Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

Inferenza nel modello di Wigner *spiked* in alta dimensione

Relatore:
Prof.
GABRIELE SICURO

Presentata da:
SARA TRAMONTANA

IV Sessione – 31 ottobre 2024
Anno Accademico 2023/2024

Indice

Introduzione

L'inferenza statistica occupa un ruolo fondamentale nelle moderne tecniche di elaborazione di *Big Data* e in diverse applicazioni di *Machine Learning*. Questa branca si concentra principalmente sulla ricostruzione del segnale o di stimare i parametri sconosciuti di un modello a partire da alcuni dati osservabili. Pierre-Simon de Laplace (1749-1827) fu uno dei primi ad utilizzare metodi probabilistici per fare inferenza su dati osservati e a perfezionare il lavoro di Thomas Bayes, che nel 1763 nella sua opera postuma *An Essay towards solving a Problem in the Doctrine of Chances* [12] aveva introdotto l'idea centrale di quella che poi sarebbe stata l'inferenza Bayesiana. A differenza della statistica classica, generalmente interessata a contesti in cui il numero di parametri è piccolo rispetto al numero dei dati, la statistica moderna ha a che fare con regimi in cui numero di parametri e quantità di dati possono essere dello stesso ordine di grandezza. A tal proposito, risultano necessarie tecniche computazionali avanzate, in modo da poter gestire la complessità dei dati e ottenere stime il più accurate possibile, definendo dunque un'opportuna metrica di errore. Questo regime ad alta-dimensione, particolarmente compatibile con metodi di inferenza Bayesiana, è strettamente legato alla meccanica statistica, sviluppatasi tra la fine del XIX secolo e l'inizio del XX secolo grazie a James Clerk Maxwell, Ludwig Boltzmann e Josiah Willard Gibbs. Questa branca della fisica si concentra sullo studio della materia a livello macroscopico, in modo da studiarne i comportamenti emergenti, scaturiti spesso da fattori esterni. Proprio questa peculiarità permette di riformulare i problemi di teoria dell'informazione, che occupano un ruolo centrale nel corso della trattazione di questa tesi, in modo da poter applicare alcuni metodi e concetti propri della meccanica statistica.

Nel primo capitolo verranno introdotte le nozioni fondamentali dell'inferenza Bayesiana, funzionali allo sviluppo della teoria del capitolo successivo, quali il teorema di Bayes e la distribuzione a posteriori, la *loss function* e l'MMSE, per poi trattare alcuni concetti cardine della teoria dell'informazione come l'entropia di Shannon di una variabile aleatoria e l'informazione mutua tra due variabili aleatorie. Inoltre si parlerà anche dell'energia libera di un sistema e della distribuzione di Gibbs–Boltzmann, sottolineando ancora una volta come la fisica statistica e l'inferenza Bayesiana siano strettamente collegate.

Il secondo capitolo sarà invece interamente dedicato allo studio del modello di Wigner *spiked* in regime Bayes-ottimale, oggetto principale della tesi, dove l'obiettivo principale sarà quello di ricostruire il segnale nascosto (i.e., lo *spike*) nel modo più accurato possibile a partire dai dati osservabili, contenenti rumore. Verranno analizzate le grandezze precedentemente introdotte nell'ottica dei modelli di campo medio, che consentono una riduzione da alta dimensione a bassa dimensione di alcune espressioni, rendendole più facili da trattare. Tale semplificazione e altre assunzioni saranno poi rigorosamente giustificate, analizzando diversi metodi matematici. Infine, si evidenzierà il fatto che l'efficienza di alcuni algoritmi di ricostruzione del segnale è strettamente legata alle proprietà spettrali delle matrici che appaiono nel problema, concentrandoci in particolare sul metodo PCA.

Capitolo 1

Inferenza Bayesiana, teoria dell'informazione e meccanica statistica

In questo primo capitolo si introducono i concetti base della teoria dell'informazione e dell'inferenza statistica, utili per comprendere il contenuto successivo. Le definizioni e i concetti esposti seguono la trattazione della Ref. [1].

1.1 Una introduzione all'inferenza statistica

Si consideri un processo generativo *casuale* $X \to Y$, dove $X \in \mathcal{X}$ sono da intendersi come parametri (o *segnale*) del processo e $Y \in \mathcal{Y}$ sono i dati generati dallo stesso. Possiamo distinguere tra due misure di probabilità, che definiremo *probabilità diretta* e *probabilità inversa*. La *probabilità diretta* $P(y \mid x) \equiv P(Y = y \mid X = x)$ caratterizza la distribuzione dei dati y a dati x fissati: si può immaginare che in questo caso il processo non si sia ancora svolto e miriamo a predire cosa accadrà. La *probabilità inversa* $P(x \mid y) \equiv P(X = x \mid Y = y)$, caratterizza invece l'informazione sul segnale in base ai dati y: si può immaginare cioè che l'interesse sia focalizzato sul segnale alla luce dei dati generati da un esperimento. L'*inferenza statistica Bayesiana* consiste proprio nel calcolare la distribuzione di probabilità

condizionata dei parametri $P(x \mid y)$ a partire dai risultati osservabili. Limiteremo la discussione di tali problemi al contesto bayesiano ottimale, ovvero al caso in cui è noto il processo di generazione dei dati, ma naturalmente non i parametri da cui tali dati dipendono. Per evidenziare ancor meglio il fatto che considereremo il caso in cui i dati sono assegnati e non possono essere modificati, definiamo la funzione di verosomiglianza.

Definizione 1.1 (Funzione di verosomiglianza)**.** La funzione di verosomiglianza $\mathcal{L}_y(x) \coloneqq P(y \mid x)$, è una funzione di probabilità condizionata, considerata come funzione del suo secondo argomento mantenendo fisso il primo. Formalmente è una funzione $\mathcal{L}_y \colon x \to P(y \mid X = x)$.

Mentre questa funzione si occupa di modellare il processo generativo dei dati, i parametri x sono sconosciuti e tutte le informazioni a priori che abbiamo su di essi sono racchiuse in $P(x)$. Questa distribuzione di probabilità, chiamata *prior* o probabilità a priori, non dipende dai dati e racchiude tutte le assunzioni fatte, riguardanti i parametri sconosciuti (i.e. segnale). Stando a quanto appena introdotto, possiamo conoscere la distribuzione a posteriori (i.e. *posterior*), utilizzando la *formula di Bayes*:

$$P(X = x \mid y) = P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}. \tag{1.1}$$

Tipicamente i parametri del modello sono considerati come il segnale di interesse e i dati osservabili come quantità affette da rumore. In questi casi, l'obiettivo è recuperare il segnale: il *signal-to-noise ratio* (SNR) quantifica tipicamente quanto questa operazione è difficile. Il rapporto segnale-rumore (SNR), è usualmente un numero puro non negativo che misura il rapporto tra la potenza del segnale rispetto alla potenza del rumore.

A differenza della statistica classica, interessata generalmente a contesti dove il numero di parametri che inferiscono n, è piccolo, rispetto al numero m di dati, nell'inferenza ad alta-dimensione sia la dimensione dei dati che dei parametri è grande. Un regime interessante è quello in cui entrambe queste quantità tendono a infinito ma in modo che il loro rapporto sia una costante di ordine uno. In

particolare l'inferenza è non triviale quando per n e m grandi, il SNR in media si mantiene finito. È proprio in questi contesti che vengono studiati i "big data" e problemi di Machine Learning, dove è stato necessario introdurre nuovi algoritmi matematici, efficaci dal punto di vista statistico ed efficienti dal punto di vista computazionale.

1.1.1 Teoria della decisione Bayesiana

Supponiamo quindi di avere un processo casuale $\boldsymbol{X} \to \boldsymbol{Y}$, disporre di dati che ha prodotto e di voler ricostruire i parametri da essi. Assumeremo che $\boldsymbol{X}$ prenda valori in uno spazio n-dimensionale $\mathcal{X}$, mentre $\boldsymbol{Y}$ prenda valori in uno spazio m-dimensionale $\mathcal{Y}$, e useremo per questo il grassetto. Appare quindi necessario definire una metrica di errore che caratterizza la qualità dell'inferenza.

Quando il segnale è *discreto*, risulta utile definire la seguente *loss function*.

Definizione 1.2 (Loss-function)**.** Sia $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y}(\boldsymbol{x}^*))$ il risultato di un certo algoritmo di ricostruzione che opera sui dati $\boldsymbol{y}(\boldsymbol{x}^*)$ prodotti da un certo segnale $\boldsymbol{x}^*$:

$$\hat{\boldsymbol{x}} = \text{algo}(\boldsymbol{y}(\boldsymbol{x}^*)),$$

si definisce *loss function* relativa a $\hat{\boldsymbol{x}}, \boldsymbol{x}$, la funzione:

$$E(\hat{\boldsymbol{x}}, \boldsymbol{x}) \coloneqq 1 - \mathbb{1}(\hat{\boldsymbol{x}} = \boldsymbol{x}).$$

La loss function sarà la funzione che deve essere minimizzata per ricostruire i parametri. Minimizzando quest'ultima si riduce al massimo la "perdita" (*loss*): questa quantità tuttavia non può essere calcolata poiché dipende da parametri sconosciuti, ovvero da $\boldsymbol{x}^*$. Il modo migliore per approssimarla è calcolando il rischio, anche detto *block error rate*, strettamente legato ad essa per mezzo del *posterior*, ovvero della probabilità a posteriori.

Definizione 1.3 (Block-error-rate)**.** Sia $P(\boldsymbol{x} \mid \boldsymbol{y})$ la distribuzione a posteriori, posso definire il rischio (chiamato anche *block error rate*) come:

$$R(\hat{\boldsymbol{x}}, \boldsymbol{y}) \coloneqq \sum_{\boldsymbol{x} \in \mathcal{X}} P(\boldsymbol{x} \mid \boldsymbol{y}) E(\hat{\boldsymbol{x}}, \boldsymbol{x}) = 1 - \sum_{\boldsymbol{x} \in \mathcal{X}} P(\boldsymbol{x} \mid \boldsymbol{y}) \mathbb{1}(\hat{\boldsymbol{x}} = \boldsymbol{x}) = 1 - P(\hat{\boldsymbol{x}} \mid \boldsymbol{y}) \quad (1.2)$$

dove $R(\hat{\boldsymbol{x}}, \boldsymbol{y})$ può essere interpretato come la probabilità a posteriori che lo stimatore $\hat{\boldsymbol{x}}$ sia sbagliato, dato $\boldsymbol{y}$.

A questo punto sorge spontanea una domanda: quale algoritmo/stimatore, minimizza questa perdita? Si nota che lo stimatore ottimale, è, nel caso della *loss function* data, la moda a posteriori:

$$\hat{\boldsymbol{x}}_{\text{opt}}(\boldsymbol{y}) \coloneqq \arg\min_{\hat{\boldsymbol{x}}} R(\hat{\boldsymbol{x}}, \boldsymbol{y}) = \arg\min_{\hat{\boldsymbol{x}}} (1 - P(\hat{\boldsymbol{x}} \mid \boldsymbol{y})) = \arg\max_{\hat{\boldsymbol{x}}} P(\hat{\boldsymbol{x}} \mid \boldsymbol{y}).$$

Definizione 1.4 (Lo stimatore MAP)**.** Lo stimatore MAP (*maximum a-posteriori*), che mi fornisce $\hat{\boldsymbol{x}}_{\text{opt}}$, è lo stimatore ottimale per minimizzare il rischio.

Il rischio ottimale associato a tale estimatore è

$$R_{\text{opt}}(\boldsymbol{y}) \coloneqq R(\hat{\boldsymbol{x}}_{\text{opt}}(\boldsymbol{y}), \boldsymbol{y}).$$

Definizione 1.5 (Contesto Bayes-ottimale)**.** Si dice che un esperimento è stato svolto in un contesto Bayes-ottimale se si conosce il processo di generazione dei dati e i dati osservabili, o equivalentemente se è nota la distribuzione a posteriori data dall'Eq. (1.1).

Come vedremo, in un contesto Bayes-ottimale è possibile ricostruire il segnale nel migliore dei modi, tramite una funzione errore che utilizza la norma di L^2.

Così come la *loss function* definita precedentemente è adeguata per un segnale discreto, una scelta più appropriata per i segnali in uno spazio numerico con la cardinalità del continuo è la loss function quadratica. Assumendo che il segnale da ricostruire $\boldsymbol{x}^*$ sia in $\mathcal{X} \subseteq \mathbb{R}^n$, essa ha la forma:

$$E(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|^2, \tag{1.3}$$

dove la norma è appunto quella della norma L^2. Il rischio a posteriori associato è chiamato *errore quadratico medio* (MSE):

$$\text{MSE}(\hat{\boldsymbol{x}}, \boldsymbol{y}(\boldsymbol{x})) = \int_{\mathcal{X}} \mathrm{d}\, P(\boldsymbol{x} \mid \boldsymbol{y}) \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2. \tag{1.4}$$

L'MSE è minimizzato dallo stimatore

$$\hat{\boldsymbol{x}}_{\text{opt}}(\boldsymbol{y}) = \int_{\mathcal{X}} \boldsymbol{x} \, \mathrm{d}\, P(\boldsymbol{x} \mid \boldsymbol{y}) =: \langle \boldsymbol{X} \rangle$$

dove $\langle \boldsymbol{X} \rangle$ sta ad indicare le *parentesi di Gibbs*. A questo punto, è possibile definire anche l'MMSE (*minimum mean square error*):

$$\begin{aligned} \mathrm{MMSE}(\boldsymbol{X}^* | \boldsymbol{Y}) &\coloneqq \int_{\mathcal{X}} \mathrm{MSE}(\hat{\boldsymbol{x}}_{\text{opt}}(\boldsymbol{y}), \boldsymbol{y}) \, \mathrm{d}\, P(\boldsymbol{y}) \\ &= \int_{\mathcal{X}} \|\boldsymbol{x} - \langle \boldsymbol{X} \rangle\|^2 \, \mathrm{d}\, P(\boldsymbol{y}) \, \mathrm{d}\, P(\boldsymbol{x} \mid \boldsymbol{y}) = \int_{\mathcal{X}} \|\boldsymbol{x}^* - \langle \boldsymbol{X} \rangle\|^2 \, \mathrm{d}\, P(\boldsymbol{x}^*) \, \mathrm{d}\, P(\boldsymbol{y} \mid \boldsymbol{x}^*) \\ &= \mathbb{E} \|\boldsymbol{X}^* - \langle \boldsymbol{X} \rangle\|^2 = \mathbb{E}_{\boldsymbol{Y}} \mathrm{Var}(\boldsymbol{X} | \boldsymbol{Y}) = \mathbb{E}_{\boldsymbol{Y}} \langle \|\boldsymbol{X}^* - \langle \boldsymbol{X} \rangle\|^2 \rangle, \quad (1.5) \end{aligned}$$

dove il primo $\mathbb{E} = \mathbb{E}_{\boldsymbol{X}^*, \boldsymbol{Y}} = \mathbb{E}_{\boldsymbol{X}^*} \mathbb{E}_{\boldsymbol{Y} | \boldsymbol{X}^*}$.

Calcolare $\hat{\boldsymbol{x}}_{\text{opt}}$ può essere molto costoso perchè ci sono integrali che riguardano quantità di dimensione n, ed è per questo motivo che molte volte si preferisce lo stimatore MAP, $\hat{\boldsymbol{x}}_{\text{MAP}} \coloneqq \arg\max_{\boldsymbol{x}} P(\boldsymbol{x} \mid \boldsymbol{y})$.

Il MMSE, oltre ad essere rilevante nelle maggior parte delle applicazioni, ha il grande vantaggio di essere più facilmente accessibile/calcolabile grazie alla stretta relazione con l'informazione mutua, o energia libera (quantità legate da una costante additiva), che definiremo successivamente.

È importante notare che la formula in Eq. (1.5) è il primo esempio in cui abbiamo utilizzato *l'identità di Nishimori*: cioè abbiamo sostituito un campione qualsiasi $\boldsymbol{X} \sim P(\bullet | \boldsymbol{y})$ con un campione avente la stessa distribuzione del segnale cercato, cioè $\boldsymbol{X}^* \sim P(\bullet | \boldsymbol{y})$.

Proposizione 1 (Identità di Nishimori)**.** *Sia* $(\boldsymbol{X}, \boldsymbol{Y})$ *una coppia di variabili casuali, con distribuzione congiunta* $P_{\boldsymbol{X}\boldsymbol{Y}}$ *e distribuzione condizionata* $P_{\boldsymbol{X}|\boldsymbol{Y}}$*. Sia* $k \geq 1$ *e siano* $\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(k)}$ *delle variabili casuali i.i.d, con distribuzione* $P_{\boldsymbol{X}|\boldsymbol{Y}}$*, denotando con* $\mathbb{E}$ *il valore atteso in relazione a* $P_{\boldsymbol{X}\boldsymbol{Y}}$ *e con* $\langle \bullet \rangle$ *la media di* $P_{\boldsymbol{X}|\boldsymbol{Y}}^{\otimes \infty}$*. Allora per ogni funzione continua e limitata* g *si ha:*

$$\mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} \langle g(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X}^{(2)}, \dots, \boldsymbol{X}^{(k)}) \rangle = \mathbb{E}_{\boldsymbol{Y}} \langle g(\boldsymbol{Y}, \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \dots, \boldsymbol{X}^{(k)}) \rangle$$

Dimostrazione. Il fatto che $P_{\boldsymbol{X}\boldsymbol{Y}} = P_{\boldsymbol{X}|\boldsymbol{Y}}P_{\boldsymbol{Y}} = P_{\boldsymbol{Y}|\boldsymbol{X}}P_{\boldsymbol{X}}$ è una diretta conseguenza della formula di Bayes in Eq. (1.1). Da questo fatto segue che:

$$\begin{aligned}\mathbb{E}\langle g(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(k)})\rangle &:= \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{X}^{(2)}|\boldsymbol{Y}}\cdots\mathbb{E}_{\boldsymbol{X}^{(k)}|\boldsymbol{Y}} g(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(K)}) \\ &= \mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{X}|\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{X}^{(2)}|\boldsymbol{Y}}\cdots\mathbb{E}_{\boldsymbol{X}^{(k)}|\boldsymbol{Y}} g(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(K)}) \\ &= \mathbb{E}_{\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{X}^{(1)}|\boldsymbol{Y}}\mathbb{E}_{\boldsymbol{X}^{(2)}|\boldsymbol{Y}}\cdots\mathbb{E}_{\boldsymbol{X}^{(k)}|\boldsymbol{Y}} g(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(K)}) \\ &=: \mathbb{E}\langle g(\boldsymbol{Y},\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\dots,\boldsymbol{X}^{(k)})\rangle.\end{aligned} \tag{1.6}$$

□

Questa identità è di fondamentale importanza, perchè è all'origine di un gran numero di semplificazioni che consentono un'analisi accurata nei problemi di inferenza ad alta dimensione, in ambito bayesiano ottimale, dove si evidenzia un ruolo simmetrico tra il segnale $\boldsymbol{X}^*$ e $\boldsymbol{X}^{(i)}$. È importante sottolineare che la validità di tale identità è garantita dal momento in cui supponiamo di essere in un contesto Bayes-ottimale.

1.2 Entropia di Shannon e mutua informazione

L'*entropia di Shannon* è un concetto matematico che può essere utilizzato per quantificare l'informazione riguardante il segnale contenuta nei dati. L'introduzione di questa quantità da parte di Claude Shannon nel 1948 in un articolo fondamentale [11] è stata di notevole importanza e ha dato avvio alla teoria dell'informazione.

Definizione 1.6 (Entropia di Shannon)**.** Sia $\boldsymbol{X} \sim P$ una variabile aleatoria casuale e discreta a valori in $\mathcal{X}$. *L'entropia di Shannon* della variabile aleatoria $\boldsymbol{X}$, è definita come:

$$H(\boldsymbol{X}) := \sum_{\boldsymbol{x}\in\mathcal{X}} P(\boldsymbol{x}) \ln \frac{1}{P(\boldsymbol{x})}$$

dove

$$h(\boldsymbol{x}) := \ln \frac{1}{P(\boldsymbol{x})}$$

è l'attesa del contenuto informativo dato dall'osservazione di $\boldsymbol{X} = \boldsymbol{x}$.

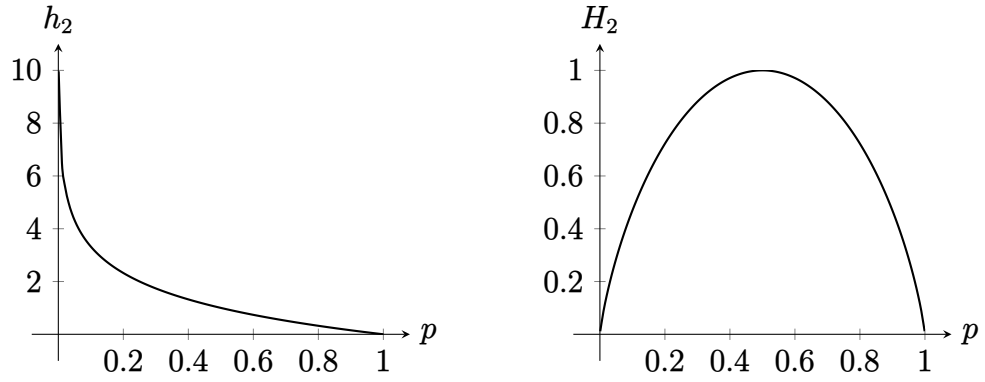

Figura 1.1: La figura mostra il contenuto informativo, in *bits*, dell'evento $X = 1$ di una variabile X bernoulliana di parametro p. A destra è invece rappresentata la funzione $H_2(X)$ per la stessa variabile, con il suo massimo ad $\frac{1}{2}$.

Sostituendo al logaritmo naturale il logaritmo in base 2 nelle formule precedenti si ottiene l'entropia $H_2(\boldsymbol{X})$ in *bits* della variabile stocastica $\boldsymbol{X}$. L'entropia di Shannon può essere interpretata come una misura dell'imprevedibilità di $\boldsymbol{X}$. Se è improbabile che si verifichi un certo risultato $\boldsymbol{x}$, questo contiene allora più informazioni se viene osservato, cioè è un sorta di *potenziale* guadagno di informazioni su $\boldsymbol{X}$: più improbabile è l'esito di una certo $\boldsymbol{x}$, maggiore sarà la tua entropia $h(\boldsymbol{x})$.

Definizione 1.7 (Entropia condizionata ad un'altra variabile aleatoria)**.** Siano $\boldsymbol{X}$ e $\boldsymbol{Y}$ due variabili aleatorie discrete a valori in $\mathcal{X}$ e $\mathcal{Y}$ rispettivamente. È possibile definire anche l'entropia condizionata di $\boldsymbol{X}$ data $\boldsymbol{Y}$:

$$H(\boldsymbol{X} \mid \boldsymbol{Y}) \coloneqq \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}} P(\boldsymbol{y}) P(\boldsymbol{x} \mid \boldsymbol{y}) \ln \frac{1}{P(\boldsymbol{x} \mid \boldsymbol{y})}.$$

Elenchiamo infine alcune proprietà dell'entropia. Sia $\boldsymbol{X} \sim P$ una variabile aleatoria discreta a valori in $\mathcal{X}$, allora:

- $H(\boldsymbol{X}) \geq 0$ ed uguale a zero se $P(\boldsymbol{x}) = 1$ per un qualche $\boldsymbol{x} \in \mathcal{X}$;
- $H(\boldsymbol{X}) \leq \ln |\mathcal{X}|$ ed è uguale a $\ln |\mathcal{X}|$ se e solo se $P(\boldsymbol{x}) = \frac{1}{|\mathcal{X}|}$ per ogni $\boldsymbol{x} \in \mathcal{X}$. Quindi la distribuzione uniforme ha entropia massima, mentre la distribuzione banale $P(\boldsymbol{x}) = \delta_{\boldsymbol{x},\boldsymbol{x}_0}$ per un qualche $\boldsymbol{x}_0 \in \mathcal{X}$, ha entropia nulla.

- *Regola della catena* per l'entropia: Siano $\boldsymbol{X}$ e $\boldsymbol{Y}$ due variabili aleatorie, allora vale che:

$$H(\boldsymbol{X}, \boldsymbol{Y}) = H(\boldsymbol{X} \mid \boldsymbol{Y}) + H(\boldsymbol{Y}) = H(\boldsymbol{Y} \mid \boldsymbol{X}) + H(\boldsymbol{X}), \tag{1.7}$$

  dove $H(\boldsymbol{X}, \boldsymbol{Y}) \coloneqq \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}\times\mathcal{Y}} P(\boldsymbol{x}, \boldsymbol{y}) \ln \frac{1}{P(\boldsymbol{x},\boldsymbol{y})}$. L'uguaglianza in Eq. (1.7) mostra che la quantità di informazione rivelata valutando simultaneamente $(\boldsymbol{X}, \boldsymbol{Y})$ è uguale all'informazione rivelata conducendo contemporaneamente due esperimenti: prima valutando l'informazione di $\boldsymbol{Y}$ e poi valutando l'informazione di $\boldsymbol{X}$ noto il valore di $\boldsymbol{Y}$.

- Sia $f\colon \mathcal{X} \to \mathbb{R}$ una funzione reale qualsiasi; allora $H(f(\boldsymbol{X}) \mid \boldsymbol{X}) = 0$. Applicando la formula precedente si ottiene: $H(\boldsymbol{X}) + H(f(\boldsymbol{X}) \mid \boldsymbol{X}) = H(f(\boldsymbol{X})) + H(\boldsymbol{X} \mid f(\boldsymbol{X}))$ da cui:

$$H(f(\boldsymbol{X})) \leq H(\boldsymbol{X}).$$

- Se $\boldsymbol{X}$ e $\boldsymbol{Y}$ sono indipendenti $H(\boldsymbol{X} \mid \boldsymbol{Y}) = H(\boldsymbol{X})$ e $H(\boldsymbol{Y} \mid \boldsymbol{X}) = H(\boldsymbol{Y})$.

- In generale vale che $H(\boldsymbol{X}, \boldsymbol{Y}) \leq H(\boldsymbol{X}) + H(\boldsymbol{Y})$, mentre l'uguaglianza vale se e solo se le due variabili sono indipendenti.

Così come abbiamo precedentemente definito l'entropia per variabili aleatorie discrete, allo stesso modo è possibile definire l'entropia per variabili aleatorie a valori reali.

Definizione 1.8 (Entropia Differenziale)**.** Si definisce l'entropia differenziale di una variabile continua $\boldsymbol{X}$ su $\mathcal{X} \subseteq \mathbb{R}^n$ nel modo seguente:

$$H(\boldsymbol{X}) \coloneqq \int_{\mathcal{X}} \mathrm{d}\, P(\boldsymbol{x}) \ln \frac{1}{P(\boldsymbol{x})}.$$

Non tutte le proprietà precedentemente definite per l'entropia di varibili aleatorie discrete si estendono correttamente a contesti continui. Tuttavia, queste nozioni vennero poi riprese e studiate da Edwin Thompson Jaynes, che ne fornì una formulazione più corretta.

1.2.1 Mutua informazione

L'informazione mutua è un concetto chiave per lo sviluppo della teoria dell'informazione. Essa è legata all'entopia appena definita da una semplice relazione.

Definizione 1.9 (Mutua informazione)**.** Siano $\boldsymbol{X}$ e $\boldsymbol{Y}$ due variabili aleatorie a valori in $\mathcal{X}$ e $\mathcal{Y}$ rispettivamente. Posso definire la mutua informazione tra $\boldsymbol{X}$ e $\boldsymbol{Y}$, come:

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &\coloneqq H(\boldsymbol{X}) + H(\boldsymbol{Y}) - H(\boldsymbol{X},\boldsymbol{Y}) \\ &= H(\boldsymbol{X},\boldsymbol{Y}) - H(\boldsymbol{X}|\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) = H(\boldsymbol{X}) - H(\boldsymbol{X}|\boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y}|\boldsymbol{X}) \geq 0. \end{aligned} \tag{1.8}$$

Come si può intuire dalla formula, essa mi indica la "quantità di informazione" ottenuta su una variabile aleatoria attraverso l'osservazione dell'altra. Elenchiamo ora alcune proprietà della mutua informazione.

- In generale vale che $I(\boldsymbol{X},\boldsymbol{Y}) \geq 0$ e l'uguaglianza vale solo nel caso in cui le due variabili aleatorie sono indipendenti.
- Per ogni coppia di funzioni misurabili $g_1 \colon \mathcal{X} \to \mathbb{R}$ e $g_2 \colon \mathcal{Y} \to \mathbb{R}$ vale che:

  $$I(g_1(\boldsymbol{X}), g_2(\boldsymbol{Y})) \leq I(\boldsymbol{X};\boldsymbol{Y}).$$

  L'ugualianza vale solo nel caso in cui entrambe le funzioni siano invertibili.
- Siano $\boldsymbol{X}$, $\boldsymbol{Y}$ e $\boldsymbol{Z}$ variabili casuali, dove $\boldsymbol{Z}$ potrebbe dipendere solo da $\boldsymbol{Y}$, allora si ha:

  $$I(\boldsymbol{X};\boldsymbol{Z}) \leq I(\boldsymbol{X};\boldsymbol{Y}).$$

  Questo significa che nessuna trasformazione dei dati può creare informazione.

1.3 L'inferenza Bayesiana come un problema disordinato di meccanica statistica

La meccanica statistica si è sviluppata tra la fine del XIX secolo e l'inizio del XX secolo. A differenza della meccanica quantistica, concentrata sulla descrizione

del comportamento della materia a livello microscopico delle singole componenti, questa branca della fisica si interessa allo studio delle quantità medie che descrivono il sistema nel suo insieme, cioè a livello macroscopico. Risulta infatti difficile descrivere un sistema complesso analizzando singolarmente, anche in modo molto preciso, tutte le sue componenti e poi cercare di combinare insieme tutte le informazioni ottenute: un sistema complesso è più convenientemente studiato nel suo insieme. Esistono infatti dei *comportamenti emergenti*, nei quali un sistema complesso esibisce proprietà macroscopiche particolari, difficilmente predicibili sulla base delle leggi che governano le sue innumerevoli componenti prese singolarmente.

Un esempio di un comportamento emergente è la *transizione di fase*, che si verifica quando un sistema complesso sperimenta un cambiamento osservabile, abbastanza improvviso, di alcune proprietà macroscopiche/globali, per effetto di un operatore esterno. Questo comportamento emerge in un "regime ad alta dimensione", tipico come appena accennato della meccanica statistica, in cui i gradi di libertà del sistema diventano innumerevoli: un regime che ricorda proprio i problemi di inferenza. In questa sezione introdurremo alcuni concetti che saranno utili per riformulare problemi di teoria dell'informazione nel linguaggio della fisica statistica, e poter utilizzare quindi i metodi di quest'ultima.

Sia $\boldsymbol{X} = (X_i)_{i=1}^n \in \mathcal{X}$, spazio n-dimensionale numerabile, una variabile aleatoria che caratterizza completamente un sistema fisico. Una sua realizzazione è detta *configurazione microscopica*, mentre le sue componenti X_i sono chiamate in fisica "gradi di libertà". Il sistema, che potrebbe dipendere anche da altri parametri fissati $\boldsymbol{y}$, si dice che è in equilibrio, se la sua distribuzione di probabilità prende la forma di *Gibbs–Boltzmann*:

$$P(\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{y}, \beta) \equiv P(\boldsymbol{x}; \boldsymbol{y}, \beta) = \frac{\exp(-\beta \mathcal{H}(\boldsymbol{x}; \boldsymbol{y}))}{\mathcal{Z}(\boldsymbol{y}, \beta)}; \tag{1.9}$$

dove $\mathcal{Z}(\boldsymbol{y}, \beta) = \sum_{\boldsymbol{x} \in \mathcal{X}} \exp(-\beta \mathcal{H}(\boldsymbol{x}; \boldsymbol{y}))$ prende il nome di *funzione di partizione* e contiene tutte le informazioni rilevanti sul sistema. Essa normalizza la distribuzione e ricorda il problema inerente il normalizzare una distribuzione di probabilità ad alta dimensione, come quella a posteriori. La funzione $\mathcal{H}(\boldsymbol{x}; \boldsymbol{y})$, invece, è detta *hamiltoniana* del sistema, cioè la sua funzione energia/costo, che ha il ruolo di

assegnare una certa energia per ogni possibile configurazione del sistema. La costante β, invece, può essere interpretata dal punto di vista fisico come una sorta di parametro di controllo che consente di regolare la "quantità di casualità" con cui una configurazione si manifesta, dovuta all'interazione del sistema con l'ambiente esterno. Definizioni analoghe possono essere date nel caso in cui $\mathcal{X}$ sia uno spazio di dimensione finita non numerabile ma misurabile in termini di densità di probabilità.

Una quantità di particolare interesse della meccanica statistica, che analizzeremo nella trattazione successiva, è l'energia libera.

Definizione 1.10 (Energia Libera)**.** Si definisce l'energia libera di un sistema di dimensione n:

$$F_n(\boldsymbol{y}, \beta) \coloneqq -\frac{1}{n\beta} \ln \mathcal{Z}(\boldsymbol{y}, \beta).$$

Questa quantità è molto importante nella meccanica statistica, poiché contiene tutte le informazioni termodinache sul modello: i punti di non analiticità del suo *limite termodinamico*, ovvero per $n \to \infty$, corrispondono alle posizioni delle transizioni di fase. Come vedremo, nel limite termodinamico, l'energia libera di diversi sistemi si concentra intorno ad un asintotico valore dato.

Chiariamo ora l'analogia tra meccanica statistica e inferenza.

Proposizione 2. *La distribuzione a posteriori data dalla formula di Bayes in Eq. (1.1), può essere pensata come una distribuzione di Gibbs–Boltzmann (cioè della forma in Eq. (1.9)) nel contesto dell'inferenza ad alta dimensione.*

Dimostrazione. La dimostrazione è immediata: basta considerare la formula in Eq. (1.9) con $\beta = 1$ e identificare nell'hamiltoniana $\mathcal{H}(\boldsymbol{x}; \boldsymbol{y}) = -\ln P(\boldsymbol{x}) - \ln P(\boldsymbol{y} \mid \boldsymbol{x})$. La funzione di partizione è allora $\mathcal{Z}(\boldsymbol{y}, 1) \equiv \mathcal{Z}(\boldsymbol{y}) = P(\boldsymbol{y})$. Si noti che questa è una delle possibili scelte. □

Segue quindi il seguente fatto.

Definizione 1.11 (Energia libera media)**.** L'energia libera media è uguale all'entropia differenziale dei dati (o entropia di Shannon in un contesto discreto) diviso

il numero dei parametri che inferiscono:

$$f_n(\lambda) \coloneqq -\frac{1}{n}\mathbb{E}\ln \mathcal{Z}(\boldsymbol{Y}) = \frac{1}{n}H(\boldsymbol{Y}). \tag{1.10}$$

Si è dunque messo in evidenza che la fisica statistica e l'inferenza Bayesiana sono strettamente collegate.

Capitolo 2

Un modello di campo medio: la magia della concentrazione in misura

In questo capitolo ci concentreremo sull'applicazione di varie tecniche e idee ispirate alla teoria dei sistemi disordinati in meccanica statistica e allo studio di un particolare modello di inferenza in alta-dimensione, il cosiddetto *modello spiked-Wigner*. Ci focalizzeremo in particolare sul regime Bayes–ottimale, in cui vale la *simmetria delle repliche*, ovvero l'autoconcentrazione dei parametri d'ordine: le loro proprietà statistiche sono correttamente e completamente descritte dalla loro media, come vedremo in seguito. Nella parte intermedia del capitolo, invece, viene enunciato e dimostrato il teorema centrale della tesi. I concetti esposti seguono la trattazione fatta nelle Ref. [1] e [2]. Infine, facendo riferimento alla Ref. [4] e alla Ref. [6] viene esposto il metodo PCA, fornendone un'analisi qualitativa.

2.1 Il modello di Wigner *spiked*

Consideriamo un vettore stocastico $\boldsymbol{X}^* = (X_i^*)_{i=1}^n \in \mathbb{R}^n$ avente componenti limitate e casuali , indipendenti e identicamente distribuite, con stessa legge P_x, che assumiamo pari e con varianza ρ: tale vettore costituirà il *segnale*. Da tale

vettore si costruisce la matrice simmetrica dei *dati* $\boldsymbol{Y} = (Y_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$, ottenuta tramite "l'osservazione del modello" come:

$$\boldsymbol{Y} = \sqrt{\frac{\lambda}{n}}\boldsymbol{X}^* \otimes \boldsymbol{X}^{*\intercal} + \boldsymbol{\Xi} \iff Y_{ij} = \sqrt{\frac{\lambda}{n}} X_i^* X_j^* + \Xi_{ij}, \quad 1 \le i,j \le n \tag{2.1}$$

dove $\boldsymbol{\Xi} = (\Xi_{ij})_{i,j=1}^n$ è una *matrice di Wigner*, ovvero simmetrica e di rumore, con elementi aventi distribuzione normale standard e i.i.d., $\Xi_{ij} = \Xi_{ji} \sim \mathcal{N}(0,1)$. Assumeremo, per semplicità di calcoli, che $\Xi_{ii} \sim \mathcal{N}(0,\sigma^2)$ con $\sigma^2 \to +\infty$ o, equivalentemente, che solo i termini fuori diagonale di $\boldsymbol{Y}$ siano accessibili [8].

Osservazione 1. La dipendenza di $\boldsymbol{Y}$ dalla matrice di segnale $\boldsymbol{X}^*$, fa in modo che si perda in $\boldsymbol{Y}$ un' informazione globale sul segno del segnale per ragioni di simmetria; si ha quindi che $P(\boldsymbol{X}^*|\boldsymbol{Y}) = P(-\boldsymbol{X}^*|\boldsymbol{Y})$, in modo che $\mathbb{E}[\boldsymbol{X}^*|\boldsymbol{Y}] = \boldsymbol{0}$.

Per quanto detto, ha senso considerare la matrice di rango uno $\boldsymbol{X}^* \otimes \boldsymbol{X}^{*\intercal} = (X_i^* X_j^*)_{i,j=1}^n$ come un segnale nascosto (chiamato "picco", i.e., *spike*) dalla matrice di Wigner $\boldsymbol{\Xi}$. Il compito dell'inferenza è proprio quello di ricostruire lo spike nel modo più accurato possibile a partire dalla matrice $\boldsymbol{Y}$, contenente rumore, e dalla conoscenza del processo di generazione dei dati.

Osservazione 2. Il compito del prefattore $\frac{1}{\sqrt{n}}$ è rendere il problema di inferenza non banale: tenendo a mente il fatto che disponiamo di $\Theta(n^2)$ osservazioni, l'inferenza non è né banale né impossibile se il SNR è $\Theta(1)$. Con lo scaling adottato, il SNR totale è dato da:

$$\frac{\#\text{osservazioni} \times \text{SNR}_{\text{obs}}}{\#\text{parametri che inferiscono}} \quad \text{ossia} \quad \frac{\frac{n(n-1)}{2} \times \frac{\rho^2\lambda}{n}}{n} = \Theta(1), \tag{2.2}$$

dove nell'espressione precedente SNR_{obs} è il SNR della singola osservazione.

Supponiamo ora di essere nel regime Bayes-ottimale, dove la probabilità *a priori* P_x è conosciuta, così come la distribuzione del segnale di disturbo: in questo contesto possiamo calcolare la probabilità *a posteriori* sfruttando il fatto che la funzione di verosomiglianza è una misura gaussiana multivariata:

$$\begin{aligned} P(\boldsymbol{x} \mid \boldsymbol{y}) &\propto \prod_{i=1}^n P_x(x_i) \frac{1}{(2\pi)^{\frac{n(n-1)}{4}}} \exp\left[-\frac{1}{2}\sum_{1\le i<j\le n}\left(y_{ij} - \sqrt{\frac{\lambda}{n}}x_i x_j\right)^2\right] \\ &= \frac{1}{\mathcal{Z}_n(\boldsymbol{y})}\prod_{i=1}^n P_x(x_i)\exp(-\mathcal{H}(\boldsymbol{x};\boldsymbol{y})); \end{aligned} \tag{2.3}$$

dove $\mathcal{H}(\boldsymbol{x};\boldsymbol{y})$ e $\mathcal{Z}_n(\boldsymbol{y})$ sono rispettivamente denominate *hamiltoniana* e *funzione di partizione* del problema.

Definizione 2.1 (Hamiltoniana del modello di Wigner spiked)**.** La funzione Hamiltoniana del modello di Wigner spiked con dati $\boldsymbol{Y} = \sqrt{\frac{\lambda}{n}}\boldsymbol{X}^* \otimes \boldsymbol{X}^{*\intercal} + \boldsymbol{\Xi}$ è

$$\mathcal{H}(\boldsymbol{x};\boldsymbol{y}) = \sum_{i<j}\left(\frac{\lambda}{2n}x_i^2 x_j^2 - y_{ij}\sqrt{\frac{\lambda}{n}}x_i x_j\right). \tag{2.4}$$

Nel contesto in analisi, quindi, la funzione di partizione è data dalla seguente funzione:

$$\mathcal{Z}_n(\boldsymbol{y}) = \int \prod_{i=1}^{n} \mathrm{d}\, P_x(x_i)\, \mathrm{e}^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{y})}\,. \tag{2.5}$$

Si noti che $\mathcal{Z}_n(\boldsymbol{y})$ *non* è $P(\boldsymbol{y})$ in questa convenzione, ma piuttosto

$$\mathcal{Z}_n(\boldsymbol{y}) = (2\pi)^{\frac{n(n-1)}{2}}\, \mathrm{e}^{\frac{1}{2}\sum_{i<j} y_{ij}^2}\, P(\boldsymbol{y}). \tag{2.6}$$

Vale la seguente Proposizione, che sarà molto importante per stimare l'informazione mutua tra dati e segnale.

Proposizione 3. *L'informazione mutua tra dati e segnale può scriversi come*

$$\frac{1}{n} I\left(\boldsymbol{X}^*; \boldsymbol{Y}\right) = f_n(\lambda) + \frac{\lambda\rho^2(n-1)}{4n}, \tag{2.7}$$

dove $f_n(\lambda) \coloneqq -\frac{1}{n}\mathbb{E}[\ln \mathcal{Z}_n(\boldsymbol{Y})]$ *è l'energia libera media.*

Dimostrazione. Dalla definizione dell'informazione mutua in Eq. (1.8) e come conseguenza della Proposizione 1.3 si ha che:

$$\begin{aligned}
\frac{I(\boldsymbol{X}^*;\boldsymbol{Y})}{n} &= \frac{1}{n}\mathbb{E}_{\boldsymbol{X}^*,\boldsymbol{Y}}\left[\ln \frac{(2\pi)^{\frac{n(n-1)}{2}}\, \mathrm{e}^{\frac{1}{2}\sum_{i<j} Y_{ij}^2}\, P(\boldsymbol{Y} \mid \boldsymbol{X}^*)}{(2\pi)^{\frac{n(n-1)}{2}}\, \mathrm{e}^{\frac{1}{2}\sum_{i<j} Y_{ij}^2}\, P(\boldsymbol{Y})}\right] \\
&= \frac{1}{n}(\mathbb{E}_{\boldsymbol{X}^*,\boldsymbol{Y}}[\ln((2\pi)^{\frac{n(n-1)}{2}}\, \mathrm{e}^{\frac{1}{2}\sum_{i<j} Y_{ij}^2}\, P(\boldsymbol{Y} \mid \boldsymbol{X}^*))] - \mathbb{E}_{\boldsymbol{Y}}[\ln \mathcal{Z}_n(\boldsymbol{Y})]) \\
&= f_n + \frac{\mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}[\ln((2\pi)^{\frac{n(n-1)}{2}}\, \mathrm{e}^{\frac{1}{2}\sum_{i<j} Y_{ij}^2}\, P(\boldsymbol{Y} \mid \boldsymbol{X}^*))]}{n} \\
&= f_n + \frac{1}{n}\mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*}\left[\ln\left(\mathrm{e}^{\sum_{i<j} -\frac{\lambda {X_i^*}^2 {X_j^*}^2}{2n} + \sqrt{\frac{\lambda}{n}} Y_{ij} X_i^* X_j^*}\right)\right]
\end{aligned}$$

$$= f_n + \frac{1}{n}\frac{n(n-1)}{2}\mathbb{E}_{\boldsymbol{X}^*}\left[-\frac{\lambda X_i^{\star 2} X_j^{\star 2}}{2n} + \frac{\lambda X_i^{\star 2} X_j^{\star 2}}{n}\right] = f_n + \frac{(n-1)}{2}\mathbb{E}_{\boldsymbol{X}^*}\left[\frac{\lambda X_i^{\star 2} X_j^{\star 2}}{2n}\right]$$
$$= f_n + \frac{\lambda(n-1)\rho^2}{4n}. \quad (2.8)$$

□

In altre parole, l'informazione mutua tra il segnale e i dati, può essere stimata calcolando l'energia libera media.

2.2 Mutua informazione nel modello di Wigner *spiked*

Le grandezze precedentemente introdotte si possono calcolare per comprendere e prevedere il comportamento degli algoritmi. In particolare, è possibile utilizzare una euristica molto efficace che descriveremo in dettaglio.

Derivare una espressione per problemi in alta dimensione è spesso possibile nel caso in cui si trattino *modelli di campo medio*. Tali formule risultano dipendere da un' ottimizzazione in un numero limitato di *parametri scalari* pur riguardando problemi di ottimizzazione in alta-dimensione. Una tale riduzione da alta dimensione a bassa dimensione è una manifestazione della concentrazione di misura tipica di tali sistemi. Presentare una formulazione del risultato dato nel seguente teorema è uno degli obiettivi di questa tesi.

Teorema 2.1 (Mutua informazione)**.** *Siano $\boldsymbol{X}^* \in \mathbb{R}^n$ e X^* due variabili aleatorie, tali per cui le componenti di $\boldsymbol{X}^*$ siano i.i.d. con distribuzione P_x e varianza ρ, sia inoltre $Z \sim \mathcal{N}(0,1)$. L'informazione mutua per il modello di Wigner spiked verifica la seguente formula:*

$$\lim_{n\to\infty} \frac{1}{n} I(\boldsymbol{X}^*, \boldsymbol{Y}) = \inf_{q\in[0,\rho]} i^{(\mathrm{RS})}(q;\lambda,\rho) \quad (2.9a)$$

dove abbiamo introdotto il potenziale replica-simmetrico

$$i^{(\mathrm{RS})}(q;\lambda,\rho) \coloneqq \frac{\lambda}{4}(q-\rho)^2 + I\left(X^*; \sqrt{\lambda q}X^* + Z\right). \quad (2.9b)$$

Il teorema mostra che il calcolo dell'informazione mutua del modello in alta dimensione descritto in Eq. (2.1), è stato ridotto ad un problema di ottimizzazione *scalare*, sotto rumore gaussiano. Il termine $I(X^*; \sqrt{\lambda q}X^* + Z)$ che compare nell'espressione in Eq. (2.9), infatti, è l'informazione mutua di un altro problema di denoising, ovvero il problema associato al processo *scalare*

$$\tilde{Y} = \frac{1}{\sigma}X^* + Z \tag{2.10}$$

dove X^* è distribuito come P_x, $Z \sim \mathcal{N}(0,1)$ e $\sigma^{-1} = \sqrt{q\lambda}$.

Osservazione 3. La presenza del prefattore $\frac{1}{n}$ nella formula 2.9a fa in modo che il limite per $n \to \infty$ sia ben definito.

2.2.1 Derivazione per mezzo del trucco di replica

Per la classe dei modelli di campo medio, dove ogni spin interagisce con molti altri, come già accenato, esistono un gran numero di potenti metodi che sono in grado di effettuare tale riduzione di dimensione, come il *metodo delle repliche*. Il metodo si basa su una semplice identità che riduce il valor atteso di una quantità logaritmica al calcolo della media della funzione di partizione di k repliche, più facile da trattare. Questa strategia, si è rivelata essere molto versatile e può essere applicata in diversi contesti.

Proposizione 4. *Vale la seguente catena di identità:*

$$\mathbb{E}[\ln \mathcal{Z}(\boldsymbol{Y})] = \lim_{k\to 0^+} \frac{\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] - 1}{k} = \lim_{k\to 0^+} \frac{\partial}{\partial k} \ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] = \lim_{k\to 0^+} \frac{\ln \mathbb{E}\left[\mathcal{Z}(\boldsymbol{Y})^k\right]}{k} \tag{2.11}$$

dove abbiamo introdotto la funzione di partizione replicata

$$\mathcal{Z}(\boldsymbol{Y})^k := \int \mathrm{d}\, P(\boldsymbol{x}_1^k) \exp\Big(- \sum_{a=1}^{k} \mathcal{H}(\boldsymbol{x}^a; \boldsymbol{Y})\Big).$$

Dimostrazione. Sia $k \in \mathbb{R}$ in un intorno dell'origine; allora:

$$\mathcal{Z}(\boldsymbol{Y})^k = \mathrm{e}^{k \ln \mathcal{Z}(\boldsymbol{Y})} = 1 + k \ln \mathcal{Z}(\boldsymbol{Y}) + o(k)$$

da cui:

$$\ln \mathcal{Z}(\boldsymbol{Y}) = \lim_{k\to 0} \frac{\mathcal{Z}(\boldsymbol{Y})^k - 1}{k}$$

Scambiando il limite $k \to 0$ con il valore atteso, otteniamo:

$$\mathbb{E}[\ln \mathcal{Z}(\boldsymbol{Y})] = \mathbb{E}\left[\lim_{k\to 0} \frac{\mathcal{Z}(\boldsymbol{Y})^k - 1}{k}\right] = \lim_{k\to 0} \frac{\partial}{\partial k} \ln \mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k].$$

□

Il trucco di replica consiste nell'utilizzare $k \in \mathbb{N}$ per il calcolo di $\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]$, oggetto tipicamente più facile da trattare, ed eseguire in seguito un *prolungamento analitico* del risultato per ottenere, tramite la Proposizione precedente, $\mathbb{E}[\ln \mathcal{Z}(\boldsymbol{Y})]$: questo approccio ha successo qualora non vi siano punti di non-analiticità.

Torniamo ora al problema del calcolo della mutua informazione del modello di Wigner spiked. Usando la notazione $\int \mathrm{d}\, P(\boldsymbol{x}_0^k) \coloneqq \int_{\mathbb{R}^{nk}} \prod_{a=1}^k \prod_{i=1}^n P_x\left(x_i^a\right) \mathrm{d}\, x_i^a$, e ricordando l'espressione dell'hamiltoniana del modello in Eq. (2.4), abbiamo che:

$$\mathcal{Z}(\boldsymbol{y})^k = \int \mathrm{d}\, P(\boldsymbol{x}_0^k) \exp\Big(-\sum_{a=1}^k \mathcal{H}(\boldsymbol{x}^a; \boldsymbol{y})\Big)$$

è la funzione di partizione di k repliche. Di conseguenza, utilizzando il metodo delle repliche

$$\lim_{n\to\infty} f_n = -\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\ln \mathcal{Z}(\boldsymbol{Y})] = -\lim_{n\to\infty} \lim_{k\to 0^+} \frac{\ln \mathbb{E}\left[\mathcal{Z}(\boldsymbol{Y})^k\right]}{nk}. \tag{2.12}$$

per cui calcoliamo

$$\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] = \mathbb{E}_{\boldsymbol{X}^*}\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}^*} \int \mathrm{d}\, P(\boldsymbol{x}_0^k) \exp\left[\sum_{i<j}\left(Y_{ij}\sqrt{\frac{\lambda}{n}}\sum_{a=1}^k x_i^a x_j^a - \frac{\lambda}{2n}\sum_{a=1}^k (x_i^a x_j^a)^2\right)\right]. \tag{2.13}$$

Integriamo le variabili Y_{ij}, che hanno distribuzione condizionatamente gaussiana

$$Y_{ij} \mid \boldsymbol{X}^* \sim \mathcal{N}\left(\sqrt{\frac{\lambda}{n}} X_i^* X_j^*, 1\right).$$

Per procedere utilizziamo la formula di integrazione gaussiana (chiamata anche trasformazione di Hubbard-Stratonovich se letta da destra a sinistra)

$$\int_{\mathbb{R}} \mathrm{d}\, z\, \exp(-az^2 + bz) = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a}\right). \tag{2.14}$$

Otteniamo dunque:

$$\begin{aligned}\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] &= \mathbb{E}_{\boldsymbol{X}^*}\left[\int \mathrm{d}P(\boldsymbol{x}_0^k)\exp\left(\sum_{i<j}\left(\frac{\lambda}{n}\sum_{a=1}^{k} x_i^a X_i^* x_j^a X_j^* - \frac{\lambda}{2n}\sum_{a\neq b}^{1,k} x_i^a x_i^b x_j^a x_j^b\right)\right)\right] \\ &= \int \mathrm{d}P(\boldsymbol{x}_0^k)\exp\left(\sum_{i<j}\left(\frac{\lambda}{n}\sum_{a=1}^{k} x_i^a x_i^0 x_j^a x_j^0 - \frac{\lambda}{2n}\sum_{a\neq b}^{1,k} x_i^a x_i^b x_j^a x_j^b\right)\right),\end{aligned} \tag{2.15}$$

dove abbiamo indicato $\boldsymbol{x}^* \equiv \boldsymbol{x}^0$ per rendere più compatta l'espressione. L'espressione precedente può essere riscritta nel modo seguente:

$$\begin{aligned}\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] &= \int \mathrm{d}\, P(\boldsymbol{x}_0^k)\exp\left(\frac{\lambda}{2n}\sum_{a\neq b}^{0,k}\sum_{i<j} x_i^a x_i^b x_j^a x_j^b\right) \\ &= \int \mathrm{d}\, P(\boldsymbol{x}_0^k)\exp\left(\frac{\lambda}{4n}\sum_{a\neq b}^{0,k}\left(\left(\sum_{i=1}^{n} x_i^a x_i^b\right)^2 - \sum_{i=1}^{n}(x_i^a x_i^b)^2\right)\right).\end{aligned} \tag{2.16}$$

Introduciamo ora un *parametro d'ordine* $\boldsymbol{q} \in \mathbb{R}^{k(k+1)}$: dalla formula di Hubbard-Stratonovich introdotta sopra (con $a = n\lambda/4, b = (\frac{\lambda}{2})\sum_i x_i^a x_i^b$) possiamo scrivere

$$\exp\left(\frac{\lambda}{4n}\sum_{a\neq b}^{0,k}\left(\sum_{i=1}^{n} x_i^a x_i^b\right)^2\right) = \left(\frac{n\lambda}{4\pi}\right)^{\frac{k(k+1)}{2}}\int \mathrm{d}\boldsymbol{q}\,\exp\left(\sum_{a\neq b}^{0,k}\left(-\frac{n\lambda q_{ab}^2}{4} + \frac{\lambda q_{ab}}{2}\sum_{i=1}^{n} x_i^a x_i^b\right)\right). \tag{2.17}$$

Perciò la precedente espressione diventa:

$$\begin{aligned}&\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k] = \\ &= \left(\frac{n\lambda}{4\pi}\right)^{\frac{k(k+1)}{2}}\int d\boldsymbol{q}\exp\left(-\frac{n\lambda}{4}\sum_{a\neq b}^{0,k} q_{ab}^2\right)\int \mathrm{d}P(\boldsymbol{x}_0^k)\prod_{i=1}^{n}\exp\left(\sum_{a\neq b}^{0,k}\left(\frac{\lambda q_{ab}}{2}x_i^a x_i^b - \frac{\lambda}{4n}(x_i^a x_i^b)^2\right)\right) \\ &= \int \mathrm{d}\,\boldsymbol{q}\exp(-nS_n(\boldsymbol{q}))\end{aligned} \tag{2.18}$$

dove, indicando con $\mathrm{d}\,P(\boldsymbol{x}) \equiv \prod_{a=0}^{k}\mathrm{d}\,P_x(x^a)$

$$S_n(\boldsymbol{q}) := -\frac{k(k+1)}{2n}\ln\left(\frac{n\lambda}{4\pi}\right) + \frac{\lambda}{4}\sum_{a\neq b}^{0,k} q_{ab}^2 - \ln\int \mathrm{d}P(\boldsymbol{x})\exp\left(\sum_{a\neq b}^{0,k}\left(\frac{\lambda q_{ab}}{2}x^a x^b - \mathcal{O}(1/n)\right)\right). \tag{2.19}$$

Supponendo ora che i limiti nell'espressione in Eq. (2.12) commutino, ovvero

$$-\lim_{n\to\infty}\lim_{k\to 0_+}\frac{\ln\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk} = -\lim_{k\to 0_+}\lim_{n\to\infty}\frac{\ln\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk}, \tag{2.20}$$

a k fissato, il limite in n potrà essere stimato grazie al metodo del punto sella

$$-\lim_{n\to\infty}\frac{\ln\mathbb{E}[\mathcal{Z}(\boldsymbol{Y})^k]}{nk}=\min_q\frac{S(q)}{k} \tag{2.21}$$

dove $S(q)\coloneqq\lim_{n\to\infty}S_n(q)$. In particolare, così facendo, da un integrale mi sono ridotta ad un limite. Nell'espressione precedente è stata fatta un' importante assunzione, ovvero $\boldsymbol{q}$ si suppone tale che $q_{ab}=q$ per tutti gli $a\neq b$. Questa assunzione, detta *di simmetria delle repliche*, mi permette di semplificare i calcoli e in particolare scrivere

$$\frac{S(q)}{k}\coloneqq(k+1)\frac{\lambda q^2}{4}-\frac{1}{k}\ln\int\mathrm{d}\,P(\boldsymbol{x})\exp\left(\frac{\lambda q}{2}\left(\left(\sum_{a=0}^k x^a\right)^2-\sum_{a=0}^k(x^a)^2\right)\right). \tag{2.22}$$

Usiamo ancora una volta la formula di Hubbard-Stratonovich (con $a=1/2$ e $b=\sqrt{\lambda q}\sum_{a=0}^k x^a$) per disaccoppiare le repliche, e sia $Z\sim\mathcal{N}(0,1)$, otteniamo:

$$\frac{S(q)}{k}=(k+1)\frac{\lambda q^2}{4}-\frac{1}{k}\ln\mathbb{E}\int\prod_{a=0}^k\mathrm{d}\,P_x(x^a)\exp\left(\sqrt{\lambda q}Zx^a-\frac{\lambda q}{2}(x^a)^2\right). \tag{2.23}$$

Le repliche sono ora disaccoppiate. Sia $X^0=X^*\sim P_x$. Mandando $k\to 0_+$, e attraverso un cambio di variabile dalla terza alla quarta riga ($z\to z-\sqrt{\lambda q}x^*$):

$$\begin{aligned}
&\frac{1}{k}\ln\mathbb{E}\int\prod_{a=0}^k\mathrm{d}\,P_x(x^a)\exp\left(\sqrt{\lambda q}Zx^a-\frac{\lambda q}{2}(x^a)^2\right)\\
&\quad=\frac{1}{k}\ln\int\frac{\mathrm{d}\,z}{\sqrt{2\pi}}\,\mathrm{d}\,P_x(x^0)\,\mathrm{e}^{-\frac{z^2}{2}+\sqrt{\lambda q}zx^0-\frac{\lambda q}{2}(x^0)^2}\left(\int\mathrm{d}\,P_x(x)\,\mathrm{e}^{\sqrt{\lambda q}zx-\frac{\lambda q}{2}x^2}\right)^k\\
&\quad=\frac{1}{k}\ln\int\frac{\mathrm{d}\,z}{\sqrt{2\pi}}\,\mathrm{d}\,P_x(x^*)\,\mathrm{e}^{-\frac{1}{2}(z-\sqrt{\lambda q}x^*)^2}\left(\int\mathrm{d}\,P_x(x)\,\mathrm{e}^{\sqrt{\lambda q}zx-\frac{\lambda q}{2}x^2}\right)^k\\
&\quad=\frac{1}{k}\ln\mathbb{E}\left[\left(\int\mathrm{d}\,P_x(x)\,\mathrm{e}^{\sqrt{\lambda q}Zx+\lambda qxX^*-\frac{\lambda q}{2}x^2}\right)^k\right]\\
&\quad=\mathbb{E}\ln\int\mathrm{d}\,P_x(x)\exp\left(\sqrt{\lambda q}Zx+\lambda qxX^*-\frac{\lambda q}{2}x^2\right)+\mathcal{O}(k).
\end{aligned} \tag{2.24}$$

Alla fine ottengo:

$$\lim_{n\to\infty}f_n(\lambda)=\min_q\left[\frac{\lambda q^2}{4}-\mathbb{E}\ln\int\mathrm{d}\,P_x(x)\exp\left(\sqrt{\lambda q}Zx+\lambda qxX^*-\frac{\lambda q}{2}x^2\right)\right] \tag{2.25a}$$

ovvero

$$\lim_{n\to+\infty}\frac{1}{n}I(\boldsymbol{X}^*,\boldsymbol{Y})=\min_q\left[\frac{\lambda q^2}{4}-\mathbb{E}\ln\int \mathrm{d}P_x(x)\exp\left(\sqrt{\lambda q}Zx+\lambda q x X^*-\frac{\lambda q}{2}x^2\right)\right]+\frac{\lambda\rho^2}{4}. \tag{2.25b}$$

Ci siamo ridotti perciò a trovare il minimo di un'espressione in una variabile *scalare*. Per interpretare meglio l'argomento del minimo nella precedente espressione, consideriamo il problema in Eq. (2.10), scegliendo $\sigma^{-1}=\sqrt{\lambda q}$

$$\tilde{Y}=\sqrt{\lambda q}X^*+Z$$

con $Z\sim\mathcal{N}(0,1)$ e $X^*\sim P_x$, mentre $\sigma>0$. La corrispondente distribuzione a posteriori di X dato Y in tale problema si può quindi scrivere nel modo seguente:

$$P_0(x\mid y)=\frac{1}{\mathcal{Z}_0(y)}P_x(x)\exp\left(-\frac{\lambda q x^2}{2}+\sqrt{\lambda q}xy\right)=\frac{1}{\mathcal{Z}_0(y)}P_x(x)\exp(-\mathcal{H}_0(x;y))$$

dove abbiamo introdotto l'hamiltoniana

$$\mathcal{H}_0(x;y)=\frac{\lambda q x^2}{2}-\sqrt{\lambda q}xy \tag{2.26}$$

e il fattore di normalizzazione

$$\mathcal{Z}_0(y)\coloneqq\int P_x(x)\,\mathrm{e}^{-\mathcal{H}_0(x;y)}\,\mathrm{d}\,x. \tag{2.27}$$

Per questo problema di denoising, l'energia libera media è:

$$f_0(\lambda q)\coloneqq-\mathbb{E}[\ln\mathcal{Z}_0(\tilde{Y})] \tag{2.28}$$

di modo che, adattando la relazione tra l'informazione mutua ed energia libera come in Eq. (2.7) al caso scalare, si ha:

$$I(X^*;\sqrt{\lambda q}X^*+Z)=f_0(\lambda q)+\frac{\rho\lambda q}{2}. \tag{2.29}$$

Ma l'espressione per $I(X^*;\sqrt{\lambda q}X^*+Z)$ coincide esattamente con l'argomento del minimo in Eq. (2.25). Dunque mettendo tutto insieme, dobbiamo provare che:

$$\lim_{n\to+\infty}f_n(\lambda)=\Phi_{\mathrm{RS}}(\lambda)\qquad\text{con}\quad\Phi_{\mathrm{RS}}(\lambda)\coloneqq\min_q\left[f_0(\lambda q)+\frac{\lambda q^2}{4}\right]. \tag{2.30}$$

Aggiungendo il termine mancante $\frac{1}{4}\rho^2\lambda$, come previsto dal legame tra l'energia libera e l'informazione mutua dimostrata sopra, ritroviamo l'espressione in Eq. (2.9) del Teorema 2.1.

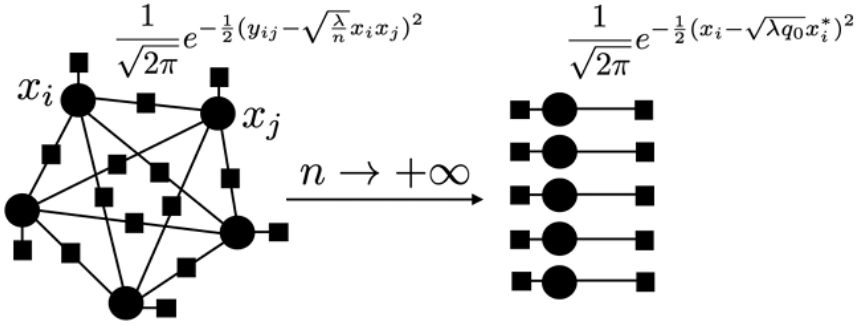

Figura 2.1: La figura, estratta dalla Ref. [1], rappresenta ciò che accade nel limite termodinamico nel problema di Wigner *spiked*: a destra è rappresentato il modello originale mentre a sinistra il problema disaccoppiato.

2.3 Concentrazione della mutua informazione

In questa sezione verrà dimostrata la concentrazione dell'enegia libera, e quindi della mutua informazione, la cui espressione è stata derivata euristicamente grazie al metodo delle repliche. A questo scopo seguiremo la strategia descritta in Ref. [3]. Per $n \to \infty$, infatti, è possibile ottenere un maggiorante e un minorante per $\lim_n \frac{1}{n} I(\boldsymbol{X}^*, \boldsymbol{Y})$, e mostrare che queste due quantità effettivamente coincidono. Questo fatto risulta particolarmente interessante perchè giustifica l'analisi *in media* fatta sopra nello studio di certi modelli probabilistici con un grande numero di parametri; l'energia libera, a meno di piccole correzioni che si annullano quando $n \to \infty$, si concentra sempre sullo stesso valore, ovvero il suo valore medio. Proprio questa informazione rende "leciti" i vari passaggi precedentemente mostrati.

Introduciamo, prima di iniziare, una notazione che utilizzeremo nelle dimostrazioni successive.

Definizione 2.2 (Media di Gibbs)**.** Siano $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \dots, \boldsymbol{X}^{(r)}$ r copie statistiche, anche dette *repliche*, della variabile n-dimensionale $\boldsymbol{X}^*$, che si assume avere misura di probabilità P e Hamiltoniana $\mathcal{H}$. In altre parole, r quantità distribuite come $\boldsymbol{X}^*$ indipendentemente tra loro, e sia $g\colon (\mathbb{R}^n)^{(r+1)} \to \mathbb{R}$ una qualsiasi funzione. Si

definisce la media di Gibbs di g la quantità

$$\langle g(\boldsymbol{X}^{(1)},\dots,\boldsymbol{X}^{(r)},\boldsymbol{X}^{*})\rangle \coloneqq \frac{\int g(\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(r)},\boldsymbol{X}^{*})\prod_{l=1}^{r}\mathrm{e}^{-\mathcal{H}(\boldsymbol{x}^{(l)})}\,\mathrm{d}\,P(\boldsymbol{x}^{(l)})}{\left(\int \mathrm{e}^{-\mathcal{H}(\boldsymbol{x})}\,\mathrm{d}\,P(\boldsymbol{x})\right)^{r}}.$$

Introduciamo, per cominciare, un *maggiorante* per l'energia libera.

Proposizione 5 (Interpolazione di Guerra)**.** *Sia $t \in [0,1]$ e sia q una variabile non-negativa. Si consideri la seguente Hamiltoniana interpolante il caso Spiked Wigner e il caso lineare:*

$$-\mathcal{H}_t(\boldsymbol{x};\boldsymbol{y},\tilde{\boldsymbol{y}}) \coloneqq \sqrt{\frac{\lambda t}{n}}\left(\sum_{i<j} x_i x_j y_{ij} - \frac{1}{2}\sqrt{\frac{t\lambda}{n}}x_i^2 x_j^2\right) + \sqrt{(1-t)\lambda q}\left(\sum_{i=1}^{n} x_i \tilde{y}_i - \frac{\sqrt{(1-t)\lambda q}}{2}x_i^2\right) \tag{2.31}$$

e i seguenti dati osservati:

$$\begin{cases} Y_{ij} = \sqrt{\frac{t\lambda}{n}}X_i^* X_j^* + \Xi_{ij}, \quad 1 \le i \le j \le n \\ \tilde{Y}_i = \sqrt{(1-t)\lambda q}X_i^* + Z_i, \quad 1 \le i \le j \le n \end{cases}$$

dove $X_i^* \sim P_x$, mentre $Z_i \sim \mathcal{N}(0,1)$ $\forall i = 1,\dots n$ e $\Xi_{ij} \sim \mathcal{N}(0,1)$ $\forall i,j = 1,\dots,n$. Allora esiste una costante positiva K tale per cui

$$f_n(\lambda) \le \Phi_{\mathrm{RS}}(\lambda) + \frac{K}{n}$$

con $\Phi_{\mathrm{RS}}(\lambda)$ definita precedentemente nell'Eq. (2.30).

Dimostrazione. Si consideri la seguente funzione:

$$\phi(t) \coloneqq -\frac{1}{n}\mathbb{E}\ln\int \mathrm{e}^{-\mathcal{H}_t(\boldsymbol{x};\boldsymbol{Y},\tilde{\boldsymbol{Y}})}\prod_{i=1}^{n}\mathrm{d}\,P_x(x_i)$$

e sia $R_{ab} \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i^{(a)}X_i^{(b)}$ per ogni coppia di repliche $\boldsymbol{X}^{(a)}$, $\boldsymbol{X}^{(b)}$, $a,b = 1,\dots,r$. Si dimostra grazie ad un'integrazione gaussiana per parti [7, pagine 147-149] che

$$\begin{aligned}\phi'(t) = \frac{\lambda}{4}\mathbb{E}\langle(R_{1\,2}-q)^2\rangle_t - \frac{\lambda}{4}q^2 - \frac{\lambda}{4n^2}\sum_{i=1}^{n}\mathbb{E}\left\langle\left(X_i^{(1)}X_i^{(2)}\right)^2\right\rangle_t \\ - \frac{\lambda}{2}\mathbb{E}\langle(R_{1*}-q)^2\rangle_t + \frac{\lambda}{2}q^2 + \frac{\lambda}{2n^2}\sum_{i=1}^{n}\mathbb{E}\left\langle\left(X_i^{(1)}X_i^{*}\right)^2\right\rangle_t\end{aligned} \tag{2.32}$$

dove $\langle \bullet \rangle_t$ è *la media di Gibbs* in Definizione 2.2 e $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{X}^*, \boldsymbol{Z}, \boldsymbol{\Xi}}$. Ora utilizzando l'identità di Nishimori, osserviamo che le coppie $(X_i^{(1)}, X_i^*)$ e $(X_i^{(1)}, X_i^{(2)})$ hanno la stessa distribuzione congiunta per ogni $i = 1, \dots, n$, quindi l'espressione precedente diventa:

$$\phi'(t) = -\frac{\lambda}{4}\mathbb{E}\langle (R_{1*} - q)^2 \rangle_t + \frac{\lambda}{4}q^2 + \frac{\lambda}{4n^2}\sum_{i=1}^{n}\mathbb{E}\langle X_i^2 X_i^{*2}\rangle_t.$$

Le componenti di X^* (e quindi anche delle repliche) sono limitate per ipotesi 2.1, in modo che $\mathbb{E}\langle X_i^2 X_i^{*2}\rangle_t$ sia finito, dunque l'ultimo termine è $\mathcal{O}(1/n)$, mentre il primo è sempre negativo quindi si ottiene:

$$\phi'(t) \leq \frac{\lambda}{4}q^2 + \frac{K}{n}$$

per qualche costante positiva K.
Integriamo sia a destra che a sinistra su t e procediamo con il *teorema fondamentale del calcolo integrale*:

$$\int_0^1 \phi'(t)\,\mathrm{d}\,t \leq \int_0^1 \left(\frac{\lambda}{4}q^2 + \frac{K}{n}\right)\mathrm{d}\,t. \tag{2.33}$$

Poiché $\phi(1) = f_n(\lambda)$ e $\phi(0) = f_0(\lambda q)$, si ottiene che per ogni $q \geq 0$ vale

$$f_n(\lambda) \leq f_0(\lambda q) + \frac{\lambda q^2}{4} + \frac{K}{n}$$

da cui segue la tesi. □

Avendo derivato un maggiorante, l'obiettivo è ora cercare un minorante per l'energia libera: a questo scopo introduciamo il potenziale di Franz–Parisi, che ci aiuterà in una stima dal basso.

Definizione 2.3 (Potenziale di Franz–Parisi)**.** Sia $m \in \mathbb{R}$ ed $\epsilon > 0$. Il potenziale di Franz–Parisi è definito come:

$$\Phi_\epsilon^n(m; \boldsymbol{X}^*) := -\frac{1}{n}\mathbb{E}_{\boldsymbol{\Xi}} \ln \int \mathbb{1}\{R_{1,*} \in [m, m+\epsilon)\}\, \mathrm{e}^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Y})} \prod_{i=1}^{n} \mathrm{d}\,P_x(x_i). \tag{2.34}$$

Il potenziale di Franz–Parisi, può essere pensato come il lavoro minimo per mantenere il sistema ad una certa distanza dalla configurazione di equilibrio. Questa definizione appena introdotta, può essere essere pensata come una funzione che

dipende dal tempo, in modo da descrivere le proprietà termodinamiche del sistema. Enunciamo a questo punto un lemma ausiliario, che non dimostriamo e che ci servirà poi nella dimostrazione della Proposizione 6.

Lemma 1. *Sia*

$$F_\ell(\boldsymbol{y}) \coloneqq \frac{1}{n} \ln \int \mathbb{1}\{R_{1,*} \in [\ell\epsilon, (\ell+1)\epsilon)\}\, \mathrm{e}^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{y})} \prod_{i=1}^{n} \mathrm{d}\, P_x(x_i) \quad \textit{con } \epsilon > 0; \tag{2.35}$$

esiste una costante $K > 0$ tale che per ogni $\gamma \geq 0$ e per ogni ℓ:

$$\mathbb{E}_{\boldsymbol{\Xi}} \left[\mathrm{e}^{\gamma (F_\ell(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})])} \right] \leq \frac{K\gamma}{\sqrt{n}}\, \mathrm{e}^{K\gamma^2/n} \tag{2.36}$$

Questo Lemma permette la dimostrazione della seguente Proposizione.

Proposizione 6. *Esiste $K > 0$ tale che per ogni $\epsilon > 0$, si ha:*

$$f_n \geq \mathbb{E}_{\boldsymbol{X}^*} \left[\min_{\substack{\ell \in \mathbb{Z} \\ |\ell|\epsilon \leq K}} \Phi_\epsilon^n(\ell\epsilon, \boldsymbol{X}^*) \right] - \frac{\ln(K/\epsilon)}{\sqrt{n}}.$$

Dimostrazione. Possiamo suddividere l'insieme dei valori di sovrapposizione $R_{1,*}$ in $2K/\epsilon$ intervalli di ampiezza ϵ per un qualche $K > 0$, essendo il supporto del segnale limitato. Questo permette la seguenze discretizzazione, dove ℓ varia sull'insieme $\{-K/\epsilon, \dots, K/\epsilon\}$:

$$\begin{aligned} -f_n &= \frac{1}{n} \mathbb{E} \ln \sum_\ell \int \mathbb{1}\{R_{1,*} \in [\ell\epsilon, (\ell+1)\epsilon)\}\, \mathrm{e}^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Y})} \prod_{i=1}^{n} \mathrm{d}\, P_x(x_i) \\ &\leq \frac{1}{n} \mathbb{E} \ln \frac{2K}{\epsilon} \max_\ell \int \mathbb{1}\{R_{1,*} \in [\ell\epsilon, (\ell+1)\epsilon)\}\, \mathrm{e}^{-\mathcal{H}(\boldsymbol{x};\boldsymbol{Y})} \prod_{i=1}^{n} \mathrm{d}\, P_x(x_i) \\ &= \mathbb{E}\Big[\max_\ell F_\ell(\boldsymbol{Y}) \Big] + \frac{\ln(2K/\epsilon)}{n}. \end{aligned} \tag{2.37}$$

Nella precedente espressione abbiamo introdotto la quantità $F_\ell(\boldsymbol{y})$ data in Eq. (2.35). Per via del Lemma 1, ogni termine $F_\ell(\boldsymbol{Y})$, si concentra sul suo valore atteso $\mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})]$ per $n \to +\infty$, dove l'aspettazione è rispetto al solo rumore Ξ in $\boldsymbol{Y}$. Dato che tutte le F_ℓ si concentrano, anche il valor atteso del massimo si concentra:

$$\mathbb{E}_{\boldsymbol{\Xi}} \Big[\max_\ell (F_\ell(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})]) \Big] \leq \frac{1}{\gamma} \ln \mathbb{E}_{\boldsymbol{\Xi}} \exp \left(\gamma \max_\ell (F_\ell(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})]) \right)$$

$$= \frac{1}{\gamma} \ln \mathbb{E}_{\boldsymbol{\Xi}} \max_{\ell} \mathrm{e}^{\gamma(F_\ell(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})])} \leq \frac{1}{\gamma} \ln \mathbb{E}_{\boldsymbol{\Xi}} \sum_{\ell} \mathrm{e}^{\gamma(F_\ell(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})])}$$
$$\leq \frac{1}{\gamma} \ln \left(\frac{2K}{\epsilon} \frac{\gamma K}{\sqrt{n}} \mathrm{e}^{\gamma^2 K/n} \right) = \frac{\ln(2K/\epsilon)}{\gamma} + \frac{1}{\gamma} \ln \frac{\gamma K}{\sqrt{n}} + \frac{\gamma K}{n}. \quad (2.38)$$

Imponendo $\gamma = \sqrt{n}$, si ha:

$$\mathbb{E}_{\boldsymbol{\Xi}} \left[\max_{\ell} (F_\ell(\boldsymbol{Y}) - \mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})]) \right] \leq \frac{\ln(K/\epsilon)}{\sqrt{n}}.$$

Inserendo questa stima nell'Eq. (2.37), si ottiene:

$$-f_n \leq \mathbb{E}_{\boldsymbol{X}^*} \max_{\ell} \mathbb{E}_{\boldsymbol{\Xi}}[F_\ell(\boldsymbol{Y})] + \frac{\ln(K/\epsilon)}{\sqrt{n}} + \frac{\ln(K/\epsilon)}{n} \leq \mathbb{E}_{\boldsymbol{X}^*} \left[\max_{\ell} \Phi_\epsilon^n(\ell\epsilon, \boldsymbol{X}^*) \right] + 2\frac{\ln(K/\epsilon)}{\sqrt{n}}.$$

Quindi:

$$f_n \geq \mathbb{E}_{\boldsymbol{X}^*} \left[\min_{\ell} \Phi_\epsilon^n(\ell\epsilon, \boldsymbol{X}^*) \right] - \frac{\ln(K/\epsilon)}{\sqrt{n}}$$

per qualche costante K, come nella tesi.

□

Per ottenere il risultato finale, otteniamo ora un minorante per il potenziale di Franz–Parisi.

Proposizione 7 (Minorante per il potenziale di Franz–Parisi)**.** *Sia* $\Phi_\epsilon^n(q; \boldsymbol{X}^*)$ *il potenziale di Franz–Parisi 2.34 per il nostro problema, con* $q \in \mathbb{R}$ *e* $\epsilon > 0$*. Esiste* $K > 0$ *tale che per qualsiasi* $\epsilon > 0$ *si ha:*

$$\Phi_\epsilon^n(q; \boldsymbol{X}^*) \geq f_0(\lambda q) + \frac{\lambda q^2}{4} - \frac{\lambda}{2}\epsilon^2 + \frac{K}{n}. \quad (2.39)$$

Dimostrazione. Si consideri la seguente Hamiltoniana interpolante, analoga a quella data in Eq. (2.31) a meno di un addendo:

$$-\mathcal{H}_t(\boldsymbol{x}; \boldsymbol{y}, \tilde{\boldsymbol{y}}) := \sqrt{\frac{\lambda t}{n}} \left(\sum_{i<j} x_i x_j y_{ij} - \frac{1}{2} \sqrt{\frac{t\lambda}{n}} x_i^2 x_j^2 \right) + \sqrt{(1-t)\lambda q} \left(\sum_{i=1}^n x_i \tilde{y}_i - \frac{\sqrt{(1-t)\lambda q}}{2} x_i^2 \right)$$
$$+ \lambda(1-t)(m-q) \sum_{i=1}^n x_i X_i^*. \quad (2.40)$$

Sia inoltre

$$\phi_{\epsilon,m}(t;\boldsymbol{X}^*) := -\frac{1}{n}\mathbb{E}_{\boldsymbol{\Xi}} \ln \int \mathrm{e}^{-\mathcal{H}_t(\boldsymbol{x};\boldsymbol{Y},\tilde{\boldsymbol{Y}})}\, \mathbb{1}\{R_{1,*} \in [m, m+\epsilon)\} \prod_{i=1}^{n} \mathrm{d}\, P_x(x_i). \quad (2.41)$$

Denotando, come prima, la media di Gibbs in Definizione 2.2 (con il vincolo aggiuntivo $\mathbb{1}\{R_{1,*} \in [m, m+\epsilon)\}$), come $\langle \bullet \rangle_t^{m,\epsilon}$ [1], indicando per brevità $\mathbb{E} \equiv \mathbb{E}_{\boldsymbol{\Xi}}$ e sviluppando i calcoli come nella Ref. [7, pagine 143-146], otteniamo:

$$\phi'_{\epsilon,m}(t;\boldsymbol{X}^*) = \frac{\lambda}{4}\mathbb{E}\langle (R_{1,2}-q)^2 \rangle_t^{m,\epsilon} - \frac{\lambda}{4}q^2 + \frac{\lambda}{2}m^2 - \frac{\lambda}{2}\mathbb{E}\langle (R_{1,*}-m)^2 \rangle_t^{m,\epsilon} + o(1).$$

Il trucco è notare che, per costruzione, data la restrizione delle sovrapposizioni delle repliche, si ha: $\mathbb{E}\langle (R_{1,*}-m)^2 \rangle_t^{m,\epsilon} \leq \epsilon^2$. Perciò

$$\begin{aligned}\phi'_{\epsilon,m}(t;\boldsymbol{X}^*) &\geq \frac{\lambda}{4}\mathbb{E}\langle (R_{1,2}-q)^2 \rangle_t^{m,\epsilon} - \frac{\lambda}{4}q^2 + \frac{\lambda}{2}m^2 - \frac{\lambda\epsilon^2}{2} + o(1) \\ &\geq -\frac{\lambda q^2}{4} + \frac{\lambda}{2}m^2 - \frac{\lambda\epsilon^2}{2} + o(1). \quad (2.42)\end{aligned}$$

È noto inoltre che $\phi_{\epsilon,m}(1;\boldsymbol{X}^*) = \Phi_\epsilon^n(m,\boldsymbol{X}^*)$ e $\phi_{\epsilon,m}(0;\boldsymbol{X}^*) \leq f_0(\lambda q)$ (anche questo per restrizione della sovrapposizione delle repliche). Dunque integrando su t, ponendo $m = q$ e applicando il *teorema fondamentale del calcolo integrale*, si ottiene la tesi:

$$\begin{aligned}\int_0^1 \phi'_{\epsilon,m}(t;\boldsymbol{X}^*)\,\mathrm{d}\,t \geq \int_0^1 \left(-\frac{\lambda q^2}{4} + \frac{\lambda}{2}m^2 - \frac{\lambda\epsilon^2}{2} + o(1)\right)\mathrm{d}\,t \Rightarrow \\ \Phi_\epsilon^n(q;\boldsymbol{X}^*) - f_0(\lambda q) \geq \frac{\lambda q^2}{4} - \frac{\lambda}{2}\epsilon^2 + \frac{K}{n}. \quad (2.43)\end{aligned}$$

□

Combinando la Proposizione 6 con l'equazione (2.39), si ha

$$f_n \geq f_0(\lambda q) + \frac{\lambda q^2}{4} - \frac{\lambda\epsilon^2}{2} - \frac{\ln(K/\epsilon)}{\sqrt{n}}.$$

In particolare, scegliendo $\epsilon = n^{-1/2}$ si ottiene il seguente teorema.

[1] Nelle stesse notazioni della Definizione 2.2:

$$\langle g(\boldsymbol{X}^{(1)},\dots,\boldsymbol{X}^{(r)},\boldsymbol{X}^*)\rangle_t^{m,\epsilon} := \frac{\int g(\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(r)},\boldsymbol{X}^*)\prod_{l=1}^r \mathrm{e}^{-\mathcal{H}(\boldsymbol{x}^{(l)})}\mathrm{d}P(\boldsymbol{x}^{(l)})\mathbb{1}\{R_{1,*}\in[m,m+\epsilon)\}}{\left(\int \mathrm{e}^{-\mathcal{H}(\boldsymbol{x})}\mathrm{d}P(\boldsymbol{x})\mathbb{1}\{R_{1,*}\in[m,m+\epsilon)\}\right)^r}.$$

Teorema 2.2 (Minorante dell'energia libera)**.** *Nelle stesse ipotesi del Teorema 2.1, vale, per una qualche costante positiva K, che:*

$$f_n(\lambda) \geq \Phi_{\mathrm{RS}}(\lambda) - \frac{\lambda}{2n} - \frac{\ln(K\sqrt{n})}{\sqrt{n}}.$$

Possiamo a questo punto combinare minorante e maggiorante, osservando che per $n \to \infty$ questi coincidono, e riproducono l'espressione in Eq. (2.25) ottenuta con il metodo euristico.

2.3.1 Relazione tra informazione mutua e MMSE

In questa sezione, come conseguenza del teorema 2.1, metteremo in evidenza un importante risultato che lega l'informazione mutua e l'MMSE, definito e studiato nel capitolo precedente (1.5).

Corollario 1. *Nelle stesse ipotesi del teorema 2.1, e per ogni (λ, ρ) tale per cui il minimo del potenziale di replica-simmetrico (2.9) sia unico:*

$$q_0(\lambda) := \arg\min_{q \in [0,\rho]} i^{(\mathrm{RS})}(q; \lambda, \rho),$$

l'MMSE della matrice del segnale verifica:

$$\lim_{n\to\infty} \frac{1}{n^2} \mathbb{E} \|\boldsymbol{X}^* \otimes \boldsymbol{X}^* - \langle \boldsymbol{X} \otimes \boldsymbol{X} \rangle\|_F^2 = \rho^2 - q_0(\lambda)$$

dove ricordiamo che $\|\cdot\|_F$ è la norma di Frobenius e $\langle \boldsymbol{X} \otimes \boldsymbol{X} \rangle = \mathbb{E}[\boldsymbol{X} \otimes \boldsymbol{X} \mid \boldsymbol{y}]$ è la media a posteriori dello spike (lo stimatore MMSE).

Per la dimostrazione di questo corollario consultare la Ref. [1, Pagine 35-36]

2.4 PCA e la transizione BBP

L'efficacia dei metodi algoritmici utilizzati in problemi di inferenza è strettamente legata all'informazione reciproca tra dati e parametri, ovvero nel nostro caso alle proprietà spettrali delle matrici che appaiono nel problema. In questo capitolo ci occuperemo di fornire una descrizione qualitativa di alcuni aspetti algoritmici

del problema, o per meglio dire, come algoritmi si comportano in contesti *Bayes-ottimali*, ovvero informati riguardo la struttura probabilistica del problema, nella ricostruzione del segnale $\boldsymbol{X}^*$ del modello in Eq. (2.1).

Nei problemi di stima di matrici a basso-rango, per esempio nel caso in cui si vuole determinare la matrice più vicina ad una osservabile $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ simmetrica e a basso rango r, viene largamente utilizzato il metodo PCA (*Principal Component Analysis*): si cerca, tra tutte le matrici di rango r, la matrice $\hat{\boldsymbol{Y}}$ che minimizza la norma di Frobenius[2] di $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$, tramite la seguente relazione:

$$\hat{\boldsymbol{Y}} = \arg \min_{\boldsymbol{\Upsilon} \text{ rango } r} \|\boldsymbol{Y} - \boldsymbol{\Upsilon}\|_{\mathrm{F}}^2 = \sum_{s=1}^{r} \lambda_s \boldsymbol{X}_s \otimes \boldsymbol{X}_s^{\intercal}, \tag{2.44}$$

dove gli $\boldsymbol{X}_s \in \mathbb{R}^n$ sono gli autovettori (normalizzati) e i λ_s gli autovalori della matrice $\boldsymbol{Y}$. È ragionevole quindi tentare di risolvere il nostro problema di ricostruzione del segnale (come nel modello di Wigner *spiked*) utilizzando la PCA, una strategia largamente conosciuta e adottata in diversi settori di ricerca.

In questo ultima sezione, seguendo la Ref. [4, Sezione 4.1.3] e la Ref. [6, Sezioni 3.2-3.3.2], ci occuperemo di rispondere alla seguente domanda: quanto sono efficaci i metodi spettrali come PCA per ricostruire il segnale nel modello di Wigner *spiked*? Come sappiamo, il modello ha la forma

$$\boldsymbol{Y} = \sqrt{\frac{\lambda}{n}} \boldsymbol{X}^* \otimes \boldsymbol{X}^{*\intercal} + \boldsymbol{\Xi} \tag{2.45}$$

dove $\Xi_{ij} \sim \mathcal{N}(0,1)$. Il problema può essere visto come una sorta di "competizione" tra la matrice di rumore $\boldsymbol{\Xi}$ e il contributo del segnale: di conseguenza, al variare del valore di λ, si osserva un cambiamento dello spettro della matrice $\boldsymbol{Y}$. In particolare, per $\lambda = 0$ lo spettro della matrice $\boldsymbol{Y} = \boldsymbol{\Xi}$ è noto, poiché in questo caso essa è una *matrice di Wigner*, ovvero una matrice simmetrica con componenti avente distribuzione gaussiana $\mathcal{N}(0,1)$. Le matrici di Wigner hanno una densità spettrale limite, tale da seguire la *legge del semicerchio di Wigner*: per $n \to \infty$, la densità degli autovalori di $\frac{\boldsymbol{\Xi}}{\sqrt{n}}$ converge ad un semicerchio centrato in 0 e di raggio

[2] Data una matrice $\boldsymbol{A} = (a_{ij})_{ij} \in \mathbb{R}^{m\times n}$, la sua norma di Frobenius è $\|\boldsymbol{A}\|_{\mathrm{F}} := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$.

$R = 2$, la cui distribuzione di probabilità associata ha una densità pari a:

$$f(x) = \frac{\sqrt{4 - x^2}}{2\pi}. \tag{2.46}$$

La presenza dello "spike" fa sì che un autovalore aggiuntivo compaia nello spettro.

Una misura di quanto correttamente il segnale può essere ricostruito, tramite il metodo PCA, è dato dalla seguente quantità:

$$\alpha(\boldsymbol{X}^*, \hat{\boldsymbol{X}}_{\text{PCA}}) = \frac{|\boldsymbol{X}^{*\intercal} \hat{\boldsymbol{X}}_{\text{PCA}}|}{\|\boldsymbol{X}^{*\intercal}\| \|\hat{\boldsymbol{X}}_{\text{PCA}}\|}$$

dove $0 \leq \alpha(\boldsymbol{X}^*, \hat{\boldsymbol{X}}_{PCA}) \leq 1$, di modo che se $\alpha(\boldsymbol{X}^*, \hat{\boldsymbol{X}}_{PCA}) = 1$ si ha una perfetta ricostruzione, e dove $\hat{\boldsymbol{X}}_{\text{PCA}}$ è il risultato di ricostruzione del segnale tramite PCA. In particolare si hanno tre regimi:

- Nel limite $\lambda \to +\infty$, il contributo di $\boldsymbol{\Xi}$ è trascurabile e $\boldsymbol{Y}$ ha solo un autovalore non nullo, avente come autovettore il segnale, quindi PCA ricostuisce perfettamente il segnale (cioè $\alpha(\boldsymbol{X}^*, \hat{\boldsymbol{X}}_{\text{PCA}}) = 1$).

- Se $\lambda > \lambda_* = \frac{n^2}{\|\boldsymbol{X}^*\|^4}$: il contributo di *rumore* è associato ad una densità spettrale a semicerchio con $n - 1$ autovalori di ampiezza $2\sqrt{\frac{n}{\lambda}}$ centrati in zero e distribuiti secondo la su menzionata legge del semicerchio di Wigner in Eq. (2.4). Il valore λ_* è critico: nel caso in cui è presente una matrice di rumore di rango 1, come nel modello in Eq. (2.1), appena $\lambda > \lambda_*$ un autovalore esce dal supporto della distribuzione di Wigner, come mostrato in figura 2.2, e il rispettivo autovettore è ben correlato con il segnale: $\alpha(\boldsymbol{X}^*, \hat{\boldsymbol{X}}_{\text{PCA}}) = \sqrt{1 - \frac{\lambda_*}{\lambda}} \pm \mathcal{O}(\frac{1}{\sqrt{n}})$. Questo fenomeno è chiamato *transizione BBP* (Baik—Ben Arous—Peché). La transizione BBP implica che recuperare il segnale nascosto $\boldsymbol{X}^*$ è possibile, a patto che questo autovalore riesca appunto a "fuoriuscire" dalla densità semicircolare.

- Se $\lambda < \lambda_*$ l'autovalore corrispondente all'autovettore correlato col segnale è nella densità semicircolare. In questo caso tutti gli autovettori di $\boldsymbol{Y}$ hanno una correlazione di ordine $\mathcal{O}(1/\sqrt{n})$ con la soluzione e recuperare il segnale è impossibile nel limite $n \to +\infty$.

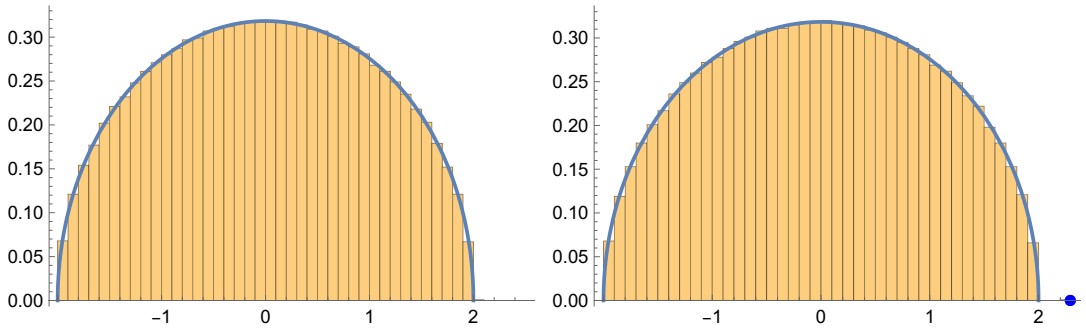

Figura 2.2: La figura mostra cosa accade nella transizione BBP: viene rappresentato lo spettro dei dati del modello di Wigner *spiked*, nel caso in cui $X_i^* \sim \mathcal{N}(0,1)$ per $\lambda < \lambda_* = 1$ (sinistra) e $\lambda = 3 > \lambda_* = 1$ (destra).

Osservazione 4. La transizione BBP, e più in generale la discussione sugli algoritmi di ricostruzione del segnale in questo capitolo, possono essere generalizzati al caso di matrici di perturbazione di rango r finito.

2.4.1 Efficienza dei vari algoritmi

L'algoritmo PCA può rivelarsi utile nella ricostruzione del segnale $\boldsymbol{X}^*$ ma soprattutto risulta ottimale nel distinguire una matrice in cui c'è segnale da una in cui non c'è, ovvero nel risolvere il cosiddetto *detection problem*, come recentemente mostrato nella Ref. [10]. Se consideriamo il caso in cui la distribuzione $P_x \sim \mathcal{N}(0,1)$, l'algoritmo PCA è ottimale anche nella ricostruzione del segnale: infatti in tal caso si ha che l'algoritmo raggiunge il minimo errore possibile, ovvero il MMSE, per $n \to +\infty$,

$$\lim_{n\to\infty} \mathrm{MSE}_n^{\mathrm{PCA}} = \lim_{n\to\infty} \mathrm{MMSE}_n$$

come dimostrato accuratamente nella Ref. [6, Sezioni 3.2-3.3.2]. Tuttavia ciò non vale in generale: nel caso di una P_x qualsiasi, l'algoritmo è sub-ottimale rispetto ad altri algoritmi polinomiali disponibili come per esempio l'AMP (*Approximate Message Passing*) [7, Sezioni 12.1-12.3.1]. Non è inoltre ovvio che *esista* un algoritmo capace di raggiungere prestazioni ottimali in un numero di passi polinomiali in ogni caso. Dalla figura 2.3, per esempio, è possibile notare come in alcuni casi

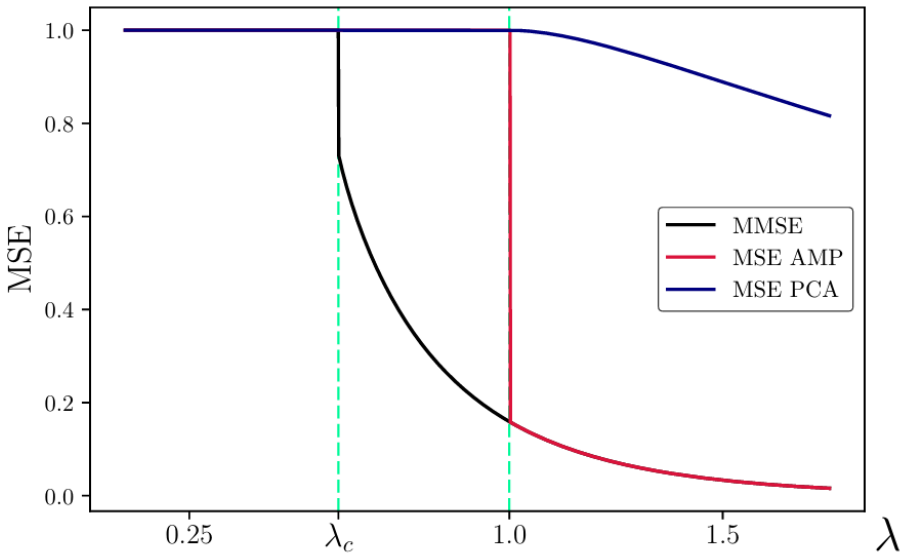

Figura 2.3: La figura, riprodotta dalla Ref. [6], rappresenta l'errore minimo raggiunto da AMP e PCA nel caso in cui P_x è dato dall'Eq. (2.47) con $p = 0.05$, da confrontarsi con la prestazione ottimale data dall'MMSE.

appaiano delle soglie critiche ulteriori oltre a λ_*. Il caso considerato corrisponde alla probabilità a priori

$$P_x = p\delta_{\sqrt{\frac{1-p}{p}}} + (1-p)\delta_{-\sqrt{\frac{p}{1-p}}}. \tag{2.47}$$

da cui $\lambda_* = 1$. Come si può vedere in figura, in questo caso, l'MMSE è uguale a 1 fino a $\lambda_c \approx 0.6$ (*regione di ricostruzione impossibile*). Per $\lambda > 1$ invece sia PCA che AMP ricostruiscono almeno in parte il segnale, e AMP in particolare raggiunge prestazioni ottimali: questa regione è chiamata *regione di ricostruzione facile*, dato che qui è possibile (parzialmente) ricostruire il segnale con successo in modo computazionalmente efficiente. Nella regione dove $\lambda_c < \lambda < 1$ è possibile costruire alcuni algoritmi non triviali in grado di recuperare il segnale, ma, al momento, non sono noti algoritmi polinomiali in grado di farlo: questa regione è appunto chiamata *regione di ricostruzione difficile*.

Conclusioni

In questa tesi è stato trattato principalmente il modello di Wigner *spiked* nel contesto dell'inferenza Bayesiana in alta dimensione, concentrandoci in particolare sull'informazione mutua tra il segnale e i dati del modello. È stato messo in evidenza, tramite alcune tecniche di meccanica statistica (come il metodo delle repliche e la concentrazione di misura), che nel limite termodinamico l'informazione mutua tra dati e segnale può essere ridotta ad un problema di ottimizzazione scalare, semplificando notevolmente l'analisi di sistemi complessi in alta dimensione. Successivamente, proprio per giustificare tale riduzione di dimensione, abbiamo cercato un *minorante* e un *maggiorante* per l'energia libera, e quindi per l'informazione mutua (quantità legate da una costante additiva): per $n \to \infty$, infatti, tali *bound* coincidono. Questa scoperta ha avuto implicazioni importanti per l'inferenza statistica, consentendo di stimare i limiti computazionali nel ricostruire in maniera ottimale il segnale nascosto, da confrontare con le prestazioni di alcuni algoritmi come PCA, a cui è dedicata la trattazione della parte finale della tesi.

Sebbene i risultati presentati chiariscano diversi aspetti del modello di Wigner *spiked*, è possibile esplorare la generalizzazione di quanto discusso: il modello, infatti, è solo un caso particolare, ma prototipico, di una vasta gamma di modelli studiati nell'inferenza Bayesiana, che a loro volta aprono la strada allo studio di nuove soluzioni algoritmiche per la gestione di grandi quantità di dati in presenza di rumore. Una naturale estensione è stata, per esempio lo studio di modelli in cui la matrice dello *spike* non è simmetrica, come nel modello di Wishart *spiked*, o la matrice di rumore non ha elementi aventi distribuzione gaussiana standard.

Bibliografia

[1] J. Barbier, *Mean-field theory of high-dimensional Bayesian inference*, ICTP Lecture notes (2019).

[2] B. Aubin, *Mean-field methods and algorithmic perspectives for high-dimensional machine learning.* PhD thesis, Institut de Physique Théorique CEA and Université Paris-Saclay, Saclay, France (2020).

[3] B. Aubin, B. Loureiro, A. Maillard, F. Krzakala, L. Zdeborová, *The spiked matrix model with generative priors (Supplementary materials)*, 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada (2019).

[4] T. Lesieur *Factorisation matricielle et tensorielle par une approche issue de la physique statistique.* PhD thesis, Université Paris Saclay (COmUE) (2017).

[5] L. Miolane *Fundamental limits of inference: A statistical physics approach.* PhD thesis, ENS & Inria Paris (2019).

[6] L. Miolane *Phase transitions in spiked matrix estimation: information-theoretic analysis*, arXiv:1806.04343 (2018).

[7] F. Krzakala e L. Zdeborová *Statistical Physics Methods in Optimization and Machine Learning* An Introduction to Replica, Cavity and Message-Passing techniques, EPFL Lecture notes (2024).

[8] A. El Alaoui e F. Krzakala, *Estimation in the Spiked Wigner Model: A Short Proof of the Replica Formula*, 2018 IEEE International Symposium on Information Theory (ISIT), 1974–1878 (2018).

[9] F. Camilli *New perspectives in statistical mechanics and high–dimensional inference.* PhD thesis, Alma Mater Studiorum –Università di Bologna e École Normale Supérieure (2023).

[10] A. Perry, A.S.. Wein, A. S. Bandeira, A. Moitra *Optimality and sub-optimality of PCA I: spiked random matrix models*, The Annals of Statistics, 46(5) (2018).

[11] C. Shannon *A mathematical theory of communication.* Disponibile all'url: http://www.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

[12] By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *An Essay towards solving a Problem in the Doctrine of Chances*, 1763