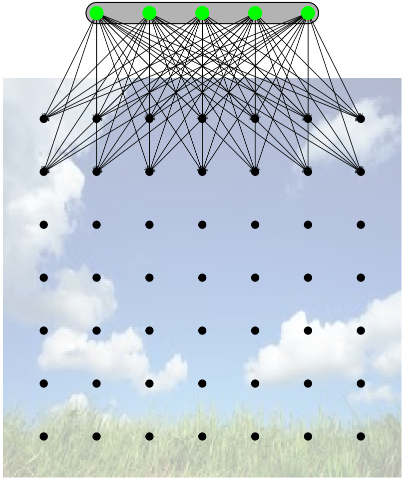
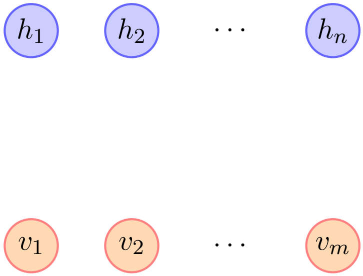
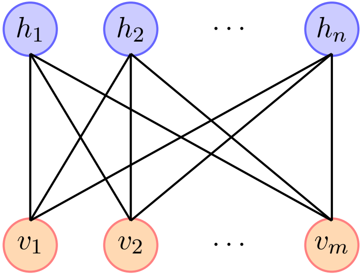
Module 19.3: Restricted Boltzmann Machines



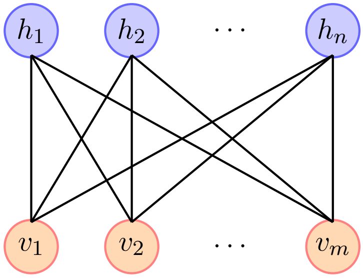
- We return back to our Markov Network containing hidden variables and visible variables



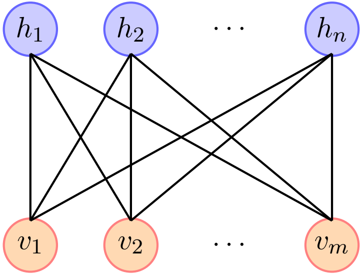
- We return back to our Markov Network containing hidden variables and visible variables
- We will get rid of the image and just keep the hidden and latent variables



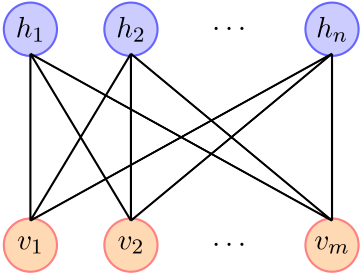
- We return back to our Markov Network containing hidden variables and visible variables
- We will get rid of the image and just keep the hidden and latent variables
- We have edges between each pair of (hidden, visible) variables.



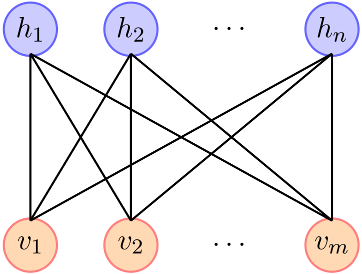
- We return back to our Markov Network containing hidden variables and visible variables
- We will get rid of the image and just keep the hidden and latent variables
- We have edges between each pair of (hidden, visible) variables.
- We do not have edges between (hidden, hidden) and (visible, visible) variables



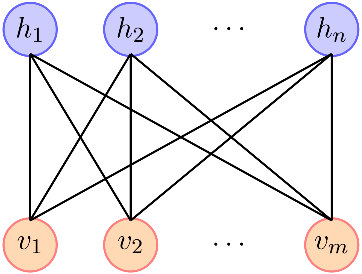
- Earlier, we saw that given such a Markov network the joint probability distribution can be written as a product of factors



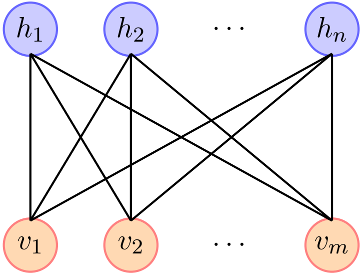
- Earlier, we saw that given such a Markov network the joint probability distribution can be written as a product of factors
- Can you tell how many factors are there in this case?



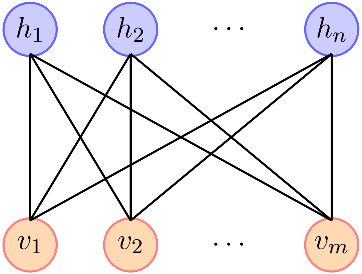
- Earlier, we saw that given such a Markov network the joint probability distribution can be written as a product of factors
- Can you tell how many factors are there in this case?
- Recall that factors correspond to maximal cliques



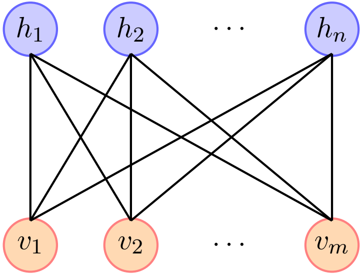
- Earlier, we saw that given such a Markov network the joint probability distribution can be written as a product of factors
- Can you tell how many factors are there in this case?
- Recall that factors correspond to maximal cliques
- What are the maximal cliques in this case?



- Earlier, we saw that given such a Markov network the joint probability distribution can be written as a product of factors
- Can you tell how many factors are there in this case?
- Recall that factors correspond to maximal cliques
- What are the maximal cliques in this case? every pair of visible and hidden node forms a clique

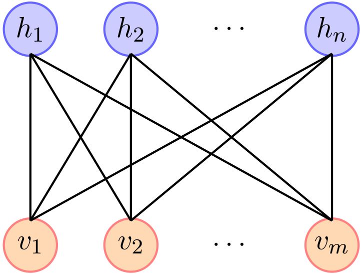


- Earlier, we saw that given such a Markov network the joint probability distribution can be written as a product of factors
- Can you tell how many factors are there in this case?
- Recall that factors correspond to maximal cliques
- What are the maximal cliques in this case? every pair of visible and hidden node forms a clique
- How many such cliques do we have? $(m \times n)$



- So we can write the joint pdf as a product of the following factors

$$P(V, H) = \frac{1}{Z} \prod_i \prod_j \phi_{ij}(v_i, h_j)$$

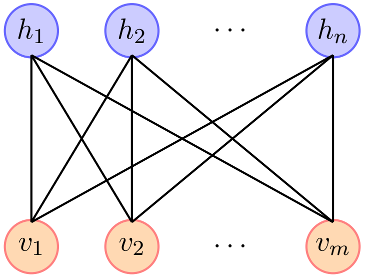


- So we can write the joint pdf as a product of the following factors

$$P(V,H) = \frac{1}{Z} \prod_i \prod_j \phi_{ij}(v_i, h_j)$$

- In fact, we can also add additional factors corresponding to the nodes and write

$$P(V,H) = \frac{1}{Z} \prod_i \prod_j \phi_{ij}(v_i, h_j) \prod_i \psi_i(v_i) \prod_j \xi_j(h_j)$$



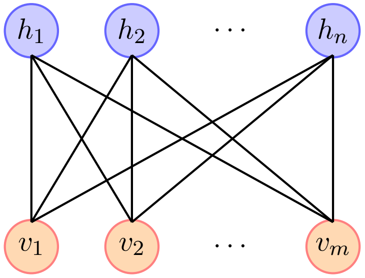
- So we can write the joint pdf as a product of the following factors

$$P(V, H) = \frac{1}{Z} \prod_i \prod_j \phi_{ij}(v_i, h_j)$$

- In fact, we can also add additional factors corresponding to the nodes and write

$$P(V, H) = \frac{1}{Z} \prod_i \prod_j \phi_{ij}(v_i, h_j) \prod_i \psi_i(v_i) \prod_j \xi_j(h_j)$$

- It is legal to do this (i.e., add factors for $\psi_i(v_i)\xi_j(h_j)$) as long as we ensure that Z is adjusted in a way that the resulting quantity is a probability distribution



- So we can write the joint pdf as a product of the following factors

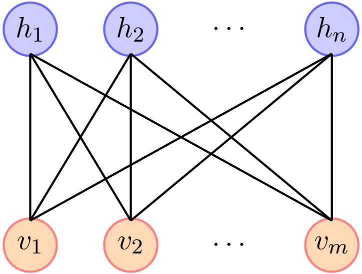
$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i, h_j)$$

- In fact, we can also add additional factors corresponding to the nodes and write

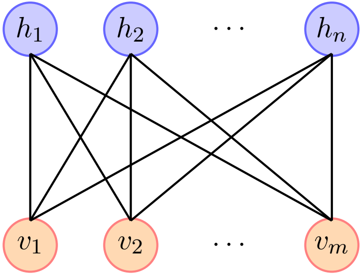
$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i, h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$

- It is legal to do this (i.e., add factors for $\psi_i(v_i)\xi_j(h_j)$) as long as we ensure that Z is adjusted in a way that the resulting quantity is a probability distribution
- Z is the partition function and is given by

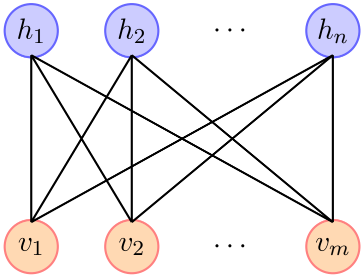
$$\sum_V \sum_H \prod_i \prod_j \phi_{ij}(v_i, h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$



- Let us understand each of these factors in more detail

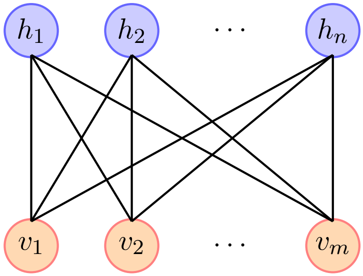


- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables



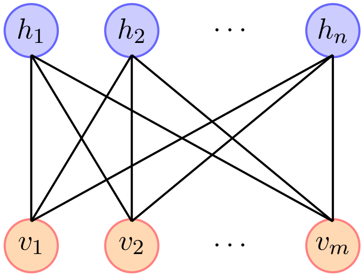
$\phi_{11}(v_1, h_1)$		
0	0	30
0	1	5
1	0	1
1	1	10

- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables
- The adjoining table shows one such possible instantiation of the ϕ_{11} function



$\phi_{11}(v_1, h_1)$		
0	0	30
0	1	5
1	0	1
1	1	10

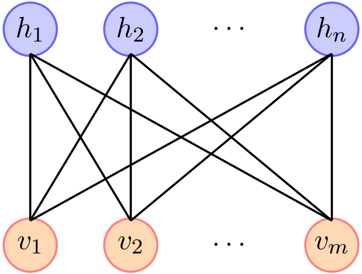
- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables
- The adjoining table shows one such possible instantiation of the ϕ_{11} function
- Similarly, $\psi_1(v_1)$ takes the value of $v_1 \in \{0, 1\}$ and gives us a number which roughly indicates the possibility of v_1 taking on the value 1 or 0



$\phi_{11}(v_1, h_1)$		
0	0	30
0	1	5
1	0	1
1	1	10

$\psi_1(v_1)$	
0	10
1	2

- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables
- The adjoining table shows one such possible instantiation of the ϕ_{11} function
- Similarly, $\psi_1(v_1)$ takes the value of $v_1 \in \{0, 1\}$ and gives us a number which roughly indicates the possibility of v_1 taking on the value 1 or 0
- The adjoining table shows one such possible instantiation of the ψ_{11} function

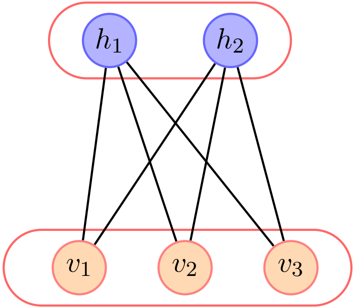


$\phi_{11}(v_1, h_1)$		
0	0	30
0	1	5
1	0	1
1	1	10

$\psi_1(v_1)$	
0	10
1	2

- Let us understand each of these factors in more detail
- For example, $\phi_{11}(v_1, h_1)$ is a factor which takes the values of $v_1 \in \{0, 1\}$ and $h_1 \in \{0, 1\}$ and returns a value indicating the affinity between these two variables
- The adjoining table shows one such possible instantiation of the ϕ_{11} function
- Similarly, $\psi_1(v_1)$ takes the value of $v_1 \in \{0, 1\}$ and gives us a number which roughly indicates the possibility of v_1 taking on the value 1 or 0
- The adjoining table shows one such possible instantiation of the ψ_{11} function
- A similar interpretation can be made for $\xi_1(h_1)$

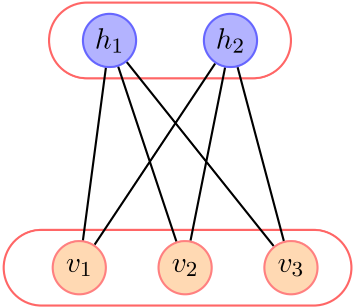
Just to be sure that we understand this correctly let us take a small example where $|V| = 3$ (i.e., $V \in \{0, 1\}^3$) and $|H| = 2$ (i.e., $H \in \{0, 1\}^2$)



$\phi_{11}(v_1,h_1)$			$\phi_{12}(v_1,h_2)$			$\phi_{21}(v_2,h_1)$			$\phi_{22}(v_2,h_2)$			$\phi_{31}(v_3,h_1)$			$\phi_{32}(v_3,h_2)$		
0	0	20	0	0	6	0	0	3	0	0	2	0	0	6	0	0	3
0	1	3	0	1	20	0	1	3	0	1	1	0	1	3	0	1	1
1	0	5	1	0	10	1	0	2	1	0	10	1	0	5	1	0	10
1	1	10	1	1	2	1	1	10	1	1	10	1	1	10	1	1	10

$\psi_1(v_1)$		$\psi_2(v_2)$		$\psi_3(v_3)$		$\xi_1(h_1)$		$\xi_2(h_2)$	
0	30	0	100	0	1	0	100	0	10
1	1	1	1	1	100	1	1	1	10

- Suppose we are now interested in $P(V =< 0,0,0 >, H =< 1,1 >)$

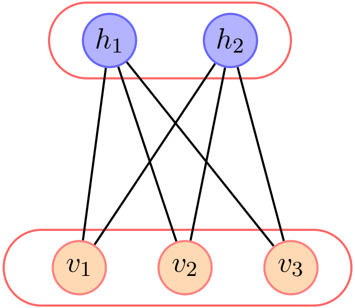


$\phi_{11}(v_1,h_1)$			$\phi_{12}(v_1,h_2)$			$\phi_{21}(v_2,h_1)$			$\phi_{22}(v_2,h_2)$			$\phi_{31}(v_3,h_1)$			$\phi_{32}(v_3,h_2)$		
0	0	20	0	0	6	0	0	3	0	0	2	0	0	6	0	0	3
0	1	3	0	1	20	0	1	3	0	1	1	0	1	3	0	1	1
1	0	5	1	0	10	1	0	2	1	0	10	1	0	5	1	0	10
1	1	10	1	1	2	1	1	10	1	1	10	1	1	10	1	1	10

$\psi_1(v_1)$		$\psi_2(v_2)$		$\psi_3(v_3)$		$\xi_1(h_1)$		$\xi_2(h_2)$	
0	30	0	100	0	1	0	100	0	10
1	1	1	1	1	100	1	1	1	10

- Suppose we are now interested in $P(V =< 0,0,0 >, H =< 1,1 >)$
- We can compute this using the following function

$$
\begin{aligned}
&P(V =< 0,0,0 >, H =< 1,1 >) \\
=&\frac{1}{Z}\phi_{11}(0,1)\phi_{12}(0,1)\phi_{21}(0,1) \\
&\phi_{22}(0,1)\phi_{31}(0,1)\phi_{32}(0,1) \\
&\psi_1(0)\psi_2(0)\psi_3(0)\xi_1(1)\xi_2(1)
\end{aligned}
$$



$\phi_{11}(v_1,h_1)$			$\phi_{12}(v_1,h_2)$			$\phi_{21}(v_2,h_1)$			$\phi_{22}(v_2,h_2)$			$\phi_{31}(v_3,h_1)$			$\phi_{32}(v_3,h_2)$		
0	0	20	0	0	6	0	0	3	0	0	2	0	0	6	0	0	3
0	1	3	0	1	20	0	1	3	0	1	1	0	1	3	0	1	1
1	0	5	1	0	10	1	0	2	1	0	10	1	0	5	1	0	10
1	1	10	1	1	2	1	1	10	1	1	10	1	1	10	1	1	10

$\psi_1(v_1)$		$\psi_2(v_2)$		$\psi_3(v_3)$		$\xi_1(h_1)$		$\xi_2(h_2)$	
0	30	0	100	0	1	0	100	0	10
1	1	1	1	1	100	1	1	1	10

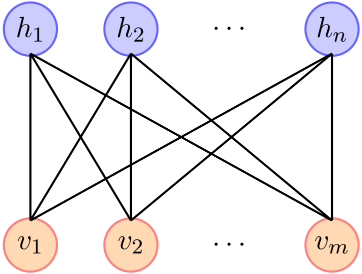
- Suppose we are now interested in $P(V =< 0,0,0 >, H =< 1,1 >)$
- We can compute this using the following function

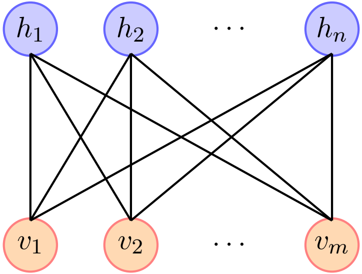
$$
\begin{aligned}
&P(V =< 0,0,0 >, H =< 1,1 >) \\
&= \frac{1}{Z}\phi_{11}(0,1)\phi_{12}(0,1)\phi_{21}(0,1) \\
&\quad \phi_{22}(0,1)\phi_{31}(0,1)\phi_{32}(0,1) \\
&\quad \psi_1(0)\psi_2(0)\psi_3(0)\xi_1(1)\xi_2(1)
\end{aligned}
$$

- and the partition function will be given by

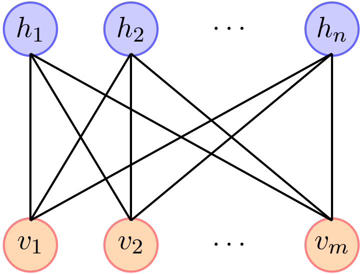
$$
\sum_{v_1=0}^{1}\sum_{v_2=0}^{1}\sum_{v_3=0}^{1}\sum_{h_1=0}^{1}\sum_{h_2=1}^{1} P(V =< v_1, v_2, v_3 >, H =< h_1, h_2 >)
$$

- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?

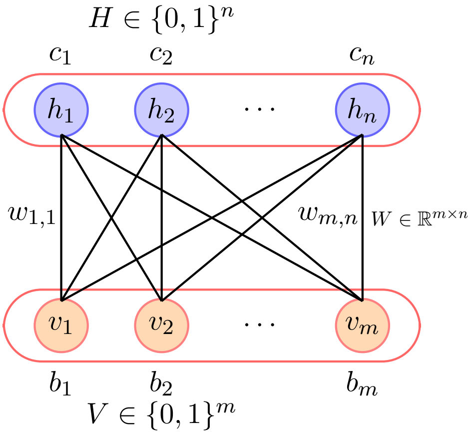




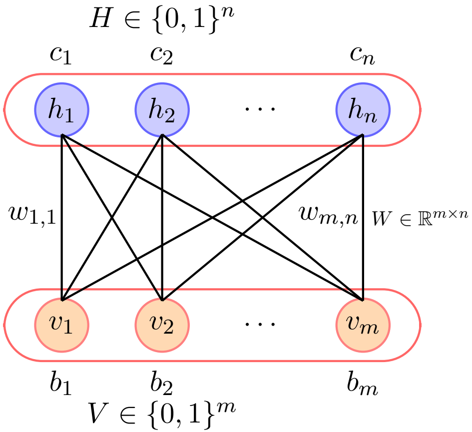
- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?
- Whenever we want to learn something what do we introduce?



- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?
- Whenever we want to learn something what do we introduce? (parameters)

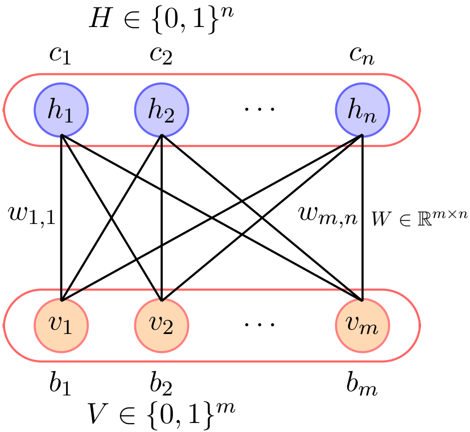


- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?
- Whenever we want to learn something what do we introduce? (parameters)
- So we will introduce a parametric form for these clique potentials and then learn these parameters

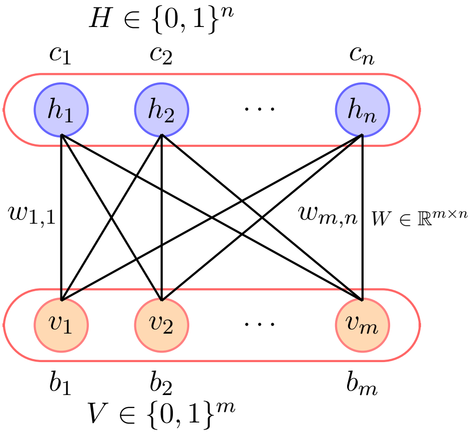


- How do we learn these clique potentials: $\phi_{ij}(v_i, h_j), \psi_i(v_i), \xi_j(h_j)$?
- Whenever we want to learn something what do we introduce? (parameters)
- So we will introduce a parametric form for these clique potentials and then learn these parameters
- The specific parametric form chosen by RBMs is

$$\phi_{ij}(v_i, h_j) = e^{w_{ij} v_i h_j}$$
$$\psi_i(v_i) = e^{b_i v_i}$$
$$\xi_j(h_j) = e^{c_j h_j}$$

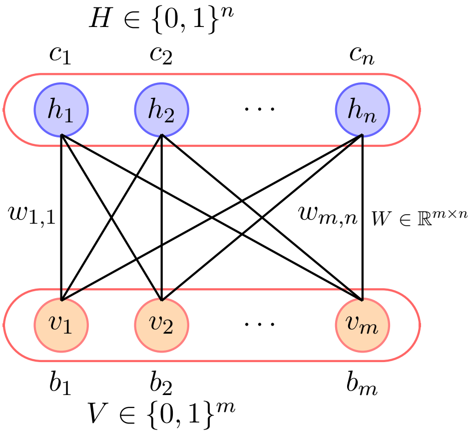


- With this parametric form, let us see what the joint distribution looks like



- With this parametric form, let us see what the joint distribution looks like

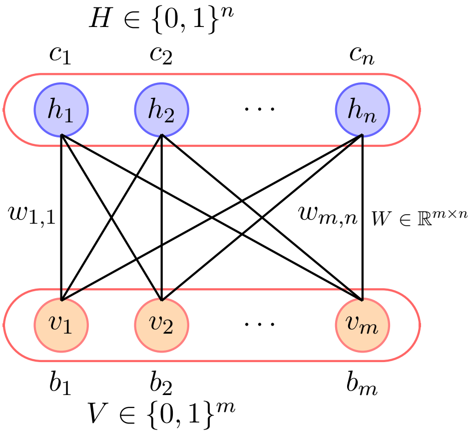
$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i, h_j) \prod_i \psi_i(v_i) \prod_j \xi_j(h_j)$$



- With this parametric form, let us see what the joint distribution looks like

$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i,h_j) \prod_i \psi_i(v_i) \prod_j \xi_j(h_j)$$

$$= \frac{1}{Z}\prod_i \prod_j e^{w_{ij}v_ih_j} \prod_i e^{b_iv_i} \prod_j e^{c_jh_j}$$

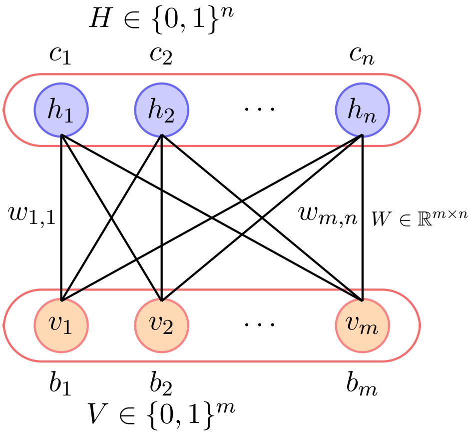


- With this parametric form, let us see what the joint distribution looks like

$$P(V,H) = \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$

$$= \frac{1}{Z}\prod_i\prod_j e^{w_{ij}v_ih_j}\prod_i e^{b_iv_i}\prod_j e^{c_jh_j}$$

$$= \frac{1}{Z}e^{\sum_i\sum_j w_{ij}v_ih_j}e^{\sum_i b_iv_i}e^{\sum_j c_jh_j}$$



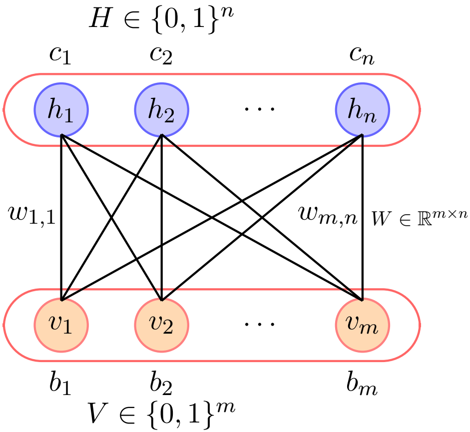
- With this parametric form, let us see what the joint distribution looks like

$$P(V,H) = \frac{1}{Z}\prod_i \prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)$$

$$= \frac{1}{Z}\prod_i \prod_j e^{w_{ij}v_ih_j}\prod_i e^{b_iv_i}\prod_j e^{c_jh_j}$$

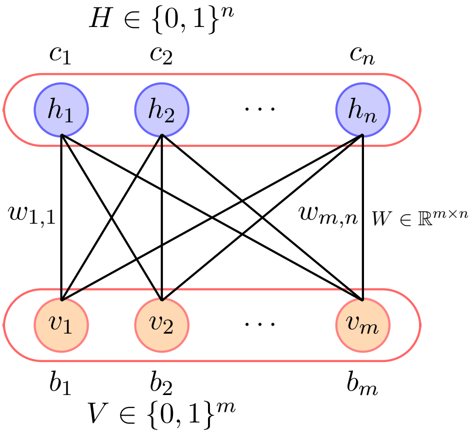
$$= \frac{1}{Z}e^{\sum_i \sum_j w_{ij}v_ih_j}e^{\sum_i b_iv_i}e^{\sum_j c_jh_j}$$

$$= \frac{1}{Z}e^{\sum_i \sum_j w_{ij}v_ih_j+\sum_i b_iv_i+\sum_j c_jh_j}$$



- With this parametric form, let us see what the joint distribution looks like

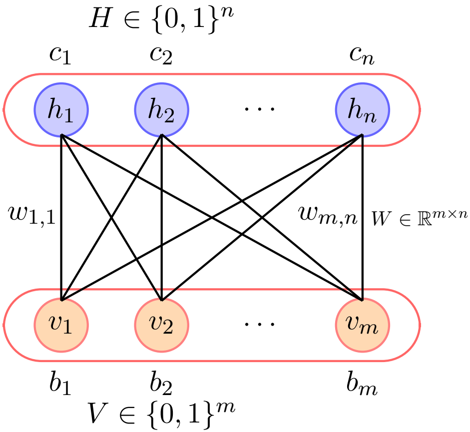
$$
\begin{aligned}
P(V,H) &= \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j) \\
&= \frac{1}{Z}\prod_i\prod_j e^{w_{ij}v_ih_j}\prod_i e^{b_iv_i}\prod_j e^{c_jh_j} \\
&= \frac{1}{Z}e^{\sum_i\sum_j w_{ij}v_ih_j}e^{\sum_i b_iv_i}e^{\sum_j c_jh_j} \\
&= \frac{1}{Z}e^{\sum_i\sum_j w_{ij}v_ih_j+\sum_i b_iv_i+\sum_j c_jh_j} \\
&= \frac{1}{Z}e^{-E(V,H)} \text{ where,}
\end{aligned}
$$



- With this parametric form, let us see what the joint distribution looks like

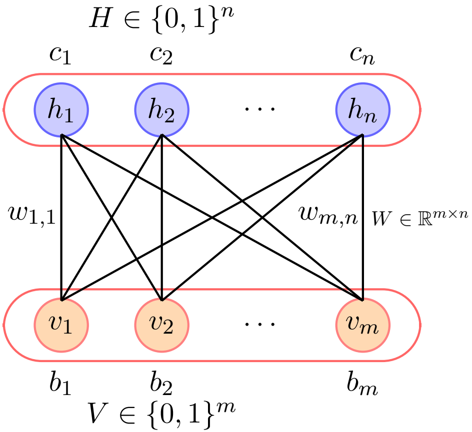
$$
\begin{aligned}
P(V,H) &= \frac{1}{Z}\prod_i\prod_j \phi_{ij}(v_i,h_j)\prod_i \psi_i(v_i)\prod_j \xi_j(h_j)\\
&= \frac{1}{Z}\prod_i\prod_j e^{w_{ij}v_ih_j}\prod_i e^{b_iv_i}\prod_j e^{c_jh_j}\\
&= \frac{1}{Z}e^{\sum_i\sum_j w_{ij}v_ih_j}e^{\sum_i b_iv_i}e^{\sum_j c_jh_j}\\
&= \frac{1}{Z}e^{\sum_i\sum_j w_{ij}v_ih_j+\sum_i b_iv_i+\sum_j c_jh_j}\\
&= \frac{1}{Z}e^{-E(V,H)} \text{ where,}
\end{aligned}
$$

$$
E(V,H) = -\sum_i\sum_j w_{ij}v_ih_j - \sum_i b_iv_i - \sum_j c_jh_j
$$



$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

- Because of the above form, we refer to these networks as (restricted) Boltzmann machines

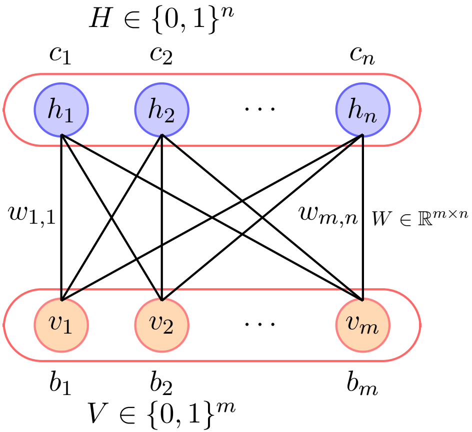


$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

- Because of the above form, we refer to these networks as (restricted) Boltzmann machines
- The term comes from statistical mechanics where the distribution of particles in a system over various possible states is given by

$$F(state) \propto e^{-\frac{E}{kt}}$$

$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$



- Because of the above form, we refer to these networks as (restricted) Boltzmann machines
- The term comes from statistical mechanics where the distribution of particles in a system over various possible states is given by

$$F(state) \propto e^{-\frac{E}{kt}}$$

which is called the Boltzmann distribution or the Gibbs distribution