Teoria dell'informazione e Meccanica Statistica

L. P.

Giugno 2007

Riporto qui una breve rassegna dell'approccio alla Meccanica Statistica mediante la teoria dell'informazione. Partiamo dalla considerazione che la probabilità p_i associata allo stato i per un sistema termodinamico *non può* essere interpretata in senso frequentistico. In effetti, anche per un sistema piccolo, il numero di stati possibili è talmente grande che, per poter valutare la frequenza relativa di occupazione dello stato i, sarebbe necessario aspettare tempi incredibilmente lunghi.

Consideriamo un sistema di Ising con N spin. Il numero di stati possibili è pari a $2^N \simeq 1000^{N/10}$. Con un milione di spin (un cubetto di lato 100), abbiamo 10^{15} stati. Supponendo di esplorare uno stato ogni picosecondo (a una frequenza di 1 GHz) dobbiamo aspettare 10^6 secondi, cioè circa 4 mesi, per passare in rassegna ciascuno stato in media una volta. Questo tempo cresce esponenzialmente con la taglia del sistema.

Viceversa le p_i possono essere interpretate come una rappresentazione della *nostra conoscenza* del sistema. Esse ci permettono di valutare i valori attesi delle osservabili X_i del sistema, tramite la relazione

$$\langle X_i \rangle = \sum_i p_i X_i. \tag{1}$$

In questo modo, invece di cercare di dimostrare un risultato analogo al teorema ergodico (secondo il quale le p_i "canoniche" corrispondono alla frequenza con cui un determinato stato viene visitato nel corso dell'evoluzione del sistema) dobbiamo giustificare la *scelta* di queste probabilità come la migliore rappresentazione possibile della nostra conoscenza del sistema.

Seguendo l'approccio di Jaynes [1, 2], questa scelta può essere giustificata interpretando in maniera diversa la proprietà di estremo del funzionale di Gibbs. Data una distribuzione di probabilità $p = (p_1, \ldots, p_n)$, il funzionale di Gibbs è definito da

$$\mathcal{S}(p) = -k \sum_i p_i \log p_i, \tag{2}$$

dove la costante k viene di solito presa uguale alla costante di Boltzmann k_B. Con questa scelta, si può mostrare che la distribuzione canonica $p_i^\mathrm{eq} = \exp(-E_i/k_\mathrm{B}T)/Z$ soddisfa il seguente principio di massimo:

> *Il funzionale di Gibbs $\mathcal{S}(p) = -k_\mathrm{B} \sum_i p_i \log p_i$ assume il valore massimo per $p = p^\mathrm{eq}$, fra tutte le distribuzioni per cui il valore medio $\langle E \rangle = \sum_i p_i E_i$ ha un valore fissato.*

Inoltre, il valore assunto dal funzionale di Gibbs per $p = p^{\mathrm{eq}}$ è uguale al valore dell'entropia termodinamica per lo stato di equilibrio termodinamico determinato da quel valore di $\langle E \rangle$.

Nell'approccio di Jaynes, questo risultato viene interpretato in termini di informazione mancante. Si richiede cioè che la scelta delle probabilità p sia tale da massimizzare l'informazione mancante sul sistema, associata alla distribuzione di probabilità, fra tutte le distribuzioni che soddisfano i vincoli di normalizzazione $\sum_i = 1$ e di valor medio dell'energia $\sum_i p_i E_i = \mathcal{E} = \text{cost}$.

1. Teorema di Shannon

In questo paragrafo, mostriamo che il funzionale di Gibbs, definito su una distribuzione di probabilità discreta $p = (p_1, \ldots, p_n)$, può essere interpretato come una misura dell'informazione mancante sulla descrizione di un evento che può verificarsi in una fra n alternative, e tali che p_i è la probabilità che si verifichi l'alternativa i. Il fatto che questa sia l'espressione dell'informazione mancante è noto come **teorema di Shannon**.

Si supponga che un evento debba verificarsi in una fra n alternative, e che all'alternativa i sia associata la probabilità p_i. Vogliamo associare alla distribuzione $p = (p_1, \ldots, p_n)$ una misura $I_n(p)$ della quantità di informazione mancante, tale che siano soddisfatti i seguenti postulati:

Irrilevanza: Alternative che hanno probabilità nulla di verificarsi non modificano l'informazione mancante. In formule:

$$I_{n+\ell}(p_1, p_2, \ldots, p_n, p_{n+1} = 0, \ldots, p_{n+\ell} = 0) = I_n(p_1, \ldots, p_n). \quad (3)$$

Monotonicità: Se tutte le alternative hanno la stessa probabilità, l'informazione mancante deve crescere al crescere del numero di alternative. Si deve cioè avere, per $n < m$,

$$I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) < I_m\left(\frac{1}{m}, \ldots, \frac{1}{m}\right). \quad (4)$$

Simmetria: L'informazione mancante non dipende da come vengono ordinate le alternative: per ogni permutazione P degli n indici, si deve avere

$$I_n\left(p_{P(1)}, \ldots, p_{P(n)}\right) = I_n(p_1, \ldots, p_n). \quad (5)$$

Additività: L'informazione mancante deve essere indipendente da come si suppone di venire a conoscere il risultato dell'evento. Supponiamo, per esempio, di raggruppare le alternative $\{1, 2, \ldots, n\}$ in m gruppi, con $m < n$, tali che il gruppo j possiede n_j elementi, e la probabilità P_j del gruppo j è data da $P_j = \sum_{i \in j} p_i$. Allora l'informazione mancante corrisponde all'ignoranza del gruppo cui appartiene l'evento che si verifica, più l'ignoranza su quale particolare evento del gruppo si sia effettivamente verificato. Indichiamo con $p(i \,|\, j)$ la probabilità condizionata che si verifichi l'evento i, ammesso che l'evento che si verifica appartiene al gruppo j. Si ha allora, per definizione di probabilità condizionata,

$$p(i \,|\, j) = \begin{cases} p_i \Big/ \displaystyle\sum_{i' \in j} p_{i'}, & \text{se } i \in j, \\ 0, & \text{altrimenti.} \end{cases} \quad (6)$$

Indichiamo con k_j il numero di elementi del gruppo j. Dobbiamo avere allora

$$I_n(p) = I_m(P) + \sum_j P_j I_{k_j}(\pi_j), \tag{7}$$

dove $\pi_j = (p(i\,|\,j))$ è la collezione delle probabilità condizionate dal gruppo j.

Continuità: $I_n(p)$ è una funzione continua delle p_i, cioè se le p_i cambiano di poco, anche $I_n(p)$ cambierà di poco.

Notiamo che la (7) implica che $I_1(1) = 0$. Consideriamo infatti due descrizioni equivalenti della stessa situazione. Abbiamo n alternative, $\{1, 2, \ldots, n\}$, ciascuna delle quali si verifica con probabilità p_i. Oppure possiamo dividere queste alternative in n gruppi, ognuno dei quali contiene un solo elemento. Abbiamo così $P_j = p_j$, $j = 1, 2, \ldots, n$, e $p(i\,|\,j) = \delta_{ij}$. D'altra parte abbiamo

$$I_n(p) = I_n(P) + \sum_j P_j I_1(1). \tag{8}$$

Ma, dato che $P = p$, questo implica

$$\sum_j p_j I_1(1) = I_1(1) = 0. \tag{9}$$

Valutiamo adesso $I_n(p)$ quando le alternative sono equivalenti. Definiamo

$$S_n = I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right). \tag{10}$$

Notiamo che si deve avere, per l'additività

$$S_{nm} = S_n + S_m. \tag{11}$$

In effetti, date nm alternative equivalenti, possiamo raggrupparle in n gruppi di m elementi l'uno. Applicando a questa situazione la (7) otteniamo

$$S_{nm} = S_n + \sum_{j=1}^{n} \frac{1}{n} S_m = S_n + S_m. \tag{12}$$

L'equazione (11) ammette evidentemente come soluzione

$$S_n = k \log n, \tag{13}$$

dove $\log n$ è il logaritmo naturale di n e k è un'arbitraria costante positiva. Ora questa è anche l'unica soluzione generale per cui vale la relazione di monotonicità.

In effetti, data la (11), possiamo scriverne la soluzione generale nella forma

$$S_n = m_1 S_{p_1} + m_2 S_{p_2} + \cdots + m_\ell S_{p_\ell}, \tag{14}$$

dove $p_1, p_2, \ldots, p_\ell$ sono i fattori primi che appaiono nello sviluppo di n, e $m_1, m_2, \ldots, m_\ell$ le relative molteplicità. Quindi è sufficiente valutare S_p per i numeri primi p. Supponiamo quindi che $S_p = k \log p$ per un certo numero primo p, e valga invece $S_q = k' \log q$ per un differente numero primo q, con $k' \neq k$. Mostriamo che, in questo caso, verrebbe violata la condizione di monotonicità. Dalla (14) otteniamo infatti l'espressione di S_n quando n è una potenza di p o di q. Si ha allora

$$S_n = \begin{cases} k \log n, & \text{se } n \text{ è una potenza di } p; \\ k' \log n, & \text{se } n \text{ è una potenza di } q. \end{cases} \tag{15}$$

Per fissare le idee, supponiamo $k' < k$. Ora è possibile trovare una potenza n di q che sia appena maggiore di una potenza m di p, e tale che $k' \log n < k \log m$:

questo contraddice la monotonicità. Approssimiamo $\log q/\log p$ mediante un numero razionale:

$$\frac{r}{s} < \frac{\log q}{\log p} < \frac{r+1}{s}. \tag{16}$$

Sappiamo che è possibile trovare un'approssimazione di questo tipo con s grande a piacere. Avremo allora

$$r \log p < s \log q < (r+1) \log p. \tag{17}$$

Scegliendo quindi $n = q^s$ e $m = p^r$, avremo $S_n = k' \log n$ e $S_m = k \log m$, e quindi $S_n - S_m = k' \log n - k \log m = (k' - k) \log m + k' \log(n/m)$. Abbiamo per ipotesi $(k' - k) \log m < 0$. Mostriamo che si può scegliere m e n in modo che il primo termine sia più piccolo (in modulo) del secondo. Si ha

$$\begin{aligned} k' \log(n/m) &= k'(\log n - \log m) = k'(s \log q - r \log p) \\ &< k' \log p, \end{aligned} \tag{18}$$

dove abbiamo sfruttato la (17). Basterà quindi scegliere m in modo che $(k - k') \log m > k' \log p$ per violare la monotonicità.

Possiamo adesso valutare $I_n(p_1, p_2, \ldots, p_n)$ per delle probabilità arbitrarie $p = (p_1, \ldots, p_n)$. Sfruttando la continuità, potremo limitarci al caso in cui le p_i sono razionali. Supponiamo quindi di scrivere $p_i = m_i/N$, dove N è un determinato intero. Allora potremo rappresentare le alternative supponendo che abbiamo a che fare con N alternative equiprobabili, di cui m_1 stanno nel primo gruppo, m_2 nel secondo, ecc. In questo modo avremo, per l'additività,

$$\begin{aligned} I_N\left(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right) &= I_n(p_1, p_2, \ldots, p_n) + \sum_{i=1}^{n} p_i I_{m_i}\left(\frac{1}{m_i}, \ldots, \frac{1}{m_i}\right) \\ &= I_n(p_1, p_2, \ldots, p_n) + \sum_{i=1}^{n} p_i \, k \log m_i. \end{aligned} \tag{19}$$

Ma

$$I_N\left(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right) = k \log N, \tag{20}$$

e quindi

$$\begin{aligned} I_n(p_1, p_2, \ldots, p_n) &= -k \sum_{i=1}^{n} p_i \log m_i + k \log N \\ &= -k \sum_{i=1}^{n} p_i \log m_i + k \sum_i p_i \log N \\ &= -k \sum_{i=1}^{n} p_i \log p_i. \end{aligned} \tag{21}$$

Evidentemente fissare la costante k equivale a fissare la base rispetto a cui vengono valutati i logaritmi. Per convenzione si sceglie di valutare i logaritmi in base 2, che corrisponde alla scelta

$$k = (\log 2)^{-1}. \tag{22}$$

L'unità di informazione corrispondente a questa scelta viene chiamata *bit*.

Notiamo che l'espressione (21) dell'informazione mancante soddisfa la seguente disuguaglianza:

$$I_n(p_1, \ldots, p_n) \leq I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = k \log n. \tag{23}$$

Questa disuguaglianza discende dalla ben nota relazione soddisfatta dal logaritmo:

$$\log x \leq x - 1. \tag{24}$$

Si ha allora, supponendo $p_i > 0$, $\forall i$,

$$\begin{aligned}
\Delta I &= I_n(p_1, \ldots, p_n) - k \log n \\
&= -k \left(\sum_{i=1}^{n} p_i \log p_i + \log n \right) = -k \sum_{i=1}^{n} p_i \log (n p_i) \\
&= k \sum_{i=1}^{n} p_i \log \left(\frac{1}{n p_i} \right) \leq k \sum_{i=1}^{n} p_i \left(\frac{1}{n p_i} - 1 \right) \\
&= k \sum_{i=1}^{n} \left(\frac{1}{n} - p_i \right) = k \left(n \frac{1}{n} - 1 \right) = 0.
\end{aligned} \tag{25}$$

Questo risultato è in accordo con la richiesta intuitiva secondo la quale l'informazione mancante è massima quando le n alternative possibili sono tutte equivalenti.

Possiamo adesso sottintendere la dipendenza da n, e otteniamo quindi che l'informazione mancante è proporzionale al funzionale di Gibbs valutato sulla distribuzione di probabilità $p = (p_1, \ldots, p_n)$:

$$I(p) = -k \sum_{i=1}^{n} p_i \log p_i. \tag{26}$$

Notiamo che, in questo contesto, il funzionale di Gibbs è anche chiamato **entropia di Shannon**.

2. Ensemble statistici e informazione mancante

Supponiamo di volere descrivere la nostra conoscenza relativa allo stato un sistema fisico che può trovarsi in uno fra n microstati mediante una distribuzione di probabilità $p = (p_1, \ldots, p_n)$. Supponiamo in particolare di conoscere il valore medio, $\mathcal{E} = \langle E \rangle$ di una funzione E_i del microstato. Quale distribuzione di probabilità rappresenterà al meglio il nostro stato di conoscenza?

Per risolvere questo problema, si introduce il seguente **principio di massimo dell'entropia di Shannon**.

> *La distribuzione p deve corrispondere al massimo valore dell'informazione mancante $I(p)$, fra tutte le distribuzioni che soddisfano i vincoli*
>
> $$\sum_{i=1}^{n} p_i = 1; \qquad \sum_{i=1}^{n} p_i E_i = \mathcal{E}. \tag{27}$$

Per giustificare questo principio, indichiamo con p^{eq} la distribuzione che lo soddisfa (per un dato valore di $\mathcal{E}$), e consideriamo un'altra distribuzione p che soddisfa i vincoli, ma che corrisponde a un valore $I(p)$ inferiore al massimo $I(p^{\mathrm{eq}})$. Allora, per definizione, la conoscenza che avremmo dello stato del sistema dalla distribuzione p sarà *più precisa* di quella corrispondente alla distribuzione p^{eq}. Ma, per ipotesi, *tutta*

la nostra conoscenza dello stato del sistema è contenuta nel valore $\mathcal{E}$ di $\langle E \rangle$, vincolo che è soddisfatto dalla p^{eq}. Quindi questa maggiore conoscenza sullo stato del sistema non è giustificata dallo stato delle nostre informazioni, ed è fuorviante.

Vediamo adesso di valutare p^{eq}. Introduciamo i moltiplicatori di Lagrange α e β per imporre i vincoli. Cerchiamo quindi il massimo dell'espressione

$$\Phi = I(p) - \alpha \sum_i p_i - \beta \sum_i p_i E_i = -\sum_i p_i \left[\log p_i + \alpha + \beta E_i\right]. \tag{28}$$

Abbiamo posto $k = 1$ per semplificare le formule. Otteniamo quindi, per ogni i,

$$\frac{\partial \Phi}{\partial p_i} = -\left[\log p_i + \alpha + \beta E_i\right] - 1. \tag{29}$$

Quindi, imponendo $\partial\Phi/\partial p_i = 0$, otteniamo

$$p_i \propto \exp(-\beta E_i). \tag{30}$$

La costante di proporzionalità viene determinata dalla condizione di normalizzazione. Definendo

$$Z = \sum_i \exp(-\beta E_i), \tag{31}$$

otteniamo

$$p_i = p_i^{\mathrm{eq}} = \frac{\mathrm{e}^{-\beta E_i}}{Z}. \tag{32}$$

D'altra parte, il valore di β viene determinato dalla condizione

$$\langle E \rangle = \sum_i p_i E_i = \mathcal{E}. \tag{33}$$

Notiamo che, se p_i ha la forma (32), si ha

$$\langle E \rangle = -\frac{\partial \log Z}{\partial \beta}. \tag{34}$$

Inoltre si ha

$$\frac{\partial \langle E \rangle}{\partial \beta} = -\left\langle \left(E - \langle E \rangle\right)^2 \right\rangle < 0. \tag{35}$$

Quindi è possibile risolvere rispetto a β l'equazione

$$-\frac{\partial \log Z}{\partial \beta} = \mathcal{E}. \tag{36}$$

Abbiamo così ottenuto che la distribuzione che soddisfa il principio di massimo dell'entropia di Shannon, e soddisfa (oltre al vincolo di normalizzazione) il vincolo $\langle E \rangle = \mathcal{E}$ è una distribuzione canonica in E, con un determinato valore di β. L'entropia di Shannon corrispondente è proporzionale all'entropia termodinamica dello stato a temperatura $k_{\mathrm{B}} T = 1/\beta$.

È immediato generalizzare questo risultato al caso in cui si dispone di informazioni sui valori medi di più quantità.

È anche utile mostrare esplicitamente che l'informazione mancante per una qualunque distribuzione p diversa dalla distribuzione canonica, ma che sia normalizzata e che abbia lo stesso valore di $\langle E \rangle$, è inferiore a quello della corrispondente distribuzione canonica. Vogliamo mostrare cioè che, se p_i soddisfa le relazioni (27), si ha

$$I(p) \leq I(p^{\mathrm{eq}}), \tag{37}$$

dove p^{eq} è definito dalla (32). Si ha in effetti

$$I(p^{\mathrm{eq}}) = -k \sum_i p_i^{\mathrm{eq}} \log p_i^{\mathrm{eq}} = k \left(\log Z + \beta \mathcal{E} \right), \tag{38}$$

dove Z è definito dalla (31) ed $\mathcal{E}$ è definito dalla (33). Si ha allora

$$\begin{aligned} \Delta I = I(p^{\mathrm{eq}}) - I(p) &= k \left(\log Z + \beta \mathcal{E} \right) + k \sum_i p_i \log p_i \\ &= k \sum_i p_i \log p_i + k \sum_i p_i \left(\log Z + \beta E_i \right) \\ &= k \sum_i p_i \log \frac{p_i}{p_i^{\mathrm{eq}}}. \end{aligned} \tag{39}$$

La quantità a secondo membro è proporzionale alla cosiddetta **divergenza di Kullback-Leibler** della distribuzione p dalla distribuzione p^{eq}, definita da

$$D_{\mathrm{KL}}(p \,|\, p^{\mathrm{eq}}) = \sum_i p_i \log \frac{p_i}{p_i^{\mathrm{eq}}}. \tag{40}$$

Questa espressione è una misura della "distanza" delle due distribuzioni. In senso matematico, però, non può essere considerata una distanza, perché non è simmetrica:

$$D_{\mathrm{KL}}(p \,|\, p^{\mathrm{eq}}) \neq D_{\mathrm{KL}}(p^{\mathrm{eq}} \,|\, p). \tag{41}$$

Tuttavia, si vede facilmente che essa non è mai negativa, e si annulla solo se $p = p^{\mathrm{eq}}$. Si ha infatti

$$\begin{aligned} D_{\mathrm{KL}}(p \,|\, p^{\mathrm{eq}}) &= -\sum_i p_i \log \frac{p_i^{\mathrm{eq}}}{p_i} \\ &\geq -\sum_i p_i \left(\frac{p_i^{\mathrm{eq}}}{p_i} - 1 \right) = -\sum_i \left(p_i^{\mathrm{eq}} - p_i \right) = 0. \end{aligned} \tag{42}$$

Bibliografia

[1] E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev.* **106** (1957) 620.
[2] E. T. Jaynes, *Probability: The Logic of Science* (Cambridge: Cambridge U. P, 2003).