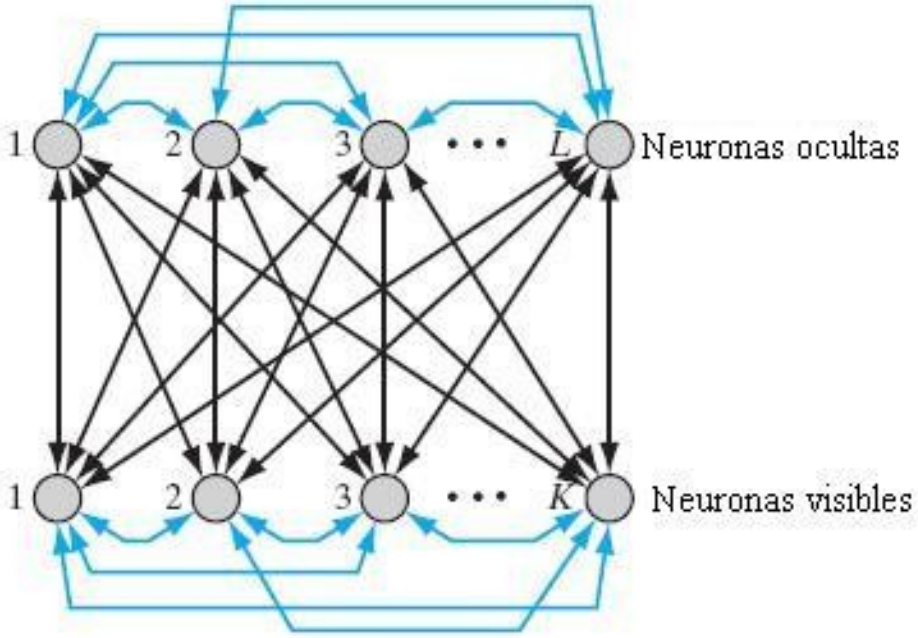
# MAQUINAS DE BOLTZMANN
# MAQUINAS DE BOLTZMANN RESTRINGIDAS

# MAQUINAS DE BOLTZMANN

Red de unidades binarias estocásticas conectadas simétricamente.




Máquina de Boltzmann
Conexiones simétricas inter e intra-capa

Patrón de interacciones entre unidades visibles (v) y ocultas (h), a través de la Función de Energía:

$$E_{\mathrm{BM}}(v,h;\theta) = -\frac{1}{2}v^T U v - \frac{1}{2}h^T V h - v^T W h - b^T v - d^T h$$

U, V con diagonal en 0's

Distribución de probabilidades sobre el espacio conjunto (v,h), vía la distribución de Boltzmann:

$$P(v,h) = \frac{1}{Z(\theta)} \exp\left(-E_{\mathrm{BM}}(v,h;\theta)\right)$$

con

$$Z(\theta) = \sum_{v_1=0}^{v_1=1} \cdots \sum_{v_D=0}^{v_D=1} \sum_{h_1=0}^{h_1=1} \cdots \sum_{h_N=0}^{h_N=1} \exp\left(-E_{\mathrm{BM}}(v,h;\theta)\right)$$

*función de partición* (normaliza la densidad).

La distribución de probabilidad conjunta da lugar a las probabilidades condicionales:

$$P(h_i \mid v, h_{\setminus i}) = \mathrm{sigmoid}\left(\sum_j W_{ji} v_j + \sum_{i' \setminus i} V_{ii'} h_{i'} + d_i\right)$$

$$P(v_j \mid h, v_{\setminus j}) = \mathrm{sigmoid}\left(\sum_i W_{ji} v_j + \sum_{j' \setminus j} U_{jj'} v_{j'} + b_j\right)$$

# INTRATABILIDAD DE LA MAQUINA DE BOLTZMANN

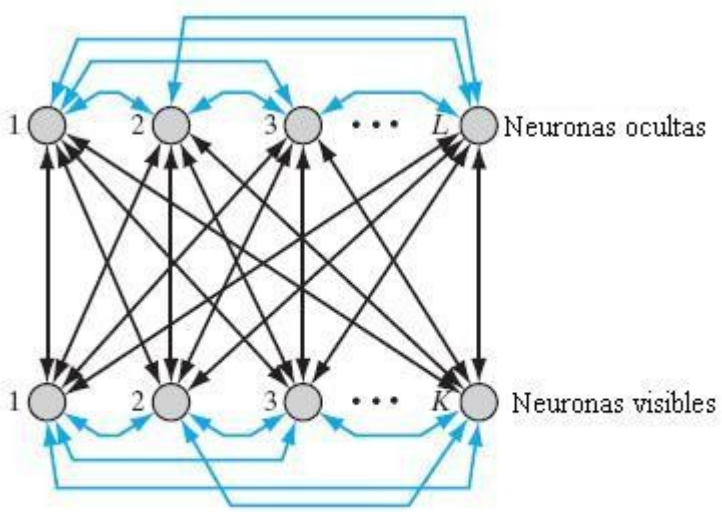
La inferencia en una MB es muy difícil, si no imposible de realizar.

Ejemplo: la probabilidad condicional de una $h_i$ dadas las visibles, $P(h_i/v)$, impone calcular la probabilidad marginal sobre el resto de las ocultas, evaluando una suma con $2^{N-1}$términos

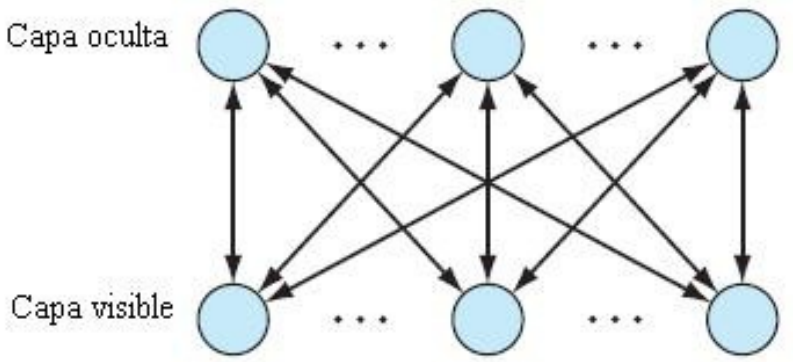
$$P(h_i \mid v) = \sum_{h_1=0}^{h_1=1} \cdots \sum_{h_{i-1}=0}^{h_{i-1}=1} \sum_{h_{i+1}=0}^{h_{i+1}=1} \cdots \sum_{h_N=0}^{h_N=1} P(h \mid v)$$

→ *surge la idea de restringir el patrón de interacciones entre visibles y ocultas.*

# MAQUINAS DE BOLTZMANN RESTRINGIDAS




Máquina de Boltzmann
Conexiones simétricas inter e intra-capa




Máquina de Boltzmann Restringida (RBM)
(sin conexiones intra-capa)

# MAQUINAS DE BOLTZMANN RESTRINGIDAS

Las más populares de la “familia”.

*U = 0, V=0 en la función de energía → grafo bipartito (dos capas de vértices: una de ocultas, otra de visibles)*

→ propiedad útil: *las distribuciones condicionales factorizan:*

$$P(h \mid v) = \prod_i P(h_i \mid v) = \prod_i \text{sigmoid}\left(\sum_j W_{ji} v_j + d_i\right)$$

$$P(v \mid h) = \prod_j P(v_j \mid h) = \prod_j \text{sigmoid}\left(\sum_i W_{ji} h_i + b_j\right)$$

Ahora las distribuciones marginales a posteriori de la forma $P(h_i/v)$ serán tratables.
*No así la función de partición.*

# APRENDIZAJE

Objetivo: maximizar el producto de probabilidades de un conjunto V de entrenamiento:

$$\underset{w}{\arg\max} \prod_{v \in V} P(v)$$

O, incrementalmente, maximizar la probabilidad logarítmica esperada para un v elegido al azar de V.

$$\underset{w}{\arg\max} \sum_{v \in V} [\log P(v)]$$

Más en detalle: encontrar un conjunto de parámetros para el modelo que maximicen aproximadamente la probabilidad logarítmica (log likelihood) de un conjunto de entrenamiento:

$$\sum_t log P(v_{:,t};\theta) = \sum_t log \sum_{h_{1,t}=0}^{h_{1,t}=1} \cdots \sum_{h_{N,t}=0}^{h_{N,t}=1} P(v_{:,t}, h_{:,t};\theta)$$

*ascenso por gradiente:*

$$\frac{\partial}{\partial \theta_i}\left(\sum_{t=1}^{T} \log p(v_{:,t})\right)$$

entrenamiento que se realiza iterativamente con el método de *máxima verosimilitud estocástica* (SML) o *divergencia contrastiva persistente* (PCD).

# MÉTODO DE DIVERGENCIA CONTRASTIVA

(*contrastive divergence*, Hinton 1999, 2002)

Versión incremental (paso a paso, CD-1):

1- Tomar un patrón de entrada, calcular las probabilidades de las unidades ocultas y con esa distribución de probabilidades generar una muestra de *h*.

2- A partir del *h* generado (*sampleado*), muestrear a su vez una reconstrucción v' de las unidades visibles (reconstrucción de la señal).

3- Muestrear un nuevo *h'* a partir de la distribución inducida por *v'*.

4- $\Delta W = \eta(vh - v'h')$ → aprendizaje hebbiano

$\Delta a = \eta(v-v')$ $\Delta b=\eta(h-h')$

5- Reducir $\eta$ e ir a 1

# Después de iterar suficiente tiempo (pasos 1 a 3, *muestreo de Gibbs*)

→ *Equilibrio térmico*: pueden cambiar los estados de la entrada y de las unidades ocultas (detectores de características), pero la probabilidad de encontrar al sistema en cualquier configuración particular es estable, máxima para los vectores del conjunto de entrenamiento (los patrones "en los que la red cree", Hinton, 2007)

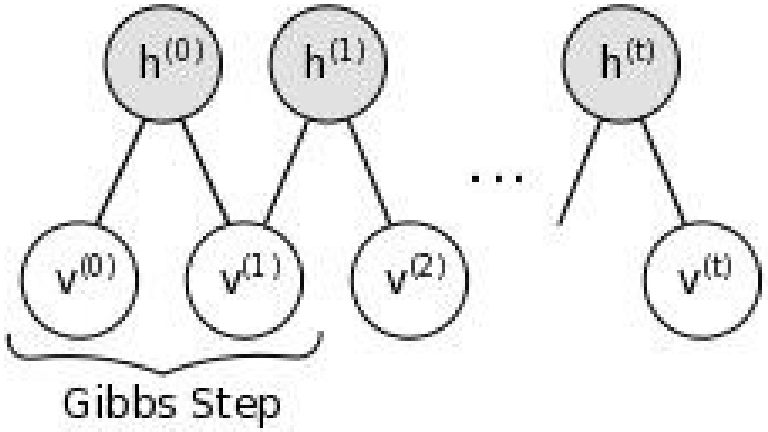
→ *Efecto promedio*: $< v_i h_j >_{datos} - < v_i h_j >_{modelo} = \partial \log P(v) / \partial w_{ij}$
→ *equivale a ascenso por gradiente según la ecuación*

$$\Delta w_{ij}(t+1) = w_{ij}(t) + \eta \, \partial \log P(v) / \partial w_{ij}$$

*Observación 1:* el muestreo de Gibbs (pasos 1 a 3) podría repetirse un número indeterminado k de veces, pero en la práctica, k=1 es suficiente.

*Observación 2*: CD aproxima el gradiente de la probabilidad logarítmica localmente alrededor del punto de entrenamiento (el training set visible).

*Observación 3:* en teoría muestreando alternadamente infinitas veces (computacionalmente prohibitivo), i.e.




con $t \rightarrow \infty$ , garantizaría la convergencia de la sucesión a una muestra exacta de *p(v,h).*

*Observación 4*:
*Neuronas ocultas: activación binaria* (detectores de características)
*Neuronas visibles: pueden ser multinomiales* → función *softmax*:
(elementos del vector de entrada, e.g. pixels de una imagen)

# MAQUINAS DE BOLTZMANN: PLAUSIBILIDAD BIOLOGICA

Aprendizaje local: la actualización de un peso entre dos neuronas depende sólo de las distribuciones de esas dos neuronas:
→ *regla de aprendizaje local*

**En términos biológicos:** regla de Hebb, las conexiones (axones y dendritas) se adaptarían sólo en base a los patrones de activación de las células que involucran (pre- y postsinápticas).

*Fire together, wire together.*

Comparar con, e.g., perceptrón multicapa con backpropagation: requiere más información que las estadísticas locales.
**En términos biológicos:** el cerebro requeriría una red secundaria de comunicación para (retro)transmitir a la red información sobre el gradiente.

Ventajas sobre modelos forward:

- Dado un vector visible, la distribución sobre vectores ocultos factoriza como un producto de distribuciones independientes para cada unidad oculta
→ pasos 2 y 3 se realizan en paralelo.

- Se pueden aprender redes más profundas concatenando *(stacking*) RMBs : activaciones *h* de un nivel son entrada para el siguiente
→ extracción jerárquica de características de nivel creciente.

# MAQUINAS DE BOLTZMANN RESTRINGIDAS: APLICACIONES

*- Reducción de dimensionalidad*
*- Clasificación*
*- Filtrado colaborativo*
*- Aprendizaje de características*
*- Modelado de tópicos (topic modeling, modelos estadísticos para predecir los tópicos o términos que ocurren en una colección de documentos)*
*- Para construir arquitecturas jerárquicas, encadenando varias RBM (deep belief networks).*