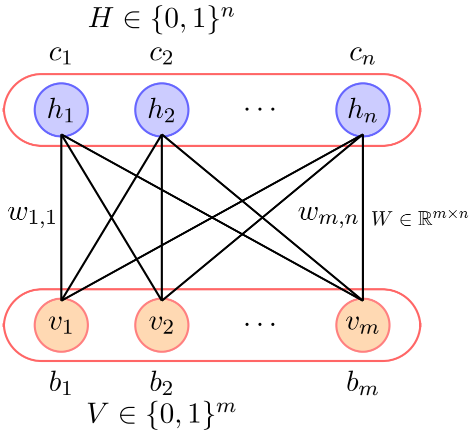
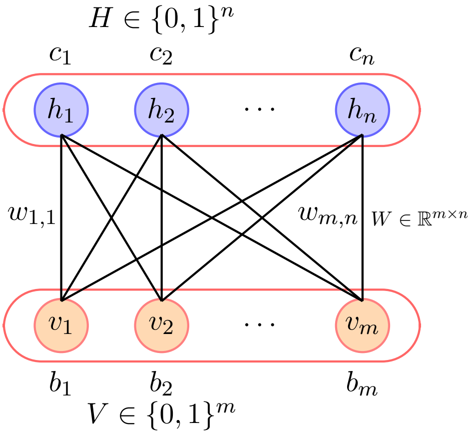
Module 19.6: Computing the gradient of the log likelihood



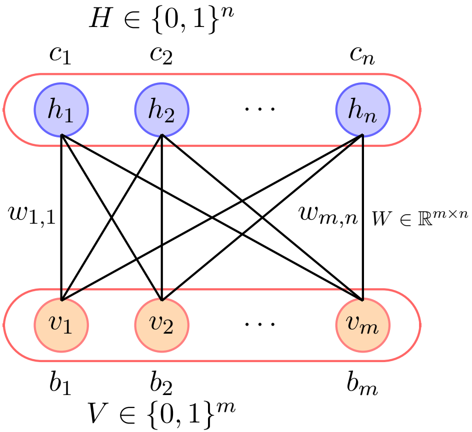
- We will just consider the loss for a single training example

$$\ln \mathscr{L}(\theta)$$



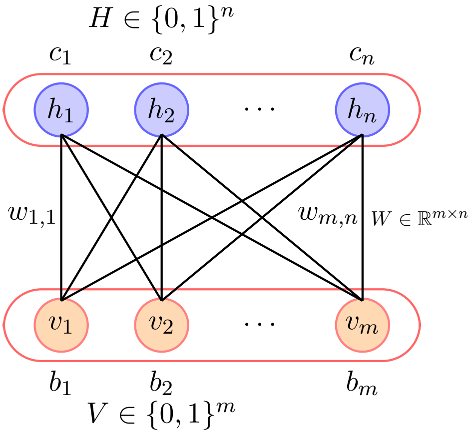
- We will just consider the loss for a single training example

$$\ln \mathscr{L}(\theta) = \ln p(V|\theta)$$



- We will just consider the loss for a single training example

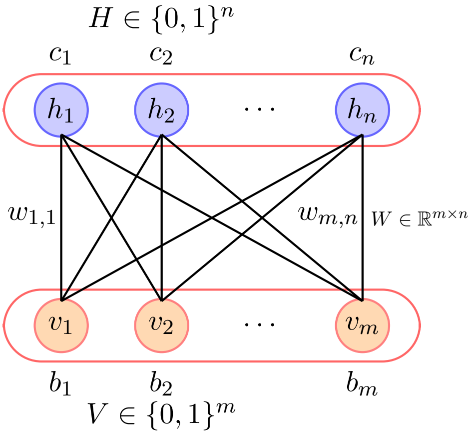
$$\ln \mathscr{L}(\theta) = \ln p(V|\theta) = \ln \frac{1}{Z} \sum_H e^{-E(V,H)}$$



- We will just consider the loss for a single training example

$$\ln \mathscr{L}(\theta) = \ln p(V|\theta) = \ln \frac{1}{Z} \sum_H e^{-E(V,H)}$$

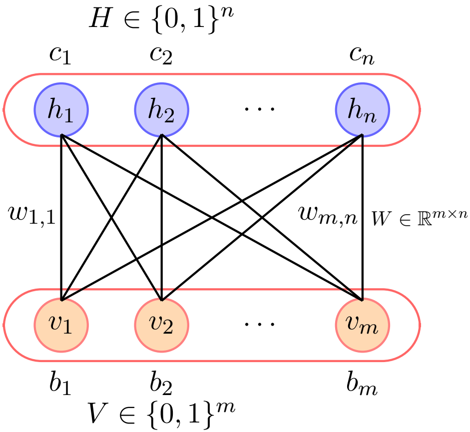
$$= \ln \sum_H e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)}$$



- We will just consider the loss for a single training example

$$\ln \mathscr{L}(\theta) = \ln p(V|\theta) = \ln \frac{1}{Z} \sum_H e^{-E(V,H)}$$
$$= \ln \sum_H e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)}$$

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta}$$

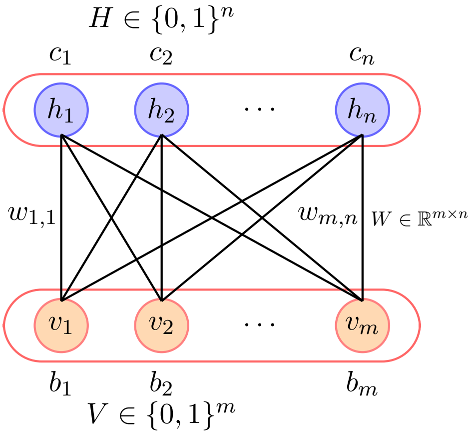


- We will just consider the loss for a single training example

$$\ln \mathscr{L}(\theta) = \ln p(V|\theta) = \ln \frac{1}{Z} \sum_H e^{-E(V,H)}$$

$$= \ln \sum_H e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)}$$

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \left(\ln \sum_H e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)} \right)$$



- We will just consider the loss for a single training example

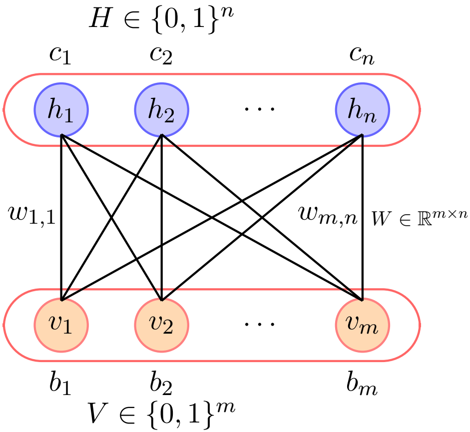
$$\ln \mathscr{L}(\theta) = \ln p(V|\theta) = \ln \frac{1}{Z}\sum_{H} e^{-E(V,H)}$$

$$= \ln \sum_{H} e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)}$$

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\ln \sum_{H} e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)}\right)$$

$$= -\frac{1}{\sum_{H} e^{-E(V,H)}} \sum_{H} e^{-E(V,H)} \frac{\partial E(V,H)}{\partial \theta}$$

$$+ \frac{1}{\sum_{V,H} e^{-E(V,H)}} \sum_{V,H} e^{-E(V,H)} \frac{\partial E(V,H)}{\partial \theta}$$



- We will just consider the loss for a single training example

$$\ln \mathscr{L}(\theta) = \ln p(V|\theta) = \ln \frac{1}{Z} \sum_H e^{-E(V,H)}$$

$$= \ln \sum_H e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)}$$

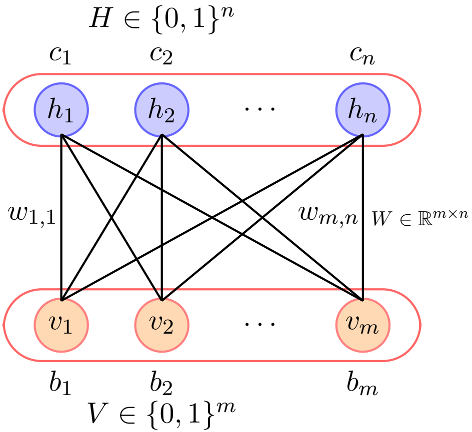
$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \left(\ln \sum_H e^{-E(V,H)} - \ln \sum_{V,H} e^{-E(V,H)} \right)$$

$$= -\frac{1}{\sum_H e^{-E(V,H)}} \sum_H e^{-E(V,H)} \frac{\partial E(V,H)}{\partial \theta}$$

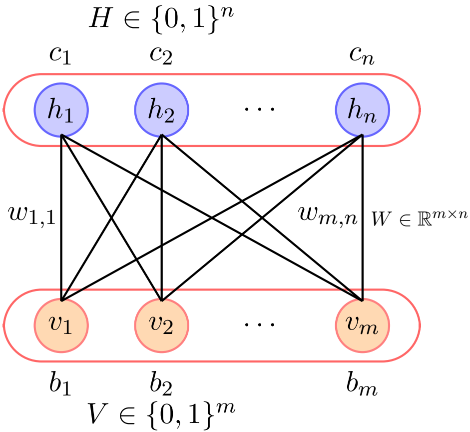
$$+ \frac{1}{\sum_{V,H} e^{-E(V,H)}} \sum_{V,H} e^{-E(V,H)} \frac{\partial E(V,H)}{\partial \theta}$$

$$= -\sum_H \frac{e^{-E(V,H)}}{\sum_H e^{-E(V,H)}} \frac{\partial E(V,H)}{\partial \theta}$$

$$+ \sum_{V,H} \frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} \frac{\partial E(V,H)}{\partial \theta}$$

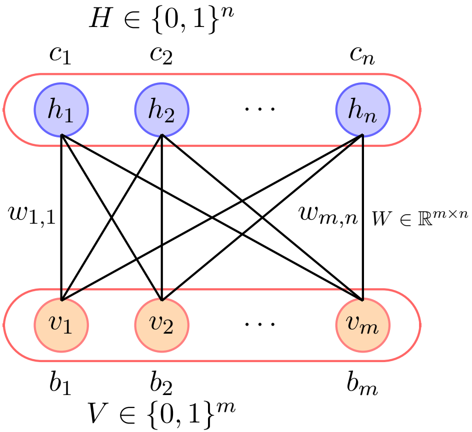

$H \in \{0,1\}^n$
c_1
c_2
c_n
h_1
h_2
$\cdots$
h_n
$w_{1,1}$
$w_{m,n}$
$W \in \mathbb{R}^{m \times n}$
v_1
v_2
$\cdots$
v_m
b_1
b_2
b_m
$V \in \{0,1\}^m$

- Now,



- Now,

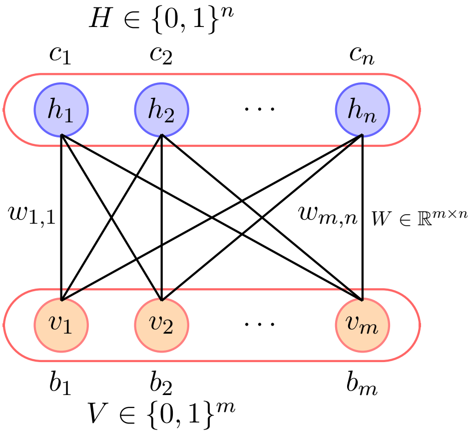
$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$



- Now,

$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$

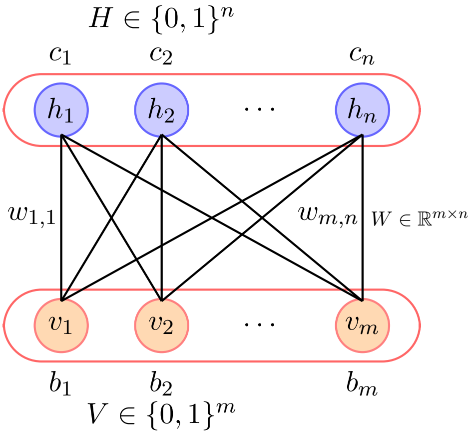
$$\frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}}$$



- Now,

$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$

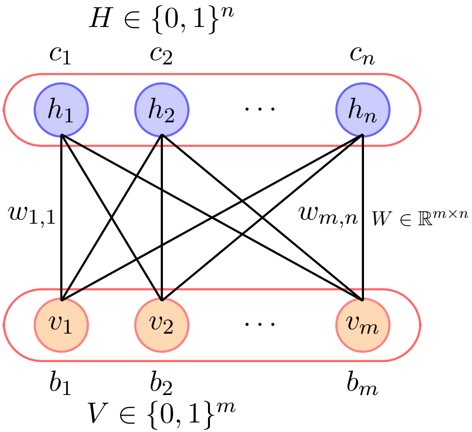
$$\frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} = \frac{\frac{1}{Z} e^{-E(V,H)}}{\frac{1}{Z}\sum_{H} e^{-E(V,H)}}$$



- Now,

$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$

$$\begin{aligned}\frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} &= \frac{\frac{1}{Z}e^{-E(V,H)}}{\frac{1}{Z}\sum_{H} e^{-E(V,H)}}\\ &= \frac{p(V,H)}{p(V)}\end{aligned}$$

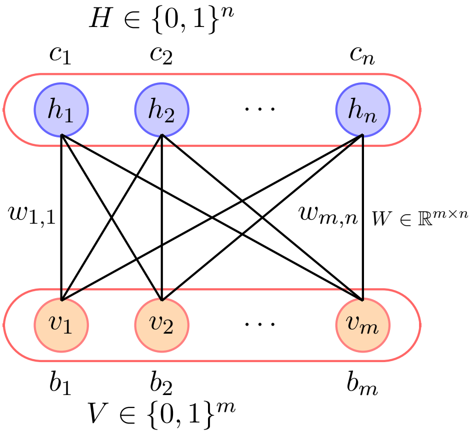


- Now,

$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$

$$\frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} = \frac{\frac{1}{Z} e^{-E(V,H)}}{\frac{1}{Z}\sum_{H} e^{-E(V,H)}}$$

$$= \frac{p(V,H)}{p(V)} = p(H|V)$$



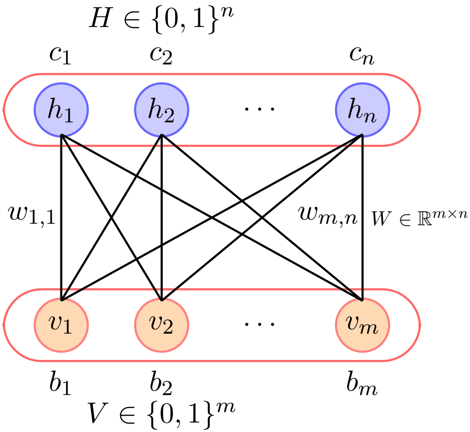
- Now,

$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$

$$\frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} = \frac{\frac{1}{Z} e^{-E(V,H)}}{\frac{1}{Z}\sum_{H} e^{-E(V,H)}}$$

$$= \frac{p(V,H)}{p(V)} = p(H|V)$$

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta}$$

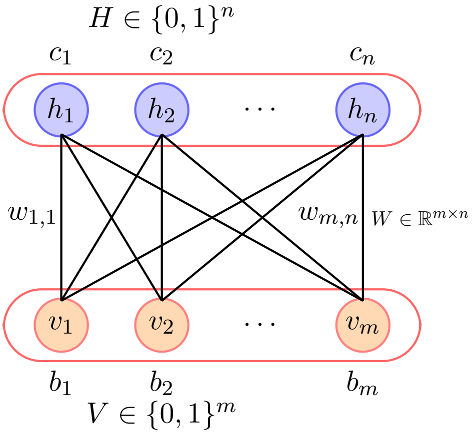


- Now,

$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$

$$\begin{aligned}\frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} &= \frac{\frac{1}{Z}e^{-E(V,H)}}{\frac{1}{Z}\sum_{H} e^{-E(V,H)}} \\ &= \frac{p(V,H)}{p(V)} = p(H|V)\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = -\sum_{H} \frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} \frac{\partial E(V,H)}{\partial \theta} \\ + \sum_{V,H} \frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} \frac{\partial E(V,H)}{\partial \theta}\end{aligned}$$



- Now,

$$\frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} = p(V,H)$$

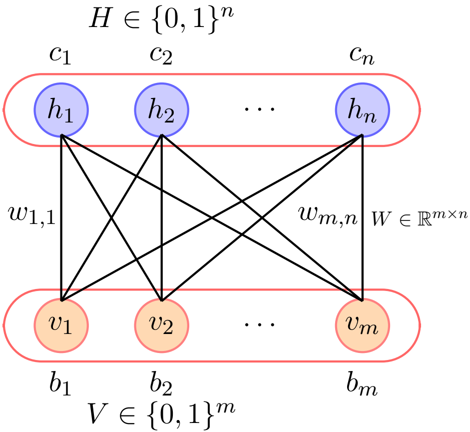
$$\frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} = \frac{\frac{1}{Z}e^{-E(V,H)}}{\frac{1}{Z}\sum_{H} e^{-E(V,H)}}$$

$$= \frac{p(V,H)}{p(V)} = p(H|V)$$

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = -\sum_{H} \frac{e^{-E(V,H)}}{\sum_{H} e^{-E(V,H)}} \frac{\partial E(V,H)}{\partial \theta}$$

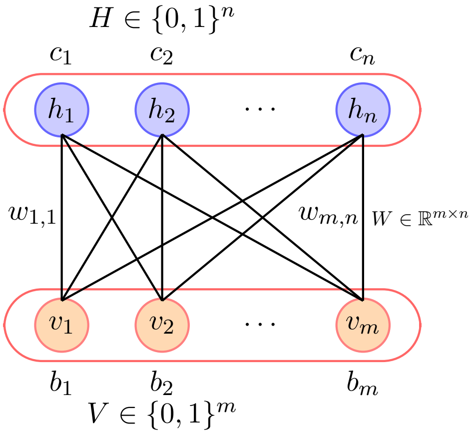
$$+ \sum_{V,H} \frac{e^{-E(V,H)}}{\sum_{V,H} e^{-E(V,H)}} \frac{\partial E(V,H)}{\partial \theta}$$

$$= -\sum_{H} p(H|V)\frac{\partial E(V,H)}{\partial \theta} + \sum_{V,H} p(V,H)\frac{\partial E(V,H)}{\partial \theta}$$



- Okay, so we have,

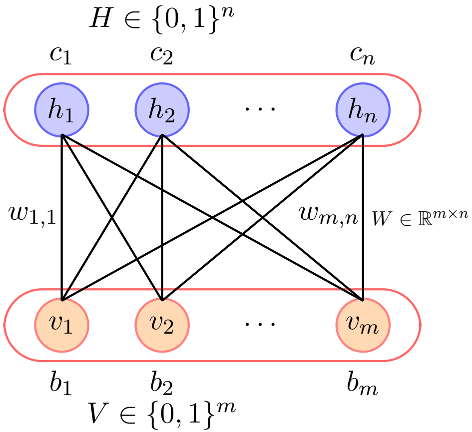
$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = -\sum_{H} p(H|V) \frac{\partial E(V,H)}{\partial \theta} + \sum_{V,H} p(V,H) \frac{\partial E(V,H)}{\partial \theta}$$



- Okay, so we have,

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = -\sum_{H} p(H|V) \frac{\partial E(V,H)}{\partial \theta} + \sum_{V,H} p(V,H) \frac{\partial E(V,H)}{\partial \theta}$$

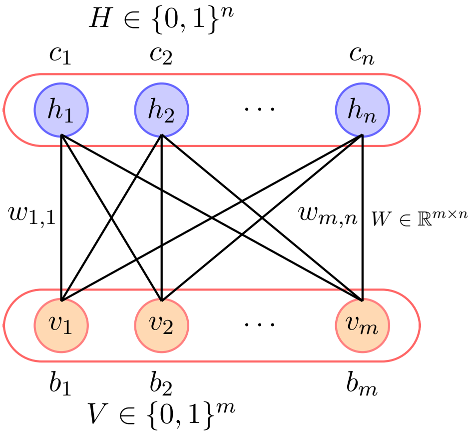
- Remember that θ is a collection of all the parameters in our model, i.e., $W_{ij}, b_i, c_j \forall i \in \{1, \ldots, m\}$ and $\forall j \in \{1, \ldots, n\}$



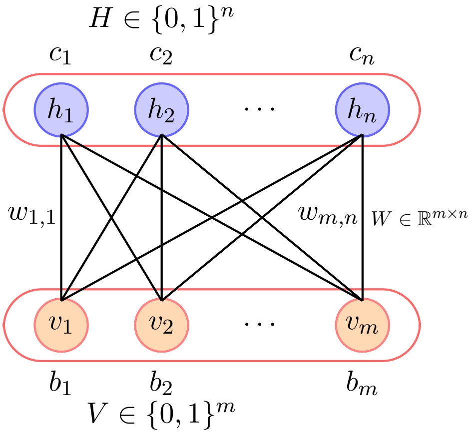
- Okay, so we have,

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} = -\sum_H p(H|V) \frac{\partial E(V,H)}{\partial \theta} + \sum_{V,H} p(V,H) \frac{\partial E(V,H)}{\partial \theta}$$

- Remember that θ is a collection of all the parameters in our model, i.e., $W_{ij}, b_i, c_j \forall i \in \{1,\dots,m\}$ and $\forall j \in \{1,\dots,n\}$
- We will follow our usual recipe of computing the partial derivative w.r.t. one weight w_{ij} and then generalize to the gradient w.r.t. the entire weight matrix W

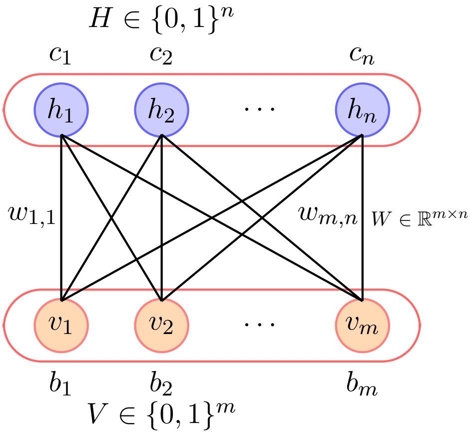


$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}}$$



$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}}$$

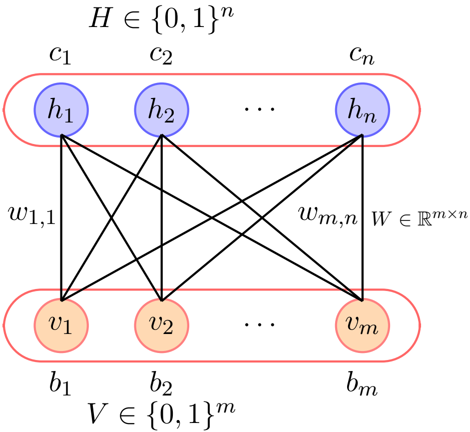
$$= -\sum_{H} p(H|V)\frac{\partial E(V,H)}{\partial w_{ij}} + \sum_{V,H} p(V,H)\frac{\partial E(V,H)}{\partial w_{ij}}$$



$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}}$$

$$= -\sum_H p(H|V)\frac{\partial E(V,H)}{\partial w_{ij}} + \sum_{V,H} p(V,H)\frac{\partial E(V,H)}{\partial w_{ij}}$$

$$= \sum_H p(H|V) h_i v_j - \sum_{V,H} p(V,H) h_i v_j$$

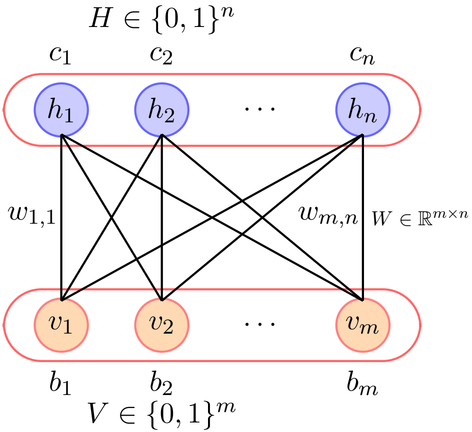


$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}}$$

$$= -\sum_{H} p(H|V) \frac{\partial E(V,H)}{\partial w_{ij}} + \sum_{V,H} p(V,H) \frac{\partial E(V,H)}{\partial w_{ij}}$$

$$= \sum_{H} p(H|V) h_i v_j - \sum_{V,H} p(V,H) h_i v_j$$

- We can write the above as a sum of two expectations



$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}}$$

$$= -\sum_H p(H|V)\frac{\partial E(V,H)}{\partial w_{ij}} + \sum_{V,H} p(V,H)\frac{\partial E(V,H)}{\partial w_{ij}}$$

$$= \sum_H p(H|V)h_i v_j - \sum_{V,H} p(V,H)h_i v_j$$

$$= \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

- We can write the above as a sum of two expectations

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

- How do we compute these expectations?

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

- How do we compute these expectations?
- The first summation can actually be simplified (we will come back and simplify it later)

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

- How do we compute these expectations?
- The first summation can actually be simplified (we will come back and simplify it later)
- However, the second summation contains an exponential number of terms and hence intractable in practice

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]$$

- How do we compute these expectations?
- The first summation can actually be simplified (we will come back and simplify it later)
- However, the second summation contains an exponential number of terms and hence intractable in practice
- So how do we deal with this ?