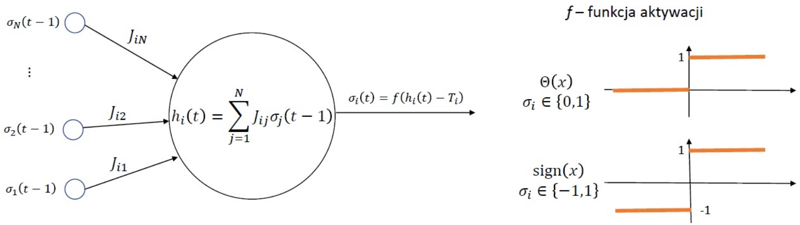
Sieci Hopfielda (sieci Hebba)

Sieci neuronowe: wykład 04

R.A. Kosiński, T. Gradowski, A. Krawiecki

Politechnika Warszawska
Wydział Fizyki

Neuron formalny McCullocha-Pittsa



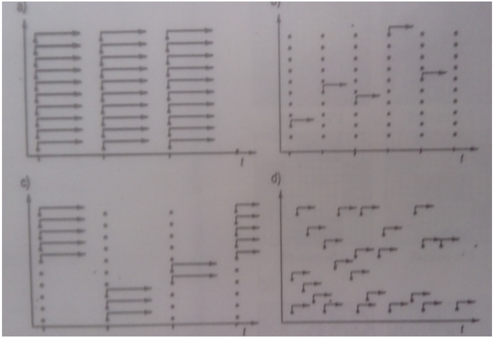
Rysunek: Neuron formalny McCullocha-Pittsa

- Sieć neuronowa składa się z $i = 1, 2, \ldots N$ neuronów dwustanowych (formalnych),
- Stan neuronu w chwili t zależy od stanu neuronów sąsiednich w chwili $t-1$ i funkcji aktywacji f,

$$\sigma_i(t) = f(h_i(t-1)), h_i(t) = \sum_{j=1}^{N} J_{ij}\sigma_j(t-1),$$

gdzie J_{ij} jest **wagą połączenia synaptycznego**, h_i jest **polem lokalnym** neuronu i,

- Stan sieci: wektor $I(t) = \{\sigma_i(t)\}$.



Rodzaje dynamiki sieci

(a) **Dynamika synchroniczna**: stan wszystkich neuronów aktualizowany jest jednocześnie, w kolejnych chwilach wg. taktów zegara,

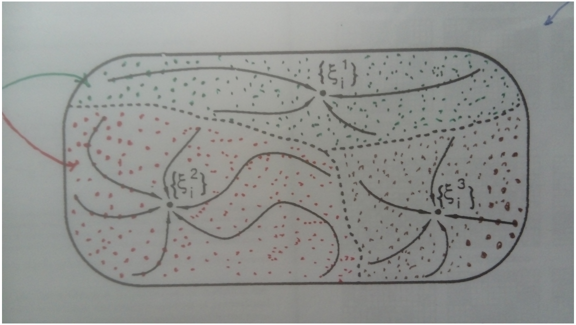
(b) **Dynamika asynchroniczna**: w każdym kroku symulacji Monte Carlo wybierany jest przypadkowo (bez powtórzeń) jeden neuron i jego stan jest aktualizowany; pełny krok dynamiki odpowiada aktualizacji N neuronów,

(c) **Dynamika synchroniczna blokowa**,

(d) **Dynamika asynchroniczna bez zegara**, ...

Atraktory sieci

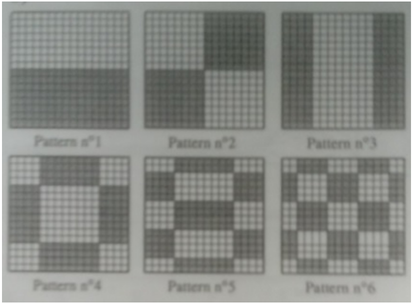
- W sieciach z dynamiką asynchroniczną jedynymi możliwymi atraktorami są **atraktory punktowe**: po dostatecznie długim czasie stan sieci się nie zmienia, $I(t) = I = \mathrm{const}$
- W sieciach z dynamiką synchroniczną moga pojawić się atraktory okresowe i chaotyczne,
- **Atraktor okresowy**: stan sieci powtarza się okresowo z okresem T, $I(t+T) = I(t)$,
- **Atraktor chaotyczny**: stan sieci zmienia się nieregularnie. W przypadku sieci o skończonej liczbie neuronów N stan sieci powtórzy się po czasie porównywalnym z liczbą różnych możliwych stanów sieci 2^N. Przy bardzo podobnych stanach początkowych $I^1(0)$, $I^2(0)$ po dostatecznie długim czasie stany $I^1(t)$, $I^2(t)$ bardzo się różnią.



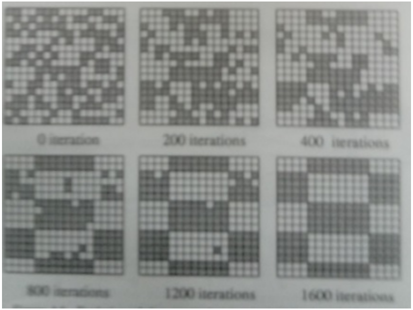
Rysunek: Schematyczne obszary przyciągania atraktorów punktowych, odpowiadające zapamiętanym wzorcom ξ^1, ξ^2, ξ^3.

Obszar przyciagania atraktora: zbiór stanów początkowych sieci, które po dostatecznie długim czasie prowadzą do osiągnięcia przez stan sieci danego atraktora.

Elementy dynamiki sieci neuronowych (5)



Rysunek: Wzorce zapamiętane.



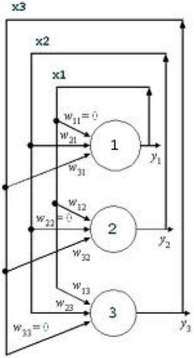
Rysunek: Rozpoznanie przez sieć wzorca nr 5.

Rozpoznanie wzorca. Po zadaniu wzorca początkowego w postaci stanów początkowych wszystkich neuronów, sieć rozpoznaje odpowiadający mu wzorzec zapamiętany, osiągając odpowiadający mu stan końcowy.

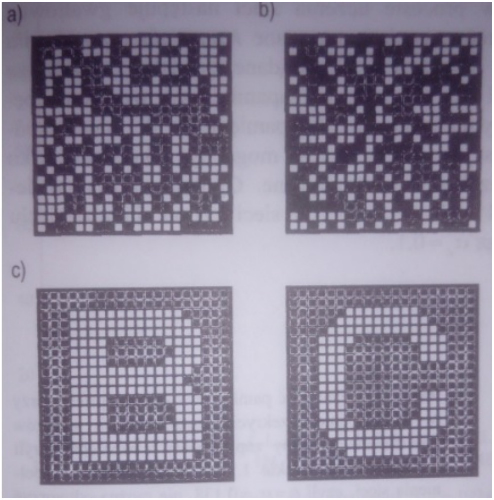
- Wzorzec zapamiętany musi być więc atraktorem punktowym sieci,
- Rozpoznanie wzorca jako jednego z wzorców zapamiętanych oznacza, że wzorzec początkowy należał do obszaru przyciagania wzorca zapamiętanego (czyli zapewne wykazywał do niego jakieś podobieństwo),
- Sieć posiada więc tzw. **pamięć asocjacyjną (skojarzeniową)**.

Model N neuronów dwustanowych, oddziałujących każdy z każdym, z symetrycznymi wagami połączeń synaptycznych.

- Neurony dwustanowe $\sigma_i \in \{-1, 1\}$, $i = 1, 2, \dots N$,
- Funkcja aktywacji $f(x) = \mathrm{sign}(x)$, progi aktywacji $T_i = 0$, $i = 1, 2, N$,
- Dynamika asynchroniczna,
- Zbiór zapamiętanych **wzorców** $I^\mu = \{\xi_i^\mu\}$, $\mu = 1, 2, \dots P$; $i = 1, 2, \dots N$; $\xi_i^\mu \in \{-1, 1\}$,
- **Przekrycia** stanu sieci z wzorcem μ: $M^\mu = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i$, $\mu = 1, 2, \dots P$,
- Wzorce **powinny być nieobciążone**, tj. prawdopodobieństwa wystąpienia $\xi_i^\mu = +1$ i $\xi_i^\mu = -1$ powinny być jednakowe, $P(\xi_i^\mu) = \frac{1}{2}\left[\delta(\xi_i^\mu - 1) + \delta(\xi_i^\mu + 1)\right]$,
- Zbiór wzrorców **powinien być nieskorelowany** (ortogonalny) $M^{\mu\mu'} = \frac{1}{N}\sum_{i=1}^{P} \xi_i^\mu \xi_i^{\mu'} = \delta\mu, \mu'$, $\mu, \mu' = 1, 2, \dots P$.



Sieci Hopfielda: definicja(2)



Rysunek: Przykłady wzorców: (a) przypadkowy nieobciążony, (b) przypadkowy obciążony - przewaga neuronów wzbudzonych, (c) - dwa wzorce skorelowane.

Sieci Hopfielda: definicja (3)

Chcemy, aby zapamiętane wzorce były przyciągającymi punktami stałymi (atraktorami punktowymi) dynamiki sieci. Aby to osiągnąć, wprowadzamy energię sieci (Hamiltonian), która osiąga minimum, gdy przekrycie stanu sieci $I = \{\sigma_i\}$ z jednym z zapamiętanych wzorców jest duże,

$$H(I) = -\frac{1}{2} N \sum_{\mu=1}^{P} (M^\mu)^2. \tag{1}$$

$$H(I) = -\frac{1}{2N} \sum_{\mu=1}^{P} \left(\sum_{i=1}^{N} \xi_i^\mu \sigma_i \right)^2 = -\frac{1}{N} \sum_{\langle i,j \rangle} \left(\sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \right) \sigma_i \sigma_j + const. \tag{2}$$

Jest to Hamiltonian **modelu Isinga** z symetrycznymi całkami wymiany

$$H(I) = -\sum_{\langle i,j \rangle} J_{ij} \sigma_i \sigma_j, \quad J_{ij} = J_{ji} = \frac{1}{N} \sum_{\mu=1}^{P} \xi_i^\mu \xi_j^\mu \tag{3}$$

Reguła Hebba uczenia sieci (zmodyfikowana, symetryczna)

Wagi połączeń synaptycznych w równ. (3) mogą powstać w wyniku procedury uczenia sieci neuronowej, polegającej na tym, że po prezentacji wzorca μ połączenie neuronów i, j zmienia się o wartość

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

Dla ułatwienia zapisu można wprowadzić

- P-wymiarowy wektor przekryć $\tilde{M} = (M^1, M^2, \ldots M^P) = \{M^\mu\}$,
- P-wymiarowy wektor wartości wzorców w węźle i $\tilde{\xi}_i = (\xi_i^1, \xi_i^2, \ldots \xi_i^P) = \{\xi_i^\mu\}$,
- Średnie po węzłach sieci, oznaczane kreską u góry, np.

$$\overline{\tilde{\xi}\, \mathrm{tgh}(\beta \tilde{\xi} \cdot \tilde{M})} = \frac{1}{N} \sum_{i=1}^{N} \tilde{\xi}_i\, \mathrm{tgh}(\beta \tilde{\xi}_i \cdot \tilde{M}) = \frac{1}{N} \sum_{i=1}^{N} \tilde{\xi}_i\, \mathrm{tgh} \left(\beta \sum_{\mu=1}^{P} \xi_i^\mu M^\mu \right)$$

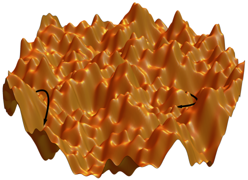
Przykład: $N = 5$, $P = 2$
$\{\xi_i^1\} = (1, -1, 1, 1, 1,)^T$
$\{\xi_i^2\} = (1, 1, 1, 1, -1)^T$

$$[J_{ij}] = \frac{1}{5}\begin{pmatrix} 2 & 0 & 2 & 2 & 0 \\ 0 & 2 & 0 & 0 & -2 \\ 2 & 0 & 2 & 2 & 0 \\ 2 & 0 & 2 & 2 & 0 \\ 0 & -2 & 0 & 0 & 2 \end{pmatrix}$$

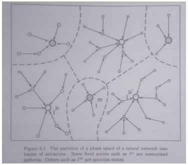
Wektor pól lokalnych dla stanu sieci
$\{\sigma_i\} = (\sigma_1, \sigma_2, \ldots \sigma_N)^T$

$$\{h_i\} = [J_{ij}]\{\sigma_i\}$$

$\{h_i^1\} = [J_{ij}]\{\xi_i^1\} = \frac{1}{5}(6, -4, 6, 6, 4)^T$
$\{h_i^2\} = [J_{ij}]\{\xi_i^2\} = \frac{1}{5}(6, 4, 6, 6, -4)^T$
Wniosek: wzorce są stabilne



Rysunek: Krajobraz energetyczny sieci Hopfielda.



Rysunek: Obszary przyciagania wzorców (minimów krajobrazu energetycznego) odpowiadają poszczególnym dolinom krajobrazu.

Rozpatrujemy stabilnośc zapamiętanych wzorców w sieci Hopfielda bez szumu

- Pole lokalne neuronu i: każdy wzorzec I^μ wnosi wkład do pola lokalnego proporcjonalny do swego przekrycia ze stanem I

$$
\begin{aligned}
h_i(I) &= \sum_{j=1}^{N} J_{ij}\sigma_j(I) = \sum_{j=1}^{N}\left(\frac{1}{N}\sum_{\mu=1}^{P}\xi_i^\mu\xi_j^\mu\right)\sigma_j(I) \\
&= \sum_{\mu=1}^{P}\frac{1}{N}\left(\sum_{j=1}^{N}\xi_j^\mu\sigma_j(I)\right)\xi_i^\mu = \sum_{\mu=1}^{P} M^\mu(I)\xi_i^\mu
\end{aligned}
$$

- W dalszym ciągu zakładamy, że stan sieci pokrywa się z wzorcem μ_0, $I = I^{\mu_0}$, którego stabilność rozpatrujemy, więc $\sigma_i(I^{\mu_0}) = \xi_i^{\mu_0}, i = 1, 2, \dots N$.

- Pole lokalne składa się ze **składnika sygnałowego** związanego ze wzorcem μ_0 i **przesłuchu** pochodzącego od pozostałych wyuczonych wzorców

$$\begin{aligned} h_i(I^{\mu_0}) &= \frac{1}{N}\sum_{j=1}^{N}\left(\xi_i^{\mu_0}\xi_j^{\mu_0}\right)\xi_j^{\mu_0} + \frac{1}{N}\sum_{\mu\neq\mu_0}\sum_{j=1}^{N}\left(\xi_i^{\mu}\xi_j^{\mu}\right)\xi_j^{\mu_0} \\ &= \xi_i^{\mu_0} + \frac{1}{N}\sum_{\mu\neq\mu_0}\sum_{j=1}^{N}\xi_i^{\mu}\xi_j^{\mu}\xi_j^{\mu_0} \end{aligned}$$

- Wzorzec μ_0 **jest stabilny**, jeżeli w każdym węźle pole lokalne ma kierunek zgodny z neuronem, czyli (wektorowy) **parametr stabilności** $\{x_i(I^{\mu_0})\}$ jest dodatni

$$x_i(I^{\mu_0}) = \xi_i^{\mu_0} h_i(I^{\mu_0}) = 1 + \frac{1}{N}\sum_{\mu\neq\mu_0}\sum_{j=1}^{N}\xi_i^{\mu}\xi_i^{\mu_0}\xi_j^{\mu}\xi_j^{\mu_0} > 0 \tag{4}$$

Fundamentalna degeneracja sieci. Jeżeli wzorzec $I = \{\sigma_i\}$ jest stabilny, to również $I' = -\{\sigma_i\}$ jest stabilny:

$$\begin{aligned} 0 < x_i(I) &= \sigma_i(I)\sum_{j=1}^{N} J_{ij}\sigma_j(I) = -\sigma_i(I)\sum_{j=1}^{N} J_{ij}(-\sigma_j(I)) \\ &= \sigma_i(I')\sum_{j=1}^{N} J_{ij}\sigma_j(I') = x_i(I') \end{aligned}$$

Sprawdzamy, jaka jest maksymalna pojemność sieci $\alpha_c = P_c/N$, przy której zapamiętany wzorzec I^1, pozostanie stabilny przy odwróceniu R neuronów (gdy sieć znajdzie się w stanie $I^{1,R}$); oznacza to, że odwrócenie stanów R neuronów nie wyprowadza sieci poza lokalne minimum energetyczne związane ze wzorcem I^1.

Parametr stabilności

$$x_i(I^{1,R}) = \xi_i^1 h_i(I^{1,R}) = x_i^s + x_i^n$$

$$x_i^s = \xi_i^1 \frac{1}{N} \sum_{j=1}^{N} \left(\xi_i^1 \xi_j^1\right) \sigma_j(I^{1,R}) = \left(1 - \frac{2R}{N}\right)$$

(ponieważ odwrócenie każdego neuronu obniża pole lokalne o 2);

$$x_i^n = \frac{1}{N} \sum_{\mu \neq \mu_0} \sum_{j=1}^{N} \xi_i^\mu \xi_i^1 \xi_j^\mu \sigma_j(I^{1,R})$$

Składnik związany z przesłuchem jest sumą NP liczb przypadkowych o wartościach ± 1, występujących z jednakowym prawdopodobieństwem, więc można (tw. de Moivre'a - Laplace'a) przybliżyć go liczbą losową z rozkładu Gaussa o średniej zero i wariancji $\sqrt{\langle (x_i^n)^2 \rangle} = \frac{1}{N}\sqrt{NP}$

$$P(x_i^n) = \frac{1}{\sqrt{2\pi \langle (x_i^n)^2 \rangle}} \exp\left[-\frac{(x_i^n)^2}{2\langle (x_i^n)^2 \rangle}\right], \quad \langle (x_i^n)^2 \rangle = \frac{P}{N} \equiv \alpha$$

Odwrócenie stanów R neuronów zdestabilizuje stan neuronu w węźle i jeżeli $x_i(I^{1,R}) = x_i^s + x_i^n < 0$.

$$\begin{aligned} P(x_i^s + x_i^n < 0) &= P(x_i^n < -x_i^s) = \int_{-\infty}^{-x_i^s} P(x_i^n)dx_i^n = \frac{1}{2} - \int_0^{x_i^s} P(x_i^n)dx_i^n \\ &= \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{x_i^s}{\sqrt{2\langle (x_i^n)^2\rangle}}\right)\right] = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{x_i^s}{\sqrt{2\alpha}}\right)\right] \qquad (5) \end{aligned}$$

Prawdopodobieństwo, że neurony we *wszystkich* N węzłach są *stabilne*

$$P^\star = (1 - P(x_i^n < -x_i^s))^N \qquad (6)$$

Żądamy, żeby zapamiętany wzorzec I^1 po odwróceniu stanów R neuronów pozostał stabilny z prawdopodobieństwem co najmniej 0.5 $\Rightarrow P^\star = 0.5$. Przy $N \to \infty$ musi być wtedy $P(x_i^n < -x_i^s) \ll 1$, więc argument funkcji błędu $\mathrm{erf}(\cdot)$ jest rzędu 1 i można przybliżyć

$$\mathrm{erf}(x) \approx 1 - \frac{1}{x\sqrt{\pi}} \exp(-x^2)$$

Z równań (5) i (6) wynika wówczas

$$P^\star = \frac{1}{2} = \left\{ \frac{1}{2} \left[1 + \mathrm{erf}\left(\frac{x_i^s}{\sqrt{2\alpha}} \right) \right] \right\}^N \approx \left\{ 1 - \frac{1}{x_i^s} \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{(x_i^s)^2}{2\alpha} \right) \right\}^N$$

$x \ll 1 \Rightarrow \ln(1-x) \approx -x \Rightarrow$

$$-\ln 2 \approx N \ln \left\{ 1 - \frac{1}{x_i^s} \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{(x_i^s)^2}{2\alpha} \right) \right\} \approx -\frac{N}{x_i^s} \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{(x_i^s)^2}{2\alpha} \right)$$

$$\sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{(x_i^s)^2}{2\alpha} \right) = \frac{x_i^s}{N} \ln 2 \Rightarrow -\left[\frac{(x_i^s)^2}{2\alpha} + \ln \frac{x_i^s}{\sqrt{2\alpha}} \right] = \ln(2\sqrt{\pi} \ln 2) - \ln N$$

N jest duże, spodziewamy się, że $\alpha \ll 1$, natomiast $x_i^s = 1 - 2R/N$ jest rzędu 1. Wobec tego po lewej stronie można zaniedbać logarytm, a po prawej stałą, jako wyrazy małe. Uzyskujemy w ten sposób

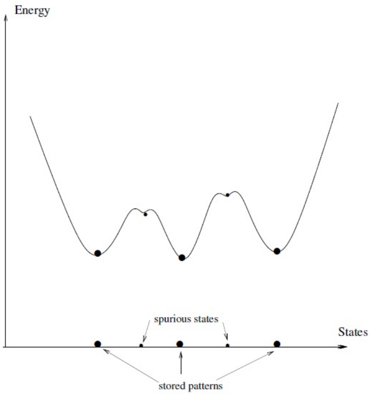
$$\frac{(x_i^s)^2}{2\alpha} = \frac{(1-2R/N)^2}{2\alpha} \approx \ln N \Rightarrow \alpha \approx \frac{(1-2R/N)^2}{2 \ln N}$$

Maksymalna pojemność sieci odpowiada sytuacji, gdy zapamiętany wzorzec zostanie zdestabilizowany przy dowolnie małym odchyleniu stanu sieci, czyli przy $R \to 0$

Ze względu na fundamentalną degenerację sieci, maksymalna liczba niezależnych zapamiętanych wzorców, które zachowują stabilność, wynosi

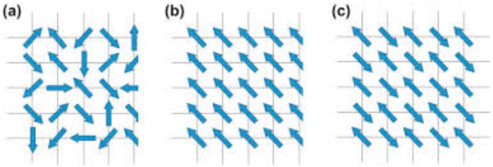
$$\alpha_c = \frac{P_c}{N} = \frac{1}{4 \ln N} \Rightarrow P_c = \frac{N}{4 \ln N} \qquad (7)$$

- Niestabilność wzorców powodowana jest zbyt dużym przesłuchem, czyli wpływem zbyt dużej liczby innych zapamietywanych wzorców.
- W procesie zapamiętywania pojawiają się spontanicznie **wzorce pasożytnicze**, odpowiadające dodatkowym minimom krajobrazu energetycznego i zwykle silnie skorelowane ze wzorcami zapamiętanymi



Rysunek: Krajobraz energetyczny sieci Hopfielda: głębokie minima odpowiadają wzorcom zapamiętanym, płytkie - pasożytniczym.

Analogie z układami magnetycznymi (1)

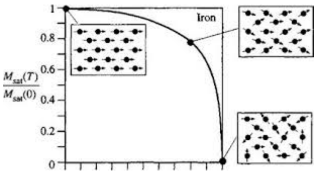


Rysunek: Sieć dwuwymiarowa: (a) faza paramagnetyczna, (b) faza ferromagnetyczna, (c) faza antyferromagnetyczna

Model Isinga: spiny dwustanowe $\sigma_i \in \{-1, 1\}$, Hamiltonian

$$H = -\sum_{\langle i,j \rangle} J_{ij}\sigma_i\sigma_j, \;\; i,j = 1,2,\ldots N.$$

- $J_{ij} = J > 0$ - dla $T < T_c$ faza ferromagnetyczna ($|M| > 0$),
- $J_{ij} = J < 0$ - na sieciach regularnych dla $T < T_c$ faza antyferromagnetyczna ($M = 0$, ale magnetyzacje podsieci przeciwne i niezerowe),
- $T > T_c$ faza paramagnetyczna ($M = 0$)



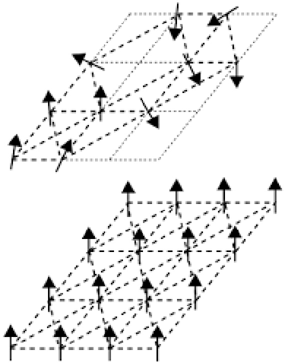
Rysunek: Magnetyzacja ferromagnetyka w funkcji temperatury

W stanie ferromagnetycznym uporządkowanie dalekiego zasięgu

Szkło spinowe

Model Edwardsa – Andersona

- Na sieci regularnej,
- $J_{ij} = \pm 1$,
  $P(J_{ij}) = r\delta(J_{ij} - 1) + (1 - r)\delta(J_{ij} + 1)$,
- Krajobraz energetyczny charakteryzuje się wieloma lokalnymi minimami,
- Dla $T < T_c$ i $r < r_c$ może pojawić się stan szkła spinowego z uporządkowaniem bliskiego zasięgu; odpowiada on osiągnieciu przez układ lokalnego (głebokiego) minimum energetycznego.

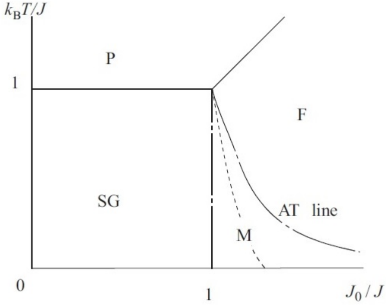


Rysunek: Porównanie typowego układu momantów magnetycznych na sieci dwuwymiarowej w fazie szkła spinowego (góra) i ferromagnetycznej (dół).

Model Sherringtona-Kirkpatricka

Model na sieci w pełni połączonej,

$$
\begin{aligned}
P\left(J_{ij}\right) &= N\left(\frac{J_0}{N}, \frac{J_1}{\sqrt{N}}\right) \\
J_{ij} &= \frac{J_0}{N} + \Delta J_{ij}, \quad i,j = 1,2\ldots N, \\
P\left(\Delta J_{ij}\right) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\Delta J_{ij})^2}{2\sigma^2}\right), \\
\sigma^2 &= \overline{\left(\Delta J_{ij}\right)^2} = \frac{J_1^2}{N},
\end{aligned}
$$



Istnieją fazy

- P: Paramagnetyczna,
- F: Ferromagnetyczna,
- SG: Szkła spinowego,
- M: Mieszana (F+SG).

Wniosek: warto rozważać również sieci neuronowe z szumem (termicznym → synaptycznym).