Diego Rovetta

Reti neuronali, macchine di Boltzmann e convergenza asintotica del *Simulated Annealing*

Corso di Matematica discreta

A.A. 2003

Reti neuronali, macchine di Boltzmann e convergenza asintotica del Simulated Annealing

Riassunto

Le reti neurali artificiali o reti neuronali sono delle macchine (o modelli computazionali) che funzionano imitando il comportamento del cervello umano. In questo lavoro verranno presentati alcuni collegamenti fra la teoria dei grafi e le reti neuronali; inoltre, si descriverà come molti problemi di ottimizzazione combinatoria possano essere risolti con un particolare tipo di rete neurale artificiale, detta macchina di Boltzmann. In seguito verrà approfondito il comportamento asintotico dell'algoritmo del Simulated Annealing, mostrando come esso, sotto particolari ipotesi, possa convergere alla soluzione ottima di un problema di ottimizzazione combinatoria.

1 Reti neurali artificiali

Una rete neurale è un grafo orientato, cioè un insieme di nodi e di archi orientati che li connettono. I nodi vengono detti unità della rete (sono l'analogo dei neuroni) e gli archi (che costituiscono l'analogo delle sinapsi) vengono detti connessioni. Ogni unità ha un certo numero di connessioni in ingresso e/o un certo numero di connessioni in uscita. Ciascuna unità costituisce un processore, un singolo dispositivo di calcolo che, ad ogni fase del calcolo, riceve i propri *input* attraverso le connessioni in ingresso, li elabora e invia l'*output* alle altre unità connesse per mezzo delle sue connessioni in uscita. I dati che le unità elaborano e si scambiano sono valori di tipo numerico: di solito, si tratta di numeri reali. Le connessioni non sono canali di trasmissione neutri, ma modificano i dati che trasmettono moltiplicandoli per un certo valore numerico che è associato a ciascuna di esse. Tale valore è detto il peso, o la forza della connessione (secondo l'analogia neuronale, il peso corrisponde alla forza di una sinapsi). Date due unità connesse i e j di una rete, indichiamo con w_{ij} il peso della connessione che va da i a j; l'aggiustamento dei pesi viene indicato come processo di apprendimento della rete. In un sistema connessionista, in ogni fase del calcolo ogni unità ha un certo valore di attivazione. I valori di attivazione delle unità della rete influenzano quelli delle unità vicine, o in modo stocastico, come nelle macchine di Boltzmann, o in modo deterministico, come nelle reti di tipo *feedforward*.

1.1 Macchine di Boltzmann

In una macchina di Boltzmann M, che è un particolare tipo di rete neuronale stocastica, il grafo orientato G che la descrive, presenta connessioni simmetriche:

$$ij \in G \Leftrightarrow ji \in G \tag{1}$$

Inoltre i pesi sono sempre vincolati in modo tale che:

$$w_{ij} = w_{ji} \tag{2}$$

Una macchina di Boltzmann M potrebbe allora essere descritta da un grafo non orientato (che potrebbe avere anelli) con archi pesati e unità della rete ai nodi; più precisamente una macchina di Boltzmann M è una coppia (G, Ω), dove $G = (V, E)$ è un grafo con n nodi ed m archi e $\Omega \subseteq \mathbb{R}^m \times \{0,1\}^n$ è un insieme di possibili stati. Per ogni stato $\omega = (w_{e1}, w_{e2}, ..., w_{em}, T_1, T_2, ..., T_n) \in \Omega$, il vettore $\mathbf{w} = (w_{e1}, w_{e2}, ..., w_{em})$ descrive i pesi da assegnare agli archi $e_1, e_2, ..., e_m$ del grafo e il vettore $(T_1, T_2, ..., T_n)$ descrive quali nodi (o unità computazionali della rete) sono attivati, dove un vertice i è attivato nello stato ω se e solo se $T_i = 1$.

La funzione di consenso della macchina di Boltzmann M è la funzione $g : \Omega \to \mathbb{R}$, data da:

$$g(\omega) = \sum_{ij \in E} w_{ij} T_i T_j \tag{3}$$

Il calcolo procede all'interno della macchina in modo stocastico e tale che il consenso venga aumentato. Così, se w_{ij} è positivo c'è una tendenza ad avere le unità i e j entrambe attivate o entrambe disattivate, mentre se il peso è negativo, c'è una tendenza ad averle con differenti attivazioni (un'unità attivata e l'altra no). Quando un peso è positivo si dice eccitatorio; in caso contrario, si dice inibitorio.

Descriviamo ora come evolve lo stato della rete quando i pesi sono fissati.

Se i è un vertice ed ω è uno stato, allora $\omega[i \to \overline{i}]$ è lo stato ottenuto da ω, cambiando T_i con $1 - T_i$ (*flipping* dell'attivazione di i). La funzione di consenso cambia nel seguente modo:

$$\Delta g_\omega(i) = g(\omega[i \to \overline{i}]) - g(\omega) \tag{4}$$

Se $\Delta g_\omega(i)$ è grande e positivo, allora è vantaggioso mutare l'attivazione di i; in caso contrario, ciò è svantaggioso. La decisione di mutare l'attivazione di i è presa stocasticamente e si basa su $\Delta g_\omega(i)$.

Il modello computazionale più semplice di macchina di Boltzmann è quello sequenziale. La macchina ha un orologio interno (*clock*) ed all'istante $t - esimo$ un nodo i_t viene scelto casualmente da una distribuzione di probabilità uniforme.

A questo punto, l'attivazione di i_t viene mutata con probabilità pari a:

$$P(\omega \to \omega[i_t \to \overline{i_t}]) = \frac{1}{1 + e^{-\beta \Delta g_\omega(i_t)}} \tag{5}$$

per qualche costante $\beta > 0$.

Si dimostra che il calcolo computazionale sequenziale converge ad una distribuzione di probabilità stazionaria su tutto l'insieme degli stati, in cui la probabilità che lo stato della macchina sia ω, è proporzionale a $e^{\beta g(\omega)}$. In una distribuzione stazionaria gli stati ad alto consenso sono i più probabili.

Il reciproco del parametro β è detto temperatura, $c = \frac{1}{\beta}$. Si dimostra che l'algoritmo di ottimizzazione del *Simulated Annealing* si comporta meglio se la temperatura viene fatta diminuire lentamente; inoltre si mostrerà in seguito che se il tempo tende all'infinito e la temperatura $c \to 0$, allora la distribuzione stazionaria limite degli stati è uniforme su tutti gli stati che massimizzano la funzione di consenso.

2 Ottimizzazione combinatoria

2.1 Problemi di ottimizzazione combinatoria

Definizione 2.1

Un problema di ottimizzazione combinatoria è un problema di massimizzazione o di minimizzazione ed è caratterizzato da un insieme di istanze.

Definizione 2.2

Un'istanza di un problema di ottimizzazione combinatoria può essere formalizzata con la coppia (S, f), dove lo spazio della soluzione S denota un insieme finito di tutte le possibili soluzioni e la funzione di costo (o funzione obiettivo) f è una mappa definita da:

$$f : S \to \mathbb{R} \tag{6}$$

Nel caso di minimizzazione, il problema si riduce a quello di individuare una soluzione $i_{opt} \in S$ che soddisfi:

$$f(i_{opt}) \leq f(i), \ \forall i \in S \tag{7}$$

Nel caso di massimizzazione, i_{opt} soddisfa:

$$f(i_{opt}) \geq f(i), \ \forall i \in S \tag{8}$$

Una soluzione i_{opt} viene detta soluzione globalmente ottima, oppure massimo (o minimo). $f_{opt} = f(i_{opt})$ denota il costo ottimo ed S_{opt} l'insieme delle soluzioni ottime.

2.2 Ricerca locale

Gli algoritmi di ricerca locale si basano sul miglioramento passo dopo passo di una soluzione, esplorando le soluzioni ad essa vicine (locali) e determinando quali fra queste ottimizzano la funzione costo. L'uso di un algoritmo di ricerca locale presuppone la definizione delle soluzioni, una funzione costo ed una struttura locale.

Definizione 2.3

Sia (S, f) un'istanza di un problema di ottimizzazione combinatoria. Una struttura locale è una funzione così definita:

$$N : S \rightarrow 2^S \tag{9}$$

che definisce per ogni soluzione $i \in S$ un insieme $S_i \in S$ di soluzioni che sono "vicine" ad i in qualche senso. L'insieme S_i è detto insieme delle soluzioni vicine ad i e ogni $j \in S_i$ è detta soluzione vicina ad i. Inoltre, si assume che:

$$j \in S_i \Longleftrightarrow i \in S_j \tag{10}$$

2.3 Equilibrio statistico

Teorema 2.1

Siano date l'istanza (S, f) di un problema di ottimizzazione statistica ed una struttura locale opportuna. Si consideri, inoltre, la distribuzione di probabilità data da:

$$P_c(\mathbf{X} = i) \stackrel{def}{=} q_i(c) = \frac{1}{N_0(c)} e^{\left(-\frac{f(i)}{c}\right)} \tag{11}$$

con X la variabile aleatoria che identifica la soluzione corrente ottenuta da un algoritmo di ricerca locale e con

$$N_0(c) = \sum_{j \in S} e^{\left(-\frac{f(j)}{c}\right)} \tag{12}$$

costante di normalizzazione.

Si dimostra che:

$$\lim_{c \rightarrow 0} q_i(c) \stackrel{def}{=} q_i^* = \frac{1}{|S_{opt}|} \chi_{(S_{opt})}(i) \tag{13}$$

in cui S_{opt} rappresenta l'insieme delle soluzioni globalmente ottime e χ la funzione caratteristica (siano dati due insiemi A e $A' \subset A$; la funzione caratteristica $\chi_{(A')} : A \rightarrow \{0, 1\}$ dell'insieme A' è definita come $\chi_{(A')}(a) = 1$, se $a \in A'$ e $\chi_{(A')}(a) = 0$, altrimenti).

Dimostrazione.

$$\begin{aligned}
\lim_{c\to 0} q_i(c) &= \lim_{c\to 0} \frac{e^{\left(-\frac{f(i)}{c}\right)}}{\sum_{j\in S} e^{\left(-\frac{f(j)}{c}\right)}} = \\
&= \lim_{c\to 0} \frac{e^{\left(\frac{f_{opt}-f(i)}{c}\right)}}{\sum_{j\in S} e^{\left(\frac{f_{opt}-f(j)}{c}\right)}} = \\
&= \lim_{c\to 0} \frac{1}{\sum_{j\in S} e^{\left(\frac{f_{opt}-f(j)}{c}\right)}} \chi_{(S_{opt})}(i) + \\
&\quad + \lim_{c\to 0} \frac{e^{\left(\frac{f_{opt}-f(i)}{c}\right)}}{\sum_{j\in S} e^{\left(\frac{f_{opt}-f(j)}{c}\right)}} \chi_{(S\setminus S_{opt})}(i) \qquad (14)
\end{aligned}$$

quindi, ricordando il limite notevole:

$$\lim_{x\to 0} e^{\frac{a}{x}} = \begin{cases} 1,\ a = 0 \\ 0,\ a < 0 \end{cases} \qquad (15)$$

si ha:

$$\lim_{c\to 0} q_i(c) = \frac{1}{|S_{opt}|} \chi_{(S_{opt})}(i) + 0 \qquad (16)$$

che completa la dimostrazione.

Il risultato di questo teorema è molto importante perché garantisce l'asintotica convergenza degli algoritmi di ricerca locale (come il *Simulated Annealing*) all'insieme delle soluzioni globalmente ottime, a patto che vengano soddisfatte le condizioni del teorema; in particolare la distribuzione stazionaria (11) deve essere calcolata per ogni valore di c.

3 Convergenza asintotica del *Simulated Annealing.*

L'algoritmo del *Simulated Annealing* può essere modellizzato usando la teoria delle catene di Markov.

3.1 Teoria delle catene di Markov

Definizione 3.1
Sia O l'insieme dei possibili esiti in un processo di estrazione. Una catena di Markov è una sequenza di estrazioni, in cui la probabilità dell'esito di una estrazione dipende soltanto dall'esito dell'estrazione precedente. Sia $\mathbf{X}(k)$ una variabile casuale che rappresenta l'esito della $k-esima$ estrazione; per ogni coppia di valori $i, j \in O$, si definisce probabilità di transizione dall'esito i all'esito j, all'estrazione k, la probabilità:

$$P_{ij}(k) = P(\mathbf{X}(k) = j | \mathbf{X}(k-1) = i) \tag{17}$$

La matrice $\mathbf{P}$, i cui elementi sono ottenuti dalla (17), è detta matrice di transizione.

Sia $a_i(k)$ la probabilità che si presenti l'esito i all'estrazione k:

$$a_i(k) = P(\mathbf{X}(k) = i) \tag{18}$$

Allora $P_{ij}(k)$ e $a_i(k)$ sono legate dalla seguente formula ricorsiva:

$$a_i(k) = \sum_l a_l(k-1) P_{li}(k) \tag{19}$$

Definizione 3.2
Una catena di Markov è detta finita se è definita su un insieme di esiti O finito.

Definizione 3.3
Una catena di Markov è detta disomogenea se le probabilità di transizione ad essa associate dipendono dal numero d'estrazione k. In caso contrario la catena di Markov è detta omogenea.

Nell'algoritmo del *Simulated Annealing*, ogni transizione corrisponde ad una estrazione e l'insieme delle soluzioni del problema di ottimizzazione combinatoria corrisponde all'insieme dei possibili esiti di ogni estrazione. Siccome nel caso del *Simulated Annealing* l'esito di un'estrazione dipende soltanto dall'esito

dell'estrazione precedente, possiamo utilizzare il concetto di catene di Markov; si tratta di catene di Markov finite perché l'insieme delle soluzioni O è finito.

Definizione 3.4

Un vettore **a** è detto stocastico se le sue componenti a_i soddisfano le seguenti condizioni:

$$a_i \geq 0, \forall i \text{ e } \sum_i a_i = 1 \tag{20}$$

Una matrice **P** è detta stocastica se le sue componenti P_{ij} soddisfano le seguenti condizioni:

$$P_{ij} \geq 0, \forall i, j \text{ e } \sum_j P_{ij} = 1, \forall i \tag{21}$$

Definizione 3.5 (Probabilità di transizione)
Sia (S, f) l'istanza di un problema di ottimizzazione combinatoria. Allora le probabilità di transizione per l'algoritmo del *Simulated Annealing* sono definite come:

$$\forall i, j \in S : P_{ij}(k) = P_{ij}(c_k) = \begin{cases} G_{ij}(c_k)A_{ij}(c_k), \, i \neq j \\ 1 - \sum\limits_{l \in S, l \neq i} P_{il}(c_k), \, i = j \end{cases} \tag{22}$$

dove $G_{ij}(c_k)$ rappresenta la probabilità di generazione, ossia la probabilità di generare una soluzione j a partire da una soluzione i, mentre $A_{ij}(c_k)$ rappresenta la probabilità di accettazione, ossia la probabilità di accettare la soluzione j, una volta che questa è stata generata dalla soluzione i.

$G_{ij}(c_k)$ e $A_{ij}(c_k)$ sono probabilità condizionate; le corrispondenti matrici $\mathbf{G}(c_k)$ e $\mathbf{A}(c_k)$ sono chiamate matrice di generazione e matrice di accettazione.

Di seguito verranno usati dei particolari insiemi di probabilità condizionate, individuati dalle seguenti definizioni.

Definizione 3.6 (Probabilità di generazione)

$$\forall i, j \in S : G_{ij}(c_k) = G_{ij} = \frac{1}{\Theta}\chi_{(S_i)}(j) \tag{23}$$

dove $\Theta = |S_i|, \forall i \in S$.

Definizione 3.7 (Probabilità di accettazione)

$$\forall i, j \in S : A_{ij}(c_k) = e^{\left(-\frac{a^+}{c_k}\right)} \tag{24}$$

dove $a^+ = f(j) - f(i)$, se $f(j) > f(i)$ e $a^+ = 0$, altrimenti; ciò nel caso in cui l'ottimizzazione sia una minimizzazione della funzione costo.

Così, le probabilità di generazione sono scelte indipendenti dal parametro di controllo c_k ed uniformi sull'insieme S_i delle soluzioni vicine ad i; si è inoltre assunto che tale insieme abbia dimensione costante, indipendentemente da i: $|S_i| = \Theta$, $\forall i \in S$.

Le definizioni di probabilità di generazione e di accettazione date da (23) e (24) corrispondono alla definizione originale del *Simulated Annealing*; queste definizioni, almeno in linea di principio, possono essere impiegate per risolvere qualunque problema di ottimizzazione combinatoria.

A questo punto vengono esposte le proprietà di convergenza asintotica dell'algoritmo *Simulated Annealing.* Quest'ultimo trova con probabilità unitaria una soluzione ottima se, dopo un certo numero sufficientemente grande di tentativi, accade che:

$$P(\mathbf{X}(k) \in S_{opt}) = 1 \tag{25}$$

Nel paragrafo successivo si dimostra che, sotto alcune ipotesi, il *Simulated Annealing* converge asintoticamente all'insieme delle soluzioni ottime:

$$\lim_{k\to\infty} P(\mathbf{X}(k) \in S_{opt}) = 1 \tag{26}$$

3.2 Distribuzione stazionaria

La dimostrazione della convergenza asintotica dell'algoritmo del *Simulated Annealing* richiede l'esistenza di un'unica distribuzione stazionaria; tale distribuzione esiste solo a certe condizioni sulle catene di Markov associate all'algoritmo.

Si dimostrerà di seguito che per le catene di Markov omogenee la distribuzione stazionaria assume la forma indicata da (11); per questo motivo si assuma che il valore del parametro di controllo c_k sia indipendente da k, ossia $c_k = c$, $\forall k$. Quindi $\mathbf{P}(k) = \mathbf{P}$, $\forall k$, che corrisponde ad una catena di Markov omogenea.

Definizione 3.8

La distribuzione stazionaria di una catena di Markov omogenea e finita con matrice di transizione $\mathbf{P}$ viene indicata con il vettore $\mathbf{q}$, le cui componenti sono date da:

$$q_i = \lim_{k\to\infty} P(\mathbf{X}(k) = i | \mathbf{X}(0) = j),\ \forall j \tag{27}$$

Se tale distribuzione esiste, allora si ha:

$$\begin{aligned} \lim_{k\to\infty} a_i(k) &= \lim_{k\to\infty} P(\mathbf{X}(k) = i) = \\ &= \lim_{k\to\infty} \sum_j P(\mathbf{X}(k) = i | \mathbf{X}(0) = j) P(\mathbf{X}(0) = j) = \\ &= q_i \sum_j P(\mathbf{X}(0) = j) = q_i \end{aligned} \tag{28}$$

Così la distribuzione stazionaria è la distribuzione di probabilità delle soluzioni dopo un numero infinito di transizioni. Inoltre si ha:

$$\begin{aligned}\mathbf{q}^T &= \lim_{k\to\infty} \mathbf{a}^T(0)\prod_{l=1}^{k}\mathbf{P}(l) = \lim_{k\to\infty} \mathbf{a}^T(0)\mathbf{P}^k = \\ &= \lim_{k\to\infty} \mathbf{a}^T(0)\mathbf{P}^{k-1}\mathbf{P} = \lim_{l\to\infty} \mathbf{a}^T(0)\mathbf{P}^l\mathbf{P} \\ &= \mathbf{qP} \end{aligned} \tag{29}$$

$\mathbf{q}$ è l'autovettore sinistro di $\mathbf{P}$ con autovalore pari ad 1. Nel caso del *Simulated Annealing*, dato che $\mathbf{P}$ dipende da c, anche $\mathbf{q}$ dipenderà da esso: $\mathbf{q} = \mathbf{q}(c)$.

Prima di dimostrare l'esistenza della distribuzione stazionaria per il *Simulated Annealing* si considerino le seguenti definizioni.

Definizione 3.9

Una catena di Markov con matrice di transizione $\mathbf{P}$ si dice irriducibile se, per ogni coppia di soluzioni $i, j \in S$ esiste una probabilità positiva di raggiungere j da i in un numero finito di transizioni:

$$\forall i, j \in S,\ \exists n \geq 1 : (P^n)_{ij} > 0 \tag{30}$$

Definizione 3.10

Una catena di Markov con matrice di transizione $\mathbf{P}$ si dice aperiodica se, per ogni soluzione $i \in S$ il massimo comune divisore $MCD(D_i) = 1$, essendo D_i l'insieme di tutti gli interi $n > 0$ tali che:

$$(P^n)_{ii} > 0 \tag{31}$$

L'intero $MCD(D_i)$ è detto periodo della soluzione i. L'aperiodicità richiede che tutte le soluzioni abbiano periodo unitario.

Lemma 3.1

Una catena di Markov con matrice di transizione $\mathbf{P}$ è aperiodica se:

$$\exists j \in S : P_{jj} > 0 \tag{32}$$

Dimostrazione.

L'irriducibilità implica che:

$$\forall i, j \in S, \exists k, l \geq 1 : (P^k)_{ij} > 0 \text{ e } (P^l)_{ji} > 0$$

Così, $(P^n)_{ii} \geq (P^k)_{ij}(P^l)_{ji} > 0$, in cui $n = k + l$.

Inoltre, $(P^{n+1})_{ii} \geq (P^k)_{ij}P_{jj}(P^l)_{ji} > 0$. Perciò $n, n+1 \in D_i$ e di conseguenza, $MCD(D_i) = 1$, dato che $1 \leq MCD(D_i) \leq MCD(n, n+1)$ e $MCD(n, n+1) = 1$, $\forall n \geq 1$.

Teorema 3.1

Sia **P** la matrice di transizione associata ad una catena di Markov finita, omogenea, irriducibile ed aperiodica. Allora esiste un vettore stocastico **q**, le cui componenti sono univocamente determinate dall'equazione:

$$\sum_j q_j P_{ji} = q_i, \, \forall i \tag{33}$$

Chiaramente il vettore **q** è la distribuzione stazionaria della catena di Markov, dato che soddisfa l'equazione (29).

Lemma 3.2

Sia **P** la matrice di transizione associata ad una catena di Markov finita, omogenea, irriducibile ed aperiodica. Allora una data distribuzione è stazionaria se le sue componenti soddisfano la seguente equazione:

$$q_i P_{ij} = q_j P_{ji}, \, \forall i, j \in S \tag{34}$$

Dimostrazione.

L'esistenza di un'unica distribuzione stazionaria **q** è assicurata dal Teorema 3.1. Pertanto la dimostrazione del teorema si riduce a mostrare che la (34) implica la (33):

$$\begin{aligned} q_i P_{ij} &= q_j P_{ji} \\ \sum_{j \in S} q_i P_{ij} &= \sum_{j \in S} q_j P_{ji} \\ q_i &= \sum_{j \in S} q_j P_{ji} \end{aligned} \tag{35}$$

essendo **P** stocastica.

In questo modo la correttezza di una distribuzione stazionaria di una catena di Markov finita ed omogenea, con matrice di transizione **P**, può essere verificata mostrando che la catena è anche aperiodica, irriducibile e che le componenti di **q** soddisfano la (34).

L'espressione (34) è detta equazione del bilancio dettagliato, mentre la (33) è detta equazione del bilancio globale.

Teorema 3.2

Siano date l'istanza (S, f) di un problema di ottimizzazione combinatoria e $\mathbf{P}(c)$, matrice di transizione associata con l'algoritmo del *Simulated Annealing*, definito da (22), (23) e (24). Siano inoltre soddisfatte le seguenti condizioni:

$$\begin{aligned} \forall i, j &\in S \; \exists p \geq 1, \, \exists l_0, \, l_1, \, ..., \, l_p \in S, \\ \text{con } l_0 &= 1, \, l_p = j \\ &\quad \text{e} \\ G_{l_k l_{k+1}} &> 0, \, k = 0, \, 1, \, ..., \, p-1. \end{aligned} \tag{36}$$

Allora la catena di Markov ha una distribuzione stazionaria $\mathbf{q}(c)$, le cui componenti sono date da:

$$q_i(c) = \frac{1}{N_0(c)} e^{-\frac{f(i)}{c}}, \forall i \in S \tag{37}$$

dove

$$N_0(c) = \sum_{j \in S} e^{-\frac{f(j)}{c}} \tag{38}$$

Dimostrazione.

Per dimostrare il teorema si deve prima mostrare che la catena di Markov così definità è irriducibile ed aperiodica; in seguito, si applichi il lemma 3.2 e si verifichi la correttezza della distribuzione stazionaria, mostrando che le componenti date dalle (37) e (38) soddisfano l'equazione del bilancio dettagliato (34).

Irriducibilità.

Date le condizioni (36) si ha:

$$\begin{aligned} (P^P)_{ij}(c) &= \sum_{k_1 \in S} \sum_{k_2 \in S} ... \sum_{k_{p-1} \in S} P_{ik_1}(c) P_{k_1 k_2}(c) ... P_{k_{p-1} j}(c) \\ &\geq G_{il_1} A_{il_1}(c) G_{l_1 l_2} A_{l_1 l_2}(c) ... G_{l_{p-1} j} A_{l_{p-1} j}(c) \\ &> 0 \end{aligned} \tag{39}$$

essendo $A_{ij} > 0 \ \forall i, j \in S$.

Aperiodicità.

Siano $i, j \in S$ con $f(i) < f(j)$ e $G_{ij} > 0$. Dalle condizioni (36) la coppia (i, j) esiste sempre, essendo $S \neq S_{opt}$. Allora $A_{ij}(c) < 1$, e quindi:

$$\begin{aligned} P_{ii}(c) &= 1 - \sum_{l \in S, l \neq i} G_{il} A_{il}(c) = \\ &= 1 - G_{ij} A_{ij}(c) - \sum_{l \in S, l \neq i, j} G_{il} A_{il}(c) > \\ &> 1 - G_{ij} - \sum_{l \in S, l \neq i, j} G_{il} = \\ &= 1 - \sum_{l \in S, l \neq i} G_{il} = 0 \end{aligned} \tag{40}$$

Pertanto:

$$P_{ii}(c) > 0 \tag{41}$$

e la catena di Markov risulta anche aperiodica (lemma 3.1).

Dal teorema 3.1, essendo la catena irriducibile ed aperiodica, esiste un'unica distribuzione stazionaria; si applichi ora il lemma 3.2 per completare la dimostrazione del teorema, mostrando che la distribuzione stazionaria definita dalle (37) e (38) è un vettore con le seguenti proprietà: *(i)* esso è stocastico e *(ii)* le sue componenti soddisfano l'equazione del bilancio dettagliato (34).

(i) Banale.

(ii) Dalla definizione (10) si ha:

$$\chi_{(S_j)}(i) = 1 \Leftrightarrow \chi_{(S_i)}(j) = 1 \tag{42}$$

Sostituendo la precedente relazione nella definizione della probabilità di generazione (23) si trova:

$$G_{ij} = G_{ji} \tag{43}$$

Quindi l'equazione del bilancio dettagliato (34) si riduce a:

$$q_i(c)A_{ij}(c) = q_j(c)A_{ji}(c),\ \forall i,j \in S \tag{44}$$

Usando la relazione precedente e le definizioni (24), (37) e (38), si ottiene:

$$q_i(c)A_{ij}(c) = \frac{1}{N_0(c)} e^{-\frac{f(i)}{c}} e^{\left(-\frac{a^+}{c}\right)} \tag{45}$$

dove $a^+ = f(j) - f(i)$, essendo $f(j) > f(i)$. Quindi:

$$\begin{aligned} q_i(c)A_{ij}(c) &= \frac{1}{N_0(c)} e^{-\frac{f(i)}{c}} e^{\left(-\frac{(f(j)-f(i))}{c}\right)} = & (46)\\ &= \frac{1}{N_0(c)} e^{-\frac{f(j)}{c}} e^{\left(-\frac{(f(i)-f(j))+(f(j)-f(i))}{c}\right)} = & (47)\\ &= \frac{1}{N_0(c)} e^{-\frac{f(j)}{c}} e^{\left(-\frac{(f(i)-f(j))}{c}\right)} = & (48)\\ &= q_j(c)A_{ji}(c),\ \forall i,j \in S & (49) \end{aligned}$$

che completa la dimostrazione del teorema.

Si noti che le (37) e (38) descrivono una distribuzione identica a quella di Boltzmann (11) e quindi richiamando il teorema 2.1, si ha:

$$\lim_{c\to 0} \mathbf{q}(c) = \mathbf{q}^* \tag{50}$$

dove le componenti di $\mathbf{q}^*$ sono date da:

$$q_i^* = \frac{1}{|S_{opt}|} \chi_{(S_{opt})}(i) \tag{51}$$

da cui si ottiene infine:

$$\lim_{c\to 0} \lim_{k\to\infty} P_c(\mathbf{X}(k) = i) = \lim_{c\to 0} q_i(c) = q_i^* \tag{52}$$

o

$$\lim_{c \to 0} \lim_{k \to \infty} P_c(\mathbf{X}(k) \in S_{opt}) = 1 \tag{53}$$

Questo risultato riflette la proprietà principale dell'algoritmo del *Simulated Annealing*, ossia l'asintotica convergenza verso la soluzione ottima; si noti che la dimostrazione resta valida (e quindi l'algoritmo converge asintoticamente ad una soluzione ottima) anche sostituendo la definizione (22) con una più generale in cui viene imposto soltanto che la probabilità di generazione sia simmetrica:

$$G_{ij} = G_{ji}, \forall i, j \in S \tag{54}$$

Per concludere, si noti che l'implementazione dell'algoritmo richiederebbe la generazione di una sequenza di catene di Markov omogenee infinitamente lunghe, al diminuire del parametro di controllo c. Questo è chiaramente impraticabile. Fortunatamente si può descrivere l'algoritmo anche come sequenza di catene di Markov omogenee di lunghezza finita e dimostrare, sotto particolari ipotesi (lenta diminuzione del parametro di controllo c), la convergenza in distribuzione all'insieme delle soluzioni ottime [1] e [2].

4 Bibliografia

[1] *Aarts, Emile - Korst, Jan,* **Simulated Annealing and Boltzmann Machines : a stochastic approach to combinatorial optimization and neural computing**, *Wiley, 1989.*

[2] *Aarts, Emile - Lenstra, Jan Karel,* **Local search in combinatorial optimization**, edited by Emile Aarts and Jan Karel Lenstra, *Wiley, 1997.*

[3] *Beineke, Lowell - Wilson, Robin (Ed.),* **Graph Connections**, *Oxford Science Pubblications.*

[4] *Madsen, Richard W.*, **Markov chains : Theory and applications**, *John Wiley, 1976.*