

Alessandro Pluchino

RETI NEURALI MEMORIE ASSOCIAТИVE e VETRI DI SPIN



Dipartimento di Fisica e Astronomia and INFN sezione di Catania

Università di Catania - Italy

Group web page: www.ct.infn.it/~cactus

EVOLUZIONE DELL' INFORMATICA E DELLA INTELLIGENZA ARTIFICIALE

GENERAZIONI						
1 ^a	2 ^a	3 ^a	4 ^a	5 ^a	6 ^a	7 ^a
TUBI ELETTRONICI	TRANSISTOR	CIRCUITI INTEGRATI	INTEGRAZIONE VLSI	SISTEMI ESPERTI	RETI NEURALI	COMPUTER BIOLOGICI
DATI				CONOSCENZA	FORME	COMPUTER QUANTISTICI
SIMBOLISMO (APPROCCIO TOP-DOWN)			CONNESSIONISMO (APPROCCIO BOTTOM-UP)			
			INTELLIGENZA NATURALE			
ARCHITETTURA VON NEUMANN				PARALLELISMO ESTESO	STRUZZURE BIOLOGICHE	

STORIA DEL CONNESSIONISMO

Anni Trenta: Norbert Wiener fonda la Cibernetica, ispirandosi alle analogie tra sistemi naturali (sistema nervoso, sistemi sensomotori o di comunicazione negli esseri viventi) e sistemi artificiali (computer, servomeccanismi, etc.);

1943 : McCulloch e Pitts (University of Chicago) propongono un modello di neurone artificiale (combinatore lineare a soglia) e la connessione di un gran numero di queste unità in reti neuronali;

1949 : D.O.Hebb (McGill University, Montreal) propone come meccanismo di apprendimento delle reti neuronali la modifica delle connessioni tra i neuroni (pesi sinaptici) secondo la regola di aumentare il peso della connessione tra due neuroni se questi sono, per più volte di seguito, contemporaneamente attivi o inattivi;

1958 : F.Rosenblatt (Cornell University) propone la macchina Perceptron (ADALine) per il riconoscimento e la classificazione di forme (Patterns);

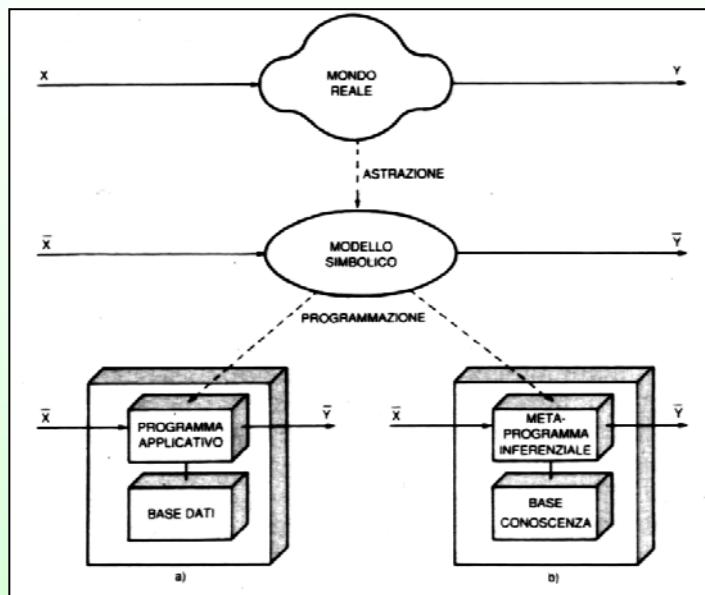
1969 : M.Minsky e S.Papert (MIT) dimostrano che il Perceptron non può risolvere molte classi di problemi (tutti quelli linearmente non separabili) e bloccano di fatto per venti anni gli ulteriori sviluppi del connessismo in favore del paradigma simbolico;

1982 : J.J.Hopfield (CalTech) propone un nuovo modello di rete neuronale a connessione totale, adatto sia alla realizzazione di memore associative, sia alla risoluzione di problemi di ottimizzazione combinatoria, sia alla modellazione di sistemi complessi fisici e biologici;

1985 : D.E.Rumelhart, G.Hinton e R.J.Williams propongono un potente algoritmo di apprendimento a retro-propagazione dell'errore (**EBP, Error Back Propagation**) che consente di superare le limitazioni del Perceptron e rilancia, alla grande, le reti neuronali e il connessismo contribuendo al fiorire di un grande numero di applicazioni commerciali.

INTELLIGENZA ARTIFICIALE e RETI NEURALI

APPROCCIO SIMBOLICO vs APPROCCIO CONNESSIONISTA



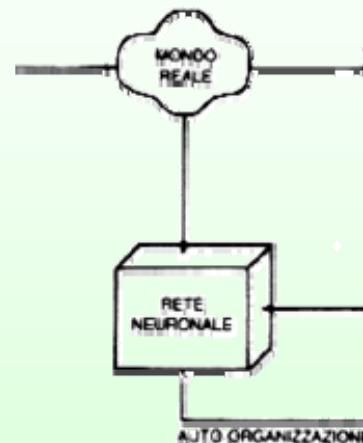
Programmazione vs Apprendimento

Rigidità vs Tolleranza

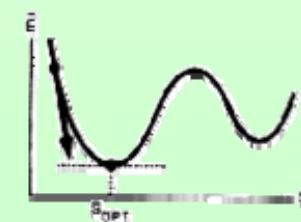
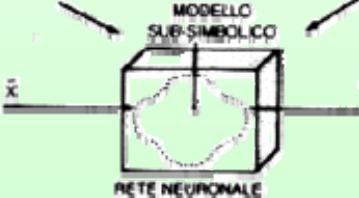
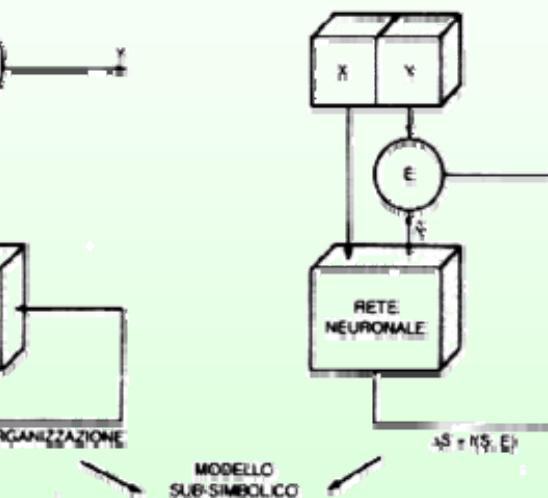
Localizzazione vs Distribuzione

Sequenzialità vs Parallelismo

**Apprendimento
'non supervised'**



**Apprendimento
'supervised'**



RETI DI NEURONI REALI E ARTIFICIALI

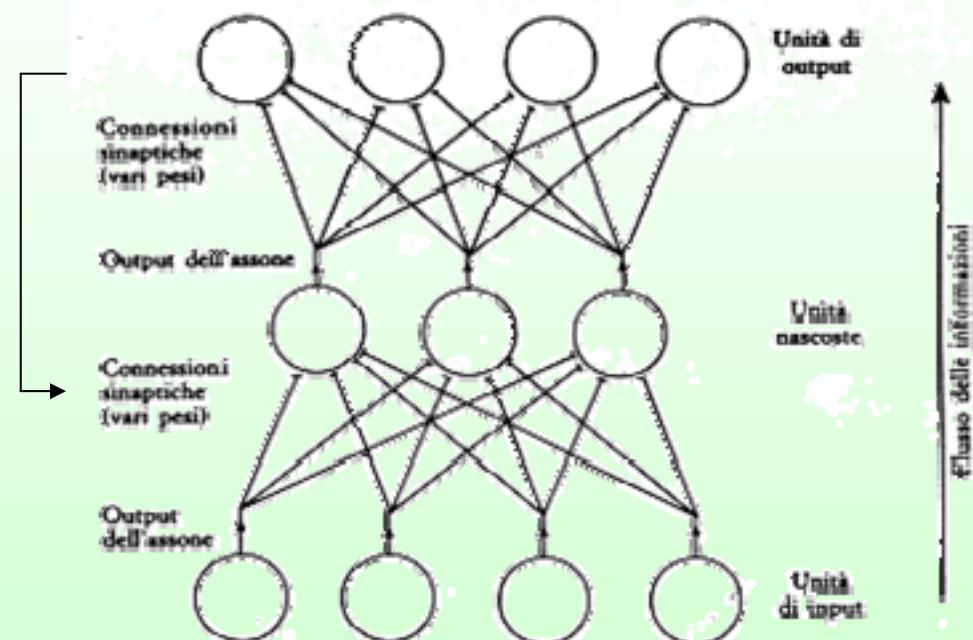
Sistema nervoso reale

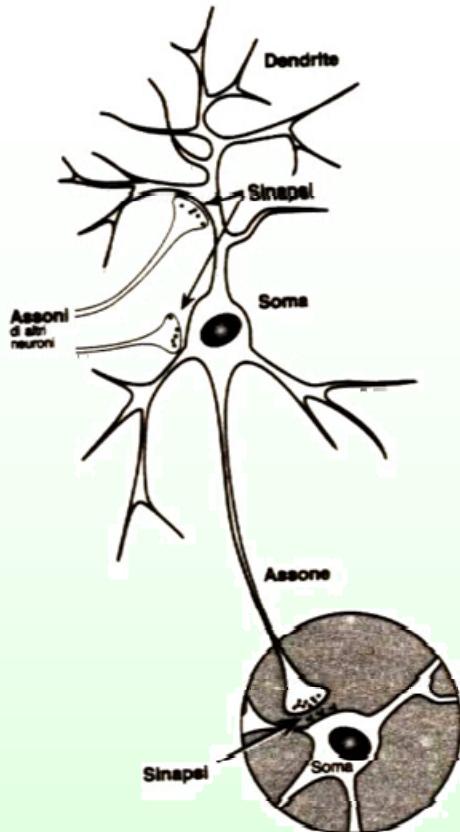
-NEURONI MOTORI
(moto-neuroni; circa 1 milione)
Agiscono in output sui muscoli

-NEURONI INTERMEDI
(inter-neuroni; 100 miliardi circa)
Ricevono l'input da altri neuroni e agiscono in output su altri neuroni

-NEURONI SENSORI
(1 milione circa) Ricevono l'input dai sensi (visivi, uditivi, olfattivi, tattili, etc.)

Rete neurale FEED-FORWARD





CARATTERISTICHE DI UN NEURONE REALE

- CORPO CELLULARE (soma)
- DENDRITI (vie di input)
- ASSONE (unica via di output)
- SINAPSI (10^4 per ogni neurone): eccitatorie o inhibitorie
- SOGLIA (varia a causa di sostanze chimiche)
- ATTIVAZIONE di tipo elettro-chimico

CARATTERISTICHE DI UN NEURONE ARTIFICIALE (Legge di attivazione)

-Stato del neurone i al tempo t : $S_i(t)$

-Input di stimolazione del neurone i :

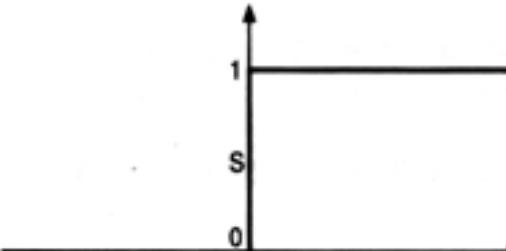
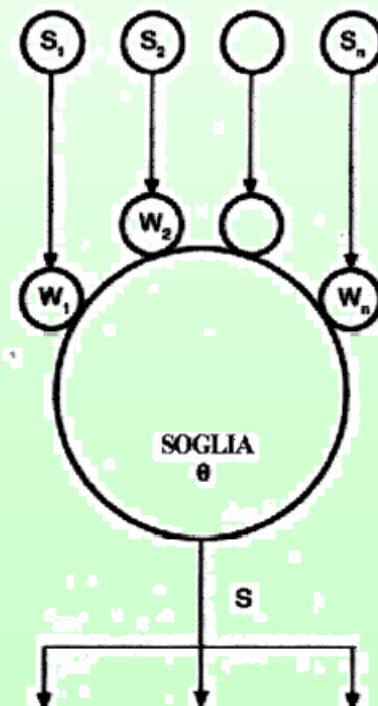
$$P_i = \sum_j (W_{ij} \cdot S_j(t) - \theta_i) \quad P = \text{potenziale di attivazione}$$

W = pesi delle connessioni; θ = soglia

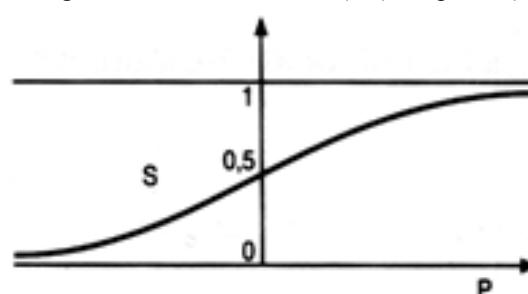
-Stato del neurone i al tempo $t+1$:

$$S_i(t+1) = F(P_i)$$

con F = funzione di trasferimento



Funzione a gradino, con valori binari (0,1) o bipolar (-1,+1)



Funzione sigmoide, con valori continui e saturazione

Leggi di apprendimento (modificano i pesi delle connessioni)

$$W_{ij}(t) \rightarrow W_{ij}(t+1) \quad \text{con} \quad W_{ij}(t+1) = W_{ij}(t) + \Delta W_{ij}$$

- **Regola Delta:** modifica i pesi in modo da far decrescere l'errore E commesso dalla rete al tempo t

$$\Delta W_{ij} = -\eta \cdot \frac{\partial E}{\partial W_{ij}} \quad (\text{apprendimento 'supervised'})$$

- **Legge di Hebb:** rinforza le connessioni tra neuroni contemporaneamente attivi

$$\Delta W_{ij} = \eta \cdot S_i(t) \cdot S_j(t) \quad (\text{apprendimento 'supervised'})$$

- **Legge di Kohonen:** ruota il vettore dei pesi W_j più vicino a quello di input X in modo da avvicinarlo ulteriormente ad X

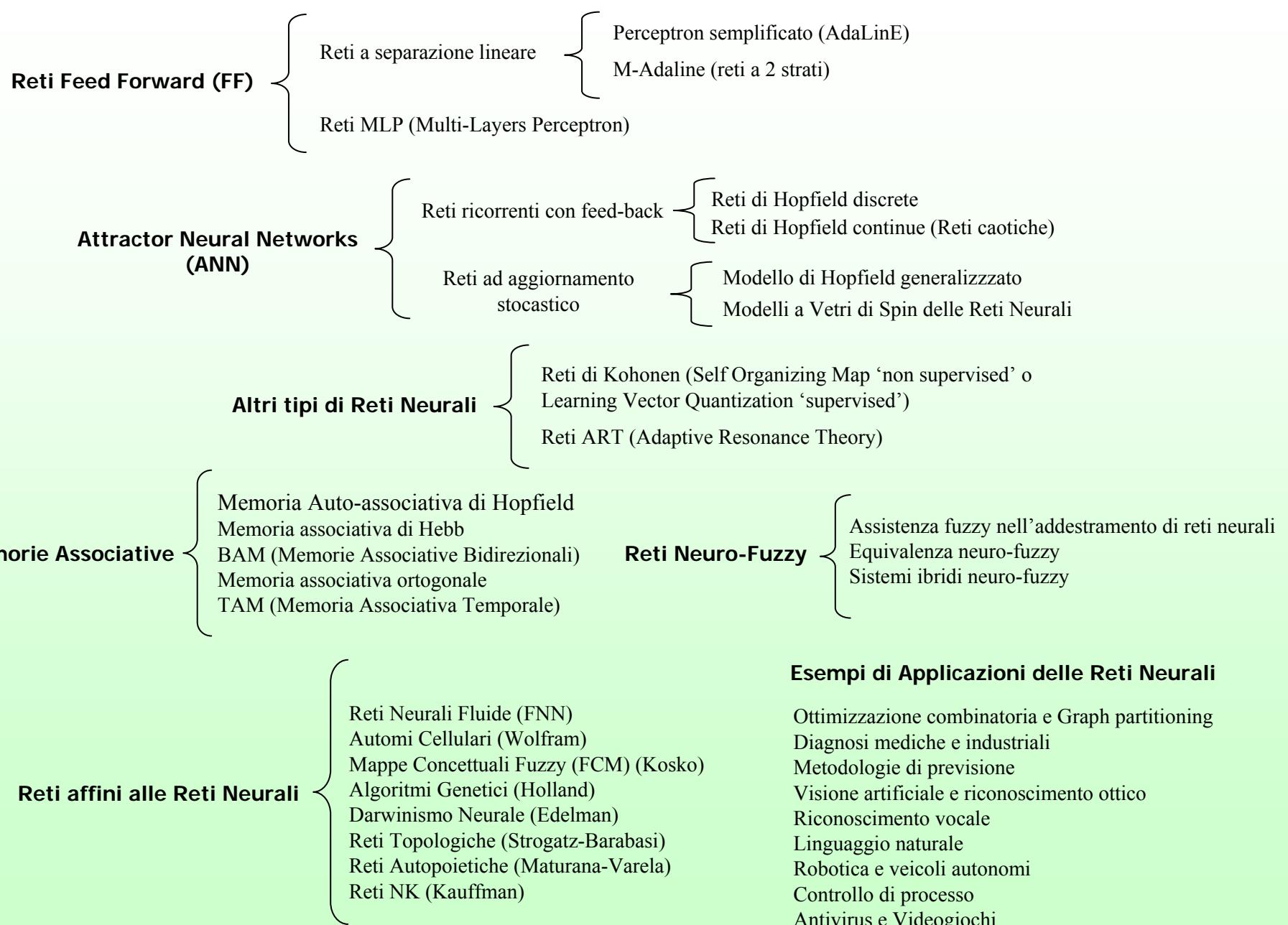
$$\Delta W_{ij} = \eta \cdot (X - W_j) \quad (\text{apprendimento 'supervised' e 'non supervised'})$$

- **Algoritmi Stocastici:** Metodo Montecarlo e Ricottura Simulata (simulated annealing)
 - **Algoritmi Evolutivi ed Algoritmi Genetici**
-

Leggi di iterazione (regolano l'ordine di attivazione dei neuroni)

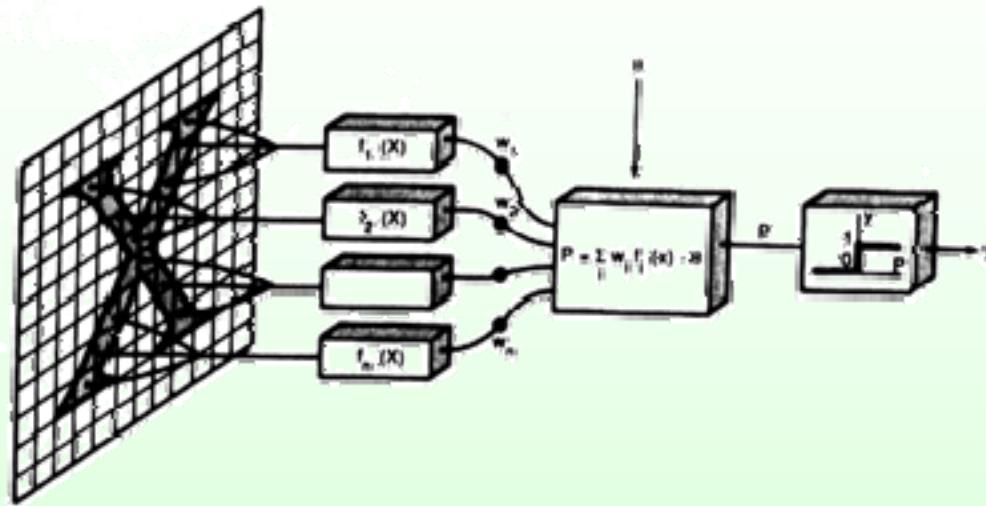
- **Parallelia** (aggiorna tutti i neuroni ad ogni passo temporale)
- **Sequenziale** (aggiorna un neurone alla volta ad ogni passo temporale scegliendolo in modo sequenziale)
- **Random** (aggiorna un neurone alla volta ad ogni passo temporale scegliendolo in modo casuale)

TIPOLOGIE DI RETI NEURALI



RETI NEURALI FEED FORWARD

Perceptron ad uno strato (Adaline) e a due strati (Multiple Adaline)



Il Perceptron Adaline (Adaptive linear element, Rosenblatt 1958) è costituito da uno strato di N neuroni binari di input X_i (eventualmente collegati ad una matrice di pixel – vedi figura – di cui calcolano una opportuna funzione $f(X)$) che alimentano, mediante dei pesi sinaptici W_i un unico neurone di output Y (combinatore lineare a soglia) secondo la tipica legge di attivazione:

$$Y(t+1) = F(P)$$

dove: $P = \sum_i (W_i \cdot X_i(t) - \theta)$ potenziale di attivazione

$F(P) = \text{sign}(P)$ funz. di trasferimento a gradino

Supponiamo di voler addestrare (in modo *supervised*) il **Perceptron** a riconoscere se un pattern di input (cioè un dato vettore \mathbf{X}) appartiene o meno ad una certa classe F . Alla presentazione di un pattern di input il neurone Y risponderà con un output pari a 1 (indicante appartenenza ad F) se $P>0$ oppure con un output pari a 0 (indicante non appartenenza ad F) se $P<0$: nel caso in cui il riconoscimento sia corretto i pesi W_i non verranno modificati, altrimenti verranno modificati con la cosiddetta **Regola Delta**:

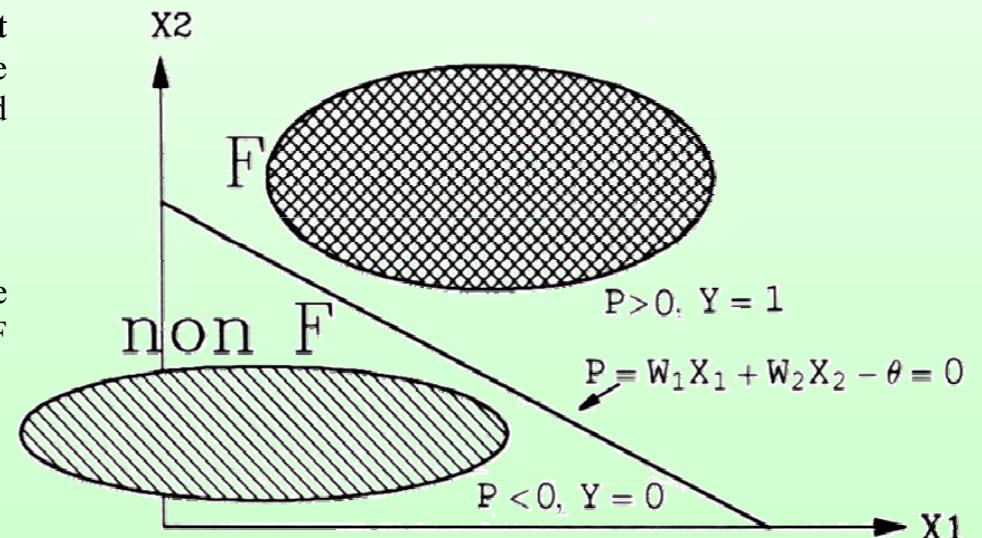
$$W_i(t+1) = W_i(t) + \Delta W_i \quad \text{con} \quad \Delta W_i = \pm \eta X_i$$

dove si prende il segno + o il segno - a seconda che occorra aumentare o diminuire il potenziale P .

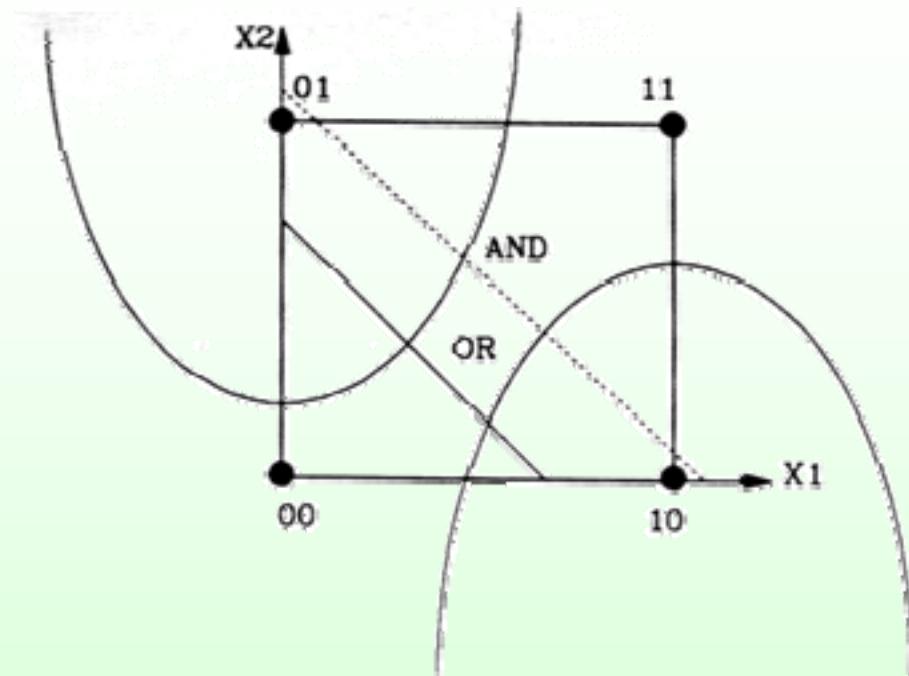
L'iterazione di questo processo conduce ad individuare il **set di pesi W_i** che consente al Perceptron di classificare correttamente ciascun pattern di input, il che equivale ad individuare l'**iper-piano** di equazione:

$$P = \mathbf{W}\mathbf{X} - \theta = 0$$

che separa (linearmente) lo spazio delle unità di input in due semi-spazi ciascuno corrispondente ad una delle due classi F e non- F (vedi a lato il caso bidimensionale).



Molto spesso però le due classi non sono linearmente separabili, come ad esempio già nel semplice caso dell'OR-esclusivo (XOR) dove, a differenza di quanto accade per l'AND e l'OR, non è possibile trovare una retta (il caso è bidimensionale, indi bastano 2 neuroni di input X_1 e X_2) che separi gli stati 10 e 01 dagli altri due.



E' questo il famoso problema della '**Separabilità lineare**' che limita le capacità del Perceptron e che non è risolvibile (per potenziali di attivazione lineari) neanche aumentando il numero di unità di output (M-Adaline).

Solo l'introduzione delle **reti multistrato MLP** (dotate cioè di uno o più strati di unità nascoste in aggiunta a quelli di input e di output) e dell'**algoritmo EBP** a retro propagazione dell'errore permise di superare questa impasse, che comunque bloccò gli sviluppi del connessionismo per più di un ventennio...

Multi Layers Perceptron: Algoritmo Error Back Propagation (EBP)

(Rumelhart, Hinton, Williams, 1985)

Consideriamo una rete a 3 strati con n neuroni di input X, q unità nascoste Z e m unità di output Y, si definisce dapprima l'errore quadratico medio E dello strato di output per ognuno dei pattern del Training Set:

$$E = \frac{1}{2} \sum_j (Y_j - D_j)^2$$

dove Y_j è l'output effettivo del neurone j-esimo e D_j l'output desiderato.

Applicando la **Regola Delta generalizzata** (o del *gradiente d'errore*):

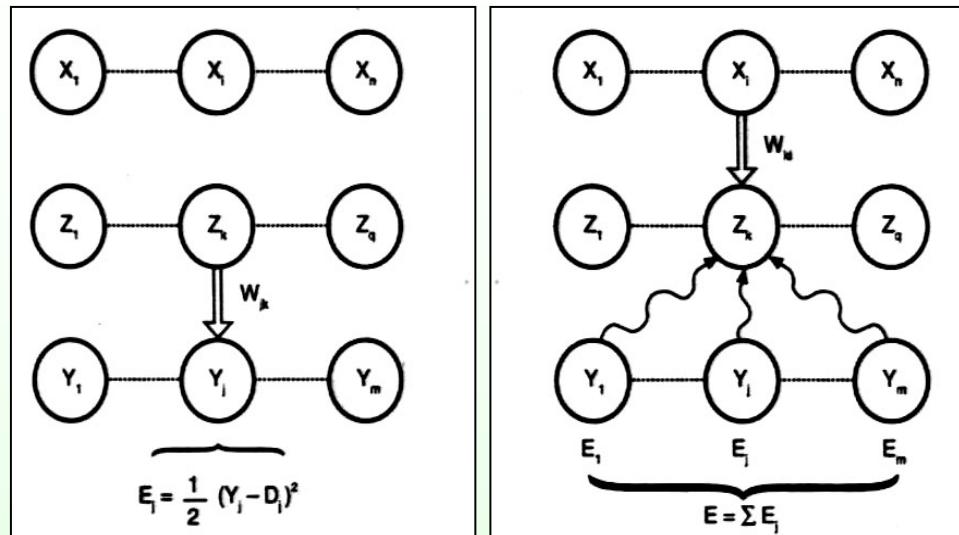
$$\Delta W_{jk} = -\eta \cdot \frac{\partial E}{\partial W_{jk}} = -\eta \cdot \frac{\partial E}{\partial Y_j} \frac{\partial Y_j}{\partial P_j} \frac{\partial P_j}{\partial W_{jk}} = -\eta (Y_j - D_j) f'(P_j) Z_k$$

vengono prima aggiornati i pesi delle connessioni tra lo strato delle unità di output e quello delle unità nascoste. Analogamente, per quanto riguarda le connessioni tra strato nascosto e strato di input si ha:

$$\Delta W_{ki} = -\eta \cdot \frac{\partial E}{\partial W_{ki}} = -\eta \cdot \frac{\partial E}{\partial Z_k} \frac{\partial Z_k}{\partial P_k} \frac{\partial P_k}{\partial W_{ki}} = -\eta \frac{\partial E}{\partial Z_k} f'(P_k) X_i$$

dove al posto degli errori noti ($Y - D$) troviamo delle derivate che occorre calcolare retro-propagando l'errore dallo strato output a ogni neurone intermedio:

$$\frac{\partial E}{\partial Z_k} = \sum_j \left(\frac{\partial E}{\partial Y_j} \frac{\partial Y_j}{\partial P_j} \frac{\partial P_j}{\partial Z_k} \right) = \sum_j (Y_j - D_j) f'(P_j) W_{jk}$$



IPER-PARABOLOIDE DELL'ERRORE

Ad ogni passo temporale (cioè ad ogni presentazione di un nuovo input) i pesi della rete vengono dunque aggiornati secondo le leggi:

$$W_{jk}(t+1) = W_{jk}(t) + \Delta W_{jk}$$

$$W_{ki}(t+1) = W_{ki}(t) + \Delta W_{ki}$$

dove, se si adotta una funzione di trasferimento
sigmoide: $Y = f(P) = 1 / (1 + e^{-kP})$ si ha:
 $f'(P) = k Y (1 - Y)$ e dunque:

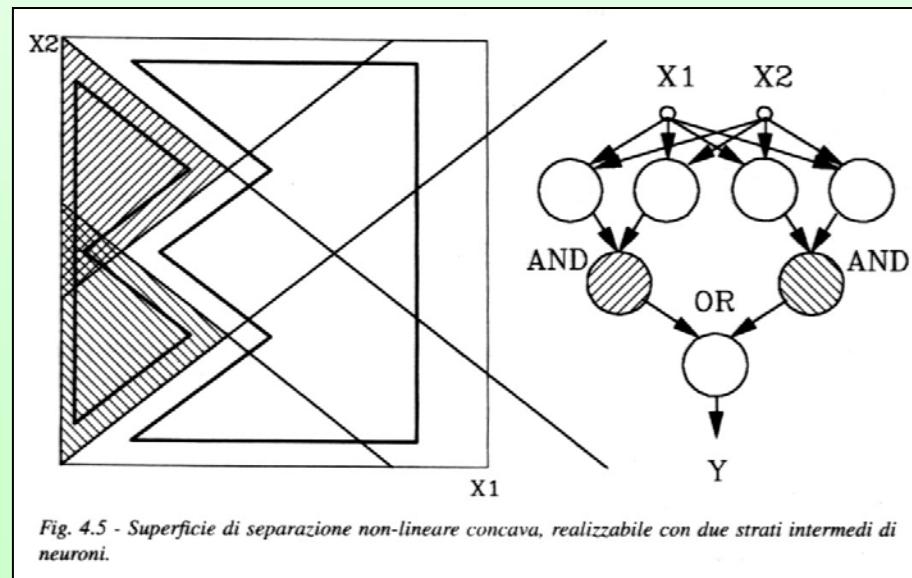
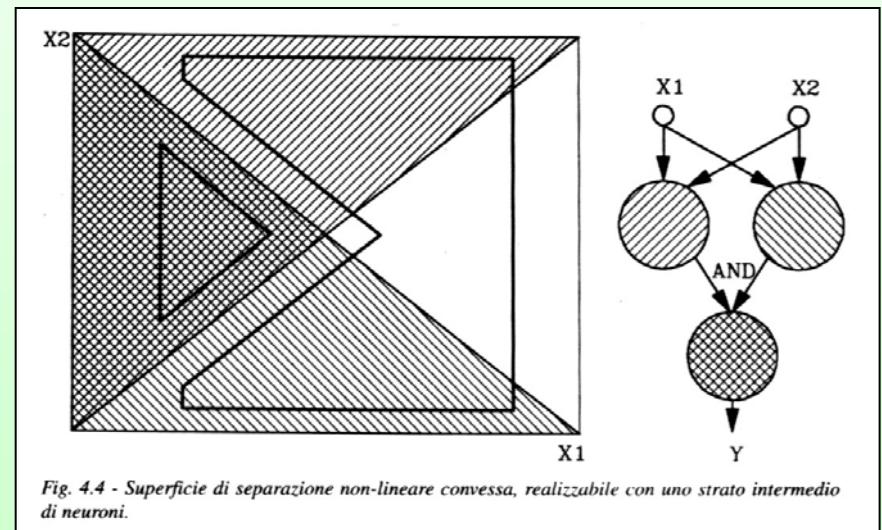
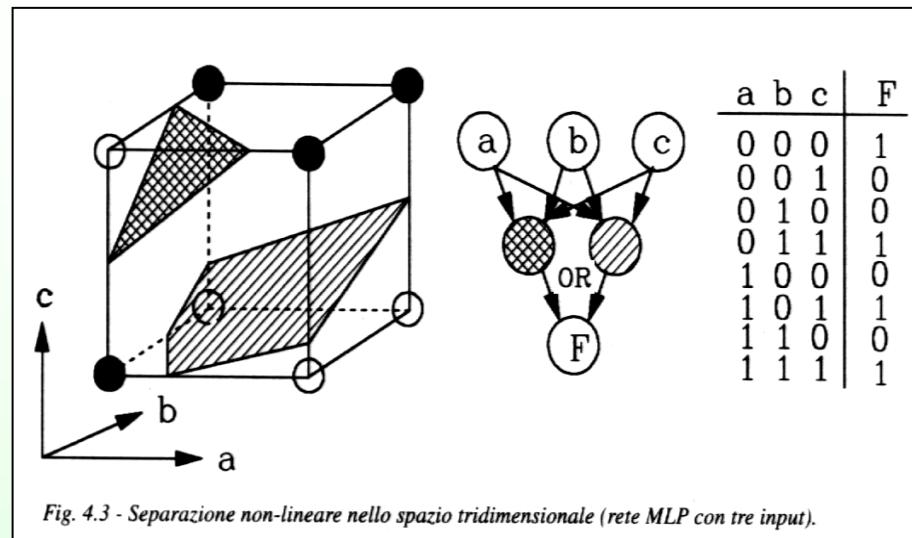
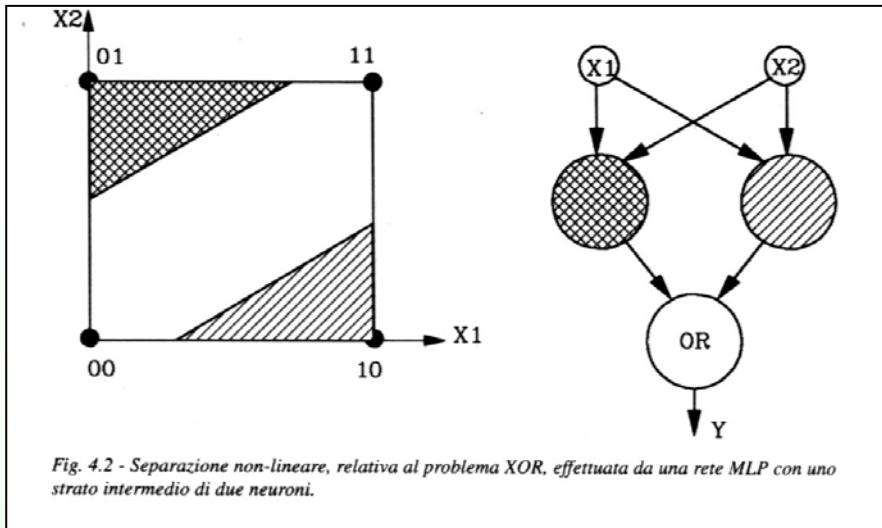
$$\Delta W_{jk} = -\eta k (Y_j - D_j) Y_j (1 - Y_j) Z_k$$

$$\Delta W_{ki} = -\eta k Z_k (1 - Z_k) (\sum_j \delta_j W_{jk}) X_i$$

$$\text{con } \delta_j = (Y_j - D_j) f'(P_j)$$



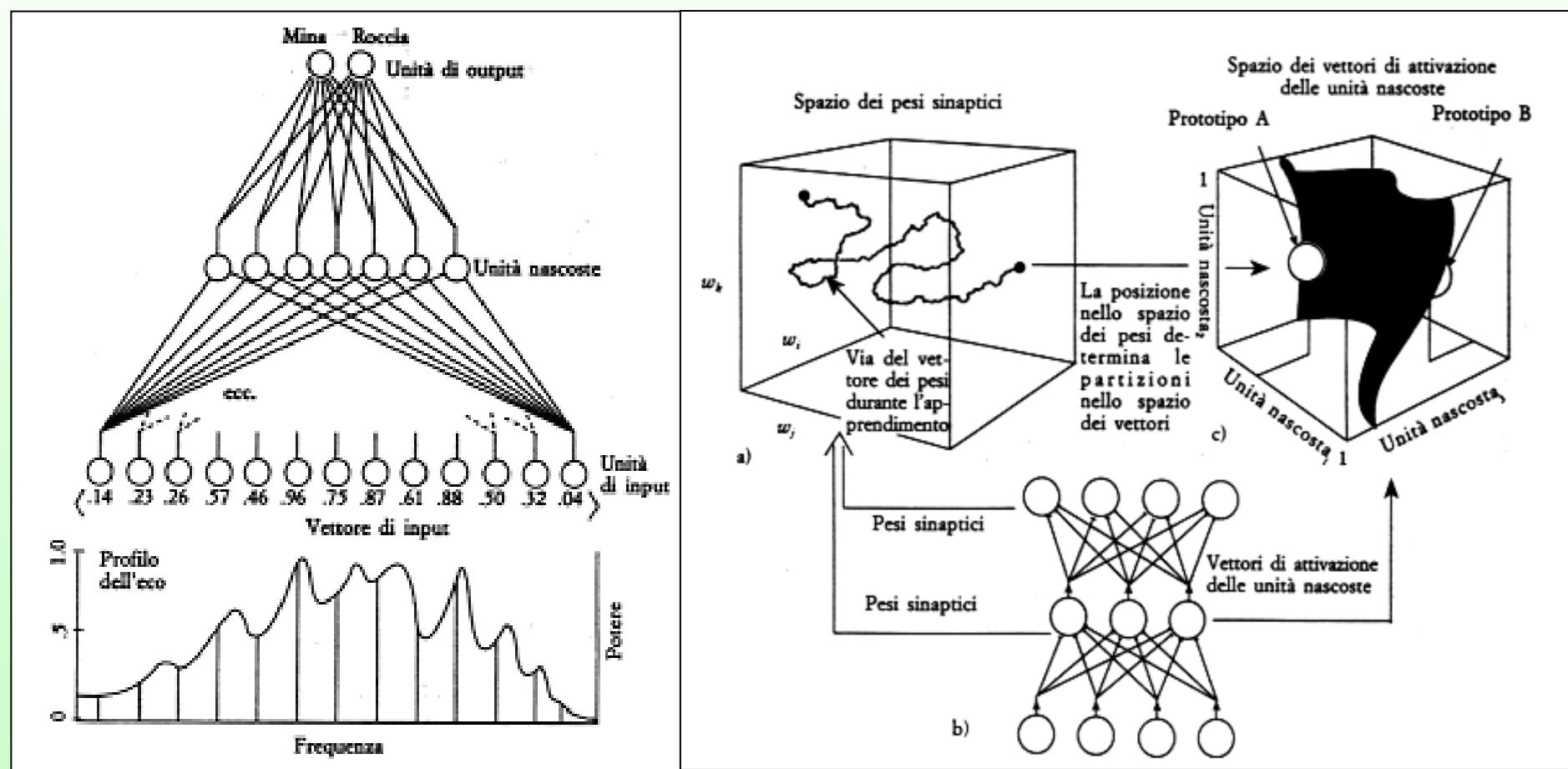
Per mezzo dell'algoritmo EPB, attraverso la retro-propagazione dell'errore, il punto rappresentativo della rete nello spazio multidimensionale dei pesi W percorre quindi una discesa lungo il gradiente di errore alla ricerca di un minimo assoluto (anche se può rimanere intrappolato in minimi relativi, per evitare i quali sono stati introdotti ulteriori accorgimenti che raffinano l'algoritmo EBP, come ad esempio l'aggiunta del termine 'momentum', l'adozione di diverse metriche dell'errore o l'utilizzo di un 'learning rate' adattativo).



RUOLO DELLE UNITÀ NASCOSTE

Nelle Reti FF ad almeno tre strati è proprio lo strato delle unità nascoste che consente di superare il problema della separabilità lineare in quanto mette la rete in grado di effettuare un partizionamento dello spazio delle unità nascoste (P.Churchland, *Modello di Attivazione del Prototipo*, 1993)

Il centroide di ogni partizione corrisponderà ad una sorta di 'prototipo' (nell'esempio qui a fianco una roccia o una mina) che la rete imparerà a riconoscere a partire anche da dati incompleti o inquinati da rumore, alla fine di un processo di apprendimento EBP che farà spostare il punto rappresentativo della rete nello spazio dei pesi sinaptici alla ricerca di un minimo dell'errore.



METODI DI OTTIMIZZAZIONE

> Metodo Montecarlo

La versione base di questo classico algoritmo ha il seguente ciclo:

1. inizializzazione casuale dei pesi sinattici e calcolo di una metrica d'errore E , funzione dei pesi (solitamente l'errore quadratico medio sul training set);
2. variazione casuale dei pesi sinattici, calcolo di E e di $\Delta E = E_{\text{nuovo}} - E_{\text{vecchio}}$:
 - 2.1. se l'errore diminuisce, cioè $\Delta E \leq 0$, accettazione dei nuovi pesi e ritorno al punto 2;
 - 2.2. se l'errore aumenta, cioè $\Delta E > 0$, rifiuto dei nuovi pesi, riadozione dei pesi precedenti e ritorno al punto 2.

Il ciclo continua per un numero prefissato di volte o finché l'errore non scende sotto un valore minimo E_{\min} prefissato. L'algoritmo è di tipo *greedy* (= avaro) perché accetta solo le variazioni di peso che diminuiscono l'errore (aumentano il guadagno) e non rinuncia al capitale (errore minimo) via via accumulato. Nella metafora della pallina pesante nel "paesaggio" dell'errore, l'algoritmo equivale a una sequenza di scosse casuali della superficie d'errore, tornando alla situazione precedente se la pallina scivola in una valle meno profonda. L'esplorazione è del tutto casuale e porta difficilmente alla scoperta del minimo assoluto.

> Ricottura simulata (Simulated Annealing)

L'algoritmo Montecarlo può essere migliorato notevolmente modificandone il ciclo nel seguente modo:

1. inizializzazione casuale dei pesi sinattici e calcolo della solita metrica d'errore E ;
2. variazione casuale dei pesi sinattici, calcolo di E e di $\Delta E = E_{\text{nuovo}} - E_{\text{vecchio}}$:
 - 2.1) se l'errore diminuisce, cioè $\Delta E \leq 0$, accettazione dei nuovi pesi e ritorno al punto 2;
 - 2.2. se l'errore aumenta, cioè $\Delta E > 0$, accettazione dei nuovi pesi solo con probabilità di Boltzmann $P(\Delta E)$, sotto specificata, e ritorno al punto 2.

Il ciclo continua per un numero prefissato di volte o finché l'errore non scende sotto un valore minimo prefissato E_{\min} . L'algoritmo non è di tipo *greedy* perché accetta anche le variazioni di peso che aumentano l'errore, sia pure nell'am-

bito di una legge di probabilità.. La probabilità P deve soddisfare la distribuzione di Boltzmann, cioè deve essere proporzionale a:

$$e^{-\Delta E/T} = e^{-(E_2 - E_1)/T} = \frac{e^{-E_2/T}}{e^{-E_1/T}} = \frac{P(E_1)}{P(E_2)}$$

dove T è un parametro denominato *temperatura*, il cui valore viene gradualmente abbassato nel tempo, in un processo analogo alla ricottura (da cui la denominazione di *ricottura simulata*). La accettazione e il rifiuto del nuovo stato di energia corrispondono rispettivamente a $P(\Delta E) > \lambda$ e $P(\Delta E) < \lambda$, dove λ è un valore fornito da un generatore di numeri casuali a distribuzione uniforme nell'intervallo (0-1). Nelle fasi iniziali del processo, T è molto grande e quindi:

$$\begin{aligned} e^{-\Delta E/T} &\rightarrow 1 \\ P(E_2) &= P(E_1) \end{aligned}$$

per cui la probabilità di scegliere uno stato della rete con $\Delta E > 0$, $E_2 > E_1$ è elevata; nelle fasi successive, T diventa sempre più piccola e quindi:

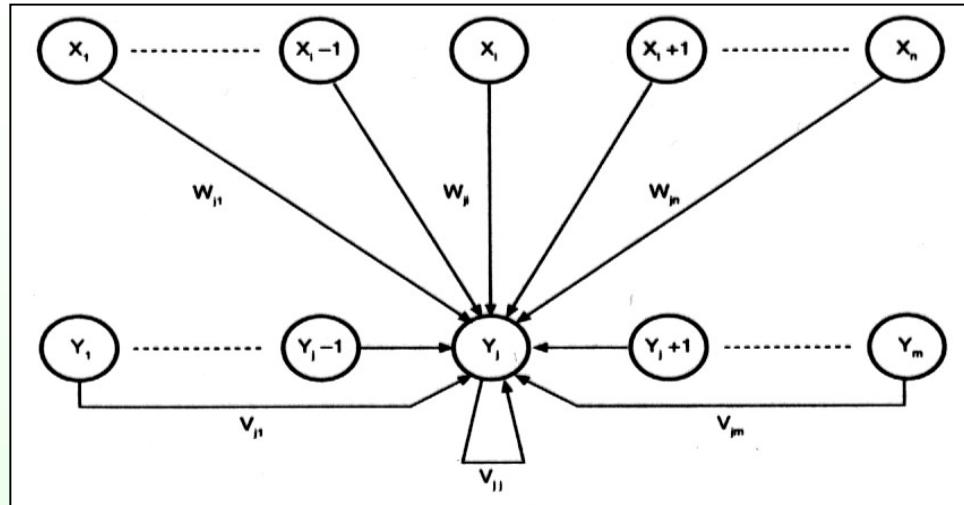
$$\begin{aligned} e^{-\Delta E/T} &\rightarrow 0 \\ P(E_2) &\rightarrow 0 \text{ per } E_2 > E_1 \end{aligned}$$

per cui gli stati con incremento dell'errore tendono sempre più ad essere rifiutati.

La ricottura simulata consente di esplorare con maggiore efficacia la superficie d'errore, non impegnandosi ad accettare preconcetti minimi apparentemente buoni, ma che in realtà predisponevano il raggiungimento futuro di minimi decisamente migliori.

Il processo è "prodigo" nelle fasi iniziali (come si è prodighi in "gioventù", alla ricerca di nuove amicizie e di nuove esperienze), per diventare poi sempre più avaro (nella "maturità" si cerca di migliorare sede la situazione in cui ci si trova). Nella metafora della pallina pesante che scende lungo la superficie d'errore alla ricerca di una valle molto profonda, il processo della ricottura simulata equivale a scorrere orizzontalmente quella superficie, prima con grande forza e poi con sempre minor vigore. La pallina allora tende a fuoriuscire dai minimi relativi iniziali poco interessanti, esplorando una grande estensione della superficie d'errore, prima di attestarsi in minimi che ne valgono la pena.

RETI DI KOHONEN (SOM – Self Organizing Map)



Sono reticolari bidimensionali di m neuroni Y_j , ciascuno connesso ad n input X_i tramite i pesi sinaptici W_{ij} variabili. In pratica realizzano una proiezione bidimensionale deformata di input multidimensionali, analoga a quella che avviene neurobiologicamente nelle mappe acustiche o in quelle somato-sensoriali.

Nelle SOM si hanno delle **connessioni intra-strato** con pesi fissati V_{jk} distribuiti a "cappello messicano", i quali individuano – per ciascun neurone - un 'vicinato' circolare di neuroni che lo eccitano e una corona circolare più esterna di neuroni che lo inibiscono.

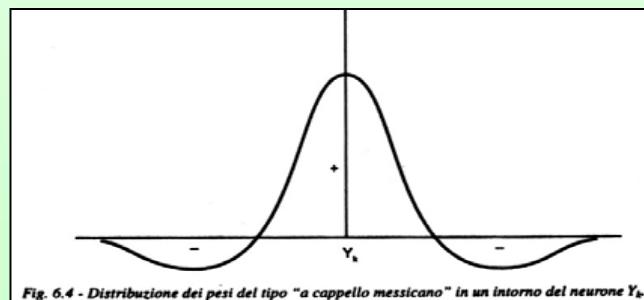


Fig. 6.4 - Distribuzione dei pesi del tipo "a cappello messicano" in un intorno del neurone Y_i

In alternativa si possono simulare le connessioni laterali intra-strato di ciascun neurone Y_j con un **vicinato topologico** di neuroni contemporaneamente attivi variabile nel tempo, i quali formano una 'bolla di attivazione' centrata sul neurone Y_j .

DINAMICA DELLA RETE DI KOHONEN (Unsupervised and competitive learning)

- Si inizializza la rete con $m \times n$ pesi sinaptici W_{ji} , costruendo quindi una matrice \mathbf{W} costituita da m vettori riga $\mathbf{W}_j = (W_{j1}, \dots, W_{jn})$ corrispondenti a ciascuno degli m neuroni della rete;
 - Si presenta in input il generico vettore colonna $\mathbf{X} = (X_1, \dots, X_n)$ e si determina quel neurone Y_j il cui vettore dei pesi \mathbf{W}_j rende massimo il prodotto scalare:
- $$\mathbf{W}_j * \mathbf{X} = \sum_i (W_{ji} X_i)$$
- Si verifica quindi una competizione tra tutti i neuroni vinta da uno solo di essi (*winner-takes-all*);
 - Il neurone vincente Y_j viene 'premiato', insieme a tutti quelli del suo vicinato, attraverso la seguente legge di aggiornamento del vettore dei pesi \mathbf{W}_j :
- $$\mathbf{W}_j(t+1) = \mathbf{W}_j(t) + \Delta \mathbf{W}_j \quad (\text{con } \Delta \mathbf{W}_j = \eta (\mathbf{X} - \mathbf{W}_j) \text{ ed } \eta: \text{tasso di apprendimento})$$
- che rappresenta una rotazione di \mathbf{W}_j verso \mathbf{X} (entrambi normalizzati) su una ipersfera unitaria.
- Il ciclo ricomincia dal punto 2.

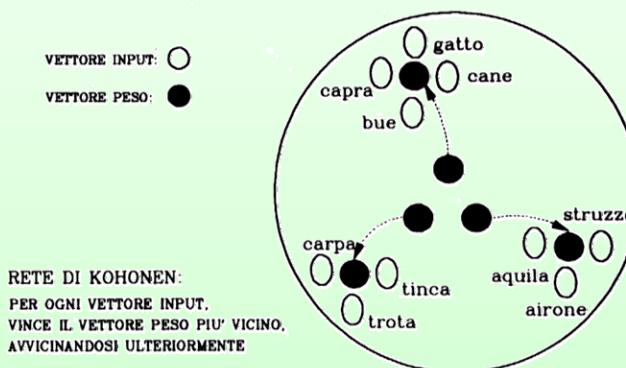
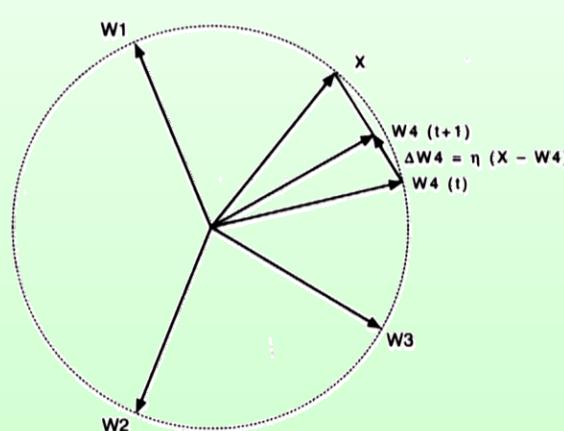


Fig. 6.6 - Competitive Learning: in un mondo n -dimensionale, i punti da classificare (bianchi) sono distribuiti, in k classi distinte, sulla superficie di una iper-sfera. I neuroni della rete con n input hanno inizialmente pesi sinaptici W (punti neri) distribuiti in modo random. L'apprendimento consiste nell'avvicinamento progressivo dei punti W ai centroidi delle classi.

Se η viene fatto diminuire nel tempo assieme alla dimensione del 'vicinato', si verifica un processo di **auto-organizzazione** nel corso del quale la rete modifica progressivamente i pesi W_{ij} in modo che la loro configurazione rifletta la distribuzione degli input \mathbf{X} . In altre parole, come già detto, i vettori \mathbf{X} vengono **proiettati** dal loro spazio n -dimensionale nello spazio bi-dimensionale della rete. Supponendo che esistano **K classi** dell'input, i relativi vettori \mathbf{X} si addenseranno in k aree diverse dell'ipersfera: nel corso di un numero molto alto di iterazioni, i vettori peso migrano progressivamente verso i **centroidi** di queste aree. Ad apprendimento concluso, per ogni \mathbf{X} presentato in input sarà attivato un vettore peso corrispondente alla classe di quell'input.

RETI DI HOPFIELD DISCRETE

Le **reti di Hopfield** costituiscono un interessante tipo di sistemi dinamici non lineari. Nella loro versione **discreta** sono costituite da N neuroni con stati bipolarì $S_i(t)=[-1,+1]$ oppure binari $S_i(t)=[0,1]$, con funzione di trasferimento a gradino e connessioni simmetriche tra ogni neurone e tutti gli altri (topologia “fully connected” senza autoconnessioni).

Si hanno dunque $N(N-1)/2$ pesi sinaptici diversi $J_{ij}=J_{ji}$ ($J_{ii}=0$) e per ogni neurone si calcola il seguente **potenziale di attivazione** P_i (corrispondente al potenziale post-sinaptico PSP dei neuroni reali) :

$$P_i(t) = \sum_j J_{ij} \cdot S_j(t) - \vartheta_i \quad \text{con soglia caratteristica } \vartheta_i$$

Poiché la **legge di attivazione** è la funzione a gradino:

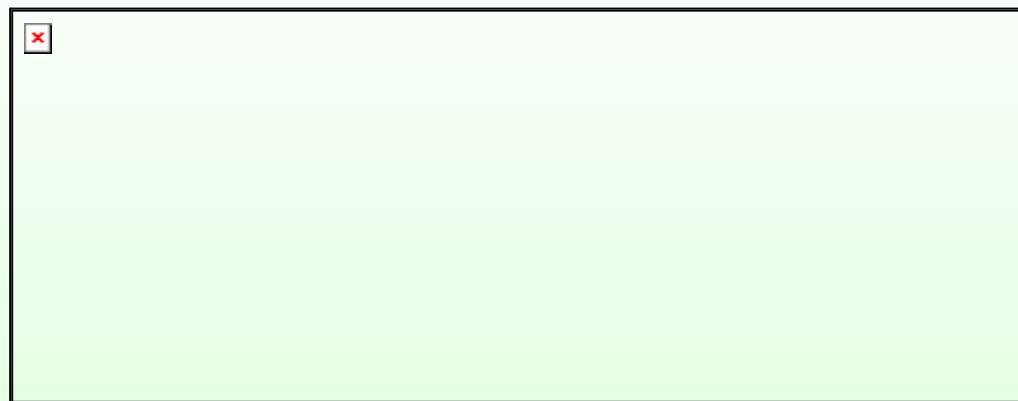
$$S_i(t+1) = \text{sign}(P_i(t))$$

si ha, per neuroni bipolarì,:

$$S_i(t+1) = +1 \quad \text{se } P_i(t) > 0$$

$$S_i(t+1) = -1 \quad \text{se } P_i(t) < 0$$

Se invece $P_i(t)=0$, si mantiene lo stato precedente.



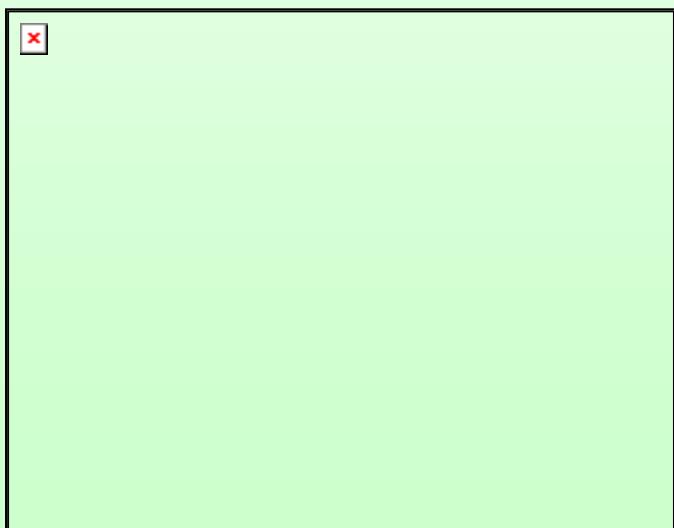
Ogni istante lo stato delle rete è rappresentato dal vertice di un **iper cubo** a N dimensioni (nel caso $N=3$ di un cubo di coordinate variabili da $(-1, -1, -1)$ a $(+1, +1, +1)$) e l'evoluzione della rete può essere vista come una sequenza di transizioni da uno dei 2^N vertici ad un altro (con attrattori a punto fisso o a ciclo limite, ma non caotici).

Si puo' verificare che la legge di attivazione delle reti di Hopfield produce una diminuzione monotona nel tempo di una **funzione energia** E (funzione di Lyapunov) cosi' definita:

$$E = -\frac{1}{2} \sum_{ij} J_{ij} S_i S_j - \sum_i \vartheta_i S_i = -\frac{1}{2} \sum_i P_i S_i$$

Infatti risulta: $\Delta E = -P_i \Delta S_i$

dove ΔS_i o è nullo o ha lo stesso segno di P_i

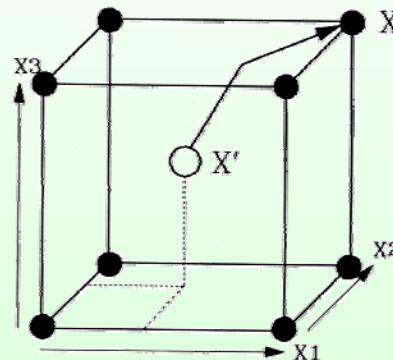


Se ora definiamo la **Capacita' di Memoria** $\alpha = p/N$ della rete come il numero di attrattori stabili a punto fisso della dinamica (p) diviso il numero di neuroni N, la rete di Hopfield puo' essere usata come **Memoria Autossociativa** se si assegnano i pesi sinaptici J_{ij} attraverso la seguente generalizzazione della **legge di Hebb** (con apprendimento statico):

$$\Delta J_{ij} = \eta \cdot S_i(t) \cdot S_j(t) \quad \forall k = 1, \dots, p \quad \longrightarrow \quad J_{ij} = \frac{1}{N} \sum_{k=1}^p F_{ki} F_{kj}$$

dove le $F_k = (F_{k1}, F_{k2}, F_{k3}, \dots, F_{kN})$ sono p vettori che rappresentano particolari configurazioni di stati neuronali, dette **patterns o forme**, che la rete deve memorizzare.

A questo punto la dinamica della rete parte da una **versione "inquinata"** o incompleta $F' = X'$ di un certo pattern $F = X$, ed evolve nel tempo verso il vertice dell'iper cubo corrispondente ad X che, per come sono stati costruiti i pesi J_{ij} , rappresenta un attrattore stabile della dinamica:



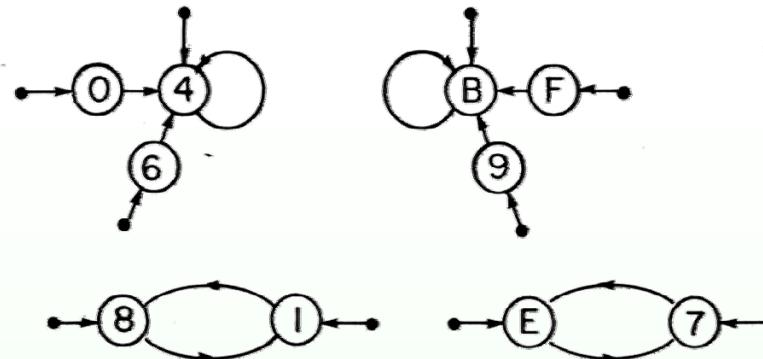
In altre parole, spostando l'attenzione sul cosiddetto **'Paesaggio dell'Energia'**, accade che partendo da un punto corrispondente al pattern incompleto F' situato dentro il bacino di attrazione di F , il sistema – la cui dinamica tende a minimizzarne l'Energia – si assesterà sul fondo della 'valle' corrispondente all'attrattore F (che è un **minimo relativo** di E).

Questo processo di minimizzazione spontanea dell'Energia rende la rete di Hopfield molto utile anche per la soluzione di problemi di **ottimizzazione combinatoria** (tipo Problema del Commesso Viaggiatore), dove all'Energia si sostituisce una **Funzione Costo** da minimizzare.

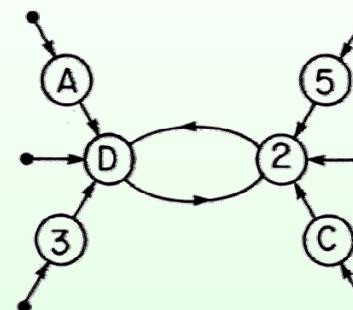
Paesaggio dell'Energia $E = -\frac{1}{2} [\sum_{ij} J_{ij} S_i S_j + \sum_i g_i S_i]$

Bacino di attrazione di F

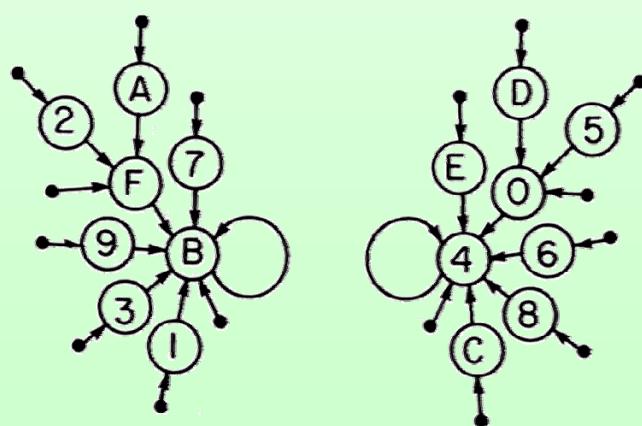
DINAMICA SINCRONA o PARALLELA
 (aggiornamento simultaneo
 dello stato di tutti i neuroni)



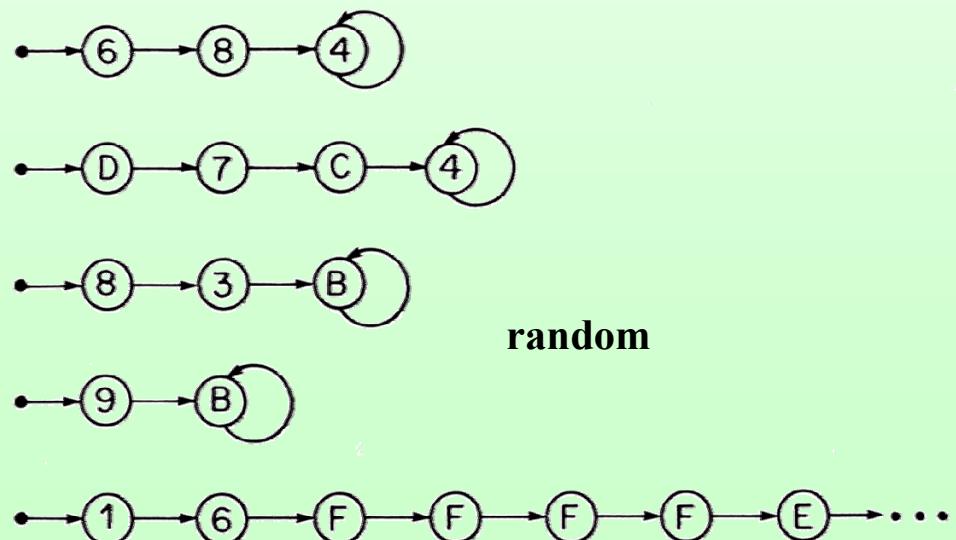
Rete di 4 neuroni
 con matrice dei pesi sinaptici J_{ij}
 simmetrica (16 stati possibili,
 contrassegnati con cifre esadecimali)



DINAMICA ASINCRONA (aggiornamento sequenziale o random dello stato di tutti i neuroni)



sequenziale



RETI DI HOPFIELD CONTINUE E RETI CAOTICHE

Nelle reti di Hopfield **continue** lo stato $S_i(t)$ di ognuno degli N neuroni della rete può assumere valori continui nell'intervallo $[0,1]$ ed evolve nel tempo con la seguente legge di attivazione:

$$S_i(t+1) = k S_i(t) + \alpha \sum_j (J_{ij} X_j(t))$$

dove k e α sono costanti positive minori di 1, J_{ij} sono i pesi sinaptici (con $J_{ii} = 0$ e $J_{ij} = J_{ji}$) e $X_j(t)$ è l'output **sigmoidale** del neurone j -esimo:

$$X_j = \frac{1}{1 + e^{-S_j(t)/T}}$$

con T = temperatura della sigmoide.

L'Energia è ancora (considerando la soglia zero):

$$E = - \sum_{ij} J_{ij} S_i S_j \quad \text{con} \quad \frac{\partial E}{\partial S_i} = - \sum_j J_{ij} S_j$$

ma non diminuisce più in modo sistematico con l'evoluzione della rete. Infatti lo stato della rete non corrisponde più ad un vertice dell'iper cubo booleano (come nelle reti di Hopfield discrete) ma è adesso un punto **interno** a tale ipercubo. La rete evolve dunque lungo una traiettoria continua nello spazio N -dimensionale interno all' ipercubo, generalmente verso attrattori stabili.

Reti Caotiche

Se nella legge di attivazione poniamo $J_{ii} = -J_i \neq 0$

ossia permettiamo ai neuroni della rete di avere dei collegamenti inibitori con se stessi (**pesi auto-sinaptici**) ecco che compaiono degli **attrattori caotici** (frattali): la rete diventa una **CNN (Chaotic Neural Network)** ed ha il pregio di visitare, nella sua evoluzione, un enorme numero di stati alla ricerca di quelli a minima energia.

Per tali motivi si è pensato di modificare la rete al fine di limitare il regime caotico alle sole fasi iniziali di ricerca per poi passare ad un regime stabile che consolidi i risultati intanto conseguiti. Si è dunque introdotto un **peso auto-sinaptico inibitorio** variabile nel tempo secondo la legge:

$$J_i(t) = J_i(0) e^{-\beta t}$$

dove β è un parametro di smorzamento, positivo e minore di 1.

In questa nuova rete (**TCNN: Transiente Chaotic Neural Network**) il peso diminuisce col tempo fino ad annullarsi dopo un tempo sufficientemente lungo e svolge la stessa funzione della temperatura nella ricottura simulata. La rete si comporta come una rete CNN nelle fasi iniziali e come una rete di Hopfield (con attrattori stabili) nelle fasi finali.

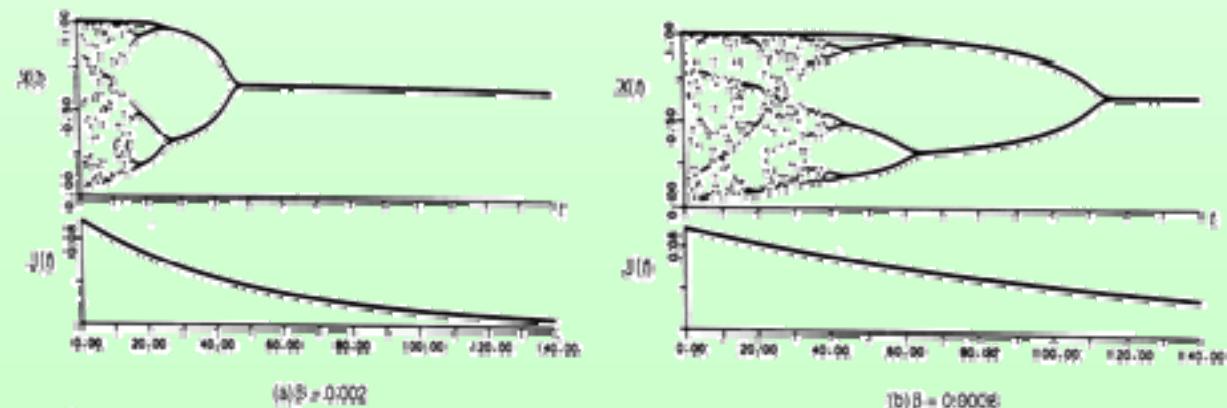


Fig. 5.8 - Evoluzione nel tempo di un singolo neurone caotico per valori diversi di β : (a) $\beta = 0.002$; (b) $\beta = 0.0008$.

Dinamica in presenza di Rumore termico per Reti di Hopfield e Reticoli di Ising

La transizione sinaptica, nei sistemi biologici, è un processo affetto da rumore e il potenziale post-sinaptico (PSP) non è determinabile con precisione. Conviene dunque introdurre nelle reti di Hopfield un **rumore termico** che trasformi il potenziale di attivazione sul neurone i -esimo in una **variabile gaussiana**:

$$\text{Prob}(P_i = P) = \frac{1}{\sqrt{2\pi}\delta} \exp\left[-\frac{(P - \bar{P}_i)^2}{2\delta^2}\right] \quad \text{dove } \bar{P}_i = \sum_j J_{ij} S_j - \vartheta_i \text{ è il PSP delle reti di Hopfield deterministiche.}$$

Dunque:

$$\text{Prob}(S_i = 1) = \int_0^\infty dP \text{Prob}(P_i = P) = \frac{1}{2} \left[1 + \text{erf}\left(\frac{\bar{P}_i}{\delta\sqrt{2}}\right) \right]$$

$$\text{con } \text{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x dt \exp(-t^2)$$

$$\text{Inoltre: } \text{Prob}(S_i = -1) = 1 - \text{Prob}(S_i = 1) =$$

$$\frac{1}{2} \left[1 - \text{erf}\left(\frac{\bar{P}_i}{\delta\sqrt{2}}\right) \right] \quad \text{e dunque:}$$

$$\text{Prob}(S_i) = \frac{1}{2} \left[1 + \text{erf}\left(\frac{P_i S_i}{\delta\sqrt{2}}\right) \right] \cong \frac{\exp(-\beta P_i S_i)}{\exp(-\beta P_i) + \exp(\beta P_i)} = \frac{1}{2} [1 + \tanh(-\beta P_i S_i)]$$

dove il fattore $\beta^{-1} = 2\sqrt{2}\delta$ indica il livello di rumore termico (per $\beta \rightarrow \infty$ si ritrova $S_i = \text{sign}(P_i)$)

Si noti che l'espressione della $\text{Prob}(S_i)$ è identica a quella che descrive la dinamica 'Single Spin Flip' di un **reticolo di spin di Ising** interagente con un 'bagno termico' alla temperatura

$T = \beta^{-1}$ (dinamica di Glauber con $k = 1$):

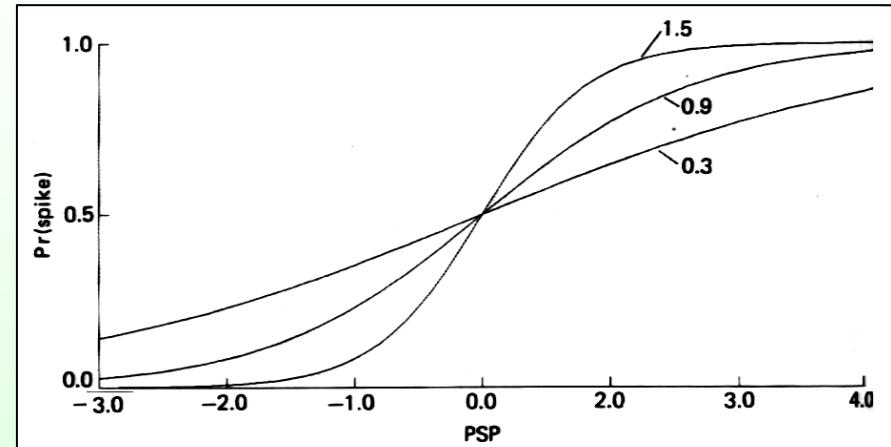


Figure 2.3: The firing probability as function of the local field (PSP) for several values of β .

$$\text{Prob}(S_i) = \frac{\exp(-\beta h_i S_i)}{\exp(-\beta h_i) + \exp(\beta h_i)} \quad \text{con} \quad h_i = \sum_j J_{ij} S_j + h^e \quad \text{e} \quad J_{ij} = J = 1$$

D'ora in avanti, in analogia con i sistemi magnetici di spin, useremo la notazione h_i per indicare il potenziale post-sinaptico P_i

Dinamica parallela (sincrona)

Gli N neuroni o spin vengono aggiornati simultaneamente, secondo dei passi temporali discreti.

La probabilità di transizione da uno stato J del sistema, specificato al tempo $(t - 1)$, ad uno stato I al tempo t, è dato dal prodotto delle probabilità di transizione dei singoli neuroni o spin:

$$W(I/J) = \prod_{i=1}^N \Pr_{ob}(S_i^I) = \frac{\exp(-\beta h_i^J S_i^I)}{\exp(-\beta h_i^J) + \exp(\beta h_i^J)}$$

e descrive una catena di Markov.

Dinamica sequenziale (asincrona)

Gli N neuroni o spin vengono aggiornati uno alla volta, in un ordine prestabilito o in modo random. Questo tipo di dinamica implica una traiettoria che **connette solo vertici adiacenti dell'ipercubo N dimensionale** che rappresenta tutti i 2^N possibili stati della rete.

Dunque la probabilità di transizione $W(I/J)$ è nulla se l'attività dello stato I differisce da quella dello stato J per più di un neurone (o se gli stati differiscono per più di uno spin).

In caso contrario la probabilità di transizione è data dal **prodotto di due termini**:

-la probabilità che il neurone (spin) k da aggiornare segua nella sequenza il neurone i che era stato aggiornato al passo temporale precedente;

-la probabilità per il singolo neurone o spin di cambiare il proprio stato.

(se la scelta dei neuroni è random si ha anche qui un processo markoviano).

-Dunque, se denotiamo con $\rho_J(t)$ la probabilità che la rete (o il sistema di spin) si trovi nello stato J al tempo t , avremo che la probabilità che la rete si trovi nello stato I al tempo $t + 1$ sarà data da:

$$\rho_I(t+1) = \sum_J W(I/J) \rho_J(t)$$

dove la somma su J è estesa ai 2^N possibili stati del sistema (o vertici dell'ipercubo N dimensionale)...

A questo punto è sufficiente che la probabilità di transizione $W(I/J)$ soddisfi la condizione cosiddetta di '**Bilancio Dettagliato**': $W(I/J)F(J) = W(J/I)F(I)$ dove F è una funzione di un singolo stato del sistema, affinché per $t \rightarrow \infty$ si abbia $\rho_I(t) \rightarrow F(I)$ cioè si ottenga una distribuzione di equilibrio.

Ed è facile verificare che questa condizione è soddisfatta, nel nostro caso, dalla **distribuzione canonica di Boltzmann-Gibbs**.

Infatti:

$$\frac{W(I/J)}{W(J/I)} = \frac{\Pr_{ob}(-S_i)}{\Pr_{ob}(S_i)} = \frac{\exp(-\beta h_i S_i)}{\exp(\beta h_i S_i)} = \exp(-2\beta h_i S_i)$$

Ricordando la definizione di Energia (che è uguale per le Reti di Hopfield e i reticolati di Ising se $J_{ij}=J$, e ponendo $h^e=0$):

$$E = -\frac{1}{2} \sum_{ij} J_{ij} S_i S_j = -\frac{1}{2} \sum_i h_i S_i \pm$$

risulta che quest'ultimo termine rappresenta nient'altro che la differenza in energia tra le due configurazioni I e J , quindi avremo:

$$W(I/J)\exp[-E(J)/kT] = W(J/I)\exp[-E(I)/kT] \quad \text{da cui segue: } \rho_I(t) \xrightarrow{t \rightarrow \infty} \rho_{BG} \approx C \cdot \exp[-E(I)/kT]$$

- Possiamo dunque applicare l'**ipotesi ergodica** e ricavare i valori medi all'equilibrio termico delle grandezze che ci interessano attraverso la classica media d'ensemble canonico:

$$\langle O \rangle = \frac{\sum_S^{2^N} O\{S\} \exp(-E\{S\}/kT)}{\sum_S^{2^N} \exp(-E\{S\}/kT)}$$

Nel caso di un **sistema di spin di Ising** la grandezza che interessa calcolare è la **magnetizzazione media per spin**, che gioca il ruolo di parametro d'ordine:

$$m \{ S \} = \frac{M \{ S \}}{N} = \frac{1}{N} \sum_{i=1}^N S_i$$

Per avere valori non nulli di m (cioè la rottura dell'ergodicità) occorre considerare un reticolo di spin ‘**fully connected**’, per il quale l’interazione J è rimpiazzata da J/N (dovendo il campo h essere una grandezza intensiva). Si ha dunque:

$$h_i = \frac{J}{N} \sum_{j, j \neq i}^N S_j \rightarrow \frac{J}{N} M \{ S \} \equiv J m \{ S \}$$

$$E \{ S \} = -\frac{J}{2N} \sum_{ij, j \neq i} S_i S_j \rightarrow -\frac{J}{2N} M^2 \{ S \} \equiv -\frac{1}{2} N J m^2 \{ S \}$$

ed è possibile dimostrare che, per una dinamica di Glauber e nel limite termodinamico, la media temporale della magnetizzazione per spin:

$$\langle m \rangle (t) = \sum_{\{S\}} m \{ S \} \rho(\{ S \}, t)$$

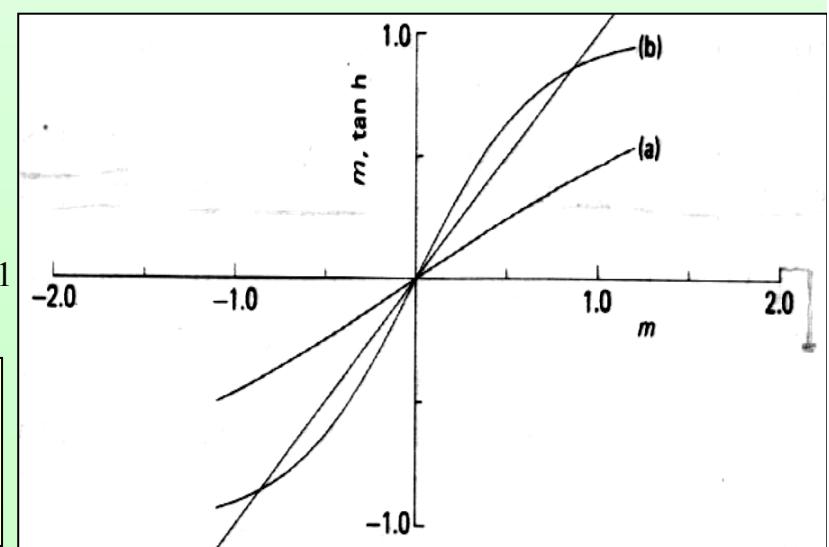
segue la legge: $\Gamma \frac{d \langle m \rangle (t)}{dt} = -\langle m \rangle (t) + \tanh\left(\frac{J \langle m \rangle (t)}{T}\right)$ (dove Γ è una costante del ciclo temporale),

le cui soluzioni all’equilibrio sono date dalla equazione autoconsistente:

$$\langle m \rangle = \tanh\left(\frac{J \langle m \rangle}{T}\right) \quad \text{che si risolve graficamente:}$$

$\left\{ \begin{array}{l} T > J \quad (\text{a}) \text{ Fase ad alta Temperatura: 1 soluz. } \langle m \rangle = 0 \\ T < J \quad (\text{b}) \text{ Fase a bassa Temperatura: 3 soluz. } \langle m \rangle = 0, \langle m \rangle = +1, \langle m \rangle = -1 \end{array} \right.$

Per $T=J$ si ha dunque una transizione di fase con conseguente rottura dell’ergodicità: la scelta del particolare attrattore ($m=1$ o $m=-1$, poiché $m=0$ è instabile) dipende dinamicamente dalle condizioni iniziali (symmetry breaking).



Lo stesso risultato si poteva ricavare cercando i minimi dell'**Energia libera** per spin $f(h, T)$, che in presenza di rumore termico sostituisce l'energia nel ruolo di funzione di Lyapunov.

Scrivendo la funzione di partizione Z (dove per completezza abbiamo reintrodotto il campo esterno $h^e = h$):

$$Z(h, T) = \sum_{\{S\}} \exp \left[\frac{J}{2NT} \left(\sum_i S_i \right)^2 + \frac{h}{T} \sum_i S_i \right] \quad \text{con } J = 1 \text{ (modello di Ising ferromagnetico)}$$

e utilizzando la trasformazione gaussiana:

$$\exp \left(\frac{A}{2} O^2 \right) = C \int_{-\infty}^{+\infty} dx \exp \left(-\frac{x^2}{2A} + xO \right)$$

$$Z(h, T) = C \int_{-\infty}^{+\infty} dx \operatorname{Tr}_S \exp \left[-N \frac{x^2}{2T} + \frac{h+x}{T} \sum_i S_i \right]$$

dove $\sum_{S=-1}^1 \exp \left(\frac{h+x}{T} S \right) = 2 \cosh \left(\frac{x+h}{T} \right)$, avremo:

$$\begin{aligned} Z(h, T) &= C \int_{-\infty}^{+\infty} dx \exp \left(-N \frac{x^2}{2T} \right) \left[2 \cosh \left(\frac{x+h}{T} \right) \right]^N = \\ &= C \int_{-\infty}^{+\infty} dx \exp \left[-N \left[\frac{x^2}{2T} - \ln \left(2 \cosh \left(\frac{x+h}{T} \right) \right) \right] \right] \end{aligned}$$

Metodo ‘**Steepest Descent**’ (si sceglie la discesa più ripida sull’ipersuperficie dell’Energia):

$$I(a) = \int \exp [N \psi(x, a)] dx \approx \sum_j C \exp [N \psi(\bar{x}_j(a), a)]$$

dove, per $N \rightarrow \infty$ gli \bar{x}_j sono i punti dell’intervallo di integrazione per cui la funzione $\psi(x, a)$ ha un **massimo** assoluto (cioè un **minimo** se l’esponenziale ha esponente negativo).

Dunque l’energia libera per spin sarà:

$$f(h, T) = -\frac{T}{N} \ln Z(h, T) \rightarrow f(h, T) = \frac{\bar{x}^2(h, T)}{2} - T \ln \left(2 \cosh \left(\frac{\bar{x}(h, T) + h}{T} \right) \right)$$

dove \bar{x} e’ un minimo (**saddle point**) per $f(x, h, T)$, cioè è soluzione dell’equazione:

$$\frac{\partial f(\bar{x}, h, T)}{\partial \bar{x}} = \bar{x} - \tanh \frac{\bar{x} + h}{T} = 0$$

$$\dots \text{ e poiché: } m = \frac{T}{N} \frac{\partial \ln Z(h, T)}{\partial h} = -\frac{\partial f}{\partial h}$$

risulta : $m(h, T) = \bar{x}(h, T)$ e dunque, reintroducendo $J (=1)$, si può riscrivere f in funzione di m :

$$f(m, h, T) = \frac{Jm^2}{2} - T \ln \left(2 \cosh \left(\frac{Jm + h}{T} \right) \right)$$

da cui, risolvendo l'eq.di saddle point $\frac{\partial f(m,h,T)}{\partial m} = 0$, ritroviamo l'equazione autoconsistente per m: $m = \tanh\left(\frac{Jm + h}{T}\right)$

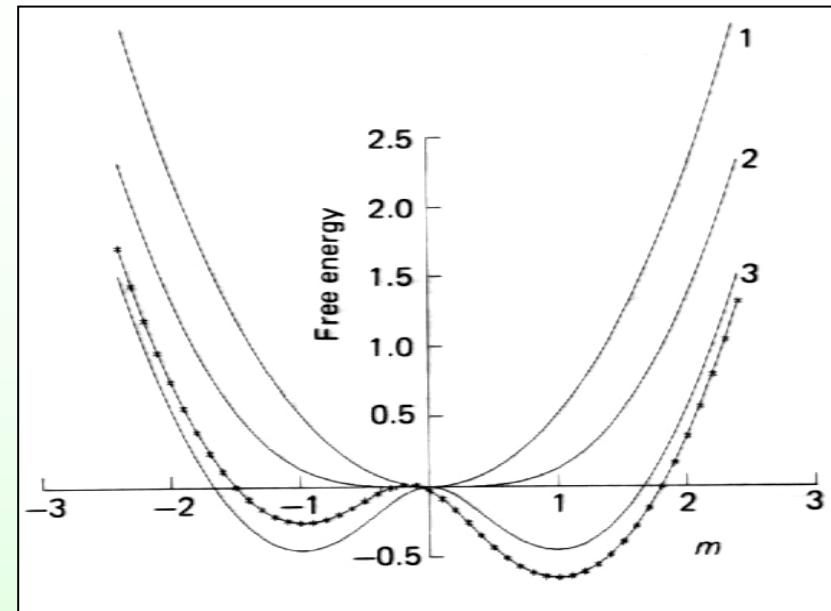
Graficamente si vede che (per $h = 0$):

- 1) $T > J$: f ha un solo minimo ad $m=0$
E' la **Fase Paramagnetica** (ergodica)
- 2) $T = J$: ci sono 3 soluzioni coincidenti
- 3) $T < J$: $m = 0$ diventa un massimo e compaiono 2 minimi degeneri ($m = +J, -J$).
E' la **Fase Ferromagnetica** (non ergodica)

Ancora una volta, $T = 1$ è la temperatura critica.

Relaxation Equation

$$\Gamma \frac{dm}{dt} = \frac{\partial f(m, h, T)}{\partial m}$$



Si vede che l'effetto di un **campo h esterno** (dotted line) provoca una rottura di simmetria e una degenerazione dei due minimi di f
Ad **alta temperatura** la dinamica è ergodica, cioè il sistema di Ising a connessione totale (e nel limite termodinamico) può trovarsi ovunque nello spazio degli stati.

A **bassa temperatura** saranno le condizioni iniziali della dinamica di Glauber (che, in condizioni di bilancio dettagliato, riduce l'energia libera ad ogni step) a determinare in quale dei 2 attrattori andrà a finire il sistema

Va precisato che in un sistema affetto da rumore termico non si può parlare di punti fissi della dinamica ma gli attrattori sono definiti da **distribuzioni di probabilità** (che all'equilibrio si riducono alla distribuzione canonica di Gibbs)

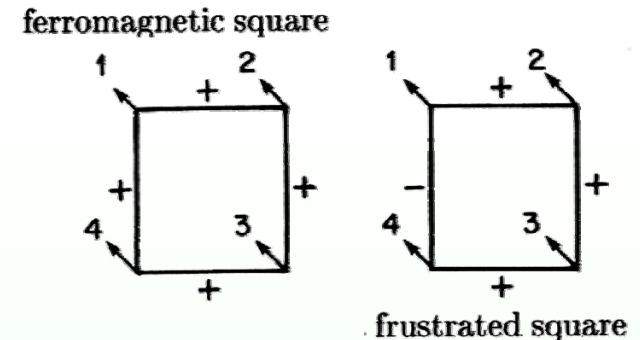
Ciascuna distribuzione è associata ad una stretta regione dello spazio degli stati (cioè una ‘valle’ della superficie energetica) in cui il sistema (a temperature sufficientemente basse) rimane **confinato** nel suo vagabondare stocastico

Queste differenti distribuzioni possono essere parametrizzate da una collezione di valori medi di certe osservabili. Nel nostro caso l'osservabile rilevante è la **magnetizzazione media m**, che come abbiamo visto gioca il ruolo di parametro d'ordine.

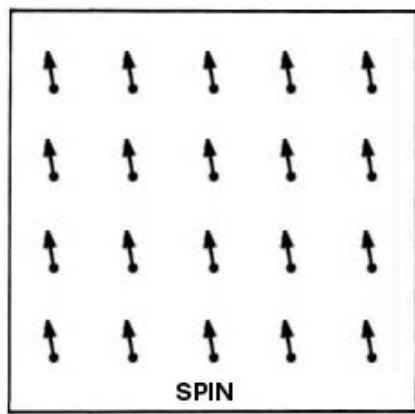
Ci sono casi, però, in cui la sola magnetizzazione non è sufficiente a descrivere il sistema di spin, a causa della presenza di “frustrazione” nelle interazioni... Sono i cosiddetti “Vetri di Spin”...

Digressione...

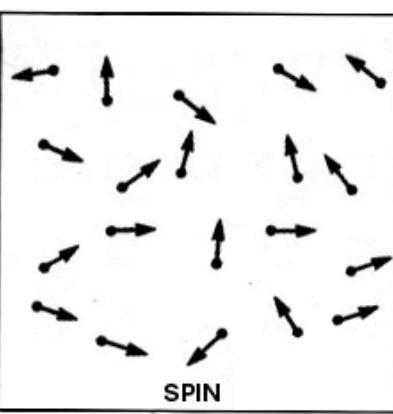
I Vetri di Spin ('Spin Glasses' SG)



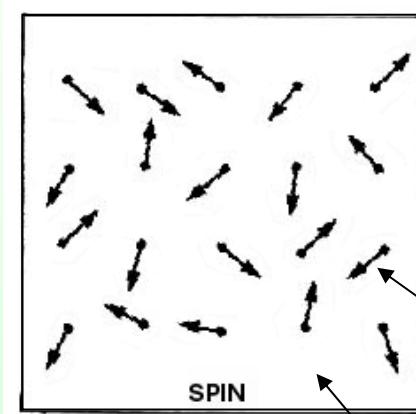
FASE
FERROMAGNETICA



FASE
PARAMAGNETICA

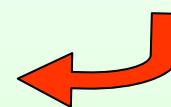


FASE
SPIN-GLASS



DISORDINE
'CONGELATO'

E
FRUSTRAZIONE



Metallo magnetico
(Ferro, Manganese).

$M \neq 0$

$M \sim 0$

$M \sim 0$

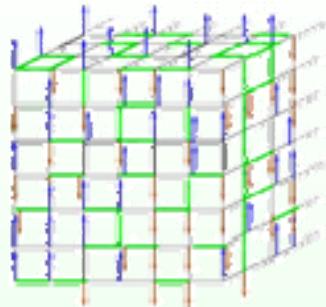
Metallo nobile
(Oro, Rame).

$$M = \frac{1}{N} \left| \sum_{i=1}^N \vec{s}_i \right|$$

Come si può distinguere
tra

la fase Paramagnetica e quella Spin Glass ?

Il modello Spin-Glass di Sherrington-Kirkpatrick



$$H = \frac{1}{2\sqrt{N}} \sum_{i,j=1}^N J_{ij} S_i S_j - H \sum_i S_i \quad \text{with } S_i = \pm 1$$

$$P(J_{ij}) = [(2\pi)^{1/2} J]^{-1} \exp\left[\frac{-(J_{ij} - J_o)^2}{2J^2}\right]$$

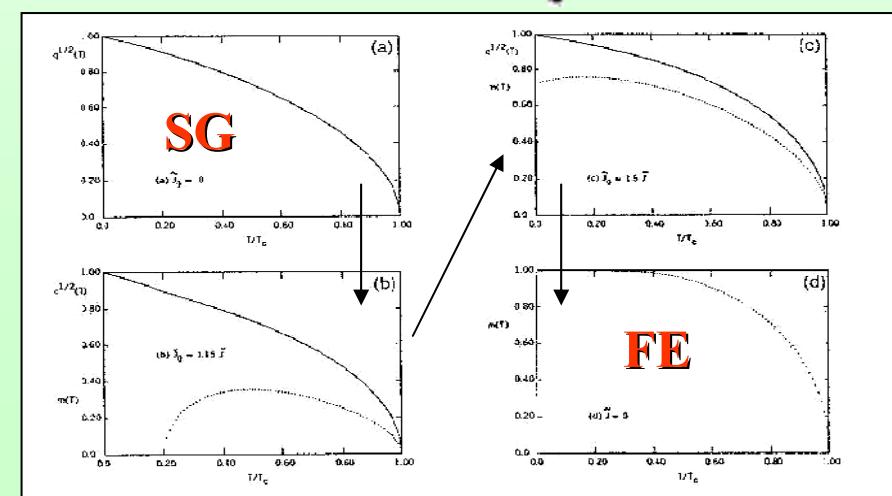
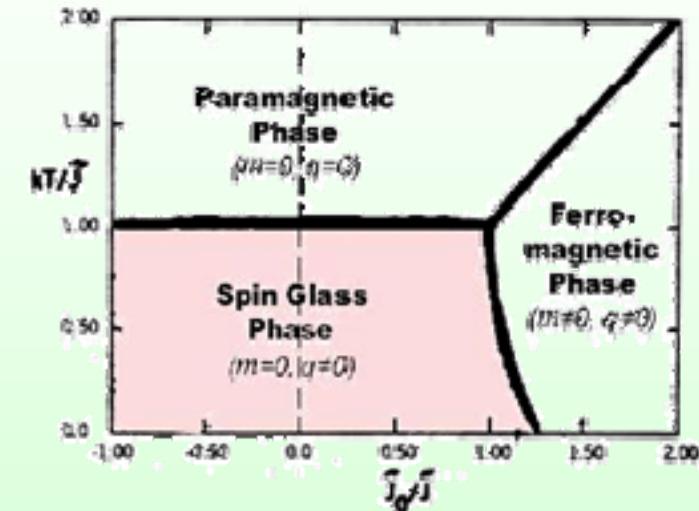
Nel modello di **Sherrington Kirpatrick (SK)**, il primo modello di **spin-glass** risolubile in approssimazione di campo medio, questo problema si risolve introducendo un **nuovo parametro d'ordine** (oltre alla magnetizzazione m) in grado di misurare il grado di '**congelamento**' degli spin nella fase spin-glass:

parametro d'ordine di Edwards Anderson

$$q_{EA} = \langle\langle S_i \rangle\rangle_d^2$$

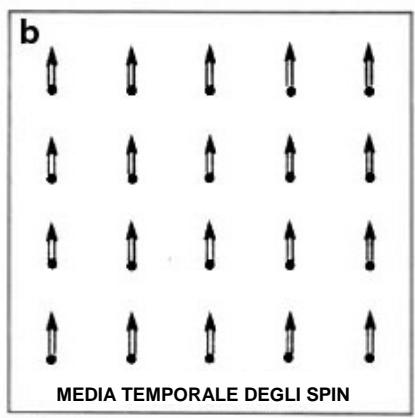
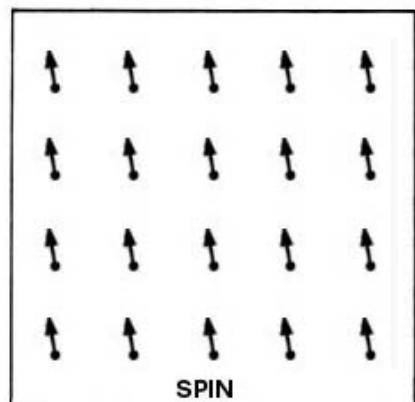
$\langle\dots\rangle$ = thermal average ; $\langle\dots\rangle_d$ = average over the spatial disorder

	m	q_{EA}
Ferromagnetic phase (FE)	$\neq 0$	$\neq 0$
Paramagnetic phase (PA)	0	0
Spin Glass phase (SG)	0	$\neq 0$



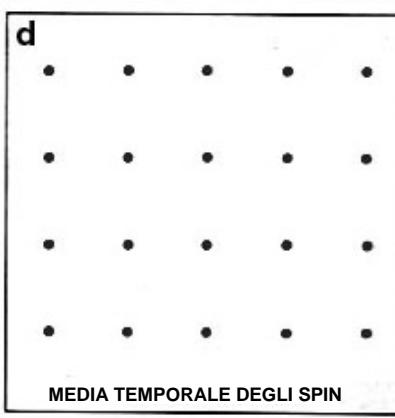
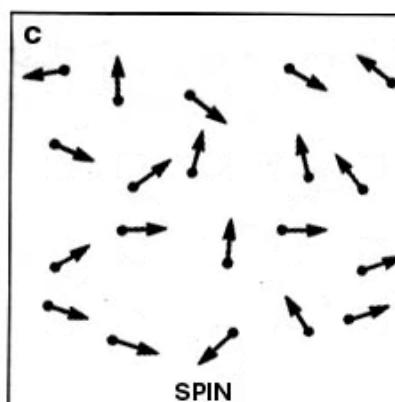
FASE FERROMAGNETICA

$$M = q \neq 0$$



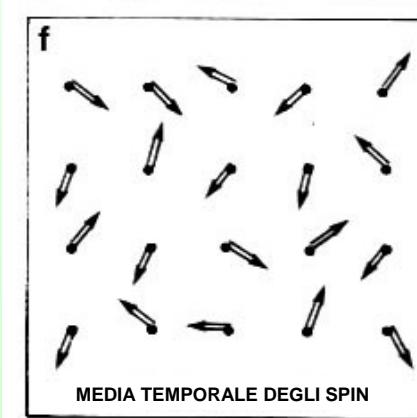
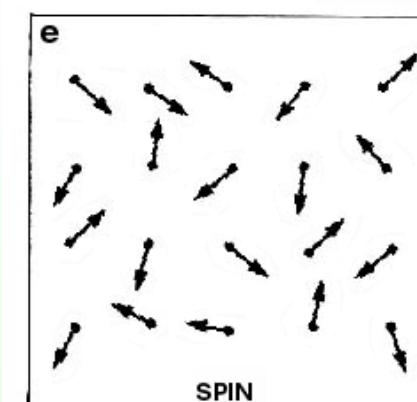
FASE PARAMAGNETICA

$$M = q \sim 0$$



FASE SPIN-GLASS

$$M \sim 0 \quad q \neq 0$$



La **frustrazione** presente nei Vetri di Spin rende impossibile minimizzare simultaneamente le energie di interazione di tutte le coppie di spin , ma allo stesso tempo produce una grande varietá di attrattori della dinamica (configurazioni degeneri ad energia minima) che possono essere utilizzati come memorie da rievocare nel contesto dei cosiddetti **modelli a Vetro di Spin delle Reti Neurali...**

Modelli a Vetro di Spin delle Reti Neurali ANN (Amit, Sompolinsky, Gutfreund, 1985-95)

Modello di Hopfield Generalizzato a bassa capacità di memoria

Assunzioni e semplificazioni

-I neuroni della rete sono elementi **discreti** a due stati $S = [+1, -1]$

-La dinamica di ogni neurone è **stocastica** (al limite, in assenza di rumore, il modello si riduce a quello classico di Hopfield con funzione di trasferimento a gradino)

-La rete comprende un numero **N** di **neuroni** molto grande (nel seguito sarà considerato infinito). I risultati sono comunque ben approssimabili nelle simulazioni anche con reti di poche centinaia di neuroni

-La rete si comporta come un **sistema dinamico stocastico** con un gran numero di elementi altamente non lineari e fortemente interagenti

-La dinamica della rete (noisy asynchronous e random) presenta delle semplici **proprietà asintotiche** ed è sufficientemente insensibile alle condizioni iniziali da poter funzionare come **memoria auto-associativa**

-Dire che la rete è in grado di memorizzare un certo numero ‘p’ di ‘patterns’ (forme) $\{\xi^\mu\}$ significa dire che, in assenza di rumore, ogni configurazione della rete del tipo:

$$S_i = \xi_i^\mu \quad \text{con } i = 1, \dots, N \quad \text{e} \quad \mu = 1, \dots, p \quad \text{è un punto fisso della dinamica}$$

-Il set di patterns $\{\xi^\mu\}$ è semplicemente correlato alla **matrice dei pesi sinaptici J** dalla solita legge di Hebb:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^p \xi_i^\mu \xi_j^\mu \quad (1) \quad \text{con} \quad J_{ii} = 0 \quad (\text{‘apprendimento statico’})$$

che è una relazione locale di tipo additivo, ossia ogni pattern ‘appreso’ aggiunge all’efficacia sinaptica della connessione J_{ij} la quantità:

$$\Delta J_{ij} = \frac{1}{N} \xi_i^\mu \xi_j^\mu$$

-La matrice dei pesi viene assunta **simmetrica**, cioè l'influenza del neurone j sul neurone i è uguale a quella del neurone i sul neurone j (e non ci sono auto-connessioni)

-Questa drastica semplificazione (che in effetti non ha un corrispettivo biologico ma trova un analogo nella simmetria delle interazioni magnetiche dei sistemi di spin) rende però possibile utilizzare, nello studio delle reti di Hopfield stocastiche, i potenti strumenti offerti dalla **meccanica statistica dei sistemi disordinati**

-A differenza di quanto accadeva in un sistema di spin di Ising, in cui $J = 1$ (caso ferromagnetico con 2 minimi dell'energia libera), qui la relazione (1) implica l'esistenza di interazioni sinaptiche inter-neuronali di segno sia positivo (eccitatorie) che negativo (inibitorie): questo significa introdurre una **frustrazione** nel sistema, che tende piuttosto a farlo avvicinare ai cosiddetti modelli a **Vetro di Spin**

-In tali sistemi, la coesistenza di **interazioni ‘congelate’** (quenched) di segno opposto e distribuite gaussianamente, impedisce la minimizzazione simultanea dell'energia per ogni coppia di spin: questa frustrazione genera dei **minimi energetici metastabili** con valori dell'energia più alti rispetto ai sistemi non frustrati, ma allo stesso tempo fa **aumentare enormemente** il numero di tali stati metastabili ($N_s \approx \exp(0.1992 N)$)

-Nella rete di Hopfield generalizzata, se il numero p di patterns si mantiene finito al tendere del numero N di neuroni della rete all'infinito, la matrice dei pesi sinaptici mette il sistema nella condizione di avere un numero contenuto di **attrattori correlati ai patterns** da memorizzare, rendendolo così in grado di rievocarli attraverso un processo dinamico di minimizzazione dell'energia libera

-Anche qui la presenza di rumore termico implica l'esistenza di attrattori sotto forma non di punti fissi della dinamica ma di **distribuzioni di probabilità** associate a regioni nello spazio degli stati connesse con i patterns memorizzati (da cui il ‘temporal averaging’, quo ultra)

-Anche qui è possibile parametrizzare tali attrattori attraverso un analogo della magnetizzazione per i sistemi di spin, e cioè attraverso il vettore \mathbf{m} degli **overlaps** dello stato istantaneo della rete con ciascuno dei p patterns memorizzati, le cui componenti sono:

$$m^\mu(t) = \frac{1}{N} \sum_{i=1}^N \xi_i^\mu S_i(t) \quad \text{con } \mu = 1, \dots, p$$

Distanza di Hamming : quantifica la diversità tra due configurazioni di spin (o neuroni) S ed S'

$$d_H = N - N_A$$

N = numero totale di spin di ciascuna configurazione

N_A = numero di spin allineati nelle due configurazioni

Overlap : quantifica la similarità tra due configurazioni S ed S'

$$M = N_A - d_H = N - 2d_H$$

$$m = \frac{M}{N}$$

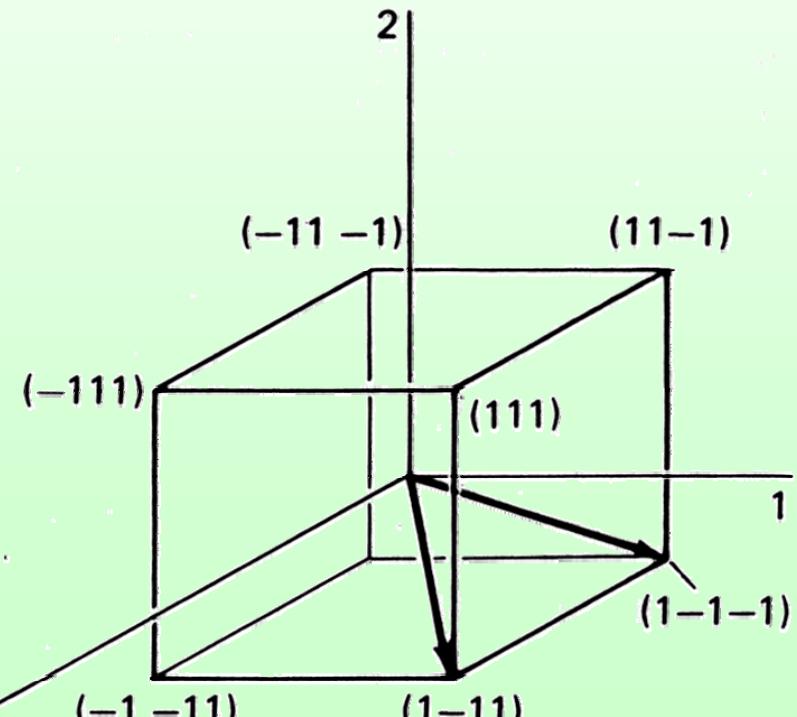
Relazione tra Overlap e Distanza di Hamming

$$d_H = \frac{1}{2} N(1 - m)$$

In analogia con la magnetizzazione si ha che: $m = \frac{1}{N} \sum_i S_i S'_i$

Overlap tra stato della rete e patterns rievocati

$$m^\mu(t) = \frac{1}{N} \sum_{i=1}^N \xi_i^\mu S_i(t) \quad \text{con } \mu = 1, \dots, p$$



Ipotesi di Campo Medio ed equazioni di auto-consistenza

$$\langle S_i \rangle = \tanh(\beta h_i) \quad \text{Temporal Averaging ('fast noise' relativo al rumore termico sulle sinapsi)}$$

$$h_i = \frac{1}{N} \sum_{j,j \neq i}^N J_{ij} \langle S_j \rangle \rightarrow h_i = \sum_{\mu=1}^p \xi_i^\mu \langle m^\mu \rangle$$

$$\text{e dunque: } \langle m^\mu \rangle = \frac{1}{N} \sum_{i=1}^N \xi_i^\mu \langle S_i \rangle \rightarrow \langle m^\mu \rangle = \frac{1}{N} \sum_{i=1}^N \xi_i^\mu \tanh \left[\beta \left(\sum_{\mu=1}^p \xi_i^\mu \langle m^\mu \rangle + h^e \right) \right]$$

che è un set di **p equazioni non lineari** (Mean Field equations) risolvibili esattamente nel limite termodinamico e che, nel caso di un unico pattern ($p=1$) con tutti gli $\xi_i = 1$, si riducono alle equazioni di campo medio del modello di Ising a connessione totale.

Self-Averaging (media sul ‘disordine congelato’)

Sempre nel limite termodinamico, se **p rimane finito**, esiste un numero finito (2^p) di possibili combinazioni dei p bit i-esimi di ciascun pattern per l'i-esimo neurone (sito), ciascuna delle quali appare un numero molto grande di volte con una frequenza proporzionale a 2^{-p} .

Dunque, fissati i patterns (che rappresentano quindi ‘**variabili congelate**’ o ‘**quenched**’), il valor medio di una certa funzione dei patterns $G(\xi^\mu)$, su tutti i siti della rete, è uguale alla media sulla distribuzione dei valori che la funzione assume per ciascun sito, eseguita come se ci fosse un gran numero di sistemi nei quali la funzione assume tutti i possibili valori in quel sito con una probabilità uguale a quella delle variabili casuali considerate:

$$\frac{1}{N} \sum_{i=1}^N G(\xi_i^\mu) \rightarrow \sum_{\xi^\mu} \Pr(\xi^\mu) G(\xi^\mu) = 2^{-p} \sum_{\xi^\mu} G(\xi^\mu) \equiv \langle G(\xi^\mu) \rangle$$

Dunque è possibile esprimere le equazioni di auto-consistenza (ricavate per una **specifica realizzazione** dei patterns) mediando il vettore degli overlaps sul ‘**disordine congelato**’ :

$$\langle m^\mu \rangle = \langle \xi^\mu \tanh \left[\beta \left(\sum_{\mu=1}^p \xi^\mu \langle m^\mu \rangle + h^e \right) \right] \rangle$$

Minimizzazione dell'Energia libera nel caso p-finito/N infinito (Capacità di memoria $\alpha=p/N$ infinitesima)

Come per il modello di Ising ‘full connected’, anche qui è possibile ottenere queste stesse equazioni di auto-consistenza attraverso lo studio degli estremi dell’**energia libera** per spin f , che come sappiamo in presenza di rumore termico (dinamica di Glauber asincrona) sostituisce l’energia nel ruolo di funzione di Lyapunov.

$$\text{Dall'energia: } H = -\frac{1}{N} \sum_{i,j} J_{ij} S_i S_j - \sum_i h_i^e S_i \quad \text{con} \quad J_{ij} = \frac{1}{N} \sum_{\mu=1}^p \xi_i^\mu \xi_j^\mu \quad \text{simmetrica,}$$

ricaviamo la **funzione di partizione** per una specifica realizzazione dei patterns:

$$\rightarrow Z = \text{Tr}_s \exp \left[\beta \left(\frac{1}{2} \sum_{i,j} J_{ij} S_i S_j + \sum_i \left(\sum_{\mu} h^{\mu} \xi_i^{\mu} \right) S_i \right) \right]$$

dove notiamo la presenza, nel contributo esterno al PSP agente su ciascun neurone, di un set di ‘**campi esterni**’ uniformi h^{μ} correlati con i patterns memorizzati (questi ‘campi’, oltre che favorire la rievocazione di un determinato pattern, possono servire in generale come meccanismo di input, per simulare degli stimoli esterni o anche per stabilire una comunicazione tra reti differenti).

Tornando alla funzione di partizione avremo, sostituendo J_{ij} e tenendo conto dei termini diagonali:

$$\rightarrow Z = e^{-\beta \frac{p}{2}} \text{Tr}_s \exp \left[\frac{\beta}{2N} \sum_{\mu} \left(\sum_i \xi_i^{\mu} S_i \right)^2 + \beta \sum_{\mu} h^{\mu} \sum_i \xi_i^{\mu} S_i \right]$$

Facendo uso della solita **trasformazione gaussiana** (in versione p-dimensionale e con $x = m^{\mu}$):

$$\rightarrow Z = (N\beta)^{\frac{p}{2}} e^{-\beta \frac{p}{2}} \int_{-\infty}^{\infty} \prod_{\mu=1}^p \frac{dm^{\mu}}{\sqrt{2\pi}} \text{Tr}_s \exp \left[-\frac{N\beta}{2} \left(\sum_{\mu} (m^{\mu})^2 + \sum_i \sum_{\mu} (m^{\mu} + h^{\mu}) \xi_i^{\mu} S_i \right) \right]$$

dove si procede come nel modello di Ising ‘full connected’ fino a giungere all’espressione finale dell’**energia libera** per spin:

$$f(\beta, m^{\mu}, h^{\mu}) = -\frac{1}{\beta N} \ln Z \quad \longrightarrow \quad f = \frac{1}{2} \sum_{\mu=1}^p (m^{\mu})^2 - \frac{1}{\beta} \ln \left(2 \cosh \left[\beta \left(\sum_{\mu} m^{\mu} \xi^{\mu} + h \right) \right] \right) >>$$

dove si è fatto uso della media (**self-average**) sul ‘disordine congelato’ dei patterns e si è fissato il campo esterno $h^{\mu} = h$

A questo punto è possibile ottenere le equazioni finali di auto-consistenza per gli overlaps medi con i patterns memorizzati semplicemente **minimizzando l'energia libera** (equazioni di saddle point):

$$\frac{\partial f(m^\mu, T, h)}{\partial m^\mu} = 0 \rightarrow \langle m^\mu \rangle = \langle \xi^\mu \tanh \left[\beta \left(\sum_{\nu=1}^p \langle m^\nu \rangle \xi^\nu + h \right) \right] \rangle$$

Le soluzioni di queste equazioni andranno studiate al variare del **rumore termico T** e di esse occorrerà anche verificare la **stabilità** attraverso lo studio della matrice:

$$A^{\mu\nu} \equiv \frac{\partial^2 f}{\partial m^\mu \partial m^\nu} \rightarrow A^{\mu\nu} = (1 - \beta) \delta^{\mu\nu} + \beta Q^{\mu\nu} \quad \text{con} \quad Q^{\mu\nu} = \langle \xi^\mu \xi^\nu \tanh^2(\beta \vec{m} \cdot \vec{\xi}) \rangle$$

Una certa soluzione sarà quindi localmente stabile solo se tutti gli **autovalori** della matrice A sono positivi.

Attrattori nel caso p-finito ($\alpha=0$)

Studiamo dunque le soluzioni delle seguenti p equazioni di campo medio eliminando, per semplicità, il campo esterno ($h = 0$):

$$m^\mu = \langle \xi^\mu \tanh \left(\beta \sum_{\nu=1}^p m^\nu \xi^\nu \right) \rangle$$

Dinamica asincrona in assenza di rumore termico (T=0 noiseless dynamics)

Retrieval States. Uno stato della rete è detto ‘**retrieval state**’ se possiede un’unica componente non nulla del vettore degli overlaps m^μ (scegliamo ad es. $m^1 = m$), ossia se è correlato solo con uno dei patterns memorizzati (in questo caso ξ_1).

Poiché nel limite $\beta \rightarrow \infty$ si ha (per $x \neq 0$) : $\tanh(\beta x) \approx \text{sign}(x)$ avremo:

- per $\mu=1 \rightarrow m = \langle\langle \xi^1 \text{sign}(m\xi^1) \rangle\rangle = \text{sign}(m) \langle\langle |\xi^1| \rangle\rangle$ (essendo $x \text{sign}(x) = |x|$)

che, poichè $|\xi^1| = 1$, diventa semplicemente: $m = \text{sign}(m)$ che ha le seguenti 2 soluzioni:

m = +1 : rappresenta uno stato attrattore perfettamente allineato col patterns n° 1

m = -1 : rappresenta uno stato attrattore in cui ogni singolo bit è anti-allineato col patterns n° 1

- per $\mu \neq 1 \rightarrow m^\mu = \langle\langle \xi^\mu \text{sign}(m\xi^1) \rangle\rangle = \text{sign}(m) \langle\langle \xi^\mu \xi^1 \rangle\rangle = 0$ (essendo i patterns non correlati)

Poiché l’indice $\mu=1$ era stato scelto arbitrariamente si avranno dunque **2p soluzioni** con un singolo overlap non nullo.

Queste soluzioni sono minimi assoluti (‘ground states’) dell’Energia (a T=0), che per essi vale:

$$E(T=0) = -\frac{1}{2} N \sum_{\mu=1}^p \left(\frac{1}{N} \sum_{i=1}^N \xi_i^\mu S_i \right)^2 + \frac{1}{2} p = -\frac{1}{2} N \sum_{\mu=1}^p (m^\mu)^2 + \frac{1}{2} p = -\frac{1}{2} N$$

essendo, nel caso p-finito, il termine $(1/2)p$ trascurabile.

Spurious States. In assenza di rumore termico le equazioni di campo medio presentano però un'altra classe di soluzioni ciascuna correlata con più di un pattern memorizzato, del tipo:

$$\vec{m} = m_n \left(\underbrace{1, 1, \dots, 1}_n, \underbrace{0, 0, \dots, 0}_{p-n}, 0 \right)$$

Sono i cosiddetti ‘spurious states’, che rappresentano stati della rete con ‘n’ overlaps non nulli tutti di uguale ampiezza m_n (sono cioè ‘miscele simmetriche’ di n patterns).

Poiché gli m_n possono cambiare di segno e distribuirsi su qualsiasi componente del vettore \mathbf{m} , il numero di possibili combinazioni diverse è:

$$2^n \binom{p}{n}$$

che è quindi anche il numero delle possibili soluzioni di questo tipo. La loro energia dipende da ‘n’ ed è data dalla relazione:

$$E_n = -\frac{1}{2} N n m_n^2 \quad \text{e poiché risulta che} \quad m_n \approx \left(\frac{2}{n \pi} \right)^{\frac{1}{2}}$$

avremo la seguente scala di energie: $E_1 < E_3 < E_5 \dots < E_\infty < \dots < E_6 < E_4 < E_2$ (con $E_\infty = -\frac{N}{\pi}$)

-L’energia delle **miscele con n dispari** decresce al descrescere di n, fino al valore $E_1 = -0.5 N$ dei retrieval states.
Questi stati sono stabili e dunque **sono attrattori** della dinamica *noiseless*.

-L’energia delle **miscele con n pari** aumenta al decrescere di n fino al valore $E_2 = -0.25 N$ ma questi stati non sono stabili (sono ‘saddle point’ dell’energia) dunque **non sono attrattori**.

Dinamica asincrona in presenza di rumore termico ($T > 0$)

Ovviamente, in generale, le nostre equazioni di campo medio per gli overlaps macroscopici:

$$m^\mu = \langle\langle \xi^\mu \tanh \left(\beta \sum_{\nu=1}^p \xi^\nu m^\nu \right) \rangle\rangle$$

hanno sempre una soluzione stabile ad overlap nullo del tipo: $m^\mu = 0$ con $\mu = 1, \dots, p$

che è tipica della fase ergodica (paramagnetica) ad alta temperatura e rappresenta evidentemente una condizione in cui non è possibile utilizzare la rete di Hopfield come memoria associativa.

E' interessante, dunque, vedere al di sotto di quale temperatura critica questa soluzione diventa instabile e si verifica la **rottura di simmetria** tipica invece della **fase non-ergodica** e necessaria per consentire la rievocazione dei patterns memorizzati.

A tal fine, moltiplicando il secondo membro delle equazioni di campo medio per m^μ , sommando su μ e utilizzando il fatto che $|\tanh x| \leq |x|$, si ottiene:

$$\sum_\mu (m^\mu)^2 = \langle\langle \left(\sum_\mu \xi^\mu m^\mu \right) \tanh \left(\beta \sum_\nu \xi^\nu m^\nu \right) \rangle\rangle \leq \beta \langle\langle \left(\sum_\mu \xi^\mu m^\mu \right) \left(\sum_\nu \xi^\nu m^\nu \right) \rangle\rangle$$

dove la non-correlazione dei patterns implica che $\langle\langle \xi^\mu \xi^\nu \rangle\rangle = 0$ tranne che per $\mu = \nu$, nel qual caso la media è uguale ad 1. Avremo allora:

$$\sum_\mu (m^\mu)^2 \leq \beta \sum_\mu (m^\mu)^2$$

dove, per $\beta < 1$ – cioè per $T > 1$ – il secondo membro si mantiene sempre minore del primo e dunque esiste solo la soluzione paramagnetica $m^\mu = 0$

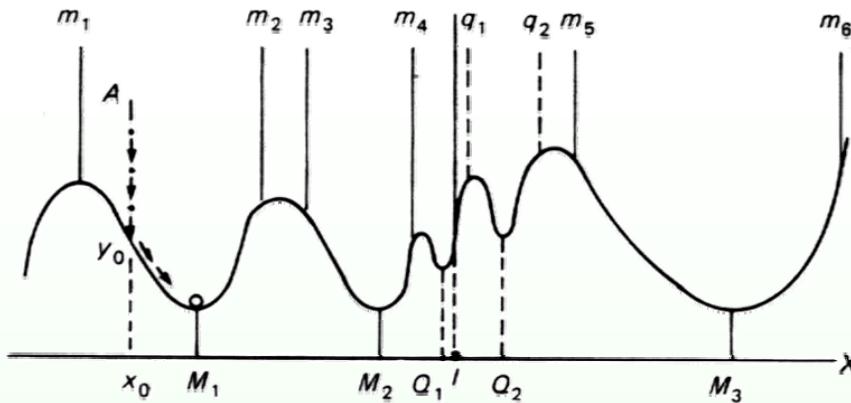


Figure 2.10: The one-dimensional landscape metaphor for associative, content addressable memory. $M_1 - M_3$ are memories, Q_1, Q_2 are spurious states, m_1, \dots, m_6 are maxima delimiting basins of attraction.

- La $T = 1$ è dunque la **temperatura critica** cercata, al di sotto della quale iniziano a comparire delle soluzioni con un numero ‘n’ di overlaps con i patterns memorizzati che diventano gradualmente non nulli (la transizione di fase è del secondo ordine) fino a raggiungere, per $T=0$, i valori ottenuti nel caso ‘noiseless’: ritroviamo dunque i desiderabili ‘**retrieval states**’ ($n = 1$) e i meno utili ‘**spurious states**’ (sotto forma di miscele simmetriche con $n > 1$).

L’analisi di **stabilità** di tali soluzioni assegna però un **ruolo positivo** al rumore termico. Infatti:

- i retrieval states rimangono stabili nell’intero range $0 < T < 1$;
- le miscele simmetriche con n pari rimangono instabili nell’intero range $0 < T < 1$;
- le miscele simmetriche con n dispari risultano instabili da $T = 1$ fino a $T = 0.461$, al di sotto della quale temperatura, invece, diventano dei minimi locali (a cominciare dalle miscele con $n = 3$);

- Dunque nella finestra $0.461 < T < 1$ solo i **retrieval states**, correlati con uno soltanto dei patterns memorizzati, **risultano stabili** e sono quindi degli attrattori per la dinamica della rete.

- Rammentiamo inoltre che, in presenza di rumore termico, non si può comunque parlare di punti fissi della dinamica poiché lo stato della rete non cessa mai di fluttuare: la rievocazione implica piuttosto una **media temporale** degli stati del sistema su una certa regione dello spazio degli stati.

Si può effettivamente parlare di rievocazione solo quando questa media è sufficientemente alta.

Il ruolo del rumore termico (“fast noise”) sulle ANN con p finito

- Non ci sono attrattori completamente stabili (punti fissi della dinamica nello spazio degli stati) ma solo ‘**distribuzioni di attrattori**’
- Se il rumore non è troppo alto, il punto rappresentativo della rete nello spazio degli stati si muove all’interno di regioni in cui l’overlap con i patterns memorizzati fluttua di poco
- I patterns memorizzati (e i loro ‘reversed states’) sono gli attrattori maggiormente stabili per tutti i livelli di rumore per i quali si ha rottura dell’ergodicità
- **All’aumentare dl livello di rumore gli ‘spurious states’ vengono gradualmente destabilizzati** (proporzionalmente al numero di patterns miscelati). Gli ultimi ad essere destabilizzati, per $T=0.46$, sono le **3-miscele simmetriche**: a questa T l’overlap medio è circa $m=0.97$
- Al di sopra della **soglia critica di rumore**, $T=1$, il sistema diventa ergodico e gli overlaps decrescono con continuità (TF del 2° ordine) a zero rendendo la rievocazione impossibile
- La **metafora del paesaggio energetico**, dei ‘retrieval states’, degli attrattori spuri e del rumore termico ha stimolato delle **speculazioni psichiatriche** sulla schizofrenia e sui disturbi del parlato: il **parlato normale**, governato da una ‘gestalt’ stabile, è associato ai retrieval states associati ai patterns memorizzati dalla rete ANN; la **schizofrenia**, che si manifesta quando il soggetto assembla diverse gestalt in un’unica, incoerente ma stabile struttura, è associata ai ‘mixtures states’; i **disturbi maniacali del parlato**, dove il soggetto salta da una singola gestalt all’altra, sarebbero invece dovuti al rumore termico, che provoca dei salti irregolari da un attrattore ad un altro

Analisi segnale rumore in ‘noiseless dynamic’

La condizione di stabilità di uno stato $\{S_i\}$ della rete è data dalla: $S_i h_i > 0 \quad (i=1,2,\dots,N)$ cioè dalla concordanza di segno (‘allineamento’) dello stato di ogni neurone col suo campo locale:

$$h_i \{S\} = \frac{1}{N} \sum_{j,j \neq i} \sum_{\mu} \xi_i^{\mu} \xi_j^{\mu} S_j$$

Se la rete si trova in uno stato che coincide con uno dei patterns memorizzati (ad es. ξ^1) avremo:

$$\xi_i^1 h_i = \frac{N-1}{N} + \frac{1}{N} \sum_{j,j \neq i} \sum_{\mu=2}^p \xi_i^1 \xi_i^{\mu} \xi_j^{\mu} \xi_j^1 > 0$$

dove compaiono due termini:

- un termine di tipo ‘**signal**’, quello per $\mu=1$, corrispondente al pattern di cui si vuole studiare la stabilità
- un termine di tipo ‘**noise**’ (‘slow noise’) che include il contributo dei rimanenti patterns memorizzati al PSP del neurone i -esimo.

Per $N \rightarrow \infty$ il primo termine è unitario, mentre il secondo contiene la somma di Np bits di valore 1 o -1 che, essendo i patterns non correlati, fluttua attorno allo zero con uno scarto quadratico medio dell’ordine di $\sqrt{p/N}$

Dunque:

- se $p \ll N$ il pattern memorizzato è stabile (è un punto fisso della dinamica ‘noiseless’) e possiede un grande bacino di attrazione. Si ha infatti:

$$\xi_i^1 h_i = 1 + R \quad \text{con} \quad |R| \approx \sqrt{\frac{p}{N}}$$

e se il sistema parte da uno stato iniziale con una frazione ‘ d ’ di ‘spin’ disallineati col pattern prescelto il ‘signal’ diventa $m_o = 1 - 2d$ mentre il rumore rimane immutato. Il campo locale sul neurone i -esimo sarà dunque $h_i = m_o + R$ e la rete rievocherà senza difficoltà il pattern memorizzato.

- se $p \approx N$, cioè quando il numero di patterns memorizzati diventa proporzionale al numero di neuroni, lo ‘slow noise’ introduce un fattore di instabilità: in tal caso i pesi sinaptici J_{ij} diventano delle variabili gaussiane totalmente non correlate (eccetto che per la simmetria $J_{ij}=J_{ji}$) e il modello di Hopfield generalizzato diventa analogo al modello a **Vetro di Spin ‘infinite range’ di Sherrington-Kirkpatrick (SK)** in cui la distribuzione delle interazioni J_{ij} abbia media zero e varianza p/N (vedremo che i due modelli diventano praticamente identici nel caso dell’introduzione di una **diluizione simmetrica** delle sinapsi nel modello di Hopfield generalizzato – che tra l’altro aumenterà anche la capacità di memoria del sistema).

Modello di Hopfield Generalizzato per $\alpha=p/N$ finito (Memory Saturation)

Quando $p=\alpha N$ non è più possibile, dunque, trascurare i piccoli **overlaps casuali** tra i patterns memorizzati (ciascuno dell'ordine di $1/\sqrt{N}$) i quali aggiungono una fonte di '**slow noise**' al già presente rumore termico ('**fast noise**'). Questo determinerà:

- una totale deformazione della distribuzione degli 'spurious state', che si 'fonderanno' tra loro in uno stato di Vetro di Spin (**Spin Glass State**) completamente senza forma
- uno spostamento della posizione degli stati stabili della dinamica (retrieval states) che si allontaneranno leggermente dai patterns memorizzati (facendo diminuire l'overlap massimo)
- una completa distruzione dei 'retrieval states' ('**black out**') al raggiungimento di una **soglia critica** di patterns memorizzati ($\alpha_c = 0.138$): oltre questa soglia (TF del 1°ordine) gli overlaps decadrono bruscamente a zero e gli attrattori (di tipo spin glass) diventeranno essenzialmente non correlati con i patterns memorizzati.

Parametri d'ordine

In questo contesto ($p=\alpha N$) è necessario introdurre, oltre ai soliti overlaps con i patterns memorizzati (la cui media temporale non nulla continuerà ad attestare le condizioni di 'retrieval'), **altri due parametri d'ordine**:

-un **parametro q** (detto di Edwards-Anderson), che mantenendosi – al contrario degli overlaps **m** – non nullo nella fase 'Spin Glass' (SG), andrà a tenere conto del 'temporal freezing' dinamico dei singoli spin non correlato con nessuna delle memorie del sistema, ma anche delle conseguenti eventuali correlazioni a lungo range tra le orientazioni dei vari spin;

-un **parametro r** ausiliario che, nel caso in cui la rete si trovi in uno stato fortemente correlato (cioè con overlaps macroscopici) con un certo numero s di patterns, serve a tenere conto del rumore casuale generato dall'accumulazione degli overlaps microscopici con i rimanenti p-s patterns (sempre nel limite termodinamico) – in totale è come se il sistema fosse sottoposto, oltre che al **rumore termico T**, ad un rumore supplementare equivalente ad una temperatura $T=\sqrt{2\alpha r}$

Replica Trick

Inoltre, non essendo più finito il numero p di patterns memorizzati al limite termodinamico, il ‘self-averaging’ sulla distribuzione dei patterns non è più sufficiente a garantire la correttezza dei risultati ottenuti per una data realizzazione dei patterns. Occorre dunque computare le proprietà all’equilibrio del sistema come se esso fosse effettivamente costituito da un ensemble di sistemi identici non interagenti (**repliche**) ciascuno con una differente realizzazione dei patterns che tenga conto della loro distribuzione di probabilità (a ciò si aggiungerà l’ipotesi di ‘**simmetria delle repliche**’).

Questa espressione dell’energia libera si ottiene utilizzando una **procedura di replicazione** simile a quella adottata da **Sherrington e Kirkpatrick** per il loro modello ‘infinite range’ di Vetro di Spin:

$$f = -\frac{1}{\beta} \lim_{n \rightarrow 0} \lim_{N \rightarrow 0} \frac{1}{nN} (\langle\langle Z^n \rangle\rangle - 1)$$

dove n è il numero di repliche e l’espressione della potenza n -esima della funzione di partizione è:

$$\langle\langle Z^n \rangle\rangle = \langle\langle Tr_{S^\rho} \exp \left[\frac{\beta}{2N} \sum_{ij\mu\rho} (\xi_i^\mu S_i^\rho)(\xi_j^\mu S_j^\rho) - \frac{1}{2} \beta p n + \beta \sum_{\nu=1}^s h^\nu \sum_{i\rho} \xi_i^\nu S_i^\rho \right] \rangle\rangle$$

dove :

ρ è l’**indice di replica**, in funzione del quale verranno espressi i parametri m , q ed r

- la traccia è sulle configurazioni di spin in ciascuna replica
- compare il termine

$$- (1/2) \beta p n$$

per tener conto dei termini di ‘autointerazione’ che stavolta sono stati inclusi nelle sommatorie per semplicità di scrittura

- compara un campo esterno h correlato con un numero finito s di ‘**condensed’ patterns** (cioè quei patterns con cui gli attrattori della rete avranno un overlap macroscopico), che nel seguito ipotizzeremo sia nullo

Anche qui, attraverso l’ipotesi di **simmetria per replica** per i parametri m , q ed r , cioè ipotizzando che:

$$m_\rho^\nu = m^\nu \quad ; \quad q_{\rho\sigma} = q \quad \rho \neq \sigma \quad ; \quad r_{\rho\sigma} = r \quad \rho \neq \sigma$$

si giunge all’espressione dell’**energia libera per spin** vista in precedenza, la quale va stavolta minimizzata in funzione di tutti e tre i parametri d’ordine. Avremo così, finalmente, le 3 equazioni:

Equazioni di saddle-point per i tre parametri d'ordine

$$m^v = \langle\langle \xi^v \tanh \beta [\sqrt{\alpha} r z + (\vec{m} \cdot \vec{\xi})] \rangle\rangle$$

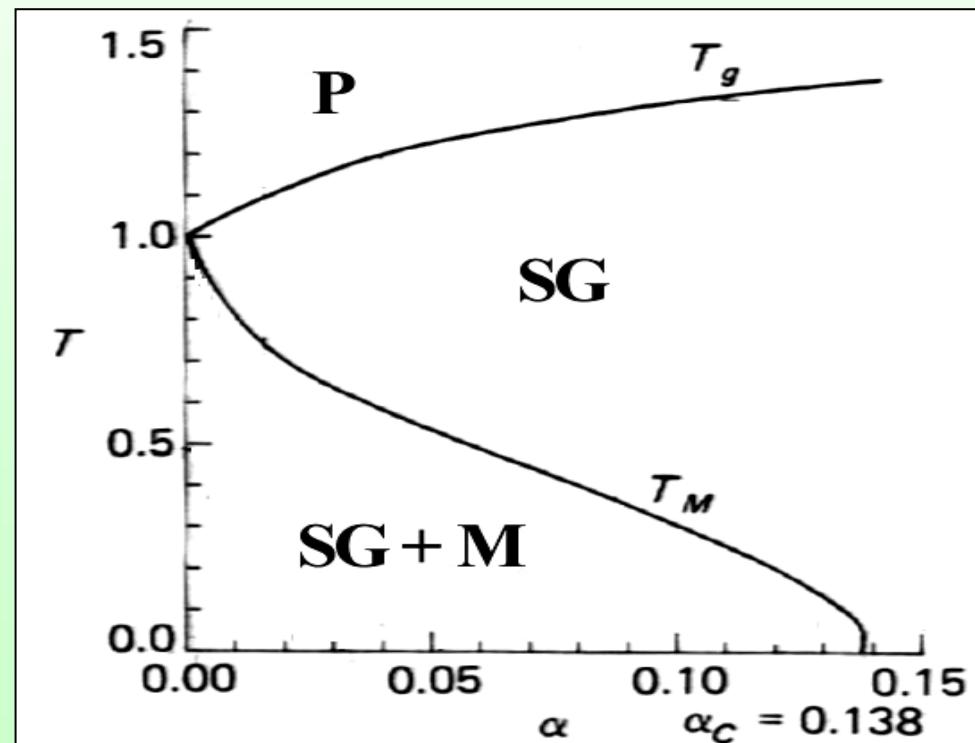
$$q = \langle\langle \tanh^2 \beta [\sqrt{\alpha} r z + (\vec{m} \cdot \vec{\xi})] \rangle\rangle$$

$$r = \frac{q}{(1 - \beta(1 - q))^2}$$

dove i vettori hanno tutti **s** componenti (relative ai ‘condensed patterns’) e dove compaiono i contributi dovuti alla **parte ferromagnetica** ($\vec{m} \cdot \vec{\xi}$) e a quella di ‘vetro di spin’ $\sqrt{\alpha} r z$

le cui soluzioni daranno luogo a **3 fasi distinte** nel diagramma $\alpha - T$:

- 1) **P - Fase Paramagnetica** (ergodica)
 $q = 0 ; r = 0 ; m^v = 0$
- 2) **SG - Fase di puro Vetro di Spin**
 $q \neq 0 ; r \neq 0 ; m^v = 0$
- 3) **SG+M - Fase mista SG + Memorie**
 $q \neq 0 ; r \neq 0 ; m^v \neq 0$



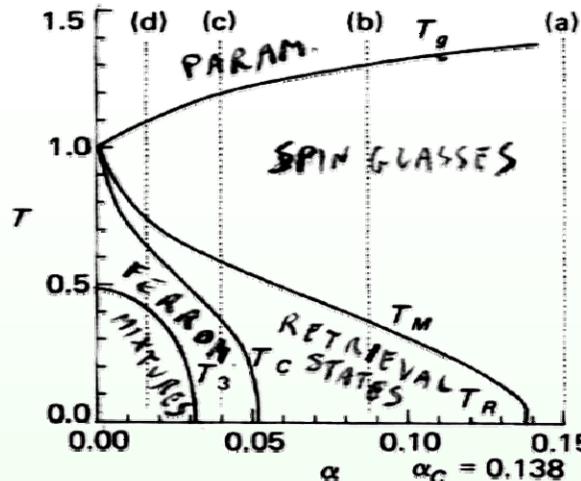


Figure 6.8: T - α phase diagram of a network near saturation. T_g is the transition temperature into the spin-glass phase vs α ; at T_M retrieval appears; at T_c the retrieval phase becomes global minimum of the free-energy; below T_3 mixtures of three patterns appear.

- ANN con un numero $p=\alpha N$ di patterns memorizzati e in assenza di ‘fast noise’ ($T=0$)
 - Se $p > 0.14 N$ non c’è rievocazione: tutti gli overlaps sono nulli ($m^\mu = 0$)
 - Immediatamente al di sotto della soglia critica ($\alpha_c = 0.14$) tutti i patterns memorizzati e i loro ‘reversed states’ sono rievocabili con una qualità migliore di 0.97 e con meno dell’ 1.5% di errori
 - Per $0.03 < \alpha < 0.14$ ci sono solo due tipi di attrattori: i patterns memorizzati (retrieval states) e gli stati ‘spin-glass’ che hanno overlaps praticamente nulli con i patterns memorizzati
 - Al di sotto di $\alpha = 0.03$ cominciano ad apparire degli spurious states
 - Per $\alpha = \alpha_c$ gli attrattori di tipo ‘retrieval’ scompaiono bruscamente e contemporaneamente: è la cosiddetta ‘blackout catastrophe’
 - Anche al di sopra della soglia critica, però, è ancora possibile rievocare dei patterns (ma non più di 3) attraverso l’introduzione di un set di ‘thresholds’ (o PSP) correlate con i patterns

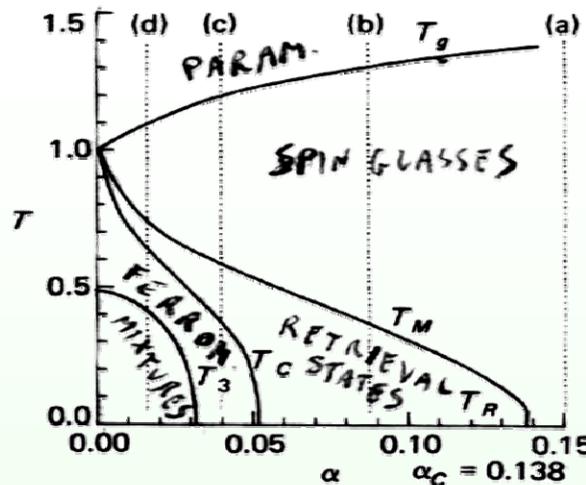


Figure 6.8: T - α phase diagram of a network near saturation. T_g is the transition temperature into the spin-glass phase vs α ; at T_M retrieval appears; at T_c the retrieval phase becomes global minimum of the free-energy; below T_3 mixtures of three patterns appear.

- ANN con un numero $p=\alpha N$ di patterns memorizzati e in presenza di ‘fast noise’ ($T>0$)
 - L’introduzione del ‘fast synaptic noise’, analogo della temperatura (rumore termico) nei sistemi magnetici, riduce la capacità di memorizzazione della rete aggiungendosi al rumore lento (‘slow quenched noise’) prodotto dagli overlaps casuali tra i patterns. Avvicinandosi a $T=1$ la capacità di memoria tende a zero e per $T>1$ la rievocazione non è più possibile
 - Come nel caso p -finito, il rumore termico elimina gli spurious states riducendo il loro α_c
 - L’aggiunta di una piccolissima quantità di rumore rende stabile la ‘replica symmetry’ per i retrieval states

Memorizzazione e rievocazione di Sequenze Temporali nelle ANN

L'idea di base, suggerita inizialmente dallo stesso Hopfield, è quella di introdurre un termine aggiuntivo **asimmetrico** nella matrice dei pesi sinaptici:

$$J_{ij} = \frac{1}{N} \sum_{\mu=1}^p \xi_i^\mu \xi_j^\mu + J^t_{ij} \quad \text{così definito:} \quad J^t_{ij} = \frac{1}{N} \sum_{\mu=1}^q \xi_i^{\mu+1} \xi_j^\mu$$

Esso rappresenta un '**transition term**' che, aggiungendosi al preesistente '**stabilizing term**', mette in relazione il pattern μ -esimo con il pattern $\mu+1$ -esimo che lo segue in una data sequenza di q patterns selezionati tra quelli memorizzati (ovviamente $q < p$).

Consideriamo dunque una rete che operi con una **dinamica asincrona** di passo temporale δt .

Se vogliamo che la rete si soffermi per un tempo $\tau = n\delta t$ su ogni pattern della sequenza temporale, in modo che la presenza in quell'attrattore possa essere riconosciuta come 'evento cognitivo', occorre introdurre un 'ritardo' τ nell'azione del termine di transizione (si parla perciò di 'slow synapses').

In presenza di **slow synapses**, infatti, anche il ‘campo locale’ (PSP) agente su ogni neurone risulterà di composto da due termini:

$$h_i(t + \delta t) = \sum_{\mu=1}^p \xi_i^\mu m^\mu(t) + \lambda \sum_{\mu=1}^q \xi_i^{\mu+1} m^\mu(t - \tau)$$

dove, come al solito, m^μ è l’overlap dello stato della rete $\{S_i\}$ con il pattern memorizzato $\{\xi_i^\mu\}$:

$$m^\mu(t) = \frac{1}{N} \sum_{i=1}^N \xi_i^\mu S_i(t)$$

Il primo termine del PSP, quello stabilizzante, dipende dall’attività dei neuroni al passo temporale precedente (cioè dalle ‘fast synapses’) attraverso gli $m^\mu(t)$, mentre il termine di transizione dipende dall’attività dei neuroni n passi temporali prima (attraverso le ‘slow synapses’).

In assenza di fluttuazioni (cioè se $p \ll N$) basta porre $\lambda > 1$ affinché la rete, sottoposta a questo PSP e partendo da un sufficiente overlap iniziale con uno dei pattern memorizzati, sia in grado di rievocare correttamente la sequenza di patterns stabilita.

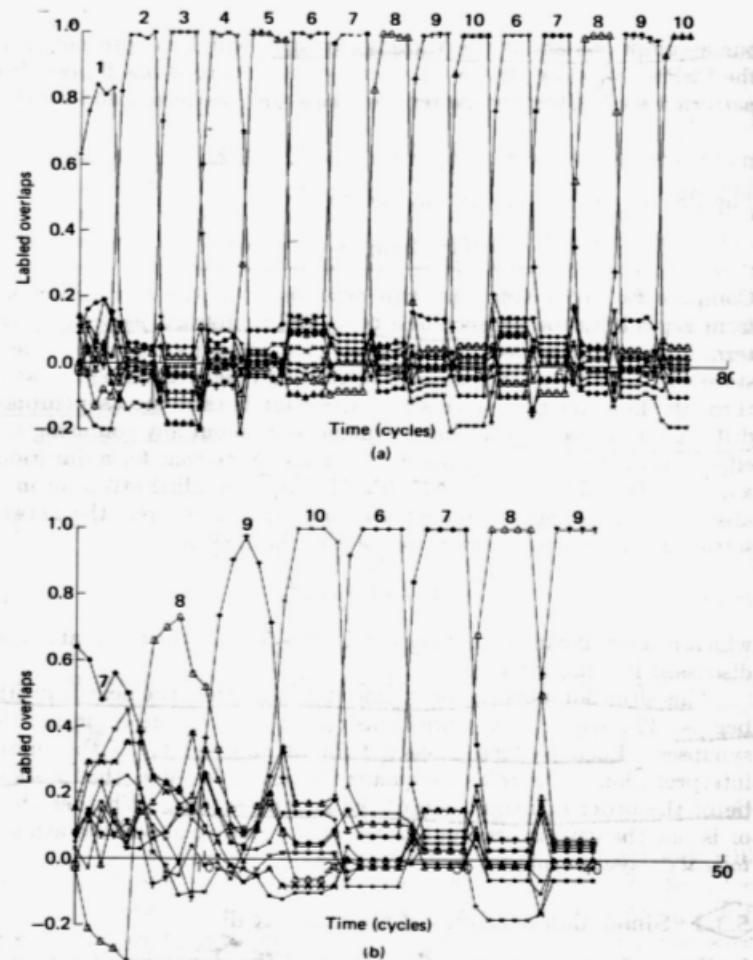


Figure 5.4: Simulation of the asynchronous dynamics of a network with $N=200$, $p=10$, $\lambda=1.5$, $\tau=5$, $T=0.6$. The sequence is 1-2-3-4-5-6-7-8-9-10-6. Curves are marked by the number of the pattern to which the particular overlap corresponds. Stimuli: (a) $m^1(0)=0.6$, (b) $m^7(0)=0.6$. Following a short transient the network follows the sequence, identifying clearly each pattern for the duration of the synaptic delay.