Università degli Studi di Milano

Dipartimento di Fisica

MECCANICA STATISTICA DEI CODICI CORRETTORI: ANALISI EURISTICA DEI CODICI DI SOURLAS A CONNETTIVITÀ FINITA

Relatore:
Prof. Sergio CARACCIOLO

Correlatore:
Dott. Pietro ROTONDO

Candidato:
Davide Maria TAGLIABUE
Matr. n. 866834

A.A. 2017-2018

Sommario

La meccanica statistica dei sistemi disordinati è stata sviluppata ben oltre il suo obiettivo iniziale di spiegare i fatti sperimentali di quei particolari magneti in cui legami ferromagnetici e antiferromagnetici sono distribuiti in modo casuale. In particolare, a partire dagli anni Ottanta ci si è accorti che una serie di problemi interdisciplinari, che includono le reti neurali, l'ottimizzazione combinatoria e l'inferenza bayesiana, possono essere studiati con i medesimi strumenti utilizzati nell'ambito dei vetri di spin.

La teoria dei codici correttori (e più in generale la teoria dell'informazione) formulata inizialmente da C. Shannon negli anni Quaranta, fornisce uno degli esempi più significativi di quanto sopra affermato. Il primo a comprendere la connessione tra gli ECC e i sistemi disordinati fu N. Sourlas nel 1989, che si accorse dell'equivalenza tra il problema di decoding negli ECC e la ricerca del ground state di una particolare classe di modelli di Ising con interazioni disordinate. Le hamiltoniane fully-connected introdotte da Sourlas furono successivamente modificate per includere il caso di connettività finita (più rilevante dal punto di vista pratico) da Y. Kabashima e D. Saad a fine anni Novanta.

Sebbene alcune proprietà di questa classe di codici correttori (che d'ora in poi chiameremo *codici di Sourlas a connettività finita*) siano state studiate in un serie di articoli da Kabashima, Saad e altri autori, non siamo riusciti a trovare nella letteratura esistente un'analisi sistematica delle performance di questi ECC, come invece succede per altri codici quali i Repetition codes, i codici BCH o i più moderni ed efficienti low-density-parity-check codes (LDPC) (che includono i Gallager codes).

In questo elaborato ci proponiamo di iniziare a colmare questo divario, introducendo una particolare classe di codici di Sourlas dotati di un semplice algoritmo di decoding di cui studieremo le performance al variare dei principali parametri in gioco.

Nel dettaglio, l'elaborato si compone di tre parti, la prima delle quali dedicata a un'analisi introduttiva della meccanica statistica dei vetri di spin e della teoria ECC. Il Capitolo 1 è un'introduzione alla teoria di campo medio delle transizioni di fase. La teoria standard di campo medio di vetri di spin è analizzata più nello specifico nel Capitolo 2, in cui sono stati riproposti sia un cenno al modello di Edward-Anderson sia i conti espliciti del celebre modello di Sherrington-Kirkpatrick. Lo studio di quest'ultimo ha permesso di fornire un'inquadratura generale del metodo delle repliche. Il Capitolo 3 si concentra su un'introduzione alla teoria ECC e al Teorema di Shannon, di cui è anche riproposta una dimostrazione. Inoltre, è riportata la teoria dei più semplici codici correttori, quali i Repetition Codes e l'Hamming Code (7,4), e una descrizione di come si possa costruire un mapping tra la teoria ECC e i modelli di vetri di spin.

La seconda parte è dedicata ai fondamenti teorici e ad un'analisi preliminare della simulazione dei codici di Sourlas a connettività finita. Il Capitolo 4 definisce gli aspetti teorici dei codici di Sourlas e una definizione formale del parametro di overlap. Il capitolo 5, invece, si concentra sull'implementazione numerica di questi codici. Nello specifico, si definisce una funzione hamiltoniana di vetri di spin e si mostra come può essere rappresentata tramite factor graph. Da ultimo, si discutono

le condizioni al contorno della simulazione, ponendo l'accento su alcune questioni fondamentali per la ricerca in questo campo.

Nella terza parte proponiamo uno studio numerico approfondito delle prestazioni della classe di codici di Sourlas introdotta nei capitoli precedenti in funzione dei principali parametri in gioco: il grado dell'interazione K, il rate R e l'errore intrinseco del canale p. Il primo risultato che mostriamo nel Capitolo 6 è un confronto con i dati numerici di Kabashima e Saad ottenuti usando l'algoritmo di decoding di belief propagation per $K = 2$ e $R = 1/2$. Nonostante il nostro algoritmo di decoding euristico sia particolarmente rudimentale, le prestazioni dei codici si equivalgono. In una seconda fase abbiamo studiato la performance dei nostri codici al variare di K, R e p, confrontandoli con: (i) i Repetition codes; (ii) i codici BCH e di Hamming e (iii) i Gallager codes. I codici di Sourlas risultano essere sistematicamente più performanti di (i), (ii) per $K \geq 3$, mentre non raggiungono (tranne in un caso particolare) le prestazioni dei codici di Gallager.

Indice

Parte I

INTRODUZIONE ALLA MECCANICA STATISTICA DEI VETRI DI SPIN E DELLA TEORIA DELLA CORREZIONE DEGLI ERRORI

Capitolo 1

Teoria di campo medio per il modello di Ising ferromagnetico

1.1 Modello di Ising

Questo primo capitolo è dedicato a una presentazione generale del *modello di Ising*, uno dei modelli più semplici di interazione in un sistema a multicorpi [17].

Iniziamo col definire un insieme di interi $\mathcal{V} = \{i\}_{i=1,\dots N}$, con $N \in \mathbb{N}$, come *reticolo*, e un suo elemento i come *sito*. Per il momento utilizziamo la notazione relativa al magnetismo, pertanto assegnamo una variabile S_i ad ogni sito: il *modello di Ising* è caratterizzato da valori binari $S_i = \pm 1$. In aggiunta, definiamo *legame* una coppia di siti (ij), e indichiamo con $\mathcal{B} = \{(ij)\}$ l'insieme di tutte le possibili coppie.

Assegnamo un'*energia di interazione* $-JS_iS_j$ per ogni coppia di $\mathcal{B}$. Nel caso del magnetismo, $S_i = 1$ indica un *up state* di spin ($\uparrow$) e $S_i = -1$ indica un *down state* di spin ($\downarrow$). Chiaramente la coppia di spin tende ad essere orientata nella stessa direzione ($\uparrow\uparrow$) o ($\downarrow\downarrow$) quando $J > 0$, dal momento che l'energia viene minimizzata: in questa configurazione lo stato è più stabile. L'interazione positiva $J > 0$ è detta *interazine ferromagnetica*, mentre quella negativa $J < 0$ è detta *interazione antiferromagnetica*.

In generale, ogni sito ha una propria energia $-hS_i$ (che corrisponde all'energia di Zeeman nel magnetismo). Pertanto, l'hamiltoniana del sistema assume la seguente forma:

$$H = -J \sum_{(ij)\in\mathcal{B}} S_iS_j - h\sum_{i=1}^{N} S_i, \tag{1.1}$$

che definsice proprio l'hamiltoniana di Ising Spin.

L'obiettivo della Meccanica Statistica consiste nel calcolare la *media termica* di un sistema fisico tramite la distibuzione di probabilità

$$P(\mathbf{S}) = \frac{\mathrm{e}^{-\beta H}}{Z}, \tag{1.2}$$

dove con $\mathbf{S} \equiv \{S_i\}$ indichiamo l'insieme degli stati di spin. L'equazione (1.2) prende il nome di *equazione di Gibbs-Boltzmann*. Mettendoci in un sistema di riferimento

tale per cui la costante di Boltzmann è unitaria (ossia $k_B = 1$, da cui segue che $\beta = 1/T$), si ottiene che la funzione di partizione Z normalizzata corrisponde a

$$Z = \sum_{S_1=\pm 1} \sum_{S_2=\pm 1} \cdots \sum_{S_N=\pm 1} \mathrm{e}^{-\beta H} = \sum_{\mathbf{S}} \mathrm{e}^{-\beta H} \,. \tag{1.3}$$

Da qui in avanti, questa notazione sarà usata in modo del tutto equivalente a

$$Z = \mathrm{Tr}\, \mathrm{e}^{-\beta H} \,. \tag{1.4}$$

1.2 Parametro d'ordine e transizione di fase

Nel modello di Ising, una delle grandezze fondamentali per la caratterizzazione di proprietà macroscopiche in interazioni ferromagnetiche è la *magnetizzazione*. Definita come

$$m = \frac{1}{N} \left\langle \sum_{i=1}^{N} S_i \right\rangle, \tag{1.5}$$

la magnetizzazione costituisce un cosiddetto *parametro d'ordine*, ossia una misura dello stato d'ordine del sistema.

In accordo con la distribuzione di G-B (1.4), per $\beta \gg 1$ gli stati a minor energia hanno una probabilità maggiore di essere popolati. Assumendo $h = 0$ nel modello di Ising, questo corrisponde ad avere un sistema in cui quasi tutti gli spin sono diretti nella stessa direzione. Pertanto gli stati di spin saranno $S_i = 1$ (oppure $S_i = -1$) per quasi tutti i siti, da cui segue che la magnetizzazione tende al valore 1 (oppure -1).

Al decrescere di β, anche gli stati a più alta energia vengono popolati con maggior probabilità e questo comporta che S_i oscilli tra i valori 1 e -1: ciò implica una diminuzione della magnetizzazione. Esiste una *temperatura critica* T_c tale per cui si ha $m \neq 0$ per $T < T_\mathrm{c}$ e $m = 0$ per $T > T_\mathrm{c}$.

Questo particolare tipo di fenomeno in un sistema macroscopico prende il nome di *transizione di fase*, caraterizzata da una acuta variazione del parametro d'ordine da valori non nulli a valori nulli. La fase con $m \neq 0$ per $T < T_\mathrm{c}$ è detta *ferromagnetica*, mentre quella con $m = 0$ per $T > T_\mathrm{c}$ *paramagnetica*.

1.3 Teoria di campo medio

1.3.1 Hamiltoniana di campo medio

In linea teorica, è possibilie calcolare il valore di aspettazione di ogni osservabile tramite la (1.2). Tuttavia, questa definizione risulta in genere troppo complicata per un conto esplicito e pertanto si utilizza la cosiddetta approssimazione di *campo medio*.

Assumiamo per comodità che nel nostro problema la magnetizzazione assuma il valor medio $m = \sum_i \langle S_i \rangle / N = \langle S_i \rangle$ e che, detta $\delta S_i = S_i - m$ la fluttuazione,

questa sia trascurabile al secondo ordine nell'energia di interazione:

$$\begin{aligned}
H &= -J \sum_{(ij)\in\mathcal{B}} S_i S_j - h \sum_{i=1}^{N} S_i = \\
&= -J \sum_{(ij)\in\mathcal{B}} (m+\delta S_i)(m+\delta S_j) - h \sum_{i=1}^{N} S_i = \\
&= -Jm^2 N_B - Jm \sum_{(ij)\in\mathcal{B}} (\delta S_i + \delta S_j) - h \sum_{i=1}^{N} S_i + \mathcal{O}(\delta^2) = \\
&= -Jm^2 N_B - 2Jm \sum_{(ij)\in\mathcal{B}} (\delta Si) - h \sum_{i=1}^{N} S_i + \mathcal{O}(\delta^2) = \\
&= -Jm^2 N_B - 2Jm \sum_{(ij)\in\mathcal{B}} (Si - m) - h \sum_{i=1}^{N} S_i + \mathcal{O}(\delta^2) = \\
&= Jm^2 N_B - (Jmz + h) \sum_{i=1}^{N} S_i + \mathcal{O}(\delta^2)
\end{aligned}$$

da cui otteniamo

$$H = Jm^2 N_B - (Jmz + h) \sum_{i=1}^{N} S_i. \tag{1.6}$$

Sono state fatte le seguenti assunzioni:

- N_B corrisponde al numero totale di coppie in $\mathcal{B}$;
- il numero z è indipendente dal sito i-esimo ed è legato a N_B dalla relazione $zN/2 = N_B$;
- il valore di aspettazione $\langle S_i \rangle$ è indipendente da i.

1.3.2 Equazione di stato

A questo punto, grazie al calcolo dell'hamiltoniana (1.6) appena svolto, sostituendo quest'ultima nell'equzione (1.4) è possibile calcolare la funzione di partizione:

$$\begin{aligned}
Z = \operatorname{Tr} \mathrm{e}^{-\beta H} &= \sum_{\mathbf{S}} \mathrm{e}^{-\beta(Jm^2 N_B - (Jmz+h)\sum_i S_i)} = \\
&= \mathrm{e}^{-\beta J m^2 N_B} \sum_{\mathbf{S}} \mathrm{e}^{\beta(Jmz+h)\sum_i S_i} = \\
&= \mathrm{e}^{-\beta J m^2 N_B} \sum_{S_1=\pm 1} \sum_{S_2=\pm 1} \cdots \sum_{S_N=\pm 1} \mathrm{e}^{\beta(Jmz+h)\sum_i S_i} = \\
&= \mathrm{e}^{-\beta J m^2 N_B} \sum_{S_1=\pm 1} \sum_{S_2=\pm 1} \cdots \sum_{S_N=\pm 1} \prod_i \mathrm{e}^{\beta(Jmz+h) S_i} = \\
&= \mathrm{e}^{-\beta J m^2 N_B} \left(\sum_{S=\pm 1} \mathrm{e}^{\beta(Jmz+h)S} \right)^N = \\
&= e^{-\beta J m^2 N_B} [2\cosh \beta(Jmz+h)]^N.
\end{aligned} \tag{1.7}$$

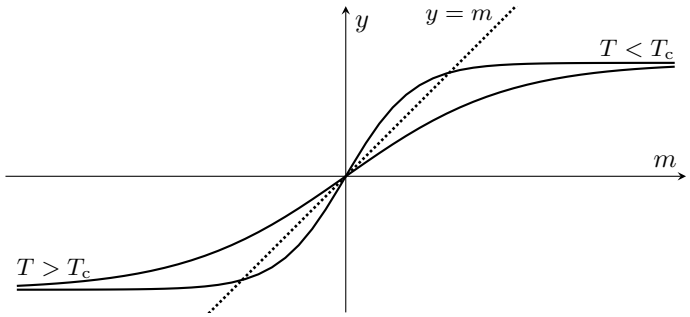


Figura 1.1: Soluzione grafica dell'equazione di stato di campo medio, in cui sono rappresentate le curve $y = m$ e $y = \tanh(\beta J m z)$. Si osservi che la soluzione $m = 0$ è sempre presente, mentre le due soluzioni $m \neq 0$ esistono solo se $\beta J z > 1$, ossia solo se $T < T_\mathrm{c}$.

Nota dunque la funzione di partizione Z, con un procedimento analogo calcoliamo la magnetizzazione:

$$\begin{aligned} m &= \mathrm{Tr}\ \frac{S_i\, \mathrm{e}^{-\beta H}}{Z} = \frac{1}{Z} \sum_{\mathbf{S}} S_i\, \mathrm{e}^{-\beta (J m^2 N_B - (Jmz+h)\sum_i S_i)} = \\ &= \frac{\mathrm{e}^{-\beta J m^2 N_B}}{Z} \sum_{\mathbf{S}} S_i\, \mathrm{e}^{\beta (Jmz+h)\sum_i S_i} = \\ &= \frac{1}{[2\cosh\beta(Jmz+h)]^N} \sum_{\mathbf{S}\setminus\{S_i\}} \prod_{j\neq i} \mathrm{e}^{\beta(Jmz+h)S_j} \sum_{S_i=\pm 1} S_i\, \mathrm{e}^{\beta(Jmz+h)S_i} = \\ &= \frac{[2\cosh\beta(Jmz+h)]^{N-1}}{[2\cosh\beta(Jmz+h)]^N}\, 2\sinh\beta(Jmz+h) = \\ &= \tanh\beta(Jmz+h). \end{aligned}$$

Otteniamo pertanto l'equazione della magnetizzazione

$$m = \tanh\beta(Jmz+h). \tag{1.8}$$

Supponendo per semplicità che $h = 0$, la soluzione dell'equazione può essere ottenuta per via grafica, come mostra la Figura 1.1. Chiaramente si ha sempre come soluzione $m = 0$; l'esistenza di ulteriori soluzioni con $m \neq 0$ dipende dalla pendenza della curva $f(m) = \tanh\beta(Jmz+h)$. Se la pendenza nell'origine è maggiore dell'unità, ossia $\beta J z > 1$, esistono ulteriori due soluzioni $(\pm m)$. É evidente che la temperatura critica sia quindi $T_\mathrm{c} = Jz$.

1.3.3 Teoria di Landau

In meccanica statistica si definisce l'*energia libera* come

$$F = -T\log Z = -NT\log[2\cosh\beta(Jmz+h)] + N_B J m^2 \tag{1.9}$$

dove per Z si è fatto uso dell'equzione (1.7). Assumendo $h = 0$ e T in un intorno stretto di T_c, ci aspettiamo che la magnetizzazione m sia vicina a zero. Pertanto

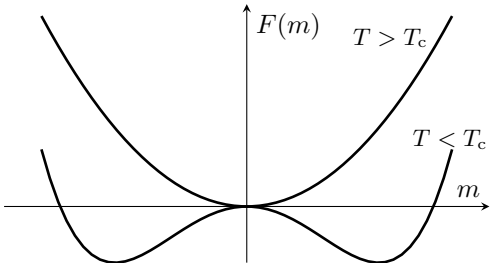


Figura 1.2: Energia libera come funzione del parametro d'ordine m: i minimi si trovano in $m \neq 0$ per $T < T_c$ oppure in $m = 0$ per $T > T_c$.

la teoria di Landau propone uno sviluppo al quarto ordine della funzione F nella variabile m:

$$F = -NT \log 2 + \frac{JzN}{2}(1 - \beta Jz)m^2 + \frac{N}{12}(Jzm)^4\beta^3 \tag{1.10}$$

da cui si nota che il segno del coefficiente di m^2 cambia in T_c.

Chiaramente se $T < T_c$ i minimi dell'energia si troveranno in $m \neq 0$, mentre per $T > T_c$ il minimo è proprio in $m = 0$ (Figura 1.2). Dal momento che la media statistica di una grandezza fisica secondo la distibuzione di G-B corrisponde al suo valore nello stato che minimizza l'energia libera, abbiamo una conferma dei risultati precedentemente trovati.

1.4 Modello infinite-range

Il modello di campo medio costituisce chiaramente un'approssimazione. Tuttavia, nel caso in cui tutte le possibili coppie di stati interagiscono tra loro, allora il problema è perfettamente risolubile. Questo modello prende il nome di *modello infinite-range*.

Riscriviamo l'hamiltoniana (1.1) come

$$H = -\frac{J}{2N}\sum_{i \neq j} S_i S_j - h\sum_{i=1}^{N} S_i, \tag{1.11}$$

dove il fattore $2N$ compare al denominatore per rendere l'hamiltoniana H una funzione estensiva, ossia $\mathcal{O}(N)$ (il numero di termini nella sommatoria è infatti pari a $N(N-1)/2$).

In accordo con la (1.4), la funzione di partizione diventa[1]

$$Z = \mathrm{Tr}\,\exp\Big(\frac{\beta J}{2N}\Big(\sum_i S_i\Big)^2 - \frac{\beta J}{2} + \beta h\sum_{i=1}^{N} S_i\Big). \tag{1.12}$$

[1] Oserviamo semplicemente che

$$\sum_{i \neq j} S_i S_j = \Big(\sum_i S_i\Big)^2 - \sum_i (S_i)^2 = \Big(\sum_i S_i\Big)^2 - N$$

Dal momento che $-\beta J/2$ è un termine $\mathcal{O}(1)$, nel limite termodinamico in cui $N \to \infty$ esso è perfettamente trascurabile. Non potendo calcolare direttamente la traccia con il termine $(\sum_i S_i)^2$ all'esponente, decomponiamo quest'ultimo con l'*integrale gaussiano*. Osservando che, in generale, si ha

$$e^{\alpha x^2/2} = \sqrt{\frac{\alpha N}{2\pi}} \int_{-\infty}^{+\infty} dm\, e^{-N\alpha m^2/2 + \sqrt{N}\alpha m x},$$

ponendo $\alpha = \beta J$ e $x = \sum_i S_i/\sqrt{N}$ e sostituendo nella (1.13) si ricava:

$$\begin{aligned} Z &= \mathrm{Tr} \sqrt{\frac{\beta JN}{2\pi}} \int_{-\infty}^{+\infty} dm\, \exp\left(-\frac{N\beta J m^2}{2} + \beta J m \sum_i S_i + \beta h \sum_i S_i\right) = \\ &= \sqrt{\frac{\beta JN}{2\pi}} \int_{-\infty}^{+\infty} dm\, \exp\left(-\frac{N\beta J m^2}{2}\right) \mathrm{Tr}\, \exp\left(\beta(Jm+h)\sum_i S_i\right) = \\ &= \sqrt{\frac{\beta JN}{2\pi}} \int_{-\infty}^{+\infty} dm\, \exp\left(-\frac{N\beta J m^2}{2}\right) [2\cosh\beta(Jm+h)]^N = \\ &= \sqrt{\frac{\beta JN}{2\pi}} \int_{-\infty}^{+\infty} dm\, \exp\left(-\frac{N\beta J m^2}{2} + N\log[2\cosh\beta(Jm+h)]\right). \end{aligned} \tag{1.13}$$

A questo punto, per valutare l'integrale di cui sopra si può far uso del *metodo di punto-sella*, riportato nell'Appendice A [23]. Infatti, sotto la condizione di limite termodinamico $N \to \infty$, l'integrale tende (a meno di una costante moltiplicativa) al valore della funzione integranda valutata nel suo massimo m_0. Detta $\mathrm{e}^{f(m)}$ l'integranda, chiaramente m_0 deve soddisfare la condizione $\frac{\partial f(m)}{\partial m}|_{m=m_0} = 0$. Scritto esplicitamente:

$$\frac{\partial}{\partial m}\left(-\frac{\beta J m^2}{2} + \log[2\cosh\beta(Jm+h)]\right) = 0$$

che porta all'equazione

$$m = \tanh\beta(Jm+h). \tag{1.14}$$

Da un confronto di quest'ultima con la (1.8) notiamo che, sostituendo nella (1.14) J con J/N e z con N, le due equazioni diventano uguali.

La grandezza m, che in questo caso è stata introdotta come variabile di integrazione, assume un significato fisico, ossia la *magnetizzazione*, in accordo con l'analogia delle equazioni (1.8) e (1.14). Osservando la prima uguaglianza della (1.13), risulta ragionevole definire m come

$$m = \frac{1}{N}\sum_i S_i, \tag{1.15}$$

il cui valore è in accordo con il valor medio della magnetizzazione nel limite termodinamico in cui $N \to \infty$. In altre parole, le fluttuazioni delle magnetizzazioni si annullano nel limite termodinamico nel modello infinity-range e quindi la teoria di campo medio restituisce un valore esatto.

Capitolo 2

Vetri di spin: definizioni preliminari e modello di Sherrington-Kirkpatrick

Nell'hamiltoniana (1.1), le interazioni tra gli spin sono uniformi nello spazio e assumono un unico valore J. Tuttavia, nel caso in cui le interazioni tra gli spin non siano uniformi, l'analisi del capitolo precedente risulta inutile (come nel caso di sistemi in cui le interazioni sono ferromagnetiche per alcuni legami e antiferromagnetiche per altri). Si introduce pertanto un nuovo modello, detto *vetri di spin*.

A differenza di un reticolo cristallino, la posizione degli atomi nel vetro non segue uno schema fisso. La peculiarità di questo materiale è che tale configurazione degli atomi apparentemente casuale non cambia nel giro di un giorno o due in un'altra configurazione atomica. Il termine vetri di spin implica proprio un'analogia tra l'orientazione degli spin e la disposizione degli atomi nel vetro: gli spin sono casualmente *congelati* nei vetri di spin.

L'obiettivo di questo capitolo è formalizzare la teoria di vetro di spin, per chiarire le condizioni di esistenza di questo stato [17].

2.1 Vetri di spin e modello di Edward-Anderson

2.1.1 Modello di Edward-Anderson

Supponiamo che per ogni coppia di spin (ij) ci sia una propria interazione J_{ij}, e supponiamo che le variabili di spin siano del tipo Ising ($S_i = \pm 1$). L'hamiltoniana, in assenza di campo magnetico esterno, ha pertanto la forma

$$H = -\sum_{(ij)\in\mathcal{B}} J_{ij} S_i S_j, \tag{2.1}$$

dove assumiamo che J_{ij} sia distributia con probabilità $P(J_{ij})$. Infatti, la casualità di J_{ij} dipende dal problema specifico e in generale è impossibile identificare la distribuzione precisa degli atomi, motivo per cui è necessario trattare il problema in termini di distribuzioni di probabilità. Un modello di questo tipo prende il nome di *modello di Edward-Anderson* [4].

Il calcolo di una grandezza fisica tramite l'hamiltoniana (2.1) ha inizio con l'operatore di traccia sulle variabili di spin $\mathbf{S} = \{S_i\}$, per un insieme fissato di J_{ij} generato dalla distribuzione di probabilità $P(J_{ij})$. Il motivo per cui eseguiamo la traccia prima su $\mathbf{S}$ per un fissato $\mathbf{J}$ è che la posizione degli atomi (portatori di spin) è casuale nello spazio, ma fissata nella scala di tempo di moto degli spin. Un disordine di questo tipo (che in letteratura viene definito come *quenched*), implica pertanto che le interazioni J_{ij} siano costanti nella scala di tempo in cui gli spin S_i subiscono delle fluttuazioni. Un disordine di questo tipo è in contrapposizione a un altro tipo di disordine, detto *annaeled*, tale per cui la scala dei tempi delle variazioni di $\mathbf{J}$ e $\mathbf{S}$ è la stessa [3]. Mentre per sistemi a temperature elevate una media annaeled sul disordine potrebbe essere corretta, per sistemi a basse temperature sarebbe sicuramente errata (dal momento che gli spin sono congelati in uno stato determinato).

Per il momento, l'energia libera assume la forma

$$F = -T \log \operatorname{Tr} \mathrm{e}^{-\beta H}, \tag{2.2}$$

che è una funzione di $\mathbf{J} \equiv \{J_i\}$. Il passo successivo è mediare la (2.2) sulla distribuzione di $\mathbf{J}$ per ottenere l'espressione generale dell'energia libera. Definiamo dunque la *media configurazionale* (d'ora in avanti indicata sempre tramite $[\dots]$) come

$$[F] = -T[\log Z] = -T \int \prod_{(ij)} d\bar{J}_{ij}\, P(J_{ij}) \log Z. \tag{2.3}$$

Il problema fondamentale è che, in teoria, ogni osservabile dipende da $\mathbf{J}$, inclusa l'energia libera. Questo è un duro scoglio, dal momento che ciò sembra suggerire che le propietà dei vetri di spin cambino al variare di $\mathbf{J}$, mentre noi vorremmo costruire una teoria più generale. Pertanto, il buon senso ci suggerisce di assumere, per sistemi sufficientemente grandi, che le proprietà fisiche non dipendano da $\mathbf{J}$. Queste grandezze prendono il nome di *grandezze self-averaging*. Quindi, se chiamiamo $f_N(\beta, \mathbf{J}) := F(\beta, \mathbf{J})/N$ l'energia libera per grado di libertà (l'energia libera è una grandezza self-averaging), in accordo con quanto appena detto avremo che

$$\lim_{N\to\infty} f_N(\beta, \mathbf{J}) = f_\infty(\beta). \tag{2.4}$$

In questo caso è chiaro che la media sul disordine di una grandezza self-averaging sia uguale al suo valore $\mathbf{J}$-indipendente, ossia

$$[f] = f_\infty(\beta). \tag{2.5}$$

Da un punto di vista analitico tutto ciò è molto positivo: il risultato che si ottiene mediando su $\mathbf{J}$ è in accordo con il valore fisico dell'osservabile. Richiedere che una grandezza sia self-averaging è fondamentalmente lo stesso che righiedere che la distribuzione di tale grandezza sia, per N sufficientemente grande, fortemente piccata attorno al suo valor medio.

2.1.2 Motodo delle repliche

La dipendenza di $[\log Z]$ da $\mathbf{J}$ è complicata e tutt'altro che semplice da gestire da un punto di vista computazionale. Una semplificazione è costituita dalla relazione [1]

$$[\log Z] = \lim_{n\to 0} \frac{[Z^n] - 1}{n}. \tag{2.6}$$

Il significato concreto di tale equazione è il seguente: si considerano n repliche del sistema originale, si valuta la media configurazionale del prodotto della loro funzione di partizione Z^n e infine si prende il limite $n \to 0$. Questa tecnica, che prende il nome di *metodo delle repliche* [18], è stata sviluppata proprio per arginare il problema di calcolare esplicitamente $[\log Z]$. Infatti, da un punto di vista computazionale lavorare con $[Z^n]$ è sicuramente molto più vantaggioso rispetto che con $[\log Z]$.

L'equazione (2.6) è sempre verificata, me è necesario sottolineare che nel nostro caso $n \in \mathbb{N}$. Più avanti verificheremo le condizioni entro cui l'equazione (2.6) abbia effettivamente senso e analizzeremo le problematiche legate al metodo delle repiche [13].

2.2 Modello di Sherrington-Kirkpatrick

2.2.1 Modello SK

La teoria di campo medio dei vetri di spin è solitamente svilppata per il *modello di Sherrington-Kirkpatrick* (SK) [20], ossia la versione infinite-range del modello di Edwards-Anderson. Definiamo pertanto l'hamiltoniana

$$H = -\sum_{i<j} J_{ij} S_i S_j - h \sum_i S_i \tag{2.7}$$

dove supponiamo che l'interazione J_{ij} sia distribuita con probabilità gaussiana

$$P(J_{ij}) = \frac{1}{J}\sqrt{\frac{N}{2\pi}} \exp\left\{-\frac{N}{2J^2}\left(J_{ij} - \frac{J_0}{N}\right)^2\right\}. \tag{2.8}$$

Osserviamo che nella (2.8) la media e la varianza sono $\mathcal{O}(1/N)$:

$$[J_{ij}] = \frac{J_0}{N}, \qquad [(\Delta J_{ij})^2] = \frac{J^2}{N}. \tag{2.9}$$

[1] Osserviamo che, in generale, vale

$$\lim_{n\to 0} \frac{x^n - 1}{n} = \lim_{n\to 0} \frac{\mathrm{e}^{n\log x} - 1}{n} = \lim_{n\to 0} \frac{1 + n\log x + \mathcal{O}(n^2) - 1}{n} = \log x.$$

2.2.2 Fuzione di partizione: metodo delle repliche

Considerando l'espressione (1.4) per Z, calcoliamo Z^n tenendo conto dell'equazione (2.7) per l'hamiltoniana:

$$
\begin{aligned}
Z^n &= \left(\sum_{\mathbf{S}} \mathrm{e}^{-\beta H}\right)^n = \left(\sum_{\mathbf{S}} \mathrm{e}^{-\beta H}\right) \cdots \left(\sum_{\mathbf{S}} e^{-\beta H}\right) = \\
&= \sum_{\mathbf{S}_1 \ldots \mathbf{S}_n} \exp\Bigg\{ \underbrace{\beta \sum_{i<j}^{N} J_{ij} S_i S_j + \cdots + \beta \sum_{i<j}^{N} J_{ij} S_i S_j}_{n} + \underbrace{\beta h \sum_{i=1}^{N} S_i + \cdots + \beta h \sum_{i=1}^{N} S_i}_{n} \Bigg\} = \\
&= \mathrm{Tr}\, \exp\left\{ \beta \sum_{i<j}^{N} J_{ij} \sum_{\alpha=1}^{n} S_i^\alpha S_j^\alpha + \beta h \sum_{i=1}^{N} \sum_{\alpha=1}^{n} S_i^\alpha \right\}.
\end{aligned}
$$

Nota Z^n e la distribuzione di probabilità (2.8), calcoliamo $[Z^n]$:

$$
\begin{aligned}
[Z^n] &= \int \left(\prod_{i<j}^{N} dJ_{ij}\, P(J_{ij}) \right) \mathrm{Tr}\, \exp\left\{ \beta \sum_{i<j}^{N} J_{ij} \sum_{\alpha=1}^{n} S_i^\alpha S_j^\alpha + \beta h \sum_{i=1}^{N} \sum_{\alpha=1}^{n} S_i^\alpha \right\} = \\
&= \int \left(\prod_{i<j} dJ_{ij}\, \frac{1}{J} \sqrt{\frac{N}{2\pi}} \exp\left\{ -\frac{N}{2J^2} \left(J_{ij} - \frac{J_0}{N} \right)^2 \right\} \right) \cdot \\
&\quad \cdot \mathrm{Tr}\, \exp\left\{ \beta \sum_{i<j} J_{ij} \sum_{\alpha} S_i^\alpha S_j^\alpha + \beta h \sum_{i} \sum_{\alpha} S_i^\alpha \right\} = \\
&= \mathrm{Tr}\, \frac{1}{J} \sqrt{\frac{N}{2\pi}} \int \left(\prod_{i<j} d\bar{J}_{ij}\, \exp\left\{ -\frac{N}{2J^2} (\bar{J}_{ij})^2 \right\} \right) \cdot \\
&\quad \cdot \exp\left\{ \beta \sum_{i<j} \left(\bar{J}_{ij} + \frac{J_0}{N} \right) \sum_{\alpha} S_i^\alpha S_j^\alpha + \beta h \sum_{i} \sum_{\alpha} S_i^\alpha \right\} = \\
&= \mathrm{Tr}\, \frac{1}{J} \sqrt{\frac{N}{2\pi}} \exp\left\{ \sum_{i<j} \frac{\beta J_0}{N} \sum_{\alpha} S_i^\alpha S_j^\alpha + \beta h \sum_{i} \sum_{\alpha} S_i^\alpha \right\} \\
&\quad \cdot \int \left(\prod_{i<j} d\bar{J}_{ij} \right) \exp\left\{ \sum_{i<j} -\frac{N}{2J^2} (\bar{J}_{ij})^2 + \beta \sum_{i<j} \bar{J}_{ij} \sum_{\alpha} S_i^\alpha S_j^\alpha \right\} = \\
&= \mathrm{Tr}\, \frac{1}{J} \sqrt{\frac{N}{2\pi}} \exp\left\{ \frac{1}{N} \sum_{i<j} \beta J_0 \sum_{\alpha} S_i^\alpha S_j^\alpha + \beta h \sum_{i} \sum_{\alpha} S_i^\alpha \right\} \\
&\quad \cdot \prod_{i<j} \int d\bar{J}_{ij}\, \exp\left\{ -\frac{N}{2J^2} (\bar{J}_{ij})^2 + \bar{J}_{ij} \left[\beta \sum_{\alpha} S_i^\alpha S_j^\alpha \right] \right\}.
\end{aligned}
$$

A questo punto non rimane che calcolare l'integrale dell'espressione di $[Z^n]$ di cui sopra. A tal proposito, ricordiamo il valore del seguente *integrale gaussiano*:

$$
\int_{-\infty}^{+\infty} dx\, \exp\Big\{ -\gamma x^2 + \lambda x \Big\} = \sqrt{\frac{\pi}{\gamma}} \exp\left\{ \frac{\lambda^2}{4\gamma} \right\},
$$

e poniamo:

- $\gamma := N/(2J^2)$
- $\lambda_{ij} := \beta \sum_\alpha S_i^\alpha S_j^\alpha$.

Detto I l'integrale, esso diventa quindi: [2]

$$\begin{aligned} I &= \prod_{i<j} \int d\bar{J}_{ij} \, \exp\left\{ -\frac{N}{2J^2}(\bar{J}_{ij})^2 + \bar{J}_{ij}\left[\beta \sum_\alpha S_i^\alpha S_j^\alpha\right]\right\} = \\ &= \prod_{i<j} \int d\bar{J}_{ij} \, \exp\left\{-\gamma(\bar{J}_{ij})^2 + \lambda_{ij}\bar{J}_{ij}\right\} = \\ &= \prod_{i<j} \sqrt{\frac{2\pi J^2}{N}} \exp\left\{\frac{\lambda_{ij}^2}{4\gamma}\right\} = \\ &= \sqrt{\frac{2\pi J^2}{N}} \exp\left\{\sum_{i<j} \frac{\lambda_{ij}^2}{4\gamma}\right\} = \\ &= \sqrt{\frac{2\pi J^2}{N}} \exp\left\{\frac{1}{N}\sum_{i<j}\frac{1}{2}\beta^2 J^2 \sum_{\alpha,\beta} S_i^\alpha S_j^\alpha S_i^\beta S_j^\beta\right\}. \end{aligned}$$

Sostituendo il valore di I nell'epsressione di $[Z^n]$, otteniamo infine:

$$[Z^n] = \mathrm{Tr}\, \exp\left\{\frac{1}{N}\sum_{i<j}\left(\frac{1}{2}\beta^2 J^2 \sum_{\alpha,\beta} S_i^\alpha S_j^\alpha S_i^\beta S_j^\beta + \beta J_0 \sum_\alpha S_i^\alpha S_j^\alpha\right) + \beta h \sum_i \sum_\alpha S_i^\alpha\right\}.$$

Riscrivendo la sommatoria per $i < j$ e per α, β con N sufficientemente grande, troviamo

$$\begin{aligned} [Z^n] = &\exp\left(\frac{N\beta^2 J^2 n}{4}\right) \mathrm{Tr}\, \exp\left\{\frac{\beta^2 J^2}{2N}\sum_{\alpha<\beta}\left(\sum_i S_i^\alpha S_i^\beta\right)^2 + \right. \\ &\left. + \frac{\beta J_0}{2N}\sum_\alpha\left(\sum_i S_i^\alpha\right)^2 + \beta h \sum_i \sum_\alpha S_i^\alpha\right\}. \end{aligned} \tag{2.10}$$

2.2.3 Riduzione tramite integrale gaussiano

In modo del tutto analogo a come abbiamo fatto nel caso dell'equazione (1.12), dobbiamo riscriverci la forma di $[Z^n]$ (2.10) in modo tale che sia poi possibile calcolarne esplicitamente la traccia.

A tal proposito, riscriviamo due dei termini esponenziali della (2.10) sfruttando l'integrale gaussiano

$$\mathrm{e}^{\gamma x^2/2} = \sqrt{\frac{\gamma N^2}{2\pi}} \int_{-\infty}^{+\infty} dm \; \mathrm{e}^{-N^2\gamma m^2/2 + N\gamma m x}\,. \tag{2.11}$$

In primis, per quanto riguarda il primo esponenziale, poniamo

- $\gamma := (\beta^2 J^2)/N$

[2] Osserviamo che $\lambda_{ij}^2 = [\beta \sum_\alpha S_i^\alpha S_j^\alpha]^2 = \beta^2 \sum_{\alpha,\beta} S_i^\alpha S_j^\alpha S_i^\beta S_j^\beta$

- $x := \sum_i S_i^\alpha S_i^\beta$,

da cui otteniamo:

$$\begin{aligned}
&\exp\left\{\frac{\beta^2 J^2}{2N}\sum_{\alpha<\beta}\left(\sum_i S_i^\alpha S_i^\beta\right)^2\right\} = \prod_{\alpha<\beta}\exp\left\{\frac{\beta^2 J^2}{2N}\left(\sum_i S_i^\alpha S_i^\beta\right)^2\right\} = \\
&= \prod_{\alpha<\beta}\sqrt{\frac{\beta^2 J^2 N}{2\pi}}\int dq_{\alpha\beta}\,\exp\left\{-\frac{N\beta^2 J^2}{2}q_{\alpha\beta}^2 + \beta^2 J^2 q_{\alpha\beta}\sum_i S_i^\alpha S_i^\beta\right\} = \\
&= \sqrt{\frac{\beta^2 J^2 N}{2\pi}}\int \prod_{\alpha<\beta} dq_{\alpha\beta}\,\exp\left\{-\frac{N\beta^2 J^2}{2}\sum_{\alpha<\beta}q_{\alpha\beta}^2 + \beta^2 J^2\sum_{\alpha<\beta}q_{\alpha\beta}\sum_i S_i^\alpha S_i^\beta\right\}.
\end{aligned} \tag{2.12}$$

Per quanto riguarda il secondo, ponendo questa volta

- $\gamma := (\beta J_0)/N$
- $x := \sum_i S_i^\alpha$,

con una procedura totalmente analoga alla precedente otteniamo:

$$\begin{aligned}
&\exp\left\{\frac{\beta J_0}{2N}\sum_{\alpha}\left(\sum_i S_i^\alpha\right)^2\right\} = \prod_{\alpha}\exp\left\{\frac{\beta J_0}{2N}\left(\sum_i S_i^\alpha\right)^2\right\} = \\
&= \prod_{\alpha}\sqrt{\frac{\beta J_0 N}{2\pi}}\int dm_\alpha \exp\left\{-\frac{N\beta J_0}{2}m_\alpha^2 + \beta J_0 m_\alpha\sum_i S_i^\alpha\right\} = \\
&= \sqrt{\frac{\beta J_0 N}{2\pi}}\int \prod_{\alpha} dm_\alpha \exp\left\{-\frac{N\beta J_0}{2}\sum_{\alpha}m_\alpha^2 + \beta J_0\sum_{\alpha} m_\alpha\sum_i S_i^\alpha\right\}.
\end{aligned} \tag{2.13}$$

A questo punto, note le relazioni (2.12) e (2.13), sostituendole nell'equazione (2.10) possiamo riscriverci $[Z^n]$ (a meno di una costante moltiplicativa) linearizzando i termini su cui agisce l'operatore di traccia:

$$\begin{aligned}
[Z^n] = &\exp\left(\frac{N\beta^2 J^2 n}{4}\right)\int\prod_{\alpha<\beta}dq_{\alpha\beta}\int\prod_\alpha dm_\alpha \\
&\cdot\exp\left\{-\frac{N\beta^2 J^2}{2}\sum_{\alpha<\beta}q_{\alpha\beta}^2 - \frac{N\beta J_0}{2}\sum_\alpha m_\alpha^2\right\} \\
&\cdot\operatorname{Tr}\exp\left\{\beta^2 J^2\sum_{\alpha<\beta}q_{\alpha\beta}\sum_i S_i^\alpha S_i^\beta + \beta\sum_\alpha(J_0 m_\alpha + h)\sum_i S_i^\alpha\right\}.
\end{aligned} \tag{2.14}$$

Se nella terza riga dell'equazione (2.14) utilizziamo il simbolo Tr per indicare la somma su un unico sito $\sum_{S_i^\alpha}$, possiamo scrivere quest'ultima come

$$\left\{\operatorname{Tr}\exp\left(\beta^2 J^2\sum_{\alpha<\beta}q_{\alpha\beta}S^\alpha S^\beta + \beta\sum_\alpha(J_0 m_\alpha + h)S^\alpha\right)\right\}^N \equiv \mathrm{e}^{N\log\operatorname{Tr}\mathrm{e}^L},$$

dove chiaramente

$$L = \beta^2 J^2 \sum_{\alpha<\beta} q_{\alpha\beta} S^\alpha S^\beta + \beta \sum_\alpha (J_0 m_\alpha + h) S^\alpha. \tag{2.15}$$

Abbiamo quindi l'espressione finale

$$\begin{aligned}[Z^n] = &\exp\left(\frac{N\beta^2 J^2 n}{4}\right) \int \prod_{\alpha<\beta} dq_{\alpha\beta} \int \prod_\alpha dm_\alpha \\ &\cdot \exp\left\{ -\frac{N\beta^2 J^2}{2} \sum_{\alpha<\beta} q_{\alpha\beta}^2 - \frac{N\beta J_0}{2} \sum_\alpha m_\alpha^2 + N \log \mathrm{Tr}\, \mathrm{e}^L \right\}.\end{aligned} \tag{2.16}$$

2.2.4 Metodo punto-sella

Ricordiamo che il nostro obiettivo, per il momento, è quello di calcolare esplicitamente il limite (2.6). I passaggi fin qui svolti ci hanno portato fino all'equazione (2.16), che sappiamo avere l'esponente con andamento $\mathcal{O}(N)$: è pertanto è possibile valutare l'integrale col già citato metodo di punto-sella (cfr. Appendice A). Nel limite termodinamico di $N \to \infty$ otteniamo

$$\begin{aligned}[Z^n] &\approx \exp\left\{ -\frac{N\beta^2 J^2}{2} \sum_{\alpha<\beta} q_{\alpha\beta}^2 - \frac{N\beta J_0}{2} \sum_\alpha m_\alpha^2 + N \log \mathrm{Tr}\, \mathrm{e}^L + \frac{N}{4}\beta^2 J^2 n \right\} \\ &\approx 1 + Nn \left\{ -\frac{\beta^2 J^2}{4n} \sum_{\alpha\neq\beta} q_{\alpha\beta}^2 - \frac{\beta J_0}{2n} \sum_\alpha m_\alpha^2 + \frac{1}{n} \log \mathrm{Tr}\, \mathrm{e}^L + \frac{1}{4}\beta^2 J^2 \right\}.\end{aligned}$$

Nel derivare la seconda riga, abbiamo calcolato il limite $n \to 0$ nell'ipotesi di $N \gg 1$, ma finito, sfruttando l'equazione (A.4) (e trascurando la costante moltiplicativa, irrilevante per i passaggi successivi).

A questo punto, tramite l'equazione (2.16) scriviamo l'energia libera $[f]$ come

$$-\beta[f] = \lim_{n\to 0} \frac{[Z^n] - 1}{nN} = \lim_{n\to 0} \left\{ -\frac{\beta^2 J^2}{4n} \sum_{\alpha\neq\beta} q_{\alpha\beta}^2 - \frac{\beta J_0}{2n} \sum_\alpha m_\alpha^2 + \frac{1}{n} \log \mathrm{Tr}\, \mathrm{e}^L + \frac{1}{4}\beta^2 J^2 \right\}. \tag{2.17}$$

In virtù del metodo punto-sella, nell'espressione precedente $q_{\alpha\beta}$ e m_α devono essere scelti in modo tale che siano estremanti della funzione tra le parentesi $\{\cdots\}$, e di conseguenza della funzione $[f]$. Pertanto, le condizioni cercate sono

$$\begin{aligned}\frac{\partial [f]}{\partial q_{\alpha\beta}} &= 0 \\ \frac{\partial [f]}{\partial m_\alpha} &= 0\end{aligned}$$

che restituiscono rispettivamente le equazioni

$$q_{\alpha\beta} = \frac{1}{\beta^2 J^2} \frac{\partial}{\partial q_{\alpha\beta}} \log \mathrm{Tr}\, \mathrm{e}^L = \frac{\mathrm{Tr}\ S_\alpha S_\beta\, \mathrm{e}^L}{Tr\, e^L} := \langle S^\alpha S^\beta \rangle_L \tag{2.18}$$

$$m_\alpha = \frac{1}{\beta J_0} \frac{\partial}{\partial m_\alpha} \log \mathrm{Tr}\, \mathrm{e}^L = \frac{Tr\, S_\alpha\, \mathrm{e}^L}{\mathrm{Tr}\, \mathrm{e}^L} := \langle S^\alpha \rangle_L. \tag{2.19}$$

2.2.5 Parametri d'ordine

Le variabili $q_{\alpha\beta}$ e m_α sono state introdotte per convenienze computazionali negli integrali gaussiani, in modo del tutto analogo al modello ferromagnetico della sezione 1.4 (era già stata utilizzata questa tecnica per ricavare l'equazione (1.13)).

Notiamo innanzitutto che l'equazione (2.18) può essere riscritta come

$$q_{\alpha\beta} = [\langle S_\alpha S_\beta \rangle] = \left[\frac{\mathrm{Tr}\ S_i^\alpha S_j^\beta \exp(-\beta \sum_\gamma H_\gamma)}{\mathrm{Tr}\ \exp(\sum_\gamma H_\gamma)} \right], \tag{2.20}$$

dove H_γ indica la γ-esima replica hamiltoniana

$$H_\gamma = -\sum_{i<j} J_{ij} S_i^\gamma S_j^\gamma - h \sum_i S_i^\gamma. \tag{2.21}$$

Per verificare ciò, iniziamo con l'osservare che il denominatore della (2.20) è proprio Z^n, che sappiamo convergere a 1 per $n \to 0$. Rimane da calcolare solo il numeratore. Il procedimento è analogo a quello sviluppato nella sottosezione 2.2.2 e porta alla grandezza

$$\left(\mathrm{Tr}\, \mathrm{e}^L\right)^{N-1} \cdot \left(\mathrm{Tr}\ S^\alpha S^\beta\, \mathrm{e}^L\right).$$

Dall'equazione (2.17) ricaviamo che $\log \mathrm{Tr}\, \mathrm{e}^L$ è $\mathcal{O}(n)$, da cu segue che $\mathrm{Tr}\, \mathrm{e}^L \to 1$ per $n \to 0$: l'espressione di cui sopra si riduce a $\mathrm{Tr}\ S^\alpha S^\beta\, \mathrm{e}^L$ nel limite $n \to 0$. Ora, dal momento che anche il denominatore della (2.18) converge a 1, otteniamo proprio che la (2.18) e la (2.20) coincidono per $n \to 0$. In modo del tutto analogo possiamo concludere che

$$m_\alpha = [\langle S_i^\alpha \rangle]. \tag{2.22}$$

Il parametro m è il già noto *parametro d'ordine ferromagnetico*, come testimonia la (2.22), mentre $q_{\alpha\beta}$ costituisce il *parametro d'ordine di vetro di spin*. Quest'ultimo può essere compreso più facilmente osservando che [3]

$$q_{\alpha\beta} = \left[\frac{\mathrm{Tr}\ S_i^\alpha\, \mathrm{e}^{-\beta H_\alpha}}{\mathrm{Tr}\, \mathrm{e}^{-\beta H_\alpha}} \frac{Tr\, S_i^\beta\, \mathrm{e}^{-\beta H_\beta}}{\mathrm{Tr}\, \mathrm{e}^{-\beta H_\beta}} \right] = [\langle S_i^\alpha \rangle \langle S_i^\beta \rangle] = [\langle S_i \rangle^2] \equiv q \tag{2.23}$$

se non distinguiamo una replica da un'altra.

Nella fase paramagnetica ad alte temperature, $\langle S_i \rangle$ si annulla per ogni sito i e dunque $m = q = 0$. La fase ferromagnetica ha una distribuzione quasi totalmente uniforme, e se scegliamo quest'ultima come direzione positiva, abbiamo che $\langle S_i \rangle > 0$ per quasi ogni sito, da cui $m > 0$ e $q > 0$.

Se ammettiamo l'esistenza dalla fase di vetri di spin (secondo il modello di Edward-Anderson o il modello SK), questa deve certamente essere caratterizzata da spin congelati. Ciò significa che in generale $\langle S_i \rangle$ non si annulla per ogni sito, proprio per il fatto che non ci sono significative fluttuazioni nel tempo. Tuttavia, il segno di $\langle S_i \rangle$ cambia da sito a sito, e questa apparente configurazione casuale è per l'appunto fissata nel tempo. Quindi, la media configurazionale di $\langle S_i \rangle$ sulla distribuzione di $\mathbf{J}$ tiene conto sia di contributi $\langle S_i \rangle > 0$ sia $\langle S_i \rangle < 0$, che suggerisce la possibilità di avere $m = [\langle S_i \rangle] = 0$. Al contrario, il parametro q sicuramente non si annulla, in quanto media di una grandezza positiva $\langle S_i \rangle^2$.

In conclusione, l'idea di base è la seguente: potrebbe esistere una fase, ossia la fase di vetri di spin, caratterizzata dai parametri d'ordine $m = 0$ e $q > 0$.

[3] Le tracce sulle repliche diverse da α e β si semplificano tra numeratore e denominatore.

2.3 Soluzione tramite replica-simmetry

2.3.1 Equazioni di stato

Come già detto, l'obiettivo finale di questo capitolo è arrivare a scrivere le equazioni di stato del sistema di vetro di spin. Ricordiamo che per ottenere l'energia libera (2.17) abbiamo sfruttato il metodo delle repliche. Tuttavia, l'introduzione di queste repliche non modifica la fisica del sistema: è semplicemente un artefizio matematico per convenienze di calcolo. Quindi è naturale assumere la *simmetria di replica* (RS), tale per cui $q_{\alpha\beta} = q$ e $m_\alpha = m$; ciò consiste sostanzialmente nel considereare ogni replica uguale al sistema di partenza.

Riscriviamo l'energia libera (2.17) tenendo conto della simmetria RS:

$$-\beta[f] = \lim_{n\to 0} \frac{\beta^2 J^2}{4n}\{-n(n-1)q^2\} - \frac{\beta J_0}{2n} n m^2 + \frac{1}{4}\beta^2 J^2 + \frac{1}{n}\log \mathrm{Tr}\, \mathrm{e}^L \,. \tag{2.24}$$

Prima di poter calcolare questo limite, abbiamo bisogno di scrivere meglio il quarto addendo del membro di destra. Ricordando l'epsressione di L (2.15), lo riscriviamo come

$$\begin{aligned}\log \mathrm{Tr}\, \mathrm{e}^L &= \log \mathrm{Tr}\ exp\left\{\beta^2 J^2 \sum_{\alpha<\beta} q_{\alpha\beta} S^\alpha S^\beta + \beta \sum_\alpha (J_0 m_\alpha + h) S^\alpha \right\} = \\ &= \log \mathrm{Tr}\ exp\left\{ \frac{\beta^2 J^2 q}{2}\left(\sum_\alpha S^\alpha\right)^2 - \frac{1}{2} n \beta^2 J^2 q + \beta(J_0 m + h)\sum_\alpha S_\alpha \right\}.\end{aligned}$$

Sfruttiamo ancora una volta l'integrale gaussiano (2.11) e, ponendo

- $\gamma := \beta^2 J^2 q$
- $x := \sum_\alpha S^\alpha$,

ricaviamo che

$$\exp\left\{ \frac{\beta^2 J^2 q}{2}\left(\sum_\alpha S^\alpha\right)^2 \right\} = \sqrt{\frac{\beta^2 J^2 q}{2\pi}} \int dz\ \exp\left\{ \frac{\beta^2 J^2 q}{2} z^2 + \beta^2 J^2 q \left(\sum_\alpha S^\alpha\right) z \right\},$$

da cui:

$$\begin{aligned}\log \mathrm{Tr}\ e^L =& \log \mathrm{Tr}\ \sqrt{\frac{\beta^2 J^2 q}{2\pi}} \int dz\ \exp\left\{ -\frac{\beta^2 J^2 q}{2} z^2 + \beta^2 J^2 q \left(\sum_\alpha S^\alpha\right) z + \right. \\ &\left. - \frac{1}{2} n \beta^2 J^2 q + \beta(J_0 m + h)\sum_\alpha S_\alpha \right\} = \\ =& \log \mathrm{Tr}\ \sqrt{\frac{\beta^2 J^2 q}{2\pi}} \int Dz \prod_\alpha \exp\left\{ \beta J \sqrt{q} S^\alpha z + \beta(J_0 m + h) S^\alpha - \frac{1}{2} n \beta^2 J^2 q \right\},\end{aligned}$$

dove abbiamo posto $Dz := dz/\sqrt{2\pi}\, \mathrm{e}^{-z^2/2}$. Chiamando $\tilde{H}(z) := J\sqrt{q} z + J_0 m + h$ per avere una scrittura più sintetica, osserviamo che

$$\mathrm{Tr} \prod_\alpha \mathrm{e}^{\beta \tilde{H}(z) S^\alpha} = \sum_{S^1 \dots S^n} \mathrm{e}^{\beta \tilde{H}(z) S^1} \cdots \mathrm{e}^{\beta \tilde{H}(z) S^n} = \{2\cosh \beta \tilde{H}(z)\}^n = \mathrm{e}^{n \log[2\cosh \beta \tilde{H}(z)]}\,.$$

Sostituiamo questa relazione nell'equazione precedente e sviluppiamo al primo ordine di n (sempre in un intorno dello 0) la funzione integranda:

$$\begin{aligned}\log \mathrm{Tr}\, \mathrm{e}^{L} &= \log \int Dz \exp\left\{ n \log[2\cosh \beta\tilde{H}(z)] - \frac{n}{2}\beta^2 J^2 q \right\} = \\ &= \log \int Dz \left(1 + n \log[2\cosh \beta\tilde{H}(z)] - \frac{n}{2}\beta^2 J^2 q + \mathcal{O}(n^2) \right) = \\ &= \log \left(1 + n \int Dz\, \log[2\cosh \beta\tilde{H}(z)] - \frac{n}{2}\beta^2 J^2 q + \mathcal{O}(n^2) \right).\end{aligned}$$

Finalmente siamo pronti per calcolare il limite (2.24):

$$\begin{aligned}-\beta[f] = \lim_{n\to 0} &-\frac{\beta^2 J^2}{4}(n-1)q^2 - \frac{\beta J_0}{2}m^2 + \frac{1}{4}\beta^2 J^2 + \\ &+ \frac{1}{n}\log\left(1 + n \int Dz\, \log[2\cosh \beta\tilde{H}(z)] - \frac{n}{2}\beta^2 J^2 q + \mathcal{O}(n^2) \right) = \\ &= \frac{\beta^2 J^2}{4}(1-q)^2 - \frac{\beta J_0}{2}m^2 + \int Dz\, \log[2\cosh \beta\tilde{H}(z)].\end{aligned} \tag{2.25}$$

Seguendo sempre lo stesso filo logico che abbiamo già mostrato nella sottosezione 2.2.4, possiamo scrivere l'equazione del moto per i parametri d'ordine m e q.

Per quanto riguarda m:

$$\frac{\partial [f]}{\partial m} = 0 \quad \rightarrow \quad m = \int Dz\, \tanh[\beta\tilde{H}(z)]. \tag{2.26}$$

Confrontando questa equazione con l'equazione di stato per un singolo spin (1.14) in cui viene posto $J = 0$, possiamo supporre che il campo interno segua una distribuzione gaussiana.

La condizione di estremizzazione rispetto a q è invece:

$$\frac{\partial [f]}{\partial q} = 0 \quad \rightarrow \quad \frac{\beta^2 J^2}{2}(q-1) + \int Dz\, \tanh[\beta\tilde{H}(z)] \cdot \frac{\beta J}{2\sqrt{q}} z = 0. \tag{2.27}$$

Integrando per parti, si arriva all'equazione finale

$$q = 1 - \int \frac{Dz}{\cosh^2[\beta\tilde{H}(z)]} = \int Dz\, \tanh^2[\beta\tilde{H}(z)]. \tag{2.28}$$

2.3.2 Diagramma di fase

Il comportamento delle equazioni (2.26) e (2.28) dipende sostanzialmente dai parametri β e J_0, restringendoci per semplicità al caso in cui non ci sia campo esterno ($h = 0$). Se la distribuzione di J_{ij} è simmetrica ($J_0 = 0$), allora abbiamo che $\tilde{H}(z) = J\sqrt{q}z$ è una funzione dispari, da cui segue che la magnetizzazione è nulla ($m = 0$): non c'è fase ferromagnetica.

Sotto queste condizioni, studiamo le proprietà del sistema vicino al punto critico in cui il parametro q è piccolo. Per far ciò, calcoliamo l'espansione in serie dell'energia libera (2.25):

$$\beta[f] = -\frac{1}{4}\beta^2 J^2 - \log 2 - \frac{\beta^2 J^2}{4}(1 - \beta^2 J^2)q^2 + \mathcal{O}(q^3). \tag{2.29}$$

La teoria di Landau, come abbiamo già visto nella sottosezione 1.3.3, ci dice che il punto critico è determinato dalla condizione di annullamento del coefficiente del termine q^2. Nel nostro caso, chiaramente abbiamo $T = J \equiv T_f$.

Nell'equazione (2.29) possiamo notare un comportamento anomalo. Infatti, abbiamo che il coefficiente di q^2 è negativo per $T > T_f$ e positivo per $T < T_f$. Ciò significa che la soluzione paramagnetica ($q = 0$) massimizza l'energia libera alle alte temperature, mentre la soluzione di vetro di spin ($q > 0$) la massimizza alle basse temperature. Questo comportamento apparentemente patologico è giustificato dal metodo delle repliche. Se osserviamo il limite (2.25) ci accorgiamo che il coefficiente di q^2 è negativo $\forall n > 1$, ma cambia di segno mandando n a 0: proprio questo causa la massimizzazione invece che la minimizzazione. Questo comportamento anomalo non si riscontra invece per il parametro m.

Una soluzione ferromagnetica ($m > 0$) potrebbe esistere se la distribuzione di J_{ij} non è simmetrica nell'origine ($J_0 > 0$). Espandendo l'equazione (2.28) e tenendo solo i termini più bassi di q e m, abbiamo che

$$q = \beta^2 J^2 q + \beta^2 J_0^2 m^2. \tag{2.30}$$

Assumendo $J_0 = 0$ e seguendo un ragionamento analogo a quello mostrato nella sottosezione 1.3.2, ritroviamo che la condizione di temperatura critica si ha per $\beta^2 J^2 = 1$, in accordo con quando esposto studiando l'energia libera: $T_f = J$.

Se $J_0 > 0$ e $m > 0$, questo implica nella (2.30) che $q = \mathcal{O}(m^2)$. Sviluppando al primo ordine l'equazione (2.26), otteniamo

$$m = \beta J_0 m + \mathcal{O}(q). \tag{2.31}$$

É ormai chiaro che il punto critico ferromagnetico, in cui m comincia ad assumere valori non nulli, è dato da $\beta J_0 = 1$, ossia $T_c = J_0$.

Abbiamo così mostrato le condizioni al contorno tra fase paramagnetica e di vetro di spin e tra fase paramagnetica e ferromagnetica. Le condizioni al contorno tra fase ferromagnetica e di vetro di spin si possono ottenere solamente risolvendo numericamente le equazioni (2.26) e (2.28).

2.3.3 Entropia negativa

Un'ulteriore prova del fallimento della simmetria RS alle basse temperature è data dal fatto che, assumendo $J_0 = 0$, il ground state ha un'entropia negativa. Mostriamo per l'appunto il verificarsi di questa anomalia; assumiamo dunque che $T \to 0$ (cioè $\beta \to \infty$) e riscriviamo l'equazione (2.28) sotto quest'ipotesi:

$$\begin{aligned} q &= 1 - \int \frac{Dz}{\cosh^2[\beta \tilde{H}(z)]} = 1 - \frac{1}{\beta J} \int Dz \frac{d}{dz} \tanh[\beta J z] \\ &\longrightarrow 1 - \int Dz 2\delta(z) = 1 - \sqrt{\frac{2}{\pi}} \frac{T}{J}. \end{aligned}$$

É quindi ragionevole porre $q = 1 - aT$ con $a = \sqrt{2/\pi}/J$.

Per ricavare l'entropia del ground state è necessario calcolare esplicitamente l'energia libera grazie all'equazione (2.25). Il primo termine porta un contributo

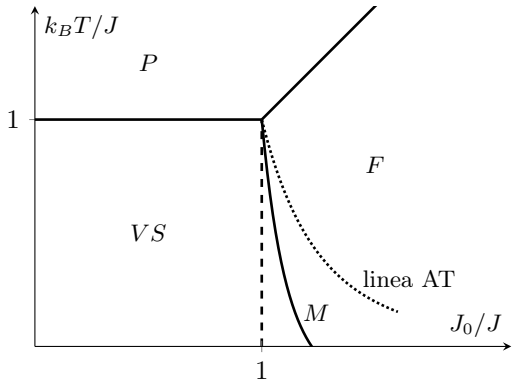


Figura 2.1: Diagramma di fase del modello SK. La linea tratteggiata è il confine tra la fase *ferromagnetica* (F) e di *vetro di spin* (VS) ed esiste solo dato l'ansatz della simmetria RS (la soluzione con simmetria RS è instabile al di sotto della linea AT: tra la fase F e quella VS ne emerge una mista). Il sistema è invece in una fase *paramagnetica* (P) nella regione ad alta temperatura.

pari a $1/2\pi$, mentre più delicato è il calcolo del secondo termine:

$$\begin{aligned}
&\int_{-\infty}^{+\infty} Dz \log[2\cosh\beta\tilde{H}(z)] = \ 2\int_{0}^{+\infty} Dz \log[2\cosh\beta\tilde{H}(z)] = \\
&= 2\int_{0}^{+\infty} Dz \log[\mathrm{e}^{\beta J\sqrt{q}z}(1+\mathrm{e}^{-2\beta J\sqrt{q}z})] = \\
&= 2\beta J\sqrt{q}\int_{0}^{+\infty} Dz\ z + 2\int_{0}^{+\infty} Dz \log(1+\mathrm{e}^{-2\beta J\sqrt{q}z}) \approx \\
&\approx \frac{2\beta J(1-aT/2)}{\sqrt{2\pi}} + 2\int_{0}^{+\infty} Dz\,\mathrm{e}^{-2\beta J\sqrt{q}z},
\end{aligned} \tag{2.32}$$

deve, per ottenere l'ultima uguaglianza, sottolineamo che

- $\sqrt{q} \approx 1 - aT/2$
- $\log(1+\mathrm{e}^{-2\beta J\sqrt{q}z}) \approx \mathrm{e}^{-2\beta J\sqrt{q}z}$.

Si dimostra con qualche passaggio extra che il secondo termine della (2.32) porta un contributo $\mathcal{O}(T^2)$ (sarà quindi trascurato da qui in avanti), mentre il primo è uguale a $-\sqrt{2/\pi}J + T/2\pi$. Dividendo l'equazione (2.25) per $-\beta$ si ottiene:

$$[f] \approx \sqrt{\frac{2}{\pi}}J + \frac{T}{2\pi}.. \tag{2.33}$$

Si ricordi che l'energia libera è definita come $F := U - TS$ (dove U indica l'energia interna): da un rapido confronto con la (2.33) si ricava $S(T=0) = -1/2\pi$. Il risultato di ottenere entropia negativa è probabilmete legato al fatto che nell'equazione (2.17) abbiamo scambiato i limiti $n \to 0$ e $N \to \infty$ (infatti, l'ordine corretto sarebbe stato $N \to \infty$ dopo $n \to 0$). Ciò nonostante, l'assunzione della simmetria RS, ossia $q_{\alpha\beta} = q$ è la maggiore responsabile di queste anomalie. Il modo in cui sia necessario rompere la simmetria di replica esula dallo scopo di questo elaborato. Sottolineamo soltanto che la soluzione al problema di equilibrio del modello di Ising venne formalizzata nel 1979 dal fisico teorico italiano Giorgio Parisi [18].

Capitolo 3

Error correcting codes

La teoria dell'informazione tramite canali rumorosi, detta *error correcting codes* (ECC), gioca un ruolo chiave nella società moderna. Alcuni dei suoi aspetti possono essere compresi e formalizzati proprio attraverso il modello di vetri di spin. Infatti, come vedremo il rumore nel canale di trasmissione può essere posto in analogia con le interazioni casuali nei vetri di spin e la sequenza di bit con la corrispondente configurazione del modello di Ising. Gli studi sulla teoria dell'informazione furono iniziati dall'ingegnere e matematico statunitense *Claude Shannon* (1916-2001), considerato a buon diritto il padre della teoria dell'informazione [19].

3.1 Error correcting codes

3.1.1 Trasmissione dell'informazione

Supponiamo di voler trasmettere un messaggio di N bit da un luogo a un altro. Per fare questo abbiamo bisogno di un cosiddetto *canale di trasmissione*. In genere un canale è soggetto a un rumore, cosa che determina alcune (o molte) differenze tra il messaggio di ingresso e quello di uscita. Il nostro obiettivo consiste proprio nel capire come recuperare il messaggio originale dall'output rumoroso.

Definiamo *encoding* (o *canale di codificazione*) il processo di conversione di un'informazione in simboli per la comunicazione o la conservazione. Spesso questo richiede di rendere ridondante il messaggio originale aggiungendo pezzi extra di informazioni. Il messaggio codificato è poi trasmesso attraverso un canale rumoroso (essendo il linguaggio macchina di tipo binario, fatto cioè di 0 e 1, il rumore del canale può mutare uno 0 in 1 o viceversa). Definiamo *decoding* il processo di recupero dell'informazione orginale dal canale rumoroso.

Un esempio di questo processo è costituito dall'aggiunta di *parity-bit*. Un parity-bit è un bit aggiunto alla stringa di un codice binario affinchè controlli che il numero di 1 all'interno di una specifica sequenza di bit sia *pari* o *dispari*. Nel caso in cui tale numero sia pari, viene aggiunto un parity-bit uguale a 0. In caso contrario uguale a 1 (in pratica viene effettuata una somma dei bit nella sequenza in modulo 2). Nel momento in cui il rumore del canale non è eccessivamente grande, con buona approssimazione si può ritenere che il messaggio di output sia privo di errori se la parità dei parity-bit coincide con la parità della stringa ricevuta. Se ciò non dovesse accadere, si riscontra un errore: la fase di correzione degli errori (*error correction*) sarà oggetto di questo capitolo.

3.1.2 Repetition codes

Introduciamo un primo esempio semplice di famiglia di algoritmi di encoding e decoding: i *Repetition codes*. Consideriamo un *binary simmetric channel* (BSC), ossia un canale in cui il rumore muta uno 0 in 1 o un 1 in 0 indipendentemente per ogni bit e secondo una data probabilità. Immaginiamo che il messaggio da inviare $\boldsymbol{\xi}$ sia una semplice stringa di 7 bit, come ad esempio

$$\boldsymbol{\xi} = 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \tag{3.1}$$

e che il rumore del canale BSC sia $p = 0.1$ (ciò significa che il rumore cambia un bit ogni 10). Come già accennato, per poter ridurre l'errore di trasmissione è necessario rendere ridondante il messaggio di input. Nel caso del repetition code R_M, l'algoritmo di encoding prevede di inviare ogni bit copiato M volte. Nel nostro caso prendiamo come esempio $M = 3$ e chiamiamo $\mathbf{J}^0$ il vettore di encoding, che sarà quindi

$$\mathbf{J}^0 = 000 \quad 000 \quad 111 \quad 000 \quad 111 \quad 111 \quad 000. \tag{3.2}$$

Possiamo pensare il rumore del canale come un vettore $\boldsymbol{\Delta}$ che viene aggiunto in modulo 2 al vettore originale ($1 + 1 = 0$ in modulo 2). In questo caso lo scegliamo uguale a

$$\boldsymbol{\Delta} = 000 \quad 001 \quad 000 \quad 000 \quad 101 \quad 000 \quad 000. \tag{3.3}$$

Se chiamiamo $\mathbf{J}$ il vettore di output, ossia $\mathbf{J} = \mathbf{J}^0 + \boldsymbol{\Delta}$, allora questo sarà uguale a

$$\begin{array}{lccccccccc} \boldsymbol{\xi} & 0 & 0 & 1 & 0 & 1 & 1 & 0 & \\ \mathbf{J}^0 & \overbrace{000} & \overbrace{000} & \overbrace{111} & \overbrace{000} & \overbrace{111} & \overbrace{111} & \overbrace{000} & + \\ \boldsymbol{\Delta} & 000 & 001 & 000 & 000 & 101 & 000 & 000 & = \\ \hline \mathbf{J} & 000 & 001 & 111 & 000 & 010 & 111 & 000 & \end{array} \tag{3.4}$$

A questo punto, $\mathbf{J}$ deve subire un processo di decoding: il vettore decodificato, che chiamiamo $\hat{\boldsymbol{\xi}}$, dovrà nuovamente avere 7 bit. La decisione ottimale di decoding (ossia quella che restituisce la minor probabilità di errore) è di trovare per ogni bit quele valore di $\boldsymbol{\xi}$ è il più probabilie, una volta noto $\mathbf{J}$. Consideriamo un singolo bit ξ_i, cui corrispondono 3 bit (tramite l'algoritmo di encoding e decoding) $\mathbf{J}_i = J_{i_1} J_{i_2} J_{i_3}$. Il *teorema di Bayes* (per una trattazione formale si rimanda alla sezione 4.1.1) garantisce che la probabilità di ottenere ξ_i una volta noto $\mathbf{J}_i$ corrisponde a[1]

$$P(\xi_i | J_{i_1} J_{i_2} J_{i_3}) = \frac{P(J_{i_1} J_{i_2} J_{i_3} | \xi_i) P(\xi_i)}{P(J_{i_1} J_{i_2} J_{i_3})}, \tag{3.5}$$

dove si definiscono:

- *prior* la probabilità $P(\xi_i)$
- *posterior* la probabilità $P(\xi_i | J_{i_1} J_{i_2} J_{i_3})$.

Quindi, noti i due posterior $P(\xi_i = 0|\mathbf{J}_i)$ e $P(\xi_i = 1|\mathbf{J}_i)$, l'algoritmo di decoding assegnerà il valore $\hat{\xi}_i = 0$ se $P(\xi_i = 0|\mathbf{J}_i) > P(\xi_i = 1|\mathbf{J}_i)$, oppure $\hat{\xi}_i = 1$ altrimenti.

[1] Ricordiamo che, dati due eventi A e B, si definisce *probabiltà condizionata* di A rispetto a B la probailità che si verifichi l'evento A, sapendo che l'evento B è verificato. In genere si indica con $P(A|B)$.

Per agevolare i conti, si assume che le probabilità di ogni prior saino uguali, ossia $P(\xi_i = 0) = P(\xi_i = 1)$. Questo implica che massimizzare $P(\xi_i|\mathbf{J}_i)$ corrisponde a massimizzare $P(\mathbf{J}_i|\xi_i)$. Ora, poichè stiamo utilizzando un canale BSC, la probabilità $P(\mathbf{J}_i|\xi_i)$ è indipendente per ogni bit, da cui

$$P(\mathbf{J}_i|\xi_i) = P(\mathbf{J}_i|\mathbf{J}_i^0(\xi_i)) = \prod_{n=1}^{M} P(J_{i_n}|J_{i_n}^0(\xi_i)), \tag{3.6}$$

dove nel nostro caso $M = 3$. Ricordando che p descrive la probabilità che ogni bit venga mutato dal rumore del canale, possiamo scrivere $P(J_{i_n}|J_{i_n}^0(\xi_i))$ come

$$P(J_{i_n}|J_{i_n}^0(\xi_i)) = \begin{cases} 1-p & \text{se } J_{i_n} = J_{i_n}^0(\xi_i) \\ p & \text{se } J_{i_n} \neq J_{i_n}^0(\xi_i) \end{cases}. \tag{3.7}$$

Il rapporto tra le due probabilità è dunque

$$\frac{P(\mathbf{J}_i|\xi_i = 1)}{P(\mathbf{J}_i|\xi_i = 0)} = \prod_{n=1}^{M} \frac{P(J_{i_n}|J_{i_n}^0(1))}{P(J_{i_n}|J_{i_n}^0(0))}, \tag{3.8}$$

dove ogni fattore $\frac{P(J_{i_n}|J_{i_n}^0(1))}{P(J_{i_n}|J_{i_n}^0(0))}$ è uguale a

- $\frac{1-p}{p}$ se $J_{i_n} = 1$
- $\frac{p}{1-p}$ se $J_{i_n} = 0$.

Introducendo infine il paramentro $\gamma \equiv \frac{1-p}{p}$ (è maggiore di 1 poichè $p = 0.1$ per ipotesi), il rapporto (3.8) sarà uguale a γ^s, con $s \in \mathbb{Z}$, tramite cui si ricava che

- $\hat{\xi}_i = 0$ se $s < 0$
- $\hat{\xi}_i = 1$ se $s > 0$.

Alla luce di quanto detto, nel caso specifico dell'esempio qui proposto, l'algoritmo di decoding è il seguente:

Sequenza ricevuta $\mathbf{J}_i$	$\frac{P(J_{i_n}\|J_{i_n}^0(1))}{P(J_{i_n}\|J_{i_n}^0(0))}$	bit decodificato $\hat{\xi}_i$
000	γ^{-3}	0
001	γ^{-1}	0
010	γ^{-1}	0
100	γ^{-1}	0
101	γ^{1}	1
110	γ^{1}	1
011	γ^{1}	1
111	γ^{3}	1

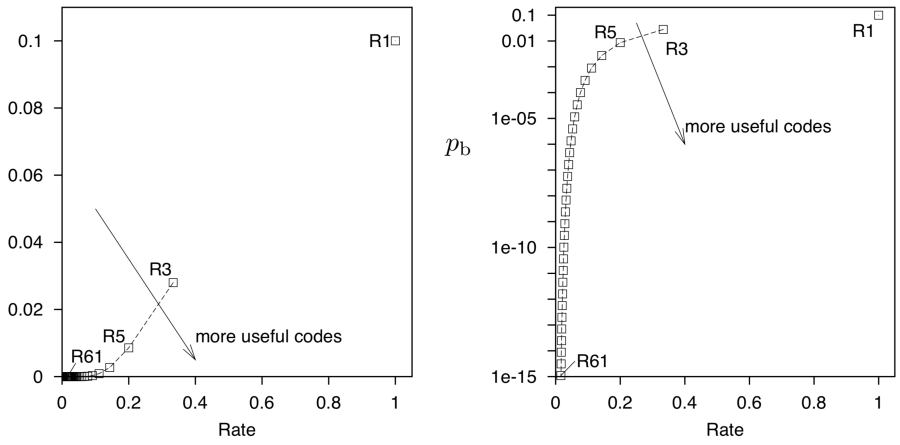


Figura 3.1: Probabilità di errore p_b in funzione del rate R per i Repetition codes (nello specifico, sono presentati da R_1 a R_{61}) in un canale BSC con $p = 0.1$. La freccia mostra l'ipotetica posizione di un codice con il miglior rendimento in termini di errore commesso e rate di trasmissione (la figura di destra è in scala logaritmica) [12].

da cui deduciamo il vettore finale

$$\begin{array}{lccccccccc} \boldsymbol{\xi} & 0 & 0 & 1 & 0 & 1 & 1 & 0 & \\ \mathbf{J}^0 & \overbrace{000} & \overbrace{000} & \overbrace{111} & \overbrace{000} & \overbrace{111} & \overbrace{111} & \overbrace{000} & + \\ \boldsymbol{\Delta} & 000 & 001 & 000 & 000 & 101 & 000 & 000 & = \\ \hline \mathbf{J} & \underbrace{000} & \underbrace{000} & \underbrace{111} & \underbrace{000} & \underbrace{010} & \underbrace{111} & \underbrace{000} & \\ \hat{\boldsymbol{\xi}} & 0 & 0 & 1 & 0 & 0 & 1 & 0 & \\ & & \star & & & \times & & & \end{array} \tag{3.9}$$

Analizzando il risultato finale, l'algoritmo ha effettivamente corretto l'errore nella seconda tripletta ($\star$), mentre non è riuscito a correggere quello nella quinta ($\times$). Infatti, nel caso del repetion code R_3, non è difficile convincersi che l'algoritmo non è in grado di risolvere il problema di due errori in una singola tripletta. É comunque importante sottolinare che se p è la probabilità di errore sul singolo bit introdotto dal canale BSC, l'algoritmo R_3 riscala l'errore come $p_b = \mathcal{O}(p^2)$: in questo caso, con $p = 0.1$ si ottiene $p_b \approx 0.03$. Chiaramente, è possibile ridurre l'errore aumentando il numero di copie M, ma come contro si appesantisce di molto il messaggio da inviare. Se infatti definiamo il *rate di trasmissione* R di un canale come

$$R := \frac{N}{N_B} \tag{3.10}$$

dove N è il numero di bit del messaggio originale e N_B il numero di bit del messaggio di encoding (in questo caso, $N_B = M \cdot N$), ci si accorge che la velocità di trasmissione $R = 1/M$ del messaggio tenda a 0 al crescere di M.

3.1.3 Hamming code

Un altra importante famiglia di algoritmi di encoding e decoding per un canale BSC è costituita dagli *Hamming codes*. In questo caso verrà presentato l'Hamming code (7,4), che trasmette $N_B = 7$ bit per ogni stringa di $N = 4$ bit. Gli $N_B - N$ bit extra sono funzioni lineari degli N bit originali: questi sono denominati *parity-bit* (sono legati alla parità della stringa originale).

Chiamiamo sempre $\boldsymbol{\xi}$ il vettore originale e $\mathbf{J}^0$ quello codificato e supponiamo di voler inviare il messaggio

$$\boldsymbol{\xi} = 1 \quad 0 \quad 0 \quad 0.$$

L'algoritmo di coding funziona come segue:

- i primi quattro bit vengono inviati uguali al messaggio originale, ossia $\xi_1\xi_2\xi_3\xi_4 = J_1^0 J_2^0 J_3^0 J_4^0$:
$$\mathbf{J}^0 = 1 \quad 0 \quad 0 \quad 0 \quad * \quad * \quad *$$
- J_5^0 è la somma dei primi tre bit $\xi_1 + \xi_2 + \xi_3$:
$$\mathbf{J}^0 = 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad * \quad *$$
- J_6^0 è la somma degli ultimi tre bit $\xi_2 + \xi_3 + \xi_4$:
$$\mathbf{J}^0 = 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad *$$
- J_7^0 è la somma del primo, terzo e quarto bit $\xi_1 + \xi_3 + \xi_4$:
$$\mathbf{J}^0 = 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1$$

dove le somme sono da intendere in modulo 2 (d'ora in avanti sarà dato come sottinteso). Il canale introduce un vettore di errore (come già specificato in precedenza, ogni bit ha una probabilità p di essere mutato), in questo caso uguale per esempio a

$$\boldsymbol{\Delta} = 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0,$$

da cui si ottiene

$\mathbf{J}^0$	1	0	0	0	1	0	1	+
$\boldsymbol{\Delta}$	0	1	0	0	0	0	0	=
$\mathbf{J}$	1	1	0	0	1	0	1	

Con $\mathbf{J} = \mathbf{J}^0 + \boldsymbol{\Delta}$ indichiamo quindi il vettore di output del canale.

L'algoritmo di decoding può essere presentato in modo grafico con l'aiuto di un'immagine. Disponiamo i bit del vettore $\mathbf{J}$ all'interno di tre circonferenze come in Figura 3.2a. Calcoliamo la parità z_i di ognuna di esse e costruiamo il vettore $\mathbf{z} = (z_1, z_2, z_3)$. Si definisce *sindrome* il vettore $\mathbf{z}$ che non contiene solo zeri: non è difficile convincersi che $\mathbf{J} = \mathbf{J}^0$ solo se la parità di ogni circonferenza è uguale a 0. Nel caso qui riportato, si ha $\mathbf{z} = (1, 1, 0)$ (le circonferenze tratteggiate hanno parità uguale a 1), sintomo del fatto che il canale ha introdotto un errore nel messaggio di input. L'obiettivo è dunque modificare il minor numero di bit affinchè la parità di ogni circonferenza sia nulla, ossia $\mathbf{z} = (0, 0, 0)$.

Studiando la Figura 3.2b, si nota che $J_2 = 1$ è l'unico bit che compare all'interno delle due circonferenze tratteggiate e all'esterno di quella a parità nulla. Ponendo infatti $J_2 = 0$, entrambe le circonferenze tratteggiate assumono parità uguale a

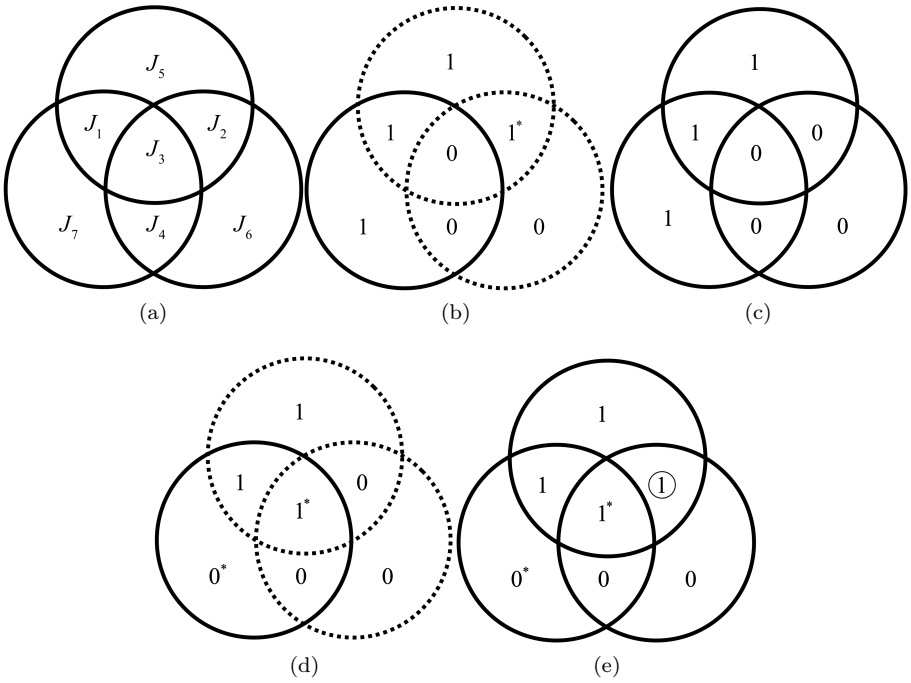


Figura 3.2: Rappresentazione grafica dell'algoritmo di decoding dell'Hamming code (7,4). Le circonferenze intere sono a parità uguale a 1, quelle tratteggiate uguale a 0. I bit contrasseganti con * sono quelli mutati dal canale.

0, e la terza continua a mantenere parità nulla: deve pertanto essere il candidato ottimale per la correzione dell'errore intodotto dal canale (Figura 3.2c).

Grazie a questo espediente geometrico è possibile individuare la presenza di un bit mutato dal canale. A seconda della sindrome $\mathbf{z}$, infatti, si può individuare quale bit sia il responsabile dell'errore:

Sindrome $\mathbf{z}$	000	001	010	100	011	101	110	111
bit errato	/	J_7	J_6	J_5	J_4	J_1	J_2	J_3

É naturale chiedersi come si comporti l'Hamming code (7,4) nel caso in cui più di un bit venga mutato dal rumore del canale. La Figura 3.2d mostra la situazione dove due bit, ossia J_3 e J_7 vengono modificati. La sindrome $\mathbf{z} = (1, 1, 0)$ suggerisce che il bit errato sia J_2, che pertanto viene modificato dall'algoritmo ottimale di decoding (infatti, supponendo che p sia piccola, l'algoritmo agisce come se non più di un bit ogni sette venga modificato dal rumore). Il risultato finale è un vettore $\mathbf{J}$ che contiene tre errori al posto che due, come in Figura 3.2e.

Chiamiamo *probabilità di errore di blocco* p_B la probabilità che uno o più bit all'interno di un blocco non siano uguali ai corrispettivi originali, ossia $p_B := P(\hat{\boldsymbol{\xi}} \neq \boldsymbol{\xi})$. La probabilità di errore p_b per ogni bit è data dunque dalla probabilità media

$$p_b = \frac{1}{N} \sum_{i=1}^{N} P(\hat{\xi}_i \neq \xi_i).$$

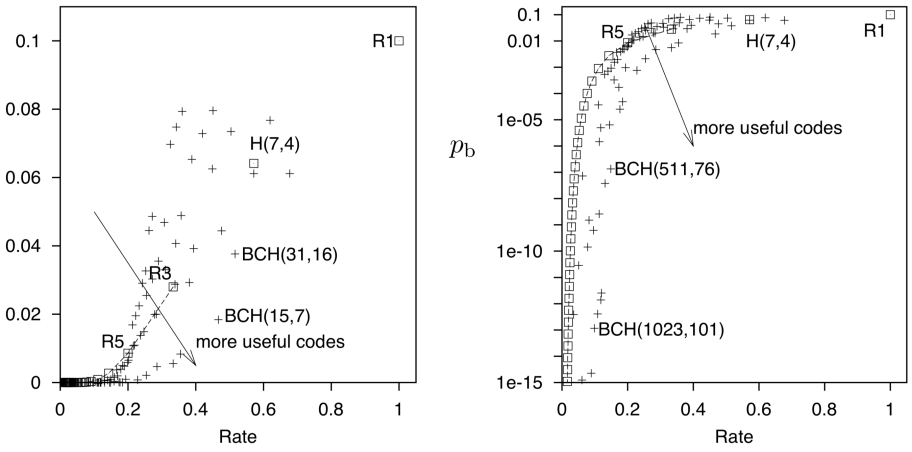


Figura 3.3: Probabilità di errore p_b in funzione del rate R per i Repetition codes, per l'Hamming code (7,4) e per gli algoritmi BCH con lunghezza dei blocchi fino a 1023 bit. Il canale utilizzato è sempre di tipo BSC con probabilità di errore $p = 0.1$ (la figura di destra è in scala logaritmica) [12].

Si può dimostrare che $p_b = \mathcal{O}(p^2)$, come nel caso del repetition code R_3. Tuttavia, il rate dell'Hamming code (7,4) è pari a $R = 4/7$, contro $R = 1/3$ dell'R_3.

3.2 ECC e vetri di spin

Per affrontare il problema dell'ECC e costruire una teoria che sia in grado di studiarlo in profondità attraverso il modello di Ising, è conveniente utilizzare i valori ± 1 al posto di 0 e 1 per identificare i bit. L'operazione di base per una sequenza di bit è l'ormai nota somma modulo 2, che corrisponde al prodotto di Ising spin se identifichiamo 0 con $S_i = 1$ e 1 con $S_i = -1$. A titolo di esempio, pensiamo $0 + 1 = 1$ come $1 \times (-1) = -1$ e analogamente $1 + 1 = 0$ come $(-1) \times (-1) = 1$. La parità di un gruppo di bit (in un qualunque *parity-check code*) corrisponderà dunque al prodotto di un'opportuna sequenza di spin.

Per costruire un modello analogo a quello di Ising, associamo un parametro di spin ξ_i per ogni sito e definiamo l'interazione tra due siti i e j del reticolo come $J^0_{ij} = \xi_i \xi_j$. Pertanto l'hamiltoniana assume la forma

$$H = -\sum_{\langle ij \rangle} \xi_i \xi_j S_i S_j, \tag{3.11}$$

il cui ground state è chiaramente dato dalla configurazione $S_i = \xi_i \ \forall i$ (o analogamente $S_i = -\xi_i \ \forall i$).

Ritornando all'error-correcting code, nel caso generale l'interazione è data da $\mathbf{J}^0 = \{J^0_{i_1 \cdots i_r} = \xi_{i_1} \cdots \xi_{i_r}\}$, con r intero, per opportune combinazioni di $\{i_1 \cdots i_r\}$. Pertanto, il vettore di dati che verrà poi immesso nel canale rumoroso sarà $\mathbf{J}^0$ invece che $\boldsymbol{\xi}$ (esattamente come mostrato nei due esempi precedenti). Il messagio di encoding è ridondante rispetto a quello originale, dal momento che il numero di interezioni N_B è maggiore del numero di spin N. Se per il momento consideriamo $r = 2$, risultano in totale $N_B = 2N$ interazioni.

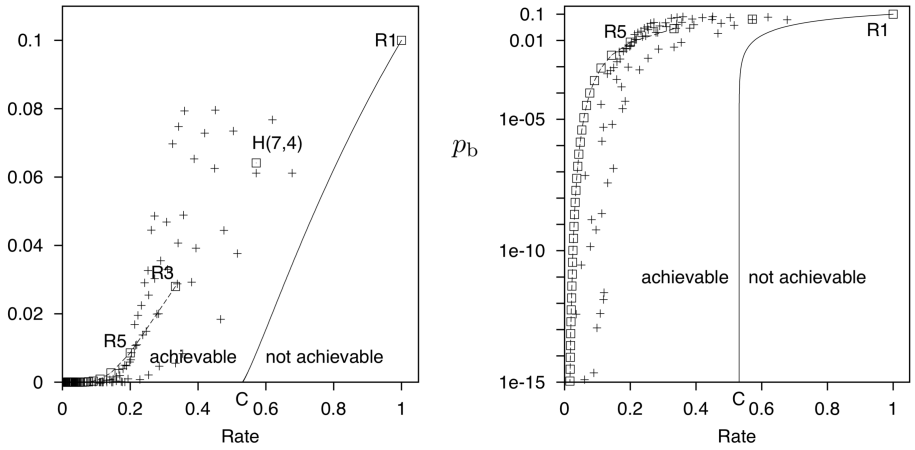


Figura 3.4: *Teorema della codifica di un canale* di Shannon. La curva continua rappresenta il limite di Shannon per un canale BSC con probabilità di errore $p = 0.1$. Rate tali che $R < C$ (con C definito in (3.13))sono teoricamente raggiungibili con parametro p_b arbitrariamente piccolo nel momento in cui la lunghezza del messaggio trasmesso tende all'infinito. I punti mostrano la performance di alcuni algoritmi, tra cui i repetition codes e l'Hamming code (7,4) (la figura di destra è in scala logaritmica). [12]

In generale, il ground state dell'hamiltoniana di un canale rumoroso sarà diverso rispetto a quello dell'hamiltoniana di un canale error-free. Tuttavia, se il rumore del canale è sufficientemente piccolo, possiamo trattare quest'ultimo come una *perturbazione* del sistema error-free: in prima approssimazione la configurazione originale di spin costituisce ancora il ground state del sistema rumoroso.

Nella teoria ECC si dimostra l'esistenza di una soglia di ridondanza del messaggio di encoding tale per cui è possibile ricostruire il messaggio originale una volta che questa viene raggiunta. Questa soglia prende il nome di *limite di Shannon* [19]. Si dimostra che è possibile una trasmissione senza errori tramite un canale BSC nel limite di una sequenza di lunghezza infinita:

$$R < C, \tag{3.12}$$

dove C è detta *capacità del canale* ed è uguale a

$$C := 1 - H_2(p) = 1 - \left[p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \right]. \tag{3.13}$$

(p indica la probabilità che un bit sia mutato dal rumore). Questo teorema garantisce che è possibile trasmettere un messaggio privo di errore in un canale BSC nel momento in cui il rate non supera la capacità propria del canale stesso ed è implementata un'opportuna procedura di encoding e decoding (la spiegazione e la dimostrazione di questo teorema sono riportate nell'Appendice B, dove compare anche una definizione più rigorosa del concetto di capacità di un canale). Un esempio di codice che satura asintoticamente il limite di Shannon è il *codice di Sourlas* [21]: si prendono tutti i possibili prodotti di r spin da N siti e, mandando $N \to \infty$ prima e $r \to \infty$ poi, si riesce a saturare il limite di Shannon.

Riconsiderando l'idea di cui sopra senza l'ipotesi di piccola probabilità di rumore, in generale si osserva che il ground state del sistema di Ising non coincide necessariamente con la configurazione originale di spin $\mathbf{J}^0$. Questo suggerisce che tale configurazione sia uno stato eccitato del sistema, cui viene associata una specifica temperatura T_p (che dipende dal tasso di errore p).

Parte II

FONDAMENTI TEORICI ED ANALISI PRELIMINARI DELLA SIMULAZIONE DEI CODICI DI SOURLAS A CONNETTIVITÀ FINITA

Capitolo 4

Meccanica statistica dei codici di Sourlas

La parte che segue propone un approfondimento riguardo alla meccanica statistica alla base dei *codici di Sourlas*, fornendo una prima descrizione generale delle grandezze fondamentali necessarie per lo studio della teoria ECC con l'utilizzo del modello di vetri di spin.

4.1 Probabilità condizionata

Supponiamo che la configurazione di spin $\boldsymbol{\xi}$ sia stata generata secondo la distribuzione di probabilità $P(\boldsymbol{\xi})$ (prior). Riprendiamo il percorso iniziato alla fine del capitolo precedente e definiamo un insieme di prodotti di r spin come

$$J^0_{i_1\cdots i_r} := \xi_{i_1}\cdots\xi_{i_r} (= \pm 1) \tag{4.1}$$

e supponiamo di inviare questo vettorore in un canale rumoroso. L'output $J_{i_1\cdots i_r}$ è mutato dal corrispondente input $J^0_{i_1\cdots i_r}$ in $-\xi_{i_1}\cdots\xi_{i_r}$ con probabilità p. La probabilità di output corretto (ossia $\xi_{i_1}\cdots\xi_{i_r}$) è quindi $1-p$.

A questo punto è ragionevole definire la probabilità di output di un canale BSC come una *probabilità condizionata*:

$$P(J_{i_1\cdots i_r}|\xi_{i_1}\cdots\xi_{i_r}) = \frac{\exp(\beta_p J_{i_1\cdots i_r}\xi_{i_1}\cdots\xi_{i_r})}{2\cosh\beta_p}, \tag{4.2}$$

dove β_p è definita come

$$\beta_p : \quad e^{2\beta_p} = \frac{1-p}{p}.$$

Osserviamo che la probabilità (4.2) è uguale a $1-p$ se $J_{i_1\cdots i_r} = \xi_{i_1}\cdots\xi_{i_r}$ ed è uguale a p se $J_{i_1\cdots i_r} = -\xi_{i_1}\cdots\xi_{i_r}$, proprio come cercato.

Se assumiamo che la (4.2) sia indipendente per ogni sito, allora la probabilità totale sarà data dal prodotto delle singole probabilità, ossia

$$P(\mathbf{J}|\boldsymbol{\xi}) = \frac{1}{(2\cosh\beta_p)^{N_B}} \exp\Big(\beta_p \sum J_{i_1\cdots i_r}\xi_{i_1}\cdots\xi_{i_r}\Big). \tag{4.3}$$

4.1.1 Formula di Bayes

Dal momento che la procedura di decoding prevede di ricostruire $\boldsymbol{\xi}$ a partire da $\mathbf{J}$, abbiamo bisogno di definire la probabilità $P(\boldsymbol{\xi}|\mathbf{J})$ (posterior). Quindi, ricordando l'espressione della *probabilità congiunta* [1]

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A), \tag{4.4}$$

troviamo la *Formula di Bayes*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_C P(B|C)P(C)}. \tag{4.5}$$

A questo punto, ricaviamo l'equazione per $P(\boldsymbol{\sigma}|\mathbf{J})$ sfruttando proprio la (4.5):

$$P(\boldsymbol{\sigma}|\mathbf{J}) = \frac{P(\mathbf{J}|\boldsymbol{\sigma})P(\boldsymbol{\sigma})}{\mathrm{Tr}_{\boldsymbol{\nu}} P(\mathbf{J}|\boldsymbol{\nu})P(\boldsymbol{\nu})}, \tag{4.6}$$

dove abbiamo identificato con $\boldsymbol{\sigma} = \{\sigma_1 \cdots \sigma_N\}$ le variabili dinamiche usate per il decoding. Per uniformare la notazione, chiamiamo $\hat{\boldsymbol{\xi}} = \{\hat{\xi_1} \cdots \hat{\xi_N}\}$ il vettore finale decodificato.

É possibile ricavare il messaggio originale $\boldsymbol{\xi}$ dalla (4.6) una volta noto il prior $P(\boldsymbol{\sigma})$. Per semplificare l'analisi teorica, assumiamo che ogni messaggio venga generato con uguale probabilità (cioè $P(\boldsymbol{\sigma})$ rimane costante), da cui segue immediatamente che

$$P(\boldsymbol{\sigma}|\mathbf{J}) = \frac{\exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \sigma_{i_1} \cdots \sigma_{i_r}\Big)}{\mathrm{Tr}_{\boldsymbol{\nu}} \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \nu_{i_1} \cdots \nu_{i_r}\Big)}. \tag{4.7}$$

Grazie all'equazione (4.7) è chiara l'analogia con il modello di vetro di spin, in cui $\mathbf{J}$ interpreta proprio il ruolo dell'interazione tra le coppie di spin.

4.1.2 MAP e MPM

Dal momento che l'equazione (4.7) è l'analogo della (1.2), già sappiamo che massimizzare $P(\boldsymbol{\sigma}|\mathbf{J})$ equivale a identificare il ground state dell'hamiltoniana

$$H = -\sum J_{i_1 \cdots i_r} \sigma_{i_1} \cdots \sigma_{i_r}. \tag{4.8}$$

Questo metodo di decoding prende il nome di *maximum a posteriori probability* (MAP). L'idea di fondo è la seguente: massimizzare $P(\mathbf{J}|\boldsymbol{\sigma})$ rispetto a $\boldsymbol{\sigma}$ è equivalente a massimizzare il posterior $P(\boldsymbol{\sigma}|\mathbf{J})$ nell'ipotesi di uniformità di $P(\boldsymbol{\sigma})$.

Un'altra procedura di decoding, ossia il *maximizer of posterior marginals* (MPM), si concentra invece sul singolo bit al posto che sulla seguenza intera. Tramite l'operatore di traccia effettuato su ogni $\sigma_{j \neq i}$ (operazione che prende il nome di *marginalizzazione*), si ricava il posterior solo per σ_i:

$$P(\sigma_i|\mathbf{J}) = \frac{\mathrm{Tr}_{\boldsymbol{\sigma}(\neq\sigma_i)} \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \sigma_{i_1} \cdots \sigma_{i_r}\Big)}{\mathrm{Tr}_{\boldsymbol{\nu}} \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \nu_{i_1} \cdots \nu_{i_r}\Big)}. \tag{4.9}$$

[1] Dati due eventi A e B, definiamo la *probabilità congiunta* $P(A, B)$ come la probabilità che si verifichino entrambi gli eventi.

Si confrontano poi le probabilità $P(\sigma_i = 1|\mathbf{J})$ e $P(\sigma_i = -1|\mathbf{J})$ e si assegna all'i-esimo bit del vettore decodificato il valore $\hat{\xi}_i = 1$ se è maggiore la prima o $\hat{\xi}_i = -1$ se è maggiore la seconda. In modo del tutto analogo possiamo riscrivere questa condizione come

$$\begin{aligned}\hat{\xi}_i =& \operatorname{sgn}\{P(\sigma_i = 1|\mathbf{J}) - P(\sigma_i = -1|\mathbf{J})\} = \operatorname{sgn} \sum_{\sigma_i = \pm 1} \sigma_i P(\sigma_i|\mathbf{J}) = \\ =& \operatorname{sgn} \frac{\operatorname{Tr}_{\boldsymbol{\sigma}} \sigma_i P(\sigma_i|\mathbf{J})}{\operatorname{Tr}_{\boldsymbol{\sigma}} P(\sigma_i|\mathbf{J})} := \operatorname{sgn}\langle\sigma_i\rangle_{\beta_p}.\end{aligned} \tag{4.10}$$

dove $\langle\sigma_i\rangle_{\beta_p}$ costituisce la magnetizzazione locale. L'equazione (4.10) suggerisce di calcolare la magnetizzazione locale a una temperatura finita $T_p = \beta_p^{-1}$, e assegna il suo segno come valore di $\hat{\xi}_i$. Per quanto riguarda invece il MAP, chiaramente la minimizzazione dell'energia porta alla condizione $\beta \to \infty$.

4.1.3 Canale gaussiano

Fino a questo momento abbiamo sempre considerato canali BSC. Un altro esempio fondamentale è il *canale gaussiano*: il messaggio di encoding $\xi_{i_1} \cdots \xi_{i_r}$ è inserito in un canale con ampiezza $J_0 \xi_{i_1} \cdots \xi_{i_r}$. L'output è distibuito in modo continuo attorno a questo input con una distribuzione gaussiana di varianza J^2:

$$P(J_{i_1 \cdots i_r}|\xi_{i_1} \cdots \xi_{i_r}) = \frac{1}{\sqrt{2\pi}J} \exp\left\{ -\frac{(J_{i_1 \cdots i_r} - J_0 \xi_{i_1} \cdots \xi_{i_r})^2}{2J^2} \right\}.$$

Ripercorrendo gli stessi conti eseguiti per il canale BSC si arriva all'equazione finale per il posterior:

$$P(\boldsymbol{\sigma}|\mathbf{J}) = \frac{\exp\Big((J_0/J^2) \sum J_{i_1 \cdots i_r} \sigma_{i_1} \cdots \sigma_{i_r} \Big)}{\operatorname{Tr}_{\boldsymbol{\eta}} \exp\Big((J_0/J^2) \sum J_{i_1 \cdots i_r} \eta_{i_1} \cdots \eta_{i_r} \Big)}. \tag{4.11}$$

Un rapido confronto con l'equazione (4.7) mostra che i posterior del canale BSC e di quello gaussiano sono analoghi nel momento in cui si sostituisce β_p con J_0/J^2.

4.2 Parametro di overlap

4.2.1 Misura della performance di decoding

É necessario inserire una misura della performance di decoding che sia in grado di descrivere quanto il messaggio decodificato $\hat{\boldsymbol{\xi}}$ sia prossimo a $\boldsymbol{\xi}$, così da avere una stima della correttezza dell'informazione decodificata rispetto a quella di input. É utile in questo caso non specificare il parametro β, cosicchè il ragionamento che segue sia valido sia per il metodo MAP sia per l'MPM.

Il prodotto tra $\hat{\xi}_i$ e il corrispettivo bit originale ξ_i, ossia $\xi_i \operatorname{sgn}\langle\sigma_i\rangle_\beta$, è uguale a 1 se $\hat{\xi}_i = \xi_i$ oppure -1 altrimenti: è chiaro che la strategia da seguire sia quella di aumentare la probabilità che tale prodotto sia 1. Infatti, nel momento in cui tale prodotto converge a 1 per ogni bit, i vettori $\hat{\boldsymbol{\xi}}$ e $\boldsymbol{\xi}$ tendono a sovrapporsi.

Implementiamo a tal proposito la funzione

$$m(\beta) = \mathrm{Tr}_{\boldsymbol{\xi}} \sum_{\mathbf{J}} P(\boldsymbol{\xi}) P(\mathbf{J}|\boldsymbol{\xi}) \xi_i \, \mathrm{sgn}\langle \sigma_i \rangle_\beta, \tag{4.12}$$

che definisce proprio il *parametro di overlp* (ossia di sovrapposizione) tra il messaggio originale e quello decodificato (per come è stata defnita, la funzione $m(\beta)$ è l'analogo discreto della media configurazionale introdotta nel Capitolo 2). Assumendo che il prior sia uniforme, ossia $P(\boldsymbol{\xi}) = 2^{-N}$, si può riscrivere la (4.12) come

$$m(\beta) = \frac{1}{2^N (2\cosh \beta_p)^{N_B}} \sum_{\mathbf{J}} \mathrm{Tr}_{\boldsymbol{\xi}} \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \sigma_{i_1} \cdots \sigma_{i_r}\Big) \xi_i \, \mathrm{sgn}\langle \sigma_i \rangle_\beta. \tag{4.13}$$

Date due stringhe di bit di ugual lunghezza, definiamo *distanza di Hamming* il numero di posizioni nelle quali i simboli corrispondenti sono diversi. In altre parole, la distanza di Hamming misura il numero di sostituzioni necessarie per convertire una stringa nell'altra. Osserviamo che l'overlap e la distanza di Hamming sono strettamente correlati tra loro: tanto minore è la distanza di Hamming, tanto maggiore è la sovrapposione. A titolo di esempio, nel caso in cui $\hat{\xi}_i = \xi_i \; \forall i$, si avrà una sovrapposizione $m(\beta) = 1$ e una distanza di Hamming nulla.

4.2.2 Limite superiore dell'overlap

É stata dunque definita un parametro di overlap (4.12) che dipende da β. Quel che vogliamo ora mostrare è che $m(\beta)$ è una funzione non monotona che ha il massimo proprio in $\beta = \beta_p$. Ricordando l'espressione di $m(\beta)$ nell'equazione (4.13), osserviamo che:

$$\begin{aligned}
m(\beta) \leq & \frac{1}{2^N (2\cosh \beta_p)^{N_B}} \sum_{\mathbf{J}} \Big| \mathrm{Tr}_{\boldsymbol{\xi}} \, \xi_i \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \xi_{i_1} \cdots \xi_{i_r}\Big) \xi_i \, \mathrm{sgn}\langle \sigma_i \rangle_\beta \Big| = \\
= & \frac{1}{2^N (2\cosh \beta_p)^{N_B}} \sum_{\mathbf{J}} \frac{\Big\{ \mathrm{Tr}_{\boldsymbol{\xi}} \, \xi_i \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \xi_{i_1} \cdots \xi_{i_r}\Big) \Big\}^2}{\Big| \mathrm{Tr}_{\boldsymbol{\xi}} \, \xi_i \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \xi_{i_1} \cdots \xi_{i_r}\Big) \Big|} = \\
= & \frac{1}{2^N (2\cosh \beta_p)^{N_B}} \sum_{\mathbf{J}} \mathrm{Tr}_{\boldsymbol{\xi}} \, \xi_i \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \xi_{i_1} \cdots \xi_{i_r}\Big) \\
& \cdot \frac{\mathrm{Tr}_{\boldsymbol{\xi}} \, \xi_i \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \xi_{i_1} \cdots \xi_{i_r}\Big)}{\Big| \mathrm{Tr}_{\boldsymbol{\xi}} \, \xi_i \exp\Big(\beta_p \sum J_{i_1 \cdots i_r} \xi_{i_1} \cdots \xi_{i_r}\Big) \Big|}.
\end{aligned}$$

Riprendendo la definizione (4.10) di $\langle\sigma_i\rangle_{\beta_p}$, constatiamo che il segno non dipende dal denominatore (è una somma di grandezze non negative), da cui:

$$\begin{aligned}
\operatorname{sgn}\langle\sigma_i\rangle_\beta &= \operatorname{sgn}\sum_{\sigma_i=\pm1}\sigma_i P(\sigma_i|\mathbf{J}) = \\
&= \operatorname{sgn}\sum_{\sigma_i=\pm1}\sigma_i\frac{\operatorname{Tr}_{\boldsymbol{\sigma}(\neq\sigma_i)}\exp\Big(\beta_p\sum J_{i_1\cdots i_r}\sigma_{i_1}\cdots\sigma_{i_r}\Big)}{\operatorname{Tr}_{\boldsymbol{\nu}}\exp\Big(\beta_p\sum J_{i_1\cdots i_r}\nu_{i_1}\cdots\nu_{i_r}\Big)} = \\
&= \operatorname{sgn}\operatorname{Tr}_{\boldsymbol{\sigma}}\sigma_i\exp\Big(\beta_p\sum J_{i_1\cdots i_r}\sigma_{i_1}\cdots\sigma_{i_r}\Big) = \\
&= \frac{\operatorname{Tr}_{\boldsymbol{\sigma}}\sigma_i\exp\Big(\beta_p\sum J_{i_1\cdots i_r}\sigma_{i_1}\cdots\sigma_{i_r}\Big)}{\Big|\operatorname{Tr}_{\boldsymbol{\sigma}}\sigma_i\exp\Big(\beta_p\sum J_{i_1\cdots i_r}\sigma_{i_1}\cdots\sigma_{i_r}\Big)\Big|}.
\end{aligned}$$

Nella catena di uguaglianze di cui sopra, nella quarta riga non compare più la funzione *segno*. Questo deriva dal fatto che

$$\operatorname{sgn}\operatorname{Tr}_{\boldsymbol{\sigma}}\sigma_i\exp(\cdots) = \frac{\operatorname{Tr}_{\boldsymbol{\sigma}}\sigma_i\exp(\cdots)}{|\operatorname{Tr}_{\boldsymbol{\sigma}}\sigma_i\exp(\cdots)|}. \tag{4.14}$$

Possiamo pertanto riscrivere la disuguaglianza precedente come:

$$\begin{aligned}
m(\beta) \leq& \frac{1}{2^N(2\cosh\beta_p)^{N_B}}\sum_{\mathbf{J}}\operatorname{Tr}_{\boldsymbol{\xi}}\xi_i\exp\Big(\beta_p\sum J_{i_1\cdots i_r}\xi_{i_1}\cdots\xi_{i_r}\Big)\operatorname{sgn}\langle\sigma_i\rangle_\beta = \\
=& m(\beta_p).
\end{aligned} \tag{4.15}$$

Abbiamo quindi mostrato che la funzione di sovrapposizione ha un massimo proprio in corrispondenza di $\beta = \beta_p$. Questo non sorprende per quento riguarda l'MPM: avevamo già mostrato come la magnetizzazione andasse calcolata proprio in corrispondenza della temperatura $T_p = \beta_p^{-1}$. Notiamo inoltre che, nonostante il MAP massimizzi il posterior $P(\boldsymbol{\xi}|\mathbf{J})$, la sua probabilità di errore per un singolo bit è maggiore che per l'MPM (valutato in $\beta = \beta_p$).

Capitolo 5

Implementazione numerica di un codice di Sourlas a connettività finita

Nel capitolo precedente abbiamo mostrato alcune generalità riguardo alla meccanica statistica dei codici di Sourlas a connettività finita. Obiettivo di questo capitolo è definire una procedura tale per cui sia possibile costruire un algoritmo di correzione degli errori mediante il modello di vetri di spin.

5.1 L'hamiltoniana H dei codici di Sourlas

5.1.1 Definizione di H

Nella sezione 4.1 è stato dimostrato come l'obiettivo dell'algoritmo di decoding (sempre nell'ipotesi di uniformità della probabilità $P(\boldsymbol{\sigma})$) sia quello di massimizzare la probabilità $P(\boldsymbol{\sigma}|\mathbf{J})$ affinchè si possano ottenere prestazioni ottimali. É stato inoltre verificato come, in virtù del modello di vetri di spin, fare ciò sia totalmente analogo ad identificare il ground state dell'hamiltoniana (4.8). Supponiamo quindi di voler inviare un messaggio $\boldsymbol{\xi}$ contenente N bit (± 1) attraverso un canale rumoroso di tipo BSC: ciò significa che ogni bit ξ_i del messaggio ha una probabilità p di essere mutato in $-\xi_i$. Ridefiniamo dunque la funzione hamiltoniana (4.8) con l'aggiunta di un campo magnetico esterno:

$$H(\boldsymbol{\sigma}) := -\sum_{i_1 \cdots i_K}^{N} J_{i_1,\cdots,i_K}\sigma_{i_1}\cdots\sigma_{i_k} - \sum_{l}^{N} h_l\sigma_l, \tag{5.1}$$

dove K indica il numero di corpi interagenti, mentre $\boldsymbol{\sigma}$ è il vettore di spin (± 1, di dimensione N) da cui dipende H. La definizione di $J_{i_1,\cdots,i_K}$ è la seguente:

$$J_{i_1,\cdots,i_K} := \begin{cases} -\xi_{i_1}\cdots\xi_{i_K} & \text{con probabilità } p \\ +\xi_{i_1}\cdots\xi_{i_K} & \text{con probabilità } 1-p \\ 0 & \nexists \text{ l'interazione tra } \xi_{i_1},\cdots,\xi_{i_K} \end{cases} \tag{5.2}$$

e in modo del tutto analogo per $\boldsymbol{h}$:

$$h_l := \begin{cases} -\xi_l & \text{con probabilità } p \\ +\xi_l & \text{con probabilità } 1-p \end{cases} \tag{5.3}$$

Questa hamltoniana ha un'interpetazione grafica che può essere ottenuta tramite un *factor graph*.

5.1.2 Rappresentazione di H tramite un factor graph

Dal momento che l'interpretazione tramite grafi dell'hamiltoniana di cui sopra è necessaria per la comprensione della simulazione del codice di Sourlas, diamo di seguito alcune definizioni fondamentali che porteranno successivamente all'introduzione del factor graph bipartito [10]:

Definizione 5.1.1. *Un grafo è una coppia ordinata $\mathcal{G} = \mathcal{G}(E, V)$, dove V è l'insieme dei vertici (o nodi) ed E l'insieme dei lati, tale per cui gli elementi di E siano coppie di elementi di V ($E \subseteq V \times V$).*

Definizione 5.1.2. *Dato un lato $e \in E$ di un grafo $\mathcal{G}$, ci sono due vertici $v_1, v_2 \in V$ tali per cui $e = (v_1, v_2)$: diciamo che v_1 e v_2 sono adiacenti. Si dice che il grafo non è direzionato se $(v_1, v_2) = (v_2, v_1)$.*

Definizione 5.1.3. *Il grado di un vertice $v \in V$ è il numero dei suoi vertici adiacenti.*

Definizione 5.1.4. *Un grafo bipartito è un grafo in cui i suoi vertici possono essere suddivisi in due insiemi disgiunti U e V tali per cui ogni vertice $u \in U$ è collegato ad un vertice $v \in V$.*

Note queste definizioni minimali, spostiamo l'attenzione verso i factor graph. Per far questo, iniziamo col considerare un insieme di variabili $\{x_1, \cdots, x_N\}$ tali che $x_i \in \mathcal{X}_i$ (dominio finito) e $f(x_1, \cdots, x_N)$ sia una funzione reale di dominio $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ (detto *spazio di configurazione*). Supponiamo ora che f sia fattorizzabile in un prodotto di M funzioni locali f_j (dette appunto *fattori*), ciascuna con insieme di variabili $S_j \subseteq \{x_1, \cdots, x_N\}$, ossia

$$f(x_1, \cdots, x_N) = \prod_{j=1}^{M} f_j(S_j). \tag{5.4}$$

Chiaramente ogni funzione f_j ha il proprio spazio di configurazione. Siamo dunque pronti per dare la seguente definizione:

Definizione 5.1.5. *Sia $f = f(x_1, \cdots, x_N)$ una funzione tale che possa essere fattorizzata come in (5.4). Un factor graph $\mathcal{G}(E, V)$ corrispondente alla funzione f è un grafo bipartito tale che:*

- *ogni variabile x_i è associata a un vertice denotato con un cerchio nero;*
- *ogni fattore f_j è associato a un quadrato bianco;*
- *se x_i è nel dominio di f_j, esiste un lato del grafo $e_{ij} = (x_i, f_j)$ che collega il cerchio x_i al quadrato f_j.*

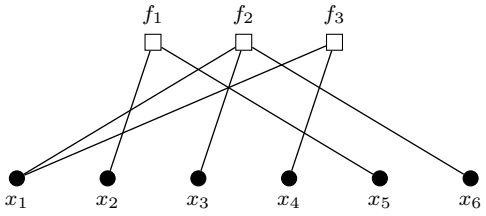


Figura 5.1: Factor graph della funzione (5.5). I fattori f_j sono rappresentati da quadrati bianchi, i vertici x_i da pallini neri.

Non è difficile convincersi del fatto che ad ogni funzione f fattorizzata corrisponda un solo factor graph $\mathcal{G}(E, V)$ e viceversa, in virtù del fatto che la funzione che mappa f nel factor graph è una funzione uno-a-uno. Un rapido esempio è il seguente: sia f una funzione di 3 variabili tale che

$$f(x_1, x_2, x_3, x_4, x_5, x_6) = f_1(x_2, x_5) f_2(x_1, x_3, x_6) f_3(x_1, x_4). \tag{5.5}$$

La Figura 5.1 rappresenta il corrispettivo factor graph.

Questo breve (ma necessario) excursus sui factor graph permette quindi sia di dare un'interpreazione grafica alla funzione hamiltoniana, sia di poterla riscrivere in modo più semplice in vista di alcuni conti che seguiranno.

Come già detto, il messaggio di ingresso è un vettore $\boldsymbol{\xi}$ di N bit, mentre con $\boldsymbol{\sigma}$ indichiamo il vettore di N variabili dell'hamiltoniana (5.1). A questo punto dobbiamo imporre alcune condizioni sul factor graph che rappresenta H.

- Assumiamo che il numero di interazioni $\boldsymbol{\mu}_i$ sia pari ad M. Ciò significa che se f è la funzione che descrive il factor graph, questa può essere scritta come il prodotto di M fattori f_j (nel grafo, dunque, devono comparire M quadrati bianchi).

- Dal momento che H descrive un'interazione di K spin, ogni interazione $\boldsymbol{\mu}_i$ deve essere della forma $\boldsymbol{\mu}_i = (\sigma_{i_1}, \cdots, \sigma_{i_K})$. Nel grafo, dunque, ogni quadrato (ossia ogni interazione $\boldsymbol{\mu}_i$) deve essere collegato con K cerchi neri (ossia con gli spin $\sigma_{i_1}, \cdots, \sigma_{i_K}$)

- Assumiamo che ogni spin σ_i interagisca con un numero N_{int} di spin fissato e uguale per ogni σ_i. Quindi, ogni cerchio nero deve essere collegato con N_{int} quadrati bianchi (cioè deve interagire con N_{int} interazioni $\boldsymbol{\mu}_i$ diverse).

Le condizioni che abbiamo posto implicano la seguente uguaglianza:

$$M \cdot K = N \cdot N_{\text{int}}. \tag{5.6}$$

Infatti, ogni cerchio nero è collegato con N_{int} quadrati bianchi, quindi ai cerchi neri corrispondono in totale $N \cdot N_{\text{int}}$ lati. Allo stesso modo, ogni quadrato bianco è collegato con K cerchi neri, per un totale di $M \cdot K$ lati. Questi due prodotti devono ovviamente coincidere. In aggiunta, richiediamo che ogni spin σ_i partecipi a interazioni tutte diverse tra loro. Ciò è possibile solo se

$$M > N_{\text{int}}. \tag{5.7}$$

Dunque, l'hamiltoniana (5.1) può essere riscritta nel seguente modo:

$$H(\boldsymbol{\sigma}) = -\sum_{i}^{M} J_{\boldsymbol{\mu}_i}\sigma_{\boldsymbol{\mu}_i} - \sum_{l}^{N} h_l\sigma_l, \tag{5.8}$$

dove $\sigma_{\boldsymbol{\mu}_i} = \prod_{j=1}^{K}\sigma_{i_j}$. Da ultimo, osserviamo che il rate R di trasmissione è pari a

$$R = \frac{K}{K + N_{\text{int}}}. \tag{5.9}$$

Infatti, $J_{\boldsymbol{\mu}}$ e $\boldsymbol{h}$ trasmettono rispettivamente M e N bit. A questo punto, tramite l'identità (5.6) si trova l'espressione di R.

5.2 Algoritmo di decoding

5.2.1 Ricerca del minimo di H

Nella sezione precedente è stata definita l'hamiltoniana dei codici di Sourlas tramite un grafo con numero di connessioni fissato a K ed uniforme.[1] Come già detto, per ottimizzare il codice è necessario minimizzare l'hamiltoniana H, ossia trovare la configurazione di spin $\hat{\boldsymbol{\xi}}$ tale per cui $H(\hat{\boldsymbol{\xi}}) = H_{\min}$. A costo di pedanteria sottolineamo questo aspetto: la discussione fin qui portata avanti si fonda sull'idea che il messaggio di ingresso $\boldsymbol{\xi}$ e il messaggio di output (che minimizza l'hamiltoniana) $\hat{\boldsymbol{\xi}}$ coincidano. Nella sezione 3.2 avevamo osservato come in realtà il minimo di H in generale non coincida con $\boldsymbol{\xi}$ proprio a causa del rumore introdotto dal canale. Tuttavia, ci mettiamo nell'ipotesi di rumore sufficientemente piccolo tale da giustificare la condizione $\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}$.

L'algoritmo qui implementato per la ricerca del minimo di H è l'*algoritmo della discesa del gradiente* [22]. In ottimizzazione e analisi numerica, il metodo di discesa del gradiente è una tecnica che consente di determinare i punti di massimo e di minimo di una funzione in più variabili. Di seguito presentiamo i passi fondamentali tramite cui ricerchiamo la configurazione $\hat{\boldsymbol{\xi}}$:

- si sceglie una configurazione di spin $\boldsymbol{\sigma}$ generica (ricordiamo $\sigma_i = \pm 1$);
- si esegue uno *scambio* (o *flip*) del primo spin, ossia $\sigma_1 \to -\sigma_1$. Chiamiamo $\boldsymbol{\sigma}_{\text{flip}}$ la configurazione con lo spin invertito, ossia $\boldsymbol{\sigma}_{\text{flip}} = (-\sigma_1\sigma_2\cdots\sigma_N)$;
- si calcola la differenza tra l'energia nella condizione finale e quella nella configurazione di partenza. Più nello specifico, si calcola $\Delta H = H(\boldsymbol{\sigma}_{\text{flip}}) - H(\boldsymbol{\sigma})$ e si studia il segno di ΔH: se $\Delta H < 0$ si sceglie $\boldsymbol{\sigma}_{\text{flip}}$ come nuova configurazione di spin, altrimenti se $\Delta H \geq 0$ si mantiene la precedente configurazione $\boldsymbol{\sigma}$
- si itera questa operazione per ogni spin σ_i e si arresta la ricerca del minimo nel momento in cui non si riescono più a trovare configurazioni di $\boldsymbol{\sigma}$ tali per cui $\Delta H < 0$.

La riscrittura della funzione H tramite la (5.8) è molto vantaggiosa proprio per il calcolo di ΔH (con la definizione (5.1) la notazione appesantirebbe il calcolo,

[1] É chiaro che, dato un messaggio di ingresso $\boldsymbol{\xi}$ di lunghezza N, la famiglia di grafi qui presa in esame costituisce una sottoclasse di tutti i possibili grafi che è possibile costruire.

andando ad influire negativamente sulle prestazioni del codice della simulazione). Riportiamo di seguito i passaggi salienti del conto. Supponiamo che $\boldsymbol{\sigma}_{\text{flip}}$ presenti un flip dell'n-esimo spin σ_n rispetto al vettore $\boldsymbol{\sigma}$. Ciò significa che è conveniente riscrivere H separando i termini che dipendono da σ_n rispetto a tutti gli altri:

$$\begin{aligned} H(\boldsymbol{\sigma}) &= -\sum_{i=1}^{M} J_{\boldsymbol{\mu}_i} \prod_{j=1}^{K} \sigma_j^{(i)} - \sum_{l=1}^{N} h_l \sigma_l = \\ &= -\sum_{i:\,\sigma_n \notin \boldsymbol{\mu}_i}^{M} J_{\boldsymbol{\mu}_i} \prod_{j=1}^{K} \sigma_j^{(i)} - \sum_{i:\,\sigma_n \in \boldsymbol{\mu}_i}^{M} J_{\boldsymbol{\mu}_i} \prod_{j\neq n}^{K} \sigma_j^{(i)} \sigma_n^{(i)} - \sum_{l\neq n}^{N} h_l \sigma_l - h_n \sigma_n. \end{aligned}$$

Ora è chiaro che, se calcolo $\Delta H = H(\boldsymbol{\sigma}_{\text{flip}}) - H(\boldsymbol{\sigma})$, i termini che non dipendono da σ_n si semplificano. Pertanto, ricordando che per $\boldsymbol{\sigma}_{\text{flip}}$ l'n-esimo spin vale $-\sigma_n$, si ottiene:

$$\begin{aligned} \Delta H &= H(\boldsymbol{\sigma}_{\text{flip}}) - H(\boldsymbol{\sigma}) = \\ &= -\sum_{i:\,\sigma_n \in \boldsymbol{\mu}_i}^{M} J_{\boldsymbol{\mu}_i} \prod_{j\neq n}^{K} \sigma_j^{(i)} (-\sigma_n^{(i)}) + \sum_{i:\,\sigma_n \in \boldsymbol{\mu}_i}^{M} J_{\boldsymbol{\mu}_i} \prod_{j\neq n}^{K} \sigma_j^{(i)} \sigma_n^{(i)} - h_n(-\sigma_n - \sigma_n) = \\ &= 2 \sum_{i:\,\sigma_n \in \boldsymbol{\mu}_i}^{M} J_{\boldsymbol{\mu}_i} \prod_{j=1}^{K} \sigma_j^{(i)} + 2h_n \sigma_n. \end{aligned}$$

Dal momento che σ_n compare in N_{int} interazioni $\boldsymbol{\mu}_i$ per definizione, la sommatoria di cui sopra è in realtà una somma su N_{int} addendi. Ma le simulazioni che stiamo trattando prevedono in genere $N_{\text{int}} = \mathcal{O}(K)$ (nel caso qui studiato, $K \leq 10$) e, poichè lavoriamo nel limite termodinamico, si ha dunque $N_{\text{int}} \ll N$. A questo punto è evidente che il calcolo di ΔH tramite l'equazione di cui sopra è molto più semplice (e quindi prestazionale a livello di codice) rispetto a quanto si riuscirebbe a fare tramite la definizione (5.1), essendo quest'ultima una sommatoria su N elementi.

Un'osservazione importante da sottolineare è la seguente: è del tutto arbitraria la scelta della condizione che interrompe l'algoritmo di ricarca del minimo. Se H fosse una funzione a valori continui, in generale si fisserebbe un $\epsilon > 0$ arbitrariamente piccolo e si arresterebbe la procedura se $\left|H(\boldsymbol{\sigma}^{(f)}) - H(\boldsymbol{\sigma}^{(i)})\right| < \epsilon$, dove $\boldsymbol{\sigma}^{(f)}$ e $\boldsymbol{\sigma}^{(i)}$ sono due vettori che differiscono per un solo scambio di spin. Tuttavia, in questo caso H è una funzione a valori discreti, dal momento che $H(\boldsymbol{\sigma}) \in \mathbb{Z}$. É pertanto più conveniente arrestare la procedura quando $\boldsymbol{\sigma}$ viene scorso per intero almeno una volta senza che nessuno scambio di spin riesca ad abbassare l'energia.

Il metodo della discesa del gradiente ha un grave difetto: può non arrivare a convergenza, ossia può non arrivare mai alla condizione $\boldsymbol{\sigma} = \hat{\boldsymbol{\xi}}$. Supponiamo infatti, per comodità, che H sia una funzione in una sola variabile reale (quindi immaginiamo che sia di dominio continuo e non discreto) a valori reali. In generale, H presentarà più minimi relativi, di cui uno sarà il minimo assoluto H_{min}. Ora, supponiamo che la configurazione iniziale $\boldsymbol{\sigma}_{\text{in}}$ da cui l'algoritmo inizia per trovare $\hat{\boldsymbol{\xi}}$ sia come in Figura 5.2. Il metodo di discesa del gradiente prevede che il punto successivo sia scelto in modo tale che $\Delta H < 0$, muovendosi però in un intorno stretto del punto $\boldsymbol{\sigma}_{\text{in}}$. Se infatti pensiamo alla procedura di cui sopra, la variazione di energia ΔH viene calcolata su due configurazioni $\boldsymbol{\sigma}$ e $\boldsymbol{\sigma}_{\text{flip}}$ che differiscono per un solo spin su N. Questo significa che, nel momento in cui $\boldsymbol{\sigma}_{\text{in}}$ si trova in una valle in cui il minimo è $H_{\text{err}} \neq H_{\text{min}}$, allora si ottiene $\boldsymbol{\sigma} \to \boldsymbol{\sigma}_{\text{err}}$ invece che $\boldsymbol{\sigma} \to \hat{\boldsymbol{\xi}}$. In

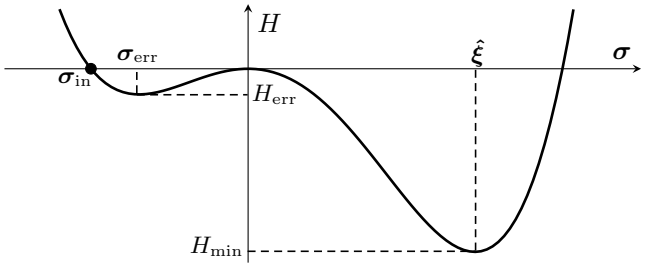


Figura 5.2: Generica funzione $H = H(\boldsymbol{\sigma})$ in un intervallo in cui sono presenti due valli, ossia due minimi relativi. La figura mostra come il metodo di *discesa del gradiente* non permetta di trovare il minimo assoluto H_{min} della funzione H nell'ipotesi in cui il dato di partenza $\boldsymbol{\sigma}_{\text{in}}$ si trovi nella valle corrispondente al minimo relativo H_{err}. Sotto questa condizione, infatti, si ottiene $\boldsymbol{\sigma}_{\text{in}} \to \boldsymbol{\sigma}_{\text{err}}$ invece che $\boldsymbol{\sigma}_{\text{in}} \to \hat{\boldsymbol{\xi}}$.

tal caso, dunque, questo metodo non è in grado di trovare il minimo assoluto della funzione H.

Tutto questo per dire che, in linea teorica, si potrebbe scegliere un dato inziale $\boldsymbol{\sigma}_{\text{in}}$ totalmente casuale. Tuttavia, più aumenta il numero di valli della funzione H e minore è la capacità del metodo di discesa del gradiente di convergere al minimo di H. Se invece si scegliesse il vettore di partenza $\boldsymbol{\sigma}_{\text{in}}$ in modo tale da trovarsi nella valle in cui è presente il minimo assoluto di H, allora la procedura fin qui descritta sarebbe perfettamente in grado di trovare quest'ultimo. Rimane comunque un ovvio problema: in generale non è noto dove si trovi il vettore $\hat{\boldsymbol{\xi}}$ che minimizza H, pertanto il problema sembra apparentemente irrisolvibile.

Non è possibile trovare una soluzione generale, però è possibile migliorare le prestazioni di questo algoritmo da un punto di vista statistico. Ora, il messaggio da inviare attraverso il canale è $\boldsymbol{\xi}$: ovviamente questo non è noto (altrimenti avrebbe poco senso costruire una procedura che permetta di ritrovarlo), bensì sono noti $J_{i_1,\cdots,i_K}$ e $\boldsymbol{h}$. Quest'ultimo è di nostro interesse. Ricordando la sua definizione in (5.3), ci si accorge che differisce dal vettore $\boldsymbol{\xi}$ per un numero di bit pari a $p \cdot N$ nel limite in cui N è sufficientemente grande (ricordando che $p \in [0, 1]$ è la probabilità di errore commessa sul singolo bit causata dal rumor del canale). In genere p è un numero piccolo, dell'ordine di 0.1: in questo caso specifico $\boldsymbol{h}$ differisce da $\boldsymbol{\xi}$ per circa un bit ogni 10. Ci aspettiamo dunque che inizializzando $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$ l'efficienza dell'algoritmo migliori rispetto ad inizializzare $\boldsymbol{\sigma}_{\text{in}}$ a una configurazione casuale.

5.2.2 Definizione del parametro di overlap m

Un primo test preliminare consiste nel verificare le condizioni al contorno della simulazione, ossia stabilire quali valori assegnare a specifici paramentri affinchè la simulazione restituisca risultati veritieri. L'hamiltoniana (5.1) dipende infatti da quattro paramentri (di cui solamente tre sono in realtà indipendenti, a causa della condizione (5.6)) che debbono pertanto essere fissati.

Il teorema di Shannon prevede la possibilità di ottenere una trasmissione priva di errori solo nel limite termodinamico, ossia solo per $N \to \infty$. Chiaramente questa condizione non può essere implementata in un algoritmo, da cui la necessità di fissare N sufficientemente grande da poter essere assunto come limite termodinamico.

Infatti, è del tutto lecito immaginare l'esistenza di certo un $N_{\text{lim}} > 0$ (il cui modulo dipenderà dall'errore p del canale, dal rate R e dal numero di interazioni K) tale che $\forall N > N_{\text{lim}}$ si possa assumere di lavorare nel limite termodinamico.

Un'altra condizione iniziale da analizzare è quella discussa nella sottosezione precedente, ossia verificare come cambino i risultati della simulazione nel caso in cui il vettore $\boldsymbol{\sigma}_{\text{in}}$ sia inizializzato ad $\boldsymbol{h}$ oppure a una configurazione casuale. Lo scopo delle prime simulazioni consiste pertanto nello stimare l'ordine di grandezza di N_{lim} e nel verificare quest'ultima condizione.

Prima di far questo, dobbiamo definire una funzione che stimi quanto il messaggio di output $\hat{\boldsymbol{\xi}}$ coincida con il messaggio sorgente $\boldsymbol{\xi}$. Quest'ultima è proprio la *funzione di overlap* m che abbiamo definito in (4.12). Per uniformare i risultati di questo elaborato alla letteratura, definiamo m come:

$$m = m(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}) := \frac{1}{N} \sum_{i=1}^{N} \xi_i \hat{\xi}_i. \tag{5.10}$$

L'overlap m così concepito restituisce un valore compreso tra -1 e 1: tanto più m tende a -1, tanto più $\hat{\boldsymbol{\xi}}$ differisce da $\boldsymbol{\xi}$ (nel limite in cui $m = -1$ i due vettori diferiscono in ogni componente). Viceversa, tanto più m tende a 1, tanto più $\hat{\boldsymbol{\xi}}$ si può sovrapporre a $\boldsymbol{\xi}$ (nel limite in cui $m = 1$ i due vettori si sovrappongono in ogni componente).

5.2.3 Prestazioni della simulazione al variare del paramentro N

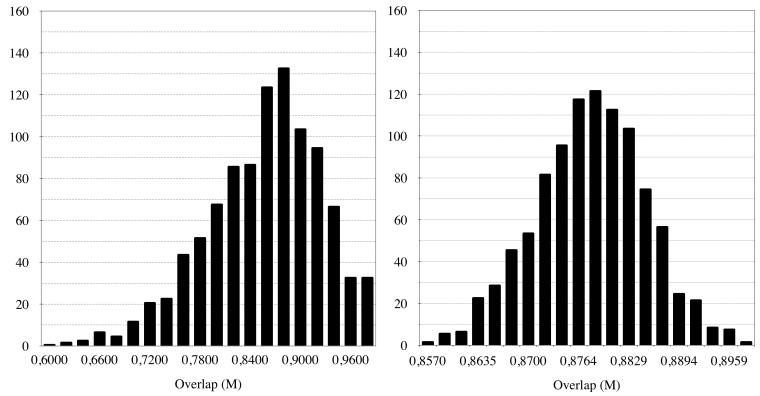
La prima delle condizioni al contorno che andiamo a studiare è quella relativa alla dimensione del vettore sorgente $\boldsymbol{\xi}$. Ricordiamo che il teorema di Shannon prevede che $N \to \infty$ per una trasmissione senza errori, da cui la necessità di stimare l'ordine di grandezza di N per assumere che il sistema si trovi nel limite termodinamico.[2]

La Figura 5.3 mostra i risultati della simulazione numerica effettuata con tre diversi valori di N per un grafo con connettività a due corpi (cioè $K = 2$) e con numero di interazioni per ogni spin pari a $N_{\text{int}} = 2$. Prima di analizzare i risultati, sottolineamo un aspetto importante: ogni istogramma raccoglie i dati di mille simulazioni, dove in ognuna di esse:

- è stato generato un messaggio $\boldsymbol{\xi}$ di N bit;
- è stata generata l'hamiltoniana H (5.1), supponendo $p = 0.1$;
- si è trovato il vettore $\hat{\boldsymbol{\xi}}$ corrispondente ground state di H;
- si è calcolato il corrispettivo parametro di overlap m tramite la (5.10).

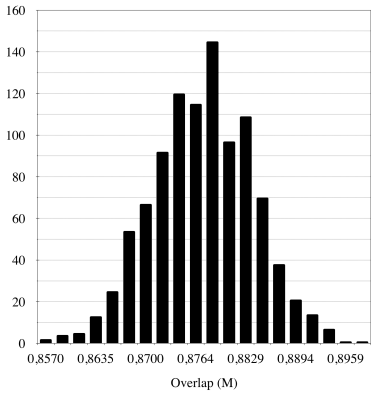
Dai mille valori di overlap ne è stata poi estrapolata la media $\bar{m}$ e la varianza σ_m^2. Andiamo dunque ad analizzare i risultati ottenuti dalla simulazione, riportati di seguito:

[2] I risultati che ora presenteremo sono stati ottenuti inizializzando $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$. La giustificazione di questa scelta (sebbene già ampiamente discussa) troverà la sua conferma nella sottosezione seguente.



(a) $N = 100$

(b) $N = 1000$



(c) $N = 10000$

Figura 5.3: Confronto tra tre distribuzioni di overlap nei casi in cui $N = 100$, $N = 1000$ e $N = 10000$. I valori medi corrispondono rispettivamente a 0.8760, 0.8784 e 0.8790: questo giustifica come la convergenza sia quasi totalmente saturata già per $N = 100$, sebbene si dovrebbe avere con $N \to \infty$ (parametri: $N_{int} = 2$, $K = 2$ e $p = 0.1$).

N	p	Media $\bar{m}$	Varianza σ_m^2
100	0.1	0.8760	$4.8 \cdot 10^{-3}$
1000	0.1	0.8784	$4.0 \cdot 10^{-4}$
10000	0.1	0.8790	$4.3 \cdot 10^{-5}$

Si evince chiaramente dagli istogrammi che la distribuzione degli overlap sia di tipo gaussiana. I dati mostrano che il valor medio $\bar{m}$ rimane invariato fino alla terza cifra decimale $\forall N$, mentre per osservare dei cambiamenti tra $N = 1000$ e $N = 10000$ bisogna andare fino alla quarta cifra decimale. Lo stesso non si può dire della varianza: se N cresce di un fattore 10, la varianza diminuisce dello stesso fattore. Questi risultati suggeriscono che l'overlap sia una grandezza *self-averaging*, di cui diamo di seguito la definizione:

Definizione 5.2.1. *Sia X una proprietà fisica di un sistema. Se tale sistema può essere interamente descritto dalla media $\bar{X}$ (intesa come media sui campioni) ed è tale per cui la varianza relativa $R_X = \sigma_X^2/\bar{X} \to 0$ per $N \to \infty$ (dove N è la taglia del sistema e σ_X^2 la sua varianza), allora X è una grandezza self-averaging.*

Se chiamiamo $\{\bar{m}_N\}$ la successione dei valori medi di ogni set di dati dipendente da N, ci aspetiamo che $\bar{m}_N \to \bar{m}_\infty$ per $N \to \infty$, dove $\bar{m}_\infty \sim 0.879$. Questo dimostra che, sebbene per raggiungere il limite termodinamico sarebbe necessario operare con $N \to \infty$, già per $N = 100$ la successione $\{\bar{m}_N\}$ ha saturato la convergenza quasi del tutto. Al contrario, come già detto un aumento di N induce una diminuzione della varianza σ_m^2 e non è difficile convincersi del fatto che, definendo in modo analogo una successione $\{\sigma_{m_N}^2\}$, si abbia $\sigma_{m_N}^2 \to 0$ per $N \to \infty$. Da ciò consegue che la varianza relativa $R_{m_N} = \sigma_{m_N}^2/\bar{m}_N \to 0$ per $N \to \infty$: allora l'overlap è una grandezza self-averaging.

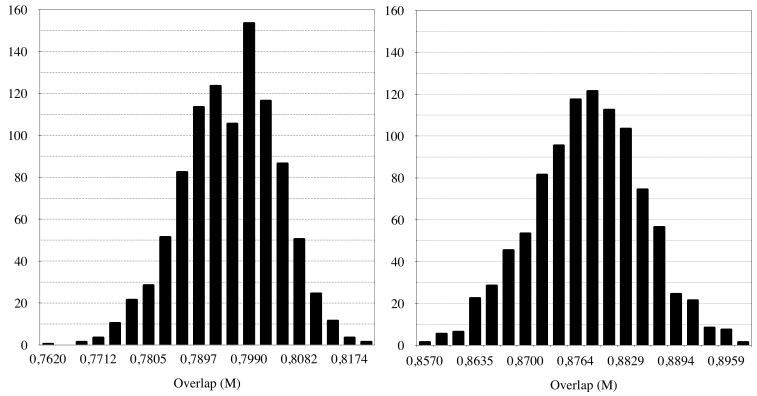
Dal momento che a noi interessa il solo valore medio $\bar{m}$ del set di 1000 dati, sarebbe sufficiente operare con $N = 100$ per avere una buona certezza che il sistema si trovi nel limite termodinamico. Tuttavia, richiedendo anche un controllo sull'errore commesso (fornito dalla varianza σ_m^2), è più opportuno scegliere come valore di lavoro $N = 1000$ o (ancora meglio) $N = 10000$.

Queste simulazioni preliminari sono state eseguite, come già detto, con $K = 2$ e $N_{\text{int}} = 2$. É possibile estendere tale discorso anche per ordini di K e di N_{int} maggiori, ma non riportiamo qui i risultati in quanto poco istruttivi e del tutto analoghi a quelli appena discussi. Un'ultima osservazione: come appena dimostrato, la scelta di $N = 1000$ o $N = 10000$ non influisce in alcun modo sul valor medio della distribuzione di overlap, quindi (a meno di un miglioramento sul controllo dell'errore) scegliere N di ordine maggiore non ha apparentemente alcun vantaggio. In realtà il problema è più sottile. In Figura 5.3 l'istogramma con $N = 100$ mostra un andamento gaussiano tagliato sul lato destro: questo comportamento è dovuto al fatto che, in più casi, l'algoritmo è riuscito a trovare il vero minimo dell'hamiltoniana. Poichè al crescere di N questo fenomeno non si ripresenta, la conclusione da trarre è una soltanto: $N = 100$ non è una buona approssimazione del limite termodinamico. Scegliendo valori dei parametri K e N_{int} maggiori di 2, il comportamento assunto in questo caso da $N = 100$ può essere riscontrato anche per $N = \mathcal{O}(1000)$. Tutto questo per dire che la scelta del parametro N dipende dai valori di K e N_{int} e verrà pertanto scelta ad hoc di volta in volta. In conclusione, possiamo affermare che in generale il limite termodinamico viene raggiunto per $N = \mathcal{O}(1000) - \mathcal{O}(10000)$ nel range di valori di K qui analizzati.

5.2.4 Confronto tra l'inizializzazione $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{\sigma}_{\text{ran}}$ e $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$

Un'altra condizione al contorno importante da verificare riguarda la scelta del vettore $\boldsymbol{\sigma}_{\text{in}}$ per applicare il metodo di discesa del gradiente. Nella sottosezione 5.1.1 abbiamo infatti esposto la problematica relativa all'inizzializzazione di $\boldsymbol{\sigma}_{\text{in}}$ a un vettore casuale (che identifichiamo con $\boldsymbol{\sigma}_{\text{ran}}$, dove *ran* sta per *random*).

Come già detto in precedenza, in generale non è noto dove si trovi la configurazione $\hat{\boldsymbol{\xi}}$ corrisponente al minimo dell'energia H. Tuttavia, affinchè il metodo del gradiente arrivi è convergenza è necessario che $\boldsymbol{\sigma}_{\text{in}}$ si trovi nel bacino di attrazione in cui il minimo è proprio $\hat{\boldsymbol{\xi}}$. Non avendo ulteriori informazioni, l'unica conclusione ovvia che possiamo trarre è la seguente: tanto più si riesce a scelgliere $\boldsymbol{\sigma}_{\text{in}}$ "vici-



(a) $\boldsymbol{\sigma}_{in} = \boldsymbol{\sigma}_{ran}$ (b) $\boldsymbol{\sigma}_{in} = \boldsymbol{h}$

Figura 5.4: Confronto tra le inizializzazioni $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{\sigma}_{\text{ran}}$ e $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$ rispetto alla distribuzione dei valori di overlap. Nel primo grafico si riscontra un valor medio pari a $\bar{m} = 0.7936$, mentre nel secondo $\bar{m} = 0.8784$: questo risultato è una conferma delle ipotesi fatte in precedenza, ossia un miglioramento delle prestazioni assumendo proprio $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$. (Parametri: $N = 1000$, $N_{int} = 2$, $K = 2$ e $p = 0.1$).

no"[3] a $\hat{\boldsymbol{\xi}}$, tanto maggiore è la probabilità che $\boldsymbol{\sigma}_{\text{in}}$ si trovi nella valle di $\hat{\boldsymbol{\xi}}$. Tutto ciò suggerisce che l'iniziallizzazione $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{\sigma}_{\text{ran}}$ sia in genere poco performante.

Avendo inviato il messaggio sorgente $\boldsymbol{\xi}$ nel canale di trasmissione, quest'utlimo è stato corrotto e ha dato origine a un vettore che contiene un certo numero di bit scambiati, ossia $\boldsymbol{h}$ (nel limite termodinamico il numero di bit scambiati è $p \cdot N$, dove p è l'ormai noto errore sul singolo bit introdotto dal canale). In generale, ci aspettiamo che sia più probabile che $\boldsymbol{h}$ si trovi nella valle di $\hat{\boldsymbol{\xi}}$ rispetto che $\boldsymbol{\sigma}_{\text{ran}}$, da cui l'idea di inizializzare $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$.

Quest'ipotesi è confermata dai risultati delle simulazioni mostrati in Figura 5.4, in cui sono state eseguite mille misure per ciascuno dei due casi in esame (i dettagli dei passaggi della simulazioni sono gli stessi della sottosezione precedente) e son stati poi costruiti due istogrammi che mostrano la distribuzione degli overlap. Di seguito riportiamo i dati numerici:

$\boldsymbol{\sigma}_{\text{in}}$	p	N	Media $\bar{m}$	Varianza σ_m^2
$\boldsymbol{\sigma}_{\text{ran}}$	0.1	1000	0.7936	$1.4 \cdot 10^{-3}$
$\boldsymbol{h}$	0.1	1000	0.8784	$4.0 \cdot 10^{-4}$

ottenuti da un grafo con $K = 2$ e $N_{\text{int}} = 2$. Come si evince dalla tabella di cui sopra, c'è una differenza sostanziale tra le due prestazioni: i risultati con l'inizializzazione $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$ sono superiori sia in termini di overlap $\bar{m}$ sia in termini di varianza σ_m^2.

Un utimo spunto di riflessione è il seguente: grazie a questi risultati possiamo ulteriormente giustificare la definizione di H in (5.1). Infatti, in generale l'hamiltoniana dei codici di Sourlas non presenta il termine dipendente dal campo esterno $\boldsymbol{h}$, ossia $-\sum h_l \sigma_l$ (cfr. [21]). Tuttavia, dal momento che stiamo facendo uso del metodo di discesa del gradiente per il calcolo di H_{min}, senza l'intervento del campo

[3] Ricordiamo che $\boldsymbol{\sigma}$ è un vettore di $\mathbb{Z}^N$, su cui è perfettamente definita una funzione di distanza.

esterno non avremmo alcuna informazione riguardo a come inizializzare $\boldsymbol{\sigma}_{\text{in}}$. La conseguente scelta di $\boldsymbol{\sigma}_{\text{ran}}$ come dato iniziale causerebbe, come appena dimostrato, un peggioramento delle prestazioni.

5.2.5 Costruzione del factor graph

Fino a questo momento abbiamo definito l'hamiltoniana dei codici di Sourlas (5.1) e ne abbiamo fornito un'interpretazione grafica tramite i factor graph (cfr. sottosezione 5.1.2). Non abbiamo però affrontato una questione rilevante, ossia come costruire un factor graph casuale secondo le restrizioni imposte, che ricordiamo essere:

- $M \cdot K = N \cdot N_{\text{int}}$,
- $M > N_{\text{int}}$.

Abbiamo già verificato che un factor graph con queste caratteristiche esiste e rimane ben definito: rimane da capire quale algoritmo possa generarlo.

La strategia che seguiremo per creare un factor graph del tutto casuale consiste fondamentalmente in due step:

- generiamo un particolare factor graph facilmente definibile tramite un algoritmo;
- eseguiamo una permutazione casuale dei lati di questo factor graph.

Riportiamo di seguito i passaggi fondamentali per la costruzione di un factor graph generico. Supponiamo fissati i parametri N, N_{int}, M e K. Quindi:

(a) colleghiamo il primo spin σ_1 con le prime N_{int} interazioni (quadratini bianchi). Dal momento che $M > N_{\text{int}}$, non può mai capitare che σ_1 venga collegato due volte con la stessa interazione.

(b) In modo analogo, colleghiamo il secondo spin σ_2 con le successive N_{int} interazioni. Se si giunge all'ultima interazione $\boldsymbol{\mu}_M$, semplicemente si ricomicia a collegare lo spin con la prima di queste, e così via.

(c) Si itera questa operazione per ogni spin σ_i. La prima delle due condizioni che abbiamo posto nella creazione del grafo non garantisce che questa procedura sia sempre fattibile. Tuttavia, una volta definiti K e N_{int} in modo arbitario, ponendo semplicemente N uguale a un multimplo intero di K e $M = N \cdot N_{\text{int}}/K$ ci si convince facilmente che la costruzione di questo grafo è sempre possibile.

(d) sSegliamo casualmente due spin diversi σ_i e σ_j: siano $\boldsymbol{\mu}_a \neq \boldsymbol{\mu}_b$ tali che σ_i sia collegato con $\boldsymbol{\mu}_a$ e σ_j con $\boldsymbol{\mu}_b$. Supponiamo inoltre che σ_i non sia collegato con $\boldsymbol{\mu}_b$ e idem per σ_j con $\boldsymbol{\mu}_a$. Scambiamo dunque questi due lati: colleghiamo σ_i con $\boldsymbol{\mu}_b$ e σ_j con $\boldsymbol{\mu}_a$.

(e) Iteriamo il passaggio precedente, dando luogo a una permutazione del grafo di partenza.

La Figura 5.5 mostra passo passo le fasi di costruzione del grafo.

Rimane soltanto da definire quanti scambi tra due lati si debbano eseguire per far sì che la permutazione del grafo iniziale dia origine a un grafo totalmente casuale. Ricordiamo che, per come viene eseguita questa simulazione, si ha che $N_{\text{int}} \ll N$. Da questo segue che il numero di lati, che ricordiamo essere $N \cdot N_{\text{int}}$, è tale per cui $N \cdot N_{\text{int}} \ll N^2$. Se assumiamo che la permutazione totale abbia esattamente N^2 scambi, nel limite termodinamico ogni lato del grafo viene scambiato almeno una volta con un altro lato (in realtà, da un punto di vista statistico ogni lato viene scambiato con un altro un numero di volte $\gg 1$). Questo garantisce che il grafo finale sia totalmente casuale.

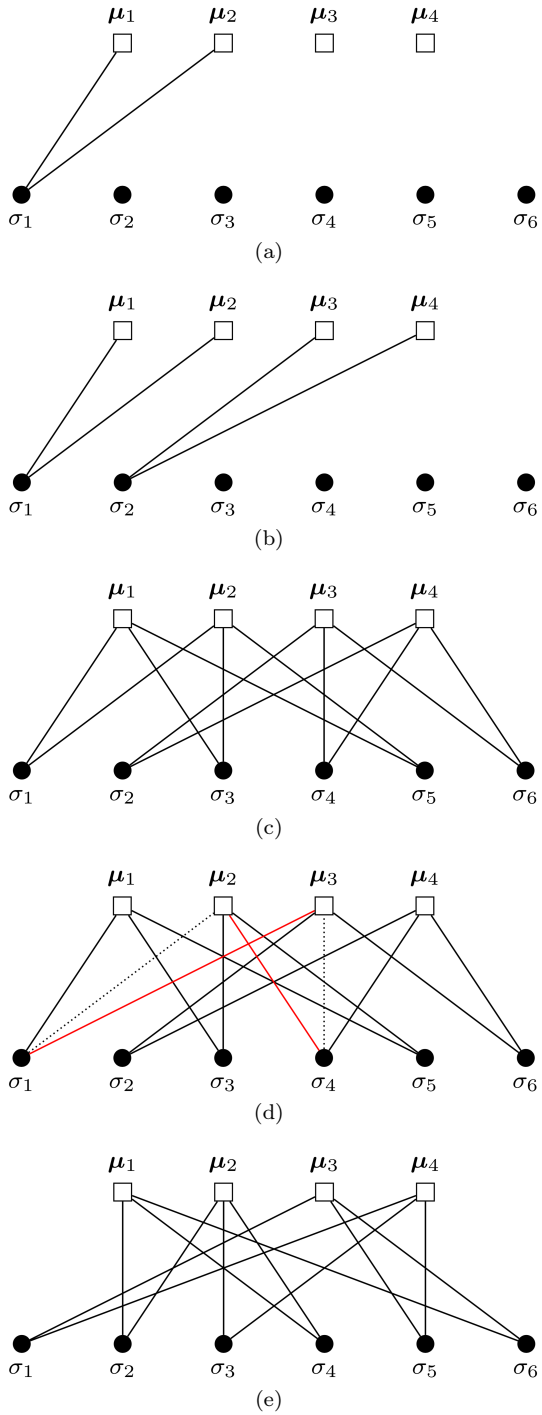


Figura 5.5: Step di costruzione di un grafo casuale con $N = 6$, $M = 4$, $K = 3$ e $N_{\text{int}} = 2$. Nella sottosezione 5.2.5 sono riportate le descrizioni dei passaggi fondamentali per la costruzione di questa tipologia di factor graph.

Parte III

STUDIO DELLE PRESTAZIONI DEI CODICI DI SOURLAS A CONNETTIVITÀ FINITA

Capitolo 6

Prestazioni dei codici di Sourlas a connettività finita

Fino a questo momento abbiamo presentato alcune generalità riguardo alla meccanica statistica dei codici di Sourlas e un'analisi delle condizioni al contorno della simulazione di quest'ultimi (ossia lo studio dei parametri da cui dipende l'hamiltoniana (5.1) di tali codici). Ci sono alcune domande fondamentali cui vorremmo dare una risposta. Supponiamo di fissare le condizioni al contorno (N, K, ecc.) Le prime domande che sorgono spontanee sono per esempio: esiste un grafo privilegiato, tale per cui le sue prestazioni siano migliori rispetto a quelle di qualunque altro grafo? Esiste un K ottimale oppure la connettività infinita è l'unica che soddisfa il limite di Shannon? Cosa accadrebbe se si potesse conoscere con assoluta certezza il minimo dell'hamiltoniana (5.1)? Un'idea delle risposte a tali domande sarà l'oggetto principale di questo capitolo.

6.1 Grafo random come soluzione subottimale

La prima serie di domande che esige una risposta approfondita è la seguente: esite un grafo ottimale? É effettivamente possibile trovarlo nell'insieme $\mathcal{G}$ di tutti i grafi costruibili con le nostre condizioni al contorno?[1] E infine, anche assumendo che ciò sia fattibile, le sue prestazioni di quanto sono superiori rispetto a quelle di un grafo generico?

Nello studio dei parametri della simulazione abbiamo sempre imposto che il factor graph fosse di tipo *random*, ossia generato in modo casuale. Tuttavia, questa assunzione potrebbe a priori inficiare i risultati finali. Non dimentichiamo che per noi un factor graph altro non è che la rappresentazione grafica dell'hamiltoniana (5.1), di cui dobbiamo poterne calcolare il minimo per recuperare il messaggio sorgente. Si potrebbe dunque pensare che una particolare geometria del grafo si traduca analiticamente in un'hamiltoniana H di cui sia più semplice calcolarne il ground state $\hat{\boldsymbol{\xi}}$.

Proviamo a supporre che esista effettivamente un factor graph la cui geometria restituisca un overlap maggiore (il parametro K è ora fissato). Ricordando la pro-

[1] Da qui in avanti, con $\mathcal{G}$ indichiamo l'insieme dei grafi ottenibili secondo le restrizioni che sono state descritte nel capitolo precedente. Nell'insieme $\mathcal{G}$ si suppone siano fissati i parametri relativi alle condizioni al contorno, ossia $\mathcal{G} = \mathcal{G}(N, N_{\text{int}}, K, M)$.

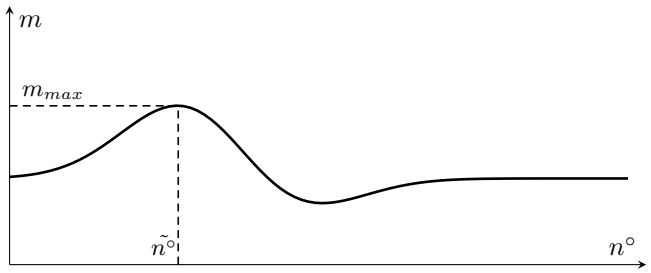


Figura 6.1: Ipotetico andamento dell'overlap m in funzione del numero di permutazione dei lati n°, ovvero $m(n^\circ)$, nell'ipotesi in cui l'andamento di m sia oscillante prima del processo di termalizzazione. Supponendo che la configurazione iniziale sia quella argomentata nella sottosezione 5.2.5, eseguendo un certo numero di scambi di lati si dovrebbe osservare un comportamento come quello qui mostrato: l'overlap raggiunge un valore massimo in corrispondeza di $\tilde{n^\circ}$, cui corrisponde uno specifico factor graph (e quindi una specifica hamiltoniana della forma (5.1)). Un comportamento come questo potrebbe giustificare l'esistenza di geometrie preferibili a quella random.

cedura mostrata nella sottosezione 5.2.5 per ottenere un particolare tipo di factor graph totalmente casuale, per ogni permutazione di una coppia di lati calcoliamo il corrispettivo valore di overlap m. In questo modo, a partire da un particolare factor graph, possiamo osservare come vari m in funzione del numero di scambi effettuati. Chiamiamo n° tale variabile, così che resti ben definita la funzione $m(n^\circ)$. É ragionevole aspettarsi che per un numero di scambi sufficentemente grande m raggiunga una condizione di equilibrio, essendo una grandezza self averaging. Tuttavia, è meno prevedibile il comportamento dell'overlap prima della stabilizzazione intorno al proprio valore medio. Sarebbe infatti possibile sia un andamento oscillante sia un crescita monotona. Il primo è più compatibile con l'ipotesi di un grafo ottimale: si potrebbe osservare un massimo di $m(n^\circ)$ in corrispondenza di uno specifico $\tilde{n^\circ}$, come mostrato in Figura 6.1. Il grafo corrispondente avrebbe prestazioni superiori rispetto a tutti gli altri analizzati e questo sarebbe il sentore dell'esistenza di geometrie speciali.

É fondamentale sottilineare questo aspetto: le simulazioni qui proposte sono di tipo *Monte Carlo*. Lo studio analitico di un'hamiltoniana della forma (5.1) è infatti un problema tutt'altro che banale da risolvere e in questo i metodi computazionali forniscono un valido strumento alternativo per analizzarlo sotto varie sfaccettature. Ciò significa che l'obiettivo della simulazione numerica di cui sopra non consiste nel calcolare l'overlap per ogni grafo di $\mathcal{G}$, dal momento che la dimensione di questo insieme è una funzione esponenziale di N e dunque non sondabile tramite metodi computazionali nella sua interezza.[2] Al contrario, si studia l'overlap in funzione di un numero di scambi sufficientemente grande da poter essere considerato rappresentativo di tutto l'insieme $\mathcal{G}$. Sotto quest'ottica, se un comportamento come quello mostrato in Figura 6.1 non si manifesta nel sottoinsieme di $\mathcal{G}$ da noi son-

[2]In genere la prestazione migliore che si intende raggiungere con un algoritmo è quella per cui il suo andamento (in questo caso parliamo in termini di tempo) sia $\mathcal{O}(\log N)$ o al più $\mathcal{O}(N)$. Già con un andamento del tipo $\mathcal{O}(N^\alpha)$, dove $\alpha > 1$, si riscontra un notevole rallentamento della procedura. Se addirittua il tempo dell'algoritmo va come $\mathcal{O}(\mathrm{e}^N)$, esso risulta praticamente inutilizzabile al crescere della taglia N del sistema.

dato, in genere si dovrebbe concludere con ragionevole certezza che esso non si manifesterebbe neppure sotto l'ipotesi di estensione del dominio della simulazione a tutto $\mathcal{G}$. Tuttavia, il sottoinsieme sondabile tramite questa simulazione Monte Carlo contiene un numero di grafi dell'ordine di $\mathcal{O}(N^\alpha)$, dove α è un numero intero non molto maggiore dell'unità (difficilmente $\alpha > 3$ o 4). Questo significa che i risultati ottenuti andranno analizzati con cautela e non avranno la pretesa di verità assoluta. Le Figure 6.2 e 6.3 mostrano la mappatura dell'overlap m in funzione del numero di scambi n° effettuati. Riportiamo di seguito i parametri utilizzati per questa simulazione:

p	N	K	n°
0.1	1000	2	$5 \cdot 10^3$
0.1	1500	3	$12 \cdot 10^3$
0.1	4000	4	$50 \cdot 10^3$

Nello specifico, fissato il factor graph di partenza descritto nella sottosezione 5.2.5, per ogni permutazione di una coppia di lati è stato calcolato il corrispettivo valore di overlap m (ottenuto come media di un set di mille dati). Sono state mappate cinque curve nei due grafici di Figura 6.2 e quattro in quello di Figura 6.3 (in cui sono stati fissati il parametro K e la dimensione del messagio sorgente N), ciascuna corrispondente a uno specifico valore di rate R. Scelto K, infatti, è possibile variare il rate di trasmissione modificando il parametro N_{in}, che ricordiamo corrispondere al numero di interazioni di ogni spin, secondo l'equazione $R = K/(K + N_{\text{int}})$.

Occupiamoci per il momento del grafico $K = 2$ (i casi $K = 3$ e $K = 4$ sono totalmente analoghi). Mostriamo di seguito i risultati fonadamentali cui siamo pervenuti.

- Ogni curva mostra un andamento monotono crescente[3] indipendentemente dal parametro di rate R. Non si riscontrano massimi o minimimi locali prima della fase di *termalizzazione* (ossia di convergenza di m a una condizione di equilibrio), dunque l'ipotesi di un grafo ottimale (cfr. Figura 6.1) è, fino a prova contraria, non giustificata.

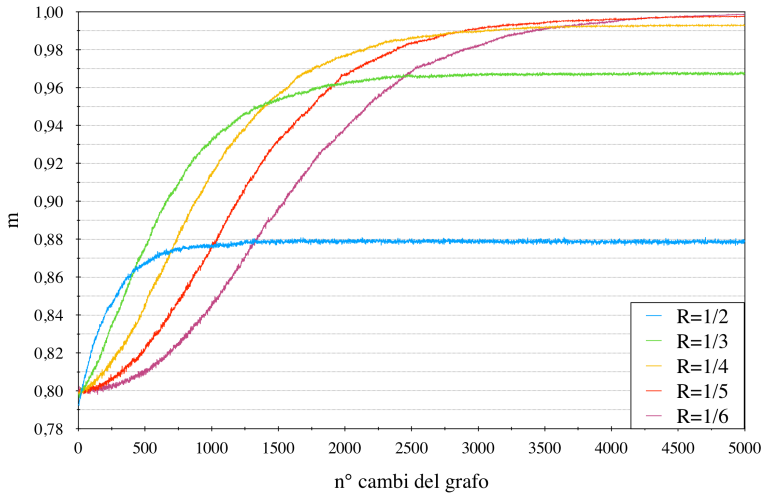
- Al crescere di n°, ogni curva converge a un valore costante c_R che dipende da R. Tanto più R si avvicina a 0, tanto più c_R si avvicina al valore 1. Una domanda interessante cui rispondere è se $c_R \to 1$ per $R \to 0$ oppure se $c_R < 1$ definitivamente.

- Il processo di termalizzazione ($m \to c_R$) dipende da R: diminuendo il rate, aumenta il numero di scambi n° necessari per la termalizzazione dell'overlap.

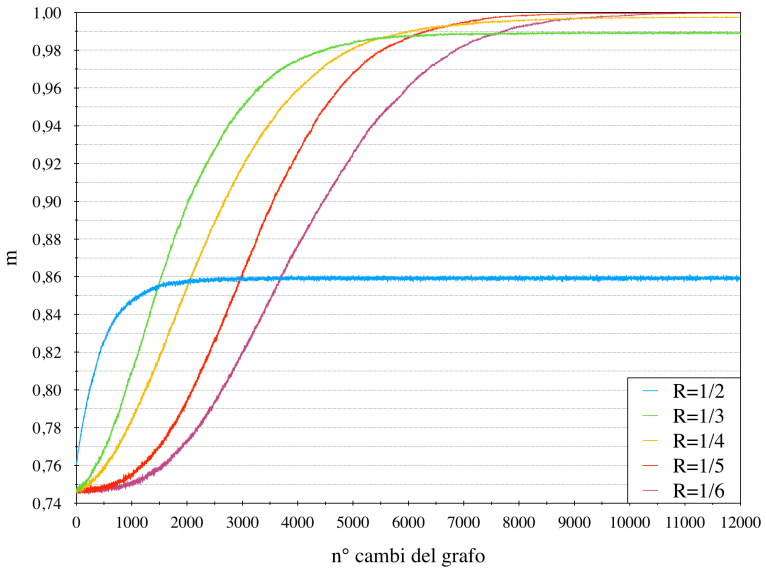
I grafici con $K = 3$ e $K = 4$ sono sostanzialmente analoghi, salvo per il fatto che la termalizzazione di m richiede un numero di scambi n° maggiore (da notare, però, che per $K = 3$ e $K = 4$ è stato utilizzato un parametro N maggiore).

La risosta alla domanda relativa all'esistenza o meno di un factor graph ottimale è dunque delicata. La crescita dell'overlap è monotona e in nessun caso si riscontrano valli di massimo o di minimo prima del processo di termalizzazine, come già

[3] In realtà, come si può notare ogni curva mostra un andamento frastagliato, dovuto a piccole oscillazioni locali. Tuttavia, mediando le oscillazioni si possono ottenere curve lisce monotone crescenti.



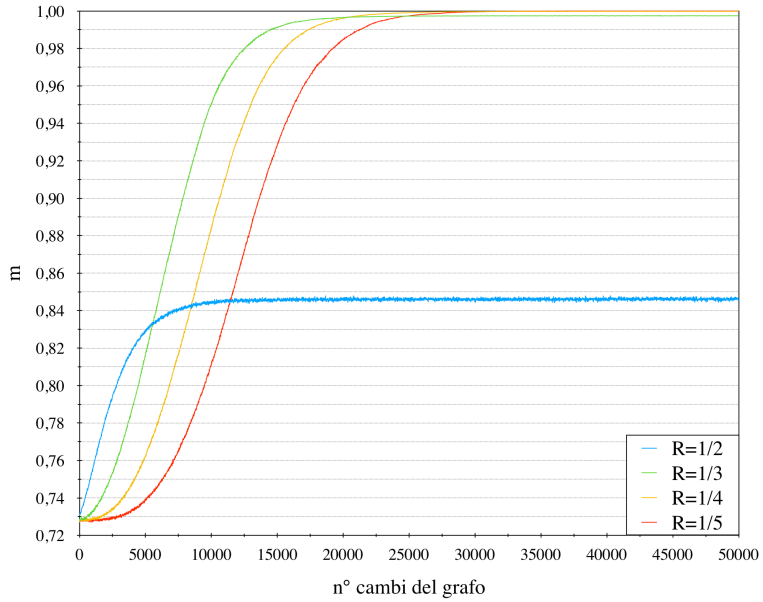
(a) $K = 2$



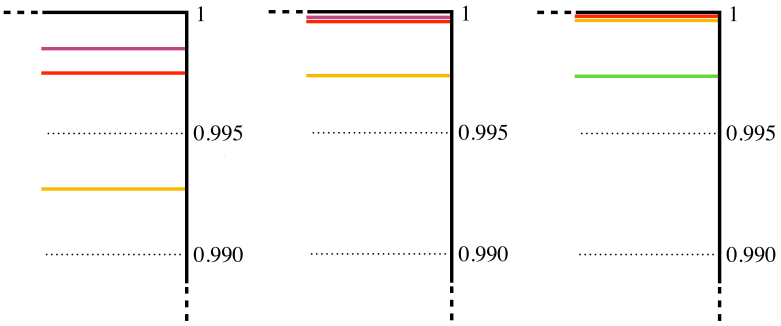
(b) $K = 3$

Figura 6.2: Andamento dell'overlap m in funzione del numero di scambi $n°$ dei lati del grafo rispettivamente nei casi con $K = 2$ e $K = 3$. Il grafo di partenza (che corrisponde al valore $n° = 0$) è quello descritto nella sottosezione 5.2.5: per ogni permutazine di una coppia di lati è stato calcolato il corrispettivo valore di overlap (media di un set di mille valori). Sono state mappate cinque curve di overlap per ognuno dei due grafici, corrispondenti a cinque valori di rate R diversi. Fissati i parametri N e K, è stato variato il rate modificando il numero di interazioni N_{int} di ogni spin, ricordando che $R = K/(K + N_{\text{int}})$.

accennato poco sopra. Ciò significa che in genere non abbiamo nessuna prova del fatto che esista un grafo la cui geometria sia privilegiata rispetto a quella degli altri. Ma il fatto di non averne alcuna prova non implica la non esistenza: questa simula-



(a) $K = 4$



(b) *Macro:* $K = 2$ (c) *Macro:* $K = 3$ (d) *Macro:* $K = 4$

Figura 6.3: La prima delle quattro figure mostra l'andamento dell'overlap m in funzione del numero di scambi $n°$ dei lati del grafo nel caso $K = 4$. Per le specifiche si rimanda alla Figura 6.2. Le successive tre sono semplicemente una macro del grafico precedente e di quelli di Figura 6.2 per facilitarne la lettura.

zione suggerisce come *probabilmente* non esista un grafo ottimale, a meno che non si riesca a sondare lo spazio dei grafi con qualche procedura più efficiente. Anche su quest'ultima cosa c'è però da far chiarezza. Abbiamo già sottolineato come l'algoritmo di discesa del gradiente abbia i suoi limiti nella ricerca del minimo di una funzione a più variabili e di come alcuni metodi più avanzati permetterebbero di trovare risultati ancora più validi. Lo stesso però non si può dire per la ricerca del factor graph ottimale, in quanto lo scoglio più grande da superare rimane sempre la dimensione dello spazio dei grafi che tende ad infinito al crescere di N con un andamento esponenziale.

Sotto quest'ottica, un factor graph random costituisce una soluzione *subottimale* (cioè che si approssima molto alla soluzione ottimale). Ci aspettiamo infatti che,

una volta fissati i parametri K e N_{int}, possa esistere un grafo con prostazioni sì superiori rispetto a tutti gli altri, ma assolutamente confrontabili con quelle di un grafo generico, al punto da poter considerare la geometria random come subottimale.

6.2 Mappatura dell'overlap in funzione di p al variare del parametro K

Uno studio dei SC con $K < \infty$ è già stato affrontato nello specifico da Y. Kabashima e D. Saad in [8], [9] e [16], sebbene con alcune sostanziali differenze. In particolare, in [9] sono stati studiati i SC a partire dall'hamiltoniana

$$H_{\text{KS}}(\boldsymbol{\sigma}) = -\sum_{i_1<\cdots<i_K}^{N} J_{i_1\cdots i_K}\sigma_{i_1}\cdots\sigma_{i_K} - F\sum_{l}^{N}\sigma_l, \tag{6.1}$$

dove $J_{i_1\cdots i_K}$ è definito come in (5.2). La Figura 6.4 mostra un confronto tra i risultati di una simulazione di Kabashima e Saad (KS) e una eseguita con l'algoritmo presentato in questo elaborato, in cui sono state mappate le curve $m^{(K)}$[4] in funzione di p al variare del parametro $K \in [2, 6]$. Nell'hamiltoniana H_{KS} è stato posto $F = 0$, che corrisponde ad avere campo esterno nullo.

Un raffronto tra le Figure 6.4a e 6.4b mette subito in luce questo fatto: per piccoli p e $K > 2$, il comportamento delle curve $m^{(K)}$ è lo stesso per entrambe le simulazioni, mentre al crescere di p quelle della seconda figura decrescono più velocemente. In realtà, per $K > 2$ le curve del grafico 6.4a mostrano le soluzioni teoriche e non quelle direttamente misurate con le simulazioni. In particolare, per $K = 5$ sono sì state fatte alcune misure, ma con condizioni iniziali vicine ai limiti previsti dalla teoria. Le differenze tra la nostra simulazione e quella in [9] non stupiscono. Verificheremo che sono perfettamente in linea con le argomentazioni proposte nella sezione 6.4, in cui giustificheremo perchè non si osservi la convergenza $p_b^{(k)} \to \Gamma_{\text{S}}$ per $K \to \infty$ nel limite termodinamico (con Γ_{S} indichiamo la curva limite di Shannon).

Discorso diverso vale invece per $K = 2$: anche nell'articolo originale [9] le condizioni iniziali non sono state fissate arbitrariamente vicine ai dati teorici. Questo permette un paragone più interessante, proposto in Figura 6.4c. Vediamo di analizzarne gli aspetti fondamentali.

- ppr $p < 0.05$ c'è una buona compatibilità tra le due curve, che restituiscono sostanzialmente gli stessi valori di overlap.
- Per p compreso tra 0.05 e 0.1, la curva di KS ha prestazioni di overalp superiori, anche se non si registrano differenze sostanziali. Possiamo dire che, salvo alcune oscillazioni, il comportamento delle due curve fino a circa $p = 0.1$ è fondamentalmente lo stesso.
- Al cresce di p oltre 0.1, la curva di KS ha una così forte decrescita a 0 che per $p > 0.18$ tende ad appiattirsi totalmente sull'asse delle ascisse. Questo comportamente non si riscontra invece nella curva della nostra simulazione, che per $p \in [0.1, 0.2]$ ha una decrescita quasi lineare.

[4] Da qui in avanti, con curve $m^{(K)}$ ci riferiamo alle curve di overlap mappete in funzione di una specifica variabile (p, R,$\cdots$) al variare del parametro K.

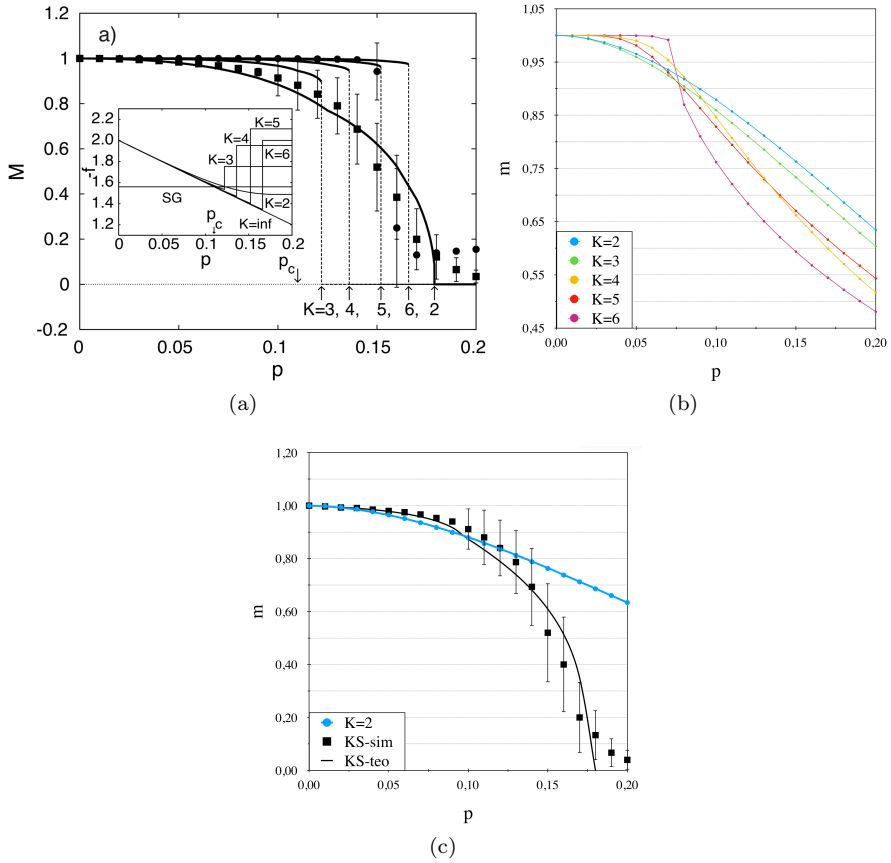


Figura 6.4: Confronto tra i risultati di una simulazione di Kabashima e Saad in [9] e una eseguita con l'algoritmo presentato in questo elaborato, in cui sono state mappate le curve $m^{(K)}$ in funzione di p al variare del parametro $K \in [2, 6]$. Nell'hamiltoniana H_{KS} (cfr. equazione (6.1)) è stato posto $F = 0$, che corrisponde ad avere campo esterno nullo. La prima e la seconda figura mostrano rispettivamente le simulazioni di KS e quelle eseguito con il nostro algoritmo. La terza, invece, è un raffronto tra le curve $m^{(2)}$ ottenute con l'hamiltoniana H_{KS} (in nero) e con la (5.1) (in azzurro).

L'unica differenza tra l'hamiltoniana H_{KS} e quella da noi definita in (5.1) sta nel fatto che la seconda possiede il termine di campo esterno $\boldsymbol{h}$. Come già argomentato nella sottosezione 5.2.4, da un lato la presenza di $\boldsymbol{h}$ costringe a minimizzare la somma di due componenti energetiche (cosa che può ulteriormente complicare la ricerca della configurazione del minimo dell'hamiltoniana), dall'altro è un valido strumento per tener traccia del messaggio di input. La Figura 6.5 mostra la differenza tra l'andamento del parametro p_b in funzione di K nel caso in cui si abbia l'inizializzazione $\boldsymbol{\sigma}_{\mathrm{in}} = \boldsymbol{h}$ (in nero) e $\boldsymbol{\sigma}_{\mathrm{in}} = \boldsymbol{\xi}$ (ossia il messaggio di input). Mentre la prima presenta un $K_{\min}$ oltre il quale le prestazioni del codici cominciano a peggiorare, la seconda ha una decrescita monotona. Questo significa che, nell'ipotesi di avere un algoritmo di decoding ancora più efficiente di quello qui implementato,

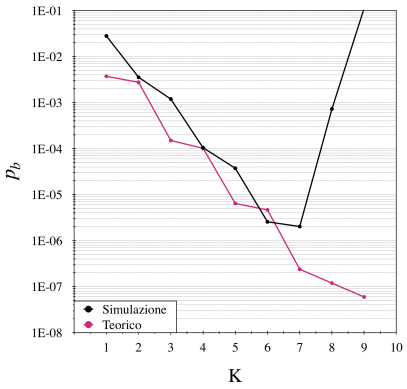


Figura 6.5: Confronto tra i risultati di una simulazione eseguita con l'algoritmo presentato in questo elaborato (in nero) e l'andamento teorico (in viola) della curva p_b mappata in funzione del numero di interazioni K (il rate è stato tenuto costante a $R = 1/4$). Nello specifico, la curva in nero è stata ottenuta con la sotilita inizializzazione $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$, mentre per quella in viola si è posto $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{\xi}$ (ossia il messaggio di input). Il fatto che nel secondo caso non si ottenga $p_b = 0 \ \forall K$ potrebbe sorprendere, ma in realtà è perfettamente in linea con quanto preannunciato nella sezione 3.2: l'errore introdotto dal canale fa sì che $\boldsymbol{\xi}$ non coincida perfettamente con il ground state dall'hamiltoniana (5.1).

per grandi K si potrebbero ottenere prestazioni dei SC ancora superiori. Inoltre, questi dati costituiscono un'ulteriore conferma del fatto che al crescere di K il bacino di attrazione del ground state si restringe, rendendo sempre più difficoltosa una sua ricerca. Un'ultima osservazione: potrebbe sorprendere il fatto di non ottenere $p_b = 0 \ \forall K$ per la curva viola, ma in realtà è perfettamente in linea con quanto preannunciato nella sezione 3.2: l'errore introdotto dal canale fa sì che $\boldsymbol{\xi}$ non coincida perfettamente con il ground state dall'hamiltoniana (5.1).

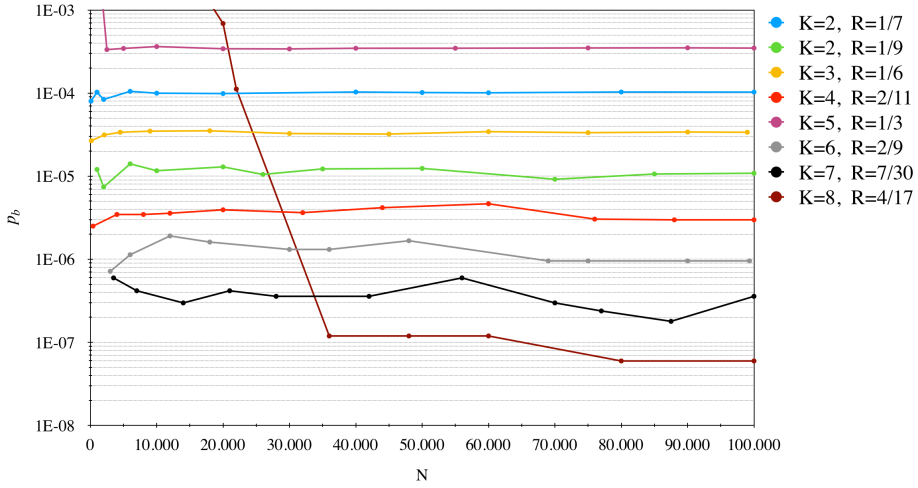
6.3 Dipendenza della termalizzazione di p_b dalla taglia del sistema

Abbiamo già affrontato la questione relativa alla taglia del sistema nella sottozezione 5.2.3, osservando come il limite termodinamico venga in genere raggiunto per $N = \mathcal{O}(1000) - \mathcal{O}(10000)$ (almeno per il numero di corpi K qui preso in considerazione). Abbiamo inoltre dato una dimostrazione qualitativa del fatto che l'overlap m sia una grandezza self-averaging.

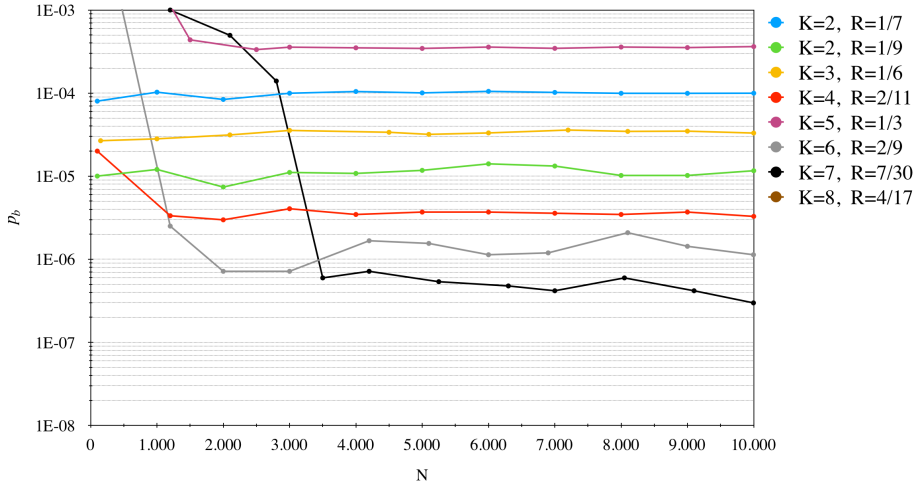
Vediamo come fornire ulteriori prove di questa proprietà dell'overlap. Ricordiamo che la probabilità di errore per ogni bit dopo la codifica del messaggio di output è definita come

$$p_b = 1 - \frac{1+m}{2}, \tag{6.2}$$

da cui si evince una corrispondenza biunivoca tra m e p_b (quindi anche quest'ultima è una grandezza self-averaging). Ci aspettiamo pertanto che mappando p_b in funzione di N si raggiunga una fase di termalizzazione, fenomeno comprovato dalla simulazione mostrata nella Figura 6.6. Come si può osservare, la velocità di terma-



(a)



(b)

Figura 6.6: Mappatura del parametro p_b (errore commesso sul singolo bit in seguito alla codifica del messaggio di output del canale) in funzione della dimensione N del sistema, eseguita al variare di K. Ogni curva mostra un processo di termalizzazione più o meno rapido a seconda del K considerato: tanto più cresce K, tanto maggiore è la dimensione minima che il sistema deve avere affinchè si possa assumere che questo si trovi nel limite termodinamico. La seconda figura è una macro della prima, con una mappatura eseguita sull'intervallo $[0, 10^3]$ al posto che $[0, 10^4]$.

lizzazione dipende fortemente dal numero K dei corpi interagenti: tanto più cresce K, tanto più aumenta la minima dimensione N del sistema tale per cui si possa assumere di lavorare nel limite termodinamico. Prima che il sistema raggiunga una condizione di equilibrio, il comportamento della curva $p_b(N)$ è in genere imprevedibile e mostra oscillazioni brusche. Una volta raggiunta la termalizzazione, invece, si riscontrano piccoli scostamenti dalla zona di equilibrio.

Il fatto che il parametro di overlap termalizzi indipendentemente da K o N_{int} è

quindi comprovato. Tuttavia, per completezza sottolineamo questo aspetto: la stima dell'errore commesso sulla misura di m nella sottosezione 5.2.3 è stata valutata tramite la varianza della distribuzione gaussiana dei mille valori calcolati. Nello specifico, quest'ultimi sono tutti stati ottenuti a partire da uno specifico grafo casuale. La varianza, dunque, tiene conto soltanto dell'errore commesso sul singolo grafo. Potrebbe infatti subire delle fluttuazioni a seconda del grafo su cui viene calcolato m e, pertanto, in generale costituisce una sottostima del reale errore cui è soggetta una misura come la nostra. Ciò nonostante, avendo precedentemente verificato che il grafo random è subottimale, da qui in avanti verrà trascurato questo aspetto e si assumerà la varianza come una buona stima dell'errore commesso.

Ricordiamo che la varianza è la misura di quanto i dati si discostano quadraticamente rispetto al valor atteso (che nel caso di una distribuzione gaussiana conincide con la media aritmetica). L'errore di quest'ultimo è invece dato dalla *deviazione standard della media* e si ricava a partire dalla varianza come

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma_X^2}{N}}. \tag{6.3}$$

6.4 Prestazioni dei codici di Sourlas a connettività finita

Nella prima sezione di questo capitolo ci siamo sostenzialmente occupati dell'eventuale esistenza (smentita fina a prova contraria) di eventuali geometrie speciali dello spazio $\mathcal{G}$ dei grafi qui presi in esame. A questo punto dobbiamo verificare le vere e proprie prestazioni dei codici di Sourlas a connettività finita. Di seguito mostreremo i risultati delle simulazioni di questi codici messe a confronto sia con i limiti teorici (previsti dal Teorema di Shannon) sia con le prestazioni di altri codici noti dalla letteratura.

6.4.1 Mappatura di p_b in funzione del rate al variare del parametro K

Mostriamo di seguito le prestazioni del codice fin qui presentato. Come già anticipato nella sezione precedente, l'overlap m e l'errore p_b sono in corrispondenza biunivoca secondo l'equazione (6.2). Di conseguenza lo studio di p_b permette di determinare le prestazioni del codice al pari dello studio di m: scegliamo il primo per uniformità alla letteratura preesistente.

In Figura 6.7 e 6.8 sono riportati i risultati delle simulazioni in cui sono state mappate le curve p_b in funzione di R al variare del parametro K. Nello specifico, lo studio di $p_b(R)$ è stato effettuato per tre valori di p (che ricordiamo essere la probabilità di errore sul singolo bit introdotta dal canale): 0.075, 0.10 e 0.16.

Esattamente come nelle simulazioni preliminari, ogni valore di p_b nel grafo è stato calcolato come media di un set di mille dati, mentre l'errore come deviazione standard della media di quest'ultimi (cfr. equazione (6.3)).[5] Nei grafici compaiono inoltre:

[5] Abbiamo già discusso il fatto che gli errori così ottenuti in genere tendono a sottostimare la reale incertezza sulle misure. Sono pertanto da intendersi come dati indicativi.

- la curva *limite di Shannon*, descritta dall'equazione

$$R = \frac{C}{1 - H_2(p_b)},$$

dove la capacità C è definita in (3.13);

- la famiglia di *Repetition Codes*: p_b varia secondo l'equazione

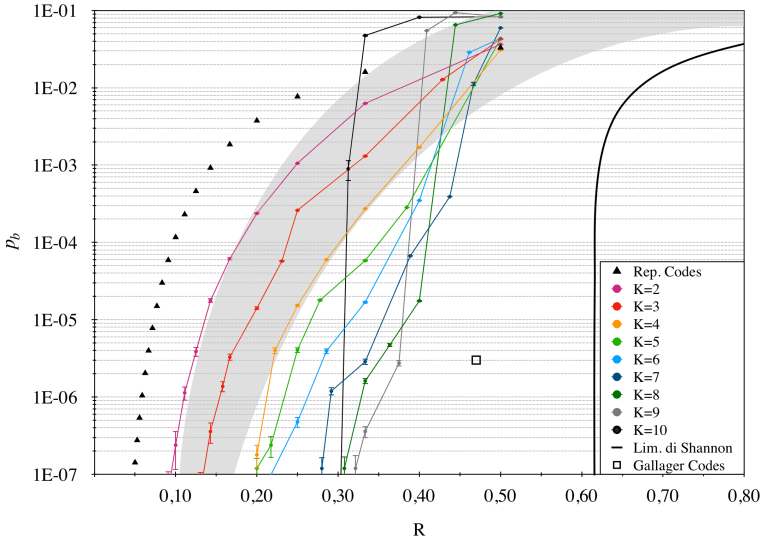
$$p_b = \sum_{i=(M+1)/2}^{M} \binom{M}{i} p^i (1-p)^{M-i},$$

dove in questo caso M indica il numero di ripetizioni;

- alcuni valori di *Gallager codes* [11]. Questi codici vennero sviluppati dall'ingegnere elettronico americano Robert G. Gallager [5] a partire dagli anni Sessanta e costituiscono uno dei massimi risultati della teoria ECC in termini prestazionali insieme ai *turbo codici* (cfr. [1], [14] e [15]);

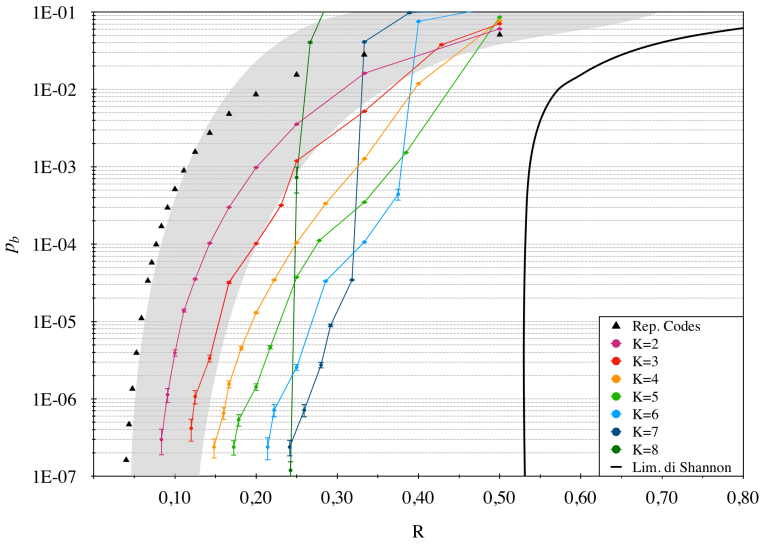
- la famiglia di codici BCH , ideati nel 1959 dal matematico francese A. Hocquenghem [7] e, in modo del tutto indipendente, nel 1960 da Raj Bose e D.K. Ray-Chaudhuri [2]. Oltre a questi sono presenti anche i *codici di Hamming* (H), tra i quali fu particolarmente rilevante l'H(7,4) (sviluppato nel 1950 [6]). Essendo molto numerosi, nelle Figure 6.7 e 6.8 compaiono come una banda grigia che ne delinea sommariamente le prestazioni.

Ricordiamo che la curva di Shannon costituisce un muro teorico invalicabile, raggiungibile solo nel limite termodinamico ($N \to \infty$). L'obiettivo ultimo di un codice

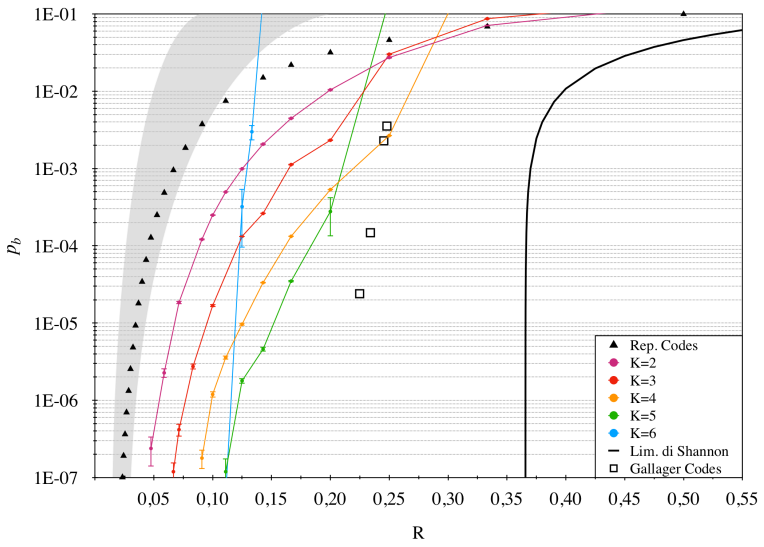


(a) $p = 0.075$

Figura 6.7: La figura mostra l'andamento di p_b in funzione del rate R al variare del parametro K nel caso in cui la probabilità di errore per il singolo bit introdotta dal canale valga $p = 0.075$ (cfr. didascalia Figura 6.8) [12].



(a) $p = 0.1$



(b) $p = 0.16$

Figura 6.8: Le due figure mostrano l'andamento di p_b in funzione del rate R al variare del parametro K nei casi in cui la probabilità di errore per il singolo bit introdotta dal canale valga rispettivamente $p = 0.1$ e $p = 0.16$. Sono stati inoltre riportati gli andamenti dei Repetion Codes, dei codici BCH e di Hamming (banda griga) e, nel secondo grafico, anche di alcuni Gallager Codes. Da ultimo, è presente anche la curva limite di Shannon, la cui equazione è descritta in (6.4.1) [11].

correttore deve quindi essere quello di approssimarsi quanto più possibile a questa curva. Il comportamento dei codici di Sourlas a connettività finità è eterogeneo e dipende fortemente dal numero K di corpi che partecipano all'interazione (da qui in avanti indicheremo con $p_b^{(K)}$ la curva p_b dipendente dal parametro K). I risultati

per i tre casi con p uguale a 0.075, 0.01 e 0.16 mostrano, salvo alcune eccezioni, comportamenti sostanzialmente simili. Vediamo di analizzarne i punti principali.

(i) I codici di Sourlas (SC) sono più efficienti dei Repetion codes (RC) solamente sotto un certo rate di soglia, che dipende da K. Ad esempio, nel caso $p = 0.1$ nessuna delle curve $p_b^{(K)}$ ha prestazioni inferiori ai RC per $R = 1/2$, mentre per $R < 1/4$ le curve $p_b^{(K)}$ giacciono definitivamente al di sotto della curva RC. Più nello specifico, ogni $p_b^{(K)}$ ha un proprio rate di soglia R_{soglia} (chiaramente dipendente da K) tale che $\forall R < R_{\text{soglia}}$ la curva $p_b^{(K)}$ in questione è più prestazionale dei RC. Il rate R_{soglia} diminuisce al crescere di K: da ciò segue che se si ricerca una trasmissione a rate vicini a $1/2$, i RC sono più performanti dei SC, ma una volta scesi al di sotto di R_{soglia} i SC sono preferibili $\forall K$ qui preso in esame.

(ii) Per quanto concerne i codici BCH (analogo per i codici di Hamming), valgono sostanzialmente le stesse considerazioni fatte per i RC, con alcune differenze. Infatti, per piccoli K (dell'ordine di 2 o 3) le prestazioni sono confrontabili con i codici BCH sostanzialmente $\forall R$, mentre al crescere di K si ripresenta il fenomeno del rate di soglia già descritto in precedenza per i RC. Per $p = 0.16$, le prestazioni dei BCH sono peggiori rispetto a quelle dei SC $\forall K$.

iii Per $p = 0.075$ e $p = 0.16$ è possibile un confronto diretto anche con i Gallager codes (GC). Ora, nel primo caso le curve $p_b^{(K)}$ hanno prestazioni al di sotto dei GC, anche se non molto distanti. Nel secondo caso, invece, accade un fenomeno interessante: $p_b^{(4)}$ ha prestazioni superiori al primo dei GC e confrontabili con quelle del secondo, rimanendo però ben al di sotto delle performance del terzo e del quarto GC. Questo risultato è molto interessante, se si pensa che i GC costituiscono lo stato dell'arte dei codici correttori ancora al giorno d'oggi.

iv Sebbene siano state mappate soltanto le prime curve della famiglia dei SC, sembra chiaro che per $K \to \infty$ non si abbia alcuna convergenza alla curva limite di Shannon. Nella sottosezione seguente approfondiremo questo aspetto.

6.4.2 Ipotesi di convergenza a una curva limite

Come abbiamo fatto notare poc'anzi, è chiaro che nel limite di $K \to \infty$ le curve $p_b^{(K)}$ non convergono al limite di Shannon. Questo sembra essere apparentemente in contrasto con i risultati ottenuti da Sourlas in [21], dove si dimostra che i SC saturano il limite di Shannon per $K \to \infty$ nel limite termodinamico. Lo stesso Sourlas, però, sottolinea come non sia a conoscenza di un algoritmo esatto che permetta di minimizzare l'hamiltoniana. Ed è proprio questo fattore la chiave di volta della questione.

Al crescere di K, in genere l'hamiltoniana presenta sempre più bacini di attrazione (cfr. 5.2.1) e, come se non bastasse, il bacino in cui è contenuto il ground state $\hat{\boldsymbol{\xi}}$ tende a restringersi sempre più. Questo complica a tal punto la ricerva di $\hat{\boldsymbol{\xi}}$ che nei fatti non si riesce più a ricostruire la convergenza al limite di Shannon prevista

da Sourlas. Dalle simulazioni si evince che ogni curva $p_b^{(K)}$ migliora le performance rispetto alle curve con K minore solo per rate minori di un certo rate di soglia.[6]

Se definiamo Γ_{S} la curva limite di Shannon, in accordo con [21] ci dovremmo aspettare

$$\lim_{K\to\infty}\lim_{N\to\infty} p_b^{(K)} = \Gamma_{\mathrm{S}}, \tag{6.4}$$

cosa che invece non accade per le motivazioni di cui sopra. In realtà, l'hamiltoniana studiata da Sourlas non è proprio quella definita in (5.1), dal momento che la prima ha campo esterno nullo ($\boldsymbol{h} = 0$). Tuttavia, almeno nel caso in cui si fa uso di un algoritmo di discesa del gradiente abbiamo verificato nella sottosezione 5.2.4 che avere campo esterno è utile per la ricerca del minimo dell'hamiltoniana. Questo fatto verrà ulteriormente giustificato più avanti.

Nonostante $p_b^{(K)} \not\to \Gamma_{\mathrm{S}}$ per $K \to \infty$, si osserva questo fattor comune per le tre simulazioni: al crescere di K, le curve $p_b^{(K)}$ si addensano attorno a un rate preciso, che indichiamo con R_∞. A questo punto si aprono due strade.

(a) Esiste una funzione limite p_b^∞ tale che $p_b^{(K)} \to p_b^\infty$ per $K \to \infty$ nel limite termodinamico. Questa ipotetica funzione, stando ai dati sperimentali, sarebbe tale da soddisfare le condizioni $p_b^\infty \to 0$ per $R \to R_\infty$ e $p_b^\infty = 0\ \forall R < R_\infty$. La convergenza di $p_b^{(K)}$, in accordo con le simulazioni, sembrerebbe dipendere da p: al crescere di p si avrebbe una convergenza più rapida.

(b) esiste un K_{tp} di "turning point" tale per cui $\forall K > K_{\mathrm{tp}}$ le curve $p_b^{(K)}$ giacciono al di sopra di $p_b^{(K_{\mathrm{tp}})}$, cioè $p_b^{(K)} > p_b^{(K_{\mathrm{tp}})}\ \forall R$. Questo equivale a dire che, fissato un certo p, esite un particolare K_{tp} oltre il quale le curve $p_b^{(K)}$ non riescono più a migliorare le prestazioni, avvicinandosi all'asse delle ordinate invece che alla curva di Shannon Γ_{S} (o all'ipotetica curva limite p_b^∞)

Le prestazioni dei SC sono fortemente influenzate dell'algoritmo di ricerca del minimo dell'hamiltoniana utilizzato e in questa sede abbiamo ampiamente esposto i limiti del metodo di discesa del gradiente da noi scelto. Per questo motivo rimandiamo a studi successivi una ricerca più approfondita sulle ipotesi di cui sopra. Vediamo però di mettere in luce un aspetto fondamentale: la ricerca del minimo dell'hamiltoniana non è un fattore che passa in secondo piano nello studio dei SC, anzi, al contrario costituisce probabilmente lo scoglio più ostico da superare nella fase di decoding.

[6] In realtà c'è un'eccezione per $K = 10$ in Figura 6.7. Tuttavia, l'andamento della curva lascia intendere che, per $p_b < 10^{-7}$, $p_b^{(10)}$ superi le performance di $p_b^{(8)}$ e $p_b^{(9)}$.

Conclusioni

Risultati

Con questo elaborato ci siamo posti l'obiettivo di approfondire alcune proprietà dei codici di Sourlas (SC) a connettività finita dal momento che, sebbene questi codici siano stati studiati in una serie di articoli (quali [8], [9], e [16]) da Y. Kabashima e D. Saad, non siamo riusciti a trovare in letteratura un'analisi sistematica delle performance degli SC al variare del numero di corpi nell'interazione K, del rate R e dell'errore intrinseco del canale p.

Seguendo la linea guida tracciata dai due ricercatori, abbiamo definito l'hamiltoniana (5.1) in analogia al modello di Ising di vetri di spin con interazioni a K corpi, di cui abbiamo dato un'interpretazione grafica tramite i factor graph (cfr. sezione 5.1). Qui è stata fatta un'analisi preliminare in vista delle simulazioni vere e proprie delle prestazioni dei SC. Nello specifico, ci siamo occupati delle seguenti tre problematiche principali.

(i) Prestazioni della simulazione al variare della dimensione N del sistema: dopo aver fornito una dimostrazione qualitativa del fatto che il parametro di overlap m sia una grandezza self-averaging, abbbiamo verificato che l'ordine di grandezza del sistema (per $K \leq 10$) debba essere $N = \mathcal{O}(1000) - \mathcal{O}(10000)$ per poter assumere di lavorare nel limite termodinamico (cfr. sottosezione 5.2.3).

(ii) Confronto tra l'inizializzazione $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{\sigma}_{\text{ran}}$ e $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$: abbiamo dimostrato che le prestazioni dei SC migliorano considerevolmente se si effettua la ricerca del minimo dell'hamiltoniana (5.1) tramite l'inizializzazione $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$ invece che $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{\sigma}_{\text{ran}}$ (cfr. 5.2.4). Abbiamo inoltre sottolineato che la difficoltà della ricerca della configurazione del minimo $\hat{\boldsymbol{\xi}}$ dell'hamiltoniana aumenta al crescere di K a causa del restringimento del bacino di attrazione in cui quest'ultimo si trova, e di come l'introduzione di un campo esterno $\boldsymbol{h}$ (che tiene traccia del messaggio di input) garantisca una probabilità maggiore di trovare $\hat{\boldsymbol{\xi}}$.

(iii) Come concretamente costruire un factor graph che sia l'equivalente di un'hamiltoniana di interazione a K corpi: abbiamo fornito un'idea di come costruire un factor graph generico, che consiste sostanzialmente nella randomizzazione dei lati di uno specifico grafo facilmente costruibile, la cui struttura è univocamente determinata dai parametri N, M, K e N_{int} in accordo con le condizioni (5.6) e (5.7) (cfr. sottosezione 5.2.5).

Definite le ottimali condizioni al contorno, siamo passati a un vero e proprio studio numerico delle prestazioni dela classe dei SC in funzione dei principali parametri in gioco: il grado di interazione K, il rate R e l'errore intriseco p per ogni bit introdotto dal canale. In primis, abbiamo cercato di capire l'effetto della geometria del factor graph sulle performance dei SC: la nostra conclusione preliminare e' che la geometria random fornisce sempre una soluzione subottimale. Come sottolineato nella sezione 6.1, le simuluzioni da noi compiute non sono in grado di dimostrare in modo formale la non esistenza di una geometria preferita, tuttavia mostrano come questo fatto sia altamente improbabile.

In una seconda fase abbiamo mostrato un confronto con i dati numerici di Kabashima e Saad tratti da [9], che costoro hanno ottenuto facendo uso dell'algoritmo di belief propagation per $K = 2$ e $R = 1/2$. Qui abbiamo dimostrato la validità dell'ipotesi di assunzione di un campo esterno non nullo nell'hamiltoniana di vetri di spin. Nello specifcio, attraverso un confronto della curva di overlap in funzione di p abbiamo verificato un fatto interessante: per $p < 0.1$ le prestazioni relative alla nostra hamiltoniana (5.1) e all'hamiltoniana (6.1) di Kabashima e Saad (dove è stato assunto $F = 0$) sostanzialmente si equivalgono, mentre per $p > 0.1$ le prestazioni di (5.1) sono decisamente superiori a quelle di (6.1) (in cui si riscontra una rapida decrescita a 0 dell'overlap, come del resto previsto dalla teoria nel caso di campo esterno nullo). Qui abbiamo dimostrato come la trasmissione del messaggio stesso, che corrisponde ad un termine di interazione a un corpo (campo magnetico esterno), permetta di utilizzare un algoritmo di decoding più rudimentale che non inficia le prestazioni del codice (cfr. metodo di discesa del gradiente, sezione 5.2.1).

Successivamente abbiamo fornito un'ulteriore prova del fatto che l'overlap sia una grandezza self-averaging, dimostrando che il processo di termalizzazione dipenda fortemente da K (cfr. sezione 6.3). Infatti, i dati riportati in Figura 6.6 verificano che la termalizzazione del sistema è tanto più lenta quanto più grande è il parametro K.

Nell'ultima fase di ricerca ci siamo occupati dello studio delle performance dei SC al variare dei parametri K, R e p. Più nel dettaglio, per i valori di p =0.075, 0.10 e 0.16 abbiamo mappato le curve $p_b^{(K)}$ in funzione del rate R al variare del numero di interazioni K, confrontando i risultati ottenuti con (i) i Repetition codes, (ii) i codici BCH e di Hamming, (iii) i Gallager codes. Abbiamo quindi verificato che i SC risultano essere sistematicamente più performanti sia di (i) sia di (ii) al di sotto di un certo rate di soglia (che dipende da K). In un caso, ossia per $p = 0.16$, i SC si sono mostrati competitivi anche con i Gallager codes, risultato notevole se si considera che quest'ultimi costituiscono (insieme ai turbo codici) lo stato dell'arte della teoria ECC. Per quanto rigurada il limite di Shannon Γ_{S}, abbiamo verificato che per $K \to \infty$ non si osserva la convergenza $p_b^{(K)} \to \Gamma_{\mathrm{S}}$, fatto ampiamente giustificato dall'oggettiva difficoltà di calcolare correttamente il ground state dell'hamiltoniana.

Proposte di sviluppi futuri

Di seguito sono riportate tre proposte in vista di sviluppi futuri della teoria dei codici di Sourlas.

(i) Nella sottosezione 6.4.2 abbiamo osservato un addensamento delle curve $p_b^{(K)}$ attorno a uno specifico rate R_∞ (il cui modulo dipende da p). Questo ci

ha portato ad ipotizzare l'esistenza di una curva limite $p_b^{(\infty)}$ che soddisfi le condizioni $p_b^{(K)} \to p_b^{(\infty)}$ e $p_b^{(\infty)} = 0 \ \forall R < R_\infty$. In questa sede non ci siamo occupati della ricerca di tale funzione, limitandoci semplicemente a un'ipotesi della sua esistenza. Rimandiamo a studi successivi una comprensione maggiore del comportamento dei SC per connettività $K \to \infty$.

(ii) Abbiamo già ampiamente sottolineato come la ricerca del minimo dell'hamiltoniana (5.1) costituisca la fase più delicata dell'intero processo di decoding. L'algoritmo implementato in questo elebarato è quello della discesa del gradiente, di cui sono stati descritti tutti i suoi limiti. Si tratta di un algoritmo sostanzialmente rudimentale che si è mostrato più che valido per una prima investigazione dei SC, tuttavia poco adatto per ulteriori ricerche più approfondite. Una valida alternativa è costituita da un riadattamento dell'algoritmo *belief propagation* (già implementato da Kabashima e Saad nelle loro ricerche) al lavoro qui proposto.

(iii) In questa sede abbiamo assunto che il parametro di connettività K fosse uniforme per ogni interazione. Questo ci ha portato a definire l'hamiltoniana di vetri di spin (5.1), che si compone di un termine di interazione a K corpi (la prima sommatoria) e uno di interazione a un corpo (la seconda sommatoria): $H = H_K + H_1$, dove H_1 è il termine energetico legato al campo esterno. Sapendo che $\boldsymbol{\sigma} = \boldsymbol{h}$ è la configurazione di minimo di H_1, abbiamo sfruttato questa informazione per calcolare il minimo di H. Ampliando quest'idea, possiamo definire una nuova hamiltoniana di interazione a K corpi come segue:

$$H_{\text{new}} = H_1 + H_2 + \cdots + H_K = \sum_{i=1}^{K} H_i, \tag{6.5}$$

dove il pedice i indica il numero di interazioni dell'hamiltoniana i-esima. Poichè la configurazione del minimo di H_1 è nota (ossia $\boldsymbol{h}$), inizializziamo $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h}$ per trovare una nuova configurazione $\boldsymbol{h'}$ che minimizzi $H_1 + H_2$. A questo punto, di nuovo inizializziamo $\boldsymbol{\sigma}_{\text{in}} = \boldsymbol{h'}$ per trovare il vettore $\boldsymbol{h''}$ che minimizzi $H_1 + H_2 + H_3$, e così via iternando il processo fino ad ottenere il ground state $\hat{\boldsymbol{\xi}}$ di H_{new}. Questo espediente di *ricerca del minimo a cascata* potrebbe limitare il problema legato al restringimento del bacino di attrazione in cui si trova $\hat{\boldsymbol{\xi}}$.

Appendice A

Metodo di punto sella

Esponiamo qui il metodo per calcolare un particolare tipo di integrale reale: il metodo di *punto-sella* [23]. Consideriamo pertanto l'integrale

$$I(N) = \int_{x_1}^{x_2} dx\, f(x) e^{Ng(x)} \tag{A.1}$$

dove supponiamo che $f, g\colon \mathbb{R} \to \mathbb{R}$ (per comodità consideriamo funzioni lisce) e $N > 0$ reale. Sia $h_N(x) := f(x)e^{Ng(x)}$ la funzione integranda.

Con un approccio naïve, possiamo immaginare che per N sufficientemente grande l'integranda h_N mostri un picco stretto e acuto, tale per cui l'integrale $I(N)$ sia totalmente dominato dall'area sottesa al picco. Siano x_0 e x_N rispettivamente il massimo assoluto della funzione g e dell'integranda h_N, con $x_0, x_N \in [x_1, x_2]$ (in questo modo rimane ben definita la successione $\{x_N\} \subset [x_1, x_2]$). Supponiamo inoltre che $f(x_0) \neq 0$. É chiaro che $\forall N > 0$ vale l'equazione:

$$h'_N(x_N) = [f'(x_N) + Nf(x_N)g'(x_N)]e^{Ng(x_N)} = 0.$$

Questo implica che $\forall N > 0$

$$f'(x_N) + Nf(x_N)g'(x_N) = 0,$$

da cui

$$g'(x_N) = -\frac{f'(x_N)}{Nf(x_N)} = \mathcal{O}\left(\frac{1}{N}\right).$$

Quindi, $g'(x_N) = 0 = g'(x_0)$ per $N \to \infty$, ossia $x_N \to x_0$. Assumiamo pertanto che x_0 sia il massimo assoluto di h_N nel limite in cui N sia sufficientemente grande.

A questo punto, cambiamo la variabile di integrazione in $x = x_0 + y/\sqrt{N}$ ed espandiamo $Ng(x)$ in potenze di y:

$$Ng(x) = Ng(x_0) + \frac{1}{2}g''(x_0)y^2 + \frac{g'''(x_0)}{6\sqrt{N}}y^3 + \cdots$$

Trattando questa espansione come un'espansione in potenze di $1/\sqrt{N}$, otteniamo

$$e^{Ng(x)} = e^{Ng(x_0)+g''(x_0)y^2/2}\left(1 + \frac{g'''(x_0)}{6\sqrt{N}}y^3 + \mathcal{O}\left(\frac{1}{N}\right)\right). \tag{A.2}$$

Seguendo lo stesso principio, abbiamo:

$$f(x) = f(x_0) \cdot \left(1 + \frac{f'(x_0)}{f(x_0)\sqrt{N}} y + \mathcal{O}\left(\frac{1}{N}\right)\right). \tag{A.3}$$

A questo punto, sostituiamo le espansioni (A.2) e (A.3) nella (A.1):

$$\begin{aligned}
I(N) = &\int_{y_1}^{y_2} \frac{dy}{\sqrt{N}} f(x_0) \cdot \left(1 + \frac{f'(x_0)}{f(x_0)\sqrt{N}} y + \mathcal{O}\left(\frac{1}{N}\right)\right) \cdot \\
&\cdot e^{Ng(x_0) + g''(x_0)y^2/2} \left(1 + \frac{g'''(x_0)}{6\sqrt{N}} y^3 + \mathcal{O}\left(\frac{1}{N}\right)\right) = \\
= &\frac{f(x_0)e^{Ng(x_0)}}{\sqrt{N}} \int_{y_1}^{y_2} dy\, e^{g''(x_0)y^2/2} \left(1 + \frac{f'(x_0)}{f(x_0)\sqrt{N}} y + \frac{g'''(x_0)}{6\sqrt{N}} y^3 + \mathcal{O}\left(\frac{1}{N}\right)\right) \approx \\
\approx &\frac{f(x_0)e^{Ng(x_0)}}{\sqrt{N}} \int_{-\infty}^{+\infty} dy\, e^{g''(x_0)y^2/2} \left(1 + \frac{f'(x_0)}{f(x_0)\sqrt{N}} y + \frac{g'''(x_0)}{6\sqrt{N}} y^3 + \mathcal{O}\left(\frac{1}{N}\right)\right) = \\
= &\frac{f(x_0)e^{Ng(x_0)}}{\sqrt{N}} \int_{-\infty}^{+\infty} dy\, e^{g''(x_0)y^2/2} \left(1 + \mathcal{O}\left(\frac{1}{N}\right)\right) = \\
= &\sqrt{\frac{2\pi}{-Ng''(x_0)}} f(x_0) e^{Ng(x_0)} \left(1 + \mathcal{O}\left(\frac{1}{N}\right)\right),
\end{aligned}$$

dove è importante sottolineare che

- $x_1 = x_0 + y_1/\sqrt{N}$, da cui chiaramente $y_1 = -\sqrt{N}(x_0 - x_1) \to -\infty$ per $N \to \infty$. In modo analogo si verifica che $y_2 \to +\infty$;
- x_0 è un punto di massimo per g, pertanto $g''(x_0) < 0$: stiamo pertanto trattando un *integrale gaussiano*.

Sintetizzando il risultato di cui sopra, abbiamo che

$$I(N) \approx \sqrt{\frac{2\pi}{-Ng''(x_0)}} f(x_0) e^{Ng(x_0)}. \tag{A.4}$$

In prima approssimazione possiamo quindi assumere che l'integrale $I(N)$ converga (a meno di una costante moltiplicativa) alla corrispettiva integranda valutata nel suo punto di massimo (ossia $I(N) \propto h_N(x_0)$).

Appendice B

Teorema della codifica di un canale

Teorema B.0.1 (Secondo teorema di Shannon)**.** *Sia dato un canale con capacità C, su cui viene trasmessa informazione con un rate R. Allora se $R < C$ esiste un codice che consente la trasmissione di informazione senza errori nel limite in cui il messaggio trasmesso ha un numero di bit che tende all'infinito.*

Dimostrazione. Supponiamo che una fonte di informazione U generi una stringa di simboli $\{a_1, \cdots, a_L\}$ ciascuno con la rispettiva probabilità $p_1, p_2, \cdots, p_L$ [17]. Assumiamo che ogni a_i sia generato singolarmente secondo la propria distribuzione di probabilità (il risultato sarà dunque una stringa della forma $a_3 a_8 a_9 \cdots$). Definiamo l'*entropia* di U come

$$H(U) := -\sum_{i=1}^{L} p_i \log_2 p_i. \tag{B.1}$$

Questa grandezza altro non è che una misura dell'incertezza sulla stringa della fonte (ad esempio, se c'è un solo simbolo che viene generato con probabilità 1 e tutti gli altri con probabilità nulla, si ottiene $H = 0$: non c'è infatti nessuna incertezza). É molto frequente l'entropia binaria $H_2(p)$: un simbolo (e.g. 0) generato con probabilità p e un altro (e.g. 1) con $1 - p$, da cui

$$H_2(p) = -p \log_2 p - (1 - p) \log_2 (1 - p). \tag{B.2}$$

Per procedere nella dimostrazione c'è bisogno di definire alcune grandezze fondamentali. La prima è l'*entropia condizionale* $H(X|Y)$, che è una misura dell'incertezza di un insieme di eventi X dato un altro evento $y \in Y$. Per una data probabilità condizionata $P(x|y)$, la grandezza che segue misura l'incertezza di X dato y:

$$H(X|y) = -\sum_{x} P(x|y) \log_2 P(x|y).$$

Si definisce dunque l'entropia condizionale come media di $H(X|y)$ sulla distribuzione di y:

$$\begin{aligned} H(X|Y) :=& \sum_y P(y) H(X|y) = \\ =& -\sum_y P(y) \sum_x P(x|y) \log_2 P(x|y) = \\ =& -\sum_{x,y} P(x,y) \log_2 P(x|y). \end{aligned} \tag{B.3}$$

Proseguendo, definiamo la *mutua informazione* come

$$I(X|Y) := H(X) - H(X|Y). \tag{B.4}$$

Il significato di questa grandezza è relativamente semplice da capire. Sia X l'insieme degli input e Y quello degli output. Ora, $H(X)$ rappresenta l'incertezza sull'input senza alcuna informazione sull'output, mentre $H(X|Y)$ corrisponde all'incertezza sull'input una volta noto l'output. La loro differenza $I(X|Y)$ può quindi essere interpretata come l'insieme di informazioni fornite dal canale.

Da ultimo, definiamo *capacità del canale* C il massimo valore della mutua informazione, ossia[1]

$$C := \max_{\{\text{input prob}\}} I(X,Y). \tag{B.5}$$

Calcoliamo la capacità di un canale BSC (in questa dimostrazione trattiamo solamente questa tipologia di canale). Supponiamo che i simboli di input siano 0 o 1 con rispettive probabilità $P(x=0) = r$ e $P(x=1) = 1-r$. Il canale ha un rumore binario simmetrico

$$\begin{aligned} P(y=0|x=0) &= P(y=1|x=1) = 1-p \\ P(y=1|x=0) &= P(y=0|x=1) = p. \end{aligned}$$

Di conseguenza la probabilità di output si può facilmente ottenere come

$$\begin{aligned} P(y=0) &= r(1-p) + (1-r)p = r + p - 2rp \\ P(y=1) &= 1 - P(y=0), \end{aligned}$$

da cui si calcolano le corrispettive entropie:

$$\begin{aligned} H(Y) =& -(r+p-2rp)\log_2(r+p-2rp) + \\ & -(1-r-p+2rp)\log_2(1-r-p+2rp) \\ H(X|Y) =& -p\log_2 p - (1-p)\log_2(1-p) = H_2(p) \\ I(X|Y) =& H(Y) - H(Y|X). \end{aligned}$$

Come detto, la capacità C del canale corrisponde al massimo della funzione $I(X|Y)$ rispetto ad r. Questo lo si raggiunge per $r = 1/2$, valore che corrisponde a un input perfettamente casuale:

$$C = \max_r I(X|Y) = 1 + p\log_2 p + (1-p)\log_2(1-p) = 1 - H_2(p). \tag{B.6}$$

[1] Le definizioni di cui sopra sono date per distribuzioni di probabilità discrete. Nel caso in cui la variabile stocastica sia continua, come ad esempio per un canale gaussiano, le definizioni rimangono analoghe, con la sola differenza di sostituire una sommatoria su x (o y) con un integrale.

Consideriamo una stringa di simboli di lunghezza N_B nella quale ogni a_i appare n_i volte (con $i = 1, \cdots, L$ e $N_B = \sum_i n_i$). Dal momento che ogni simbolo è generato in modo indipendente con probabilità p_i, nel limite di N_B sufficientemente grande si può assumere (in accordo con la legge dei grandi numeri) che $\forall \epsilon > 0,\ \exists N_B$ tale che $|n_i/N_B - p_i| < \epsilon$. Quindi, la probabilità p_{tip} che a_i appaia n_i volte nella stringa è pari a

$$\begin{aligned} p_{\text{tip}} &= p_1^{n_1} \cdots p_L^{n_L} \approx p_1^{N_B p_1} \cdots p_L^{N_B p_L} = \\ &= 2^{N_B (p_1 \log_2 p_1 + \cdots p_L \log_2 p_L)} = 2^{-N_B H(U)}. \end{aligned}$$

Una stringa di questo tipo è detta *sequenza tipica*.

Il numero di sequenze tipiche di lunghezza N_B corrisponde al numero di modi in cui distribuire n_i simboli a_i tra tutti gli N_B simboli della stringa, che corrisponde a

$$N_{\text{tip}} = \frac{N_B!}{n_1! \cdots n_L!}. \tag{B.7}$$

Poichè siamo nell'approssimazione di N_B e n_i $(\forall i)$ sufficientemente grandi, è possibile approssimare la (B.7) con la *formula di Stirling*:

$$\begin{aligned} \log_2 N_B \approx & N_B (\log_2 N_B - \log_2 \mathrm{e}) - \sum_{i=1}^{L} n_i (\log_2 n_i - \log_2 \mathrm{e}) = \\ = & - N_B \sum_{i=1}^{L} p_i \log_2 p_i = N_B H(U), \end{aligned}$$

da cui si ricava che

$$N_{\text{tip}} = 2^{N_B H(U)} = (p_{\text{tip}})^{-1}. \tag{B.8}$$

Restringiamoci al caso binario, dove $a_1 = 0$ e $a_2 = 1$ (ossia $L = 2$): indichiamo con X l'insieme degli input e con Y quello degli output, entrambi composti da stringhe di lunghezza N_B. Ora, il *random coding* è un metodo di codifica di un canale nel quale una stringa è scelta arbitrariamente da una sequenza tipica in X. Concretamente, se un messaggio sorgente ha una lunghezza N (ci sono in totale 2^N messaggi sorgenti diversi), si assegna una stringa di simboli per ogni messaggio sorgente scegliendo una sequenza tipica di lunghezza N_B dall'insieme X. É chiaro che ci sono $2^{N_B H(U)}$ sequenze tipiche in X, di cui solo 2^N rappresentano effettivamente un messaggio sorgente.

L'obiettivo finale è dimostrare che si può ottenere una probabilità di decoding corretto arbitrariamente vicina a 1 nel limite di $N_B \to \infty$. Questo è possibile solo nel momento in cui un singolo $x \in X$ corrisponde a un unico output $y \in Y$ (noto y, ci sono in generale $2^{N_B H(X|Y)}$ input x che potrebbero corrispondere). Per stimare la probabilità che ciò accada, osserviamo innanzitutto che la probabilità che una sequenza tipica di lunghezza N_B corrisponda effettivamente a una stringa sorgente di lunghezza N è

$$\frac{2^N}{2^{N_B H(X)}} = 2^{-N_B [H(X) - R]},$$

dal momento che 2^N stringhe sono scelte tra $2^{N_B H(X)}$. Quindi, la probabilità che una sequenza tipica di lunghezza N_B non sia un messaggio sorgente è

$$1 - 2^{-N_B [H(X) - R]}.$$

Richiediamo che nessuna delle $2^{N_B H(X|Y)}$ sequenze tipiche di lunghezza N_B sia un messaggio sorgente eccetto una. Questa probabilità corrisponde a

$$p_{\text{corretta}} = \left[1 - 2^{-N_B[H(X)-R]}\right]^{2^{N_B H(X|Y)}-1} \approx 1 - 2^{-N_B[H(X)-H(X|Y)-R]}.$$

Il massimo di p_{corretta} si ottiene sostituendo $H(X) - H(X|Y)$ con il suo massimo, che è proprio la capacità del canale C:

$$\max(p_{\text{corretta}}) = 1 - 2^{-N_B[C-R]}, \tag{B.9}$$

che converge a 1 per $N_B \to \infty$ sotto l'ipotesi $R < C$. La tesi del teorema è dunque dimostrata. É importante sottolineare questo fatto: la dimostrazione della tesi prevede che $N_B \to \infty$, mentre non viene posta alcuna condizione sulla lunghezza N del messaggio sorgente. Tuttavia, se si assume N finito e $N_B \to \infty$, si ha che $R \to 0$, che è ovviamente un risultato privo di interesse. Quindi, è chiaro che una trasmissione senza errori debba prevedere che anche $N \to \infty$ affinchè il rate non sia nullo. □

Bibliografia

[1] C. Berrou, A. Glavieux, and P. Thitimajshima. Near shannon limit error-correcting coding and decoding: Turbo-codes. 1. In *Communications, 1993. ICC'93 Geneva. Technical Program, Conference Record, IEEE International Conference on*, volume 2, pages 1064–1070. IEEE, 1993.

[2] R. C. Bose and D. K. Ray-Chaudhuri. On a class of error correcting binary group codes. *Information and control*, 3(1):68–79, 1960.

[3] T. Castellani and A. Cavagna. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05012, 2005.

[4] S. F. Edwards and P. W. Anderson. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

[5] R. Gallager. Low-density parity-check codes. *IRE Transactions on information theory*, 8(1):21–28, 1962.

[6] R. W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[7] A. Hocquenghem. Codes correcteurs d'erreurs. *Chiffres*, 2(2):147–56, 1959.

[8] Y. Kabashima and D. Saad. Belief propagation vs. tap for decoding corrupted messages. *EPL (Europhysics Letters)*, 44(5):668, 1998.

[9] Y. Kabashima and D. Saad. Statistical mechanics of error-correcting codes. *EPL (Europhysics Letters)*, 45(1):97, 1999.

[10] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang. The factor graph approach to model-based signal processing. *Proceedings of the IEEE*, 95(6):1295–1322, 2007.

[11] D. J. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE transactions on Information Theory*, 45(2):399–431, 1999.

[12] D. J. MacKay and D. J. Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[13] M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[14] A. Montanari. Turbo codes: The phase transition. *The European Physical Journal B-Condensed Matter and Complex Systems*, 18(1):121–136, 2000.

[15] A. Montanari and N. Sourlas. The statistical mechanics of turbo codes. *The European Physical Journal B-Condensed Matter and Complex Systems*, 18(1):107–119, 2000.

[16] T. Murayama, Y. Kabashima, D. Saad, and R. Vicente. Statistical physics of regular low-density parity-check error-correcting codes. *Physical Review E*, 62(2):1577, 2000.

[17] H. Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[18] G. Parisi. A sequence of approximated solutions to the sk model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980.

[19] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[20] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[21] N. Sourlas. Spin-glass models as error-correcting codes. *Nature*, 339(6227):693, 1989.

[22] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[23] R. Wong. *Asymptotic approximations of integrals*, volume 34. SIAM, 2001.