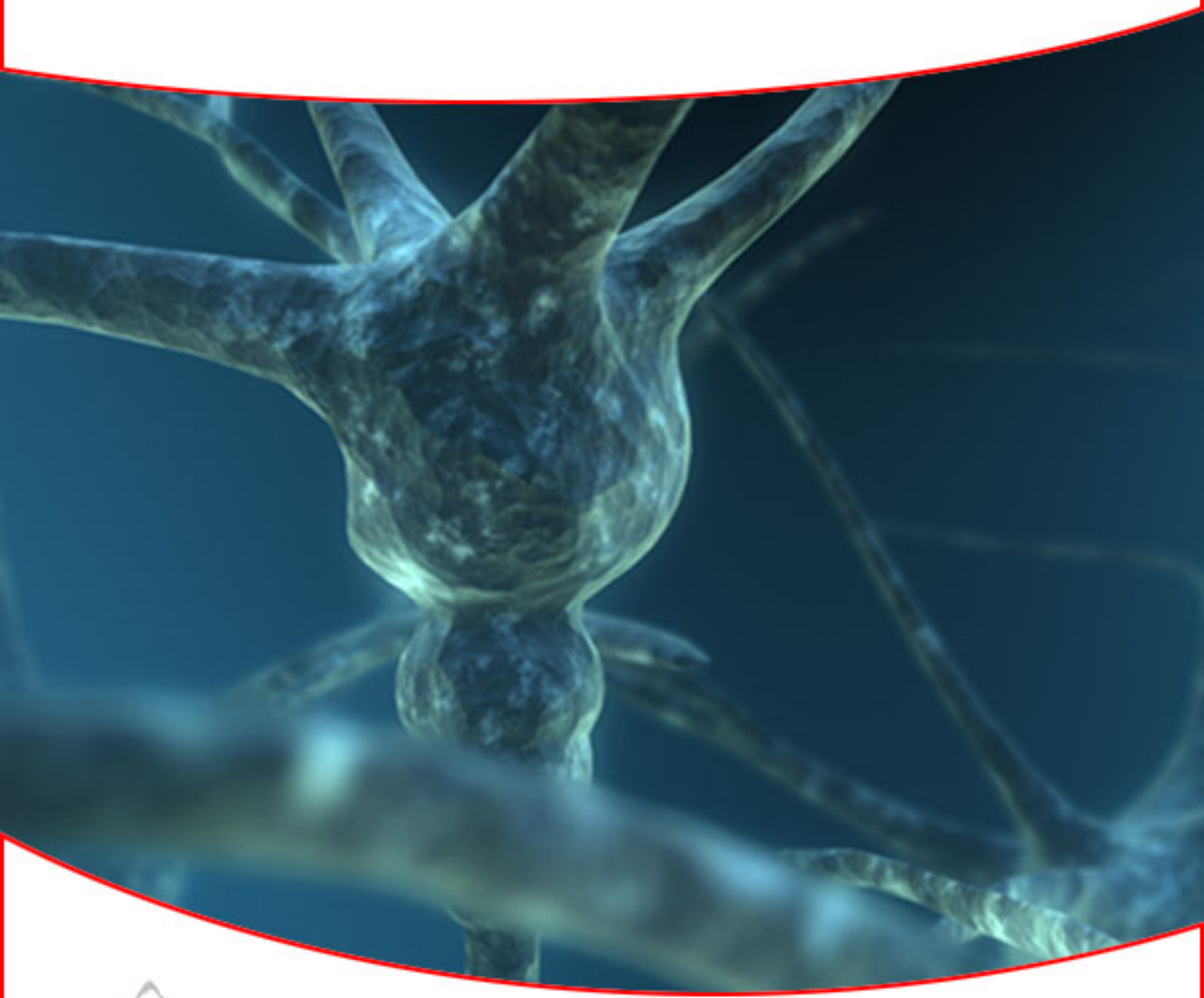
Neuroscienze.net

Journal of Neuroscience, Psychology and Cognitive Science



LA MACCHINA DELLA MENTE

Parte II[a]. Modelli di reti nervose.

Autore: Renato Nobili
Dipartimento di Fisica "G.Galilei" – Università di Padova.

Sommario: Questa parte contiene una breve rassegna di alcuni modelli di reti nervose ispirati al cosiddetto *paradigma connessionista*. L'idea che sta alla base di questo tipo di ricerca è che il gran numero di connessioni sinaptiche che si osservano tra i neuroni del sistema nervoso, unitamente alla variabilità dei coefficienti di trasmissione sinaptica secondo regole di rinforzo del tipo simile a quelle ipotizzate da Hebb (1949), renda possibile la realizzazione di processi informazionali sostanzialmente diversi e per certi aspetti più "intelligenti" di quelli seriali effettuabili dai comuni calcolatori. Modelli più evoluti, capaci di simulare alcuni comportamenti oscillatori delle reti neuronali, pur rientrando parzialmente nel paradigma connessionista, saranno trattati nella terza parte dell'articolo.

2.1. Vecchi e nuovi perceptron

Fin dai primi tentativi di modellizzazione delle reti nervose risultò chiaro (Barlow, 1959) che l'equiparazione del cervello al calcolatore elettronico, sebbene affascinante e ricca d'implicazioni generali, era fondamentalmente errata e fuorviante. Negli anni '60, per l'arretratezza della tecnologia elettronica e dell'informatica e l'insufficienza delle conoscenze neurofisiologiche, non era facile né conveniente tentare di sviluppare una teoria del calcolo parallelo, come fu fatto invece per il calcolo seriale a partire dagli approcci di Turing (1936) e von Neumann (1951). Dopo la morte del secondo (1956), sebbene gli automi cellulari progettati dal grande matematico fornissero un primo esempio di calcolo parallelo digitale (von Neumann-Burks, 1966), la tecnologia elettronica era lontana dal permetterne un'implementazione hardware. I tentativi di fondare una teoria del calcolo parallelo furono deboli e caratterizzati da uno spirito più ingegneristico che logico-matematico, e continuarono ad esserlo per molti anni. Ancora oggi non esiste una teoria sistematica generale dei processi paralleli. Questo ritardo si spiega solo in parte col fatto che l'implementazione di processi di calcolo parallelo simili a quelli che sono effettuati dai sistemi nervosi animali incontra difficoltà pratiche che la tecnologia attuale non è ancora in grado di superare. Quello che sembra ancora mancare è una comprensione teorica delle reali potenzialità del calcolo parallelo. Una cosa, tuttavia, cominciava ad apparire chiara: mentre i processi seriali trovavano la loro naturale via di sviluppo nella tecnologia elettronica digitale (utilizzo di unità di calcolo a stati discreti) quelli paralleli sembravano trovarla in quella analogica (utilizzo di procedimenti fisici con variabili di stato continue).

Il primo importante esempio di processo parallelo di tipo analogico fu proposto da Rosenblatt (1959) col modello del *perceptron*. Il perceptron è una rete formata da unità funzionali elementari simili a quelle di McCulloch e Pitts (vedi parte I). In realtà, queste unità costituiscono modelli di neuroni reali in un senso piuttosto vago. Per evitare confusioni con le reti neuronali reali quelle formate da quelle unità funzionali ideali sono usualmente chiamate *reti neurali*, sebbene tali componenti elementari continuino ad essere chiamati "neuroni". Bisogna aggiungere che questa terminologia ha contribuito a nobilitare scientificamente una disciplina di modesto spessore matematico (la teoria delle reti neurali) che nonostante le attese e le pretese iniziali non è mai stata in grado di fornire autentici modelli dei sistemi nervosi reali. Una rete neurale è costituita da uno o più strati d'unità neurali connessi in cascata, in modo che i neuroni di ciascuno strato (neuroni sorgente), se v'è più di uno strato, agiscano su quelli dello strato successivo (neuroni bersaglio) - ma non su quelli del medesimo strato o degli eventuali strati precedenti - attraverso connessioni (sinapsi) che permettono l'immissione di un segnale nell'unità bersaglio.

Lo stato interno d'ogni unità neurale è rappresentato dal *potenziale postsinaptico* (o semplicemente *potenziale* dell'unità neurale) che si esprime matematicamente come una combinazione algebrica lineare degli ingressi sinaptici. I coefficienti di questa combinazione algebrica (coefficienti di trasmissione sinaptica o pesi) possono assumere valori sia positivi che negativi e sono generalmente soggetti a variazioni. La rete è atta a processare un insieme di segnali entranti in parallelo attraverso una frazione di ingressi sinaptici (complessivamente chiamati *stimolo*) producendo un insieme di segnali uscenti in parallelo attraverso gli "assoni" dei neuroni (complessivamente chiamati *risposta*).

Ciò che si richiede a un perceptron è che sia "addestrabile". Vale a dire che, dopo l'applicazione di ripetuti abbinamenti di stimoli di varia forma dotati di certe caratteristiche comuni C_1, C_2, …, C_n (ad esempio suoni vocalici pronunciati da persone diverse), con corrispondenti risposte desiderate r_1, r_2,…,r_n (ad esempio segnali che attivano la scrittura di vocali alfabetiche), sia capace di generare la risposta r_i quando lo stimolo possiede le caratteristiche C_i.

Le procedure di apprendimento possono consistere di aggiustamenti della soglia neuronale – cioè il valore del potenziale oltre il quale il neurone passa dallo stato 0 allo stato 1 - o per correzione dei coefficienti di trasmissione sinaptica W_{ij} (pesi) tra coppie di neuroni i, j collegati cascata. Noi ci occuperemo delle procedure di questo secondo tipo. Se W_{ij} hanno valori casuali, un qualsiasi stimolo sarà tradotto dal perceptron in una risposta apparentemente casuale. I pesi dovranno essere modificati dal processo in modo che gli stimoli d'ogni determinata classe evochino in corrispondenza le risposte desiderate.

Ora, per i perceptron a strato singolo sono possibili relazioni stimolo-risposta piuttosto semplici e di scarso interesse applicativo. Invece, per quelli costituiti da più strati, pur di aumentare abbastanza il numero degli strati, sono possibili, *in linea di principio*, relazioni stimolo-risposta di qualsivoglia complessità.

Purtroppo questi modelli hanno un grave difetto generale: non esiste alcun metodo effettivo per calcolare le correzioni dei pesi sulla base dei parametri di stato della rete e del loro corso temporale. Più precisamente, mentre per i perceptron a strato singolo le regole per la correzione dei pesi risultarono facili da determinare, per quelli a molti strati non si riesce a trovare per i neuroni degli strati intermedi (*unità nascoste*) regole efficaci che siano concretamente implementabili mediante procedure dipendenti dallo stesso processo d'apprendimento. Così l'intervento correttivo si dà come una possibilità matematica che non è realizzabile in pratica. Molti anni dopo si sarebbe capito che la difficoltà era imputabile proprio all'infinita ripidità delle soglie a gradino delle unità neurali. Ricerche ulteriori, svolte per oltre un decennio, principalmente negli Stati Uniti da Widrow e i suoi allievi (Widrow e Lehr, 1990), ma anche in Italia da Caianiello e collaboratori (De Luca e Ricciardi, 1981), hanno contribuito in vario modo all'analisi dei processi paralleli implementabili con dispositivi simili al perceptron. Altre, di carattere più specialistico, hanno cercato di perfezionare quell'idea. Purtroppo il successo fu alquanto scarso, in parte per le difficoltà dei calcolatori di allora a simulare il comportamento dei modelli neurali formati da un gran numero di neuroni, in parte perché a quell'epoca nessuno era in grado di formulare ipotesi attendibili circa la struttura e il funzionamento delle reti nervose reali. E' giusto comunque menzionare quelli che hanno avuto un certo riscontro nelle reti nervose reali: le matrici di neuroni a connessioni rientranti di Reichardt (1956) che permettono di effettuare l'autocorrelazione dell'informazione dei dati forniti in ingresso; le reti ad inibizione laterale, che hanno fornito modelli attendibili di alcune reti neuronali come i neuroni motori della corteccia cerebrale o quelli che raccolgono l'informazione visiva negli occhi di molte specie d'insetti (Varju, 1962); le reti che organizzano temporalmente i segnali della rete nervosa cerebellare (Breitenberg, 1962). Un'esposizione dettagliata su questi argomenti è reperibile nel libro di De Luca e Ricciardi (1981).

Il filone di ricerca basato sull'idea del perceptron non ebbe allora un grande sviluppo e fu tralasciato per oltre un decennio dopo che nel 1969 Minsky e Papert dimostrarono che i dispositivi di quel genere non sono nemmeno in grado di apprendere ad effettuare l'operazione logica XOR (per una rete a due ingressi e un'uscita l'operatore XOR, generalmente indicato col simbolo ^, è

definito dalle equazioni 1^1 = 0^0 = 0, 1^0 = 0^1 = 1, dove 1 significa VERO e 0 significa FALSO). Mettendo in evidenza i limiti e i difetti del perceptron, questi due autori determinarono il dirottamento dei finanziamenti della ricerca sull'intelligenza artificiale (IA) verso l'approccio digitale-seriale.

Negli anni recenti la teoria del perceptron, sotto il nuovo nome di teoria delle *reti a retropropagazione*, ha avuto un'improvvisa ripresa. La storia è ricominciata quando Rumelhart, Hinton e Williams (1986), sull'onda del successo del modello di Hopfield, di cui parleremo più avanti, riesaminando la vecchia teoria, si accorsero che, rimpiazzando i neuroni con risposta a gradino con neuroni a risposta sigmoidale, si potevano ottenere perceptron multistratificati capaci di supportare processi d'apprendimento rapidamente convergenti nonché straordinariamente efficienti. Il successo dell'implementazione dello XOR fu salutato come un segnale di rivincita da tutti i vecchi studiosi del perceptron. Fu allora chiaro che proprio l'assunzione della risposta a gradino, perciò, in definitiva, lo stesso paradigma Booleano che aveva ispirato McCulloch e Pitts (parte I), aveva bloccato la teoria dei perceptron per più di vent'anni!

La strategia d'apprendimento dei nuovi perceptron è basata su un'ipotetica propagazione a ritroso, attraverso la cascata degli strati neurali, di segnali ricavati in uscita come differenze comparative tra le risposte effettive e quelle volute. La correzione dei pesi è fatta in modo proporzionale agli effetti esercitati localmente da questi segnali su ogni neurone degli strati intermedi (unità nascoste). Il procedimento richiede che le correzioni dei pesi siano, almeno approssimativamente*, proporzionali alle derivate delle risposte dei neuroni che formano le unità nascoste* (ciò spiega l'insuccesso dei neuroni con risposte a gradino poiché la derivata di questo tipo di risposta è ovunque zero tranne che in un punto, dove prende un valore infinito). Tuttavia, questo modello presenta un difetto simile a quello dei vecchi perceptron: non si sa indicare un metodo di cablaggio e dei meccanismi locali atti a produrre le giuste variazioni dei pesi per rendere effettiva la retropropagazione. A differenza dai vecchi perceptron, il procedimento di variazione dei pesi è ora matematicamente ben definito, ma la difficoltà ad immaginare un procedimento fisico che sia capace di rendere effettiva la retropropagazione non permette di considerare questo sistema come un modello neurologicamente compatibile. Nonostante queste difficoltà concettuali, è possibile simulare i processi d'apprendimento di queste reti a retropropagazione con procedimenti ricorsivi implementabili in un calcolatore.

Le simulazioni al calcolatore dei nuovi perceptron hanno prodotto risultati spettacolari. Reti addestrate a produrre suoni vocali sintetici, che all'inizio rispondono in modo balbettante, imparano a parlare; altre a riconoscere figure; altre a compiere operazioni di controllo complesse (ad esempio governare in retromarcia un camion con rimorchio; Widrow e Lehr, 1990); altre riescono a catturare la legge sottostante ad un insieme di stimoli (ad esempio estraggono risposte invarianti rispetto a gruppi di trasformazioni operanti sugli stimoli); altre a comprimere l'informazione eliminando la ridondanza di messaggi applicati come stimoli all'ingresso ecc. Il loro frequente impiego nei processi di controllo degli impianti industriali dimostra nel modo più convincente quale grado d'efficienza e affidabilità questi modelli abbiano potuto raggiungere.

La spiegazione di questo comportamento, che può essere definito *intelligente*, è molto semplice: la procedura d'apprendimento per retropropagazione abilita la rete a stabilire corrispondenze univoche tra classi di stimoli possibili e corrispondenti risposte possibili. Se la propagazione dei segnali attraverso la rete durante la fase d'apprendimento procede sempre in avanti, vale a dire se ogni neurone manda segnali solo ai neuroni dello strato successivo, e non vi sono connessioni tra neuroni di uno stesso strato, la risposta d'ogni neurone è una funzione univoca e continua delle risposte fornite dai neuroni degli strati precedenti. Pertanto le relazioni stimolo-risposta tra i neuroni del primo strato e quelli dell'ultimo sono mappe univoche e continue. Pertanto, dato che la risposta d'ogni neurone è una funzione monotona e smussata dell'ingresso sinaptico, la dipendenza delle uscite dagli ingressi non presenta fenomeni critici (multistabilità, isteresi, oscillazioni, comportamenti caotici ecc.).

L'interpretazione matematica di questo fenomeno è di facile comprensione. L'insieme degli stimoli possibili entranti per gli *N* ingressi si caratterizza matematicamente come *spazio N*-dimensionale se è possibile definire una "distanza" naturale tra coppie di stimoli possibili. Ciò si verifica, ad esempio, se tra stimoli diversi possono stabilirsi relazioni variabili con continuità (criteri di dissimilarità, reciproca trasformabilità per operazioni di tipo geometrico o per deformazioni, classificabilità secondo determinate leggi ecc.). Queste relazioni sono così proiettate nello spazio *M*-dimensionale delle risposte a *M* uscite dove, in generale, risultano organizzate in modi diversi secondo i valori assunti dai pesi durante le procedure d'addestramento.

Ora, la proprietà veramente importante esibita dalle reti a retropropagazione è la seguente: affinché la rete apprenda a mappare uno spazio di stimoli in uno di risposte è sufficiente che essa sia sottoposta ad un campionario sufficientemente rappresentativo di corrispondenze stimolo-risposta. Sotto queste condizioni, si genera una mappa non lineare tra lo spazio degli stimoli e quello delle risposte. Se i profili di risposta delle unità neurali non fossero sigmoidali ma semplicemente lineari, si otterrebbe una semplice corrispondenza tra regioni convesse dello spazio *N*-dimensionale degli stimoli e regioni convesse dello spazio *M*-dimensionale delle risposte. Grazie ai profili sigmoidali la mappa invece si deforma, si ripiega e si schiaccia in certe regioni in modi complicati. Così zone convesse dello spazio degli stimoli possono essere mappate in zone non convesse; zone con uguali dimensioni in zone di dimensioni diverse; zone non intersecantesi in zone intersecantesi ecc. Se il numero delle unità nascoste è sufficientemente grande, si possono stabilire corrispondenze molto ben approssimate tra certe regioni dello spazio degli stimoli e regioni dello spazio delle risposte di forma arbitraria. Le mappe così create vincolano ad una corrispondenza simile tutte le altre possibili relazioni stimoli-risposta. In altri termini la rete cattura la "legge" che correla tra loro gli ingressi esemplificati dal campionario iniziale di corrispondenze. Si è potuto dimostrare che con una rete formata da un solo strato d'unità nascoste sufficientemente numeroso si riesce ad approssimare mappe di qualsiasi complessità (Girosi e Poggio, 1990). In particolare, può accadere che la rete riesca a mappare tutti gli stimoli di una stessa classe a risposte molto vicine, così da avere, con buona approssimazione, una corrispondenza di tipo classe $\rightarrow$ punto.

Una notevole proprietà delle reti a retropropagazione, che ha importanti applicazioni pratiche, riguarda la possibilità di ottenere la *compressione dell'informazione*. La questione è rilevante anche dal punto di vista neurologico perché vari esperimenti di psicologia sperimentale (Hick, 1952; Hyman, 1953; Leonard, 1961) hanno dimostrato che un operatore umano addestrato a scegliere in modo consapevole le giuste risposte da dare a segnali occorrenti probabilisticamente con varie frequenze statistiche possiede un limite intrinseco nella sua velocità di risposta che è di circa sette bit per secondo (il magico numero sette); ciò in modo indipendente dalla mansione, dalla distribuzione statistica degli stimoli e dipendente soltanto dalla quantità d'informazione elaborata. Tutto accade come se il flusso dell'informazione nervosa elaborata della coscienza operazionale fosse trasportato da un unico canale di comunicazione di capacità limitata (Welford, 1959).

La questione della compressione dell'informazione è in stretta relazione con un noto teorema della teoria della comunicazione di Shannon (1949). In generale, i messaggi d'ogni specie possiedono una frazione d'informazione eccedente o *ridondanza*. Questa è tanto maggiore quanto maggiore è il numero di relazioni tra le parti del messaggio (ripetizioni, simmetrie, derivabilità del testo da un programma di lunghezza ridotta ecc.). Se la ridondanza fosse nulla, il minimo errore di copiatura in un processo di elaborazione del messaggio, o il minimo rumore nella linea di trasmissione, renderebbe il messaggio totalmente indecifrabile. Grazie alla ridondanza i messaggi affetti da piccoli errori possono essere completamente ricostruiti. Ad esempio, si può ricostruire agevolmente il testo di una frase da cui mancano alcune lettere alfabetiche perché il linguaggio ordinario ha una ridondanza di circa il 60%. D'altronde, la ridondanza dei messaggi acustici o visivi ordinari è piuttosto elevata a causa dei forti vincoli e dal gran numero di relazioni interne che le leggi fisiche impongono alle diverse parti di un suono o di un'immagine. Se la linea che trasmette i segnali che veicolano i messaggi di questo tipo è poco disturbata, una buona parte della ridondanza naturale originaria può essere eliminata senza perdita d'informazione. Per fare ciò bisogna

codificare i messaggi alla sorgente in modo più economico, in modo da eliminare gran parte della loro ridondanza (*compressione dell'informazione*), ed applicare quindi la trasformazione inversa ai segnali giunti a destinazione, in modo da recuperare la ridondanza originaria (*decompressione dell'informazione*). È chiaro che la codificazione ottimale di un messaggio, ma anche quella più rischiosa, è quella che elimina tutta la ridondanza.

Per ridurre la ridondanza, ossia comprimere l'informazione, si possono applicare i seguenti criteri: codificare in forma abbreviata le parti di messaggio che ricorrono più frequentemente, riservando le forme di codificazione più lunghe a quelle più improbabili; codificare separatamente l'elenco delle parti prive di relazioni e le regole che descrivono le relazioni tra le parti; codificare un programma generatore del messaggio invece del messaggio ecc. L'importanza della compressione dell'informazione nella teoria della comunicazione deriva dal fatto che i canali di trasmissione reali (linee elettriche, cavi telefonici, bande di frequenza delle portanti radiotelevisive, fibre ottiche ecc.) non hanno capacità illimitate. Per intensi flussi d'informazione essi si comportano come colli di bottiglia stretti.

I procedimenti di compressione e decompressione sono relativamente facili da calcolare quando si abbia una buona conoscenza delle forme di ridondanza caratteristiche dei messaggi da trasmettere, ma i calcoli possono diventare proibitivi se le forme di ridondanza sono complesse e imprevedibili. Inutile dire che la telefonia mobile, la trasmissione e la memorizzazione delle immagini digitali si avvale ampiamente di queste tecniche.

Ora, i segnali uscenti da uno strato d'unità nascoste di una rete a retropropagazione ben addestrata possono costituire un flusso d'informazione compressa che si produce in modo automatico. Nel caso più semplice questo effetto si ottiene con una rete a tre stadi: uno d'ingresso, uno d'uscita (entrambi con lo stesso numero N di terminali) e uno stadio intermedio formato dalle uscite di un numero K di unità nascoste sensibilmente inferiore a N. Addestrando questo sistema a fornire in uscita gli stessi segnali applicati in ingresso, purché la ridondanza originale sia maggiore di $1-K/N$, l'informazione che attraversa lo strato delle unità nascoste è conservata. Sperimentalmente si trova che un flusso d'informazione digitale (formata da sequenze di 0 e 1) entranti attraverso N linee parallele in una rete a retropropagazione con unità nascoste a stati continui si conserva se il numero delle linee intermedie è superiore a circa $K = \log_2 N$.

Recenti ricerche di simulazione hanno dimostrato che, a parità di ingressi e uscite, un processo di apprendimento funziona meglio se viene prima decomposto, ove sia possibile, in sottoprocessi indipendenti e quindi implementato in una rete costituita di molte sottoreti di moderata connettività, piuttosto che in un'unica rete di grande connettività (Fogelmann, 1990). Il fattore limitante decisivo, riguardo l'impiego delle reti a retropropagazione, è il tempo di convergenza dei processi di apprendimento, che cresce esponenzialmente coll'aumentare della connettività, e che si riduce notevolmente col frazionamento di questa.

2.2. La retropropagazione nella corteccia cerebrale

Uno degli aspetti più interessanti delle reti a retropropagazione sta in una certa loro somiglianza con l'organizzazione nervosa della corteccia cerebrale, sebbene nelle prime manchi l'analogo delle connessioni tra unità dello stesso stadio, tipiche della struttura corticale. Si sa, infatti, che ogni distinta area corticale (ogni area è caratterizzata da un'omogeneità citoarchitettonica specifica - ve ne sono circa trenta per la sola corteccia visiva umana) proietta le fibre dei propri neuroni efferenti su poche altre aree (in media tre o quattro) in modo approssimativamente topografico, e che queste proiezioni sono invariabilmente reciprocate da connessioni decorrenti in senso inverso.

Così il flusso d'informazione nervosa che nasce dagli organi sensoriali, si smista a cascata attraverso un reticolo di aree e areole corticali variamente specializzate, talvolta connesse trasversalmente, fino a sfociare nelle aree associative del sistema limbico e delle aree premotorie, prefrontali e frontali. Questo flusso è accompagnato da un controflusso che ripercorre simultaneamente a ritroso tutti gli stadi della cascata.

Il modello a retropropagazione suggerisce che una delle funzioni del controflusso sia di indurre le modificazioni sinaptiche che sono necessarie per determinare l'apprendimento adattativo di precise corrispondenze funzionali tra stimoli sensoriali e risposte associative e premotorie. Si spiegherebbe così, nei suoi tratti essenziali, la formazione delle facoltà percettive e della memoria a lungo termine.

Ma il controflusso in questione potrebbe avere anche altre funzioni, non interpretabili dai modelli finora studiati. In particolare potrebbe operare non solo per i processi formativi durevoli del genere appena considerato, ma anche per processi volatili, che richiedono solo eccitazioni transitorie e reversibili della rete neurale, come, ad esempio, quelli proposti da Grossberg (1987) per spiegare l'*attenzione selettiva*. Grossberg ha ipotizzato che questa importante funzione cerebrale, ampiamente studiata dagli psicologi sperimentali, sia generata da segnali di controflusso provenienti dagli stadi terminali del sistema percettivo e retroagenti su tutti gli stadi intermedi di questo processo. Tale retroazione interverrebbe a modulare le capacità di filtraggio ed elaborazione di ciascun'area corticale favorendo, stadio per stadio, la promozione delle componenti informazionali più significative e l'eliminazione di quelle irrilevanti (rispetto alle funzioni finali del processo percettivo). In effetti, vari esperimenti psicometrici dimostrano che ogni tipo d'attenzione selettiva dipende fortemente dalla particolare preparazione o aspettativa motoria del soggetto (Berlucchi, 1989). Per Walter Freeman (2000) il controflusso è l'espressione neurofisiologica della facoltà mentale detta *intenzionalità* che caratterizza il funzionamento cerebrale dei vertebrati superiori. Gli esperimenti di quest'autore hanno dimostrato che essa ha sua sede primaria nel sistema limbico e governa ogni forma d'acquisizione dell'informazione sensoriale.

Un'altra interessante idea riguardante la retropropagazione è stata proposta da Okajima (1990). Quest'autore attribuisce alle mappe topografiche corticali funzioni di trasduzione e trasformazione integrale simili a trasformate di Fourier locali. Quest'ipotesi, che ha il principale referente sperimentale nei risultati di Hubel e Wiesel (1962-1977) sulle risposte a stimoli ottici dei neuroni delle aree visive, risale ai lavori di Schwartz (1976) il quale ha dimostrato che attraverso una mappa logaritmico-polare del campo visivo, abbastanza simile a quella che la retina proietta effettivamente sull'area visiva primaria, e una successiva trasformata di Fourier, ogni immagine visiva può tradursi in una risposta caratteristica invariante per rotazione e dilatazione e, sotto certe condizioni, anche per trasformazione prospettica.

Se, come ha ipotizzato B.Cavanagh (1978), la mappa logaritmico-polare fosse preceduta da un'altra trasformazione di Fourier, si potrebbe ottenere un risultato invariante anche per traslazione piana. Quest'informazione terminale sarebbe infine utilizzata nel riconoscimento degli stimoli visivi. Naturalmente è difficile pensare che le reti nervose animali possano fare le trasformate di Fourier; si può invece immaginare che lo facciano in modi approssimativi o approssimativamente equivalenti, magari attraverso trucchi ingegnosi non ancora scoperti. Un indizio di questo è fornito dal fatto che i neuroni complessi della corteccia visiva sembrano rispondere agli stimoli in modi che assomigliano a trasformate di Fourier locali. È certo che la percezione visiva umana è effettivamente invariante per ampie rototraslazioni, dilatazioni e trasformazioni prospettiche dell'immagine. Bisogna tuttavia dire che l'elaborazione dell'informazione visiva nei cervelli dei mammiferi, specie in quelli umani, è un processo assai complesso, com'è riccamente descritto e commentato nel libro *Image and Brain* di Stephen Kosslyn (1996).

Okajima ha ipotizzato che la retropropagazione fornisca, stadio per stadio, la trasformazione inversa rispetto a quella che ha luogo attraverso la propagazione diretta. Ma poiché - riferendoci al modello matematico - le trasformazioni dirette rimuovono parte dell'informazione sensoriale, la completa rigenerazione, mediante trasformazioni inverse, di stati di eccitazione corticale simili a quelli formatisi durante la percezione originale, sarebbe possibile solo se l'informazione rimossa venisse reintegrata stadio per stadio. Si può tuttavia notare che quest'ipotesi, o una simile opportunamente migliorata, sembra accordarsi con quei risultati di fisiologia e psicologia della percezione che dimostrano come l'evocazione di ricordi visivi sia accompagnata da un'eccitazione corticale che procede a ritroso dalle aree associative temporali e parietali fino alle aree visive

secondarie (Farah, 1989; Kosslyn, 1996). Questo avviene come se l'evocazione mentale di immagini visive (ricordi, pensieri, sogni) comportasse la ricostruzione a ritroso di stati d'eccitazione corticale simili a quelli occorsi durante l'esperienza percettiva diretta. Si potrebbe dunque ipotizzare che la memoria percettiva (cognitiva o descrittiva) sia distribuita su tutte le aree corticali deputate al processamento dell'informazione sensoriale, anche quelle più periferiche, e non soltanto, come si è ritenuto in passato, su aree associative terminali specializzate per tale funzione, o in centri specifici del cervello.

2.3. I modelli olografici

I modelli olografici meritano solo un breve cenno perché si sono dimostrati sostanzialmente sbagliati. Negli aspetti concernenti la neurofisiologia l'approccio olografico si è principalmente ispirato ai risultati della ricerca trentennale condotta da Lashley (1950) sulla localizzazione dei ricordi nei cervelli animali. Gli esperimenti del celebre neurofisiologo sembravano indicare in modo inequivocabile che nel cervello dei mammiferi non esistono punti apicali di convergenza dell'informazione sensoriale e che la memoria di un evento sembra disseminata in modo piuttosto uniforme su tutta la corteccia cerebrale. L'evidenza derivava principalmente dal fatto che, dopo avere praticato incisioni in tutti modi possibili sulla corteccia di piccoli mammiferi addestrati a riconoscere degli oggetti, non si potevano notare apprezzabili perdite di memoria cognitiva.

Dopo la scoperta dell'olografia, prima in teoria (Gabor, 1959) e poi nella pratica (Leith e Upatnieks, 1962), diversi studiosi notarono una certa analogia tra le proprietà descritte da Lashley e quella degli ologrammi. In questi, infatti, l'informazione relativa a ogni dettaglio di un'immagine è sparpagliata uniformemente sull'intera estensione di una lastra fotografica nella forma della traccia lasciata dalla figura di interferenza tra l'onda luminosa originata da una comune sorgente laser e riflessa da una parte da un oggetto immagine (che fornisce l'immagine da olografare) e da un'altra da uno specchio o un altro oggetto che genera l'onda "chiave". È noto che investendo la lastra fotografica con l'onda chiave si produce, per diffrazione attraverso la traccia fotografica, un'onda identica a quella originariamente emessa dall'oggetto immagine. L'intensità ottica dell'immagine olografica è dovuta all'efficacia dell'interferenza costruttiva tra le componenti ondulatorie in coincidenza di fase diffratte dall'ologramma. Se l'onda evocatrice ha relazioni di fase genericamente diverse da quelle dell'onda chiave appropriata, al posto dell'immagine olografica si ha, a causa dell'interferenza distruttiva delle onde secondarie, un'emissione luminosa d'intensità debole e statisticamente uniforme.

L'evocazione delle immagini olografiche può prodursi anche in modo associativo, usando come onda chiave la radiazione filtrata attraverso un altro ologramma. Sembrò così che le ben note proprietà associative della memoria trovassero nel modello olografico la loro più naturale interpretazione.

Il primo modello olografico di memoria associativa, basato sull'ipotesi che nella corteccia cerebrale abbiano luogo propagazioni di tipo ondulatorio, fu proposto da Van Heerden nel 1962, e riproposto in sede neurofisiologica da Pribram (1969) e Barret (1969). Una versione temporale del modello olografico, basata sull'idea di un'integrazione temporale dei segnali emessi da una miriade di oscillatori e riferita all'apparente attività oscillatoria di origine subcorticale, fu proposta da Longuet-Higgins nel 1968. A questa fece immediatamente seguito un articolo di Gabor (1968) che propose un'interpretazione analogica del modello olografico-temporale e dimostrò come le operazioni tipiche dei processi di registrazione e riproduzione degli ologrammi (le trasformate di Fourier) possono trasportarsi nell'ambito di una teoria dell'integrazione di segnali stocastici o di aspetto stocastico (noise-like). Una migliore formulazione matematica di questo approccio, basata sulle operazioni coniugate di convoluzione e correlazione, venne fornita da Borsellino e Poggio (1974). Tuttavia, sebbene recentemente nella corteccia cerebrale si siano rilevati fenomeni di propagazione ondulatoria (onde di calcio lente nel tessuto gliale), oggi la maggioranza degli studiosi è concorde nel ritenere che l'attività elettrica macroscopica che si osserva nella corteccia dipenda da processi di sincronizzazione di circuiti eccitatori-inibitori di origine subcorticale (Elul, 1972) o

intracorticale (Freeman, 2000). Alcuni autori ritengono che questi fenomeni abbiano un ruolo essenziale nella sintesi corticale di capacità di risposta dipendenti da proprietà globali della stimolazione sensoriale (Eckhorn e Reitboeck, 1989). Risultati più recenti attribuiscono alle attività oscillatorie delle reti nervose ruoli ancora più interessanti. Ma di quest'argomento, e delle scoperte e idee avanzate in tempi più recenti circa la natura e la funzione dell'attività oscillatoria della corteccia cerebrale, ci occuperemo nella terza parte dell'articolo.

2.4. Alcuni aspetti dell'organizzazione corticale

Rivolgiamo ora l'attenzione ad alcuni aspetti del funzionamento delle reti nervose cerebrali che sono stati trascurati nei precedenti paragrafi, che riguardano il problema del processamento parallelo dell'informazione nervosa e il carattere distribuito e associativo della memoria.

Tornando a considerare le proprietà delle reti a retropropagazione multistratificate sommariamente descritte nel paragrafo 2.1, possiamo notare che ogni neurone è il punto di convergenza di segnali provenienti dai neuroni d'uno strato precedente e contemporaneamente d'emissione a ventaglio di segnali diretti ai neuroni di uno stadio seguente. Sono invece assenti le interconnessioni tra neuroni di uno stesso stadio. Anche le connessioni decorrenti in senso inverso sono ignorate, per quanto la teoria dell'apprendimento per retropropagazione le richieda per la formazione della memoria. La ragione di quest'ultima circostanza è piuttosto semplice: nessuno ha ancora trovato un modo di implementare in concreto le procedure di correzione dei pesi, sebbene queste siano matematicamente ben definite. Perciò anche questi modelli, come i vecchi perceptron, devono ritenersi modelli di reti nervose insufficienti e incompleti. Ora è opportuno osservare che non solo i neuroni con risposte a gradino, tipici dei perceptron, ma anche quelli con risposta sigmoidale, caratteristici delle reti a retropropagazione, appaiono in ultima analisi come semplici decisori a soglia. La gradualità delle risposte dei secondi è importante ai fini della realizzazione del processo d'apprendimento, ma è scarsamente rilevante per quanto riguarda il buon funzionamento della rete, una volta che i pesi sinaptici siano stati aggiustati a valori adatti.

La decisione dicotomica è l'operazione più elementare che si possa effettuare localmente in un processo parallelo suddiviso in stadi, mentre invece si deve ritenere che le unità colonnari della corteccia, considerate come sistemi locali di convergenza e divergenza simultanea di alcune migliaia di segnali nervosi, effettuino processi paralleli di tipo più complesso. La condizione che un processo locale o globale abbia un carattere più parallelo che seriale comporta che l'organizzazione spaziale del sistema e i fenomeni d'interazione sincronica tra le parti abbiano un ruolo tanto importante quanto l'ordinamento temporale delle relazioni causali tra le varie fasi del processo. Ora, da un punto di vista fisico-funzionale, si può osservare che nell'interazione tra le parti di un sistema suddiviso in stadi, che deve funzionare in modo spazialmente e temporalmente ben organizzato, a causa degli inevitabili fattori erratici di ritardo nella propagazione dei segnali nervosi, la sincronizzazione degli effetti si produce tanto meglio quanto più la struttura è simmetrica rispetto alla permutazione di parti omologhe. Le asimmetrie delle connessioni, infatti, tendono generalmente favorire la formazione di treni di segnali circolari permanenti forti consumatori d'energia ma di scarsa utilità funzionale. Per questa ragione è lecito aspettarsi che le connessioni tra i neuroni di uno stesso strato neuronale siano con buona approssimazione simmetriche rispetto allo scambio di neuroni omologhi.

Per studiare meglio il problema, e cogliere alcune importanti differenze tra lo stato delle cose reali e le proprietà generalmente assunte nei modelli, è opportuno fare una breve digressione in ambito neurofisiologico per analizzare con maggior dettaglio certe proprietà della corteccia cerebrale.

La corteccia cerebrale, che nel cervello umano ha mediamente lo spessore di circa due millimetri, è suddivisa in varie aree (un centinaio per il cervello umano) distinguibili per le diverse tessiture e architetture neuronali. Le aree corticali possono essere analizzate sia in senso verticale che orizzontale. Verticalmente, un'area appare suddivisa in strati neurali, tipicamente sei, differenziati per taglia e organizzazione funzionale. Orizzontalmente, invece, esse appaiono

piuttosto omogenee. Del resto, l'estensione di un'area è definita proprio sulla base del criterio di omogeneità citoarchitettonica. Tuttavia un esame più attento rivela che l'organizzazione orizzontale ha una componente più fine; la rete nervosa appare suddivisa in strutture colonnari a loro volta connesse secondo schemi di varia forma e disposizione: bande, raggruppamenti più o meno regolari, ecc. Ogni struttura colonnare contiene migliaia di neuroni di vari tipi e dimensioni e ha un diametro medio di alcune centinaia di micrometri. Interessa qui rilevare che le strutture colonnari hanno proprietà simili a quelle che è naturale richiedere ad un circuito elettronico per funzionare come unità locale di calcolo analogico parallelo. In particolare, neuroni omologhi, vale a dire disposti allo stesso modo entro queste unità, sembrano connessi in modi approssimativamente simmetrici.

Per agevolare la descrizione delle complesse architetture nervose corticali, metteremo ora in evidenza alcune proprietà semplificative generali. Nel sistema nervoso centrale non si trovano neuroni che siano simultaneamente eccitatori verso un tipo di neurone ed inibitori verso un altro tipo. Ogni neurone è ben caratterizzato come eccitatorio o inibitorio verso ogni altro neurone con cui possa entrare in contatto. I neuroni *piramidal*i, gli unici dotati d'afferenze ed efferenze extracorticali, sono eccitatori e costituiscono le unità di riferimento della corteccia. Gli altri possono essere chiamati *interneuroni*, poiché svolgono funzioni ausiliarie tra e per i piramidali. I piramidali hanno una struttura abbastanza tipica, caratterizzata dalla presenza di una formazione dendritica, che si espande ampiamente fuori del corpo cellulare, fornita di rigogliose ramificazioni alla radice (dendrita basale) e all'estremità superiore (dendrita apicale). Il tronco dendritico riceve principalmente segnali eccitatori d'origine esterna: fibre talamiche che portano informazione sensoriale di tipo specifico (fibre specifiche) o fibre di piramidali dello stadio precedente della cascata corticale, sia direttamente o attraverso interneuroni eccitatori (cellule stellate-spinose). Queste ultime raccolgono segnali provenienti dall'esterno convogliandoli sulle spine degli alberi dendritici di alcuni piramidali limitrofi mediante terminazioni assoniche ascendenti, a forma di coda di cavallo. Il dendrita basale riceve principalmente segnali eccitatori da alcune migliaia di piramidali omologhi o giacenti a livelli inferiori. Il dendrita apicale riceve principalmente segnali eccitatori da fibre talamiche aspecifiche e dai piramidali d'altre aree che inviano segnali di retropropagazione.

La maggior parte degli interneuroni agisce inibitoriamente sui piramidali. In particolare le cosiddette *cellule a canestro* si caratterizzano per esercitare intense azioni inibitorie sui piramidali limitrofi; ognuna di esse avvolge coi suoi terminali assonici i corpi a forma di piramide di alcune di queste cellule con innervazioni cestiformi. Esse ricevono segnali eccitatori principalmente dai piramidali della stessa unità colonnare e in misura minore da quelle adiacenti; cosicché all'interazione sinergica eccitatoria dei piramidali si contrappone, per il loro tramite, un'azione inibitoria antagonista, in parte autogena e in parte d'origine laterale. Un altro tipo di interneuroni inibitori sono le *cellule a candelabro*. Le loro tipiche diramazioni assoniche rivolte all'insù innervano direttamente gli assoni di alcuni piramidali limitrofi agendo come interruttori che bloccano l'emissione dei segnali (Asanuma e Crick, 1987).

Considerando che la sola interazione eccitatoria tra i piramidali - omologhi e non - di una medesima unità colonnare porterebbe rapidamente il sistema al parossismo eccitatorio, si comprende quanto siano essenziali per il buon funzionamento della rete nervosa gli interneuroni inibitori. Ma affinché l'attività inibitoria abbia l'effetto di limitare quella eccitatoria, senza tuttavia impedirla del tutto, bisogna che la dipendenza dell'azione inibitoria sui piramidali cresca con legge di potenza maggiore di quella eccitatoria. Poiché le cellule a canestri sono eccitate dagli stessi piramidali, a tal fine è sufficiente che l'azione inibitoria si trasmetta con legge di potenza maggiore di uno. A questo proposito è interessante considerare il fatto che, a causa di un meccanismo d'attivazione a doppio sito, l'azione esercitata sui recettori presenti nei corpi dei piramidali dei segnali inibitori prodotti dalle cellule a canestri dipenda quadraticamente dalla concentrazione del neurotrasmettitore inibitorio, l'acido γ-amminobutirrico (GABA) (Borman e Clapham, 1985). Circa le possibili modalità di funzionamento delle unità colonnari e delle interazioni tra unità diverse non

si è ancora raggiunta una comprensione soddisfacente. Le ricerche di Hübel e Wiesel sulle risposte dei neuroni della corteccia visiva primaria a stimoli visivi indicano che le complessità delle interazioni locali aumentano dal quarto strato (quello di arrivo dei segnali talamici) verso gli strati superiori e inferiori, cioè secondo una struttura approssimativa a doppio cono. È lecito presumere che le proprietà più importanti dipendano in modo essenziale dal sincronismo dei processi d'interazione locale. Tuttavia, anche i fenomeni d'ordinamento temporale potrebbero avere grande importanza. Sarebbe perciò poco prudente ritenere che il gioco di sinergismi e antagonismi entro le strutture colonnari produca risultati più rilevanti dal punto di vista dei processi paralleli e meno da quello dei processi seriali.

2.5. Stabilità strutturale, multimodalità e isterèsi

Rivolgiamo ora l'attenzione al problema delle modellizzazione delle strutture corticali locali, privilegiando gli aspetti che possono riguardare maggiormente la funzionalità e i comportamenti tipici dei processi paralleli.

I modelli di reti a retropropagazione descritti nei precedenti paragrafi inducono facilmente a considerare le aree corticali come filtri analizzatori dei flussi d'informazione nervosa (sensoriale, motoria, centrale ecc.) e codificatori delle loro componenti significative. Questi sistemi sarebbero capaci di mappare con continuità, nei modi più convenienti, lo "spazio degli stimoli possibili" in "spazi di risposte corrispondenti". Ciò che viene ignorato in questa visuale è la possibilità di avere risposte s*tabili*, *multimodali* e *isteresiche*.

Per *stabilità* delle risposte intendiamo la loro sostanziale invariabilità rispetto a piccole e arbitrarie variazioni degli stimoli. Per *multimodalità* la possibilità che allo stesso stimolo possano corrispondere più risposte diverse. Per *isteresi* la tendenza di ogni risposta a permanere a dispetto anche di ampie variazioni dello stimolo e a cambiare bruscamente quando lo stimolo è variato oltre certi limiti.

La multimodalità è comunemente osservata negli esperimenti sulla percezione (figure ambigue, interpretazioni dipendenti dal contesto, ecc.). Se le risposte sono multimodali, l'isteresi si manifesta nel seguente modo: per transitare da una risposta stabile a un'altra bisogna variare gli stimoli oltre certi limiti critici; e quando si tenta di riottenere la risposta precedente, i limiti critici della transizione inversa risultano dislocati altrove rispetto ai precedenti. Da un punto di vista matematico l'isteresi è una proprietà generale dei sistemi *strutturalmente stabili*, cioè tali che un'arbitraria perturbazione non ne distrugga le proprietà (comportamenti, varietà delle risposte, condizioni di funzionamento ecc.). Questa non è solo una condizione matematica cui un modello deve soddisfare per avere senso, ma anche una condizione fisicamente necessaria per il buon funzionamento di una rete nervosa reale. Essa richiede che la rete sopporti perturbazioni e persino danni di una certa entità senza subire gravi conseguenze funzionali.

Stabilità, multimodalità e isteresi sono proprietà che emergono con evidenza da tutti gli esperimenti di psicologia: dall'analisi delle capacità di discriminazione percettiva di dettagli elementari alla percezione gestaltica delle forme complesse, dall'evocazione mnemonica alla cognizione concettuale. Esse si presentano come regole dominanti e straordinariamente importanti di tutti i processi mentali. Ora, è interessante considerare che queste proprietà sono anche caratteristiche dei sistemi non lineari costituiti di parti uguali interagenti in modi almeno approssimativamente simmetrici e sottoposte all'azione di parametri di controllo. Siamo così portati in modo naturale a studiare modelli di reti nervose capaci di funzionare come sistemi multistabili a controllo multiparametrico. A questo proposito conviene osservare che la ragione essenziale per cui la stabilità, la multimodalità e l'isteresi non si manifestano nei perceptron e nei modelli a retropropagazione, considerati in precedenza, è sostanzialmente dovuta sia al carattere elementare e unimodale delle unità di processo parallelo locale (decisori a soglia) sia all'assenza di retroazioni positive tra stadi successivi in condizioni di funzionamento effettivo. Dovremmo dunque attenderci progressi decisivi rimpiazzando i decisori semplici delle unità nascoste con *decisori complessi* capaci di comportamenti multimodali.

Se i fenomeni di multimodalità e isteresi osservati nei sistemi nervosi reali non dipendano solo da interazioni neurali aventi luogo localmente nei singoli strati (multistabilità locale), ma anche da fenomeni di retroazione di ciascuno stadio sui precedenti (multistabilità globale) dovremmo aspettarci che il gioco delle interdipendenze funzionali tra livelli di multistabilità locali e globali sia capace di generare processi complessivi di straordinaria ricchezza. Abbiamo già potuto considerare come le reti a retropropagazione siano capaci di memorizzare semplici corrispondenze dirette tra certi stimoli e certe risposte. Abbiamo anche intravisto la possibilità che modelli più appropriati siano capaci di spiegare funzioni di memoria assai più complesse: l'attenzione selettiva, la ricostruzione dello stimolo originale mediante retropropagazione, la dipendenza dal contesto dei ricordi (fenomeno che potrebbe interpretarsi come un effetto isteresico), l'organizzazione automatica delle contestualizzazioni dei ricordi secondo criteri di similarità e peculiarità distintive (capacità di categorizzazione), la produzione di pseudoricordi o chimere (evocazioni oniriche o fantastiche dotate di significati non banali) ecc. Cerchiamo dunque di descrivere con una certa generalità alcuni modelli che riescono ad interpretare le proprietà ora descritte.

2.6. Un modello generale di decisore complesso

Da un punto di vista matematico un insieme di n neuroni eccitatori (piramidali), controllati da afferenze esterne, interagenti direttamente tra essi e anche indirettamente per il tramite di interneuroni inibitori, costituisce un sistema multistabile che può essere matematicamente descritto da un sistema di n equazioni non lineari dipendente da parametri di controllo. In tale rappresentazione lo stato di eccitazione (frequenza di scarica) del generico piramidale i a un certo istante t è rappresentato da una variabile reale positiva $y_i(t)$. Essa può ritenersi una funzione istantanea $F[v_i(t)]$ del potenziale postsinaptico $v_i(t)$ che dipende da tutte le afferenze, eccitatorie e inibitorie del piramidale i. Poiché la frequenza di scarica di un piramidale si annulla per depolarizzazioni negative e inoltre non può superare un certo valore massimo, determinato dal periodo di refrattarietà dei potenziali d'azione (*spikes*), $F(v)$ avrà una forma sigmoidale generalmente asimetrica. Possiamo convenzionalmente assumere che i valori di F si estendano tra un minimo 0 e un massimo 1. Un'espressione tipica è

$$F(v) = \frac{1}{1+\exp(-4\lambda v)}, \qquad (1)$$

dove $\lambda > 0$ è la pendenza della sigmoide nel suo punto di flesso v=0. La forma precisa di questa funzione non è rilevante, poiché il funzionamento dei sistemi multistabili interessa assai più per i loro aspetti qualitativi piuttosto che per quelli quantitativi. Il potenziale postsinaptico v_i del piramidale i dipende a sua volta dagli stati d'eccitazione di tutti i neuroni afferenti negli istanti immediatamente precedenti e da un eventuale stimolo esterno x_i direttamente applicato al piramidale che rappresenta il parametro di controllo dell'unità nervosa. Assumiamo per semplicità che v_i dipenda a sua volta dalle stesse attività dei piramidali (direttamente oppure tramite gli interneuroni) con ritardi temporali che per semplicità potremo assumere pari a un multiplo di un certo intervallo elementare τ (ritardo sinaptico). Potremo pertanto esprimere i potenziali postsinaptici al tempo $t+\tau$ nella forma $v_i(t+\tau) = g_i\,[y(t), y(t-\tau), y(t - 2\tau),\ldots]+x_i$, dove $g_i\,(y)$ sono opportune funzioni dipendenti dagli accoppiamenti sinaptici in gioco. Si otterrà così un sistema di equazioni che descrivono la dipendenza dello stato di eccitazione di tutti i piramidali della rete, vale a dire dello *stato della rete* all'istante $t+\tau$, che indicheremo semplicemente con $y(t+\tau)$, dagli stati di tutti i neuroni (piramidali e non) agli istanti t, $t-\tau$, $t-2\tau$, …Otterremo così il sistema di equazioni

$$y_i(t+\tau) = F\left\{g_i\left[y(t), y(t-\tau), y(t-2\tau),\ldots\right]+x_i\right\};\quad i=1,2,\ldots,n$$

Le soluzioni di questo sistema descrivono l'evoluzione temporale dello stato della rete per certi valori dei parametri di controllo x_i. In generale, in corrispondenza a fissati valori di questi parametri, il sistema ammetterà una moltitudine di soluzioni generalmente corrispondenti a stati stabili o instabili, o a comportamenti oscillatori e persino caotici di vario genere. Tuttavia, se sono soddisfatte certe condizioni di simmetria almeno approssimativa degli accoppiamenti tra i neuroni,

il sistema ammetterà almeno uno stato stabile per arbitrari valori dei parametri di controllo. In tal caso, agendo su questi parametri, si potranno determinare brusche transizioni tra stati stabili. In questo caso si può dimostrare con tutta generalità che esiste una specie di funzione $U(y, x)$, dipendente dallo stato della rete e dai parametri di controllo, (funzione di Ljiapunov) che ha i minimi relativi in corrispondenza degli stati stabili. Essa rappresenta un paesaggio di *buche di potenziale* nello spazio degli stati $y \equiv \{y_i\}$. Agendo sui parametri di controllo la buca può deformarsi in modo tale da determinare il brusco "versamento" di uno stato da una configurazione stabile a un'altra. Si spiega in questo modo la multistabilità del sistema.

2.7. Il modello di Hopfield

Il più semplice modello di rete neurale del tipo ora descritto è quello di J.J.Hopfield. La rete neurale di Hopfield è formata da un solo strato di neuroni che interagiscono simmetricamente tra loro, ma senza auto-interazione. Nella prima versione di questo modello (1982) si assume che, al variare del potenziale postsinaptico, i neuroni producano la risposta a gradino tipica del modello di MacCulloch e Pitts; in quella successiva (1984) si assume che la risposta dei neuroni abbia la forma di una sigmoide simmetrica. Nel primo caso la risposta di un neurone viene posta uguale alla funzione a gradino di Heaviside: $F(x)=0$ per $x < 0$; $F(0)=1/2$; $F(x)=1$ per $x > 0$. Nel secondo caso si assume per $F(x)$ l'espressione della formula (1). In entrambi i casi si assume che i potenziali postsinaptici v_i dipendano linearmente dallo stato della rete y secondo l'equazione:

$$v_i = \sum_j W_{ij} y_j , \qquad (2)$$

dove le quantità W_{ij} (pesi sinaptici) possono assumere entrambi i segni e soddisfano alla condizione di simmetria $W_{ij} = W_{ji}$ e di non auto-interazione $W_{ii}=0$. Le equazioni della rete diventano allora

$$y_i(t+\tau) = F\left[\sum_j W_{ij} y_j(t) + x_i\right]. \qquad (3)$$

Per $x_i >> 1$ o $x_i << 0$ lo stato di eccitazione del neurone i tende ad evolvere nel tempo verso un valore u_i uguale a 1 o a 0 per il modello con le risposte a gradino o molto vicino a 0 e 1 per il modello a risposta sigmoidale se la costante λ che appare nella formula (1) è sufficientemente grande.

L'idea di Hopfield fu di attribuire alla multistabilità funzioni di memoria. Infatti, ogni stato stabile della rete costituisce un attrattore per gli stati iniziali prossimi ad esso. Inizializzando la rete ad uno di questi, lo stato evolverà rapidamente verso l'attrattore. Pertanto, se si riesce a fare in modo che certi stati di eccitazione della rete siano *appresi* come stati stabili, ciascuno di questi potrà ritenersi un ricordo evocabile da una parte del ricordo (*content-addressable memory*). In tale modo la rete potrà funzionare come una memoria auto-associativa. Per ottenere questo risultato bisogna trovare il modo di (1) imporre alla rete di assumere uno stato voluto, (2) di agire sui pesi W in modo che questo stato diventi un attrattore, in modo, cioè, che si formi una buca di potenziale in corrispondenza di questo stato.

Per forzare un neurone i ad assumere un certo stato u_i basta agire sui parametri di controllo facendo aumentare il valore di x_i se si desidera che la risposta del neurone i sia 1 o prossima a 1, facendolo invece diminuire, portandolo magari a valori negativi, se si desidera che la risposta sia 0.

Imponendo simili valori estremi ai parametri di controllo si può forzare la rete a memorizzare uno stato $u=(u_1, u_2, \ldots, u_n)$, $u_i=0,1$, cioè uno stato costituito da un insieme di neuroni massimamente attivi e dall'insieme complementare di neuroni silenti.

Per ottenere la memorizzazione dello stato si può applicare una regola di variazione dei pesi sinaptici simile a quella proposta da Hebb nel 1949: le connessioni sinaptiche tra due neuroni devono rinforzarsi se i neuroni fanno la stessa cosa (se sono entrambi accesi o entrambi spenti), indebolirsi se fanno cose opposte (se uno è acceso mentre l'altro è spento). Possiamo descrivere matematicamente questa condizione stabilendo la condizione che le variazioni dei pesi ΔW_{ij} siano legate agli stati forzati u_i, u_j dei neuroni i e j dalla relazione

$$\Delta W_{ij} = \alpha(u_i - 1/2)(u_j - 1/2) ,$$

dove α è una quantità positiva dipendente dalla durata della stimolazione.

Queste regole sembrano in buon accordo coi fenomeni di *long term potentiation* (LTP) (Bliss e Lomo, 1973; Levy e Steward, 1979) e *long term depression* (LTD) (Stanton e Sejnowski, 1989) rilevati studiando gli effetti di stimolazioni combinate sui neuroni piramidali dell'ippocampo che, com'è noto, è una delle strutture del cervello implicate nella formazione dei ricordi. Il fenomeno della LTP si manifesta come rinforzo (temporaneo) del contatto sinaptico quando una scarica sinaptica coglie il neurone bersaglio in uno stato di depolarizzazione di ampiezza maggiore di circa 20 mV, mentre quello della LTD consiste nell'indebolimento del contatto quando la scarica lo coglie in uno stato di depolarizzazione inferiore a detto valore. Tenendo conto che la depolarizzazione di un neurone è sufficientemente alta solo se il neurone è attivo, si deduce che la regola di Hebb-Hopfield sarebbe completamente confermata, almeno per le sinapsi eccitatorie dei piramidali, se, oltre alle proprietà rilevate, avvenisse anche il rinforzo sinaptico di neuroni inattivi.

Il funzionamento della rete di Hopfield come memoria associativa si spiega nel seguente modo. Se nella (3) si pone x_i =0 e gli stati memorizzati sono rappresentati da vettori ortogonali dello spazio *N*-dimensionali delle variabili *y*, gli stati stabili della rete sono proporzionali agli autovettori della matrice dei pesi. Essi coincidono con i minimi della funzione

$$E(y) = -\frac{1}{2}\sum_{ij} W_{ij} y_i y_j , \qquad (4)$$

che può interpretarsi come una funzione di energia potenziale. È dunque naturale che l'evoluzione temporale della (3) converga questi minimi, che funzionano pertanto come bacini d'attrazione per gli stati della rete. Se gli stati memorizzati non sono ortogonali, gli autovettori di *W* non sono più proporzionali agli stati memorizzati. Ciò nonostante, grazie alla marcata non linearità compressiva delle equazioni dinamiche (3), se da un certo momento in poi si pone x_i =0, gli stati tendono comunque ad evolvere uno stato memorizzato. Ciò è possibile perché l'equazione non lineare

$$y_i = F\left[\sum_j W_{ij} y_j\right] \qquad (5)$$

differisce in modo sostanziale dall'equazione lineare agli autovalori

$$\sum_j W_{ij} y_j = \alpha y_i .$$

Questo significa che la rete riesce a discriminare pattern che differiscono tra loro soltanto in parte.

Il richiamo associativo degli stati memorizzati nella rete di Hopfield funziona nel seguente modo. Una volta formatasi la matrice dei pesi *W*, le unità della rete sono inizializzate applicando uno stimolo esterno transitorio rappresentato da certi valori x_i dei parametri di controllo. All'istante iniziale la rete comincia ad evolvere secondo la legge dinamica imposta dell'equazione (3), e da questo momento in poi lo stato tende a convergere verso lo stato memorizzato più prossimo allo stimolo inizializzatore.

Si può notare come il modello di Hopfield sia simmetrico rispetto allo scambio di 0 con 1. Questo ha come conseguenza che creando la memoria di uno stato *u* si ottiene automaticamente anche quello dello stato complementare u^c, che si ottiene scambiando tra loro gli 0 e gli 1 di *u*. Pertanto addestrando la rete a memorizzare un insieme di pattern si ottiene inevitabilmente anche la memorizzazione dei corrispondenti pattern negativi. Questa proprietà, che a prima vista può sembrare utile e interessante, è in realtà la causa di un inconveniente molto serio: la formazione non desiderata di combinazioni di stati positivi e negativi dotati di notevole stabilità (stati spuri).

All'epoca in cui Hopfield propose questo modello erano già noti gli importanti risultati teorici ottenuti da Giorgio Parisi (1979) circa le proprietà di un modello matematicamente equivalente, che andava sotto il nome di *vetro di spin* (S.Kirkpatrick e D.Sherrington, 1978). Più tardi fu possibile evidenziare importanti proprietà d'auto-organizzazione degli stati stabili di questi sistemi, in particolare dei ricordi impressi con criteri hebbiani. Precisamente la loro tendenza a organizzarsi spontaneamente secondo uno schema *ultrametrico*, una sorta d'ordinamento gerarchico ad albero continuo, conformemente al criterio d'ascendenza verso gradi di similarità crescenti (Parisi, 1983; Mèzard *et al*., 1983; H.Gutfreund, 1988). Questo comportamento trova spiegazione nel fatto che la

facilità di transizione da un ricordo all'altro è tanto maggiore quanto più questi ricordi possiedono una porzione comune, ossia quanto maggiore è il numero dei neuroni eccitati in comune nei ricordi evocati. Si noti che, per modelli di reti di tipo più generale, non accade necessariamente che un insieme di ricordi, o stati stabili della rete, siano collegati da relazioni di reciproca accessibilità secondo un'organizzazione gerarchica. Ad esempio le strutture topologiche delle relazioni di similarità riscontrate in certe analisi statistiche assomigliano a distanze tra punti di uno spazio multidimensionale. La proprietà mostrata dalle reti di Hopfield dipende essenzialmente dal fatto che, avendo forzato la rete ad assumere certi ricordi, non si può impedire che si generino, proprio a causa del carattere combinatorio-lineare delle azioni postsinaptiche, stati stabili di similarità intermedie (stati spuri) (J.Hopfield, D.I.Feinstein e R.G.Palmer, 1983; D.J.Amit, H.Gutfreund e H.Sompolinsky, 1985, 1987). Ed è proprio per intercalazione di questi stati spuri tra i ricordi autentici che si forma un sistema complessivo di stati stabili organizzati secondo una topologia ultrametrica. A prima vista, la formazione degli stati spuri potrebbe essere interpretata come una specie di capacità di generare spontaneamente contenuti nuovi; ma a un'analisi più attenta ci si rende conto che tali presunte creazioni sono in realtà chimere illogiche prive di sensate relazioni con i ricordi autentici.

Purtroppo il modello di Hopfield è apparso poco convincente ai neurofisiologi e a tutti coloro che nel frattempo cercavano modelli più realistici. I difetti strutturali più spesso denunciati, a parte quelli relativi alla formazione degli stati spuri, sono la precisa *simmetria degli accoppiamenti* e l'*ultraconnessionismo*, cioè la condizione che ogni neurone sia connesso con tutti gli altri. Tuttavia queste critiche possono essere facilmente respinte in primo luogo facendo valere la dimostrazione matematica che il modello di Hopfield è strutturalmente stabile per moderate violazioni della simmetria dei coefficienti d'accoppiamento, e in secondo luogo proponendolo solo come modello di unità colonnari corticali (nelle quali, infatti, è ravvisabile sia la connessione massimale che una simmetria approssimativa degli accoppiamenti tra neuroni omologhi) o d'interazioni neuronali a raggio d'azione limitato. D'altronde non è poi tanto che la limitata connettività tra i neuroni di due aree della corteccia cerebrale implichi che non si possa avere un grado di connettività assai maggiore di quella che va da una fibra uscente da un'area a all'area irradiata dall'espansione dalle diramazioni della fibra nell'area bersaglio. Se gli ingressi di uno stesso segnale in un'area corticale vengono moltiplicati e applicati a distanze medie dell'ordine di grandezza del raggio d'interazione neuronale, gli effetti del segnale sull'altra area possono coprire in modo abbastanza uniforme l'intera area.

A parere dell'Autore, i difetti più rilevanti del modello sono invece i seguenti: 1) la simmetria tra azioni inibitorie ed eccitatorie; 2) l'eccessiva semplicità nella rappresentazione dei ritardi temporali; 3) l'improponibilità della regola di Hebb per le variazioni dei coefficenti sinaptici ad effetto inibitorio.

Riguardo al primo punto, c'è da osservare che i potenziali postsinaptici del modello di Hopfield sono combinazioni lineari delle attività dei neuroni afferenti, con coefficienti positivi per le azioni eccitatorie e negativi per quelle inibitorie; mentre, a causa della molteplicità dei siti di attivazione dei canali sinaptici (da due a tre, secondo la natura del neurotrasmettitore), si deve assumere che gli effetti postsinaptici siano funzioni non lineari delle ampiezze dei segnali afferenti. Di conseguenza, poiché gli interneuroni intervengono moltiplicativamente nella trasmissione dei segnali, sarebbe naturale attendersi una sostanziale asimmetria tra gli effetti delle afferenze eccitatorie dirette (piramidale $\rightarrow$ piramidale) e quelli delle afferenze inibitorie indirette (piramidale $\rightarrow$ cellula a canestri $\rightarrow$ piramidale).

Tale asimmetria potrebbe avere un ruolo importante nell'impedire che, a causa del carattere sinergico dell'interazione eccitatoria, le attività dei neuroni raggiungano rapidamente il loro livello di saturazione (mentre invece, nel modello di Hopfield, è proprio la saturazione che limita l'attività neurale). Ciò sarebbe in accordo col fatto che le frequenze di sparo dei piramidali osservate *in vivo* superano di rado i 100 *spikes* al secondo, mentre il valore di saturazione oltrepassa il migliaio di spikes al secondo.

Riguardo al secondo punto, l'assunzione di un comune tempuscolo di ritardo sinaptico τ per tutti i neuroni contrasta chiaramente col fatto che gli interneuroni inibitori introducono ritardi suppletivi rispetto alle interazioni eccitatorie dirette. D'altronde è noto che l'improvvisa stimolazione di una popolazione di neuroni corticali genera in un primo momento una breve e intensa fase eccitatoria, alla quale fa seguito una fase inibitoria di durata maggiore, al termine della quale il sistema generalmente si spegne o continua ad oscillare.

Per quanto riguarda il terzo punto, bisogna considerare che le variazioni dei pesi sinaptici secondo la regola di Hebb non è applicabile se il bersaglio sinaptico di un piramidale è un interneurone, che può trovarsi in uno stato di depolarizzazione generalmente indipendente da quello dei piramidali a cui afferisce.

Dunque, se si ritiene che l'asimmetria e lo sfasamento temporale tra le componenti eccitatorie e inibitorie, unitamente al carattere non hebbiano della stimolazione inibitoria dei piramidali, abbiano importanti ruoli funzionali nelle reti nervose reali, allora è particolarmente significativo rilevare che questi ruoli funzionali non possono essere adeguatamente interpretati dal modello di Hopfield.

2.8. Generi prossimi e differenze specifiche

Nonostante i difetti elencati, il modello di Hopfield ha suscitato molto interesse tra i fisici, i quali in pochi anni hanno saputo esibire una profusione di risultati teorici su varianti e aspetti particolari (A.Crisanti e H.Sompolinsky, 1987; K.E.Kürten e J.W.Clark, 1987; B.Lautrup, 1988 ecc.). In particolare è stata studiata la possibilità di farlo funzionare come sistema di memoria auto-organizzante. Si è così scoperto un procedimento, della *ricottura* (annealing), che per un certo periodo è sembrato adeguato a sfruttare le risorse del modello. Si tratta, in breve, di simulare una specie di "riscaldamento" della rete stimolando i neuroni con segnali rapidamente fluttuanti in modi casuali attorno a certi valori medi di uno stimolo fisso assunto come frammento inizializzatore di un ricordo. In queste condizioni lo stato della rete transita erraticamente entro un insieme di stati, generalmente instabili, costituenti una specie di *intorno* topologico comune a uno o più stati stabili. Aumentando l'ampiezza delle fluttuazioni, ossia la "temperatura" della rete, l'intorno si espanderà abbracciando nuovi stati stabili fino a invadere, oltre una certa "temperatura critica", l'intero spazio degli stati possibili. Diminuendo la temperatura esso si restringerà in un intorno più piccolo dal quale alcuni stati stabili potranno restare esclusi dal precedente intorno. Chiaramente lo stato memorizzato, essendo più stabile degli altri, continuerà a rimanere confinato in questo intorno. Infine, alla temperatura nulla, l'intorno si sarà ridotto ad un singolo stato stabile che con grande probabilità coinciderà con lo stato memorizzato.

Evidentemente, la probabilità che al diminuire della temperatura il sistema resti confinato in uno piuttosto che un altro intorno, dipenderà dai valori medi dei segnali stimolatori e saranno favorite le restrizioni ad intorni di stati stabili maggiormente prossimi a quei valori medi. Pertanto, a ogni definita temperatura, lo spazio degli stati si presenta suddiviso in un insieme d'intorni non comunicanti. Questa suddivisione si *ramifica* in suddivisioni via via più fini in corrispondenza di temperature via via minori, fino a ridursi, a temperatura nulla, alla semplice collezione degli stati stabili. Invece, sopra la temperatura critica, tutti i rami convergono al tronco. L'organizzazione gerarchica degli stati stabili, secondo relazioni di maggiore o minore similitudine, consiste proprio in questa diramazione degli intorni durante i processi di "ricottura".

Il procedimento di recupero di un ricordo può effettuarsi nel seguente modo: prima "riscaldando" la rete, sotto l'influenza di uno stimolo medio evocatore (frammento di ricordo), a una temperatura tale che l'intorno possa "catturare" lo stato stabile corrispondente al ricordo completo; poi "raffreddandola" fino a ottenere il confinamento del ricordo completo entro un'intorno abbastanza ristretto.

La prima idea che sta alla base di questo procedimento è che esso dovrebbe favorire il disincaglio dello stato fluttuante dagli attrattori spuri. La seconda è che in esso si può ravvisare una tendenza all'organizzazione dei ricordi che ricorda la *categorizzazione per generi prossimi e*

differenze specifiche (Aristotele, 306-367 a.C.). Infatti, interpretando la risalita verso il tronco, determinata dal riscaldamento, come la fase d'individuazione del genere prossimo di un insieme di memorie simili, e la discesa verso un ramo terminale, determinata dal raffreddamento, come fase che di recupero di una differenza specifica, è evidente che un "riscaldamento" moderato, sotto l'effetto di uno stimolo medio iniziatore, applicato a una rete che già si trova in un certo stato avrà l'effetto di farla accedere con maggiore probabilità a un genere prossimo di questo stato.

Si potrebbe intravvedere in ciò una specie di fenomeno di *permanenza del contesto*. Purtroppo questo interessante comportamento delle reti di Hopfield ha un serio difetto: anche se i ricordi sono "ortogonali", cioè privi di parti comuni, la probabilità di evocare chimere insensate, invece che ricordi autentici, continua a rimanere non trascurabile. Così la pretesa capacità categorizzante si rivela in realtà piuttosto illusoria.

2.9. Alla ricerca della memoria categorizzante

Ponendo a confronto il comportamento del modello di Hopfield con quello dei sistemi nervosi reali, si possono osservare alcune analogie ma anche differenze piuttosto rilevanti. Che la memoria animale, e umana in particolare, non sia un semplice archivio di dati indirizzati, ma possieda una specie di capacità di organizzazione automatica dei ricordi, secondo schemi di categorizzazione abbastanza indipendenti dalle modalità di acquisizione, traspare con grande evidenza da tutta la letteratura scientifica sul linguaggio: dalle indagini della psicologia gestaltica e cognitivista alle ricerche sull'intelligenza artificiale (C.Cornoldi, 1978; H.Gardner, 1988). Se la percezione del genere prossimo è una funzione mentale essenziale nel riconoscimento delle forme, altrettanto lo è il discernimento delle differenze specifiche. La prontezza con cui la mente umana è in grado di "percepire" anche i generi più astratti e le più sottili differenze tra i dati della percezione non ha bisogno di essere documentata. Altrettanto evidente è la tendenza più o meno pronunciata dell'evocazione di ricordi alla permanenza dell'informazione di contesto.

Ma il comportamento della memoria animale diverge in vari punti da quello esibito dal modello descritto nel paragrafo precedente, e rivela persino aspetti apparentemente contraddittori. In primo luogo il processo di "ricottura" delle reti di Hopfield deve effettuarsi in modo *adiabatico*, cioè in tempi molto lunghi rispetto alla durata media di una fluttuazione, mentre i ricordi della memoria animale reale hanno tempi d'evocazione confrontabili con le durate delle fasi d'eccitazione della rete nervosa ($\cong$ 1/10 sec.). Inoltre nel modello l'individuazione del genere prossimo è ottenibile solo calcolando la media temporale degli stati erraticamente fluttuanti, mentre il recupero di una differenza specifica può essere ottenuto sottraendo tale media dallo stato finale della rete ricotta e raffreddata. Ciò significa che in pratica, l'effettivo funzionamento di questo tipo di memoria richiede apparati di calcolo ausiliari che non si sa né a quali strutture nervose dovrebbero fare riferimento né mediante quali strutture accessorie della rete neurale potrebbero essere implementati.

Un problema simile si presenta in relazione alla registrazione dei ricordi. Alcuni autori ritengono che l'organizzazione dei ricordi appartenenti ad un genere comune cominci dalla memorizzazione di un prototipo, o comunque avvenga in relazione alla memorizzazione di uno stereotipo. Verrebbe da pensare che il cervello registri solo differenze specifiche tra gli stimoli nuovi e quelli già memorizzati. Ciò troverebbe una conferma nel fatto che, nel riconoscimento delle forme, il processo evocativo sembra funzionare in modo tale da evidenziare le novità neutralizzando gli aspetti costanti, o rimuovendo la componente relativa al genere, ad esempio inibendo per assuefazione la percezione del contesto. Questo tipo di funzionamento, che nei modelli finora considerati potrebbe ottenersi in modi alquanto artificiosi, è stato formalizzato da T.Kohonen e E.Oja (1986, 1987), i quali hanno assunto l'esistenza di procedure ricorsive di *ortogonalizzazione* preliminare degli stimoli da ricordare rispetto a quelli già registrati.

Un'altra rilevante discrepanza è rinvenibile nel fatto che nelle reti nervose reali gli stimoli agiscono in modi immediati e transitori, in contrasto con la necessità della loro permanenza nei processi di ricottura artificiali. Inoltre le categorizzazioni della memoria reale non sono

necessariamente *ad albero*, ad esempio quelle della memoria linguistica possono avere strutture *a rizoma*, *a reticolo*, ecc.: vale a dire che una stessa differenza specifica può fare capo a più generi prossimi diversi, come se uno stesso frammento di un ricordo appartenesse a topologie diverse. In altri termini le relazioni di transitività tra intorni di stati stabili d'una rete reale devono appartenere a topologie intermedie tra quelle ultrametriche e quelle metriche, con la possibilità di un controllo multiparametrico degli accessi ai generi prossimi e delle riduzioni alle differenze specifiche.

Poiché tutti questi difetti sono concomitanti ai difetti strutturali messi in evidenza alla fine del paragrafo precedente, si presenta in modo naturale il problema di capire se reti neurali meno difettive dal punto di vista strutturale ammettano comportamenti più soddisfacenti. Cercando di correlare le proprietà strutturali che dovrebbe avere un modello più evoluto di quello di Hopfield (asimmetria tra gli effetti postsinaptici eccitatori e quelli inibitori, anticipazione dei primi sui secondi, regole neurologicamente plausibili di variazione dei coefficienti sinaptici per la componente inibitoria, ecc.) con alcune proprietà attendibili delle reti reali (transitorietà della stimolazione, rapidità del processo di categorizzazione, selezione delle pure differenze specifiche, ecc.) si giungerebbe a prospettare, invece del procedimento di "ricottura" su descritto, il seguente principio di funzionamento: *la stimolazione della rete ha un effetto rapido e transitorio; essa innesca l'accesso immediato al genere prossimo nella fase eccitatoria, cui segue la rapida selezione della differenza specifica nella fase inibitoria* (senza bisogno di simulare processi di ricottura, calcolare valori medi ed effettuare sottrazioni). Ciò significa che nella fase eccitatoria dovrebbero attivarsi spontaneamente funzioni sinergiche estensive (tipo unione insiemistica) e in quella inibitoria funzioni anergiche intensive (tipo intersezione insiemistica).

Un modello di rete multistabile, ad effetti eccitatori lineari ed effetti inibitori quadratici e ritardati, capace di funzionare secondo questo principio, dovrebbe avere le seguenti caratteristiche: 1) sostanziale riduzione degli stati spuri; 2) organizzazione dei ricordi in modi dipendenti dalla struttura dei confini di criticità nello spazio degli stimoli possibili; in tal modo si trova che la teoria dei punti critici (V.I.Arnold A.Varchenko e S.Goussein-Zadè, 1986), comunemente nota come teoria delle catastrofi, diviene lo strumento matematico più adeguato allo studio delle forme di categorizzazione; 3) formazione preliminare dei generi prossimi come attrattori dello spazio degli stati; 4) successiva trasformazione degli attrattori in repulsori e formazione, attorno a questi, di un nuovo sistema di attrattori, corrispondenti alle differenze specifiche, in reciproca competizione isteresica.

2.10. La quantizzazione vettoriale di Kohonen

Gli stati di una rete neurale possono appartenere ad uno spazio discreto o continuo. Il vantaggio del discreto sul continuo sta nel fatto che l'elaborazione di segnali veicolati da variabili discrete è molto più affidabile di quella veicolata da variabili continue. Un sistema dinamico a stati discreti possiede una stabilità intrinseca e una resistenza al rumore incomparabilmente maggiori di uno a stati continui. Per contro, un sistema a stati continui tende a perdere informazione durante la sua evoluzione temporale: un errore trascurabile delle condizioni iniziali può amplificarsi esponenzialmente col trascorrere del tempo, cosicché alla fine l'informazione veicolata dallo stato viene completamente perduta.

D'altronde, il significato di un messaggio non è altro che l'insieme delle operazioni che l'apparato ricevente può esercitare nell'ambiente grazie all'utilizzo di quel messaggio. L'informazione contenuta in un messaggio sensoriale, catturata da un cervello animale, si traduce in ultima analisi in un possibile programma d'azione utile all'animale. Possiamo dire che la quantità utile d'informazione trasportata dal messaggio è determinata dai procedimenti di codificazione che hanno luogo negli stadi terminali del processo d'elaborazione dell'informazione nervosa. È chiaro che per ottenere una codifica ottimale dell'informazione sensoriale è necessario che persino lo stadio d'ingresso dell'informazione sensoriale sia organizzato in funzione del suo modo d'utilizzo nello stadio terminale.

In questa parte dello scritto non entreremo nel merito di come questo possa accadere nei cervelli animali. Limitiamoci ad osservare che per l'attivazione dei programmi di comportamento da parte dell'apparato effettore l'animale utilizza solo una piccola frazione del flusso d'informazione sensoriale. Anzi, secondo i risultati della psicologia sperimentale, servono solo pochi bit per secondo. Questo ci fa pensare che il modo più conveniente d'utilizzare l'input sensoriale comporti una sorta di *quantizzazione* dell'informazione nervosa. Alcuni modelli di reti nervose studiati da Kuvo Kohonen risolvono questo problema.

In un certo senso il modello di Hopfield a stati discreti quantizza l'informazione sensoriale in modo naturale. Sebbene l'informazione che giunge alle unità neurali sia veicolata da segnali variabili con continuità, i loro effetti sulla rete sono in ogni caso rappresentati da un insieme finito di 0 e 1. Tuttavia, come sottolinea Kohonen (1988), in realtà né i neuroni né le sinapsi sono elementi di memoria bistabili. Bisogna però osservare che il modello di Hopfield con i neuroni a risposta sigmoidale, pur non costituendo a stretto rigore un sistema a stati discreti, approssima tuttavia molto bene un sistema a stati discreti. Nelle condizioni di funzionamento stabilite da Hopfield le risposte delle unità neurali sono in gran parte molto prossime a 0 o a 1. Tutto sommato, per l'uso che in generale viene fatto dei dati memorizzati, si tratta di una discrepanza dal livello di perfetta quantizzazione perfettamente tollerabile. Chiaramente la forte non-linearità compressiva determinata dalle risposte sigmoidali rende possibile qualcosa di molto simile alla quantizzazione.

Il fenomeno della quantizzazione o quasi quantizzazione dell'informazione nervosa è strettamente collegato con la tendenza spontanea all'auto-organizzazione delle strutture nervose animali. Questo argomento è stato oggetto di ampie ricerche nel corso degli anni 80 e, dati i limiti di questo articolo, per un'ampia rassegna su questi argomenti possiamo consultare ad esempio il trattato di Hykin (1994), la collezione di articoli del Gruppo Nazionale di Bioingegneria (a cura di Biondi et al., 1991), le Lecture Notes dell'Istituto di Santa Fe (Hertz et al, 1991).

Kohonen ha studiato il problema della quantizzazione degli stati delle reti neurali, rappresentato come uno spazio vettoriale, introducendo la non linearità nello stesso processo dinamico che porta gli stati della rete a convergere verso un insieme ristretto di possibili stati finali. L'idea che sta alla base del procedimento di Kohonen è il principio del *vincitore-piglia-tutto* (*winner takes all*). Le reti di Kohonen hanno una struttura planare e contengono due tipi di connessioni sinaptiche a raggio d'azione limitato: connessioni in avanti dalle sorgenti primarie dell'eccitazione, e quelle interne alla rete che producono auto-retroazione positiva ed inibizione laterale. L'interazione tra le unità dipende dalla distanza e può caratterizzarsi come segue: 1) un'eccitazione laterale a corto raggio d'azione; 2) una penombra d'azione inibitoria laterale; 3) un'area d'eccitazione più debole che circonda la penombra inibitoria. Il profilo dell'azione di ciascuna unità su quelle circostanti ha pertanto la forma di un cappello messicano.

In una fase iniziale la rete, attivata da uno stimolo applicato a un insieme numeroso di segnali portati da N fibre d'ingresso, produce un insieme molto meno numeroso di segnali di risposta da M fibre in uscita. Grazie alle loro interazioni interne le unità neurali agiscono le une sulle altre in modo da favorire l'aumento d'ampiezza dei segnali in uscita delle unità che già emettevano un segnale di ampiezza maggiore riducendo quella delle unità che emettevano segnali di ampiezza minore. In questo modo la rete evolve verso uno stato in cui un solo segnale d'uscita, o un sottoinsieme di segnali in uscita, raggiunge il valore massimo.

L'applicazione più interessante di questo modello è stata la progettazione di reti capaci di auto-organizzarsi sotto l'azione dei segnali d'ingresso senza l'intervento di un supervisore esterno. Il metodo si avvale del fatto che, inizializzando la rete con valori casuali dei pesi sinaptici, il processo di quantizzazione vettoriale porta in ogni caso alla formazione di segnali d'uscita che entrano in competizione tra loro per occupare il loro spazio secondo un criterio ottimale. In corrispondenza, lo spazio dei possibili segnali d'ingresso risulta suddiviso in gruppi separati che fanno capo a diversi segnali in uscita. In questo modo un insieme di stimoli parzialmente sovrapposti, che in altri modelli di reti suscitano risposte parzialmente sovrapposte, vengono alla fine ordinati in gruppi distinti. Il modello di Kohonen è inoltre confortato dai riscontri relativi alla

tendenza all'organizzazione spontanea delle reti nervose nelle aree corticali, in particolare in quelle visive (Kohonen, 1993).

2.11. Il modello di Morita

Un aspetto generale della ricerca sulle reti neurali sviluppatasi impetuosamente nel corso degli anni 80 è stato la frustrazione delle aspettative iniziali. Le difficoltà incontrate nel tentare d'indicare procedimenti di cablaggio e regole d'interazione adeguate alle funzioni richieste dalle reti o nell'integrare in un'unica teoria generale le varie funzioni rappresentate dai modelli, la scarsa plausibilità neurologica dei modelli proposti, ma ancor più il fatto che l'efficienza delle reti neurali si rivelava inferiore alle aspettative, hanno infine determinato una crisi in sordina di questo settore. Come era già accaduto alla fine degli anni '60, questo stato di cose ha determinato lo spostamento della ricerca informatica sulle tecniche d'elaborazione mediante processi digitali seriali, in particolare nel settore della compressione dell'informazione audio e video. I ricercatori che si sono resi conto che si trattava di un vero e proprio fallimento del paradigma connessionista non sono stati molti, ancora meno quelli che hanno cercato di capire le profonde ragioni della crisi. A questo proposito merita una particolare attenzione l'analisi critica del modello di Hopfield condotta da Morita (1993) e le soluzioni da egli proposte per uscire dall'impasse.

Come è stato chiarito nei paragrafi 2.7, 2.8 e 2.9, l'unico serio inconveniente del modello di Hopfield è la formazione degli stati spuri. Come è stato detto in precedenza, gli stati spuri sono combinazioni di memorie negative e positive, in prossimità dei quali la funzione energia (4, par. 2.7) possiede minimi relativi che soddisfano all'equazione (5) sebbene la rete non sia mai stata addestrata a memorizzarli. Per quanto la formazione di queste chimere abbia qualcosa di simile con l'attività onirica, esse in realtà disturbano seriamente il funzionamento della memoria associativa al punto di renderla inutilizzabile.

Esperimenti di simulazione e considerazioni teoriche (Amari e Maginu, 1988) hanno messo in evidenza che la piaga degli stati spuri comincia a diventare seria quando il rapporto r tra numero di memorie M e numero di neuroni N supera il valore dell'1 o 2 percento. La formazione degli stati spuri limita la capacità di memoria della rete ben oltre il massimo teorico, calcolabile sulla base della quantità d'informazione contenuta nella matrice dei pesi. Il problema sembra infatti essere dovuto più alla fallacia della dinamica della rete che alle sue proprietà statiche. Si verifica, infatti, che, una volta inizializzato con un frammento di memoria, di grandezza percentuale inferiore ad un certo valore critico dipendente da r, lo stato della rete, dopo un temporaneo avvicinamento verso la memoria corretta (valutabile come grado di sovrapposizione tra lo stato corrente e stato memorizzato), comincia ad allontanarsi da quello su cui dovrebbe convergere e finisce per essere intrappolato da uno stato spurio. L'effetto dipende dalla grandezza del frammento inizializzatore, ma se $r > 0.15$ nemmeno un frammento pari al 99% dello stato memorizzato sfugge a tale drammatico destino. Un'accurata analisi sulle cause di questo fenomeno ha rivelato che non è tanto la profondità dei minimi relativi del paesaggio dell'energia potenziale a determinare il problema, quanto piuttosto il fatto che il numero degli stati spuri cresce più rapidamente del numero M degli stati memorizzati. Se gli stati memorizzati sono ortogonali, le buche di potenziale sono ben separate, molto profonde e localizzate esattamente in corrispondenza di questi stati. Tuttavia, al crescere di r le buche delle memorie corrette sono circondate sempre più fittamente da nugoli di buche spurie che confinano con le prime attraverso una fitta trama di valli molto strette. Per quanto siano meno profonde delle prime, a causa del loro numero elevato esse esercitano un'azione dispersiva sulla dinamica della rete. A meno che non si trovi molto vicino al minimo di uno stato memorizzato, lo stato che all'inizio tendeva ad avvicinarsi ad uno stato memorizzato, giunto in prossimità del suo bersaglio, subisce una miriade di deviazioni da parte della corona di attrattori spuri.

L'idea che ha permesso a Morita di aggirare l'ostacolo è nata dalle sue osservazioni su un gran numero di simulazioni al calcolatore. Egli ha notato che, nel corso dell'evoluzione dello stato, in corrispondenza dell'azione attrattiva degli stati spuri, i potenziali postsinaptici d'alcuni neuroni

raggiungevano valori molto positivi o molto negativi più rapidamente di altri. Erano proprio gli stati di questi neuroni che andavano a formare gli 0 e gli 1 degli stati spuri ed era dunque proprio questa la tendenza che bisognava contrastare se si voleva favorire la convergenza dello stato verso la memoria corretta. Per fare questo Morita aggiunse una regola generale alla legge d'evoluzione del sistema: diminuire sensibilmente il valore dei potenziali che raggiungono valori troppo elevati. Gli stratagemmi per ottenere questo risultato possono essere diversi, in particolare si possono introdurre profili di risposta dei neuroni di forma alternata, con un tratto centrale sigmoidale.

I risultati che si ottengono con questo metodo sono spettacolari. Invece di venire assorbito e intrappolato, lo stato corrente della rete rimbalza sugli stati spuri fino a trovare la strada per convergere verso la memoria corretta. Fino valori di $r = M/N = 0.3\text{-}0.4$, tutti i frammenti di memoria di misura superiore al 15% inizializzano l'evoluzione dello stato verso il bersaglio corretto.

Morita ha applicato le sue simulazioni anche al caso in cui gli stati memorizzati sono suddivisi in gruppi di stati con larghe percentuali di sovrapposizione all'interno di ciascun gruppo e con scarsa sovrapposizione tra stati di gruppi diversi. In queste condizioni si può osservare che una volta inizializzata la rete con un frammento di uno di tali stati memorizzati, lo stato corrente evolve in una prima fase verso il "baricentro" del gruppo e in una fase successiva si diparte dal baricentro per andare a bersaglio sullo stato memorizzato corretto. *È proprio questo il comportamento che si richiede ad una memoria auto-organizzante capace di evocare ricordi prima per generi prossimi e poi per differenze specifiche.*

A giudizio dell'Autore è questo il risultato più brillante ottenuto nell'ambito del paradigma connessionista. Si deve tuttavia osservare che la tecnica di Morita, basata sull'adozione di procedure correttive della dinamica piuttosto che di cablaggi e interazioni neurali più sofisticati, suggerisce che risultati ancora migliori possano ottenersi con reti neurali dotate di dinamiche non lineari più sofisticate. Rimane in particolare da chiarire se dinamiche di tipo oscillatorio non offrano metodi più efficienti per ottenere sistemi dotati di memoria associativa auto-categorizzante.

Padova 15.10.2004

Bibliografia

1. S.Amari and K.Maginu (1988) Statistical Neurodynamics of Associative Memory. *Neural Networks,* 1:63-73.
2. D.J.Amit, H.Gutfreund e H.Sompolinsky (1985) *Phys. Rev. A*, 32:1007-18; (1987) 35:2293-2303; (1987) *Ann. of Phys*., 173:30-67.
3. M.Arbib (1968) *La mente, le macchine e la matematica*., Ed. Boringhieri, Bologna.
4. Aristotele (1973) Categorie, in *Opere*, vol.I, Ed.Laterza, Bari.
5. V.Arnold, A.Varchenko e S.Goussein-Zadè (1986) *Singularitè des application diffèrentiables* - voll.1, 2, Ed MIR, Mosca.
6. C.Asanuma e F.Crick, in *Parallel Distributed Processing*, J.L. McClelland e D.E.Rumelhart Eds., Vol. 2, 333-371, (1986).
7. H.B.Barlow, in *Cybernetics* C.R.Evans e A.D.J.Robertson Eds., p. 183-207, Butterworths (London, 1968).
8. G.Berlucchi (1990) *Acta Psychologica* in *Brain and Reading*, C.von Euler Ed., McMillan London.
9. E.Biondi, P.Morasso e V.Tagliasco Editori (1991) *Neuroscienze e Scienze dell'Artificiale: Dal Neurone all'Intelligenza.* Pàtron Editore, Bologna.
10. T.V.Bliss and T.Lomo (1973) *J. Physiol. Lond*. 232:331-356.
11. J.Borman and D.E.Clapham (1985) *Proc. Natl. Acad. Sci.* 82:2168-72.
12. A.Borsellino and T.Poggio (1972) *Kibernetik* 13:10.
13. B.Cavanagh (1978) *Perception* 7:167-177.

14. C.Cornoldi (1978) *Modelli della memoria*. Ed. Giunti Barbera, Firenze.
15. A.Crisanti and H.Sompolinsky (1987) *Phys. Rev. A*, 36:4922-39.
16. A.De Luca e L.M.Ricciardi (1981*) Introduzione alla Cibernetica* – Franco Angeli Ed. Milano.
17. R.Eckorn and H.J. Reitboeck (1989), 99-111; C.M.Gray, P.König, A.K.Engel and W.Singer, 82-98 in *Synergetics of Cognition*, H.Haken and M.Stadler Eds., Springer-Verlag (Berlin).
18. M.Farah (1989) TINS, 12:395-399.
19. F.Fogelman (1990) *Int. Neur. Net. Conf.* (*INNC-90*). Tutorial paper.
20. W.J.Freeman (2000) *Come pensa il cervello*. Einaudi Ed., Torino
21. H.Gardner (1988) *La nuova scienza della mente*. Ed. Feltrinelli, Milano.
22. D.Gabor (1968) *Nature* 217a:548.
23. G.Girosi and T.Poggio (1990) *Science* 247:978-82; (1990) *Biol. Cybern*. 63:169-176.
24. S.Grossberg (1988) *Neural Networks and Natural Intelligence*. The MIT Press Ed.
25. R.Elul (1972) *Int. Rev. Neurobiol*. (15).
26. D.O.Hebb (1949) *The Organization of Behavior*. John Wiley Ed. (1949); trad. ital. (1975) *L'organizzazione del comportamento*, Ed. F.Angeli.
27. P.van Heerden (1973) *Appl. Opt*, 2:387.
28. W.E.Hick (1952) *The Quarterly Journal of Experimental Psychology*, 4:11-26.
29. S.Hykin (1994) *Neural Networks. A Comprehensive Foundation*. Macmillan College Pub.Co.
30. J.J.Hopfield (1982) *Proc. Natl. Acad. Sci. USA*, 79:2554-58; (1982) 81:3088-92.
31. J.Hopfield, D.I.Feinstein and R.G.Palmer, *Nature*, 304:158-59.
32. J.Hertz, A.Krogh and R.G.Palmer (1991) *Introduction to the Theory of Neural Computation*. Addison-Wesley Pub.Co.
33. D.H.Hubel (1989) *Occhio, cervello e visione*. Ed. Zanichelli, Bologna.
34. R.Hyman (1953) *Journal of Experimental Psychology*, 45:188-196.
35. S.Kirkpatrik and D.Sherrington (1978) *Phys. Rev. B*, 17: 4384.
36. K.Kohonen (1988) An introduction to Neural Computing, *Neural Networks*, 1:3-16.
37. T.Kohonen and E.Oja (1976) *Biol. Cybernetics*, 21:85-95.
38. T.Kohonen (1993) Physiological interpretation of self-organizing map algorithm. *Neural Networks*, 6:895-905.
39. S.M.Kosslyn (1996) *Image and Brain*, The MIT Press.
40. K.E.Kürten and J.W.Clark (1986) *Phys. Lett*. 114A: 413-418.
41. K.S. Lashley (1950) In Search of the Engram. *Symposia of the Soc. of Experimental Biology*, 4:454-482.
42. J.A.Leonard (1961) Choice reaction time experiments and information theory, in *Information Theory, Fourth London Symposium*, C.Cherry Editor, Butterworths, London.
43. W.B.Levy and O.Steward (1983) *Neurosciences* 8:791-797.
44. H.C.Longuet-Higgins (1968) *Nature*, 217a:547-548.
45. S.W.McCulloch & W.Pitts (1953), *Bull. Math. Biophysics*, 5:115-133.
46. M.Mèzard, G.Parisi, N.Sourlas, G.Toulouse & M.Virasoro (1984), *Phys. Rev. Lett*., 52:1156-1159.
47. M.Morita (1993) Associative Memory with Nonmonotone Dynamics. *Neural Networks*, 6:115-126.
48. R.G.M.Morris, E.R.Kandel and L.R.Squire (1988) *TINS, special issue* 11:125-127.
49. J.von Neumann (1951) *The General and Logical Theory of Automata, in Cerebral Mechanisms and Behavior*, Hixon Symposium, J.Wiley.
50. J.von Neumann (1966), *Theory of Self--Reproducing Automata*. University of Illinois Press (Urbana).
51. J.von Neumann (1958) *The Computer and the Brain*. Yale University Press.
52. K.Okajima (1990) *Proc. INNC-90* (*Int. Neur. Net. Conf.*), Kluver Acad. Pub., vol.II, 504—507.
53. G.Parisi (1983) *Phys. Rev. Lett*. 50:1946-1948.
54. D.E.Rumelhart, G.E.Hinton & R.J.Williams (1986), *Nature*, 323:533-536.

55. E.L.Schwartz, (1981) *Perception*, 10:455-468.
56. C.Shannon (1971) *La teoria matematica delle comunicazioni*. Ed. Eta Compas.
57. P.K.Stanton and T.J.Sejnowski (1989) *Nature* 339:215-218.
58. A.M.Turing (1936) On Computable Numbers with Applications to the Entscheidungsproblem, *Proc.London. Math. Soc*. 42:230-265.
59. A.T.Welford (1959) *The Quarterly Journal of Experimental Psychology*, 11(4):193-208.
60. B.Widrow & M.A.Lehr (1990) *Proc. of the IEEE. Special issue on neural networks, I*, 78:1415-1442.