NOTE DI TERMODINAMICA STATISTICA

Marco Favretti

September 22, 2011

Contents

1 Richiamo di termodinamica 3

2 Termodinamica statistica 11
2.1 Entropia e informazione 12
2.1.1 Entropia relativa 14
2.1.2 Necessità dell'entropia relativa nel caso di infinite alternative 17
2.1.3 Principio della Massima Entropia (PME) e della Minima Informazione di Discriminazione (PMDI) 18
2.2 Il PME nella termodinamica di equilibrio 20
2.2.1 Geometria dei problemi di estremo vincolato 21
2.2.2 Il caso di k osservabili 23
2.2.3 Significato di λ: la temperatura 24
2.3 Il primo principio della Termodinamica 25
2.3.1 L'energia libera 26
2.3.2 Entropia del gas ideale 27

3 Alcune applicazioni del PME 29
3.0.3 La distribuzione normale tramite il PME 29
3.0.4 Informazione ridondante e Inferenza 31
3.0.5 Evoluzione della densità di probabilità lungo il flusso di un campo vettoriale 35
3.0.6 Il modello di Ising e le reti neurali. 36

4 Approfondimenti 39
4.1 Legge dei grandi numeri e entropia 39
4.1.1 Il teorema di concentrazione dell'entropia 43
4.1.2 Legame tra entropia relativa e test χ^2 46
4.1.3 Entropia congiunta ed entropia condizionata 47

5 Termodinamica di non equilibrio 51
5.1 Preliminari matematici 51
5.1.1 Processi stocastici 51
5.1.2 Catene di Markov 52
5.1.3 Legge dei grandi numeri per le Catene di Markov ergodiche . 54
5.1.4 Proprietà di Equipartizione Asintotica 55
5.1.5 Entropia di una catena di Markov 56
5.1.6 Tasso di Crescita dell'Entropia 57

6 Elementi di teoria delle grandi deviazioni 61
6.1 Premessa 61
6.2 Deviazioni moderate e grandi deviazioni. 63
6.2.1 Grandi deviazioni su lanci ripetuti di una moneta 64
6.3 Primi teoremi sulle grandi deviazioni 65
6.3.1 Il Teorema di Sanov 67
6.4 Grandi deviazioni per misure di coppia 69
6.4.1 Grandi deviazioni per catene di Markov 70
6.5 Teorema di Gartner-Ellis 72

7 Applicazione alla termodinamica di non equilibrio 75
7.0.1 Calcolo di valori medi di osservabili associati alle traiettorie . 76
7.0.2 Fluttuazioni dei valori medi 78
7.1 Gas su reticolo 80
7.1.1 Grandi deviazioni per la densità 81
7.1.2 Teoria idrodinamica 82

8 Appendici 85
8.1 Struttura geometrica soggiacente al PME 85
8.1.1 Funzioni generatrici di sottovarietà Lagrangiane. 85
8.1.2 Famiglie generatrici di sottovarietà Lagrangiane 86
8.1.3 La funzione di Lagrange come famiglia generatrice 88

1 Richiamo di termodinamica

La Termodinamica è una teoria fisica utile a descrivere il comportamento macroscopico di sistemi complessi (tipicamente sistemi di $\approx 10^{23}$ elementi), un dominio ove la meccanica finito–dimensionale è inadatta allo scopo. Il punto di partenza dell Termodinamica è la constatazione che per tali sistemi la formulazione della meccanica del teorema della forze vive non è più valida. Bisogna introdurre altri concetti: 1) lo stato 'energetico' di un sistema è descritto dalla sua *energia interna*, una grandezza più generale dell'energia cinetica e che tiene conto del moto microscopico oltre che macroscopico degli elementi del sistema, 2) è possibile variare l'energia interna del sistema oltre che agendo con delle forze esterne sul sistema (lavoro meccanico, come nella meccanica finito–dimensionale) anche mettendo semplicemente il sistema *a contatto* con un altro sistema; in questo modo viene variato il moto microscopico degli elementi del sistema e si dice che è stata fornita energia al sistema sotto forma di *calore*.

Il primo principio della termodinamica. Il primo principio della termodinamica è una legge fisica che afferma la conservazione del'energia nel senso allargato che abbiamo introdotto: *la variazione di energia interna di un sistema è pari al calore ricevuto* ΔQ *meno il lavoro* ΔL *effettuato dal sistema sull'ambiente circostante*

$$\Delta E = \Delta Q - \Delta L.$$

Osservazione: per convenzione, consideriamo *positivo* il calore *ricevuto* dal sistema dall'ambiente circostante e *positivo* il lavoro *effettuato* dal sistema sull'ambiente circostante.

Il primo principio della termodinamica va interpretato come la legge di Newton $f = ma$: postulata l'esistenza della forza f, definite indipendentemente la massa m e l'accelerazione a, la legge di Newton 'lega' f a ma, non è la definizione di f. Allo stesso modo, il primo principio della Termodinamica non è la definizione dell'energia interna, ma afferma l'equivalenza di calore e lavoro e conseguentemente 'ristabilisce' la conservazione dell'energia.

In altre parole, il primo principio afferma che calore e lavoro sono quantità scalari aventi la stessa natura per tutti i sistemi e che quindi possono essere sommate algebricamente; il coefficiente di proporzionalità (universale) è l'equivalente meccanico del calore.

Il secondo principio della termodinamica[1]

La termodinamica nacque dall'esigenza di progettare delle macchine termiche che potessero sfruttare il lavoro meccanico prodotto dall'espansione del vapore riscaldato.

[1] La breve esposizione del secondo principio della termodinamica che facciamo qui è tratta dal bellissimo articolo di E.T. Jaynes [Jaynes], del quale si consiglia la lettura.

Non stupisce quindi che S. Carnot schematizzasse un *motore termico* come una macchina che assorbe il calore Q_1da una sorgente a temperatura uniforme t_1, compie un lavoro L sull'ambiente e cede una parte Q_2 del calore ricevuto ad una sorgente a temperatura t_2 *più bassa* di t_1 (l'ambiente) *senza* variare la propria energia interna. Applicando il primo principio al sistema motore, il quale opera *senza* variare la propria energia interna, abbiamo che

$$\Delta E = 0 \quad \Rightarrow \quad L = Q_1 - Q_2.$$

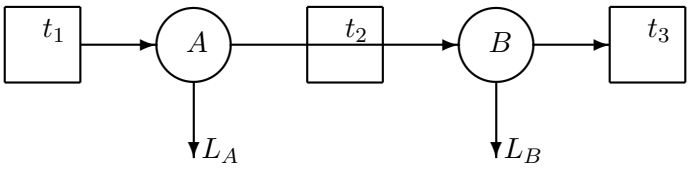
Si definisce *efficienza* del motore il rapporto

$$e = \frac{L}{Q_1} = \frac{Q_1 - Q_2}{Q_1}, \qquad 0 \leq e \leq 1.$$

Il motore è detto *reversibile* se può funzionare anche in senso inverso, prelevando il calore Q_2 dalla sorgente a temperatura t_2 e scaricando il calore Q_1 a temperatura t_1 assorbendo il lavoro L (frigorifero o pompa di calore). Lo scopo della progettazione era di massimizzare il lavoro utile L e minimizzare il calore disperso Q_2. Carnot sintetizza il risultato degli esperimenti effettuati e lo enuncia in un principio, che fino ad oggi non è mai stato smentito dall'esperienza

Principio di Carnot. *Nessun motore termico è più efficiente di un motore termico reversibile operante tra le stesse temperature t_1 e t_2.*

Carnot enuncia quindi una legge fisica, che costituisce una delle forme in cui si esprime il secondo principio della termodinamica. Come si intuisce facilmente, una violazione del principio di Carnot, e cioè la costruzione di un motore che fornisse, a parità di condizioni, un lavoro $L' > L$ permetterebbe di usare la parte L del lavoro ottenuto per far funzionare il motore reversibile al contrario ripristinando le condizioni di partenza, ed usare la parte $L' - L$ per altri scopi. Avremmo così realizzato il moto perpetuo, producendo del lavoro dal nulla. Con lo stesso ragionamento, si deduce che tutti i motori reversibili operanti tra le stesse temperature hanno la stessa efficienza, indipendentemente dalle loro particolarità costruttive, e dipendente solo dalle temperature di lavoro t_1 e t_2.



L'argomento seguente, dovuto a Lord Kelvin permette di introdurre la scala assoluta delle temperature a partire dal Principio di Carnot. Consideriamo un motore termico reversibile C costituito da due motori reversibili A e B collegati in serie (vedi figura): il motore A assorbe il calore Q_1 a temperatura t_1, fornisce il lavoro L_A e cede il calore Q_2 a temperatura $t_2 < t_1$ al motore B. Il motore B assorbe Q_2, fornisce il lavoro L_B e cede il calore Q_3 allla sorgente a temperatura $t_3 < t_2$. Le efficienze dei tre motori A, B e C sono rispettivamente

$$e_A(t_1, t_2) = \frac{L_A}{Q_1}, \qquad e_B(t_2, t_3) = \frac{L_B}{Q_2}, \qquad e_C(t_1, t_3) = \frac{L_A + L_B}{Q_1}, \tag{1.1}$$

Pertanto, la funzione efficienza $e(t_1, t_2)$ (universale, per il Principio di Carnot) deve soddisfare la seguente relazione funzionale dedotta da (1.1)

$$e(t_1, t_3) = e(t_1, t_2) + e(t_2, t_3) - e(t_1, t_2)e(t_2, t_3)$$

o, in una forma più simmetrica, ponendo $f_{12} = f(t_1, t_2) := 1 - e(t_1, t_2)$

$$f_{13} = f_{12} f_{23}.$$

Ne segue che, ad esempio, $e(t_2, t_3)$ si puo' scrivere come

$$e(t_2, t_3) = 1 - f_{23} = 1 - \frac{f_{13}}{f_{12}} = 1 - \frac{\theta(t_3)}{\theta(t_2)},$$

essendo t_1 arbitraria e considerando il solo motore B. Ora, tenendo t_2 fissa e variando t_3, misuriamo l'efficienza $e(t_3)$ e costruiamo la scala assoluta di temperatura

$$\frac{\theta(t_3)}{cost} = 1 - e(t_3).$$

Tale scala è per costruzione definita a meno di una costante. Inoltre, dalle relazioni $(1.1)_1$ e $(1.1)_3$ assieme si deduce che $\theta(t)$ è funzione *monotona crescente* di t.

In base al principio di Carnot, l'efficienza di un motore *reversibile* è una funzione universale, indipendente dalle particolarità costruttive del motore e pertanto il rapporto $\theta(t_3)/\theta(t_2)$può servire per definire una *scala universale delle temperature* (gradi Kelvin), indipendente dalla sostanza termometrica usata,

$$T(t) = c\theta(t), \quad c > 0.$$

L'efficienza di un motore termico *reversibile*, definita dalla (1.1), è allora pari a $(T_2 < T_1)$

$$e_r(T_1, T_2) = 1 - \frac{Q_2}{Q_1} \overset{rev}{\equiv} 1 - \frac{T_2}{T_1}. \tag{1.2}$$

Il passo successivo, con l'introduzione della funzione entropia, fu compiuto da Clausius. Un generico sistema termodinamico può essere considerato come un motore termico *non reversibile*, quindi con efficienza $e < e_r$. Pertanto, per un generico sistema termodinamico, che assorbe e cede calore a diverse temperature, dalla (1.2) abbiamo che

$$\frac{Q_1}{T_1} - \frac{Q_2}{T_2} \overset{irr}{\leq} 0 \tag{1.3}$$

ove l'uguaglianza vale solo nel caso di un motore termico reversibile. Ricordiamo che abbiamo posto come ipotesi al'inizio del nostro sviluppo che il motore non modificasse il proprio stato termico ($\Delta E = 0$). Se consideriamo un arbitrario processo purchè *ciclico* ($\Delta E_{tot} = 0$) compiuto dal sistema, nel quale esso scambia successivamente il calore Q_i (sottointendendo il segno + o − a seconda che il calore sia assorbito o ceduto) con la sorgente $i, i = 1 \ldots, n$ a temperatura T_i, la relazione precedente si generalizza facilmente a

$$\sum_{i=1}^{n} \frac{Q_i}{T_i} \leq 0.$$

Se ora immaginiamo, con un processo di limite spesso utilizzato in fisica, che l'indice discreto i diventi un indice continuo $s \in [0,1]$ e che il sistema scambi la quantità di calore $dQ = \frac{dQ}{ds} ds$ con la sorgente s a temperatura $T(s)$, la relazione precedente diviene

$$\int_0^1 \frac{1}{T(s)} \frac{dQ}{ds} ds = \int_{\mathcal{C}} \frac{dQ}{T} \leq 0 \tag{1.4}$$

ove $\mathcal{C}$ intende un arbitrario processo ciclico. Nel limite di un processo ciclico *reversibile*[2], la relazione precedente è soddisfatta *come eguaglianza* e questo ci dice che la quantità sotto integrale è un *differenziale esatto*; esiste cioè una funzione S delle variabili macroscopiche[3] detta da Clausius *entropia* tale che

$$S_B - S_A = \int_A^B \frac{dQ}{T}$$

ove $A \to B$ è un processo *reversibile* (ora non ciclico) e gli stati A e B sono stati di *equilibrio*.

Se ora consideriamo un processo ciclico tale che $A \overset{rev}{\to} B \overset{irr}{\to} A$ abbiamo che la (1.4) vale per l'intero ciclo e quindi che

$$\int_A^B \frac{dQ}{T} \leq S_B - S_A$$

ovvero che la differenza di entropia rappresenta un limite superiore raggiunto solo nel caso di una trasformazione ciclica.

La conseguenza celebre di questa relazione è che per un sistema *isolato*, per esempio l'Universo, per il quale $dQ = 0$, si ha che *nel passare da uno stato di equilibrio ad un altro, l'entropia non può diminuire*

$$S_{finale} \geq S_{iniziale}.$$

Concludiamo questa brevissima presentazione del secondo principio illustrando il contributo datone da J.W. Gibbs. Il secondo principio dell Termodinamica dice 'solo' in che direzione avviene spontaneamente un processo per un sistema isolato, senza poter specificare lo stato finale o la velocità $\frac{dS}{dt}$ con la quale avviene il processo.

Nella sua formulazione del secondo principio, Gibbs fa una affermazione più forte di quella data da Clausius e precisamente: Sono equivalenti

[2]Una certa cura deve essere posta nell'utilizzare il concetto astratto di processo reversibile, che deve essere considerato come il limite di processi irreversibili, fisicamente realizzabili. Abbiamo assimilato un generico sistema ad un motore termico irreversibile, il cui lavoro prodotto può essere nullo. In tale caso $L = Q_1 - Q_2 = 0$ e la relazione (1.3), soddisfatta in senso stretto dice che $T_1 > T_2$, ovvero che il passaggio di una quantità di calore Q_1 avviene 'spontaneamente' *solo* da una sorgente a temperatura più alta ad una a temperatura più bassa. Ora, possiamo mantenere la quantità di calore ceduto Q_1 costante facendo tendere $T_1 - T_2 \to 0^+$ prendendo delle sorgenti opportune (si pensi alla relazione $\Delta Q = cm\Delta T$ ove c è il calore specifico e m la massa). In questo modo, nel limite $T_1 - T_2 \to 0^+$ la (1.3) è soddisfatta come uguaglianza e possiamo qualificare come *reversibile* il processo. Ne segue che per un processo reversibile, T è la temperatura *comune* del sistema e della sorgente e gli stati che attraversa il sistema, compresi gli stati iniziale e finale, sono stati di *equilibrio*.

[3]Le variabili macroscopiche di un sistema sono le grandezze, misurabili sperimentalmente, che definiscono lo stato macroscopico del sistema: volume, pressione, temperatura, energia interna, magnetizzazione, numero di moli, ecc. La loro scelta e individuazione dipende dallo sperimentatore.

a) *un sistema isolato raggiunge spontaneamente lo stato macroscopico di equilibrio di entropia* massima *tra quelli permessi dai vincoli macroscopici, ovvero per energia interna e volume fissati*,

b) *un sistema isolato raggiunge spontaneamente lo stato macroscopico di equilibrio di energia interna* minima *tra quelli permessi dai vincoli macroscopici, ovvero per entropia e volume fissati*

Si osservi che lo stato finale del sistema prescritto da Gibbs è (e non potrebbe essere altrimenti) uno stato *macroscopico* del sistema, che può essere 'generato' da milioni di stati *microscopici* diversi del sistema, sui quali il secondo principio non dice nulla. Lo stato microscopico può essere indagato solo usando le leggi (reversibili) della meccanica particellare e la relazione tra dinamica microscopica ed osservazioni macroscopiche deve far intervenire la statistica.

Formulazione di Gibbs del primo principio

Si tratta di una formulazione del principio di conservazione dell'energia (primo principio) che tiene conto della definizione di entropia data dal secondo principio combinandole assieme. Poichè per una trasformazione *reversibile* $dQ = TdS$, il primo principio si riscrive (qui P è la pressione e V il volume)

$$\Delta Q = T\Delta S = \Delta E + \Delta L = \Delta E + P\Delta V$$

da cui, usando i differenziali

$$dS = \frac{1}{T}de + \frac{P}{T}dV. \tag{1.5}$$

Applicazione. Calcoliamo l'entropia di un gas ideale, definito come un fluido che obbedisce alle equazioni di stato

$$PV = nRT, \qquad E = nc_vT$$

ove R è la costante dei gas, n è in numero di moli e c_v è la capacità termica specifica a volume costante. Integriamo la relazione di Gibbs (primo principio) attraverso una successione di due trasformazioni reversibili, la prima costituita da un riscaldamento a volume costante da T_1 a T_2 e la seconda da un'espansione isoterma da V_1a V_2. La (1.5) si riscrive

$$dS = \frac{1}{E}nc_vdE + \frac{nRT}{VT}dV = n(c_v\frac{dE}{E} + R\frac{dV}{V}) \tag{1.6}$$

che, integrata, porge (in ogni trasformazione solo uno dei due termini è non nullo)

$$S(E_2, V_2) - S(E_1, V_1) = nc_v \ln(\frac{E_2}{E_1}) + nR\ln(\frac{V_2}{V_1}).$$

Osservazione. L'espressione trovata dipende da n, ma non definisce la forma della dipendenza funzionale di S da n. Infatti nella trasformazione complessiva il numero di moli non è variato. Per determinare tale dipendenza, consideriamo l'integrale generale della (1.5) applicata al gas ideale, ove pensiamo S come funzione di E, V e n

$$S(E, V, n) = nc_v \ln E + nR\ln V + f(n) + c.$$

Per determinare il fattore integrante $f(n)$ possiamo usare l'ipotesi che S è funzione positivamente 1–omogenea delle variabili. Si trova allora (imponendo la condizione $S(\lambda X) = \lambda S(X)$, ponendo $n = 1$ e poi sostituendo λ con n) che f ha la forma $f(n) = nf(1) - n\ln n(c_v + R)$ pertanto l'espressione completa dell'entropia è la funzione positivamente 1–omogenea e concava (verificarlo)

$$S(E, V, n) = nc_v \ln(\frac{E}{n}) + nR \ln(\frac{V}{n}).$$

Legge delle adiabatiche. Calcoliamo l'equazione di stato per un gas sottoposto ad una trasformazione reversibile e *adiabatica*, ovvero $dQ = 0$. Dalla ultima eguaglianza della (1.6), essendo $dS = 0$ si ottiene

$$\frac{dE}{E} = -\frac{R}{c_v}\frac{dV}{V}$$

da cui, integrando

$$\ln \frac{E_2}{E_2} = -\frac{R}{c_v} \ln \frac{V_2}{V_1}$$

che si riscrive nelle formulazioni equivalenti

$$EV^{\frac{R}{c_v}} = cost(S), \qquad TV^{\frac{R}{c_v}} = cost(S), \qquad PV^{\gamma} = cost(S).$$

Trasformazione di Legendre e potenziali termodinamici

Abbiamo visto nella formulazione precedente del secondo principio che energia e entropia possono essere usati entrambi per esprimere la condizione variazionale. Sotto opprtune ipotesi, che non indagheremo qui, si puo' supporre che l' entropia, come funzione delle variabili macroscopiche, sia *convessa*. La convessità di S, oltre a definire un *unico* stato di equilibrio termodinamico nel caso di vincoli macroscopici lineari, permette di usare, come variabili indipendenti che descrivono lo stato macroscopico del sistema, una qualsiasi coppia tra le variabili S, E, T, V, P. Vediamo ad esempio come passare dalle variabili (S, V) alle (T, V). Il procedimento è analogo alla trasformazione che in Meccanica permette di passare dalle variabili Lagrangiane $(q, \dot{q}) \approx (V, S)$ alle variabili Hamiltoniane $(q, p) \approx (V, T)$. (Per un discorso piu' approfondito sulla Trasformazione di Legendre, vedasi [Cardin].)

Dalla funzione $E(S, V)$, convessa, definiamo come sopra $T(S, V) = \partial E/\partial S$ e, usando l'ipotesi $\partial^2 E/\partial S^2 > 0$ di convessità possiamo invertire per ogni valore fissato di V la $T(S, V)$ e ottenere $S = \hat{S}(T, V)$. Introduciamo ora la funzione 'Hamiltoniana' associata alla Lagrangiana $E(S, V)$

$$-\Psi(T, V) = T\hat{S}(T, V) - E(\hat{S}(T, V), V), \text{ ovvero } \Psi(T, V) = E(\hat{S}, V) - T\hat{S},$$

detta *energia libera* del sistema. Le sue derivate parziali sono

$$\frac{\partial \Psi}{\partial V} = \frac{\partial E}{\partial V} + \frac{\partial E}{\partial \hat{S}}\frac{\partial S}{\partial V} - T\frac{\partial \hat{S}}{\partial V} = \frac{\partial E}{\partial V} \equiv -P(T, V) \tag{1.7}$$

$$\frac{\partial \Psi}{\partial T} = \frac{\partial E}{\partial \hat{S}}\frac{\partial \hat{S}}{\partial T} - \hat{S} - \frac{\partial \hat{S}}{\partial T}T = -\hat{S}(T, V). \tag{1.8}$$

Infine,

$$\frac{\partial^2 \Psi}{\partial T^2} = -\frac{\partial \hat{S}}{\partial T} = -(\frac{\partial T}{\partial S})^{-1} = -(\frac{\partial^2 E}{\partial S^2})^{-1} < 0$$

pertanto l'energia libera è funzione strettamente concava $\partial^2 \Psi / \partial T^2 < 0$ della temperatura. Riscriviamo il primo principio combinato con il secondo per una trasformazione isoterma reversibile

$$dL = dQ - dE = TdS - dE = d(TS - E) = -d\Psi.$$

Pertanto, il lavoro fornito del sistema è pari alla sua variazione di energia libera cambiata di segno. Per trasformazioni isoterme, quindi, l'energia libera svolge la stessa funzione dell'energia potenziale in un sistema puramente meccanico.

2 Termodinamica statistica

La fisica è stata la prima disciplina che si è posta il problema di dare una descrizione, in termini qualitativi e anche quantitativi, del comportamento dei sistemi complessi, ovvero sistemi costituiti da un gran numero (tipicamente 10^{23}) di elementi non ulteriormente divisibili. Non stupisce quindi che, ai nostri giorni, un numero crescente di discipline, dall'economia, alla biologia, alla linguistica, che si affacciano allo studio di sistemi sempre più complessi, utilizzino dei concetti come quello di microstato, macrostato, entropia che sono stati originariamente elaborati all'interno della fisica. Vediamo ora come addentrarci nello studio dei sistemi complessi intesi come aggregati di moltissimi componenti elementari, prendendo come esempio paradigmatico il caso del sistema costitituito da particelle materiali puntiformi della fisica classica, anche se le idee soggiacenti saranno facilmente generalizzabili. Uno stato *microscopico* del sistema è l'assegnazione di posizione e velocità di ogniuna delle particelle componenti. La specificazione dello stato microscopico, spesso impossibile, contiene moltissima informazione sullo stato del sistema, la maggior parte della quale è *inutile* per descrivere a livello *macroscopico* il sistema. Quest'ultimo viene specificato assegnando i valori (numeri reali) di pochi parametri macroscopici. La scelta dei parametri macroscopici è onere e onore dello sperimentatore, in una maniera che sarà precisata tra breve. Diciamo che lo sperimentatore ha definito un insieme *completo* di parametri macroscopici quando egli ha il controllo completo del sistema a livello macroscopico, ovvero, agendo sperimentalmente sui parametri macroscopici, è in grado di riprodurre ogni volta che sia richiesto, lo stesso stato macroscopico.

Vediamo quindi che con le dizioni 'micro' e 'macro' abbiamo di fatto affermato che un sistema complesso ha due livelli di lettura o di descrizione, il livello microscopico e quello macroscopico. Un punto cruciale è che nel passaggio dal livello microscopico a quello macroscopico vi è una perdita di informazione sul sistema, perdita che può essere innocua e anzi positiva avendo cura che in realtà venga abbandonata solo l'informazione ridondante per la riproducibilità del comportamento macroscopico. Detto in altre parole, ad ogni stato macroscopico del sistema corrispondono moltissimi stati microscopici fra loro equivalenti.

Il progetto di dedurre le leggi che descrivono la dinamica a livello macroscopico (la Termodinamica nel nostro esempio iniziale) da quelle delle dinamica microscopica costituisce l'ambito di indagine della Meccanica Statistica. Per fare questo essa realizza il passaggio dal livello microscopico a quello macroscopico, dicendo che le proprietà macroscopiche si ottengono come *valori medi*, rispetto ad una densità di probabilità, di opportune funzioni di stato, o osservabili, definite sull'insieme degli stati microscopici. Il problema quindi è di *individuare la densità di probabilità* o *stato statistico* che meglio descrive lo stato macroscopico del sistema. L'oggetto principale di indagine di questo capitolo è di presentare una ricetta per determinare

lo stato statistico. Questa ricetta è nota come Principio della massima Entropia di Gibbs–Jaynes. Fin d'ora pero' possiamo intuire che sono centrali i concetti di stato statistico e di informazione associata ad uno stato statistico, ed è ad essi che ci rivolgiamo nel prossimo paragrafo.

2.1 Entropia e informazione

La nozione di informazione associata ad una distribuzione di probabilità discreta finita è stata elaborata da Claude Shannon in un celebre articolo del 1948. E' importante sottolineare che la nozione di entropia contenuta nel teorema di Shannon sotto riportato ha una portata assolutamente generale, ben più vasta dell'ambito della fisica. Per collegarci a quanto detto in precedenza pensiamo all'insieme degli stati microscopici del sistema come ad un insieme finito di alternative $A_1, \ldots, A_n$ mutualmente esclusive. Una distribuzione di probabilità discreta $p_1, \ldots, p_n$ definisce quindi per ogni i la probabilità che sia realizzata l'alternativa A_i. Nella dimostrazione si ragiona per induzione su n, e si suppone quindi che la funzione matematica $H_n(p)$ che descrive l'*incertezza* associata ad una distribuzione discreta $p_1, \ldots, p_n$ abbia la stessa forma funzionale per ogni n. Inoltre si fa l'ipotesi che ognuna delle alternative A_i possa essere decomposta in più alternative parziali, anch'esse mutualmente esclusive, ovvero $A_k = A_1^{(k)} \cup \ldots A_m^{(k)}$ e $p_k = p_1^{(k)} + \ldots + p_m^{(k)}$. Nella dimostrazione si prova che, stabiliti assiomaticamente un certo numero di requisiti che ragionevolmente deve soddisfare una misura dell'incertezza riguardo al realizzarsi di un insieme di alternative descritta da una distribuzione di probabilità p, tali requisiti determinano *univocamente* la forma funzionale che deve avere la funzione $H(p)$.

Teorema 2.1.1 (C.Shannon) *Sia $H_n(p)$ una funzione definita, per ogni n, sul dominio $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$. Supponiamo che*

1) $H_n(p_1, \ldots, p_n)$ è funzione continua nelle p_i per ogni n,

2) $H_n(p_1, \ldots, p_n)$ è funzione simmetrica dei suoi argomenti,

3) $H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n)$,

4) $H_n(p_1, \ldots, p_n) \leq H_n(\frac{1}{n}, \ldots, \frac{1}{n})$,

5) posto $p_k = p_1^{(k)} + \ldots + p_m^{(k)}$ con $\sum_{k=1}^n \sum_{l=1}^m p_l^{(k)} = 1$ si ha che

$$H_{nm}(p_1^{(1)}, \ldots, p_m^{(n)}) = H_n(p_1, \ldots, p_n) + \sum_{k=1}^n p_k H_m(\frac{p_1^{(k)}}{p_k}, \ldots, \frac{p_m^{(k)}}{p_k}),$$

allora

$$H_n(p_1, \ldots, p_n) = -\lambda \sum_{k=1}^n p_k \ln p_k. \tag{2.1}$$

Commentiamo brevemente i requisiti assiomaticamente introdotti dal teorema di Shannon. I requisiti 1) e 2) di continuità e di indipendenza dal numero con cui etichetto le varie alternative sono naturali. Il requisito 3) esprime il fatto che considerare in più un'alternativa che so essere irrealizzabile non deve far aumentare

l'incertezza legata al sistema mentre il requisito 4) introduce l'ipotesi che l'incertezza sia massima quando tutte le possibili alternative sono egualmente probabili. Infine, il requisito 5) afferma che se so che l'alternativa A_k che si realizza con probabilità p_k è in realtà decomposta in m alternative $A_l^{(k)}$, allora l'incertezza associata alla distribuzione di probabilità è pari all'incertezza riguardo al realizzarsi delle alternative $A_1, \ldots, A_n$ più l'incertezza, pesata con peso p_k dovuta al dover scegliere tra le sotto–alternative $A_1^{(k)}, \ldots, A_m^{(k)}$ alle quali assegno la probabilità condizionata $p_1^{(k)}/p_k, \ldots, p_m^{(k)}/p_k$.

Dimostrazione. Per l'ipotesi 1) di continuità di H è sufficiente mostrare che (2.1) vale su un sottoinsieme denso di $[0,1]$, e quindi considerare distribuzioni di probabilità della forma

$$p_i = \frac{q_i}{\sum_{i=1}^n q_i},$$

con q_i interi. Poniamo, per semplificare, $h(n) = H_n(\frac{1}{n}, \ldots, \frac{1}{n})$. Dalle condizioni 3) e 4) considerate nel caso particolare $p_i = 1/n$ segue subito che per ogni naturale n

$$h(n) \leq h(n+1)$$

ovvero che, nel caso di alternative egualmente probabili, l'incertezza deve essere una funzione *monotona non decrescente* di n. In effetti questo requisito puo' essere introdotto al posto di 3) e 4). Ora, nel caso piu' semplice che tutte le alternative o sotto–alternative siano egualmente probabili, ovvero $p_i = 1/n$ e $p_i^{(k)} = 1/mn$, dal requisito 5) segue subito che

$$h(mn) = h(m) + h(n)$$

e se scriviamo $n = m^{r-1}$ otteniamo che h deve soddisfare alla condizione

$$h(m^r) = rh(m).$$

Una *possibile* soluzione è $h(n) = \lambda \ln(n)$. Dobbiamo ora che quella trovata è l'unica soluzione, almeno nel caso di alternative tutte egualmente probabili, e poi che essa permette di calcolare la H nel caso sufficentemente generale di p_i razionali, $p_i = \frac{q_i}{\sum_{i=1}^n q_i}$. Per fare questo usiamo un lemma preliminare:

Lemma 2.1.1 *Siano t,s due interi maggiori o eguali a* 2. *Allora, per ogni intero n è possibile trovare un intero m che soddisfa alla disuguaglianza*

$$s^m \leq t^n < s^{m+1} \quad \textit{ovvero} \quad \frac{m}{n} \leq \frac{\ln t}{\ln s} < \frac{m+1}{n}. \tag{2.2}$$

Dimostrazione. La funzione potenza $m \mapsto s^m$ di base s maggiore o uguale a due è monotona crescente su $\mathbb{Z}$. L'insieme ordinato $\{m \in \mathbb{Z} \ : \ s^m \leq t^n\}$ connesso e inferiormente illimitato, è superiormente limitato e quindi ammette massimo m^*. Allora $m^* + 1$ appartiene al complementare dell'insieme appena introdotto e quindi vale la tesi per m^*. □

Poichè $h(n)$ è monotona non decrescente, segue subito che $h(s^m) \leq h(t^n) \leq h(s^{m+1})$ ovvero che

$$mh(s) \leq nh(t) \leq (m+1)h(s)$$

che possiamo riscrivere come

$$\frac{m}{n} \leq \frac{h(t)}{h(s)} \leq \frac{m+1}{n}. \tag{2.3}$$

Confrontando le disuguaglianze (2.2) e (2.3) si conclude che

$$\Big|\frac{h(t)}{h(s)} - \frac{\ln(t)}{\ln(s)}\Big| \leq \frac{1}{n} \quad \text{ovvero} \quad \Big|\frac{h(t)}{\ln(t)} - \frac{h(s)}{\ln(s)}\Big| \leq \frac{h(s)}{n\ln(t)}$$

ovvero che, per l'arbitrarietà di n, *deve* essere $h(n) \equiv \lambda \ln(n)$, la costante essendo equivalente alla scelta della base dei logaritmi.

Ora dobbiamo far vedere che il calcolo di $H_n(p)$ con p razionale si puo' esprimere tramite $h(n)$. Posto $\sum_{i=1}^n q_i = Q$, possiamo considerare le n alternative A_i come decomposte ognuna in q_i alternative elementari *tutte equiprobabili*, per cui le Q sotto–alternative elementari risultano avere probabilità $1/Q$. Applichiamo allora la regola 5) e otteniamo

$$h(Q) = H_n(p_1, \ldots, p_n) + \sum_{i=1}^n p_i h(q_i).$$

Poichè $q_i = p_i Q$, segue subito che

$$h(Q) = H_n(p_1, \ldots, p_n) + \sum_{i=1}^n p_i h(p_i) + \sum_{i=1}^n p_i h(Q)$$

da cui $H_n(p_1, \ldots, p_n) = -\sum_{i=1}^n p_i h(p_i) = -\lambda \sum_{i=1}^n p_i \ln(p_i)$. □

Proprietà di H. Con l'estensione continua $0 \ln 0 = 0$ la funzione $H(p)$ risulta definita in tutto il primo ottante $\mathbb{R}^n_+ = \{p \in \mathbb{R}^n : p_i \geq 0\}$ e ivi infinitamente derivabile. Inoltre

$$HessH(p) = -Diag[\frac{1}{p_i}]$$

quindi $H(p)$ è una funzione concava all'interno del dominio e ha massimo assoluto in $p_i = \frac{1}{e}$ ove vale $H(\frac{1}{e}) = h(n) = \lambda \ln n$.

Nel caso più semplice, quando si è confrontati a due alternative A_1, A_2 equiprobabili, l'incertezza associata a questa distribuzione discreta vale $h(2) = \lambda 2 \ln 2$. Se si prende $2 \ln 2$ come unità di misura dell'incertezza associata ad una scelta tra due alternative assolutamente equivalenti, allora $\lambda = 1$.

2.1.1 Entropia relativa

Nel Capitolo precedente abbiamo visto che la funzione entropia di Shannon ben si presta a descrivere l'incertezza associata a una distribuzione di probabilità. Ora, prendendo un'altra direzione esploriamo un'altra possibilità di introdurre una quantità, l'entropia relativa, che costituisce una generalizzazione della nozione di entropia. Per fare questo abbiamo bisogno di alcune nozioni tipiche dell'approccio soggettivo (di B. de Finetti) alla teoria della probabilità. Indichiamo come è d'uso, con AB e $A + B$ il prodotto e la somma logica di due proposizioni. Indichiamo inoltre con

$$p(A|B)$$

la probabilità che A è vera sapendo che B è vera. Nella teoria assiomatica, la nozione equivalente è quella di probabilità condizionata

$$p(A|B) = \frac{mis(A \cap B)}{mis(B)} \quad \text{posto che} \quad mis(B) \neq 0.$$

Da requisiti di coerenza logica della teoria soggettiva discendono le leggi della somma e del prodotto (qui $\bar{A}$ indica la negazione logica di A)

$$p(A|B) + p(\bar{A}|B) = 1 \quad \text{somma}$$

$$p(AB|C) = p(B|C)p(A|BC) = p(A|C)p(B|AC) \quad \text{prodotto}$$

del tutto analoghe a quelle della teoria assiomatica. Nella legge del prodotto, l'ultima eguaglianza esprime l'indifferenza nel considerare prima A o B. Da tale uguaglianza segue subito la regola di Bayes

$$p(A|BC) = p(A|C)\ \frac{p(B|AC)}{p(B|C)}.$$

La regola di Bayes esprime come la plausibilità in una certa affermazione viene ad essere modificata dalla conoscenza della plausibilità di nuove affermazioni o risultati di esperimenti. Vediamo come. Nella regola di Bayes intendiamo $A = H$ come un'ipotesi da discriminare se vera o falsa, $B = D$ come il dato o l'informazione ottenibile da un esperimento e $C = X$ come l'insieme delle conoscenze o postulati validi all'interno della teoria che descrive l'esperimento, in un certo senso tutto quello che sappiamo già essere vero. Nella regola di Bayes, allora

$$p(H|DX) = p(H|X)\ \frac{p(D|HX)}{p(D|X)} = p(H|X)\mathcal{L}(H)$$

il primo fattore al secondo membro $p(H|X)$ esprime la plausibilità a priori, ovvero sulla base delle sole conoscenze della teoria mentre il termine $p(H|DX)$ esprime la plausibilità di H noti X e D. Il termine $\mathcal{L}(H)$, detto *verosimiglianza* di H (likelyhood) esprime il rapporto tra la plausibilità del dato D includendo o meno H tra i postulati della teoria e quindi misura la rilevanza dell'ipotesi H rispetto all'esperimento. La regola di Bayes indica quindi come aggiornare la nostra soggettiva percezione di plausibilità di una affermazione sulla base di nuove conoscenze.

Siano ora H_1 e H_2 due ipotesi da discriminare come vere o false sulla base dello stesso esperimento o dato D. Applicando separatamente la regola nei due casi, facendo il rapporto e successivamente prendendo il logaritmo (la scelta della base è ininfluente) si ottiene

$$\ln \frac{p(H_1|DX)}{p(H_2|DX)} = \ln \frac{p(H_1|X)}{p(H_2|X)} + \ln \frac{p(D|H_1X)}{p(D|H_2X)}.$$

I primi due rapporti esprimono la "preferenza" o "favore" (gli "odds" dei bookmakers) di H_1 rispetto a H_2 come vera dopo l'esperimento e prima dell'esperimento. La loro differenza

$$\mathcal{G} = \ln \frac{p(D|H_1X)}{p(D|H_2X)}$$

è il logaritmo del rapporto tra le verosimiglianze delle due ipotesi (log-likelyhood ratio) ed esprime l'importanza del dato D per discriminare H_1 rispetto a H_2, ovvero il *guadagno di evidenza (o di informazione) ottenibile dall'esperimento* per la discriminazione di H_1 rispetto a H_2.

Supponiamo ora di considerare l'informazione relativa alla conoscenza dello stato nel quale si trova il nostro sistema. Come nel Teorema di Shannon, consideriamo i possibili dati $D = A_i$, $i = 1, \ldots, n$ ove le A_i sono un insieme esaustivo di alternative mutualmente esclusive, ovvero $p(A_i A_j|X) = \delta_{ij}$. Se supponiamo che H_1 o che H_2 sia vera, le plausibilità dei vari stati sono nei due casi

$$p_i = p(A_i|H_1X), \quad i = 1, \ldots, n, \quad q_i = p(A_i|H_2X), \quad i = 1, \ldots, n.$$

Il guadagno medio di evidenza per discriminare H_1 rispetto a H_2 ottenibile dalla determinazione dello stato in cui si trova il sistema (nell'ipotesi che H_1 sia vera) è

$$\mathcal{G} = \sum_{i=1}^{n} p(A_i|H_1X) \ln \frac{p(A_i|H_1X)}{p(A_i|H_2X)} = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} = D(p\|q).$$

La quantità trovata, espressa con altri simboli di uso corrente,

$$D(p\|q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

ha vari nomi: guadaglio di informazione, entropia relativa, divergenza di Kullback–Leibler, informazione di discriminazione. Vediamo di interpretare ancora meglio il suo significato e la sua utilità. Supponiamo che H_2 sia un'ipotesi che è *sempre vera* dati i postulati X della teoria, ovvero che la sua considerazione non aggiunga nulla a quanto già noto dalla teoria. Allora

$$q_i = p(A_i|H_2X) = p(A_i|X) \quad \forall A_i, \ i = 1, \ldots, n$$

rappresenta l'assegnazione di probabilità (da qui in poi usiamo il linguaggio standard) dello stato A_i sulla base delle conoscenze della teoria,

$$p_i = p(A_i|H_1X) \quad i = 1, \ldots, n,$$

rappresenterebbe l'assegnazione di probabilità degli stati se H_1 fosse vera e $D(p||q)$ il guadagno medio di informazione di discriminazione dell'ipotesi H_1 dato dalla distribuzione di probabilità p_i stessa.

Proprietà di D. L'entropia relativa $D(p||q)$ verifica le proprietà

$$D(p\|q) \geq 0 \ \forall p, q \quad \text{e} \quad D(p\|q) = 0 \iff p \equiv q$$

inoltre, dalla concavità di $H(p)$, si vede subito che $D(p\|q)$ è funzione convessa nelle p per q fissato.

Dimostrazione: Pensiamo a $D(p\|q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$ come funzione di p per q fissato. Cerchiamo il minimo di $D(q)$ vincolato al sottospazio $\sum_i p_i = 1$. usando il Metodo dei Moltiplicatori di Lagrange (vedi oltre) cerchiamo l'estremo libero della funzione

$$G(p, \mu) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} - \mu(\sum_i p_i - 1).$$

La soluzione unica è $\hat{p}_i = q_i e^{\mu-1}$ e μ si determina imponendo il vincolo di normalizzazione, ottenendo $\mu = 1$. Verifichiamo che $\hat{p}_i = q_i$ è un minimo di D. Dalla condizione sufficiente di estremo di secondo ordine si ha

$$Hess(G)(q) = Hess(D(p\|q))_{|p=q},$$

e

$$Hess(G)(q)_{ij} = \frac{\delta_{ij}}{q_i} \in Sym^+. \tag{2.4}$$

Anche se D non è funzione simmetrica dei suoi argomenti e non soddisfa la disuguaglianza triangolare, è corretto pensare a D come una misura della 'distanza' tra due distribuzioni. Si ottiene una quantità simmetrica in p, q ponendo

$$J(p:q) = D(p\|q) + D(q\|p)$$

Tale quantità introdotta da Kullback con il nome di *divergenza* di p e q è una misura della difficoltà di discriminare tra due ipotesi.

2.1.2 Necessità dell'entropia relativa nel caso di infinite alternative

La definizione di entropia associata a una distribuzione di probabilità relativa ad un insieme finito $A_1, \ldots, A_n$ non è immediatamente estendibile al caso in cui n è infinito oppure al caso in cui le alternative costituiscano un insieme che ha la cardinalità del continuo per il motivo seguente. Se le alternative sono indicate da $x \in X \subset \mathbb{R}^d$ e la loro probabilità con $p(x)$, noi vogliamo poter rinominare le alternative tramite un diffeomorfismo $f : Y \to X$, $x = f(y)$ senza che questo modifichi l'incertezza o entropia associata alla d.d.p. p. La condizione di normalizzazione, unita al teorema di cambiamento di variabile negli integrali d–dimensionali impongono la forma di $p(y)$ (qui dx indica la misura di Lebesgue in $\mathbb{R}^d$)

$$1 = \int_X p(x)dx = \int_Y p(f(y))F(y)dy = \int_Y \tilde{p}(y)dy, \qquad F(y) := |\det f'(y)|$$

ovvero $\tilde{p}(y) = (p \circ f)(y)F(y)$. Con questa definizione naturale l'entropia viene a dipendere da f, ovvero dal modo in cui si etichettano le varie alternative, il che è inaccettabile nella teoria. Si ha infatti che

$$H(p) = -\int_X p(x)\ln(p(x))dx = -\int_Y \tilde{p}(y)\ln(p(f(y)))dy = -\int_Y \tilde{p}(y)\ln[\frac{\tilde{p}(y)}{F(y)}]dy$$

quindi

$$H(p) = H(\tilde{p}) + \int_Y \tilde{p}(y)\ln(F(y))dy.$$

Il problema è risolubile scegliendo di misurare l'incertezza non assoluta, ma rispetto ad una fissata d.d.p. di riferimento q attraverso l'entropia relativa. Si ha allora che

$$D(p\|q) = \int_X p\ln[\frac{p}{q}]dx = \int_Y \tilde{p}\ln[\frac{\tilde{p}F}{F\tilde{q}}]dy = \int_Y \tilde{p}\ln[\frac{\tilde{p}}{\tilde{q}}]dy = D(\tilde{p}\|\tilde{q}).$$

Una giustificazione piu' profonda della necessità di usare l'entropia relativa nel caso di infinite alternative è contenuta in [Jaynes].

Un caso in cui il PMDI si riduce al PME

E' immediato constatare che se X ha misura d-dimensionale finita e la d.d.p. a priori q è la distribuzione uniforme $q(x) = 1/|X|$, ove $|X|$ indica la misura d-dimensionale di X , il PMDI si riduce al PME. Infatti

$$D(p\|q) = \int_X p(x) \ln(p(x)|X|)dx = -H(p) + \ln|X|.$$

2.1.3 Principio della Massima Entropia (PME) e della Minima Informazione di Discriminazione (PMDI)

Nei paragrafi precedenti abbiamo definito l'incertezza associata ad una distribuzione di probabilità tramite la funzione di Shannon $H(p)$ e il guadagno di informazione per la discriminazione di un'ipotesi dato dal considerare una distribuzione p al posto di una distribuzione q. Poichè intuitivamente, l'incertezza deve diminuire quando l'informazione aumenta, le due nozioni devono essere collegate. Questo si vede facilmente supponendo che la distribuzione deducibile dai fatti noti sia la distribuzione uniforme $1_i = \frac{1}{n}$ per ogni i, che esprime l'assenza di informazione (completa ignoranza) sullo stato del sistema. Allora vale la relazione

$$D(p\|1) = \sum_{i=1}^{n} p_i \ln(p_i n) = -H(p) + \ln n.$$

che esprime il fatto che $D(p\|1) + H(p) = \ln n = cost$. A questo punto risultano perlomeno ragionevolmente fondati i due principi metodologici seguenti, che esprimono il requisito che la distribuzione di probabilità che scegliamo per descrivere lo stato statistico di un sistema *non deve aggiungere altra informazione oltre a quella disponibile sotto forma di vincoli sulla d.d.p.*

Proposizione 2.1.1 (PME) *p deve rendere massima l'incertezza $S(p)$ associata alla d.d.p. stessa accogliendo contemporaneamente l'informazione disponibile sullo stato statistico del sistema sotto forma di vincoli sulla d.d.p ovvero*

$$H(p) = -\sum_{i=1}^{n} p_i \ln p_i \quad \rightarrow \quad max \quad ove \quad F(p) = c;$$

Proposizione 2.1.2 (PMDI) *p deve rendere minimo il guadagno di informazione per la discriminazione di un'ipotesi rispetto alla distribuzione q che esprime la conoscienza attuale del sistema accogliendo contemporaneamente l'informazione disponibile sullo stato statistico del sistema sotto forma di vincoli sulla d.d.p*

$$D(p\|q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} \quad \rightarrow \quad min \quad ove \quad F(p) = c.$$

Il PME è dovuto a E.T. Jaynes mentre il PMDI è dovuto a S. Kullback. Entrambi sono delle estensioni a casi molto più complessi del Principio di ragione insufficente di Laplace che, in sostanza, afferma che: *in assenza di informazioni* non vi sono ragioni per supporre che il verificarsi di A abbia probabilità diversa dal verificarsi di $\bar{A}$ e quindi bisogna assegnare ad entrambi probabilità $\frac{1}{2}$.

Il PME ha un corollario che ne giustifica la sua applicazione alla fisica e che, in particolare, permette di ritrovare *tutti* i risultati della termodinamica classica o assiomatica

Proposizione 2.1.3 (Entropia fisica) *L'entropia fisica S associata ad un sistema coincide con il valore assunto dall'entropia di Shannon in corrispondenza del massimo globale vincolato definito dal PME*

$$S(c) \equiv H(p(c))$$

ove

$$p(c) \qquad : \qquad S(p(c)) = max\ S(p) \quad e \quad F(p(c)) = c.$$

L'entropia fisica S viene ad essere quindi funzione dei valori assunti dai vincoli e quindi dai dati disponibili sul sistema. Ne segue che l'entropia fisica del sistema non è definita una volta per tutte ma che il suo valore dipende dalla quantità di informazione disponibile. Sperimentatori in possesso di diversa informazione possono quindi assegnare entropie diverse allo stesso sistema. Infatti, più sono numerosi vincoli, ovvero più numerosi sono gli osservabili che si prendono in considerazione per la determinazione dello stato statistico del sistema e minore sarà l'entropia o incertezza che assegnamo al sistema, perchè restringiamo sempre pù il dominio nelle p sul quale cerchiamo il massimo di H. Allo stesso modo, vincoli che non restringono tale sottoinsieme (per esempio associati a osservabili che sono funzionalmente dipendenti da altri già considerati) sono ininfluenti per la determinazione di S.

Prima di passare alla ricostruzione della termodinamica classica sulla base del PME, vediamo alcuni semplici conseguenze del principio stesso.

a. *Assenza di informazioni sul sistema.*

In questo caso l'unico vincolo da considerare è $\sum_{i=1}^{n} p_i = 1$. Il problema di estremo vincolato del PME ha soluzione $p_i = \frac{1}{n}$ coerentemente con il Principio di ragione insufficiente e con il teorema di Shannon, e l'entropia fisica del sistema è

$$S_n = H(\frac{1}{n}) = -\sum_{i=1}^{n} \frac{1}{n} \ln 1n = \ln n.$$

b. *Sistema ottenuto riunendo due sottosistemi $\mathcal{S}_1$ e $\mathcal{S}_2$ aventi rispettivamente n_1 e n_2 stati.*

Se consideriamo simultaneamente i due sistemi, ogni stato del sistema complessivo sarà assegnato specificando lo stato in cui si trova ognuno dei due sottosistemi. Esso avrà quindi $n_1 n_2$ stati e in assenza di informazioni la sua entropia fisica sarà

$$S_{1+2} = \ln(n_1 n_2) = \ln n_1 + \ln n_2 = S_{n_1} + S_{n_2}.$$

Qesto risultato è vero più in generale: supponiamo che, nel sistema ottenuto riunendo i due sottosistemi $\mathcal{S}_1$ e $\mathcal{S}_2$, il fatto di sapere che $\mathcal{S}_1$è nello stato i non dia nessuna informazione sullo stato j di $\mathcal{S}_2$, ovvero che $p_{ij} = p_i p_j$. Questo fatto, che vedremo anche più avanti in relazione alla nozione di temperatura, si esprime anche dicendo che non vi è correlazione tra gli stati, o anche che l'energia di interazione tra i sottosistemi è trascurabile. Si ha allora subito che

$$S_{1+2} = -\sum_{i,j=1}^{n,m} p_i p_j \ln(p_i p_j) = -\sum_{i}^{n} p_i \ln(p_i) - \sum_{j=1}^{m} p_j \ln(p_j) = S_1 + S_2.$$

2.2 Il PME nella termodinamica di equilibrio

In questo capitolo vediamo come l'applicazione del Principio della massima entropia permetta di generare tutte le relazioni che costituiscono la termodinamica di equilibrio. Quest'ultima altro non è che la descrizione di un sistema complesso all'equilibrio, ovvero in uno stato in cui i valori delle sue variabili macroscopiche sono costanti. Nella termodinamica di equilibrio, quindi, il tempo non compare.

Il primo passo nell'applicazione del PME è l'individuazione dell'insieme degli *stati microscopici* nei quali si puo' trovare il nostro sistema fisico. Per la successiva trattazione statistica, supporremo che essi costituiscano un insieme esaustivo di alternative mutualmente esclusive. Inoltre, per non infilarci subito in un contesto troppo oneroso dal punto di vista matematico, supporremo che tale insieme di stati sia *finito*. Questo per due ragioni: la prima è che la generalizzazione ad un numero infinito di stati non aggiunge alcuna nuova legge fisica se non il metter in luce la necessità di sostituire il PME col il PMDI; l'entropia di Shannon $H(p)$, a differenza dell'entropia relativa $D(p||q)$ infatti non è una grandezza invariante per riparametrizzazioni e questo problema non può essere aggirato nel caso infinito dimensionale. La seconda è che il modello a stati discreti ha un suo interesse e significato fisco autonomo, in quanto esso appare spontaneamente se si pensa ad una discretizzazione in celle di lato ϵ dello spazio delle fasi di un sistema Hamiltoniano oppure se si introduce una descrizione quantistica di un sistema confinato in un volume finito. In tal caso infatti gli stati quantici sono in numero finito e i valori possibili per un osservabile (per esempio l'energia) sono quantità discrete (spettro).

Come detto nell'introduzione, la scelta di un insieme *completo* di osservabili, tale cioè da poter garantire la riproducibilità do un generico stato macroscopico controllando l'insieme delle variabili macroscopiche è onere e onore dello sperimentatore. Questa assunzione di principio non è priva di conseguenze: essa infatti introduce degli elementi di soggettività nella descrizione di un sistema fisico, nel senso che diversi sperimentatori, considerando diversi insiemi di variabili macroscopiche, possono assegnare al sistema una diversa funzione entropia, che risulta essere quindi dipendente dallo sperimentatore. Questa apparente contraddizione con il modo di procedere della fisica si può sanare enunciando il postulato (dovuto a Jaynes), che deve essere poi verificato sperimentalmente, che sperimentatori diversi che hanno accesso alle *stesse informazioni* sul sistema, devono giungere alla *stessa descrizione* dell' entropia del sistema. Chiarito questo aspetto, possiamo enunciare la scelta (soggettiva) di considerare un solo osservabile, l'energia, per la descrizione del sistema. Inoltre supponiamo che l'energia del sistema nello stato i dipenda da dei parametri, specificatamente il volume v e il numero di particelle n, che qui trattiamo come variabili reali positive $(v, n) \in \mathbb{R}^2_+$, il cui valore è noto allo sperimentatore con precisione arbitraria. Supponiamo quindi che sia noto il valore dell'energia

$$E_i = E_i(v, n)$$

per ognuno degli stati microscopici $i = 1, \ldots, n$ e che l'unica informazione disponibile sullo stato macroscopico sia il valore medio (macroscopico) dell'energia. Il PME ci permette di determinare la d.d.p. $p_i, i = 1, \ldots, n$ che rende massima l'entropia $H(p)$

e che soddisfa ai vincoli *lineari* seguenti

$$\sum_{i=1}^{n} p_i - 1 = 0, \qquad \sum_{i=1}^{n} p_i E_i(v,n) - e = 0. \tag{2.1}$$

ove $e \in \mathcal{F} = [\min_i E_i(v,n), \max_i E_i(v,n)]$. Nel prossimo paragrafo esaminiamo il problema matematico soggiacente al PME.

2.2.1 Geometria dei problemi di estremo vincolato

Il Principio della Massima Entropia è, dal punto di vista matematico, un problema di estremo vincolato. La tecnica standard per affrontare tali problemi è il Metodo dei Moltiplicatori di Lagrange, che esponiamo brevemente.

Sia X spazio vettoriale, $\mathcal{P}$ un aperto di X,

$$U : \mathcal{P} \to \mathbb{R}, \quad U = U(p)$$

una funzione (la funzione obbiettivo o utilità nei problemi di ottimizzazione o l'energia potenziale in meccanica) sufficientemente regolare della quale vogliamo trovare i punti critici ristretti al vincolo definito da $F : \mathcal{P} \to Y$, con Y spazio vettoriale. Nel seguito avremo bisogno di considerare vincoli dipendenti da parametri nella forma

$$F : \mathcal{P} \times V \to Y, \quad F = F(p,v),$$

con V spazio vettoriale. Tutte le condizioni che enunceremo, quindi, saranno da intendersi come valide per ogni fissato valore $v \in V$. Supponiamo inoltre che $\dim X = n$, $\dim Y = k < n$, $\dim V = m$ e che il vincolo abbia rango massimo

$$rk\ F_p(p,v) = max = \dim Y = k \quad \forall v \in V.$$

Sia $\mathcal{F} = Im(F) \subset Y$. Introduciamo la funzione di Lagrange

$$G : \mathcal{F} \times V \times \mathcal{P} \times Y^* \to \mathbb{R},$$

$$G(c,v,p,\lambda) = U(p) - \langle \lambda, F(p,v) - c \rangle$$

ove $\langle\ ,\ \rangle$ è la dualità tra Y e Y^*.

1. Condizioni necessarie di estremo. Sia ora $v \in V$ fissato. Il punto $\tilde{p} \in \mathcal{P}$ è un punto di estremo ($dU(\tilde{p}) = 0$) sulla sottovarietà $\mathcal{P}_c = \{p \in \mathcal{P} : F(p,v) = c\}$ se e solo se la coppia $(\tilde{p}, \lambda) \in \mathcal{P} \times Y^*$ soddisfa al sistema di equazioni

$$\begin{cases} G_p(c,v,\tilde{p},\lambda) = U_p(\tilde{p}) - \langle \lambda, F_p(\tilde{p},v) \rangle = 0, \\ G_\lambda(c,v,\tilde{p},\lambda) = F(\tilde{p},v) - c = 0. \end{cases}$$

Osserviamo inoltre che se vale la condizione precedente sul rango (vincoli fra loro indipendenti) allora il moltiplicatore di Lagrange λ associato al problema di estremo vincolato è unico, $\lambda = \tilde{\lambda}(c,v)$.

2. Condizioni sufficienti di massimo. Se la matrice di secondo ordine valutata nel punto di estremo vincolato $(\tilde{p}, \tilde{\lambda})$

$$G_{pp}(c,v,\tilde{p},\lambda) = U_{pp} - \langle \lambda, F_{pp}(p,v) \rangle$$

verifica la condizione

$$G_{pp}(h,h) < 0 \qquad \forall h \in \ker F_p(\tilde{p}, v),$$

allora il punto di estremo vincolato è un *massimo* locale.

Introduciamo quindi la funzione di Lagrange relativa al problema adottando d'ora in poi la nomenclatura del paragrafo precedente, per cui identifichiamo $c = e$ e $v = (v, n) \in \mathbb{R}^2_+$

$$G : \mathcal{F} \times \mathbb{R}_+ \times \mathbb{R}^2_+ \times \mathbb{R}^n_+ \times \mathbb{R}^2 \to \mathbb{R}, \quad G = G(e; v; p; \mu, \lambda)$$

$$G(e; v; p; \mu, \lambda) = H(p) - \mu(\sum_{i=1}^{n} p_i - 1) - \lambda(\sum_{i=1}^{n} p_i E_i(v) - e). \tag{2.2}$$

Notiamo che i vincoli sono tra loro indipendenti se, come supponiamo, $E_i \neq E_j$ per qualche i, j. Ne consegue che moltiplicatori λ, μ associati ai punti di estremo vincolato sono univocamente determinati; inoltre, come vedremo, il fatto che i vincoli siano lineari e che la funzione $H(p)$ sia convessa, conduce ad avere un solo punto di estremo, che sarà un massimo, deducibile applicando le condizioni sufficenti di massimo. Abbiamo quindi il sistema di equazioni (qui $i = 1, \ldots, n$)

$$\begin{cases} G_{p_i} = -1 - \ln p_i - \lambda E_i(v) - \mu = 0, \\ G_\lambda = \sum_{i=1}^n p_i E_i(v) - e = 0, \\ G_\mu = \sum_{i=1}^n p_i - 1 = 0. \end{cases} \tag{2.3}$$

La prima equazione, risolta, permette di esprimere il punto di estremo vincolato in funzione dei moltiplicatori (λ, μ)

$$\tilde{p}_i(\lambda, \mu, v) = e^{-\lambda E_i(v) - (\mu+1)}, \quad i = 1, \ldots, n.$$

Ora imponiamo il vincolo di normalizzazione alla soluzione

$$\sum_{i=1}^{n} \tilde{p}_i = e^{-(\mu+1)} \sum_{i=1}^{n} e^{-\lambda E_i} = 1 \;\Rightarrow\; e^{(\mu+1)} = \sum_{i=1}^{n} e^{-\lambda E_i} =: Z(\lambda; v)$$

e vediamo che

$$\tilde{\mu}(v, \lambda) = \ln Z(\lambda; v) - 1$$

ove si è introdotta la *funzione di partizione*

$$Z(\lambda; v) = \sum_{i=1}^{n} e^{-\lambda E_i(v)}. \tag{2.4}$$

La distribuzione di probabilità che rende massima l'entropia puo' così essere espressa nella forma di Gibbs o *canonica*

$$\tilde{p}_i(\lambda, v) = \frac{e^{-\lambda E_i(v)}}{\sum_{i=1}^n e^{-\lambda E_i(v)}} = \frac{e^{-\lambda E_i(v)}}{Z(\lambda; v)}. \tag{2.5}$$

Proseguiamo cercando ora di eliminare il moltiplicatore λ associato al vincolo sull'energia. Con pochi facili passaggi si vede che

$$\sum_{i=1}^{n} \tilde{p}_i E_i(v) = e \;\Leftrightarrow\; -\frac{\partial}{\partial \lambda} \ln Z(\lambda; v) = e. \tag{2.6}$$

2.2.2 Il caso di k osservabili

La formulazione generale del problema si ha supponendo che, oltre al vincolo di normalizzazione, siano presenti $k \leq n$ vincoli lineari del tipo $(2.1)_2$ descritti da una matrice $A(v) \in Mat(k,n)$, per cui (2.1) diviene

$$\sum_{i=1}^{n} p_i - 1 = 0, \qquad \sum_{i=1}^{n} A_i^\alpha p_i - f^\alpha = 0, \qquad rkA = k. \tag{2.7}$$

Introdotto il vettore $\lambda \in \mathbb{R}^k$ dei Moltiplicatori di Lagrange, la soluzione di massima entropia si scrive

$$\tilde{p}_i(\lambda, v) = \frac{e^{-(A(v)^T\lambda)_i}}{Z(\lambda; v)} = \frac{e^{-A_i^\alpha \lambda_\alpha}}{Z(\lambda; v)} \tag{2.8}$$

ove i moltiplicatori si devono determinare dall'analoga di (2.6)

$$\sum_{i=1}^{n} A_i^\alpha \tilde{p}_i = f^\alpha \;\Leftrightarrow\; -\frac{\partial}{\partial \lambda_\alpha} \ln Z(\lambda; v) = f^\alpha. \tag{2.9}$$

Il problema è quindi ricondotto al problema di inversione di una mappa gradiente. Una condizione sufficente per l'invertibilità globale della mappa gradiente $(2.9)_2$ su un dominio convesso è che la sua matrice jacobiana sia *definita* (vedi Teorema **??** nei Preliminari Matematici).

Mostriamo che in effetti la matrice Jacobiana associata a $\nabla \ln Z$ è *definita positiva*. Con pochi passaggi si mostra che

$$-\frac{\partial^2 \ln Z(\lambda; v)}{\partial \lambda_\alpha \partial \lambda_\beta} = \mathbb{E}_{\tilde{p}}(A^\alpha A^\beta) - \mathbb{E}_{\tilde{p}}(A^\alpha)\mathbb{E}_{\tilde{p}}(A^\beta) = \mathbb{E}_{\tilde{p}}(A^\alpha A^\beta) - f^\alpha f^\beta = K_{\alpha\beta}$$

ove $K_{\alpha\beta}$ è detta matrice delle *covarianze* degli osservabili A^α e A^β. Nel caso $k = 1$ abbiamo semplicemente la varianza dell'energia. La K è certamente simmetrica. Inoltre, avendo a che fare con vincoli lineari non è restrittivo supporre che $f^\alpha \equiv 0$ $\forall \alpha$. Preso $u \in \mathbb{R}^k$, non nullo, si ha che

$$(u, Ku) = \sum_{\alpha\beta} \mathbb{E}_{\tilde{p}}(A^\alpha A^\beta) u^\alpha u^\beta = \mathbb{E}_{\tilde{p}}(\sum_{\alpha\beta} A^\alpha A^\beta u^\alpha u^\beta) = \sum_i [(A^T u)_i]^2 \tilde{p}_i$$

e quindi

$$(u, Ku) = 0 \quad \Leftrightarrow \quad A^T u = 0 \quad \Leftrightarrow \quad u = 0$$

perchè A^T è iniettiva (A è suriettiva per ipotesi). Questo conclude la dimostrazione e dice che i moltiplicatori λ_α hanno un'unica espressione in funzione della coordinate della varietà base.

Calcoliamo ora il valore della funzione entropia in corrispondenza del massimo (2.5). Si ha con semplici passaggi che

$$\tilde{G}(e, v, \tilde{p}, \tilde{\mu}, \tilde{\lambda}) = H(\tilde{p}) = \ln Z(\tilde{\lambda}(e, v); v) + \tilde{\lambda}(e, v)e =: \tilde{S}(e, v).$$

La funzione $\tilde{S}(e, v)$ trovata

$$\tilde{S}(e, v) = \ln Z(\tilde{\lambda}(e, v); v) + \tilde{\lambda}(e, v)e \tag{2.10}$$

esprime quindi l'*entropia fisica* del sistema in esame.

2.2.3 Significato di λ: la temperatura

Nella distribuzione di massima entropia di Gibbs compare il moltiplicatore λ associato al vincolo (informazione) sull'energia media del sistema. Il suo valore caratterizza la distribuzione di probabilità sistema nel senso che, se λ è alto, sono favoriti gli stati a bassa energia, mentre se λ è basso, il sistema tende a essere trovato in stati di energia E_i elevata. Inoltre vale la semplice formula

$$\frac{p_i}{p_j} = e^{-\lambda(E_i - E_j)}$$

per cui, se λ è grande, un piccolo aumento nell'energia del livello è sufficiente per diminuire drasticamente la probabilità di trovare il sistema in quello stato. Possiamo interpretare in senso fisico tale moltiplicatore? Consideriamo due sistemi $\mathcal{S}_1$ e $\mathcal{S}_2$ aventi rispettivamente n e m stati e consideriamo per entrambi il solo osservabile energia. In assenza di altre informazioni sul comportamento del sistema ottenuto riunendo i due sotto sistemi, possiamo solo dire che esso ha $n \times m$ stati e che l'energia dello stato (i, j) sarà (per il principio di conservazione dell'energia, supponendo nulla o trascurabile l'energia di interazione tra i sottosistemi)

$$E_{ij} = E_i + E_j, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m$$

La funzione di partizione per il sistema complessivo $\mathcal{S}_{1+2}$ ha pertanto la forma di prodotto delle funzioni di partizione dei due sottosistemi:

$$Z_{1+2}(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{m} e^{-\lambda(E_i + E_j)} = Z_1(\lambda) Z_2(\lambda)$$

e, di conseguenza, la probabilità di trovare il sistema complessivo nello stato (i, j) è

$$p_{ij}(\lambda) = \frac{e^{-\lambda(E_i + E_j)}}{Z_{1+2}(\lambda)} = \frac{e^{-\lambda E_i}}{Z_1(\lambda)} \frac{e^{-\lambda E_j}}{Z_2(\lambda)} = p_i(\lambda) p_j(\lambda)$$

coerentemente con l'ipotesi fatta che i due sistemi sono statisticamente scorrelati (indipendenti). Ricordiamo che la probabilità (marginale) di trovare il sottosistema $\mathcal{S}_1$ in uno uno stato i è la somma su j delle probabilità p_{ij}, ovvero

$$p_i = \sum_{j=1}^{m} p_{ij} = \frac{e^{-\lambda E_i}}{Z_1(\lambda)}.$$

Sia ora λ_1 il moltiplicatore di Lagrange che descrive la d.d.p. del sistema $\mathcal{S}_1$ considerato singolarmente tramite

$$p_i(\lambda_1) = \frac{e^{-\lambda_1 E_i}}{Z_1(\lambda_1)}.$$

L'analisi svolta si riassume nell'enunciato seguente:

La probabilità di trovare il sistema $\mathcal{S}_1$ nello stato i è la stessa prima e dopo la sua riunione con il sistema $\mathcal{S}_2$ se e solo se $\lambda = \lambda_1$.

Questo fatto si interpreta fisicamente dicendo che uguale λ significa uguale *temperatura* e che due sistemi descritti da una d.d.p. avente uguale λ si trovano in *equilibrio termico*. Inoltre, uno dei due sistemi può essere considerato come un

bagno termico o serbatoio di calore (vedi Jaynes per maggiori precisazioni). Per completare l'interpretazione dobbiamo mostrare che dobbiamo proprio *identificare* λ con $1/T$. Questo ultimo passo si compie utilizzando la definizione di temperatura come fattore integrante che rende la uno–forma dQ esatta, ovvero dicendo che $dS = dQ/T$ è esatta. Per brevità omettiamo quest'ultima derivazione, che il lettore interessato può trovare ad esempio in [Schrödinger].

La discussione sin qui svolta è completamente generale e riflette l'assunzione che riunenendo due (sotto)sistemi non si abbiano altre informazioni oltre a quelle che derivano dal considerare i due sotto-sistemi separatamente. Tale ipotesi può essere o meno fondata in certi sistemi fisici, che però qui non discutiamo. Invece, il ragionamento si estende senza variazioni al caso di più osservabili e possiamo quindi introdurre il concetto generalizzato di eguale k-temperatura per l'osservabile k. In particolare, quanto detto si applica senza difficoltà al potenziale chimico μ.

2.3 Il primo principio della Termodinamica

La formula (2.10) esprime l'entropia fisica associata ad un sistema come il valore di H corrispondente alla soluzione di un problema di estremo vincolato in cui il vincolo è

$$e = \mathbb{E}_p[E] = \sum_{i=1}^{n} E_i(v) p_i$$

Cosa succede se cambiamo di poco l'osservabile E o la soluzione p_i? Differenziando la relazione sopra scritta otteniamo

$$de = \sum_{i=1}^{n} \frac{\partial E_i}{\partial v} p_i dv + \sum_{i=1}^{n} E_i(v) dp_i = \mathbb{E}_p[\frac{\partial E}{\partial v}]dv + dQ = -dW + dQ.$$

Il primo termine si interpreta come lavoro fatto dal sistema e il secondo come calore fornito al sistema fisico. Come cambia il valore di $S(e, v) = H(\hat{p}(e, v))$?

In prima approssimazione, dobbiamo considerare il differenziale della funzione stessa. Se calcoliamo quindi le derivate parziali, usando la (2.6) otteniamo

$$\tilde{S}_e = \frac{\partial \tilde{S}}{\partial e} = \tilde{\lambda}(e, v) + e\frac{\partial \tilde{\lambda}}{\partial e} + \frac{\partial \ln Z}{\partial \lambda}\frac{\partial \tilde{\lambda}}{\partial e} = \lambda,$$

e inoltre allo stesso modo

$$\tilde{S}_v = \frac{\partial \ln Z}{\partial \lambda}\frac{\partial \tilde{\lambda}}{\partial v} + \frac{\partial \ln Z}{\partial v} + e\frac{\partial \tilde{\lambda}}{\partial v} = \frac{\partial \ln Z}{\partial v} = -\tilde{\lambda}\mathbb{E}_{\tilde{p}}(\frac{\partial E}{\partial v})$$

ove, con l'ultimo simbolo $\mathbb{E}_p(A)$ indichiamo il valore medio dell'osservabile A rispetto alla d.d.p. p. L'ultima relazione trovata si dettaglia tenendo conto che con v abbiamo indicato un generico parametro, e che nel nostro caso $v = (v, n)$, rispettivamente volume e numero di particelle. Abbiamo quindi, oltre alla relazione precedente, ove v indica il volume, l'analoga relazione

$$\tilde{S}_n = \frac{\partial \ln Z}{\partial n} = -\tilde{\lambda}\mathbb{E}_{\tilde{p}}(\frac{\partial E}{\partial n}).$$

Queste ultime quantità introdotte hanno un significato fisico importante:

$$-\mathbb{E}_{\tilde{p}}(\frac{\partial E}{\partial v}) = \frac{\partial \ln Z}{\partial v} = P(e, v, n)$$

è il (valore medio) della *pressione* esercitata sul sistema. Invece

$$-\mathbb{E}_{\tilde{p}}(\frac{\partial E}{\partial n}) = \frac{\partial \ln Z}{\partial n} = \mu(e, v, n)$$

è il *potenziale chimico* o affinità del sistema. Mettendo tutto assieme otteniamo la relazione

$$d\tilde{S}(e, v, n) = +\tilde{\lambda}de + \tilde{\lambda}Pdv + \tilde{\lambda}\mu dn. \tag{2.11}$$

che esprime il primo principio nella forma di Gibbs se interpretiamo il moltiplicatore λ come inverso della temperatura ottenendo

$$TdS = de + Pdv + \mu dn. \tag{2.12}$$

2.3.1 L'energia libera

In questo paragrafo vediamo un'interessante legame tra la funzione di Lagrange, l'entropia relativa e l'energia libera. Supponiamo che l'informazione sul sistema sia rappresentata da k vincoli aggiuntivi oltre al vincolo di normalizzazione come in (2.7). Scriviamo la funzione di Lagrange nella forma (2.2)

$$G(p, \lambda, \mu, c) = H(p) - \lambda \cdot (Ap - c) - \mu(\sum_i p_i - 1)$$

La soluzione del problema di estremo vincolato è data da (2.8). Sia $\hat{\lambda} = \hat{\lambda}(c)$ il vettore dei moltiplicatori determinato da (2.9). L'entropia del sistema si esprime con l'ovvia generalizzazione di (2.10)

$$S(c) = G(\hat{p}, \hat{\lambda}, \mu, c) = H(\hat{p}) = \ln Z(\hat{\lambda}(c)) + \hat{\lambda}(c) \cdot c.$$

Nel dominio $D = \{p \in \mathbb{R}^n_+ \ : \ \sum_i p_i = 1\}$ e per $c \in \mathbb{R}^n$ fissato consideramo la seguente funzione

$$\Delta G(p) = G(\hat{p}, \hat{\lambda}, \mu, c) - G(p, \hat{\lambda}, \mu, c) = \ln Z(\hat{\lambda}) - H(p) + \hat{\lambda} \cdot Ap. \tag{2.13}$$

Si ha allora

$$\begin{aligned}
\Delta G(p) &= \ln Z(\hat{\lambda}) - H(p) + \hat{\lambda} \cdot Ap = \\
&= \ln Z(\hat{\lambda}) - H(p) + \sum_{l,i} \hat{\lambda}_l A_{li} p_i = \\
&= \ln Z(\hat{\lambda}) - H(p) + \sum_i p_i \ln e^{-\sum_l A_{li} p_i} = \\
&= \ln Z(\hat{\lambda}) - H(p) + \sum_i p_i \ln e^{-(A^T \hat{\lambda})_i} = \\
&= (\sum_i p_i) \ln Z(\hat{\lambda}) + \sum_i p_i \ln p_i + \sum_i p_i \ln e^{-(A^T \hat{\lambda})_i} = \\
&= -\sum_i p_i \ln \frac{e^{-(A^T \hat{\lambda})_i}}{Z(\hat{\lambda})} + \sum_i p_i \ln p_i = \\
&= \sum_i p_i \ln \frac{p_i}{\hat{p}_i} = D(p||\hat{p})
\end{aligned}$$

quindi l'entropia relativa misura, per c o, equivalentemente, per $\hat{\lambda}(c)$ fissato la differenza tra il valore di G corrispondente al massimo (libero, non vincolato) $\hat{p}$ e quello corrispondente ad un'altra scelta per $p \in D$.

Sia ora $k = 1$, e $A = E$ l'energia, per cui l'energia media del sistema è $e = \mathbb{E}_p(E)$. Introducendo la funzione *energia libera* (è la trasformata di Legendre dell'entropia)

$$\psi = \psi(\lambda, p) = \mathbb{E}_p(E) - \frac{H(p)}{\lambda}$$

che dà l'energia libera fisica del sistema in esame per $p = \hat{p}$ e $\lambda = \hat{\lambda}$,

$$\hat{\psi} = \psi(\hat{\lambda}, \hat{p}) = e - Ts = e - \frac{\ln Z(\hat{\lambda}) + \hat{\lambda} e}{\hat{\lambda}} = -\frac{\ln Z(\hat{\lambda})}{\hat{\lambda}}$$

si ha che

$$\Delta G = D(p\|\hat{p}) = \hat{\lambda}[\psi(\hat{\lambda}, p) - \psi(\hat{\lambda}, \hat{p})] = \hat{\lambda}[\psi - \hat{\psi}] \geq 0. \tag{2.14}$$

Otteniamo quindi che $\hat{p}$, soluzione del PME, è anche soluzione del problema di minimo vincolato per l'energia libera ψ. Si noti che la relazione (2.14) appena trovata permette di associare un dato quantitativo all'enunciato *qualitativo* del Secondo Principio della Termodinamica che dice: lo stato di equilibrio termodinamico del sistema è quello che rende minima l'energia libera per fissata energia e.

2.3.2 Entropia del gas ideale

Il conto che presentiamo è uno dei primi esempi di applicazione del formalismo della Massima Entropia e come tale si trova in tutti i manuali di Meccanica Statistica, ai quali si rimanda per ulteriori approfondimenti. Consideriamo un gas ideale monoatomico, visto come un insieme di N particelle puntiformi di massa m non interagenti contenute in un sottoinsieme Q di $\mathbb{R}^3$ di volume finito V. Lo spazio delle fasi (posizione e velocità) della particella singola è $\Gamma = Q \times \mathbb{R}^3$ e $\Gamma_N = Q^N \times \mathbb{R}^{3N}$ quello del sistema di N particelle. L'energia del gas nella configurazione $x = (q, v) \in \Gamma_N$ coincide con la sua energia cinetica

$$E(x) = \frac{1}{2} m |v|^2 = \frac{1}{2} m \sum_{i=1}^{N} v_i^2.$$

Cerchiamo la d.d.p. $p(x)$ che descrive dal punto di vista statistico il gas a partire dall'informazione macroscopica sull'energia media del gas, e quindi dai vincoli

$$e = \int_{\Gamma_N} E(x) p(x) d^{6N} x, \qquad \int_{\Gamma_N} p(x) d^{6N} x = 1.$$

La soluzione del PME calcolata a partire dall'entropia

$$H(p) = -\int_{\Gamma_N} p(x) \ln p(x) d^{6N} x$$

si ottiene con lo stesso procedimento del caso discreto e ha la forma

$$p(x) = \frac{e^{-\lambda E(x)}}{Z(\lambda)}$$

Il punto cruciale è calcolare la funzione di partizione $Z(\lambda)$ usando il Teorema di Fubini

$$\begin{aligned} Z(\lambda) &= \int_{\Gamma_N} e^{-\lambda E(x)} d^{6N}x = V^N \int_{\mathbb{R}^{3N}} e^{-\lambda \frac{m}{2}|v|^2} d^{3N}v \\ &= V^N [\int_{\mathbb{R}} e^{-\lambda \frac{m}{2} t^2} dt]^{3N} \\ &= V^N \Big(\frac{2\pi}{\lambda m}\Big)^{\frac{3N}{2}} \end{aligned}$$

da cui

$$\ln Z = N \ln V - \frac{3}{2} N \ln \lambda + \frac{3}{2} N \ln \frac{2\pi}{m}.$$

Il valore del moltiplicatore si ottiene dal vincolo

$$e = -\frac{\partial \ln Z(\lambda)}{\partial \lambda} \quad \text{da cui} \quad e = \frac{3}{2} N \frac{1}{\lambda} = \frac{3}{2} N k_B T.$$

Se indichiamo con n il numero di moli, N_A il numero di Avogadro, $N = nN_A$, e $R = N_A k_B$ la costante dei gas, otteniamo la relazione

$$e = \frac{3}{2} nRT.$$

L'entropia del gas ideale è data da ($\lambda(e) = \frac{3N}{2e}$)

$$\begin{aligned} S(e) &= \ln Z(\lambda(e)) + \lambda(e) e = N \ln V - \frac{3}{2} N \ln \frac{3N}{2e} + \frac{3N}{2} \\ &= N \ln V + \frac{3}{2} N \ln e - \frac{3}{2} N \ln \frac{3N}{2} + \frac{3N}{2}. \end{aligned}$$

Distribuzione di Maxwell-Boltzmann

Se consideriamo la soluzone di massima entropia trovata

$$p(x) = p(q, v) = \frac{e^{-\lambda \frac{m}{2}|v|^2}}{V^N [\int_{\mathbb{R}} e^{-\lambda \frac{m}{2} t^2} dt]^{3N}}$$

vediamo che essa risulta fattorizzata nel prodotto di N distribuzioni di probabilità $p_i(q, v)$ della particella singola e che quest'ultima è il prodotto della distribuzione uniforme sulla posizione per la distribuzione gaussiana tridimensionale sulla velocità. Riscriviamo allora quest'ultima usando la relazione$\lambda = 1/k_b T$

$$dp = p(v) d^3 v = \frac{e^{-\frac{mv^2}{2k_B T}}}{(\frac{2\pi k_B T}{m})^{\frac{3}{2}}} d^3 v$$

Se ora passiamo a coordinate polari sferiche e riscriviamo l'elemento di volume come

$$d^3 v = v^2 dv d\theta d\phi \sin\theta, \quad 0 < \theta \ < \pi, \quad 0 < \phi < 2pi$$

si ottiene, dopo l'integrazione nelle variabili angolari, la distribuzione del *modulo* della velocità, detta distribuzione di Maxwell-Boltzmann (è una distribuzione χ^2)

$$p(|v|) = \frac{4\pi}{(\frac{2\pi k_B T}{m})^{\frac{3}{2}}} v^2 e^{-\frac{mv^2}{2k_B T}} d|v|.$$

3 Alcune applicazioni del PME

3.0.3 La distribuzione normale tramite il PME

[1]Fino ad ora abbiamo considerato una sola variabile aleatoria, x. Se generalizziamo al caso di più variabili $y_1, \ldots, y_m$ la distribuzione di probabilità che si riferisce a questo nuovo problema sarà la d.d.p. congiunta $p(y_1, \ldots, y_m)$. Nell'applicazione del PME si introduce l'entropia

$$H = -\int_{\mathbb{R}^m} p(y_1, \ldots, y_m)\, \ln p(y_1, \ldots, y_m)\, dy_1 \ldots dy_m;$$

i vincoli sulla d.d.p. sono sotto forma di valori medi F_k di diverse funzioni $f_k(y_1, \ldots, y_m)$, con $1 \leq k \leq l$, e le relazioni da imporre sono

$$F_k = \int_{\mathbb{R}^m} p(y_1, \ldots, y_m)\, f_k(y_1, \ldots, y_m)\, dy_1 \ldots dy_m.$$

Consideriamo ora un caso che generalizza l'esempio precedente: supponiamo che i vincoli per la d.d.p. siano rappresentati dalle correlazioni

$$r_{ij} = \int_{\mathbb{R}^m} p(y_1, \ldots, y_m)\, y_i y_j\, d^m y \qquad i, j = 1, \ldots, m$$

dove $r_{ij} = r_{ji}$. La d.d.p. che realizza la massima entropia è

$$\begin{aligned} p(\mathbf{y}) &= \frac{1}{Z(\lambda)}\, e^{-\sum_{i,j=1}^{m} \lambda_{ij}\, y_i y_j} = \\ &= \frac{1}{Z(\lambda)}\, e^{-\sum_i \lambda_{ii}\, {y_i}^2 - \sum_{i \neq j} (\lambda_{ij} + \lambda_{ji})\, y_i y_j}. \end{aligned} \tag{3.1}$$

Per semplificare questa espressione introduciamo la matrice simmetrica $\Lambda \in Sym(m)$ tale che:

$$\begin{cases} (\Lambda)_{ii} = \lambda_{ii} = \frac{1}{2}(\lambda_{ii} + \lambda_{ii}) & \forall i \\ (\Lambda)_{ij} = \frac{1}{2}(\lambda_{ij} + \lambda_{ji}) & i \neq j \end{cases}$$

Con queste notazioni si puo' scrivere (3.1) come:

$$p(\mathbf{y}) = \frac{1}{Z(\Lambda)} e^{-(\mathbf{y}^t \Lambda \mathbf{y})}. \tag{3.2}$$

Esprimiamo ora la funzione di partizione come integrale m dimensionale:

$$Z(\Lambda) = \int_{\mathbb{R}^m} e^{-(\mathbf{y}^t \Lambda \mathbf{y})} d^m y;$$

[1]Il contenuto di questo paragrafo è una rielaborazione della stesura originale contenuta nella tesi di A. Mariotti, LTM 2007

per definizione Λ è una matrice simmetrica, quindi diagonalizzabile. Sia R la matrice ortogonale ($\det R = 1$) tale che $R^t \Lambda R = D$ con $D = Diag[a_i]$ matrice diagonale. Con un cambio di variabile $y := Rx$ si trova:

$$\begin{aligned} Z(\Lambda) &= \int_{\mathbb{R}^m} e^{-(R^t \mathbf{x}^t \Lambda R \mathbf{x})} \, |\det R| \, d^m x = \\ &= \int_{\mathbb{R}^m} e^{-\sum_i a_i {x_i}^2} d^m x = \qquad \text{(Fubini)} \\ &= \prod_{i=1}^{m} \int_{\mathbb{R}} e^{-(a_i {x_i}^2)} dx_i = \\ &= \prod_{i=1}^{m} \int_{\mathbb{R}} \frac{1}{\sqrt{a_i}} \, e^{-z^2} dz = \\ &= \frac{\pi^{\frac{m}{2}}}{\prod_{i=1}^{m} \sqrt{a_i}} = \\ &= \pi^{\frac{m}{2}} (\det \Lambda)^{-\frac{1}{2}} \end{aligned} \tag{3.3}$$

Rimane ora da calcolare la matrice dei moltiplicatori di Lagrange;

$$\begin{aligned} r_{ij} &= -\frac{\partial}{\partial \lambda_{ij}} \ln Z(\Lambda) = \\ &= -\sum_{lk} \frac{\partial \ln Z}{\partial \Lambda_{lk}} \frac{\partial \Lambda_{lk}}{\partial \lambda_{ij}} = \\ &= -\frac{1}{2} \left(\frac{\partial \ln Z}{\partial \Lambda_{ij}} + \frac{\partial \ln Z}{\partial \Lambda_{ji}} \right). \end{aligned} \tag{3.4}$$

mentre

$$r_{ii} = -\frac{\partial \ln Z}{\partial \Lambda_{ii}}.$$

Dall'espressione appena trovata per $Z(\Lambda)$ calcoliamo

$$-\frac{\partial \ln Z}{\partial \Lambda_{ij}} = \frac{1}{2} \frac{1}{\det \ \Lambda} \frac{\partial \det \ \Lambda}{\partial \Lambda_{ij}}$$

da cui, introducendo la matrice dei complementi algebrici

$$C(\Lambda)_{ij} = (-1)^{i+j} \det(\hat{\Lambda}_{ij}) = \det \Lambda (\Lambda^{-1})_{ij}$$

e lo sviluppo del determinante secondo la riga i-esima

$$\det(\Lambda) = \sum_{j=1}^{m} \Lambda_{ij} \, C(\Lambda)_{ij},$$

otteniamo

$$-\frac{\partial \ln Z}{\partial \Lambda_{ij}} = \frac{1}{2} (\Lambda^{-1})_{ij}$$

da cui si ha immediatamente

$$r_{ii} = -\frac{\partial \ln Z}{\partial \Lambda_{ii}} = \frac{1}{2} (\Lambda^{-1})_{ii}.$$

Ricordando che Λ^{-1} è una matrice simmetrica perchè Λ lo è, si ha

$$\begin{aligned} r_{ij} &= -\frac{1}{2}\left(\frac{\partial \ln Z}{\partial \Lambda_{ij}} + \frac{\partial \ln Z}{\partial \Lambda_{ji}}\right) = \\ &= \frac{1}{2}\Big(\frac{1}{2}(\Lambda^{-1})_{ij} + \frac{1}{2}(\Lambda^{-1})_{ji}\Big) = \\ &= \frac{(\Lambda^{-1})_{ji}}{2}, \end{aligned} \tag{3.5}$$

e ricaviamo la relazione esistente tra matrice delle correlazioni e moltiplicatori

$$\Lambda = \frac{1}{2}\, r^{-1}. \tag{3.6}$$

Con le precedenti notazioni la funzione di partizione (3.3) diventa allora

$$\begin{aligned} Z(\Lambda) &= \pi^{\frac{m}{2}} (\det \Lambda)^{-\frac{1}{2}} = \\ &= \pi^{\frac{m}{2}} \left(\frac{1}{2}\right)^{-\frac{m}{2}} (\det r)^{\frac{1}{2}} = \\ &= (2\pi)^{\frac{m}{2}} (\det r)^{\frac{1}{2}}, \end{aligned} \tag{3.7}$$

e la d.d.p. è la gaussiana m dimensionale

$$p\,(\mathbf{y}) = \frac{e^{-\,\mathbf{y}^T \Lambda \mathbf{y}}}{Z(\Lambda)} = \frac{e^{-\frac{1}{2}\,\mathbf{y}^T r^{-1} \mathbf{y}}}{(2\pi)^{\frac{m}{2}} (\det r)^{\frac{1}{2}}} \tag{3.8}$$

che non è più fattorizzabile! Calcoliamo ora l'entropia realizzata da questa d.d.p.

$$\begin{aligned} H_{max} &= \ln Z + \sum_{i,j} \lambda_{ij} r_{ij} = \\ &= \ln Z + \mathrm{tr}(\Lambda r) = \\ &= \ln Z + \mathrm{tr}\left(\frac{1}{2}\, r^{-1} r\right) = \\ &= \ln Z + \frac{m}{2} = \\ &= \frac{m}{2} \ln \pi + \frac{1}{2} \ln \frac{1}{|\Lambda|} + \frac{m}{2} = \\ &= \ln \sqrt{(\pi e)^m \frac{1}{|\Lambda|}}. \end{aligned}$$

3.0.4 Informazione ridondante e Inferenza

Abbiamo visto che la distribuzione di probabilità di massima entropia, è determinata univocamente delle informazioni sul sistema di cui disponiamo inizialmente. Ad ognuna di esse viene infatti associato un moltiplicatore di Lagrange che influisce sul risultato finale. Come si comporta il metodo della massima entropia nel caso di informazioni ridondanti, ovvero ricavate ad esempio a partire da un insieme di vincoli alcuni dei quali sono funzionalmente dipendenti dai rimanenti? Dimostriamo che in questo caso il moltiplicatore da associare è nullo, quindi la distribuzione

di probabilità non viene minimamente modificata considerando informazioni ridondanti. Ciò equivale infatti ad aggiungere al nostro problema dei vincoli che non restringono il dominio di H, e di conseguenza non possono influire sulla d.d.p. che rende massima H.
Diamo una giustificazione del fatto che a un'informazione ridondante è necessariamente associato un moltiplicatore nullo in un contesto sufficientemente generale per i nostri scopi. Supponiamo di voler calcolare quella distribuzione di probabilità p_i che rende massima l'entropia

$$H(p) = \sum_{i=1}^{n} p_i \ln p_i$$

rispettando il vincolo

$$\sum_{i=1}^{n} p_i f_i = a.$$

Sia $\hat{p}$ la soluzione a questo problema. Per quanto visto in (2.5), si ha:

$$\hat{p}_i = \frac{e^{-\lambda_a f_i}}{Z(\lambda_a)} = \frac{e^{-\lambda_a f_i}}{\sum_{i=1}^{n} e^{-\lambda_a f_i}}$$

dove λ_a è il moltiplicatore di Lagrange associato al vincolo a

$$a = -\frac{\partial}{\partial \lambda_a} \ln Z(\lambda_a).$$

Consideriamo un altro osservabile $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ e supponiamo che l'informazione relativa a g sia ridondante, ovvero che sia possibile esprimere il suo valore medio utilizzando i dati di cui già disponiamo:

$$b := \mathbb{E}_{\hat{p}}[g] = \sum_{i=1}^{n} g_i \hat{p}_i(\lambda_a) = \sum_{i=1}^{n} g_i \frac{e^{-\lambda_a f_i}}{Z(\lambda_a)}.$$

Applichiamo ora il principio della massima entropia al problema che considera entrambi i vincoli:

$$\sum_{i=1}^{n} f_i \, p_i = a \qquad \sum_{i=1}^{n} g_i p_i = \; b \; = \mathbb{E}_{\hat{p}}[g]$$

Definendo una nuova funzione di Lagrange e imponendo la condizione necessaria di estremo si ricava la d.d.p.

$$p_i = \frac{e^{-\lambda f_i - \mu g_i}}{Z(\lambda, \mu)} = \frac{e^{-\lambda f_i - \mu g_i}}{\sum_{i=1}^{n} e^{-\lambda f_i - \mu g_i}}.$$

Ora si tratta di dimostrare che l'informazione ridondante, data dal vincolo posto per g, corrisponde a un moltiplicatore nullo, $\mu = 0$. Imponendo i vincoli si trova:

$$\frac{\sum_{i=1}^{n} g_i e^{-\lambda f_i - \mu g_i}}{\sum_{i=1}^{n} e^{-\lambda f_i - \mu g_i}} = b = \frac{\sum_{i=1}^{n} g_i e^{-\lambda_a f_i}}{\sum_{i=1}^{n} e^{-\lambda_a f_i}}$$

$$\frac{\sum_{i=1}^{n} f_i e^{-\lambda f_i - \mu g_i}}{\sum_{i=1}^{n} e^{-\lambda f_i - \mu g_i}} = a = \frac{\sum_{i=1}^{n} f_i e^{-\lambda_a f_i}}{\sum_{i=1}^{n} e^{-\lambda_a f_i}}$$

quindi una possibile soluzione per i moltiplicatori è data da

$$\begin{cases} \lambda = \lambda_a, \\ \mu = 0. \end{cases} \tag{3.9}$$

Nel caso in cui le due funzioni f e g siano indipendenti la mappa

$$(\lambda, \mu) \mapsto \left(-\frac{\partial \ln Z(\lambda, \mu)}{\partial \lambda}, -\frac{\partial \ln Z(\lambda, \mu)}{\partial \mu} \right)$$

è un diffeomorfismo (sulla base del Teorema **??**), quindi (3.9) è l'*unica soluzione del problema*

$$a = -\frac{\partial \ln Z(\lambda, \mu)}{\partial \lambda}; \qquad b = -\frac{\partial \ln Z(\lambda, \mu)}{\partial \mu}.$$

Detto diversamente, più un dato di avvicina all'essere "irrilevante", più il suo moltiplicatore tenderà a zero, contribuendo minimamente nella struttura dello spettro.
Il livello di conoscenza del problema non corrisponde alla quantità di dati osservati disponibili, ma all'effettiva informazione che riusciamo a ricavare da essi, trascurando tutto ciò che risulta ridondante. Questa distinzione non rappresenta tuttavia un ostacolo: sarà proprio l'applicazione del PME ad occuparsene, associando automaticamente ad ogni informazione ridondante un moltiplicatore nullo, che non influisca in alcun modo sulla soluzione finale.

Il calcolo dei moltiplicatori ha in definitiva un duplice scopo: è fondamentale per poter determinare la distribuzione di probabilità di massima entropia, e allo stesso tempo indica il "peso" dei dati iniziali sulla soluzione.

Analisi delle serie temporali

Consideriamo un vettore aleatorio $\mathbf{x} = (x_1, x_2, x_3)$ e supponiamo noti i valori di r_0 e r_1 ove

$$r_0 = \mathbb{E}_p[x_i^2],\ i = 1, 2, 3, \qquad r_1 = \mathbb{E}_p[x_1 x_2] = \mathbb{E}_p[x_2 x_3]$$

Cerchiamo una stima di r_2 applicando in questo caso semplice il formalismo della massima entropia. Possiamo considerare la matrice delle covarianze del vettore aleatorio $\mathbf{x}$, nella quale abbiamo fissato $r_2 = \tilde{r}_2$ il valore dedotto da r_0 e r_1 con il PME.

$$r = \begin{pmatrix} r_0 & r_1 & \tilde{r}_2 \\ r_1 & r_0 & r_1 \\ \tilde{r}_2 & r_1 & r_0 \end{pmatrix}$$

La distribuzione di probabilità di massima entropia, data una matrice simmetrica di covarianza R, è:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} |R|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{x}^t r^{-1} \mathbf{x}}$$

Sia Λ la matrice dei moltiplicatori di Lagrange associata a questo problema,

$$\Lambda = \begin{pmatrix} \lambda_1 & \lambda_{12} & 0 \\ \lambda_{12} & \lambda_2 & \lambda_{23} \\ 0 & \lambda_{23} & \lambda_3 \end{pmatrix}$$

dove $\lambda_{13} = \lambda_{31} = 0$ per quanto visto in precedenza (informazione ridondante), e ricordiamo che tra le due matrici definite vale la relazione:

$$r = \frac{1}{2}\Lambda^{-1}$$

Per ricostruire la matrice delle correlazioni è quindi necessario invertire quella dei moltiplicatori, calcolando

$$(\Lambda^{-1})_{ij} = \frac{1}{\det \Lambda}(-1)^{i+j} \det(\hat{\Lambda}_{ji}).$$

Nel nostro caso:

$$\frac{1}{2}\Lambda^{-1} = \frac{1}{2}\frac{1}{\det \Lambda}\begin{pmatrix} \lambda_2\lambda_3 - \lambda_{23}^2 & -\lambda_{12}\lambda_3 & \lambda_{12}\lambda_{23} \\ -\lambda_{12}\lambda_3 & \lambda_1\lambda_3 & -\lambda_1\lambda_{23} \\ \lambda_{12}\lambda_{23} & -\lambda_1\lambda_{23} & \lambda_1\lambda_2 - \lambda_{12}^2 \end{pmatrix} = r$$

Per la matrice Λ^{-1} valgono quindi le uguaglianze:

$$\lambda_2\lambda_3 - \lambda_{23}^2 = \lambda_1\lambda_3 = \lambda_1\lambda_2 - \lambda_{12}^2 = r\, 2 \det \Lambda$$

$$\lambda_{12}\lambda_3 = \lambda_1\lambda_{23}$$

Vogliamo calcolare il valore di $\tilde{r}_2 = \frac{\lambda_{12}\lambda_{23}}{2 \det \Lambda}$ sapendo che

$$\frac{\lambda_1\lambda_3}{2 \det \Lambda} = r_0 \qquad \text{e} \qquad -\frac{\lambda_{12}\lambda_3}{2 \det \Lambda} = r_1$$

per cui si ricava facilmente la stima

$$\tilde{r}_2 = \frac{r_1^2}{r_0}.$$

Questo esempio è il punto di partenza per l'appicazione del PME all'analisi delle serie temporali, che si incontrano in economia, telecomunicazioni, geofisica ecc. Si considera un vettore aleatorio $\mathbf{x} = (x_1, \ldots, x_N)$ che descrive un processo stocastico a tempo discreto, per cui $x_k = X(k\tau)$ ove τ è un intervallo di tempo fissato, *stazionario* e, non è restrittivo, a media nulla. Sono noti i valori medi delle correlazioni fino all'intervallo (lag) $m << N$, ovvero i valori medi degli osservabili

$$x_1^2, x_1x_2, \ldots x_1x_m$$

che esprimono la correlazione tra la v.a. che descrive il fenomeno al tempo t e al tempo $t + k\tau$, $i = 1, \ldots, m$. Si vuole determinare la d.d.p. $p(x_1, \ldots, x_N)$ di massima entropia sulla base dell'informazioni sulle correlazioni fino al lag m. Nelle applicazioni, la d.d.p. trovata si usa per stimare (è un'inferenza) le correlazioni per valori di lag superiori a m (vedi esempio precedente ove $m = 2$ e $N = 3$).

3.0.5 Evoluzione della densità di probabilità lungo il flusso di un campo vettoriale

Queste considerazioni sono il punto di partenza per la trattazione statistica dei sistemi meccanici Hamiltoniani, nella Meccanica Statistica. Consideriamo il Problema di Cauchy definito da un campo vettoriale dipendente dal tempo $g(x,t)$, $x \in \mathbb{R}^n$:

$$\dot{x} = g(x,t), \quad x(0) = X \in \mathcal{P} \subset \mathbb{R}^n. \tag{3.10}$$

Il flusso $\Phi_g(X,t)$ associato al campo vettoriale definisce una famiglia di diffeomorfismi a t fissato e quindi un moto $x(X,t) := \Phi_g(X,t)$ di $\mathcal{P}$. Assegnamo, a $t = 0$, la probabilità che il sistema si trovi nello stato (configurazione) $X \in \mathcal{P}$. Questa sarà descritta da una funzione (densità di probabilità) $\rho^* : \mathcal{P} \to [0,1]$ tale che

$$\int_{\mathcal{P}} \rho^* \; d_x v = 1.$$

E' intuitivo pensare a ρ^* come a una sorta di massa associata ai punti di $\mathcal{P}$, che deve evolvere mantenendo verificata la relazione $\int_{\mathcal{P}_t} \rho = 1$. Si ha quindi, applicando il teorema del trasporto

$$0 = \frac{d}{dt} \int_{\mathcal{P}_t} \rho \; d_x v = \int_{\mathcal{P}_t} [\dot{\rho} + \rho \; div \; e] d_x v = \int_{\mathcal{P}_t} [\dot{\rho} + \rho \; div \; g] d_x v$$

da cui, se possiamo supporre $\mathcal{P}$ arbitrario, si ha la relazione in forma puntuale, detta equazione di Liouville,

$$\dot{\rho} + \rho \; div \; g = \rho' + div(\rho g) = 0 \tag{3.11}$$

che esprime l'equazione di continuità per la probabilità. Un caso di rilevante interesse è quello dei campi vettoriali hamiltoniani, $g = X_H$, ove $H(x)$, $x = (q,p)$, è la funzione di Hamilton, o hamiltoniana. Essi verificano la condizione $div \; X_H = 0$. Si ha allora, applicando la formula di variazione del volume, che (Teorema di Liouville):

$$vol(\mathcal{P}) \text{ è } \textit{costante nel tempo}$$

e che, per un sistema Hamiltoniano,

$$\dot{\rho} = \rho' + grad \; \rho \cdot X_H = \rho' + \frac{d}{dt}(\rho \circ \Phi_{X_H}) = \rho' + \mathcal{L}_{X_H} \rho = \rho' + \{H, \rho\} = 0,$$

ovvero la probabilità è un integrale primo del moto (è costante lungo le traiettorie del sistema).

Due problemi

A. L'equazione di Liouville (3.11) è un'equazione alle derivate parziali lineare per la densità di probabilità incognita $\rho(x,t)$, computazionalmente pesante da risolvere e inoltre necessita della specificazione del dato iniziale ρ^*, problema anch'esso non banale. Nell'approccio numerico alla soluzione del problema di Cauchy (3.10) con dato iniziale incerto, calcolandone la soluzione $x(t, X_i)$, $i = 1, \ldots, N$ per un numero

N finito di dati iniziali scelto campionando $\mathcal{P}$ in base alla ρ^*, si puo' disporre della densità di probabilità *empirica*

$$\rho_N(x,t) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x(t, X_i))$$

ove δ è la delta di Dirac. Problema: qual'è l'errore che si commette usando ρ_N invece della soluzione esatta dell'equazione di Liouville? O meglio, quale è la perdita di informazione determinata dall'uso delle distribuzione empirica? La soluzione di questo problema fa uso della *entropia relativa*

$$D(\rho_N \| \rho) = \int_{\mathcal{P}} \rho_N \ln \frac{\rho_N}{\rho}\ d_x v$$

e permette di stimare il numero N minimo per avere una perdita di informazione accettabile.

B. Supponiamo di avere a che fare con la più generale equazione di Fokker–Planck (**??**)

$$\rho' + div(\rho g) = k\Delta\rho$$

che si riduce alla equazione di Liouville per $k = 0$. Introduciamo nel modello anche un'incertezza sul campo vettoriale g per esempio considerandolo dipendente da un parametro non noto $g_\epsilon = g(x, t, \epsilon)$. In generale, le soluzioni dell'equazione di Fokker–Planck per g e g_ϵ saranno differenti, ma ci interessa vederne l'impatto solo per quanto riguarda la stima del *valor medio* di uno o più osservabili di interesse per il modello in esame.

Problema: Dette ρ e ρ_ϵ le due soluzioni (per semplicità con lo stesso dato iniziale) e indicato con $A : \mathcal{P} \to \mathbb{R}$ l'osservabile, dobbiamo calcolare la quantità

$$\delta A(\epsilon, t) = \int_{\mathcal{P}} A(x)[\rho(x,t) - \rho_\epsilon(x,t)]\ d_x v.$$

La soluzione è il contenuto del Teorema di Fluttuazione–Dissipazione.

Corollario. Siano ρ e ρ_k le soluzioni rispettivamente dell'equazione di Liouville (3.11) e di Fokker–Plank (**??**). Calcolare

$$\frac{d}{dt} D(\rho_k \| \rho) = \frac{d}{dt} \int_{\mathcal{P}} \rho_k \ln \frac{\rho_k}{\rho}\ d_x v$$

per stimare la perdita di informazione su ρ dovuta alla presenza del termine di rumore stocastico di ampiezza proporzionale a k.

3.0.6 Il modello di Ising e le reti neurali.

In questo breve capitolo presentiamo una panoramica di modelli, tutti di grandissimo interesse per le applicazioni e oggetto di studio teorico approfondito. Si tratta di modelli costituiti da un gran numero N di costituenti elementari (atomi o neuroni) i quali si possono per semplicità trovare in un numero finito n (qui $n = 2$ per semplicità) di stati o alternative. E' quindi immediato il collegamento con quanto visto al capitolo sull'interpretazione frequentista del PME. Lo spazio campionario dei possibili stati (sono in numero finito!) della rete è $\Omega = \{A_1, A_2\}^N$ ove $A_i = +1, -1$

rappresenta il momento magnetico di *spin* per gli atomi e $A_i = 1, 0$ rappresenta lo stato del neurone (attivato o silente). Vogliamo determinare la probabilità $p(\omega)$ dello stato $\omega \in \Omega$. Per applicare il PME dobbiamo specificare gli osservabili rilevanti per il sistema. Indichiamo con x_i la componente i-esima del vettore (stringa) $\omega = (A_{i_1}, \ldots, A_{i_N})$, $A_{i_j} = 1, 2$.

Reti neurali

Gli osservabili sono $\sigma_i(\omega) = x_i$ stato del neurone i, $i = 1, \ldots, N$ e $\sigma_{ij}(\omega) = x_i x_j$, prodotto degli stati dei neuroni i e j, $i, j = 1, \ldots, N$. Noti i valori medi degli osservabili, la distribuzione di massima entropia è (indici ripetuti si intendono sommati)

$$p(\omega) = \frac{1}{Z(h.J)} e^{-h_i \sigma_i(\omega) - J_{ij} \sigma_i(\omega) \sigma_j(\omega)};$$

i moltiplicatori h_i hanno il significato fisico di intensità del campo locale di attivazione e J_{ij} di intensità della connessione sinaptica. Ovviamente, il valore dei moltiplicatori è univocamente determinato dalla specificazione dei valori medi.

Modello di Ising

Usiamo le notazioni dell'esempio precedente, ove ora σ_i è il valore dello spin dell'atomo nel sito i del reticolo di N atomi. L'osservabile (dipendente dal parametro τ) che si introduce, e che ha il significato di energia del reticolo è l' Hamiltonana di Ising

$$H_\tau(\omega) = -\sum_{i,j} \sigma_i(\omega) \sigma_j(\omega) - \tau \sum_i \sigma_i(\omega)$$

ove il parametro τ è l'intensità di un campo magnetico esterno. La distribuzione di massima entropia, noto il valore medio del'energia, è

$$p(\omega) = \frac{1}{Z(\beta)} e^{\beta[\sum_{i,j} \sigma_i(\omega)\sigma_j(\omega) - \tau \sum_i \sigma_i(\omega)]}$$

e $\beta = 1/kT$ ha il significato di temperatura. Un modello intermedio tra le reti neurali e il reticolo di Ising è quello dei *vetri di spin* o spin–glasses, nei quali la d.d.p. che descrive il sistema ha la stessa forma di quella delle reti neurali, solo che i parametri h_i e J_{ij} si suppongono *assegnati* in maniera casuale. Per esempio $Prob(J_{ij})$ è una distribuzione gaussiana con varianza che dipende dalla distanza dei siti i e j.

4 Approfondimenti

4.1 Legge dei grandi numeri e entropia

Viene illustrata ora la fondamentale connessione che esiste fra il PME e le frequenze osservate in N prove ripetute. Sia dato un esperimento che consiste in N osservazioni indipendenti di un evento casuale nel quale ad ogni osservazione si verifichi una delle n alternative $\Omega = \{A_1, A_2, ...A_n\}$ esaustive e mutualmente esclusive. Detto $\Omega_N = \{A_1, \ldots, A_n\}^N$ lo spazio campionario dei possibili risultati, $card\ \Omega_N = n^N$, ciascun risultato di tale esperimento è del tipo $\omega = (A_{i_1}, \ldots, A_{i_N}) \in \Omega_N$ con $i_1, \ldots, i_N \in \{1, \ldots, n\}$. Se Ω è dotato di una distribuzione di probabilità $p = (p_1, \ldots, p_n)$, $p_i = Prob(A = A_i)$, l'ipotesi di osservazioni (estrazioni) indipendenti equivale a dotare Ω_N della probabilità prodotto, per cui

$$\mathbb{P}(\omega) = Prob(A_1 = A_{i_1}, \ldots, A_n = A_{i_n}) = \prod_{k=1}^{N} Prob(A_k = A_{i_k}). \tag{4.1}$$

Ogni ω definisce un insieme di numeri interi (si dice anche il 'tipo' di ω) che contano il numero di volte in cui compare A_i in ω:

$$N_i(\omega) \geq 0 \quad \text{con} \quad \textstyle\sum_{i=1}^{n} N_i(\omega) = N.$$

E' utile introdurre la funzione vettoriale $X : \Omega_N \to Q_N$, $X = (X_1, \ldots, X_n)$,

$$X_i(\omega) = \frac{N_i(\omega)}{N}, \qquad \sum_{i=1}^{n} X_i(\omega) = 1, \tag{4.2}$$

ove abbiamo indicato con

$$Q_N = \{q \in \mathbb{Q}^N, q_i = \frac{N_i}{N}, \sum_{i=1}^{n} N_i(\omega) = N\}, \tag{4.3}$$

l'insieme dei possibili vettori di frequenze dei risultati in N osservazioni. Tramite la X possiamo riscrivere (4.1) come

$$\mathbb{P}(\omega) = \prod_{i=1}^{N} {p_i}^{N_i} = \prod_{i=1}^{N} {p_i}^{NX_i(\omega)}.$$

Ricordiamo inoltre che per le variabili *binomiali*[1] X_i in (4.2) vale

$$\mathbb{E}[X_i] = \sum_{\omega \in \Omega_N} X_i(\omega)\mathbb{P}(\omega) = p_i, \tag{4.4}$$

[1]Ricordiamo che solo $n-1$ delle variabili binomiali X_i sono indipendenti in virtù della $(4.2)_2$.

$$var(X_i) = \mathbb{E}[(X_i - p_i)^2] = \frac{p_i(1-p_i)}{N}, \quad \forall i. \tag{4.5}$$

Introduciamo l'insieme, definito per $q \in Q_N$,

$$B(q) = \{\omega \in \Omega_N \ : \ X_i(\omega) = q_i\}, \quad q = (q_1, \ldots, q_n). \tag{4.6}$$

Gli elementi di $B(q)$ sono tutte le possibili sequenze di N osservazioni indipendenti che danno luogo allo stesso vettore di frequenze q e si vede immediatamente che su di essi $\mathbb{P}$ è costante. La variabile aleatoria vettoriale X induce una partizione di Ω_N in classi di equivalenza rispetto alla relazione di equivalenza $\omega \sim \omega'$ se e solo se $X(\omega) = X(\omega')$ e che tali classi sono proprio i sottinsiemi $B(q)$ definiti in (4.6).La cardinalità di $B(q)$ è espressa dal coefficiente multinomiale

$$card(B(q)) = W(q) = \frac{N!}{\prod_{i=1}^n (Nq_i)!}.$$

La distribuzione di probabilità indotta da X sull'insieme *finito* Q_N è la distribuzione multinomiale

$$\eta(q) = \mathbb{P}_X(q) = \mathbb{P}(B(q)) = \frac{N!}{\prod_{i=1}^n (Nq_i)!} \prod_{i=1}^N {p_i}^{NX_i(\omega)} = W(q) \prod_{i=1}^N {p_i}^{Nq_i}. \tag{4.7}$$

Vale infatti

$$\sum_{q \in Q_N} \eta(q) = (p_1 + p_2 + \ldots + p_n)^N = 1. \tag{4.8}$$

Se, in particolare la distribuzione su Ω è quella uniforme, $p_i = \frac{1}{n}$, la distribuzione indotta sullo spazio prodotto Ω_N è ancora uniforme

$$\mathbb{P}(\omega) = \prod_{i=1}^N {p_i}^{NX_i(\omega)} = \prod_{i=1}^n \left(\frac{1}{n}\right)^{N_i} = \left(\frac{1}{n}\right)^{\sum_{i=1}^n N_i} = \frac{1}{n^N} = \frac{1}{card\ \Omega_N}$$

e quella indotta su Q_N è definita dal coefficiente multinomiale $W(q)$

$$\eta(q) = \mathbb{P}(B(q)) = \frac{W(q)}{card\ \Omega_N} = \frac{N!}{\prod_{i=1}^n (Nq_i)!} \frac{1}{n^N}. \tag{4.9}$$

Osserviamo inoltre che, per ogni $\omega \in B(q)$

$$\begin{aligned} \mathbb{P}(\omega) &= \prod_{i=1}^n {p_i}^{NX_i(\omega)} = e^{N \sum_i X_i(\omega) \ln p_i} = \\ &= e^{N \sum_i X_i \ln(p_i \frac{X_i}{X_i})} = e^{N \sum_i X_i \ln X_i} e^{-\sum_i X_i \ln(\frac{X_i}{p_i})} = \\ &= e^{-N(H(X(\omega)) + D(X(\omega)|p))} = e^{-N(H(q) + D(q\|p))} \leq e^{-NH(q)} \end{aligned}$$

che dipende solo, come ci aspettavamo, da q. Essendo l'entropia relativa una quantità sempre non negativa e nulla solo se $q = p$, concludiamo subito che $\mathbb{P}(\omega)$ è massima se $\omega \in B(p)$ e che $\mathbb{P}(\omega)$ tende a zero per N tendente all'infinito per ogni ω. Come si comporta la distribuzione indotta η su Q_N? Troviamo subito che

$$\eta(q) = \mathbb{P}(B(q)) = W(q)\mathbb{P}(\omega) = W(q) e^{-N(H(q) + D(q\|p))}. \tag{4.10}$$

Ci interessiamo ora al comportamento limite di η al tendere di N all'infinito. Preliminarmente, osserviamo che $card\ \Omega_N = n^N$ mentre

$$card\ Q_N < (N+1)^n \tag{4.11}$$

dal momento che $Nq_i \in \{0, \dots, N\}$ per ogni i. Quindi il numero di possibili sequenze cresce esponenzialmente con N mentre il numero di possibili vettori frequenze cresce al più polinomialmente con N. Per poter avanzare dobbiamo trovare il modo di stimare i coefficienti multinomiali $W(q)$ in (4.10). Usiamo l'approssimazione di Stirling per i fattoriali

$$\ln N! \sim (N+\frac{1}{2})\ln N - N + \frac{1}{2}\ln 2\pi \qquad \text{per} \quad N \to +\infty.$$

Si ottiene

$$\begin{aligned}
\ln W(q) &= \ln \frac{N!}{(Nq_1)!\cdots(Nq_n)!} = \ln N! - \sum_{i=1}^{n} \ln (Nq_i)! = \\
&\sim (N+\frac{1}{2})\ln N - N - \sum_{i=1}^{n}[(Nq_i+\frac{1}{2})\ln(Nq_i) - Nq_i] - \frac{n-1}{2}\ln 2\pi = \\
&= N\ln N + \frac{1}{2}\ln N - \sum_{i=1}^{n}(Nq_i+\frac{1}{2})(\ln N + \ln q_i) - \frac{n-1}{2}\ln 2\pi = \\
&= \frac{1}{2}\ln N - N\sum_{i=1}^{n} q_i \ln q_i - \frac{n}{2}\ln N - \frac{1}{2}\sum_{i=1}^{n}\ln q_i - \frac{n-1}{2}\ln 2\pi = \\
&= NH(q) - \frac{n-1}{2}\ln 2\pi N - \frac{1}{2}\sum_{i=1}^{n}\ln q_i
\end{aligned}$$

e quindi

$$W(q) \ \sim \ e^{NH(q)} \frac{1}{(2\pi N)^{\frac{n-1}{2}}} \prod_{i=1}^{n} \frac{1}{\sqrt{q_i}}. \tag{4.12}$$

Pertanto dalla (4.10) nel limite $N \to +\infty$,

$$\eta(p) \sim \frac{1}{(2\pi N)^{\frac{n-1}{2}}} \prod_{i=1}^{n} \frac{1}{\sqrt{p_i}}, \quad \eta(q) \sim \frac{1}{(2\pi N)^{\frac{n-1}{2}}} \prod_{i=1}^{n} \frac{1}{\sqrt{q_i}} e^{-ND(q\|p)} \quad \forall q \neq p, \tag{4.13}$$

e inoltre

$$\frac{\eta(q)}{\eta(q')} \ \sim \ \prod_{i=1}^{n} \sqrt{\frac{q'_i}{q_i}} e^{-N(D(q\|p)-D(q'\|p))}, \quad \frac{\eta(q)}{\eta(p)} \ \sim \prod_{i=1}^{n} \sqrt{\frac{p_i}{q_i}}\ e^{-ND(q\|p)}. \tag{4.14}$$

Il rapporto $(4.14)_1$ è zero o infinito a seconda che q sia più lontano o più vicino a p nella "distanza" definita dall'entropia relativa. Nel caso particolare $(4.14)_2$, si vede quindi che, nel limite asintotico, la probabilità di ottenere un vettore di frequenze *diverso* da p tende a zero e che l'entropia relativa è la funzione che si presenta spontaneamente per descrivere questo comportamento. Possiamo studiare in maniera più fine il limite asintotico di $\eta(p) = \mathbb{P}(B(p)) \sim 1/(2\pi N)^{\frac{n-1}{2}}$? Questo

è possibile ricorrendo alla Legge dei Grandi Numeri in forma debole. Si vedrà che, *anche se $B(p)$ asintoticamente ha misura nulla, non appena "ingrossiamo" un po' $B(p)$, ogni sequenza ω viene ad essere ivi contenuta $\mathbb{P}$-quasi certamente* (sequenze tipiche). Definiamo a questo proposito

$$B(q,\varepsilon) = \{\omega \in \Omega_N \ : \ |X_i(\omega) - q_i| < \varepsilon, \ \forall i\}$$

e

$$C(q_i,\varepsilon) = \{\omega \in \Omega_N \ : |X_i(\omega) - q_i| \geq \varepsilon\} \quad i = 1,\dots,n.$$

Dalla legge debole dei grandi numeri applicata alle variabili aleatorie positive X_i e da (4.5) si ha

$$\mathbb{P}(C(p_i,\varepsilon)) \leq \frac{var(X_i)}{\varepsilon^2} = \frac{p_i(1-p_i)}{N\varepsilon^2} \leq \frac{1}{4N\varepsilon^2} \quad \forall i$$

e quindi, *comunque preso* $\varepsilon > 0$, essendo $\Omega_N = B(p,\varepsilon) \cup (\bigcup_{i=1}^n C(p_i,\varepsilon))$,

$$\mathbb{P}(B(p,\varepsilon)) \geq 1 - \sum_{i=1}^n \mathbb{P}(C(p_i,\varepsilon)) \to 1 \quad \text{per} \quad N \to +\infty.$$

Una descrizione sostanzialmente equivalente del comportamento appena trovato per $B(p)$ è fornita dal Teorema di Shannon seguente, che mette nuovamente in luce il ruolo naturale dell'entropia nel descrivere il comportamento di un sistema a finite alternative nel limite asintotico del numero di osservazioni.

Teorema 4.1.1 (Shannon) *Sia $p_i > 0$ per ogni $i = 1,\dots,n$. Per ogni $\varepsilon > 0$ esiste un $N(\varepsilon,p)$ tale che per ogni $N > N(\varepsilon,p)$ si ha*

1. $e^{N(H(p)-\varepsilon')} \leq card\ B(p,\varepsilon) \leq e^{N(H(p)+\varepsilon')}$,
2. $e^{-N(H(p)+\varepsilon')} \leq \mathbb{P}(\omega) \leq e^{-N(H(p)-\varepsilon')}$, *per ogni* $\omega \in B(p,\varepsilon)$
3. $\lim_{N\to+\infty} \mathbb{P}(B(p,\varepsilon)) = 1$

ove ε' è piccolo con ε.

Dimostrazione. La *3.* è conseguenza immediata della legge debole dei grandi numeri. Mostriamo *2.* Ragionando come prima si ha

$$\begin{aligned} \mathbb{P}(\omega) &= \prod_{i=1}^n {p_i}^{NX_i(\omega)} = e^{N\sum_i X_i(\omega)\ln p_i \pm N\sum_i p_i \ln p_i} = \\ &= e^{-NH(p)} e^{N\sum_i (X_i(\omega)-p_i)\ln p_i} = e^{-NH(p)} e^{N\gamma} \end{aligned}$$

quindi, se $\omega \in B(p,\varepsilon)$, possiamo stimare

$$|\gamma| \leq \sum_i |X_i(\omega) - p_i||\ln p_i| \leq \varepsilon \sum_i |\ln p_i| =: \varepsilon'.$$

da cui segue subito *2.* Per mostrare il punto *1.* possiamo usare il risultato appena trovato e stimare

$$1 \geq \mathbb{P}(B(p,\varepsilon)) \geq e^{-N(H(p)+\varepsilon')} card\ B(p,\varepsilon)$$

da cui

$$card\ B(p,\varepsilon) \leq e^{N(H(p)+\varepsilon')}.$$

Infine, da *3.*, ovvero dalla legge debole dei grandi numeri, per ogni $c \in (0,1)$, esiste $N = N(\varepsilon, p)$ sufficientemente elevato per cui $1 \geq \mathbb{P}(B(p,\varepsilon)) \geq 1-c$. Allora,

$$1 - c \leq \mathbb{P}(B(p,\varepsilon)) \leq card\ (B(p,\varepsilon))e^{-N(H(p)-\varepsilon')}$$

da cui, prendendo ε' pari a $\varepsilon' - \varepsilon''$,

$$card\ (B(p,\varepsilon)) \geq (1-c)e^{-N(H(p)-\varepsilon')}e^{N\varepsilon''} \geq e^{-N(H(p)-\varepsilon')}.$$

□

Osservazione. La *2.* del Teorema di Shannon è detta *Proprietà di equipartizione asintotica* e si scrive comunemente come

$$\lim_{N\to+\infty} \frac{1}{N} \ln \frac{1}{\mathbb{P}(\omega)} = H(p), \quad \text{con probabilità } 1.$$

Essa vale per processi stocastici più generali (processi ergodici) di quelli a variabili i.i.d. qui considerati, ove è una semplice conseguenza della legge debole dei grandi numeri. Essa afferma in pratica, che, per successioni sufficientemente lunghe, con probabilità 1 una successione appartiene a $B(p,\varepsilon)$. Esse sono dette *sequenze tipiche*. Si noti che questo non vuol dire che le sequenze in $B(p,\varepsilon)$ siano la maggioranza delle sequenze possibili. Infatti sulla base del teorema appena dimostrato si ha la stima asintotica

$$\frac{card\ B(p,\varepsilon)}{card\ \Omega_N} \sim e^{N(H(p)-\ln n)}$$

che tende a zero se $H(p) < \ln n$ ovvero se p non è la distribuzione uniforme!.

4.1.1 Il teorema di concentrazione dell'entropia

Abbiamo visto nel paragrafo precedente che, se è nota la probabilità a priori p, la quasi totalità delle sequenze ottenibili ha un vettore di frequenze che sta in un intorno arbitrariamente piccolo di p purchè si prendano sequenze sufficientemente lunghe. Tutto questo vale anche in presenza di *vincoli* sulle frequenze ottenibili q, vincoli rappresentanti dell'informazione supplementare sulle condizioni di realizzazione dell'esperimento, ma per il momento non nonsideriamo questo caso più generale. Il problema che ci ponamo ora è il seguente: avendo ottenuto un certo vettore di frequenze $q \neq p$, stimare la probabilità che la differenza ottenuta sia dovuta a) a una "fluttuazione statistica" che ha prodotto una sequenza $\omega \notin B(p,\varepsilon)$, oppure b) che tale differenza sia dovuta ad altre cause non note (per esempio, la "vera" distribuzione a priori è un'altra oppure vi sono dei vincoli sulla realizzazione dell'esperimento dei quali non abbiamo tenuto conto). Tale problema è il problema classico della Teoria delle grandi deviazioni, che cerca di quantificare la probabilità di eventi "rari". Nel seguito vedremo che il Teorema di Concentrazione dell'entropia di Jaynes, del quale diamo una dimostrazione sommaria, da una risposta a tale problema.

Abbiamo introdotto in precedenza lo spazio di probabilità (Q_N, η) finito dotato della distribuzione di probabilità multinomiale (4.7), la cui espressione asintotica

è data da $(4.13)_2$. Ora, in un certo senso, intuitivamente chiaro, vogliamo considerare la distribuzione $\eta(q)$ definita sul sottoinsieme finito Q_N di $\mathbb{Q}^n$ come la 'quantizzazione' di una densità che indichiamo ancora con η definita sul *simplesso* $(n-1)$-dimensionale che 'estende' Q_N

$$\Delta_{n-1} = \{q \in \mathbb{R}^n_+ \ : \ \sum_i q_i = 1\} \subset \mathbb{R}^n.$$

E' immediato constatare che Δ_{n-1} è varietà $(n-1)$-dimensionale la cui misura $(n-1)$-dimensionale è

$$\lambda_{n-1}(\Delta_{n-1}) = \frac{\sqrt{n}}{(n-1)!2^{\frac{n-1}{2}}}.$$

Pensiamo quindi a $\eta(q)$ come il valore medio di η data da $(4.13)_2$ sulla 'cella' centrata in q di misura $\varepsilon = \lambda_{n-1}(\Delta_{n-1})/card\ Q_N$. L'insieme di queste celle costituisce un ricoprimento di Δ_{n-1}. Si ha quindi, usando (un po' sportivamente...) il teorema della media che, per (4.8),

$$\varepsilon = \varepsilon \sum_{q \in Q_N} \eta(q) \sim \int_{\Delta_{n-1}} \eta(q) d\lambda_{n-1}(q) \tag{4.15}$$

Per calcolare questo integrale usiamo il metodo di Laplace (o sviluppo asintotico degli integrali), ovvero sostituiamo a $D(\cdot\|p)$ il suo sviluppo di Taylor di D sapendo che $D(\cdot\|p)$ ha un minimo in $q = p$ ove vale zero, per cui

$$D(q\|p) = \frac{1}{2} D_{qq}(p\|p)[q-p, q-p] + \mathcal{O}(3) = \frac{1}{2} \sum_{i=1}^n \frac{(q_i - p_i)^2}{p_i} + \mathcal{O}(3)$$

e sostituiamo $\sqrt{p_i}$ a $\sqrt{q_i}$ nella funzione integranda. Dobbiamo quindi calcolare, trascurando[2] i termini di ordine $\mathcal{O}(3)$,

$$\frac{1}{(2\pi N)^{\frac{n-1}{2}}} \int_{\Delta_{n-1}} e^{-\frac{N}{2}\sum_{i=1}^n \frac{(q_i-p_i)^2}{p_i}} \prod_{i=1}^n \frac{1}{\sqrt{p_i}} d\lambda_{n-1}(q). \tag{4.16}$$

Con il cambio di variabile

$$\mathbf{q} = L(q-p), \quad L = Diag[\frac{1}{\sqrt{p_i}}], \quad \det L = (\prod_{i=1}^n p_i)^{-\frac{1}{2}},$$

si vede che

$$\Delta_{n-1} = \{\mathbf{q} \in \mathbb{R}^n_+ \ : \sum_i \sqrt{p_i}\mathbf{q}_i = 0\} \equiv \mathbb{R}^{n-1}$$

pertanto l'integrale precedente è l'integrale di una gaussiana $(n-1)$-dimensionale *concentrata sempre più attorno all'origine* (al punto p nelle coordinate originali) che si fattorizza con il teorema di Fubini in

$$\varepsilon \sim \frac{1}{(2\pi N)^{\frac{n-1}{2}}} \Big[\int_{\mathbb{R}} e^{-\frac{Nx^2}{2}} dx \Big]^{n-1} = \frac{1}{(2\pi N)^{\frac{n-1}{2}}} \Big(\frac{2\pi}{N}\Big)^{\frac{n-1}{2}} = \frac{1}{N^{n-1}}.$$

[2] per fare una dimostrazione più rigorosa (e molto più lunga) si puo' stimare il termine $\mathcal{O}(q^3)$ con un piccolo quadrato $\mathcal{O}(q^3) \approx \epsilon\mathcal{O}(q^2)$.

Prima di proseguire, osserviamo che il conto appena svolto permette di dare una stima asintotica di *card* Q_N

$$card\ Q_N = \varepsilon^{-1}\lambda_{n-1}(\Delta_{n-1}) \sim N^{n-1}\frac{\sqrt{n}}{(n-1)!2^{\frac{n-1}{2}}} \sim N^{n-1}$$

più fine di (4.11). Vediamo ora due applicazioni.

Teorema di concentrazione dell'entropia

Consideriamo l'entropia relativa come variabile aleatoria non negativa

$$D(\cdot\|p) : Q_N \to [0, +\infty)$$

e calcoliamo la d.d.p. indotta. Sia, per ogni $d > 0$,

$$\mathcal{D} = \{q \in Q_N \ : \ D(q\|p) \leq d\}.$$

Allora, da (4.10) e da (4.15),

$$Prob\{D(\cdot\|p) \in [0, d]\} = \eta(\mathcal{D}) = \sum_{q\in\mathcal{D}} \eta(q)$$

Siamo interessati al comportamento asintotico della quantità appena scritta. Nel limite asintotico, possiamo: approssimare l'insieme discreto $\mathcal{D}$ con il sottoinsieme $(n-1)$–dimensionale,

$$\mathcal{D} = \{q \in \mathbb{R}^n_+ \ : \ \sum_i q_i = 1, D(q\|p) \leq d\} \subset \Delta_{n-1}$$

e sostituire la sommatoria con l'integrale rispetto alla misura di Lebesgue $n-1$ dimensionale come fatto in precedenza. Pertanto

$$\eta(\mathcal{D}) \sim \frac{1}{\varepsilon}\int_{\mathcal{D}} \eta(q) d\lambda_{n-1}(q).$$

Inoltre usiamo ancora il metodo di Laplace. Si vede che, rispetto al conto precedente, cambia solo il dominio che ora è

$$\mathcal{D} = \{\mathbf{q} \in \mathbb{R}^n_+ \ : \ 0 \leq \sum_i \mathbf{q}_i^2 \leq 2d, \ \sum_i \sqrt{p_i}\mathbf{q}_i = 0\} \equiv \mathcal{B}_{n-1}(\sqrt{2d})$$

ove $\mathcal{B}_{n-1}(\sqrt{2d})$ è la palla di raggio $\sqrt{2d}$ in $\mathbb{R}^{n-1}$. Usando il metodo di 'integrazione per sfere' ci si riduce a

$$\eta(\mathcal{D}) \sim \frac{1}{\varepsilon(2\pi N)^{\frac{n-1}{2}}}\int_{\mathcal{B}_{n-1}(\sqrt{2d})} e^{-\frac{N\mathbf{q}^2}{2}} d\lambda_{n-1}(\mathbf{q}) = \frac{s_{n-2}}{\varepsilon(2\pi N)^{\frac{n-1}{2}}}\int_0^{\sqrt{2d}} e^{-\frac{Nr^2}{2}} r^{n-2} dr$$

ove $s_{n-2} = \frac{2\ \pi^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})}$ è la misura rispetto a $d\lambda_{n-2}$ della sfera $(n-2)$–dimensionale. Infine, con il cambio di variabile $Nr^2 = 2t$, dopo qualche passaggio facile si arriva a

$$Prob\{D(\cdot\|p) \in [0, d]\} = \eta(\mathcal{D}) \sim \frac{1}{\Gamma(\frac{n-1}{2})}\int_0^{Nd} e^{-t} t^{\frac{n-1}{2}-1} = F(Nd)$$

ove $F(x)$ è la funzione di distribuzione di una variabile χ^2 con $n-1$ gradi di libertà

$$F(x) = \frac{\int_0^x e^{-t} t^{\frac{n-1}{2}-1} dt}{\int_0^\infty e^{-t} t^{\frac{n-1}{2}-1} dt} = \frac{\int_0^x e^{-t} t^{\frac{n-1}{2}-1} dt}{\Gamma(\frac{n-1}{2})}$$

Pertanto:

Teorema 4.1.2 (E.T. Jaynes) *Asintoticamente, $ND(\cdot\|p)$ è variabile aleatoria distribuita come χ^2_{n-1}.*

Applicazioni.

Teorema di Sanov

Sia E un sottoinsieme di misura piena di Δ_{n-1} sufficentemente regolare, per esempio coincidente con la chiusura del suo interno. Vogliamo stimare la probabilità di ottenere un vettore di frequenze in $Q_N \cap E$. Dobbiamo calcolare

$$\eta(Q_N \cap E) \sim \eta(E) = \varepsilon^{-1} \int_E \eta(q) d\lambda_{n-1}(q).$$

Sia $q^* \in E$ il punto di *minima* distanza in entropia relativa di E da p, $D(q^*\|p) = min\{D(q|\ p)$ for all $q \in E\}$. Allora, applicando ancora il metodo di Laplace, possiamo scrivere

$$D(q|\ p) = D(q^*\|p) + \frac{1}{2} \sum_{i=1}^n \frac{(q_i - q_i^*)^2}{p_i} + \mathcal{O}(3)$$

e quindi ci riconduciamo a calcolare, vedi (4.16), l'integrale di una gaussiana $(n-1)$-dimensionale concentrata attorno a q^*

$$\frac{e^{-ND(q^*\|p)}}{\varepsilon(2\pi N)^{\frac{n-1}{2}}} \int_E e^{-\frac{N}{2}\sum_{i=1}^n \frac{(q_i - q_i^*)^2}{p_i}} \prod_{i=1}^n \frac{1}{\sqrt{q_i^*}} d\lambda_{n-1}(q)$$

Nel limite asintotico, la differenza tra l'integrale su E e quello sull'intero simplesso diviene trascurabile per cui otteniamo l'enunciato del Teorema di Sanov

$$\eta(E) \sim \prod_{i=1}^n \sqrt{\frac{p_i}{q_i^*}} e^{-ND(q^*\|p)} \sim e^{-ND(q^*\|p)}.$$

4.1.2 Legame tra entropia relativa e test χ^2

Abbiamo visto che l'entropia relativa esprime la "distanza" tra due distribuzioni di probabilità e che $D(\varphi||p)$ misura la probabilità che il vettore di frequenze φ sia ottenuto in N prove ripetute. Questo problema è anche quello che si affronta e risolve attraverso il test del χ^2. Il legame tra i due argomenti è immediatamente visibile: sviluppiamo in serie di Taylor la funzione di φ $D(\varphi||p)$, tenendo conto che ha un minimo in p e che ivi è nulla. Allora, dalla (2.4),

$$D(\varphi||p) = \frac{1}{2} Hess(D)(\varphi - p)^2 + \mathcal{O}((\varphi - p)^3) = \frac{1}{2} \sum_i \frac{1}{q_i} (\varphi_i - p_i)^2 + \mathcal{O}((\varphi - p)^3)$$

per cui

$$2ND(\varphi||p) = \sum_i \frac{(N\varphi_i - Np_i)^2}{Nq_i} + \mathcal{O}((\varphi - p)^3) = \chi^2_{n-1} + N\mathcal{O}((\varphi - p)^3).$$

ovvero, nel limite asintotico $N \to +\infty$, $2ND$ è variabile aleatoria distribuita come χ^2_{n-1}.

Tavole di contingenza

Siano $X : \Omega \to \mathcal{X}$ e $Y : \Omega \to \mathcal{Y}$ due variabili aleatorie, ove $\mathcal{X} = \{x_1, \ldots, x_n\}$ e $\mathcal{Y} = \{y_1, \ldots, y_m\}$. Siano p_i e p_j le distribuzioni di probabilità associate e p_{ij} la loro d.d.p. congiunta. Se le variabili aleatorie sono indipendenti, i.e. $p_{ij} = p_i p_j$, allora si ha che

$$H(XY) = \sum_{ij} p_{ij} \ln p_{ij} = H(X) + H(Y).$$

E' ragionevole introdurre come misura della dipendenza tra le variabili la quantità simmetrica non negativa

$$H(X) + H(Y) - H(XY) = \sum_{ij} p_{ij} \ln \frac{p_{ij}}{p_i p_j} = D(p_{XY} \| p_X p_Y) =: I(X, Y)$$

comunemente indicata come mutua informazione tra X e Y. Ragionando come prima, possiamo considerare lo sviluppo di Taylor di D attorno al suo minimo P_{XY} e ottenere

$$I(X, Y) = \frac{1}{2} \sum_{ij} \frac{(p_{ij} - p_i p_j)^2}{p_i p_j} + \mathcal{O}(3).$$

Se eseguiamo N osservazioni ripetute delle due variabili X e Y (lancio di due dadi a n e m facce), e approssimiamo $p_i \approx \frac{N_i}{N}$, $p_j \approx \frac{N_j}{N}$ e $p_{ij} \approx \frac{N_{ij}}{N}$, per cui $p_i p_j \approx N_i N_j N^{-1}$, la relazione precedente porge

$$2NI(X, Y) = 2ND(p_{XY} \| p_X p_Y) = \sum_{ij} \frac{N^2 (p_{ij} - p_i p_j)^2}{N p_i p_j} = \sum_{ij} \frac{(N_{ij} - N^{-1} N_i N_j)^2}{N^{-1} N_i N_j}$$

L'ultima relazione è l'espressione del test di χ^2 per saggiare l'ipotesi che le variabili X e Y siano indipendenti.

4.1.3 Entropia congiunta ed entropia condizionata

Nel precedente paragrafo abbiamo definito l'entropia associata ad una singola variabile casuale X. E' immediato ora estendere tale definizione al caso di due o più variabili casuali X_k a valori in un insieme finito $\Omega = \{A_1, \ldots, A_n\}$ di alternative mutualmente esclusive. La distribuzione di probabilità congiunta $p(i_1, \ldots, i_N)$ definisce quindi la probabilità che l'esito di un esperimento aleatorio sia $\omega = (i_1, \ldots, i_N) \in \Omega_N = \Omega^N$, dove $i_k \in \Omega$ rappresenta il valore assunto dalla variabile casuale X_k. Consideriamo quindi la variabile aleatoria vettoriale $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ a valori in Ω^N

Definizione 4.1 (Entropia congiunta) *L' entropia congiunta è data da:*

$$H(\boldsymbol{X}) = H(X_1, \ldots, X_N) = -\sum_{\omega \in \Omega_N} p_{\boldsymbol{X}}(\omega) \log p_{\boldsymbol{X}}(\omega).$$

Ad esempio, nel caso $N = 2$, ponendo $i_1 = i$ e $i_2 = j$, si ha esplicitamente:

$$H(X_1, X_2) = -\sum_{i,j \in \Omega} p(i,j) \log p(i,j).$$

Supponiamo ora sia noto che le variabili casuali $X_1, X_2, \ldots, X_{N-1}$ assumano rispettivamente i valori $i_1, i_2, \ldots, i_{N-1}$. Qual è l'entropia rimanente su X_N, con tali informazioni? Studiamo il caso particolare $N = 2$. Supponiamo quindi che la variabile casuale X_1 assuma il valore $i_1 \in \Omega$. Allora l'entropia associata ad X_2 è:

$$H(X_2|X_1 = i_1) = -\sum_{i_2 \in \Omega} p(i_2|i_1) \log p(i_2|i_1)$$

dove

$$p(i_2|i_1) = P\{X_2 = i_2 | X_1 = i_1\}.$$

Definiamo l'entropia condizionata di X_2 supposto noto il valore di X_1 come il valor medio su X_2 dell'entropia della probabilità condizionata, ovvero

$$\begin{aligned} H(X_2|X_1) &= \sum_{i_1 \in \Omega} p(i_1) H(X_2|X_1 = i_1) \\ &= -\sum_{i_1 \in \Omega} p(i_1) \sum_{i_2 \in \Omega} p(i_2|i_1) \log p(i_2|i_1) \\ &= -\sum_{i_1, i_2} p(i_1, i_2) \log p(i_2|i_1). \end{aligned}$$

Siamo ora in grado di dare la definizione dell'entropia condizionata nel caso generale di N variabili casuali.

Definizione 4.2 (Entropia Condizionata) *Date le variabili casuali $\{X_t\}_{t \in T}$, l'entropia condizionata della variabile casuale X_t è:*

$$\begin{aligned} H(X_t|X_{t-1}, \ldots, X_1) &= \sum_{i_1, \ldots, i_{t-1}} p(i_1, \ldots, i_{t-1}) H(X_t|i_1, \ldots, i_{t-1}) \\ &= -\sum_{i_1, \ldots, i_t} p(i_1, \ldots, i_t) \log p(i_t|i_1, \ldots, i_{t-1}). \end{aligned}$$

Entropia congiunta ed entropia condizionata sono fortemente correlate. Cio' è chiaramente dimostrato nel seguente teorema, il quale afferma che la misura dell'incertezza associata ad N variabili casuali congiunte è pari alla somma dell'incertezza associata ad ogni v.c. X_t supposto noto il valore assunto dalle v.c. precedenti $X_{t-1}, X_{t-2}, \ldots, X_1$.

Teorema 4.1 (Regola della catena)

$$H(X_1, X_2, \ldots, X_N) = \sum_{t=2}^{N} H(X_t|X_{t-1}, \ldots, X_1) \tag{4.17}$$

Dimostrazione La regola della catena si dimostra facilmente per induzione sul numero N di variabili casuali. Per $N = 2$ si ha:

$$\begin{aligned} H(X_2, X_1) &= -\sum_{i_1,i_2\in\Omega} p(i_1,i_2)\log p(i_1,i_2) \\ &= -\sum_{i_1,i_2\in\Omega} p(i_1,i_2)\log p(i_2|i_1)p(i_1) \\ &= -\sum_{i_1,i_2\in\Omega} p(i_1,i_2)\log p(i_2|i_1) + \sum_{i_1} p(i_1)\log p(i_1) \\ &= H(X_2|X_1) + H(X_1). \end{aligned} \tag{4.18}$$

Per ipotesi induttiva supponiamo la tesi sia verificata per $N-1$ variabili casuali. Allora, posto $X_1 = (X_1, \ldots, X_{N-1})$ e $X_2 = X_N$ ed usando una generalizzazione immediata di (4.18), si ha

$$\begin{aligned} H(X_1,\ldots,X_N) &= H(X_1,\ldots,X_{N-1}) + H(X_N|X_{N-1},\ldots,X_1) \\ &= \sum_{t=1}^{N-1} H(X_t|X_{t-1},\ldots,X_1) + H(X_N|X_{N-1},\ldots,X_1) \\ &= \sum_{t=1}^{N} H(X_t|X_{t-1},\ldots,X_1) \end{aligned}$$

□

Osserviamo che l'entropia condizionata non gode della proprietà di simmetria. In generale quindi, se X ed Y sono due variabili casuali a valori in $\mathcal{A}$, sia ha

$$H(X|Y) \neq H(X|Y)$$

Tuttavia vale la seguente uguaglianza

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

5 Termodinamica di non equilibrio

5.1 Preliminari matematici

In questo capitolo[1] intendiamo studiare, senza entrare nel dettaglio, alcuni aspetti introduttivi della teoria dei processi stocastici. Ricordiamo che tale teoria si occupa dello studio di fenomeni casuali che evolvono nel tempo.
Precisamente rivolgiamo la nostra attenzione ad una particolare famiglia di processi, detta *catene di Markov* dal nome del matematico russo che per primo ne definì formalmente la teoria. Agli inizi del '900 egli si volse infatti allo studio delle grandezze aleatorie dipendenti, e nello specifico sviluppo' la teoria di questa particolare sotto classe dei processi stocastici a tempi discreti, caratterizzati dal fatto che lo stato futuro dipende unicamente dallo stato del sistema all'istante presente ed è indipendente dal passato. Come vedremo alla fine di questo capitolo, Markov dimostro' che l'indipendenza tra le variabili aleatorie non è una condizione necessaria per l'applicazione della Legge dei Grandi Numeri.

5.1.1 Processi stocastici

Sia $(\Omega, \mathcal{F}, P)$ uno spazio di probabilità e sia $(\mathcal{A}, \alpha)$ uno spazio misurabile. Una funzione misurabile $X : \Omega \longrightarrow \mathcal{A}$ si dice variabile casuale a valori in $(\mathcal{A}, \alpha)$. Si dice che X è una variabile casuale discreta se $X(\Omega)$ è finito.
Un Processo Stocastico è definito come una famiglia di variabili aleatorie $\{X_t\}$, con $t \in T$. In particolare il processo è detto *a tempo discreto* se l'insieme T è un insieme discreto (ad esempio $T = \{0, 1, 2, \ldots\}$ oppure $T = \{0, 1, 2, \ldots, m\}$), *a parametro continuo* se T è un intervallo non singolare dei numeri reali (ad esempio $T = [0, 1]$).
I processi stocastici sono inoltre classificati a seconda della natura delle variabili aleatorie $\{X_t\}$; si parlerà quindi di *processi discreti* se esse sono variabili a valori discreti, di *processi continui* se sono continue. I valori che possono essere assunti dalle variabili aleatorie sono detti *stati* ed indichiamo con $\mathcal{A}$ l'*insieme degli stati* $\{A_1, A_2, \ldots\}$; qualora risulti $X_t = A_i$ diremo quindi che il sistema al tempo t si trova nello stato A_i. D'ora in avanti assumeremo che:

- $T = \{1, 2, \ldots, N\}$;
- il processo stocastico $\{X_t\}$ sia discreto;
- l'insieme $\mathcal{A}$ degli stati del processo sia finito ($card\ \mathcal{A} = n$).

[1] Il contenuto di questo capitolo segue essenzialmente la tesi per la LTM di D. Vanzelli, AA 2007-08

Senza perdita di generalità rappresenteremo gli stati di un processo con delle etichette numeriche, in questo modo avremo $\mathcal{A} = \{1, 2, \ldots, n\}$.
Ad ogni istante di tempo $t \in T$ il sistema si troverà quindi in un determinato stato appartenente ad un insieme di stati esaustivi e tra loro mutuamente esclusivi. Indichiamo un generico risultato dell'esperimento aleatorio con

$$\omega = (i_1, i_2, \ldots, i_N)$$

e definiamo lo spazio campionario del processo stocastico, ovvero lo spazio in cui assumono valori le variabili casuali $(X_1, \ldots, X_N)$

$$\Omega = \{\omega = (i_1, \ldots, i_N) \mid i_t \in \mathcal{A} \ \ \forall\, t \in T\}.$$

5.1.2 Catene di Markov

Le Catene di Markov sono particolari processi stocastici $\{X_t\}_{t\in T}$ che verificano alcune proprietà.

Definizione 5.1 (Proprietà di Markov) *Un processo stocastico verifica la proprietà di Markov se per ogni istante di tempo $t \in T$ e per ogni sequenza di stati $i, i_1, \ldots, i_{t-1}, j \in \mathcal{A}$, risulta:*

$$P\{X_{t+1} = j | X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i\} = P\{X_{t+1} = j | X_t = i\}.$$

Questa proprietà indica che la probabilità condizionata di un evento futuro dati lo stato attuale e tutti gli eventi passati dipende esclusivamente dallo stato attuale del processo e non dai precedenti; in pratica è come se il processo avesse una perdita di memoria riguardo al suo passato.

Definizione 5.2 (Catena di Markov) *Un processo stocastico discreto $\{X_t\}_{t\in T}$, avente un insieme di stati $\mathcal{A}$ finito, è detto Catena di Markov se verifica la proprietà di Markov.*

Una catena di Markov si dice finita se T è un insieme finito. Una catena di Markov finita $\{X_t\}_{t\in T}$ è ben definita se sono noti:

- il vettore distribuzione di probabilità iniziale $p_0 = (p_0(1), \ldots, p_0(n))$ definito da
$$p_0(i) = P\{X_0 = i\} \ \ \forall i \in \mathcal{A};$$
- le probabilità di transizione da uno stato i ad uno stato j all'istante t
$$p_t(j|i) = P\{X_{t+1} = j \mid X_t = i\}$$
per ogni coppia di stati $i, j \in \mathcal{A}$ e per ogni tempo $t \in T$.

Per ogni fissato $t \in T$, le probabilità di transizione si possono rappresentare con una matrice $n \times n$

$$P_t = [p_t(j|i)]_{i,j\in\mathcal{A}}.$$

Si osservi che le matrici P_t sono non-negative ed in particolare sono matrici stocastiche per righe, in quanto ogni riga ha somma 1.
Allora diciamo che una catena di Markov è ben definita se sono assegnate le d.d.p. iniziali $p_0(i)$ e $N-1$ matrici delle transizioni $P_1, \ldots, P_{N-1}$.

Definizione 5.3 (Catena di Markov Omogenea) *Una catena di Markov si dice omogenea se le matrici di transizione non dipendono dal parametro tempo t, ovvero se per ogni coppia di stati $i, j \in \mathcal{A}$ verifica la seguente proprietà:*

$$p_t(j|i) = p(j|i) \qquad \forall t \in T.$$

In quanto segue, senza ulteriori precisazioni, ci riferiremo sempre a catene di Markov omogenee e a volte indicheremo le probabilità di transizione $p(j|i)$ con p_{ij}. Le catene di Markov omogenee godono di una notevole semplificazione: esse infatti sono ben definite una volta assegnate le d.d.p. iniziali $p_0(i)$ ed una sola matrice stocastica $P = [p_{ij}]$ che definisce le probabilità di transizione in ogni istante.

In un processo di Markov omogeneo si definisce *probabilità di transizione in k passi*, da uno stato i ad uno stato j, la probabilità che il sistema si trovi nello stato j all'istante $t + k$, posto che il processo si trovava nello stato i all'istante t. Tali probabilità sono indicate con

$$p_{ij}^{(k)} = P\{X_{t+k} = j | X_t = i\}.$$

Le probabilità di transizione p_{ij} corrisponde quindi al caso particolare di probabilità di transizione in un passo.

Si puo' dimostrare che se la d.d.p. al tempo k è $p^{(k)}$, allora la d.d.p. al tempo $k + 1$ è

$$p^{(k+1)} = \sum_{i \in \mathcal{A}} p^{(k)}(i) p_{ij}.$$

Osservando che il termine destro della relazione rappresenta la definizione del prodotto fra matrici, si ha che, in forma matriciale

$$P^{(n)} = P^{(n-1)} P$$

in cui $P^{(n)}$ è la matrice i cui elementi sono le probabilità di transizione al passo n-esimo. La relazione rappresenta l'equazione di Chapman-Kolgomorov che, applicata iterativamente, porta a

$$P^{(n)} = P^n$$

ovvero, la matrice delle probabilità al passo n-esimo è la potenza n-esima della matrice di transizione di partenza.
Appare dunque chiaro quanto affermato sopra: una catena di Markov è completamente specificata se sono note (p_0, P).

Osserviamo inoltre che dalla formula di Bayes segue facilmente che per una catena di Markov la distribuzione di probabilità congiunta dell'esito di un esperimento aleatorio $\omega = (i_1, \ldots, i_N)$ di una catena di Markov puo' essere scritta come:

$$p(i_1, \ldots, i_N) = p(i_1) p(i_2|i_1) \ldots p(i_N|i_{N-1}).$$

Definiamo ora alcune proprietà che caratterizzano gli stati dei processi markoviani.

Definizione 5.4 (Catena di Markov Ergodica) *Una matrice stocastica P è detta ergodica se esiste $k \in \mathbb{N}$ tale che le probabilità di transizione in k passi $p_{ij}^{(k)}$ sono positive per ogni i e j. Una catena di Markov omogenea si dice ergodica se puo' essere definita da una distribuzione iniziale p_0 e da una matrice ergodica.*

Definizione 5.5 *Una d.d.p. π è detta una* distribuzione stazionaria *per la catena di Markov se*

$$\pi P = \pi.$$

Osserviamo che il vettore di d.d.p. stazionaria π è autovettore sinistro della matrice di transizione, corrispondente all'autovalore $\lambda = 1$. In pratica l'applicazione, anche ripetuta, della matrice P a π non modifica la distribuzione di probabilità stazionaria.

Enunciamo ora un risultato fondamentale nella Teoria delle Catene di Markov, senza pero' addentrarci nella dimostrazione.

Teorema 5.1 (Teorema Ergodico per le Catene di Markov [2]) *Sia data una catena di Markov ergodica definita dalla matrice di transizione P. Allora esiste una unica distribuzione di probabilità stazionaria $\pi = (\pi_1, \ldots, \pi_n)$ tale che*

$$\lim_{N \to +\infty} p_{ij}^{(N)} = \pi_j.$$

Quindi se una catena di Markov è ergodica, allora la distribuzione stazionaria è unica, e per ogni distribuzione iniziale p_0 la distribuzione di probabilità della variabile casuale X_N tende alla distribuzione stazionaria per N che tende all'infinito. In pratica una catena di Markov ergodica tende, al crescere di N, a dimenticare le condizioni iniziali del processo (proprietà di perdita progressiva di memoria).

5.1.3 Legge dei grandi numeri per le Catene di Markov ergodiche

Le catene di Markov ergodiche sono particolarmente importanti in quanto ad esse si puo' estendere la Legge dei Grandi Numeri. L'importanza di tale risultato deriva dal fatto che una catena di Markov è costituita da una successione di variabili casuali non indipendenti.

Sia data una catena di Markov ergodica $\{X_t\}_{t \in T}$ con distribuzione stazionaria π e matrice di transizione P. Per ogni esito $\omega = (i_1, \ldots, i_N)$ di un esperimento aleatorio, definiamo le variabili casuali $N_i(\omega)$ uguali al numero di occorrenze dello stato i e le variabili casuali $N_{ij}(\omega)$ uguale al numero di occorrenze della coppia ordinata di stati (i, j) in ω. Il seguente teorema è un'estensione della Legge dei Grandi Numeri alle Catene di Markov ergodiche[3]

Teorema 5.2 *Per ogni $\epsilon > 0$ vale*

[3] Una dimostrazione si puo' trovare in *Probability Theory: An Introductory Course*, Yakov G. Sinai.

$$\lim_{N\to+\infty} P\left\{\left|\frac{N_i(\omega)}{N} - \pi_i\right| \leq \epsilon\right\} = 0,$$
$$\lim_{N\to+\infty} P\left\{\left|\frac{N_{ij}(\omega)}{N} - \pi_i p_{ij}\right| \leq \epsilon\right\} = 0.$$

In pratica, dalla legge dei grandi numeri consegue che:

$$\lim_{N\to+\infty} \frac{N_i(\omega)}{N} = \pi_i \qquad \lim_{N\to+\infty} \frac{N_{ij}(\omega)}{N} = \pi_i p_{ij}$$

e si deduce così un'importante proprietà delle catene di Markov. Definiamo la frequenza dello stato $i \in \mathcal{A}$ in un esperimento ω come il rapporto tra il numero di occorrenze di i in ω ed il numero di stati di ω, ovvero

$$f_i(\omega) = \frac{N_i(\omega)}{N}$$

e analogamente definiamo la frequenza della coppia ordinata di stati $(i, j) \in \mathcal{A}^2$ come il rapporto tra il numero di occorrenze ed il numero dei digrammi in ω, ovvero

$$f_{ij}(\omega) = \frac{N_{ij}(\omega)}{N-1}.$$

Allora, per ogni fissato $\omega \in \Omega$, nel limite di catene di lunghezza infinita, la probabilità di transizione da uno stato i ad uno stato j è pari al rapporto tra le frequenze $f_{ij}(\omega)$ e le frequenze $f_i(\omega)$, ovvero

$$\lim_{N\to+\infty} \frac{f_{ij}(\omega)}{f_i(\omega)} = p_{ij}. \tag{5.1}$$

Enunciamo ora, senza dimostrarlo[4], un risultato di grande importanza nella teoria dell'informazione.

Teorema 5.3 (Riduzione condizionata dell'entropia) *Siano $\{X_t\}_{t\in T}$ variabili casuali a valori nello spazio degli stati $\mathcal{A}$, allora*

$$H(X_{t+1}|X_t, \ldots, X_1) \leq H(X_{t+1}|X_t, \ldots, X_2)$$

e vale l'uguaglianza se e solo se X_{t+1} e $(X_t, \ldots, X_1)$ sono indipendenti.

In pratica l'incertezza associata ad una variabile casuale, in media, decresce quando è noto l'esito di altre variabili casuali.

5.1.4 Proprietà di Equipartizione Asintotica

Nella teoria dell'informazione, l'analogo della Legge dei Grandi Numeri è la *Proprietà di Equipartizione Asintotica (AEP).*
Consideriamo una successione $\{X_t\}_{t\in T}$ di variabili casuali discrete definite su uno stesso spazio di probabilità $(\Omega, \mathcal{F}, P)$. Supponiamo che le X_t siano indipendenti

[4]per una dimostrazione si veda *Elements of Information Theory*, Cover – Thomas

e identicamente distribuite (i.i.d.). Ricordiamo che la Legge dei Grandi Numeri afferma che per variabili casuali i.i.d. con probabilità p, la media campionaria tende alla media probabilistica, per N grande. L'AEP afferma che la quantità $-\frac{1}{N}\log p(\omega)$ tende all'entropia di p, dove $p(\omega)$ è la probabilità di osservare la sequenza $\omega = (i_1, \ldots, i_N)$.

Teorema 5.4 (Proprietà di Equipartizione Asintotica) *Date $X_1, \ldots, X_N$ variabili casuali i.i.d. $\sim p(\omega)$, allora*

$$\lim_{N\to+\infty} -\frac{1}{N}\log p(\omega) = H(p) \tag{5.2}$$

con probabilità 1.

Dimostrazione
Per ipotesi le v.c. X_t sono i.i.d., allora $p(\omega) = p(i_1)\cdot p(i_2)\cdot \ldots \cdot p(i_N)$. Osserviamo inoltre che, essendo le X_t v.c. indipendenti, anche le v.c. definite da $f(X_t) = \log p(X_t)$ sono indipendenti. Segue:

$$\begin{aligned}
\lim_{N\to+\infty} -\frac{1}{N}\log p(\omega) &= \lim_{N\to+\infty} -\frac{1}{N}\sum_{t=1}^{N}\log p(i_t) \\
&= -E[\log p_i] \\
&= -\sum_{i=1}^{N} p_i \log p_i \\
&= H(p)
\end{aligned}$$

dove nella seconda uguaglianza abbiamo usato la Legge dei Grandi Numeri per v.c. indipendenti.

□

Definizione 5.6 (Sequenze Tipiche) *Per ogni fissato $\epsilon \geq 0$ definiamo sequenze tipiche $\omega_\epsilon = (i_1, \ldots, i_N)$ con $\omega_\epsilon \in \Omega_\epsilon \subseteq \Omega$, sequenze che soddisfano alla seguente proprietà:*

$$e^{-N(H+\epsilon)} \leq p(\omega_\epsilon) \leq e^{-N(H-\epsilon)}.$$

Dal teorema AEP segue che per ogni fissato $\epsilon \geq 0$ le sequenze tipiche sono equiprobabili ed hanno probabilità che tende ad 1 per N sufficientemente grande, ovvero

$$\lim_{N\to+\infty} p(\omega_\epsilon) = 1 \qquad \forall\ \omega_\epsilon \in \Omega_\epsilon.$$

5.1.5 Entropia di una catena di Markov

Vediamo ora come dal teorema della legge dei grandi numeri per le catene di Markov (5.2) e dal teorema AEP si deriva l'equazione dell'entropia associata ad una catena di Markov. Per fare cio' abbiamo prima bisogno di studiare l'andamento della funzione

$\log p(\omega)$ per sequenze tipiche ω nel caso di catene di Markov ergodiche. Ricordando che le v.c. Ricordando che le v.c. $N_{ij}(\omega)$ sono definite come il numero di occorrenze della coppia di stati (i, j) in ω, possiamo scrivere

$$p(\omega) = p_0 \prod_{i,j \in \mathcal{A}} p_{ij}^{N_{ij}(\omega)} = \exp\left\{ \log p_0 + \sum_{i,j} p_{ij}^{N_{ij}(\omega)} \right\}$$

e quindi

$$\log p(\omega) = \log p_0 + \sum_{i,j} p_{ij}^{N_{ij}(\omega)}.$$

Dal teorema (5.2), per sequenze tipiche ω vale $N_{ij}(\omega)/N \to \pi_i p_{ij}$. Quindi, per tali ω si ha

$$-\frac{1}{N} \log p(\omega) = -\frac{1}{N} \log p_0 - \sum_{i,j} \frac{N_{ij}(\omega)}{N} \log p_{ij} \to -\sum_{i,j} \pi_i p_{ij} \log p_{ij}.$$

Si ottiene così una funzione molto simile alla funzione entropia introdotta da Shannon. Appare dunque naturale la seguente

Definizione 5.7 (Entropia di una catena di Markov ergodica) *Data una catena di Markov ergodica definita da una distribuzione stazionaria π e da una matrice delle transizioni P, si definisce entropia della catena di Markov*

$$H(\pi, P) = -\sum_{i,j \in \mathcal{A}} \pi_i p_{ij} \log p_{ij}. \tag{5.3}$$

5.1.6 Tasso di Crescita dell'Entropia

In questo paragrafo mostreremo che l'entropia associata ad una catena di Markov $H(\pi, P)$ tende asintoticamente ad una quantità $H(\mathcal{X})$, che chiameremo *tasso di crescita dell'entropia* (in inglese *entropy rate*).
Data una sequenza di N variabili casuali, è naturale chiedersi quale sia l'andamento della funzione entropia al crescere del numero delle variabili casuali. Definiamo allora il tasso entropico come segue

Definizione 5.8 *Il tasso di crescita dell'entropia di un processo stocastico $\{X_t\}_{t \in T}$ è dato da:*

$$H(\mathcal{X}) = \lim_{N \to +\infty} \frac{1}{N} H(X_1, \ldots, X_N) \tag{5.4}$$

quando il limite esiste.

Il tasso entropico così definito assume il significato di entropia media del singolo esito dell'esperimento aleatorio ω.
Diamo ora una seconda definizione di tasso entropico.

Definizione 5.9 *Definiamo tasso di crescita dell'entropia, l'entropia condizionata dell'ultima variabile casuale, supposte note le precedenti*

$$H'(\mathcal{X}) = \lim_{N \to +\infty} H(X_N | X_{N-1}, \ldots, X_1) \tag{5.5}$$

quando il limite esiste.

I due oggetti appena introdotti corrispondono a due differenti nozioni di tasso entropico. Dimostriamo ora un importante risultato valido per i processi stocastici stazionari.

Teorema 5.5 *Per un processo stocastico stazionario i limiti (5.4) e (5.5) esistono e sono uguali, ovvero*

$$H(\mathcal{X}) = H'(\mathcal{X})$$

La dimostrazione di questo teorema richiede prima la dimostrazione di due risultati intermedi.

Lemma 5.6 *Per un processo stocastico stazionario, la probabilità condizionata*

$$H(X_N|X_{N-1}, \ldots, X_1)$$

è una funzione decrescente di N ed ha limite $H'(\mathcal{X})$.

Dimostrazione Per la regola della catena (teorema (5.3)) si ha

$$H(X_{t+1}|X_t, \ldots, X_1) \leq H(X_{t+1}|X_t, \ldots, X_2) = H(X_t|X_{t-1}, \ldots, X_1)$$

dove l'uguaglianza segue dall'ipotesi di stazionarietà della catena.
Poichè $H(X_t|X_{t-1}, \ldots, X_1)$ definisce una successione decrescente di termini non negativi, essa ammette limite, $H'(\mathcal{X})$.

□

Dimostriamo ora un semplice risultato dell'analisi.

Lemma 5.7 (Somma di Cesàro) *Sia a_n una successione convergente $\lim_{n\to+\infty} a_n = \alpha$ e sia b_n la successione definita come la media aritmetica degli a_i, ossia $b_n = \sum_{i=1}^{n} a_i$. Allora b_n è una successione convergente e vale $\lim_{n\to+\infty} b_n = \alpha$.*

Dimostrazione Poichè a_n è una successione convergente, esiste un $n_\epsilon \in \mathbb{N}$ tale che

$$|a_n - \alpha| \leq \epsilon \qquad \forall n \geq n_\epsilon.$$

Allora $\forall\ n \geq n_\epsilon$ possiamo scrivere

$$\begin{aligned} |b_n - \alpha| &= \left|\frac{1}{n}\sum_{i=1}^{n} a_i - \alpha\right| = \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - \alpha)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}|a_i - \alpha| = \frac{1}{n}\sum_{i=1}^{n_\epsilon}|a_i - \alpha| + \frac{1}{n}\sum_{n_\epsilon}^{n}|a_i - \alpha| \\ &\leq \frac{1}{n}\sum_{i=1}^{n_\epsilon}|a_i - \alpha| + \epsilon. \end{aligned}$$

Osservando che, per n abbastanza grande, il primo termine converge a 0, concludiamo che

$$|b_n - \alpha| \leq 2\epsilon$$

e quindi la successione b_n converge ad α.

□

Abbiamo ora tutti gli strumenti per dimostrare il teorema (5.5)

Dimostrazione

Per la regola della catena si ha

$$\frac{1}{N}H(X_1,\dots,X_N) = \frac{1}{N}\sum_{t=2}^{N} H(X_t|X_{t-1},\dots,X_1).$$

Per la seconda definizione data del tasso entropico, sappiamo che le probabilità condizionate tendono, nel limite per $N \to +\infty$ al tasso entropico $H'(\mathcal{X})$.
Inoltre, per il lemma (5.7), la successione definita come la media aritmetica delle entropie condizionate converge anch'essa ad $H'(\mathcal{X})$. Allora, usando il lemma (5.6) si conclude

$$H(\mathcal{X}) = \lim_{N\to+\infty}\frac{1}{N}H(X_1,\dots,X_N) = \lim_{N\to+\infty} H(X_N|X_{N-1},\dots,X_1) = H'(\mathcal{X}).$$

□

In generale $H(\mathcal{X})$ non esiste. Tuttavia esso è ben definito per processi stocastici stazionari, ed in particolare per le catene di Markov.

Teorema 5.8 *Per una catena di Markov stazionaria, il tasso entropico è dato da*

$$H(\mathcal{X}) = H(X_2|X_1).$$

Dimostrazione

Dal teorema (5.5) sappiamo che

$$H(\mathcal{X}) = H'(\mathcal{X})$$

da cui

$$\lim_{N\to+\infty} H(X_N|X_{N-1},\dots,X_1) = \lim_{N\to+\infty} H(X_N|X_{N-1}) = H(X_2|X_1).$$

□

Teorema 5.9 *Sia $\{X_t\}_{t\in T}$ una catena di Markov ergodica con distribuzione stazionaria π e matrice delle transizioni P. Allora il tasso di crescita dell'entropia tende asintoticamente all'entropia $H(\pi,P)$ della catena di Markov:*

$$\lim_{N\to+\infty} H(\mathcal{X}) = H(\pi,P).$$

Dimostrazione Per il teorema (5.8) si ha che

$$H(\mathcal{X}) = H(X_2|X_1).$$

Dalla definizione di entropia condizionata segue

$$\begin{aligned}\lim_{N\to+\infty} H(\mathcal{X}) &= \lim_{N\to+\infty} H(X_2|X_1)\\ &= \lim_{N\to+\infty} -\sum_{i,j\in\mathcal{A}} p_i p_{ij}\log p_{ij}\\ &= -\sum_{i,j\in\mathcal{A}} \pi_i p_{ij}\log p_{ij}.\end{aligned}$$

□

6 Elementi di teoria delle grandi deviazioni

6.1 Premessa

I risultati che esponiamo qui sono validi anche in ipotesi più deboli. Poichè lo scopo di queste pagine è di arrivare rapidamente ad esporre alcune delle idee alla base della teoria delle grandi deviazioni, sacrifichiamo la generalità a favore della semplicità di esposizione.

Se vogliamo studiare il comportamento asintotico della media campionaria, intesa come la variabile aleatoria

$$\bar{X}_n = \frac{1}{n}\sum_i X_i$$

in tuttà generalità, possiamo dire che se le variabili hanno media $E(X_i) = \mu$, allora anche $E(\bar{X}_n) = \mu$; quindi se abbiamo a disposizione molte 'realizzazioni' $\{x_1, \ldots, x_n\}$ delle v.a. X_i, le medie campionarie

$$\bar{x}_n = \frac{1}{n}\sum_i x_i.$$

avranno 'in media' il valore μ. Nelle applicazioni, abbiamo a disposizione solo una realizzazione. Cosa possiamo dire della differenza

$$|\bar{x}_n - \mu| = |\frac{1}{n}\sum_i x_i - \mu|?$$

Cominciamo con una definizione. Sia X_i successione di v.a. dotate di media finita. Non è restrittivo allora supporre che abbiano tutte la stessa media μ.

Definizione 6.1.1 (Legge debole dei grandi numeri) *La successione X_i soddisfa alla* legge debole dei grandi numeri *se converge in probabilità alla v.a. costante μ, $X_i \xrightarrow{P} \mu$, ovvero se $\forall \varepsilon > 0$*

$$\lim_{n\to\infty} P(\{\omega \in \Omega \ : \ |\bar{X}_n(\omega) - \mu)| > \varepsilon\}) = 0.$$

Quindi, purchè si prenda un campione sufficientemente numeroso, possiamo avere una probabilità inferiore a un valore fissato (ma non nulla!) che la media del campione differisca dalla media μ per più di ε. Il seguente risultato permette di stimare $P(|\bar{X}_n - \mu| > \varepsilon)$.

Proposizione 6.1.1 (Disuguaglianza di Chebyshev) *Sia X v.a. di media $E(X) = \mu$ e varianza $var(X) = \sigma^2$ finita. Allora, $\forall \varepsilon > 0$*

$$P(|X - \mu| \geq \varepsilon) \leq \frac{var(X)}{\varepsilon^2}$$

Dimostrazione. Diamo la dimostrazione nel caso in cui X sia v.a. assolutamente continua di densità f_X. La dimostrazione nel caso discreto segue le stesse linee, ma è meno evidente. Non è restrittivo, inoltre, supporre che $\mu = 0$. Allora

$$\begin{aligned} var(X) &= \int_{\mathbb{R}} x^2 f_X(x)dx \geq \int_{-\infty}^{-\varepsilon} x^2 f_X(x)dx + \int_{\varepsilon}^{+\infty} x^2 f_X(x)dx \\ &\geq \varepsilon^2 [\int_{-\infty}^{-\varepsilon} f_X(x)dx + \int_{\varepsilon}^{+\infty} f_X(x)dx\ ] \\ &= \varepsilon^2 P(|X - \mu| \geq \varepsilon). \end{aligned}$$

□

Vediamo una condizione sufficiente per una successione di variabili aleatorie per soddisfare la legge debole dei grandi numeri.

Proposizione 6.1.2 *Sia X_i successione di variabili i.i.d. aventi momento secondo finito. Allora X_i soddisfa la legge debole dei grandi numeri.*

Dimostrazione. Poniamo $E(X_i) = \mu$, $Var(X_i) = \sigma^2$. Allora $E(\bar{X}_n) = \mu$ e

$$var(\bar{X}_n) = var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(\frac{1}{n}X_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) = \frac{\sigma^2}{n}$$

ma allora, usando la disuguaglianza di Chebyshev sopra scritta per la v.a. $\bar{X}_n$

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \tag{6.1}$$

e quindi

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0 \qquad \forall \varepsilon.$$

□

La stima trovata in (6.1) è piuttosto rozza. Il seguente teorema, che descrive un aspetto fondamentale (da qui l'aggettivo centrale che compare nel titolo) del comportamento asintotico di una somma di variabili aleatorie, permette di raffinarla. In pratica, la distribuzione di probabilità che descrive $\bar{X}_n$ si "avvicina" alla distribuzione normale all' aumentare di n e quindi possiamo usare la distribuzione normale per migliorare la (6.1).

Sia X_i successione di variabili *i.i.d* aventi momento secondo finito. Da X_n attraverso il procedimento detto di centratura e normalizzazione otteniamo una v.a

$$Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$$

di media nulla e varianza unitaria. Vale allora la seguente versione del Teorema Centrale del Limite

Teorema 6.1.1 (Teorema Centrale del Limite) *La successione Z_n di v.a. converge in distribuzione a $Y \sim N(0,1)$ ovvero*

$$\lim_{n\to\infty} P(Z_n \in [a,b]) = F_Y(b) - F_Y(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{y^2}{2}} dy. \tag{6.2}$$

6.2 Deviazioni moderate e grandi deviazioni.

Abbiamo ora gli strumenti per poter parlare di deviazioni dal valore atteso della media. Per semplicità supponiamo che X_i sia successione di variabili i.i.d. di media nulla e varianza unitaria (non è restrittivo). Allora, posto $S_n = \sum_{i=1}^n X_i$ vale

$$Z_n = \sqrt{n}\bar{X}_n = \frac{1}{\sqrt{n}} S_n.$$

Dato $\varepsilon > 0$, la quantità

$$P(|S_n| > \sqrt{n}\varepsilon) = P(|Z_n| > \varepsilon) = P(|\bar{X}_n| > \frac{\varepsilon}{\sqrt{n}})$$

è la probabilità di deviazioni moderate, mentre

$$P(|S_n| > n\varepsilon) = P(|\bar{X}_n| > \varepsilon) = P(|Z_n| > \sqrt{n}\varepsilon)$$

è la probabilità di *grandi* deviazioni. Usiamo il Teorema Centrale del Limite per stimare la probabilità di grandi deviazioni. Sia $A = [-\varepsilon, \varepsilon]$ e $A_n = [-\sqrt{n}\varepsilon, \sqrt{n}\varepsilon]$ e indichiamo con A^c, A_n^c i loro complementari. Allora

$$\lim_{n\to\infty} P(|\bar{X}_n| > \varepsilon) = \lim_{n\to\infty} P(|Z_n| > \sqrt{n}\varepsilon) = \frac{1}{\sqrt{2\pi}} \int_{A_n^c} e^{-\frac{x^2}{2}} dx = \sqrt{\frac{n}{2\pi}} \int_{A^c} e^{-\frac{ny^2}{2}} dy$$

ove, nell'ultima eguaglianza abbiamo usato il cambio di variabili $x = \sqrt{n}y$. L'ultimo integrale si può stimare con il *Principio di Laplace*

$$\text{se} \qquad \int_A e^{-\varphi(x)} dx < \infty, \qquad\qquad \lim_{n\to\infty} \frac{1}{n} \ln \int_A e^{-n\varphi(x)} dx = -\inf_{x\in A} \varphi(x)$$

ottenendo

$$\lim_{n\to\infty} \frac{1}{n} \ln P(|\bar{X}_n| > \varepsilon) = \lim_{n\to\infty} \frac{1}{n} \ln \sqrt{\frac{n}{2\pi}} + \lim_{n\to\infty} \frac{1}{n} \ln \int_{A^c} e^{-\frac{ny^2}{2}} dy$$

e quindi

$$\lim_{n\to\infty} \frac{1}{n} \ln P(|\bar{X}_n| > \varepsilon) = \lim_{n\to\infty} \frac{1}{n} \ln P(\bar{X}_n \in A^c) = -\inf_{y\in A_c} \{\frac{y^2}{2}\} = -\frac{\varepsilon^2}{2}.$$

Il risultato trovato ha la forma di una stima di grandi deviazioni e la funzione $I(y) = \frac{y^2}{2}$ è la *rate function*. Si enuncia anche nella forma

$$P(|\bar{X}_n| > \varepsilon) = P(|S_n| > n\varepsilon) \sim e^{-n\frac{\varepsilon^2}{2}}$$

ove il simbolo $\sim$ indica qui uguaglianza nel limite logaritmico ovvero

$$f(n) \sim g(n) \quad \Longleftrightarrow \quad \lim_{n\to\infty} \frac{1}{n} \ln f(n) = \lim_{n\to\infty} \frac{1}{n} \ln g(n).$$

Esercizio. Mostrare che se X_i sono i.i.d. con media μ e varianza σ^2 vale la stima precedente di grandi deviazioni con $I(y) = (y-\mu)^2/2\sigma^2$.

6.2.1 Grandi deviazioni su lanci ripetuti di una moneta

Il risultato trovato in precedenza di basa sullo sviluppo asintotico dell'integrale gaussiano. Qui invece arriviamo ad un'altra stima di tipo grandi deviazioni usando lo sviluppo asintotico del fattoriale

$$n! = n^n e^{-n}\sqrt{2\pi n}(1 + \mathcal{O}(1/n)).$$

Siano X_i variabili i.i.d. a valori in $\{0, 1\}$ con legge $f_X(0) = f_X(1) = 1/2$. Allora $E(X_i) = 1/2$ e $E(S_n) = n/2$. Se considero n lanci indipendenti si ha immediatamente che

$$P(S_n = k) = \frac{1}{2^n}\binom{n}{k} = \frac{1}{2^n}\frac{n!}{k!(n-k)!}$$

Dato $\varepsilon \geq 1/2$, vogliamo calcolare la stima di grandi deviazioni

$$\lim_{n\to\infty} \frac{1}{n} \ln P(|\bar{X}_n| > \varepsilon) = \lim_{n\to\infty} \frac{1}{n} \ln P(|S_n| > n\varepsilon).$$

Vale

$$P(S_n > n\varepsilon) = \sum_{k>\varepsilon n} P(S_n = k) = \sum_{k>\varepsilon n} \frac{1}{2^n}\binom{n}{k}.$$

Se n è sufficientemente elevato possiamo approssimare $n\varepsilon$ con l'intero immediatamente successivo. Allora, posto

$$Q_n(\varepsilon) = max_{k>n\varepsilon}\binom{n}{k} \equiv \binom{n}{n\varepsilon}$$

si ha la stima rozza ma sufficentemente accurata per concludere

$$\frac{1}{2^n}Q_n(\varepsilon) \leq P(S_n > n\varepsilon) \leq (n+1)\frac{1}{2^n}Q_n(\varepsilon).$$

Prima di calcolare $Q_n(\varepsilon)$, osserviamo che la coppia di diseguaglianze scritte implica l'eguaglianza logaritmica

$$P(S_n > n\varepsilon) \sim \frac{1}{2^n}Q_n(\varepsilon).$$

Usando l'approssimazione di Sitrling del fattoriale sopra scritta si ha facilmente

$$Q_n(\varepsilon) = \frac{[\varepsilon(1-\varepsilon)n2\pi]^{-\frac{1}{2}}}{\varepsilon^{\varepsilon n}(1-\varepsilon)^{n(1-\varepsilon)}}$$

da cui

$$\lim_{n\to\infty} \frac{1}{n} \ln \frac{1}{2^n}Q_n(\varepsilon) = -\ln 2 - \varepsilon \ln \varepsilon - (1-\varepsilon)\ln(1-\varepsilon) = I(\varepsilon).$$

Si ha quindi la stima di grandi deviazioni per la media in n lanci

$$P(S_n > n\varepsilon) = P(\bar{X}_n > \varepsilon) \sim e^{-nI(\varepsilon)}$$

con rate function

$$I(\varepsilon) = -\ln 2 - \varepsilon \ln \varepsilon - (1-\varepsilon)\ln(1-\varepsilon).$$

Osservazione. Poichè si ha $I(\varepsilon) = I(1-\varepsilon)$, non è difficile mostrare che lo stesso risultato vale per $P(S_n < n\varepsilon)$, con $\varepsilon < 1/2$.

Interpretazione di $I(\varepsilon)$

Indichiamo con $q = (1/2, 1/2)$ la distribuzione uniforme su $\{0, 1\}$ e con $p(\varepsilon) = (\varepsilon, 1-\varepsilon)$ la distribuzione indotta da $\varepsilon \in [0, 1]$. Allora si ha che

$$I(\varepsilon) = D(p(\varepsilon)\|q) = \sum_{i=1}^{2} p_i \ln(\frac{p_i}{q_i})$$

ove $D(p||q)$ è l'entropia relativa di p rispetto a q. Si tratta di una funzione definita sul simplesso $\sum_i p_i = 1$, convessa, non negativa avente minimo vincolato in $p = q$, ovvero in $\varepsilon = 1/2$ nel nostro caso. Inoltre, posto $f(p) = D(p||q)$ si ha

$$grad\ f(q) = 0, \qquad Hess(f)(q) = Diag[1/q_i]$$

e quindi lo sviluppo di Taylor di I attorno a $\varepsilon = 1/2 + a$ si scrive

$$I(\frac{1}{2} + a) = \frac{1}{2} I''(\frac{1}{2}) a^2 = \frac{1}{2} \sum_{i=1}^{2} \frac{1}{\frac{1}{2}} a^2 = 2a^2.$$

Vediamo quindi che, se ci limitiamo a grandi deviazioni "di piccola ampiezza" la rate function diventa "universale" e pari a quella che si ottiene applicando il teorema Centrale del Limite

$$P\Big(|S_n| > n(1/2 + a)\Big) = P(|\bar{X}_n| > 1/2 + a) \sim e^{-n2a^2}$$

mentre in generale dobbiamo aspettarci che la rate function dipenda dalla natura della successione X_i. In effetti gli oggetti interessanti della teoria della grandi deviazioni sono le rate function e la loro interpretazione in termini di entropia, energia libera eccetera nelle loro applicazioni alla termodinamica statistica.

6.3 Primi teoremi sulle grandi deviazioni

La teoria delle grandi deviazioni si presenta come una collezione di teoremi. Un primo risultato che permette di calcolare le rate function è il Teorema di Cramer per v.a. scalari

Teorema 6.3.1 (Cramer) *Sia X_i successione di variabili reali i.i.d., $X_1 \sim p_X(x)$ Sia*

$$Z(\lambda) = E(e^{\lambda X_1}) = \int_{\mathbb{R}} e^{-\lambda x} p_X(x) dx < +\infty.$$

Allora, per ogni $\varepsilon > E(X_1)$ si ha

$$\lim_{n\to\infty} \frac{1}{n} \ln P(S_n > n\varepsilon) = -I(\varepsilon)$$

ove I è la trasformata di Legendre di Z

$$I(\varepsilon) = sup_{\lambda\in\mathbb{R}}\ \{\varepsilon\lambda - \ln Z(\lambda)\}.$$

La funzione $\Lambda(\lambda) = \ln Z(\lambda)$ è detta funzione generatrice dei momenti. Infatti si ha

$$\Lambda'(0) = E(X), \quad \Lambda''(0) = var(X) > 0.$$

Dalla convessità di Λ si ha che I è convessa e ha minimo in $m = E(X)$ ove vale zero. Inoltre I ha insiemi di sottolivello compatti (vedi oltre, per la definizione formale di good rate function). Vediamo l'applicazione del teorema di Cramer alla variabile del lancio della moneta. La funzione di partizione è

$$Z(\lambda) = \frac{1}{2}e^{\lambda 1} + \frac{1}{2}e^{\lambda 0} = \frac{1}{2}(e^{\lambda} + 1).$$

La sua trasformata di Legendre si calcola determinando l'unico punto critico di $\ln Z(\lambda)$ che è $\hat{\lambda} = \ln(\varepsilon/1-\varepsilon)$ da cui

$$I(\varepsilon) = \ln 2 + \varepsilon \ln \varepsilon + (1-\varepsilon)\ln(1-\varepsilon) = D(p(\varepsilon)\|\frac{1}{2})$$

coincidente con quella trovata in precedenza. Una dimostrazione del teorema di Cramer nel caso di una variabile a valori discreti si ottiene rapidamente dal Teorema di Sanov, che vedremo in seguito. Il Teorema di Cramer ammette una generalizzazione al caso di variabili non i.i.d., il Teorema di Gartner–Ellis (vedi oltre). Prima di procedere, dobbiamo formalizzare la nozione di Principio di grandi deviazioni e di rate function

Definizione 6.3.1 (Rate function) *Sia (M, d) spazio metrico. Una funzione $I : M \to [0, +\infty)$ è detta*

a) rate function se ha insiemi di sottolivello I_a chiusi per ogni $a \in [0, +\infty)$ ove

$$I_a = \{x \in M : I(x) \leq a\}$$

b) good rate function se ha insiemi di sottolivello I_a compatti per ogni $a \in [0, +\infty)$.

Definizione 6.3.2 (Principio di grandi deviazioni) *Una successione di variabili aleatorie $X_i : \Omega \to (M, d)$ soddisfa a un principio di grandi deviazioni con velocità a_n (ove $a_n \to +\infty$) e rate function I se per ogni insieme boreliano $A \subset M$*

$$lim\ sup \frac{1}{a_n} \ln P(X_n \in A) \leq -inf_{x \in cl(A)}\ I(x),$$

$$lim\ inf \frac{1}{a_n} \ln P(X_n \in A) \leq -inf_{x \in int(A)}\ I(x),$$

Uno strumento utile per semplificare il calcolo di rate function è il seguente, detto Principio di contrazione

Proposizione 6.3.1 (Principio d contrazione) *Sia $X_i : \Omega \to (M, d)$ successione di v.a. che soddisfa a un principio di grandi deviazioni con velocità a_n e good rate function I su M; sia $f : (M, d) \to (N, \delta)$ funzione continua tra spazi metrici. Allora la successione di v.a. $Y_i = f(X_i)$ soddisfa ad un principio di grandi deviazioni con velocità a_n e good rate function J su N ove*

$$J(y) = inf\{I(x) : f(x) = y\}.$$

Dimostrazione. Sia $A \subset N$ un chiuso. Allora

$$\begin{aligned} & lim\ sup\ \frac{1}{a_n} \ln P(f(X_n) \in A) = lim\ sup\ \frac{1}{a_n} \ln P(X_n \in f^{-1}(A)) \\ \leq \quad & -inf_{\ x \in f^{-1}(A)}\ I(x) = -inf_{y \in A}\ inf_{\ x \in f^{-1}(y)}\ I(x) = -inf_{\ y \in A} J(y) \end{aligned}$$

ove abbiamo usato il fatto che $f^{-1}(A)$ è chiuso. Analogo ragionamento se A è aperto. Rimane da verificare che J è good rate function. Vale

$$\begin{aligned} J_a &= \{y \in N : J(y) \leq a\} = \{y \in N : inf_{\ x \in f^{-1}(y)} I(x) \leq a\} \\ &= f(\{x \in M : I(x) \leq a\}) = f(I_a) \end{aligned}$$

quindi J_a è compatto poichè I_a lo è e f è continua.

□

Un altro risultato utile che costituisce una generalizzazione del Principio di Laplace è il Lemma di Varadhan

Proposizione 6.3.2 (Lemma di Varadhan) *Sia X_i successione di v.a. $X_i \sim \mu_i$ soddisfacente ad un principio di grandi deviazioni su (M, d) con velocità n e good rate function I. Sia $F : M \to \mathbb{R}$ funzione continua e inferiormente limitata. Allora*

$$\lim \frac{1}{n} \ln \int_M e^{nF(x)} \mu_n(dx) = sup_{x \in M}(F(x) - I(x)).$$

6.3.1 Il Teorema di Sanov

Il Teorema di Sanov generalizza il risultato trovato in precedenza sulle grandi deviazioni nel caso del lancio della moneta al caso di variabili i.i.d. a valori in uno spazio finito $\Gamma = \{a_1, \ldots, a_r\}$ detto alfabeto. Sia

$$\mathcal{M}(\Gamma) = \{\nu = (\nu_1, \ldots, \nu_r) \ : \ \nu_k \in [0,1],\ \sum \nu_k = 1\}$$

l'insieme delle d.d.p. su Γ. $\mathcal{M}(\Gamma)$ diventa spazio metrico separabile e completo con la distanza in variazione totale

$$d(\mu, \nu) = \frac{1}{2} \sum_{i=1}^{r} |\mu_r - \nu_r|.$$

Sia ora X_i successione di v.a. i.i.d. definite in Ω e a valori in Γ. Per ogni n indichiamo con L_n^X la v.a. detta *misura empirica* associata a $X_1, \ldots X_n$

$$L_n^X = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

ove δ_{X_i} è la misura di Dirac su Γ concentrata in X_i. Ogni L_n^X è una v.a. definita in Ω e a valori in $\mathcal{M}(\Gamma)$: dato $\omega \in \Omega$ si ha $X_i(\omega) = x_i$ e per ogni $k \in \Gamma$

$$(L_n^{X(\omega)})(k) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}(k) = \frac{1}{n} \#\{x_i = k\} = \frac{n(k)}{n}$$

ove $\sum_{k=1}^{r} n(k) = n$.

Teorema 6.3.2 (Sanov) *Sia X_i successione i.i.d. a valori in Γ, e sia $\rho \in \mathcal{M}(\Gamma)$ la distribuzione delle X_i. La successione L_n^X di v.a. a valori in $\mathcal{M}(\Gamma)$ soddisfa un principio di grandi deviazioni con velocità n e good rate function su $\mathcal{M}(\Gamma)$*

$$I(\nu) = D(\nu\|\rho) = \sum_{k=1}^{r} \nu_k \ln \frac{\nu_k}{\rho_k}.$$

Quindi, per ogn $A \subset \mathcal{M}(\Gamma)$

$$lim\ sup \frac{1}{n} \ln P(L_n^X \in A) \leq -inf_{\nu \in cl(A)}\ I(\nu),$$

$$lim\ inf \frac{1}{n} \ln P(L_n^X \in A) \leq -inf_{\nu \in int(A)}\ I(\nu).$$

Osservazione 1. L'entropia relativa è una good rate function. Vale $D(\nu\|\rho) \geq 0$ e si annulla solo per $\nu = \rho$. Infatti, posto $\phi(x) = x \ln x$, ϕ è strettamente convessa e dalla disuguaglianza di Jensen sulle funzioni convesse $\phi(\sum \lambda_i x_i) \leq \sum \lambda_i \phi(x_i)$ si ha

$$D(\nu\|\rho) = \sum \rho_k \phi(\frac{\nu_k}{\rho_k}) \geq \phi(\sum \rho_k \frac{\nu_k}{\rho_k})) = \phi(1) = 0.$$

Inoltre $D(\nu\|\rho)$ è convessa in ν. Il teorema di Prohorov permette di mostrare che i suoi insiemi di sottolivello sono compatti.

Osservazione 2. Prima di dimostrare il teorema mostriamo che il T. di Sanov implica il T. di Cramer nel caso di v.a. a valori in uno spazio *finito* (alfabeto) attraverso il principio di contrazione. Sia $\Gamma = \{a_1, \ldots, a_r\}$ e definiamo la funzione continua $f : \mathcal{M}(\Gamma) \to (\mathbb{R}, |\cdot|)$ tra spazi metrici

$$f(\nu) = E_\nu(a) = \sum_{k=1}^{r} a_k \nu_k.$$

Il trasformato tramite f della v.a. L_n^X è la v.a. reale

$$f \circ L_n^X = E_{L_n^X}(a) = \frac{1}{n} \sum_{k=1}^{r} a_k \delta_{X_k} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n$$

Poichè f è continua, possiamo applicare il Principio di Contrazione per ottenere la rate function delle v.a. $f \circ L_n^X$. Si ha

$$\begin{aligned} J(y) &= inf_{\nu \in f^{-1}(y)}\ I(\nu) = inf\{I(\nu) : \nu \in \mathcal{M}(\Gamma), \quad E_\nu(a) = y\} \\ &= inf\{D(\nu\|\rho) : \nu \in \mathcal{M}(\Gamma), \quad E_\nu(a) = y\} = sup_{\lambda \in \mathbb{R}}\{\lambda y - \ln Z(\lambda)\} \end{aligned}$$

ove $Z(\lambda)$ vale

$$Z(\lambda) = E(e^{\lambda X}) = \sum_{k=1}^{r} \rho_k e^{\lambda a_k}.$$

L'ultima eguaglianza $inf = sup$ si mostra per calcolo diretto. Infatti $D(\nu\|\rho)$ ha minimo assoluto vincolato su $E_\nu(a) = y$ per

$$\hat{\nu}_k = \nu(\hat{\lambda}) = \rho_k \frac{e^{\hat{\lambda} a_k}}{Z(\hat{\lambda})}$$

ove $\hat{\lambda} = \hat{\lambda}(y)$ è univocamente determinato dal vincolo $\sum a_k \hat{\nu}_k = y$ e tale minimo vale

$$D(\hat{\nu}\|\rho) = inf\{D(\nu\|\rho) : \nu \in \mathcal{M}(\Gamma), \quad E_\nu(a) = y\} = -\ln Z(\hat{\lambda}) + \hat{\lambda} y.$$

Basta mostrare che il sup di $g(\lambda) = \lambda y - \ln Z(\lambda)$ si ha per $\lambda = \hat{\lambda}(y)$. Infatti

$$g'(\lambda) = y - \frac{d}{d\lambda} \ln Z(\lambda) = y - E_{\nu(\lambda)}(a) = 0 \quad \text{per} \quad \lambda = \hat{\lambda}(y)$$

Inoltre

$$g''(\lambda) = -\frac{d^2}{d\lambda^2} \ln Z(\lambda) = -var_{\nu(\lambda)}(a) < 0$$

quindi $\hat{\lambda}(y)$ è l'unico massimo di $g(\lambda)$.

□

Dimostrazione del T. di Sanov La dimostrazione segue lo schema combinatoriale già visto nel caso dei lanci della moneta.

□

6.4 Grandi deviazioni per misure di coppia

Consideriamo ancora il caso di v.a. a valori in un alfabeto finito Γ e ci interessiamo agli eventi costituiti da coppie ordinate di elementi di Γ che indichiamo con

$$st = (a_s, a_t) \in \Gamma^2 = \Gamma \times \Gamma.$$

Come prima indichiamo con

$$\mathcal{M}^2(\Gamma) = \{\nu = (\nu_{11}, \ldots, \nu_{rr}) : \nu_{st} \in [0,1] \quad \sum \nu_{st} = 1\}$$

l'insieme delle d.d.p. su Γ^2 che è uno spazio metrico separabile e completo con la distanza in variazione totale introdotta sopra. Una $\nu \in \mathcal{M}(\Gamma)$ induce una d.d.p su Γ^2 $\nu \times \nu$ data da $\nu_{st} = \nu_s \nu_t$ (misura prodotto) mentre una d.d.p. su Γ^2 ne induce due (le sue marginali)

$$\nu_{\cdot s} = \sum_{t=1}^{r} \nu_{ts}, \qquad \nu_{s\cdot} = \sum_{t=1}^{r} \nu_{st}$$

Sia X_i una successione di v.a. a valori in Γ. Ci interessiamo a realizzazioni finite $X_1, \ldots, X_n$ del processo con condizioni periodiche $X_{n+1} = X_1$. In analogia con le misure empiriche introdotte in precedenza, definiamo le misure empiriche di coppia

$$L_n^2 = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i X_{i+1}}$$

Esse sono v.a. a valori nel sottoinsieme $\check{\mathcal{M}}^2(\Gamma)$ delle misure sul prodotto Γ^2 che hanno marginali uguali $\nu_{\cdot s} = \nu_{s\cdot} = \nu_\cdot$. Dato $\omega \in \Omega$ si ha $X_i(\omega) = x_i$ e per ogni $st \in \Gamma^2$ si ha

$$(L_n^2)(st) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i x_{i+1}}(st) = \frac{1}{n} \#\{(x_i, x_{i+1}) = (a_s, a_t)\} = \frac{n(st)}{n}$$

ove $\sum_{s,t=1}^{r} n(st) = n$.

Il teorema ergodico di Birkhoff o anche la legge forte dei grandi numeri permettono di mostrare che, se le X_i sono i.i.d. con $X_i \sim \rho$, allora la distanza in variazione totale tra le L_n^2 e la misura prodotto $\rho \times \rho$ tende a zero P-quasi certamente

$$\lim_{n \to +\infty} d(L_n^2, \rho \times \rho) = 0.$$

Vale infatti per misure di coppia su $\check{\mathcal{M}}^2(\Gamma)$ l'analogo del teorema di Sanov per misure empiriche. Lo enunciamo in una forma particolare, usando particolari sottoinsiemi di $\check{\mathcal{M}}^2(\Gamma)$: per ogni $a > 0$ indichiamo con $B(\rho \times \rho, a)$ la palla chiusa di raggio a in $(\check{\mathcal{M}}^2(\Gamma), d)$ e con $B^c(\rho \times \rho, a)$ il suo complementare.

Teorema 6.4.1 (Sanov per misure di coppia) *Sia X_i successione i.i.d. con $X_i \sim \rho$. Allora la successione L_n^2 di v.a. a valori in $\check{\mathcal{M}}^2(\Gamma)$ soddisfa un principio di grandi deviazioni con velocità n e good rate function su $\check{\mathcal{M}}^2(\Gamma)$*

$$I(\nu) = D(\nu \| \nu_{\cdot} \times \rho) = \sum_{s,t=1}^{r} \nu_{st} \ln(\frac{\nu_{st}}{\nu_{\cdot s} \rho_t}). \tag{6.3}$$

In particolare

$$\lim_{n \to +\infty} \frac{1}{n} \ln P(L_n^2 \in B^c(\rho \times \rho, a)) = -inf_{\nu \in B^c(\rho \times \rho, a)} \; I(\nu).$$

6.4.1 Grandi deviazioni per catene di Markov

Il teorema di Sanov per le misure di coppia empiriche si estende alle variabili non i.i.d. descritte da catene di Markov tempo-omogenee (a stati finiti) a valori in Γ. Sia allora P la matrice stocastica ad elementi positivi

$$P_{st} = Prob(X_{i+1} = a_t \mid X_i = a_s)$$

e π l'unica distribuzione stazionaria per P, ovvero $P^T \pi = \pi$. Consideriamo realizzazioni della catena con $X_1 \sim \pi$ e condizioni periodiche $X_{n+1} = X_1$. Il punto chiave per usare i risultati precedenti è di esprimere la probabilità di una traiettoria $\omega_n = (i_1, \ldots, i_n)$ con $i_k \in \Gamma$ tramite le misure L_n^2. Si ha infatti

$$\begin{aligned} P^X(\omega_n) &= \pi_{i_1} P_{i_1 i_2} \ldots P_{i_{n-1} i_n} = \frac{\pi_{i_1}}{P_{i_n i_1}} e^{\sum_{k=1}^{n} \ln P_{i_k i_{k+1}}} \\ &= \frac{\pi_{i_1}}{P_{i_n i_1}} e^{n \sum_{s,t \in \Gamma^2} L_n^2(st) \ln P_{st}} \end{aligned}$$

ove $L_n^2(st)$ è la misura di coppia su $\check{\mathcal{M}}^2(\Gamma)$ associata alla realizzazione ω_n. Se ora consideriamo la stessa traiettoria come realizzata in un processo i.i.d. con $X_i \sim \pi$, la relazione precedente è ancora valida con

$$P_{st} = Prob(X_{i+1} = a_t \mid X_i = a_s) = \pi_t$$

e quindi

$$Q^X(\omega_n) = \prod_{k=1}^{n} \pi_{i_k} = e^{n \sum_{s,t \in \Gamma^2} L_n^2(st) \ln \pi_t}.$$

Il loro rapporto ha la forma di una derivata di Radon-Nikodym

$$\frac{dP^X}{dQ^X}(\omega_n) = \mathcal{O}(1)e^{n\varphi(L_n^2(\omega_n))}$$

ove $\mathcal{O}(1)$ esprime il prefattore limitato $\frac{\pi_{i_1}}{P_{i_n i_1}}$ e φ indica la funzione su $\check{\mathcal{M}}^2(\Gamma)$

$$\varphi(\nu) = \sum_{s,t=1}^{r} \nu_{st} \ln(\frac{P_{st}}{\pi_t}).$$

Teorema 6.4.2 (T. di Sanov per misure di coppia su catene di Markov) *Sia X_i successione di v.a. descritta dalla catena di Markov a stati finiti (P, π). Allora la successione L_n^2 di v.a. a valori in $\check{\mathcal{M}}^2(\Gamma)$ soddisfa un principio di grandi deviazioni con velocità n e good rate function su $\check{\mathcal{M}}^2(\Gamma)$*

$$I_P(\nu) = D(\nu \| \nu_\cdot \times P) = \sum_{s,t=1}^{r} \nu_{st} \ln(\frac{\nu_{st}}{\nu_{\cdot s} P_{st}}).$$

ove $(\nu_\cdot \times P)_{st} = \nu_{\cdot s} P_{st}$.

Dimostrazione Usiamo la relazione tra probabilità di una traiettoria ω_n nella catena e nel processo i.i.d. data dalla derivata di radon-Nikodym sopra scritta. Allora, per ogni $A \subset \check{\mathcal{M}}^2(\Gamma)$ si ha che

$$\begin{aligned} \frac{1}{n} \ln P^X(L_n^2 \in A) &= \frac{1}{n} \ln \int_A P^X(L_n^2 \in d\nu) \\ &= \frac{1}{n}\mathcal{O}(1) + \frac{1}{n} \ln \int_A e^{n\varphi(\nu)} Q^X(L_n^2 \in d\nu) \end{aligned}$$

Ora, il processo X_i i.i.d. con $X_i \sim \pi$ soddisfa un principio di grandi deviazioni con velocità n e good rate function vedi (6.3)

$$I_\pi(\nu) = D(\nu \| \nu_\cdot \times \rho) = \sum_{s,t=1}^{r} \nu_{st} \ln(\frac{\nu_{st}}{\nu_{\cdot s} \pi_t}).$$

Inoltre, il secondo integrale scritto sopra è nella forma adatta per poter applicare il Lemma di Varadhan poichè φ è inferiormente limitata e continua. Quindi

$$\begin{aligned} \frac{1}{n} \ln P^X(L_n^2 \in A) &= sup_{\nu \in A} \{\varphi(\nu) - I_\pi(\nu)\} \\ &= sup_{\nu \in A} \sum_{s,t=1}^{r} \nu_{st} \ln(\frac{P_{st}\nu_{\cdot s}}{\nu_{st}}) \\ &= -inf_{\nu \in A} \sum_{s,t=1}^{r} \nu_{st} \ln(\frac{\nu_{st}}{P_{st}\nu_{\cdot s}}) = -inf_{\nu \in A}\ I_P(\nu) \end{aligned}$$

□

Possiamo ora studiare le grandi deviazioni per misure empiriche su catene di Markov. In generale, data una funzione reale f definita sull'alfabeto Γ, la media empirica di f su una realizzazione $\omega_n = (i_1, \ldots, i_n)$ del processo è data da

$$\bar{f}_n = \frac{1}{n}\sum_{k=1}^{n} f(i_k) = \sum_{k=1}^{r} f(i_k) L_n^X(k)$$

quindi $\bar{f}_n$ è successione di variabili aleatorie reali e $\bar{f}_n = F(L_n^X)$ con F continua. Per il Principio di contrazione, se L_n soddisfa ad un principio di grandi deviazioni con velocità n e rate function $I(\nu)$, anche la successione $\bar{f}_n$ lo soddisfa con rate function su $\mathbb{R}$

$$J(y) = inf\{I(\nu) : F(\nu) = \sum_{k=1}^{r} \nu_k f(a_k) = E_\nu(f(\Gamma)) = y\}.$$

L'argomento appena esposto permette di ricavare un principio di grandi deviazioni per misure empiriche su catene di Markov. Consideriamo infatti la funzione continua che trasforma una d.d.p. su Γ^2 nella sua marginale

$$c : \check{\mathcal{M}}^2(\Gamma) \to \mathcal{M}(\Gamma), \quad c(\nu) = \nu_\cdot$$

Allora, il teorema di Sanov dimostrato in precedenza per le misure di coppia su catene di Markov assieme al principio di contrazione permette di stabilire il seguente risultato

Teorema 6.4.3 (T. di Sanov per misure empiriche su catene di Markov) *Sia X_i successione di v.a. descritta dalla catena di Markov a stati finiti (P, π). Allora la successione L_n^X di v.a. a valori in $\mathcal{M}(\Gamma)$ soddisfa un principio di grandi deviazioni con velocità n e good rate function su $\mathcal{M}(\Gamma)$*

$$J_P(\mu) = inf\{I_P(\nu) : c(\nu) = \nu_\cdot = \mu\}.$$

Osservazione. La good rate function ora determinata non è esplicita come nel caso delle successioni i.i.d. e come nel caso delle misure di coppia. Una riscrittura più esplicita della $J_P(\mu)$ si ottiene dall'applicazione del Teorema di Gartner–Ellis, che vale per successioni di variabili dipendenti generali.

6.5 Teorema di Gartner-Ellis

Esistono versioni del Teorema di Gartner-Ellis leggermente più generali di quella qui esposta, ma di più elaborata esposizione.

Teorema 6.5.1 (Gartner-Ellis) *Sia X_i successione di v.a. di densità $X_i \sim \mu_i$ a valori in $\mathbb{R}^d$. Se la funzione generatrice logaritmica dei momenti $Z(\lambda)$ su $\mathbb{R}^d$*

$$\begin{aligned} Z(\lambda) &= \lim_{n\to+\infty} \frac{1}{n} \ln E(e^{n\langle\lambda, X_n\rangle}) \\ &= \lim_{n\to+\infty} \frac{1}{n} \ln \int_{\mathbb{R}^d} e^{n\langle\lambda, y\rangle} \mu_n(dy) \end{aligned}$$

esiste finita ed è differenziabile, allora la successione X_i soddisfa ad un principio di grandi deviazioni con velocità n e good rate function data dalla Trasformata di Legendre di $Z(\lambda)$

$$I(x) = sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, x\rangle - Z(\lambda)\}$$

Associato al Teorema di Gartner-Ellis vi è il seguente utile risultato

Teorema 6.5.2 *Supponiamo che la funzione generatrice logaritmica dei momenti esista finita per ogni $\lambda \in \mathbb{R}^d$. Allora sono equivalenti*

1. *X_n converge esponenzialmente ad un $\hat{z} \in \mathbb{R}^d$, ovvero per ogni $\varepsilon > 0$ esistono $C > 0$ e $N \in \mathbb{N}$ tali che*

$$P_n(|X_n - \hat{z}| < \varepsilon) \leq e^{-nC} \qquad \forall\ n > N;$$

2. *$Z(\lambda)$ è differenziabile in 0 e*

$$grad_\lambda Z(0) = \hat{z}$$

3. *la good rate function $I(x)$ ha un unico minimo in $z = \hat{z}$.*

Osservazione 1. nel caso in cui le v.a. siano definite tutte sullo stesso spazio di probabilità, la convergenza esponenziale in 1. implica la covergenza quasi certa se $\sum_n e^{-nC}$ è finita.

Osservazione 2. Ovviamente è possibile trattare il caso di successioni i.i.d. di variabli aleatorie con il Teorema di Gartenr-Ellis appena esposto. Vediamo quindi come i risultati precedenti possano essere considerati applicazioni particolari del Teorema di Gartner- Ellis.

Caso di variabili i.i.d. $X_i \sim \rho$ a valori in un alfabeto finito Γ.

Consideriamo le medie empiriche $L_n^X = n^{-1}\sum_{i=1}^n \delta_{X_i}$ a valori in $\mathcal{M}(\Gamma) \subset \mathbb{R}^r$. Dato $\lambda \in \mathbb{R}^r$ vale

$$e^{n\langle\lambda, L_n^X\rangle} = e^{\sum_{k=1}^r \lambda_k n_k}$$

ove $\sum_{k=1}^r n_k = n$. Ora, per l'ipotesi $X_i \sim \rho$, i.i.d. si ha, posto $\nu = (n^{-1}n_1, \ldots, n^{-1}n_r)$

$$P(n_1, \ldots, n_r) = P(L_n^X = \nu) = \prod_{k=1}^r \rho_k^{n_k}$$

e quindi

$$E(e^{n\langle\lambda, L_n^X\rangle}) = \sum_{n_1,\ldots,n_r} e^{\sum_{k=1}^r \lambda_k n_k} \prod_{k=1}^r (\rho_k)^{n_k} = \sum_{n_1,\ldots,n_r} \prod_{k=1}^r (\rho_k e^{\lambda_k})^{n_k} = \Big[\sum_{k=1}^r \rho_k e^{\lambda_k}\Big]^n$$

La funzione generatrice logaritmica dei momenti si scrive

$$Z(\lambda) = \lim_{n\to+\infty} \frac{1}{n} \ln E(e^{n\langle\lambda, L_n^X\rangle}) = \ln\Big[\sum_{k=1}^r \rho_k e^{\lambda_k}\Big].$$

La sua trasformata di Legendre coincide con l'entropia relativa $D(\nu\|\rho)$ come visto in precedenza nell'Osservazione 2.

Caso di una catena di Markov a stati finiti

Consideriamo le misure di coppia L_n^2, intese come v.a. a valori in $\check{\mathcal{M}}^2(\Gamma) \subset Mat(r) \subset \mathbb{R}^{r^2}$. Dato un $\lambda \in Mat(r)$ ragionando come prima si ha che

$$E(e^{n\langle \lambda, L_n^2 \rangle}) = \sum_{k,l=1}^{r} \pi_k Q_{kl}^n(\lambda) e^{\lambda_{lk}}$$

ove abbiamo introdotto la matrice

$$Q_{kl}(\lambda) = P_{kl} e^{\lambda_{kl}}.$$

Applicando il Teorema di Perron-Frobenius sulle matrici non negative, si calcola

$$Z(\lambda) = \lim_{n \to +\infty} \frac{1}{n} \ln E(e^{n\langle \lambda, L_n^2 \rangle}) = \ln \rho(\lambda)$$

ove $\rho(\lambda)$ è l'autovalore reale positivo di modulo massimo di $Q(\lambda)$ detto anche raggio spettrale della matrice. La rate function è la trasformata di Legendre e vale

$$I(\nu) = sup_{\lambda \in Mat(r)} \{\nu \cdot \lambda - \ln \rho(\lambda)\}, \qquad \nu \in \check{\mathcal{M}}^2(\Gamma)$$

ove $\nu \cdot \lambda = tr(\lambda^T \nu)$ indica il prodotto scalare tra matrici. Si dimostra (vedi F. den Hollander, *Large Deviations*) che la rate function trovata coincide con quella determinata in precedenza.

Per completezza riportiamo qui l'enunciato del Teorema di Perron-Frobenius sulle matrici non negative

Teorema 6.5.3 (Perron-Frobenius) *Sia A matrice ad elementi non negativi e tale che per ogni i, j esiste una potenza A^m di A che ha $(A^m)_{ij} > 0$. Allora A ha un autovalore reale ρ, detto raggio spettrale di A tale che*

1. *ogni altro autovalore di A ha modulo non superiore a ρ*
2. *gli autovettori destro e sinistro di A sono positivi e hanno autospazi di dimensione uno*
3. *per ogni vettore u positivo e per ogni $i = 1, \dots dim\ A$ si ha*

$$\lim_{n \to +\infty} \frac{1}{n} \ln(A^n u)_i = \lim_{n \to +\infty} \frac{1}{n} \ln((A^T)^n u)_i = \ln \rho.$$

7 Applicazione alla termodinamica di non equilibrio

Uno dei problemi centrali della termodinamica di non equilibrio è la descrizione dell'irreversibilità dei fenomeni. Qui ci restringiamo alla descrizione di sistemi in uno stato di non equilibrio stazionario, ovvero sistemi nei quali sono presenti correnti (di carica, di massa, di calore), quindi non isolati, ma la cui descrizione è indipendente dal tempo. Si tratta di uno scenario intermedio tra i sistemi all'equilibrio e quelli fuori dall'equilibrio nel quale possiamo sperare di utilizzare il formalismo della termodinamica di equilibrio sostituendo agli stati le traiettorie del sistema.

Consideriamo sistemi descritti da catene di Markov (π, P) a stati finiti in Γ ove P è matrice stocastica non negativa e π è l"unica distribuzione stazionaria per P, $P^T\pi = \pi$. Inoltre supporremo sempre che la distribuzione iniziale della catena sia π (catena stazionaria). Data una realizzazione della catena o *traiettoria*[1]

$$\omega_n = (i_0, \dots, i_n) = (X(t_0) = i_0, X(t_1) = i_1, \dots, X(t_n) = i_n)$$

indichiamo con

$$\omega_n^- = (i_n, \dots, i_0) = (X(t_0) = i_n, X(t_1) = i_{n-1}, \dots, X(t_n) = i_0)$$

la traiettoria ricavata da ω_n per inversione temporale. Vale allora

$$P(\omega_n) = \pi_{i_0} P_{i_0 i_1} \dots P_{i_{n-1} i_n}, \qquad P(\omega_n^-) = \pi_{i_n} P_{i_n i_{n-1}} \dots P_{i_1 i_0}$$

Una catena di Markov stazionaria si dice *reversibile* se per ogni n e per ogni ω_n si ha $P(\omega_n) = P(\omega_n^-)$. Una condizione necessaria e sufficente è che valga per $n = 1$, detta condizione di *bilancio dettagliato*

$$P(X(t_0) = i, X(t_1) = j) = \pi_i P_{ij} = \pi_j P_{ji} = P(X(t_0) = j, X(t_1) = i).$$

Se vale questa condizione, la matrice delle probabilità di transizione a tempo invertito coincide con P

$$P^-_{ji} = P(X(t_0) = i | X(t_1) = j) = P_{ji}.$$

mentre in generale essa dipende dal tempo ed è data da

$$P^-_{ji}(t) = \frac{\pi_i(t) P_{ij}}{\pi_j(t+1)}.$$

[1] L'insieme delle traiettorie coincide con l'insieme delle possibili applicazioni tra $[0, n] \subset \mathbb{N}$ e Γ e quindi è un insieme finito di cardinalità $|\Gamma|^n = r^n$.

7.0.1 Calcolo di valori medi di osservabili associati alle traiettorie

Sia ε_i l'energia del sistema nello stato microscopico $i \in \Gamma$ e sia J_{ij} la (variazione di) corrente nel passaggio sel sistema dallo stato i allo stato j. La corrente è descritta da una matrice antisimmetrica $J = -J^T$ (per esempio pensiamo a $J_{ij} = \varepsilon_j - \varepsilon_i$). L'energia associata ad una traiettoria o energia integrata è

$$\varepsilon(\omega_n) = \sum_{k=0}^{n} \varepsilon_{i_k}$$

e il suo valore medio è una quantità estensiva con n

$$E(\varepsilon) = \sum_{\omega_n} \varepsilon(\omega_n) P(\omega_n) = n \sum_{k=1}^{r} \pi_k \varepsilon_k.$$

Analogamente, la corrente associata ad una traiettoria è

$$J(\omega_n) = \sum_{k=0}^{n} J_{i_k i_{k+1}}$$

e il suo valore medio è

$$E(J) = \sum_{\omega_n} J(\omega_n) P(\omega_n) = n \sum_{k,l=1}^{r} \pi_k P_{kl} J_{kl} = nJ \cdot \mathcal{P}.$$

ove abbiamo introdotto la probabilità congiunta $\mathcal{P}_{ij} = \pi_i P_{ij}$ e $J \cdot \mathcal{P} = \sum_{k,l=1}^{r} \mathcal{P}_{kl} J_{kl}$ è il prodotto scalare tra matrici.

Definiamo poi l'*entropia del sistema nell'intervallo temporale* $[0, n]$ come il valor medio sulle traiettorie dell'osservabile $-\ln P(\omega_n)$ per cui

$$H(n) = -\sum_{\omega_n} P(\omega_n) \ln P(\omega_n)$$

e definiamo il *tasso di produzione di entropia* del sistema nell'intervallo temporale $[0, n]$ come il valor medio sulle traiettorie dell'osservabile $W(\omega_n) = \ln P(\omega_n) - \ln P(\omega_n^-)$ per cui

$$\sigma(n) = -\sum_{\omega_n} P(\omega_n) W(\omega_n) = -\sum_{\omega_n} P(\omega_n) \ln \frac{P(\omega_n)}{P(\omega_n^-)}.$$

Definiamo allora entropia del sistema e produzione di entropia del sistema le quantità

$$H = \lim \frac{1}{n} H(n), \qquad \sigma = \lim \frac{1}{n} \sigma(n).$$

Proposizione 7.0.1 *L'entropia di una catena di Markov stazionaria* (P, π) *vale*

$$H = H(P, \pi) = -\sum_{k,l=1}^{r} \pi_k P_{kl} \ln P_{kl}$$

Dimostrazione Partendo dall'espressione sopra scritta per $H(n)$

$$\begin{aligned} H(n) &= -\sum_{\omega_n} P(\omega_n)\ln P(\omega_n) = -\sum_{i_0,\ldots,i_n} \pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-1}i_n}\ln \pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-1}i_n} \\ &= -\sum_{i_0,\ldots,i_n} \pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-1}i_n}\ln[\pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-2}i_{n-1}} + \ln P_{i_{n-1}i_n}] \\ &= -\sum_{i_0,\ldots,i_{n-1}} \pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-2}i_{n-1}}\ln[\pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-2}i_{n-1}}] + \\ &\quad -\sum_{i_0,\ldots,i_n} \pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-1}i_n}P_{i_{n-1}i_n}\ln P_{i_{n-1}i_n} \\ &= H(n-1) - \sum_{i_{n-1}i_n} \pi_{i_{n-1}}P_{i_{n-1}i_n}\ln P_{i_{n-1}i_n} \end{aligned}$$

ove nell'ultima riga abbiamo usato il fatto che $P^t\pi = \pi$. Dalla relazione ricorsiva trovata si trova la quantità estensiva con n

$$H(n) = H(0) - n\sum_{k,l=1}^{r} \pi_k P_{kl}\ln P_{kl} = -\sum_k \pi_k \ln \pi_k - n\sum_{k,l=1}^{r} \pi_k P_{kl}\ln P_{kl}$$

da cui dividendo per n si ha la tesi.

□

Osservazione L'espressione trovata per l'entropia è valida anche se la distribuzione iniziale della catena non è la distribuzione stazionaria π. Per dimostrarlo si usa il teorema ergodico per le catene di Markov. (vedi dispense M.F.).

Proposizione 7.0.2 *Il tasso di produzione di entropia di una catena di Markov stazionaria (P,π) vale*

$$\sigma = \sigma(\pi,P) = \frac{1}{2}\sum_{k,l=1}^{r} \pi_k P_{kl} - \pi_l P_{lk}\ln\frac{\pi_k P_{kl}}{\pi_l P_{lk}} = \sum_{k,l=1}^{r} \pi_k P_{kl}\ln\frac{\pi_k P_{kl}}{\pi_l P_{lk}}.$$

Dimostrazione. Partendo dall'espressione per $\sigma(n)$ e moltiplicando e dividendo nell'argomento del logaritmo per $\pi_{i_{n-1}}\ldots\pi_i$ si ha

$$\begin{aligned} \sigma(n) &= \sum_{i_0,\ldots,i_n} \pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-1}i_n}\ln[\frac{\pi_{i_0}P_{i_0i_1}\ldots P_{i_{n-1}i_n}}{\pi_{i_n}P_{i_ni_{n-1}}\ldots P_{i_1i_0}}] \\ &= n\sum_{kl=1}^{r} \pi_k P_{kl}\ln\frac{\pi_k P_{kl}}{\pi_l P_{lk}} \end{aligned}$$

e la tesi segue subito dividendo per n.

□

Osservazione Siano P e Q matrici stocastiche e π distribuzione di probabilità. L'entropia relativa di Q rispetto a P è

$$D_\pi(Q\|P) = \sum_{k,l} \pi_k Q_{kl}\ln\frac{Q_{kl}}{P_{kl}}$$

Allora la produzione di entropia si scrive come

$$\sigma = D_\pi(P\|P^T) = \sum_{k,l} \pi_k P_{kl} \ln \frac{P_{kl}}{P_{lk}}$$

7.0.2 Fluttuazioni dei valori medi

Sia (π, P) catena di Markov stazionaria. Consideriamo un generico osservabile definito sulle traiettorie; data $\varphi : \Gamma \to \mathbb{R}$ consideriamo la v.a. che sulla traiettoria $\omega_n = (i_0, \dots, i_n)$ assume il valore

$$\varphi_n(\omega_n) = \frac{1}{n} \sum_{k=0}^{n} \varphi(i_k).$$

Per il teorema ergodico di Birkhoff, vale

$$\lim_{n\to+\infty} \varphi_n = E_\pi(\varphi).$$

Vogliamo applicare il Teorema di Gartner-Ellis alla successione di v.a. scalari non i.i.d. φ_n descritte dalla densità $P(\omega_n)$. Dobbiamo calcolare, per $\lambda \in \mathbb{R}$

$$\begin{aligned} E(e^{n\langle\lambda,\varphi_n\rangle}) &= E(e^{\lambda\sum_{k=0}^{n}\varphi(i_k)}) = \sum P(\omega_n) e^{\lambda\varphi_n(\omega_n)} \\ &= \sum_{i_0,\dots,i_n} \pi_{i_0} P_{i_0 i_1} \dots P_{i_{n-1} i_n} [\prod_{k=0}^{n} e^{\lambda\varphi(i_k)}] \\ &= \sum_{l=1}^{r} (P(\lambda)^{nT}\pi)_l \end{aligned}$$

ove abbiamo posto

$$P(\lambda)_{kl} = P_{kl} e^{\lambda\varphi(i_k)}.$$

La matrice $P(\lambda)$ è ad elementi positivi e quindi ad essa si applica il Teorema di Perron-Frobenius. Pertanto,

$$Z(\lambda) = \lim_{n\to+\infty} \frac{1}{n} \ln E(e^{n\langle\lambda,\varphi_n\rangle}) = \lim_{n\to+\infty} \frac{1}{n} \ln \sum_{l=1}^{r} (P(\lambda)^{nT}\pi)_l = \ln \rho(\lambda).$$

Applicazione

1. Consideriamo l'osservabile definito sulle traiettorie dato dal tasso di produzione di entropia

$$W_n(\omega_n) = \frac{1}{n} \ln \frac{P(\omega_n)}{P(\omega_n^-)}.$$

Vogliamo applicare il Teorema di Gartner-Ellis alla successione di v.a. scalari non i.i.d. W_n descritte dalla densità $P(\omega_n)$. Dobbiamo calcolare per $\lambda \in \mathbb{R}$

$$\begin{aligned} E(e^{n\langle\lambda,W_n\rangle}) &= E(\Big[\frac{P(\omega_n)}{P(\omega_n^-)}\Big]^\lambda) \\ &= \sum_{i_0,\dots,i_n} \pi_{i_0} P_{i_0 i_1} \dots P_{i_{n-1} i_n} [\frac{\pi_{i_0} P_{i_0 i_1} \dots P_{i_{n-1} i_n}}{\pi_{i_n} P_{i_n i_{n-1}} \dots P_{i_1 i_0}}]^\lambda \\ &= \sum_{i_0,\dots,i_n} \prod_{k=0}^{n} \pi_{i_k} P_{i_k i_{k+1}} [\frac{\pi_{i_k} P_{i_k i_{k+1}}}{\pi_{i_{k+1}} P_{i_{k+1} i_k}}]^\lambda. \end{aligned}$$

Definiamo

$$A(\lambda)_{kl} = P_{kl}[\frac{\pi_k P_{kl}}{\pi_l P_{lk}}]^\lambda$$

che è matrice ad elementi positivi se P lo è. Per il teorema di Perron-Frobenius $A(\lambda)$ ha raggio spettrale $\rho = \rho(\lambda)$. Per 3. del teorema di Gartner–Ellis, non è difficile mostrare (vedi Qians) che la funzione generatrice logaritmica dei momenti, che qui chiamiamo energia libera, vale

$$Z(\lambda) = \lim_{n\to+\infty} \frac{1}{n} \ln E(e^{n\langle\lambda, W_n\rangle}) = \ln \rho(\lambda)$$

Quindi la successione delle v.a. W_n soddisfa un principio di grandi deviazioni si $\mathbb{R}$ con good rate function su $\mathbb{R}$

$$I(x) = sup_{\lambda\in\mathbb{R}}\{\lambda x - \ln \rho(\lambda)\}.$$

2. Ripetiamo il ragionamento precedente per la v.a. $A_n(\omega) = -\frac{1}{n}\ln P(\omega)$ detta *auto-processo*. Si ha allora

$$\begin{aligned} E(e^{n\langle\lambda, A_n\rangle}) &= E(e^{-\lambda \ln P(\omega)}) = E(P(\omega)^{-\lambda}) \\ &= \sum P(\omega)P(\omega)^{-\lambda} = \sum P(\omega)^{1-\lambda} \\ &= \sum_{i_0,\dots,i_n} \pi_{i_0}^{1-\lambda} P_{i_0 i_1}^{1-\lambda} \dots P_{i_{n-1} i_n}^{1-\lambda} = \sum_{l=1}^{r} (P(\lambda)^{nT}\pi(\lambda))_l \end{aligned}$$

ove abbiamo posto, come prima

$$P(\lambda)_{kl} = P_{kl}^{1-\lambda}, \qquad \pi(\lambda)_k = \pi_k^{1-\lambda}.$$

Pertanto

$$Z(\lambda) = \lim_{n\to+\infty} \frac{1}{n} \ln E(e^{n\langle\lambda, A_n\rangle}) = \ln \rho(\lambda)$$

Teorema di Fluttuazione per W_n

Enunciamo una proprietà di simmetria della rate function trovata, che è una versione del teorema di Fluttuazione di Gallavotti-Cohen.

Teorema 7.0.4 *L'energia libera e la rate function della famiglia di v.a. W_n hanno le seguenti proprietà*

$$Z(\lambda) = Z(-\lambda - 1), \qquad I(x) = I(-x) - x.$$

L'interpretazione di questo risultato è la seguente (vedi Qians)

$$\frac{Prob\ (W_n = z)}{Prob\ (W_n = -z)} = e^{nz}.$$

Vale inoltre (vedi Qians)

Teorema 7.0.5 *Per P-quasi ogni traiettoria ω_n della catena stazionaria (π, P)*

$$\lim_{n\to+\infty} W_n(\omega_n) = \sigma(\pi, P) = D_\pi(P\|P^T) = \sum_{k,l} \pi_k P_{kl} \ln \frac{P_{kl}}{P_{lk}}.$$

e

$$\lim_{n\to+\infty} A_n(\omega_n) = H(\pi, P) = \sum_{k,l} \pi_k P_{kl} \ln P_{kl}.$$

7.1 Gas su reticolo

Con gas su reticolo intendiamo sistemi stocastici costituiti da un numero finito di particelle indistinguibili che si muovono sui siti di un reticolo d-dimensionale. Lo stato microscopico o configurazione del sistema è specificato dando il numero di particelle che occupano un dato sito (n oppure $0, 1$ se vale un principio di esclusione). L'interazione tra le particelle viene descritta legando le probabilità di transizione tra siti vicini o anche lontani alla configurazione. Sistemi non isolati sono descritti tramite l'operazione di creazione o distruzione di particelle al bordo del reticolo e l'effetto di campi esterni è descritto da asimmetrie nelle probabilità di transizione in una direzione fissata del reticolo. In alternativa a serbatoi di particelle, si possono mettere condizioni periodiche al bordo. In questo paragrafo, seguendo (GJL), considereremo processi di Markov a tempo continuo e stati discreti. La differenza con quanto fatto in precedenza è quindi che ora $t \in \mathbb{R}$.

Sia Λ un dominio regolare connesso di $\mathbb{R}^d$ e $\Lambda_n = n\Lambda/\mathbb{Z}^d$ il reticolo a $|\Lambda_n|$ siti di passo 1 su Λ. Se $X = \{0, 1\}$ oppure $X = \mathbb{N}$ indica il numero di particelle che possono occupare un sito, lo spazio delle configurazioni del reticolo è X^{Λ_n}. Denotiamo con $\eta \in X^{\Lambda_n}$ una configurazione e con η_x il numero di particelle nel sito $x \in \Lambda_n$ nella configurazione η. Allora il numero totale di particelle nel reticolo e la densità di particelle in una configurazione $\eta \in X^{\Lambda_n}$ sono espresse da

$$\nu_n(\eta) = \sum_{x \in \Lambda_n} \eta_x, \qquad \rho_n = \frac{\nu_n(\eta)}{|\Lambda_n|} = \frac{1}{|\Lambda_n|} \sum_{x \in \Lambda_n} \eta_x.$$

Dati due siti $x, y \in \Lambda_n$ e una configurazione η, indichiamo con $\sigma^{x,y}\eta$ la configurazione che si ottiene da η spostando una particella da x a y, quindi

$$\sigma^{x,y}\eta_y = \eta_y + 1, \qquad \sigma^{x,y}\eta_x = \eta_x - 1.$$

Se il sito di partenza o di arrivo non appartiene al reticolo, nella nuova configurazione $\sigma^{x,y}\eta$ vi è creazione o rispettivamente distruzione di una particella. Indichiamo con

$$c_{x,y}(\eta) \in [0, 1]$$

la frazione rispetto al totale di particelle che passano dal sito x al sito y in un intervallo di tempo dt (jump rate o probabilità di transizione). La dinamica è reversibile se $c_{x,y}(\eta) = c_{y,x}(\eta)$. Una condizione sulla dinamica fisicamente significativa è quella di soddisfare un bilancio dettagliato rispetto ad una misura di Gibbs definita da un Hamiltoniana $\mathcal{H}$ sulle configurazioni (qui l'Hamiltoniana ingloba la temperatura inversa) e a una probabilità di transizione "iniziale" c^0 che è supposta reversibile

$$c^0_{x,y}(\eta) = e^{\mathcal{H}(\eta) - \mathcal{H}(\sigma^{x,y}\eta)} c^0_{x,y}(\sigma^{x,y}\eta), \qquad x, y \in \Lambda_n.$$

La condizione di bilancio dettagliato per salti tra siti interni e siti appartenenti al bordo $\partial\Lambda_n$ del reticolo si descrive introducendo il *potenziale chimico* $\lambda_0 : \partial\Lambda_n \to \mathbb{R}$ e quindi

$$c^0_{x,y}(\eta) = e^{(\mathcal{H}(\eta) - \mathcal{H}(\sigma^{x,y}\eta) - \lambda_0(y))} c^0_{x,y}(\sigma^{x,y}\eta), \qquad x \in \Lambda_n, y \notin \Lambda_n.$$

Nel seguito supporremo per semplicità che il potenziale chimico sia costante.

Indichiamo con $\mu_n(\eta)$ la probabilità che il sistema si trovi nella configurazione η. L'evoluzione della probabilità è descritta dal generatore infinitesimo L_n del processo di Markov, l'analogo infinito dimensionale della matrice stocastica P. Senza entrare in dettaglio, possiamo per analogia accettare che, sotto certe condizioni, il processo ammetta un'unica misura μ_n *invariante* per L_n. Per esempio, per i sistemi sopra accennati, l'unica misura invariante è la distribuzione gran-canonica

$$\mu_{n,\lambda_0}(\eta) = \frac{1}{Z_n(\lambda_0)} e^{-\mathcal{H}(\eta)+\lambda_0\nu_n(\eta)}$$

ove si è posto $\Sigma_{n,k} = \{\eta \in X^{\Lambda_n} : \nu_n(\eta) = k\}$ e

$$Z_n(\lambda) = \sum_{k\geq 0} e^{\lambda k} \sum_{\eta\in\Sigma_{n,k}} e^{-\mathcal{H}(\eta)}.$$

7.1.1 Grandi deviazioni per la densità

Ripetendo quanto fatto nelle applicazioni precedenti del Teorema di Gartner-Ellis, consideriamo la successione di v.a. aleatorie reali ρ_n definite su X^{Λ_n} aventi densità μ_{n,λ_0} e introduciamo la funzione generatrice logaritmica dei momenti

$$p(\lambda) = \lim_{n\to+\infty} \frac{1}{|\Lambda_n|} \ln E(e^{|\Lambda_n|\langle\lambda,\rho_n\rangle}) = \lim_{n\to+\infty} \frac{1}{|\Lambda_n|} \ln E(e^{\lambda\nu_n}).$$

SI dimostra (vedi GJL) che la successione soddisfa un principio di grandi deviazioni su $\mathbb{R}$ con velocità $|\Lambda_n|$ e good rate function data dalla trasformata di Legendre di $p(\lambda)$ (energia libera)

$$f(\rho) = sup_{\lambda\in\mathbb{R}} \{\lambda\rho - p(\lambda)\}. \tag{7.1}$$

Pertanto, scrivendo in maniera informale si ha

$$P_n(\rho_n \approx \rho) \sim e^{-|\Lambda_n| f(\rho)}$$

ove $\approx$ indica appartenenza ad un intorno di ρ in $\mathbb{R}$.

Grandi deviazioni per il profilo della densità

Il risultato trovato è classico. L'innovazione di GJL è di stabilire un risultato tipo teorema di grandi deviazioni non per la densità globale ma per il profilo di densità macroscopica $\rho(u)$ ove $u = x/n$ è un punto del dominio $\Lambda \subset \mathbb{R}^d$ opportunamente riscalato con n. Cominciamo introducendo la densità empirica: ad ogni configurazione microscopica $\eta \in X^{\Lambda_n}$ associamo un profilo di densità *macroscopico*

$$\pi_n(u) = \pi_n(\eta; u), \qquad u \in \Lambda$$

che è implicitamente definito come segue: per ogni funzione liscia $G : \Lambda \to \mathbb{R}$

$$\langle\pi_n, G\rangle = \int_\Lambda \pi_n(u)G(u)du = \frac{1}{n^d} \sum_{x\in\Lambda_n} G(x/n)\eta_x.$$

Quindi π_n è successione di v.a. a valori nell'insieme delle funzioni non negative e sommabili su Λ e descritte dalla densità μ_n introdotta sopra. Nella teoria di grandi

deviazioni macroscopica elaborata da GJL si mostra che la successione π_n soddisfa un principio di grandi deviazioni espresso informalmente come

$$P_n(\pi_n \approx \rho) \sim e^{-n^d \mathcal{F}(\rho)}.$$

Qui $\rho \approx \rho'$ indica che le medie delle due densità sono vicine su ogni intorno macroscopico piccolo e $\mathcal{F}(\rho)$ è il funzionale convesso e locale

$$\mathcal{F}(\rho) = \int_\Lambda f(\rho(u))du \tag{7.2}$$

ove f è l'energia libera introdotta in (7.1). La formula (7.2) permette di introdurre la naturale estensione della nozione di energia libera per un sistema termodinamico arbitrariamente fuori dall'equilibrio. In casi generali la misura invariante μ_n non è nota esplicitamente.

7.1.2 Teoria idrodinamica

L'approccio di GJL è quindi composto da vari passi: 1. descrizione della dinamica microscopica su reticolo, 2. scrittura del generatore del processo di Markov, 3. determinazione della distribuzione stazionaria (in generale non esplicita a parte alcuni modelli), 4. prova di un risultato di grandi deviazioni per il profilo di densità empirico, 5. scrittura dell'energia libera macroscopica per il sistema.

Il risultato successivo ottenuto dalla teoria il seguente: fissato un profilo di densità $\hat{\rho}(t,u)$, $(t,u) \in [0,T] \times \Lambda$, vogliamo calcolare la probabilità che la densità empirica π_n sia in un intorno di $\hat{\rho}$. GJL mostrano che tale probabilità è esponenzialmente piccola se $\hat{\rho}$ *non è* soluzione della PDE (legge di bilancio)

$$\partial_t \rho = div[\frac{1}{2} D(\rho) grad\ \rho - \chi(\rho)E] \tag{7.3}$$

ove D è la matrice di diffusione e E è un campo esterno di drift. Il risultato di grandi deviazioni ha la forma

$$Prob(\pi_n \approx \hat{\rho}) \sim e^{-n^d \mathcal{I}(\hat{\rho})} \tag{7.4}$$

ove la good rate function si ottiene mediante integrazione spazio-temporale su $[0,T] \times \Lambda$ di $\nabla H = gradH$

$$\mathcal{I}(\hat{\rho}) = \mathcal{I}_{[0,T]}(\hat{\rho}) = \frac{1}{2} \int_0^T dt \int_\Lambda du \langle \nabla H, \chi(\hat{\rho}) \nabla H \rangle$$

e ∇H è l'*extra campo esterno gradiente* necessario per avere il profilo dato $\hat{\rho}$ come soluzione della PDE di bilancio sopra scritta:

$$\partial_t \hat{\rho} = div[\frac{1}{2} D(\hat{\rho}) grad\ \hat{\rho} - \chi(\hat{\rho})(E + \nabla H)].$$

$\mathcal{I}_{[0,T]}(\hat{\rho})$ ha in significato fisico di lavoro fatto dal campo ∇H per portare la densità del sistema pari a $\hat{\rho}$.

Equazione di Hamilton-Jacobi

Consideriamo la seguente situazione fisica: il sistema a $t = -\infty$ è descritto dal profilo di densità $\bar{\rho}(u)$ che è soluzione stazionaria di (7.3). Come visto sopra, le soluzioni di $\bar{\rho}(u)$ sono le densità che hanno probabilità massima di essere osservate nel senso della teoria delle grandi deviazioni, ma altri profili di densità sono osservabili. Supponiamo quindi che il sistema descritto da $\bar{\rho}$ a $t = -\infty$ si trovi a $t = 0$ descritto dal profilo di densità ρ. Vogliamo determinare la traiettoria più probabile seguita dal sistema nella creazione di questa fluttuazione. Sulla base di (7.4) è ragionevole pensare che essa rende minimo $\mathcal{I}_{[-\infty,0]}$ tra tutte le traiettorie producibili con un extra campo ∇H opportuno e che rispettano le 'condizioni al bordo'

$$\hat{\rho} \in CB \quad \leftrightarrow \qquad \hat{\rho} = \bar{\rho} \quad \text{a} \quad t = -\infty, \qquad \hat{\rho} = \rho \quad \text{a} \quad t = 0.$$

Rimane quindi definito un funzionale $V(\rho)$ dello stato finale a $t = 0$ che pesa il lavoro minimo necessario per creare la fluttuazione di densità

$$V(\rho) = inf_{\hat{\rho} \in CB}\, \mathcal{I}_{[-\infty,0]}(\hat{\rho})$$

GJL mostrano che V è soluzione (non unica) di un'equazione di Hamilton-Jacobi infinito-dimensionale

$$\frac{1}{2}\langle \nabla \frac{\delta V}{\delta \rho}, \chi(\rho) \nabla \frac{\delta V}{\delta \rho} \rangle + \langle \nabla \frac{\delta V}{\delta \rho}, \nabla \cdot [\frac{1}{2} D(\rho) \nabla \rho - \chi(\rho) E] \rangle = 0.$$

ove le parentesi angolari indicano integrazione spaziale sul dominio Λ.

8 Appendici

8.1 Struttura geometrica soggiacente al PME

I principi variazionali, che sono alla base di molte formulazioni di teorie fisiche, affermano che l'evoluzione del sistema (nella dinamica) o la configurazione di equilibrio del sistema (nella statica) hanno la proprietà di rendere estremale una certa funzione definita sull'insieme delle possibili evoluzioni / configurazioni del sistema. In questo Capitolo vediamo come i principi variazionali abbiano una struttura geometrica comune ad essi soggiacente e che quindi è possibile ritrovare ogni volta che si adotti una formulazione della teoria a partire da un principio variazionale. Il principio della Massima Entropia ricade quindi all'interno delle teorie della statica a formulazione variazionale. La nozione chiave per descrivere la struttura geometrica è quella di funzione generatrice o famiglia generatrice (o di Morse) di sottovarietà Lagrangiane. Tale nozione, che appare già nei lavori di Lie, Jacobi e nel trattato di Meccanica di Whittaker, si colloca quindi all'interno della geometria simplettica, della quale diamo un brevissimo cenno.

8.1.1 Funzioni generatrici di sottovarietà Lagrangiane.

Sia M una varietà liscia di dimensione pari e ω una due–forma chiusa e non degenere. La coppia (M, ω) definisce una varietà *simplettica*. Una sottovarietà $S \subset M$ tale che $2 \dim S = \dim M$ e che $\omega_{|S} = 0$ è detta sottovarietà *Lagrangiana* di M.

Per esempio, data una varietà Q, il suo fibrato cotangente T^*Q equipaggiato con la due–forma canonica $\omega = dp \wedge dq$, dove (q, p) è una carta locale di T^*Q, è una varietà simplettica. Data una funzione $G : Q \to \mathbb{R}$, il suo grafico

$$\Lambda_G = graph\ (dG) = \{(q, dG(q)) : q \in Q\} \subset T^*Q$$

definisce una sottovarietà Lagrangiana di T^*Q. Infatti: $dim\Lambda_G = dimQ$ e $\omega_{|S} = dG \wedge dq$ per cui $d\omega_{|S} = 0$. Nel caso del grafico di un differenziale la sottovarietà lagrangiana è *trasversale alle fibre* di $\pi : T^*Q \to Q$ poichè $\pi_{|\Lambda_G}$ è un diffeomorfismo.

Esempio. Per illustrare le nozioni sopra introdotte consideriamo il caso semplice, ma molto istruttivo $Q = \mathbb{R}$. Allora il suo fibrato cotangente T^*Q coincide con il piano $\mathbb{R}^2$ riferito alle coordinate (q, p). Per comprendere la struttura simplettica associata dobbiamo pensare a $\mathbb{R}^2 \equiv T^*Q$ come ad una varietà; il suo spazio tangente nel punto $(q, p) \in \mathbb{R}^2$ (le coordinate nel piano) è dato dall'insieme dei vettori del piano di componenti $(\dot{q}, \dot{p})$ che pensiamo applicati in $(q, p) \in \mathbb{R}^2$ e quindi coincide con il piano $\mathbb{R}^2$ stesso ad ogni punto $(q, p) \in \mathbb{R}^2$. L'azione della due–forma simplettica $\omega(q, p) = dp \wedge dq$ sulle coppie $a = (\dot{q}, \dot{p})$, $b = (\dot{q}', \dot{p}')$ di vettori tangenti in $(q, p) \in \mathbb{R}^2$

è

$$\omega(q,p)(a,b) = (dp \wedge dq)(a,b) = \langle \dot{p}', \dot{q} \rangle - \langle \dot{p}, \dot{q}' \rangle = \dot{p}', \dot{q} - \langle \dot{p}, \dot{q}'$$

ove con $\langle\ ,\ \rangle$ abbiamo denotato l'azione delle uno–forme sui vettori. Se immergiamo $\mathbb{R}^2$ in $\mathbb{R}^3$ e introduciamo i vettori $\bar{a} = (\dot{q}, \dot{p}, 0)$, $\bar{b} = (\dot{q}', \dot{p}', 0)$ e $e_3 = (0,0,1)$ allora si ha che

$$\omega(q,p)(a,b) = \bar{a} \times \bar{b} \cdot e_3$$

e quindi il valore della due forma simplettica sulla coppia di vettori tangenti a e b coincide con il valore dell'*area* (con segno, perchè $\omega(q,p)(a,b) = -\omega(q,p)(b,a)$) del parallelogramma formato dai vettori associati $\bar{a}$, $\bar{b}$.

Sottovarietà Lagrangiane. E' facile vedere che ogni curva $\mathcal{L}$ del piano (sottovarietà di $\mathbb{R}^2$ uno–dimensionale) definisce una sottovarietà Lagrangiana poichè il suo spazio tangente $T\mathcal{L}$ in un suo punto è la retta tangente alla curva nel punto stesso e quindi ogni coppia di vettori (necessariamente paralleli!) di esso definisce un parallelogramma di area nulla. Quindi $\omega_{|\mathcal{L}} = 0$ e $\mathcal{L}$ è Lagrangiana. I punti dove la tangente alla curva è verticale sono i punti ove $\mathcal{L}$ *non* è trasversale alle fibre di $\pi : \mathbb{R}^2 \to \mathbb{R}$, $\pi(q,p) = q$ e il loro insieme costituisce la caustica associata a $\mathcal{L}$.

8.1.2 Famiglie generatrici di sottovarietà Lagrangiane

Consideriamo ora una *famiglia* liscia G di funzioni su Q dipendenti da k parametri reali (da qui appunto il nome di 'famiglia')

$$G : Q \times U \longrightarrow \mathbb{R}, \quad U \subset \mathbb{R}^k$$

alla quale resta associata la fibrazione triviale $\pi : Q \times U \to Q$. Indichiamo con

$$\mathcal{E} = \{(q,u) \in Q \times U : G_u(q,u) = 0\} \tag{8.1}$$

l'insieme critico di G, ovvero l'insieme dei punti critici della derivata di G rispetto alla direzione di fibra (qui G_u denota la derivata parziale di G rispetto a u).

La famiglia di funzioni G si dice *famiglia generatrice* (o di Morse) se 0 è valore *regolare* di G_u, ovvero se dG_u è suriettiva (ha rango massimo su $\mathcal{E}$):

$$rk\ \Big[dG_u\Big]_{|\mathcal{E}} = rk\ \Big[G_{uq}\ G_{uu}\Big]_{|\mathcal{E}} = max = dim\ U = k. \tag{8.2}$$

In tal caso $\mathcal{E}$ è una sottovarietà regolare di $Q \times U$ di dimensione pari a $dim\ Q$.

L'utilizzo delle famiglie generatrici in Fisica e in Meccanica Analitica è motivato dal seguente importante risultato (vedi [Weinstein, [**?**] Lect. 6])

Proposizione 8.1.1 *Sia G famiglia generatrice. Allora*

$$\Lambda = \{(q, G_q(q,u)) \in T^*Q \quad : \quad (q,u) \in \mathcal{E}\}$$

è una sottovarietà Lagrangiana di T^*Q,generata da G.

Dimostrazione. Introduciamo la mappa

$$i : \mathcal{E} \subset Q \times U \to T^*Q, \quad i(q,u) = (q, G_q(q,u))$$

per cui risulta $\Lambda = i(\mathcal{E})$. Dico che poichè dG_u è suriettiva su $\mathcal{E}$, i è iniettiva. Allora Λ è una sottovarietà immersa di T^*Q di dimensione pari a $dim\mathcal{E} = dimQ$. Inoltre mostreremo che $i^*\omega = \omega_\Lambda = 0$, quindi la sottovarietà Λ è Lagrangiana. Mostriamo che $Ti : \ker dG_u \to T(T^*Q)$ è iniettiva. In componenti, per ogni $(q,u) \in \mathcal{E}$ si ha

$$Ti(q,u)\begin{pmatrix} \dot{q} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} \dot{q} \\ G_{qq}\dot{q} + G_{qu}\dot{u} \end{pmatrix} = \begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix}$$

e Ti è iniettiva se e solo se $\ker Ti = \{0\}$ ovvero se, posto $e = (q,u)$

$$Ti(e)\dot{e} = 0 \quad \text{e} \quad dG_u(e)\dot{e} = 0 \;\Rightarrow\; \dot{e} = 0.$$

In componenti si ha

$$\begin{pmatrix} \dot{q} \\ G_{qq}\dot{q} + G_{qu}\dot{u} \\ G_{uq}\dot{q} + G_{uu}\dot{u} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \;\Rightarrow\; \begin{pmatrix} \dot{q} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Si vede immediatamente che il sistema precedente è equivalente a

$$\begin{pmatrix} G_{qu} \\ G_{uu} \end{pmatrix}\begin{pmatrix} \dot{u} \\ \dot{u} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \;\Rightarrow\; \dot{u} = 0.$$

Ma la mappa lineare appana introdotta è iniettiva poichè la sua matrice trasposta coincide con la matrice associata a dG_u, che è suriettiva. Quindi Λ è sottovarietà di T^*Q. Per la commutatività delle operazioni di differenziazione esterna e di pull-back possiamo scrivere

$$i_*\omega = i_*d\theta = d(i_*\theta), \quad \text{ove} \quad \theta = pdq.$$

Calcoliamo $\theta_\Lambda = i_*\theta$. Per definizione,

$$\langle \theta_\Lambda, \dot{e}\rangle = \langle i_*\theta, \dot{e}\rangle = \langle \theta(i(e)), Ti(e)\dot{e}\rangle = \langle G_q(e)dq, Ti(e)\dot{e}\rangle = G_q(e)dq$$

Pertanto, la uno–forma canonica ristretta alla sottovarietà Λ è

$$\theta_\Lambda = G_q(e)dq = G_q(e)dq + G_u(e)du = dG(q,u) \tag{8.3}$$

dal momento che, se $e \in \mathcal{E}$, $G_u(e) = 0$. Quindi θ_Λ coincide con un differenziale e quindi è chiusa

$$\omega_\Lambda = d(\theta\Lambda) = d(dG(q,u)) = \sum_{j=1}^{n} \frac{\partial^2 G(q,u)}{\partial q_i \partial q_j} dq_j \wedge dq_i = 0. \tag{8.4}$$

□

Osservazione 1. Se la condizione sul rango (8.2) che definisce la famiglia generatrice è soddisfatta dalla sottomatrice quadrata G_{uu}, ovvero

$$rk\Big[dG_{uu}\Big]_{|\mathcal{E}} = max = k, \tag{8.5}$$

allora il teorema della funzione implicita permette di affermare che, localmente, $\mathcal{E}$ è il grafico di una funzione $g : W \subset Q \to U$, $u = g(q)$. Inoltre, ponendo $\tilde{G} = G \circ g$ si ha che

$$d\tilde{G}(q) = G_q(q, g(q)) + G_u(q, g(q))g_q(q) = G_q(q, g(q)), \quad \text{per } q \in W. \tag{8.6}$$

Questi risultati sono riassunti nella seguente

Proposizione 8.1.2 *Se vale la condizione* (8.5) *i parametri u possono essere localmente eliminati quindi, sopra* W, Λ *coincide il grafico di* $d\tilde{G}$ *e pertanto è sottovarietà Lagrangiana localmente trasversale alle fibre di* π.

Se G è famiglia generatrice, l'insieme $\Gamma \subset Q$ dei punti $q \in Q$ tali che nei punti corrispondenti sulla sottovarietà $(q,p) \in \Lambda$ la sottovarietà stessa non è trasversale alle fibre di π (ovvero cade la condizione (8.5)) è detto la *caustica* di Λ ed è definito dalla condizione

$$\Gamma = \{q \in Q : \ \det G_{uu}(q,u) = 0, \ \text{e} \ (q,u) \in \mathcal{E}\}. \tag{8.7}$$

Osservazione 2. Nel caso di una sottovarietà Lagrangiana Λ localmente trasversle alle fibre di T^*Q, la condizione di chiusura diventa

$$\frac{\partial^2 \tilde{G}(u)}{\partial q_i \partial q_j} dq_j \wedge dq_i = 0 \qquad i.e. \qquad \tilde{G}_{ij} = \tilde{G}_{ji}.$$

8.1.3 La funzione di Lagrange come famiglia generatrice

L'apparato esposto nelle sezioni precedenti ci permette di mostrare che la funzione di Lagrange associata ad un problema di estremo vincolato è una particolare famiglia generatrice. Per vederlo è sufficiente fare le identificazioni seguenti:

$$q = (c,v), \ u = (p,\lambda), \ Q = \mathcal{F} \times V, \ U = \mathcal{P} \times Y^*$$

e considerare la fibrazione $\pi(c,v,p,\lambda) = (c,v)$. Allora, in relazione alla famiglia generatrice

$$G(c,v,p,\lambda) = U(p) - \langle \lambda, F(p,v) - c \rangle \tag{8.8}$$

i) l'insieme critico $\mathcal{E}$ in (8.1) costituisce l'insieme delle soluzioni dei problemi di estremo vincolato per valori fissati (c,v)

$$\mathcal{E} = \{(c,v,p,\lambda) \in \mathcal{F} \times V \times \mathcal{P} \times Y^* : \quad G_p = 0, \ G_\lambda = 0\}. \tag{8.9}$$

ii) la condizione che G sia una famiglia generatrice si legge

$$q = (c,v), \qquad u = (p,\lambda)$$

$$rk \begin{bmatrix} G_{qu} & G_{uu} \end{bmatrix}_{|\mathcal{E}} = rk \begin{pmatrix} 0 & -\langle \lambda, F_{pv} \rangle & G_{pp} & -F_p \\ -\mathbb{I}_k & F_v & -F_p^T & 0 \end{pmatrix}_{|\mathcal{E}} = k + n$$

ove $\mathbb{I}_k$ è la matrice identità in in $\mathbb{R}^k$ and $G_{pp} \in Sym(n)$ è la matrice Hessiana di G rispetto alle variabili p

$$G_{pp} = U_{pp}(p) - \langle \lambda, F_{pp}(p,v) \rangle. \tag{8.10}$$

iii) la sottovarietà Lagrangiana $\Lambda_G \subset \mathcal{F} \times V \times Y^* \times V^*$ generata dalla famiglia G è, vedi Prop. 8.1.1,

$$\Lambda = \{(c,v,G_c,G_v) = (c,v,\lambda, -\langle \lambda, F_v(p,v) \rangle : \quad (c,v,p,\lambda) \in \mathcal{E}\} \tag{8.11}$$

e viene ad essere quindi l'oggetto geometrico naturale legato al considerare una *famiglia* di problemi di estremo nei quali variano i parametri e i valori dei vincoli;

iv) la condizione locale per l'eliminazione di tutti i parametri è

$$rk\ G_{uu}|_{\mathcal{E}} = rk \begin{pmatrix} G_{pp} & -F_p \\ -F_p^T & 0 \end{pmatrix}_{|\mathcal{E}} = max = k + n. \tag{8.12}$$

v) La uno-forma canonica ristretta alla sottovarietà λ in (8.3) è

$$\theta_\Lambda = G_q(q,u)dq = G_c dc + G_v dv = \lambda dc - \langle \lambda, F_v(p,v)\rangle dv.$$

Nel caso in cui la sottovarietà sia trasversale, possiamo esprimere $u = \tilde{u}(q)$ e inoltre, per la particolare forma della famiglia generatrice (8.8), la famiglia generatrice liberata dai parametri *coincide con la funzione obbiettivo U valutata nel punto di estremo*, ovvero :

$$\tilde{G}(q) = G(q, \tilde{u}(q)) = U(\tilde{p}(c,v)) =: S(c,v) \tag{8.13}$$

e la uno–forma canonica si scrive

$$\theta_\Lambda = \tilde{G}_q(q)dq = \tilde{\lambda}(c,v)dc - \langle \tilde{\lambda}(c,v), F_v(\tilde{p}(c,v),v)\rangle dv. \tag{8.14}$$

La proposizione seguente lega la trasversalità della sottovarietà Lagrangiana alle condizioni sufficienti di massimo locale e quindi dà una condizione (solo sufficiente) di trasversalità locale. In pratica, se il punto di massimo locale è deducibile dalle condizioni sufficienti di secondo ordine , la sottovarietà Lagrangiana è trasversale. Questa condizione sarà sempre verificata se i vincoli sono *al più lineari* nelle p e $U(p)$ è funzione concava nelle p come nel caso del Principio della Massima Entropia, vedi oltre.

Proposizione 8.1.3 *Se la matrice simmetrica $G_{pp} \in Sym(n)$ in* (8.10) *è* definita (negativa o positiva) *su* $\ker F_p \subset \mathbb{R}^n$ *per ogni* $(c,v,p,\lambda) \in \mathcal{E}$*, allora la matrice quadrata G_{uu} in* (8.12) *ha rango massimo e quindi tutti parametri $u = (p,\lambda)$ possono essere eliminati.*

Proof. We have to show that $G_{uu}z = 0$ implies that $z = 0$, where $z \in \mathbb{R}^{k+n}$. Now, setting $z^T = (w,v) \in \mathbb{R}^{k+n}$, we have that

$$G_{uu}z = 0 \quad \Longleftrightarrow \quad \begin{cases} d\zeta\ v = 0, \\ d\zeta^T\ w + G_{mm}v = 0. \end{cases}$$

Now, if $v = 0$, the equation $d\zeta^T w = 0$ implies that $w = 0$ since $d\zeta^T$ is injective ($d\zeta$ is surjective by hypothesis) and we are done. As a reductio ad absurdum hypothesis, suppose that $v \neq 0$ and $d\zeta v = 0$, i.e. $v \in \ker d\zeta$. The equation $d\zeta v = 0$ implies $v^T d\zeta^T = 0^T$; by multiplying the second equation by $v^T \neq 0$ we get

$$v^T d\zeta^T w + v^T G_{mm} v = v^T G_{mm} v = 0 \quad \text{with } v \neq 0$$

which is an absurdum since G_{mm} is a (positive on negative) definite matrix on $\ker d\zeta(m)$. □

Note bibliografiche

Per la parte generale di meccanica dei continui si è seguito essenzialmente il testo M.E. Gurtin *An Introduction to Continuum Mechanics* Academic Press 1981; per alcuni argomenti (teorema delle forze vive, tensore di Piola–Kirkhhoff, disuguaglianza di Clausius–Duhem) si è seguito in parte il testo T. Manacorda *Introduzione alla termomeccanica dei continui* Quaderni dell'Unione Matematica Italiana Ed. Pitagora, Bologna 1979. Infine, una esposizione rigorosa e raffinata delle basi della Meccanica dei Continui è contenuta negli *Appunti di Istituzioni di Fisica matematica 1* di A. Marzocchi, Università cattolica del S. Cuore, Brescia, disponibili sulla homepage di A. Marzocchi.
Per il cenno del secondo principio della termodinamica si è seguito l'articolo E.T. Jaynes *The evolution of Carnots's Principle* (vedi http://bayes.wustl.edu/), mentre Per la parte sul PME si sono utilizzati gli scritti di E.T. Jaynes, per esempio l'articolo (vedi http://bayes.wustl.edu/) E.T.Jaynes *Information Theory and Statistical Mechanics I*, Physical Review 106 (1957), p.620-630, contenuto nel bel volume E.T. Jaynes *Papers on Probability, Statistics and Statistical Physics*, Kluwer, e inoltre il testo classico di E. Schrödinger *Termodinamica Statistica*.
Le considerazioni sull'uso dell'entropia nello studio delle P.D.E. sono tratte da L.C. Evans *Entropy and Partial Differential Equations* disponibile sulla home page di L.C. Evans.

Nota. Queste dispense sono ancora in una forma provvisoria, che deve essere sperimentata a lezione. Probabilmente, non tutto quanto in esse è contenuto verrà presentato, mentre sicuramente alcune parti potranno essere ulteriormente approfondite, a seconda dell'interesse degli studenti. Ringrazio fin d'ora quanti vorranno segnalarmi errori e incongruenze in quanto scritto.

M.F.