Vidres de spins i xarxes neuronals

Fèlix Ritort* **i August Romeo†**

Introducció

Un dels interessos més grans de la física és l'estudi de les propietats dels sistemes macroscòpics amb molts graus de llibertat. D'aquesta branca de la física s'encarrega la mecànica estadística, l'objectiu de la qual és la derivació de les propietats globals d'un sistema conegudes les interaccions al nivell de les seves partícules constituents. La mecànica estadística, encara que es va iniciar ja fa molt temps amb els primers resultats de L. Boltzmann i J. W. Gibbs ara farà uns cent anys, ha anat obrint el seu camp d'influència i actualment comprèn un gran nombre de temes de recerca que van des de l'estat sòlid fins a la turbulència dels fluids passant per tota una fenomenologia molt vasta. La mecànica estadística s'ha revelat també de molta utilitat en permetre un llenguatge comú amb la teoria quàntica de camps fins al punt de parlar-se avui en dia de teoria estadística dels camps per fer referència a totes dues.

En aquest treball presentem les idees principals que hi ha darrere d'una de les branques més recents de la mecànica estadística anomenada teoria dels vidres de spins –*spin glasses*. El seu fi és estudiar els fenòmens cooperatius en sistemes desordenats. Els fenòmens cooperatius i també especialment les transicions de fase són un dels èxits més recents de la mecànica estadística, que es va iniciar a meitat d'aquest segle i ha assolit el grau de màxima completesa amb la teoria del grup de renormalització desenvolupada per K. G. Wilson. Les tècniques de la mecànica estadísitica han estat també aplicades a una nova ciència interdisciplinària que es basa en les xarxes neuronals. Aquests models, encara que apareguts originalment per altres motius, poden ser tractats amb els mètodes propis dels processos estocàstics, de manera que una bona part del formalisme estadístic per a reticles de spins hi resulta adequada. De fet, en considerar certs tipus de xarxa apareix una fase de vidre de spins, i el coneixement de les propietats d'aquesta es fa imprescindible per entendre'n el funcionament.

Què és una transició de fase?

La transició de fase és un dels fenòmens més interessants de la física. Primer, perquè es troba a la vida diària en molts casos diferents aparentment inconnexos i segon, perquè es pot estudiar usant models d'interacció microscòpica, que permeten fer prediccions precises sobre el comportament global del sistema macroscòpic que poden ser comprovades tant experimentalment com numèricament per mitjà d'ordinadors.

A continuació, introduirem els conceptes generals del que és una transició de fase. A la secció següent explicarem què és un vidre de spins per passar a parlar de quin tipus peculiar de transició de fase presenten aquests vidres.

Potser la forma més senzilla d'explicar què és una transició de fase és a partir d'un exemple quotidià qualsevol. I no cal anar gaire lluny per recordar el fenomen de la congelació de l'aigua. En efecte, sabem que l'aigua per sobre de la temperatura de 0°C és líquida. A l'estat líquid les molècules d'aigua es poden moure més o menys lliurement. En augmentar la temperatura l'agitació tèrmica dóna energia cinètica a les molècules de tal forma que les partícules poden escapar al lligam de les forces de Van der Waals. A la pressió de 1 atm, per sobre dels 100 °C l'aigua està en estat gasós. Pel contrari, si baixem la temperatura, l'energia cinètica de les molècules disminueix i aquestes queden lligades per les forces intermoleculars en posicions fixes. Aquest és l'estat sòlid i l'únic moviment de les molècules es limita a ràpides vibracions harmòniques al voltant de la posició d'equilibri. El punt interessant és que entre aquestes tres fases hi ha dues temperatures ben precises en les quals la matèria experimenta canvis dràstics passant d'una fase a l'altra. Per al cas de l'aigua aquest canvi es manifesta per l'existència d'una calor latent. És a dir, per passar de líquid a sòlid cal treure calor a l'aigua mantenint la temperatura invariable. Aquesta extracció de calor és la mínima necessària perquè totes les molècules d'aigua quedin bloquejades en posicions precises.

Aquesta transició (on hi ha calor latent) rep el nom de *transició de primer ordre*. Aquí l'entropia S sofreix una discontinuïtat finita. Aquesta magnitud pot obtenir-se fent una derivada primera respecte a la temperatura d'un cert potencial termodinàmic que, en les nostres consideracions, serà l'energia lliure de Helmholtz F, i de la qual parlarem més endavant. En transicions de primer ordre, doncs, hi ha salts finits en derivades primeres del potencial termodinàmic que donaran lloc a divergències en les derivades d'ordre dos o més alt.

*Fèlix Ritort Dipartimento di Fisica, Università di Roma II, Departament de Física Fonamental, U B

†August Romeo Departament de Matemàtica Aplicada i Anàlisi, Facultat de Matemàtiques, U B

A la natura, però, hi ha altres tipus de transicions que, tot i que són més rares que les de primer ordre, s'han revelat de màxima importància per entendre els fenòmens cooperatius en general. Parlem de les transicions *contínues*. En aquest cas, no hi calor latent però certes magnituds com la calor específica divergeixen. Les transicions (contínues) de segon ordre es caracteritzen per la continuïtat de les derivades primeres del potencial termodinàmic i la presència de singularitats en magnituds que en són derivades segones. Com a conseqüència, les derivades d'ordre superior tindran divergències. La calor específica resulta ser (llevat de constants) la derivada segona del potencial termodinàmic respecte a la temperatura.

Pot ser que el sistema canviï bruscament d'estructura sense cedir o prendre calor. L'exemple més comú és un sistema ferromagnètic. Un sistema ferromagnètic es pot imaginar com un conjunt de petits imants microscòpics que interaccionen entre ells a causa de la seva proximitat. Per simplificar pensarem que els imants poden prendre dues direccions possibles (amunt o avall). En aplicar un camp magnètic tots els imants s'orienten en la direcció del camp aplicat i donen lloc a una imantació global. A temperatures molt altes l'agitació tèrmica fa que tots els imants fluctuïn independentment i s'orientin aleatòriament cap amunt o cap avall. En aquest cas el sistema té una imantació nul·la. A mesura que disminuïm la temperatura l'agitació tèrmica disminueix i les interaccions entre els imants esdevenen cada vegada més importants fins al punt que per sota d'una certa temperatura crítica (que per al ferro és de l'ordre dels 1000 °C) el sistema inevitablement no pot resistir la forta interacció entre tots els imantets, que els fa orientar-se en la mateixa direcció. El sistema té ara una imantació no nul·la i està ordenat.

Un indici experimental que assenyala una transició de fase *de primer ordre* en un ferromagnet és la divergència de la susceptibilitat magnètica, és a dir, la resposta del sistema a un petit camp aplicat h. Com que en el punt crític la magnetització pot prendre un valor o el seu oposat, aquesta experimenta una discontinuïtat a camp nul passant de $-m$ si el camp h és negatiu a m si el camp és positiu. D'altra banda, la magnetització pot expressar-se com la derivada d'una certa energia lliure respecte al camp aplicat, i la susceptibilitat com la derivada d'ordre superior. En termes matemàtics, la discontinuïtat finita en la derivada primera –magnetització– dóna lloc a una divergència en la derivada segona –susceptibilitat.

En realitat, la transició de fase és deguda (tant en el cas de l'aigua, i com en els sistemes magnètics, com en tants altres casos) a la competició entre dues tendències a la natura: una tendència entròpica per la qual els sistemes volen desordenar-se al màxim (l'exemple més clar és un sistema de partícules lliures) i una tendència energètica per la qual els sistemes volen configurar-se en posicions energèticament favorables (i d'aquestes n'hi ha poques). La temperatura és el paràmetre que regula quina és la tendència dominant. A temperatures altes l'efecte entròpic és dominant. A temperatures baixes ho és l'energètic.

Més exactament, i a l'hora de quantificar-ho tot, els dos termes que competeixen són l'energia U i l'entropia S, i el que cal buscar és la configuració microscòpica que minimitza l'energia lliure

$$F = U - TS, \tag{1}$$

on T és la temperatura. Aquesta relació té una gran importància, ja que ens mostra, entre altres coses, que per a $T = 0$ els mínims de F i els de U coincideixen. És d'esperar que això continuï passant per a T prou petites, però no quan ens allunyem prou de $T = 0$. Cal tenir-ho molt en compte, perquè a l'hora de cercar els punts corresponents a transicions de fase haurem de detectar els canvis en la natura dels extrems de la funció F: aparició o desaparició de màxims i mínims, extrems locals que passen a ser absoluts i a l'inrevés, etc.

Vidres de spins: generalitats teòriques

Podem ara intentar explicar què és un vidre de spins –*spin glass*– i quins tipus de fenòmens cooperatius el caracteritzen. En primer lloc, reprendrem la nostra discussió sobre els sistemes ferromagnètics. Tal com hem vist, aquests estan constituïts per una infinitat de petits imants que interaccionen com si fossin dos dipols. Aquests petits imants (tal com pensava a l'època Ampère) sabem avui en dia que corresponen a una propietat quàntica dels àtoms que rep el nom de spin. Quan dos spins d'un sistema magnètic tendeixen a orientar-se segons la mateixa direcció parlem d'una interacció ferromagnètica. En el cas més general, però, aquesta interacció pot ser també antiferromagnètica. Aleshores els spins que interaccionen volen posar-se en direccions contràries. Suposem, doncs, un sistema magnètic a temperatura baixa. Quan la interacció és ferromagnètica és fàcil que tots els spins es posin d'acord a minimitzar l'energia global d'interacció. Cal només que tots s'orientin en la mateixa direcció (tots amunt o tots avall). Si la interacció és antiferromagnètica és més difícil aconseguir acontentar els spins. En efecte, imaginem tres spins magnètics interaccionant entre ells antiferromagnèticament (figura 1). És impossible aconseguir una configuració en què els spins de qualsevol parella estiguin orientats en direccions contràries. Diem que el sistema està *frustrat*. En aquest cas no hi ha una única solució energèticament òptima com en el cas del ferromagnet. No obstant això, el sistema troba configuracions estables on l'energia és mínima. D'aquestes configuracions estables pot haver-n'hi un nombre molt elevat, però, i el que és més im-

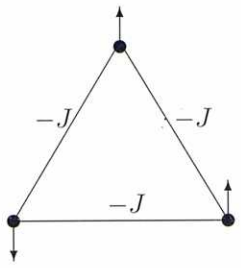

Figura 1: *Tres spins frustrats. La unió a la part superior dreta està "insatisfeta". En un sistema macroscòpic hi ha una frustració que implica un nombre molt elevat de spins*

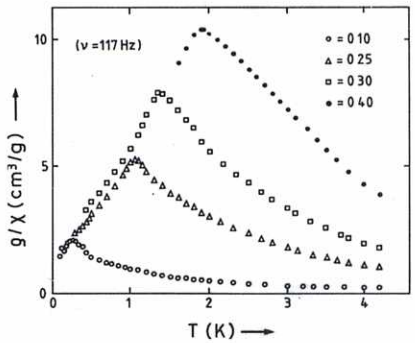

Figura 2: *Pic de la susceptibiliat en un vidre de spins metàl·lic. Encara que no hi hagi divergència això és un indici de l'existència d'una transició de fase*

portant, estan relacionades l'una amb l'altra per una operació de simetria (generalment estan relacionades amb les simetries geomètriques de la xarxa de spins). Fins aquí hem explicat les idees bàsiques del que és una transició de fase i les característiques més àmpliament acceptades de com és la fase ordenada en els sistemes magnètics.

El problema dels vidres de spins s'inicia aleshores a principi dels setanta amb la recerca experimental de certs aliatges metàl·lics (CuMn, AuFe, AgMn) on s'introdueixen impureses magnètiques en una concentració de l'ordre d'un tant per cent baix. Les propietats d'aquests sistemes eren molt interessants perquè mostraven indicis d'una ordenació a temperatures molt baixes ja que s'observava un pic en la susceptibilitat magnètica (figura 2). Aquest pic, però, no era una divergència, la qual cosa feia sospitar que aquesta transició era de naturalesa diversa. Més exactament, no podia tractar-se d'una transició de primer ordre. Encara més estrany, la calor específica no mostrava ni pic, només un màxim molt ampli i desplaçat cap a temperatures més elevades que aquella en la qual s'observava el pic en la susceptibilitat.

En què diferien aquests sistemes dels sistemes magnètics més comuns abans esmentats (ferros o antiferros)? En aquests sistemes la interacció entre els spins veïns es fa via electrons de conducció. A mitjan dels anys seixanta Ruddermann, Kasuya, Kittel i Yosida van demostrar que la interacció via electrons de conducció entre impureses magnètiques fluctuava en intensitat i variava de signe (és a dir, podia ser de tipus ferro o antiferro) segons la distància de separació entre les impureses (figura 3).

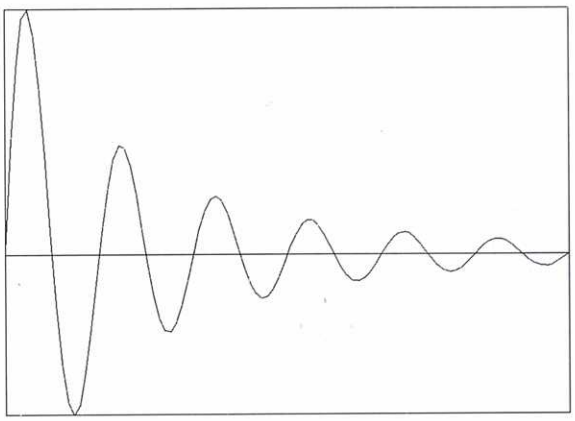

Figura 3: *Interacció de tipus RKKY entre impureses magnètiques en un vidre de spins. Canvia alternativament de signe (ferro o antiferro) i la seva intensitat minva com el cub de la distància. És, per tant, una interacció de llarg abast*

Com que les impureses se situen dins l'aliatge metàl·lic de forma aleatòria, les distàncies entre aquestes varien també aleatòriament. La conseqüència de tot això és que els spins del sistema interaccionen no de forma ferromagnètica, no de forma antiferromagnètica, sinó aleatòriament de forma ferro i antiferro. Si ja amb el sistema antiferromagnètic és difícil determinar quina és la configuració energèticament favorable per a temperatures baixes, en un cas desordenat la complexitat augmenta considerablement.

És a dir, el sistema no és només frustrat sinó també desordenat. I aquests són els ingredients bàsics d'un vidre de spins: desordre i frustració.

El 1975 es proposa un primer model de vidre de spins que es pot tractar exactament de forma analítica. El model, que es pot entendre com una aproximació de camp mitjà (és a dir, suposar que no hi ha correlacions entre els spins) rep el nom de model de Sherrington-Kirkpatrick. En aquest model es recorria al truc de la rèplica, una tècnica matemàtica enormement útil però de la qual es coneix ben poc pel que fa a la seva fonamentació. Usant una certa hipòtesi de simetria es resol el model analíticament encara que es demostra que aquesta hipòtesi és incorrecta, ja que té com a resultat una solució inestable. En general, el model donava

l'existència d'una transició de fase contínua amb característiques molt semblants al tipus de transició que es troba experimentalment.

La resolució definitiva arriba el 1979 quan G. Parisi proposa un trencament de simetria molt particular que resol les inestabilitats trobades en tots els treballs precedents. La significació física d'aquest trencament de simetria roman fosca fins a l'any 1983 en què el mateix G. Parisi en dóna la interpretació correcta. Segons aquesta, la transició que té lloc en un vidre de spins es caracteritza perquè el sistema passa d'una fase desordenada o paramagnètica a una fase amb moltes configuracions d'equilibri possibles. És a dir, la fase vidre de spins està formada per una infinitat d'estats d'equilibri separats per barreres molt altes d'energia. En termes més tècnics es diu que l'ergodicitat es trenca. Això vol dir que en un vidre de spins macroscòpic cal esperar temps enormement grans (de l'ordre d'escales astronòmiques) perquè el sistema visiti les diferents configuracions d'equilibri. A diferència dels sistemes magnètics normals, aquests estats d'equilibri no estan relacionats per cap operació de simetria (com succeeix per als ferros i els antiferros). El 1984 la teoria dels vidres de spins arriba al seu punt culminant amb el treball de M. Mézard, G. Parisi, N. Sourlas, G. Toulouse i M. A. Virasoro. Usant com a punt de partida el trencament de simetria proposat per G. Parisi l'any 1979, troben que els estats d'equilibri d'una fase de vidre de spins s'ordenen de forma jeràrquica molt similarment a la manera com s'organitzen les espècies animals i vegetals en arbres genealògics (figura 4).

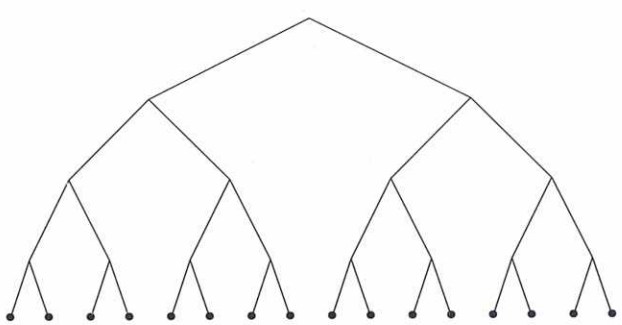

Figura 4: *Els estats d'equilibri d'un vidre de spins s'organitzen jeràrquicament ocupant els extrems inferiors d'una estructura d'arbre. L'exemple mostrat correspon a un arbre binari.*

És a dir, encara que no hi ha cap simetria que connecti els diferents estats entre si, aquests estan ordenats de forma jeràrquica. Aquesta propietat d'ordenació dels estats rep el nom d'*ultrametricitat*.

I poc més de fonamental s'ha fet des d'aleshores, excepte estendre totes aquestes idees a l'estudi d'altres fenòmens de la física on el desordre i la frustració tenen una funció primordial (el cas de les xarxes neuronals n'és un exemple).

Ha de quedar clara la natura peculiar d'aquesta transició. En disminuir la temperatura per sota del seu valor crític, el sistema minimitza l'energia d'interacció global per tal de vèncer l'efecte entròpic que tendeix a desordenar el sistema. A diferència dels sistemes magnètics normals el vidre de spins és un sistema desordenat i molt complex però així i tot el vidre de spins troba la seva configuració estable interna òptima. Aquesta configuració òptima és completament desordenada. Hi ha moltíssimes configuracions òptimes que el sistema pot escollir i aquestes, lluny de ser aleatòries o simètricament organitzades, estan distribuIdes de forma jeràrquica. I això no deixa de ser una sorpresa.

Xarxes neuronals

Actualment molts físics veuen l'estudi de les xarxes neuronals –*Neural Networks*– com una parcella de la teoria dels vidres de spins, ja que hi ha importants tipus de xarxes equiparables a sistemes de spins amb soroll estocàstic, sobretot el model de Hopfield i les seves variants. Habitualment, aquesta classe de models són emprats per al reconeixement d'imatges –òptiques, acústiques, etc.– expressables per mitjà de seqüències binàries. Els segments de codi binari susceptibles de ser reconeguts (per exemple la seqüència d'uns i zeros que queda en "digitalitzar" una fotografia de mida fixada) s'acostumen a anomenar *patterns*.

Hi ha moltes altres classes de xarxa d'elements interconnectats que processen informació, amb la finalitat de classificar, predir, traduir, deduir regles o senzillament avaluar funcions lògiques, on l'estocasticitat pot ser absent del funcionament. Aquests models plantegen problemes igualment difícils relatius a les prestacions, les capacitats d'emmagatzemar i generalitzar, i l'entrenament o l'aprenentage –que pot ser de caire estocàstic o no. Les xarxes neuronals són adequades per resoldre problemes on el detall va creixent per aproximacions successives, és a dir, en passos ordenats jeràrquicament. Hi ha models de xarxa neuronal potencialment capaços de classificar i generalitzar, que poden fer-ho fins i tot sense coneixement d'unes regles, per exemple poden aprendre a pronunciar text en anglès, a partir d'exemples.

Per aquestes raons, les xarxes neuronals estan passant de ser vistes com un terreny interdisciplinari dependent de la física, la biologia, la informàtica i, per descomptat, les matemàtiques, a considerar-se una nova ciència amb entitat pròpia, com ho prova la multitud de butlletins i revistes científiques aparegudes sobre el tema en els darrers anys.

Els orígens de l'interès en les xarxes neuronals són el desig d'entendre els principis de funcionament de la ment humana i les necessitats de realitzar tasques com-

plexes, per a les quals els ordinadors seqüencials són inadequats. La primera font de motivació ve del desig de molts científics de comprendre millor els mecanismes pels quals el cervell processa informació. El treball amb xarxes neuronals permet la simulació de processos difícils d'experimentar sobre cervells reals, i la realització de models per conceptes i mecanismes que no es dedueixen directament a partir d'observacions. El famós estudi de Hubel i Wiesel sobre el còrtex visual estriat del gat va mostrar que algunes neurones estan especialitzades a detectar *patterns* d'imatges amb una certa orientació, i que les neurones adjacents detecten el mateix amb orientacions lleugerament canviades. La llarga durada d'aquest procés d'estructuració "topològica" del còrtex va fer impossible d'investigar-lo en el propi entorn biològic. En canvi, uns models de xarxa neuronal varen permetre l'estudi per simulació d'alguns d'aquests mecanismes.

Hi ha molts problemes que necessiten el reconeixement de *patterns* òptics o acústics complicats, no determinats per regles lògiques simples, per exemple el vol d'una mosca que evita obstacles o el moviment per un laberint sense tocar cap paret. La solució en ordinadors tradicionals, pels mètodes d'intel·ligència artificial, ha donat com a resultat els *sistemes experts*, que són, però, massa lents per identificar imatges o paraules prou de pressa. No cal esperar que en un futur proper la velocitat dels microprocessadors augmenti en ordres de magnitud. Per reduir el temps s'ha tornat a la idea de procés en paral·lel, on diverses operacions es fan alhora en diferents elements.

En el present, es consideren les xarxes neuronals com realitzacions prototípiques de processos distribuïts en paral·lel (Rumelhart, 1986a). Els elements processadors d'una xarxa neuronal són relativament simples, n'hi ha un gran nombre, i estan molt interconnectats. El problema principal és trobar la millor estructura de connexions i després trobar els valors de les intensitats que realitzin la tasca. Això es fa sovint per un procés selectiu basat en l'aprenentatge –*learning*– a partir de casos particulars.

L'actual expansió del treball en xarxes neuronals, ja sigui com a objectiu de recerca o com a font de beneficis per a la indústria, és deguda als seus avantatges. L'operació en paral·lel és molt més ràpida i, per tant, més rendible en teoria. La necessitat de hardware especial per implementar aquests principis ha portat al desenvolupament i la construcció de xips amb unitats en paral·lel. A més, les solucions tenen una robustesa que les acostuma a fer molt tolerants als errors. Mentre que una equivocació en un bit destrossa un programa convencional, l'operació d'una xarxa neuronal generalment no es veu afectada quan unes poques unitats o connexions fallen.

Primers models

L'any 1943 W. McCulloch i W. Pitts, (McCulloch, 1943), proposaren una teoria sobre el processament d'informació per elements binaris, semblants a una versió simplificada de les neurones biològiques. En una xarxa amb N d'aquests elements, cada un pot prendre dos valors $n_i = 0$ o 1, $i = 1 \ldots N$. Per simular l'activitat de les neurones reals, se suposa que els canvis d'estat tenen lloc a intervals de temps discrets $t = 0, 1, 2, \ldots$ El nou estat de cada unitat i ve donat per la influència de la resta a través de la combinació lineal:

$$h_i(t) = \sum_j \omega_{ij} n_j(t). \qquad (2)$$

ω és una matriu on els coeficients –o *pesos*– descriuen les intensitats sinàptiques entre cada parella ij de neurones. $h_i(t)$ representa el potencial sinàptic (o el camp) produït sobre la unitat i per l'acció de les altres. Considerant h com a input i n com output, les propietats del sistema estan determinades per la relació funcional entre els $n_i(t+1)$ i els $h_i(t)$. En el cas més simple, s'acostuma a usar una llei del tipus

$$n_i(t+1) = \Theta(h_i(t) - \theta_i), \qquad (3)$$

on θ_i és un cert llindar d'activació i Θ és la funció esglaó, $\Theta(x) = 0$ per a $x \leq 0$ i 1 per a $x > 0$.

McCulloch i Pitts mostraren com aquestes xarxes poden fer en principi qualsevol càlcul, de forma semblant a un ordinador digital o la seva abstracció matemàtica, la *màquina de Turing*. L'equivalent al programa seria la matriu ω, que governa el procés. No obstant això, els passos no s'executen seqüencialment, sinó en paral·lel dins de cada unitat. De fet, pot dir-se que el codi font es redueix a una sola instrucció, la resultant de combinar les eq. (2) i (3). Aquesta petitesa està compensada per l'ús d'un gran nombre d'unitats processadores en comptes d'una de sola. En el cervell humà pot haver-hi fins a quantitats de l'ordre de 10^{11} neurones.

El programador d'una xarxa de McCulloch-Pitts ha d'escollir la ω adequada per fer una certa tasca *cognitiva* donada. Aquesta mena de tasca pot ser qualsevol problema que comporti el procés digital o analògic d'informació, com el reconeixement de *patterns* en imatges òptiques o acústiques. La primera solució de caràcter general fou donada el 1961 per Eduardo Caianiello, en forma d'algorisme "d'aprenentatge" per trobar les connexions sinàptiques d'una xarxa neuronal. El mètode, anomenat originalment equació 'mnemònica', incorpora el principi conegut com la regla de Hebb.

Regla de Hebb

Per explicar-la amb facilitat, cal avançar idees relatives al model de Hopfield. Suposem que ara els dos estats de les unitats tenen valors ± 1 en lloc de $0, 1$, i diguem-los s_i. Tindrem $s_i = 2n_i - 1$. Usant la relació entre les

funcions Θ i sgn –funció signe– la combinació de (2) i (3) esdevé una llei de tipus

$$s_i(t+1) = \mathrm{sgn}\left(\sum_{j=1}^{N} \omega_{ij} s_j(t) - \bar{\theta}_i\right), \qquad (4)$$

on les $\bar{\theta}_i$ depenen de les θ_i i de ω. De totes maneres suposarem $\bar{\theta}_i = 0$, que és el que s'acostuma a prendre. Volem ara escollir ω de manera que partint d'un estat $s = (s_1, \ldots, s_N)$ la xarxa evolucioni cap a algun dels *patterns* preestablerts $\{\xi^\mu, \mu = 1, \ldots, p\}$, –cada ξ^μ denota $(\xi_1^\mu, \ldots, \xi_N^\mu)$. Concretament es desitja que l'estat s de la xarxa acabi formant el ξ^μ més proper o semblant a s, i que aquest *pattern* reconegut quedi assenyalat per la persistència de la xarxa un cop que hagi arribat a formar-lo. Perquè això funcioni és necessari, doncs (i entre altres coses), que les configuracions corresponents a s igual a algun dels ξ^μ siguin estables sota evolució temporal.

En el cas especial que el conjunt de *patterns* donats es reduis a un sol ξ ($p = 1$) hi ha una senzilla solució, que és

$$\omega_{ij} = \frac{1}{N} \xi_i \xi_j \qquad (5)$$

Gràcies al fet que $\xi_i^2 = 1$, aquesta relació permet comprovar immediatament per mitjà de (4) (amb $\bar{\theta} = 0$) que

$$s(t) = \xi \;\Rightarrow\; s(t+1) = \xi, \qquad (6)$$

i així per a tot t successiu.

La llei de Hebb, que s'usa quan $p > 1$, és la generalització de (5):

$$\omega_{ij} = \frac{1}{N} \sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu \qquad (7)$$

Ara, però, cada ξ^μ no compleix (6) sempre. Un càlcul senzill basat en l'aplicació de (4) (per a $\bar{\theta} = 0$) mostra que apareixen termes addicionals que no teníem quan $p = 1$. Aquestes contribucions podrien alterar els signes de manera que cada ξ^μ deixés de ser estable sota evolució temporal. Una avaluació estadística d'aquests nous termes ens permet dir que la propietat d'estabilitat es manté si $\sqrt{(p-1)/N}$ és proper a zero, és a dir, si p és petit en comparació amb N.

Perceptrons

El 1960 Frank Rosenblatt i el seu grup estudiaren un model particular de xarxa que anomenaren *perceptró*. La versió més senzilla consta de dues capes de neurones, que representen entrada i sortida. Les neurones de la capa sortida reben senyals sinàptiques de la capa entrada, però no a l'inrevés, i a cada capa no es comuniquen entre elles (figura 5). El flux d'informació es propaga direccionalment i en un sol sentit –suposadament cap endavant– i per això es diu que aquestes xarxes són de classe *feed-forward*.

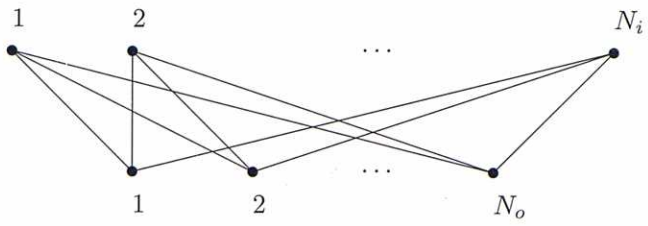

Figura 5: *Xarxa feed-forward per a un perceptró de dues capes: entrada de N_i unitats, i sortida de N_o. Les connexions no són dobles, sinó que només tenen lloc en sentit d'entrada a sortida*

Aquest grup de recerca generà un algorisme iteratiu –conegut com a *perceptron learning rule*– per construir la matriu sinàptica ω de tal manera que un cert *pattern* binari, llegit com a entrada, es transformi en el desitjat en sortida, i fins i tot va ser capaç de demostrar-ne la convergència. Alguns anys més tard arribà la crítica de M. Minsky i S. Papert (Minsky, 1988), assenyalant que aquella demostració només servia per a la classe de problemes resolubles mitjançant perceptrons de dues capes. No contents amb això, mostraren problemes molt senzills que no es poden resoldre amb aquests models. El més conegut és la realització de la porta lògica de la disjunció exclusiva, on entrada i sortida són dues i una unitats binàries, respectivament, de tal manera que el resultat sigui 1 quan una sola de les entrades és 1, i 0 en els altres casos. Com que es tracta d'un problema corrent en disseny d'ordinadors, i que es resol amb facilitat en arquitectures convencionals, aquesta constatació va ser un cop prou destructiu a la idea del perceptró.

Xarxes neuronals i reticles de spins

Un altre desenvolupament arrenca de quan William Little notà la semblança entre una xarxa neuronal de McCulloch-Pitts i un reticle de spins a l'estil dels models d'Ising. En aquests sistemes, cada spin s_i en un lloc i pot prendre només dues orientacions: $s_i = +1$ (cap amunt) i $s_i = -1$ (cap avall). L'analogia amb una xarxa neuronal es fa identificant cada spin amb una neurona i associant l'orientació $s_i = +1$ a l'estat actiu $n_i = 1$ i l'orientació $s_i = -1$ a l'estat pasiu $n_i = 0$, que no és més que el canvi d'elements binaris a elements tipus signe abans comentat.

La idea fou continuada pel mateix Little i per John Hopfield, que estudià com un sistema d'aquesta naturalesa pot servir per emmagatzemar i recuperar informació. A la pràctica, es recuperen *patterns* per persistència de configuracions sota l'evolució temporal de la xarxa. Els models de Little i de Hopfield només difereixen en la manera en què té lloc el canvi al llarg del temps. En el model de Little, tots els spins o neurones són actualitzats alhora en passar de t a $t+1$, mentre que en el de Hopfield són actualitzats seqüencialment,

un per cada nou t, ja sigui per ordre de posició o a l'atzar. Encara que aquest segon mètode dóna facilitats per a l'anàlisi teòrica i la simulació, està de fet abandonant la característica bàsica de les xarxes neuronals: l'operació simultània de moltes unitats en paral·lel.

L'analogia amb sistemes de spins ha donat un gran profit, gràcies als avenços en la comprensió de les propietats termodinàmiques de sistemes de spins en desordre, particularment xarxes amb fase de vidre de spins, a la qual s'ha arribat tan sols en els anys vuitanta. Per tal d'aplicar aquest cúmul de resultats, cal substituir la llei d'evolució determinista (4) per una regla estocàstica, on el valor de $s_i(t+1)$ és assignat aleatòriament i segons una probabilitat que és funció de h_i i que depèn també d'un paràmetre $\beta = 1/T$, on T fa el paper de temperatura. Aquesta T, però, no s'ha d'interpretar com la temperatura física de la xarxa, sinó com una forma d'introduir soroll estocàstic que permeti d'aplicar les tècniques de la mecànica estadística. La funció probabilitat és tal que en el límit $T \to 0$ reapareix la llei (4). Normalment, això es fa prenent

$$P_\beta(s_i(t+1) = \pm 1; h_i) = \frac{1}{1 + e^{\mp 2\beta h_i(t)}}, \qquad (8)$$

on ara

$$h_i(t) = \sum_j \omega_{ij} s_j(t). \qquad (9)$$

S'estudiaran les situacions d'equilibri dinàmic, que s'ateny quan l'estat de qualsevol neurona ja no canvia en mitjana al llarg del temps. Aquesta suposició porta a una certa forma per a la funció de distribució de la probabilitat $P[s]$ que les neurones de la xarxa estiguin en l'estat $s = (s_1, \ldots, s_N)$, la qual resulta ser una distribució de Boltzmann de l'estil

$$P[s] = \frac{1}{Z} e^{-\beta E[s]}, \qquad (10)$$

$$E[s] = -\frac{1}{2} \sum_{ij} \omega_{ij} s_i s_j \qquad (11)$$

$E[s]$ és la funció energia de la xarxa, i U en (1) ve a ser el seu valor mitjà. Com que ω és simètrica, es tracta d'una funció definida negativa. Si, com és habitual, els seus coeficients estan donats per la llei de Hebb, veiem que quan $s = \xi^\nu$, per a algun ν, ateny un mínim absolut. Poden, però, haver-hi mínims locals que no corresponguin a cap ξ^ν. $\frac{1}{Z}$ apareix com un factor de normalització a determinar que fa que la suma de totes les probabilitats sigui 1, de la qual cosa surt

$$Z = \sum_s e^{-\beta E[s]}.$$

Podem veure Z com la funció de partició canònica, que es relaciona amb l'energia lliure de Helmholtz F (1) per mitjà de $Z = e^{-\beta F}$, d'on

$$F = -T \ln Z.$$

Per a processos a T fixada, aquesta té el paper d'un potencial termodinàmic. Semblantment, l'entropia S de la xarxa es defineix com

$$S = -\frac{\partial F}{\partial T}.$$

Com que normalment Z creix amb T i acostuma a ser molt més gran que 1, l'energia lliure F és negativa i decreix amb T. D'aquí que l'entropia sigui definida positiva i només s'anul·li per a $T = 0$.

La mitjana (*tèrmica*) de l'energia en prendre els valors corresponents a totes les possibles configuracions de s amb probabilitat donada per (10) és

$$< E >= \sum_s E[s]P[s] = -\frac{\partial}{\partial \beta} \ln Z.$$

Usant un truc matemàtic consistent a afegir termes de tipus "font o camp extern" a l'expressió $E[s]$ i prenent derivades respecte a aquests en zero poden també calcular-se quantitats com:

- valor mitjà de l'activitat de la neurona i-èsima $< s_i >$,
- correlacions entre parells de neurones $< s_i s_j >$, totalment anàlogues a les correlacions entre dos spins,
- correlacions entre tres o més neurones, etc.

El paral·lelisme amb el comportament col·lectiu dels reticles de spins s'obté, a part de per l'existència de dos valors possibles en cada element, per la interpretació de la funció energia (11). Mirant-ho com una interacció entre spins, aquesta pot tenir alhora termes ferro i antiferromagnètics segons els possibles signes dels ω_{ij}, que en general no seran tots iguals. De fet, si pensem que els ξ^μ han estat donats aleatòriament, llavors com a conseqüència d'aplicar la llei de Hebb les ω_{ij} també seran aleatòries. Això ens porta novament a la presència d'interaccions ferro i antiferromagnètiques aleatòriament barrejades, que és un dels trets dels vidres de spins.

Considerada com a sistema per trobar el *pattern* ξ més pròxim a un s donat, la xarxa "funcionarà bé" en situacions en que, a l'equilibri dinàmic, els mínims de E associats als ξ^μ siguin els únics mínims estables de F. Aquesta imatge pot espatllar-se si la temperatura és prou gran perquè l'estocasticitat confongui l'evolució cap als ξ^μ. Es tractaria del cas, abans comentat, on la tendència entròpica predomina sobre la de reducció de l'energia.

Com abans hem dit, la llei de Hebb condueix a una bona recuperació dels ξ^μ quan p es petit en comparació amb N. Aquesta qualitat es descriu pel paràmetre $\alpha = \frac{p}{N}$, també anomenat *capacitat*, perquè mesura el nombre de *patterns* "en memòria" amb relació a la mida de la xarxa. És per això que els augments de p, i per tant de α, tenen un efecte semblant a fer créixer la temperatura, encara que no sigui del tot equivalent.

Sense necessitat d'haver arribat a un *scenario* de desordre tèrmic total, cal també considerar casos amb presència de mínims de F que ja no ho siguin els corresponents als ξ^μ, i la possibilitat que esdevinguin més estables que els primers. De fet pot ser que aquests perdin l'estabilitat "global", i passin a ser tan sols metastables.

El resultat a gran escala dependrà, doncs, dels valors i relacions entre T i α. Aquests dos paràmetres permeten construir un diagrama de fases on es tenen quatre regions, associades a diferents tipus de comportament del sistema:

- **P**, fase paramagnètica. Tota mena d'ordre queda destruït perquè la temperatura aquí és alta. En termes magnètics, hi ha una "desalineació" completa entre els spins.

- **F**, fase ferromagnètica, a temperatures baixes i α no excessivament gran. En aquesta fase, els ξ^μ que són mínims de l'energia, són també mínims estables de F de manera que es poden veure com els seus "estats fonamentals". Es diu que hi ha ordre a llarg abast perquè totes les parelles ij de spins estan alineats segons un sentit preferit, definit pel signe de la interacció per a cada terme –és a dir, de cada ω_{ij}. Notem que, aquí, l'ús de la paraula "ferromagnètica" no vol dir que totes les interaccions siguin del mateix signe, sinó que, als estats energèticament òptims, *tots* els spins queden orientats segons els signes de les interaccions.

- **F+SG**, fase de *barreja* ferromagnètica i de vidre de spins, a temperatures una mica per sobre de les de la fase **F**. Els mímims corresponents als ξ^μ s'han fet ara metastables. Els estats on el sistema va finalment a parar, diem que són de vidre de spins, perquè presenten frustració respecte a les alineacions ideals donades per ω, i perquè n'hi ha molts de possibles, separats per barreres energètiques altes.

- **SG**, fase purament de vidre de spins, situada entre la fase **P** i la **F+SG**. Com l'anterior, però amb la diferència que ara els ξ^μ ja no poden recuperar-se amb cap mena d'estabilitat, ni tan sols com a estats metastables.

- Existeix encara una altra fase, on està trencada la *simetria de rèplica*, a temperatures molt baixes. Es tracta de la regió on el *truc de rèplica* esmentat a la secció 3 deixa de ser vàlid. Encara que la seva importància teòrica és inqüestionable, el seu efecte sobre el funcionament d'una xarxa en les condicions habituals és més aviat petit, i per parlar-ne hauríem d'endinsar-nos més en els aspectes matemàtics, cosa que queda fora de l'objectiu d'aquest treball.

A $T = 0$, es produeix una transició discontínua (mirant variacions de α) de **F** a **F+SG** per a $\alpha \simeq 0.051$. Després hi ha una altra transició discontínua, de **F+SG** a **SG**, que és el col·lapse d'una fase "magnetitzada" on encara es recorden els ξ^μ, a una "sense magnetització" on hi ha amnèsia o confusió total. Aquesta té lloc per a un valor de α anomenat $\alpha_c(0) \simeq 0.138$. El mateix passa a T més altes, amb els valors de les $\alpha_c(T)$ de transició disminuint a mesura que T creix, fins a arribar a $T = 1$, on $\alpha_c(1) = 0$. Per sobre de $T = 1$ apareix la fase **P** i només hi ha el pas d'aquesta a **SG**. La recuperació dels ξ^μ és ja impossible.

Pel que fa a la separació entre les fases, **P** i **SG**, la tenim a $T = 1$ per a $\alpha = 0$ i a T lleugerament més grans en augmentar α. La divisòria entre les dues regions és una línea de transició de fase *de segon ordre*.

Xarxes multicapa

En els darrers anys, l'interès per xarxes *feed-forward* de diverses capes ha tornat a l'actualitat. El motiu ha estat el redescobriment d'un algorisme per trobar la matriu sinàptica en xarxes amb més de dues capes, és a dir, amb capes intermèdies o *amagades*, tal com la que es mostra a la figura 6.

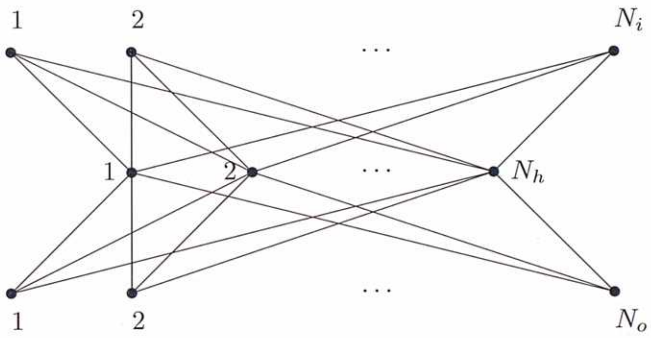

Figura 6: *Xarxa feed-forward de tres capes: entrada de N_i unitats, una capa amagada de N_h, i sortida de N_o neurones. En general es poden dissenyar xarxes d'aquesta classe amb qualsevol nombre de capes intermèdies*

La potència del mètode, actualment conegut com *error backpropagation*, fou reconeguda (una dècada i mitja més tard de la seva invenció) per diversos grups investigadors, especialment pel de Rumelhart (San Diego, CA), que el llençà a la fama amb el seu article a *Nature* (Rumelhart, 1986b). Aquest algorisme d'aprenentatge es basa en la modificació iterativa de les ω_{ij} a fi que, per a cada exemple conegut, el senyal de sortida difereixi el menys possible del desitjat. L'objectiu s'aconsegueix aplicant el mètode del *gradient*, que dóna els valors adequats de les variacions $\delta\omega_{ij}$. Ja que el funcionament de la xarxa és la realització d'una funció altament no-lineal –com ho és (3)– entre entrada i sortida, el mètode s'aplica molts cops fins que se n'observa convergència.

Backpropagation és un d'entre molts possibles algorismes d'aprenentage de la categoria anomenada *supervised learning*, on a cada pas la xarxa ajusta els

pesos comparant la sortida real amb l'esperada. Encara que matemàticament aquests algorismes són seriosos i eficients, la seva realització biològica sembla molt improbable, i només són viables quan la sortida desitjada es coneix amb precisió. Altres formes d'aprenentatge basades en principis de premi i sanció –*reward and penalty*– i d'evolució i selecció estan sent objecte d'estudi continuat.

Problemes oberts i conclusions

Hi ha molts problemes oberts en la teoria dels vidres de spins encara que creiem que no seria oportú comentarlos tots per la manca d'espai. En recordarem un d'important. La teoria presentada a la secció 3 és una teoria de camp mitjà, és a dir, és una aproximació del que succeeix en la realitat. S'ha revelat útil per descobrir aquest nou tipus de fenomen cooperatiu però caldria saber-hi treballar en qualsevol cas general usant les tècniques estàndard de la mecànica estadística. Val a dir que aquestes són la teoria de les pertorbacions i la teoria del grup de renormalització. La primera és útil per conèixer les propietats termodinàmiques generals lluny del punt crític. La segona és especialment apropiada per entendre les propietats crítiques (tipus de transició, valors dels exponents...).

La teoria dels vidres de spins ha experimentat un munt d'aplicacions diferents. Això és degut al fet que les característiques principals dels sistemes que estudia (desordre i frustració) feien inviables les tècniques usuals d'anàlisi. Entre les aplicacions destacarem tot l'estat sòlid que va des de l'estudi dels superconductors a l'estudi dels polímers passant pels fenòmens de conducció elèctrica tals com els fenòmens de localització. També s'ha aplicat a l'estudi de problemes d'optimització matemàtica i ha tingut un èxit especial en l'estudi de sistemes biològics (xarxes neuronals, evolució prebiòtica, conformació de proteïnes...). I des de fa molt poc es comença a fer recerca en la implicació d'aquesta teoria en economia i en sociologia.

Possiblement la forma millor d'acabar aquesta presentació és usar l'analogia per fer veure com els vidres de spins poden ser útils per entendre certs fenòmens sociològics. Suposem una massa social (per exemple, els habitants de Barcelona). Tenim la tendència a pensar que som éssers individuals i que l'evolució de la gran col·lectivitat (diguem-li ciutat) és una cosa abstracta i inassequible per a cadascun de nosaltres. Res més lluny de la realitat. En efecte, cadascun de nosaltres està lligat a qualsevol persona de la ciutat per una cadena d'amics que potser és de l'ordre de 5 persones com a màxim. Ja que si cadascun de nosaltres té de mitjana de l'ordre de 30 coneguts està connectat amb 30^5 persones diferents a través d'una cadena màxima de 5 persones. I això és de l'ordre de 10^6 persones que és també de l'ordre del nombre dels habitants de Barcelona. Així doncs és molt probable que agafant dues persones a l'atzar d'una ciutat hi hagi una cadena de com a màxim 5 persones que els connecta. I és aquesta gran connectivitat la que fa que un acte de decisió o opinió personal pugui arribar a qualsevol punt de la massa social. O bé que un acte de decisió personal es vegi influït pel d'una altra persona arbitrària de la massa. La configuració de la massa social (val a dir, tots els paràmetres sociològics que la caracteritzen) és el resultat d'un conflicte constant entre decisions personals i decisions d'altres persones que conflueixen en cadascun de nosaltres. D'alguna forma estem frustrats ja que no existeix mai una solució a un problema que afecti tots els habitants de la ciutat que satisfaci tothom. I la teoria dels vidres de spins ens recorda que la solució òptima a un problema col·lectiu no és única. Més aviat, n'hi ha moltes i de molt diferents, totes igual de bones. I no seria estrany que totes fossin organitzades d'alguna forma jeràrquica. Concretar això és una tasca que pertoca a la sociologia.

Tot el que hem explicat és un exemple i està molt lluny de poder-se quantificar. Per tant, ha de romandre com una simple analogia, que és el que és. La teoria dels vidres de spins és per si sola una contribució molt important a la mecànica estadística. El temps dirà si permet obrir una via de coneixement alternativa a tants i tants problemes diferents que, després d'haver-se classificat amb el nom de complexos, s'han demostrat intractables usant tècniques d'estudi estàndard.

Agraïments

Volem donar les gràcies a Giorgio Parisi –ara a *I Università di Roma, "La Sapienza"*– tant pel que ha contribuït a la nostra formació en aquestes matèries com per l'extens suport rebut de part seva. F.R. agraeix una beca de la CE.

Referències

D.E. RUMELHART, J.L. MCCLELLAND and the PDP Research Group, *Parallel Distributed Processing, Vols 1 and 2*, MIT Press (1986).

M. MINSKY i S. PAPERT, *Perceptrons: An Introduction to Computational Geometry*, MIT Press, Cambridge (1988).

D.E. RUMELHART, G.E. HINTON i R.J. WILLIAMS, *Nature*, **323**, 533 (1986).

M. MÉZARD, G. PARISI i M. A. VIRASORO, *Spin Glass Theory and Beyond*, World Scientific, 1987.

G. PARISI, *Field Theory, Disorder and Simulations*, World Scientific, Singapore 1992.