Alma Mater Studiorum · Università di Bologna

DIPARTIMENTO DI MATEMATICA
Corso di Laurea Triennale in Matematica

MODELLO TEACHER-STUDENT PER LA REGRESSIONE LINEARE

Relatore:
Prof.ssa Federica Gerace
Correlatore:
Dott. Théo Marchetta

Presentata da:
Claudia Magli

Anno Accademico 2024–2025

A chi, con passione e semplicità,
mi ha fatto incontrare la Matematica.

Indice

Introduzione

Negli ultimi anni, il rapido progresso tecnologico ha offerto un terreno fertile per la crescita e lo sviluppo del machine learning; ad oggi, molte delle attività quotidiane che compiamo sono supportate da sistemi intelligenti: dallo sblocco del telefono tramite riconoscimento facciale, all'ambito medico con la diagnostica anticipata.
Questa tesi propone di affrontare lo studio del modello teacher-student per la regressione lineare, un paradigma teorico che consente di analizzare in maniera controllata i meccanismi di apprendimento supervisionato. In tale modello, un "teacher" genera i dati secondo una regola nota, ma non accessibile allo "student", che ha il compito di inferire questa regola a partire da un insieme finito di esempi.
Il primo capitolo rappresenta una panoramica sul machine learning, dalla sua nascita all'utilizzo odierno mediante sistemi che, imitando le funzioni del cervello umano, riescono ad apprendere ed elaborare informazioni per fare previsioni su dati mai visti prima. Il modello si presta naturalmente all'analisi meccanico-statistica, permettendo di utilizzare concetti fondamentali della fisica dei sistemi disordinati, come l'energia libera, la funzione di partizione, e il replica trick, temi trattati ed approfonditi nel secondo capitolo. Dopo un'introduzione sul machine learning e sulle reti neurali, la tesi esplora il calcolo dell'energia libera e delle equazioni di saddle point associate al modello, arrivando a dedurre, nel terzo capitolo, un'espressione analitica dell'errore di generalizzazione, che rappresenta una misura chiave delle performance predittive dello student. Nel quarto ed ultimo capitolo, tramite

simulazioni numeriche in ambiente Python, si confrontano i valori dell'errore di generalizzazione e dei parametri d'ordine calcolati teoricamente con quelli ottenuti empiricamente, analizzando la dipendenza dal rapporto tra numero di esempi e dimensione del problema. Questa duplice impostazione, teorica e computazionale, consente non solo di validare il modello, ma anche di comprenderne i limiti e i punti di forza, fornendo un contributo utile alla comprensione dei meccanismi fondamentali che governano l'apprendimento nei sistemi complessi ad alta dimensionalità.

Capitolo 1

Il Machine Learning

1.1 Introduzione al machine learning

Intelligenza artificiale (AI) è un termine generico che si riferisce a sistemi o macchine che imitano l'intelligenza umana.
Un sottoinsieme dell'intelligenza artificiale è il Machine Learning (ML) che si occupa di creare sistemi che imparano a svolgere determinati compiti come la traduzione di un testo o la generazione di contenuto grafico, direttamente da esempi raccolti in vasti archivi di dati, senza la necessità di istruzioni esplicite da parte dei programmatori [1]; fonda le sue radici negli anni '50 del secolo scorso con la pubblicazione da parte di Alan Turing, matematico, logico, crittografo e filosofo britannico, dell'articolo "Computing Machinery and Intelligence". Turing voleva capire quanto una macchina potesse essere vicina all'essere umano e per cercare una risposta inventò il "test di Turing", ancora oggi utilizzato per capire se una macchina può essere considerata intelligente. Il test è ispirato al "gioco dell'imitazione" in cui un interrogante attraverso le risposte, talvolta truccate, di due persone di sesso opposto, deve capire chi tra i due è la donna e chi l'uomo; a questo punto, sostituendo la donna con una macchina si ripete lo stesso gioco. Se il numero di volte in cui l'interrogante associa il sesso ai due personaggi è pressoché simile al precedente, allora si potrebbe considerare la macchina intelligente in quanto

indistinguibile da un essere umano [2].
Nel corso degli anni sono state presentate diverse versioni del test sia per l'evoluzione del concetto di macchina intelligente, ma anche perché alcune macchine, pur non essendo obiettivamente intelligenti, lo superavano: è il caso di Eliza, macchina ideata per emulare uno psicoterapeuta [3].
Durante la più grande competizione per la conduzione di test di Turing, organizzata nel 2012 a Milton Keynes, il chatterbot "Eugene Goostman" programmato da ricercatori russi per rispondere come un ragazzino ucraino, convinse il 29% dei giudici presenti [4]. Come evidenziato nell'articolo pubblicato da Geopop [5], alcuni studi recenti sostengono che la versione 4.5 di ChatGPT passa il test di Turing con l'80% di probabilità e, confrontando le risposte del chatterbot con quelle fornite da oltre 100.000 individui di 52 Paesi diversi, gli studiosi non hanno rilevato sostanziali differenze. ChatGPT-4 mostra comportamenti e caratteristiche psicologiche che, a livello statistico, non si distinguono da quelli di una persona reale.

1.2 Le reti neurali

Comprendere come il cervello umano svolge le sue attività è da sempre un obiettivo condiviso da diverse comunità scientifiche. Geoffrey Hinton, psicologo e informatico britannico-canadese, è noto come uno dei padri fondatori dell'intelligenza artificiale moderna, insieme a Yann LeCun e Yoshua Bengio. Nel 2018, i tre scienziati hanno ricevuto il Premio Turing per i loro contributi pionieristici nello sviluppo del deep learning e delle reti neurali artificiali [6].
Il nostro cervello contiene circa 85 miliardi di neuroni, collegati tra loro attraverso 10^{14} connessioni sinaptiche, che costituiscono l'unità funzionale del tessuto nervoso. I neuroni hanno due proprietà fondamentali:

- la conducibilità: capacità di trasmettere impulsi elettrici lungo l'assone.
- l'eccitabilità: capacità di elaborare l'informazione a seguito dell'eccitazione dovuta ad uno stimolo e di generare una risposta.

Si ritiene che tali fenomeni siano resi possibili dalla plasticità sinaptica, ovvero dalla capacità del nostro cervello di modificare l'intensità delle connessioni tra i neuroni, nonché di formarne di nuove o eliminarne di esistenti. In particolare, prima di qualsiasi esperienza, la struttura delle connessioni sinaptiche è determinata geneticamente. In seguito all'apprendimento, tale struttura si modifica rafforzando le connessioni tra i neuroni che tendono ad attivarsi simultaneamente. Un esempio classico di questo meccanismo si osserva nell'apprendimento di una nuova abilità motoria, come suonare uno strumento musicale: con la pratica ripetuta, i circuiti neurali coinvolti nell'esecuzione dei movimenti diventano progressivamente più efficienti, grazie al rafforzamento delle connessioni sinaptiche tra i neuroni attivati con maggiore frequenza [7].
Si è dimostrato che l'apprendimento tramite esempi può essere descritto ed analizzato per mezzo della meccanica statistica, dal momento che non siamo interessati a studiare le proprietà del singolo neurone, ma piuttosto quelle collettive, quali l'apprendimento o la memoria. In particolare in questa disciplina le reti neurali biologiche vengono modellizzate attraverso le cosiddette **reti neurali artificiali**.
Nel 1943 gli studiosi McCulloch e Pitts proposero il primo modello di neurone artificiale, in cui la risposta di un neurone a uno stimolo esterno o output viene generata attraverso una somma del prodotto degli stimoli esterni o input per l'intensità della connessione sinaptica o peso [8].

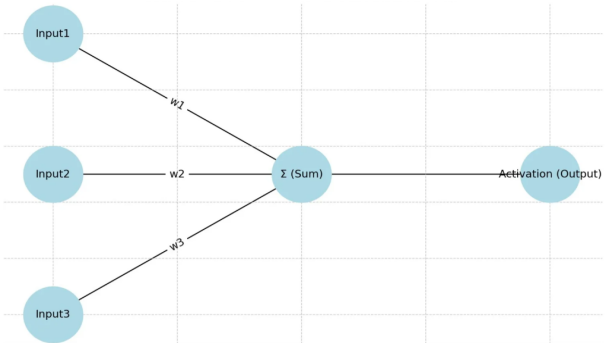


Figura 1.1: Schema del neurone di McCulloch-Pitts [8]

Nel 1958 Frank Rosenblatt, con l'articolo "The perceptron: a probabilistic model for information storage and organization in the brain" [9], presentò

il percettrone, ad oggi considerato una delle forme più semplici di neurone artificiale in cui M input vengono mappati in un unico output.
Gli elementi principali di una rete neurale artificiale sono i nodi o unità, più comunemente noti come neuroni. All'interno della rete possiamo individuare tre diversi tipi di strati con cui i neuroni si connettono per processare le informazioni.

- **input layer**: progettato per captare informazioni dall'esterno che la rete deve imparare, riconoscere e processare;
- **hidden layers**: connessione tra l'input layer e l'output layer che consente di estrarre dai dati quelle caratteristiche o "features" dell'input che sono rilevanti per svolgere un determinato compito;
- **output layer**: ultimo strato utilizzato per mostrare ciò che la rete ha imparato.

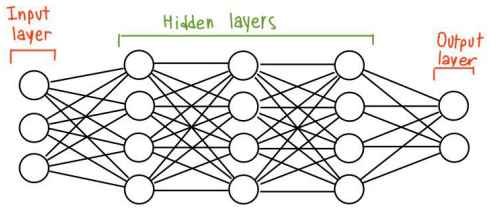
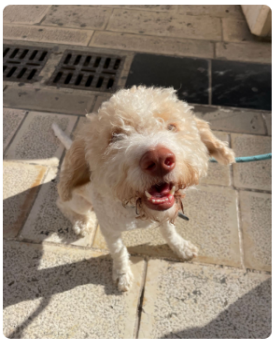


Figura 1.2: Rete neurale artificiale

Se le connessioni tra i nodi non formano cicli parliamo di rete neurale feed-forward, altrimenti si parla di reti neurali ricorrenti.
Il riconoscimento delle immagini è una delle numerose applicazioni delle reti neurali artificiali; in input viene fornita un'immagine successivamente convertita in una matrice di pixel, i dati vengono poi elaborati attraverso gli strati nascosti, in cui vengono estratte caratteristiche rilevanti. Ogni livello successivo elabora combinazioni più complesse delle informazioni precedenti. Infine, lo strato di output restituisce una probabilità associata a ciascuna

classe possibile (es. “cane”), indicando quale sia l’ipotesi più probabile. Questo meccanismo è alla base di molti sistemi di computer vision, dai filtri fotografici fino alle auto a guida autonoma.



What a cheerful pup! This dog looks like a Lagotto Romagnolo — a breed known for its curly coat and exceptional truffle-hunting skills. The dog's happy expression, curly fur, and light brown nose are all characteristic traits. Is there anything you'd like me to do with the image, such as enhancing it, identifying features, or creating a fun version (like a cartoon or painting)?

Fai una domanda

+ Strumenti

ChatGPT può commettere errori. Assicurati di verificare le informazioni importanti. Vedi Preferenze sui cookie.

Figura 1.3: Riconoscimento immagine Chat-Gpt 4o

È proprio in questi contesti che vengono studiati i “big data” e problemi di Machine Learning, dove è necessario introdurre nuovi algoritmi matematici, efficaci dal punto di vista statistico ed efficienti dal punto di vista computazionale.

Capitolo 2

Cenni di Meccanica Statistica

La meccanica statistica nasce a cavallo tra il diciannovesimo e il ventesimo secolo, come branca della fisica che si occupa dello studio dei sistemi nella loro interezza, piuttosto che attraverso le singole parti che li compongono. Negli ultimi anni, la meccanica statistica ha ricoperto un ruolo significativo nello studio dei problemi di apprendimento automatico, permettendo di comprendere meglio in quali condizioni un sistema possa raggiungere buone prestazioni [10].
In questo capitolo introduciamo gli strumenti concettuali e matematici della meccanica statistica che saranno fondamentali per analizzare il comportamento dei sistemi di Machine Learning.

2.1 Inferenza Bayesiana

Teorema 2.1.1 (Bayes)**.** *Siano x, y eventi non trascurabili, vale:*

$$P(\boldsymbol{X} = x \mid y) := P(x \mid y) = \frac{P(x)\, P(y \mid x)}{P(y)}$$

Considerando y come un set di dati $\mathbf{D}$ ed x come il sistema $\mathbf{M}$ che li ha generati, possiamo scrivere:

$$P(\mathbf{M}|\mathbf{D}) \propto P(\mathbf{D}|\mathbf{M})P(\mathbf{M})$$

Il membro di sinistra è chiamato **probabilità a posteriori**, in quanto determina la probabilità che sia il sistema **M** ad aver generato il set di dati **D**. Nel membro di destra sono presenti due parti: $P(\mathbf{M})$ chiamata **probabilità a priori**, questa non dipende dai dati e rappresenta l'ipotesi iniziale circa il modello e la **funzione di verosimiglianza** $P(\mathbf{D}|\mathbf{M})$, che quantifica come l'ipotesi iniziale circa il modello **M** si modifica una volta osservato il set di dati **D**.
Questa espressione è particolarmente utile nel contesto del Machine Learning, nel quale l'obiettivo è di inferire, a partire dai dati osservati, la distribuzione dei pesi della rete; in base al dataset si cerca di recuperare la distribuzione a posteriori dei parametri del modello. Tale prospettiva è adottata anche nell'ambito della Meccanica Statistica, in cui il punto di vista probabilistico permette di interpretare l'apprendimento automatico come un processo di inferenza sui modelli che meglio descrivono i dati osservati.

Definizione 2.1.1 (Funzione di verosimiglianza)**.** *La funzione di verosimiglianza*

$$\mathcal{L}_y(x) := P(y \mid x)$$

è una funzione di probabilità condizionata, considerata come funzione del suo secondo argomento mantenendo fisso il primo. Formalmente è una funzione:

$$\mathcal{L}_y : x \mapsto P(y \mid \boldsymbol{X} = x)$$

Nel contesto del ML, questa funzione di verosimiglianza $\mathcal{L}_y(x)$ rappresenta la probabilità di ottenere una certa etichetta y dato un input $\mathbf{X}$. In pratica, descrive la relazione tra i dati a disposizione e l'informazione che vogliamo predire. In questo modo abbiamo la possibilità di modellare l'incertezza e di costruire algoritmi capaci di imparare dai dati per fare previsioni su nuovi esempi in modo coerente con l'esperienza passata.

2.2 Problema di apprendimento supervisionato

L'**apprendimento supervisionato** è un approccio dell'apprendimento automatico in cui ad un sistema vengono forniti degli esempi composti da coppie input-output, con l'obiettivo di permettergli di apprendere una regola che associ correttamente l'uno all'altro e fare in modo che lo stesso sia in grado di costruire un modello che generalizzi la relazione tra dati in ingresso e le corrispondenti etichette, al fine di effettuare previsioni su casi non visti. Ad esempio, si può fornire al sistema l'immagine di un gatto (input) associata alla sua etichetta "gatto" (output), affinché impari a riconoscere e catalogare immagini simili in futuro. Consideriamo dunque il set di dati $D = \{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^{M}$, che consiste in M esempi tali che $\mathbf{X} \sim \mathcal{N}(0, 1_d)$, dove con d indico la dimensione del dato in ingresso.

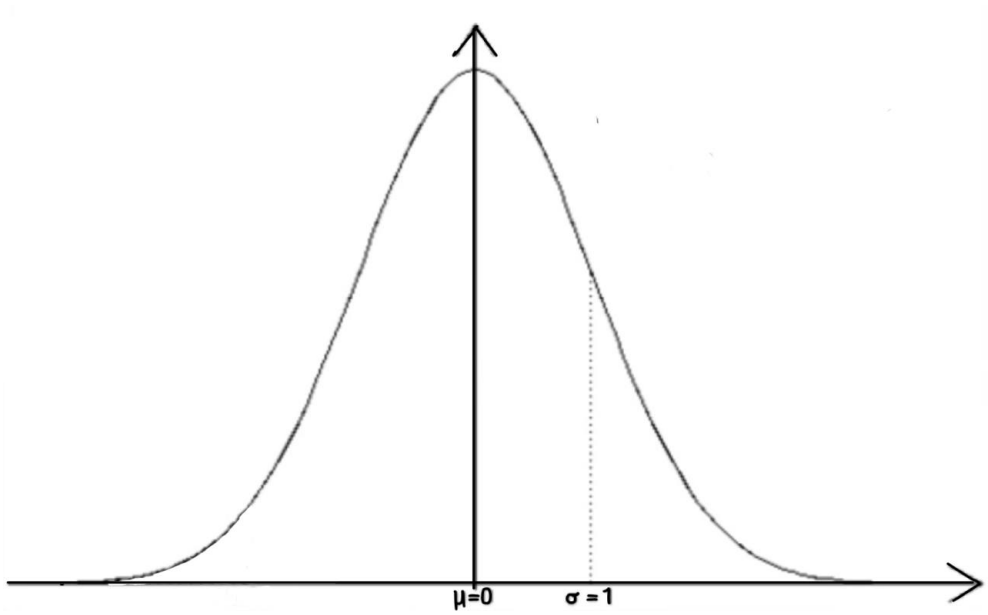


Figura 2.1: Funzione di densità della normale standard

Le etichette (label) y^μ sono generate da una rete neurale a singolo layer che

chiamiamo teacher dato da:

$$y^\mu = \mathbf{W}^* \mathbf{X}^\mu + \sigma^* \xi$$

dove indichiamo con $\mathbf{W}^*$ i pesi del teacher e con $\sigma^* \xi$ l'eventuale rumore Gaussiano presente durante il processo di etichettamento dei dati eseguito dal teacher. Poiché il teacher assegna un'etichetta a ciascun input in modo probabilistico, a causa della presenza di rumore nel processo generativo, la variabile y^μ è aleatoria. Se si assume che il rumore sia gaussiano, $\xi \sim \mathcal{N}(0, 1)$, allora la legge secondo cui il teacher genera y^μ è data da una Gaussiana con media $\mathbf{W}^* \mathbf{X}^\mu$ e varianza σ^{*^2}:

$$y^\mu \sim \mathcal{N}(\mathbf{W}^* \mathbf{X}^\mu, \sigma^{*^2})$$

Consideriamo dunque una rete neurale a singolo layer, che chiamiamo student, descritta da:

$$\hat{y}^\mu = \mathbf{W} \ \mathbf{X}^\mu$$

dove indichiamo con $\mathbf{W}$ i pesi dello student.
L'obiettivo del conto che svilupperò è caratterizzare le performance di generalizzazione dello student, quindi dell'estimatore:

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \left[\mathcal{L}(y - \hat{y}) + \frac{\lambda}{2} ||\mathbf{W}||_2^2 \right] \tag{2.1}$$

la funzione $\mathcal{L}$ prende il nome di **loss function**, utilizzata in apprendimento automatico per misurare la distanza tra l'etichetta reale e quella inferita,

$$\mathcal{L}(y - \hat{y}) = \frac{1}{2} \sum_{\mu=1}^{M} (y^\mu - \hat{y^\mu})^2.$$

Per chiarire visivamente il problema dell'apprendimento supervisionato, si consideri la figura:

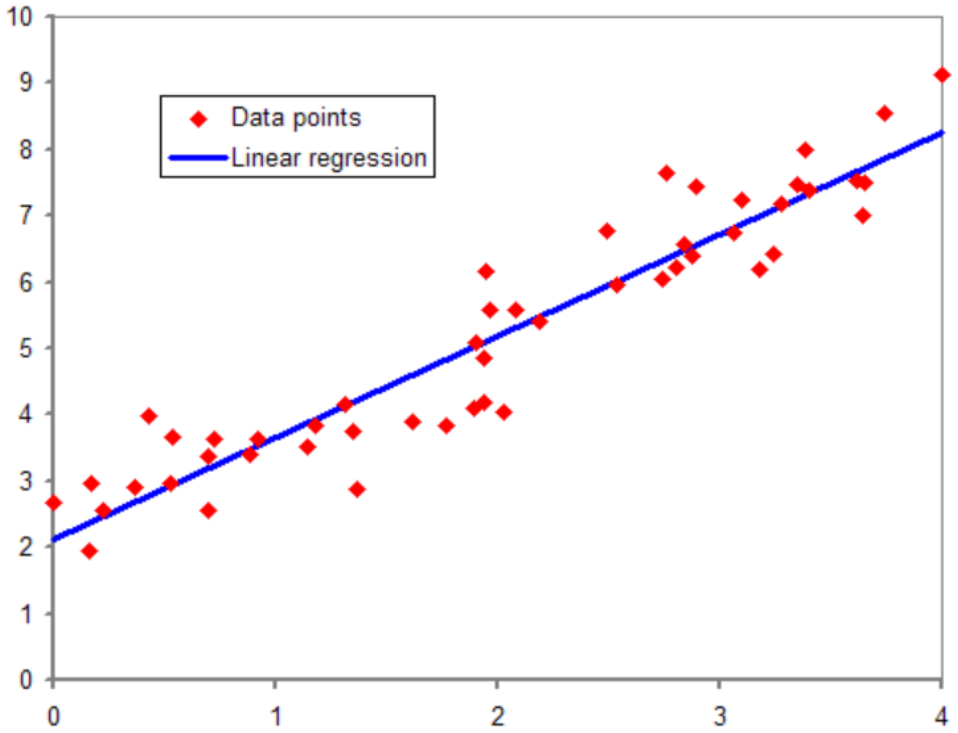


Figura 2.2: Regressione lineare

I punti rossi rappresentano le osservabili reali (input-output del dataset), mentre la linea blu rappresenta il modello appreso, ovvero lo student, che può essere una rete neurale a singolo layer o un semplice modello di regressione lineare; rappresenta la soluzione ottimale trovata, ovvero la configurazione di $\mathbf{W}$ che meglio approssima i dati nel senso della minima perdita. L'obiettivo del modello è quindi trovare una configurazione dei pesi che minimizzi la loss function.

2.3 Energia libera, funzione di partizione e replica trick

Una volta chiarito il problema dell'apprendimento supervisionato e descritto il modello attraverso cui lo student cerca di avvicinarsi alle etichette fornite dal teacher, è naturale chiedersi come valutare, da un punto di vista teorico, le capacità di generalizzazione del sistema.
In questo contesto, gli strumenti della meccanica statistica si rivelano particolarmente utili. Un concetto chiave che emerge è quello di **energia libera**, una quantità che gioca un ruolo fondamentale sia nella fisica dei sistemi disordinati sia nell'analisi teorica dei modelli di apprendimento.
In **termodinamica**, l'energia libera rappresenta la differenza tra l'energia interna di un sistema e il prodotto tra temperatura ed entropia:

$$F = U - TS$$

ed è collegata alla quantità di lavoro utile che un sistema può svolgere.
Nell'**apprendimento automatico**, l'energia libera misura l'equilibrio tra l'adattamento ai dati (energia) e la complessità del modello (entropia), indicato con $(\Omega, \mathcal{F}, P)$ lo spazio di probabilità, vale:

$$F[P] = \mathbb{E}_P[H] - kTS[P]$$

H è un'Hamiltoniana fissata, che pensiamo come una variabile aleatoria associata allo spazio campionario Ω, $k \in \mathbb{R}^+$ è la costante di Boltzmann e T è la temperatura.
Minimizzare l'energia libera equivale, nel contesto dell'apprendimento automatico, a minimizzare la funzione di costo (loss) sui dati osservati, bilanciando al contempo la complessità del modello, per ottenere una buona generalizzazione. Questo approccio si collega anche a:

Teorema 2.3.1 (Principio variazionale)**.**

$$\inf_P F[P] = -\frac{1}{\beta} \log \mathcal{Z}(\beta)$$

dove $\mathcal{Z}(\beta) = \sum_{i\in\Omega} e^{-\beta h_i}$ *e prende il nome di* ***funzione di partizione****. Si dimostra che il minimo viene raggiunto nella distribuzione di Boltzmann-Gibbs.*

In natura i sistemi fisici tendono a minimizzare la loro energia libera, quindi computare la stessa ci permette di avere tutte le informazioni sul sistema di interesse.
Il nostro obiettivo è quindi calcolare l'energia libera nel modello teacher-student, che coerentemente con quanto visto si traduce nel voler calcolare:

$$\mathcal{F} = \lim_{N\to+\infty} \frac{1}{N} \mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M} [\log \mathcal{Z}].$$

Nell'espressione precedente compare una media rispetto al dataset $\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M$, che risulta necessaria poiché i dati sono considerati come variabili aleatorie, la cui distribuzione è nota nel modello; di conseguenza, anche l'energia libera diventa una variabile aleatoria, essendo funzione dei dati. Per ottenere una descrizione generale del comportamento del sistema, indipendente dalla specifica realizzazione dei dati, è fondamentale considerare il valore medio rispetto alla loro distribuzione. Questa media consente di caratterizzare il comportamento tipico del sistema, coerentemente con la proprietà di self-averaging dell'energia libera (tendenza di una variabile aleatoria a coincidere con il suo valore medio): nel limite termodinamico, le fluttuazioni rispetto alla media tendono a scomparire.
$\mathcal{Z}$ indica la funzione di partizione, definita come la somma su tutte le possibili configurazioni degli accoppiamenti sinaptici del peso di Boltzmann-Gibbs, in altre parole è il posterior descritto nel framework bayesiano:

$$\mathcal{Z} = \int d\mathbf{W}\ P(\mathbf{W}) \prod_{\mu=1}^{M} P_{\text{out}} \left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i X_i^\mu \right)$$

Nel limite precendente N indica la dimensione del sistema, e in particolare consideriamo il cosiddetto **limite termodinamico** $N \to +\infty$, che porterà all'emergere di un comportamento non intuitivo del sistema, tipico dell'alta dimensionalità.

Presa in generale una funzione $\Omega(\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M)$ vale :

$$\Phi_Q = \frac{1}{N}\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\log(\Omega(\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M)]$$
$$\Phi_A = \frac{1}{N}\log \mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[(\Omega(\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M)]$$

dove Φ_Q prende il nome di "media quenched" e Φ_A quello di "media annealed". Anche se Φ_A è indubbiamente più facile da calcolare, le due medie non coincidono; infatti, dal teorema di Jensen si ha: $\Phi_A \geq \Phi_Q$, allora Φ_A fornisce solo un limite superiore di Φ_Q, e descrive il comportamento di eventi rari o atipici, non quello tipico del sistema. Per questo motivo, nella nostra analisi è fondamentale calcolare Φ_Q, in quanto è questa quantità a descrivere il comportamento medio tipico del modello rispetto alla distribuzione dei dati. Poiché la valutazione diretta di Φ_Q è analiticamente complessa, si fa uso del **replica trick**, una tecnica sviluppata nella fisica dei sistemi disordinati, in particolare nello studio degli spin glass, per trattare la difficoltà analitica posta dalla media del logaritmo della funzione di partizione, come accade appunto nella definizione Φ_Q. Questa tecnica sfrutta la seguente identità formale:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\log \mathcal{Z}] = \lim_{n\to 0}\frac{1}{n}\log\left(\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n]\right) \tag{2.2}$$

La funzione di partizione replicata è quindi:

$$\mathcal{Z}^n = \int \prod_{a=1}^n d\mathbf{W}^a \; P(\mathbf{W}^a) \prod_{a=1}^n \prod_{\mu=1}^M P_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{N}}\sum_{i=1}^N W_i^a X_i^\mu\right) \tag{2.3}$$

dove abbiamo indicato con n il numero di repliche, con M il numero di esempi e con N la dimensione del problema.
Il metodo delle repliche, sebbene potente, non è matematicamente rigoroso nel suo complesso: la quantità $\langle \mathcal{Z}^n \rangle$ che andremo a calcolare ha senso, in prima istanza, solo per $n \in \mathbb{N}$, poiché rappresenta la media su n repliche indipendenti dello stesso sistema. Tuttavia, per procedere nel contesto del *replica trick*, estenderemo formalmente questa espressione a valori continui

di n, fino ad assumere il limite $n \to 0$, pur consapevoli della mancanza di rigore matematico di tale passaggio.

Capitolo 3

Apprendimento in un modello Teacher-Student

In questo capitolo sviluppiamo l'analisi teorica del modello **teacher–student** in alta dimensione, con l'obiettivo di legare in modo esplicito l'errore di generalizzazione alle quantità d'ordine che emergono dalla descrizione replica–symmetry (RS).
La soluzione della (2.1) può essere espressa come la media della seguente misura di Gibbs:

$$\pi_\beta(\mathbf{W}|\{\mathbf{X}^\mu, y^\mu\}) = \frac{1}{\mathcal{Z}} e^{-\beta\left[\sum_{\mu=1}^{M}(\mathcal{L}(y^\mu, \mathbf{X}^\mu\mathbf{W}) + \frac{\lambda}{2}||\mathbf{W}||_2^2)\right]}$$

per $\beta \to \infty$. A questo punto si applica il **replica trick**, esso afferma che la distribuzione della densità di energia libera $log\mathcal{Z}$ si concentra, nel limite ad alta dimensione, attorno a un valore che dipende solo dalla funzione di partizione replicata media. È interessante notare che $\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^{M}}[\mathcal{Z}^n]$ può essere calcolata esplicitamente per $n \in \mathbb{N}$. Questo valore atteso risulterà essere una funzione di quantità scalari, che rappresentano quantità misurabili da monitorare durante l'apprendimento, come la norma dei pesi dello student o l'allineamento tra i pesi dello student e quelli del teacher. L'energia libera risulterà essere: $\mathcal{F} = \underset{m,v,q,\hat{m},\hat{v},\hat{q}}{\text{extr}} \phi(m, r, q, \hat{m}, \hat{r}, \hat{q})$.
I parametri che minimizzano l'espressione sono le soluzioni delle equazioni di saddle point.

Nell'ultima parte del capitolo viene presentato l'errore di generalizzazione $\mathcal{E}_{gen} = \mathbb{E}_{\{\mathbf{X}^{new}, y^{new}, \hat{y}^{new}\}}[(y^{new} - \hat{y}^{new})^2]$, utile per quantificare quanto la rete sia in grado di generalizzare ad un nuovo set di esempi, quanto ha imparato sul training set.
Qui di seguito, discuterò gli step principali del calcolo mentre tutti i dettagli si trovano in appendice A.

3.1 Calcolo dell'energia libera

Coerentemente con quanto spiegato nel capitolo precedente, ci preoccupiamo di calcolare l'energia libera nel modello teacher-student, utilizzando l'identità formale (2.2), in cui la (2.3) è la funzione di partizione replicata.
Esaminando il valore atteso di $\mathcal{Z}^n$ sul dataset $D = \{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M$, assumendo che gli input $\{\mathbf{X}^\mu\}_{\mu=1}^M$ siano variabili aleatorie indipendenti e identicamente distribuite, tali che $\mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}[\mathbf{X}^\mu] = 0$ e $\mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}[(\mathbf{X}^\mu)^2] = 1$, possiamo quindi scrivere:

$$
\begin{aligned}
&\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \\
&= \mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M}\left[\int \prod_{a=1}^n d\mathbf{W}^a \; P(\mathbf{W}^a) \prod_{a=1}^n \prod_{\mu=1}^M P_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i^a X_i^\mu\right)\right] = \\
&= \mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}\left[\prod_{\mu=1}^M \int dy^\mu \; P(y^\mu) \int \prod_{a=1}^n d\mathbf{W}^a \; P(\mathbf{W}^a) \prod_{a=1}^n P_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i^a X_i^\mu\right)\right]
\end{aligned}
$$

Dato che:

$$
y^\mu \sim P_{out}^*\left(y^\mu \,|\, \mathbf{X}^\mu, \mathbf{W}^*\right) = P_{out}^*\left(y^\mu \,\middle|\, \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i^* X_i^\mu\right) \tag{3.1}
$$

e:

$$
P(y^\mu) = \int d\mathbf{W}^* \; P_{\text{out}}^*\left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^N W_i^* X_i^\mu\right) P(\mathbf{W}^*). \tag{3.2}
$$

Possiamo quindi considerare il teacher come una replica aggiuntiva, ovvero: dove:

$$P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}W_i^aX_i^\mu\right)=\begin{cases}P^*_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}W_i^*X_i^\mu\right) & a=0\\ P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}W_i^aX_i^\mu\right) & a>0\end{cases}$$

e:

$$P(\mathbf{W}^a)=\begin{cases}P^*(\mathbf{W}^*) & a=0 \quad \text{prior del teacher}\\ P(\mathbf{W}^a) & a>0\end{cases}$$

Secondo questa notazione possiamo ridurre l'espressione del valore atteso studiato a:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n]=\mathbb{E}_{\{X^\mu\}_{\mu=1}^M}\left[\int\prod_{\mu=1}^{M}dy^\mu\int\prod_{a=0}^{n}d\mathbf{W}^a\,P(\mathbf{W}^a)\times\right.$$
$$\left.\times\prod_{\mu=1}^{M}\prod_{a=0}^{n}P_{\text{out}}\left(y^\mu\,\middle|\,\frac{1}{\sqrt{N}}\sum_{i=1}^{N}W_i^aX_i^\mu\right)\right]$$

Effettuando un cambiamento di variabile $\lambda_\mu^a=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}W_i^aX_i^\mu$, si ha:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n]=\int\prod_{\mu=1}^{M}dy^\mu\int\prod_{\mu=1}^{M}\prod_{a=0}^{n}d\lambda_\mu^a\int\prod_{a=0}^{n}d\mathbf{W}^a\,P(\mathbf{W}^a)\times$$
$$\times\prod_{\mu=1}^{M}\prod_{a=0}^{n}P_{\text{out}}(y^\mu|\lambda_\mu^a)\prod_{\mu=1}^{M}\mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}\left[\prod_{a=0}^{n}\delta\left(\lambda_\mu^a-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}W_i^aX_i^\mu\right)\right]$$

Nell'espressione precedente si è inoltre applicata la definizione di δ di Dirac, per la quale:

$$f(x)=\int d\lambda f(\lambda)\delta(x-\lambda). \tag{3.3}$$

In accordo con la definizione di trasformata di Fourier otteniamo:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n]=\int\prod_{\mu=1}^{M}dy^\mu\int\prod_{\mu=1}^{M}\prod_{a=0}^{n}\frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi}\exp\left(i\sum_{a=0}^{n}\hat{\lambda}_\mu^a\lambda_\mu^a\right)$$
$$\times\int\prod_{a=0}^{n}d\mathbf{W}^a\,P(\mathbf{W}^a)\prod_{\mu=1}^{M}P_{\text{out}}(y^\mu|\lambda_\mu^a)\prod_{\mu=1}^{M}\mathbb{E}_{\mathbf{X}^\mu}\left[\exp\left(-\frac{i}{\sqrt{N}}\sum_{i=1}^{N}X_i^\mu\sum_{a=0}^{n}\hat{\lambda}_\mu^aW_i^a\right)\right]$$

Possiamo calcolare il valore atteso sulle $\{\mathbf{X}^\mu\}_{\mu=1}^M$ dell'esponenziale complesso. Infatti, essendo le $\{\mathbf{X}^\mu\}_{\mu=1}^M$ variabili Gaussiane, tale quantità si riduce ad un integrale Gaussiano con parte lineare complessa, la cui soluzione è nota:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int \prod_{\mu=1}^M dy^\mu \int \prod_{\mu=1}^M \prod_{a=0}^n \frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi} \exp\left(i\sum_{a=0}^n \hat{\lambda}_\mu^a \lambda_\mu^a\right) \quad \times$$
$$\times \int \prod_{a=0}^n d\mathbf{W}^a\, P(\mathbf{W}^a) \prod_{\mu=1}^M P_{\text{out}}(y^\mu|\lambda_\mu^a) \prod_{\mu=1}^M \exp\left(-\frac{1}{2}\sum_{a,b=0}^n \hat{\lambda}_\mu^a \hat{\lambda}_\mu^b \frac{1}{N}\sum_{i=1}^N W_i^a W_i^b\right)$$

A questo punto introduciamo la definizione di overlap tra le repliche del sistema:

$$q_{ab} = \frac{1}{N}\sum_{i=1}^N W_i^a W_i^b$$

Introducendo questa definizione di overlap per mezzo delle delta di Dirac e la loro corrispondente trasformata di Fourier, otteniamo:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int \prod_{\mu=1}^M dy^\mu \int \prod_{\mu=1}^M \prod_{a=0}^n \frac{d\lambda_\mu^a \hat{\lambda}_\mu^a}{2\pi} \exp\left(i\sum_{a=0}^n \hat{\lambda}_\mu^a \lambda_\mu^a\right) \quad \times$$
$$\times \int \prod_{a=0}^n d\mathbf{W}^a\, P(\mathbf{W}^a) \prod_{\mu=1}^M \prod_{a=0}^n P_{\text{out}}\left(y^\mu|\lambda_\mu^a\right) \int \prod_{a\leq b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi} \quad \times$$
$$\times \exp\left(-N\sum_{a\leq b} i\hat{q}_{ab}q_{ab} + i\sum_{a\leq b}\hat{q}_{ab}\sum_{i=1}^N W_i^a W_i^b\right) \prod_{\mu=1}^M \exp\left(-\frac{1}{2}\sum_{a,b=0}^n \hat{\lambda}_\mu^a \hat{\lambda}_\mu^b q_{ab}\right)$$

Sfruttando l'ipotesi che gli input siano indipendenti e identicamente distribuiti (i.i.d.), è possibile fattorizzare sull'indice degli esempi μ. Questo implica che ogni esempio è statisticamente equivalente agli altri e può essere trattato in modo simmetrico, senza che vi sia la necessità di mantenere un'etichetta distinta. Inoltre, nella sommatoria l'indice i rappresenta le componenti dei vettori dei pesi sinaptici, ed è fatto variare da 1 ad N, dove N è la dimensione

del problema.

$$\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int \prod_{a\leq b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi} \times \\ \times \exp\left(-N\sum_{a\leq b} i\hat{q}_{ab}q_{ab} + NG_S(\{\hat{q}_{ab}\}) + \alpha N G_E(\{q_{ab}\})\right) \tag{3.4}$$

dove:

- $\Psi(\{q_{ab}\}, \{\hat{q}_{ab}\})$ **termine di traccia**:

$$\Psi(\{q_{ab}\}, \{\hat{q}_{ab}\}) = -\sum_{a\leq b} i\hat{q}_{ab}q_{ab}$$

  raccoglie l'integrazione su tutte le configurazioni accessibili del sistema, ovvero rappresenta la media sull'insieme delle variabili aleatorie.

- $G_S(\{\hat{q}_{ab})\}$ **termine entropico**:

$$G_S(\{\hat{q}_{ab})\}) = \log \int \prod_{a=0}^{n} dW^a\, P(W^a) \exp\left(\sum_{a\leq b} i\hat{q}_{ab}W^aW^b\right)$$

  misura la quantità di configurazioni microscopiche compatibili con un certo stato macroscopico: non dipende dalla loss function ed è anzi funzione delle variabili coniugate.

- $G_E(\{q_{ab}\})$ **termine energetico**:

$$G_E(\{q_{ab}\}) = \log \int dy \int \prod_{a=0}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \exp\left(i\sum_{a=0}^{n} \hat{\lambda}^a\lambda^a\right) \times \\ \times \prod_{a=0}^{n} P_{\text{out}}(y|\lambda^a) \exp\left(-\frac{1}{2}\sum_{a,b=0}^{n} \hat{\lambda}^a\hat{\lambda}^b q_{ab}\right)$$

  deriva dall'interazione tra i pesi dello student e i dati osservati; si chiama così in quanto è il termine che contiene l'informazione sull'energia del sistema che, nei problemi di machine learning, è rappresentato dalla loss function.

Si considera un numero di dati pari a $M := \alpha N$, questo perché ci si aspetta che il numero di dati scali linearmente con N, in quanto bisogna individuare la giusta configurazione tale per cui, date N connessioni sinaptiche, la predizione della rete coincide con l'etichetta per ogni esempio $\mu = 1, ..., M$. In altre parole, dati M vincoli si devono scegliere N sinapsi, ragion per cui M deve essere almeno proporzionale al numero di sinapsi, ecco perché $M = \alpha N$. Insieme, questi termini contribuiscono a descrivere l'equilibrio statistico del modello.

3.2 Replica Symmetric

Il calcolo procede con le assunzioni di replica symmetric (RS): si assume cioè che tutte le repliche siano statisticamente equivalenti e interagiscano tra loro nello stesso modo, senza alcuna differenza tra una e l'altra:

$$
\begin{aligned}
q_{00} &= r_0 & i\hat{q}_{00} &= \hat{r}_0 \\
q_{aa} &= r & i\hat{q}_{aa} &= -\frac{1}{2}\hat{r} \\
q_{a0} &= m & i\hat{q}_{a0} &= \hat{m} \\
q_{ab} &= q & i\hat{q}_{ab} &= \hat{q} \quad (a \neq b,\ a, b \geq 1)
\end{aligned}
$$

La precedente assunzione risulta ragionevole nel sistema considerato, in quanto la replica consiste semplicemente nella creazione di n copie distinte e indipendenti del medesimo modello. In assenza di ulteriori vincoli o interazioni specifiche, non vi è alcuna ragione a priori per cui una replica debba essere trattata in modo diverso dalle altre. Tuttavia, è importante sottolineare che questa simmetria non è garantita in generale. Esistono infatti modelli nei quali si verifica una rottura spontanea della simmetria tra repliche; un esempio classico è rappresentato dal percettrone binario, in cui la struttura della funzione di costo induce la selezione di sottoinsiemi di repliche preferenziali [10]. Nel caso specifico della regressione lineare mediata da una rete neurale

a singolo layer (scenario teacher-student), grazie alla convessità della funzione di costo (tipicamente loss quadratica) la simmetria tra le repliche risulta conservata.
In accordo con le assunzioni di RS possiamo riscrivere i termini della (3.4) come segue:

- **Termine di traccia**

$$\psi(\{q_{ab}\},\{\hat{q}_{ab}\}) = -r_0\hat{r}_0 + \frac{1}{2}nr\hat{r} - nm\hat{m} - \frac{1}{2}n(n-1)q\hat{q} \tag{3.5}$$

- **Termine entropico**

$$G_S(\{\hat{q}_{ab}\}) = \log \int \mathcal{D}z \int d\boldsymbol{W}^0\, P(\boldsymbol{W}^0)\, \exp\left(\hat{r}_0(\boldsymbol{W}^0)^2\right) \times \\ \times \left(\int d\boldsymbol{W}\, P(\boldsymbol{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \hat{m}\boldsymbol{W}^0\boldsymbol{W} + \sqrt{\hat{q}}z\mathbf{W}\right)\right)^n \tag{3.6}$$

- **Termine energetico**

$$G_E(\{q_{ab}\}) = \log \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi}\, \exp\left(-\frac{1}{2}r_0(\hat{\lambda}^0)^2 + i\hat{\lambda}^0\lambda^0\right) \quad \times \\ \times P_{\text{out}}(y|\lambda^0)\left(\int \frac{d\lambda d\hat{\lambda}}{2\pi}\, \exp\left(i\hat{\lambda}\lambda\right)\, P_{\text{out}}(y|\lambda) \times \right. \\ \left. \times \exp\left(-\frac{1}{2}(r-q)\left(\hat{\lambda}\right)^2 - m\hat{\lambda}\hat{\lambda}^0 + i\sqrt{q}z\hat{\lambda}\right)\right)^n \tag{3.7}$$

È importante sottolineare che, in questa fase, siamo riusciti a fattorizzare il termine elevato a n, ottenendo una forma che semplifica notevolmente i calcoli successivi. Infatti, l'ansatz Replica Simmetrico è solo una delle possibili scelte della continuazione analitica per $n \to 0$. Se questo ansatz non risulta corretto, allora bisogna iniziare a rompere la simmetria delle repliche a più livelli: si parla infatti di one-step replica symmetry breaking (1-RSB), 2-RSB,..., K-RSB. Questo risultato non è affatto scontato nella teoria delle

repliche, dove spesso si è costretti a eseguire nuove somme sulle repliche; un caso è nuovamente quello del percettrone binario, ovvero una rete neurale a singolo layer con pesi sinaptici che possono assumere come valori solo $+1, 1$ e non un continuo di valori. Tale rete cerca poi di risolvere un problema di classificazione binario, ovvero cerca di assegnare un dato di esempi a due classi distinte [10].

3.3 Limite $n \to 0$

Consideriamo lo sviluppo dei termini di traccia, entropico ed energetico approssimati fino all'ordine $O(n)$. I calcoli si trovano in maggiore dettaglio nell'Appendice A; si riportano di seguito i risultati:

- **Termine di traccia**

$$\psi(\{q_{ab}\}, \{\hat{q}_{ab}\}) \simeq -r_0\hat{r}_0 + \frac{1}{2}nr\hat{r} - nm\hat{m} + \frac{1}{2}nq\hat{q} \tag{3.8}$$

- **Termine entropico**
  Attraverso il cambio di varibile $z \to z - \frac{\hat{m}}{\sqrt{\hat{q}}}\mathbf{W}^0$, possiamo separare il teacher dallo student ottenendo dunque:

$$\begin{aligned} G_S(\{\hat{q}_{ab}\}) = \log \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right) + \\ + n \frac{\int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\left(\frac{\hat{m}^2}{\hat{q}} - 2\hat{r}_0\right)(\boldsymbol{W^0})^2\right)}{\int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right)} \times \\ \times \log\left(\int dW\, P(W) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})W^2 + \sqrt{\hat{q}}zW\right)\right) \end{aligned} \tag{3.9}$$

- **Termine energetico**
  Separiamo il teacher dallo student con il seguente cambio di variabile:

$$z \to z + \frac{m\hat{\lambda}^0}{i\sqrt{q}}$$

e integriamo sia su $\hat{\lambda}^0$ che su $\hat{\lambda}$, ottenendo:

$$G_E(\{q_{ab}\}) = \log \frac{1}{\sqrt{r_0 - \frac{m^2}{q}}} \int \mathcal{D}z \int dy \int d\lambda^0 \exp\left(-\frac{1}{2} \frac{\left(\lambda^0 + \frac{m}{\sqrt{q}} z\right)^2}{r_0 - \frac{m^2}{q}} \right) \times$$
$$\times P_{\text{out}}(y|\lambda^0) + n \frac{1}{\sqrt{r_0 - \frac{m^2}{q}}} \int \mathcal{D}z \int dy \int d\lambda^0 \exp\left(-\frac{1}{2} \frac{\left(\lambda^0 + \frac{m}{\sqrt{q}} z\right)^2}{r_0 - \frac{m^2}{q}} \right) \times$$
$$\times P^*_{\text{out}}(y|\lambda^0) \log \frac{1}{\sqrt{r-q}} \int d\lambda \, \exp\left(-\frac{1}{2} \frac{(\lambda + \sqrt{q} z)^2}{r-q} \right) P_{\text{out}}(y|\lambda)$$

Per semplificare questa espressione, eseguiamo il seguente cambio di variabili:

$$\lambda^0 \to \frac{\lambda^0 + \frac{m}{\sqrt{q}} z}{\sqrt{r_0 - \frac{m^2}{q}}} \qquad \text{e} \qquad \lambda \to \frac{\lambda + \sqrt{q} z}{\sqrt{r-q}}$$

Quindi otteniamo:

$$G_E(\{q_{ab}\}) = n \int \mathcal{D}z \int dy \int D\lambda^0 \, P^*_{\text{out}}\left(y \middle| \lambda^0 \sqrt{r_0 - \frac{m^2}{q}} + \frac{m}{\sqrt{q}} z \right) \times$$
$$\times \log \left(\int D\lambda \, P_{\text{out}}\left(y \middle| \lambda \sqrt{r-q} + \sqrt{q} z \right) \right) \tag{3.10}$$

Combinando traccia, termine entropico ed energetico, otteniamo la seguente espressione:

$$\begin{aligned}\phi = &-r_0\hat{r}_0 + \frac{1}{2}nr\hat{r} - nm\hat{m} + \frac{1}{2}nq\hat{q} + \log\int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right) + \\ &+ n\frac{\int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\left(\frac{\hat{m}^2}{\hat{q}} - 2\hat{r}_0\right)(\boldsymbol{W^0})^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W^0}z\right)}{\int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right)} \times \\ &\times \log\left(\int dW\, P(W) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})W^2 + \sqrt{\hat{q}}zW\right)\right) + \\ &+ n\int \mathcal{D}z \int dy \int D\lambda^0\, P^*_{\text{out}}\left(y\middle|\lambda^0\sqrt{r_0 - \frac{m^2}{q}} + \frac{m}{\sqrt{q}}z\right) \times \\ &\times \log\left(\int D\lambda\, P_{\text{out}}\left(y\middle|\lambda\sqrt{r-q} + \sqrt{q}z\right)\right)\end{aligned}$$

Per evitare divergenze nel limite $n \to 0$, il termine $\mathcal{O}(1)$ in n deve annullarsi, cioè deve valere:

$$-r_0\hat{r}_0 + \log\int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right) = 0$$

Questo è vero se e solo se $\hat{r}_0 = 0$. Quindi abbiamo:

$$\begin{aligned}\phi = &\frac{1}{2}nr\hat{r} - nm\hat{m} + \frac{1}{2}nq\hat{q} + \\ &+ n\int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}(\boldsymbol{W^0})^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W^0}z\right) \times \\ &\times \log\left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \sqrt{\hat{q}}z\mathbf{W}\right)\right) + \\ &+ n\alpha\int \mathcal{D}z \int dy \int D\lambda^0\, P^*_{\text{out}}\left(y\middle|\lambda^0\sqrt{r_0 - \frac{m^2}{q}} + \frac{m}{\sqrt{q}}z\right) \times \\ &\times \log\left(\int D\lambda\, P_{\text{out}}\left(y\middle|\lambda\sqrt{r-q} + \sqrt{q}z\right)\right)\end{aligned}$$

Nell'applicare il replica trick, dividiamo il membro di destra per n, permettendoci di prendere il limite $n \to 0$ e di avere un comportamento non banale:

$$\phi = \frac{1}{2} r\hat{r} - m\hat{m} + \frac{1}{2} q\hat{q} + G_S(\hat{m}, \hat{r}, \hat{q}) + \alpha G_E(m, r, q) \tag{3.11}$$

Ricapitolando, l'obiettivo ultimo del calcolo è determinare l'energia libera del sistema ($\mathcal{F}$), che conseguentemente all'applicazione del replica trick risulta essere:

$$\mathcal{F} = \lim_{N\to\infty} \lim_{n\to 0} \frac{\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] - 1}{Nn} \tag{3.12}$$

Per conti svolti dettagliatamente in Appendice A, in cui si ricorre all'approssimazione per saddle point, si ha che:

$$\begin{aligned} \mathcal{F} &= \lim_{N\to\infty} \lim_{n\to 0} \frac{1}{Nn} \operatorname*{extr}_{m,v,q,\hat{m},\hat{v},\hat{q}} (Nn\phi(m, r, q, \hat{m}, \hat{r}, \hat{q})) = \\ &= \operatorname*{extr}_{m,v,q,\hat{m},\hat{v},\hat{q}} \phi(m, r, q, \hat{m}, \hat{r}, \hat{q}) \end{aligned} \tag{3.13}$$

Nel punto di sella, la condizione di consistenza $\hat{r}_0 = 0$ fissa il valore di r_0 per $n = 0$, vale: $r_0 = \mathbb{E}[(\mathbf{W}^0)^2]$.

3.4 Equazioni di saddle point

Risolviamo l'equazione (3.13) riprendendo i tre termini principali della (3.11), assumiamo che $v = r - q$ e $\hat{v} = \hat{r} + \hat{q}$ e dopo varie semplificazioni (appendice A) otteniamo:

$$\begin{aligned} \phi(m, v, q, \hat{m}, \hat{v}, \hat{q}) &= \frac{1}{2} v(\hat{v} - \hat{q}) + \frac{1}{2} q\hat{v} - m\hat{m} + \\ &+ \frac{1}{2} \frac{\hat{m}^2}{\beta\nu + \hat{v}} + \frac{1}{2} \frac{\hat{q}}{\beta\nu + \hat{v}} - \frac{1}{2} \log(\beta\nu + \hat{v}) + \\ &+ \alpha \left[\log \frac{\sigma}{\sqrt{2\pi(v + \sigma^2)}} - \frac{\beta}{2(v + \sigma^2)} \left[\hat{\sigma}^2 + \left(r_0 - \frac{m^2}{q} \right) + \left(\frac{m - q}{\sqrt{q}} \right)^2 \right] \right] \end{aligned} \tag{3.14}$$

Come riportato formalmente in appendice A dall'equazione (39) all'equazione (48), sviluppando gli integrali gaussiani secondo la definizione si ottiene l'espressione (3.14). Focalizzandoci sul problema di regressione lineare con regolarizzazione $L2$, si ha che la distribuzione a priori dei pesi è:

$$P_{out}(y|\sqrt{r - q}\lambda + \sqrt{q}z) = e^{-\beta \mathcal{L}(y, \hat{y})},$$

dove:

$$\mathcal{L}(y,\hat{y}) = \frac{1}{2\sigma^2}(y - \sqrt{r-q}\lambda + \sqrt{q}z)^2$$

è la loss function intesa, dunque, come una funzione quadratica.
Siamo interessati a studiare il comportamento del nostro sistema al **limite termodinamico**, nel quale $\beta \to \infty$. Per garantire che l'energia libera non diverga né si annulli, è necessario che i termini che vi compaiono siano lineari in β. A questo scopo, effettuiamo un cambio di variabile:

$$\begin{aligned} \hat{v} &\to \beta\hat{v} & v &\to \beta^{-1}v \\ \hat{m} &\to \beta\hat{m} & m &\to m \\ \hat{q} &\to \beta^2\hat{q} & q &\to q \end{aligned} \tag{3.15}$$

Calcolando:

$$\lim_{\beta\to\infty} \frac{1}{\beta}\phi(m,v,q,\hat{m},\hat{v},\hat{q}) \tag{3.16}$$

Otteniamo:

$$\begin{aligned} \mathcal{F}(m,v,q,\hat{m},\hat{v},\hat{q}) := &-\frac{v}{2}\hat{q} + \frac{1}{2}q\hat{v} - m\hat{m} + \frac{1}{2}\frac{\hat{m}^2}{\nu+\hat{v}} + \frac{1}{2}\frac{\hat{q}}{\nu+\hat{v}} - \\ &- \frac{\alpha}{2(v+\sigma^2)}\left[\hat{\sigma}^2 + \left(r_0 - \frac{m^2}{q}\right) + \left(\frac{m-q}{\sqrt{q}}\right)^2\right] \end{aligned} \tag{3.17}$$

Determiniamo i 6 parametri $m, q, v, \hat{m}, \hat{q}, \hat{v}$, imponendo le condizioni di stazionarietà rispetto agli stessi della funzione $\mathcal{F}$:

$$\frac{\partial\mathcal{F}}{\partial\hat{q}}(m,v,q,\hat{m},\hat{v},\hat{q}) = 0 \quad \Rightarrow \quad v = \frac{1}{\nu+\hat{v}}$$

$$\frac{\partial\mathcal{F}}{\partial\hat{m}}(m,v,q,\hat{m},\hat{v},\hat{q}) = 0 \quad \Rightarrow \quad m = \frac{\hat{m}}{\nu+\hat{\nu}}$$

$$\frac{\partial\mathcal{F}}{\partial\hat{v}}(m,v,q,\hat{m},\hat{v},\hat{q}) = 0 \quad \Rightarrow \quad q = \frac{\hat{q}+\hat{m}^2}{(\nu+\hat{v})^2}$$

$$\frac{\partial \mathcal{F}}{\partial q}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = 0 \quad \Rightarrow \quad \hat{v} = \alpha \frac{1}{v + \sigma^2}$$

$$\frac{\partial \mathcal{F}}{\partial m}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = 0 \quad \Rightarrow \quad \hat{m} = \alpha \frac{1}{v + \sigma^2}$$

$$\frac{\partial \mathcal{F}}{\partial v}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = 0 \quad \Rightarrow \quad \hat{q} = \alpha \frac{\hat{\sigma}^2 + r_0 + q - 2m}{(v + \sigma^2)^2}$$

Le precedenti identità rappresentano le relazioni tra le variabili d'interesse e forniscono un quadro completo del regime asintotico del modello teacher-student.

3.5 Errore di generalizzazione

Per una rete neurale l'**errore di generalizzazione** è un'osservabile che ci consente di quantificare quanto la rete è riuscita ad imparare correttamente il task e quindi a generalizzare quanto imparato sul training set ad un nuovo set di esempi. Indicata con y^{new} l'architettura del teacher e con $\hat{y}^{new}$ quella del modello di ML considerato, vale:

$$\mathcal{E}_{gen} = \mathbb{E}_{\{\mathbf{X}^{new}, y^{new}, \hat{y}^{new}\}}[(y^{new} - \hat{y}^{new})^2] \tag{3.18}$$

Sostituiamo i valori corrispondenti di y^{new} e $\hat{y}^{new}$

$$\mathbb{E}_{\{\mathbf{W}, \xi, \mathbf{W}^*, \mathbf{X}^{new}\}} \left[\left(\frac{\mathbf{W}^{*T} \mathbf{X}^{\text{new}}}{\sqrt{N}} + \hat{\sigma}\, \xi - \frac{\mathbf{W}^T \mathbf{X}^{\text{new}}}{\sqrt{N}} \right)^2 \right]$$

Calcolando la media rispetto ad X^{new}, ξ, $\mathbf{W}$ e $\mathbf{W}^*$, come riportato in appendice B, otteniamo:

$$\mathcal{E}_{gen} = r_0 + \hat{\sigma}^2 + q - 2m$$

dove gli addendi della somma del membro di destra sono le soluzioni delle equazioni di saddle point per α fissato e hanno una precisa interpretazione fisica, ovvero:

- q è la norma dei pesi che risolvono la regressione lineare;

- m è il teacher-student overlap, ovvero il parametro d'ordine che esprime quanto lo student è riuscito a capire quale sia la regola con la quale le etichette sono state assegnate agli input, in altre parole quanto bene lo student sta riuscendo a capire chi è il teacher. Questo parametro può quindi darci informazioni su quanto bene lo student riuscirà a generalizzare.

Capitolo 4

Analisi empirica e confronto con la teoria nel modello teacher-student

L'obiettivo principale di questa sezione è confrontare i risultati ottenuti empiricamente attraverso simulazioni numeriche con le predizioni teoriche derivate da un sistema di equazioni analizzate nel capitolo precedente.

4.1 Descrizione del codice

Si realizza un codice Python suddiviso in più blocchi funzionali:

- **Generazione dei dati**: si genera un dataset $\mathbf{X} \in \mathbb{R}^{n \times p}$, ciascuno rappresentante un input di dimensione. Gli n vettori di dimensione p sono campionati indipendentemente da una distribuzione gaussiana standard.

- **Campionamento del teacher**: si estraggono i pesi $\mathbf{W}^*$ del teacher da una distribuzione gaussiana con varianza $\sigma^2_{teacher}$ assegnata.

- **Campionamento delle etichette**: si determinano le etichette come somma pesata del dato in input con i pesi del teacher, a cui viene aggiunto del rumore Gaussiano.
- **Apprendimento da parte dello student**: lo student è un regressore lineare con regolarizzazione $L2$ (Ridge regression) che cerca di approssimare le etichette generate dal teacher.
- **Valutazione degli errori**: si calcola sia l'errore sul training set, sia quello di generalizzazione, cioè la distanza media quadratica tra l'output dello student e quello del teacher su un test.

In parallelo, si effettua un calcolo teorico dell'errore di generalizzazione, insieme ad altre quantità correlate come (m, q, v), risolvendo numericamente un sistema di equazioni a punto fisso, derivate dal formalismo statistico del modello. Il sistema viene iterato fino a quando non si raggiunge la convergenza, con una tolleranza stabilita in anticipo. È stato implementato un ciclo su diversi valori del rapporto $\alpha = \frac{N}{p}$, che misura il numero di dati rispetto alla dimensione dell'input. Per ciascun valore di α, è stato ripetuto l'esperimento empirico un numero n di volte, per diverse realizzazioni del training set. Questo ci consente di mediare gli esperimenti numerici rispetto alla distribuzione dei dati e quindi ottenere delle predizioni circa il caso di apprendimento tipico, così come descritto dalla media quenched del conto di replica. In questo modo, gli esperimenti numerici diventano direttamente confrontabili con le predizioni teoriche date dalla teoria delle repliche (chiaramente più n è grande, più la media empirica delle simulazioni numeriche approssima il valore di aspettazione sulla distribuzione dei dati in input).

4.2 Descrizione dei grafici

In questa sezione ci focalizziamo sull'analisi dell'errore di generalizzazione, osservando come esso dipenda da alcuni parametri chiave. Il parametro ν, che controlla la regolarizzazione, influisce direttamente sulla distribuzione a priori dei pesi sinaptici e modifica la forma delle soluzioni delle equazioni di saddle point, alterando così i valori dei parametri d'ordine come m, q, v e, di conseguenza, anche l'errore teorico. Variazioni di ν permettono quindi di studiare come la regolarizzazione possa ridurre fenomeni di overfitting e migliorare la generalizzazione del modello.
Nel grafico riportato nella figura 4.1 è possibile confrontare l'errore empirico di generalizzazione (in rosso) e l'errore teorico (in blu) in funzione di α. All'aumentare di α l'errore decresce, dimostrando che la generalizzazione di un modello è tanto più precisa quanti più sono i dati forniti. Dal grafico riportato in Figura 4.1 si osserva la presenza di un picco della funzione errore per $\alpha = 1$. Possiamo, infatti, considerare la regressione lineare come un sistema di p equazioni lineari (y pari alla predizione della rete per ogni p) che si vogliono risolvere avendo a disposizione N incognite, ovvero i pesi della rete.
Se $\alpha = 1$, vuol dire che $p = N$ cioè il numero di equazioni è esattamente uguale al numero di incognite, quindi questo sistema ammette una ed una sola soluzione.
Invece, se $\alpha < 1$, il sistema lineare è sotto-determinato, questo significa che si hanno più incognite che equazioni e quindi tante soluzioni del sistema, ovvero tante configurazioni possibili dei pesi che risolvono y uguale alla predizione per ogni p nel training set. Infine, quando $\alpha > 1$, il sistema è sovradeterminato, ovvero si hanno più equazioni che incognite e quindi il sistema non ammette soluzione, cioè non esiste una configurazione dei pesi per cui y è uguale alla predizione per ogni p nel training set. Il picco ad $\alpha = 1$ è dovuto all'overfitting, infatti si ha una sola soluzione che fitta correttamente tutti gli esempi del training set; in quanto tale, potrebbe essere molto sensibile

al rumore delle label, portando ad overfitting. Per rimuovere questo picco e quindi migliorare le performance di generalizzazione, dobbiamo aumentare l'intensità della regolarizzazione; quando invece si ha più di una soluzione, la regolarizzazione $L2$ aiuta a selezionare, tra tutte le soluzioni possibili, quelle che hanno norma minima, in modo da evitare problemi di overfitting.

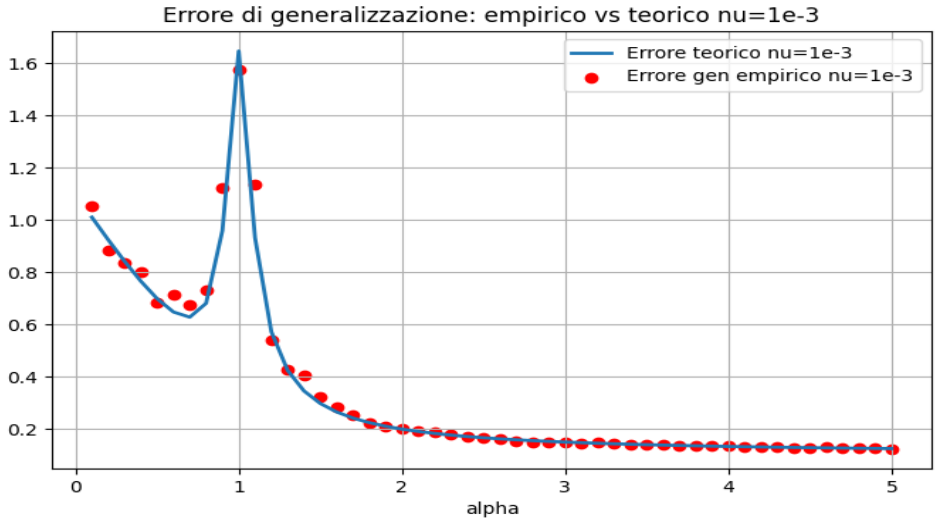


Figura 4.1: Errore empirico VS errore teorico nu=1e-3

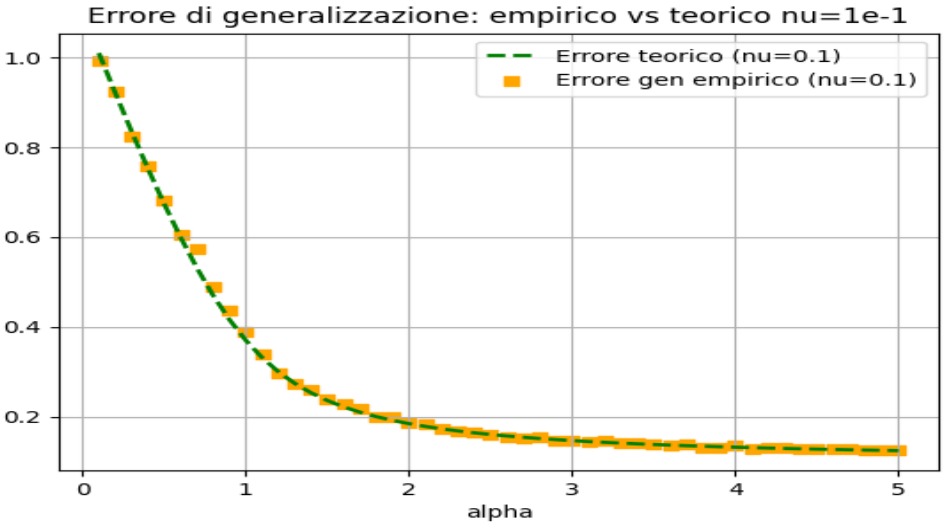


Figura 4.2: Errore empirico VS errore teorico nu=1e-1

Le figure precedenti mostrano inoltre che, l'errore di generalizzazione diminuisce all'aumentare del numero di dati, ma non diventa mai zero. Questo è dovuto al fatto che, avendo rumore nelle label, non si riuscirà mai ad inferire esattamente il teacher (condizione di perfect recovery). L'errore di generalizzazione raggiunge quindi un plateau dettato dal rumore nelle label, non si può fare meglio di così (l'errore satura esattamente all'intensità del rumore).

In accordo con le assunzioni di RS fatte nella sezione 3.2, il parametro **m** misura la correlazione tra i pesi dello student e quelli del teacher. Si nota dal grafico riportato in seguito che, sia a livello teorico che a livello empirico, una volta superata una certa soglia di α, lo student è in grado di replicare in maniera accurata quanto fatto dal teacher:

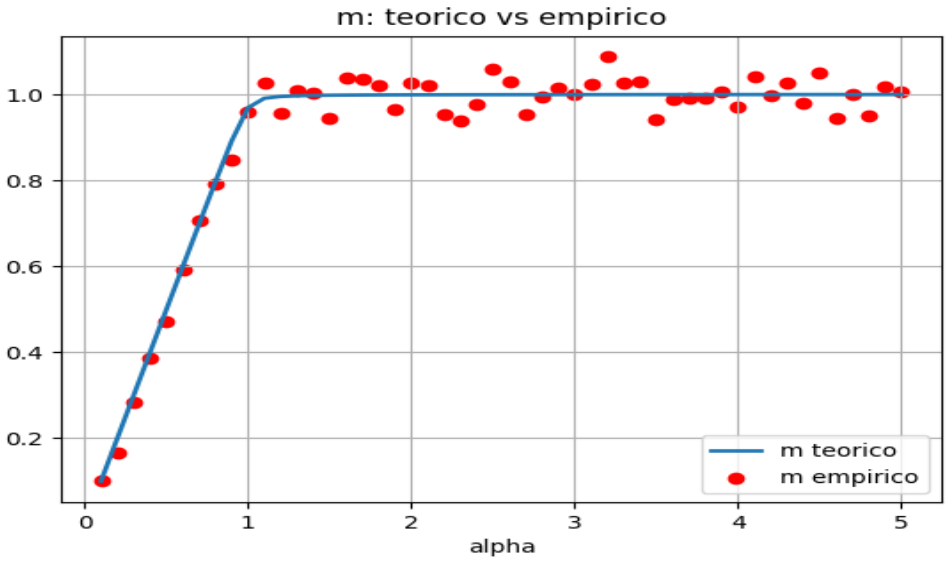


Figura 4.3: m empirico VS m teorico

La figura 4.4 rappresenta l'andamento empirico e teorico di q al variare di α. Notiamo in particolare che per valori piccoli di α la funzione cresce rapidamente, fino a raggiungere un picco per $\alpha \approx 1$. Dopo tale soglia, q decresce gradualmente fino a stabilizzarsi: il modello riesce a rappresentare i dati in modo più efficiente, senza necessità di pesi eccessivamente grandi.

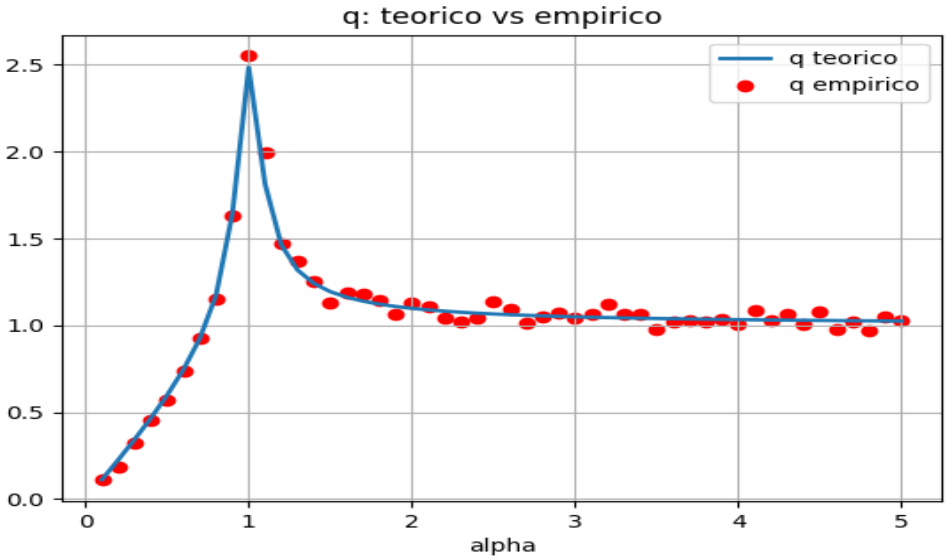


Figura 4.4: q empirico VS q teorico

Conclusioni

In questa tesi, ci siamo concentrati sul modello teacher-student per la regressione lineare, cercando di capire a fondo, sia con la teoria che con la pratica, le capacità di generalizzazione di una rete neurale a singolo layer, quando addestrata a risolvere un task di regressione lineare. L'introduzione del concetto di energia libera ha permesso di descrivere l'equilibrio tra adattamento ai dati osservati e complessità del modello. Attraverso derivate dell'energia libera possiamo calcolare tutte le osservabili fondamentali che descrivono il processo di apprendimento di una rete neurale, in relazione al numero di dati nel training set come, ad esempio, l'overlap tra il teacher e lo student o l'errore di generalizzazione, fondamentale per misurare l'abilità del modello nel generalizzare quanto ha imparato a fare sui dati di training a dati mai visti durante la fase di apprendimento. I risultati teorici sono stati testati attraverso un esperimento computazionale in cui un modello di regressione Ridge apprende dai dati prodotti dal teacher, consentendo un confronto diretto tra prestazioni empiriche e teoriche; l'analisi dei grafici ha mostrato un'ottima corrispondenza tra le stesse, confermando la validità del modello. Come ci aspettavamo è emerso che l'aumentare del numero di esempi, rispetto alla dimensionalità dell'input, porta a un miglioramento significativo della capacità di generalizzazione del modello. L'analisi dei grafici ha inoltre mostrato che l'errore di generalizzazione presenta un picco in corrispondenza di $\alpha = 1$, dove il modello tende a sovradattarsi ai dati del training set (overfitting). Tale fenomeno può essere ridotto introducendo una regolarizzazione adeguata, che stabilizza la soluzione e ne riduce la sensibilità al rumore. Inoltre, l'errore di

generalizzazione si riduce all'aumentare del numero di dati, ma si stabilizza su un valore di plateau determinato dall'intensità del rumore presente nelle label.

Questa tesi, anche se si limita ad un modello lineare semplice, offre un punto di partenza per studiare scenari più complessi, come modelli non lineari o reti neurali profonde; inoltre mostra chiaramente come strumenti presi in prestito dalla fisica possano essere usati con successo per capire e analizzare problemi di apprendimento automatico.

Appendice A

Energia libera

Una rete neurale a singolo strato, data la sequenza $\{\mathbf{X}^\mu\}_{\mu=1}^M$, restituisce l'output y^μ, dato da:

$$y^\mu = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i X_i^\mu \tag{1}$$

dove W sono i pesi sinaptici e ξ il parametro del rumore. Inoltre vale:

$$y^\mu \sim P_{\text{out}}^* \left(y^\mu \,\middle|\, \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i X_i^\mu \right) = \int d\xi \, P(\xi) \, \delta \left(y^\mu - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i X_i^\mu \right) \tag{2}$$

$$y^\mu \sim P_{out}^* \left(y^\mu \,|\, \mathbf{X}^\mu, \mathbf{W}^* \right) = P_{out}^* \left(y^\mu \,\middle|\, \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i^* X_i^\mu \right) \tag{3}$$

L'obiettivo dello studente è inferire l'insieme dei pesi sinaptici $\mathbf{W}$ generati dai y^μ. Lo studente parte da una distribuzione a priori $P(\mathbf{W})$ sul possibile modello, e l'aggiorna alla luce dei dati di addestramento tramite la verosimiglianza $P_{\text{out}}(y^\mu|\mathbf{X}^\mu, \mathbf{W})$, calcolando la distribuzione a posteriori nel seguente modo:

$$P(\mathbf{W}|\mathbf{X}^\mu, y^\mu) \propto \prod_{\mu=1}^{M} P_{\text{out}}(y^\mu|\mathbf{X}^\mu, \mathbf{W}) \, P(\mathbf{W}) = \prod_{\mu=1}^{M} P_{\text{out}} \left(y^\mu \Big| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i X_i^\mu \right) P(\mathbf{W}) \tag{4}$$

Calcolo dell'energia libera mediante il metodo delle repliche

$$\mathcal{F} = \beta\varphi = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M} [\log \mathcal{Z}]$$

dove $\mathcal{Z}$ è la *funzione di partizione*, cioè la somma su tutte le possibili configurazioni dei pesi sinaptici con il peso di Boltzmann:

$$\mathcal{Z} = \int d\mathbf{W}\ P(\mathbf{W}) \prod_{\mu=1}^{M} P_{\text{out}} \left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i X_i^\mu \right) \tag{5}$$

Passo 1: Replichiamo il volume: sfruttiamo il *replica trick* (dato dalla formula $\mathbb{E}[\log \mathcal{Z}] = \lim_{n\to 0} \frac{1}{n} \log \left(\mathbb{E}[\mathcal{Z}^n] \right)$)

$$\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M} [\log \mathcal{Z}] = \lim_{n\to 0} \frac{1}{n} \log \left(\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M} [\mathcal{Z}^n] \right)$$

La funzione di partizione replicata è quindi:

$$\mathcal{Z}^n = \int \prod_{a=1}^{n} d\mathbf{W}^a\ P(\mathbf{W}^a) \prod_{a=1}^{n} \prod_{\mu=1}^{M} P_{\text{out}} \left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i^a X_i^\mu \right) \tag{6}$$

Considerando l'attesa rispetto al dataset $D = \{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M$, possiamo scrivere:

$$\begin{aligned} \mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M} [\mathcal{Z}^n] &= \mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M} \left[\int \prod_{a=1}^{n} d\mathbf{W}^a\ P(\mathbf{W}^a) \prod_{a=1}^{n} \prod_{\mu=1}^{M} P_{\text{out}} \left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i^a X_i^\mu \right) \right] \\ &= \mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M} \left[\prod_{\mu=1}^{M} \int dy^\mu\ P(y^\mu) \int \prod_{a=1}^{n} d\mathbf{W}^a\ P(\mathbf{W}^a) \prod_{a=1}^{n} P_{\text{out}} \left(y^\mu \middle| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i^a X_i^\mu \right) \right] \end{aligned} \tag{7}$$

In accordo con il modello T-S,

$$P(y^\mu) = \int d\mathbf{W}^*\, P(y^\mu, \mathbf{W}^*) \tag{8}$$

$$= \int \mathbf{dW}^*\, P(y^\mu|\mathbf{W}^*)P(\mathbf{W}^*) \tag{9}$$

$$= \int d\mathbf{W}^*\, P^*_{\text{out}}(y^\mu|X^\mu, \mathbf{W}^*)P(W^*) \tag{10}$$

$$= \int d\mathbf{W}^*\, P^*_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i^* X_i^\mu\right) P(\mathbf{W}^*) \tag{11}$$

Sostituendo l'Eq.(11) nell'espressione del volume replicato, possiamo trattare l'insegnante come una replica extra, ovvero:

$$P_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i^a X_i^\mu\right) = \begin{cases} P^*_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i^* X_i^\mu\right) & a = 0 \\ P_{\text{out}}\left(y^\mu \middle| \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i^a X_i^\mu\right) & a > 0 \end{cases}$$

e:

$$P(\mathbf{W}^a) = \begin{cases} P^*(\mathbf{W}^*) & a = 0 \quad \text{prior del teacher} \\ P(\mathbf{W}^a) & a > 0 \end{cases}$$

ottenendo così:

$$\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \mathbb{E}_{\{X^\mu\}_{\mu=1}^M}\left[\int \prod_{\mu=1}^{M} dy^\mu \int \prod_{a=0}^{n} d\mathbf{W}^a\, P(\mathbf{W}^a)\times \right.$$
$$\left. \times \prod_{\mu=1}^{M}\prod_{a=0}^{n} P_{\text{out}}\left(y^\mu \,\middle|\, \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i^a X_i^\mu\right)\right]$$

Passo 2: calcolo del valore atteso sugli input
Assumiamo che gli input $\{\mathbf{X}^\mu\}_{\mu=1}^M$ siano variabili aleatorie indipendenti e identicamente distribuite con:

$$\mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}[\mathbf{X}^\mu] = 0 \quad \text{e} \quad \mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}\left[(\mathbf{X}^\mu)^2\right] = 1$$

Allora:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \mathbb{E}_{\{X^\mu\}_{\mu=1}^M}\left[\int\prod_{\mu=1}^M dy^\mu \int\prod_{a=0}^n d\mathbf{W}^a\, P(\mathbf{W}^a) \times\prod_{\mu=1}^M\prod_{a=0}^n P_{\text{out}}\left(y^\mu\middle|\frac{1}{\sqrt{N}}\sum_{i=1}^N W_i^a X_i^\mu\right)\right] \tag{12}$$

Per poter calcolare la media, definiamo i campi locali:

$$\lambda_\mu^a = \frac{1}{\sqrt{N}}\sum_{i=1}^N W_i^a X_i^\mu$$

Inseriamo questa definizione tramite la delta di Dirac, ottenendo:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int\prod_{\mu=1}^M dy^\mu \int\prod_{\mu=1}^M\prod_{a=0}^n d\lambda_\mu^a \int\prod_{a=0}^n d\mathbf{W}^a\, P(\mathbf{W}^a)\prod_{\mu=1}^M\prod_{a=0}^n P_{\text{out}}(y^\mu|\lambda_\mu^a)\times \times\prod_{\mu=1}^M \mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}\left[\prod_{a=0}^n\delta\left(\lambda_\mu^a - \frac{1}{\sqrt{N}}\sum_{i=1}^N W_i^a X_i^\mu\right)\right] \tag{13}$$

Convertiamo l'espressione tramite la sua trasformata di Fourier:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int\prod_{\mu=1}^M dy^\mu \int\prod_{\mu=1}^M\prod_{a=0}^n \frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi}\exp\left(i\sum_{a=0}^n\hat{\lambda}_\mu^a\lambda_\mu^a\right)\int\prod_{a=0}^n d\mathbf{W}^a\, P(\mathbf{W}^a)\times \times\prod_{\mu=1}^M P_{\text{out}}(y^\mu|\lambda_\mu^a)\prod_{\mu=1}^M\mathbb{E}_{\mathbf{X}^\mu}\left[\exp\left(-\frac{i}{\sqrt{N}}\sum_{i=1}^N X_i^\mu\sum_{a=0}^n\hat{\lambda}_\mu^a W_i^a\right)\right] \tag{14}$$

Concentriamoci sulla media sui pattern di input:

$$
\begin{aligned}
&\mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}\left[\exp\left(\frac{i}{\sqrt{N}}\sum_{i=1}^N X_i^\mu \sum_{a=0}^n \hat{\lambda}_\mu^a W_i^a\right)\right] \sim \\
&\sim \mathbb{E}_{\{\mathbf{X}^\mu\}_{\mu=1}^M}\left[1 - \frac{i}{\sqrt{N}}\sum_{i=1}^N X_i^\mu \sum_{a=0}^n \hat{\lambda}_\mu^a W_i^a \quad - \frac{1}{2N}\sum_{i=1}^N (X_i^\mu)^2 \left(\sum_{a=0}^n \hat{\lambda}_\mu^a W_i^a\right)^2\right] = \\
&= 1 - \frac{1}{2N}\sum_{i=1}^N \mathbf{E}_{X^\mu}\left[(X_i^\mu)^2\right]\left(\sum_{a=0}^n \hat{\lambda}_\mu^a W_i^a\right)^2 \sim \\
&\sim \exp\left(-\frac{1}{2N}\sum_{i=1}^N \left(\sum_{a=0}^n \hat{\lambda}_\mu^a W_i^a\right)^2\right) = \\
&= \exp\left(-\frac{1}{2}\sum_{a,b=0}^n \hat{\lambda}_\mu^a \hat{\lambda}_\mu^b \cdot \frac{1}{N}\sum_{i=1}^N W_i^a W_i^b\right)
\end{aligned}
\tag{15}
$$

Possiamo ora reinserire questa espressione nella funzione di partizione mediata, ottenendo:

$$
\begin{aligned}
\mathbb{E}_{\{\mathbf{X}^\mu, y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int \prod_{\mu=1}^M dy^\mu \int \prod_{\mu=1}^M \prod_{a=0}^n \frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi} \exp\left(i\sum_{a=0}^n \hat{\lambda}_\mu^a \lambda_\mu^a\right) \int \prod_{a=0}^n d\mathbf{W}^a \, P(\mathbf{W}^a) \times \\
\times \prod_{\mu=1}^M P_{\text{out}}(y^\mu|\lambda_\mu^a) \prod_{\mu=1}^M \exp\left(-\frac{1}{2}\sum_{a,b=0}^n \hat{\lambda}_\mu^a \hat{\lambda}_\mu^b \frac{1}{N}\sum_{i=1}^N W_i^a W_i^b\right)
\end{aligned}
\tag{16}
$$

A questo punto, possiamo inserire la definizione dell'overlap:

$$
q_{ab} = \frac{1}{N}\sum_{i=1}^N W_i^a W_i^b
$$

Usando la delta di Dirac:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int \prod_{\mu=1}^{M} dy^\mu \int \prod_{\mu=1}^{M}\prod_{a=0}^{n} \frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi} \exp\left(i\sum_{a=0}^{n}\hat{\lambda}_\mu^a\lambda_\mu^a\right)\times$$
$$\times \int \prod_{a=0}^{n} d\mathbf{W}^a\, P(\mathbf{W}^a) \prod_{\mu=1}^{M}\prod_{a=0}^{n} P_{\text{out}}(y^\mu|\lambda_\mu^a)\times$$
$$\times \int \prod_{a\le b} dq_{ab}\delta\left(\sum_{i=1}^{N} W_i^a W_i^b - q_{ab}N\right) \prod_{\mu=1}^{M} \exp\left(-\frac{1}{2}\sum_{a,b=0}^{n}\hat{\lambda}_\mu^a\hat{\lambda}_\mu^b q_{ab}\right) \tag{17}$$

usando la trasformata di Fourier otteniamo:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int \prod_{\mu=1}^{M} dy^\mu \int \prod_{\mu=1}^{M}\prod_{a=0}^{n} \frac{d\lambda_\mu^a \hat{\lambda}_\mu^a}{2\pi} \exp\left(i\sum_{a=0}^{n}\hat{\lambda}_\mu^a\lambda_\mu^a\right)\times$$
$$\times \int \prod_{a=0}^{n} d\mathbf{W}^a\, P(\mathbf{W}^a) \prod_{\mu=1}^{M}\prod_{a=0}^{n} P_{\text{out}}\left(y^\mu|\lambda_\mu^a\right) \int \prod_{a\le b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi}\times$$
$$\times \exp\left(-N\sum_{a\le b} i\hat{q}_{ab}q_{ab} + i\sum_{a\le b}\hat{q}_{ab}\sum_{i=1}^{N} W_i^a W_i^b\right) \prod_{\mu=1}^{M} \exp\left(-\frac{1}{2}\sum_{a,b=0}^{n}\hat{\lambda}_\mu^a\hat{\lambda}_\mu^b q_{ab}\right) \tag{18}$$

Fattorizzando sia sull'indice i che sull'indice μ. otteniamo:

$$\mathbb{E}_{\{\mathbf{X}^\mu,y^\mu\}_{\mu=1}^M}[\mathcal{Z}^n] = \int \prod_{a\le b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi} \exp\left(-N\sum_{a\le b} i\hat{q}_{ab}q_{ab} + NG_S(\{\hat{q}_{ab}\}) + \alpha N G_E(\{q_{ab}\})\right) \tag{19}$$

dove G_S è la cosiddetta parte entropica ed è data da:

$$G_S(\{\hat{q}_{ab}\}) = \log \int \prod_{a=0}^{n} dW^a\, P(W^a) \exp\left(\sum_{a\le b} i\hat{q}_{ab}W^a W^b\right) \tag{20}$$

e G_E rappresenta invece la cosiddetta parte energetica ed è data da:

$$G_E(\{q_{ab}\}) = \log \int dy \int \prod_{a=0}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \exp\left(i \sum_{a=0}^{n} \hat{\lambda}^a \lambda^a \right) \prod_{a=0}^{n} P_{\text{out}}(y|\lambda^a) \times \\ \times \exp\left(-\frac{1}{2} \sum_{a,b=0}^{n} \hat{\lambda}^a \hat{\lambda}^b q_{ab} \right)$$

Replica Symmetry Assumption

Per procedere, imponiamo l'assunzione di simmetria replica (RS) sulle matrici q_{ab} e $\hat{q}_{ab}$:

$$\begin{aligned} &q_{00} = r_0 && i\hat{q}_{00} = \hat{r}_0 \\ &q_{aa} = r && i\hat{q}_{aa} = -\frac{1}{2}\hat{r} \\ &q_{a0} = m && i\hat{q}_{a0} = \hat{m} \\ &q_{ab} = q && i\hat{q}_{ab} = \hat{q} \quad (a \neq b,\, a, b \geq 1) \end{aligned} \tag{21}$$

possiamo quindi riscrivere la funzione di partizione come segue:

$$\begin{aligned} \psi(\{q_{ab}\}, \{\hat{q}_{ab}\}) &= -\sum_{a \leq b} i\hat{q}_{ab} q_{ab} = \\ &= -\sum_{a=0}^{n} i\hat{q}_{aa} q_{aa} + \frac{1}{2} \sum_{a \neq b} i\hat{q}_{ab} q_{ab} = \\ &= -i\hat{q}_{00} q_{00} - \sum_{a=1}^{n} i\hat{q}_{aa} q_{aa} - \sum_{a=1}^{n} i\hat{q}_{a0} q_{a0} - \frac{1}{2} \sum_{\substack{a \neq b \\ a,b=1}}^{n} i\hat{q}_{ab} q_{ab} = \\ &= -r_0 \hat{r}_0 + \frac{1}{2} n r \hat{r} - n m \hat{m} - \frac{1}{2} n(n-1) q \hat{q} \end{aligned}$$

Termine entropico

$$
\begin{aligned}
G_S(\{\hat{q}_{ab}\}) &= \log \int \prod_{a=0}^{n} d\mathbf{W}^a \, P(\mathbf{W}^a) \exp\left(\sum_{a\leq b} i\hat{q}_{ab} \mathbf{W}^a \mathbf{W}^b \right) = \\
&= \log \int \prod_{a=0}^{n} d\mathbf{W}^a \, P(\mathbf{W}^a) \exp\left(\sum_{a=0}^{n} i\hat{q}_{aa} (\mathbf{W}^a)^2 + \frac{1}{2} \sum_{a\neq b} i\hat{q}_{ab} \mathbf{W}^a \mathbf{W}^b \right) = \\
&= \log \int \prod_{a=0}^{n} d\mathbf{W}^a \, P(\mathbf{W}^a) \exp\Bigg(i\hat{q}_{00} (\mathbf{W^0})^2 + \sum_{a=1}^{n} i\hat{q}_{aa} (\mathbf{W}^a)^2 + \\
&\quad + \sum_{a=1}^{n} i\hat{q}_{a0} \mathbf{W}^a \mathbf{W^0} + \frac{1}{2} \sum_{\substack{a\neq b \\ a,b=1}}^{n} i\hat{q}_{ab} \mathbf{W}^a \mathbf{W}^b \Bigg)
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
&= \log \int \prod_{a=0}^{n} d\mathbf{W}^a \, P(\mathbf{W}^a) \exp\Bigg(\hat{r}_0 (\mathbf{W^0})^2 - \frac{1}{2}\hat{r} \sum_{a=1}^{n} (\mathbf{W}^a)^2 + \hat{m} \sum_{a=1}^{n} \mathbf{W}^a \mathbf{W^0} + \\
&\quad + \frac{1}{2}\hat{q} \sum_{a\neq b} \mathbf{W}^a \mathbf{W}^b \Bigg) = \\
&= \log \int \prod_{a=0}^{n} d\mathbf{W}^a \, P(\mathbf{W}^a) \exp\Bigg(\hat{r}_0 (\mathbf{W^0})^2 - \frac{1}{2}\hat{r} \sum_{a=1}^{n} (\mathbf{W}^a)^2 + \hat{m} \sum_{a=1}^{n} \mathbf{W}^a \mathbf{W^0} + \\
&\quad + \frac{\hat{q}}{2} \left(\left(\sum_{a=1}^{n} \mathbf{W}^a \right)^2 - \sum_{a=1}^{n} (\mathbf{W}^a)^2 \right) \Bigg) = \\
&= \log \int \prod_{a=0}^{n} d\mathbf{W}^a \, P(\mathbf{W}^a) \exp\Bigg(\hat{r}_0 (\mathbf{W^0})^2 - \frac{1}{2}(\hat{r}+\hat{q}) \sum_{a=1}^{n} (\mathbf{W}^a)^2 + \hat{m} \sum_{a=1}^{n} \mathbf{W}^a \mathbf{W}^0 + \\
&\quad + \frac{\hat{q}}{2} \left(\sum_{a=1}^{n} \mathbf{W}^a \right)^2 \Bigg)
\end{aligned}
\tag{23}
$$

Applicando la trasformazione di Hubbard-Stratonovich otteniamo:

$$
\begin{aligned}
G_S(\{\hat{q}_{ab}\}) = \log \int \mathcal{D}z \int \prod_{a=0}^{n} d\mathbf{W}^a \, P(\mathbf{W}^a) \exp\Bigg(& \hat{r}_0 (\mathbf{W^0})^2 - \frac{1}{2}(\hat{r}+\hat{q}) \sum_{a=1}^{n} (\mathbf{W}^a)^2 + \\
& + \hat{m} \sum_{a=1}^{n} \mathbf{W}^a \mathbf{W^0} + \sqrt{\hat{q}} z \sum_{a=1}^{n} \mathbf{W}^a \Bigg)
\end{aligned}
\tag{24}
$$

Dopo la trasformazione, si può fattorizzare sul singolo indice di replica a:

$$G_S(\{\hat{q}_{ab}\}) = \log \int \mathcal{D}z \int d\boldsymbol{W}^0 \, P(\boldsymbol{W}^0) \, \exp\left(\hat{r}_0 (\boldsymbol{W}^0)^2\right) \times \\ \times \left(\int d\boldsymbol{W} \, P(\boldsymbol{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \hat{m}\boldsymbol{W}^0\boldsymbol{W} + \sqrt{\hat{q}} z \mathbf{W} \right) \right)^n \tag{25}$$

—

Termine energetico

$$G_E(\{q_{ab}\}) = \log \int dy \int \prod_{a=0}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \, \exp\left(i \sum_{a=0}^{n} \lambda^a \hat{\lambda}^a \right) \times \prod_{a=0}^{n} P_{\text{out}}(y|\lambda^a) \times \\ \times \exp\left(-\frac{1}{2} \sum_{a,b=0}^{n} \hat{\lambda}^a \hat{\lambda}^b q_{ab} \right) \tag{26}$$

Sviluppo del termine nel caso simmetrico:

$$= \log \int dy \int \prod_{a=0}^{n} \frac{d\lambda^a \hat{\lambda}^a}{2\pi} \exp\left(i \sum_{a=0}^{n} \lambda^a \hat{\lambda}^a \right) \prod_{a=0}^{n} P_{\text{out}}(y|\lambda^a) \times \\ \times \exp\left(-\frac{1}{2}(\hat{\lambda}^0)^2 q_{00} - \sum_{a=1}^{n} \frac{1}{2}(\hat{\lambda}^a)^2 q_{aa} - \sum_{a=1}^{n} \hat{\lambda}^a \hat{\lambda}^0 q_{a0} - \frac{1}{2} \sum_{a \neq b=1}^{n} \hat{\lambda}^a \hat{\lambda}^b q_{ab} \right) = \tag{27}$$

Scrivendo in modo compatto con r_0, r, m, q:

$$= \log \int dy \int \prod_{a=0}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \exp\left(i\hat{\lambda}^0 \lambda^0 + i \sum_{a=0}^{n} \lambda^a \hat{\lambda}^a \right) \prod_{a=0}^{n} P_{\text{out}}(y|\lambda^a) \times \\ \times \exp\left(-\frac{1}{2} r_0 (\hat{\lambda}^0)^2 - \frac{1}{2} r \sum_{a=1}^{n} (\hat{\lambda}^a)^2 - m \sum_{a=1}^{n} \hat{\lambda}^a \hat{\lambda}^0 - \frac{1}{2} q \sum_{a \neq b=1}^{n} \hat{\lambda}^a \hat{\lambda}^b \right) \tag{28}$$

Il che ci porta alla forma compatta finale:

$$= \log \int dy \int \prod_{a=0}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \exp\left(i\hat{\lambda}^0 \lambda^0 + i \sum_{a=0}^{n} \lambda^a \hat{\lambda}^a \right) \prod_{a=0}^{n} P_{\text{out}}(y|\lambda^a) \times \\ \times \exp\left(-\frac{1}{2} r_0 (\hat{\lambda}^0)^2 - \frac{1}{2}(r-q) \sum_{a=1}^{n} (\hat{\lambda}^a)^2 - m \sum_{a=1}^{n} \hat{\lambda}^a \hat{\lambda}^0 - \frac{1}{2} q \left(\sum_{a=1}^{m} \hat{\lambda}^a \right)^2 \right) \tag{29}$$

Applichiamo la trasformazione di Hubbard-Stratonovich per linearizzare il termine quadratico. Questo ci permette di riscrivere il termine energetico come segue:

$$\begin{aligned} G_E(\{q_{ab}\}) = \log \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \exp\left(-\frac{1}{2} r_0 (\hat{\lambda}^0)^2 + i\hat{\lambda}^0\lambda^0\right) P_{\text{out}}(y|\lambda^0) \times \\ \times \int \prod_{a=1}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \exp\left(i \sum_{a=1}^{n} \hat{\lambda}^a \lambda^a\right) \prod_{a=1}^{n} P_{\text{out}}(y|\lambda^a) \times \\ \times \exp\left(-\frac{1}{2}(r-q) \sum_{a=1}^{n} (\hat{\lambda}^a)^2 - m \sum_{a=1}^{n} \hat{\lambda}^a \hat{\lambda}^0 + i\sqrt{q} z \sum_{a=1}^{n} \hat{\lambda}^a\right) \end{aligned} \tag{30}$$

Grazie alla simmetria tra le repliche, possiamo ora fattorizzare rispetto all'indice a:

$$\begin{aligned} G_E(\{q_{ab}\}) = \log \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \exp\left(-\frac{1}{2} r_0 (\hat{\lambda}^0)^2 + i\hat{\lambda}^0\lambda^0\right) P_{\text{out}}(y|\lambda^0) \times \\ \times \left(\int \frac{d\lambda d\hat{\lambda}}{2\pi} \exp\left(i\hat{\lambda}\lambda\right) P_{\text{out}}(y|\lambda) \exp\left(-\frac{1}{2}(r-q)\left(\hat{\lambda}\right)^2 - m\hat{\lambda}\hat{\lambda}^0 + i\sqrt{q} z \hat{\lambda}\right)\right)^n \end{aligned} \tag{31}$$

Prendiamo ora il limite $n \to 0$, considerando solo i termini fino all'ordine $O(n)$, espandendo i tre contributi principali:

Termine di traccia:

$$\begin{aligned} \psi(\{q_{ab}\}, \{\hat{q}_{ab}\}) &= -r_0\hat{r}_0 + \frac{1}{2} n r \hat{r} - n m \hat{m} - \frac{1}{2} n(n-1) q \hat{q} \simeq \\ &\simeq -r_0\hat{r}_0 + \frac{1}{2} n r \hat{r} - n m \hat{m} + \frac{1}{2} n q \hat{q} \end{aligned} \tag{32}$$

Termine entropico:

$$\begin{aligned} G_S(\{\hat{q}_{ab}\}) = \log \int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0 (\boldsymbol{W^0})^2\right) \times \\ \times \left(\int d\boldsymbol{W}\, P(\boldsymbol{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\boldsymbol{W}^2 + \hat{m}\boldsymbol{W^0}\boldsymbol{W} + \sqrt{\hat{q}} z \boldsymbol{W}\right)\right)^n \end{aligned} \tag{33}$$

Applicando il seguente cambio di variabile:

$$z \quad \rightarrow \quad z - \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W^0}$$

disaccopiamo il teacher dallo student, ottenendo:

$$G_S(\{\hat{q}_{ab}\}) = \log \int \frac{dz}{\sqrt{2\pi}} \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0 (\boldsymbol{W^0})^2\right) \times$$
$$\times \exp\left(-\frac{1}{2}\left(z - \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W^0}\right)^2\right) \left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \sqrt{\hat{q}} z \mathbf{W}\right)\right)^n \tag{34}$$

Semplificando i termini esponenziali:

$$G_S(\{\hat{q}_{ab}\}) = \log \int \frac{dz}{\sqrt{2\pi}} \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}\left(\boldsymbol{W^0}\right)^2 + \hat{r}_0\left(\boldsymbol{W^0}\right) + \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W}^0 z\right) \times$$
$$\times \exp\left(n \log\left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \sqrt{\hat{q}} z \mathbf{W}\right)\right)\right) \tag{35}$$

poiché vogliamo prendere il limite $n \to 0$, possiamo espandere l'esponenziale che è lineare in n:

$$\begin{aligned} G_S(\{\hat{q}_{ab}\}) \sim \log \int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}(\boldsymbol{W^0})^2 + \hat{r}_0(\boldsymbol{W^0})^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W}^0 z\right) \times \\ \times \left(1 + n \log\left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \sqrt{\hat{q}} z \mathbf{W}\right)\right)\right) = \\ = \log \int \mathcal{D}z \int d\boldsymbol{W}^0\, P(\boldsymbol{W}^0) \exp\left(-\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}(\boldsymbol{W}^0)^2 - 2\hat{r}_0(\boldsymbol{W}^0)^2\right) + \\ + n \int \mathcal{D}z \int d\boldsymbol{W}^0\, P(\boldsymbol{W}^0) \exp\left(-\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}(\boldsymbol{W}^0)^2 - 2\hat{r}_0(\boldsymbol{W}^0)^2\right) \times \\ \times \log\left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \sqrt{\hat{q}} z \mathbf{W}\right)\right) = \\ = \log(A + nB) \sim \log(A) + n\frac{B}{A} \end{aligned} \tag{36}$$

Focalizziamoci su A:

$$
\begin{aligned}
A &= \int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\left(\frac{\hat{m}^2}{\hat{q}} - 2\hat{r}_0\right)(\boldsymbol{W^0})^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W^0} z\right) = \\
&= \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\left(\frac{\hat{m}^2}{\hat{q}} - 2\hat{r}_0\right)(\boldsymbol{W^0})^2\right) \times \\
&\times \int \frac{dz}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W^0} z\right) = \\
&= \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\left(\frac{\hat{m}^2}{\hat{q}} - 2\hat{r}_0\right)(\boldsymbol{W^0})^2\right) \exp\left(\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}(\boldsymbol{W^0})^2\right) = \\
&= \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right)
\end{aligned}
\tag{37}
$$

quindi:

$$
\begin{aligned}
G_S(\{\hat{q}_{ab}\}) &= \log A + n\frac{B}{A} = \\
&= \log \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right) + \\
&\quad + n \int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\left(\frac{\hat{m}^2}{\hat{q}} - 2\hat{r}_0\right)(\boldsymbol{W^0})^2\right) \times \\
&\quad \times \log\left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \sqrt{\hat{q}} z \mathbf{W}\right)\right) \times \\
&\quad \times \frac{1}{\int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0(\boldsymbol{W^0})^2\right)}
\end{aligned}
$$

Termine energetico:

$$
G_E(\{q_{ab}\}) = \log \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \exp\left(-\frac{1}{2} r_0 (\hat{\lambda}^0)^2 + i\hat{\lambda}^0\lambda^0\right) P^*_{\text{out}}(y|\lambda^0) \times
$$

$$
\times \left(\int \frac{d\lambda d\hat{\lambda}}{2\pi} \exp(i\hat{\lambda}\hat{\lambda}) P_{\text{out}}(y|\lambda) \exp\left(-\frac{1}{2}(r-q)(\hat{\lambda})^2 - m\hat{\lambda}\hat{\lambda}^0 + i\sqrt{q} z \hat{\lambda}\right)\right)^n
$$

dobbiamo separare il teacher dallo student e lo facciamo con il seguente cambio di variabile:

$$z \to z + \frac{m\hat{\lambda}^0}{i\sqrt{q}}$$

$$G_E(\{q_{ab}\}) = \log \int \frac{dz}{\sqrt{2\pi}} \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \exp\left(-\frac{1}{2}\left(z - i\frac{m\hat{\lambda}^0}{\sqrt{q}}\right)^2\right) \times$$
$$\times \exp\left(-\frac{1}{2} r_0 (\hat{\lambda}^0)^2 + i\hat{\lambda}^0\lambda^0\right) P^*_{\text{out}}(y|\lambda^0) \times$$
$$\times \left(\int \frac{d\lambda d\hat{\lambda}}{2\pi} \exp(i\hat{\lambda}\lambda) P_{\text{out}}(y|\lambda) \exp\left(-\frac{1}{2}(r-q)(\hat{\lambda})^2 + i\sqrt{q}z\hat{\lambda}\right)\right)^n =$$
$$= \log \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \exp\left(-\frac{1}{2}\left(r_0 - \frac{m^2}{q}\right)(\hat{\lambda}^0)^2 + i\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)\hat{\lambda}^0\right) \times$$
$$\times P^*_{\text{out}}(y|\lambda^0) \exp\left(n \log\left(\int \frac{d\lambda d\hat{\lambda}}{2\pi} P_{\text{out}}(y|\lambda) \exp\left(-\frac{1}{2}(r-q)(\hat{\lambda})^2 + i(\lambda + \sqrt{q}z)\hat{\lambda}\right)\right)\right)$$

Analogamente a quanto fatto per il termine entropico, possiamo espandere il termine esponenziale:

$$G_E(\{q_{ab}\}) \sim \log \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \exp\left(-\frac{1}{2}\left(r_0 - \frac{m^2}{q}\right)(\hat{\lambda}^0)^2 + \right.$$
$$\left. + i\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)\hat{\lambda}^0\right) P^*_{\text{out}}(y \mid \lambda^0)\left(1 + n \log\left(\int \frac{d\lambda\, d\hat{\lambda}}{2\pi} P_{\text{out}}(y \mid \lambda) \times\right.\right.$$
$$\left.\left. \times \exp\left(-\tfrac{1}{2}(r-q)\hat{\lambda}^2 + i(\lambda + \sqrt{q}z)\hat{\lambda}\right)\right)\right) =$$
$$= \log \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \exp\left(-\frac{1}{2}\left(r_0 - \frac{m^2}{q}\right)(\hat{\lambda}^0)^2 + i\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)\hat{\lambda}^0\right) \times$$
$$\times P^*_{\text{out}}(y \mid \lambda^0) + n \int \mathcal{D}z \int dy \int \frac{d\lambda^0 d\hat{\lambda}^0}{2\pi} \times$$
$$\times \exp\left(-\frac{1}{2}\left(r_0 - \frac{m^2}{q}\right)(\hat{\lambda}^0)^2 + i\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)\hat{\lambda}^0\right) \times$$
$$\times P^*_{\text{out}}(y \mid \lambda^0) \log\left(\int \frac{d\lambda\, d\hat{\lambda}}{2\pi} P_{\text{out}}(y \mid \lambda) \exp\left(-\tfrac{1}{2}(r-q)\hat{\lambda}^2 + i(\lambda + \sqrt{q}z)\hat{\lambda}\right)\right)$$

Nota che possiamo integrare sia su $\hat{\lambda}^0$ che su $\hat{\lambda}$:
- $\hat{\lambda}^0$**:**

$$\int \frac{d\hat{\lambda}^0}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(r_0 - \frac{m^2}{q}\right)(\hat{\lambda}^0)^2 + i\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)\hat{\lambda}^0\right) = \\ = \sqrt{\frac{2\pi}{r_0 - \frac{m^2}{q}}} \exp\left(-\frac{1}{2}\frac{\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)^2}{r_0 - \frac{m^2}{q}}\right)$$

- $\hat{\lambda}$**:**

$$\int \frac{d\hat{\lambda}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(r-q)(\hat{\lambda})^2 + i(\lambda + \sqrt{q}z)\hat{\lambda}\right) = \sqrt{\frac{2\pi}{r-q}} \exp\left(-\frac{1}{2}\frac{(\lambda + \sqrt{q}z)^2}{r-q}\right)$$

Quindi, alla fine otteniamo:

$$G_E(\{q_{ab}\}) = \log\left[\frac{1}{\sqrt{r_0 - \frac{m^2}{q}}} \int \mathcal{D}z \int dy \int d\lambda^0 \exp\left(-\frac{1}{2}\frac{\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)^2}{r_0 - \frac{m^2}{q}}\right) \times \right. \\ \left. \times P_{\text{out}}(y|\lambda^0)\right] + n\frac{1}{\sqrt{r_0 - \frac{m^2}{q}}} \int \mathcal{D}z \int dy \int d\lambda^0 \exp\left(-\frac{1}{2}\frac{\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)^2}{r_0 - \frac{m^2}{q}}\right) \times \\ \times P^*_{\text{out}}(y|\lambda^0) \times \log\left[\frac{1}{\sqrt{r-q}} \int d\lambda \, \exp\left(-\frac{1}{2}\frac{(\lambda + \sqrt{q}z)^2}{r-q}\right) P_{\text{out}}(y|\lambda)\right] = \\ = \log(C + nD)$$

Focalizziamoci su C:

$$C = \frac{1}{\sqrt{\left(r_0 - \frac{m^2}{q}\right)}} \int \mathcal{D}z \int dy \int d\lambda^0 \exp\left(-\frac{1}{2}\frac{\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)^2}{\left(r_0 - \frac{m^2}{q}\right)}\right) P^*_{\text{out}}(y|\lambda^0)$$

Assumiamo che le probabilità siano normalizzate:

$$\int dy\ P^*_{\text{out}}(y|\lambda^0) = 1$$

Quindi:

$$C = \frac{1}{\sqrt{\left(r_0 - \frac{m^2}{q}\right)}} \int \mathcal{D}z \int d\lambda^0 \exp\left(-\frac{1}{2}\frac{\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)^2}{\left(r_0 - \frac{m^2}{q}\right)}\right)$$

Poi, eseguendo il seguente cambio di variabile:

$$\lambda^0 \to \frac{\lambda^0 + \frac{m}{\sqrt{q}}z}{\sqrt{r_0 - \frac{m^2}{q}}}$$

Otteniamo infine:

$$C = \int \mathcal{D}z \int D\lambda^0 = 1$$

La parte energetica è quindi data da:

$$\begin{aligned} G_E(\{q_{ab}\}) = \log(1 + nD) \sim nD = \\ = n\frac{1}{\sqrt{\left(r_0 - \frac{m^2}{q}\right)}} \int \mathcal{D}z \int dy \int d\lambda^0 \exp\left(-\frac{1}{2}\frac{\left(\lambda^0 + \frac{m}{\sqrt{q}}z\right)^2}{\left(r_0 - \frac{m^2}{q}\right)}\right) P^*_{\text{out}}(y|\lambda^0) \times \\ \times \log\left(\frac{1}{\sqrt{r-q}} \int d\lambda\ \exp\left(-\frac{1}{2}\frac{(\lambda + \sqrt{q}z)}{r-q}\right) P_{\text{out}}(y|\lambda)\right) \end{aligned}$$

Per semplificare questa espressione, eseguiamo il seguente cambio di variabili:

$$\lambda \to \frac{\lambda + \sqrt{q}z}{\sqrt{r-q}}$$

Quindi otteniamo:

$$
\begin{aligned}
G_E(\{q_{ab}\}) &= \log(C + nD) \sim \\
&\sim nD = \\
&= n \int \mathcal{D}z \int dy \int D\lambda^0 \, P_{\text{out}} \left(y \middle| \lambda^0 \sqrt{r_0 - \frac{m^2}{q}} - \frac{m}{\sqrt{q}} z \right) \times \\
&\quad \times \log \left(\int D\lambda \, P_{\text{out}} \left(y \middle| \lambda \sqrt{r - q} - \sqrt{q} z \right) \right) = \\
&= n \int \mathcal{D}z \int dy \int D\lambda^0 \, P^*_{\text{out}} \left(y \middle| \lambda^0 \sqrt{r_0 - \frac{m^2}{q}} + \frac{m}{\sqrt{q}} z \right) \times \\
&\quad \times \log \left(\int D\lambda \, P_{\text{out}} \left(y \middle| \lambda \sqrt{r - q} + \sqrt{q} z \right) \right)
\end{aligned}
$$

Riepilogo:

Combinando traccia, termine entropico ed energetico, otteniamo la seguente espressione per l'azione $\phi(\cdot)$:

$$
\begin{aligned}
\phi = &-r_0 \hat{r}_0 + \frac{1}{2} n r \hat{r} - n m \hat{m} + \frac{1}{2} n q \hat{q} + \log \int d\boldsymbol{W^0} \, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0 (\boldsymbol{W^0})^2 \right) + \\
&+ n \frac{\int \mathcal{D}z \int d\boldsymbol{W^0} \, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2} \left(\frac{\hat{m}^2}{\hat{q}} - 2\hat{r}_0 \right) (\boldsymbol{W^0})^2 + \frac{\hat{m}}{\sqrt{\hat{q}}} \boldsymbol{W^0} z \right)}{\int d\boldsymbol{W^0} \, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0 (\boldsymbol{W^0})^2 \right)} \times \\
&\times \log \left(\int d\mathbf{W} \, P(\mathbf{W}) \exp\left(-\frac{1}{2} (\hat{r} + \hat{q}) W^2 + \sqrt{\hat{q}} z \mathbf{W} \right) \right) + \\
&+ n \int \mathcal{D}z \int dy \int D\lambda^0 \, P^*_{\text{out}} \left(y \middle| \lambda^0 \sqrt{r_0 - \frac{m^2}{q}} + \frac{m}{\sqrt{q}} z \right) \times \\
&\times \log \left(\int D\lambda \, P_{\text{out}} \left(y \middle| \lambda \sqrt{r - q} + \sqrt{q} z \right) \right)
\end{aligned}
$$

Per evitare divergenze nel limite $n \to 0$, il termine $\mathcal{O}(1)$ in n deve annullarsi:

$$
-r_0 \hat{r}_0 + \log \int d\boldsymbol{W^0} \, P(\boldsymbol{W^0}) \exp\left(\hat{r}_0 (\boldsymbol{W^0})^2 \right) = 0
$$

Questo è vero se e solo se $\hat{r}_0 = 0$. Quindi abbiamo:

$$\begin{aligned}\phi = &\frac{1}{2}nr\hat{r} - nm\hat{m} + \frac{1}{2}nq\hat{q} + \\ &+ n\int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}(\boldsymbol{W^0})^2 + \frac{\hat{m}}{\sqrt{\hat{q}}}\boldsymbol{W^0}z\right) \times \\ &\times \log\left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}(\hat{r}+\hat{q})\mathbf{W}^2 + \sqrt{\hat{q}}z\mathbf{W}\right)\right) + \\ &+ n\alpha \int \mathcal{D}z \int dy \int D\lambda^0\, P^*_{\text{out}}\left(y\middle|\lambda^0\sqrt{r_0 - \frac{m^2}{q}} + \frac{m}{\sqrt{q}}z\right) \times \\ &\times \log\left(\int D\lambda\, P_{\text{out}}\left(y\middle|\lambda\sqrt{r-q} + \sqrt{q}z\right)\right)\end{aligned}$$

Indicata con $\mathcal{F}$ l'energia libera e mettendo insieme tutte le argomentazioni trattate fino a questo punto, vale:

$$\begin{aligned}\mathcal{F} &= \lim_{N\to\infty}\lim_{n\to 0}\frac{1}{Nn}\underset{m,v,q,\hat{m},\hat{v},\hat{q}}{\text{extr}}(Nn\phi(m,r,q,\hat{m},\hat{r},\hat{q})) = \\ &= \underset{m,v,q,\hat{m},\hat{v},\hat{q}}{\text{extr}}\phi(m,r,q,\hat{m},\hat{r},\hat{q})\end{aligned} \tag{38}$$

Ci occupiamo quindi di risolvere un problema di massimo. Nel punto sella, la condizione di consistenza $\hat{r}_0 = 0$ fissa il valore di r_0 per $n = 0$. Infatti:

$$\left.\frac{\partial\phi}{\partial\hat{r}_0}\right|_{\hat{r}_0=0} = 0 \quad \Rightarrow \quad -r_0^0 + \left.\frac{\partial G_S(n=0)}{\partial\hat{r}_0}\right|_{\hat{r}_0=0} = 0$$

Ci focalizziamo sulla derivata:

$$\left.\frac{\partial G_S(n=0)}{\partial\hat{r}_0}\right|_{\hat{r}_0=0} = \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0})(\boldsymbol{W^0})^2 = \mathbb{E}[(\boldsymbol{W^0})^2]$$

$$\Rightarrow \quad r_0 = \mathbb{E}[(\boldsymbol{W^0})^2]$$

Riduciamo i vari termini:

$$
\begin{aligned}
G_S(\hat{m},\hat{r},\hat{q}) &= \int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\mathbf{W}^{0T}\frac{\hat{m}^2}{\hat{q}}\boldsymbol{W^0} + \mathbf{W}^{0T}\frac{\hat{m}}{\sqrt{\hat{q}}}z\right)\times \\
&\quad \times \log\left(\int d\mathbf{W}\, P(\mathbf{W}) \exp\left(-\frac{1}{2}\mathbf{W}^T(\hat{r}+\hat{q})\mathbf{W} + \mathbf{W}^T\sqrt{\hat{q}}z\right)\right) = \\
&= \int \mathcal{D}z \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(-\frac{1}{2}\mathbf{W}^{0T}\frac{\hat{m}^2}{\hat{q}}\boldsymbol{W^0} + \mathbf{W}^{0T}\frac{\hat{m}}{\sqrt{\hat{q}}}z\right)\times \\
&\quad \times \log\left(\frac{1}{\sqrt{\beta\nu+\hat{v}}}\exp\left(\frac{1}{2}\frac{\hat{q}}{\beta\nu+\hat{v}}z^Tz\right)\right) =
\end{aligned}
\tag{39}
$$

$$
\begin{aligned}
&= \int d\boldsymbol{W^0}\, P(\boldsymbol{W^0}) \exp\left(\frac{1}{2}\frac{\hat{m}^2}{\hat{q}}\mathbf{W}^{0T}\mathbf{W}^0\right)\times \\
&\quad \times \int dz \exp\left(-\frac{1}{2}\left(z-\frac{\hat{m}}{\sqrt{\hat{q}}}\mathbf{W}^0\right)^2\frac{1}{2}\frac{\hat{q}}{\beta\nu+\hat{v}}z^Tz\right) + \log\frac{1}{\sqrt{\beta\nu+\hat{v}}} = \\
&= \int d\mathbf{W}^0 \exp\left(-\frac{1}{2}\mathbf{W}^{0T}\mathbf{W}^0\right)\left[\frac{1}{2}\frac{\hat{q}}{\beta\nu+\hat{V}}\left(\mathbf{W}^{0T}\frac{\hat{m}^2}{\hat{q}}\mathbf{W}^0+1\right)\right] + \\
&\quad + \log\left(\frac{1}{\sqrt{\beta\nu+\hat{\nu}}}\right) \\
&\Rightarrow G_S(\hat{m},\hat{r},\hat{q}) = \frac{1}{2}\frac{\hat{m}^2}{\beta\nu+\hat{v}} + \frac{1}{2}\frac{\hat{q}}{\beta\nu+\hat{v}} - \frac{1}{2}\log(\beta\nu+\hat{v})
\end{aligned}
\tag{40}
$$

$$
\begin{aligned}
G_E(m,r,q) &= \int \mathcal{D}z \int dy \int \mathcal{D}\lambda^0 P^*_{\text{out}}\left(y \mid \sqrt{r_0-\frac{m^2}{q}}\lambda^0 + \frac{m}{\sqrt{q}}z\right)\times \\
&\quad \times \log\left(\int \mathcal{D}\lambda\, P_{\text{out}}\left(y \mid \sqrt{r-q}\lambda + \sqrt{q}z\right)\right) = \\
&= \int \mathcal{D}z \int dy \int \mathcal{D}\lambda^0 \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(y-\sqrt{r_0-\frac{m^2}{q}}\lambda^0 - \frac{m}{\sqrt{q}}z\right)^2\right)\times \\
&\quad \times \log\left(\int \mathcal{D}\lambda \exp\left(-\frac{\beta}{\sigma^2}\frac{1}{2}\left(y-\sqrt{r-q}\lambda-\sqrt{q}z\right)^2\right)\right) = *
\end{aligned}
\tag{41}
$$

Considero ora esclusivamente l'integrale in $\mathcal{D}\lambda$:

$$\begin{aligned}
&\int \mathcal{D}\lambda \exp\left(-\frac{\beta}{\sigma^2}\frac{1}{2}\left(y-\sqrt{r-q}\lambda-\sqrt{q}z\right)^2\right) = \\
&= \int \frac{d\lambda}{\sqrt{2\pi}} \exp\left(-\frac{\lambda^2}{2}-\frac{\beta}{2\sigma^2}\left(y-\sqrt{r-q}\lambda-\sqrt{q}z\right)^2\right) = \\
&= \frac{1}{\sqrt{2\pi v}} \int d\lambda \exp\left(-\frac{1}{2v}\left(\lambda+\sqrt{q}z\right)^2-\frac{\beta}{2\sigma^2}\left(y-\lambda\right)^2\right) = \\
&= \frac{1}{\sqrt{2\pi v}} \int d\lambda \exp(-\beta\left(\frac{1}{2v}(\lambda+\sqrt{q}z)^2+\frac{1}{2\sigma^2}(y-\lambda)^2\right) = **
\end{aligned} \tag{42}$$

Dove nei passaggi precedenti abbiamo effettuato due cambi di variabile:

$$\lambda \to \frac{1}{\sqrt{v}}(\lambda+\sqrt{q}z)$$
$$v \to \beta^{-1}v$$

In generale per le saddle point equation con $\beta \to +\infty$ vale:

$$\int dx \exp(-\beta f(x)) \to \sqrt{\frac{1}{\beta f''(x^*)}} \exp(-\beta f(x^*))$$

quindi:

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}}\left(\frac{1}{2\nu}(\lambda+\sqrt{q}z)^2+\frac{1}{2\sigma^2}(y-\lambda)^2\right)$$
$$\Rightarrow \lambda^* = \frac{v\sigma^2}{v+\sigma^2}\left(\frac{y}{\sigma^2}-\frac{\sqrt{q}z}{v}\right) = \frac{yv-\sigma^2\sqrt{q}z}{v+\sigma^2}$$

$$\begin{aligned}
** = \frac{1}{\sqrt{2\pi v}} \cdot \sqrt{\frac{1}{\frac{v+\sigma^2}{v\sigma^2}}} \exp\Bigg(-\beta\Bigg(\frac{1}{2v}\left(\frac{y\nu-\sigma^2\sqrt{q}z}{v+\sigma^2}+\sqrt{q}z\right)^2 + \\
+\frac{1}{2\sigma^2}\left(y-\frac{yv-\sigma^2\sqrt{q}z}{v+\sigma^2}\right)^2\Bigg)\Bigg) = \\
= \frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}} \exp\left(-\beta\left(\frac{1}{2v}\left(\frac{v(y-\sqrt{q}z)^2}{v+\sigma^2}\right)^2+\frac{1}{2\sigma^2}\left(\frac{\sigma^4(y-\sqrt{q}z)^2}{(v+\sigma^2)^2}\right)\right)\right) = \\
= \frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}} \exp\left(-\beta\frac{(y-\sqrt{q}z)^2}{2(v+\sigma^2)}\right)
\end{aligned} \tag{43}$$

$$* = \int \mathcal{D}z \int dy \int \mathcal{D}\lambda^0 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - \sqrt{r_0 - \frac{m^2}{q}}\lambda^0 - \frac{m}{\sqrt{q}}z\right)^2\right) \times$$
$$\times \left(\log \frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}} - \beta\frac{(y-\sqrt{q}z)^2}{2(v+\sigma^2)}\right) \tag{44}$$

Considero l'integrale in dy:

$$\int dy \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}\left(y - \sqrt{r_0 - \frac{m^2}{q}} - \frac{m}{\sqrt{q}}z\right)^2\right) \times$$
$$\times \left[\log\left(\frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}}\right) - \beta\left(\frac{(y-\sqrt{q}z)^2}{2(v+\sigma^2)}\right)\right] =$$
$$= \log\frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}}\left[1 - \frac{\beta}{2(v+\sigma^2)}\left(\hat{\sigma}^2 + \left(\sqrt{r_0 - \frac{m^2}{q}}\lambda_0 + \frac{m}{\sqrt{q}}z - \sqrt{q}z\right)^2\right)\right] \tag{45}$$

Integriamo in $D\lambda^0$:

$$\int \frac{d\lambda^0}{\sqrt{2\pi}} \log\frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}}\left[1 - \frac{\beta}{2(v+\sigma^2)}\left(\hat{\sigma}^2 + \left(\sqrt{r_0 - \frac{m^2}{q}}\lambda_0 + \frac{m}{\sqrt{q}}z - \sqrt{q}z\right)^2\right)\right] \tag{46}$$

$$\sqrt{r_0 - \frac{m^2}{q}}\lambda^0 \to T$$

$$T \to T - \frac{m}{\sqrt{q}}z$$

$$\int \frac{dT}{\sqrt{2\pi\left(r_0 - \frac{m^2}{q}\right)}} \quad e^{-\frac{\left(T - \frac{m}{\sqrt{q}}z\right)^2}{2\left(r_0 - \frac{m^2}{q}\right)}} \log\frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}}\left[1 - \frac{\beta}{2(v+\sigma^2)}\left(\hat{\sigma}^2 + (T - \sqrt{q}z)^2\right)\right] =$$
$$= \log\frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}} - \frac{\beta}{2(v+\sigma^2)}\left[\hat{\sigma}^2 + r_0 - \frac{m^2}{q} + \left(\frac{m-q}{\sqrt{q}}z\right)^2\right]$$

Integrando infine in $\mathcal{D}z$ otteniamo:

$$\int \mathcal{D}z \left[\log \frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}} - \frac{\beta}{2(v+\sigma^2)} \left[\hat{\sigma}^2 + r_0 - \frac{m^2}{q} + \left(\frac{m-q}{\sqrt{q}} z \right)^2 \right] \right] \quad (47)$$

$$\Rightarrow G_E(m,v,q) = \log \frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}} - \frac{\beta}{2(v+\sigma^2)} \left[\hat{\sigma}^2 + \left(r_0 - \frac{m^2}{q} \right) + \left(\frac{m-q}{\sqrt{q}} \right)^2 \right] \quad (48)$$

Ricordando che:

$$\begin{aligned} \phi(m,v,q,\hat{m},\hat{v},\hat{q}) &= \frac{1}{2} v(\hat{v}-\hat{q}) + \frac{1}{2} q\hat{v} - m\hat{m} + \\ &+ \frac{1}{2} \frac{\hat{m}^2}{\beta\nu + \hat{v}} + \frac{1}{2} \frac{\hat{q}}{\beta\nu + \hat{v}} - \frac{1}{2} \log(\beta\nu + \hat{v}) + \\ &+ \alpha \left[\log \frac{\sigma}{\sqrt{2\pi(v+\sigma^2)}} - \frac{\beta}{2(v+\sigma^2)} \left[\hat{\sigma}^2 + \left(r_0 - \frac{m^2}{q} \right) + \left(\frac{m-q}{\sqrt{q}} \right)^2 \right] \right] \end{aligned} \quad (49)$$

Effettuo un cambiamento di variabile per rendere tutti i termini lineari in β, (il termine entropico lo è già grazie ad osservazioni fatte precedentemente):

$$\begin{aligned} \hat{v} &\to \beta\hat{v} & v &\to \beta^{-1} v \\ \hat{m} &\to \beta\hat{m} & m &\to m \\ \hat{q} &\to \beta^2 \hat{q} & q &\to q \end{aligned} \quad (50)$$

Applicando il metodo del punto sella, otteniamo l'espressione della funzione libera valutata nei valori ottimali dei parametri d'ordine:

$$\begin{aligned} &\lim_{\beta\to\infty} \frac{1}{\beta} \phi(m,v,q,\hat{m},\hat{v},\hat{q}) = \\ &= -\frac{v}{2}\hat{q} + \frac{1}{2} q\hat{v} - m\hat{m} + \frac{1}{2} \frac{\hat{m}^2}{\nu + \hat{v}} + \frac{1}{2} \frac{\hat{q}}{\nu + \hat{v}} - \frac{\alpha}{2(v+\sigma^2)} \left[\hat{\sigma}^2 + \left(r_0 - \frac{m^2}{q} \right) + \left(\frac{m-q}{\sqrt{q}} \right)^2 \right] \end{aligned} \quad (51)$$

Imponendo la condizione di stazionarietà sulla funzione energia libera $\mathcal{F}(m,q,v,\hat{m},\hat{q},\hat{v})$, ovvero ponendo uguali a zero le derivate parziali rispetto a ciascun parametro d'ordine, otteniamo:

$$\frac{\partial \mathcal{F}}{\partial \hat{q}}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = -\frac{v}{2} + \frac{1}{2(\nu + \hat{v})} = 0 \quad \Rightarrow \quad v = \frac{1}{\nu + \hat{v}}$$

$$\frac{\partial \mathcal{F}}{\partial \hat{m}}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = -m + \frac{\hat{m}}{\nu + \hat{v}} = 0 \quad \Rightarrow \quad m = \frac{\hat{m}}{\nu + \hat{v}}$$

$$\frac{\partial \mathcal{F}}{\partial \hat{v}}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = \frac{1}{2}q + \frac{1}{2}\frac{\hat{m}^2 + \hat{q}}{(\nu + \hat{v})^2} = 0 \quad \Rightarrow \quad q = \frac{\hat{q} + \hat{m}^2}{(\nu + \hat{v})^2}$$

$$\frac{\partial \mathcal{F}}{\partial q}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = \frac{1}{2}\hat{q} - \frac{\alpha}{2(v + \sigma^2)}\left[\frac{m^2}{q^2} - \frac{m^2}{q^2} + 1\right] = 0 \quad \Rightarrow \quad \hat{v} = \alpha\frac{1}{v + \sigma^2}$$

$$\frac{\partial \mathcal{F}}{\partial m}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = -\hat{m} - \frac{\alpha}{2(v + \sigma^2)}\left[-\frac{2m}{q} + \frac{2(m - q)}{q}\right] = 0 \quad \Rightarrow \quad \hat{m} = \alpha\frac{1}{v + \sigma^2}$$

$$\frac{\partial \mathcal{F}}{\partial v}(m, v, q, \hat{m}, \hat{v}, \hat{q}) = -\frac{1}{2}\hat{q} - \frac{\alpha}{2(v + \sigma^2)^2}\left[\hat{\sigma}^2 + r_0 + q - 2m\right] = 0 \Rightarrow$$
$$\Rightarrow \hat{q} = \alpha\,\frac{\hat{\sigma}^2 + r_0 + q - 2m}{(v + \sigma^2)^2}$$

Appendice B

Errore di generalizzazione

Calcoliamo l'errore di generalizzazione secondo la definizione degli scarti quadratici:

$$\mathcal{E}_{gen} = \mathbb{E}_{\{\mathbf{X}^{\text{new}}, y^{new}, \hat{y}^{new}\}}[(y^{new} - \hat{y}^{new})^2]$$

Sostituendo la definizione di y^{new} e $\hat{y}^{new}$ e sviluppando il quadrato di binomio, otteniamo:

$$\mathbb{E}_{\{\mathbf{W},\xi,\mathbf{W}^*,\mathbf{X}^{\text{new}}\}} \left[\left(\frac{\mathbf{W}^{*^T}\mathbf{X}^{\text{new}}}{\sqrt{N}} + \hat{\sigma}\,\xi - \frac{\mathbf{W}^T\mathbf{X}^{\text{new}}}{\sqrt{N}} \right)^2 \right] =$$

$$= \mathbb{E}_{\{\mathbf{W},\xi,\mathbf{W}^*,\mathbf{X}^{\text{new}}\}} \left[\frac{\mathbf{W}^{*^T}\mathbf{X}^{\text{new}}\mathbf{X}^{\text{new}T}\mathbf{W}^*}{N} + \hat{\sigma}^2\,\xi^2 + \frac{\mathbf{W}^T\mathbf{X}^{\text{new}}\mathbf{X}^{\text{new}T}\mathbf{W}}{N} \right.$$

$$\left. + \frac{2\mathbf{W}^{*^T}\mathbf{X}^{\text{new}}}{\sqrt{N}}\hat{\sigma}\,\xi - \frac{2\mathbf{W}^{*^T}\mathbf{X}^{\text{new}}\mathbf{X}^{\text{new}T}\mathbf{W}}{N} - 2\hat{\sigma}\,\xi\,\frac{\mathbf{W}^T\mathbf{X}^{\text{new}}}{\sqrt{N}} \right]$$

Calcoliamo ora la media $\mathbf{X}^{\text{new}}$, variabile aleatoria di media nulla e varianza unitaria (passo 2 appendice A):

$$\Rightarrow \quad \mathbb{E}_{\{\mathbf{W},\xi,\mathbf{W}^*\}} \left[\frac{\mathbf{W}^{*^T}\mathbf{W}^*}{N} + \hat{\sigma}^2\,\xi^2 + \frac{\mathbf{W}^T\mathbf{W}}{N} - \frac{2\mathbf{W}^{*T}\mathbf{W}}{N} \right]$$

Considerando ora che per costruzione $\xi \sim \mathcal{N}(0,1)$, *si ha:*

$$\mathbb{E}[\xi^2] = \mathbb{V}(\xi) + \mathbb{E}[\xi]^2 = 1$$

$$\Rightarrow \quad \mathbb{E}_{\{\mathbf{W},\mathbf{W}^*\}}\left[\frac{\mathbf{W}^{*T}\mathbf{W}^*}{N} + \hat{\sigma}^2 + \frac{\mathbf{W}^T\mathbf{W}}{N} - \frac{2\mathbf{W}^{*T}\mathbf{W}}{N}\right] = r_0 + \hat{\sigma}^2 + q - 2m$$

Bibliografia

[1] URL: https://www.oracle.com/it/artificial-intelligence/machine-learning/what-is-machine-learning.

[2] URL: https://ellycode.com/it/blog/il-test-di-turing-e-il-gioco-dellimitazione/.

[3] URL: https://it.wikipedia.org/wiki/ELIZA_(chat_bot).

[4] URL: https://www.ai4business.it/intelligenza-artificiale/test-di-turing-tutto-quello-che-bisogna-sapere/.

[5] URL: https://www.geopop.it/chatgpt-supera-il-test-di-turing-lo-studio-personalita-sovrapponibile-a-quella-umana/.

[6] URL: https://it.wikipedia.org/wiki/Geoffrey_Hinton.

[7] Alessia Centorame, Federica Finetti e Lucia Pitton. *ATTI DEGLI ISTOLOGI, A.A. 2021/2022.*

[8] URL: https://www.nexsoft.it/reti-neurali-struttura-funzioni-applicazioni-mondo-it/.

[9] F.Rosenblatt. *The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain.*

[10] A. Engel e C.Van den Broeck. *Statistical Mechanics of Learning.*