Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Matematica
Corso di Laurea Magistrale in Matematica

Transizioni di fase nell'allenamento delle Macchine di Boltzmann Ristrette

Relatore:
Prof. Daniele Tantari

Presentata da:
Francesco Tosello

Anno Accademico 2021/2022

Sommario

In questa tesi ci siamo occupati delle Macchine di Boltzmann Ristrette con prior binari come modelli di apprendimento non supervisionato, analizzando il ruolo del numero di neuroni nascosti in relazione alla quantità di esempi necessari per un proficuo addestramento. Abbiamo simulato uno scenario insegnante-studente e calcolato l'efficienza della macchina sotto l'assunzione di simmetria di replica per studiare la localizzazione della soglia critica oltre la quale l'apprendimento è possibile. I nostri risultati confermano la congettura secondo la quale, in assenza di correlazione tra i pesi della macchina che genera i dati, la soglia critica non dipende dal numero di unità nascoste (purché sia finito) e quindi dalla complessità dei dati. La presenza di correlazione invece riduce sensibilmente l'ammontare di esempi necessari per l'apprendimento. Abbiamo mostrato che questo effetto si accentua al crescere del numero di unità nascoste. L'intera analisi è corredata da simulazioni numeriche che ne corroborano i risultati.

Abstract

In this thesis, we dealt with Restricted Boltzmann Machines with binary priors as models of unsupervised learning, analyzing the role of the number of hidden neurons on the amount of examples needed for successful training. We simulated a teacher-student scenario and calculated the efficiency of the machine under the assumption of replica symmetry to study the location of the critical threshold beyond which learning begins. Our results confirm the conjecture that, in the absence of correlation between the weights of the data-generating machine, the critical threshold does not depend on the number of hidden units (as long as it is finite) and thus on the complexity of the data. Instead, the presence of correlation significantly reduces the amount of examples needed for training. We have shown that this effect becomes more pronounced as the number of hidden units increases. The entire analysis is supported by numerical simulations that corroborate the results.

Indice

Introduzione

Tutti noi apprendiamo continuamente concetti e nessi causali. Sappiamo che in questo processo giocano un ruolo cruciale le reti neuronali presenti nella corteccia cerebrale, formate da miliardi di unità interagenti fra loro. Negli ultimi decenni, la crescente potenza dei calcolatori ha permesso lo sviluppo dei modelli di apprendimento automatico ispirati a queste reti biologiche, le cosiddette *neural networks* o reti neuronali artificiali [MP43; Hop82; Smo86]. Il successo di questi modelli nell'emulare abilità umane quali la memoria, la capacità di distinguere e riprodurre le immagini, la facoltà del linguaggio, ecc. non è stato ancora seguito da una pari comprensione del loro funzionamento [Hua21]. Il nostro lavoro apporta un piccolo contributo al tentativo della comunità scientifica di colmare questa lacuna. Abbiamo approfondito quali fattori influenzino l'addestramento di una di queste reti neuronali artificiali, la cosiddetta Macchina di Boltzmann Ristretta (RBM); questa fornisce un semplice prototipo di modelli più complessi e si presta bene ad uno studio teorico [SH09; HOT06; Bar+12].

Una rete neuronale artificiale si può descrivere tramite un grafo; i vertici sono variabili dette neuroni e i pesi degli archi, detti pesi sinaptici, stabiliscono la relazione tra ciascun neurone e quelli a lui collegati. La RBM è un grafo bipartito: i neuroni sono divisi in due strati che svolgono funzioni differenti e le interazioni avvengono solamente tra neuroni di strati distinti. I neuroni detti visibili sono i recettori degli stimoli esterni, mentre quelli del cosiddetto strato nascosto identificano le caratteristiche dei dati. I pesi delle sinapsi che collegano i due strati rappresentano i nessi tra gli stimoli e le caratteristiche che questi segnalano [DF21].
Noi analizzeremo la RBM come modello di apprendimento non supervisionato [LBH15]. In questo contesto, l'allenamento della rete avviene tramite la stimolazione con una grande quantità di esempi dei concetti che le si vogliono insegnare, in risposta ai quali vengono modificati i pesi sinaptici per assimilare il contenuto informativo presente nei dati. Matematicamente parlando, oltre al grafo è data anche la distribuzione di probabilità sui

possibili stati dei neuroni visibili, condizionata al valore dei pesi sinaptici; questa stabilisce quanto ogni possibile stimolo sia spiegabile a partire dai concetti noti alla macchina. L'obiettivo dell'addestramento è modificare i pesi sinaptici affinché questa distribuzione approssimi la reale distribuzione degli esempi: una buona approssimazione significa che le sinapsi sono riuscite ad estrapolare le relazioni principali presenti nei dati esaminati.

È stato osservato che l'addestramento del modello presenta una transizione di fase [HT16]: finché non gli viene fornito un quantitativo sufficiente di esempi i pesi sinaptici che inferisce sono casuali, cioè non hanno alcuna relazione con i concetti presenti nei dati, questo significa che non impara; l'apprendimento comincia solo al superamento di una determinata soglia critica. Stabilire con metodi teorici la localizzazione di questa soglia è fondamentale sia per poter tarare a priori la quantità di dati da fornire alla rete neuronale, sia per capire se e come sia possibile abbassarla, cioè quali fattori concorrano alla sua determinazione.
Per calcolare questa soglia critica abbiamo valutato le prestazioni del modello in un contesto controllato, nel quale il numero e le caratteristiche dei concetti presenti negli esempi sono stabiliti a priori: il paradigma insegnante-studente [Bar+18; Dec+21]. Questo sfrutta la possibilità di usare la RBM anche in modalità generativa: una volta che le sue sinapsi sono state allenate, dalla distribuzione di probabilità si possono estrarre degli esempi di stimoli, nei quali sono presenti i concetti appresi. L'esperimento teorico si svolge nel modo seguente. Innanzitutto si costruisce una RBM ausiliaria che ha il ruolo di insegnante, le cui sinapsi sono preimpostate su certi valori; campionando la sua distribuzione si generano degli esempi che conterranno l'informazione essenziale sui concetti a lei noti, con leggere variazioni casuali in ciascuno di essi. Dopodiché si allena la RBM studente, della quale siamo interessati a valutare il processo di apprendimento, fornendole i dati generati dall'insegnante affinché provi ad inferire i concetti originari. A questo punto, confrontando le sinapsi dell'insegnante con quelle della studente si deduce l'esito dell'allenamento.

Il lavoro che abbiamo precedentemente citato analizzava la transizione di fase nel caso in cui l'insegnante prova a comunicare un singolo concetto. Sulla quella scia, in altre ricerche [Bar+17; HWH19] è stato investigato se il fenomeno si presenti allo stesso modo in presenza di due concetti da inferire; è stato osservato che la soglia critica rimane invariata in assenza di correlazione tra i due concetti, mentre se questa aumenta è necessario un minor numero di esempi per dare inizio all'apprendimento. Nonostante sia stata proposta la congettura secondo la quale in assenza di correlazione la transizione di fase non dipende

dal numero dei concetti considerati, nella letteratura manca ancora un'analisi completa del ruolo di questo parametro sulla localizzazione della soglia critica.
Tuttavia, il numero di concetti che vengono veicolati è importante per almeno due motivi: oltre a determinare le capacità della macchina, influisce pesantemente sulla complessità computazionale del processo di allenamento, poiché è collegato al numero di neuroni nascosti necessari per l'inferenza. In questo studio ci siamo posti l'obiettivo di estendere i risultati citati poc'anzi: verificare la congettura dato un numero finito arbitrario di concetti da assimilare, e investigare più approfonditamente il ruolo della correlazione nella determinazione della soglia critica. Per farlo abbiamo interpretato il problema di inferenza del paradigma insegnante-studente tramite le lenti della meccanica statistica: una disciplina che applica le tecniche proprie della statistica per trattare grandi sistemi di particelle interagenti fra loro, fornendo degli strumenti efficaci nel valutarne il comportamento complessivo quando l'analisi diretta della dinamica microscopica è troppo complicata [Nis01]. Nel nostro caso, al posto delle particelle ci sono enormi quantità di neuroni (nello strato visibile) e ingenti volumi di esempi da processare. La prospettiva macroscopica propria di questa branca della fisica e della matematica ci ha permesso di superare le difficoltà connesse alla dinamica dell'allenamento, complicata da processi di interazione non-lineare [CKS05]. La nostra analisi conferma la congettura sopra citata, inoltre mostra che in presenza di correlazione tra i concetti anche il loro aumento in numero favorisce l'apprendimento del modello, permettendo l'allenamento con volumi di esempi molto più bassi.

Questo elaborato è stato scritto con due finalità: innanzitutto presentare una sintesi del lavoro svolto e dei risultati ottenuti; in secondo luogo fornire un'introduzione graduale a questa tematica di ricerca e alle tecniche utilizzate, in modo che sia fruibile da un lettore senza competenze specifiche in materia di meccanica statistica o di apprendimento automatico. Infatti, la ricerca in questo ambito è nel pieno dello sviluppo e si inserisce all'intersezione di settori scientifici diversi, perciò è ancora carente di manuali di riferimento ad ampio spettro.
Per questo motivo, il primo capitolo contiene un'introduzione alla meccanica statistica e all'inferenza, nella quale delineiamo i concetti principali a partire da alcuni esempi portanti, affinché il lettore possa acquisire dimestichezza, in un contesto semplificato, con gli strumenti che verranno utilizzati più avanti - da ciò segue che gli esempi presentati sono da considerarsi parte integrante del testo. Nel secondo capitolo presentiamo alcuni esempi di reti neuronali già consolidati nella letteratura, concentrandoci sul modello di Hopfield

per illustrare le tecniche di analisi che abbiamo utilizzato nella nostra ricerca, in particolare la linearizzazione gaussiana, il metodo di Laplace e il metodo delle repliche [MPV86]. Concludiamo il capitolo descrivendo brevemente i vari modelli di RBM e mettendoli a confronto con il modello di Hopfield, in modo da facilitare la comprensione del loro inserimento all'interno del quadro precedentemente descritto. L'ultimo capitolo è interamente dedicato all'esposizione del nostro lavoro; può essere letto indipendentemente del resto se si è già a proprio agio con le tecniche della ricerca in questo settore. Questo comincia con una breve introduzione al tema che ha lo scopo di motivare il calcolo della pressione della distribuzione inversa all'interno del paradigma insegnante-studente. Dopodiché, ricaviamo una formula variazionale di questa pressione tramite il metodo delle repliche ed esplicitiamo le relative equazioni di auto-consistenza. Infine, ci dedichiamo a ricavare la soglia critica oggetto di questo studio e ne presentiamo le caratteristiche, accompagnando l'analisi con l'esposizione dei risultati numerici che abbiamo ottenuto.

Concludiamo questa introduzione con un paio di note: una sul rigore matematico dei contenuti, l'altra sul linguaggio utilizzato.
Abbiamo già accennato al fatto che in questo ambito la ricerca è interdisciplinare; perciò si sviluppa in una commistione di rigore matematico, intuizione fisica e verifiche sperimentali. Pertanto, alcune delle tecniche che utilizzeremo nel seguito non hanno, ad oggi, alcuna giustificazione matematica e ciononostante permettono di ottenere dei risultati in accordo con le evidenze empiriche. Il loro utilizzo ha permesso spesso di anticipare dei risultati che sono stati dimostrati matematicamente solamente svariati decenni più tardi - si pensi ad esempio alla soluzione con rottura di simmetria di replica del modello SK, dovuta a Giorgio Parisi [Par80b; Par80a]. Consci dell'importanza di separare nettamente il dominio matematico da quello euristico, sarà nostra cura segnalare i passaggi privi di una giustificazione rigorosa: saranno indicati con un punto esclamativo (ad esempio $\overset{!}{=}$, $\overset{!}{\sim}$) o espressamente discussi.
La seconda puntualizzazione riguarda la scelta dell'autore di utilizzare ove possibile le traduzioni italiane al posto dei termini inglesi che compaiono nella letteratura scientifica. Questa decisione è dovuta alla convinzione che oltre a rendere più scorrevole e uniforme il testo, principalmente scritto in italiano, facilitino l'utilizzo didattico dell'elaborato favorendo una comprensione più immediata e profonda dei concetti introdotti. Abbiamo comunque inserito i termini originali in appositi incisi o note a piè di pagina, per fare in modo che il lettore inesperto possa stabilire un parallelo con la letteratura di riferimento.

Capitolo 1

Elementi di meccanica statistica

La meccanica statistica ha avuto origine dallo studio di sistemi di molecole interagenti (gas, magneti, ecc.). Prevedere il comportamento di questi sistemi a partire dalle leggi microscopiche che regolano le interazioni dei singoli atomi o molecole non è fattibile, nemmeno numericamente, a causa delle dimensioni dei sistemi risultanti; per esempio si consideri che in una goccia d'acqua ci sono circa 10^{24} molecole. Per superare questo ostacolo è stato necessario riuscire a condensare le informazioni sulle singole interazioni in poche leggi descriventi le dinamiche macroscopiche. Questo è possibile con l'ausilio della statistica - da cui il nome della disciplina - introducendo una distribuzione di probabilità sulle configurazioni microscopiche del sistema e studiando le medie delle proprietà di interesse.
Con questo approccio emerge una caratteristica essenziale dei sistemi complessi: in certi casi i parametri del sistema presentano delle soglie critiche attraversando le quali il comportamento macroscopico cambia radicalmente; si pensi al passaggio dell'acqua da stato solido a liquido quando la temperatura supera gli $0\,^{\circ}\mathrm{C}$; questo fenomeno prende il nome di transizione di fase ed è uno degli aspetti cruciali sui quali si sofferma l'analisi dei sistemi in questa disciplina.
Negli ultimi decenni queste tecniche si sono rivelate molto versatili e le loro applicazioni si sono estese dallo studio dei sistemi fisici ai sistemi economici, alle dinamiche sociali e perfino all'apprendimento automatico. In questo lavoro ci occuperemo proprio di quest'ultima evoluzione: studieremo sistemi di tanti neuroni interagenti e ne ricaveremo una descrizione di alcuni sistemi di apprendimento automatico, così da poterne studiare un'importante transizione di fase riguardante la capacità di questi di imparare dai dati forniti.

Questo capitolo serve a richiamare i concetti principali e l'impostazione di base di questa disciplina nonché a stabilire delle notazioni e introdurre alcuni modelli basilari a cui faremo riferimento nel seguito. La trattazione comincia con la formulazione matematica del concetto di sistema in meccanica statistica: lo studio si sviluppa a partire da una funzione, che rappresenta l'energia di ogni configurazione, dalla quale si ricaverà la distribuzione di probabilità precedentemente citata, cioè la distribuzione del sistema all'equilibrio. Dopodiché introduciamo alcune quantità fondamentali, tra cui l'energia libera, e dei teoremi che ci permetteranno di derivare da questa le proprietà macroscopiche di interesse, le cosiddette osservabili. Successivamente definiamo il concetto di limite termodinamico: siccome lavoreremo con enormi quantità di corpi interagenti l'analisi dell'energia libera verrà svolta nel limite di infiniti corpi; questo richiederà delle tecniche più elaborate ma permetterà di semplificare i risultati e inquadrare le transizioni di fase come discontinuità di alcune osservabili significative. Introdotto il formalismo proprio della meccanica statistica spieghiamo come sia possibile, nel caso di sistemi descrivibili tramite particolari processi stocastici, ricavare la distribuzione di equilibrio a partire dalla dinamica microscopica e viceversa; questo esula dal campo proprio della disciplina ma ci sarà utile nei capitoli successivi. Infine, nell'ultima sezione richiamiamo alcuni elementi di statistica che ci saranno utili per lo studio teorico dei modelli nel capitolo 3 e li armonizzeremo con il formalismo del presente elaborato.

I prerequisiti per la lettura di questo capitolo sono la conoscenza di un minimo di analisi in più variabili e di un'infarinatura di teoria della probabilità che arrivi a menzionare la legge dei grandi numeri. Chi fosse digiuno di questi argomenti può cominciare con il classico [Rud76] - o [De 92] se si preferisce l'italiano - per quanto riguarda l'analisi e [Pas20; KS07; Dur19] per un'introduzione generale alla teoria della probabilità; nella sezione 1.3 introdurremo le catene di Markov, chi avesse bisogno di un sostegno può consultare [Nor97; EK05] per una trattazione più approfondita.
I concetti di meccanica statistica che utilizzeremo in questa tesi - a grandi linee la teoria di campo medio e qualche tecnica di calcolo - verranno introdotti dalle basi, tuttavia indichiamo qui alcune referenze utili: tra quelle che trattano quasi tutti gli argomenti di questo capitolo consigliamo [CKS05; MM09], la seconda ha un taglio più computazionale; il modello di Curie-Weiss è presente anche in [FV17] che ne offre una trattazione rigorosa; altri due libri che si concentrano in particolare sull'apprendimento artificiale sono [Nis01] e il recente [Hua21], il cui autore ha contribuito a scrivere un articolo del quale parleremo approfonditamente nel capitolo 3. La sezione 1.4 è un calco della parte iniziale delle note

Figura 1.1: Qui è rappresentata schematicamente una catena di N spin interagenti.

di Jean Barbier [Bar19].
Queste referenze verranno richiamate nel corso del capitolo per sopperire all'assenza di molte dimostrazioni, omesse in quanto esulano dallo scopo della tesi e sono generalmente elementari.

1.1 Descrizione di un sistema

Come abbiamo anticipato nell'introduzione del capitolo, i sistemi che studieremo saranno composti da un insieme di nodi (o individui, agenti, ecc.) che potranno trovarsi in diversi stati, a ciascuno dei quali corrisponderà un'energia dipendente dalle interazioni che si sviluppano in quelle condizioni. Ad esempio, riprendendo l'esempio dell'acqua, i nodi saranno le singole molecole e l'energia sarà data dalle loro interazioni elettromagnetiche e nucleari. In generale chiameremo Λ l'insieme dei **nodi**, ciascuno dei quali potrà trovarsi in uno degli **stati** presenti nell'insieme χ; lo **spazio delle configurazioni** del sistema - indicate solitamente con $\sigma, \tau, \dots$ - sarà $\Sigma \coloneqq \chi^\Lambda$ e ci riferiremo allo stato di un nodo in una configurazione usando i pedici, ad esempio $\sigma_i, i \in \Lambda$. La funzione che esprime la dipendenza dell'energia dalla configurazione si dice **hamiltoniana del sistema** e si indica con $\mathcal{H}\colon \Sigma \to \mathbb{R}$; questa, di volta in volta dipenderà anche da alcuni parametri che indicheremo a seguito del punto e virgola oppure in pedice, ad esempio: $\mathcal{H}_N(\sigma;\theta), \sigma \in \Sigma$.

Esempio 1.1 (Modello di Ising unidimensionale) — Questo modello è stato studiato da Ernst Ising e Wilhelm Lenz, circa un secolo fa, per comprendere il fenomeno del magnetismo a partire dall'allineamento dei momenti di dipolo degli atomi.
Si consideri una catena di $N \in \mathbb{N}_+$ nodi affiancati, quindi $\Lambda = \{1, \dots, N\}$; associamo a ogni sito uno spin a valori in $\chi = \{+1, -1\}$ che rappresenta il verso del momento di dipolo; nel seguito useremo i termini nodo e spin in modo intercambiabile. Ciascuno spin interagisce con i propri vicini tendendo ad allinearvisi siccome in questo stato l'energia del sistema è minore. Possiamo modellare matematicamente il sistema con la seguente hamiltoniana:

$$\mathcal{H}_N(\sigma) = -\sum_{i=1}^{N-1} \sigma_i \sigma_{i+1}$$

dove il pedice N indica la **taglia** del sistema.
In aggiunta, gli spin risentono di un campo esterno h che tende ad orientarli in una delle due direzioni; per tenere conto anche di questo effetto si può ampliare l'hamiltoniana così:

$$\mathcal{H}_N(\sigma; h) = -\sum_{i=1}^{N-1} \sigma_i \sigma_{i+1} - h \sum_{i=1}^{N} \sigma_i$$

◇

Data questa descrizione del sistema, bisogna stabilire le probabilità di osservare il sistema in ciascuno stato. Più avanti tratteremo alcuni casi in cui è possibile ricavare questa distribuzione direttamente dalla dinamica microscopica; per il momento ci limitiamo allo studio prettamente meccanico-statistico, che costruisce la distribuzione di probabilità sullo spazio delle configurazioni a partire dalla loro energia, secondo la massima: i sistemi fisici privilegiano gli stati di minima energia. Assumiamo che Λ e χ siano insiemi finiti, così lo sono anche le configurazioni e possiamo definire una probabilità discreta, dipendente dal parametro $\beta > 0$:

$$\mathrm{P}_\beta(\sigma) \coloneqq \mathcal{Z}^{-1} \exp(-\beta\, \mathcal{H}(\sigma)) \tag{1.1}$$

dove il termine $\mathcal{Z}$ è il fattore di normalizzazione, detto anche **funzione di partizione**; esplicitamente, in due notazioni equivalenti:

$$\mathcal{Z} \coloneqq \sum_{\sigma \in \Sigma} \exp(-\beta\, \mathcal{H}(\sigma)) \equiv \mathrm{Tr}_\sigma \, \exp(-\beta\, \mathcal{H}(\sigma))$$

Il parametro β rappresenta l'inverso della **temperatura**, indicata con T, e condiziona la variabilità del sistema: ad alte temperature (ossia per $\beta \to 0^+$) la distribuzione tende ad essere uniforme sugli stati poiché vi è una maggiore libertà di movimento e quindi il sistema è disordinato; viceversa, per $\beta \to +\infty$ il sistema tende a "congelarsi" negli stati di minima energia - che prendono il nome di **stati fondamentali** (in inglese *ground states*) - sui quali si distribuirà una probabilità uniforme. Così costruita, P_β in (1.1) è detta **distribuzione di Boltzmann-Gibbs** per $\mathcal{H}$ a temperatura β.

Nota 1.2 — Volendo essere precisi (1.1) non è una distribuzione, bensì una funzione di distribuzione: l'equivalente di una densità per distribuzioni su spazi discreti. Data una funzione di distribuzione $\mathrm{P}\colon \Sigma \to [0,1]$ la relativa distribuzione è così definita: $\tilde{\mathrm{P}}(A) \coloneqq \sum_{\sigma \in A} \mathrm{P}(\sigma)\ \forall A \in \mathcal{P}(\Sigma)$, dove la notazione $\mathcal{P}$ indica l'insieme delle parti. Viceversa, la funzione di distribuzione si ottiene valutando la distribuzione sui singoletti: $\mathrm{P}(\sigma) = \tilde{\mathrm{P}}(\{\sigma\})$.

Siccome lavoreremo con distribuzioni discrete, per le quali vi è una corrispondenza biunivoca con le relative funzioni di distribuzione, definiremo le prime tramite queste ultime e d'ora in poi ci permetteremo di identificarle.
Inoltre, a rigore, per ogni funzione di probabilità che considereremo dovremmo usare una notazione distinta (ad esempio $\mu, \nu, \dots$), tuttavia ci permetteremo l'abuso di notazione di impiegare sempre la notazione P e distinguere le varie distribuzioni dal nome delle variabili che compariranno. Ad esempio indicheremo la funzione di distribuzione della variabile aleatoria ξ con $\mathrm{P}(\xi)$, nonostante questo possa generare una piccola confusione tra la funzione di distribuzione e il valore che la stessa assume in un punto. Sarebbe più preciso stabilire la notazione P_ξ per la funzione di distribuzione e $\mathrm{P}_\xi(\hat{\xi})$ per la sua valutazione nel punto $\hat{\xi}$, però sarà utile destinare i pedici prevalentemente ai parametri, come abbiamo già visto per la distribuzione di Boltzmann-Gibbs. Ci auguriamo che l'interpretazione delle notazioni sarà agevolata dal contesto; ciononostante sarà nostra cura specificare il significato degli argomenti e dei pedici qualora riterremo che possano generare confusione.

Visto che la distribuzione di Boltzmann-Gibbs si esprime in termini di esponenziali, è bene tenere a mente alcune utili proprietà raccolte nel prossimo lemma. D'ora in avanti indicheremo con log il logaritmo a base esponenziale e con $\sinh, \cosh, \tanh$ le funzioni iperboliche.

Lemma 1.3 — *Valgono le seguenti identità* $\forall x \in \mathbb{R}$*:*

1. $\sum_{\sigma=\pm 1} e^{x\sigma} = 2\cosh(x)$
2. $\sinh(x) + \cosh(x) = e^x$
3. $\cosh(x)^2 - \sinh(x)^2 = 1$
4. $\partial_x \tanh(x) = 1 - \tanh(x)^2$
5. $\log\cosh(x) \underset{x\to 0^+}{\sim} \frac{x^2}{2}$
6. $\cosh(x)\tanh(x)^2 = \sinh(x)\tanh(x)$

Possiamo studiare il comportamento del sistema tramite i valori attesi delle **osservabili**: funzioni $\mathcal{O}\colon \Sigma \to \mathbb{R}$ che ne rappresentano certe proprietà; ad esempio $\mathcal{H}$ è l'osservabile relativa all'energia interna. La media rispetto alla distribuzione di Boltzmann-Gibbs, anche detta **media termica**, si indica

$$< \mathcal{O} >_\beta := \mathrm{Tr}_\sigma \, \mathrm{P}_\beta(\sigma)\, \mathcal{O}(\sigma)$$

Con questa notazione possiamo definire alcune quantità importanti:

- **energia interna** $\mathrm{U}_\beta \coloneqq \langle \mathcal{H} \rangle_\beta$
- **entropia** $\mathrm{S}_\beta \coloneqq \langle -\log(\mathrm{P}_\beta) \rangle_\beta = -\mathrm{Tr}_\sigma \, \mathrm{P}_\beta(\sigma) \log(\mathrm{P}_\beta(\sigma))$
- **pressione** $\mathrm{A}_\beta \coloneqq \log(\mathcal{Z}_\beta)$
- **energia libera** $\mathrm{F}_\beta \coloneqq -\frac{1}{\beta} \mathrm{A}_\beta = -\frac{1}{\beta} \log(\mathcal{Z}_\beta)$

Si noti che l'entropia è una quantità definibile per ogni distribuzione di probabilità discreta (usando l'ultima espressione): ne misura l'imprevedibilità; equivalentemente, misura il disordine nel sistema che essa rappresenta. Nel caso della distribuzione di Boltzmann-Gibbs dipende da $\mathcal{H}$ indirettamente. La pressione e l'energia libera sono due quantità cruciali, sostanzialmente equivalenti fra loro, dalle quali si possono ricavare le medie di altre osservabili. Richiamiamo un risultato fondamentale di termodinamica che mette in relazione queste quantità.

Teorema 1.4 — *Vale la seguente relazione fondamentale della termodinamica:*

$$\mathrm{F}_\beta = \mathrm{U}_\beta - \frac{1}{\beta} \mathrm{S}_\beta$$

Dimostrazione. La dimostrazione segue direttamente dalle definizioni. □

Avendo definito queste osservabili possiamo enunciare una proprietà caratterizzante della distribuzione di Boltzmann-Gibbs che la rende naturale nella descrizione dei fenomeni fisici: massimizza l'entropia a energia media del sistema fissata. Il risultato è un caso particolare del seguente teorema.

Teorema 1.5 — *Siano $\{\mathcal{O}_i\}_{i=1}^M$ delle osservabili su uno spazio discreto Σ a valori reali. La distribuzione di probabilità, senza esiti trascurabili, che massimizza l'entropia condizionatamente ai vincoli $\langle \mathcal{O}_i \rangle = c_i$, $i = 1, \dots, M$, dove $\langle \bullet \rangle$ indica la media e $c_i \in \mathbb{R}$ sono delle costanti, si esprime in questa forma:*

$$\mathrm{P} = \frac{1}{\mathcal{Z}} \exp\left(\sum_i \lambda_i \, \mathcal{O}_i \right)$$

nella quale i parametri $\lambda_i \in \mathbb{R}$ sono fissati in modo da rispettare i vincoli (risolvendo il sistema di equazioni) e $\mathcal{Z}$ è il fattore di normalizzazione.

Dimostrazione. Il risultato si ottiene applicando la tecnica dei moltiplicatori di Lagrange alla funzione entropia, considerando come variabili le probabilità dei vari esiti (si veda ad esempio [CKS05, sez. 13.2]). Per un risultato rigoroso si considera la funzione entropia definita su ${\mathbb{R}_+}^d$ (con d opportuno), affinché sia di classe $\mathcal{C}^1$. Inoltre bisogna ricordarsi di imporre l'ulteriore vincolo $\sum_{\sigma \in \Sigma} \mathrm{P}(\sigma) = 1$. □

Corollario 1.6 — *Data un'hamiltoniana $\mathcal{H}$ su uno spazio di configurazioni Σ, la distribuzione di Boltzmann-Gibbs (1.1) massimizza l'entropia a energia interna media fissata. In altre parole*

$$\mathrm{P}_\beta = \underset{\langle \mathcal{H} \rangle_{\mathrm{P}} = E}{\operatorname{argmax}} \, \mathrm{S}(\mathrm{P})$$

con $E \in \mathrm{Imm}(\mathcal{H})$ fissata e β scelto in modo da ottenere $\langle \mathcal{H} \rangle_\beta = E$.

Il calcolo diretto delle osservabili spesso non è agevole, pertanto si ricorre a dei metodi variazionali. Consideriamo un'hamiltoniana dipendente da un parametro λ:

$$\tilde{\mathcal{H}} \colon \Sigma \times [0,1] \ni (\sigma, \lambda) \mapsto \tilde{\mathcal{H}}_\lambda(\sigma) \in \mathbb{R} \tag{1.2}$$

Fissato il valore di λ possiamo calcolare rispetto all'hamiltoniana risultante le medie delle osservabili, l'energia libera e le altre quantità definite in precedenza; aggiungiamo il pedice λ per indicare questa dipendenza. Con queste notazioni otteniamo un teorema che mette in relazione la sensibilità alle perturbazioni (in λ) di questa hamiltoniana con le fluttuazioni dell'energia libera del sistema perturbato.

Teorema 1.7 (della risposta lineare) — *Sia $\tilde{\mathcal{H}}$ l'hamiltoniana definita in (1.2). Allora:*

$$\partial_\lambda \mathrm{F}_{\beta,\lambda} = < \partial_\lambda \tilde{\mathcal{H}} >_{\beta,\lambda}$$

Inoltre, detta $\mathcal{O}$ un'osservabile qualunque, vale anche

$$\partial_\lambda \langle \mathcal{O} \rangle_{\beta,\lambda} = -\beta \ \mathrm{Cov} \left[\mathcal{O}; \partial_\lambda \tilde{\mathcal{H}} \right]_{\beta,\lambda}$$

dove Cov *indica la covarianza rispetto a $\mathrm{P}_{\beta,\lambda}$: la distribuzione di Boltzmann-Gibbs di $\tilde{\mathcal{H}}$.*

Dimostrazione. Anche in questo caso per la dimostrazione è sufficiente sviluppare il conto. Si tratta fondamentalmente di usare il fatto che derivando $\log \mathcal{Z}$ si ottiene una media termica. □

Il teorema acquisisce rilevanza quando, a partire da un sistema iniziale descritto da $\mathcal{H}$, costruiamo un'hamiltoniana perturbata nella forma:

$$\tilde{\mathcal{H}}(\sigma, \lambda) = \mathcal{H}(\sigma) + \lambda \, \mathcal{O}(\sigma) \tag{1.3}$$

dove $\mathcal{O}$ è l'osservabile oggetto del nostro studio; infatti, in questo caso $\partial_\lambda \tilde{\mathcal{H}} = \mathcal{O}$ e otteniamo il seguente corollario.

Corollario 1.8 — *Considerando il sistema* (1.3) *valgono:*

$$\langle \mathcal{O} \rangle_{\beta,\lambda=0} = \partial_\lambda \Big|_{\lambda=0} \mathrm{F}_{\beta,\lambda}$$
$$\mathrm{Var}\,[\mathcal{O}]_{\beta,\lambda=0} = -\frac{1}{\beta}\partial_\lambda^2 \Big|_{\lambda=0} \mathrm{F}_{\beta,\lambda}$$

dove Var *è la varianza rispetto a* $\langle \bullet \rangle$.

Dimostrazione. È sufficiente usare il teorema precedente e calcolare i due risultati in $\lambda = 0$. Si osservi che calcolare media e varianza in $\lambda = 0$ significa calcolarle rispetto a $\mathcal{H}$. □

Pure l'energia interna e l'entropia si ricavano dalla conoscenza dell'energia libera di un sistema.

Corollario 1.9 — *Valgono le relazioni:* $\mathrm{U}_\beta = \partial_\beta(\beta\,\mathrm{F}_\beta)$, $\mathrm{S}_\beta = \beta^2 \partial_\beta\,\mathrm{F}_\beta$.

Dimostrazione. Applico il teorema 1.7 con $\tilde{\mathcal{H}} = 0 + \beta\,\mathcal{H}$, dove β è il parametro perturbativo e la temperatura è fissata a $T = 1$. Così facendo le quantità calcolate rispetto a $\tilde{\mathcal{H}}$ non sono altro che quelle calcolate rispetto a $\mathcal{H}$ alla giusta temperatura. Dalla prima equazione del teorema ottengo il primo risultato. Il secondo segue dal teorema 1.4. □

Noti questi due corollari risulta evidente come la conoscenza di un sistema discenda dalla conoscenza della sua energia libera (e delle derivate di questa). Pertanto, lo studio dei sistemi che esamineremo si concentrerà sull'ottenere un'espressione quanto più esplicita possibile per l'energia libera o, equivalentemente, per la pressione del sistema.

Esempio 1.10 (Modello a spin indipendenti) — Consideriamo il modello di sistema senza interazioni:

$$\mathcal{H}_0(\sigma) = -\sum_{i=1}^{N} h_i \sigma_i \quad h_i \in \mathbb{R}\ \forall i \tag{1.4}$$

Un'osservabile importante del sistema, che giocherà un ruolo cruciale nelle analisi dei prossimi capitoli, è la **magnetizzazione**: rappresenta il grado di polarizzazione del sistema in uno dei due versi.

$$m_N(\sigma) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i \tag{1.5}$$

Alla luce del teorema 1.7, per studiare la magnetizzazione possiamo considerare l'hamiltoniana ottenuta perturbando il sistema (1.4) con $-\lambda N m_N$:

$$\mathcal{H}_\lambda(\sigma) = -\sum_{i=1}^{N} (h_i + \lambda)\sigma_i$$

Per calcolare la funzione di partizione ricorriamo ad una fattorizzazione sui singoli spin della sommatoria sulle configurazioni; questo è il prototipo di una tecnica di calcolo imprescindibile che verrà utilizzata in varie occasioni. Dunque si calcola (omettendo l'indicazione la dipendenza da λ in pedice):

$$\begin{aligned}
\mathcal{Z}_{N,\beta} &= \mathrm{Tr}_\sigma e^{\beta \sum_{i=1}^N (h_i+\lambda)\sigma_i} \\
&= \mathrm{Tr}_\sigma \left(\prod_{i=1}^{N-1} e^{\beta(h_i+\lambda)\sigma_i} \right) \cdot e^{\beta(h_N+1)\sigma_N} \\
&= \sum_{\substack{\sigma_1=\pm 1 \\ \dots \\ \sigma_{N-1}=\pm 1}} \left(\prod_{i=1}^{N-1} e^{\beta(h_i+\lambda)\sigma_i} \right) \cdot \sum_{\sigma_N=\pm 1} e^{\beta(h_N+\lambda)\sigma_N} \\
&= \sum_{\substack{\sigma_1=\pm 1 \\ \dots \\ \sigma_{N-1}=\pm 1}} \left(\prod_{i=1}^{N-1} e^{\beta(h_i+\lambda)\sigma_i} \right) \cdot 2 \cdot \cosh(\beta(h_N+\lambda)) \\
&= \mathcal{Z}_{N-1,\beta}\, 2\cosh(\beta(h_N+\lambda))
\end{aligned}$$

Per induzione su N si ottiene

$$\mathcal{Z}_{N,\beta} = 2^N \cdot \prod_{i=1}^N \cosh(\beta(h_i+\lambda))$$

Da questo possiamo calcolare l'energia libera:

$$\mathrm{F}_{N,\beta} = -\frac{1}{\beta}\left(N\log(2) + \sum_i \log(\cosh(\beta h_i + \beta\lambda)) \right)$$

e sfruttando il corollario 1.8 ottenere media e varianza della magnetizzazione rispetto al sistema originario $\mathcal{H}_0$:

$$\begin{aligned}
\langle m_N \rangle_{\beta,0} &= -\frac{1}{N}\,\partial_\lambda\Big|_{\lambda=0} \mathrm{F}_{N,\beta} = \frac{1}{N}\sum_i \tanh(\beta h_i) \\
\mathrm{Var}\,[m_N]_{\beta,0} &= -\frac{1}{N^2\beta}\partial_\lambda^2\Big|_{\lambda=0} \mathrm{F}_{N,\beta} = \frac{1}{N^2}\sum_i (1-\tanh(\beta h_i)^2)
\end{aligned}$$

Nel caso in cui $h_i = h \in \mathbb{R}\ \forall i$ le espressioni si semplificano e notiamo che il valore della magnetizzazione si concentra sulla sua media termica per $N \to +\infty$. ◇

1.2 Limite termodinamico e transizioni di fase

Nell'esempio precedente abbiamo visto che all'aumentare della taglia del sistema la magnetizzazione progressivamente si concentra (probabilisticamente parlando). Questo fenomeno

di concentrazione delle osservabili nel limite di sistemi di grandi dimensioni accade frequentemente: se l'osservabile è **intensiva**, ossia non cresce con la taglia del sistema, e il sistema non si trova in una condizione critica, le fluttuazioni tendono ad annullarsi secondo il teorema del limite centrale quindi l'osservabile è approssimabile con un valore deterministico.
Precisamente, perturbando un sistema con $\lambda N\,\mathcal{O}$ (quantità **estensiva** come il resto dell'hamiltoniana) e chiamando $\mathrm{F}_{N,\lambda,\beta}$ l'energia libera risultante, se $\mathrm{F}_{N,\lambda,\beta}/N$ e le sue prime due derivate in λ hanno un limite finito per $N \to +\infty$ allora grazie al corollario 1.8 valgono

$$\begin{aligned} \langle \mathcal{O} \rangle_{\beta,0} &= \partial_\lambda \Big|_{\lambda=0} \frac{\mathrm{F}_{N,\lambda,\beta}}{N} \xrightarrow[N\to+\infty]{} C \in \mathbb{R} \\ \mathrm{Var}\,[\mathcal{O}]_{\beta,0} &= -\frac{1}{N\beta} \partial_\lambda^2 \Big|_{\lambda=0} \frac{\mathrm{F}_{N,\lambda,\beta}}{N} \xrightarrow[N\to+\infty]{} 0 \end{aligned} \tag{1.6}$$

Il processo di passaggio al limite $N \to +\infty$ prende il nome di **limite termodinamico** proprio perché si stabilizzano le proprietà macroscopiche - storicamente termodinamiche come la pressione, energia interna, ecc. Se un'osservabile soddisfa le due equazioni qui sopra, cioè converge in $L^2(\Sigma)$ alla sua media, si dice che **automedia** nel limite termodinamico (in inglese l'osservabile si dice *self-averaging*); questa convergenza implica quella in probabilità e in distribuzione, perciò giustifica il fatto che si studi il comportamento medio come campione di ciascuna realizzazione microscopica del sistema, poiché le fluttuazioni non sono rilevanti vista la taglia del sistema.
Sebbene non fosse esplicitato, notiamo che l'osservabile stessa può dipendere da N, come nel caso della magnetizzazione m_N; tuttavia, nel limite termodinamico è importante lavorare con quantità intensive poiché solo queste possono avere un limite finito. Indicheremo con le lettere minuscole le controparti intensive delle quantità che studieremo, tenendo sottintesa la dipendenza dalla taglia del sistema; ad esempio:

$$\mathrm{a}_\beta \coloneqq \frac{\mathrm{A}_\beta}{N} \qquad \mathrm{f}_\beta \coloneqq \frac{\mathrm{F}_\beta}{N} \qquad \mathrm{u}_\beta \coloneqq \frac{\mathrm{U}_\beta}{N} \qquad \mathrm{s}_\beta \coloneqq \frac{\mathrm{S}_\beta}{N}$$

Osserviamo che i risultati visti in precedenza valgono similmente per le versioni intensive grazie alla linearità.
Come si evince dalle equazioni (1.6) il comportamento delle osservabili dipende da quello dell'energia libera nel limite termodinamico; talvolta si dice che il sistema è ben posto quando l'energia libera intensiva ha un limite finito, perché questo rende possibile lo studio del sistema.

Aggiungiamo un breve commento sulle motivazioni e le conseguenze tecniche del limite termodinamico. Matematicamente parlando, il processo di limite accentua i comportamenti e fa emergere le discontinuità e le rotture di simmetria - osservate in natura - che altrimenti non sarebbero individuabili. Si pensi al tempo di dimezzamento dell'uranio ^{238}U ($\approx 4{,}5 \times 10^9$ anni): di fatto è un intervallo così grande che in natura il fenomeno non si osserva, eppure resta un numero finito. Per ovviare al problema è utile considerare dei limiti per dichiarare che il processo richiede un tempo infinito e quindi è da considerarsi impossibile. Il ragionamento è in stretta analogia con l'introduzione del concetto di insieme di misura nulla nella teoria della probabilità: sono entrambi *escamotage* per far combaciare la nostra esperienza con i calcoli.

Vedremo nel prossimo esempio che il comportamento di alcune osservabili specifiche permette di distinguere i vari stati del sistema e descrivere le transizioni di fase. Inoltre, questo esempio sarà il prototipo di ragionamenti simili che faremo nei prossimi capitoli.

Esempio 1.11 (Modello di Curie-Weiss) — Questo modello, nel seguito abbreviato anche con CW, è la versione *infinite-range*[1] del modello di Ising definito nell'esempio 1.1; si faccia attenzione al fatto che spesso la dicitura "modello di Ising" comprende un qualunque sistema con spin in $\{\pm 1\}$, incluso questo modello infinite-range. In questo caso tutti i nodi interagiscono tra di loro:

$$\mathcal{H}_N(\sigma) = -\frac{1}{2N}\sum_{i,j=1}^{N}\sigma_i\sigma_j - h\sum_{i=1}^{N}\sigma_i \tag{1.7}$$

Oltre a questa descrizione microscopica se ne può dare un'altra tramite un'osservabile macroscopica: la magnetizzazione definita in (1.5); ciò permette tra l'altro di palesare la natura estensiva dell'energia interna.

$$\mathcal{H}_N(\sigma) = -\frac{N}{2}m_N(\sigma)^2 - hNm_N(\sigma)$$

La magnetizzazione è la quantità macroscopica che ci permetterà di distinguere gli stati del sistema in stati ordinati - che esibiscono cioè una qualche simmetria, regolarità, uniformità nel comportamento dei singoli componenti - e disordinati. In generale, le osservabili che permettono una descrizione macroscopica del sistema e di distinguerne qualitativamente il comportamento si dicono **parametri d'ordine** - il più delle volte in verità con questo

[1] Potremmo tradurlo con "a portata infinita" o "ad interazione illimitata", si usa quando l'interazione non è ristretta ai soli nodi più prossimi bensì coinvolge tutti gli altri nodi.

termine ci si riferisce alle loro medie termiche; generalmente sono indicatori statistici, ad esempio la media aritmetica dello stato degli spin nel caso della magnetizzazione. Abbiamo già detto che si ottengono derivando l'energia libera (o equivalentemente la pressione) rispetto ad alcune variabili: in questo caso la magnetizzazione corrisponde alla derivata in h. Quando queste derivate presentano delle discontinuità si dice che il sistema esibisce una **transizione di fase**. Questa classificazione è dovuta a Paul Ehrenfest e distingue le transizioni in base al loro **ordine**: l'ordine della discontinuità; ad esempio, se la prima derivata è continua ma la seconda no si dirà che il sistema presenta una transizione di fase del second'ordine in quel parametro.

Studiamo dunque la pressione del modello di Curie-Weiss. Questo fa parte di una classe di modelli detti *di campo medio* poiché il suo comportamento nel limite termodinamico si approssima, usando il principio di Gibbs[2], con modelli a spin indipendenti in cui le interazioni di uno spin con gli altri nodi sono condensate in un campo esterno, che in un certo senso rappresenta l'effetto medio del quale il sito risente. Infatti, dall'approssimazione si ricava che la sua pressione è ottenibile come soluzione di un principio variazionale, il cui funzionale discende dalla pressione del modello spiegato nell'esempio 1.10.

Teorema 1.12 — *Considerando il modello di Curie-Weiss* (1.7)*, si ottiene la seguente pressione intensiva nel limite termodinamico:*

$$\alpha(\beta,h) \coloneqq \lim_{N\to+\infty} \mathrm{a}_{N,\beta,h} = \sup_{x\in\mathbb{R}} \mathrm{a}^{var}_{\beta,h}(x)$$
$$dove\ \ \mathrm{a}^{var}_{\beta,h}(x) \coloneqq \log(2\cosh(\beta x+\beta h)) - \frac{\beta x^2}{2} \tag{1.8}$$

Il punto di massimo $\bar{x}(\beta,h) \coloneqq \mathrm{argsup}_{x\in\mathbb{R}}\, \mathrm{a}^{var}_{\beta,h}(x)$ *è unico su* $(\mathbb{R}_+\times\mathbb{R})\setminus T$ *mentre assume due valori simmetrici su* $T \coloneqq \{\beta>1 \wedge h=0\}$*. Inoltre, lungo la linea* T *il modello presenta una transizione di fase del prim'ordine rispetto al parametro* h *e nel punto* $(\beta=1,h=0)$ *una transizione del second'ordine rispetto al parametro* β*.*

Dimostrazione. Che la pressione si esprima come soluzione del problema variazionale è provato in [MM09; FV17]; per ottenere l'equivalenza tra la versione delle referenze e quella da noi proposta è necessario esprimere la distribuzione di Boltzmann-Gibbs tramite delle distribuzioni di tipo Bernoulli.

[2]Il principio variazionale di Gibbs consiste nel definire la distribuzione di Boltzmann-Gibbs come quella che minimizza un certo funzionale, detto energia libera di Gibbs, che generalizza la definizione dell'osservabile F_β. Ulteriori informazioni a riguardo si trovano in [MM09, sez. 4.4].

Per studiare il sup deriviamo (1.8) e otteniamo un'equazione che il punto di massimo deve soddisfare:

$$\tanh(\beta(\bar{x}+h)) = \bar{x} \tag{1.9}$$

Questa si chiama **equazione di auto-consistenza** poiché esprime la quantità in esame in modo implicito, tramite una funzione di sé stessa; una tale espressione si presta ad algoritmi iterativi di punto fisso finalizzati alla sua soluzione. Nel caso $h = 0$ la funzione $\mathrm{a}^{\mathrm{var}}_{\beta,0}$ è simmetrica in x. Il grafico dell'equazione di auto-consistenza è presentato in figura 1.2; osserviamo che per $\beta > 1$ ci sono due soluzioni simmetriche; la biforcazione al variare di β è mostrata in figura 1.3. Per quanto riguarda il caso $h \neq 0$ notiamo che $\mathrm{a}^{\mathrm{var}}_{\beta,h}(x) = \mathrm{a}^{\mathrm{var}}_{\beta,0}(x+h) + \beta x h + \frac{\beta h^2}{2}$ perciò bisogna traslare il caso precedente e aggiungere un termine lineare che rompe la simmetria; si ottiene un solo massimo per qualsiasi valore di β. I conti dettagliati si possono trovare in sempre in [MM09], in figura 1.4 c'è il grafico riassuntivo di $\bar{x}$ in funzione dei parametri.
Verifichiamo che il modello presenti la transizione di fase studiando le derivata in h eβ della pressione nel limite termodinamico.

$$\begin{aligned}
\partial_h \alpha &= d_h \, \mathrm{a}^{\mathrm{var}}_{\beta,h}(\bar{x}(\beta,h)) = \partial_h \, \mathrm{a}^{\mathrm{var}}_{\beta,h}|_{\bar{x}} + \partial_x|_{\bar{x}} \, \mathrm{a}^{\mathrm{var}}_{\beta,h} \;\; \partial_h \bar{x} \\
&= \beta \tanh(\beta(\bar{x}+h)) + 0 = \beta \bar{x} \\
\partial_\beta \alpha &= \ldots = \frac{\bar{x}^2}{2} + \bar{x} h
\end{aligned}$$

Qui abbiamo calcolato la derivata del limite termodinamico della pressione, invece il comportamento della magnetizzazione è dato dal limite termodinamico della derivata. Infatti, procedendo in modo analogo a quanto fatto nell'esempio 1.10 si può interpretare il termine di campo esterno dell'hamiltoniana come una perturbazione e, grazie ai risultati della sezione 1.1, troviamo che al di fuori della linea critica T vale $\langle m_N \rangle_{\beta,h} = \frac{1}{\beta} \partial_h \, \mathrm{a}_{N,\beta,h}$. L'analisi del comportamento del sistema sulla linea critica T è più complicata e a noi non necessaria - infatti per evidenziare la discontinuità è sufficiente studiare i limiti all'approcciarsi della linea critica da ambo le parti - perciò a riguardo faremo solamente qualche breve accenno più avanti. Nel prossimo lemma sono descritte le condizioni stanti le quali vale l'equivalenza tra derivata del limite e limite delle derivate. Anticipiamo il fatto che qui le ipotesi sono verificate, perciò il parametro d'ordine del modello di Curie-Weiss (la magnetizzazione) coincide con $\bar{x}$. Ciò prova che il modello di Curie-Weiss esibisce le transizioni di fase enunciate, infatti $\bar{x}$ e discontinua in h lungo la linea T, mentre in β, scelto uno qualunque dei due rami, la discontinuità è nella derivata (si guardi anche la figura 1.3). ◇

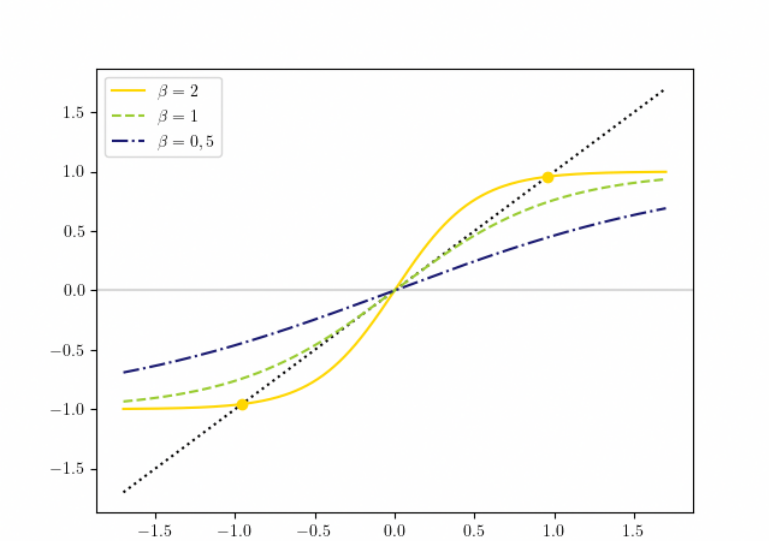


Figura 1.2: In figura è presentato il grafico dei due membri dell'equazione di autoconsistenza (1.9) con $h = 0$, in funzione di $\bar{x}$. La linea punteggiata è l'identità, le altre raffigurano la tangente iperbolica al variare di β. Si osservi che per $\beta > 1$ le intersezioni con la retta $\bar{x}$ sono due e sono simmetriche.

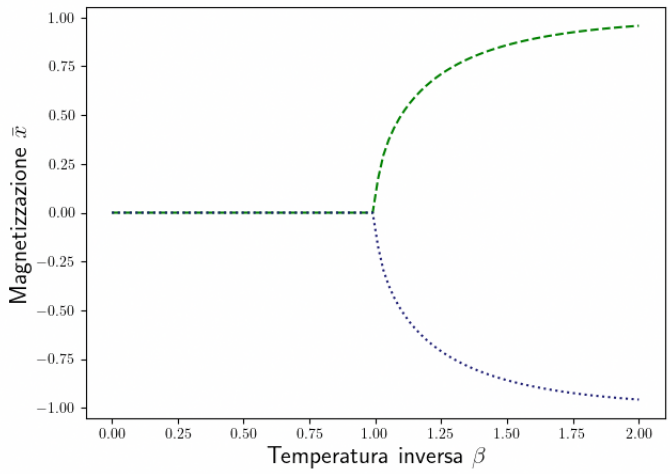


Figura 1.3: Grafico della biforcazione del punto di massimo $\bar{x}$ al variare di β per $h = 0$. Le due linee rappresentano rispettivamente la soluzione non-negativa e quella non-positiva, che per $\beta \leq 1$ coincidono, come si vede anche dalla figura 1.2.

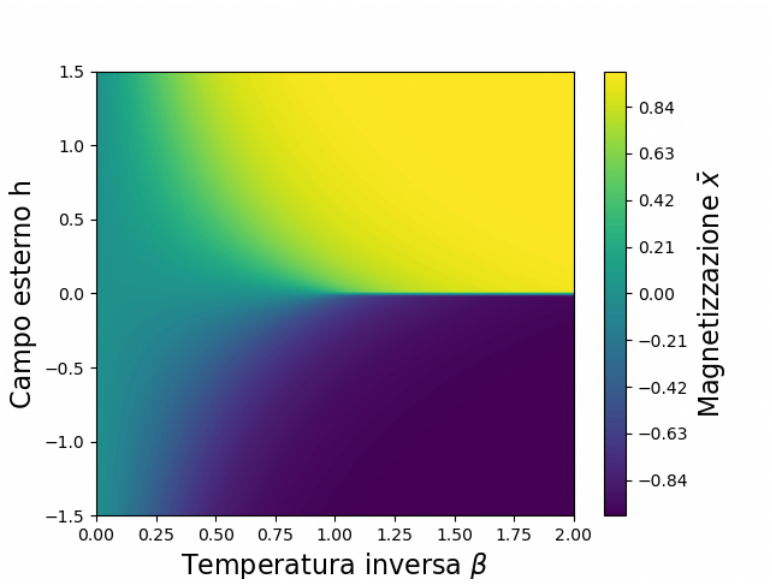


Figura 1.4: Grafico di $\bar{x}$ al variare di β e h in cui si può osservare la transizione di fase. Le due soluzioni distinte di (1.9) sono ottenute inizializzando il metodo numerico di ricerca della soluzione con valori di segno distinto.

Lemma 1.13 — *Data una sequenza di funzioni convesse $g_N\colon \mathbb{R} \to \mathbb{R}$ convergenti puntualmente a g e scelto un punto x_0 in cui sia g che tutte le g_N siano differenziabili, allora lì si possono scambiare il limite e la derivata:*

$$g'_N(x_0) \underset{N\to+\infty}{\to} g'(x_0)$$

Dimostrazione. Manipolando la definizione di funzione convessa si ricava che il rapporto incrementale è una funzione monotona crescente al variare del punto finale, da cui segue che fissato $h > 0$ vale $g'_N(x_0) \leq (g_N(x_0+h) - g_N(x_0))/|h|$, sfruttando anche la differenziabilità di g_N. Prendendone il $\limsup_{N\to+\infty}$ si ottiene $\limsup_{N\to+\infty} g'_N(x_0) \leq (g(x_0+h) - g(x_0))/|h|$ per via della convergenza puntuale. Passando al limite $h \to 0^+$ si ricava che $\limsup_{N\to+\infty} g'_N(x_0) \leq g'(x_0)$ grazie alla differenziabilità di g. Ripetendo il ragionamento per $h < 0$ si prova anche $g'(x_0) \leq \liminf_{N\to+\infty} g'_N(x_0)$. L'ipotesi di differenziabilità di g è cruciale per ottenere che $\limsup_{N\to+\infty} g'_N(x_0) \leq \liminf_{N\to+\infty} g'_N(x_0)$, da cui si ricava l'asserto per le semplici proprietà di lim sup e lim inf. ◇

I conti nella dimostrazione del teorema 1.12 forniscono anche un significato fisico al punto di massimo e alla transizione di fase che evidenzia; infatti abbiamo dimostrato che

$$\langle m_N \rangle_{\beta,h} \underset{N\to+\infty}{\to} \bar{x}(\beta, h)$$

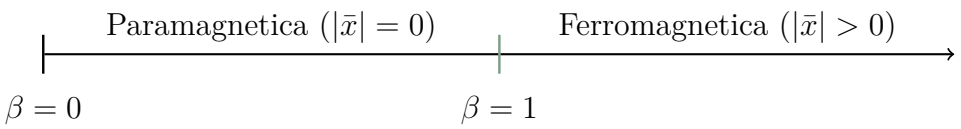


Figura 1.5: Questo è il diagramma di fase (unidimensionale, per $h = 0$) del modello di Curie-Weiss. Sono indicate le due regioni, separate dal punto critico, e il valore assoluto della magnetizzazione ($|\bar{x}|$) in ciascuna di esse.

Ciò significa che il massimizzatore $\bar{x}$ rappresenta la magnetizzazione media del sistema, quindi la transizione di fase illustrata dal grafico in figura 1.4 è in effetti il grafico della magnetizzazione media: per temperature basse avrà un valore non nullo e tenderà ad allinearsi con il campo esterno, cambiando segno bruscamente se lo cambia il parametro h. Questa identificazione tra parametri di estremizzazione e parametri d'ordine è ricorrente in meccanica statistica, perciò con un abuso di notazione spesso si usa lo stesso termine per entrambi. Sottolineiamo che la capacità di allineamento con il campo esterno di questo modello assomiglia fortemente al comportamento osservato nei magneti, punto di partenza dello studio di Ising.

Infine, questo esempio ci permette di introdurre un ultimo concetto fondamentale della disciplina: il diagramma di fase. Restringiamoci al caso $h = 0$. Sapendo che il grafico in figura 1.3 rappresenta la magnetizzazione media del sistema, possiamo ricavarne uno schema che distingue le diverse fasi del sistema al variare dei parametri: ciò che prende il nome di **diagramma di fase** - talvolta al plurale: diagramma delle fasi; ogni regione del diagramma si distinguerà per valori diversi dei parametri d'ordine. Nel caso in esame l'unico parametro è la temperatura (o equivalentemente β) perciò il diagramma sarà unidimensionale: si veda la figura 1.5. Osserviamo che si possono distinguere due regioni: quella ad alte temperature viene detta paramagnetica perché non c'è un orientamento uniforme degli spin - infatti la magnetizzazione è nulla; quella a basse temperature si dice ferromagnetica e lì la magnetizzazione assume un valore diverso da zero; il punto, o più in generale l'iperpiano, che separa regioni diverse del diagramma di fase si dice **critico** o di transizione. Ricordiamo infine che talvolta la magnetizzazione viene chiamata anche parametro d'ordine ferromagnetico proprio per sottolineare che il suo valore distingue le due regioni del diagramma di fase. Per riassumere, in questo esempio abbiamo verificato che studiare le discontinuità dell'energia libera intensiva nel limite termodinamico, rispetto alle variabili relative ai parametri d'ordine, fornisce la conoscenza

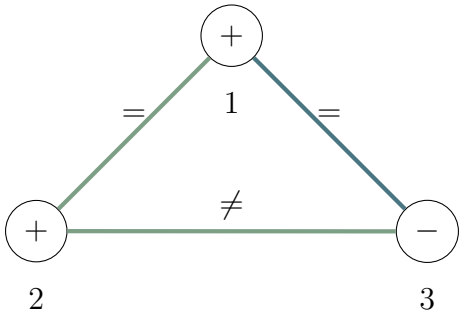


Figura 1.6: Qui è rappresentato il sistema descritto nell'esempio 1.14. All'interno di ogni nodo è indicato un possibile valore assegnato allo spin e accanto agli archi il vincolo locale prescritto dal sistema. Come si evince dal diagramma, lo spin 2 è allineato correttamente con gli altri due, mentre 1 e 3 sono discordi invece che avere lo stesso segno; per questo il sistema è frustrato.

del comportamento del sistema, quantomeno nelle sue manifestazioni macroscopiche; da questa si può costruire il diagramma di fase del modello che ne sintetizza il comportamento al variare dei parametri. ◇

Prima di passare ad una breve analisi della dinamica microscopica dei sistemi, presentiamo un esempio di sistema con frustrazione: un concetto che tornerà utile nel capitolo 3.

Esempio 1.14 — Consideriamo un sistema di tre spin tutti interagenti fra loro, illustrato in figura 1.6; si tratta di una variante del modello di Ising unidimensionale dell'esempio 1.1 in cui i nodi terminali sono anch'essi connessi. Supponiamo che i coefficienti di interazione abbiano tutti modulo unitario ma non lo stesso segno; ne risulta che a meno di permutazioni l'hamiltoniana ha questa forma: $\mathcal{H}(\sigma) = \sigma_1\sigma_2 + \sigma_1\sigma_3 - \sigma_2\sigma_3$.
Come si evince dalla figura, non è possibile trovare una configurazione che sia concorde simultaneamente con tutti i vincoli (nel senso che gli spin sono allineati in presenza di un coefficiente positivo e discordi altrimenti) perché questi vincoli locali sono in conflitto tra loro quando considerati nel contesto globale; in questo caso si dice che il sistema è frustrato o presenta **frustrazione**. In questo caso non è possibile minimizzare contemporaneamente tutti i contributi energetici locali, perciò la ricerca dello stato fondamentale dovrà tenere conto dello stato complessivo del sistema; in questo senso la frustrazione dei sistemi è una caratteristica fondamentale dei sistemi complessi.
Per concludere osserviamo che la presenza o meno di frustrazione, ovvero l'influenza

della struttura globale sulle interazioni locali, come accade spesso in matematica dipende fortemente dalla topologia del sistema. ◇

1.3 Dinamica microscopica ed equilibrio

Abbiamo già detto che la meccanica statistica si occupa dello studio di sistemi all'equilibrio da un punto di vista macroscopico poiché la trattazione microscopica della dinamica è impraticabile e poco rilevante. Tuttavia, rimane da chiarire come sia possibile stabilire la probabilità di osservare il sistema in ciascuna configurazione a partire dalla dinamica dei singoli componenti. Viceversa, per gli esperimenti numerici ci tornerà utile lo studio del problema opposto: data una distribuzione di equilibrio, come posso costruire una dinamica aleatoria che vi converga? In questa sezione approfondiremo la relazione tra il punto di vista macroscopico e quello microscopico nel caso di sistemi descrivibili tramite catene di Markov, cominciando proprio dall'introduzione di questi concetti di probabilità.

Un processo a tempo discreto su $\Sigma \coloneqq \{1, \dots, D\}$ è una collezione di variabili aleatorie $(X_t)_{t\in\mathbb{N}}$ a valori nello **spazio degli stati** Σ, dove $\operatorname{Card}\Sigma = D \in \mathbb{N}_+$. In linea di massima con gli indici $i, j, \dots$ indicheremo degli stati del sistema, quindi numeri in Σ. Questo processo si dice **catena di Markov** (di ordine uno) se lo stato futuro dipende solamente da quello presente e non anche da tutti quelli passati: $\mathrm{P}(X_{t+1} = j_{t+1}|X_t = j_t) = \mathrm{P}(X_{t+1} = j_{t+1}|X_t = j_t, X_{t-1} = j_{t-1}, \dots, X_0 = j_0)\ \forall t$. Quando le probabilità di transizione $\mathrm{P}(X_{t+1} = j_{t+1}|X_t = j_t)$ sono indipendenti da t la catena si dice **omogenea** - abbreviato nel seguito con **CMO**; in questo caso si dice **matrice di transizione** della catena di Markov la matrice così costruita:

$$\mathrm{P}_{jk} \coloneqq \mathrm{P}(X_{t+1} = k|X_t = j) \quad \forall j, k \in \Sigma, \forall t \in \mathbb{N}_+$$

Come già detto nella nota 1.2, lavorando con spazi discreti possiamo definire una probabilità direttamente a partire dagli esiti; nel seguito rappresenteremo le distribuzioni come **vettori stocastici**, cioè vettori la cui somma delle componenti sia 1: la probabilità associata a ciascuno stata è data dalla relativa componente nel vettore. Similmente possiamo definire le **matrici stocastiche**: matrici le cui righe sono vettori stocastici. Con questa terminologia osserviamo che le matrici di transizione sono per definizione matrici stocastiche.
Il prossimo lemma fornisce l'informazione cruciale che ci permette di concentrare il nostro studio proprio sulle matrici di transizione.

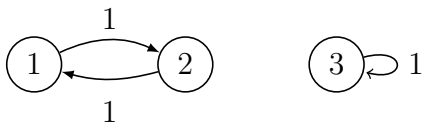


Figura 1.7: Qui è rappresentato il sistema dell'esempio 1.16 tramite un grafo: i nodi $1, 2, 3$ sono connessi da archi, a fianco dei quali è scritta la probabilità di effettuare il passaggio (direzionato) tra gli stati che collegano.

Lemma 1.15 — *La legge di una CMO è univocamente determinata dalla sua matrice di transizione* P *e dalla distribuzione* μ^0 *del suo stato iniziale* X_0*, tramite la formula*

$$\mathrm{P}(X_t = j_t, X_{t-1} = j_{t-1}, \dots, X_0 = j_0) = \mu^0_{j_0} \, \mathrm{P}_{j_0 j_1} \cdots \mathrm{P}_{j_{t-1} j_t}$$

Indicando con μ^t *il vettore stocastico che rappresenta la distribuzione di* X_t*, si ottiene la formula:* $\mu^{t+1} = \mu^t \, \mathrm{P}$ *, valida* $\forall t$*.*

Dimostrazione. La prima equazione segue dalla regola della catena per la probabilità condizionata e dalla proprietà di Markov; la seconda si ottiene marginalizzando la prima. □

Esempio 1.16 — Si consideri, sullo spazio $\Sigma = \{1, 2, 3\}$, una CMO con matrice di transizione

$$\mathrm{P} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

In figura 1.7 è illustrata la dinamica tra i tre stati del sistema, indicando a fianco di ciascun cambiamento di stato la probabilità che questo si verifichi (in questo esempio basilare ci sono solo eventi certi o impossibili). È evidente che se il sistema viene inizializzato sullo stato 3 vi rimarrà perennemente, altrimenti non lo raggiungerà mai. Diversamente, un sistema che parte dagli stati 1 o 2 continuerà perpetuamente ad alternarsi in modo ciclico tra questi. La verifica di queste intuizioni si ottiene svolgendo i conti tramite il lemma 1.15. ◇

Nell'esempio qui sopra abbiamo visto che il comportamento del sistema, in accordo col lemma 1.15, dipende dallo stato iniziale. Tuttavia, alcuni tipi di sistemi perdono questa dipendenza dalla condizione iniziale nel limite di tempi di evoluzione lunghi; ciò permette di semplificarne lo studio, ma affinché sia possibile vanno esclusi i comportamenti osservati nell'esempio 1.16.

Una CMO con matrice di transizione P si dice **irriducibile** se $\forall j, k \in \Sigma \exists r \in \mathbb{N}_+$ tale che $(\mathrm{P}^r)_{jk} > 0$, in un certo senso vuol dire che lo spazio degli stati è connesso: permette di passare da ogni stato a qualunque altro; si dice che uno stato $j \in \Sigma$ è **aperiodico** se $\exists r \in \mathbb{N}_+$ tale che $(\mathrm{P}^{r+s})_{jj} > 0 \, \forall s \in \mathbb{N}_+$, cioè se è possibile fermarsi in quello stato; infine, si dice che la matrice di transizione - e quindi la CMO - sono **ergodiche** quando $\exists r \in \mathbb{N}_+$ tale che $(\mathrm{P}^r)_{jk} > 0 \forall j, k \in \Sigma$.
Noi ci concentreremo sui sistemi ergodici: come evidenzia il prossimo lemma sono proprio quelli che non presentano le caratteristiche dell'esempio discusso sopra.

Lemma 1.17 — *Una CMO irriducibile con almeno uno stato aperiodico è ergodica. Viceversa, una CMO ergodica è irriducibile e tutti i suoi stati sono aperiodici.*

Dimostrazione. Per verificare la prima affermazione si chiami $\bar{m}_{ij}$ il minimo esponente tale che $(\mathrm{P}^{\bar{m}_{ij}})_{ij} > 0$ e sia $\bar{m} \coloneqq \max_{ij} \bar{m}_{ij}$; siccome lavoriamo in dimensione finita il massimo esiste. Sia ora k uno stato aperiodico e sia $n_k \in \mathbb{N}_+$ tale che $(\mathrm{P}^{n_k+p})_{kk} > 0 \, \forall p \in \mathbb{N}_+$. Allora

$$(\mathrm{P}^{n_k+2\bar{m}})_{ij} = \sum_{h,l} (\mathrm{P}^{\bar{m}})_{ih} (\mathrm{P}^{n_k})_{hl} (\mathrm{P}^{\bar{m}})_{lj} \geq (\mathrm{P}^{\bar{m}})_{ik} (\mathrm{P}^{n_k})_{kk} (\mathrm{P}^{\bar{m}})_{kj} > 0 \quad \forall i, j$$

La prima disuguaglianza vale perché le matrici di transizione hanno elementi non-negativi, la seconda per le ipotesi fatte.
Per quanto riguarda il viceversa l'irriducibilità segue direttamente dalla definizione, proviamo l'aperiodicità. Sia $r \in \mathbb{N}_+$ tale che P^r sia positiva (ipotesi di ergodicità). Allora per ogni $s \in \mathbb{N}_+$ sappiamo che $(\mathrm{P}^{r+s})_{ii} = \sum_k (\mathrm{P}^s)_{ik} (\mathrm{P}^r)_{ki} > 0$ perché per essere nullo dovrebbero essere nulli tutti gli elementi della riga i di P^s, ma questo non è possibile poiché è una matrice stocastica. □

Per enunciare il teorema cruciale di questa sezione necessitiamo di un'altra definizione: una distribuzione π si dice **invariante** per una matrice stocastica P quando il corrispondente vettore stocastico ne è autovettore sinistro di autovalore uno: $\pi \, \mathrm{P} = \pi$. Osserviamo che per una CMO con matrice P un tale stato è stazionario; nella terminologia propria della meccanica statistica si dice che è uno **stato di equilibrio** del sistema: la probabilità di osservarlo in ciascuna delle sue configurazioni rimarrà costante nel tempo. Come vedremo a breve, lo studio dei sistemi in questi stati, ossia all'equilibrio, può essere affrontato con gli strumenti esposti nelle sezioni precedenti poiché la distribuzione invariante racchiude tutta l'informazione sul sistema eliminando la necessità di considerarne la dinamica.

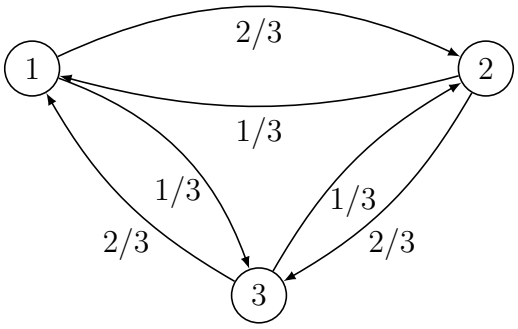


Figura 1.8: Questa è una raffigurazione della CMO a tre stati dell'esempio 1.18. Si noti che la probabilità di rimanere fermi in ciascuno stato è nulla.

Esempio 1.18 — Si consideri una CMO a tre stati rappresentata in figura 1.8. La sua matrice di transizione e il quadrato di questa sono le seguenti:

$$\mathrm{P} = \begin{pmatrix} 0 & 2/3 & 1/3 \\ 1/3 & 0 & 2/3 \\ 2/3 & 1/3 & 0 \end{pmatrix} \qquad \mathrm{P}^2 = \begin{pmatrix} 4/9 & 1/9 & 4/9 \\ 4/9 & 4/9 & 1/9 \\ 1/9 & 4/9 & 4/9 \end{pmatrix}$$

Osservando P^2 si vede immediatamente che la matrice è ergodica. Notando la simmetria della matrice si trova subito un suo vettore invariante: $(1/3, 1/3, 1/3)$. ◇

Il prossimo teorema motiva lo studio dei sistemi modellabili tramite CMO ergodiche sulla base delle proprietà delle loro distribuzioni invarianti.

Teorema 1.19 (di Perron e Frobenius) — *Una CMO $(X_t)_{t\in\mathbb{N}}$ irriducibile possiede una distribuzione invariante π; inoltre, ogni distribuzione invariante ha componenti strettamente positive. Se è ergodica, lo stato invariante è unico e X_t converge a π in distribuzione, a prescindere dalla distribuzione iniziale:* $\mathrm{P}(X_t = j) \underset{t\to+\infty}{\to} \pi_j \quad \forall j \in \Sigma$.

Dimostrazione. Il teorema è in realtà enunciabile in un contesto più generale. Per una dimostrazione nel linguaggio delle catene di Markov si veda ad esempio [Nor97, sez. 1.7]; precisiamo che nella referenza il teorema non porta questo nome. □

Grazie a questo teorema sappiamo che per CMO ergodiche la dinamica tenderà all'unico stato di equilibrio del sistema - un processo chiamato termalizzazione, in fisica. A priori non è banale stabilire la distribuzione invariante partendo dalla matrice di transizione, tuttavia esiste una condizione più semplice da verificare che implica l'invarianza (non

vale il viceversa, si veda l'esempio 1.18): si dice che una distribuzione π e una matrice stocastica P sono in **bilancio dettagliato** se vale la condizione

$$\pi_j \, \mathrm{P}_{jk} = \pi_k \, \mathrm{P}_{kj} \quad \forall j, k \in \Sigma \tag{1.10}$$

Lemma 1.20 — *Se una distribuzione π e una matrice stocastica* P *sono in bilancio dettagliato, allora la distribuzione è invariante per la matrice.*

Dimostrazione. Sommando su j l'equazione (1.10) si ottiene $(\pi \, \mathrm{P})_k = \pi_k$. □

Tenendo a mente la finalità fisica dalla quale siamo partiti possiamo riassumere quanto detto come segue: se consideriamo un sistema di nodi interagenti la cui evoluzione è regolata da leggi microscopiche che rispettano i postulati delle CMO ergodiche, nel limite di una lunga evoluzione ($t \to +\infty$) il suo stato tenderà ad un equilibrio probabilistico indipendente dallo stato iniziale. Inoltre, da questo stato di equilibrio si ricavano anche le medie temporali delle osservabili - utili ad esempio quando il sistema cambia stato troppo velocemente per effettuare una misurazione (si pensi al movimento delle molecole d'acqua in un recipiente). Dunque l'equilibrio finale è sufficiente per studiare il comportamento del sistema nel lungo periodo.

Teorema 1.21 (ergodico di Birkhoff) — *Data una CMO ergodica $(X_t)_{t\in\mathbb{N}}$ e detta π la sua distribuzione invariante, per ogni funzione $G\colon \Sigma \to \mathbb{R}$ vale*

$$\frac{1}{t} \sum_{k=0}^{t-1} G(X_t) \underset{t\to+\infty}{\overset{q.c.}{\to}} \langle G \rangle_\pi \tag{1.11}$$

dove q.c. indica una convergenza quasi certamente e $\langle G \rangle_\pi \coloneqq \sum_{j\in\Sigma} \pi_j G(j)$ la media termica rispetto a π.
In particolare, nel limite quasi ogni traiettoria visiterà tutti gli stati, secondo le proporzioni indicate da π.

Dimostrazione. La dimostrazione del teorema principale si trova ad esempio in [Nor97]. Per l'ultima osservazione è sufficiente considerare G la funzione indicatrice dello stato $j \in \Sigma$. Li visita tutti poiché π non ha componenti nulle, secondo il teorema 1.19. □

Di seguito esibiamo un esempio di sistema fisico che possiamo rappresentare con una CMO ergodica e studiare all'equilibrio con le tecniche viste.

Esempio 1.22 — Si consideri un sistema di $N \in \mathbb{N}_+$ persone che devono scegliere tra due possibilità, anche dette opinioni: $\chi = \{\pm 1\}$; per analogia con il formalismo delle

sezioni precedenti indicheremo le configurazioni del sistema con $\sigma \in \Sigma \coloneqq \{\pm 1\}^N$ invece che con i numeri cardinali. La dinamica del sistema è dovuta ai confronti tra gli individui, a seguito dei quali possono cambiare opinione; indichiamo lo stato del sistema con $(\sigma^t)_{t\in\mathbb{N}_+}$ (precedentemente abbiamo usato X_t).
Possiamo costruire questa dinamica probabilistica in modo tale che soddisfi i requisiti di una CMO: stabiliamo la distribuzione iniziale e la matrice di transizione, dalle quali segue la legge del processo stocastico secondo il lemma 1.15. Supponiamo che lo stato iniziale σ^0 sia uniforme su Σ. Assumiamo che l'aggiornamento delle scelte avvenga in modo sequenziale, ad ogni tempo scegliendo in modo uniforme l'individuo tra gli N presenti che cambierà la sua scelta. La probabilità del cambio di opinione dipenderà dalla divergenza della scelta rispetto alla media degli individui restanti; sia m^t la media del gruppo e sia $m_i^t \coloneqq m^t - \frac{\sigma_i^t}{N} = \frac{1}{N}\sum_{j\neq i}\sigma_j^t$ la media degli altri. Per ogni stato attuale σ^t fissato, modelliamo la probabilità che l'individuo cambi la sua scelta con

$$\mathrm{P}(\sigma_i^{t+1} = -\sigma_i^t | \sigma^t) \coloneqq \begin{cases} 1 & \text{se } m_i^t\sigma_i^t < 0 \\ \exp(-2\beta|m_i^t|) & \text{altrimenti} \end{cases}$$

dove $\beta \in \mathbb{R}_+$ giocherà il ruolo della temperatura inversa e ci permetterà di controllare la tendenza degli individui ad andare controcorrente, ossia ad avere un'opinione diversa dalla media. Siccome un solo individuo alla volta può cambiare scelta, stabiliamo la seguente notazione: $\sigma^{t!j}$ indicherà lo stato ottenuto da σ^t invertendo la decisione dell'individuo j-esimo, cioè $\sigma^{t!j}_i = \sigma_i^t \iff i \neq j$. Data la probabilità qui sopra possiamo costruire la matrice di transizione $\mathrm{P}_{\rho\tau} = \mathrm{P}(\sigma^{t+1} = \tau | \sigma^t = \rho)$ (quadrata, di dimensione $D = 2^N$); questa sarà definita per casi:

$$\mathrm{P}_{\rho\tau} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_i (1 - \mathrm{P}(\sigma_i^{t+1} = -\sigma_i^t | \sigma^t = \rho)) & \text{se } \tau = \rho \\ \dfrac{1}{N}\,\mathrm{P}(\sigma_i^{t+1} = -\sigma_i^t | \sigma^t = \rho) & \text{se } \tau = \rho^{!i}, i \in \{1,\dots,N\} \\ 0 & \text{altrimenti} \end{cases}$$

La CMO così ottenuta è ergodica grazie al lemma 1.17; infatti, è aperiodica poiché ci saranno sempre degli individui allineati con la maggioranza che avranno quindi la possibilità di rimanere della stessa opinione, e irriducibile siccome la probabilità di cambiare una qualunque opinione non è mai nulla, quindi attraverso una successione di cambi si può raggiungere qualunque stato. Ciò significa che la dinamica del sistema converge verso una distribuzione invariante. Si verifica che la matrice di transizione è in bilancio dettagliato con la seguente distribuzione - che quindi per il lemma 1.20 è lo stato di equilibrio del

sistema:

$$\pi_\sigma := \frac{\exp(\frac{\beta}{N}\sum_{i\neq j}\sigma_i\sigma_j)}{\sum_{\sigma\in\Sigma}\exp(\frac{\beta}{N}\sum_{i\neq j}\sigma_i\sigma_j)}$$

Si osservi che questa è la distribuzione di equilibrio del modello di Curie-Weiss: esempio 1.11 in assenza di campo esterno. Abbiamo dimostrato che la dinamica di lungo periodo del modello riguardante il comportamento delle opinioni in un contesto di gruppo può essere studiata tramite l'analisi dell'equilibrio del modello di Curie-Weiss. Ciò permette di rispondere a interrogativi sul comportamento dinamico del sistema, ad esempio circa l'emergenza o meno di una polarizzazione nelle opinioni, tramite le tecniche spiegate nelle sezioni precedenti. Nella fattispecie, a questa domanda si risponde analizzando il comportamento della magnetizzazione: come abbiamo visto nell'esempio 1.11, se $h = 0$ la magnetizzazione media sarà nulla al di sopra di una certa temperatura critica e sotto a questa si approssimerà al valore non nullo $\bar{x}$. Tuttavia, siccome è presente una biforcazione (figura 1.3) il sistema assumerà una magnetizzazione positiva o negativa a seconda della specifica traiettoria aleatoria; la traiettoria può anche passare da stati con un certo segno della magnetizzazione a stati di segno opposto, tuttavia la probabilità di questo salto decresce in modo esponenziale all'aumentare della taglia del sistema. In figura 1.9 è raffigurato il risultato di un esperimento numerico che evidenzia il comportamento della magnetizzazione del sistema durante l'evoluzione di una dinamica in fase ferromagnetica. ◇

Prima di procedere è necessario un breve commento sulla relazione che sussiste tra il limite temporale che conduce all'equilibrio ($t \to +\infty$) e il limite termodinamico ($N \to +\infty$). In questa sezione ci siamo occupati di sistemi a taglia finita, che sono ergodici ossia convergono ad un equilibrio in cui il sistema ha una probabilità non nulla di visitare ciascuno stato. Per studiare la transizione di fase abbiamo eseguito il limite termodinamico dopo aver considerato il limite temporale, conservando quindi l'ergodicità. Se invece applichiamo prima il limite termodinamico, per quanto detto poc'anzi la probabilità di osservare un cambio del segno della magnetizzazione del sistema tenderà a zero, cioè il sistema rimarrà concentrato attorno ad uno dei due valori possibili a seconda del suo stato iniziale; ciò significa che l'ergodicità non vale per sistemi di taglia infinita (o molto grande, nella pratica); si dice che il modello presenta una **rottura dell'ergodicità**.

Abbiamo visto come passare dalla descrizione microscopica di un sistema a quella macroscopica; il processo inverso si rivelerà utile nelle simulazioni numeriche, quando sarà necessario effettuare un campionamento - cioè estrarre delle realizzazioni - a partire da una

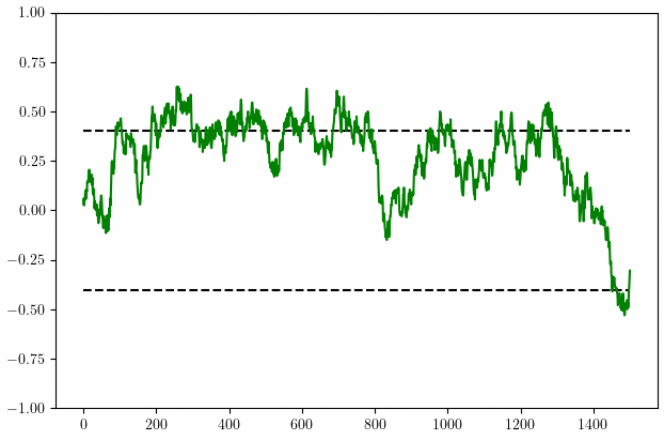


Figura 1.9: Questo grafico illustra la magnetizzazione media del sistema per una dinamica aleatoria impostata come descritto nell'esempio 1.22. Il sistema ha $N = 40$ spin e $\beta = 1{,}06$ (quindi è nella fase ferromagnetica); la dinamica è sequenziale con 1500 iterazioni. Le linee tratteggiate rappresentano il valore teorico di riferimento per la magnetizzazione media, calcolato risolvendo l'equazione di auto-consistenza (1.9).

certa distribuzione di probabilità data. Infatti, per sistemi di grandi dimensioni diventa impraticabile calcolare direttamente la probabilità di ogni singola estrazione, sia perché il numero di stati cresce esponenzialmente e pure perché il calcolo dell'energia di ogni singolo stato diventa dispendioso.
Il teorema 1.21 di Birkhoff permette di calcolare le medie rispetto ad una distribuzione di equilibrio sfruttando la legge dei grandi numeri; per far questo è sufficiente trovare una CMO che converga alla distribuzione scelta e calcolarne la dinamica a partire da un qualunque stato iniziale. Questi metodi per il campionamento da una distribuzione di probabilità vengono chiamati **Markov Chain Monte Carlo**, abbreviato MCMC. Concludiamo la sezione introducendo un paio di dinamiche generali che possono essere adattate per convergere ad una qualunque distribuzione di Boltzmann-Gibbs data.

Esempio 1.23 — Data un'hamiltoniana $\mathcal{H}$ e fissata una temperatura inversa β costruiamo una CMO ergodica, detta **dinamica di Glauber**, che ha come equilibrio la distribuzione di Boltzmann-Gibbs relativa a questa hamiltoniana. Scegliamo arbitrariamente uno stato iniziale X_0. A partire dallo stato X_t, per ottenere lo stato successivo operiamo come segue: scegliamo uniformemente uno spin i, calcoliamo la differenza di energia con lo stato attuale $\Delta\mathcal{H} := \mathcal{H}(X_t{}^{!i}) - \mathcal{H}(X_t)$ e con una certa probabilità accettiamo il cambiamento ponendo $X_{t+1} := X_t{}^{!i}$, altrimenti manteniamo $X_{t+1} := X_t$. Abbiamo usato la notazione $\bullet^{!i}$, già

introdotta in precedenza, per indicare il cambiamento di un singolo spin. Nella dinamica di Glauber la probabilità di accettare l'inversione di spin è

$$\frac{1}{1+e^{\beta\Delta\mathcal{H}}} = \frac{1}{2}(1+\tanh(-\frac{\beta}{2}\Delta\mathcal{H})) \tag{1.12}$$

Ciò significa che favoriamo il passaggio a configurazioni con minore energia, in accordo con il principio generale della fisica. Notiamo che per determinare la dinamica è sufficiente calcolare la differenza energetica tra gli stati invece che l'intera energia; questo può ridurre sensibilmente la complessità computazionale dell'algoritmo. Per trovare la distribuzione invariante esplicitiamo la matrice di transizione:

$$\mathrm{P}_{\rho\tau} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_i \frac{e^{\beta(\mathcal{H}(\rho^{!i})-\mathcal{H}(\rho))}}{1+e^{\beta(\mathcal{H}(\rho^{!i})-\mathcal{H}(\rho))}} & \text{se } \tau = \rho \\ \dfrac{1}{N}\dfrac{1}{1+e^{\beta(\mathcal{H}(\rho^{!i})-\mathcal{H}(\rho))}} & \text{se } \tau = \rho^{!i}, i \in \{1,\dots,N\} \\ 0 & \text{altrimenti} \end{cases} \tag{1.13}$$

Questa matrice è chiaramente aperiodica poiché c'è una probabilità non nulla di restare in qualunque stato; è irriducibile perché è sempre possibile invertire uno spin arbitrario e quindi, siccome c'è un cammino finito tra due stati qualunque, la probabilità di eseguire proprio quelle transizioni è positiva; pertanto la matrice è ergodica grazie al lemma 1.17. Verifichiamo che questa matrice di transizione è in bilancio dettagliato con la distribuzione di Boltzmann-Gibbs prescelta. L'unico caso non banale è quello in cui $\tau = \rho^{!i}$ per un certo indice i; detta $\Delta\mathcal{H}$ la differenza energetica tra τ e ρ otteniamo la seguente catena di identità che prova l'asserto.

$$\begin{aligned} \frac{\exp(-\beta\,\mathcal{H}(\rho))}{1+e^{\beta\Delta\mathcal{H}}} &= \frac{\exp(-\beta\,\mathcal{H}(\rho^{!i}))}{1+e^{-\beta\Delta\mathcal{H}}} \\ \frac{e^{\beta\Delta\mathcal{H}}}{1+e^{\beta\Delta\mathcal{H}}} &= \frac{1}{1+e^{-\beta\Delta\mathcal{H}}} \end{aligned}$$

Un metodo che produce un risultato equivalente più rapidamente è il cosiddetto **Metropolis-Hastings**, ottenuto con la stessa procedura della dinamica appena descritta ma sostituendo la probabilità di accettare l'inversione dello spin con la seguente espressione: $\exp(-\beta\max\{0,\Delta\mathcal{H}\})$. In pratica possiamo distinguere due casi: se il cambiamento abbassa l'energia lo accettiamo sicuramente, altrimenti solo con una certa probabilità dipendente dalla temperatura e dal divario energetico. ◇

1.4 Applicazione a problemi di inferenza

Fino ad ora abbiamo analizzato il comportamento dei modelli al variare dei parametri (temperatura, intensità delle interazioni, ecc.); questo viene detto **problema diretto**. Altresì rilevante è lo studio del **problema inverso**: conoscendo il comportamento del sistema, è possibile ricavare i valori dei parametri che producono tale dinamica? Studiare anche questa prospettiva è importante perché la conoscenza dei parametri permette di comprendere e replicare il funzionamento del modello. Nella fattispecie, nei capitoli successivi ci occuperemo di modelli di apprendimento automatico che hanno proprio il compito di inferire i criteri con i quali sono stati generati certi dati.
Siccome rispetto al problema diretto dovremo lavorare con ruoli invertiti tra parametri e variabili che indicano lo stato del modello (si parte dalle ultime per ricavare i primi), al fine di evitare confusioni ci riferiremo alle impostazioni iniziali del modello diretto con il nome di **segnale** - indicato con x - e al suo comportamento con il nome generico **dati** (prodotti) - d'ora in poi indicati con y.
Nel seguito ci focalizzeremo su problemi di inferenza e descriveremo come si possano trattare attraverso le lenti della meccanica statistica; in questi casi il processo diretto è la generazione dei dati - anche detti esempi - che avviene secondo una certa legge probabilistica $\mathrm{P}(y|x)$; l'obiettivo dell'inferenza, cioè del problema inverso, sarà ricavare il segnale con il quale era verosimilmente impostato il modello quando ha emesso i dati in esame.
Siccome durante l'inferenza il segnale è incognito assumeremo che sia una variabile aleatoria e ci serviremo della regola di Bayes per calcolare la cosiddetta **distribuzione a posteriori**:

$$\mathrm{P}(x|y) = \frac{\mathrm{P}(x)\,\mathrm{P}(y|x)}{\mathrm{P}(y)} = \frac{\mathrm{P}(x)\,\mathrm{P}(y|x)}{\sum_x \mathrm{P}(x)\,\mathrm{P}(y|x)} \tag{1.14}$$

Per fare questo dobbiamo postulare una probabilità congiunta sull'insieme del segnale e dei dati oppure, equivalentemente, la distribuzione del segnale - detta ***prior***[3] - e la distribuzione condizionata dei dati in funzione del segnale - detta **verosimiglianza**; questi sono i due termini che appaiono al numeratore dell'equazione (1.14), il termine $\mathrm{P}(y)$ al denominatore si dice **evidenza**. Mentre la verosimiglianza si ricava dalla conoscenza del funzionamento del modello, il prior può non essere noto - in particolare qualora il segnale sia in realtà deterministico e la sua aleatorietà sia dovuta esclusivamente alla

[3]Termine inglese che in italiano significa precedente, a priori. Sta a indicare il fatto che la distribuzione deve essere nota prima di effettuare l'analisi, in un certo senso è preconcetta e rappresenta una polarizzazione del segnale.

nostra ignoranza dello stesso; in questi casi può essere assunto uniforme - massimizzando l'entropia, ossia l'incertezza riguardo alla sua distribuzione - e si semplifica dall'espressione (1.14). Osserviamo incidentalmente che in inferenza classica, quando vi sono molti dati rispetto al segnale da recuperare, la verosimiglianza domina il prior che può quindi essere tralasciato, a prescindere dalla sua distribuzione.
La risoluzione del problema inverso, ossia l'inferenza del segnale, avverrà tramite la ricerca dei massimi (in x) della verosimiglianza (in inglese *maximum likelihood estimation*). Osserviamo che massimizzare la verosimiglianza è equivalente a massimizzarne il logaritmo - detto *log-likelihood* - poiché questa è una funzione monotona crescente. Se lo spazio dei valori del segnale è discreto e non troppo grande possiamo cercare direttamente l'$\operatorname{argmax}_x P(x|y)$ - procedimento chiamato **massimizzazione a posteriori**, o MAP; nel caso continuo o in grandi dimensioni introdurremo in seguito tecniche più raffinate.
Per concretizzare il discorso si consideri questo semplice esempio di processo generativo: fissata una parola di partenza se ne scrivano altre ad essa correlate man mano che sovvengono spontaneamente. Un'altra persona può risolvere il problema inverso cercando di indovinare la parola originaria a partire da quelle che sono state scritte; per farlo cercherà di individuare le parole date le quali la verosimiglianza di quelle scritte è massima. Naturalmente non sempre questo è fattibile; o meglio, talvolta nell'indovinare non possiamo fare di meglio che scegliere casualmente, perché non siamo in possesso di abbastanza informazioni per orientare la decisione. Da questo esempio si evince l'utilità di comprendere anche in quali condizioni il processo inverso sia possibile e come vari la qualità dell'inferenza all'aumentare dei dati disponibili. In questo contesto si può parlare di transizione di fase quando il ritmo del recupero di informazioni varia bruscamente al superamento di una certa soglia di dati a disposizione; si pensi alla dinamica del richiamo dei ricordi dalla memoria: capita di non riuscire a ricordare qualcosa fino a che improvvisamente un indizio "sblocca il ricordo".
Al di là della terminologia, il formalismo mutuato dalla meccanica statistica fornisce delle tecniche molto efficaci per studiare i problemi inversi di grandi dimensioni; infatti le tecniche classiche sono spesso inapplicabili per ragioni di complessità computazionale quando la cardinalità dello spazio delle configurazioni del segnale è elevata - si tenga presente che nei problemi riguardanti l'allenamento delle reti neuronali il numero di variabili da inferire è addirittura dell'ordine di quello dei dati, se non superiore.
Dunque, tornando alla massimizzazione della verosimiglianza, interpretiamo la distribuzione a posteriori (1.14) nella forma di Boltzmann-Gibbs: sia

$$\mathcal{H}(x;y) \coloneqq -\log(\mathrm{P}(x)) - \log(\mathrm{P}(y|x))$$

così possiamo definire

$$\mathrm{P}(x|y,\beta) \coloneqq \frac{\exp(-\beta\,\mathcal{H}(x;y))}{\Omega(y,\beta)} \tag{1.15}$$

dove Ω è il fattore di normalizzazione (in x) e β è un parametro aggiuntivo che ci permetterà di controllare la variabilità del processo di inferenza, spesso detto **rumore** del sistema (più si abbassa e più l'informazione si disperde); per $\beta = 1$ si ottiene la distribuzione a posteriori originaria. La distribuzione, considerata a y fissato, privilegia gli stati di maggiore probabilità congiunta e per il corollario 1.6 massimizza l'entropia del segnale sotto certi vincoli. La funzione di partizione non è altro che la quantità che prima abbiamo chiamato evidenza; facciamo presente che questa, se non anche l'hamiltoniana, può dipendere dal numero $N \in \mathbb{N}_+$ di parametri che bisogna inferire. Come abbiamo già detto, ci interesseremo al caso di sistemi con un numero enorme di gradi di libertà del segnale, perciò in analogia con i sistemi fisici visti fin ora chiameremo sempre limite termodinamico il limite per $N \to +\infty$.

Un'importante differenza con quanto visto nelle sezioni precedenti è che qui il parametro y dell'hamiltoniana è una variabile aleatoria su un certo spazio di probabilità. Questo deve essere distinto dallo spazio degli stati su cui è presente la distribuzione di Boltzmann-Gibbs, che rappresenta l'equilibrio di una dinamica; ricordiamo che la relativa media, indicata con $\langle\bullet\rangle$, si dice termica. Invece, le medie rispetto alla distribuzione dei parametri aleatori si dicono **medie sul disordine** e le indicheremo con la notazione $\overline{\bullet}$; in contesti analoghi talvolta ci riferiremo ai parametri aleatori come al disordine del sistema. Si tenga presente che il termine disordine usato in questa accezione va tenuto distinto dallo stesso termine usato come contrario di ordine, regolarità, simmetria come abbiamo fatto in precedenza, ad esempio nella spiegazione del concetto di entropia. Inoltre, è opportuna un'ulteriore precisazione terminologica riguardo alle varie tipologie di disordine, il cui significato risulterà maggiormente chiaro nel capitolo 3. Generalmente, in meccanica statistica quando un sistema dipende dal disordine, quindi è nella forma (1.15), si dice che il disordine è **temprato**[4]. Nel caso specifico dei problemi inversi la questione è più delicata, perché siamo in presenza di due distribuzioni di probabilità. Nel problema diretto si dice che il disordine (nel segnale originario) è **piantato** nel sistema; quindi anche i dati da

[4]Il nome viene dal processo di tempra dei metalli, la cui struttura viene fissata tramite un brusco raffreddamento; l'analogia sta nel fatto che questi parametri sono fissati nel modello e durante l'evoluzione restano congelati. In inglese si usa il termine *quenched disorder* (e per la media *quenched average*). Nel caso in cui i parametri siano liberi di variare assieme alle configurazioni del sistema si parla di *annealed disorder*; quest'ultimo termine è anch'esso un prestito dalla metallurgia e significa cotta di nuovo o riscaldata.

esso generati, che conservano questa impronta, questo condizionamento, quando vengono intesi come disordine della probabilità inversa verranno chiamati disordine piantato. La distinzione è meramente concettuale e non sostanziale; ciononostante, spesso si assume che il disordine temprato abbia una distribuzione il più uniforme possibile, mentre nel caso del problema inverso ciò non solo non è possibile ma non avrebbe senso: il disordine piantato contiene l'informazione relativa al problema diretto dal quale deriva.
Nei casi di sistemi disordinati saremo interessati a calcolare le medie delle osservabili rispetto a entrambe le distribuzioni contemporaneamente (termica e sul disordine); tuttavia, la media termica ha la precedenza poiché l'evoluzione del modello e quindi il raggiungimento dell'equilibrio avvengono a dati fissati. Analogamente alle sezioni precedenti, dove abbiamo definito e studiato l'energia libera del sistema $\mathrm{F}_\beta(y)$, ora che y è una quantità aleatoria definiamo l'**energia libera temprata/piantata** come la sua media sul disordine:

$$\hat{\mathrm{F}}_\beta \coloneqq \overline{\mathrm{F}_\beta(y)} = -\frac{1}{\beta}\overline{\log(\Omega(y,\beta))}$$

Nel seguito, in caso di sistemi con con parametri aleatori utilizzeremo prevalentemente la versione temprata (o piantata, a seconda del problema; qui non faremo differenze) dell'energia libera e di altre quantità e osservabili. Come fatto nel corollario 1.8, a partire dall'energia libera temprata potremo calcolare la media sul disordine di varie quantità utili; sarà sufficiente scambiare la derivata rispetto ai parametri usati per le perturbazioni con la media temprata (è lecito se l'energia libera è sufficientemente regolare). Similmente a quanto visto nella sezione 1.2 rispetto alla distribuzione di Boltzmann-Gibbs, se le quantità temprate si concentrano sulla loro media (disordinata) nel limite termodinamico si dice che **automediano**. Anche in questo caso la proprietà di concentrazione delle quantità di interesse giustifica lo studio delle quantità medie in rappresentanza delle singole impostazioni dei parametri. Come nel caso senza disordine, un sistema con parametri aleatori è ben posto quando l'energia libera intensiva temprata automedia nel limite termodinamico.
Possiamo dare un significato allo studio dell'energia libera in un contesto di inferenza osservando che si tratta dell'entropia dei dati: ponendo $\beta = 1$ si ottiene

$$\overline{\mathrm{F}_1(y)} = -\overline{\log(\mathrm{P}(y))} = \sum_y -\mathrm{P}(y)\log(\mathrm{P}(y)) = H(y) \tag{1.16}$$

Ricordando che l'entropia misura l'imprevedibilità del sistema - o in altre parole la nostra ignoranza riguardo ad esso - si nota che studiare l'entropia è un modo per analizzare quanto è efficace il processo di inferenza, quanto riusciamo a scoprire sui dati.

Capitolo 2

Modelli di apprendimento artificiale

Con la dicitura "tecniche di apprendimento artificiale" ci si riferisce a dei modi di programmare i calcolatori affinché siano in grado di *imparare*, come fa un bambino, dai segnali che riceve interagendo con l'ambiente. Non a caso il campo di ricerca è spesso chiamato anche *intelligenza artificiale*: si tratta di istruire la macchina affinché sia in grado di simulare l'intelligenza umana. Ai fini della nostra esposizione è sufficiente una descrizione per esempi dell'intelligenza: pensiamo alla capacità di riconoscere e classificare gli oggetti, la facoltà di interagire tramite il linguaggio, l'abilità di sintetizzare e rielaborare le informazioni; tutti questi sono ambiti nei quali gli algoritmi di intelligenza artificiale sono giunti a livelli di prestazioni comparabili con quelle umane. Queste attività necessitano di estrapolare l'informazione dai dati forniti, strutturarla e stabilire delle relazioni tra i concetti; i modelli di apprendimento artificiale cercano di sintetizzare queste operazioni in una formulazione matematica, utile sia per ricavare degli algoritmi operativi che per comprenderne meglio il funzionamento. In questa trattazione ci concentreremo prevalentemente sulla modellazione matematica del problema, accennando solo brevemente alle implementazioni algoritmiche e alle loro ottimizzazioni.

Prima di entrare nel vivo della materia trovo necessaria una puntualizzazione: la potenza di questi *sistemi intelligenti* e la loro rapidissima evoluzione reclamano delle urgenti e inevitabili considerazioni di carattere etico e filosofico circa l'opportunità e l'indirizzo della ricerca in questi ambiti; vista la natura tecnica di questo elaborato non tratteremo qui la questione - anche perché sull'argomento si sono espresse personalità di ben altro calibro, da Isaac Asimov a Shoshana Zuboff [Zub19]; tuttavia invito i lettori di questo scritto ad informarsi, interrogarsi e confrontarsi su questi temi, poiché le decisioni che

prenderemo come collettività senz'altro condizioneranno le nostre vite negli anni a venire.

Tornando ai modelli matematici dell'apprendimento, iniziamo col dire che i primi ebbero origine da imitazioni del funzionamento biologico del cervello, precisamente delle reti neuronali: ci si interessò alla loro capacità di evolvere per imparare e se ne cominciarono a studiare delle versioni artificiali; i modelli elaborati tramite questa analogia vanno sotto il nome di **reti neuronali artificiali**[5]. Inizieremo la trattazione con la descrizione del modello di neurone di McCulloch e Pitts, spiegando il suo funzionamento e la dinamica neuronale deterministica che ne scaturisce. Questa dinamica, in condizioni opportune, presenta degli stati attrattori che possono essere sfruttati per impostare il sistema affinché "ricordi" certe informazioni: data una condizione iniziale tenderà a convergere verso la *memoria* che più gli assomiglia; il modello che si ottiene però svolge una funzione mnemonica, non compiutamente di apprendimento nel senso descritto poc'anzi. Considerando la versione stocastica di questo modello la dinamica diventa aleatoria; ciononostante si riesce comunque ad ottenere una convergenza probabilistica verso un equilibrio che può essere studiato come un sistema di meccanica statistica, detto modello di Hopfield, con le tecniche richiamate in precedenza. Lo studio di questo modello ci sarà utile anche per introdurre delle tecniche, come il metodo delle repliche, che utilizzeremo nel seguito. Successivamente esporremo il modello di apprendimento il cui studio occuperà la parte restante dell'elaborato: la cosiddetta Macchina di Boltzmann Ristretta; ne forniremo una breve descrizione e spiegheremo in quale misura sia un modello di apprendimento artificiale.

Questo capitolo vuole servire come introduzione ai modelli di apprendimento per chi abbia già familiarità con la meccanica statistica, altrimenti funge anche da introduzione a certe tecniche di calcolo. Di seguito indichiamo alcune referenze di accompagnamento per il lettore che volesse costruirsi una conoscenza più solida in materia. Le prime due sezioni descrivono dei modelli ben consolidati nella manualistica, mentre le macchine di Boltzmann trattate nell'ultima sono tutt'ora un argomento vivo in letteratura, quindi il materiale disponibile è più frammentato. Consigliamo [CKS05; Mac03] per dei libri incentrati sui temi che tratteremo; sul modello di Hopfield si possono consultare anche [Tal11; Nis01] - dei quali il primo è impostato da un punto di vista matematico; per quanto riguarda l'ultimo argomento, recentemente è stato pubblicato [Hua21], altrimenti bisogna affidarsi a degli

[5]Spesso ci si riferisce a questi col nome di *reti neurali* - calcato sull'inglese *neural networks*; tuttavia si tratta di un uso improprio poiché si riferiscono al sistema neuronale, non al sistema nervoso. Enciclopedia Treccani. it. URL: https://www.treccani.it/vocabolario/neurale (visitato il 01/09/2022)

articoli riassuntivi come [DF21]. Una prima presentazione delle tecniche che introdurremo nel corso del capitolo si trova in queste referenze; fa eccezione la teoria delle distribuzioni, della quale faremo un uso molto limitato e la cui conoscenza non è indispensabile per la comprensione del procedimento; ciononostante, man mano che affronteremo i vari argomenti daremo indicazioni più specifiche per un eventuale approfondimento. Oltre a queste referenze didattiche segnaliamo anche un utile numero sull'intelligenza artificiale del periodico italiano Ithaca [Mat20], che affronta queste tematiche con spirito tecnico-divulgativo per mezzo del contributo di vari ricercatori.

2.1 Neurone di McCulloch e Pitts

Siccome il primo modello si basa sul funzionamento della rete neuronale nel cervello, innanzitutto richiamiamo alcuni rudimenti di biologia necessari alla comprensione. All'interno del cervello vi è una fitta rete di cellule dette *neuroni* - dell'ordine di 10^{10} unità - che si scambiano impulsi elettrici. Ciascun neurone è connesso a circa 10^4 sui simili tramite delle giunzioni dette *sinapsi* - che possono essere inibitorie o eccitatorie - e, a seconda dei segnali che riceve, potrà trasmettere a sua volta un impulso elettrico verso gli altri neuroni ai quali è connesso.
Trascurando i restanti particolari biologici, possiamo già descrivere la proposta di modellazione avanzata dal neuropsicologo W. S. McCulloch e dal matematico W. Pitts nel 1943 [MP43]. Siano $\sigma_1, \dots, \sigma_N$ i **neuroni** della rete; possono essere attivi o inattivi: $\sigma_i \in \{+1, -1\} \forall i$. Nel seguito ci riferiremo ai neuroni anche usando soltanto il loro indice. Sia $J_{ij} \in \mathbb{R}\, \forall i, j$ l'efficacia della **sinapsi** tra i neuroni σ_i e σ_j, il segno ne indica la tipologia; in assenza di collegamento si avrà $J_{ij} = 0$, ad esempio assumeremo che $J_{ii} = 0 \forall i$. Il segnale che il neurone i riceve sarà calcolato come $\sum_{j=1}^{N} J_{ij}\sigma_j = J_i \cdot \sigma$, dove il membro di destra è stato scritto in notazione matriciale pensando a J come matrice e σ come vettore. Questo neurone ha una soglia U_i al di sopra della quale si attiva dal segnale entrante, perciò il suo nuovo stato si esprime mediante la formula

$$\sigma_i(t + \Delta t) \coloneqq \operatorname{sgn}(J_i \cdot \sigma(t) - U_i) \tag{2.1}$$

Questo è detto modello di **neurone di McCulloch-Pitts**, talvolta abbreviato con *neurone MP*. Da un punto di vista matematico in realtà possiamo rimuovere le soglie U dal modello, aggiungendo un neurone fittizio $\sigma_\star$ con sinapsi $J_{i\star} = U_i$ che inizialmente sia attivo - vi resterà perpetuamente a patto di assumere che $\operatorname{sgn}(0) = 1$. Dunque, d'ora in poi assumeremo che U sia il vettore nullo. Biologicamente i neuroni si aggiornano con tempistiche differenti, spesso

anche in parallelo. Per semplicità, nel modello si indicizzano i tempi di aggiornamento con $t \in \mathbb{N}_0$ e si considerano due modalità estremali di avanzamento del sistema: **in parallelo**, dato lo stato $\sigma(t)$ del sistema si aggiorna ogni neurone contemporaneamente tramite (2.1), quindi $\sigma(t) \rightsquigarrow \sigma(t+1)$; **sequenzialmente**, dato lo stato $\sigma(t)$ si sceglie uniformemente un indice $i \in \{1, \dots, N\}$ e si aggiorna quel neurone con (2.1), da cui $\sigma(t) \rightsquigarrow \sigma_i(t+1)$. Con entrambe le scelte otteniamo una dinamica su $\Sigma = \{\pm 1\}^N$, dipendente dalla matrice J, in cui i neuroni si aggiornano con la regola:

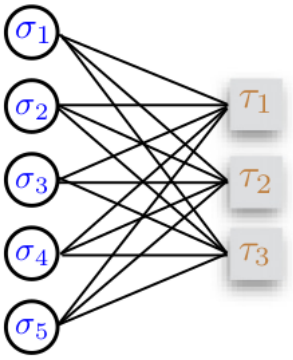
$$\sigma_i(t+1) \coloneqq \operatorname{sgn}(J_i \cdot \sigma(t)) \tag{2.2}$$

La struttura della rete neuronale, condensata nella matrice J, influenza pesantemente il comportamento del sistema e di conseguenza anche le tecniche utilizzate per studiarlo. Una tipologia di reti molto studiate sono quelle *a strati* (in inglese *layered*) raffigurate in figura 2.1a; in questo caso i neuroni hanno interazioni solamente con quelli di strati diversi, ad esempio nelle reti *feed-forward* dove la struttura dei neuroni forma un grafo diretto aciclico, quindi in cui la relazione causale tra neuroni ha solo una direzione. Noi studieremo invece le **reti neuronali ricorrenti** (*recurrent* in inglese), dove possono essere presenti dei cicli, perciò il segnale emesso da un neurone può indirettamente influenzare sé stesso; per queste si tenga presente l'illustrazione in figura 2.1b.

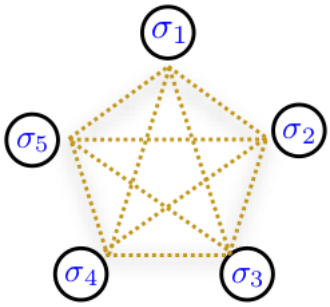
Esempio 2.1 (Sistema con sinapsi uniformi) — Consideriamo la dinamica in parallelo definita da (2.2) con $J_{ij} = (1 - \delta_{ij})\frac{J}{N}$, $J \in \mathbb{R}_*$, cioè assumiamo che l'efficacia e il tipo di sinapsi siano uniformi. Questa si riduce a

$$\sigma_i(t+1) = \operatorname{sgn}(J) \operatorname{sgn}\Big(\frac{1}{N}\sum_j \sigma_j(t)\Big)$$

che è un'espressione indipendente da i; del resto, per simmetria, non poteva essere altrimenti. Riprendendo la notazione per la magnetizzazione definita nell'esempio 1.10 - in questo contesto detta anche **attività neuronale media** - possiamo scrivere equivalentemente $\sigma_i(t+1) = \operatorname{sgn}(J)\operatorname{sgn}(m_N(t))$. Per evitare casi patologici assumiamo che N sia dispari, cosicché la magnetizzazione non possa mai annullarsi. Nel seguito trascureremo questi casi patologici poiché ai fini dello studio delle quantità nel limite termodinamico saranno ininfluenti.
Se $J > 0$, a seconda del segno della magnetizzazione al tempo iniziale $t = 0$, in un solo passo il sistema si polarizzerà facendo assumere a tutti i neuroni lo stesso stato - diremo anche che i neuroni *si allineano*. Questi stati uniformi sono di equilibrio per la dinamica poiché se $\sigma_i = +1 \forall i$ (il caso opposto è analogo) allora la magnetizzazione avrà lo stesso



(a) Questa è una rete neuronale a due strati; i neuroni del primo sono indicati con la lettera σ mentre quelli del secondo con τ. Si noti che non sono presenti collegamenti tra i neuroni dello stesso strato. Si tratta di una RBM con 5 nodi visibili e 3 nascosti.



(b) Questa è una rete neuronale ricorrente, nella quale tutti i neuroni sono connessi tra loro. Si tratta di una schematizzazione del modello di Hopfield che vedremo nella sezione 2.2.

Figura 2.1: Qui sono rappresentati i grafi di due reti neuronali: i neuroni sono i nodi e le sinapsi sono i collegamenti tra questi. Le figure sono prese da [Bar+18].

segno e quindi nei tempi successivi lo stato rimarrà invariato. In generale gli stati del sistema $\bar{\sigma} \in \Sigma$ che sono un punto fisso per la dinamica (2.2) si dicono **attrattori**; in simboli $\sigma \mapsto \cdots \mapsto \bar{\sigma} \mapsto \bar{\sigma} \mapsto \ldots$.
D'altra parte, se $J < 0$ appare un cambio di segno e la dinamica inverte la magnetizzazione: $\sigma_i(t+1) = -\operatorname{sgn}(m_N(t))$. Ciò fa si che al primo aggiornamento il sistema si polarizzerà sullo stato opposto rispetto al caso con sinapsi eccitatorie e, successivamente, il suo stato si alternerà tra $\sigma = +1$ e $\sigma = -1$, infatti $\forall t$:

$$\sigma_i(t+2) = -\operatorname{sgn}(m_N(t+1)) = -\operatorname{sgn}(\sum_j \sigma_j(t+1)) = -\operatorname{sgn}(\sum_j -\operatorname{sgn}(m_N(t))$$
$$= \operatorname{sgn}(N \operatorname{sgn}(m_N(t))) = \operatorname{sgn}(m_N(t)) = -\sigma_i(t+1)$$

Quando la dinamica converge verso degli stati che una volta raggiunti si alternano in modo ciclico si dice che presenta un **ciclo attrattore** di periodo k, dove k è la lunghezza del ciclo; in simboli $\sigma \mapsto \cdots \mapsto \bar{\sigma}_1 \mapsto \cdots \mapsto \bar{\sigma}_k \mapsto \bar{\sigma}_1 \mapsto \ldots$. ◇

La dinamica di questi sistemi, per quanto deterministica una volta fissate le interazioni, può essere molto complicata da descrivere; ciononostante esistono dei teoremi che ne

garantiscono la convergenza verso degli attrattori. Per enunciare questi risultati abbiamo bisogno di definire cosa si intende per **funzione di Lyapunov** della dinamica: una funzione $L\colon \Sigma \to \mathbb{R}$ che sia limitata dal basso e che sia monotona decrescente all'evolvere della dinamica, cioè $L(\sigma(t+1)) \leq L(\sigma(t)) \forall t$. Si osservi che se il sistema raggiunge uno stato di minimo della funzione di Lyapunov questo sarà un punto fisso della dinamica, cioè un attrattore; tuttavia possono esistere attrattori che non siano punti di minimo per la funzione.

Teorema 2.2 (Convergenza di dinamiche in parallelo) — *Se la matrice J è simmetrica allora la funzione*

$$L(\sigma) \coloneqq -\sum_i |J_i \cdot \sigma|$$

è una funzione di Lyapunov per la dinamica in parallelo (2.2) e il sistema evolverà verso un attrattore o un ciclo attrattore di periodo 2.

Dimostrazione. La prova è una verifica, si può trovare una dimostrazione in [CKS05, sez. 3.2]. Rimarchiamo il fatto che è essenziale lavorare con sistemi finiti affinché la dimostrazione funzioni. □

Teorema 2.3 (Convergenza di dinamiche sequenziali) — *Se la matrice J è simmetrica e sulla diagonale ci sono solo valori non-negativi allora la funzione*

$$L(\sigma) \coloneqq -\frac{1}{2}\sum_{i,j} \sigma_i J_{ij} \sigma_j = -\frac{1}{2}\sigma^T \cdot J \cdot \sigma \tag{2.3}$$

è una funzione di Lyapunov per la dinamica sequenziale (2.2), indipendentemente dall'ordine di aggiornamento dei neuroni, e il sistema evolverà verso un attrattore.

Dimostrazione. Come per la prova precedente la dimostrazione si trova in [CKS05, sez. 3.2]. Qui la condizione di auto-interazioni non-negative è essenziale. □

Nella prossima sezione analizzeremo un'applicazione dei neuroni MP al problema di memorizzare, tramite la dinamica, un segnale dato; di impostare cioè il modello, affinché converga verso alcuni stati prestabiliti. In alternativa, a partire dai neuroni MP si possono costruire modelli più complessi di reti neuronali. Un esempio è il *percettrone multistrato*: un tipo di rete feed-forward che può essere allenata a imitare - o meglio, approssimare - il comportamento di certe funzioni non lineari [CKS05].

2.2 Modello di Hopfield per la memorizzazione

Prima di addentrarci nella fattispecie del modello facciamo una breve digressione per chiarire in che senso la convergenza verso uno stato prestabilito può essere interpretata come una funzione mnemonica. Nell'uso comune ci riferiamo alla memoria come alla capacità di richiamare alla mente alcune informazioni precedentemente assimilate. In questi termini si può dire che alcuni sistemi fisici presentano una capacità mnemonica: indipendentemente dalla loro condizione iniziale, tramite la loro dinamica ritornano sempre ad uno stato predeterminato. Si pensi ad una molla che nonostante le sollecitazioni tende a spostarsi verso lo stato di equilibrio, oppure ad un magnete che lasciato libero si orienta sempre nella stessa direzione (dipendente dai campi magnetici dei quali risente). Dunque i neuroni MP, la cui dinamica converge verso degli attrattori, sono dei dispositivi di memoria se siamo in grado di configurare l'efficacia delle sinapsi in modo tale da rendere attrattori degli stati predefiniti.

Supponiamo di voler memorizzare nel sistema di neuroni gli stati $\xi^\mu \in \Sigma, \mu = 1, \dots, P$; li chiameremo vettori di informazione o ***pattern***[6]. Il numero $P \in \mathbb{N}_+$ è detto **carico** del sistema e rappresenta la quantità di informazione che si cerca di apprendere; nel caso in cui sia una quantità estensiva, con lo stesso termine si indica direttamente il rapporto $\alpha \coloneqq \frac{P}{N}$.

2.2.1 Definizione e soluzione in basso carico

Si dice che il sistema si trova in basso carico quando il numero di memorie P è finito, perciò quando il rapporto $\frac{P}{N}$ tende a zero nel limite termodinamico. Risolviamo innanzitutto il caso $P = 1$ cercando un'opportuna matrice simmetrica con diagonale non-negativa J che permetta di recuperare l'informazione $\xi \in \Sigma$. Dovendo scegliere l'efficacia delle sinapsi vogliamo massimizzarla tra i siti che assumono lo stesso valore in ξ, cioè tra i neuroni la cui attività nella memoria in questione è altamente correlata; sia quindi $J_{ij} \coloneqq \frac{1}{N}\xi_i\xi_j$ la matrice delle interazioni. Consideriamo il sistema con dinamica sequenziale, nel caso di dinamica in parallelo il ragionamento sarebbe simile. La funzione di Lyapunov associata è

[6]Per tradurre questo termine dall'inglese, in italiano utilizziamo vari nomi a seconda del contesto: *modello*, *motivo* (di una melodia o di una tessitura), *schema* o *matrice*. Astraendo il significato potremo dire che indica un'informazione strutturata, magari regolare, presa come paradigma di un concetto o di un oggetto. Nell'elaborato preferiremo la voce inglese, poiché useremo il termine modello per riferirci alla formalizzazione matematica di un fenomeno e gli altri termini richiamano contesti troppo specifici.

quella dell'equazione (2.3), cioè

$$L(\sigma) = -\frac{1}{2N}\sum_{i,j}\xi_i\sigma_i\xi_j\sigma_j$$

Prima di continuare si osservi che l'assenza di auto-interazioni $J_{ii} = 0 \forall i$ cambia la funzione solo di una costante siccome uno stato è sempre positivamente correlato a sé stesso, pertanto non influisce sulla localizzazione dei minimi. Più avanti escluderemo la presenza di auto-interazioni ma per quanto appena detto il modello resterà equivalente. Verifichiamo che gli stati $\pm\xi$ sono punti fissi della dinamica mostrando che minimizzano la funzione di Lyapunov: $L(\pm\xi) = -\frac{1}{2N}\sum_{i,j}(\pm\xi_i)^2(\pm\xi_j)^2 = -\frac{N^2}{2N} = -\frac{N}{2} \leq L(\sigma)\forall\sigma$. Notiamo che oltre a ξ abbiamo reso attrattore anche il suo opposto; questi stati si dicono **memorie fondamentali**. In questo caso semplice possiamo analizzare la dinamica più dettagliatamente, a partire dall'equazione (2.2) che diventa

$$\sigma_i(t+1) = \operatorname{sgn}(\xi_i)\operatorname{sgn}(\frac{1}{N}\xi\cdot\sigma(t)) = \xi_i\operatorname{sgn}(\xi\cdot\sigma(t))$$

Possiamo riscrivere la dinamica in termini delle variabili $\tau_i \coloneqq \sigma_i\xi_i$ - un procedimento detto *trasformazione di Gauge* (si veda [Nis01, sez. 4.2]) - ottenendo $\tau_i(t+1) = \operatorname{sgn}(\frac{1}{N}\sum_j \tau_j(t)) = \operatorname{sgn}(m_N(t))$ dove abbiamo indicato con m_N la magnetizzazione nelle variabili τ. Da quest'equazione si evince che, a seconda della magnetizzazione dello stato iniziale $\tau(0)$, la dinamica convergerà verso $+1$ o -1. In termini delle variabili σ significa che gli stati possono convergere verso l'attrattore ξ o il suo opposto, a seconda del segno della seguente quantità:

$$\mathrm{q}(\sigma,\xi) \coloneqq \frac{1}{N}\sigma\cdot\xi \in [-1,1] \tag{2.4}$$

Questa quantità si dice **sovrapponibilità** - in inglese *overlap* - tra gli stati σ e ξ e misura il grado di similarità tra i due stati; infatti $\mathrm{q} = 1 \iff \sigma = \xi$ e analogamente $\mathrm{q} = -1 \iff \sigma = -\xi$. Riassumendo: il sistema così impostato ha due soli attrattori, verso i quali convergono tutti gli stati iniziali dirigendosi all'uno o all'altro a seconda della similarità. Ciò significa che il sistema riesce a ricondurre un "ricordo" σ alla "memoria" ξ se gli è sufficientemente simile.

Il caso generale di P finito si affronta in modo analogo. Innanzitutto assumiamo che i pattern da memorizzare siano ortogonali fra loro: $\frac{1}{N}\xi^\mu\cdot\xi^\nu = \delta_{\mu,\nu}$; così facendo non si perde generalità perché possiamo sempre preventivamente effettuare una scomposizione in una base ortogonale, al limite aumentando il numero di pattern da memorizzare. In questo caso l'efficacia delle sinapsi si imposta in modo simile al precedente, secondo la cosiddetta

regola di Hebb.

$$J_{ij} \coloneqq \frac{1}{N}\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \tag{2.5}$$

In questo caso la funzione di Lyapunov si scrive

$$\begin{aligned} L(\sigma) &= -\frac{1}{2}\sum_{i,j}\frac{1}{N}\sum_{\mu}\sigma_i\xi_i^\mu\xi_j^\mu\sigma_j = -\frac{1}{2N}\sum_{\mu}(\sum_{i}\sigma_i\xi_i^\mu)^2 \\ &= -\frac{N}{2}\sum_{\mu}\mathrm{q}(\sigma,\xi^\mu)^2 = -\frac{1}{2}\sum_{\mu}(\sigma\cdot\frac{\xi^\mu}{\sqrt{N}})^2 \end{aligned} \tag{2.6}$$

Verifichiamo che le memorie fondamentali ξ^μ sono suoi minimi. Siano $\Xi^\mu \coloneqq \frac{\xi^\mu}{\sqrt{N}}$ i pattern normalizzati - rispetto alla norma associata al prodotto scalare di $\mathbb{R}^N$ usato fino ad ora. Siccome posso completarli ad una base ortonormale vale $\sum_{\mu=1}^{P}(\sigma\cdot\Xi^\mu)^2 \leq \sum_{\mu=1}^{N}(\sigma\cdot\Xi^\mu)^2 = \sigma\cdot\sigma\ \forall\sigma\in\Sigma$, con cui calcoliamo

$$\begin{aligned} L(\sigma) &\geq -\frac{1}{2}\sigma\cdot\sigma = -\frac{N}{2} \\ L(\xi^\nu) &= -\frac{1}{2N}\sum_{\mu}(\xi^\nu\cdot\xi^\mu)^2 = -\frac{1}{2N}N^2 = -\frac{N}{2} \quad \forall\nu = 1,\dots,P \end{aligned}$$

Dunque, tramite la regola di Hebb abbiamo costruito un sistema dinamico deterministico che memorizza più pattern. Si dice anche *modello di memoria associativa* perché la convergenza verso le memorie dipende dalla loro similarità con lo stato iniziale. Facciamo presente che non sempre la rete è in grado di recuperare la memoria; infatti, oltre alle memorie fondamentali e i loro opposti esistono altri stati attrattori che non corrispondono ad alcun ricordo, detti **stati misti**. La situazione si complica se i pattern non sono ortogonali poiché in quel caso anche le memorie fondamentali possono destabilizzarsi e non fungere più da attrattori.

Quando abbiamo modellato la dinamica di un neurone con l'equazione (2.1) abbiamo trascurato alcuni dettagli biologici e l'interazione con l'ambiente; per considerare anche questi fattori aggiungiamo una componente di fluttuazione stocastica dipendente dalla variabile aleatoria $z_i(t)$, dove $\{z_i(t)\}_{i=1,\dots,N;t\in\mathbb{N}_+}$ saranno variabili indipendenti e identicamente distribuite - d'ora in avanti abbreviato con i.i.d.. La dinamica aleatoria del neurone diventa

$$\sigma_i(t+1) \coloneqq \mathrm{sgn}(J_i\cdot\sigma(t) - U_i + \frac{z_i(t)}{\beta}) \tag{2.7}$$

con $\beta\in\mathbb{R}_+$ parametro che regola la sua variabilità, chiamato **rumore** del sistema o **temperatura inversa** in analogia con il formalismo della meccanica statistica visto nel

capitolo 1. Si noti che per $\beta \to +\infty$ si ottiene il sistema deterministico appena studiato; al contrario, per $\beta \to 0^+$ il rumore diventa talmente forte che le configurazioni delle sinapsi risultano essere irrilevanti: i neuroni si attivano e disattivano con probabilità uniforme. Le tecniche utilizzate per lo studio della dinamica deterministica non sono appropriate in questo contesto perché questa formulazione non ammette punti fissi: c'è sempre la possibilità che un neurone cambi stato; perciò usiamo gli strumenti della sezione 1.3: interpretiamo il sistema come una catena di Markov omogenea e ricaviamo la distribuzione di equilibrio del sistema; conoscendo questa potremo studiare in quali circostanze le memorie fondamentali diventano configurazioni ad alta probabilità, l'equivalente degli stati attrattori nel caso deterministico che svolgono il ruolo di ricordi. Prima di cominciare l'analisi è necessario fare qualche ipotesi: innanzitutto abbiamo già spiegato che non perdiamo generalità considerando $J_{ii} = 0 \forall i$; inoltre richiederemo anche $U_i = 0 \forall i$ e assumeremo che la distribuzione delle $z_i(t)$, indicata con μ_z, abbia densità $(1 - \tanh(z)^2)/2$ - un'analisi analoga si può fare in generale per distribuzioni simmetriche. In queste condizioni possiamo esprimere la matrice di transizione tramite la funzione di ripartizione di μ_z: $F(z) = (1 + \tanh(z))/2 = 1/(1 + e^{-2z})$. Infatti, definendo $h_i(\sigma) \coloneqq J_i \cdot \sigma$, una volta scelto il neurone i da aggiornare si può esprimere la probabilità di cambiare il suo stato nel modo seguente, con x variabile dicotomica:

$$\begin{aligned} \mathrm{P}(\sigma_i(t+1) = x | \sigma(t)) &= \mathrm{P}(x[h_i(\sigma(t)) + \beta^{-1} z_i(t)] > 0) = \mathrm{P}(x z_i(t) > -x\beta h_i(\sigma(t))) \\ &= \mathrm{P}(z_i(t) < x\beta h_i(\sigma(t))) = F(x\beta h_i(\sigma(t))) \end{aligned}$$

L'ipotesi di simmetria è stata sfruttata nella penultima uguaglianza. Osserviamo che la dinamica neuronale qui descritta è quella di Glauber delineata nell'esempio 1.23; per accorgersene è sufficiente considerare l'hamiltoniana $\mathcal{H} = -\sum_{j,k=1}^{N} J_{jk}\sigma_j\sigma_k/2$, infatti sfruttando la simmetria di J_{ij} si ricava che la differenza energetica corrispondente è

$$\Delta\,\mathcal{H} = -\frac{1}{2}\left(\sum_{j\neq i} J_{ji}\sigma_j(-2\sigma_i) + \sum_{k\neq i} J_{ik}\sigma_k(-2\sigma_i)\right) = 2\sum_j J_{ij}\sigma_i\sigma_j = 2\sigma_i h_i$$

Da questa identificazione sappiamo che la dinamica converge alla distribuzione di Boltzmann-Gibbs relativa alla seguente funzione energia:

$$\mathcal{H}(\sigma) \coloneqq -\frac{1}{2}\sum_{i,j=1}^{N} \sigma_i J_{ij}\sigma_j = -\frac{1}{2N}\sum_{i,j,\mu} \xi_i^\mu \xi_j^\mu \sigma_i\sigma_j + \frac{P}{2} \simeq L(\sigma) \tag{2.8}$$

Notiamo che è la funzione di Lyapunov (2.6); si tenga presente che in questa formulazione non sono presenti le auto-interazioni, cioè $J_{ij} = \sum_\mu (1 - \delta_{ij})\xi_i^\mu \xi_j^\mu / N$, mentre nel calcolare

l'altra equazione le avevamo incluse. Il fatto che le funzioni energia nel caso deterministico e in quello stocastico coincidano non è casuale: entrambe le dinamiche sono state progettate per minimizzare proprio quella funzione. Il modello meccanico-statistico di equilibrio dato da questa hamiltoniana prende il nome di **modello di Hopfield**; come già anticipato, lo studieremo nel limite termodinamico per ricavarne le capacità mnemoniche. Ricordiamo che la media rispetto alla distribuzione di equilibrio prende il nome di media termica. Le caratteristiche del modello dipendono dal carico del sistema, perciò differenzieremo l'analisi in casi e sfrutteremo l'occasione per introdurre man mano delle tecniche che torneranno utili nel seguito dell'elaborato. Nel caso in cui $P = 1$, trattando la controparte deterministica abbiamo già visto che tramite una trasformazione di Gauge l'hamiltoniana diventa un modello di Curie-Weiss con interazioni uniformi e senza campo esterno: l'esempio 1.11 sulla linea di transizione di fase; ciò significa che la sovrapponibilità con il pattern, che in questo caso è il parametro d'ordine del sistema, ottenuto dalla magnetizzazione media invertendo la trasformazione di Gauge, tenderà a ± 1 solamente al di sotto di una certa temperatura critica, altrimenti si approssimerà a zero. Già da questo primo caso emerge una transizione di fase nel recupero della memoria, dipendente dal rumore presente nelle comunicazioni sinaptiche. Ora supponiamo che $P \in \mathbb{N}_+$ sia fissato, cosicché il rapporto P/N tenda a 0 nel limite termodinamico; in questo caso si parla di basso carico del sistema e possiamo trascurare il termine $P/2$ nell'hamiltoniana. Prima di calcolare l'energia libera del sistema semplifichiamo l'espressione della funzione di partizione. Da (2.8) (rinominando la variabile muta ρ in σ) ricaviamo

$$\mathcal{Z}_{N,\beta,\{\xi^\mu\}} = \sum_{\sigma\in\Sigma} \exp\left(\frac{\beta}{2N}\sum_\mu (\sum_i \xi_i^\mu \sigma_i)^2\right) \tag{2.9}$$

Sottolineiamo il fatto che la presenza dell'elevamento alla seconda rende il sistema interagente: l'energia risente del grado di accoppiamento di spin diversi. Se questo termine fosse lineare il sistema sarebbe l'unione di spin non interagenti e la somma esterna si potrebbe fattorizzare sui singoli neuroni rendendo agevole il calcolo:

$$\sum_{\sigma\in\Sigma} \exp\left(\sum_i \dots \sigma_i\right) = \sum_{\sigma\in\Sigma}\prod_i \exp\left(\dots \sigma_i\right) = \prod_i \sum_{\sigma_i=\pm 1} \exp\left(\dots \sigma_i\right)$$

Si tratta in sostanza della differenza tra l'esempio 1.10 e l'esempio 1.11. Alla luce di questa osservazione introduciamo un procedimento detto **linearizzazione gaussiana**, o di **Hubbard-Stratonovich**, che ci permette di rendere gli elevamenti alla seconda trattabili come campi esterni; si basa sulla seguente proprietà della distribuzione normale.

Lemma 2.4 — *Sia* $\mathrm{N}(0,\sigma^2)$ *la distribuzione gaussiana con media nulla e deviazione standard* $\sigma \in \mathbb{R}_+$*, la cui densità è*

$$\gamma(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{z^2}{2\sigma^2})$$

Indicando con $\mathbb{E}_{z \overset{d.}{\sim} \mathrm{N}(0,\sigma^2)} [\bullet]$ *la media rispetto a questa distribuzione, per ogni* $a \in \mathbb{R}$ *vale*

$$\exp(\frac{\sigma^2 a^2}{2}) = \mathbb{E}_{z \overset{d.}{\sim} \mathrm{N}(0,\sigma^2)} [e^{az}] \tag{2.10}$$

L'equazione si esprime anche nella forma seguente, dove $b \in \mathbb{R}_+$*:*

$$\exp(b\,\frac{a^2}{2}) = \int \mathrm{d}z\, \frac{\sqrt{b}}{\sqrt{2\pi}} \exp(baz - \frac{bz^2}{2}) = \mathbb{E}_{z \overset{d.}{\sim} \mathrm{N}(0,1/b)} [\exp(baz)] \tag{2.11}$$

L'equazione (2.10) *vale anche per* $a \in \mathbb{C}$*, intendendo le medie separatamente sulle parti reale e immaginaria dei numeri complessi.*

Dimostrazione. Si tratta di un conto che sfrutta il completamento del quadrato:

$$\begin{aligned}
\mathbb{E}_{z \overset{d.}{\sim} \mathrm{N}(0,\sigma^2)} [e^{az}] &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{z^2}{2\sigma^2} + az)\, \mathrm{d}z \\
&= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(z + a\sigma^2)^2}{2\sigma^2} + \frac{\sigma^2 a^2}{2})\, \mathrm{d}z \\
&= \exp(\frac{\sigma^2 a^2}{2})
\end{aligned}$$

La seconda equazione si ottiene dalla prima ponendo $\sigma^2 = 1$ e usando $a = a\sqrt{b}$, poi applicando il cambio di variabile $z \mapsto \sqrt{b}z$.
Nel caso di numeri complessi lo stesso risultato si ottiene sfruttando il teorema dei residui per poter fare un opportuno cambio di variabili e ricondursi ad una gaussiana standard di media unitaria (moltiplicata per la costante richiesta). Un libro sul quale si può approfondire il calcolo dei residui è [Car95]. □

Per applicarlo alla funzione di partizione (2.9) dobbiamo prima raccogliere la sommatoria su μ e successivamente introdurre una variabile di integrazione per ciascun termine $\mu = 1, \dots, P$; chiamandole m^μ e usando l'equazione (2.11) con $b = \beta N$ e $a = 1/N \sum_i \xi_i^\mu \sigma_i$ si trova

$$\mathcal{Z} = \sum_{\sigma \in \Sigma} \prod_\mu \int \mathrm{d}m^\mu \sqrt{\frac{\beta N}{2\pi}} \exp\left(\beta m^\mu \sum_i \xi_i^\mu \sigma_i - \frac{\beta N}{2}(m^\mu)^2\right)$$

Ora uniamo gli integrali e sfruttiamo la linearità per ottenere una forma più semplice; a questa applichiamo la fattorizzazione su i come mostrato sopra.

$$\begin{aligned}
\mathcal{Z} &= \sum_{\sigma\in\Sigma}\int_{\mathbb{R}^P}(\prod_\mu \mathrm{dm}^\mu)\left(\frac{\beta N}{2\pi}\right)^{\frac{P}{2}}\prod_\mu \exp\left(-\frac{\beta N}{2}(m^\mu)^2\right)\exp(\beta m^\mu\sum_i \xi_i^\mu\sigma_i)\\
&= \left(\frac{\beta N}{2\pi}\right)^{\frac{P}{2}}\int_{\mathbb{R}^P}(\prod_\mu \mathrm{dm}^\mu)\exp\left(-\frac{\beta N}{2}\sum_\mu(m^\mu)^2\right)\sum_{\sigma\in\Sigma}\exp(\sum_i\sigma_i\sum_\mu\beta m^\mu\xi_i^\mu)\\
&= \left(\frac{\beta N}{2\pi}\right)^{\frac{P}{2}}\int_{\mathbb{R}^P}(\prod_\mu \mathrm{dm}^\mu)\exp\left(-\frac{\beta N}{2}\sum_\mu(m^\mu)^2\right)\prod_i 2\cosh(\sum_\mu\beta m^\mu\xi_i^\mu)\\
&= \left(\frac{\beta N}{2\pi}\right)^{\frac{P}{2}}\int_{\mathbb{R}^P}(\prod_\mu \mathrm{dm}^\mu)\exp\left(-\frac{\beta N}{2}\sum_\mu(m^\mu)^2+\sum_i\log(2\cosh(\sum_\mu\beta m^\mu\xi_i^\mu))\right)
\end{aligned}$$

Chiamando $C_{N,P}$ la costante davanti all'integrale, m il vettore di componenti m^μ e indicando con $\circ$ il prodotto scalare su $\mathbb{R}^P$, si esprime nella forma

$$\mathcal{Z} = C_{N,P}\int \mathrm{d}\{\mathrm{m}^\mu\}\exp\left(-N\frac{\beta}{2}\|m\|_\circ^2+\sum_i\log(2\cosh(\beta m\circ\xi_i))\right) \tag{2.12}$$

Il calcolo esplicito di questo integrale è complicato, tuttavia ai nostri fini è sufficiente ricavare il suo comportamento asintotico per $N\to+\infty$; per fare questo introduciamo col prossimo lemma una tecnica chiamata **metodo di Laplace**, integrazione su punto di sella (*saddle-point integration* in inglese), o *steepest descent* nel caso più generale.

Lemma 2.5 — *Siano $P\in\mathbb{N}_+$ e $G_N\colon\mathbb{R}^P\to\mathbb{R}$ una successione di funzioni indicizzata da $N\in\mathbb{N}_+$. Chiamiamo $\hat{x}_N$ il massimo di G_N e $\hat{x}$ il massimo di G che è il limite uniforme della successione, assumendo che questi siano unici. Sotto opportune ipotesi di limitatezza, indicando con C una costante, vale la formula asintotica*

$$\mathrm{I_N}[G_N] \coloneqq \int_{\mathbb{R}^P}\mathrm{dx}\, e^{NG_N(x)} \underset{N\to+\infty}{\sim} CN^{-P/2}e^{NG_N(\hat{x}_N)}$$

perciò in particolare

$$\lim_{N\to+\infty}\frac{1}{N}\log\mathrm{I_N}[G_N] = \lim_{N\to+\infty}G_N(\hat{x}_N) = G(\hat{x}) \tag{2.13}$$

Nel caso in cui la funzione G sia a valori complessi, sotto ulteriori ipotesi, vale comunque il limite (2.13) interpretando $\hat{x}$ come un punto estremale per G, ossia un punto nel quale si annullano tutte le derivate - indicato con $\mathrm{argextr}_G$.

Spiegazione. Il caso più semplice è quello in cui, invece di una successione di funzioni, G_N è una funzione di variabile reale indipendente da N. In questo caso la dimostrazione è rigorosa e sfrutta lo sviluppo di Taylor attorno al massimo per ricavare il termine dominante. I dettagli si possono trovare nell'appendice di [CKS05]. Nel caso uniforme sono necessarie alcune ipotesi specifiche riguardo la limitatezza delle funzioni e la loro regolarità. Alcune versioni rigorose di questo risultato si possono trovare in [Alb+16, teo. A.1] e [ER82]. Il caso generale è più delicato e sfrutta risultati di analisi complessa, pertanto rimandiamo il lettore a testi più specialistici quali [Erd65; Won01; But07]. △

Nota 2.6 — Il lettore si sarà accorto che non abbiamo specificato precisamente tutte le ipotesi necessarie per la validità del lemma 2.5. Questo è dovuto al fatto che nelle analisi che seguiranno, soprattutto nel capitolo 3, utilizzeremo dei metodi intrinsecamente non rigorosi che renderebbero vana un'eccessiva precisione in questi passaggi: in ogni caso la validità dei nostri studi dovrà essere confermata dai risultati degli esperimenti numerici; perciò nell'applicare questo lemma ci interessa il significato qualitativo delle espressioni che scriviamo.
A questo si aggiunga che in presenza di molteplici punti estremali non si riesce sempre ad individuare il massimo globale per via analitica; inoltre, i modelli che studieremo sono orientati alla costruzione di algoritmi ed è noto che spesso la ricerca dei massimi per via numerica è condizionata dal punto di partenza della ricerca.

Nel caso in esame, per poter applicare il lemma è necessario che l'argomento dell'esponenziale in (2.12) converga uniformemente. Otterremo questo risultato imponendo che le memorie siano estratte in modo aleatorio, infatti se le componenti dei pattern sono tra loro i.i.d. possiamo applicare la legge dei grandi numeri. Questa ipotesi di aleatorietà non è drastica come può sembrare poiché la nostra analisi del modello sarà incentrata sull'efficienza teorica e quindi andrebbe comunque svolta rispetto a memorie ignote e variegate. Assumiamo quindi che i pattern siano indipendenti e che le componenti di ciascun pattern siano tra loro i.i.d. secondo $\xi_i^\mu \overset{d.}{\sim} p^\mu\,\delta_{+1} + (1-p^\mu)\,\delta_{-1}$ per certi $p^\mu \in [0,1], \mu = 1,\dots,P$ fissati; con la scelta $p^\mu = \frac{1}{2} \forall \mu$ si ottengono dei pattern distribuiti in modo uniforme. In analogia con quanto visto nella sezione 1.4 indicheremo con $\overline{\bullet}$ la media sulle 2^P configurazioni di disordine temprato $(\xi^1,\dots,\xi^P)$ rispetto alla distribuzione introdotta sopra (identica per ogni indice i) e manterremo la denominazione di media termica $\langle\bullet\rangle$ per quella fatta sullo stato dei neuroni. Considerando l'equazione (2.12) in

vista dell'applicazione del lemma 2.5 identifichiamo

$$\mathrm{G_N}(m^1,\dots,m^P) \coloneqq \frac{-\beta}{2}\|m\|_\circ^2 + \frac{1}{N}\sum_i \log(2\cosh(\beta m \circ \xi_i))$$

$$\mathrm{G}(m^1,\dots,m^P) \coloneqq \frac{-\beta}{2}\|m\|_\circ^2 + \overline{\log(2\cosh(\beta m \circ \xi))} \tag{2.14}$$

Queste quantità, quando discendono dal calcolo dell'energia libera di un modello, prendono il nome di **pseudo energia libera** o talvolta azione (ci si riferisce indifferentemente a quella dipendente da N o al suo limite). Come accennavamo sopra, la seconda equazione è ottenuta dalla prima applicando la legge dei grandi numeri sul disordine temprato. Chiamando $\hat{m}_N$ i punti estremali delle $\mathrm{G_N}$ e $\hat{m}$ quello di G applichiamo il lemma 2.5 all'equazione (2.12) ottenendo l'espressione per l'energia libera intensiva del modello:

$$\mathrm{f}_{\beta,\{\xi^\mu\}} = \lim_{N\to+\infty} -\frac{1}{\beta}\frac{1}{N}\log\mathcal{Z} \overset{!}{=} -\frac{1}{\beta}\lim_{N\to+\infty} G_N(\hat{m}_N) = -\frac{1}{\beta}G(\hat{m}) \tag{2.15}$$

I punti estremali possono essere molteplici ma devono tutti soddisfare le equazioni di auto-consistenza - dette anche equazioni di punto di sella o **condizioni stazionarie** - ottenute imponendo l'annullamento delle derivate:

$$\hat{m}_N^\mu = \frac{1}{N}\sum_i \xi_i^\mu \tanh(\beta \hat{m}_N \circ \xi_i) \quad \forall \mu = 1,\dots,P \tag{2.16}$$

$$\hat{m}^\mu = \overline{\xi^\mu \tanh(\beta \hat{m} \circ \xi)} \quad \forall \mu = 1,\dots,P \tag{2.17}$$

Precisiamo che le condizioni stazionarie sono P equazioni; qui sopra sono semplicemente scritte in due forme: una dipendente da N e l'altra è il suo limite per $N \to +\infty$; l'equivalenza si ottiene grazie alla legge dei grandi numeri (la distribuzione è i.i.d. sugli spin). Per dare un significato alle soluzioni delle condizioni stazionarie si procede come nella parte finale dell'esempio 1.11: fissato $\nu \in \{1,\dots,P\}$, perturbiamo l'hamiltoniana del modello di Hopfield (2.8) con un termine $-\lambda N q_N(\xi^\nu,\sigma)$ così da poter ottenere questa sovrapponibilità (definita in (2.4)) come derivata dell'energia libera; poi, scambiamo il limite termodinamico e la derivata. Innanzitutto esplicitiamo la funzione di partizione che estende quella presentata in (2.9):

$$\mathcal{Z}_{N,\lambda,\nu} = \sum_{\sigma\in\Sigma} \exp\left(\frac{\beta}{2N}\sum_\mu (\sum_i \xi_i^\mu \sigma_i)^2 + \lambda\beta \sum_i \xi_i^\nu \sigma_i\right)$$

Svolgendo conti analoghi a quelli presentati poc'anzi nel caso $\lambda = 0$ (la linearizzazione di questo nuovo termine non è necessaria), si ricava l'andamento asintotico dell'energia libera intensiva:

$$\mathrm{f}_{\lambda,N} \underset{N\to+\infty}{\overset{!}{\sim}} \frac{\|\hat{m}_N\|^2}{2} - \frac{1}{\beta N}\sum_i \log(2\cosh(\beta \hat{m}_N \circ \xi_i + \lambda\beta\xi_i^\nu))$$

Qui $\hat{m}_N$ è il punto estremale dell'espressione e dipende esso stesso da λ oltre che da N, pertanto l'espressione qui sopra è nella forma $T_N(\lambda, \hat{m}_N(\lambda))$. Adesso calcoliamo le derivate dell'energia libera rispetto al parametro che abbiamo aggiunto: come visto nella dimostrazione del teorema 1.12, per definizione di punto estremale vale

$$d_\lambda|_0 T_N = \partial_\lambda|_0 T_N = -\frac{1}{N}\sum_i \tanh(\beta \hat{m}_N \circ \xi_i)\xi_i^\nu = -\hat{m}_N^\nu$$

dove nell'ultimo passaggio abbiamo usato l'equazione di auto-consistenza (2.16), siccome per $\lambda = 0$ otteniamo proprio il punto estremale ricavato nel caso precedente. D'altra parte, per il corollario 1.8 sappiamo che $\langle q_N(\xi^\nu, \sigma)\rangle = -d_\lambda|_0 \, \mathrm{f}_{\lambda,N}$ perciò, assumendo di poter scambiare i limiti con le derivate, si ottiene

$$\langle q_N(\xi^\nu, \sigma)\rangle \underset{N\to+\infty}{\overset{!}{\sim}} -d_\lambda|_0 \, \mathrm{f}_{\lambda,N} \underset{N\to+\infty}{\sim} -d_\lambda|_0 T_N \underset{N\to+\infty}{\overset{!}{\sim}} \hat{m}_N^\nu \underset{N\to+\infty}{\to} \hat{m}^\nu$$

Da questo si evince che nel limite termodinamico $\hat{m}^\nu$ segnala la sovrapponibilità con la memoria ξ^ν, cioè coincide col parametro d'ordine $\langle q_N(\xi^\nu, \sigma)\rangle$ - detto **magnetizzazione di Mattis** - che misura la qualità del riconoscimento di quel pattern. Come già detto nel caso dell'esempio 1.11, ci permetteremo di identificare le variabili che descrivono i punti estremali con i parametri d'ordine che rappresentano.
Tra l'altro, le stesse condizioni stazionarie ottenute per $\hat{m}$ si possono trovare per le magnetizzazione di Mattis partendo dalla dinamica stocastica (2.7): scriviamo l'evoluzione temporale di questa magnetizzazione

$$\begin{aligned}\mathbb{E}_{\mathrm{P}(\sigma(t+1)|\sigma(t))}\left[q_N(\sigma(t+1), \xi^\mu)\right] &= \frac{1}{N}\sum_i \xi_i^\mu \, \mathbb{E}_{\mathrm{P}(\sigma_i(t+1)|\sigma(t))}\left[\sigma_i(t+1)\right] \\ &= \frac{1}{N}\sum_i \xi_i^\mu \tanh(\beta \sum_\mu \frac{\xi_i^\mu}{N}\sum_j \xi_j^\mu \sigma_j(t))\end{aligned}$$

e ne facciamo la media rispetto alla distribuzione d'equilibrio per $\sigma(t)$, ottenendo

$$\langle q_N(\sigma, \xi^\mu)\rangle = \frac{1}{N}\sum_i \xi_i^\mu \langle \tanh(\beta q_N(\xi^\mu, \sigma) \circ \xi_i)\rangle$$

Specifichiamo che nel membro di sinistra dell'ultima equazione avremmo ottenuto la media della magnetizzazione rispetto alla distribuzione di $\sigma(t+1)$, ma trattandosi di un equilibrio coincide con la media sulla distribuzione di $\sigma(t)$.
Riconosciamo l'equazione di auto-consistenza (2.16), dove al posto delle variabili $\hat{m}_N^\mu$ sono presenti proprio le sovrapponibilità $q_N(\sigma, \xi^\mu)$ e il tutto è mediato rispetto alla distribuzione di equilibrio. Come vedremo successivamente - durante l'analisi del sistema in alto carico,

dopo le equazioni (2.30) - nel limite termodinamico non solo le magnetizzazioni di Mattis tendono al punto estremale $\hat{m}_N$, ma siccome automediano pure le medie delle osservabili da queste dipendono tendono al loro valore nel punto estremale; ne segue che nel limite termodinamico l'equazione appena scritta coincide esattamente con (2.17).

Finita la piccola digressione sul significato delle equazioni di auto-consistenza concentriamoci sulla ricerca delle soluzioni di (2.17) in termini del vettore $\hat{m}$; ci restringeremo al caso di un disordine uniforme, cioè $p^\mu = 1/2 \forall \mu$. Il caso più semplice da analizzare è quello in cui viene memorizzato un solo pattern, assumiamo che sia il primo (altrimenti li permutiamo): $\hat{m}^1 = m, \hat{m}^\mu = 0 \forall \mu \geq 2$; in questo caso le equazioni per gli indici $\mu = 2, \dots, P$ sono banalmente vere, quella per $\mu = 1$ va verificata:

$$m = \overline{\xi^1 \tanh(\beta m \xi^1)} = \tanh(\beta m) \tag{2.18}$$

Così scritta si riconosce l'equazione di auto-consistenza del modello di Curie-Weiss, perciò dall'esempio 1.11 ricaviamo il comportamento del modello: a temperature sufficientemente basse (precisamente $\beta > 1$) il modello converge verso uno stato stabile con le sovrapponibilità prescritte; quando $\beta \to +\infty$ lo stato stabile tende ad un perfetto recupero della prima memoria fondamentale (o del suo opposto) indipendentemente dalla condizione iniziale. Questo risultato è coerente con quanto visto nel caso $P = 1$ e prova che il modello di Hopfield funziona come sistema di memoria associativa. All'abbassarsi della temperatura anche alcuni stati misti diventano stabili; ad esempio, tra le soluzioni delle condizioni stazionarie appaiono quelle nella forma $\hat{m} = (m, \dots, m, 0, \dots, 0)$ che recuperano un numero dispari di memorie, oppure soluzioni non uniformi. Osserviamo quindi che una piccola variabilità nella dinamica (rumore β poco superiore a 1) è utile a destabilizzare questi stati misti e a rendere più agevole il recupero delle memorie fondamentali. Per una trattazione più approfondita delle soluzioni di queste equazioni rimandiamo a [CKS05, sezione 21.1].

2.2.2 Soluzione in alto carico

L'ultimo caso che rimane da trattare è quello di alto carico, quando $P = \alpha N$ con $\alpha \in \mathbb{R}_+$. Consideriamo da subito un sistema con disordine temprato, perciò assumeremo che i pattern siano estratti con distribuzione uniforme come fatto in precedenza. Come visto nella sezione 1.4, il nostro obiettivo è calcolare l'energia libera temprata $\overline{\mathrm{f}_{\beta,N}(\{\xi^\mu\}_{\mu=1}^P)} = -1/(\beta N)\overline{\log \mathcal{Z}(\{\xi^\mu\}, N, \beta)}$ nel limite termodinamico. Tuttavia, il logaritmo all'interno della media sul disordine rende il calcolo di difficile esecuzione. Per ovviare

a questo problema ricorreremo ad una tecnica chiamata **trucco delle repliche**[7]; prima di specializzarla nel caso del modello di Hopfield la introduciamo in generale, poiché risulterà utile anche in seguito. Supponiamo di considerare un sistema di hamiltoniana $\mathcal{H}(\sigma; y)$, la cui funzione di partizione verrà indicata con $\mathcal{Z}(y)$ e la media sul disordine $\mathbb{E}_y[\bullet]$. Il calcolo si basa sulla seguente approssimazione per $\mathbb{R}_+ \ni k \to 0^+$: $\mathcal{Z}^k \sim 1 + k \log \mathcal{Z}$, da cui segue che $\log \mathbb{E}_y\left[\mathcal{Z}^k\right] \sim \log(1 + k\,\mathbb{E}_y[\log \mathcal{Z}]) \sim k\,\mathbb{E}_y[\log \mathcal{Z}]$ e quindi

$$\mathbb{E}_y[\log \mathcal{Z}] = \lim_{k \to 0^+} \frac{1}{k} \log \mathbb{E}_y\left[\mathcal{Z}^k\right] \tag{2.19}$$

Così siamo riusciti a portare la media all'interno del logaritmo, rendendo il calcolo più agevole. Precisiamo che fino a questo punto i passaggi sono legittimi poiché il valor medio non è altro che un integrale normalizzato e che l'equazione appena scritta è applicabile su una qualunque distribuzione di y e una qualunque funzione $\mathcal{Z}$ sufficientemente regolare. Rimane da esprimere la potenza della funzione di partizione in una forma più semplice, per far questo assumeremo $k \in \mathbb{N}_+$. Una volta ottenuta un'espressione semplificata, il passaggio non rigoroso consisterà nel calcolo del limite per $\mathbb{R}_+ \ni k \to 0^+$. Nonostante questa tecnica non abbia una giustificazione formale, al momento ha sempre condotto a dei risultati verificati numericamente o, in qualche caso, matematicamente dimostrati per altre vie. Se $k \in \mathbb{N}_+$ possiamo esprimere il prodotto di sommatorie come un'unica sommatoria su più indici

$$\mathcal{Z}^k = \sum_{\sigma^1, \ldots, \sigma^k \in \Sigma} \exp\left(-\beta \sum_{\gamma=1}^{k} \mathcal{H}(\sigma^\gamma; y)\right)$$

A questo punto bisogna calcolare la media sul disordine. Supponiamo di riuscire a esprimere $\mathbb{E}_y\left[\exp(-\beta \sum_{\gamma=1}^k \mathcal{H}(\sigma^\gamma; y))\right] = \exp(-\beta \Phi(\sigma^1, \ldots, \sigma^k))$ con una certa Φ che aggrega le k **repliche** del sistema originario; allora il limite termodinamico dell'energia libera intensiva temprata assume la forma

$$\lim_{N \to +\infty} \lim_{k \to 0^+} \frac{-1}{\beta k N} \log \sum_{\{\sigma^\gamma\}_{\gamma=1}^k} \exp(-\beta \Phi(\sigma^1, \ldots, \sigma^k)) \tag{2.20}$$

Scambiando i due limiti - tenendo presente che nemmeno questo in generale è un passaggio matematicamente giustificato - ci riconduciamo al calcolo della funzione di partizione di un sistema senza disordine con hamiltoniana $\Phi(\{\sigma^\gamma\}_{\gamma=1,\ldots,k})$, perciò per risolverlo possiamo

[7] Dall'inglese *replica trick*. Con un po' di malizia il termine trick si potrebbe tradurre anche con imbroglio, a memoria del fatto che il metodo non è rigoroso. Per una trattazione approfondita di questa tecnica e una discussione delle condizioni di applicabilità rimandiamo a [MPV86].

usare le tecniche viste in precedenza e successivamente calcolare il limite in k - augurandoci che sia possibile e che il risultato sia corretto.

Passiamo ora al caso specifico del modello di Hopfield. Al fine di concentrare l'attenzione sulle tecniche a noi utili semplificheremo il calcolo assumendo - similmente al caso in basso carico - che solamente il primo pattern venga memorizzato; il caso del recupero di un numero finito di memorie viene trattato in [CKS05]. Questa restrizione del campo di indagine ci permetterà comunque di distinguere le diverse fasi del sistema.
Come visto sopra partiamo dall'equazione (2.8) e cerchiamo la relativa funzione Φ; in un primo momento calcoliamo la media sul disordine solo relativamente alle memorie $\xi^\mu, \mu > 1$.

$$\overline{\exp\left[-\beta\sum_{\gamma=1}^{k}\left(-\frac{1}{2N}\sum_{\mu}(\sum_{i}\sigma_i^\gamma\xi_i^\mu)^2+\frac{P}{2}\right)\right]} = e^{-\frac{\beta}{2}kP}\overline{\prod_{\mu}\exp\left(\frac{\beta}{2N}\sum_{\gamma}(\sum_{i}\sigma_i^\gamma\xi_i^\mu)^2\right)}$$

Siccome abbiamo scelto il disordine in modo uniforme, il prodotto in μ fattorizza nel termine con ξ^1 (che non ha la media sul disordine) moltiplicato per $P-1$ termini uguali. In questi $P-1$ termini possiamo linearizzare il quadrato rimanente con il metodo di Hubbard-Stratonovich descritto nel lemma 2.4

$$\begin{aligned}
&\overline{\int \mathrm{d}\{\mathrm{z}^\gamma\}(2\pi)^{-k/2}\exp\left(\sum_\gamma z^\gamma\sqrt{\frac{\beta}{N}}\sum_i\sigma_i^\gamma\xi_i^\mu-\frac{(z^\gamma)^2}{2}\right)} = \\
&= (2\pi)^{-k/2}\int \mathrm{d}\{\mathrm{z}^\gamma\}\, e^{-\frac{1}{2}\|z\|^2}\overline{\exp\left(\sqrt{\frac{\beta}{N}}\sum_i\xi_i^\mu\sum_\gamma z^\gamma\sigma_i^\gamma\right)} \\
&= (2\pi)^{-k/2}\int \mathrm{d}\{\mathrm{z}^\gamma\}\, e^{-\frac{1}{2}\|z\|^2}\prod_i\cosh\left(\sqrt{\frac{\beta}{N}}\sum_\gamma z^\gamma\sigma_i^\gamma\right) \\
&= (2\pi)^{-k/2}\int \mathrm{d}\{\mathrm{z}^\gamma\}\exp\left(-\frac{1}{2}\|z\|^2+\sum_i\log\cosh\sqrt{\frac{\beta}{N}}\sum_\gamma z^\gamma\sigma_i^\gamma\right)
\end{aligned}$$

Per ricavare la seconda equazione abbiamo scambiato l'integrale con la media (che è discreta) e per arrivare alla terza abbiamo nuovamente sfruttato l'uniformità della distribuzione del disordine esprimendola tramite il coseno iperbolico. Ricomponendo i pezzi e introducendo la media sul disordine del primo pattern ξ^1 - sempre indicata con $\overline{\bullet}$ perché a questo punto la media sugli altri è stata calcolata - con le notazioni introdotte nell'equazione (2.20) la

funzione Φ del modello di Hopfield è

$$\Phi(\sigma^1,\dots,\sigma^k) = \frac{1}{2}kP + \frac{Pk}{2\beta}\log(2\pi) - \frac{1}{2N}\sum_\gamma \overline{\left(\sum_i \sigma_i^\gamma \xi_i^1\right)^2} \\ - \frac{P-1}{\beta}\log\int \mathrm{d}\{\mathrm{z}^\gamma\}\exp\left(-\frac{\|z\|^2}{2} + \sum_i \log\cosh\sqrt{\frac{\beta}{N}}\sum_\gamma z^\gamma\sigma_i^\gamma\right)$$

Prima di procedere oltre facciamo alcune precisazioni. Innanzitutto si osservi che, per una configurazione σ fissata, la variabile aleatoria $1/N\sum_i \sigma_i\xi_i^\mu$ ha media nulla e varianza $1/\sqrt{N}$. Questo significa che nel caso precedente, siccome i pattern erano un numero finito, i termini relativi alle memorie che non venivano recuperate ($\mu > 1$) non erano determinanti nel limite termodinamico; infatti, rispetto alla configurazione del sistema questi pattern risultavano disallineati e quindi del tutto aleatori. Ora invece, poiché si sommano un numero di contributi di ordine N, anche queste interazioni disordinate devono essere prese in considerazione. In più, se il sistema si allinea con una memoria (la prima nel nostro caso), rispetto a questa la sovrapponibilità convergerà verso un valore deterministico di ordine 1. Alla luce di questa differenza abbiamo separato questo pattern dagli altri proprio per evidenziare il suo contributo. Cerchiamo ora di semplificare l'espressione, partendo dall'integrale nell'ultimo termine. A tal fine facciamo un'ulteriore assunzione non rigorosa[8]: sostituiamo il logaritmo del coseno iperbolico con la sua espansione per $N \to +\infty$ (si veda il lemma 1.3)

$$\log\cosh\sqrt{\frac{\beta}{N}}\sum_\gamma z^\gamma\sigma_i^\gamma \underset{N\to+\infty}{\sim} \frac{\beta}{2N}\sum_{\gamma,\tilde\gamma} z^\gamma z^{\tilde\gamma}\sigma_i^\gamma\sigma_i^{\tilde\gamma}$$

quindi l'integrale diventa

$$\int \mathrm{d}\{\mathrm{z}^\gamma\}\exp\left(-\frac{\|z\|^2}{2} + \frac{\beta}{2N}\sum_{\gamma,\tilde\gamma} z^\gamma z^{\tilde\gamma}\sum_i \sigma_i^\gamma\sigma_i^{\tilde\gamma}\right) = \\ = \int \mathrm{d}\{\mathrm{z}^\gamma\}\exp(-\frac{1}{2}z^{\mathrm{T}}(\mathrm{Id} - \beta\,\mathrm{Q})z)$$

con la notazione $\mathrm{Q}_{\gamma\tilde\gamma} \coloneqq \frac{1}{N}\sum_i \sigma_i^\gamma\sigma_i^{\tilde\gamma}$ per la matrice di sovrapposizione delle repliche e usando la scrittura $z^{\mathrm{T}}Az$ per indicare la norma di z rispetto alla forma bilineare definita dalla matrice A. Un integrale gaussiano in questa forma si può esprimere più esplicitamente, come mostra il seguente lemma.

[8] Avendo una sommatoria con N termini non basta semplicemente espandere ogni termine per $N \to +\infty$ come si potrebbe fare nel caso finito. Procediamo lo stesso, confortati dal fatto che l'espansione di ciascun termine risulta in un fattore di ordine $1/N$, perciò il contributo della sommatoria è di ordine 1.

Lemma 2.7 — *Data una matrice $d \times d$, simmetrica e definita positiva A, vale la seguente uguaglianza:*

$$\int_{\mathbb{R}^d} \mathrm{d}\{\mathrm{z_k}\}\, e^{-z^{\mathrm{T}} A z/2} = \sqrt{\frac{(2\pi)^d}{\det A}}$$

Dimostrazione. Non è altro che un cambio di variabili per un integrale gaussiano multidimensionale. Si veda ad esempio [Pas20], oppure [CKS05, app. D.1]. □

Nel nostro caso si ottiene $(2\pi)^{k/2} \det(\mathrm{Id} - \beta\, \mathrm{Q})^{-1/2}$; sostituendo il risultato nell'equazione (2.20) si trova

$$\overline{\mathrm{f}_\beta(\{\xi^\mu\})} = \frac{\alpha}{2} - \lim_{k\to 0^+} \frac{1}{\beta k} \lim_{N\to+\infty} \frac{1}{N} \log \sum_{\{\sigma^\gamma\}_{\gamma=1}^k} \exp\left(\frac{\beta}{2N} \sum_\gamma \overline{\left(\sum_i \sigma_i^\gamma \xi_i^1 \right)^2} \right) \det(\mathrm{Id} - \beta\, \mathrm{Q})^{-(P-1)/2} \quad (2.21)$$

Il calcolo del limite in questa forma non è affatto agevole per via della sommatoria sulle configurazioni, perciò ci serviamo della teoria delle distribuzioni in modo tale da esprimere l'argomento del logaritmo tramite un integrale, al quale potremo successivamente applicare il metodo del punto di sella. Della teoria in realtà è sufficiente sfruttare una rappresentazione integrale della delta di Dirac, che viene richiamata anche in [CKS05, appendice F]; tuttavia chi fosse interessato ad un'introduzione approfondita può consultare i manuali [Hör63; Rud91]. Nel seguito esponiamo brevemente questi concetti per chi non ne fosse a conoscenza.

Fissiamo un punto $\bar{x} \in \mathbb{R}^d$. Si può definire la delta di Dirac in $\bar{x}$, indicata con $\delta_{\bar{x}}$, come un funzionale su uno spazio di funzioni $\{\varphi\colon \mathbb{R}^d \to \mathbb{R}\}$ (sufficientemente regolari, ad esempio $\mathcal{C}^\infty_{\mathrm{comp.}}$) che agisce in questo modo: $\varphi \overset{\delta_{\bar{x}}}{\mapsto} \varphi(\bar{x})$. I funzionali che agiscono su questo insieme di funzioni si dicono *distribuzioni*, quindi formalmente la delta di Dirac è una distribuzione. In quanto segue, l'*azione* di una distribuzione su una funzione - cioè la valutazione della prima sulla seconda - sarà indicata con le parentesi angolari, ad esempio $\langle \delta_{\bar{x}} | \varphi \rangle$; ricordiamo inoltre che lo spazio delle distribuzioni è dotato naturalmente di una struttura di $\mathbb{R}$-spazio vettoriale indotta. A partire da funzioni sufficientemente regolari g si può ottenere una distribuzione che agisce come un integrale: $\varphi \mapsto \int g\varphi$, per questo talvolta le distribuzioni si chiamano anche funzioni generalizzate. La delta di Dirac non permette questo tipo di rappresentazione, tuttavia si può immaginare che sia la distribuzione associata ad una funzione speciale nulla quasi ovunque con un picco localizzato nel punto $\bar{x}$. La concezione della delta di Dirac come impulso puntuale a noi torna utile per raggruppare i termini di

una sommatoria a seconda del valore di qualche parametro; ad esempio supponiamo che l'osservabile f dipenda solo dal valore della magnetizzazione del sistema: $f(\sigma) = f(\mathrm{m}(\sigma))$, allora vale

$$\sum_{\sigma\in\Sigma} f(\sigma) = \sum_{\sigma\in\Sigma} \int \mathrm{dm}\, \delta_{\mathrm{m}(\sigma)}(m) f(m) = \int \mathrm{dm}\, f(m) \sum_{\sigma\in\Sigma} \delta_{\mathrm{m}(\sigma)}(m)$$

Così facendo abbiamo spostato la dipendenza dalla configurazione sulla distribuzione. Il punto di svolta sta nel fatto che in un certo senso possiamo rappresentare anche la delta di Dirac con un integrale - rappresentazione detta *di Fourier* - e quindi riuscire a rimuovere la dipendenza della funzione dalle specifiche configurazioni.

Lemma 2.8 — *Fissato un punto $\bar{x} \in \mathbb{R}^d$ e data una funzione sufficientemente regolare $f\colon \mathbb{R}^d \to \mathbb{R}$, indicando con i l'unità immaginaria vale l'uguaglianza*

$$f(\bar{x}) = \langle \delta_{\bar{x}} | f \rangle = \iint \frac{\mathrm{dm}\, \mathrm{d}\hat{m}}{2\pi} e^{i\hat{m}(\bar{x}-m)} f(m)$$

Per un risultato rigoroso si può prendere ad esempio f nello spazio di Schwartz (vedi [Hör63]).

Precisiamo che questo non implica che la delta si possa rappresentare tramite come l'azione dell'integrale contro la funzione $\int \frac{\mathrm{d}\hat{m}}{2\pi} e^{i\hat{m}(\bar{x}-\bullet)}$, poiché quest'ultimo integrale non è convergente e perciò non ha significato; ciononostante la sua azione sulle funzioni è ben definita, quindi è una distribuzione ben posta. Altrimenti detto: la distribuzione delta di Dirac, interpretata come una misura, non ha una derivata di Radon-Nikodym.
Prima di procedere con la dimostrazione osserviamo che grazie a questo lemma possiamo riscrivere l'espressione precedente in modo formale:

$$\sum_{\sigma\in\Sigma} f(\mathrm{m}(\sigma)) = \iint \frac{\mathrm{dm}\, \mathrm{d}\hat{m}}{2\pi} f(m) \sum_{\sigma\in\Sigma} e^{i\hat{m}(\mathrm{m}(\sigma)-m)} \tag{2.22}$$

Generalmente, le variabili m e $\hat{m}$ così introdotte si diranno rispettivamente **parametri d'ordine** e **parametri d'ordine coniugati** - questi ultimi non hanno a che fare con l'operazione di coniugio dei numeri complessi, si potrebbero anche chiamare parametri ausiliari.
Il lettore digiuno di teoria delle distribuzioni può semplicemente saltare la dimostrazione del lemma 2.8 senza che questo infici la comprensione futura.

Dimostrazione. La dimostrazione sfrutta sostanzialmente la trasformata di Fourier di una distribuzione, in particolare il fatto che $2\pi\, \delta_0 = \mathcal{F}(1)$ come distribuzioni, dove $\mathcal{F}(1)$ indica

la trasformata della funzione costante uguale a 1; questo è un fatto noto che si ottiene dal teorema di inversione della trasformata di Fourier. A questo punto è sufficiente traslare le distribuzioni nel punto desiderato; indichiamo con $\mathrm{T}_{\bar{x}}$ l'operatore di traslazione che agisce come segue su una distribuzione u: $\langle \mathrm{T}_{\bar{x}}\, u | f \rangle \coloneqq \langle u | f(\bullet + \bar{x}) \rangle$. Otteniamo $\delta_{\bar{x}} = \mathrm{T}_{\bar{x}}\, \delta_0 = \frac{1}{2\pi}\, \mathrm{T}_{\bar{x}}\, \mathcal{F}(1)$. Concludiamo calcolando

$$\langle \mathrm{T}_{\bar{x}}\, \mathcal{F}(1) | f \rangle = \langle \mathcal{F}(1) | f(\bullet + \bar{x}) \rangle = \langle 1 | \int \mathrm{dm}\, e^{-i \bullet m} f(m + \bar{x}) \rangle = \iint \mathrm{dm}\, \mathrm{d}\hat{\mathrm{m}}\, e^{-i\hat{\mathrm{m}}(m - \bar{x})} f(m)$$

□

Fatta questa breve digressione possiamo tornare al calcolo dell'energia libera ripartendo dall'equazione (2.21). Applichiamo il lemma 2.8 introducendo delle delte di Dirac per le osservabili

$$\mathrm{m}_\gamma \coloneqq \frac{1}{N} \sum_i \xi_i^1 \sigma_i^\gamma \quad \forall \gamma \qquad\qquad \mathrm{q}_{\gamma\tilde{\gamma}} \coloneqq \frac{1}{N} \sum_i \sigma_i^\gamma \sigma_i^{\tilde{\gamma}} \quad \forall \gamma, \tilde{\gamma} \tag{2.23}$$

chiamate, similmente a quanto già fatto, magnetizzazioni di Mattis e sovrapponibilità delle repliche (o *overlap*); nel seguito indicheremo con m e q rispettivamente il vettore delle magnetizzazioni e la matrice delle sovrapponibilità. Facciamo presente che la matrice degli overlap sarà per costruzione simmetrica con degli 1 sulla diagonale; per questo si sarebbe potuto evitare l'inserimento di alcune delte di Dirac: al posto di $\mathrm{q}_{\gamma,\gamma}$ si sarebbe potuto mettere direttamente un uno e l'espressione risultante si sarebbe potuta semplificare in modo da richiedere solo la parte triangolare (superiore o inferiore) della matrice dei parametri; in ogni caso queste scelte non influiscono sul risultato. Così facendo e poi operando il cambio di variabili $\hat{\mathrm{m}}_\gamma \mapsto -N\hat{\mathrm{m}}_\gamma, \hat{\mathrm{q}}_{\gamma\tilde{\gamma}} \mapsto -N\hat{\mathrm{q}}_{\gamma\tilde{\gamma}}$ otteniamo la seguente espressione per l'energia libera

$$\overline{\mathrm{f}_\beta} = \frac{\alpha}{2} - \lim_{k \to 0^+} \frac{1}{\beta k} \lim_{N \to +\infty} \frac{1}{N} \log \int \cdots \int \mathrm{dm}\, \mathrm{d}\hat{\mathrm{m}}\, \mathrm{dq}\, \mathrm{d}\hat{\mathrm{q}}\, \overline{\mathcal{D}(\mathrm{m}, \mathrm{q})} \exp\left(\frac{\beta N}{2} \sum_\gamma \mathrm{m}_\gamma^2 \right) \det(\mathrm{Id} - \beta\, \mathrm{q})^{-(P-1)/2}$$

dove il termine $\mathcal{D}$, detto *densità degli stati*, contiene la sommatoria sulle repliche ed è così definito

$$\mathcal{D}(\mathrm{m}, \mathrm{q}) \coloneqq \sum_{\{\sigma^\gamma\}_{\gamma=1}^k} \left(\frac{N}{2\pi} \right)^{k^2+k} \exp\left(-iN \sum_\gamma \hat{\mathrm{m}}_\gamma \left(\frac{1}{N} \sum_j \xi_j^1 \sigma_j^\gamma - \mathrm{m}_\gamma \right) - iN \sum_{\gamma,\tilde{\gamma}} \hat{\mathrm{q}}_{\gamma\tilde{\gamma}} \left(\frac{1}{N} \sum_j \sigma_j^\gamma \sigma_j^{\tilde{\gamma}} - \mathrm{q}_{\gamma\tilde{\gamma}} \right) \right)$$

Per semplificare questo termine si possono inizialmente raccogliere i termini che non dipendono dalle repliche. Negli altri, la sommatoria fattorizza su j e per $N \to +\infty$ si può applicare la legge dei grandi numeri come visto nel caso in basso carico (ad esempio per arrivare alle condizioni stazionarie (2.16) e (2.17)) sfruttando $\prod_j \cdots = \exp(N\,\overline{\log \ldots})$; così ottengo che nel limite si comporta come

$$\mathcal{D}(\mathrm{m},\mathrm{q}) \underset{N\to+\infty}{\sim} \left(\frac{N}{2\pi}\right)^{k^2+k} \exp\left(iN\sum_{\gamma}\hat{\mathrm{m}}_\gamma\,\mathrm{m}_\gamma + iN\sum_{\gamma,\tilde{\gamma}}\hat{\mathrm{q}}_{\gamma\tilde{\gamma}}\,\mathrm{q}_{\gamma\tilde{\gamma}}\right) \exp\left(N\overline{\log\sum_{\sigma^1,\ldots,\sigma^k=\pm1}\exp\left(-i\sum_{\gamma}\hat{\mathrm{m}}_\gamma\xi\sigma^\gamma - i\sum_{\gamma,\tilde{\gamma}}\hat{\mathrm{q}}_{\gamma\tilde{\gamma}}\sigma^\gamma\sigma^{\tilde{\gamma}}\right)}\right)$$

dove la media $\overline{\bullet}$ si intenderà rispetto alla distribuzione uniforme di un singolo spin dicotomico ξ. Inoltre, vista la convergenza della densità degli stati verso questa media, il disordine che ulteriore sul primo pattern che gli era applicato nell'ultima espressione per l'energia libera è superfluo. A questo punto possiamo applicare il metodo di Laplace del lemma 2.5 per calcolare il primo limite presente nell'espressione dell'energia libera, ottenendo

$$\overline{\mathrm{f}_\beta} = \lim_{k\to0^+}\operatorname{extr} G_k(\mathrm{m},\hat{\mathrm{m}},\mathrm{q},\hat{\mathrm{q}}) \tag{2.24}$$

dove

$$G_k(\mathrm{m},\hat{\mathrm{m}},\mathrm{q},\hat{\mathrm{q}}) \coloneqq \frac{\alpha}{2} - \frac{i}{\beta k}\sum_{\gamma}\hat{\mathrm{m}}_\gamma\,\mathrm{m}_\gamma - \frac{i}{\beta k}\sum_{\gamma,\tilde{\gamma}}\hat{\mathrm{q}}_{\gamma\tilde{\gamma}}\,\mathrm{q}_{\gamma\tilde{\gamma}} - \frac{1}{2k}\sum_{\gamma}\mathrm{m}_\gamma^2 - \frac{1}{\beta k}\overline{\log\sum_{\sigma^1,\ldots,\sigma^k=\pm1}\exp\left(-i\sum_{\gamma}\hat{\mathrm{m}}_\gamma\xi\sigma^\gamma - i\sum_{\gamma,\tilde{\gamma}}\hat{\mathrm{q}}_{\gamma\tilde{\gamma}}\sigma^\gamma\sigma^{\tilde{\gamma}}\right)} + \frac{\alpha}{2\beta k}\log\det(\mathrm{Id}-\beta\,\mathrm{q}) \tag{2.25}$$

Estremizzando G si trovano le seguenti condizioni stazionarie:

$$\begin{aligned}\mathrm{m}_\gamma &= \overline{\langle\sigma^\gamma\xi\rangle_{\mathcal{G}}}\\ \mathrm{q}_{\gamma\tilde{\gamma}} &= \overline{\langle\sigma^\gamma\sigma^{\tilde{\gamma}}\rangle_{\mathcal{G}}}\\ \hat{\mathrm{m}}_\gamma &= i\beta\,\mathrm{m}_\gamma\\ \hat{\mathrm{q}}_{\gamma\tilde{\gamma}} &= \frac{i\alpha\beta}{2}(\mathrm{Id}-\beta\,\mathrm{q})^{-1}_{\gamma,\tilde{\gamma}}\end{aligned} \tag{2.26}$$

dove $\langle\bullet\rangle_{\mathcal{G}}$ indica la media termica a temperatura 1 rispetto all'hamiltoniana

$$\mathcal{G}(\sigma) \coloneqq i\sum_{\gamma}\hat{\mathrm{m}}_\gamma\xi\sigma^\gamma + i\sum_{\gamma,\tilde{\gamma}}\hat{\mathrm{q}}_{\gamma\tilde{\gamma}}\sigma^\gamma\sigma^{\tilde{\gamma}}$$

Innanzitutto notiamo che $\mathrm{q}_{\gamma,\gamma} = 1$ e questo concorda con la definizione data in (2.23). Inoltre osserviamo che i parametri coniugati condividono il fattore moltiplicativo i derivante dall'espressione di Fourier delle delte di Dirac. Questo fa sì che l'espressione dell'energia (2.25) resti a valori reali, conservando il suo significato fisico. Infatti, in generale introducendo le delte di Dirac come nell'equazione (2.22) si ottiene l'estremizzazione di un funzionale nella forma $\log f(m) + i\hat{m}(x - m)$ e imponendo l'annullamento della sue derivate si ottengono le equazioni: $f'(m)/f(m) = i\hat{m}$ e $m = x$; sapendo che x è un numero reale e parimenti f è a valori reali, ne segue che i parametri coniugati sono per loro natura numeri immaginari. Perciò, ai fini dei nostri ragionamenti, possiamo considerare tutt'uno i termini $i\hat{m}$ inglobando l'unità immaginaria all'interno della variabile (che quindi diventa una variabile reale) sia nell'espressione del funzionale da estremizzare che (di conseguenza) nelle condizioni stazionarie; se consideriamo il funzionale solamente nel suo punto estremale questa modifica non produce alterazioni. Siccome l'interesse fisico si concentra sull'energia libera e sul valore dei parametri d'ordine (non coniugati) nel punto estremale questa identificazione non cambia i risultati, però permette di lavorare solamente con numeri reali semplificando un po' la trattazione. Nel caso specifico delle condizioni stazionarie (2.26), vista la semplice forma dell'equazione relativa al parametro $\hat{\mathrm{m}}_\gamma$, possiamo addirittura sostituirla nelle altre eliminando in toto la dipendenza da quei parametri nell'espressione di G_k.
Ciononostante, questa forma non è adeguata per il calcolo del limite perché a priori tutti i parametri di integrazione potrebbero avere un valore diverso e valutarne la media risulta complicato. Per ovviare al problema si può ridurre la ricerca del punto estremale ad un sottospazio con maggiore simmetria; precisamente, visto che in origine le repliche erano identiche, si può assumere che nel punto estremale i parametri $\mathrm{m}_\gamma, \mathrm{q}_{\gamma\tilde{\gamma}}, \hat{\mathrm{q}}_{\gamma\tilde{\gamma}}$ non dipendano dagli indici di replica; $\hat{\mathrm{m}}_\gamma$ non è presente perché lo consideriamo già sostituito. Precisamente assumeremo che

$$\begin{aligned} \mathrm{m}_\gamma &= \mathrm{m} \\ \mathrm{q}_{\gamma\tilde{\gamma}} &= \delta_{\gamma\tilde{\gamma}} + \mathrm{q}(1 - \delta_{\gamma\tilde{\gamma}}) \\ \hat{\mathrm{q}}_{\gamma\tilde{\gamma}} &= \frac{i\alpha\beta^2}{2}(R\,\delta_{\gamma\tilde{\gamma}} + r(1 - \delta_{\gamma\tilde{\gamma}})) \end{aligned} \tag{2.27}$$

dove ora m, q indicano dei numeri reali invece che un vettore e una matrice. Questa ipotesi, comunemente utilizzata per risolvere i calcoli risultanti dal replica trick, prende il nome di **ipotesi di simmetria di replica** o *replica simmetry ansatz*[9], in futuro abbreviata

[9]Il termine ansatz è un prestito dal tedesco e in matematica viene usato per indicare un'assunzione

anche con *ipotesi RS*. In questo caso si può direttamente cercare il punto estremale di f_β nel sottospazio, che equivale a inserire l'ipotesi di simmetria nell'equazione (2.25) e poi estremizzare rispetto ai parametri ridotti. L'espressione valutata in un punto con questa simmetria si semplifica: i termini con R si elidono vicendevolmente - perciò considereremo solo il parametro r al posto di $\hat{\mathrm{q}}$ - quello con la media sul disordine può essere gestito tramite la linearizzazione gaussiana e il determinante vedremo che si calcolerà in modo agevole. Tra l'altro i parametri $\mathrm{m}, \mathrm{q}, r$ possono essere messi in relazione con i parametri d'ordine del sistema analogamente a come abbiamo fatto nel caso in basso carico per la magnetizzazione, oppure tramite un semplice ragionamento che esporremo una volta finiti questi calcoli. Si trova che m è la sovrapponibilità media del sistema con la prima memoria (quella che cerchiamo di recuperare), quindi il parametro ferromagnetico; q è la sovrapponibilità media di due repliche distinte del sistema, perciò indica se il sistema si "congela" o meno e si chiama parametro vetroso - o *spin-glass*; r rappresenta il contributo medio dei pattern che non vengono recuperati. Dunque $\overline{\mathrm{f}_\beta^{\mathrm{RS}}} = \lim_{k\to 0^+} \operatorname{extr} G_k^{\mathrm{RS}}(\mathrm{m}, \mathrm{q}, r)$ con

$$
\begin{aligned}
G_k^{\mathrm{RS}}(\mathrm{m}, \mathrm{q}, r) \coloneqq \frac{\alpha}{2} + \frac{\mathrm{m}^2}{2} + \frac{\alpha\beta}{2}\,\mathrm{q}\,r(k-1) \\
- \frac{1}{\beta k}\overline{\log \sum_{\sigma^1,\dots,\sigma^k=\pm 1} \exp\left(\beta\,\mathrm{m}\,\xi \sum_\gamma \sigma^\gamma + \frac{\alpha\beta^2}{2} r (\sum_\gamma \sigma^\gamma)^2 - \frac{\alpha\beta^2}{2} r k\right)} \\
+ \frac{\alpha}{2\beta k} \log\det(\mathrm{Id} - \beta\,\mathrm{q})
\end{aligned}
$$

ma il termine con il disordine si semplifica (usando il lemma 2.4 e fattorizzando sulle repliche) in

$$
\begin{aligned}
- \frac{\alpha\beta^2}{2} r k + \overline{\log \int \underline{\mathrm{dz}} \sum_{\sigma^1,\dots,\sigma^k=\pm 1} \exp\left(\beta\,\mathrm{m}\,\xi \sum_\gamma \sigma^\gamma + \sqrt{\alpha r}\beta z \sum_\gamma \sigma^\gamma\right)} \\
= -\frac{\alpha\beta^2}{2} r k + \overline{\log \int \underline{\mathrm{dz}} \left(2\cosh(\beta\,\mathrm{m}\,\xi + \sqrt{\alpha r}\beta z)\right)^k} \qquad (2.28)
\end{aligned}
$$

dove l'integrale è gaussiano, espresso tramite la notazione formale $\underline{\mathrm{dz}} \coloneqq \mathrm{dz}\exp(-z^2/2)/\sqrt{2\pi}$. Per il determinante si può usare il prossimo risultato, da cui segue che $\log\det(\mathrm{Id} - \beta\,\mathrm{q}) = \log(1 - \beta + \beta\,\mathrm{q}(1-k)) + (k-1)\log(1-\beta+\beta\,\mathrm{q})$.

fatta per risolvere un problema che risulta essere appropriata una volta nota la soluzione. Ad esempio, nel cercare una soluzione a certe equazioni differenziali si assume che sia esprimibile in una specifica forma (questo l'ansatz); nel caso si riesca a trovare l'ipotesi risulta verificata. Naturalmente, nel caso in cui fatto un certo ansatz non si riesca a trovare una soluzione accettabile questo va rigettato.

Lemma 2.9 — *Una matrice di dimensione $d \in \mathbb{N}_+$ nella forma $M_{ij} = a\,\delta_{ij} + b(1-\delta_{ij})$ ha gli autovalori: $bd + (a-b)$ con molteplicità 1 e $a-b$ con molteplicità $d-1$.*

Dimostrazione. Si scrive come $M_{ij} = b + (a-b)\,\delta_{ij}$ perciò ha come autospazi quello generato da $(1,\dots,1)$ e $\{x \in \mathbb{R}^d \,|\, \sum_i x_i = 0\}$ che è un iperpiano per definizione. □

Prima di proseguire osserviamo che a posteriori, sapendo di dover imporre l'ipotesi di simmetria di replica, avremmo potuto svolgere il calcolo applicandola subito nell'equazione (2.21), in modo tale da introdurre solamente un numero ridotto di delte di Dirac sui parametri (2.27); così facendo avremmo ottenuto lo stesso funzionale di pseudo energia libera $G_k^{\mathrm{RS}}(\mathrm{m},\mathrm{q},r)$ da estremizzare. Arrivati a questo punto assumiamo di poter scambiare il limite con la valutazione nel punto estremale, ad esempio supponendo che la successione di funzioni G_k converga uniformemente e che i punti estremali si comportino in modo sufficientemente regolare. Per calcolare il limite si tenga conto che il termine ottenuto dal determinante si può sviluppare in serie di Taylor al prim'ordine ottenendo

$$\frac{1}{k}\log\det(\mathrm{Id} - \beta\,\mathrm{q}) \underset{k\to 0^+}{\sim} \log(1-\beta+\beta\,\mathrm{q}) - \frac{\beta\,\mathrm{q}}{1-\beta+\beta\,\mathrm{q}}$$

mentre per l'integrale gaussiano in (2.28) si può applicare il limite (2.19) presentato nel contesto del trucco delle repliche (l'attuale media sul disordine è discreta e indipendente dal limite, non corrisponde al disordine usato nel trucco delle repliche). Pertanto otteniamo l'energia libera come punto estremale della funzione

$$\begin{aligned} G^{\mathrm{RS}}(\mathrm{m},\mathrm{q},r) \coloneqq \frac{\alpha}{2} + \frac{\mathrm{m}^2}{2} + \frac{\alpha\beta}{2} r(1-\mathrm{q}) \\ - \frac{1}{\beta}\int \underline{\mathrm{dz}}\, \overline{\log\left(2\cosh(\beta\,\mathrm{m}\,\xi + \sqrt{\alpha r}\beta z)\right)} \\ + \frac{\alpha}{2\beta}\left(\log(1-\beta+\beta\,\mathrm{q}) - \frac{\beta\,\mathrm{q}}{1-\beta+\beta\,\mathrm{q}}\right) \end{aligned} \tag{2.29}$$

Derivando questa espressione per calcolare le condizioni stazionarie - o equivalentemente valutando il limite $k \to 0^+$ delle equazioni (2.26) sotto l'ipotesi di simmetria di replica - e sfruttando l'integrazione per parti, il lemma 1.3 e il fatto che la funzione tanh sia dispari per applicare il cambio di variabili $z \mapsto \xi z$, si ottengono le seguenti:

$$\begin{aligned} \mathrm{m} &= \int \underline{\mathrm{dz}}\, \tanh\left(\beta\,\mathrm{m} + \sqrt{\alpha r}\beta z\right) \\ \mathrm{q} &= \int \underline{\mathrm{dz}}\, \tanh^2\left(\beta\,\mathrm{m} + \sqrt{\alpha r}\beta z\right) \\ r &= \frac{q}{(1-\beta+\beta q)^2} \end{aligned} \tag{2.30}$$

Prima di analizzare il diagramma di fase del sistema alla luce del significato attribuibile a questi parametri, proponiamo brevemente un ulteriore metodo per comprendere intuitivamente che questi indicano il valore dei parametri d'ordine nel limite termodinamico.
Descriviamo questo metodo nel caso di un sistema dipendente da un solo parametro d'ordine per una maggiore chiarezza espositiva; il metodo si estende facilmente al caso generale di sistemi analizzabili con il metodo di Laplace dipendenti da vari parametri, come nel caso dell'equazione (2.20). Come esempio di riferimento si può prendere il modello di Curie-Weiss, la cui pressione si può calcolare anche con il metodo di Laplace.
Dato un sistema di hamiltoniana $\mathcal{H}(\mathrm{m}(\sigma))$ (dove qui m è un qualunque parametro d'ordine), l'idea è di utilizzare l'equazione (2.22) per introdurre gli integrali e poi usare il metodo di Laplace enunciato nel lemma 2.5. Se g è un'osservabile dipendente solo da un parametro d'ordine intensivo m: $g(\sigma) = g(\mathrm{m}(\sigma))$ che automedia nel limite termodinamico, possiamo calcolare la sua media termica nel limite termodinamico come segue

$$\begin{aligned}\langle g\rangle &= \sum_{\sigma\in\Sigma} g(\mathrm{m}(\sigma))\frac{\exp(-N\,\mathcal{H}(\mathrm{m}(\sigma)))}{\mathcal{Z}} \\ &= \iint \mathrm{dm}\,\mathrm{d}\hat{\mathrm{m}}\, g(m)\frac{\exp(-N\,\mathcal{H}(m))}{\mathcal{Z}}\sum_{\sigma\in\Sigma}\frac{e^{i\hat{m}(\mathrm{m}(\sigma)-m)}}{2\pi} \\ &\equiv \iint \mathrm{dm}\,\mathrm{d}\hat{\mathrm{m}}\, g(m)\frac{e^{NG(m,\hat{m})}}{\mathcal{Z}} \underset{N\to+\infty}{\approx} g(m^\star)\frac{e^{NG(m^\star,\hat{m}^\star)}}{\mathcal{Z}}\end{aligned} \tag{2.31}$$

Nel caso in cui g sia la funzione costante 1 si ottiene che $\mathcal{Z} \underset{N\to+\infty}{\approx} e^{NG(m^\star,\hat{m}^\star)}$, perciò otteniamo che $\langle g\rangle \underset{N\to+\infty}{\to} g(m^\star)$ dove $m^\star$ è il punto di massimo di $G(m,\hat{m})$; questa funzione G è esattamente la pseudo energia libera (a meno di costanti) che si ottiene esprimendo la funzione di partizione $\mathcal{Z}$ tramite il metodo di Laplace.
In sostanza, sotto buone ipotesi l'integrazione su punto di sella concentra il valore dell'integrale nel punto di massimo della pseudo energia libera, perciò le medie delle osservabili si ottengono valutandole proprio in questo punto.

Conoscendo il significato dei vari parametri delle equazioni (2.30), possiamo studiare il diagramma di fase del sistema. Innanzitutto notiamo che la soluzione paramagnetica $\mathrm{m} = \mathrm{q} = r = 0$ esiste sempre. Siccome siamo interessati alla fase di recupero caratterizzata da $\mathrm{m} > 0$, si studiano le equazioni di auto-consistenza per capire a che condizioni emerge. Una parte di analisi teorica si trova in [CKS05], tuttavia per ottenere il diagramma delle fasi completo si ricorre a delle simulazioni numeriche.
Tra l'altro bisogna aggiungere una considerazione riguardo alla validità del diagramma ottenuto per via analitica: per ricavarlo abbiamo sfruttato la procedura formale (non

rigorosa) delle repliche e abbiamo pure imposto l'ipotesi di simmetria; a priori non ci sono garanzie della validità del metodo, per questo gli esperimenti numerici sono cruciali per corroborarne la validità. Ne emerge che i risultati analitici sono validi in quasi tutte le zone del diagramma, fa eccezione una piccola regione nella fase di recupero a bassissime temperature: in questa parte i risultati euristici non coincidono con quelli predetti dalla teoria, perciò in questi casi è da rigettare l'ipotesi di simmetria di replica; si dice che il sistema subisce una **rottura di simmetria di replica**, o *replica simmetry breaking* (RSB). Nonostante questa discrepanza, nel caso del modello di Hopfield le predizioni qualitative del modello con simmetria di replica sono assimilabili al comportamento effettivo. In generale le differenze possono anche essere significative, il caso più noto è quello del modello di Sherrington e Kirkpatrick per il quale Giorgio Parisi propose una soluzione corretta [Par80b; Par80a] imponendo una simmetria più granulare rispetto all'assunzione di simmetria di replica che abbiamo presentato sopra.
Soffermiamoci ora sul commento del diagramma risultante dalle simulazioni, raffigurato in figura 2.2. Per temperature troppo alte l'unica fase presente è quella paramagnetica nella quale il sistema è troppo disordinato per poter convergere verso degli stati significativi; qui l'energia libera non ha minimi stabili.
Abbassando la temperatura compare una fase vetrosa $\mathrm{m} = 0; \mathrm{q}, r > 0$, caratterizzata dalla comparsa di minimi locali che però rappresentano degli stati non sufficientemente correlati con le memorie di partenza; a queste condizioni l'interferenza tra i ricordi è ancora troppo alta per poter recuperare delle informazioni utili.
Continuando a diminuire il disordine si incontra una transizione di fase del prim'ordine verso la fase di richiamo, suddivisa a sua volta in una fase metastabile e una stabile, nella quale $\mathrm{m}, \mathrm{q}, r > 0$ perciò è possibile che avvenga la convergenza verso la memoria prescelta. Nella fase metastabile gli stati di richiamo sono ancora dei minimi locali dell'energia libera, perciò la convergenza dipende dal punto iniziale del sistema; solo nella fase di richiamo stabile, anche detta ferromagnetica, diventano dei minimi assoluti permettendo l'effettivo funzionamento del modello di Hopfield come memoria associativa.
Avendo delineato le caratteristiche generali del diagramma di fase, una prima osservazione possibile riguarda il comportamento del sistema in basso carico, ossia per $\alpha = 0$: sia considerando le equazioni (2.30) che guardando la figura 2.2 si evince che il sistema ha solamente due fasi (paramagnetica e ferromagnetica) separate da una transizione di fase; rimarchiamo anche che questa analisi coincide con quella svolta in basso carico, a partire dall'equazione (2.18) della magnetizzazione.
D'altra parte segnaliamo anche che quando il carico del sistema diventa eccessivo la fase

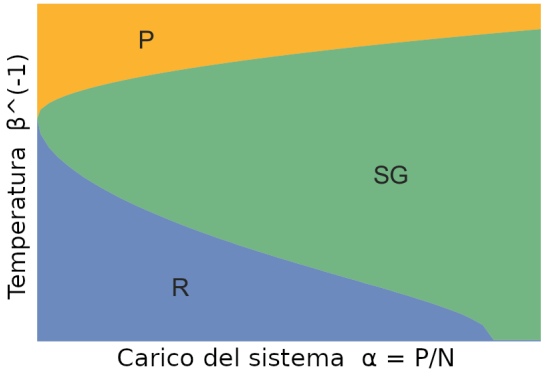


Figura 2.2: Questo è il diagramma delle fasi del modello di Hopfield al variare del carico α (sull'asse delle ascisse) e della temperatura β^{-1} (sulle ordinate). I regimi sono paramagnetico (P), vetroso (SG) e di richiamo (R). Non è indicata la demarcazione tra la fase di richiamo meta-stabile e quella stabile. La figura è presa da [Mat20].

di recupero scompare totalmente: data una taglia il modello di Hopfield è in grado di memorizzare solo un certo numero di pattern, oltre il quale i ricordi sovraccaricano il sistema al punto da renderlo non funzionante.

2.3 Macchina di Boltzmann per l'apprendimento

Nella sezione precedente abbiamo visto che il modello di Hopfield è un modello di memoria associativa più che di apprendimento, in quanto necessita una preimpostazione degli stati attrattori. Ora introduciamo un modello chiamato Macchina di Boltzmann - in seguito anche BM - proposto inizialmente in [HSA84; HSA85]; questo inferisce autonomamente l'informazione caratterizzante contenuta in una serie di esempi che gli vengono forniti - per questo si tratta di un modello di apprendimento non supervisionato, a differenza dei percettroni multistrato citati alla fine della sezione 2.1.
Il modello impara costruendosi una rappresentazione interna dei dati sotto forma di caratteristiche, identificate con i neuroni, e pesi assegnati alle sinapsi che le collegano, i quali dipendono dalle correlazioni tra queste. Dunque, il sistema identifica le qualità principali e apprende le correlazioni sottostanti agli esempi. Ciò fa sì che una volta terminato l'allenamento, anche detto processo di **apprendimento**, si possa studiare la configurazione della rete per analizzare i dati forniti. In questo senso è un modello simile alla *analisi delle componenti principali* (PCA), tuttavia opera in modo non-lineare e quindi ha una versatilità decisamente maggiore. Inoltre, grazie alla forma della rappresentazione interna

dei dati, in una fase successiva questo tipo di rete consente di produrre dei dati simili a quelli che ha ricevuto durante l'allenamento, rendendolo un modello utilizzabile anche per la cosiddetta **fase generativa** o di **richiamo** - *retrieval* in inglese.
Questa seconda fase rimanda al funzionamento del modello di Hopfield, tuttavia in questo caso non si tratta semplicemente di recuperare delle informazioni memorizzate, bensì di simulare dei dati verosimili sulla base delle conoscenze apprese. In questo senso la BM apprende in modo molto più simile al cervello umano di quanto non facesse il modello di Hopfield. Facciamo anche presente che la capacità di concentrarsi sulla struttura dei dati permette alla BM di operare con quantità di esempi (e quindi di caratteristiche) notevolmente più ingenti rispetto al modello di Hopfield, che abbiamo visto avere problemi di recupero quando il carico della rete è eccessivo (si faccia riferimento alla figura 2.2 e alla discussione che la segue); questo proprio perché si concentra sulla generalizzazione invece che sulla ripetizione meccanica.

I dati d'ora in poi saranno rappresentati come configurazioni di neuroni $\sigma_1, \dots, \sigma_N \in \{\pm 1\}$. Alla macchina di Boltzmann vengono forniti degli esempi $\{\sigma^a\}_{a=1}^M$ e questa cerca di approssimarli: tenta di riprodurne efficacemente la distribuzione congiunta. Per scegliere la famiglia di distribuzioni entro la quale la macchina cerca di compiere questa approssimazione si utilizza il principio di massima entropia con il vincolo di riprodurre le medie e le correlazioni dei neuroni. Come abbiamo visto nel teorema 1.5 la distribuzione P che risolve questo problema vincolato si può esprimere nella forma di Boltzmann-Gibbs con la funzione energia di un modello di Ising $\mathcal{H}(\sigma) \coloneqq -\sum_{i<j} \xi_{ij}\sigma_i\sigma_j - \sum_i \theta_i\sigma_i$; la temperatura è fissata a $\beta = 1$ perché sono già presenti i parametri liberi in ogni variabile. Dunque il problema si riduce a individuare i valori ottimali dei parametri ξ_{ij}, θ_i che producano le medie e correlazioni presenti nei dati.
All'interno del paradigma statistico richiamato nella sezione 1.4 questo è il problema inverso di un modello di Ising (dove si studia la distribuzione di equilibrio a parametri fissati, sebbene aleatori); infatti, l'hamiltoniana e la distribuzione scritte sopra sono da intendersi condizionate rispetto ai parametri di interazione e campo esterno, che assumiamo aleatori non conoscendoli a priori. Il modello così costruito prende il nome di **Macchina di Boltzmann** ed è una generalizzazione del più noto modello di Sherrington e Kirkpatrick. Il collegamento rigoroso tra i due è stato proposto in [Bar+15; Pan15], dove viene precisata la distribuzione di probabilità da assegnare ai parametri aleatori e ne viene calcolata la pressione; in figura 2.3 presentiamo il diagramma di fase di questo modello diretto per un confronto con quello del modello di Hopfield illustrato nella figura 2.2.

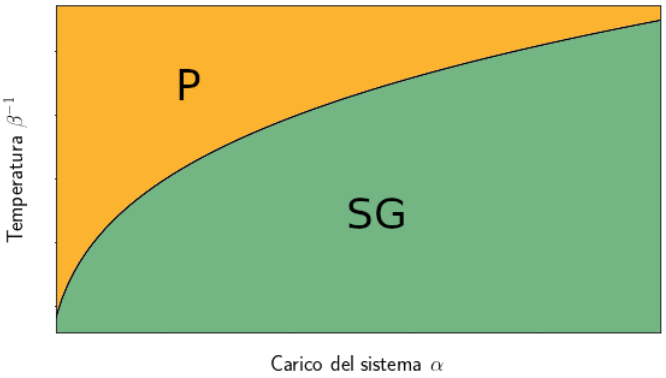


Figura 2.3: Abbiamo rappresentato schematicamente il diagramma di fase di una macchina di Boltzmann al variare del grado di sollecitazione $\alpha = M/N$ e della temperatura β^{-1} (per ottenere il grafico bisogna normalizzare le variabili aleatorie e scorporare la dipendenza da β). La linea rappresenta la transizione di fase tra il regime paramagnetico (P) e quello vetroso (SG). Si noti l'assenza di una fase di richiamo.

Per risolvere il problema inverso, cioè allenare la macchina, si può usare il teorema di Bayes assumendo i parametri uniformi e massimizzare il logaritmo della verosimiglianza $\mathcal{V}(\xi_{ij}, \theta_i) \coloneqq \sum_{a=1}^{M} \sum_{i<j} \xi_{ij}\sigma_i^a\sigma_j^a + \sum_i \theta_i\sigma_i^a - M \log \mathcal{Z}(\xi_{ij}, \theta_i)$, dove $\mathcal{Z}$ indica la funzione di partizione di P. Sottolineiamo che questo approccio equivale a minimizzare la divergenza di Kullback-Leibler[10] della distribuzione P rispetto a quella bersaglio $1/M \sum_a \delta_{\sigma^a}$.
Vista la difficoltà di trattare analiticamente $\mathcal{V}$ per via della funzione di partizione, si può ricorrere ad un algoritmo iterativo di ascesa del gradiente che aggiorna i parametri nel modo seguente: $\Delta\xi_{ij} = 1/M \sum_a \sigma_i^a\sigma_j^a - \langle\sigma_i\sigma_j\rangle_{\mathcal{H}}$ (e $\Delta\theta_i$ analogamente), dove $\langle\bullet\rangle_{\mathcal{H}}$ è la media termica rispetto alla distribuzione calcolata con i parametri non incrementati; il primo termine si dice *gradiente positivo* mentre il secondo è chiamato *gradiente negativo*. Notiamo che la regola di aggiornamento è simile a quella di Hebb (2.5) vista per il modello di Hopfield.

Per fare in modo che il modello possa approssimare la distribuzione dei dati più fedelmente è necessario poter riprodurre anche le correlazioni di ordine maggiore. Una

[10]La divergenza di Kullback-Leibler dalla distribuzione discreta $\mathcal{Q}$ alla distribuzione $\mathcal{P}$ sullo stesso spazio di probabilità, anche chiamata *entropia relativa* di $\mathcal{P}$ rispetto a $\mathcal{Q}$, è così definita: $\mathrm{D}_{\mathrm{KL}}(\mathcal{P}|\mathcal{Q}) \coloneqq \sum_x \mathcal{P}(x) \log\left(\frac{\mathcal{P}(x)}{\mathcal{Q}(x)}\right)$. Misura l'errore commesso approssimando un fenomeno di distribuzione $\mathcal{P}$ con la distribuzione $\mathcal{Q}$, si tratta in sostanza di una forma di distanza tra due distribuzioni. Tuttavia, nonostante sia spesso chiamata distanza, non è simmetrica; conserva però la proprietà di essere non-negativa, e nulla se e soltanto se le due distribuzioni coincidono. La nozione si generalizza anche al caso di distribuzioni assolutamente continue, si veda ad esempio [CKS05].

possibilità è quella di aggiungere alla rete dei neuroni che non recepiscono direttamente i dati ma si aggiornano solamente di riflesso, aumentando così i gradi di libertà del sistema. I neuroni iniziali si diranno **visibili** mentre quelli aggiuntivi **nascosti** e li indicheremo con $\tau_1, \dots, \tau_P \in \{\pm 1\}$; l'hamiltoniana che ne deriva è analoga a quella presentata poc'anzi e la relativa distribuzione di Boltzmann-Gibbs $\mathrm{P}(\sigma, \tau)$ dipende da tutti i neuroni. Siccome gli esempi sono configurazioni di N spin e vengono in contatto solamente con le unità visibili, la distribuzione approssimante dovrà avere la stessa dimensione e sarà ottenuta mediante marginalizzazione sui nodi nascosti: $\mathrm{P}(\sigma) \coloneqq \sum_\tau \mathrm{P}(\sigma, \tau)$.
Il modello così costruito risulta essere difficilmente utilizzabile e analizzabile a causa dell'interconnessione di tutti i neuroni; per semplificarlo si possono separare ulteriormente i ruoli dei nodi visibili e di quelli nascosti azzerando le interazioni tra neuroni dello stesso tipo. Quella ottenuta è una rete neuronale a due strati senza auto-interazioni, lo strato che interagisce con l'esterno si dice visibile e l'altro nascosto. Una prima formulazione di questo modello è stata proposta in [Smo86]. Da un punto di vista matematico abbiamo a che fare con una struttura di grafo bipartito non orientato (a differenza delle reti feed-forward), la cui hamiltoniana si esprime in questa forma:

$$\mathcal{H}(\sigma, \tau) \coloneqq - \sum_{\substack{1 \le i \le N \\ 1 \le \mu \le P}} \xi_i^\mu \sigma_i \tau_\mu - \sum_{i=1}^{N} \theta_i \sigma_i - \sum_{\mu=1}^{P} \tilde{\theta}_\mu \tau_\mu \tag{2.32}$$

Si noti che abbiamo leggermente cambiato notazione per i pesi, mettendo l'indice del neurone nascosto in apice invece che in pedice. Il modello appena definito prende il nome di **Macchina di Boltzmann Ristretta**, abbreviato con RBM dall'inglese *Restricted Boltzmann Machine*. La separazione dei due tipi di unità ha il vantaggio di permettere una fattorizzazione sui nodi nascosti della distribuzione marginale su quelli visibili:

$$\mathrm{P}(\sigma) = \mathcal{Z}^{-1}\, e^{\sum_i \theta_i \sigma_i} \prod_\mu 2 \cosh\left(\sum_i \xi_i^\mu \sigma_i + \tilde{\theta}_\mu \right)$$

Anche qui $\mathcal{Z}$ è la funzione di partizione relativa all'hamiltoniana (2.32); equivalentemente si può pensare come il fattore di normalizzazione di quest'ultima espressione; in ogni caso dipende dai parametri del modello.
Da questa forma si intuisce anche la correttezza del ragionamento iniziale che prevedeva per le RBM maggiori capacità di approssimazione, infatti espandendo quest'ultima formula per piccoli valori dei parametri di interazione si ottiene un'hamiltoniana efficace contenente interazioni di qualsiasi ordine tra i neuroni visibili. Difatti è stato dimostrato che le RBM sono approssimatori universali [LB08; MA11]: qualunque distribuzione di probabilità

su $\{\pm 1\}^N$ può essere approssimata arbitrariamente bene, nel senso della divergenza di Kullback-Leibler, da una RBM con un numero sufficiente di unità nascoste.
Un secondo vantaggio della stratificazione delle unità risiede nel fatto che, condizionatamente ad uno strato, i neuroni dell'altro sono indipendenti nella distribuzione di probabilità. Questo permette un metodo algoritmico di allenamento della rete, detto *contrastive divergence* [Hin02], che consiste in una variante approssimata dell'algoritmo di ascesa del gradiente visto per le BM: si effettua un campionamento alternato - che grazie alla fattorizzazione avviene in parallelo sulle unità - dello strato nascosto e di quello visibile tramite un metodo MCMC (si veda la sezione 1.3) chiamato *campionamento di Gibbs*. Per ulteriori informazioni circa questo metodo di allenamento delle RBM si può consultare [Hin10]. Altri algoritmi che si basano sul cosiddetto *cavity method* sono descritti in [Méz17; Hua21].
Una possibile generalizzazione è la cosiddetta versione *profonda*, nella quale vengono aggiunti strati nascosti a catena ottenendo le *Deep Boltzmann Machine* [SH09; Alb+21]. Queste sono utili per trattare separatamente diversi ordini di correlazione e permettono una modellazione ancora più approfondita e fedele. Le RBM e le loro varianti sono utilizzate per risolvere i compiti più variegati: dalla categorizzazione delle immagini a allo studio della struttura delle proteine [Tub18].

Nel contesto di questa tesi è più rilevante concentrarci sullo studio teorico della distribuzione di probabilità e generalizzare in un'altra direzione: introducendo dei prior non uniformi sui neuroni per simulare meglio eventuali asimmetrie che i segnali esterni possono presentare. Esprimiamo la distribuzione di probabilità sui neuroni moltiplicando il fattore di Boltzmann-Gibbs ottenuto dall'hamiltoniana (2.32) per due prior (ben distinti dal prior uniforme sui parametri): $\mathrm{P}_v(\sigma)$ sui nodi visibili e $\mathrm{P}_n(\tau)$ su quelli nascosti. Siccome questi fattori sono arbitrari, possiamo inglobarvi le parti relative ai campi esterni senza perdita di generalità. Si ottiene la forma seguente - che volendo si può chiamare *RBM generalizzata*:

$$\mathrm{P}(\sigma,\tau) \coloneqq \mathcal{Z}^{-1}\ \mathrm{P}_v(\sigma)\, e^{\sum_{i,\mu} \xi_i^\mu \sigma_i \tau_\mu}\ \mathrm{P}_n(\tau) \tag{2.33}$$

Se nel prior i neuroni nascosti sono indipendenti, la proprietà di fattorizzazione rimane valida e si ottiene la forma generale per la distribuzione marginale sui nodi visibili

$$\mathrm{P}(\sigma) = \mathcal{Z}^{-1}\,\mathrm{P}_v(\sigma) \prod_\mu \mathbb{E}_{\mathrm{P}_n(\tau_\mu)} \left[e^{\tau_\mu (\sum_i \xi_i^\mu \sigma_i)} \right] = \mathcal{Z}^{-1}\,\mathrm{P}_v(\sigma) \prod_\mu \psi(\sum_i \xi_i^\mu \sigma_i) \tag{2.34}$$

dove $\psi(x) \coloneqq \mathbb{E}_{\mathrm{P}_n(\tau_\mu)}\left[e^{\tau_\mu x}\right]$ sarà una funzione dipendente dal prior scelto per le unità nascoste. Il modello così costruito può essere specializzato in contesti diversi a seconda degli

obiettivi, scegliendo di volta in volta lo spazio di appartenenza delle variabili neuronali e i loro prior. Una sinossi delle impostazioni più comuni si può trovare in [DF21].

Lasciamo al prossimo capitolo la trattazione del caso con unità nascoste dicotomiche e analizziamo brevemente il caso di unità nascoste a valori reali con prior indipendenti distribuiti secondo una normale standard; ne ricaveremo un utile parallelismo con il modello di Hopfield studiato nella sezione precedente. In questo caso dal lemma di linearizzazione gaussiana 2.4 otteniamo $\psi(x) = e^{x^2/2}$; possiamo inoltre supporre l'assenza di campo esterno $\tilde{\theta}$ sui neuroni nascosti siccome la gaussiana è centrata; assumiamo infine un prior uniforme sulle unità visibili dicotomiche e quindi verosimilmente anche l'assenza di campo esterno θ. Con queste ipotesi si ottiene la distribuzione così esprimibile:

$$\mathrm{P}(\sigma) = \mathcal{Z}^{-1} \exp\left(\sum_{\mu} \frac{1}{2}\left(\sum_{i} \xi_i^\mu \sigma_i\right)^2\right) = \tilde{\mathcal{Z}}^{-1} \exp\left(\sum_{i<j} \sigma_i \sigma_j \sum_{\mu} \xi_i^\mu \xi_j^\mu\right)$$

Nell'ultimo passaggio abbiamo fatto confluire i termini con $i = j$ nella funzione di partizione.
In questa forma è evidente che si tratta della distribuzione di equilibrio del modello di Hopfield definito dall'hamiltoniana (2.8); l'unica differenza riguarda il parametro β che regola il rumore: qui è incorporato nei parametri ξ_i^μ. Talvolta il modello dato dalla distribuzione in (2.34) viene detto *Hopfield generalizzato*.
In questa analogia i pesi che collegano i neuroni visibili a quelli nascosti ξ_i^μ sono proprio le memorie del modello di Hopfield. La differenza cruciale tra i due modelli - cioè tra l'apprendimento e la memorizzazione - risiede nel processo di allenamento: nel caso della RBM le memorie vengono allenate partendo dagli esempi con gli algoritmi descritti sopra, mentre nel modello di Hopfield i pattern verso i quali convergere sono predeterminati. L'equivalenza dei due modelli (a parametri fissati) può essere sfruttata in entrambi i versi: da un lato permette di utilizzare l'algoritmo di campionamento di Gibbs per calcolare l'evoluzione della dinamica del modello di Hopfield più rapidamente, dall'altro fornisce un ponte con la teoria meccanico-statistica di analisi all'equilibrio che può essere usata per valutare progressivamente il successo dell'allenamento [Bar+12]. Per una disamina dello stato attuale della letteratura scientifica su questa equivalenza si può consultare [MA21].

Capitolo 3

Transizioni di fase nell'apprendimento delle RBM

In questo capitolo ci occuperemo delle Macchine di Boltzmann Ristrette (d'ora in poi RBM), introdotte nella sezione 2.3 come modelli di apprendimento artificiale. Queste sono in grado di analizzare dei dati e inferirne le correlazioni, permettendo così sia di riprodurne la distribuzione, sia di studiare le caratteristiche dei dati forniti. Il modello che presenteremo è un tipo di rete neuronale che nelle applicazioni viene utilizzato sia a sé stante che per pre-allenare delle reti profonde [SH09].
Nel seguito presenteremo lo studio teorico che abbiamo svolto, mirato a valutare quali fattori influiscono sui ritmi dell'apprendimento e le conferme sperimentali che abbiamo ottenuto. I risultati che esibiremo confermano e estendono quelli ricavati in [HWH19; Bar+17; Dec+21], fornendo una prospettiva più ampia che tiene conto anche delle dimensioni della rete neuronale. In particolare abbiamo verificato che all'aumentare degli esempi studiati il processo di apprendimento attraversa una transizione di fase; dopodiché abbiamo investigato la sua dipendenza dalla dimensione della rete e dalla correlazione presente nei dati per capire come sia possibile rendere l'apprendimento più efficiente.

Cominceremo la trattazione con un breve richiamo del modello e un'introduzione al problema teorico, inserendolo nel contesto dell'inferenza bayesiana e della meccanica statistica. In questa chiave il problema si riduce allo studio della pressione di una distribuzione di Boltzmann-Gibbs, dalla quale potremo ricavare l'andamento di alcuni indicatori della qualità dell'apprendimento. Una parte consistente del capitolo sarà dedicata al calcolo di questa pressione tramite il *replica trick*; il risultato è esposto nel risultato 3.9. Nell'ultima parte

forniamo una stima dall'alto della soglia critica e verifichiamo che questa diminuisce in presenza di una correlazione non-nulla tra i dati, in misura proporzionale anche al numero di neuroni nascosti della rete (quelli responsabili di individuare delle caratteristiche distintive).

Il presente capitolo è stato progettato affinché sia direttamente fruibile da un lettore con una conoscenza generica della materia, per questo ribadiremo rapidamente alcune notazioni e qualche concetto già espressi nei capitoli precedenti; rimandiamo a questi per delle spiegazioni più approfondite. Poiché il nostro lavoro estende il risultato di [HWH19] faremo spesso riferimento a questo articolo. Altre referenze utili saranno elencate nel corso della trattazione.
Trattandosi di una disciplina a cavallo tra la teoria e le applicazioni, durante i calcoli abbiamo sfruttato tecniche come il *replica trick* e il metodo di Laplace che non sono stati ancora del tutto formalizzati e basano la loro affidabilità sulle conferme sperimentali. Per questo motivo allo studio teorico abbiamo affiancato delle simulazioni numeriche che corroborano le nostre previsioni; la parte principale del codice che abbiamo scritto è disponibile al seguente indirizzo: `https://github.com/finedust/codici-tesi-magistrale`. Ciononostante abbiamo cercato per quanto possibile di mantenere un buon livello di rigore matematico, segnalando esplicitamente i passaggi la cui giustificazione è al momento solamente euristica.

3.1 Introduzione al problema

Cominciamo questa sezione richiamando il modello di apprendimento artificiale detto Macchina di Boltzmann Ristretta (d'ora in poi RBM), per una definizione più dettagliata ci si riferisca alla sezione 2.3.
Da un punto di vista matematico la RBM si rappresenta come un grafo non direzionato, bipartito in due strati di vertici distinti - d'ora in avanti detti strato visibile e strato nascosto - che non presenta interazioni all'interno dello stesso strato. Chiamando neuroni i vertici e sinapsi gli archi, secondo la classificazione data nella sezione 2.1 rientra nella categoria delle reti neuronali ricorrenti a due strati. Si faccia riferimento alla figura 2.1a per un'illustrazione schematica della struttura. Per formalizzare questa rappresentazione si chiamino $\sigma_1, \dots, \sigma_N$ le unità dello strato visibile, $\tau_1, \dots, \tau_P$ quelle nascoste e $\xi_i^\mu, 1 \leq i \leq N, 1 \leq \mu \leq P$ i pesi delle sinapsi che le connettono. Con queste notazioni il modello è dato da una distribuzione di probabilità sui neuroni visibili dipendente dai pesi delle sinapsi, ottenuta come marginalizzazione di una distribuzione congiunta sui neuroni visibili e nascosti;

richiamiamo qui la forma generale già scritta in (2.34).

$$\mathrm{P}(\sigma|\xi) \coloneqq \mathrm{P}_v(\sigma)\, \mathcal{Z}^{-1} \prod_{\mu} \psi\left(\beta \sum_i \xi_i^\mu \sigma_i\right)$$

In questa espressione P_v è una distribuzione di probabilità[11] sulle configurazioni dei neuroni visibili; svolge una funzione di polarizzazione delle unità e viene chiamata prior. Analogamente, la funzione $\psi(x) \coloneqq \mathbb{E}_{\mathrm{P}_n(\tau_\mu)}\left[e^{\tau_\mu x}\right]$ dipende dal prior sulle unità nascoste $\mathrm{P}_n(\tau_\mu)$ (lo stesso per ogni neurone). Poi $\mathcal{Z}$ è il fattore di normalizzazione dell'esponenziale, anche detto funzione di partizione, e β è il parametro che controlla la temperatura inversa del sistema (più si abbassa e più il sistema diventa disordinato e uniforme).
A seconda dello spazio dei valori assegnabili alle variabili neuronali e ai pesi (numeri reali, valori dicotomici, ecc.), il modello presenta comportamenti differenti; si vedano per esempio [Bar+17; Bar+18] per un confronto tra varie impostazioni del modello. Nel seguito ci occuperemo del caso in cui sia le unità che i pesi possono assumere valori in $\chi \coloneqq \{+1, -1\}$ quindi le configurazioni saranno $\sigma \in \Sigma_v, \tau \in \Sigma_n$ con $\Sigma_v \coloneqq \chi^N, \Sigma_n \coloneqq \chi^P$. Inoltre assumiamo che i neuroni non siano polarizzati, perciò prendiamo i prior uniformi e assorbiamo le loro normalizzazioni all'interno del fattore $\mathcal{Z}$; nel nostro caso $\psi(x) = \cosh(x)$, al quale aggiungiamo un fattore di normalizzazione dell'argomento $N^{-1/2}$ - più avanti verrà chiarito il motivo dell'esponente non unitario. Ricapitolando, la distribuzione del modello di RBM che analizzeremo ha la forma seguente:

$$\mathrm{P}(\sigma|\xi) \coloneqq \mathcal{Z}(\xi)^{-1} \prod_{\mu=1}^{P} \cosh\left(\frac{\beta}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^\mu \sigma_i\right) \tag{3.1}$$

La struttura del modello si può interpretare in questo modo: i neuroni visibili sono i recettori degli stimoli esterni; i neuroni nascosti invece sono delle unità ausiliarie che permettono di sintetizzare le caratteristiche dei dati, di raggruppare diverse informazioni provenienti da vari recettori; infine, le sinapsi sono il ponte tra questi due agenti, la conoscenza che permette di associare gli stimoli alle caratteristiche delle quali sono indicatori, potremmo dire che sono i concetti noti alla macchina; nel seguito verrà usato il termine pattern - il cui significato è spiegato nella nota a pagina 41 - per indicare ciascun vettore $(\xi_1^\mu, \ldots, \xi_N^\mu)$ contenente la suscettibilità agli stimoli di ciascuna caratteristica.
Lavoreremo con una distribuzione di probabilità sia perché nel funzionamento cerebrale

[11] A rigore qui con la notazione P ci stiamo riferendo alla funzione densità della distribuzione di probabilità, quindi alla probabilità dei singoli esiti. Si veda la nota 1.2 per una discussione più precisa della differenza; d'ora in poi assumeremo di trattare distribuzioni per le quali questa identificazione è possibile e ci riferiremo indistintamente ad esse o alla loro densità.

l'associazione di uno stimolo ad una caratteristica non è univoca, sia per tenere conto delle possibilità di errore. In quest'ottica, la probabilità congiunta sui nodi visibili e nascosti condizionata ad uno stato delle sinapsi modella la verosimiglianza di effettuare una determinata associazione una volta fissati i concetti noti alla macchina; la marginale sui neuroni visibili si può interpretare come indicatore di quanto un certo dato è rappresentabile tramite le categorie che la RBM conosce. Nel nostro caso la marginale (3.1) privilegia le configurazioni dei neuroni nelle quali le coppie (unità visibile, unità nascosta) sono allineate se $\xi_i^\mu = +1$ e viceversa hanno segno contrario se $\xi_i^\mu = -1$.
Si chiama allenamento o fase di apprendimento della RBM il processo attraverso il quale in base ai dati forniti ($\sigma^a \in \Sigma, a = 1, \dots, M$) si selezionano i valori dei pesi ξ_i^μ che meglio descrivono la batteria di dati fornita (in inglese *training set*). L'individuazione dei pesi ottimali corrisponde in modo indiretto all'apprendimento delle correlazioni tra i dati, perciò permette, successivamente, di generare nuovi dati simili a quelli analizzati; quest'ultimo viene detto processo di richiamo e avviene tramite il campionamento dalla distribuzione (3.1).
I metodi di allenamento sono vari e usano misuratori differenti per valutare il successo dell'apprendimento; ciononostante hanno in comune la caratteristica di non necessitare di alcuna informazione al di fuori della batteria di dati, per questo si dice che il modello rientra nella categoria dell'apprendimento non supervisionato. Un esempio di algoritmo per l'apprendimento è quello al quale abbiamo accennato nella sezione 2.3 che usa la *contrastive divergence*, questo minimizza la divergenza di Kullback-Leibler dalla distribuzione ottenuta con il modello a quella empirica dei dati. Poiché l'analisi di questi algoritmi è troppo involuta e complicata, per lo studio teorico ci si concentra su dei misuratori afferenti alla disciplina dell'inferenza statistica, dei quali a breve discuteremo.

In questo capitolo abbiamo l'obiettivo di analizzare quali fattori influenzano la capacità di apprendimento della RBM. Se volessimo valutare le capacità complessive della macchina di riprodurre dei dati simili a quelli ricevuti, potremmo allenarla, farle generare dei dati, e infine misurare la differenza tra la distribuzione della batteria di allenamento e quella generata dalla macchina. Invece, siccome siamo interessati a valutare solamente le prestazioni della fase di apprendimento, disgiunta da quella di richiamo, abbiamo la necessità di studiare il modello in un contesto controllato, nel quale i dati forniti alla macchina si categorizzano tramite concetti noti in partenza, così da poterli confrontare con quelli successivamente appresi.
Per far questo introduciamo un paradigma detto **insegnante-studente**. Supponiamo di

generare dei dati tramite una RBM ausiliaria impostata con la matrice dei pesi sinaptici $\xi^\star$ e usare questa batteria per allenare la macchina originaria. In questo modo possiamo valutare la qualità dell'apprendimento confrontando i pesi ottenuti dall'allenamento, che chiameremo d'ora in poi ξ, con quelli originari $\xi^\star$. Ripetendo il procedimento al variare dei pesi di partenza si ottiene una buona stima delle capacità del modello.
Nel contesto più generale dell'inferenza, delineato nella sezione 1.4, il metodo insegnante-studente consiste in due processi: innanzitutto un insegnante genera degli esempi tramite il modello prescelto dipendente da un segnale aleatorio (una distribuzione di probabilità condizionata), in modo tale che l'insieme degli esempi contenga un contenuto informativo sul segnale impostato; successivamente uno studente cerca di inferire il segnale originario osservando gli esempi e sfruttando la regola di Bayes (1.14) che fornisce una distribuzione di probabilità sullo spazio del segnale; confrontando il segnale inferito con quello originario si misura quanto lo studente ha imparato, cioè l'efficacia del modello di inferenza.
Nel caso del modello RBM che ci siamo posti di analizzare supponiamo di generare in modo indipendente una batteria di dati $\sigma^a \overset{d.}{\sim} \mathrm{P}(\bullet|\xi^\star)$, $a = 1, \dots, M$ con la distribuzione (3.1) condizionata a una matrice dei pesi $\xi^\star \in \chi^{N\times P}$; qui e in seguito $\bullet$ indicherà l'argomento di una funzione. Il rapporto $\alpha \coloneqq M/N$ verrà chiamato **grado di sollecitazione** del sistema o, impropriamente, anche carico dello stesso; è un indicatore di quante stimolazioni riceve la rete neuronale in rapporto al numero dei recettori che possiede. Nonostante la terminologia non aiuti, il grado di sollecitazione va tenuto distinto dal carico inteso come quantità di concetti da apprendere, che nel nostro caso coincideranno con il numero P dei nodi nascosti del sistema. Quest'ultimo è il significato da tenere presente per stabilire il parallelo con il modello di Hopfield sviluppato nella sezione 2.3; il quel caso il carico del sistema era definito come il numero delle memorie da immagazzinare in rapporto ai neuroni disponibili. Siccome qui P è finito, la RBM lavora in basso carico (di concetti).
I pesi sinaptici saranno variabili aleatorie affinché successivamente sia possibile fare una media per valutare le prestazioni del modello su un campione di dati arbitrario. Precisamente stabiliamo che siano estratti dei vettori $(\xi_i^1, \dots, \xi_i^P)$ in modo indipendente e identicamente distribuito al variare dell'indice del nodo visibile i; la distribuzione su χ^P sarà quella che massimizza l'entropia fissate la media nulla e la matrice di correlazione $\mathcal{Q}$. Nel seguito non sarà necessario fare riferimento alla forma esplicita - che si può trovare come descritto nel teorema 1.5. La scelta di questa distribuzione spiega anche il fattore di normalizzazione dell'argomento del coseno iperbolico nell'equazione (3.1): fa in modo che l'argomento, escluso il moltiplicatore β, sia una variabile aleatoria con varianza unitaria.
Nella fase di allenamento della RBM sfrutteremo la regola di Bayes e sceglieremo i pesi otti-

mali in base alla distribuzione inversa $\mathrm{P}(\xi|\{\sigma^a\}_{a=1}^M)$, anche detta distribuzione a posteriori; i criteri in base ai quali effettueremo la scelta saranno descritti nel prossimo paragrafo. Assumendo nella fase di inferenza di non possedere informazioni circa distribuzione dei pattern (oltre alla loro natura dicotomica), usiamo la regola di Bayes con prior uniformi (non si faccia confusione con i prior sulle unità, anch'essi uniformi) ottenendo la presunta distribuzione dei pesi sinaptici, cioè la probabilità inversa:

$$\begin{aligned}\mathrm{P}(\xi|\{\sigma^a\}_{a=1}^M) &= \Omega(\{\sigma^a\})^{-1} \prod_{a=1}^{M} \mathrm{P}(\sigma^a|\xi) \\ &= \Omega(\{\sigma^a\})^{-1} \prod_{a=1}^{M} \frac{\prod_{\mu=1}^{P} \cosh\left(\frac{\beta}{\sqrt{N}} \sum_{i=1}^{N} \xi_i^\mu \sigma_i^a\right)}{\mathcal{Z}(\xi)}\end{aligned} \tag{3.2}$$

dove Ω è il fattore di normalizzazione (in ξ) e la produttoria in a è una conseguenza dell'indipendenza dei dati.

Si faccia attenzione alle notazioni: nonostante i dati siano stati generati sulla base dei pesi $\xi^\star$ in questa formula compare il termine $\mathrm{P}(\sigma^a|\xi)$; qui ξ è solamente una variabile di inferenza muta che indica l'argomento della funzione densità $\mathrm{P}(\bullet|\{\sigma^a\})$ e viene usata come matrice dei valori rispetto ai quali condizionare la probabilità diretta (3.1). Si tenga anche presente che la funzione di partizione $\mathcal{Z}(\xi)$ della probabilità diretta non dipende da σ e quindi nemmeno dagli esempi $\{\sigma^a\}$, dipende invece da ξ. In sintesi si consideri che il problema inverso è in un certo senso il duale di quello precedente: gli spin dell'hamiltoniana sono quelli indicati con ξ e gli esempi $\{\sigma^a\}$ fungono da disordine.

Cogliamo l'occasione per puntualizzare una notazione: quando indicheremo una scrittura all'interno delle parentesi graffe, come nel caso di $\{\sigma^a\}$, si intenderà la collezione di quelle notazioni al variare degli indici presenti nel loro insieme di riferimento, o al variare dei soli indice in pedice quando presenti. In questo modo l'esempio precedente indica $\sigma^1, \dots, \sigma^M$, mentre una sommatoria su $\{\xi_i^\mu\}$ è da intendersi su tutte le configurazioni possibili dell'insieme di variabili $\{\xi_1^1, \xi_2^1, \dots, \xi_1^2, \dots, \xi_N^P\}$; invece la sommatoria su $\{\xi_i^\mu\}_\mu$ è sulle configurazioni delle sole variabili $\xi_i^1, \dots, \xi_i^P$, a indice i fissato.

Prima di proseguire aggiungiamo un commento sulla forma della distribuzione del problema duale (3.2). Se il termine $\mathcal{Z}(\xi)$ si fattorizzasse sui vari pattern la distribuzione si potrebbe scrivere come prodotto indicizzato da μ delle distribuzioni di P RBM con un solo nodo nascosto. Ciò vuol dire che il fattore di partizione $\mathcal{Z}$ svolge un ruolo cruciale nell'accoppiare l'attività dei vari nodi nascosti, rendendoli interconnessi e cooperativi. Le prime ricerche [Dec+21] mostrano che in assenza di correlazione si può interpretare come un termine che forza i pattern inferiti ad essere ortogonali, quindi a specializzarsi su segnali diversi.

Per completare il quadro del paradigma insegnante-studente dobbiamo scegliere una tecnica di allenamento della RBM e un misuratore che ci permetta di confrontare ξ con $\xi^\star$ per valutare la qualità dell'apprendimento.
Una strategia di allenamento piuttosto intuitiva consiste nel selezionare i pattern la cui probabilità rispetto alla distribuzione a posteriori è massima: $\xi \coloneqq \operatorname{argmax} \mathrm{P}(\bullet|\{\sigma^a\})$; questa tecnica prende il nome di massimizzazione a posteriori, o MAP. Purtroppo, questa non è una strada praticabile per sistemi di grandi dimensioni poiché lo spazio degli esiti ha cardinalità 2^{NP} quindi una ricerca estensiva diventa troppo onerosa. Si consideri che le reti neuronali vengono generalmente utilizzate per analizzare ingenti quantità di dati; la taglia del sistema deve essere proporzionata affinché la macchina abbia sufficiente margine per adattarsi ai dati senza sovraccaricarsi (situazione che accade ad esempio nella fase vetrosa del modello di Hopfield in alto carico).
Per ovviare a questo inconveniente si può considerare una strategia più limitata: invece che massimizzare la distribuzione congiunta si può massimizzare la marginale per ogni peso della matrice, cioè scegliere $\forall \mu, i\ \xi_i^\mu \coloneqq \operatorname{argmax} \sum_{\xi|\xi_i^\mu=\bullet} \mathrm{P}(\xi|\{\sigma^a\})$; equivalentemente questa scelta si può esprimere tramite il segno della media del peso: $\operatorname{sgn}\langle \xi_i^\mu \rangle_{\mathrm{P}(\xi|\{\sigma^a\})}$ (si tenga presente che se una funzione dipende solamente da una variabile valutarne la media rispetto alla relativa marginale o alla distribuzione congiunta produce lo stesso risultato). Ammesso di essere in grado di calcolare facilmente le distribuzioni marginali, questo calcolo non presenta problemi di scalabilità perché la massimizzazione è fatta rispetto ad una variabile dicotomica. Questa strategia viene chiamata **massimizzazione delle marginali a posteriori** o più sinteticamente MPM, per ulteriori dettagli si può consultare [Nis01, p.81].
La misura di errore che le viene comunemente associata si calcola separatamente per ogni pattern che si vuole inferire, anzi per ogni coppia $(\xi^{\star\mu}, \xi^\nu)$ al variare di μ, ν; infatti, per via della simmetria del problema presente in tutte le distribuzioni che utilizziamo, è possibile ad esempio che il secondo pattern inferisca correttamente il primo pattern originario, o viceversa. Per ciascuna di queste coppie l'indicatore è la media, rispetto ai possibili pattern originari e ai dati che vengono generati a partire da questi, della sovrapponibilità (anche detta overlap, all'inglese) tra il pattern originario e quello inferito; la sovrapponibilità, definita dall'equazione (2.4), non è altro il prodotto scalare sullo spazio χ^N, opportunamente normalizzato affinché stia nell'intervallo $[-1, 1]$. Nello specifico caso della sovrapponibilità tra lo stato del sistema ξ e una variabile fissata $\xi^\star$ si parla più propriamente di magnetizzazione di Mattis.

Un'ulteriore criterio di scelta dei pesi consiste nell'estrazione di ξ secondo la distribuzione di probabilità inversa; ci riferiremo a questa come alla **strategia del campionamento**. A differenza dei precedenti questo metodo non è deterministico perciò la misura di errore dovrà tenerne conto. Solitamente si considerano le stesse sovrapponibilità usate per il MPM, però in questo caso la media viene effettuata anche sulla distribuzione a posteriori oltre che sui dati e sui pattern originari. Siccome la sovrapponibilità è una somma sui vari indici i anche in questo caso si ottiene una media sulle distribuzioni marginali.
Di seguito riassumiamo gli indicatori usati nella valutazione dell'apprendimento nel caso del MPM e del campionamento; sono da intendersi al variare degli indici $1 \leq \mu, \nu \leq P$.

$$\begin{aligned} &\frac{1}{N}\sum_i \mathbb{E}_{\mathrm{P}(\xi^\star,\{\sigma^a\})}\left[\xi_i^{\star\mu}\,\mathrm{sgn}\langle\xi_i^\nu\rangle_{\mathrm{P}(\xi_i^\nu|\{\sigma^a\})}\right] \\ &\frac{1}{N}\sum_i \mathbb{E}_{\mathrm{P}(\xi^\star,\{\sigma^a\})}\left[\xi_i^{\star\mu}\langle\xi_i^\nu\rangle_{\mathrm{P}(\xi_i^\nu|\{\sigma^a\})}\right] \end{aligned} \tag{3.3}$$

In queste espressioni ci siamo permessi un abuso di notazione per indicare le marginali di $\mathrm{P}(\xi|\{\sigma^a\})$: abbiamo sfruttato il nome ξ_i^ν dato alla variabile della funzione di densità per indicare che questa è da intendersi come funzione esclusivamente di quella variabile, perciò marginalizzata sulle altre. In seguito continueremo ad usare questa convenzione augurandoci che si evincerà facilmente dal contesto la distribuzione di probabilità alla quale ci staremo riferendo.
Prima di procedere aggiungiamo un'annotazione utile nel contesto di un problema inverso con modelli provenienti dalla meccanica statistica. In questo caso la probabilità diretta presenta un parametro di temperatura (inversa) $\beta_{gen.}$ che regola il disordine del sistema. Dunque, anche la probabilità inversa possiederà lo stesso parametro. Tuttavia in fase di inferenza non è detto che sia nota la temperatura usata in fase di generazione dei dati, perciò a priori andrebbero distinte. Ciò significa che in questo caso nella distribuzione inversa compare solo il parametro $\beta_{inf.}$; il parametro $\beta_{gen.}$ è invece presente nella distribuzione dei dati, quindi nel disordine, perciò entra in gioco nel momento in cui si calcolano gli indicatori di inferenza (come quelli appena visti). Quando si studia un sistema supponendo che $\beta_{gen.} = \beta_{inf.}$ si dice che questo si trova sulla *linea di Nishimori*. Noi concentreremo il nostro studio in questo caso, per questo non abbiamo distinto le temperature fin dall'inizio.

Dunque, per procedere con l'analisi del modello che ci siamo posti di compiere dobbiamo innanzitutto calcolare le quantità presenti nelle equazioni (3.3). Per la nostra analisi ci concentreremo sulla seconda equazione (senza la funzione segno), relativa all'allenamento effettuato tramite la strategia del campionamento, anche detta strategia dell'*inferenza*

bayesiana in certi contesti.
Il calcolo diretto delle medie non è fattibile a causa delle dimensioni dei sistemi coinvolti, perciò applichiamo le tecniche di meccanica statistica viste nei capitoli precedenti. Come abbiamo spiegato nella sezione 1.4, possiamo interpretare la distribuzione inversa (3.2) come una distribuzione di Boltzmann-Gibbs con disordine piantato (*planted disorder* in inglese); in questo modo le quantità che vogliamo calcolare sono proprio le medie delle magnetizzazioni dei vari pattern su quelli originali.
Un primo modo per calcolare le medie delle osservabili sfrutta il teorema della risposta lineare 1.7 e ricava la loro media dalle derivate dell'energia libera dopo un'opportuna perturbazione dell'hamiltoniana del sistema. Altrimenti, siccome stiamo analizzando un sistema sulla linea di Nishimori - e in questi casi è noto che le magnetizzazioni automediano (vedi ad esempio [CKS05]) - possiamo utilizzare l'approccio descritto nel caso del modello di Hopfield in (2.31): calcoliamo l'energia libera del modello tramite il metodo di Laplace facendo in modo che le magnetizzazioni si possano esprimere tramite i parametri di integrazione e poi le valutiamo nel punto estremale che otterremo. In realtà a posteriori sappiamo che alcuni parametri naturali saranno proprio le magnetizzazioni stesse, quindi basterà considerare il loro valore nel punto estremale ottenuto con le condizioni stazionarie.
Entrambi i metodi che abbiamo descritto si basano sul calcolo dell'energia libera del modello (o equivalentemente della pressione, come abbiamo visto nel capitolo 1); la sezione 3.2 è interamente dedicata a questo scopo. Precisiamo che per ottenere le magnetizzazioni con il teorema della risposta lineare bisognerebbe perturbare l'hamiltoniana originaria e calcolare l'energia libera di questa; la differenza tra le due è un termine lineare nella perturbazione, avente come coefficiente proprio la magnetizzazione prescelta, quindi la magnetizzazione si ottiene comunque come una coordinata del punto di minimo dell'energia libera non perturbata.
Inoltre osserviamo che lavorando nell'ipotesi di sistemi con un grande numero di neuroni calcoleremo tutte le quantità di interesse nel limite termodinamico ($N \to +\infty$); infatti, i modelli artificiali che cercano di emulare il funzionamento del cervello umano vengono impostati con un numero elevato di unità, proprio per approssimare le dimensioni del sistema neuronale (si stima che nel cervello umano siano presenti $1{,}6 \times 10^{10}$ neuroni).

3.2 Calcolo della pressione

Per quanto detto fin'ora sappiamo che abbiamo bisogno di esprimere l'energia libera del modello (3.2); precisamente, essendo questo un modello disordinato, vogliamo calcolarne

la media sul disordine nel limite termodinamico. Per evitare un ulteriore fattore $-\beta$ calcoleremo invece la pressione, che sappiamo esserle sostanzialmente equivalente, nella sua versione intensiva:

$$\mathrm{a}_{P,\beta} \coloneqq \lim_{N\to+\infty} \frac{1}{N} \overline{\log \Omega_{P,\beta}(\{\sigma^a\})}^{\mathrm{P}(\xi^\star,\{\sigma^a\})}$$

Abbiamo indicato al pedice i parametri principali al variare dei quali saremo interessati a studiare l'andamento della pressione; a questi si aggiunge un ulteriore parametro, in questa espressione implicito, che è il grado di sollecitazione α del sistema. Procedendo con i conti daremo per scontate queste dipendenze. La notazione $\overline{\bullet}^{\mathrm{P}(\xi^\star,\{\sigma^a\})}$ indica la media sul disordine, che sarà calcolata tramite la distribuzione condizionale: $\mathrm{P}(\xi^\star,\{\sigma^a\}) = \mathrm{P}(\xi^\star)\,\mathrm{P}(\{\sigma^a\}|\xi^\star)$. Come abbiamo visto nella sezione 2.2.2, siccome calcolare la media del logaritmo è complicato sfrutteremo il replica trick, perciò calcoleremo la pressione in questo modo:

$$\begin{aligned}
\mathrm{a}_{P,\beta} &= \lim_{N\to+\infty} \lim_{k\to 0^+} \frac{1}{kN} \log \overline{\Omega_{P,\beta}(\{\sigma^a\})^k}^{\mathrm{P}(\xi^\star,\{\sigma^a\})} = \\
&= \lim_{N\to+\infty} \lim_{k\to 0^+} \frac{1}{kN} \log \sum_{\xi^\star} \mathrm{P}(\xi^\star) \sum_{\{\sigma^a\}_{a=1}^M} \left(\prod_a \mathrm{P}(\sigma^a|\xi)\right) \Omega_{P,\beta}(\{\sigma^a\})^k
\end{aligned} \tag{3.4}$$

L'obiettivo dei prossimi calcoli sarà di semplificare la funzione di partizione all'interno del logaritmo esprimendola come integrale, ricalcando quanto fatto per il modello di Hopfield nella sezione 2.2. Iniziamo riarrangiando i termini in modo da poter fattorizzare l'espressione sulle varie repliche.

$$\begin{aligned}
\overline{\Omega_{P,\beta}(\{\sigma^a\})^k}^{\mathrm{P}(\xi^\star,\{\sigma^a\})} &= \sum_{\xi^\star} \mathrm{P}(\xi^\star) \sum_{\{\sigma^a\}_{a=1}^M} \left(\prod_a \mathrm{P}(\sigma^a|\xi^\star)\right) \times \left(\sum_\xi \prod_a \mathrm{P}(\sigma^a|\xi)\right)^k \\
&= \sum_{\xi^\star} \mathrm{P}(\xi^\star) \sum_{\{\sigma^a\}} \left(\prod_a \mathrm{P}(\sigma^a|\xi^\star)\right) \times \left(\sum_{\{\xi^{\gamma\mu}\}_{\mu,\gamma}} \prod_{\gamma,a} \mathrm{P}(\sigma^a|\xi^\gamma)\right) \\
&= \sum_{\xi^\star,\{\xi^{\gamma\mu}\}} \mathrm{P}(\xi^\star) \sum_{\{\sigma^a\}} \prod_a \left(\mathrm{P}(\sigma^a|\xi^\star) \prod_\gamma \mathrm{P}(\sigma^a|\xi^\gamma)\right) \\
&= \sum_{\xi^\star,\{\xi^{\gamma\mu}\}} \mathrm{P}(\xi^\star) \prod_{a=1}^M \sum_\sigma \left(\mathrm{P}(\sigma|\xi^\star) \prod_\gamma \mathrm{P}(\sigma|\xi^\gamma)\right) \\
&= \sum_{\xi^\star,\{\xi^{\gamma\mu}\}} \mathrm{P}(\xi^\star) \left(\sum_\sigma \mathrm{P}(\sigma|\xi^\star) \prod_\gamma \mathrm{P}(\sigma|\xi^\gamma)\right)^M
\end{aligned} \tag{3.5}$$

L'ultimo passaggio è dovuto al fatto che il termine è indipendente dall'indice di replica (come è naturale aspettarsi). Prima di continuare è il caso di soffermarsi un momento a commentare l'espressione della probabilità diretta

$$\mathrm{P}(\sigma|\xi) = \mathcal{Z}(\xi)^{-1} \prod_{\mu} \cosh(\frac{\beta}{\sqrt{N}} \sum_{i} \xi_i^{\mu} \sigma_i) \tag{3.6}$$

Nel caso in cui ci sia un solo nodo nascosto ($P = 1$) la funzione di partizione $\mathcal{Z}$ non dipende dal pattern ξ, infatti tramite una trasformazione di Gauge si ottiene l'equivalenza $\mathrm{Tr}_\sigma \cosh(\beta/\sqrt{N} \sum_i \xi_i \sigma_i) \overset{\sigma_i \mapsto \xi_i \sigma_i}{=} \mathrm{Tr}_\sigma \cosh(\beta/\sqrt{N} \sum_i \sigma_i)$. Ricordiamo che Tr_σ è una notazione equivalente per esprimere la somma sulle configurazioni, in questo caso dei nodi visibili; capiterà di indicare in apice il numero di variabili, ad esempio $\mathrm{Tr}_\sigma^{(N)}$.
D'altra parte, per $P \geq 2$ non è possibile procedere analogamente poiché tutte le sinapsi sono accoppiate con lo stesso σ; a ben guardare questa interdipendenza è esplicita nel modello e non eludibile: ogni nodo visibile influenza tutti quelli nascosti contemporaneamente. Questo collegamento non solo fa dipendere la funzione di partizione da ciascun pattern, ma precisamente dall'insieme di questi. Sebbene non sia possibile fattorizzare i contributi di ciascuna sinapsi, possiamo esprimere $\mathcal{Z}$ in funzione delle correlazioni tra questi (cioè le sovrapponibilità) e ciò faciliterà notevolmente i calcoli che seguiranno.
Anticipiamo al risultato un lemma che verrà usato anche più avanti.

Lemma 3.1 — *Dati $P \in \mathbb{N}_+$ e $x_1, \ldots, x_P \in \mathbb{R}$ vale l'uguaglianza*

$$\prod_{\mu=1}^{P} \cosh(x_\mu) = \frac{1}{2^P} \mathrm{Tr}_\tau^{(P)} \cosh\left(\sum_{\mu} \tau_\mu x_\mu\right)$$

Dimostrazione. La dimostrazione si può fare per induzione. Il caso $P = 1$ discende dalla parità della funzione coseno iperbolico. Il caso generale segue dalla seguente proprietà: $\cosh(x)\cosh(y) = (\cosh(x+y) + \cosh(x-y))/2$; infatti assunta l'ipotesi induttiva per $P-1$ possiamo scrivere

$$\begin{aligned}\prod_{\mu=1}^{P} \cosh(x_\mu) &= \frac{1}{2^{P-1}} \sum_{\tau_1,\ldots,\tau_{P-1}=\pm 1} \cosh\left(\sum_{\mu=1}^{P-1} \tau_\mu x_\mu\right) \cosh(x_\mu) \\ &= \frac{1}{2^P} \sum_{\tau_1,\ldots,\tau_{P-1}=\pm 1} \left(\cosh\left(\sum_{\mu=1}^{P-1} \tau_\mu x_\mu + x_P\right) + \cosh\left(\sum_{\mu=1}^{P-1} \tau_\mu x_\mu - x_P\right)\right)\end{aligned}$$

da cui segue la tesi. □

Teorema 3.2 — *La funzione di partizione del modello diretto con N nodi visibili e P nodi nascosti è esprimibile in queste forme:*

$$\begin{aligned}\mathcal{Z}(\xi) &= 2^{N-P}\mathrm{Tr}_\tau^{(P)} \exp\left(\frac{\beta^2}{2}\sum_{\mu,\nu=1}^{N}\tau_\mu\tau_\nu\frac{1}{N}\sum_i \xi_i^\mu\xi_i^\nu\right)\\ &= 2^{N-P}e^{P\beta^2/2}\mathrm{Tr}_\tau^{(P)} \exp\left(\beta^2\sum_{1\le\mu<\nu\le P}\tau_\mu\tau_\nu\frac{1}{N}\sum_i \xi_i^\mu\xi_i^\nu\right)\end{aligned}$$

Dimostrazione. Prima usiamo un paio di volte il lemma 3.1 per trasformare la somma sugli N spin visibili σ in una somma sui P spin nascosti τ, dopodiché possiamo sfruttare la già citata espansione del termine $\log\cosh$ (si veda il lemma 1.3).

$$\begin{aligned}\mathcal{Z}(\xi) &= \mathrm{Tr}_\sigma^{(N)}\prod_\mu \cosh\left(\frac{\beta}{\sqrt{N}}\sum_i \sigma_i\xi_i^\mu\right) = \mathrm{Tr}_\sigma^{(N)}\frac{1}{2^P}\mathrm{Tr}_\tau^{(P)}\cosh\left(\sum_\mu\tau_\mu\frac{\beta}{\sqrt{N}}\sum_i\sigma_i\xi_i^\mu\right)\\ &= \frac{2^N}{2^P}\mathrm{Tr}_\tau^{(P)}\frac{1}{2^N}\mathrm{Tr}_\sigma^{(N)}\cosh\left(\sum_i\sigma_i\sum_\mu\tau_\mu\frac{\beta}{\sqrt{N}}\xi_i^\mu\right) = 2^{N-P}\mathrm{Tr}_\tau^{(P)}\prod_i\cosh\left(\sum_\mu\tau_\mu\frac{\beta}{\sqrt{N}}\xi_i^\mu\right)\\ &= 2^{N-P}\mathrm{Tr}_\tau^{(P)}\exp\left(\sum_i\frac{\beta^2}{2N}\left(\sum_\mu\tau_\mu\xi_i^\mu\right)^2\right) = 2^{N-P}\mathrm{Tr}_\tau^{(P)}\exp\left(\frac{\beta^2}{2}\sum_{\mu,\nu=1}^{N}\tau_\mu\tau_\nu\frac{1}{N}\sum_i\xi_i^\mu\xi_i^\nu\right)\end{aligned}$$

□

Osserviamo che a parte il fattore moltiplicativo si può interpretare come la funzione di partizione di un sistema di Curie-Weiss a P spin (visti nell'esempio 1.11), senza campo esterno e i cui coefficienti di interazione sono un multiplo delle sovrapponibilità tra i pattern.
Inoltre, per facilitare il calcolo del limite termodinamico semplifichiamo il fattore 2^N dove possibile: si osservi che la distribuzione di equilibrio (3.2), la quale presenterebbe M fattori 2^N derivanti dalla probabilità diretta su σ^a, resta invariata se assorbiamo questi fattori nella funzione di partizione; possiamo cioè moltiplicare ogni $\mathrm{P}(\sigma^a|\xi)$ per 2^N a patto di fare lo stesso nella funzione di partizione Ω. Questo naturalmente non è possibile nella probabilità diretta perché non sarebbe più normalizzata, ma nell'inversa si può semplificare perché la stessa è presente anche al denominatore. Nell'equazione finale di (3.5) questo significa moltiplicare ciascun termine $\mathrm{P}(\sigma|\xi^\gamma)$ per 2^N, ad esempio al denominatore usando $2^{-N}\,\mathcal{Z}(\xi^\gamma)$ al posto di $\mathcal{Z}(\xi^\gamma)$.
Per semplicità raccogliamo le funzioni di normalizzazione in un unico fattore $\mathcal{W} \coloneqq \mathcal{Z}(\xi^\star)\prod_\gamma 2^{-N}\,\mathcal{Z}(\xi^\gamma)$. Sempre per semplicità di scrittura, d'ora in poi indicheremo la media sui pattern originari con $\mathbb{E}_{\xi^\star}$ intendendola applicata a ciò che segue.

Ricapitolando, possiamo esprimere la media sul disordine della funzione di partizione replicata $\overline{\Omega_{P,\beta}^k}$ in questo modo:

$$\mathbb{E}_{\xi^\star} \sum_{\{\xi^{\gamma\mu}\}} \mathcal{W}^{-M} \left(\mathrm{Tr}_\sigma \prod_\mu \left(\cosh(\frac{\beta}{\sqrt{N}} \sum_i \xi_i^{\star\mu} \sigma_i) \prod_\gamma \cosh(\frac{\beta}{\sqrt{N}} \sum_i \xi_i^{\gamma\mu} \sigma_i) \right) \right)^M$$

Fino a questo punto la trattazione ha seguito la stessa linea intrapresa in [HWH19] nel caso $P = 2$. Tuttavia la presenza dei coseni iperbolici rende il calcolo piuttosto involuto e difficilmente generalizzabile per un generico $P \in \mathbb{N}_+$. L'idea che ci ha permesso di evitare queste difficoltà consiste nel ritornare all'espressione dei coseni iperbolici tramite esponenziali; infatti ricordiamo che questi nel caso di una RBM dicotomica appaiono come marginalizzazione sui nodi nascosti. Dunque, usiamo una variabile dicotomica $\tau_\bullet$ per l'espansione di ogni coseno iperbolico (come da lemma 1.3, punto 1). Il termine che comincia con la somma sulle configurazioni all'interno della parentesi esterna verrà indicato con $\mathcal{N}$; per quanto abbiamo appena detto si può scrivere in questo modo:

$$\begin{aligned}
\mathcal{N} \coloneqq \mathrm{Tr}_\sigma^{(N)} \prod_\mu \Bigg[& \left(\frac{1}{2} \sum_{\tau_{\star\mu}=\pm 1} \exp(\frac{\beta}{\sqrt{N}} \sum_i \xi_i^{\star\mu} \sigma_i \tau_{\star\mu}) \right) \\
& \times \prod_\gamma \left(\frac{1}{2} \sum_{\tau_{\gamma\mu}=\pm 1} \exp(\frac{\beta}{\sqrt{N}} \sum_i \xi_i^{\gamma\mu} \sigma_i \tau_{\gamma\mu}) \right) \Bigg] \\
= 2^{-P(k+1)} \mathrm{Tr}_\sigma^{(N)} & \sum_{\substack{\{\tau_{\star\mu}=\pm 1\}_\mu \\ \{\tau_{\gamma\mu}=\pm 1\}_{\mu,\gamma}}} \exp\left(\frac{\beta}{\sqrt{N}} \sum_i \sigma_i \sum_\mu \left(\xi_i^{\star\mu} \tau_{\star\mu} + \sum_\gamma \xi_i^{\gamma\mu} \tau_{\gamma\mu} \right) \right)
\end{aligned}$$

Ora, siccome nel nostro studio vogliamo confrontare i pattern originali con quelli inferiti, manipoliamo l'equazione affinché spariscano le variabili relative al disordine (le σ dei nodi visibili) e compaiano le sovrapponibilità tra i pattern - per il termine $\mathcal{W}$ è stato già fatto con il teorema 3.2; per far questo fattorizziamo l'espressione sui σ_i e poi sviluppiamo il coseno iperbolico.

$$\begin{aligned}
\mathcal{N} &= 2^{-P(k+1)} \sum_{\substack{\{\tau_{\star\mu}=\pm 1\}_\mu \\ \{\tau_{\gamma\mu}=\pm 1\}_{\mu,\gamma}}} 2^N \prod_i \cosh\left(\sum_\mu \left(\xi_i^{\star\mu} \tau_{\star\mu} + \sum_\gamma \xi_i^{\gamma\mu} \tau_{\gamma\mu} \right) \right) \\
&= 2^{-P(k+1)+N} \sum_{\substack{\{\tau_{\star\mu}=\pm 1\}_\mu \\ \{\tau_{\gamma\mu}=\pm 1\}_{\mu,\gamma}}} \exp\left(\sum_i \frac{\beta^2}{2N} \left(\sum_\mu (\xi_i^{\star\mu} \tau_{\star\mu} + \sum_\gamma \xi_i^{\gamma\mu} \tau_{\gamma\mu}) \right)^2 \right)
\end{aligned}$$

$$
\begin{aligned}
&= 2^{-P(k+1)+N} \sum_{\substack{\{\tau_{\star\mu}=\pm 1\}_\mu \\ \{\tau_{\gamma\mu}=\pm 1\}_{\mu,\gamma}}} \exp\left(\frac{\beta^2}{2N}\sum_{\mu,\nu}\sum_i\left(\xi_i^{\star\mu}\tau_{\star\mu}+\sum_\gamma \xi_i^{\gamma\mu}\tau_{\gamma\mu}\right)^2\right) \\
&= 2^{-P(k+1)+N} \sum_{\substack{\{\tau_{\star\mu}=\pm 1\}_\mu \\ \{\tau_{\gamma\mu}=\pm 1\}_{\mu,\gamma}}} \exp\left(\frac{\beta^2}{2}\sum_{\mu,\nu}\frac{1}{N}\sum_i\Big(\xi_i^{\star\mu}\xi_i^{\star\nu}\tau_{\star\mu}\tau_{\star\nu}\right. \\
&\qquad \left. +\sum_\gamma(\xi_i^{\star\mu}\xi_i^{\gamma\nu}\tau_{\star\mu}\tau_{\gamma\nu}+\xi_i^{\gamma\mu}\xi_i^{\star\nu}\tau_{\gamma\mu}\tau_{\star\nu})+\sum_{\gamma,\tilde\gamma}\xi_i^{\gamma\mu}\xi_i^{\tilde\gamma\nu}\tau_{\gamma\mu}\tau_{\tilde\gamma\nu})\right) \\
&= 2^{-P(k+1)+N}e^{\beta^2Pk/2} \sum_{\substack{\{\tau_{\star\mu}=\pm 1\}_\mu \\ \{\tau_{\gamma\mu}=\pm 1\}_{\mu,\gamma}}} \exp\left(\frac{\beta^2}{2}\sum_{\mu,\nu}\tau_{\star\mu}\tau_{\star\nu}\frac{1}{N}\sum_i\xi_i^{\star\mu}\xi_i^{\star\nu}\right. \\
&\qquad +\beta^2\sum_{\mu,\nu;\gamma}\tau_{\star\mu}\tau_{\gamma\nu}\frac{1}{N}\sum_i\xi_i^{\star\mu}\xi_i^{\gamma\nu}+\beta^2\sum_{\mu<\nu;\gamma}\tau_{\gamma\mu}\tau_{\gamma\nu}\frac{1}{N}\sum_i\xi_i^{\gamma\mu}\xi_i^{\gamma\nu} \\
&\qquad\qquad \left.+\beta^2\sum_{\mu,\nu;\gamma<\tilde\gamma}\tau_{\gamma\mu}\tau_{\tilde\gamma\nu}\frac{1}{N}\sum_i\xi_i^{\gamma\mu}\xi_i^{\tilde\gamma\nu}\right)
\end{aligned}
$$

Abbiamo ottenuto un'espressione per la funzione di partizione totale in termini degli accoppiamenti tra i pattern: $\overline{\Omega_{P,\beta}^k} = \mathbb{E}_{\xi^\star}\ \sum_{\{\xi^{\gamma\mu}\}}(\mathcal{N}\,/\,\mathcal{W})^M$ con $\mathcal{N}$ e $\mathcal{W}$ esprimibili in funzione di variabili nella forma $\mathrm{q}(\xi^{\cdots},\xi^{\cdots})$.
Ora, analogamente a quanto fatto nella sezione 2.2 sfruttiamo il lemma 2.8 per rappresentare il termine $(\mathcal{N}\,/\,\mathcal{W})^M$ tramite un integrale. Introduciamo una delta di Dirac in rappresentazione di Fourier per ogni sovrapponibilità presente, esclusa quella tra i pattern originari perché per $N\to+\infty$ tende ad un valore predeterminato. Di seguito elenchiamo per ogni sovrapponibilità il nome della relativa variabile di integrazione; come visto nella sezione 2.2, quelle di sinistra si diranno parametri d'ordine, quelle di destra (con il cappuccio $\hat{\bullet}$) parametri d'ordine coniugati o ausiliari. Sempre in quella sezione è presente una rassicurazione sul fatto che, nonostante l'introduzione delle delte di Dirac introduca dei numeri complessi, la pressione che stiamo calcolando resta a valori reali.

$$
\begin{aligned}
&m^{\star\gamma\mu\nu},\hat m^{\star\gamma\mu\nu} \quad \text{per } \frac{1}{N}\sum_j \xi_j^{\star\mu}\xi_j^{\gamma\mu} && \forall 1\le\mu,\nu\le P; 1\le\gamma\le k \\
&q^{\gamma\gamma\mu\nu},\hat q^{\gamma\gamma\mu\nu} \quad \text{per } \frac{1}{N}\sum_j \xi_j^{\gamma\mu}\xi_j^{\gamma\mu} && \forall 1\le\mu<\nu\le P; 1\le\gamma\le k \\
&q^{\gamma\tilde\gamma\mu\nu},\hat q^{\gamma\tilde\gamma\mu\nu} \quad \text{per } \frac{1}{N}\sum_j \xi_j^{\gamma\mu}\xi_j^{\tilde\gamma\mu} && \forall 1\le\mu,\nu\le P; 1\le\gamma<\tilde\gamma\le k
\end{aligned}
\tag{3.7}
$$

In totale sono $2(P^2k+P(P-1)k/2+P^2k(k-1)/2) = kP(2P+Pk-1)$. Dall'integrazione

risulta

$$\overline{\Omega^k_{P,\beta}} = \mathbb{E}_{\xi^\star} \sum_{\{\xi^{\gamma\mu}\}} \int\cdots\int \begin{smallmatrix}\{\mathrm{dm}^{\star\gamma\mu\nu}\}\\ \{\mathrm{d}\hat{\mathrm{m}}^{\star\gamma\mu\nu}\}\end{smallmatrix}_{\substack{1\le\mu,\nu\le P\\ 1\le\gamma\le k}} \begin{smallmatrix}\{\mathrm{dq}^{\gamma\gamma\mu\nu}\}\\ \{\mathrm{d}\hat{\mathrm{q}}^{\gamma\gamma\mu\nu}\}\end{smallmatrix}_{\substack{1\le\mu<\nu\le P\\ 1\le\gamma\le k}} \begin{smallmatrix}\{\mathrm{dq}^{\gamma\tilde{\gamma}\mu\nu}\}\\ \{\mathrm{d}\hat{\mathrm{q}}^{\gamma\tilde{\gamma}\mu\nu}\}\end{smallmatrix}_{\substack{1\le\mu,\nu\le P\\ 1\le\gamma<\tilde{\gamma}\le k}}$$
$$(2\pi)^{-kP(2P+Pk-1)/2} \exp\left(i\sum_{\mu,\nu}\sum_{\gamma} \hat{m}^{\star\gamma\mu\nu}\left(\frac{1}{N}\sum_j \xi_j^{\star\mu}\xi_j^{\gamma\nu} - m^{\star\gamma\mu\nu}\right)\right.$$
$$\left. + i\sum_{\mu<\nu}\sum_{\gamma} \hat{q}^{\gamma\gamma\mu\nu}\left(\frac{1}{N}\sum_j \xi_j^{\gamma\mu}\xi_j^{\gamma\nu} - q^{\gamma\gamma\mu\nu}\right) + i\sum_{\mu,\nu}\sum_{\gamma<\tilde{\gamma}} \hat{q}^{\gamma\tilde{\gamma}\mu\nu}\left(\frac{1}{N}\sum_j \xi_j^{\gamma\mu}\xi_j^{\tilde{\gamma}\nu} - \hat{q}^{\gamma\gamma\mu\nu}\right)\right)$$
$$+ \exp(M\log(\mathcal{N}) - M\log(\mathcal{W}))$$

dove $\mathcal{N}$ e $\mathcal{W}$ si intendono calcolati nelle variabili di integrazione ausiliarie di (3.7). Osserviamo per inciso che in $\mathcal{Z}(\xi^\gamma)$ bisogna usare la seconda notazione esposta nel teorema 3.2 perché il parametro $q^{\gamma\gamma\mu\nu}$ è definito solo per $\mu < \nu$. Per semplificare le notazioni d'ora in poi indicheremo le variabili di integrazione con $\{\mathrm{d}\ldots\}$. Inoltre osserviamo che possiamo operare una dilatazione di ogni parametro coniugato $y \mapsto Ny$; la trasformazione interessa solo il primo esponenziale e fa apparire nell'integrale anche un termine moltiplicativo N per ogni cambio di variabili.
L'introduzione degli integrali ha fatto sì che la somma su $\xi^{\gamma\mu}$ interessi solo alcuni termini, escludendo in particolare tutto il secondo esponenziale contenente $\mathcal{N}$ e $\mathcal{W}$. Nonostante non sia immediato da verificare, anche per quanto riguarda la media su $\xi^\star$ si ha un effetto analogo: nel limite termodinamico si può calcolare separatamente per il primo e il secondo esponenziale, e in quest'ultimo la dipendenza addirittura scompare. La prova di questa affermazione verrà fornita a breve.
In conclusione, si arriva a esprimere l'equazione (3.4) dalla quale siamo partiti in una forma adatta all'applicazione del metodo di Laplace illustrato nella sezione 2.2 - anche qui supponendo arbitrariamente che sia legittimo scambiare i limiti in k e N:

$$\mathrm{A}_{P,\beta,\alpha} = \lim_{k\to 0^+}\lim_{N\to+\infty} \frac{1}{kN}\log\int\cdots\int\{\mathrm{d}\ldots\}\left(\frac{N}{2\pi}\right)^{kP(2P+Pk-1)/2} \exp(NG_{N,k}) \tag{3.8}$$

con

$$\begin{aligned} G_{N,k} &\coloneqq G_{(\mathrm{Q})} + G_{(\mathrm{S})} + \alpha G_{(\mathrm{E})} \\ G_{(\mathrm{Q})} &\coloneqq -i\left(\sum_{\mu,\nu}\sum_{\gamma} m^{\star\gamma\mu\nu}\hat{m}^{\star\gamma\mu\nu} + \sum_{\mu<\nu}\sum_{\gamma} q^{\gamma\gamma\mu\nu}\hat{q}^{\gamma\gamma\mu\nu} + \sum_{\mu,\nu}\sum_{\gamma<\tilde{\gamma}} q^{\gamma\tilde{\gamma}\mu\nu}\hat{q}^{\gamma\tilde{\gamma}\mu\nu}\right) \end{aligned} \tag{3.9}$$

$$G_{(\mathrm{S})} \coloneqq \frac{1}{N} \log \mathbb{E}_{\xi^\star} \sum_{\{\xi^{\gamma\mu}\}} \prod_j \exp\left(i \sum_{\mu,\nu} \sum_\gamma \hat{m}^{\star\gamma\mu\nu} \xi_j^{\star\mu} \xi_j^{\gamma\nu} + i \sum_{\mu<\nu} \sum_\gamma \hat{q}^{\gamma\gamma\mu\nu} \xi_j^{\gamma\mu} \xi_j^{\gamma\nu} + i \sum_{\mu,\nu} \sum_{\gamma<\tilde{\gamma}} \hat{q}^{\gamma\tilde{\gamma}\mu\nu} \xi_j^{\gamma\mu} \xi_j^{\tilde{\gamma}\nu} \right)$$
$$G_{(\mathrm{E})} \coloneqq \log \mathcal{N} - \log \mathcal{W} \tag{3.10}$$

Ricordiamo che $\alpha = M/N$ rappresenta il grado di sollecitazione del sistema di inferenza, talvolta chiamato carico. Il termine $G_{N,k}$ si dice **pseudo energia libera** mentre $G_{(\mathrm{S})}$ e $G_{(\mathrm{E})}$ sono rispettivamente il **termine entropico** e il **termine energetico**.
Riprendiamo ora la dimostrazione del fatto che la media su $\xi^\star$ interessi solo il termine entropico. Per cominciare osserviamo che fino a questo punto non abbiamo ancora sfruttato il fatto che la media non interessi il termine energetico (a parte per la comodità di scriverlo separatamente). Ora dovremo applicare il metodo di Laplace e in ultima analisi calcolare il limite termodinamico della pseudo energia libera. Dunque è sufficiente dimostrare che al limite $G_{(\mathrm{E})}$ non dipende dalla media sul disordine; racchiudiamo il risultato nel prossimo lemma.

Lemma 3.3 — *Sia $(X_i)_{\mathbb{N}_+}$ una successione di variabili aleatorie i.i.d. a valori in $\mathbb{R}^P$, con $P \in \mathbb{N}_+$, la cui distribuzione abbia media nulla e matrice di covarianza* C. *Siano $(f_N)_{\mathbb{N}_+}$ una successione di funzioni tale che $f_N(X_1, \dots, X_N)$ converga in distribuzione e $g\colon \mathbb{R}^{P\times P} \to \mathbb{R}$ una funzione continua. Allora vale il seguente limite:*

$$\lim_{N\to+\infty} \mathbb{E}\left[f_N(X_1,\dots,X_N)\, g(\{\frac{1}{N}\sum_i X_i^\mu X_i^\nu\}_{\mu,\nu}) \right] = g(\mathrm{C}) \lim_{N\to+\infty} \mathbb{E}\left[f_N(X_1,\dots,X_N) \right]$$

Nel caso in esame le variabili aleatorie sono i vettori $\xi_i^\star$, la cui matrice di correlazione abbiamo stabilito essere $\mathcal{Q}$; la funzione g corrisponde al termine energetico (che non dipende da N al di fuori delle sommatorie $1/N \sum_j \xi_j^{\star\mu} \xi_j^{\star\nu}$, si noti anche che α è una costante); la funzione f comprende tutti gli altri termini della pseudo energia libera e vedremo in seguito che converge in distribuzione perché riusciremo ad eliminare la sua dipendenza da N. Dunque l'applicazione del metodo del punto di sella all'integrale in (3.8) risulta corretta calcolando la media $\mathbb{E}_{\xi^\star}$ esclusivamente sul termine entropico e sostituendo $\mathcal{Q}_{\mu\nu}$ al posto di $1/\sum_j \xi_j^{\star\mu} \xi_j^{\star\nu}$ nell'espressione del termine energetico. Esibiamo ora la prova del risultato.

Dimostrazione. Innanzitutto per la legge dei grandi numeri, applicabile grazie all'ipotesi di variabili i.i.d., vale la convergenza in $\mathbb{L}^2$: $1/N \sum_i X_i^\mu X_i^\nu \to \mathrm{C}_{\mu\nu}$. Allora, per continuità di

g vale anche la convergenza in probabilità $g(\{1/N \sum_i X_i^\mu X_i^\nu\}_{\mu,\nu}) \to g(\mathrm{C})$. Per concludere è sufficiente applicare il teorema di Slutsky considerando $f_N(X_1, \dots, X_N)$ come una successione di variabili aleatorie, intesa al variare di N. □

Ripartendo dall'equazione (3.8), prima di applicare il metodo di Laplace imponiamo un'ipotesi di simmetria di replica, analogamente a quanto visto per il modello di Hopfield nella sezione 2.2. Abbiamo già visto nella sezione 2.2.2 che possiamo imporre l'ipotesi di simmetria prima di usare il metodo del punto di sella ottenendo lo stesso risultato. Siccome il modello che analizziamo è sulla linea di Nishimori ci aspettiamo che l'assunzione di simmetria di replica sia soddisfatta [Nis01; MM09]. Come di consueto assumiamo che il massimo della pseudo energia libera si abbia in un punto nel quale i parametri non dipendono dalla replica γ alla quale si riferiscono, cioè supponiamo che

$$m^{\star\mu\nu} := m^{\star\gamma\mu\nu} \qquad \hat{m}^{\star\mu\nu} := i\hat{m}^{\star\gamma\mu\nu} \qquad \forall\gamma;\mu,\nu \tag{3.11}$$

$$q^{=\mu\nu} := q^{\gamma\gamma\mu\nu} \qquad \hat{q}^{=\mu\nu} := i\hat{q}^{\gamma\gamma\mu\nu} \qquad \forall\gamma;\mu<\nu \tag{3.12}$$

$$q^{\neq\mu\nu} := q^{\gamma\tilde{\gamma}\mu\nu} \qquad \hat{q}^{\neq\mu\nu} := i\hat{q}^{\gamma\tilde{\gamma}\mu\nu} \qquad \forall\gamma<\tilde{\gamma};\mu,\nu \tag{3.13}$$

Nello stabilire questa notazione abbiamo anche inglobato l'unità immaginaria nei parametri coniugati per semplificare le scritture; come è stato spiegato nella sezione 2.2.2 questo non comporta alterazioni nel calcolo dell'energia libera né nel valore dei parametri d'ordine non coniugati nel punto estremale. Con questo vincolo i parametri da considerare sono $P(5P-1)/2$ (tra normali e ausiliari). Allo scopo di facilitare il confronto con [HWH19] elenchiamo le identificazioni tra i parametri usati in quell'articolo nel caso $P = 2$ e quelli appena definiti per il presente elaborato.

$$\begin{aligned} m^{\star 11} &= T_1, & m^{\star 22} &= T_2 \\ m^{\star 12} &= \tau_1, & m^{\star 21} &= \tau_2 \\ q^{\neq 11} &= q_1, & q^{\neq 22} &= q_2 \\ q^{=12} &= R, & q^{\neq 12}, q^{\neq 21} &= r \end{aligned}$$

Si osservi che nell'articolo viene fatta l'assunzione ulteriore, verificabile sperimentalmente, $q^{\neq\mu\nu} = q^{\neq\nu\mu}$ che noi non faremo per evitare confusioni; questa discrepanza non altera i risultati ma va tenuta presente perché nel caso $P = 2$ la nostra trattazione ha 9 parametri, non 8.

Sotto queste ipotesi i termini della pseudo energia libera definita poc'anzi si semplificano

come segue: in (3.9) le sommatorie non dipendono più dalle repliche quindi si sostituiscono con un fattore moltiplicativo; in (3.10) le funzioni di partizioni delle repliche in $\mathcal{W}$ sono tutte uguali, quindi il loro prodotto si può sostituire con un elevamento alla k; per quanto riguarda gli altri termini il vantaggio dovuto alla simmetria è meno visibile e risulterà evidente più avanti. Anticipiamo che questa assunzione permette di raccogliere un fattore moltiplicativo davanti alla pseudo energia libera che compenserà il fattore $1/k$ e quindi farà in modo che il limite in $k \to 0^+$ non diverga.
Per procedere applichiamo il metodo di Laplace (si veda il lemma 2.5) all'equazione (3.8) ottenendo

$$\mathrm{A}_{P,\beta,\alpha} = \lim_{k\to 0^+} \frac{1}{k} \operatorname{extr} \lim_{N\to+\infty} {G_{N,k}}^{\mathrm{RS}} \tag{3.14}$$

Il termine $(N/2\pi)^{\cdots}$ non è presente perché tende a zero nel limite termodinamico; extr indica il punto estremale - ossia il punto in cui si annullano tutte le derivate - della funzione alla quale converge uniformemente la pseudo energia libera sotto l'assunzione di simmetria di replica, i cui argomenti sono i parametri definiti in (3.11).

Nota 3.4 — Utilizziamo la notazione punto estremale e non punto di massimo perché l'individuazione del punto dominante non è affatto semplice. Oltre alle considerazioni fatte nella nota 2.6, in questo caso l'ulteriore difficoltà è dovuto principalmente all'incertezza generata dal metodo delle repliche, come già noto a partire dai lavori di Giorgio Parisi sul modello SK [Par80b; Par80a]. Infatti la successione di funzioni ${G_{N,k}}^{\mathrm{RS}}$ dipende sia da N che da k e per ottenere l'equazione (3.8) abbiamo scambiato i due limiti, una procedura per niente innocente: basti pensare al fatto che la funzione x^k cambia concavità quando k attraversa il valore 1. Può quindi capitare che al posto del punto di massimo sia necessario considerare il punto di minimo per ottenere un risultato corretto. Pertanto, d'ora in poi ci riferiremo al punto estremale della funzione sottintendendo che il punto corretto va scelto euristicamente tra quelli estremali della funzione.

Per chiarezza espositiva richiamiamo qui i tre addendi (dei quali il terzo va moltiplicato per α) di cui si compone ${G_k}^{\mathrm{RS}}$.

$$G^{\mathrm{RS}}_{(\mathrm{Q})} \coloneqq -k \sum_{\mu,\nu} m^{\star\mu\nu} \hat{m}^{\star\mu\nu} - k \sum_{\mu<\nu} q^{=\mu\nu} \hat{q}^{=\mu\nu} - \frac{k(k-1)}{2} \sum_{\mu,\nu} q^{\neq\mu\nu} \hat{q}^{\neq\mu\nu} \tag{3.15}$$

$$G^{\mathrm{RS}}_{(\mathrm{S})} \coloneqq \lim_{N\to+\infty} \frac{1}{N} \log \mathbb{E}_{\xi^\star} \sum_{\{\xi^{\gamma\mu}\in\Sigma\}_{\mu,\gamma}} \prod_{j} \tag{3.16}$$

$$\exp\left(\sum_{\mu,\nu} \hat{m}^{\star\mu\nu} \xi_j^{\star\mu} \sum_{\gamma} \xi_j^{\gamma\nu} + \sum_{\mu<\nu} \hat{q}^{=\mu\nu} \sum_{\gamma} \xi_j^{\gamma\mu} \xi_j^{\gamma\nu} + \sum_{\mu,\nu} \hat{q}^{\neq\mu\nu} \sum_{\gamma<\tilde{\gamma}} \xi_j^{\gamma\mu} \xi_j^{\tilde{\gamma}\nu} \right)$$

$$G_{(\mathrm{E})}^{\mathrm{RS}} \coloneqq \log \frac{\mathcal{N}^{\mathrm{RS}}}{2^{P-N}\,\mathcal{Z}(\xi^\star)} - k \log \mathcal{Z}(q^{=\mu\nu}) \tag{3.17}$$

dove

$$\mathcal{N}^{\mathrm{RS}} \coloneqq \sum_{\substack{\{\tau_{\star\mu}=\pm 1\}_\mu \\ \{\tau_{\gamma\mu}=\pm 1\}_{\mu,\gamma}}} \exp\left(\frac{\beta^2}{2}\sum_{\mu,\nu}\tau_{\star\mu}\tau_{\star\nu}\,\mathcal{Q}_{\mu\nu} + \beta^2 \sum_{\mu,\nu} m^{\star\mu\nu}\tau_{\star\mu}\sum_{\gamma}\tau_{\gamma\nu}\right. \tag{3.18}$$

$$\left. + \beta^2 \sum_{\mu<\nu} q^{=\mu\nu} \sum_{\gamma}\tau_{\gamma\mu}\tau_{\gamma\nu} + \beta^2 \sum_{\mu,\nu} q^{\neq\mu\nu} \sum_{\gamma<\tilde{\gamma}}\tau_{\gamma\mu}\tau_{\tilde{\gamma}\nu}\right)$$

$$\mathcal{Z}(q^{=\mu\nu}) \coloneqq \mathrm{Tr}_\tau^{(P)} \exp\left(\beta^2 \sum_{1\le\mu<\nu\le P} \tau_\mu\tau_\nu q^{=\mu\nu}\right) \tag{3.19}$$

e precisiamo che il termine $2^{P-N}\,\mathcal{Z}(\xi^\star)$, oltre a non dipendere da N, non dipende nemmeno da $\xi^\star$ ma solamente da $\mathcal{Q}$, perché abbiamo effettuato la sostituzione nell'applicazione del lemma 3.3.

Prima di continuare dimostriamo che il termine $G_{(\mathrm{S})}^{\mathrm{RS}}$ non dipende da N e che quindi il limite termodinamico è pleonastico - in sostanza è già stato calcolato quando abbiamo applicato il metodo di Laplace. Osserviamo che si possono fattorizzare sui vari neuroni visibili (indicizzati da j) sia la somma in $\xi^{\gamma\mu} \in \Sigma$ che il successivo valore atteso in $\xi^{\star\mu} \in \Sigma$ poiché le rispettive distribuzioni sono indipendenti per ipotesi; tra l'altro, siccome sono anche identicamente distribuite, i termini risultanti sono tutti uguali.

$$\begin{aligned}
G_{(\mathrm{S})}^{\mathrm{RS}} &= \frac{1}{N}\log \mathbb{E}_{\xi^\star} \prod_j \sum_{\{\xi^{\gamma\mu}\in\{\pm1\}\}_{\mu,\gamma}} \\
&\qquad \exp\left(\sum_{\mu,\nu}\hat{m}^{\star\mu\nu}\xi_j^{\star\mu}\sum_\gamma \xi^{\gamma\nu} + \sum_{\mu<\nu}\hat{q}^{=\mu\nu}\sum_\gamma \xi^{\gamma\mu}\xi^{\gamma\nu} + \sum_{\mu,\nu}\hat{q}^{\neq\mu\nu}\sum_{\gamma<\tilde{\gamma}}\xi^{\gamma\mu}\xi^{\tilde{\gamma}\nu}\right) \\
&= \frac{1}{N}\log \prod_j \mathbb{E}_{\xi_j^\star} \sum_{\{\xi^{\gamma\mu}\in\{\pm1\}\}_{\mu,\gamma}} \\
&\qquad \exp\left(\sum_{\mu,\nu}\hat{m}^{\star\mu\nu}\xi_j^{\star\mu}\sum_\gamma \xi^{\gamma\nu} + \sum_{\mu<\nu}\hat{q}^{=\mu\nu}\sum_\gamma \xi^{\gamma\mu}\xi^{\gamma\nu} + \sum_{\mu,\nu}\hat{q}^{\neq\mu\nu}\sum_{\gamma<\tilde{\gamma}}\xi^{\gamma\mu}\xi^{\tilde{\gamma}\nu}\right) \\
&= \log \mathbb{E}_{\xi^\star} \sum_{\{\xi^{\gamma\mu}\}} \exp\left(\sum_{\mu,\nu}\hat{m}^{\star\mu\nu}\xi^{\star\mu}\sum_\gamma \xi^{\gamma\nu} + \sum_{\mu<\nu}\hat{q}^{=\mu\nu}\sum_\gamma \xi^{\gamma\mu}\xi^{\gamma\nu} + \sum_{\mu,\nu}\hat{q}^{\neq\mu\nu}\sum_{\gamma<\tilde{\gamma}}\xi^{\gamma\mu}\xi^{\tilde{\gamma}\nu}\right)
\end{aligned}$$

Nota 3.5 — Nell'ultima riga il valore atteso e la sommatoria sono da intendersi rispettivamente sulla distribuzione di $(\xi^{\star 1},\dots,\xi^{\star P}) \in \{\pm1\}^P$ e su $\{\xi^{\gamma\mu}\in\{\pm1\}\}_{\mu,\gamma}$. D'ora in avanti adotteremo questa convenzione, rimpiazzando la precedente.

A questo punto, per procedere supponiamo di poter scambiare il limite in k con il calcolo del punto estremale e cerchiamo di riscrivere i termini in una forma più agevole per il calcolo del limite. Partiamo proprio dal termine appena calcolato $G^{\mathrm{RS}}_{(\mathrm{S})}$. Per eliminare la dipendenza dalle repliche dobbiamo raccogliere la somma in γ, però l'ultimo termine ha pure una dipendenza da $\tilde{\gamma}$. Risolviamo questa complicanza utilizzando un'equivalenza algebrica e la linearizzazione gaussiana. Siccome anche più avanti avremo la necessità di effettuare un calcolo simile (si osservi l'analogia con l'equazione (3.18)), unifichiamo il risultato nei lemmi seguenti.

Lemma 3.6 — *Dati $a_1, \dots, a_k, b_1, \dots, b_k \in \mathbb{R}$ vale*

$$\sum_{1\le\gamma<\tilde{\gamma}\le k} a_\gamma b_{\tilde{\gamma}} = \frac{1}{4}\left(\sum_{\gamma=1}^{k} a_\gamma + b_\gamma\right)^2 - \frac{1}{4}\left(\sum_{\gamma=1}^{k} a_\gamma\right)^2 - \frac{1}{4}\left(\sum_{\gamma=1}^{k} b_\gamma\right)^2 - \frac{1}{2}\sum_{\gamma=1}^{k} a_\gamma b_\gamma$$

Dimostrazione. Si usano le equivalenze

$$\sum_{\gamma<\tilde{\gamma}} a_\gamma b_{\tilde{\gamma}} = \frac{1}{2}\left(\sum_{\gamma,\tilde{\gamma}} a_\gamma b_{\tilde{\gamma}} - \sum_{\gamma} a_\gamma b_\gamma\right)$$

$$\left(\sum_{\gamma=1}^{k} a_\gamma + b_\gamma\right)^2 - \left(\sum_{\gamma=1}^{k} a_\gamma\right)^2 - \left(\sum_{\gamma=1}^{k} b_\gamma\right)^2 = \sum_{\gamma,\tilde{\gamma}} (a_\gamma + b_\gamma)(a_{\tilde{\gamma}} + b_{\tilde{\gamma}}) - a_\gamma a_{\tilde{\gamma}} - b_\gamma b_{\tilde{\gamma}}$$

□

Lemma 3.7 — *Siano $\lambda \in \mathbb{R}_+$ e $\tau_\star, \tau_1, \dots, \tau_k \in \{\pm 1\}^P$. Considero le notazioni usate per i parametri definiti nelle equazioni (3.11) semplicemente dei numeri reali arbitrari fissati - quindi è ininfluente utilizzare le notazioni dei parametri d'ordine o di quelli coniugati. Allora, intendendo μ, ν indici in $1, \dots, P$ e $\gamma, \tilde{\gamma}$ indici in $1, \dots, k$, vale l'uguaglianza*

$$\exp\left(\lambda \sum_{\mu,\nu} m^{\star\mu\nu}\tau_{\star\mu}\sum_{\gamma}\tau_{\gamma\nu} + \lambda\sum_{\mu<\nu} q^{=\mu\nu}\sum_{\gamma}\tau_{\gamma\mu}\tau_{\gamma\nu} + \lambda\sum_{\mu,\nu} q^{\neq\mu\nu}\sum_{\gamma<\tilde{\gamma}}\tau_{\gamma\mu}\tau_{\tilde{\gamma}\nu}\right)$$
$$= \exp\left(-k\frac{\lambda}{2}\sum_{\mu} q^{\neq\mu\mu}\right)\int\cdots\int\left(\prod_{\mu<\nu}\underline{\mathrm{dz}_\mu}\ \underline{\mathrm{dz}_{\mu\nu}}\right)\prod_{\gamma}\exp\mathcal{L}_\lambda(\tau_{\gamma 1}, \dots, \tau_{\gamma P}; \{\tau_{\star\mu}\}, \{z_\mu, z_{\mu\nu}\})$$

con gli integrali gaussiani di variabile reale espressi tramite la notazione formale $\underline{\mathrm{dz}} := \mathrm{dz}\exp(-z^2/2)/\sqrt{2\pi}$ *e la funzione $\mathcal{L}_\lambda$ (che è in pratica un'hamiltoniana su $\{\pm 1\}^P$) definita*

come segue.

$$\mathcal{L}_\lambda(\sigma_1,\dots,\sigma_P;\{\tau_{\star\mu}\},\{z_\mu,z_{\mu\nu}\}) \coloneqq \sum_{\mu<\nu}\lambda\left(q^{=\mu\nu}-\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}\right)\sigma_\mu\sigma_\nu + \sum_\mu \sigma_\mu\left(\sum_\nu \lambda m^{\star\nu\mu}\tau_{\star\nu}+\sqrt{\lambda q^{\neq\mu\mu}-\lambda\sum_{\nu\neq\mu}\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}}z_\mu + \sum_{\nu>\mu}\sqrt{\lambda\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}}z_{\mu\nu}+\sum_{\nu<\mu}\sqrt{\lambda\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}}z_{\nu\mu}\right) \tag{3.20}$$

Dimostrazione. Innanzitutto usiamo il lemma 3.6 per sviluppare l'ultimo termine dell'argomento dell'esponenziale e manipoliamo le sommatorie per accorpare alcuni termini grazie alle simmetrie, ottenendo il seguente risultato che presenta solamente due tipi di fattori elevati al quadrato:

$$\begin{aligned}\lambda\sum_{\mu,\nu}q^{\neq\mu\nu}\sum_{\gamma<\tilde\gamma}\tau_{\gamma\mu}\tau_{\tilde\gamma\nu} = {}& \frac{\lambda}{2}\sum_\mu\left(q^{\neq\mu\mu}-\sum_{\nu\neq\mu}\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}\right)\left(\sum_\gamma\tau_{\gamma\mu}\right)^2\\ &+\frac{\lambda}{2}\sum_{\mu<\nu}\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}\left(\sum_\gamma\tau_{\gamma\mu}+\tau_{\gamma\nu}\right)^2\\ &-k\frac{\lambda}{2}\sum_\mu q^{\neq\mu\mu}-\lambda\sum_{\mu<\nu}\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}\sum_\gamma\tau_{\gamma\mu}\tau_{\gamma\nu}\end{aligned}$$

A questo punto si linearizza ciascuno dei termini elevati al quadrato con il metodo di Hubbard-Stratonovich introdotto nel lemma 2.4, in particolare usando l'equazione (2.10) relativa a una gaussiana standard ($\sigma=1$). Ad esempio uno dei termini sarà

$$\exp\left(\frac{1}{2}\left(\sqrt{\lambda\frac{q^{\neq12}+q^{\neq21}}{2}}\sum_\gamma(\tau_{\gamma1}+\tau_{\gamma2})\right)^2\right) = \int_{\mathbb{R}}\underline{\mathrm{dz}_{12}}\exp\left(z_{12}\sqrt{\lambda\frac{q^{\neq12}+q^{\neq21}}{2}}\sum_\gamma(\tau_{\gamma1}+\tau_{\gamma2})\right)$$

In totale andranno inseriti $P(P+1)/2$ integrali con le rispettive variabili di integrazione, che chiameremo $z_\mu, z_{\mu\nu}$ come nell'enunciato del lemma.

Moltiplicando i vari integrali (fatti su variabili distinte) si ottiene

$$\int\cdots\int\left(\prod_{\mu<\nu}\underline{\mathrm{d}z_\mu}\ \underline{\mathrm{d}z_{\mu\nu}}\right)\exp\left(\lambda\sum_{\mu,\nu}m^{\star\mu\nu}\tau_{\star\mu}\sum_\gamma\tau_{\gamma\nu}+\lambda\sum_{\mu<\nu}q^{=\mu\nu}\sum_\gamma\tau_{\gamma\mu}\tau_{\gamma\nu}\right.$$
$$-k\frac{\lambda}{2}\sum_\mu q^{\neq\mu\mu}+\sum_\mu z_\mu\sqrt{\lambda q^{\neq\mu\mu}-\lambda\sum_{\nu\neq\mu}\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}}\sum_\gamma\tau_{\gamma\mu}$$
$$\left.+\sum_{\mu<\nu}z_{\mu\nu}\sqrt{\lambda\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}}\sum_\gamma(\tau_{\gamma\mu}+\tau_{\gamma\nu})-\lambda\sum_{\mu<\nu}\frac{q^{\neq\mu\nu}+q^{\neq\nu\mu}}{2}\sum_\gamma\tau_{\gamma\mu}\tau_{\gamma\nu}\right)$$

Riarrangiando i termini e raccogliendo la sommatoria in γ si ottiene l'espressione presentata nell'enunciato. □

Siccome nel seguito useremo il lemma sia con i parametri d'ordine che con i loro coniugati, differenziamo le notazioni per la funzione $\mathcal{L}$, che diventerà $\mathcal{L}^{(\mathrm{O})}$ nel primo caso e $\mathcal{L}^{(\mathrm{C})}$ nel secondo.

Nota 3.8 — Si osservi che gli argomenti delle radici presenti nell'espressione di $\mathcal{L}_\lambda$ non è detto che siano positivi, perché i parametri con i quali lavoriamo possono essere dei numeri reali arbitrari. Per questo è necessario interpretare il termine $\exp\mathcal{L}_\lambda$ presente nell'enunciato del lemma 3.7 come un esponenziale complesso. Di conseguenza, l'integrale rispetto alle variabili gaussiane reali andrà fatto separatamente sulla parte reale e quella immaginaria di questo esponenziale. Ciononostante, siccome il membro di sinistra dell'equazione è certamente a valori reali, anche il risultato dell'integrazione sarà un numero reale. Questo vuol dire che l'integrale della parte immaginaria in realtà è nullo e ci permette di conservare il significato fisico dell'espressione che stiamo trattando.

Riprendendo da dove ci siamo interrotti per introdurre i lemmi, possiamo esprimere il termine entropico sfruttando il lemma 3.7 e, dopo aver usato la linearità dell'integrale, fattorizzare sulle repliche:

$$\begin{aligned}G^{\mathrm{RS}}_{(\mathrm{S})} &= \log\mathbb{E}_{\xi^\star}\sum_{\{\xi^{\gamma\mu}\}}\exp\left(-k\frac{1}{2}\sum_\mu\hat{q}^{\neq\mu\mu}\right)\\ &\qquad\int\cdots\int\left(\prod_{\mu<\nu}\underline{\mathrm{d}z_\mu}\ \underline{\mathrm{d}z_{\mu\nu}}\right)\prod_\gamma\exp\mathcal{L}_1^{(\mathrm{C})}(\{\xi^{\gamma\mu}\}_\mu;\{\xi^{\star\mu}\},\{z_\mu,z_{\mu\nu}\})\\ &= -k\frac{1}{2}\sum_\mu\hat{q}^{\neq\mu\mu}+\log\mathbb{E}_{\xi^\star}\int\cdots\int\left(\prod_{\mu<\nu}\underline{\mathrm{d}z_\mu}\ \underline{\mathrm{d}z_{\mu\nu}}\right)\\ &\qquad\left(\sum_{\{\xi^\mu=\pm1\}_\mu}\exp\mathcal{L}_1^{(\mathrm{C})}(\{\xi^\mu\}_\mu;\{\xi^{\star\mu}\},\{z_\mu,z_{\mu\nu}\})\right)^k\end{aligned}\tag{3.21}$$

Come si può vedere siamo riusciti a rimpiazzare la dipendenza dalle repliche con una semplice dipendenza da k, molto più agevole nel calcolo del limite. Sottolineiamo che per quanto visto nella nota 3.8 l'integrale gaussiano una volta calcolato risulta in un numero reale.
Occupiamoci ora del termine energetico $G_{(\mathrm{E})}^{\mathrm{RS}}$ definito in (3.17). In questo caso l'unica dipendenza dalle repliche si ha nel termine $\mathcal{N}^{\mathrm{RS}}$; un importante vantaggio della nostra trattazione rispetto a quella in [HWH19] risiede nel fatto che grazie alla scelta di scrivere le probabilità in termini di esponenziali, l'espressione della pseudo energia libera presenta una notevole simmetria, perciò possiamo sviluppare il termine energetico con lo stesso metodo seguito per quello entropico. Infatti, riprendendo la definizione di $\mathcal{N}^{\mathrm{RS}}$ data in (3.18), osserviamo che possiamo applicare sempre il lemma 3.7 ottenendo:

$$\begin{aligned}
\mathcal{N}^{\mathrm{RS}} = & \sum_{\substack{\{\tau_{\star\mu}=\pm1\}_\mu \\ \{\tau_{\gamma\mu}=\pm1\}_{\mu,\gamma}}} \exp\left(\frac{\beta^2}{2}\sum_{\mu,\nu}\tau_{\star\mu}\tau_{\star\nu}\,\mathcal{Q}_{\mu\nu}\right)\exp\left(-k\frac{\beta^2}{2}\sum_{\mu}q^{\neq\mu\mu}\right) \\
& \int\cdots\int\left(\prod_{\mu<\nu}\underline{\mathrm{dz}_\mu}\;\underline{\mathrm{dz}_{\mu\nu}}\right)\prod_{\gamma}\exp\mathcal{L}_{\beta^2}^{(\mathrm{O})}(\tau_{\gamma1},\dots,\tau_{\gamma P};\{\tau_{\star\mu}\},\{z_\mu,z_{\mu\nu}\}) \\
= & \exp\left(-k\frac{\beta^2}{2}\sum_{\mu}q^{\neq\mu\mu}\right)\sum_{\{\tau_{\star\mu}=\pm1\}_\mu}\exp\left(\frac{\beta^2}{2}\sum_{\mu,\nu}\tau_{\star\mu}\tau_{\star\nu}\,\mathcal{Q}_{\mu\nu}\right) \\
& \int\cdots\int\left(\prod_{\mu<\nu}\underline{\mathrm{dz}_\mu}\;\underline{\mathrm{dz}_{\mu\nu}}\right)\left(\sum_{\{\tau_\mu=\pm1\}_\mu}\exp\mathcal{L}_{\beta^2}^{(\mathrm{O})}(\{\tau_\mu\}_\mu;\{\tau_{\star\mu}\},\{z_\mu,z_{\mu\nu}\})\right)^k
\end{aligned}$$

Riconosciamo che il termine precedente l'integrale è lo stesso che è presente in $\mathcal{Z}(\xi^\star)$ nella forma data dal teorema 3.2 (cambia solo la variabile muta: in un caso $\tau_{\star\mu}$ mentre nell'altro τ_μ). In effetti l'insieme dei due termini (escludendo il fattore moltiplicativo davanti al denominatore, che tanto si semplifica) forma una media ponderata sulle variabili $\tau_{\star\mu}$ che è la discendente del termine $\mathrm{P}(\sigma|\xi^\star)$ di (3.5); siccome possiamo interpretarla come una media rispetto a una distribuzione di Boltzmann-Gibbs, definiamo

$$\mathcal{M}_\star(\tau_{\star1},\dots,\tau_{\star P}) \coloneqq -\frac{\beta}{2}\sum_{\mu,\nu}\tau_{\star\mu}\tau_{\star\nu}\,\mathcal{Q}_{\mu\nu} \tag{3.22}$$

e indichiamo con $\mathbb{E}_{\mathcal{M}_\star}$ la relativa media termica a temperatura inversa β, definita dall'equazione (1.1).

Così facendo possiamo esprimere il termine energetico come

$$G_{(\mathrm{E})}^{\mathrm{RS}} = -k \log \mathcal{Z}(q^{=\mu\nu}) - k\frac{\beta^2}{2}\sum_{\mu} q^{\neq\mu\mu} + \log \mathbb{E}_{\mathcal{M}_\star} \int\cdots\int \left(\prod_{\mu<\nu} \underline{\mathrm{dz}_\mu}\ \underline{\mathrm{dz}_{\mu\nu}}\right) \left(\sum_{\{\tau_\mu=\pm 1\}_\mu} \exp \mathcal{L}_{\beta^2}^{(\mathrm{O})}(\{\tau_\mu\}_\mu; \{\tau_{\star\mu}\}, \{z_\mu, z_{\mu\nu}\})\right)^k \tag{3.23}$$

Per portare a compimento il calcolo della pressione dell'equazione (3.14) bisogna dividere la pseudo energia libera per k e portare il risultato al limite. La maggior parte dei termini presentano un fattore k moltiplicativo; fanno eccezione i due con gli integrali derivanti dal lemma 3.7: per valutarli faremo ricorso al limite (2.19), che si applica alle equazioni (3.21) e (3.23) perché le integrazioni presenti non sono altro che delle medie rispetto a variabili gaussiane. Dunque, raccogliendo tutti i termini e calcolando il limite in $k \to 0^+$ si ottiene l'espressione finale per la pressione del sistema.

Risultato 3.9 — *Consideriamo il problema inverso della RBM esposto nella sezione 3.1. Siano quindi N il numero di neuroni visibili, P quelli nascosti e $\alpha \coloneqq M/N$ il grado di sollecitazione del sistema, dove M è il numero di esempi forniti. Supponiamo inoltre che la temperatura del processo generativo e di quello di inferenza siano entrambe pari a β^{-1}. In queste ipotesi la distribuzione di equilibrio del problema inverso è data dall'equazione (3.2), i cui parametri sono estratti in modo i.i.d. dalla distribuzione (3.1), a sua volta dipendente da un vettore aleatorio $\xi^\star$ con media nulla e matrice di correlazione $\mathcal{Q}$.*
Utilizzando il trucco delle repliche e assumendo l'ipotesi di simmetria di replica esposta in (3.11), la media sul disordine della pressione intensiva del modello (nel limite termodinamico) si può ottenere come segue.

$$\begin{aligned} \mathrm{a}_{P,\beta,\alpha,\mathcal{Q}} = \operatorname{extr} \frac{1}{2}\sum_{\mu,\nu} q^{\neq\mu\nu}\hat{q}^{\neq\mu\nu} - \sum_{\mu,\nu} m^{\star\mu\nu}\hat{m}^{\star\mu\nu} - \sum_{\mu<\nu} q^{=\mu\nu}\hat{q}^{=\mu\nu} \\ - \frac{1}{2}\sum_{\mu}\hat{q}^{\neq\mu\mu} + \mathbb{E}_{\xi^\star}\,\mathbb{E}_{\{z_\mu,z_{\mu\nu}\}} \log \mathcal{Z}(\mathcal{L}^{(C)}) \\ - \frac{\alpha\beta^2}{2}\sum_{\mu} q^{\neq\mu\mu} + \alpha\,\mathbb{E}_{\mathcal{M}_\star}\,\mathbb{E}_{\{z_\mu,z_{\mu\nu}\}} \log \mathcal{Z}(\mathcal{L}^{(O)}) - \alpha \log \mathcal{Z}(q^{=\mu\nu}) \end{aligned} \tag{3.24}$$

dove

- *gli indici μ e ν si intendono variabili tra 1 e P, tenendo eventualmente conto di ulteriori vincoli esplicitati di volta in volta.*

- extr *sta a indicare che l'intera espressione dev'essere valutata in un suo punto estremale (dove si annullano le derivate) rispetto ai* $P(5P-1)$ *parametri reali:* $\{m^{\star\mu\nu}\}_{\mu,\nu}$, $\{q^{=\mu\nu}\}_{\mu<\nu}$, $\{q^{\neq\mu\nu}\}_{\mu,\nu}$ *(p. d'ordine) e* $\{\hat{m}^{\star\mu\nu}\}_{\mu,\nu}$, $\{\hat{q}^{=\mu\nu}\}_{\mu<\nu}$, $\{\hat{q}^{\neq\mu\nu}\}_{\mu,\nu}$ *(p. d'ordine ausiliari o coniugati). Per quanto riguarda la scelta del punto nel caso in cui fossero molteplici si veda la nota 3.4.*

- $\mathbb{E}_{\xi^\star}$ *indica la media rispetto al disordine (piantato, vedi la sezione 3.1): un vettore* P*-dimensionale a valori in* $\{\pm 1\}$ *di media nulla e matrice di correlazione* $\mathcal{Q}$*.*

- $\mathbb{E}_{\mathcal{M}_\star}$ *rappresenta la media termica (a temperatura inversa* β*) rispetto all'hamiltoniana* $\mathcal{M}_\star$ *definita in (3.22) e dipendente da* $\mathcal{Q}$*.*

- $\mathbb{E}_{\{z_\mu, z_{\mu\nu}\}}$ *è una media sulle variabili aleatorie* $\{z_\mu, z_{\mu\nu}\}_{\mu<\nu}$ *indipendenti e identicamente distribuite come delle gaussiane standard.*

- $\mathcal{Z}(q^{=\mu\nu})$ *è la funzione di partizione definita dall'equazione (3.19).*

- $\mathcal{Z}(\mathcal{L}^{(C)})$ *e* $\mathcal{Z}(\mathcal{L}^{(O)})$ *sono le funzioni di partizione così definite*

$$\mathcal{Z}(\mathcal{L}^{(C)}) \coloneqq \sum_{\{\xi^\mu=\pm1\}_\mu} \exp \mathcal{L}_1^{(C)}(\{\xi^\mu\}_\mu; \{\xi^{\star\mu}\}, \{z_\mu, z_{\mu\nu}\})$$
$$\mathcal{Z}(\mathcal{L}^{(O)}) \coloneqq \sum_{\{\tau_\mu=\pm1\}_\mu} \exp \mathcal{L}_{\beta^2}^{(O)}(\{\tau_\mu\}_\mu; \{\tau_{\star\mu}\}, \{z_\mu, z_{\mu\nu}\})$$

- $\mathcal{L}_1^{(C)}$ *e* $\mathcal{L}_{\beta^2}^{(O)}$ *sono due versioni dell'hamiltoniana definita in (3.20), nel primo caso usando i parametri coniugati e nel secondo quelli d'ordine.*

- *i logaritmi delle funzioni di partizione* $\mathcal{Z}(\mathcal{L}^{(C)})$ *e* $\mathcal{Z}(\mathcal{L}^{(O)})$ *sono da intendersi come il ramo principale del logaritmo naturale complesso poiché le hamiltoniane* $\mathcal{L}_1^{(C)}$ *e* $\mathcal{L}_{\beta^2}^{(O)}$ *sono a valori complessi. Tuttavia, l'integrale della loro parte immaginaria rispetto alle variabili gaussiane si annulla per quanto detto nella nota 3.8, quindi la funzione da estremizzare è a valori reali.*

Spiegazione. La spiegazione è stata data nel corso del capitolo, qui ricapitoliamo i passaggi principali. Dopo aver impostato il trucco delle repliche siamo partiti dall'equazione (3.4), semplificando la funzione di partizione Ω di qualche fattore 2^N, come spiegato dopo il teorema 3.2. Grazie all'intuizione di scrivere la distribuzione del modello tramite esponenziali, evitando così i coseni iperbolici, siamo riusciti a esprimere la funzione di partizione replicata $\overline{\Omega_{P,\beta}^k}$ in una forma agevole per procedere con il calcolo. Dopodiché

abbiamo introdotto delle delte di Dirac in rappresentazione di Fourier, applicato il metodo di Laplace sugli integrali risultanti e imposto l'ipotesi di simmetria di replica. Infine, dopo aver espresso la pseudo energia libera risultante ${G_{N,k}}^{\mathrm{RS}}$ in una forma più pratica per mezzo del lemma 3.7, abbiamo calcolato il limite rimanente come in (2.19), ottenendo l'enunciato di questo risultato. △

Commentiamo brevemente il risultato. Innanzitutto notiamo che le dipendenze della pressione da α e da β risiedono solamente negli ultimi tre termini, la prima delle quali è semplicemente lineare. In più, grazie allo stratagemma dell'espressione della probabilità inversa tramite esponenziali, l'espressione che abbiamo ottenuto presenta una notevole simmetria (fa eccezione il termine con $\mathcal{Z}(q^{=\mu\nu})$). In particolare si confronti con l'espressione ricavata in [HWH19] per il caso $P = 2$; nonostante siano espresse in forma diversa le due equazioni sono equivalenti.

Affinché questo risultato sia utilizzabile numericamente è necessario individuare il punto estremale nel quale l'espressione (3.24) dev'essere valutata. Per trovarlo cerchiamo i punti nei quali si annullano tutte le derivate rispetto ai parametri.
Siccome sono presenti svariate funzioni di partizione la prima osservazione utile riguarda le loro derivate: per una generica funzione di partizione $\mathcal{Z}_\lambda = \mathrm{Tr}_\sigma e^{c\,\mathcal{H}(\sigma;\lambda)}$ vale $\partial_\lambda \log \mathcal{Z}_\lambda = c\langle \partial_\lambda \mathcal{H}\rangle$, dove la media termica si intende rispetto alla stessa distribuzione esponenziale. Pertanto, innanzitutto calcoliamo le derivate della funzione $\mathcal{L}_\lambda$ nella sua forma generale rispetto ai parametri di nostro interesse. A seguire anticipiamo un lemma che ci sarà utile più avanti per svolgere i conti.

Lemma 3.10 — *Considerando l'hamiltoniana $\mathcal{L}_\lambda$ definita in (3.20), dipendente dai parametri $\{m^{\star\mu\nu}, q^{=\mu\nu}, q^{\neq\mu\nu}\}$. Valgono le seguenti identità:*

$$\begin{aligned}
\partial_{m^{\star\iota\kappa}} \mathcal{L}_\lambda &= \lambda \sigma_\kappa \tau_{\star\iota} \\
\partial_{q^{=\iota\kappa}} \mathcal{L}_\lambda &= \lambda \sigma_\iota \sigma_\kappa \\
\partial_{q^{\neq\iota\kappa}} \mathcal{L}_\lambda &= (2\,\delta_{\iota,\kappa} - 1)\frac{\lambda}{4}\left(\sigma_\iota A_\iota^{-1} z_\iota + \sigma_\kappa A_\kappa^{-1} z_\kappa\right) \\
&\qquad - (1-\delta_{\iota,\kappa})\frac{\lambda}{2}\sigma_\iota\sigma_\kappa + (1-\delta_{\iota,\kappa})\frac{\lambda}{4}(\sigma_\iota + \sigma_\kappa) B_{\iota\kappa}^{-1} z_{\iota\kappa}
\end{aligned}$$

dove si intende $z_{\iota\kappa}$ o $z_{\kappa\iota}$ a seconda dell'ordinamento degli indici (le variabili sono definite

solo per $\iota < \kappa$) e con le seguenti notazioni:

$$\begin{aligned} A_\iota &\coloneqq \sqrt{\lambda q^{\neq \iota\iota} - \frac{\lambda}{2}\sum_{\nu\neq\iota}(q^{\neq \iota\nu} + q^{\neq \nu\iota})} \quad \forall \iota \\ B_{\iota\kappa} &\coloneqq \sqrt{\frac{\lambda}{2}(q^{\neq \iota\kappa} + q^{\neq \kappa\iota})} \quad \forall \iota \neq \kappa \end{aligned} \tag{3.25}$$

Dimostrazione. Le prime due equazioni seguono da un semplice conto. Per quanto riguarda la derivata in $q^{\neq \iota\kappa}$ analizziamo il contributo di ciascuno dei cinque termini presenti. Il primo termine restituisce $-\lambda/2\sigma_\iota\sigma_\kappa(1-\delta_{\iota,\kappa})$ perché è presente solo quando $\iota \neq \kappa$. Il termine con $m^{\star\mu\nu}$ si annulla sempre. Introduciamo ora le notazioni definite in (3.25). Nel terzo termine bisogna distinguere i casi in cui gli indici sono uguali o distinti: se sono uguali si ottiene $\frac{\lambda}{2}\sigma_\iota A_\iota^{-1} z_\iota$, altrimenti bisogna considerare sia il caso in cui l'indice della sommatoria μ è uguale a ι, sia il caso in cui è κ; nel primo caso si ottiene $-\sigma_\iota \frac{\lambda}{4} A_\iota^{-1} z_\iota$, l'altro è analogo. L'apporto degli ultimi due termini è diverso da zero solamente quando $\iota \neq \kappa$, nel qual caso ciascuno dei due pezzi presenta una sola coppia di indici della sommatoria con un valore non nullo; per via della simmetria i due contributi si distinguono solamente per l'indice si σ. Ad esempio se $\iota < \kappa$ il primo di questi sarà pari a $\sigma_\iota \frac{\lambda}{4} B_{\iota\kappa}^{-1} z_{\iota\kappa}$. In ogni caso, siccome vanno sommati, il loro apporto congiunto sarà simmetrico e non dipenderà dall'ordinamento. L'unica attenzione da prestare riguarda la variabile $z_{\iota\kappa}$ che è definita solo per $\iota < \kappa$, perciò è necessario distinguere i due casi, almeno concettualmente. Raccogliendo e riordinando i termini si ottiene l'asserto. □

Lemma 3.11 — *Data una funzione limitata (in modulo) $f\colon \mathbb{R} \to \mathbb{C}$, vale la seguente identità per integrali gaussiani di variabile reale:*

$$\int f(z) z \,\underline{\mathrm{dz}} = \int f'(z)\,\underline{\mathrm{dz}}$$

dove l'apice indica la derivata e abbiamo usato la notazione formale $\underline{\mathrm{dz}}$ introdotta nel lemma 3.7.

Dimostrazione. Il risultato segue direttamente dall'integrazione per parti con i fattori $f(z)$ e $(-e^{-z^2/2})'$ poiché all'infinito il termine esponenziale porta tutto a zero. □

Applicando questi risultati alla pressione (3.24) si arriva ad un sistema di condizioni stazionarie, raccolte nel prossimo teorema, che devono essere soddisfatte dai parametri affinché individuino un punto estremale.

Teorema 3.12 — *Facendo riferimento al risultato 3.9 e con le stesse notazioni lì introdotte, il punto estremale dell'espressione (3.24) per la pressione intensiva del modello, i cui parametri sono*

$$\{m^{\star\mu\nu}, \hat{m}^{\star\mu\nu}\}_{\mu,\nu}, \{q^{=\mu\nu}, \hat{q}^{=\mu\nu}\}_{\mu<\nu}, \{q^{\neq\mu\nu}, \hat{q}^{\neq\mu\nu}\}_{\mu,\nu}$$

deve soddisfare le seguenti equazioni di auto-consistenza:

$$\begin{aligned}
m^{\star\mu\nu} &= \mathbb{E}_{\xi^\star}\, \mathbb{E}_{\{z\}} \left[\xi^{\star\mu} \langle \xi_\nu \rangle_{\mathcal{L}^{(O)}}\right] \\
q^{=\mu\nu} &= \mathbb{E}_{\xi^\star}\, \mathbb{E}_{\{z\}} \left[\langle \xi_\mu \xi_\nu \rangle_{\mathcal{L}^{(O)}}\right] \\
q^{\neq\mu\nu} &= \mathbb{E}_{\xi^\star}\, \mathbb{E}_{\{z\}} \left[\langle \xi_\mu \rangle_{\mathcal{L}^{(O)}} \langle \xi_\nu \rangle_{\mathcal{L}^{(O)}}\right] \\
\hat{m}^{\star\mu\nu} &= \alpha\beta^2\, \mathbb{E}_{\mathcal{M}_\star}\, \mathbb{E}_{\{z\}} \left[\tau_{\star\mu} \langle \tau_\nu \rangle_{\mathcal{L}^{(O)}}\right] \\
\hat{q}^{=\mu\nu} &= \alpha\beta^2 \left(\mathbb{E}_{\mathcal{M}_\star}\, \mathbb{E}_{\{z\}} \left[\langle \tau_\mu \tau_\nu \rangle_{\mathcal{L}^{(O)}}\right] - \langle \tau_\mu \tau_\nu \rangle_{q^{=\mu\nu}}\right) \\
\hat{q}^{\neq\mu\nu} &= \alpha\beta^2\, \mathbb{E}_{\mathcal{M}_\star}\, \mathbb{E}_{\{z\}} \left[\langle \tau_\mu \rangle_{\mathcal{L}^{(O)}} \langle \tau_\nu \rangle_{\mathcal{L}^{(O)}}\right]
\end{aligned} \tag{3.26}$$

dove $\mathbb{E}_{\{z\}}$ è un'abbreviazione per il valore atteso sulle variabili gaussiane $\{z_\mu, z_{\mu\nu}\}_{\mu<\nu}$, $\tau_\star$ indica l'argomento di $\mathcal{M}_\star$ e le medie termiche, indicate con $\langle\bullet\rangle$, hanno in pedice la notazione che indica la funzione di partizione alla quale si riferiscono: se $\mathcal{Z}(\star) = \mathrm{Tr}_\sigma e^{\mathcal{H}(\sigma)}$ allora $\langle f \rangle_\star \coloneqq \mathrm{Tr}_\sigma f(\sigma) e^{\mathcal{H}(\sigma)}$.

Anche in questo caso rimarchiamo la simmetria delle equazioni relative ai due tipi di parametri. Si può verificare che le equazioni sono equivalenti a quelle presentate in [HWH19]. Da queste equazioni possiamo anche inferire il significato dei parametri d'ordine: magnetizzazione, correlazione e sovrapponibilità. Di seguito forniamo la dimostrazione del teorema.

Dimostrazione. Per ottenere queste espressioni bisogna calcolare le derivate di (3.24) (ignorando extr) e porle uguali a zero.
Cominciamo dal calcolo della derivata rispetto al parametro $\hat{m}^{\star\iota\kappa}$ - usiamo gli indici ι e κ per evitare confusioni. Scambiando come al solito la derivata con le medie, in prima battuta si ottiene $-m^{\star\iota\kappa} + \mathbb{E}_{\xi^\star}\, \mathbb{E}_{\{z\}} \left[\langle \partial_{\hat{m}^{\star\iota\kappa}} \mathcal{L}_1^{(\mathrm{C})} \rangle_{\mathcal{L}^{(\mathrm{C})}}\right] = 0$. Questo perché nell'espressione (3.24) i termini contenenti i parametri d'ordine e quelli contenenti i parametri d'ordine ausiliari sono ben distinti (esclusi i primi tre termini dove compaiono assieme, nei quali però la dipendenza è palese). Adesso è sufficiente applicare il lemma 3.10 per sapere che $\partial_{\hat{m}^{\star\iota\kappa}} \mathcal{L}_1^{(\mathrm{C})} = \xi^\kappa \xi^{\star\iota}$. Mettendo insieme i pezzi si ottiene il risultato. Il caso della derivata rispetto a $m^{\star\iota\kappa}$ è analogo, in più compare un termine β^2 derivando $\mathcal{L}_{\beta^2}^{(\mathrm{O})}$.
Analoghi calcoli portano al risultato per le derivate in $\hat{q}^{=\iota\kappa}$ e $q^{=\iota\kappa}$; in quest'ultimo caso bisogna derivare anche il termine ulteriore $\log \mathcal{Z}(q^{=\mu\nu})$ per il quale è sufficiente ricordare

la proprietà delle funzioni di partizione richiamata sopra al lemma 3.10.
Infine calcoliamo la derivata rispetto a $q^{\neq \iota\kappa}$, quella rispetto al parametro ausiliario è analoga. Saltando qualche passaggio simile a quelli già visti, concentriamoci sul termine $\alpha\,\mathbb{E}_{\mathcal{M}_\star}\,\mathbb{E}_{\{z\}}\langle\partial_{q^{\neq\iota\kappa}}\,\mathcal{L}^{(\mathrm{O})}_{\beta^2}\rangle_{\mathcal{L}^{(\mathrm{O})}}$. Applicando il solito lemma e tralasciando per il momento il moltiplicatore α e la media $\mathbb{E}_{\mathcal{M}_\star}$ otteniamo

$$(2\,\delta_{\iota,\kappa}-1)\frac{\beta^2}{4}\left(A_\iota^{-1}\,\mathbb{E}_{\{z\}}\left[\langle\tau_\iota\rangle_{\mathcal{L}^{(\mathrm{O})}}z_\iota\right]+A_\kappa^{-1}\,\mathbb{E}_{\{z\}}\left[\langle\tau_\kappa\rangle_{\mathcal{L}^{(\mathrm{O})}}z_\kappa\right]\right)$$
$$-(1-\delta_{\iota,\kappa})\frac{\beta^2}{2}\langle\tau_\iota\tau_\kappa\rangle_{\mathcal{L}^{(\mathrm{O})}}+(1-\delta_{\iota,\kappa})\frac{\beta^2}{4}B_{\iota\kappa}^{-1}\,\mathbb{E}_{\{z\}}\left[\langle\tau_\iota+\tau_\kappa\rangle_{\mathcal{L}^{(\mathrm{O})}}z_{\iota\kappa}\right]$$

dove naturalmente nei termini A e B bisogna considerare $\lambda=\beta^2$. Ora usiamo il lemma 3.11 per semplificare i termini con gli integrali gaussiani. L'integrazione sarebbe multidimensionale ma grazie all'indipendenza delle variabili possiamo svolgerla a pezzi e applicare il lemma solamente rispetto all'indice che ci interessa (che è uno solo ogni volta). Ad esempio per il primo termine si ottiene $\mathbb{E}_{\{z\}}\left[\partial_{z_\iota}\langle\tau_\iota\rangle_{\mathcal{L}^{(\mathrm{O})}}\right]$. Prima di scrivere il risultato richiamiamo una proprietà elementare della media termica (la verifica è un semplice calcolo, per una dimostrazione si veda ad esempio la prova del lemma 3.13): se f è un'osservabile, la media termica è calcolata rispetto ad una distribuzione nella forma $\mathcal{Z}^{-1}\,e^{\mathcal{H}}$ e x è un parametro (solo) di $\mathcal{H}$, vale l'identità: $\partial_x\langle f\rangle=\langle f\,\partial_x\,\mathcal{H}\rangle-\langle f\rangle\langle\partial_x\,\mathcal{H}\rangle$. Da questo segue che è necessario calcolare le derivate di $\mathcal{L}^{(\mathrm{O})}_{\beta^2}$ rispetto alle variabili gaussiane. Dalla definizione di $\mathcal{L}$ data nell'equazione (3.20) si trova che $\partial_{z_\iota}\,\mathcal{L}^{(\mathrm{O})}_{\beta^2}=\tau_\iota A_\iota$ e $\partial_{z_{\iota\kappa}}\,\mathcal{L}^{(\mathrm{O})}_{\beta^2}=\tau_\iota B_{\iota\kappa}+\tau_\kappa B_{\iota\kappa}$ (qui si ha per forza $\iota<\kappa$). Ricomponendo i pezzi, abbiamo provato che

$$\mathbb{E}_{\{z\}}\langle\partial_{q^{\neq\iota\kappa}}\,\mathcal{L}^{(\mathrm{O})}_{\beta^2}\rangle_{\mathcal{L}^{(\mathrm{O})}}=-(1-\delta_{\iota,\kappa})\frac{\beta^2}{2}\langle\tau_\iota\tau_\kappa\rangle_{\mathcal{L}^{(\mathrm{O})}}$$
$$+(2\,\delta_{\iota,\kappa}-1)\frac{\beta^2}{4}\left(A_\iota^{-1}\,\mathbb{E}_{\{z\}}\left[A_\iota-A_\iota\langle\tau_\iota\rangle^2_{\mathcal{L}^{(\mathrm{O})}}\right]+A_\kappa^{-1}\,\mathbb{E}_{\{z\}}\left[A_\kappa-A_\kappa\langle\tau_\kappa\rangle^2_{\mathcal{L}^{(\mathrm{O})}}\right]\right)$$
$$+(1-\delta_{\iota,\kappa})\frac{\beta^2}{4}B_{\iota\kappa}^{-1}\,\mathbb{E}_{\{z\}}\left[B_{\iota\kappa}\langle(\tau_\iota+\tau_\kappa)^2\rangle_{\mathcal{L}^{(\mathrm{O})}}-B_{\iota\kappa}\langle\tau_\iota+\tau_\kappa\rangle^2_{\mathcal{L}^{(\mathrm{O})}}\right]$$

I termini A e B si semplificano (la vita ci sorride! sì, l'ho scritto sul serio: se leggi questa riga per favore trova il modo di farmelo sapere, sarò molto contento di scoprire che qualcuno ha davvero controllato i conti di questa tesi). Ricordando che $\langle\tau_\iota+\tau_\kappa\rangle^2_{\mathcal{L}^{(\mathrm{O})}}=\left(\langle\tau_\iota\rangle_{\mathcal{L}^{(\mathrm{O})}}+\langle\tau_\kappa\rangle_{\mathcal{L}^{(\mathrm{O})}}\right)^2$ alcuni termini si semplificano e ricaviamo una forma più semplice:

$$\mathbb{E}_{\{z\}}\langle\partial_{q^{\neq\iota\kappa}}\,\mathcal{L}^{(\mathrm{O})}_{\beta^2}\rangle_{\mathcal{L}^{(\mathrm{O})}}=\frac{\beta^2}{2}\left(\delta_{\iota,\kappa}-\mathbb{E}_{\{z\}}\left[\langle\tau_\iota\rangle_{\mathcal{L}^{(\mathrm{O})}}\langle\tau_\kappa\rangle_{\mathcal{L}^{(\mathrm{O})}}\right]\right)$$

A questo risultato dobbiamo solo anteporre $\alpha\,\mathbb{E}_{\mathcal{M}_\star}$ e sommarlo ai contributi ottenuti derivando gli altri termini dell'espressione (3.24) per ottenere la relativa equazione di auto-consistenza. □

3.3 Analisi delle transizioni di fase

Come abbiamo descritto alla fine della sezione 3.1, gli indicatori tramite i quali valutiamo l'apprendimento sono le sovrapponibilità tra i pattern originari e quelli inferiti: quando sono non-nulle indicano che il sistema ha imparato a ricostruire quei determinati pesi sinaptici Sappiamo dal ragionamento costruito attorno all'equazione (2.31) che per come abbiamo definito i parametri in (3.7) questi indicatori si studiano equivalentemente valutando i parametri di magnetizzazione $m^{\star\mu\nu}$, sempre sottintendendo di lavorare con $N >> 1$. I valori di questi parametri sono determinati dalle equazioni di auto-consistenza (3.26). Tra l'altro, queste equazioni confermano l'intuizione sul significato da attribuire alle variabili poiché sono espresse proprio come sovrapponibilità medie tra i pattern, secondo una specifica distribuzione che terrà conto del modello nel suo complesso; il ragionamento è simile a quello visto per il modello di Curie-Weiss nell'esempio 1.11.
In questa sezione, guidati dai risultati già noti nel caso elementare di una sola unità nascosta ($P = 1$), innanzitutto investigheremo la presenza di una transizione di fase nel comportamento del sistema al variare del grado di sollecitazione α, cioè l'esistenza di una soglia critica da superare per rendere possibile l'apprendimento. Dopo averne ricavato un'espressione generale discuteremo la sua variabilità in dipendenza di alcuni parametri cruciali: il rumore presente nel sistema, la correlazione dei dati originari e il numero di concetti da inferire (che nel nostro caso coincide con il numero di neuroni nascosti del sistema). Questo ci permetterà di fornire un quadro piuttosto generale sulle condizioni ottimali di allenamento di una RBM dicotomica. Affiancheremo la presentazione di questi risultati originali con vari grafici, ottenuti tramite delle simulazioni numeriche, che corroborano le nostre previsioni teoriche.

3.3.1 Calcolo della soglia critica

Osservando le equazioni di auto-consistenza ricaviamo che per $\alpha \approx 0$ le soluzioni saranno molto vicine alla condizione nulla, poiché i parametri coniugati sono minori di $\alpha\beta^2$ in valore assoluto, quindi anche la media termica rispetto a $\mathcal{L}^{(\mathrm{C})}$ sarà pressoché uniforme. Come vedremo, anche numericamente si evince che per carichi troppo bassi il modello non riesce ad accumulare informazione sufficiente e le magnetizzazioni sono nulle. Dunque, per stabilire una stima dall'alto per la transizione di fase studiamo la pressione cercando di capire quando la soluzione nulla non sia più il giusto punto estremale da prendere in considerazione.
L'espressione della pressione tramite un principio variazionale discende dall'utilizzo del

metodo di Laplace (lemma 2.5) all'interno del contesto del replica trick. Entrambi i metodi forniscono risultati non rigorosi, perciò basandosi solamente sull'analisi teorica che abbiamo svolto non è possibile stabilire con certezza la natura di questi punti estremali (si veda anche la nota 3.4). Per risolvere questo inconveniente, una possibilità è quella di analizzare l'hessiana della pressione e studiarne i cambi di segno degli autovalori augurandosi di riuscire a ricavare informazioni sufficienti per individuare correttamente la soglia critica. Invece, la strada che noi percorreremo, estendendo il ragionamento fatto anche in [HWH19], consiste nel ricavare in quali condizioni il punto desiderato (zero nel nostro caso) è un punto fisso delle equazioni di auto-consistenza (3.26). Questo coincide con la condizione per avere una soluzione stabile risolvendo le equazioni con un metodo iterativo, perciò è verosimile che porti al risultato cercato.
Nel caso di un'equazione vettoriale $x = f(x)$, la condizione per la stabilità di un punto fisso è che il differenziale di f nel punto abbia tutti gli autovalori minori strettamente di uno; viceversa, affinché sia instabile è sufficiente che uno di questi sia maggiore strettamente di uno. Nel nostro caso la funzione di punto fisso a valori vettoriali sarà data dai membri di destra delle condizioni stazionarie elencate in (3.26). Per entrambi i metodi è cruciale il calcolo delle derivate nel punto zero dei membri di destra delle condizioni stazionarie.
Precisiamo che a priori, nonostante il nostro interesse principale siano le magnetizzazioni di Mattis, le equazioni (3.26) si presentano in una forma non separabile: le variabili sono tutte collegate e non è possibile analizzare ciascuna equazione separatamente. L'unica dipendenza dai parametri d'ordine (coniugati e non) è all'interno delle medie termiche, perciò cominciamo con l'espansione di queste in serie di potenze al prim'ordine. Risolviamo innanzitutto il caso generale, al second'ordine perché nella nostra espressione sono presenti delle radici.

Lemma 3.13 — *Consideriamo un sistema su uno spazio χ^P e intendiamo gli indici variabili in $1, \dots, P$. Definiamo la seguente hamiltoniana dipendente dai parametri $\{J_{\mu\nu}\}$ e $\{h_\mu\}$*

$$\mathcal{L}(\tau) = -\sum_{\mu<\nu} J_{\mu\nu}\tau_\mu\tau_\nu - \sum_\mu h_\mu\tau_\mu$$

Nel limite di parametri molto vicini a zero e considerando la media termica a temperatura inversa β valgono le seguenti espansioni al second'ordine in serie di potenze:

$$\langle\tau_\lambda\rangle \sim \beta h_\lambda + \beta^2 \sum_{\iota\neq\lambda} h_\iota J_{\iota\lambda} \tag{3.27}$$

$$\langle\tau_\kappa\tau_\lambda\rangle \sim \beta J_{\kappa\lambda} + \beta^2 h_\kappa h_\lambda + \beta^2 \sum_{\iota\neq\kappa,\lambda} J_{\iota\kappa}J_{\iota\lambda} \tag{3.28}$$

con l'accortezza di interpretare gli indici dei parametri J tenendo il più piccolo a sinistra, scambiandoli se necessario: i parametri di interazione sono definiti solo per $\mu < \nu$.

Dimostrazione. Consideriamo una generico osservabile f che non dipenda dai parametri dell'hamiltoniana e calcoliamo le derivate della sua media termica rispetto a $x, y \in \{J_{\mu\nu}, h_\mu\}$. Indichiamo come di consueto con $\mathcal{Z}$ la funzione di partizione dell'hamiltoniana in questione e con Tr_τ la somma sulle configurazioni.
Come abbiamo già accennato nella dimostrazione del teorema 1.7, una proprietà importante della distribuzione di Boltzmann-Gibbs è il suo comportamento rispetto alla derivazione. Infatti è facile osservare che

$$\begin{aligned}\partial_x\langle f\rangle &= \mathrm{Tr}_\tau f(\tau)\left(\frac{-\beta e^{-\beta\,\mathcal{L}(\tau)}\partial_x\,\mathcal{L}(\tau)}{\mathcal{Z}} - \frac{e^{-\beta\,\mathcal{L}(\tau)}\partial_x \mathrm{Tr}_\rho e^{-\beta\,\mathcal{L}(\rho)}}{\mathcal{Z}^2}\right)\\ &= -\beta\left(\langle f\partial_x\,\mathcal{L}\rangle - \langle f\rangle\langle\partial_x\,\mathcal{L}\rangle\right)\end{aligned}$$

Sfruttando questo risultato si prova anche l'espressione per le derivate seconde:

$$\begin{aligned}\partial_{y,x}\langle f\rangle = \beta^2\big(&\langle f\partial_x\,\mathcal{L}\,\partial_y\,\mathcal{L}\rangle - \langle f\partial_x\,\mathcal{L}\rangle\langle\partial_y\,\mathcal{L}\rangle - \langle f\partial_y\,\mathcal{L}\rangle\langle\partial_x\,\mathcal{L}\rangle\\ &+2\langle f\rangle\langle\partial_x\,\mathcal{L}\rangle\langle\partial_y\,\mathcal{L}\rangle - \langle f\rangle\langle\partial_x\,\mathcal{L}\,\partial_y\,\mathcal{L}\rangle\big)\end{aligned}$$

Osserviamo però che quando l'hamiltoniana viene valutata con parametri nulli è anch'essa identicamente nulla perciò la relativa media termica è uniforme. In questo caso gli spin sono tutti indipendenti tra loro e hanno media nulla, da cui segue anche che la media delle derivate di $\mathcal{L}$ è nulla e per le scelte di f a noi utili anche la media di quest'ultima si annulla. Dunque, calcolando le derivate appena ottenute in zero (a parametri nulli) si ottiene:

$$\begin{aligned}\partial_x\Big|_0\langle f\rangle &= -\beta\langle f\,\partial_x\Big|_0\,\mathcal{L}\rangle_{u.}\\ \partial_{y,x}\Big|_0\langle f\rangle &= \beta^2\langle f\,\partial_x\Big|_0\,\mathcal{L}\,\partial_y\Big|_0\,\mathcal{L}\rangle_{u.}\end{aligned}$$

dove abbiamo indicato con $\langle\bullet\rangle_{u.}$ la media uniforme. A questo punto si può procedere al calcolo dello sviluppo nei casi di nostro interesse. Per $f(\tau) = \tau_\lambda$ si ottiene

$$\begin{aligned}\langle f\rangle \sim 0 &+ \beta\sum_\mu\langle\tau_\lambda\tau_\mu\rangle_{u.}h_\mu + \beta\sum_{\mu<\nu}\langle\tau_\lambda\tau_\mu\tau_\nu\rangle_{u.}J_{\mu\nu}\\ &+\frac{\beta^2}{2}\left(\sum_{\kappa,\mu}\langle\tau_\lambda\tau_\kappa\tau_\mu\rangle_{u.}h_\kappa h_\mu + \sum_{\iota<\kappa;\mu<\nu}\langle\tau_\lambda\tau_\iota\tau_\kappa\tau_\mu\tau_\nu\rangle_{u.}J_{\iota\kappa}J_{\mu\nu}\right.\\ &\qquad\left.+2\sum_{\kappa;\mu<\nu}\langle\tau_\lambda\tau_\kappa\tau_\mu\tau_\nu\rangle_{u.}h_\kappa J_{\mu\nu}\right)\end{aligned}$$

Un'attenta analisi dell'espressione qui sopra, alla luce del fatto che abbiamo a che fare con variabili di spin il cui quadrato fa sempre uno, prova l'asserto. □

Tornando alla linearizzazione delle equazioni di auto-consistenza ribadiamo che i valori attesi sui pattern originari e sulle variabili gaussiane si possono scambiare con le derivate rispetto ai parametri (sostanzialmente grazie al teorema della convergenza dominata). Quindi per ottenere la matrice jacobiana che cerchiamo è sufficiente sostituire le medie termiche con le loro linearizzazioni secondo il lemma 3.13.
Cominciamo con la prima condizione stazionaria, quella relativa a $m^{\star\mu\nu}$, esplicitando il termine $\langle \xi^\nu \rangle_{\mathcal{L}_1^{(\mathrm{C})}}$; nel farlo ricordiamo che la temperatura della media di Boltzmann-Gibbs è fissata a 1 e che l'hamiltoniana di riferimento è quella definita nell'equazione (3.20), con i parametri ausiliari e $\lambda = 1$. Per maggiore chiarezza, evitiamo di indicare i termini con le variabili gaussiane che tanto si eliderebbero una volta applicato il relativo valore atteso (siccome sono centrate e indipendenti).

$$\begin{aligned}\langle \xi^\nu \rangle_{\mathcal{L}_1^{(\mathrm{C})}} &\sim \sum_\iota \hat{m}^{\star\iota\nu}\xi^{\star\iota} + \sum_{\kappa\neq\nu}\left(\sum_\iota \hat{m}^{\star\iota\kappa}\xi^{\star\iota}\right)\left(\hat{q}^{=\kappa\nu} - \frac{\hat{q}^{\neq\kappa\nu}+\hat{q}^{\neq\nu\kappa}}{2}\right)\\ &= \sum_\iota \xi^{\star\iota}\left(\hat{m}^{\star\iota\nu} + \sum_{\kappa\neq\nu}\hat{m}^{\star\iota\kappa}\left(\hat{q}^{=\kappa\nu} - \frac{\hat{q}^{\neq\kappa\nu}+\hat{q}^{\neq\nu\kappa}}{2}\right)\right)\end{aligned}$$

Prima di sostituire questa espressione nell'equazione di auto-consistenza osserviamo che per lo sviluppo al prim'ordine è sufficiente conservare solo il primo termine all'interno della parentesi ($\hat{m}^{\star\iota\nu}$) perché l'altro contiene una moltiplicazione di parametri distinti e quindi ha un andamento del second'ordine. Dunque, la linearizzazione del membro di destra della prima condizione stazionaria in (3.26) è

$$\mathbb{E}_{\xi^\star}\,\mathbb{E}_{\{z\}}\,\xi^{\star\mu}\langle \xi^\nu \rangle_{\mathcal{L}_1^{(\mathrm{C})}} \sim \sum_\iota \hat{m}^{\star\iota\nu}\,\mathbb{E}_{\xi^\star}\left[\xi^{\star\mu}\xi^{\star\iota}\right] \tag{3.29}$$

Sottolineiamo che in questa espressione compaiono solamente i parametri ausiliari $\hat{m}^{\star\mu\nu}$. Osserviamo che l'equazione relativa a questi parametri coniugati è analoga a quella appena risolta: si distingue per il fattore moltiplicativo $\alpha\beta^2$, per la distribuzione delle variabili $\tau_\star$ (omologhe di $\xi^\star$) e per la presenza dei parametri d'ordine, con la relativa hamiltoniana $\mathcal{L}^{(\mathrm{O})}$ valutata con $\lambda = \beta^2$ (da non confondere con la temperatura inversa della media termica che invece resta pari a uno). Pertanto la risoluzione segue gli stessi passaggi e alla fine si ottiene

$$\alpha\beta^2\,\mathbb{E}_{\mathcal{M}_\star}\,\mathbb{E}_{\{z\}}\,\tau_{\star\mu}\langle \tau_\nu \rangle_{\mathcal{L}_1^{(\mathrm{O})}} \sim \alpha\beta^2\sum_\iota \beta^2 m^{\star\iota\nu}\,\mathbb{E}_{\mathcal{M}_\star}\left[\tau_{\star\mu}\tau_{\star\iota}\right] \tag{3.30}$$

Da queste due equazioni si evince che a differenza del sistema originario la forma linearizzata al prim'ordine presenta una struttura meno interdipendente: le magnetizzazioni dipendono solamente dai rispettivi parametri coniugati e viceversa. Allora la parte relativa alle magnetizzazioni di nostro interesse dell'equazione di punto fisso si può scrivere nella forma $x = g(y); y = h(x)$. In questo caso per avere la stabilità di un punto (x_0, y_0) è necessario che x_0 sia stabile per $x = g(h(x))$ (e pure che soddisfi una condizione analoga in y_0); in questo caso la relativa matrice jacobiana è notoriamente data dal prodotto delle matrici delle due funzioni g e h.
Quindi una condizione sufficiente per l'instabilità della soluzione nulla è che il prodotto delle matrici jacobiane di (3.29) e (3.30) valutate in zero abbia un autovalore maggiore (strettamente) di uno. Ricordiamo che $\mathcal{Q}$ è la matrice di correlazione dei vettori $\xi^\star$ definita nella sezione 3.1 e chiamiamo $\mathcal{R}$ la matrice di correlazione dei vettori $\tau_\star$ rispetto alla media termica definita a partire dall'hamiltoniana in (3.22). Chiamiamo $\mathcal{J}$ e $\hat{\mathcal{J}}$ le due matrici jacobiane appena citate (valutate in zero). Queste sono matrici di dimensione $P^2 \times P^2$, esprimibili in questo modo: $\mathcal{J}^{\iota\kappa}_{\mu\nu} = \delta_{\nu\kappa}\,\mathcal{Q}_{\mu\iota}$ e $\hat{\mathcal{J}}^{\iota\kappa}_{\mu\nu} = \delta_{\nu\kappa}\,\mathcal{R}_{\mu\iota}$, dove in pedice abbiamo indicato l'indice di riga (componente della funzione) e in apice quello relativo alla colonna (variabile di derivazione). Intendiamo le variabili ordinate come in (3.26). Calcolando il loro prodotto si trova la matrice $(\mathcal{J}\hat{\mathcal{J}})^{\iota\kappa}_{\mu\nu} = \alpha\beta^4\,\delta_{\nu\kappa}(\mathcal{Q}\,\mathcal{R})_{\mu\iota}$. Questa si può esprimere come una matrice a blocchi di dimensione $P \times P$; ciascun blocco sarà una matrice scalare $c\,\mathrm{Id}$, il cui fattore moltiplicativo è $c = \alpha\beta^4(\mathcal{Q}\,\mathcal{R})_{\mu\nu}$ dove (μ, ν) è la posizione del blocco. Tra gli autovalori di questa matrice ci sono anche quelli della matrice $\alpha\beta^4(\mathcal{Q}\,\mathcal{R})$; infatti basta considerare gli autovettori di quest'ultima, diciamo $v = (v_1, \ldots, v_P)$, e costruire gli autovettori della matrice grande inserendo $P - 1$ zeri dopo ogni posizione: $\tilde{v} \coloneqq (v_1, \ldots, 0, v_2, \ldots, 0, v_P, \ldots, 0)$. In conclusione, la matrice della quale è necessario studiare gli autovalori risulta essere

$$\mathcal{S} = \alpha\beta^4\,\mathcal{Q}\,\mathcal{R} \tag{3.31}$$

La condizione sufficiente per l'instabilità della soluzione nulla, quindi del superamento della transizione di fase, si ottiene imponendo che il suo autovalore massimo $\lambda^{\mathcal{S}}_{\max}$ sia maggiore di uno. Con questa notazione ricaviamo la stima sulla soglia critica per l'apprendimento della RBM binaria:

$$\alpha_{\text{crit.}} \leq \frac{1}{\beta^4 \lambda^{\mathcal{S}}_{\max}} \tag{3.32}$$

3.3.2 Risultati in assenza di correlazione

Adesso che abbiamo ricavato un'espressione esplicita per la soglia critica di apprendimento ne analizziamo la variabilità rispetto ai parametri del sistema: temperatura, dimensioni della rete, correlazione dei pesi sinaptici. Cominciamo con il caso più semplice, supponendo che gli stati originari delle sinapsi siano tra loro indipendenti, cioè assumiamo che $\mathcal{Q} = \mathrm{Id}$. Nella recente letteratura scientifica per questo caso particolare è emersa la congettura che sostiene che la soglia critica non dipenda dal numero di unità nascoste della rete, purché questo sia finito [Bar+17; HWH19]. In particolare, dalle osservazioni empiriche e qualche calcolo effettuato per $P = 2$ risulta essere sempre pari a β^{-4}. Osserviamo brevemente che questa dipendenza dalla temperatura è qualitativamente ragionevole: più il rumore è alto (β si abbassa) e maggiore sarà il numero di esempi che la rete dovrà analizzare per iniziare ad apprendere i concetti.
La nostra analisi prova questa congettura, inoltre fornisce un'espressione semplificata per la pressione del modello e le equazioni di auto-consistenza. Il primo effetto dell'assenza di correlazioni è quello di rendere le medie su $\mathbb{E}_{\xi^\star}$ e $\mathbb{E}_{\mathcal{M}_\star}$ uniformi, quindi in particolare a variabili indipendenti. Allora la matrice in (3.31) diventa $\alpha\beta^4\,\mathrm{Id}$. Questo vuol dire che i suoi autovalori sono tutti pari a $\alpha\beta^4$ e quindi la soglia critica, nel caso senza correlazione, è $\alpha_{\text{crit.}} = \beta^{-4}$ come previsto dalla congettura.
Inoltre, questo fa sì che emerga una soluzione particolarmente simmetrica delle condizioni stazionarie, indipendente dal numero di unità nascoste della rete, racchiusa nel prossimo risultato. Peraltro questo risultato coincide con quello trovato in [HWH19] nel caso $P = 2$.

Teorema 3.14 — *Consideriamo le equazioni di auto-consistenza descritte nel teorema 3.12. Se* $\mathcal{Q} = \mathrm{Id}$*, queste ammettono una soluzione nella forma*

$$\begin{aligned} m^{\star\mu\nu} &= \delta_{\mu\nu}\, m & q^{=\mu\nu} &= 0 & q^{\neq\mu\nu} &= \delta_{\mu\nu}\, m \\ \hat{m}^{\star\mu\nu} &= \delta_{\mu\nu}\, \hat{m} & \hat{q}^{=\mu\nu} &= 0 & \hat{q}^{\neq\mu\nu} &= \delta_{\mu\nu}\, \hat{m} \end{aligned} \tag{3.33}$$

dove i parametri $m, \hat{m}$ *devono soddisfare le equazioni ridotte:*

$$\begin{aligned} m &= \mathbb{E}_z \left[\tanh(\hat{m} + \sqrt{\hat{m}}z)\right] \\ \hat{m} &= \alpha\beta^2\, \mathbb{E}_z \left[\tanh(\beta^2 m + \beta\sqrt{m}z)\right] \end{aligned} \tag{3.34}$$

nelle quali $\mathbb{E}_z$ *indica la media rispetto ad una variabile reale distribuita secondo una normale standard.*

Dimostrazione. Per dimostrarlo studiamo come si semplificano le equazioni (3.26). Useremo le stesse notazioni del teorema precedente.

Supponendo che la soluzione sia in questa forma, la funzione $\mathcal{L}_\lambda$ definita in (3.20) diventa: $\mathcal{L}_\lambda(\{\sigma\};\{\tau_\star\},\{z_\mu, z_{\mu\nu}\}) = \sum_\mu \sigma_\mu(\lambda m\tau_{\star\mu} + \sqrt{\lambda m} z_\mu)$. Questo significa che le medie termiche rispetto alle hamiltoniane $\mathcal{L}^{(\mathrm{O})}$ e $\mathcal{L}^{(\mathrm{C})}$ sono indipendenti sugli spin. Una prima conseguenza di questo fatto è che le equazioni per $q^{=\mu\nu}$ e $q^{\neq\mu\nu}$ coincidono per ogni coppia $\mu < \nu$ (lo stesso vale anche per $\hat{q}^{=\mu\nu}$ e $\hat{q}^{\neq\mu\nu}$ siccome la media termica su $q^{=\mu\nu}$ diventa uniforme). Calcoliamo la media termica degli spin rispetto a $\mathcal{L}^{(\mathrm{O})}$ (il caso di $\mathcal{L}^{(\mathrm{C})}$ è analogo). Per un modello a spin indipendenti (si veda l'esempio 1.10) sappiamo che la media è data dalla tangente iperbolica del campo esterno, quindi nel nostro caso $\langle \tau_\mu \rangle_{\mathcal{L}^{(\mathrm{O})}} = \tanh(\beta^2 m\tau_{\star\mu} + \beta\sqrt{m} z_\mu)$. Ciò significa che l'equazione per $\hat{m}^{\star\mu\nu}$ diventa

$$\delta_{\mu\nu}\,\hat{m} = \alpha\beta^2\,\mathbb{E}_{\mathcal{M}_\star}\,\mathbb{E}_{\{z\}}\left[\tau_{\star\mu}\tanh(\beta^2 m\tau_{\star\nu} + \beta\sqrt{m} z_\nu)\right]$$

Per quanto riguarda il membro di destra, se $\mu \neq \nu$ i termini $\tau_{\star\mu}$ e tanh si separano per l'indipendenza della media e quindi si annullano perché la distribuzione di $\tau_{\star\mu}$ è uniforme; questo verifica la condizione imposta dalla delta di Dirac nel membro di sinistra. Altrimenti, per simmetria il membro di destra non dipende dall'indice μ; in più possiamo operare il cambio di variabili $z_\mu \mapsto z_\mu \tau_{\star\mu}$ e sfruttare il fatto che tanh è una funzione dispari per eliminare anche la dipendenza da $\tau_{\star\mu}$, ottenendo così l'equazione dell'enunciato. Il caso dell'equazione per $m^{\star\mu\nu}$ è del tutto analogo e fornisce l'altra equazione di (3.34). Le equazioni per $q^{\neq\mu\nu}$ e $\hat{q}^{\neq\mu\nu}$ si semplificano operando lo stesso cambio di variabili in z e risultano in delle equazioni analoghe, nelle quali la tangente iperbolica è elevata al quadrato. Ad esempio, per $\hat{q}^{\neq\mu\nu}$ si trova l'equazione

$$\hat{m} = \alpha\beta^2\,\mathbb{E}_z\left[\tanh(\beta^2 m + \beta\sqrt{m} z)^2\right]$$

Verifichiamo che questa in realtà coincide con l'altra equazione che abbiamo precedentemente ricavato per $\hat{m}$. Utilizziamo ancora la notazione formale $\underline{\mathrm{dz}}$ introdotta nel lemma 3.7. Partendo dall'equazione appena scritta operiamo il cambio di variabili $z \mapsto z - \beta\sqrt{m}$ nell'integrale; si ottiene l'espressione seguente, scritta in due forme equivalenti, la seconda delle quali deriva dalla prima a seguito un'ulteriore cambio di variabili $z \mapsto -z$ ricordando l'antisimmetria della funzione tangente iperbolica.

$$\alpha\beta^2 e^{-\frac{m\beta^2}{2}} \int \underline{\mathrm{dz}}\, e^{\beta\sqrt{m}z} \tanh(\beta\sqrt{m}z)^2 = \alpha\beta^2 e^{-\frac{m\beta^2}{2}} \int \underline{\mathrm{dz}}\, e^{-\beta\sqrt{m}z} \tanh(\beta\sqrt{m}z)^2$$

Facendo una media aritmetica delle due espressioni equivalenti appena esposte se ne ottiene una terza a loro equivalente:

$$\alpha\beta^2 e^{-\frac{m\beta^2}{2}} \int \underline{\mathrm{dz}} \cosh(\beta\sqrt{m}z) \tanh(\beta\sqrt{m}z)^2 = \alpha\beta^2 e^{-\frac{m\beta^2}{2}} \int \underline{\mathrm{dz}} \sinh(\beta\sqrt{m}z) \tanh(\beta\sqrt{m}z)$$

dove nell'ultimo passaggio abbiamo sfruttato il lemma 1.3. Analogamente a quanto fatto prima, possiamo interpretare il seno iperbolico come una media aritmetica di due espressioni, equivalenti tramite il cambio di variabili $z \mapsto -z$, una delle quali è

$$\alpha\beta^2 e^{-\frac{m\beta^2}{2}} \int \underline{\mathrm{dz}}\, e^{\beta\sqrt{m}z} \tanh(\beta\sqrt{m}z)$$

Tramite il cambio di variabili $z \mapsto z + \beta\sqrt{m}$ si ottiene l'equazione in (3.34) per $\hat{m}$. Un ragionamento analogo vale per m. □

Abbiamo già detto che queste equazioni non dipendono dal numero di unità nascoste, quindi forniscono un altra conferma del fatto che in assenza di correlazione il numero di unità nascoste sia ininfluente ai fini del comportamento del modello. Inoltre osserviamo che in questi punti l'espressione di $\mathcal{L}$ è sempre a valori reali, quindi il ragionamento fatto nella nota 3.8 si semplifica ulteriormente. Soprattutto possiamo notare che questa soluzione corrisponde ad un punto in cui magnetizzazione e sovrapponibilità tra le repliche coincidono, come previsto per un sistema che si trovi sulla linea di Nishimori. In più si verifica sperimentalmente che questa è proprio la soluzione sulla quale il sistema si stabilizza negli esperimenti numerici, quindi corrisponde effettivamente al punto estremale dell'energia libera.

Possiamo fare un'ulteriore osservazione riguardante la pressione del sistema quando si trova in corrispondenza di queste soluzioni. Nella prova del teorema 3.14 abbiamo visto che la funzione $\mathcal{L}_\lambda$ diventa a spin indipendenti, pertanto la sua funzione di partizione si può scrivere come prodotto delle funzioni di partizione dei sistemi con un solo spin; precisamente, nel caso di $\mathcal{L}^{(\mathrm{O})}_{\beta^2}$ (l'altro è analogo) si ottiene: $\mathcal{Z}(\mathcal{L}^{(\mathrm{O})}) = \prod_\mu 2\cosh(\beta^2 m\tau_{\star\mu} + \beta\sqrt{m}z_\mu)$. Allora, sfruttando l'indipendenza delle medie e la loro uniformità rispetto ai vari indici ricaviamo $\mathbb{E}_{\mathcal{M}_\star}\mathbb{E}_{\{z_\mu, z_{\mu\nu}\}} \log \mathcal{Z}(\mathcal{L}^{(\mathrm{O})}) = \mathbb{E}_{\tau_\star}\mathbb{E}_z \log(2\cosh(\beta^2 m\tau_\star + \beta\sqrt{m}z))$, dove la media su $\tau_\star$ è quella uniforme su uno spin dicotomico e quella su z è rispetto ad una normale standard. La dipendenza dalla media su $\tau_\star$ si può eliminare tramite un cambio di variabili nell'integrale gaussiano. Poi la funzione di partizione $\mathcal{Z}(q^{=\mu\nu})$ è pari a 2^P perché la relativa media termica è uniforme. Dunque, nel complesso la pressione del modello si può esprimere come

$$\begin{aligned}\mathrm{a}_{P,\beta,\alpha,\mathrm{Id}} = P\ \mathrm{extr}\ & -\frac{1}{2}m\hat{m} - \frac{1}{2}\hat{m} + \mathbb{E}_z \log\Big(2\cosh(\hat{m} + \sqrt{\hat{m}}z)\Big) \\ & - \frac{\alpha\beta^2}{2}m + \alpha\,\mathbb{E}_z \log\Big(2\cosh(\beta^2 m + \beta\sqrt{m}z)\Big) - \alpha\log 2\end{aligned}$$

Si osservi che questa espressione dipende linearmente dal numero di nodi nascosti del sistema. Da questa espressione si evince chiaramente che le soluzioni delle condizioni

stazionarie non dipendono dal numero di neuroni nascosti, in conformità con quanto visto nel teorema 3.14.

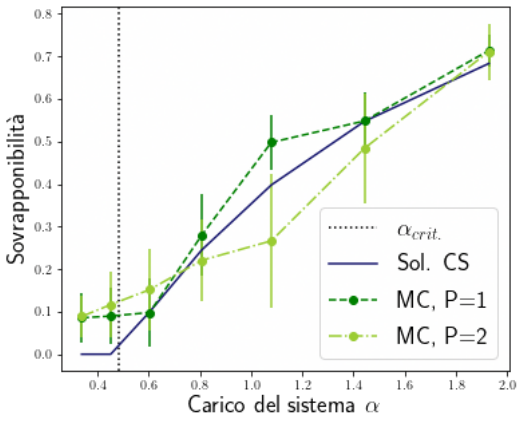


Figura 3.1: Qui è raffigurato il confronto, descritto alla fine della sezione 3.3.2, tra l'overlap del primo pattern ottenuto con il metodo Monte Carlo (linee tratteggiate) e la soluzione della prima equazione di (3.34) (linea continua). La temperatura inversa è fissata a $\beta = 1{,}2$. Le estrazioni del metodo MC, impostato con $N = 300$, sono effettuate dopo 7500 iterazioni; le barre di errore rappresentano il risultato medio su 40 repliche.

Per verificare queste previsioni teoriche abbiamo implementato numericamente il paradigma di analisi insegnante-studente. Innanzitutto abbiamo fissato una distribuzione di probabilità uniforme sullo spazio dei pattern originari $\xi^\star$; poi, usando il metodo di campionamento Metropolis-Hastings descritto nell'esempio 1.23, abbiamo estratto degli esempi $\{\sigma^a\}$ a partire da questo segnale, tramite la distribuzione diretta (3.1); successivamente abbiamo campionato, sempre con l'algoritmo MH, le variabili di inferenza ξ usando la distribuzione della RBM inversa (3.2), utilizzando gli esempi precedentemente estratti come disordine piantato; infine abbiamo calcolato le sovrapponibilità tra il segnale piantato nel sistema e quello inferito.
Il confronto tra i risultati di questo esperimento, effettuato al variare di P, e il parametro m ottenuto tramite la risoluzione numerica delle equazioni (3.34) è illustrato in figura 3.1. Nel grafico è rappresentata solo la sovrapponibilità del primo pattern inferito con il primo segnale originario. Per ottenere una comparazione più precisa abbiamo ripetuto il campionamento del problema inverso diverse volte con gli stessi esempi e fatto una media dei risultati ottenuti.

Possiamo riassumere l'analisi appena conclusa, riguardante il caso di RBM i cui pesi sinaptici siano inizializzati senza correlazione tra i pattern, affermando che il numero di nodi nascosti non influisce sulla soglia critica oltre la quale inizia l'apprendimento; in un certo senso la macchina tratta indipendentemente ciascun nodo nascosto, di conseguenza aumentando il loro numero è in grado di imparare più concetti con la stessa facilità.

3.3.3 Effetti della correlazione

Passiamo ora ad analizzare il caso in cui sia presente una correlazione non nulla tra i concetti impiantati nella RBM responsabile della generazione dei dati. Il comportamento della RBM che fa inferenza dipende dalla scelta della matrice $\mathcal{Q}$, come si evince dall'equazione (3.31). Assumiamo $P > 1$, altrimenti l'analisi seguente perde significato. Vista l'arbitrarietà della matrice, il caso generale è troppo variegato per essere analizzato teoricamente; tuttavia può essere all'occorrenza risolto numericamente. Noi ci concentreremo sul caso in cui tutte le correlazioni tra pattern distinti sono pari a un numero reale $q \in (0, 1]$; cioè nel caso in cui la matrice di correlazione sia nella forma

$$\mathcal{Q}_{\mu\nu} = \delta_{\mu\nu} + q(1 - \delta_{\mu\nu})$$

Precisiamo che abbiamo dovuto scegliere un valore di correlazione q non-negativo per evitare fenomeni di frustrazione del sistema (vedi l'esempio 1.14). La simmetria di questa matrice si riflette in un'analoga simmetria della matrice $\mathcal{R}$: la matrice delle correlazioni rispetto alla distribuzione di Boltzmann-Gibbs relativa all'hamiltoniana $\mathcal{M}_\star$. Infatti, nel caso in esame l'hamiltoniana si semplifica:

$$\mathcal{M}_\star(\{\tau_\star\}) = -\frac{\beta}{2} \sum_{\mu,\nu} \tau_{\star\mu}\tau_{\star\nu} \left(\delta_{\mu\nu} + q(1 - \delta_{\mu\nu})\right) = -P\frac{\beta}{2} - \beta q \sum_{\mu<\nu} \tau_{\star\mu}\tau_{\star\nu} \approx -\beta q \sum_{\mu<\nu} \tau_{\star\mu}\tau_{\star\nu}$$

Nell'ultimo passaggio abbiamo scartato un termine costante che non influisce sulla distribuzione. Riconosciamo che questa è l'hamiltoniana del modello di Curie-Weiss presentato nell'esempio 1.11, basta inglobare il coefficiente all'interno della temperatura inversa (che diventa quindi $\beta^2 q$). Che la correlazione non dipenda dagli spin indicati segue immediatamente dalla simmetria del modello; dunque, chiamiamo r il valore fuori dalla diagonale di $\mathcal{R}$.
Il prodotto di due matrici con un valore sulla diagonale e un altro fuori conserva questa simmetria; in questo caso la matrice (3.31) è la seguente:

$$\mathcal{S}_{\mu\nu} = \left(1 + (P-1)rq\right)\delta_{\mu\nu} + \left(q + r + (P-2)rq\right)(1 - \delta_{\mu\nu})$$

Usiamo il lemma 2.9 per calcolare gli autovalori, ottenendo:

$$\begin{cases} (1 + rq - q - r) + P(q + r + (P-2)rq) & \text{con molteplicità } 1 \\ 1 + rq - q - r & \text{con molteplicità } P - 1 \end{cases}$$

Per trovare l'autovalore massimo dobbiamo stabilire il segno di $q + r + (P-2)rq$. Questo dipende da r, che è la correlazione di due spin diversi rispetto alla media termica ricavata dall'hamiltoniana $\mathcal{M}_\star$. Trattandosi di un modello di campo medio con interazioni ferromagnetiche (cioè i coefficienti dei termini di interazione sono positivi), è noto che la correlazione tra due spin è sempre non-negativa, come conseguenza delle disuguaglianze GKS [Gri67]. Ciò significa che r è sempre positivo, quindi l'autovalore massimo è quello che ha sempre molteplicità 1 e vale

$$\lambda_{\max}^{\mathcal{S}} = (P-1)^2 rq + (P-1)(q+r) + 1$$

In primo luogo osserviamo che questo risultato estende sia quello presentato poc'anzi nel caso senza correlazione, che quello ottenuto in [HWH19] nel caso $P = 2$; infatti, in quest'ultima circostanza si può calcolare facilmente che $r = \tanh(\beta^2 q)$. Nell'articolo citato era già stato evidenziato che la correlazione incide sulla rapidità di apprendimento migliorandone nettamente le prestazioni. Avendo ricavato l'espressione per P generico possiamo dire di più: a P fissato, la correlazione facilita l'apprendimento e lo fa tanto di più quanti sono i nodi nascosti, secondo una relazione polinomiale. Si tenga presente che anche il numero di pattern da apprendere varia con P, perciò significa che la presenza di più concetti correlati tra loro facilita la macchina nell'individuazione di queste relazioni. In figura 3.2 è presentato l'andamento dell'autovalore massimo al variare degli altri parametri liberi; ricordiamo che la soglia critica gli è inversamente proporzionale (con un ulteriore fattore β^{-4}). Invece, in figura 3.3 abbiamo illustrato le linee di transizione di fase della RBM, al variare della correlazione q e del numero di pattern P. Fino a quando il grado di sollecitazione risulta inferiore alla soglia critica, cioè nella parte superiore dei diagrammi, il sistema si trova in uno stato paramagnetico; la parte inferiore dei diagrammi invece corrisponde alla fase di richiamo. Osserviamo che è sufficiente l'aggiunta di qualche nodo nascosto o un piccolo aumento della correlazione per abbassare notevolmente la soglia critica, quindi migliorare le prestazioni del modello. Questi diagrammi si possono confrontare con quello della macchina di Boltzmann diretta presentati in figura 2.3: la differenza ragguardevole è che la fase vetrosa della BM nel caso delle RBM è sostituita da una fase di richiamo.

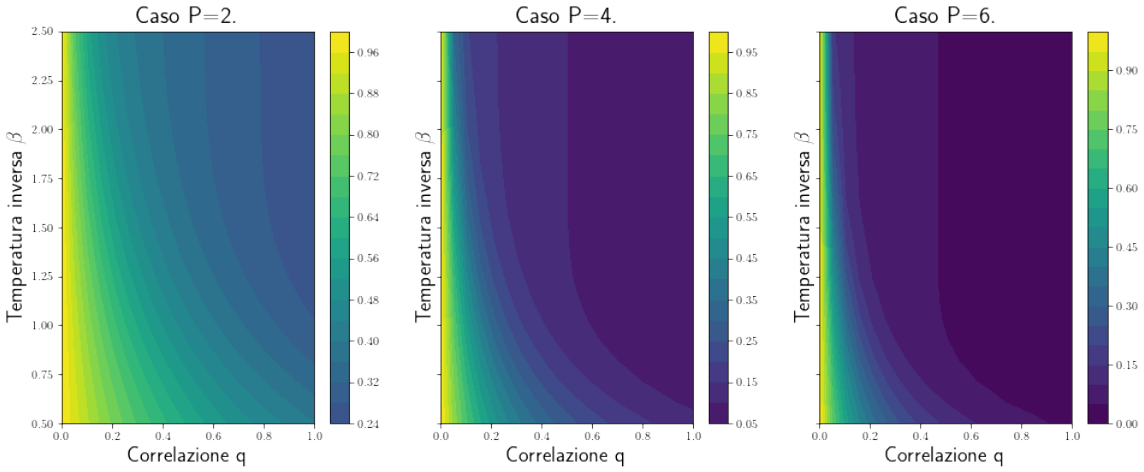


Figura 3.2: Qui è rappresentato il valore dell'autovalore massimo $\lambda^{\mathcal{S}}_{\max}$ al variare della temperatura inversa β, della correlazione q e del numero di nodi nascosti P.

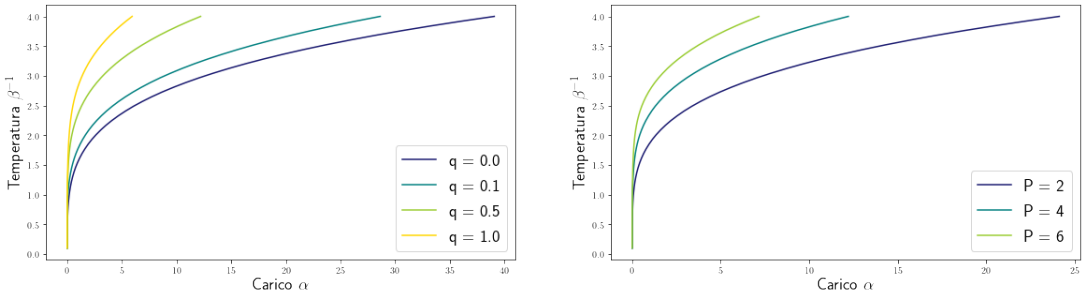


Figura 3.3: I due grafici qui presentati confrontano varie linee di transizione di fase per il modello RBM che abbiamo analizzato nel capitolo 3. Nel grafico a sinistra abbiamo fissato a $P = 4$ il numero di nodi nascosti facendo variare la correlazione q. Nel grafico di destra la correlazione è fissata a $q = 0{,}5$ e varia P.

Conclusioni (e prossimi passi)

Il presente studio aveva l'obiettivo di verificare l'esistenza di una transizione di fase nell'apprendimento del modello di rete neuronale artificiale detto Macchina di Boltzmann Ristretta e, nel caso fosse presente, di analizzarne le qualità. In particolare volevamo ottenere una conferma teorica della congettura presentata in [Bar+17; HWH19], riguardante la RBM a valori in ± 1. Infatti, negli articoli citati è stato ipotizzato che in assenza di correlazione tra i concetti insegnati alla macchina, il numero di esempi da fornirle prima che questa cominci ad apprendere non vari all'aumentare del numero di nodi nascosti che ha a disposizione per categorizzare le informazioni.
Per svolgere questo studio abbiamo utilizzato il paradigma insegnante-studente affinché fosse possibile confrontare, secondo indicatori certi, i risultati del modello con le informazioni effettivamente presenti all'interno dei dati sui quali è stato allenato. Abbiamo interpretato il problema con le tecniche della meccanica statistica, per via dell'ingente numero di neuroni con i quali vengono solitamente impostati questi modelli di apprendimento automatico; questo ci ha condotti a calcolare la pressione del modello, tramite qualche proficua intuizione e il metodo delle repliche, ricavando infine gli indicatori cercati.

I nostri risultati confermano l'esistenza della transizione di fase e pure il fatto che in assenza di correlazione questa non dipenda dal numero di neuroni nascosti ma solamente dal rumore presente nei dati. Inoltre, avendo ricavato una stima piuttosto generale per la localizzazione della soglia critica, siamo stati in grado di analizzare il comportamento del modello anche nel caso in cui i concetti sono tutti correlati tra loro (nella stessa misura). In questo caso emerge un contributo importante della correlazione nel facilitare il processo di apprendimento; inoltre, a dispetto delle aspettative, la generalità del nostro studio ci permette di affermare che il numero di neuroni nascosti del sistema influisce fortemente sulla curva di apprendimento accentuando l'effetto della correlazione.

Va sottolineato il fatto che la formalizzazione matematicamente rigorosa di alcune

tecniche utilizzate nello studio (e in generale nella disciplina, in particolare il metodo di Laplace e il trucco delle repliche) non è ancora completa; questo apre senz'altro la possibilità di approfondire il lavoro per chiarire le basi teoriche del risultato. *In primis* si potrebbe delucidare la natura dei punti estremali del funzionale dal quale si ricava la pressione, confrontando il metodo da noi seguito con una dettagliata analisi della matrice hessiana.
Contemporaneamente, la nostra analisi è supportata da un ventaglio di simulazioni numeriche che sarebbe utile ampliare, includendo ad esempio le simulazioni Monte Carlo in presenza di correlazione, per verificare le previsioni teoriche sulla dipendenza dal numero di unità nascoste; si potrebbe indagare anche il comportamento del modello quando il numero di concetti impiantati nei dati differisce dal numero di nodi nascosti utilizzati per l'inferenza.
Oltre a queste direttrici di ricerca, un paio di generalizzazioni particolarmente immediate dello studio che abbiamo svolto sono l'introduzione di un prior non uniforme sui pattern originari durante il processo di inferenza (estendendo quando presentato in [Hua21]) e la distinzione delle temperature del processo generativo e di quello di inferenza, ossia l'analisi del modello fuori dalla linea di Nishimori. Oppure sarebbe interessante verificare la teoria della rottura di simmetria di permutazione descritta in [HWH19] (nel caso $P = 2$) e estendere il presente lavoro analizzando l'influenza del tipo di unità della RBM, similmente a quanto fatto in [Bar+18].

Bibliografia

[MP43] W. S. McCulloch e W. Pitts. «A logical calculus of the ideas immanent in nervous activity». In: *The bulletin of mathematical biophysics* 5 (1943). DOI: 10.1007/BF02478259.

[Hör63] Lars Hörmander. *Linear Partial Differential Operators*. 3ª ed. Heidelberg: Springer Berlin, 1963. DOI: 10.1007/978-3-642-46175-0.

[Erd65] A. Erdélyi. *Asymptotic expansions*. Dover, 1965.

[Gri67] Robert B. Griffiths. «Correlations in Ising Ferromagnets. I». In: *Journal of Mathematical Physics* 8.3 (1967). DOI: 10.1063/1.1705219.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis*. 3ª ed. New York: McGraw-Hill, 1976.

[Par80a] Giorgio Parisi. «A sequence of approximated solutions to the S-K model for spin glasses». In: *Journal of Physics A: Mathematical and General* 13.4 (1980). DOI: 10.1088/0305-4470/13/4/009.

[Par80b] Giorgio Parisi. «The order parameter for spin glasses: a function on the interval 0-1». In: *Journal of Physics A: Mathematical and General* 13.3 (1980). DOI: 10.1088/0305-4470/13/3/042.

[ER82] Richard S. Ellis e Jay S. Rosen. «Laplace's Method for Gaussian Integrals with an Application to Statistical Mechanics». In: *The Annals of Probability* 10.1 (1982).

[Hop82] J. J. Hopfield. «Neural networks and physical systems with emergent collective computational abilities». In: *Proceedings of the national academy of sciences* 79.8 (1982). DOI: 10.1073/pnas.79.8.2554.

[HSA84] Geoffrey E. Hinton, Terrence J. Sejnowski e David H. Ackley. *Boltzmann Machines: Constraint Satisfaction Networks that Learn*. Rapp. tecn. Pittsburgh, PA, U.S.A.: Carnegie-Mellon University, Department of Computer Science, 1984.

[HSA85] Geoffrey E. Hinton, Terrence J. Sejnowski e David H. Ackley. «A learning algorithm for Boltzmann machines». In: *Cognitive science* 9.1 (1985). DOI: 10.1016/S0364-0213(85)80012-4.

[MPV86] M. Mézard, G. Parisi e M. Virasoro. *Spin Glass Theory and Beyond*. Lecture Notes in Physics 9. World Scientific, 1986. DOI: 10.1142/0271.

[Smo86] Paul Smolensky. «Information Processing in Dynamical Systems: Foundations of Harmony Theory. Foundations». In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1. MIT Press, 1986. URL: https://stanford.edu/~jlmcc/papers/PDP/Volume%201/Chap6_PDP86.pdf (visitato il 05/12/2022).

[Rud91] Walter Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991.

[De 92] Giuseppe De Marco. *Analisi Due: secondo corso di analisi matematica per l'università*. it. A cura di Decibel editrice. 2 voll. Bologna: Zanichelli editore, 1992.

[Car95] Henri Paul Cartan. *Elementary Theory of Analytic Functions of One Or Several Complex Variables*. Dover Publications, 1995.

[Nor97] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. DOI: 10.1017/CBO9780511810633.

[Nis01] Hidetoshi Nishimori. *Statistical Physics of Spin Glasses and Information Processing. An Introduction*. International series of monographs on physics. Oxford University Press, 2001.

[Won01] R. Wong. *Asymptotic Approximations of Integrals*. Classics in Applied Mathematics. Society for Industrial e Applied Mathematics, 2001.

[Hin02] Geoffrey E. Hinton. «Training Products of Experts by Minimizing Contrastive Divergence». In: *Neural computation* 14.8 (2002). DOI: 10.1162/089976602760128018.

[Mac03] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. 6ª ed. Cambridge University Press, 2003. URL: http://www.inference.phy.cam.ac.uk/mackay/itila/ (visitato il 05/12/2022).

[CKS05] A. C. C. Coolen, R. Kuehn e P. Sollich. *Theory of Neural Information Processing Systems*. Oxford University Press, 2005.

[EK05] Stewart N. Ethier e Thomas G. Kurtz. *Markov Processes. Characterization and Convergence*. Wiley Series in Probability and Statistics. Wiley, 2005.

[HOT06] Geoffrey E. Hinton, Simon Osindero e Yee-Whye Teh. «A Fast Learning Algorithm for Deep Belief Nets». In: *Neural Computation* 18.7 (2006). DOI: 10.1162/neco.2006.18.7.1527.

[But07] R.W. Butler. *Saddlepoint Approximations with Applications*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2007. DOI: 10.1017/CBO9780511619083.

[KS07] Leonid Koralov e Yakov G. Sinai. *Theory of Probability and Random Processes*. 2ª ed. Universitext. Heidelberg: Springer Berlin, 2007. DOI: 10.1007/978-3-540-68829-7.

[LB08] Nicolas Le Roux e Yoshua Bengio. «Representational power of restricted boltzmann machines and deep belief networks». In: *Neural computation* 20.6 (2008). DOI: 10.1162/neco.2008.04-07-510.

[MM09] Marc Mézard e Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009. DOI: 10.1093/acprof:oso/9780198570837.001.0001.

[SH09] Ruslan Salakhutdinov e Geoffrey Hinton. «Deep Boltzmann Machines». In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. Vol. 5. Proceedings of Machine Learning Research. PMLR, 2009, pp. 448–455. URL: https://proceedings.mlr.press/v5/salakhutdinov09a.html (visitato il 14/11/2022).

[Hin10] Geoffrey E. Hinton. *A Practical Guide to Training Restricted Boltzmann Machines*. 2010. URL: https://www.cs.toronto.edu/~hinton/absps/guideTR.pdf (visitato il 14/11/2021).

[MA11] Guido Montufar e Nihat Ay. «Refinements of Universal Approximation Results for Deep Belief Networks and Restricted Boltzmann Machines». In: *Neural Computation* 23.5 (2011). DOI: 10.1162/NECO_a_00113.

[Tal11] Michel Talagrand. *Mean Field Models for Spin Glasses. Basic Examples*. Vol. 1. Heidelberg: Springer Berlin, 2011. DOI: 10.1007/978-3-642-15202-3.

[Bar+12] Adriano Barra et al. «On the equivalence of Hopfield networks and Boltzmann Machines». In: *Neural Networks* 34 (2012). DOI: 10.1016/j.neunet.2012.06.003.

[Bar+15] Adriano Barra et al. «Multi-Species Mean Field Spin Glasses. Rigorous Results». In: *Annales Henri Poincaré* 16 (2015). DOI: 10.1007/s00023-014-0341-5.

[LBH15] Yann LeCun, Yoshua Bengio e Geoffrey Hinton. «Deep learning». In: *Nature* 521 (2015). DOI: 10.1038/nature14539.

[Pan15] Dmitry Panchenko. «The free energy in a multi-species Sherrington–Kirkpatrick model». In: *The Annals of Probability* 43.6 (2015). DOI: 10.1214/14-AOP967.

[Alb+16] Diego Alberici et al. «Limit Theorems for Monomer-Dimer Mean-Field Models with Attractive Potential». In: *Communications in Mathematical Physics* 346 (2016). DOI: 10.1007/s00220-015-2543-1.

[HT16] Haiping Huang e Taro Toyoizumi. «Unsupervised feature learning from finite data by message passing: Discontinuous versus continuous phase transition». In: *Physical Review E* 94 (6 2016). DOI: 10.1103/PhysRevE.94.062310.

[Bar+17] Adriano Barra et al. «Phase transitions in restricted Boltzmann machines with generic priors». In: *Physical Review E* 96 (4 2017). DOI: 10.1103/PhysRevE.96.042156.

[FV17] Sacha Friedli e Yvan Velenik. *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge University Press, 2017. DOI: 10.1017/9781316882603. URL: https://www.unige.ch/math/folks/velenik/smbook/ (visitato il 05/12/2022).

[Méz17] Marc Mézard. «Mean-field message-passing equations in the Hopfield model and its generalizations». In: *Physical Review E* 95 (2 2017). DOI: 10.1103/PhysRevE.95.022117.

[Bar+18] Adriano Barra et al. «Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors». In: *Physical Review E* 97 (2 2018). DOI: 10.1103/PhysRevE.97.022310.

[Tub18] Jérôme Tubiana. «Restricted Boltzmann machines: from compositional representations to protein sequence analysis». Tesi di dott. Université Paris sciences et lettres, 2018. URL: https://tel.archives-ouvertes.fr/tel-02183417 (visitato il 05/12/2022).

[Bar19] Jean Barbier. *Mean-field theory of high-dimensional Bayesian inference*. 2019. URL: https://www.sissa.it/sites/default/files/CoursePisa.pdf (visitato il 14/11/2021).

[Dur19] Rick Durrett. *Probability. Theory and Examples*. 5ª ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. URL: https://services.math.duke.edu/~rtd/PTE/PTE5_011119.pdf (visitato il 05/12/2022).

[HWH19] Tianqi Hou, K. Y. Michael Wong e Haiping Huang. «Minimal model of permutation symmetry in unsupervised learning». In: *Journal of Physics A: Mathematical and Theoretical* 52.41 (set. 2019). DOI: 10.1088/1751-8121/ab3f3f.

[Zub19] Shoshana Zuboff. *Il capitalismo della sorveglianza. Il futuro dell'umanità nell'era dei nuovi poteri*. it. Luiss University Press, 2019.

[Mat20] Dipartimento di Matematica e Fisica "Ennio De Giorgi" dell'Università del Salento, cur. *Ithaca. Viaggio nella Scienza* 16 (2020): *Intelligenza artificiale*. it. URL: http://ithaca.unisalento.it/ (visitato il 05/12/2022).

[Pas20] Andrea Pascucci. *Teoria della Probabilità. Variabili aleatorie e distribuzioni*. it. Springer Milano, 2020. DOI: 10.1007/978-88-470-4000-7.

[Alb+21] Diego Alberici et al. «The Solution of the Deep Boltzmann Machine on the Nishimori Line». In: *Communications in Mathematical Physics* 387 (2021). DOI: 10.1007/s00220-021-04165-0.

[Dec+21] Aurelien Decelle et al. «Inverse problems for structured datasets using parallel TAP equations and restricted Boltzmann machines». In: *Scientific Reports* 11.19990 (2021). DOI: 10.1038/s41598-021-99353-2.

[DF21] Aurélien Decelle e Cyril Furtlehner. «Restricted Boltzmann machine: Recent advances and mean-field theory». In: *Chinese Physics B* 30.4 (2021). DOI: 10.1088/1674-1056/abd160.

[Hua21] Haiping Huang. *Statistical Mechanics of Neural Networks*. Springer Singapore, 2021. DOI: 10.1007/978-981-16-7570-6.

[MA21] Chiara Marullo e Elena Agliari. «Boltzmann Machines as Generalized Hopfield Networks: A Review of Recent Results and Outlooks». In: *Entropy (Basel, Switzerland)* 23 (1 2021). DOI: 10.3390/e23010034.