2. Memoria asociativa: Modelo de Hopfield

2.1. Introducción

- El primer modelo de neuronas fue propuesto por Pitts and McCulloch en 1943. Segun ellos una "neurona formal" es una puerta lógica con dos posibles estados internos, activo o apagado.

  $s_i = 1$ (activo) $s_i = 0$ (apagado)

  La neurona recibe unas pocas entradas que son las salidas de otras neuronas. Estas entradas se suman, $h_i = \sum_j J_{ij} s_j$, y el estado de la neurona se determina por comparación con cierto umbral θ_i.

$$s_i = \begin{cases} 1 & h_i > \theta_i \\ 0 & h_i < \theta_i \end{cases} \tag{48}$$

- Se puede demostrar que todas las operaciones lógicas (NOT, AND, OR, XOR) se pueden conseguir con este tipo de neuronas formales, elegiendo los pesos y umbrales convenientemente. (**Ejercicio**)

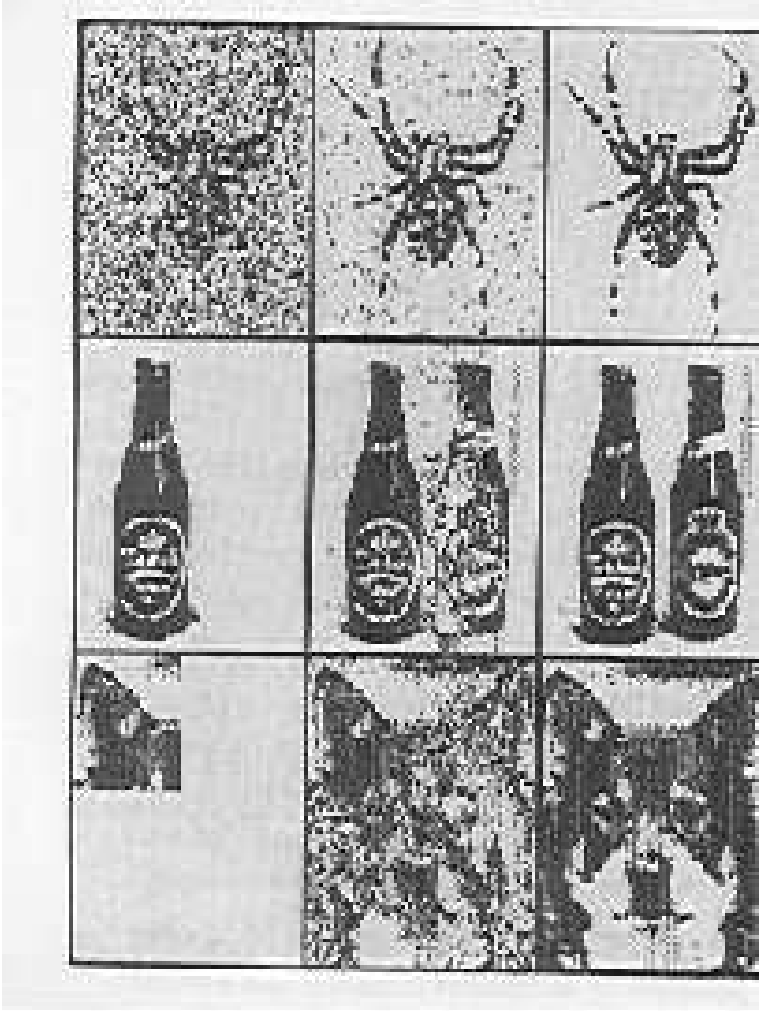
- El modelo de Hopfield es un modelo mecánico-estadístico de red neuronal basado en el modelo de McCulloch-Pitts, que tiene la propiedad de memoria asociativa. Con este término se denota la capacidad que tiene la red de neuronas de recuperar la información correspondiente a un determinado número de patrones, que previamente ha almacenado, a partir del conocimiento parcial o imperfecto de dicha información, sin conocer la localización precisa donde ha sido almacenada.

- El modelo de Hopfield está constituido por una red Λ_d d-dimensional, en cuyos nudos hay N neuronas representadas por variables de estado $s_{\mathbf{x}} = \pm 1$. Se asume que durante un proceso previo de aprendizaje la red es capaz de aprender un conjunto de P patrones $\boldsymbol{\xi}^\mu \equiv \{\xi^\mu_{\mathbf{x}} = \pm 1, \mathbf{x} \in \Lambda_d\}$, $\nu = 1, \ldots, P$. En lo siguiente consideraremos que los patrones son elegidos de forma aleatoria e independientes entre sí tomando sus elementos, es decir $\xi^\mu_{\mathbf{x}}$, el valor $+1$ y -1 con igual probabilidad. Entonces, la energía configuracional del sistema viene descrita por el hamiltoniano

$$\mathcal{H}_{\mathbf{J}}(\mathbf{s}) = -\frac{1}{2} \sum_{\substack{\mathbf{x},\mathbf{y} \\ \mathbf{x} \neq \mathbf{y}}} J_{xy} s_{\mathbf{x}} s_{\mathbf{y}} + \sum_{\mathbf{x}} \theta_{\mathbf{x}} s_{\mathbf{x}}. \tag{49}$$

  con

$$J_{xy} = \frac{1}{N} \sum_{\mu=1}^{P} \xi^\mu_{\mathbf{x}} \xi^\mu_{\mathbf{y}} \tag{50}$$



que se conoce como regla de Hebb. Aquí $\theta_{\mathbf{x}} \equiv -\sum_{\mu=1}^{P} h^{\mu}\xi_{\mathbf{x}}^{\mu}$ representa el umbral de excitación asociado a cada neurona y $\mathbf{h} \equiv (h^1, ..., h^P)$ es un vector constante de P componentes.

El estudio del modelo de Hopfield depende sustancialmente del comportamiento de la relación $\alpha \equiv P/N$ en el límite $N \to \infty$ como veremos a continuación.

2.2. Función de partición en el límite $\alpha \to 0$

- Dado que las variables J_{xy} son fijas (desorden *quenched*), el problema de evaluar la función de partición se reduce a calcular

$$Z = \sum_{\mathbf{s}} e^{-\beta \mathcal{H}_{\mathbf{J}}(\mathbf{s})} = e^{-\frac{1}{2}\beta P} \sum_{\mathbf{s}} \exp\left[\frac{\beta}{2N} \sum_{\mu=1}^{P} \left\{ 2Nh^{\mu} \sum_{\mathbf{x}} \xi_{\mathbf{x}}^{\mu} s_{\mathbf{x}} + \left(\sum_{\mathbf{x}} \xi_{\mathbf{x}}^{\mu} s_{\mathbf{x}} \right)^2 \right\}\right], \tag{51}$$

donde $\sum_{\mathbf{s}} \equiv \sum_{s_{\mathbf{x}_1}=\pm 1} \dots \sum_{s_{\mathbf{x}_N}=\pm 1}$. Utilizando la integración gaussiana $\int_{-\infty}^{+\infty} dz \exp\left(-az^2 + bz\right) = \left(\frac{\pi}{a}\right)^{1/2} \exp\left(\frac{b^2}{4a}\right)$, con $a = \frac{\beta N}{2}$ y $b = \beta \sum_{\mathbf{x}} \xi_{\mathbf{x}}^{\mu} s_{\mathbf{x}}$, se tiene

$$Z = e^{-\frac{1}{2}\beta P} \left(\frac{\beta N}{2\pi}\right)^{P/2} \int d\mathbf{m} e^{-\frac{1}{2}\beta N \mathbf{m}^2} \prod_{\mathbf{x}} \left[\sum_{s_{\mathbf{x}}=\pm 1} \exp\left\{\beta(\mathbf{m} + \mathbf{h}) \cdot \boldsymbol{\xi}_{\mathbf{x}} s_{\mathbf{x}}\right\} \right] \tag{52}$$

donde $\mathbf{m} \equiv (m^1, \dots, m^P)$, con m^{μ} $\mu = 1, \dots P$ variables mudas de integración y $\mathbf{m}^2 \equiv \sum_{\mu=1}^{P} (m^{\mu})^2$. La sumatoria, dentro del productorio, puede evaluarse de forma sencilla obteniendo

$$Z = \left(\frac{\beta N}{2\pi}\right)^{P/2} \int d\mathbf{m} \exp\left\{-\beta N f(\mathbf{m})\right\} \tag{53}$$

con $f(\mathbf{m}) = \frac{\mathbf{m}^2}{2} - \frac{1}{\beta N} \sum_{\mathbf{x}} \ln\left[2\cosh\left\{\beta(\mathbf{m} + \mathbf{h}) \cdot \boldsymbol{\xi}_{\mathbf{x}}\right\}\right] + \frac{P}{2N}$.

- En el límite $\alpha \to 0$, con P finito y $N \to \infty$, la integral dada por (53) puede ser evaluada mediante el método del punto de silla, que consiste en desarrollar la función $f(\mathbf{m})$ en serie de Taylor alrededor del mínimo y, dado que dicha función aparece multiplicada por N en el argumento de una exponencial, sólo van a contribuir al valor de la integral de forma importante los términos dominantes del desarrollo. Así se obtiene para la energía libre del sistema la expresión

$$\frac{F}{N} = -\frac{1}{N\beta} \ln Z = f(\mathbf{m}_0) + \mathcal{O}\left(\frac{\ln N}{N}\right), \tag{54}$$

donde $\mathbf{m}_0$ es solución a la ecuación

$$\left.\frac{\partial f(\mathbf{m})}{\partial \mathbf{m}}\right|_{\mathbf{m}=\mathbf{m}_0} = 0, \tag{55}$$

que da

$$\mathbf{m}_0 = \frac{1}{N} \sum_{\mathbf{x}} \boldsymbol{\xi}_{\mathbf{x}} \tanh \left\{ \beta (\mathbf{m}_0 + \mathbf{h}) \cdot \boldsymbol{\xi}_{\mathbf{x}} \right\}, \tag{56}$$

que es la solución de campo medio del modelo para $\alpha \to 0$. El resultado dado por (54) permite definir la energía libre del sistema como $F = Nf(\mathbf{m})$. Por otra parte se tiene que

$$\sum_{\mathbf{x}} \boldsymbol{\xi}_{\mathbf{x}} \langle s_{\mathbf{x}} \rangle = \lim_{\mathbf{h} \to 0} \frac{1}{\beta} \frac{\partial}{\partial \mathbf{h}} \ln Z, \tag{57}$$

que, utilizando (54) y (56), implica $\mathbf{m}_0 = N^{-1} \sum_{\mathbf{x}} \boldsymbol{\xi}_{\mathbf{x}} \langle s_{\mathbf{x}} \rangle$, ecuación que constituye la solución estacionaria de campo medio del modelo. Se puede demostrar que la sumatoria sobre todas las neuronas que aparece en (56) se puede sustituir por un promedio sobre la distribución de patrones $\boldsymbol{\xi}^{\mu}$, quedando

$$\mathbf{m}_0 = \left\langle \boldsymbol{\xi}_{\mathbf{x}} \langle s_{\mathbf{x}} \rangle \right\rangle_{\xi} \tag{58}$$

- Se llama modelo de Mattis cuando se tiene:

$$P(\boldsymbol{\xi}) = \prod_{\mu, \mathbf{x}} p(\xi_{\mathbf{x}}^{\nu}) \tag{59}$$

con

$$p(\xi_{\mathbf{x}}^{\nu}) = \frac{1}{2} \delta(\xi_{\mathbf{x}}^{\nu} - 1) + \frac{1}{2} \delta(\xi_{\mathbf{x}}^{\nu} + 1) \tag{60}$$

Los estados estacionarios interesantes desde el punto de vista de la propiedad de memoria asociativa son aquellos tales que $\mathbf{m} = (m^1, 0, \ldots, 0)$ (estados de Mattis o estados puros). Hay $2P$ de estos estados y son solucion a la ecuación:

$$m^1 = \tanh(\beta m^1) \tag{61}$$

Además podemos encontrar estados mezcla de la forma $\mathbf{m} = (m^1, \ldots, m^p)$, con $m^{\mu} \neq 0$.

Ejercicio: Demostrar que:

$$\sum_{\mathbf{x}} \boldsymbol{\xi}_{\mathbf{x}} \langle s_{\mathbf{x}} \rangle = \lim_{\mathbf{h} \to 0} \frac{1}{\beta} \frac{\partial}{\partial \mathbf{h}} \ln Z. \tag{62}$$

Ejercicio: Calcular el punto crítico en el modelo de Mattis para la aparición de estados de mattis.

Ejercicio: Demostrar que para $T \to 0$ se tiene $E = -\frac{1}{2}\mathbf{m}^2$ y $\mathbf{m} = \langle\langle \boldsymbol{\xi}\, \mathrm{sgn}(\boldsymbol{\xi} \cdot \mathbf{m})\rangle\rangle$

Ejercicio Estudiar la estabilidad local de las soluciones *punto de silla* para el modelo de Mattis. Para ello introducir la matriz

$$Q^{\mu\nu} \equiv \langle \xi^\mu \xi^\nu \tanh^2(\beta \mathbf{m} \cdot \xi)\rangle_\xi$$

Ver que en el caso particular de estados mezcla simétricos, $\mathbf{m} = m_n(1, 1, \ldots, 0)$, solo hay 3 tipos de autovalores que cerca del punto crítico tienen el comportamiento:

$$\begin{array}{l} \lambda_1 \approx 2t \\ \lambda_2 \approx \frac{2t}{3n-2} \\ \lambda_3 \approx \frac{-4t}{3n-2} \end{array} \tag{63}$$

con $t = T_c - T$.

Estudiar la estabilidad de dichas soluciones para $T \to 0$. Calcular el valor de temperature T_n por debajo del cual las soluciones con n impar son localmente estables.

- Pese a construirse similarmente al modelo de EA, el modelo de mattis no presenta comportamiento vidrio de espin pues la elección de la regla de Hebb con un número finito de patrones impide cualquier tipo de frustración en el sistema
- Veremos que el tomar un número infinito de patrones induce comportamiento vidrio de espín a temperaturas bajas por la aparición de una infinitud de mínimos locales separados por barreras infinitas

2.3. Red de Hopfield para α finito

- En este caso $P = \alpha N$ con $N \to \infty$ y $\alpha \neq 0$, por lo que el método seguido en la sección anterior no es aplicable: el método del punto de silla no puede llevarse a cabo pues $f(\mathbf{m})$ explota en el límite termodinámico; cf. (53). Si evaluamos la energía por neurona a partir de (49), entonces se debe hacer la hipótesis de que sólo un determinado número k (finito) de patrones tiene solapamientos $\mathcal{O}(1)$ en $N \to \infty$, mientras que el resto de los solapamientos han de ser $\mathcal{O}(1/\sqrt{N})$ para que dicha energía no diverja; estos k patrones van a ser lo que van a contribuir al umbral $\theta_{\mathbf{x}}$ en la forma dada por (49). Ahora hay que poner más cuidado a la hora de promediar sobre la distribución de patrones. En este caso la energía libre del sistema, dado que las sinapsis son fijas, se obtiene en la forma

$$\frac{F}{N} = -\frac{1}{N\beta} \langle \ln Z \rangle_{\xi} \,. \tag{64}$$

- El obtener ahora la energía libre tiene la dificultad matemática de evaluar el promedio del logaritmo. Para llevarlo a cabo se utiliza la técnica de réplicas que consiste en utilizar la relación:

$$\ln Z = \lim_{n \to 0} \frac{Z^n - 1}{n}, \tag{65}$$

por lo que el problema se reduce a evaluar $\langle Z^n \rangle_{\xi}$. Como antes ($P$ finito), tras introducir las integraciones gaussianas, obtenemos

$$\langle Z^n \rangle_{\xi} = e^{-\frac{1}{2}\beta n P} \left(\frac{\beta N}{2\pi} \right)^{nP/2} \left\langle \sum_{\mathbf{s}^1 \ldots \mathbf{s}^n} \int \left(\prod_{\rho=1}^{n} \prod_{\mu=1}^{P} dm_{\rho}^{\mu} \right) \; X(\mathbf{m}, \vec{s}, \boldsymbol{\xi}) \right\rangle_{\xi}, \tag{66}$$

donde

$$\begin{aligned} X(\mathbf{m}, \vec{s}, \boldsymbol{\xi}) = \; & \exp \left\{ -\tfrac{1}{2}\beta N \sum_{\rho=1}^{n} \sum_{\mu=1}^{P} (m_{\rho}^{\mu})^2 \right\} \\ & \times \exp \left\{ \beta \sum_{\rho=1}^{n} \sum_{\mathbf{x}} \left[\sum_{\mu > k} m_{\rho}^{\mu} \xi_{\mathbf{x}}^{\mu} s_{\mathbf{x}}^{\rho} + \sum_{\nu=1}^{k} (m_{\rho}^{\nu} + h^{\nu}) \xi_{\mathbf{x}}^{\nu} s_{\mathbf{x}}^{\rho} \right] \right\}, \end{aligned} \tag{67}$$

los índices ρ y σ denotan n réplicas del sistema y $\vec{s} \equiv (s^1, \ldots, s^n)$. Introduciendo la representación integral $1 = \int_{\Re} dq_{ab} \delta(q_{ab} - N^{-1} \sum_{\mathbf{x}} s_{\mathbf{x}}^a s_{\mathbf{x}}^b)$, y la representación integral de la delta de Dirac, se pueden evaluar las

integrales sobre las variables $m_\rho^\mu,\ \rho = 1,\ldots,n$ y $\mu > k$, ya que para estas variables la integral factoriza en forma de integrales gaussianas que no llevan ninguna dependencia en las variables neuronales. Así (66) queda de la forma

$$\begin{aligned}\langle Z^n\rangle_\xi &= e^{-\beta P n/2}\int\left(\prod_{\rho=1}^{n}\prod_{\nu=1}^{k} dm_\rho^k\right)\int\left(\prod_{\rho\neq\sigma} dq_{\rho\sigma}dr_{\rho\sigma}\right)e^{-\beta N\Phi}\\ \Phi = &\ \frac{1}{2}\sum_{\rho=1}^{n}\sum_{\nu=1}^{k}(m_\rho^\nu)^2 + \frac{\alpha}{2\beta}\mathrm{Tr}\ln\left\{(1-\beta)\,\mathbf{I}-\beta\mathbf{q}\right\} + \frac{\alpha\beta}{2}\sum_{\substack{\rho,\sigma\\ \rho\neq\sigma}} r_{\rho\sigma}q_{\rho\sigma}\ ,\\ &-\tfrac{1}{\beta}\left\langle \ln \mathrm{Tr}_{\vec{s}} e^{\beta H_\xi(\vec{s})}\right\rangle_\xi\end{aligned} \tag{68}$$

con

$$H_\xi(\vec{s}) \equiv \sum_{\rho=1}^{n}\left[\frac{\alpha\beta}{2}\sum_{\substack{\sigma=1\\ \sigma\neq\rho}}^{n} r_{\rho\sigma}s^\rho s^\sigma + \sum_{\nu=1}^{k}(m_\rho^\nu + h^\nu)\xi^\nu s^\rho\right]. \tag{69}$$

En la expresión de Φ que aparece en (68), el promedio $\langle ...\rangle_\xi$ denota promedio sobre la distribución de los $\nu = 1,\ldots,k$ patrones condensados. $\mathbf{I}$ es la matriz identidad con elementos $\delta_{\rho\sigma}$ y $\mathbf{q}$ es la matriz con elementos $q_{\rho\sigma}$ si $\rho \neq \sigma$ y ceros para $\rho = \sigma$. La energía libre por espín es entonces

$$\begin{aligned}\tfrac{F}{N} = &\ \tfrac{\alpha}{2} + \tfrac{\alpha}{2\beta n}\mathrm{Tr}\ln\left[(1-\beta)\,\mathbf{I}-\beta\mathbf{q}\right] +\\ &+\tfrac{1}{2n}\sum_{\rho=1}^{n}\left[\sum_{\nu=1}^{k}\left(m_\rho^\nu\right)^2 + \alpha\beta\sum_{\substack{\sigma=1\\ \sigma\neq\rho}}^{n} r_{\rho\sigma}q_{\rho\sigma}\right] - \tfrac{1}{n\beta}\left\langle \ln \mathrm{Tr}_{\vec{s}} e^{\beta H_\xi(\vec{s})}\right\rangle_\xi .\end{aligned} \tag{70}$$

- Como en el caso de P finito, la energía interna puede ser ahora evaluada en $N \to \infty$ utilizando la aproximación del punto de silla, que este caso está definido por las condiciones

$$\frac{\partial\Phi}{\partial m_\rho^\nu} = 0, \quad \frac{\partial\Phi}{\partial r_{\rho\sigma}} = 0, \quad \frac{\partial\Phi}{\partial q_{\rho\sigma}} = 0, \tag{71}$$

ecuaciones que dan significado físico a los parámetros de orden m_ρ^ν, $q_{\rho\sigma}$ y $r_{\rho\sigma}$, esto es

$$m_\rho^\nu = \frac{1}{N}\left\langle \sum_{\mathbf{y}} \xi_{\mathbf{y}}^\nu \left\langle s_{\mathbf{y}}^\rho\right\rangle\right\rangle_\xi \tag{72}$$

como una medida del solapamiento del estado del sistema en la réplica ρ con un patrón almacenado dado, y

$$q_{\rho\sigma} = \frac{1}{N}\left\langle \sum_{\mathbf{y}} \langle s_{\mathbf{y}}^{\rho}\rangle \langle s_{\mathbf{y}}^{\sigma}\rangle \right\rangle_{\xi}, \quad r_{\rho\sigma} = \frac{1}{\alpha}\sum_{\mu=k+1}^{P} \langle m_{\rho}^{\mu} m_{\sigma}^{\mu}\rangle_{\xi}, \tag{73}$$

que en el caso de simetría de réplicas ($q_{\rho\sigma} = q \;\; \rho \neq \sigma$ y $q_{\rho\rho} = 0$) reducen a los conocidos parámetros de orden de EA y AGS, respectivamente.