = \langle \xi^\mu \xi^1 \tanh(\beta m) \rangle_{\boldsymbol{\xi}} = \delta_{\mu 1} \tanh(\beta m) \, . \tag{2.90}$$

Dunque l'Ansatz (2.44) produce effettivamente una soluzione dell'equazione di punto a sella se il valore dell'overlap soddisfa:

$$m = \tanh(\beta m) \, . \tag{2.91}$$

Questa equazione è esattamente analoga all'equazione di autoconsistenza (2.37) già trovata applicando la teoria di campo medio al Modello di Hopfield. Dunque, per gli stati ferromagnetici, si ha la transizione descritta nella sezione 2.3 tra una fase con stati di memoria stabili corrispondenti ai pattern immagazzinati (in cui il numero di bit corretti dipende dalla temperatura), a $T < T_c$, e una fase in cui gli stati stabili presenti sono randomici, a $T > T_c$, con $T_c = 1$.

Cerchiamo ora soluzioni del tipo

$$\mathbf{m} = m_n(\underbrace{1,\dots,1}_{n volte},\underbrace{0,\dots,0}_{(p-n) volte}) \quad n > 1\,, \tag{2.92}$$

ovvero stati misti simmetrici. Possiamo espandere l'equazione di punto a sella per piccoli $|\mathbf{m}|$, sfruttando le proprietà (2.86) - (2.88), ottenendo

$$\begin{aligned} m_\mu &= \langle \boldsymbol{\xi} \tanh(\beta \boldsymbol{\xi}\cdot\mathbf{m})\rangle_{\boldsymbol{\xi}} = \langle \xi^\mu \beta \boldsymbol{\xi}\cdot\mathbf{m}\rangle_{\boldsymbol{\xi}} - \frac{1}{3}\langle \xi^\mu(\beta\boldsymbol{\xi}\cdot\mathbf{m})^3\rangle_{\boldsymbol{\xi}} + \mathcal{O}(\mathbf{m}^5) \\ &= \sum_\nu \beta\langle \xi^\mu\xi^\nu\rangle_{\boldsymbol{\xi}} - \frac{1}{3}\beta^3\sum_{\nu,\rho,\lambda}\langle \xi^\mu\xi^\nu\xi^\rho\xi^\lambda\rangle_{\boldsymbol{\xi}} m_\nu m_\rho m_\lambda + \mathcal{O}(\mathbf{m}^5) \\ &= \beta m_\mu + \beta^3 m_\mu\left(-\mathbf{m}^2 + \frac{2}{3}m_\mu^2\right) + \mathcal{O}(\mathbf{m}^5) \\ &= (1+\tau)m_\mu + (1+\tau)^3 m_\mu\left(-\mathbf{m}^2 + \frac{2}{3}m_\mu^2\right) + \mathcal{O}(\mathbf{m}^5) \\ &= (1+\tau)m_\mu + m_\mu\left(-\mathbf{m}^2 + \frac{2}{3}m_\mu^2\right) + \mathcal{O}(\mathbf{m}^5, \tau\mathbf{m}^3)\,, \end{aligned} \tag{2.93}$$

dove si è introdotto il parametro $\tau = \beta - 1$ e nell'ultimo passaggio $(1+\tau)^3$ è stato espanso e troncato al primo termine. Si ottiene quindi

$$m_\mu\left(\tau - \mathbf{m}^2 + \frac{2}{3}m_\mu^2\right) = 0\,, \tag{2.94}$$

che dà come soluzioni

$$m_\mu = 0\,, \qquad m_\mu = \pm\sqrt{\frac{3}{2}(\mathbf{m}^2 - \tau)} = \pm a\,. \tag{2.95}$$

Per esplicitare il valore di a notiamo che se il vettore $\mathbf{m}$ ha n componenti non nulle, allora $\mathbf{m}^2 = n^2 a$, che sostituito nella (2.95) dà

$$a = \left(\frac{3}{3n-2}\right)^{\frac{1}{2}}\tau^{\frac{1}{2}}\,. \tag{2.96}$$

Dunque, possiamo sempre riarrangiare i pattern immagazzinati con permutazioni e riflessioni $\xi^\mu \to -\xi^\mu$ in modo che gli stati misti acquisiscano la forma

$$\mathbf{m} = m_n(\underbrace{1,\dots,1}_{n volte},\underbrace{0,\dots,0}_{(p-n) volte}) , \qquad m_n = \left(\frac{3}{3n-2}\right)^{\frac{1}{2}} (\beta-1)^{\frac{1}{2}} . \tag{2.97}$$

Notiamo che sostituendo soluzioni di questo tipo nell'equazione di punto a sella si ottiene

$$0 = \langle \xi^\mu \tanh(\beta m_n M)\rangle_{\boldsymbol{\xi}} \qquad \text{se } \mu > n \tag{2.98}$$

$$m_n = \langle \xi^\mu \tanh(\beta m_n M)\rangle_{\boldsymbol{\xi}} \qquad \text{se } \mu < n \tag{2.99}$$

dove si è posto

$$M = \sum_{\nu=1}^{n} \xi^\nu . \tag{2.100}$$

La (2.98) è automaticamente soddisfatta, mentre la (2.99) ci dà una condizione sul valore di m_n:

$$m_n = \frac{1}{N}\langle M \tanh(\beta m_n M)\rangle_{\boldsymbol{\xi}} . \tag{2.101}$$

Vogliamo ora studiare la stabilità di queste soluzioni. Per farlo occorre verificare che siano punti di minimo dell'energia libera, calcolando gli autovalori della matrice Hessiana, che devono risultare tutti positivi per un punto di minimo. In questo caso risulta immediato calcolare l'Hessiana della pseudo-energia libera (2.84) piuttosto che quella dell'energia libera (2.68), poiché conosciamo $g(\mathbf{m}) = g(\mathbf{m}, \mathbf{x}(\mathbf{m}))$ solo in maniera implicita; possiamo farlo perché nei punti a sella le due funzioni coincidono. Calcoliamo dunque l'Hessiana nei punti dati dalle soluzioni della forma

$$\mathbf{m} = m_n(\underbrace{1,\dots,1}_{n volte},\underbrace{0,\dots,0}_{(p-n) volte}) \tag{2.102}$$

con n generico (che comprende quindi sia gli stati misti simmetrici se $n > 1$, sia gli stati ferromagnetici, se $n = 1$)

$$\begin{aligned} A_{\mu\nu} = \frac{\partial^2 \widetilde{f}}{\partial m_\mu \partial m_\nu} &= \delta_{\mu\nu} - \beta\langle \xi^\mu \xi^\nu [1 - \tanh^2(\beta \boldsymbol{\xi}\cdot\mathbf{m})]\rangle_{\boldsymbol{\xi}} \\ &= \delta_{\mu\nu} - \beta(\delta_{\mu\nu} - Q_{\mu\nu}) , \end{aligned} \tag{2.103}$$

dove si è posto

$$Q_{\mu\nu} = \langle \xi^\mu \xi^\nu \tanh^2(\beta \boldsymbol{\xi}\cdot\mathbf{m})]\rangle_{\boldsymbol{\xi}} . \tag{2.104}$$

Gli elementi diagonali dell'Hessiana sono

$$A_{\mu\mu} = \beta(1-q) = d \quad \text{con} \quad q = Q_{\mu\mu} = \langle \tanh^2(\beta \boldsymbol{\xi}\cdot\mathbf{m})]\rangle_{\boldsymbol{\xi}} , \tag{2.105}$$

mentre gli elementi fuori dalla diagonale sono nulli per $\mu, \nu > n$, e per $\mu, \nu < n$ hanno la forma

$$\beta \langle \xi^1 \xi^2 \tanh^2(\beta m_n M) \rangle_{\boldsymbol{\xi}} = \beta Q \,, \tag{2.106}$$

dove si è sfruttata la proprietà di *self-averaging*, per cui la media sugli $\boldsymbol{\xi_i}$ non dipende dai pattern scelti, dunque per il prodotto $\xi^\mu \xi^\nu$ possiamo sceglierne due qualsiasi, per esempio $\xi^1 \xi^2$. Dunque la matrice avrà la forma

$$\left(\begin{array}{ccccc|ccc} d & \beta Q & \dots & \dots & \beta Q & & & \\ \beta Q & \ddots & \ddots & & \vdots & & & \\ \vdots & \ddots & \ddots & \ddots & \vdots & & 0 & \\ \vdots & & \ddots & \ddots & \beta Q & & & \\ \beta Q & \dots & \dots & \beta Q & d & & & \\ \hline & & & & & d & & \\ & & & & & & \ddots & \\ & & 0 & & & & \ddots & 0 \\ & & & & & 0 & & \ddots \\ & & & & & & & d \end{array}\right)$$

Diagonalizzando questa matrice si ottengono tre sottospazi, cui corrispondono i seguenti autovettori con i rispettivi autovalori:

I. $\mathbf{x} = (1, \dots, 1, 0, \dots, 0)$ $\quad \lambda_1 = 1 - \beta(1-q) + \beta(n-1)Q$

II. $\mathbf{x} = (0, \dots, x_{n+1}, \dots, x_p)$ $\quad \lambda_2 = 1 - \beta(1-q)$

III. $\mathbf{x} = (x_1, \dots, x_n, 0, \dots, 0), \sum_\mu x_\mu = 0$ $\quad \lambda_3 = 1 - \beta(1-q) - \beta Q$

dove l'autovalore λ_1 è non degenere, l'autovalore λ_2 ha degenerazione $(p-n)$, e infine λ_3 ha degenerazione $(n-1)$.

Per dare un'interpretazione fisica a queste soluzioni notiamo che la funzione di partizione data da (2.55) e (2.56), tenendo conto della densità degli stati ottenuta in (2.65), può essere scritta come

$$Z = \int d\mathbf{m}\, e^{N\beta \frac{\mathbf{m}^2}{2}} \mathcal{D}(\mathbf{m}) = \left(\frac{N}{2\pi}\right)^{\frac{p}{2}} \frac{1}{\sqrt{\det(H)}} \int d\mathbf{m}\, e^{N\beta \frac{\mathbf{m}^2}{2}} e^{-Ng(\mathbf{m})} \,, \tag{2.107}$$

che trascurando i pre-fattori e ricordando l'espressione per la pseudo-energia libera (2.84), diventa

$$Z \simeq \int d\mathbf{m}\, e^{-N\beta \widetilde{f}(\mathbf{m})} \,. \tag{2.108}$$

Possiamo ora sviluppare $\widetilde{f}(\mathbf{m})$ intorno al suo punto di minimo $\mathbf{m}^*$ (dato dalle soluzioni della forma (2.102)) per applicare l'approssimazione del punto a sella, scrivendo

$$\widetilde{f}(\mathbf{m}) = \widetilde{f}(\mathbf{m}^*) + \frac{1}{2}(\mathbf{m} - \mathbf{m}^*)^T \mathbf{A} (\mathbf{m} - \mathbf{m}^*) + \mathcal{O}(|\mathbf{m} - \mathbf{m}^*|^3) \,. \tag{2.109}$$

Quindi risulta

$$Z \simeq e^{-N\beta\widetilde{f}(\mathbf{m}^*)} \int d\mathbf{m}\, e^{-N\frac{\beta}{2}(\mathbf{m}-\mathbf{m}^*)^T\mathbf{A}(\mathbf{m}-\mathbf{m}^*)}\,. \tag{2.110}$$

Il secondo termine ha la forma di un integrale gaussiano, con media $\mathbf{m}^*$ e varianza $\mathbf{A}^{-1}$: l'Hessiana $\mathbf{A}$ fornisce quindi una misura delle fluttuazioni intorno alla soluzione $\mathbf{m}^*$.

In particolare, l'autovettore I. descrive fluttuazioni degli n pattern che formano lo stato misto, ed essendo tutte della stessa ampiezza possiamo vederle come fluttuazioni del valore di m_n. L'autovettore II. descrive invece fluttuazioni sugli $(n-p)$ pattern che non sono correlati allo stato misto corrispondente alla configurazione del sistema, e introducono quindi una sovrapposizione anche con altri pattern. Infine l'autovettore III. descrive fluttuazioni delle n memorie che contribuiscono a formare lo stato misto, come I., tuttavia in questo caso le fluttuazioni sono di entità diverse (e possono avere anche segni diversi): ovvero, lo stato misto non sarà più simmetrico, perché le n memorie che lo formano non contribuiranno più tutte con lo stesso peso.

Si può dimostrare che vicino a $T = 1$ l'unica soluzione localmente stabile è quella con $n = 1$, corrispondente dunque agli stati ferromagnetici. Gli stati misti con n pari sono instabili ad ogni temperatura, mentre quelli con n dispari diventano stabili a una certa temperatura $0 < T_n < 1$. Ricorrendo a metodi numerici, si ottengono i seguenti valori:

$$T_3 = 0.461, \qquad T_5 = 0.385, \qquad T_7 = 0.345 \tag{2.111}$$

e in generale T_3 risulta la più alta tra le temperature critiche corrispondenti agli stati misti. Dunque, per $0.461 < T < 1$ gli unici stati stabili risultano quelli ferromagnetici, corrispondenti esattamente ai pattern che vogliamo immagazzinare nel network (o al loro inverso): come anticipato, il rumore ha un ruolo fondamentale nell'eliminare gli stati misti e garantire un buon funzionamento della memoria associativa. Per una trattazione più approfondita, si rimanda al paper di Amit, Gutfreund e Sompolinsky [13].

2.4.4 Diagramma di fase nel caso generale

Per una trattazione approfondita del caso generale, in cui $\alpha = \frac{p}{N}$ assume valori finiti, ovvero il numero di memorie immagazzinate scala proporzionalmente con il numero di unità che costituiscono il network, si rimanda a [16] e [17]. Verranno qui esposte solamente le linee principali.

Supponiamo di voler richiamare il pattern $\mu = 1$. Ci aspettiamo quindi che m_μ valga 1 per $\mu = 1$, e che sia piccolo, circa $\frac{1}{\sqrt{N}}$, per gli altri pattern. Il punto di partenza è l'equazione di autoconsistenza derivante dalla teoria di campo medio (2.34), che sostituita

nella definizione di overlap (2.42) porta a

$$\begin{aligned} m_\nu &= \frac{1}{N}\sum_{i=1}^{N}\xi_i^\nu \tanh\left(\beta\sum_\mu \xi_i^\mu m_\mu\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\xi_i^\nu\xi_i^1 \tanh\left[\beta\left(m_1 + \xi_i^\nu\xi_i^1 m_\nu + \beta\sum_{\substack{\mu\neq 1\\ \mu\neq\nu}}\xi_i^\mu\xi_i^1 m_\mu\right)\right] , \end{aligned} \tag{2.112}$$

dove l'ultimo termine nell'argomento della tangente iperbolica è il *crosstalk term* che tiene conto della sovrapposizione tra pattern diversi, e nel caso in cui α è finito non può essere trascurato. A causa della presenza di termini di correlazione tra i pattern è necessario introdurre due nuove variabili:

$$r = \frac{1}{\alpha}\sum_{\nu\neq 1} m_\nu^2 \tag{2.113}$$

$$q_{\mu\nu} = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\xi_i^\nu \, . \tag{2.114}$$

Il parametro r esprime l'overlap quadratico medio della configurazione del sistema con le memorie non richiamate. Il fattore $\frac{1}{\alpha} = \frac{N}{p}$ lo rende effettivamente una media sui p overlap quadratici, e cancella la dipendenze da $\frac{1}{\sqrt{N}}$ degli m_μ. Il parametro q è invece una sorta di prodotto scalare tra i pattern che ne esprime l'overlap; è noto come parametro di Edwards-Anderson nello studio degli spin glass [18].

La soluzione di campo medio consiste in un sistema accoppiato di tre equazioni per i parametri che descrivono il sistema $m = m_1$, r e q, detti anche parametri d'ordine

$$m = \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tanh[\beta(m + \sqrt{\alpha r}z)] \tag{2.115}$$

$$r = \frac{q}{[1-\beta(1-q)]^2} \tag{2.116}$$

$$q = \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tanh^2[\beta(m + \sqrt{\alpha r}z)] \tag{2.117}$$

con

$$z = \sum_{\mu\neq\nu,1}\xi_i^\mu\xi_i^1 m_\mu \, . \tag{2.118}$$

Si può notare che nella soluzione di campo medio lo scalare (2.113), il vettore (2.43) e la matrice (2.114) che definiscono i tre parametri d'ordine si riducono tutti a degli scalari.

Nel caso $T \to 0$ si ha che $q \to 1$, ed è utile definire la quantità non divergente

$$C = \beta(1-q) \, . \tag{2.119}$$

In questo caso le espressioni si semplificano notevolmente

$$m = \mathrm{erf}(\frac{m}{\sqrt{2\alpha r}}) \tag{2.120}$$

$$r = \frac{1}{(1-C)^2} \tag{2.121}$$

$$C = \sqrt{\frac{2}{\pi\alpha r}} e^{-\frac{m^2}{2\alpha r}} \, . \tag{2.122}$$

Ponendo per semplicità

$$y = \frac{m}{\sqrt{2\alpha r}} \, , \tag{2.123}$$

dalle tre equazioni precedenti si ottiene

$$y \left(\sqrt{2\alpha} + \frac{2}{\sqrt{\pi}} e^{-y^2} \right) = \mathrm{erf}(y) \, , \tag{2.124}$$

che può essere risolta graficamente come riportato in Figura 2.3. Con metodi numerici

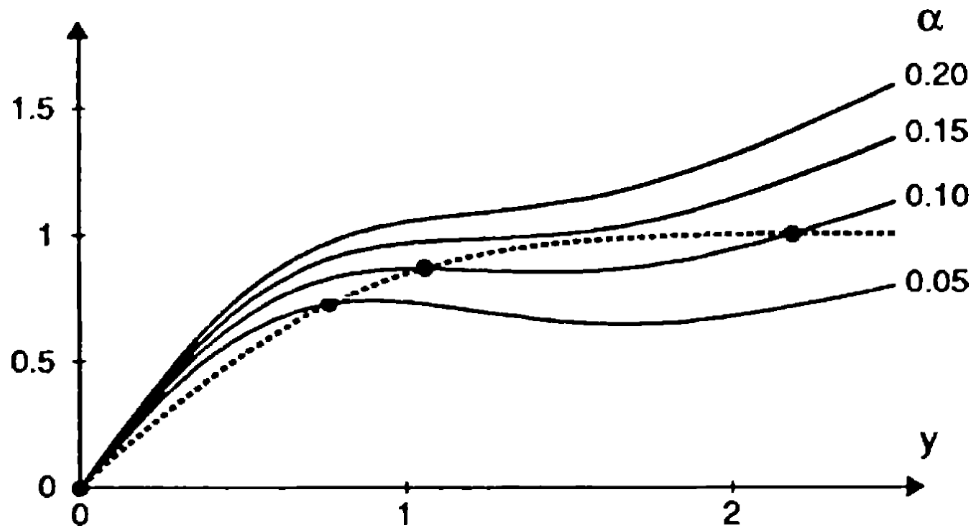

Figura 2.3: Soluzione grafica dell'equazione (2.124). Le linee continue mostrano l'andamento del primo membro per vari valori di α, mentre la linea tratteggiata mostra l'andamento del secondo membro, $\mathrm{erf}(y)$.

si ottiene il valore di α a cui le soluzioni non banali $m \neq 0$ scompaiono:

$$\alpha_c \approx 0.138 \, . \tag{2.125}$$

In questo punto si passa in maniera discontinua da un valore $m = 0.97$ (che ricordando la (2.40) indica una memoria quasi perfettamente funzionante) a $m = 0$ (ovvero una memoria inutile).

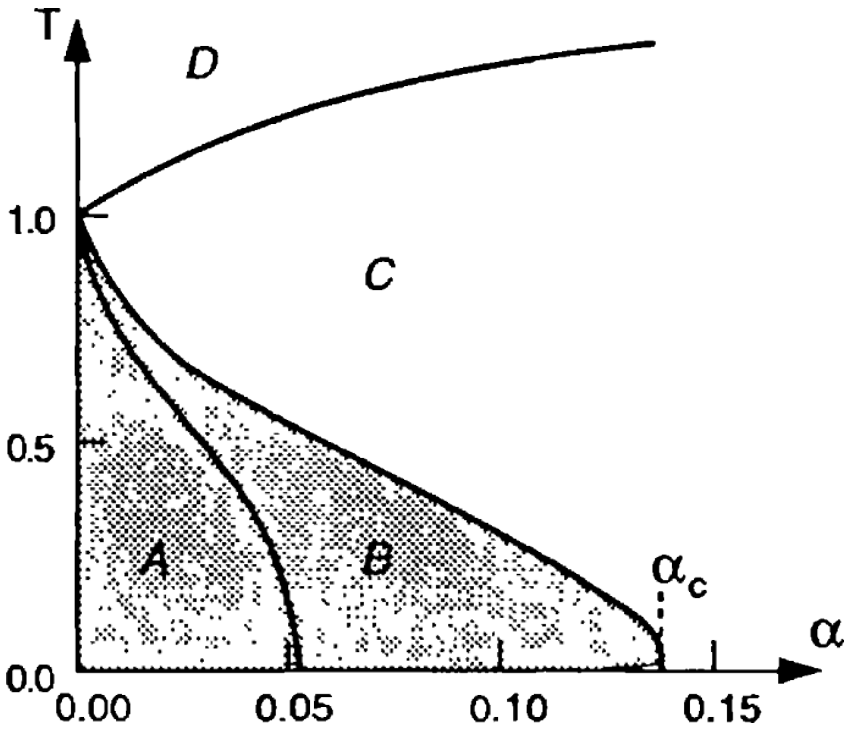

Figura 2.4: Diagramma di fase del Modello di Hopfield, ottenuto da Amit et al. [14]. Sono riportate le varie regioni in cui, al variare di α e T, sono presenti diversi tipi di stati stabili. L'area ombreggiata indica la regione in cui gli stati ferromagnetici sono stabili.

In Figura 2.4 è riportato il diagramma di fase completo del Modello di Hopfield, in funzione di T e α; al variare del livello di rumore e del rapporto tra pattern immagazzinati e numero di unità, sono riportati i diversi tipi di stati stabili che si formano.

I valori $T_c = 1$ e $\alpha_c = 0.138$ pongono i limiti della regione AB in cuingli stati ferromagnetici sono stabili. In questa regione sono presenti anche stati di spin glass, che nella regione B risultano più stabili rispetto a quelli ferromagnetici in quanto hanno una energia minore, mentre nella regione A, delimitata dal valore $\alpha = 0.051$, gli stati ferromagnetici sono minimi globali. Per bassi valori di α e T sono presenti anche gli stati misti, che hanno sempre valori in energia maggiori rispetto agli stati ferromagnetici, e risultano quindi meno stabili. Per ogni tipo di stati misti, si hanno diversi valori delle intercette con gli assi che delimitano le rispettive regioni di stabilità, ma sempre minori dei valori $T_c = 1$ e $\alpha_c = 0.138$ che delimitano la regione di stabilità degli stati ferromagnetici.

Lungo la frontiera della regione AB m diventa nullo in modo discontinuo (transizione di fase di primo ordine), tranne sull'asse delle temperature $\alpha = 0$, in cui la transizione al valore $m = 0$ avviene in modo continuo, come già analizzato nella sezione 2.3 (transizione di fase di secondo ordine).

Nella regione C sono presenti come stati stabili soltanto gli stati di spin glass, che non sono correlati con nessuna delle memorie. Alzando la temperatura, nella regione D svaniscono anche gli stati di spin glass e l'unico stato stabile presente è la soluzione banale $\langle S_i \rangle = 0$.

Dunque, in conclusione, la regione utile in cui gli stati ferromagnetici sono i minimi globali, ovvero il sistema lasciato libero di evolvere tenderà verso una configurazione corrispondente a uno dei pattern immagazzinati (o al suo inverso) risulta la regione A, delimitata sugli assi dai valori $T_c = 1$ e $\alpha = 0.051$.

Capitolo 3

Estensioni del Modello di Hopfield

La trattazione svolta fino a questo punto riguarda la forma più semplice del Modello di Hopfield. Esso, grazie proprio alla sua semplicità e al fatto che offre la possibilità di una analisi teorica profonda, è diventato il punto di partenza per numerosi studi successivi. I miglioramenti apportati da tali studi al modello originario mirano da un lato a renderlo maggiormente plausibile sul piano biologico, e dall'altro anche a favorire una più semplice realizzabilità sul piano pratico di implementazioni hardware.

Il Modello di Hopfield originario presenta infatti vari aspetti in contrasto con le evidenze biologiche, come per esempio la simmetria delle connessioni sinaptiche, il fatto che esse possano passare da eccitatorie a inibitorie nel corso del processo di apprendimento, e i livelli di attività del sistema (legati al numero di neuroni nello stato *firing*) che sono decisamente superiori rispetto a quelli osservati nella corteccia. D'altro canto, le implementazioni hardware, sia circuitali che ottiche, hanno fatto notevoli passi avanti apportando modifiche a questo modello; per un compendio si veda [19].

Generalmente ciò su cui si agisce è la forma delle connessioni sinaptiche J_{ij}, poiché come è già stato ampiamente discusso sono esse a contenere l'immagazzinamento dell'informazione nel network. Si può distinguere tra una prima generazione di estensioni del Modello di Hopfield, che propongono una formula chiusa per i J_{ij}, e una seconda generazione di estensioni che prevedono invece algoritmi iterativi di apprendimento, più lenti ma più potenti.

Verrà di seguito presentata una prima generalizzazione del Modello di Hopfield che non agisce sulla forma delle connessioni sinaptiche bensì sul tipo di valori che possono assumere le unità costituenti il network, seguita da alcune estensioni di prima generazione e da cenni alle estensioni di seconda generazione, che in entrambi i casi riguardano i coefficienti sinaptici.

Il fatto rilevante riguardo le estensioni del Modello di Hopfield è che esse conservano molte delle caratteristiche emergenti dal modello originario: si potrebbe infatti pensare che le proprietà della memoria associativa ricavate nel Modello di Hopfield siano il risultato della ultra-semplificazione di tale modello, tuttavia la resistenza di molte di queste caratteristiche a una vasta gamma di modificazioni è indicativa del fatto che tale

modello, lungi dall'essere esaustivo, potrebbe aver individuato direzioni interessanti su cui procedere.

3.1 Unità a valori continui

Un primo miglioramento del Modello di Hopfield consiste in una generalizzazione del neurone di McCulloch-Pitts in modo da rendere l'output di ogni unità una variabile continua, come suggerito dallo stesso Hopfield in un articolo successivo a quello in cui descriveva il modello originario [20]. Si tratta di una descrizione più realistica sul piano biologico, e a volte risulta più conveniente anche nelle implementazioni hardware.

Consideriamo l'output di ogni unità come una funzione continua dell'input

$$V_i = g(u_i) = g\left(\sum_j J_{ij} V_j\right) , \tag{3.1}$$

dove la variabile continua V_i corrisponde agli stati S_i del caso discreto, e l'input u_i corrisponde ai campi locali h_i. Tale funzione è detta anche funzione di attivazione ed è solitamente non lineare, in quanto è utile usare funzioni che raggiungano valori di saturazione fissati per grandi valori di $|u|$; scelte tipiche sono $\tanh(\beta u)$ corrispondente all'intervallo $[-1, +1]$ oppure $f_\beta(u)$, come definita in (2.28), corrispondente all'intervallo [0, 1]. In particolare con la prima di queste due scelte u_i e V_i corrispondono esattamente a $\langle h_i \rangle$ e $\langle S_i \rangle$ e si ottengono le equazioni di campo medio descritte nella sezione 2.3. È dunque possibile usare unità a valori continui per risolvere il problema di campo medio per un network stocastico a temperatura finita.

L'aggiornamento dello stato delle unità del network può essere non solo sincrono o asincrono come accennato in precedenza, ma anche continuo nel tempo: in questa terza possibilità tutte le unità cambiano il proprio output secondo (3.1) non in step temporali discreti bensì in maniera continua, oltre che simultanea, così come gli input u_i variano in maniera continua. Questo tipo di aggiornamento risulta utile per le implementazioni circuitali. Il cambiamento in modo continuo può essere descritto da un sistema di equazioni differenziali

$$\tau_i \frac{dV_i}{dt} = -V_i + g(u_i) \tag{3.2}$$

con τ_i costanti di tempo opportune. È immediato vedere che all'equilibrio ($\frac{dV_i}{dt} = 0$) gli stati che soddisfano la (3.1) sono stabili, e risultano dunque attrattori secondo la regola dinamica (3.2). Lo stesso risultato si ottiene anche ragionando in termini degli input con una regola dinamica analoga alla (3.2)

$$\tau_i \frac{du_i}{dt} = -u_i + \sum_j J_{ij} g(u_i) . \tag{3.3}$$

Anche in questo caso parallelamente alla descrizione dinamica è possibile seguire una descrizione mediante una funzione energia, che risulta

$$\mathcal{H} = -\frac{1}{2}\sum_{ij} J_{ij} V_i V_j + \sum_i \int_0^{V_i} g^{-1}(V) dV \tag{3.4}$$

ed è possibile dimostrare che la regola dinamica appena definita minimizza questa funzione.

3.2 Sinapsi asimmetriche

La simmetria delle connessioni sinaptiche ($J_{ij} = J_{ji}$) è un requisito fondamentale del Modello di Hopfield, poiché senza di essa non sarebbe valida la condizione di *detailed balance* e non sarebbe quindi possibile definire una funzione energia e sfruttare il formalismo della meccanica statistica all'equilibrio. Tuttavia si tratta di una condizione altamente inverosimile dal punto di vista biologico; è infatti noto che per un dato neurone le sinapsi sono di un solo tipo (eccitatorie o inibitorie) e con il requisito della simmetria questo implicherebbe due popolazioni di neuroni non connesse tra loro [21]. Per questo motivo vari tentativi vanno nella direzione di generalizzare il Modello di Hopfield al caso di sinapsi asimmetriche, ovvero tali che il valore della connessione sinaptica nella direzione $i \to j$ non sia correlato con il valore della connessione nella direzione $j \to i$.

Circuiti asimmetrici hanno in generale un repertorio di comportamenti più ampio rispetto alla convergenza a uno stato stabile presa in considerazione nel capitolo precedente. Si possono generare attrattori dipendenti dal tempo, dovuti allo sfasamento nella risposta di una sequenza di connessioni quando sono attraversate in direzioni opposte; nello spazio delle configurazioni la traiettoria del sistema può descrivere orbite periodiche o quasi-periodiche, oppure traiettorie caotiche. Se l'asimmetria è casuale (ovvero consideriamo come coefficienti sinaptici variabili random distribuite normalmente e indipendenti, per cui vale $\langle J_{ij} J_{ji} \rangle = 0$ per la definizione di asimmetria) e non troppo forte può giocare il ruolo del rumore ed essere sfruttata ai fini del funzionamento del network. Per esempio, per $\alpha = \frac{p}{N}$ finito, la presenza di asimmetria random può eliminare la presenza di stati di spin glass a $T > 0$ [22] [23], tuttavia introduce anche fluttuazioni che rallentano la convergenza a un attrattore [24].

L'asimmetria può essere ottenuta tramite un procedimento detto *dilution*, che consiste nel rimuovere in maniera randomica una frazione delle connessioni; se questa operazione è effettuata in maniera indipendente per J_{ij} e J_{ji} si ottengono connessioni asimmetriche.

Se la frazione di connessioni rimosse è finita si parla di *weak dilution*. Indicando con c la concentrazione relativa delle connessioni rimanenti dopo il processo ogni coefficiente sinaptico J_{ij} assumerà il valore dato dalla regola di Hebb (2.7) con probabilità c e sarà invece nullo con probabilità $1 - c$, ovvero, definendo una variabile C_{ij} che vale 1 se è

presente una connessione non nulla tra i neuroni i e j ed è nulla altrimenti, avremo

$$J_{ij} = C_{ij} J_{ij}^{Hebb} \, . \tag{3.5}$$

Dunque l'input che riceve l'unità i è dato da

$$h_i = \sum_{j=1}^{N} C_{ij} J_{ij}^{Hebb} S_j \, , \tag{3.6}$$

che applicando la teoria di campo medio diventa

$$\langle h_i \rangle = c \sum_{j=1}^{N} J_{ij}^{Hebb} \langle S_j \rangle \, . \tag{3.7}$$

Questa espressione conduce esattamente all'equazione di autoconsistenza (2.34) se si riscala la temperatura di un fattore c; ovvero possiamo continuare ad applicare i risultati del capitolo precedente semplicemente riscalando la temperatura.

Se in seguito al processo di *dilution* rimane solo una frazione infinitesima delle connessioni originarie si parla invece di *strong dilution*. Indicando con K il numero medio di connessioni da e verso ciascuna unità si può ricavare la capacità di immagazzinamento ripetendo tutta la trattazione riportata nella sezione 2.1.3 semplicemente sostituendo K a N. Si otterranno risultati analoghi a quelli del capitolo precedente, in particolare ragionando sull'overlap m_μ si può ottenere un'equazione di autoconsistenza analoga alla (2.120):

$$m_\mu = \mathrm{erf}\left(\frac{m_\mu}{\sqrt{2\alpha'}} \right) \, , \tag{3.8}$$

con $\alpha' = \frac{p}{K}$. Questa equazione può essere risolta graficamente trovando un valore critico $\alpha'_c = \frac{2}{\pi}$ al di sopra del quale scompaiono le soluzioni non banali $m_\mu \neq 0$. Tuttavia in questo caso la transizione $m_\mu \to 0$ per $\alpha' \to \alpha'_c$ avviene in modo continuo (transizione di fase di secondo ordine), mentre nel caso di un network totalmente connesso trattato nel capitolo precedente la transizione per $\alpha \to \alpha_c$ è discontinua (transizione di fase di primo ordine). Per una trattazione più approfondita si veda [16].

3.2.1 Associazione temporale

Per studiare circuiti asimmetrici invece che considerare traiettorie singole è utile definire fasi del sistema [25], intese come medie fra più traiettorie (o realizzazioni del sistema al variare del tempo) di quantità dinamiche sul rumore stocastico per $t \to \infty$: a differenza dei circuiti simmetrici, in cui le fasi sono stazionarie, nel caso asimmetrico possono dipendere dal tempo ed esiste una temperatura critica al di sotto della quale si ha un comportamento periodico. In questo caso l'attività delle singole unità continua a essere

casuale, ma l'attività globale di una parte macroscopica del circuito consiste in oscillazioni coerenti, con le quali si è tentato di spiegare la presenza di oscillazioni presenti nei sistemi nervosi reali.

Questo tipo di comportamento è stato sfruttato anche per riprodurre l'associazione temporale, in cui il sistema deve ricostruire una sequenza temporalmente ordinata di memorie (come può avvenire per esempio quando cantiamo una canzone o recitiamo una poesia). A questo fine si sfrutta proprio il fatto che, essendo gli attrattori dinamici, possiamo fare in modo di organizzare il flusso del sistema come una sequenza temporalmente ordinata di transizioni rapide tra stati quasi-stabili corrispondenti a memorie individuali.

In un modello di questo tipo la matrice sinaptica è costituita da una parte simmetrica corrispondente alle regola di Hebb (2.7), che garantisce la stabilità dei singoli pattern per tempi brevi, e una componente asimmetrica che codifica l'ordine temporale delle memorie data da

$$J_{ij}^{asymm} = \frac{1}{N} \sum_{\mu=1}^{p} S_i^{\mu+1} S_j^{\mu} \, . \tag{3.9}$$

In questo modo i campi locali a un certo istante t sono dati da

$$h_i(t) = \frac{1}{2} \sum_{j=1}^{N} J_{ij}(S_j(t)+1) + \frac{\lambda}{2} \sum_{j=1}^{N} J_{ij}^{asymm}(S_j(t-\tau)+1) \, , \tag{3.10}$$

dove τ è un tempo di ritardo nella risposta. Affinché l'associazione temporale funzioni è infatti necessario che le componenti simmetrica e asimmetrica lavorino su scale temporali diverse, cosicché le transizioni da una memoria all'altra non intacchino il recupero della singola memoria. λ è invece la forza relativa dell'input sinaptico ritardato: esiste un valore critico λ_c, solitamente dell'ordine dell'unità, al di sotto del quale tutte le memorie rimangono stabili, mentre per $\lambda > \lambda_c$ il sistema rimane in ogni memoria per un tempo τ e poi passa alla memoria successiva, terminando nell'ultima memoria della sequenza oppure ripartendo dalla prima se le memorie sono organizzate in maniera ciclica.

L'idea di sfruttare connessioni sinaptiche asimmetriche del tipo (3.9) per ottenere un flusso del sistema attraverso stati metastabili era stata proposta dallo stesso Hopfield [7] ed è stata ripresa in vari lavori aggiungendo il tempo di ritardo da Kleinfeld [26], Sompolinsky e Kanter [27], Peretto e Niez [28].

Un modo alternativo per immagazzinare una sequenza di pattern può essere quello di utilizzare valori di soglia $\theta_i(t)$ variabili nel tempo ciascuno in maniera dipendente dall'evoluzione temporale dell'unità S_i corrispondente, come proposto da Horn e Usher [29], in modo tale da inibire un neurone che è stato nello stato *firing* troppo a lungo, oppure attribuire al rumore un ruolo attivo nel processare sequenze temporali, come suggerito da Buhmann e Schulten [30]. Un'ulteriore applicazione dell'associazione temporale, come proposto da Amit [31], può anche essere quella di contare segnali esterni che non sono legati ai pattern.

3.3 Pattern correlati

Nel capitolo precedente è stato evidenziato come il limite fondamentale per la capacità di immagazzinamento del network sia dato dalla presenza del *crosstalk term*, dovuto alla sovrapposizione tra i pattern, nel campo locale (2.9) sentito da ciascuna unità. Questa sovrapposizione costituisce un problema considerevole nel caso in cui i pattern siano correlati; esiste una soluzione generale a questo problema, nota come *projection method* o *pseudo-inverse approach*, che può essere applicata a un qualunque insieme di $p < N$ pattern linearmente indipendenti.

Possiamo definire la matrice di sovrapposizione $\mathbf{Q}$ con elementi

$$q_{\mu\nu} = \frac{1}{N} \sum_{i=1}^{N} \xi_i^\mu \xi_i^\nu \,, \tag{3.11}$$

equivalenti al parametro di Edwards-Anderson (2.114); notiamo che il concetto di correlazione tra pattern non ha nulla a che vedere con la dipendenza lineare: $q_{\mu\nu}$ è alto, e dunque i pattern μ e ν sono correlati, se un grande numero di unità assume lo stesso stato nei due pattern. Se i pattern sono linearmente indipendenti allora $\det \mathbf{Q} \neq 0$ ed è possibili sfruttare l'inversa $\mathbf{Q}^{-1}$ per definire la matrice delle connessioni sinaptiche come

$$J_{ij} = \frac{1}{N} \sum_{\mu\nu} \xi_i^\mu (\mathbf{Q}^{-1})_{\mu\nu} \xi_j^\nu \,. \tag{3.12}$$

Si può mostrare [16] che questa forma della matrice sinaptica richiama correttamente i pattern memorizzati considerando una configurazione del sistema corrispondente a un certo pattern immagazzinato ν ($S_i = \xi_i^\nu \quad \forall i$) e verificando che la regola dinamica (2.2) lascia invariata tale configurazione:

$$S_i' = \sum_j J_{ij} \xi_j^\nu = \xi_i^\nu \quad \forall i \,. \tag{3.13}$$

Questa espressione può essere vista come un'equazione agli autovalori, con $\boldsymbol{\xi}^\nu$ autovettore di autovalore 1 della matrice J_{ij}, che può essere vista come una matrice di proiezione sul sottospazio generato dai pattern memorizzati; da qui deriva il nome *projection method*.

I vantaggi di questo metodo rispetto alla Regola di Hebb originaria sono robustezza alla sovrapposizione tra i pattern, recupero di memorie senza errore, e capacità di immagazzinamento maggiore [32] [33], tuttavia lo svantaggio è la perdita della località. Se infatti la Regola di Hebb (2.7) ha un significato biologico in quanto mette in relazione solo i neuroni pre-sinaptico e post-sinaptico, al contrario la (3.12) richiede la conoscenza dello stato di tutte le unità del network, il che non è biologicamente plausibile. La località può essere recuperata se si rinuncia ad avere una formula chiusa per le connessioni sinaptiche: esiste un algoritmo iterativo che converge alla stessa forma della matrice sinaptica data dal *projection method* usando solo informazioni locali, come verrà accennato in seguito.

3.3.1 *Sparse coding*

Benché nella trattazione svolta precedentemente gli stati *firing* e *not firing* siano stati trattati ai fini dell'immagazzinamento dei pattern semplicemente come due stati astratti, equivalenti ed ugualmente probabili, essi non sono equivalenti né per quanto riguarda le implementazioni pratiche né dal punto di vista biologico. È infatti noto che i livelli di attività nella corteccia sono di gran lunga inferiori al 50%: per costruire un modello più plausibile dovremo introdurre un bias dei valori assunti dalle unità verso lo stato -1. Ci si riferisce ai modelli in cui la maggior parte delle unità sono *not firing* in termini di biased pattern (che possono essere visti come un caso particolare di pattern correlati) o *sparse coding*. Nelle applicazioni pratiche risulta conveniente usare 0, 1 al posto di +1, -1 come valori assunti dalle unità [34], ed è immediato che in questo modo la regola dinamica (2.2) si riduce a contare le unità *firing*, dunque dal punto di vista pratico lo *sparse coding* rende il procedimento molto più veloce. Inoltre usare 0, 1 come valori ammessi ha il vantaggio di evitare il passaggio delle sinapsi da eccitatorie a inibitorie, che è ammesso dalla Regola di Hebb originaria ma risulta biologicamente implausibile.

Il primo modello di questo tipo è stato proposto da Willshaw [35], considerando un sistema sincrono con un'architettura a due strati, generalizzato successivamente a un network ricorrente totalmente connesso con dinamica asincrona da Golomb, Rubin, Sompolinski [36]. La forma proposta per i coefficienti sinaptici è

$$J_{ij} = \Theta\left(\sum_{\mu=1}^{p}(S_i^\mu + 1)(S_j^\mu + 1)\right) , \tag{3.14}$$

con

$$\Theta(x) = \begin{cases} 1 & \text{se } x > 0 \\ 0 & \text{altrimenti} . \end{cases} \tag{3.15}$$

Questo modello risulta poco efficiente per immagazzinare pattern random e non correlati come quelli considerati nel capitolo precedente, ma si rivela invece molto efficace per i biased pattern ora in esame. Aggiungendo un termine inibitorio uniforme, ovvero una costante negativa, alla (3.14) si ottengono fasi del sistema parzialmente ordinate, con un'alta correlazione con i pattern immagazzinati, a livelli di attività tipici dello *sparse coding*.

Lo *sparse coding* è tornato un argomento di grande interesse con più recenti sviluppi e applicazioni pratiche nel campo del machine learning e pattern recognition [19].

3.4 *Learning within bounds*

Per le applicazioni reali è inevitabile utilizzare algoritmi di apprendimento non lineari. Il modo più semplice per ottenerli è usare la Regola di Hebb (2.7) ma prendere una funzione non lineare del risultato come valore per i coefficienti sinaptici.

Alcuni esempi sono la discretizzazione dei valori ammessi per i coefficienti sinaptici, che può essere utile per costruire circuiti usando un numero fissato di resistenze standardizzate, il *clipping*, che consiste nel restringere i valori ammessi per le connessioni sinaptiche a un certo range fissato $|J_{ij}| \leq A$, e il caso estremo della binarizzazione, in cui i valori assunti dai J_{ij} sono ristretti a $+1$ e -1 a seconda del segno del risultato della Regola di Hebb.

Il *clipping* può essere sfruttato per ridurre il *crosstalk term* presente nella (2.9), che aumenta non solo a causa della correlazione tra i pattern come menzionato in precedenza, ma anche all'aumentare del numero di pattern immagazzinati, e diventa quindi problematico nell'apprendimento successivo di pattern, portando al breakdown se si supera il valore critico della capacità di immagazzinamento. Se confiniamo i valori dei coefficienti sinaptici all'intervallo $[-A, A]$ e consideriamo un rate di acquisizione λ, il rapporto segnale-rumore non cresce oltre $\frac{\lambda\sqrt{N}}{A}$ indipendentemente dal numero p di pattern, ostacolando il breakdown. I limiti del dominio $[-A, A]$ devono essere scelti accuratamente, in modo che siano paragonabili al livello di rumore che causa il breakdown, poiché se sono troppo grandi non impediscono il breakdown, mentre se sono troppo piccoli sopprimono l'immagazzinamento di qualsiasi memoria.

Risulta così possibile realizzare in maniera efficiente l'apprendimento successivo di pattern usando la Regola di Hebb (2.7) in modo incrementale, secondo un procedimento noto come *learning within bounds* [37] [38] che consiste nell'aggiornare i coefficienti sinaptici all'aggiunta di una nuova memoria secondo

$$J_{ij}^{new} = \begin{cases} z & \text{se } |z| \leq A \\ J_{ij}^{old} & \text{se } |z| > A \end{cases} \tag{3.16}$$

con

$$z = J_{ij}^{old} + \lambda \xi_i^\mu \xi_j^\mu \, .$$

Una memoria di questo tipo è definita palinsesto, poiché non c'è un vero e proprio valore critico della capacità di immagazzinamento a cui avviene il breakdown, bensì aggiungere nuove memorie semplicemente cancella gradualmente quelle più vecchie, mentre i pattern appena imparati possono essere richiamati senza errore. Questo modello è stato collegato al funzionamento della memoria a breve termine.

3.5 Cenni ad algoritmi di apprendimento iterativi

Il processo di apprendimento consiste nel costruire la matrice delle connessioni sinaptiche. Le estensioni di seconda generazione del Modello di Hopfield descrivono questo processo mediante algoritmi iterativi che tentano di organizzare lo spazio delle configurazioni in bacini attorno ai pattern da immagazzinare in modo tale che risultino attrattori. La maggior parte di questi algoritmi possono essere formulati in termini di una funzione

energia definita non sullo spazio delle configurazioni del sistema, bensì sullo spazio delle configurazioni della matrice sinaptica: minimizzando tale funzione i coefficienti sinaptici convergono ai valori necessari per l'immagazzinamento dei pattern. La funzione energia tiene conto dei vincoli che deve soddisfare la matrice sinaptica e assume valore nullo se tutti i vincoli sono soddisfatti, mentre in caso contrario è positiva e il suo valore è una misura della violazione dei vincoli. Una forma generale per tale funzione energia è

$$\mathcal{H}(J_{ij}) = \sum_{i=1}^{N} \sum_{\mu=1}^{p} V(h_i^\mu S_i^\mu) \, . \tag{3.17}$$

Sono possibili varie scelte per la funzione $V(x)$, che determinano algoritmi diversi.

Il vantaggio di formulare il problema in termini di una funzione energia è che consente di sfruttare i metodi della meccanica statistica all'equilibrio per determinare le proprietà della matrice sinaptica tali da soddisfare i vincoli presenti. Nonostante l'interesse teorico di questo tipo di approccio, la sua applicabilità sul piano biologico è controversa per quanto riguarda la modalità con cui i pattern da immagazzinare vengono presentati al network nel processo di apprendimento, oltre al fatto che la distinzione temporale tra fase di apprendimento e fase di recupero delle memorie è puramente artificiale. Un'altra problematica che può sorgere nella descrizione mediante una funzione energia è la presenza di eventuali minimi locali in cui il sistema può rimanere intrappolato senza raggiungere il minimo globale dato dal valore nullo dell'energia. Analogamente alla trattazione seguita nel Modello di Hopfield, si può ottimizzare questo problema introducendo una quantità opportuna di rumore termico, come proposto da Kirkpatrick et al. [39].

Primi esempi di algoritmi di apprendimento iterativi sono quelli che fanno riferimento al *perceptron algorithm* [40], o al cosiddetto *Adaline algorithm* (da *Adaptive Linear Learning*) [41] [42], che converge allo stesso risultato del *projection method* descritto precedentemente. Nel primo caso la funzione $V(x)$ in (3.17) viene scelta in modo tale da soddisfare il vincolo $h_i^\mu S_i^\mu > k \quad \forall i, \mu$, dove k è una costante positiva, mentre nel secondo caso il vincolo imposto è più stringente: $h_i^\mu S_i^\mu = k \quad \forall i, \mu$. In entrambi i casi la costante k è definita normalizzando a 1 gli elementi diagonali della matrice sinaptica.

Inoltre è stato sviluppato da Gardner un metodo per il calcolo della capacità indipendente dal particolare algoritmo di apprendimento usato [43], che può quindi essere usato per valutare la performance di un particolare algoritmo di apprendimento in base a quanto il risultato si discosta dal limite fissato da Gardner. Gardner ha trovato come capacità nel caso di pattern non correlati $\alpha = \frac{p}{N} = 2$ (di gran lunga superiore al valore critico del Modello di Hopfield $\alpha_c = 0.138$), e ha dimostrato che per pattern correlati la capacità aumenta, fino a divergere nel caso $k = 0$ se la sovrapposizione tra i pattern m tende a 1.

Conclusioni

L'analogia tra il Modello di Ising e il Modello di Hopfield costituisce un interessante esempio di come gli strumenti della meccanica statistica possano essere applicati anche a problemi in campo biologico, e in particolare alle reti neurali.

Un formalismo come quello delle transizioni di fase, tipicamente oggetto della meccanica statistica e già in campo fisico applicabile a una vasta classe di fenomeni, risulta uno strumento molto utile anche esteso ai neural network. In questo nuovo contesto il modello di Ising si riconferma un punto di riferimento, soprattutto mediante la teoria di campo medio che permette di semplificare notevolmente il problema. Dunque esso non solo è largamente sfruttato in meccanica statistica ma mostra come la meccanica statistica possa essere applicata anche a fenomeni che appartengono ad ambiti diversi da quello in cui è nata e si è sviluppata, come esemplificato dal Modello di Hopfield.

Si tratta di un modello idealizzato, che descrive un network molto semplice costituito da unità che possono assumere solo due stati. Questo modello consente di riprodurre il funzionamento di una memoria associativa, ovvero in grado di recuperare memorie immagazzinate come attrattori della dinamica o equivalentemente come minimi dell'energia; ciò avviene implementando la Regola di Hebb in una formula per i coefficienti sinaptici, e definendo un'opportuna regola dinamica oppure una funzione energia esattamente analoga all'Hamiltoniana del Modello di Ising. Tuttavia accanto ai *retrieval states* sono presenti anche stati spuri come stati stabili, che compromettono il corretto funzionamento del network. L'evoluzione del sistema è caratterizzata dalla presenza di rumore, che può essere efficacemente trattato in maniera analoga alla temperatura e può essere sfruttato per ottenere una migliore performance del network: lavorando in intervalli di temperatura opportuni è infatti possibile eliminare gli stati misti come stati stabili, in quanto essi scompaiono al di sopra di una certa temperatura critica caratteristica dei vari stati ma sempre inferiore alla temperatura che definisce la regione di stabilità dei *retrival states*.

I risultati principali sono ottenibili mediante la teoria di campo medio tanto per il Modello di Ising quanto per quello di Hopfield. Nel primo caso si ha una transizione di fase per la magnetizzazione: a $T = T_c = \frac{J\gamma}{k_B}$ il sistema passa da una fase disordinata ($T > T_c$) con magnetizzazione nulla a una fase ordinata ($T < T_c$) in cui sono presenti soluzioni non banali $m = \pm m_0$. Analogamente, il Modello di Hopfield presenta una transizione di fase a $T = T_c = 1$ da una memoria utile ($T < T_c$) a una memoria

inutile ($T > T_c$). Il corretto funzionamento di questo sistema come memoria associativa dipende non solo dalla temperatura ma anche dal numero di pattern immagazzinati: il valore della capacità di immagazzinamento $\alpha = \frac{p}{N} = 0.051$ delimita la regione in cui i *retrieval states* sono minimi globali, mentre per $0.051 < \alpha < 0.138$ essi sono stabili, ma i minimi globali del sistema sono dati dagli stati di spin glass. Per $\alpha > 0.138$ gli unici stati stabili rimangono quelli di spin glass.

Nonostante la sua semplicità questo modello risulta utile in quanto presenta caratteristiche che resistono a vari miglioramenti apportati dalle estensioni successive, motivate dalla doppia esigenza di una maggiore plausibilità sul piano biologico e di facilitare le implementazioni hardware. È possibile continuare ad applicare la teoria di campo medio e si possono recuperare i risultati del modello originario considerando network con unità a valori continui o network asimmetrici, che costituiscono modellizzazioni più realistiche del sistema nervoso. Il *projection method* fornisce una soluzione al problema per i pattern correlati e si ottengono una capacità di immagazzinamento maggiore e il recupero di memorie senza errore. Con lo *sparse coding* è inoltre possibile ottenere sinapsi che non cambino da eccitatorie a inibitorie nel corso del processo di apprendimento e ottenere livelli di attività più vicini a quelli realmente osservati nella corteccia, nonché rendere le simulazioni più veloci ed efficaci.

Queste estensioni del Modello di Hopfield da un lato esibiscono comportamenti analoghi a meccanismi di funzionamento del sistema nervoso, e potrebbero dunque costituire una chiave interpretativa per fenomeni quali la presenza di oscillazioni nella corteccia e l'associazione temporale, che possono essere ricondotte al comportamento periodico di network asimmetrici, e la memoria a breve termine, che può essere ricondotta al meccanismo dell'apprendimento successivo di pattern mediante la tecnica del *learning within bounds* . Dall'altro lato, hanno aperto la strada a innumerevoli realizzazioni hardware e simulazioni, in particolare a partire dagli anni 90 e 2000, che costituiscono elementi importanti nel campo dell'intelligenza artificiale.

Ricordando il contesto generale da cui siamo partiti, risulta evidente come la ricerca nell'ambito dei neural network si ponga al crocevia di diversi piani conoscitivi e disciplinari. Essa muove i primi passi dal piano biologico con il tentativo di applicare strumenti e mezzi d'indagine formali tipici del linguaggio fisico (in questo caso la meccanica statistica), prende poi vita come campo di ricerca a sé stante, i cui risultati possono infine declinarsi da una parte sul piano pratico nell'ambito dell'intelligenza artificiale, e dall'altra parte sul piano teorico fornendo linee guida per comprendere i meccanismi di funzionamento del sistema nervoso. Nonostante le semplificazioni fatte in questo piano siano più o meno ad hoc, motivate spesso dall'intuizione e dal desiderio di trattare il problema in modo quantitativo, come si è mostrato sono stati fatti passi avanti per incorporare più dettagli biologici in modelli risolvibili analiticamente. Il gap tra i modelli idealizzati disponibili e la realtà biologica è ancora grande. Non è chiaro, per esempio, a quale livello di organizzazione del sistema nervoso siano da applicare questi modelli, se all'intera corteccia costituita da 10^{11} neuroni, o a sotto-aree (il che porrebbe problemi

non banali, perché eventuali sotto-aree non sono sistemi dinamici isolati); inoltre non disponiamo di criteri per distinguere tra i vincoli imposti dalla biologia che sono funzionalmente importanti o che possono essere trascurati. Si tratta di un campo di ricerca che sicuramente deve crescere ancora molto prima che i modelli dei neural network possano fornire risultati predittivi sul piano biologico, ma sicuramente è in grado di individuare direzioni interessanti su cui procedere.

Bibliografia

[1] Bialek, W. (2012). *Biophysics: searching for principles*. Princeton University Press.

[2] Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117, 500.

[3] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5, 115.

[4] Cragg, B. G., & Temperley, H. N. V. (1955). Memory: the analogy with ferromagnetic hysteresis. *Brain*, 78, 304.

[5] Caianiello, E. R. (1961). Outline of a theory of thought-processes and thinking machines. *J. Theor. Biol.*, 1, 204.

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

[7] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci.*, 79, 2554.

[8] Huang, K. (1963). *Statistical mechanics*. New York—London.

[9] Griffiths, R. B. (1964). Peierls proof of spontaneous magnetization in a two-dimensional Ising ferromagnet. *Phys Rev*, 136, A437.

[10] Greiner, W., Neise, L., & Stöcker, H. (2012). *Thermodynamics and statistical mechanics*. Springer Science & Business Media.

[11] Krapivsky, P. L., Redner, S., & Ben-Naim, E. (2010). *A kinetic view of statistical physics*. Cambridge University Press.

[12] Glauber, R. J. (1963). Time dependent statistics of the Ising model. *J. Math. Phys.*, 4, 294.

[13] Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Spin-glass models of neural networks. *Phys. Rev. A*, 32, 1007.

[14] Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *Ann. Phys.*, 173, 30.

[15] Coolen, A. C., Kühn, R., & Sollich, P. (2005). *Theory of neural information processing systems*. OUP Oxford.

[16] Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Weasley.

[17] Geszti, T. (1990). *Physical models of neural networks*. Singapore: World Scientific.

[18] Mezard, M., Parisi, G., & Virasoro, M. A. (1987). *Spin Glass Theory and Beyond*, World Scientific.

[19] Palm, G. (2013). Neural associative memories and sparse coding. *Neural Netw.*, 37, 165.

[20] Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci.*, 81, 3088.

[21] Shinomoto, S. (1987). A cognitive and associative memory. *Biol. Cybern.*, 57, 197.

[22] Hertz, J. A., Grinstein, G., & Solla, S. A. (1986). Memory networks with asymmetric bonds. *AIP Conf. Proc.*, 151, 212.

[23] Crisanti, A., & Sompolinsky, H. (1987). Dynamics of spin systems with randomly asymmetric bonds: Langevin dynamics and a spherical model. *Phys. Rev. A*, 36, 4922.

[24] Parisi, G. (1986). Asymmetric neural networks and the process of learning. *J. Phys. A. Math. Gen.*, 19, L675.

[25] Sompolinsky, H. (1988). Statistical mechanics of neural networks. *Phys. Today*, 41, 70.

[26] Kleinfeld, D. (1986). Sequential state generation by model neural networks. *Proc. Natl. Acad. Sci.*, 83, 9469.

[27] Sompolinsky, H., & Kanter, I. (1986). Temporal association in asymmetric neural networks. *Phys. Rev. Lett.*, 57, 2861.

[28] Peretto, P., & Niez, J. J. (1986). Long term memory storage capacity of multiconnected neural networks. *Biol. Cybern.*, 54, 53.

[29] Horn, D., & Usher, M. (1989). Neural networks with dynamical thresholds. *Phys. Rev. A*, 40, 1036.

[30] Buhmann, J., & Schulten, K. (1987). Noise-driven temporal association in neural networks. *Europhys. Lett.*, 4, 1205.

[31] Amit, D. J. (1988). Neural networks counting chimes. *Proc. Natl. Acad. Sci.*, 85, 2141.

[32] Kanter, I., & Sompolinsky, H. (1987). Associative recall of memory without errors. *Phys. Rev. A*, 35, 380.

[33] Personnaz, L., Guyon, I., & Dreyfus, G. (1985). Information storage and retrieval in spin-glass like neural networks. *J. Phys. Let.*, 46, 359.

[34] Tsodyks, M. V., & Feigel'Man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhys. Lett.*, 6, 101.

[35] Willshaw, D. J., Buneman, O. P., & Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, 222, 960.

[36] Golomb, D., Rubin, N., & Sompolinsky, H. (1990). Willshaw model: Associative memory with sparse coding and low firing rates. *Phys. Rev. A*, 41, 1843.

[37] Parisi, G. (1986). A memory which forgets. *J. Phys. A. Math. Gen.*, 19, L617.

[38] Nadal, J. P., Toulouse, G., Changeux, J. P., & Dehaene, S. (1986). Networks of formal neurons and memory palimpsests. *Europhys. Lett.*, 1, 535.

[39] Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220, 671.

[40] Rosenblatt, F. (1962). *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Spartan Books.

[41] Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*, 4, 96.

[42] Diederich, S., & Opper, M. (1987). Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.*, 58(9), 949.

[43] Gardner, E. (1988). The space of interactions in neural network models. *J. Phys. A. Math. Gen.*, 21, 257.

Ringraziamenti

Giunti in fondo a un percorso, è il momento di guardarsi indietro e vedere dall'alto il mosaico emerso da questi tre anni, di cui riuscivo a vedere solo un pezzettino per volta mentre ci ero immersa. Il tratto costante di ogni tassello sono le persone, persone senza le quali non sarei mai arrivata dove sono ora. Ammirando questo mosaico un grande grazie va a tutte le persone che mi hanno supportata e sopportata in questi anni di sfide e decisioni importanti.

Ringrazio innanzitutto le persone che mi hanno accompagnata nel tratto finale. In primo luogo la mia relatrice Elisa Ercolessi, per la disponibilità e l'impegno non solo nel seguirmi in questo lavoro, ma anche nel ricoprire il suo incarico all'interno del dipartimento, in cui è un punto di riferimento per molti studenti. Ugualmente importante è stato l'aiuto della mia correlatrice Barbara Bravi, che ha fornito l'ispirazione fondamentale per questo lavoro e ne ha seguito lo sviluppo con grandi pazienza e dedizione, nonostante una tesi di dottorato da consegnare.

Grazie alla mia famiglia che ha sempre creduto in me e mi ha sostenuta in questo percorso.

Grazie a tutti gli amici che allo stesso modo hanno creduto in me, a partire da chi mi ha spinta in giorni ormai lontani a provare test che non avrei preso minimamente in considerazione e a chi mi è stato letteralmente affianco.

Grazie ai compagni di corso, compagni di chiacchiere e sventure che hanno reso più liete le mie giornate; grazie in particolare a Caterina e Lucia, cassa di risonanza di questi anni, consapevoli della bellezza di continuare a raccontarsi pezzi di vita nonostante ognuno prenda la propria strada.

Un grazie immenso che non so esprimere al Collegio, luogo dalle grandi sfide e dalle grandi persone, terreno fertile e accogliente in cui sentirsi a casa; grazie alla famigliola dei fisici e a Lisa che se ne è sempre presa cura. Non avrei saputo immaginare un ambiente migliore in cui trascorrere questi anni.

Grazie, infine, a Luca, che sa sempre essermi accanto.