CS7015 (Deep Learning) : Lecture 20

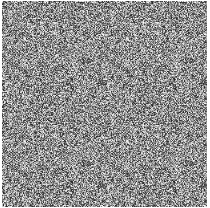
Markov Chains, Gibbs Sampling for Training RBMs, Contrastive Divergence for training RBMs

Mitesh M. Khapra

Department of Computer Science and Engineering
Indian Institute of Technology Madras

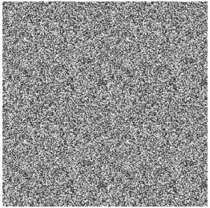
Module 20.1 : Markov Chains

- Let us first begin by restating our goals



$X \in R^{1024}$

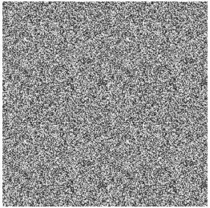
- Let us first begin by restating our goals
- **Goal 1**: Given a random variable $X \in R^n$, we are interested in drawing samples from the joint distribution $P(\mathbf{X})$



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

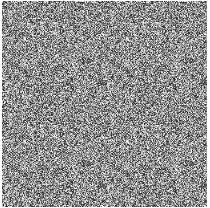
- Let us first begin by restating our goals
- **Goal 1**: Given a random variable $X \in R^n$, we are interested in drawing samples from the joint distribution $P(\mathbf{X})$
- **Goal 2**: Given a function $f(X)$ defined over the random variable X, we are interested in computing the expectation $\mathbb{E}_{P(X)}[f(X)]$



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

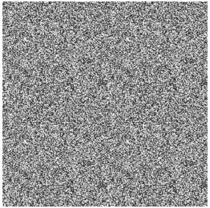
- Let us first begin by restating our goals
- **Goal 1**: Given a random variable $X \in R^n$, we are interested in drawing samples from the joint distribution $P(\mathbf{X})$
- **Goal 2**: Given a function $f(X)$ defined over the random variable X, we are interested in computing the expectation $\mathbb{E}_{P(X)}[f(X)]$
- We will use Gibbs Sampling (class of Metropolis-Hastings algorithm) to achieve these goals



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

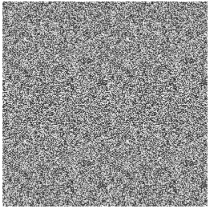
- Let us first begin by restating our goals
- **Goal 1**: Given a random variable $X \in R^n$, we are interested in drawing samples from the joint distribution $P(\mathbf{X})$
- **Goal 2**: Given a function $f(X)$ defined over the random variable X, we are interested in computing the expectation $\mathbb{E}_{P(X)}[f(X)]$
- We will use Gibbs Sampling (class of Metropolis-Hastings algorithm) to achieve these goals
- We will first understand the intuition behind Gibbs Sampling and then understand the math behind it



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

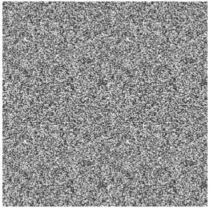
- Suppose instead of a single random variable $X \in R^n$, we have a chain of random variables $X_1, X_2, \ldots, X_K$ each $X_i \in R^n$



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

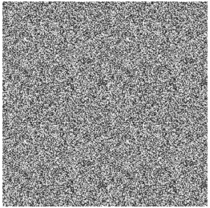
- Suppose instead of a single random variable $X \in R^n$, we have a chain of random variables $X_1, X_2, \ldots, X_K$ each $X_i \in R^n$
- The i here corresponds to a time step



$X \in R^{1024}$

$$\mathbb{E}_{P(X)}[f(X)]$$

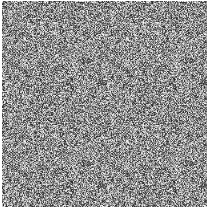
- Suppose instead of a single random variable $X \in R^n$, we have a chain of random variables $X_1, X_2, \ldots, X_K$ each $X_i \in R^n$
- The i here corresponds to a time step
- For example, X_i could be a n-dimensional vector containing the number of customers in a given set of n restaurants on day i



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

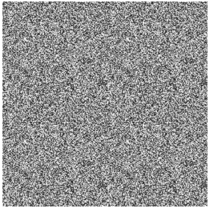
- Suppose instead of a single random variable $X \in R^n$, we have a chain of random variables $X_1, X_2, \ldots, X_K$ each $X_i \in R^n$
- The i here corresponds to a time step
- For example, X_i could be a n-dimensional vector containing the number of customers in a given set of n restaurants on day i
- In our case, X_i could be a 1024 dimensional image sent by our friend on day i



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

- Suppose instead of a single random variable $X \in R^n$, we have a chain of random variables $X_1, X_2, \ldots, X_K$ each $X_i \in R^n$
- The i here corresponds to a time step
- For example, X_i could be a n-dimensional vector containing the number of customers in a given set of n restaurants on day i
- In our case, X_i could be a 1024 dimensional image sent by our friend on day i
- For ease of illustration we will stick to the restaurant example and assume that instead of actual counts we are interested only in binary counts (high=1, low=0)



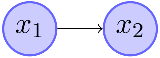
$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

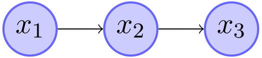
- Suppose instead of a single random variable $X \in R^n$, we have a chain of random variables $X_1, X_2, \ldots, X_K$ each $X_i \in R^n$
- The i here corresponds to a time step
- For example, X_i could be a n-dimensional vector containing the number of customers in a given set of n restaurants on day i
- In our case, X_i could be a 1024 dimensional image sent by our friend on day i
- For ease of illustration we will stick to the restaurant example and assume that instead of actual counts we are interested only in binary counts (high=1, low=0)
- Thus $X_i \in \{0, 1\}^n$

- On day 1, let X_1 take on the value x_1 (x_1 is one of the possible 2^n vectors)

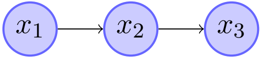
- On day 1, let X_1 take on the value x_1 (x_1 is one of the possible 2^n vectors)
- On day 2, let X_2 take on the value x_2 (x_2 is again one of the possible 2^n vectors)



- On day 1, let X_1 take on the value x_1 (x_1 is one of the possible 2^n vectors)
- On day 2, let X_2 take on the value x_2 (x_2 is again one of the possible 2^n vectors)
- One way of looking at this is that the state has transitioned from x_1 to x_2



- On day 1, let X_1 take on the value x_1 (x_1 is one of the possible 2^n vectors)
- On day 2, let X_2 take on the value x_2 (x_2 is again one of the possible 2^n vectors)
- One way of looking at this is that the state has transitioned from x_1 to x_2
- Similarly, on day 3, if X_3 takes on the value x_3 then we can say that the state has transitioned from x_1 to x_2 to x_3



- On day 1, let X_1 take on the value x_1 (x_1 is one of the possible 2^n vectors)
- On day 2, let X_2 take on the value x_2 (x_2 is again one of the possible 2^n vectors)
- One way of looking at this is that the state has transitioned from x_1 to x_2
- Similarly, on day 3, if X_3 takes on the value x_3 then we can say that the state has transitioned from x_1 to x_2 to x_3
- Finally, on day n, we can say that the state has transitioned from x_1 to x_2 to x_3 to $\ldots x_n$

- We may now be interested in knowing what is the most likely value that the state will take on day i given the states on day 1 to day $i-1$

$$x_1 \to x_2 \to x_3 \to \cdots \to x_i$$

- We may now be interested in knowing what is the most likely value that the state will take on day i given the states on day 1 to day $i-1$
- More formally, we may be interested in the following distribution

$$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1})$$

$$x_1 \to x_2 \to x_3 \to \cdots \to x_i$$

$$x_1 \to x_2 \to x_3 \to \cdots \to x_i$$

- We may now be interested in knowing what is the most likely value that the state will take on day i given the states on day 1 to day $i-1$
- More formally, we may be interested in the following distribution

$$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1})$$

- Now suppose the chain exhibits the following Markov property

$$\begin{aligned} P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) \\ = P(X_i = x_i | X_{i-1} = x_{i-1}) \end{aligned}$$

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow \cdots \rightarrow x_i$$

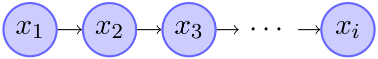
- We may now be interested in knowing what is the most likely value that the state will take on day i given the states on day 1 to day $i-1$
- More formally, we may be interested in the following distribution

$$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1})$$

- Now suppose the chain exhibits the following Markov property

$$\begin{aligned} P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) \\ = P(X_i = x_i | X_{i-1} = x_{i-1}) \end{aligned}$$

- In other words, given the previous state X_{i-1}, X_i is independent of all preceding states



- We may now be interested in knowing what is the most likely value that the state will take on day i given the states on day 1 to day $i-1$
- More formally, we may be interested in the following distribution

$$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1})$$

- Now suppose the chain exhibits the following Markov property

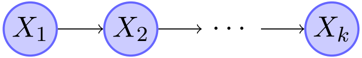
$$\begin{aligned} P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) \\ = P(X_i = x_i | X_{i-1} = x_{i-1}) \end{aligned}$$

- In other words, given the previous state X_{i-1}, X_i is independent of all preceding states
- Can we draw a graphical model to encode this independence assumption ?

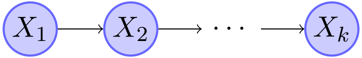
- In this graphical model, the random variables are $X_1, X_2, \ldots, X_k$

- In this graphical model, the random variables are $X_1, X_2, \ldots, X_k$
- We will have a node corresponding to each of these random variables

- In this graphical model, the random variables are $X_1, X_2, \ldots, X_k$
- We will have a node corresponding to each of these random variables
- What will be the edges in the graph ?



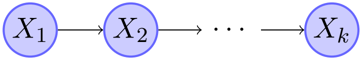
- In this graphical model, the random variables are $X_1, X_2, \ldots, X_k$
- We will have a node corresponding to each of these random variables
- What will be the edges in the graph ?
- Well, each node only depends on its predecessor, so we will just have an edge between successive nodes



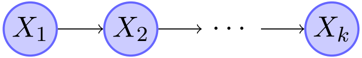
- This property ($X_i \perp\!\!\!\perp X_1^{i-2} | X_{i-1}$) is called the Markov property

- This property ($X_i \perp\!\!\!\perp X_1^{i-2} | X_{i-1}$) is called the Markov property
- And the resulting chain $X_1, X_2, \ldots, X_k$ is called a Markov chain

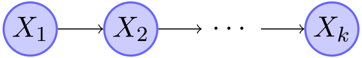
$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$



- This property ($X_i \perp\!\!\!\perp X_1^{i-2} | X_{i-1}$) is called the Markov property
- And the resulting chain $X_1, X_2, \ldots, X_k$ is called a Markov chain
- Further, since we are considering discrete time steps, this is called a discrete time Markov Chain



- This property ($X_i \perp\!\!\!\perp X_1^{i-2} | X_{i-1}$) is called the Markov property
- And the resulting chain $X_1, X_2, \ldots, X_k$ is called a Markov chain
- Further, since we are considering discrete time steps, this is called a discrete time Markov Chain
- Further, since X_i's take on discrete values this is called a discrete time discrete space Markov Chain



- This property ($X_i \perp\!\!\!\perp X_1^{i-2} | X_{i-1}$) is called the Markov property
- And the resulting chain $X_1, X_2, \ldots, X_k$ is called a Markov chain
- Further, since we are considering discrete time steps, this is called a discrete time Markov Chain
- Further, since X_i's take on discrete values this is called a discrete time discrete space Markov Chain
- Okay, but why are we interested in Markov chains? (we will get there soon! for now let us just focus on these definitions)

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$

- Recall that each $X_i \in \{0,1\}^n$

- Let us delve a bit deeper into Markov Chains and define a few more quantities

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_k$$

- Recall that each $X_i \in \{0,1\}^n$

- Let us delve a bit deeper into Markov Chains and define a few more quantities
- Let us assume $2^n = l$ (*i.e.*, X_i can take l values)

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$

- Recall that each $X_i \in \{0,1\}^n$

- Let us delve a bit deeper into Markov Chains and define a few more quantities
- Let us assume $2^n = l$ ($i.e.$, X_i can take l values)
- How many values do we need to specify the distribution

$$P(X_i = x_i | X_{i-1} = x_{i-1})?$$

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_k$$

- Recall that each $X_i \in \{0,1\}^n$

- Let us delve a bit deeper into Markov Chains and define a few more quantities
- Let us assume $2^n = l$ (*i.e.*, X_i can take l values)
- How many values do we need to specify the distribution

$$P(X_i = x_i | X_{i-1} = x_{i-1})? \quad (l^2)$$

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$

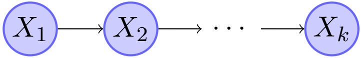
- Recall that each $X_i \in \{0, 1\}^n$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- Let us delve a bit deeper into Markov Chains and define a few more quantities
- Let us assume $2^n = l$ (*i.e.*, X_i can take l values)
- How many values do we need to specify the distribution

$$P(X_i = x_i | X_{i-1} = x_{i-1})? \quad (l^2)$$

- We can represent this as a matrix $T \in l \times l$ where the entry $T_{a,b}$ of the matrix denotes the probability of transitioning to state b from state a (*i.e.*, $P(X_i = b | X_{i-1} = a)$)



- Recall that each $X_i \in \{0,1\}^n$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- Let us delve a bit deeper into Markov Chains and define a few more quantities
- Let us assume $2^n = l$ ($i.e.$, X_i can take l values)
- How many values do we need to specify the distribution

$$P(X_i = x_i | X_{i-1} = x_{i-1})? \quad (l^2)$$

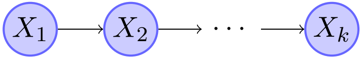
- We can represent this as a matrix $T \in l \times l$ where the entry $T_{a,b}$ of the matrix denotes the probability of transitioning to state b from state a ($i.e.$, $P(X_i = b | X_{i-1} = a)$)
- The matrix T is called the transition matrix

$$X_1 \to X_2 \to \cdots \to X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- We need to define this transition matrix T_{ab}, *i.e.*,

$$P(X_i = b | X_{i-1} = a) \quad \forall a, b \quad \forall i$$



X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- We need to define this transition matrix T_{ab}, *i.e.*,

$$P(X_i = b | X_{i-1} = a) \quad \forall a, b \quad \forall i$$

- Why do we need to define this $\forall i$?

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- We need to define this transition matrix T_{ab}, *i.e.*,

$$P(X_i = b | X_{i-1} = a) \quad \forall a, b \quad \forall i$$

- Why do we need to define this $\forall i$? Well, because this transition probabilities may be different for different time steps

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- We need to define this transition matrix T_{ab}, *i.e.*,

$$P(X_i = b | X_{i-1} = a) \;\; \forall a, b \;\; \forall i$$

- Why do we need to define this $\forall i$? Well, because this transition probabilities may be different for different time steps
- For example, the transition in the number of customers may be different from Friday to Saturday (weekend) as compared to from Sunday to Monday(weekday)

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- We need to define this transition matrix T_{ab}, *i.e.*,

$$P(X_i = b | X_{i-1} = a) \quad \forall a, b \quad \forall i$$

- Why do we need to define this $\forall i$? Well, because this transition probabilities may be different for different time steps
- For example, the transition in the number of customers may be different from Friday to Saturday (weekend) as compared to from Sunday to Monday(weekday)
- Thus, for a Markov chain $X_1, X_2, \ldots, X_k$ we will have k such transition matrices $T_1, T_2, \ldots, T_k$

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- However, for this discussion we will assume that the Markov chain is time homogeneous

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- However, for this discussion we will assume that the Markov chain is time homogeneous
- What does that mean? It means that

$$T_1 = T_2 = \cdots = T_k = T$$

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

- However, for this discussion we will assume that the Markov chain is time homogeneous
- What does that mean? It means that

$$T_1 = T_2 = \cdots = T_k = T$$

- In other words

$$P(X_i = b | X_{i-1} = a) = T_{ab} \quad \forall a, b \quad \forall i$$

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$

X_{i-1}	X_{i-2}	T_{ab}
1	1	0.05
1	2	0.06
$\vdots$	$\vdots$	$\vdots$
1	l	0.02
2	1	0.03
2	2	0.07
$\vdots$	$\vdots$	$\vdots$
2	l	0.01
$\vdots$	$\vdots$	$\vdots$
l	1	0.1
l	2	0.09
$\vdots$	$\vdots$	$\vdots$
l	l	0.21

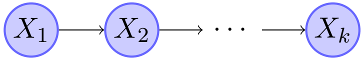
- However, for this discussion we will assume that the Markov chain is time homogeneous
- What does that mean? It means that

$$T_1 = T_2 = \cdots = T_k = T$$

- In other words

$$P(X_i = b | X_{i-1} = a) = T_{ab} \quad \forall a, b \quad \forall i$$

- The transition matrix does not depend on the time i and hence such a Markov Chain is called *time homogeneous*

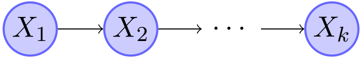


- Now suppose the starting distribution at time step 0 is given by μ^0)

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_k$$

- Now suppose the starting distribution at time step 0 is given by μ^0)
- Just to be clear μ^0 is a 2^n dimensional vector such that

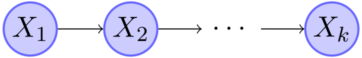
$$\mu_a^0 = P(X_0 = a)$$



- Now suppose the starting distribution at time step 0 is given by μ^0)
- Just to be clear μ^0 is a 2^n dimensional vector such that

$$\mu_a^0 = P(X_0 = a)$$

- μ_a^0 is the probability that the random variable takes on the value a among all the possible 2^n values

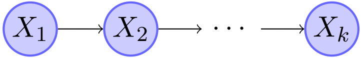


- Now suppose the starting distribution at time step 0 is given by μ^0)
- Just to be clear μ^0 is a 2^n dimensional vector such that

$$\mu_a^0 = P(X_0 = a)$$

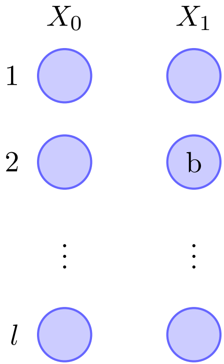
- μ_a^0 is the probability that the random variable takes on the value a among all the possible 2^n values
- Given μ^0 and T how will you compute μ^k where

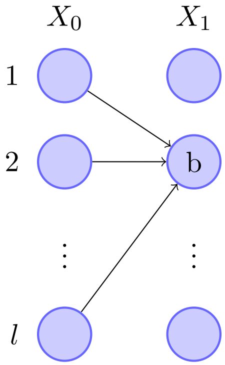
$$\mu_a^k = P(X_k = a)$$



- Now suppose the starting distribution at time step 0 is given by μ^0)
- Just to be clear μ^0 is a 2^n dimensional vector such that
$$\mu_a^0 = P(X_0 = a)$$
- μ_a^0 is the probability that the random variable takes on the value a among all the possible 2^n values
- Given μ^0 and T how will you compute μ^k where
$$\mu_a^k = P(X_k = a)$$
- μ^k is again a 2^n dimensional vector whose a^{th} entry tells us the probability that X_k will take on the value a among all the possible 2^n values

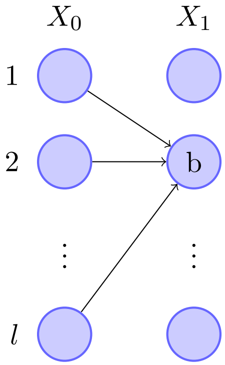
- Let us consider $P(X_1 = b)$





- Let us consider $P(X_1 = b)$

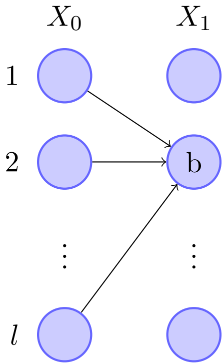
$$P(X_1 = b) = \sum_a P(X_0 = a, X_1 = b)$$



- Let us consider $P(X_1 = b)$

$$P(X_1 = b) = \sum_{a} P(X_0 = a, X_1 = b)$$

- The above sum essentially captures all the paths of reaching $X_1 = b$ irrespective of the value of X_0

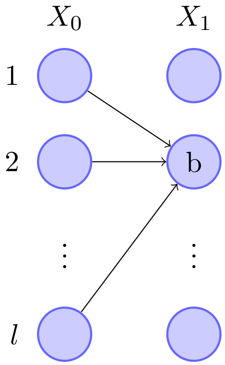


- Let us consider $P(X_1 = b)$

$$P(X_1 = b) = \sum_a P(X_0 = a, X_1 = b)$$

- The above sum essentially captures all the paths of reaching $X_1 = b$ irrespective of the value of X_0

$$P(X_1 = b) = \sum_a P(X_0 = a, X_1 = b)$$

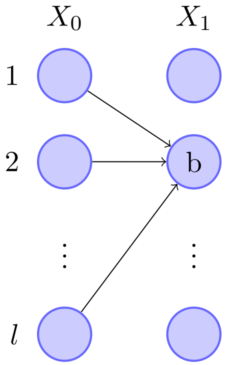


- Let us consider $P(X_1 = b)$

$$P(X_1 = b) = \sum_a P(X_0 = a, X_1 = b)$$

- The above sum essentially captures all the paths of reaching $X_1 = b$ irrespective of the value of X_0

$$\begin{aligned} P(X_1 = b) &= \sum_a P(X_0 = a, X_1 = b) \\ &= \sum_a P(X_0 = a) P(X_1 = b | X_0 = a) \end{aligned}$$



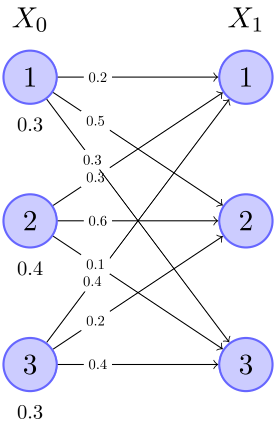
- Let us consider $P(X_1 = b)$

$$P(X_1 = b) = \sum_a P(X_0 = a, X_1 = b)$$

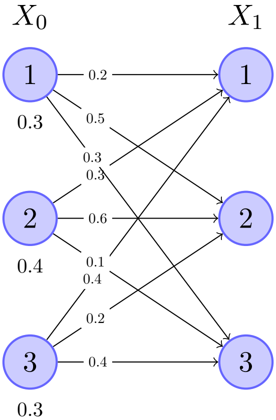
- The above sum essentially captures all the paths of reaching $X_1 = b$ irrespective of the value of X_0

$$\begin{aligned} P(X_1 = b) &= \sum_a P(X_0 = a, X_1 = b) \\ &= \sum_a P(X_0 = a) P(X_1 = b | X_0 = a) \\ &= \sum_a \mu_a^0 T_{ab} \end{aligned}$$

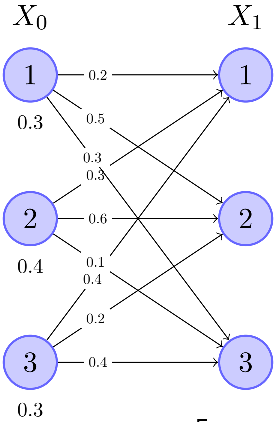
- Let us see if there is a more compact way of writing the distribution $P(X_1)$ ($i.e.$, of specifying $P(X_1 = b)\ \forall b$)



- Let us see if there is a more compact way of writing the distribution $P(X_1)$ (*i.e.*, of specifying $P(X_1 = b)\ \forall b$)
- Let us consider a simple case when $l = 3$ (as opposed to 2^n)

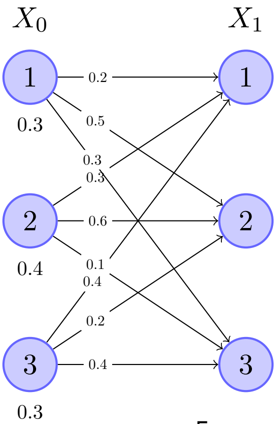


- Let us see if there is a more compact way of writing the distribution $P(X_1)$ ($i.e.$, of specifying $P(X_1 = b)\ \forall b$)
- Let us consider a simple case when $l = 3$ (as opposed to 2^n)
- Thus, $\mu^0 \in R^3$ and $T \in R^{3\times3}$



$$\mu^0 T = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix} \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.3 & 0.6 & 0.1 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}$$

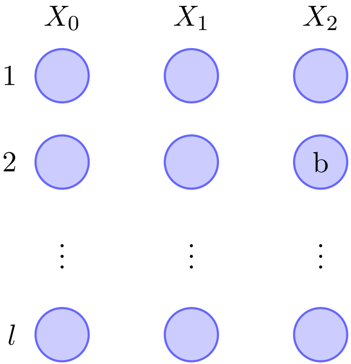
- Let us see if there is a more compact way of writing the distribution $P(X_1)$ ($i.e.$, of specifying $P(X_1 = b)\ \forall b$)
- Let us consider a simple case when $l = 3$ (as opposed to 2^n)
- Thus, $\mu^0 \in R^3$ and $T \in R^{3\times 3}$
- What does the product $\mu^0 T$ give us ?



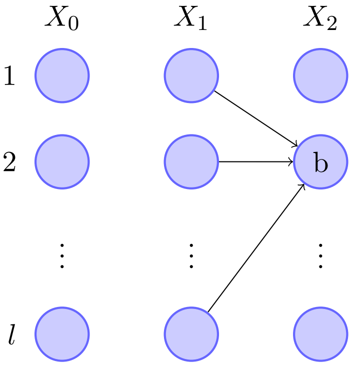
$$\mu^0 T = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix} \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.3 & 0.6 & 0.1 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}$$

$$= \begin{bmatrix} 0.3 & 0.45 & 0.25 \end{bmatrix}$$

- Let us see if there is a more compact way of writing the distribution $P(X_1)$ (*i.e.*, of specifying $P(X_1 = b)\ \forall b$)
- Let us consider a simple case when $l = 3$ (as opposed to 2^n)
- Thus, $\mu^0 \in R^3$ and $T \in R^{3\times 3}$
- What does the product $\mu^0 T$ give us ?
- It gives us the distribution μ_1! (the b^{th} entry of this vector is $\sum_a \mu_a^0 T_{ab}$ which is $P(X_1 = b)$)

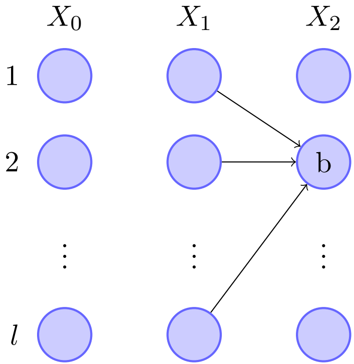


- Let us consider $P(X_2 = b)$



- Let us consider $P(X_2 = b)$

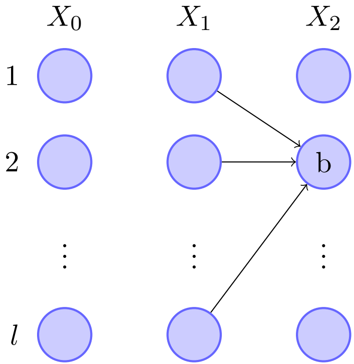
$$P(X_2 = b) = \sum_a P(X_1 = a, X_2 = b)$$



- Let us consider $P(X_2 = b)$

$$P(X_2 = b) = \sum_a P(X_1 = a, X_2 = b)$$

- The above sum essentially captures all the paths of reaching $X_2 = b$ irrespective of the value of X_1

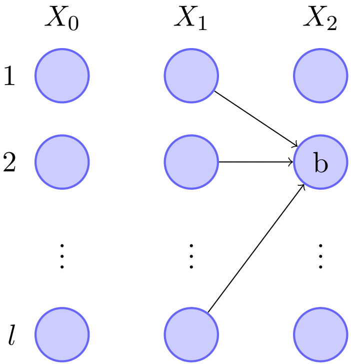


- Let us consider $P(X_2 = b)$

$$P(X_2 = b) = \sum_a P(X_1 = a, X_2 = b)$$

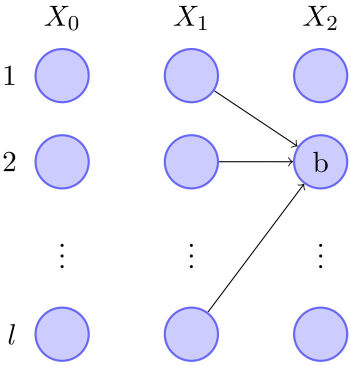
- The above sum essentially captures all the paths of reaching $X_2 = b$ irrespective of the value of X_1

$$P(X_2 = b) = \sum_a P(X_1 = a, X_2 = b)$$



- Let us consider $P(X_2 = b)$
$$P(X_2 = b) = \sum_a P(X_1 = a, X_2 = b)$$
- The above sum essentially captures all the paths of reaching $X_2 = b$ irrespective of the value of X_1

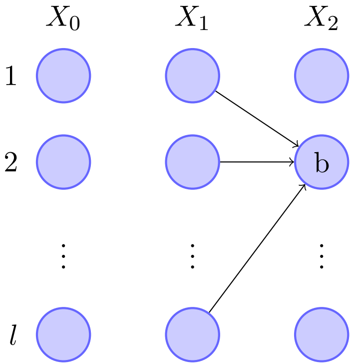
$$\begin{aligned} P(X_2 = b) &= \sum_a P(X_1 = a, X_2 = b) \\ &= \sum_a P(X_1 = a) P(X_2 = b | X_1 = a) \end{aligned}$$



- Let us consider $P(X_2 = b)$
$$P(X_2 = b) = \sum_a P(X_1 = a, X_2 = b)$$
- The above sum essentially captures all the paths of reaching $X_2 = b$ irrespective of the value of X_1

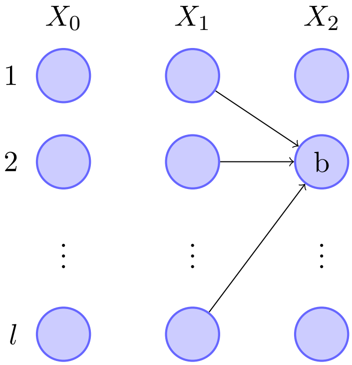
$$\begin{aligned} P(X_2 = b) &= \sum_a P(X_1 = a, X_2 = b) \\ &= \sum_a P(X_1 = a)P(X_2 = b|X_1 = a) \\ &= \sum_a \mu_a^1 T_{ab} \end{aligned}$$

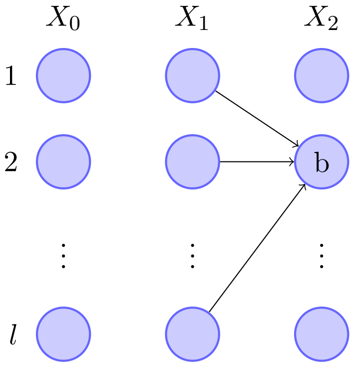
- Once again we can write $P(X_2)$ compactly as



- Once again we can write $P(X_2)$ compactly as

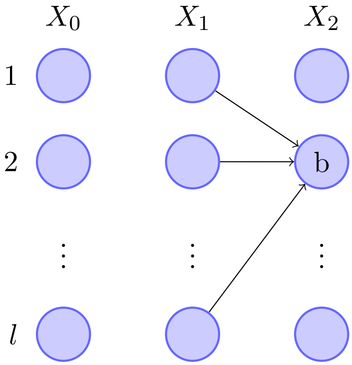
$$P(X_2) = \mu^1 T$$





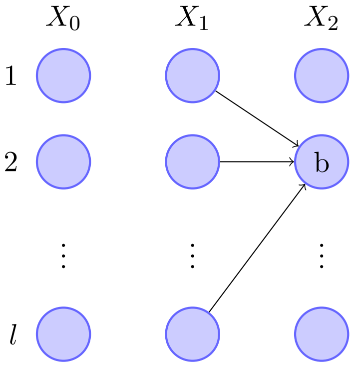
- Once again we can write $P(X_2)$ compactly as

$$P(X_2) = \mu^1 T = (\mu^0 T)T$$



- Once again we can write $P(X_2)$ compactly as

$$P(X_2) = \mu^1 T = (\mu^0 T)T = \mu^0 T^2$$

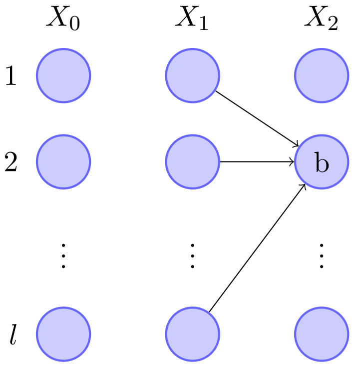


- Once again we can write $P(X_2)$ compactly as

$$P(X_2) = \mu^1 T = (\mu^0 T)T = \mu^0 T^2$$

- In general,

$$P(X_k) =$$

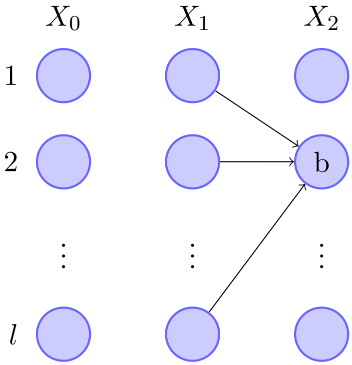


- Once again we can write $P(X_2)$ compactly as

$$P(X_2) = \mu^1 T = (\mu^0 T)T = \mu^0 T^2$$

- In general,

$$P(X_k) = \mu^0 T^k$$



- Once again we can write $P(X_2)$ compactly as

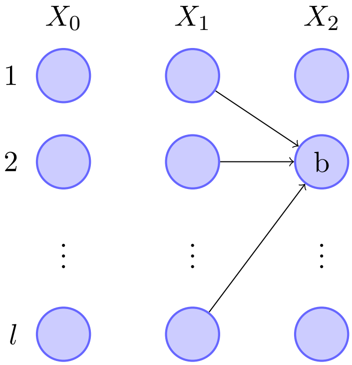
$$P(X_2) = \mu^1 T = (\mu^0 T)T = \mu^0 T^2$$

- In general,

$$P(X_k) = \mu^0 T^k$$

- Thus the distribution at any time step can be computed by finding the appropriate element from the following series

$$\mu^0 T^1, \mu^0 T^2, \mu^0 T^3, \ldots, \mu^0 T^k, \ldots$$



- Once again we can write $P(X_2)$ compactly as

$$P(X_2) = \mu^1 T = (\mu^0 T)T = \mu^0 T^2$$

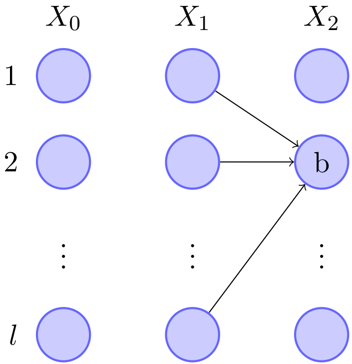
- In general,

$$P(X_k) = \mu^0 T^k$$

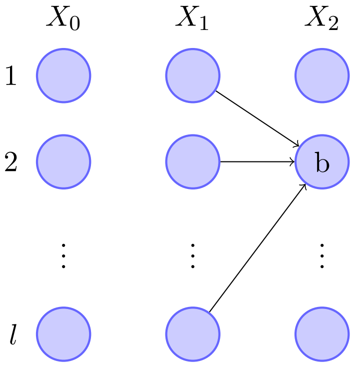
- Thus the distribution at any time step can be computed by finding the appropriate element from the following series

$$\mu^0 T^1, \mu^0 T^2, \mu^0 T^3, \ldots, \mu^0 T^k, \ldots$$

- Note that this is still computationally expensive because it involves a product of $\mu^0(2^n)$ and $T^k(2^n \times 2^n)$ (but later on we will see that we do not need this full product)

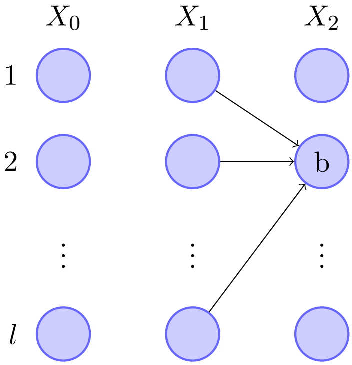


- If at a certain time step t, μ^t reaches a distribution π such that $\pi T = \pi$



- If at a certain time step t, μ^t reaches a distribution π such that $\pi T = \pi$
- Then for all subsequent time steps

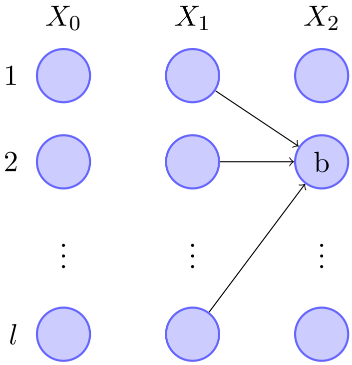
$$\mu^j = \pi (j \geq t)$$



- If at a certain time step t, μ^t reaches a distribution π such that $\pi T = \pi$
- Then for all subsequent time steps

$$\mu^j = \pi (j \geq t)$$

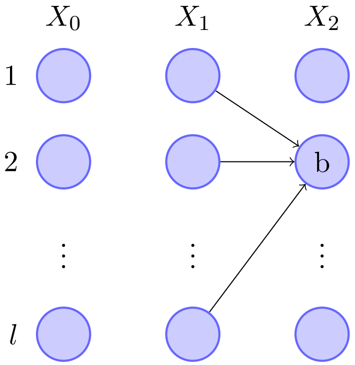
- π is then called the stationary distribution of the Markov chain



- If at a certain time step t, μ^t reaches a distribution π such that $\pi T = \pi$
- Then for all subsequent time steps

$$\mu^j = \pi (j \geq t)$$

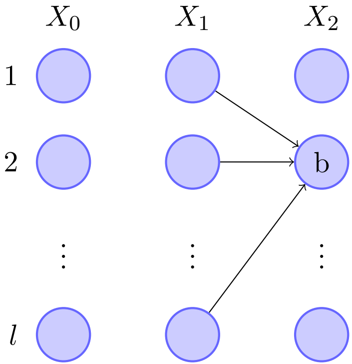
- π is then called the stationary distribution of the Markov chain
- $X_t, X_{t+1}, X_{t+2}, \ldots$ will all follow the same distribution π



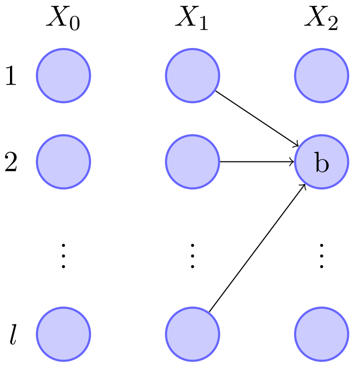
- If at a certain time step t, μ^t reaches a distribution π such that $\pi T = \pi$
- Then for all subsequent time steps

$$\mu^j = \pi (j \geq t)$$

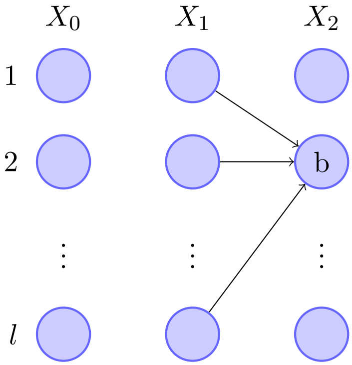
- π is then called the stationary distribution of the Markov chain
- $X_t, X_{t+1}, X_{t+2}, \ldots$ will all follow the same distribution π
- In other words, if we have $X_t = x_t, X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}$ and so on then we can think of x_t, x_{t+1}, x_{t+2} as samples drawn from the same distribution π (this is a crucial property and we will return back to it soon)



- **Important:** If we run a Markov Chain for a large number of time steps then after a point we start getting samples $x_t, x_{t+1}, x_{t+2}, \ldots$ which are essentially being drawn from the stationary distribution (**Spoiler Alert:** one of our goals was to draw samples from a very complex distribution)

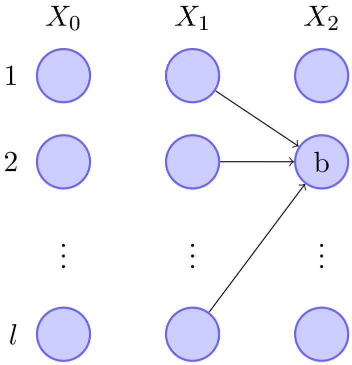


- **Important:** If we run a Markov Chain for a large number of time steps then after a point we start getting samples $x_t, x_{t+1}, x_{t+2}, \ldots$ which are essentially being drawn from the stationary distribution (**Spoiler Alert:** one of our goals was to draw samples from a very complex distribution)
- What do we mean by run a Markov Chain for a large number of time steps ?

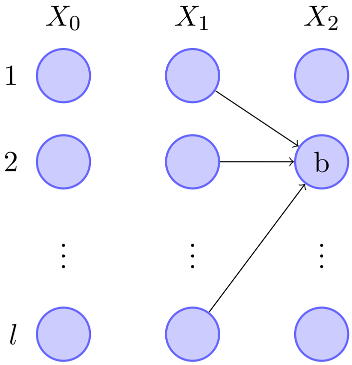


- **Important:** If we run a Markov Chain for a large number of time steps then after a point we start getting samples $x_t, x_{t+1}, x_{t+2}, \ldots$ which are essentially being drawn from the stationary distribution (**Spoiler Alert:** one of our goals was to draw samples from a very complex distribution)
- What do we mean by run a Markov Chain for a large number of time steps ?
- It means we start drawing a sample $X_0 \sim \mu^0$ and then continue drawing samples

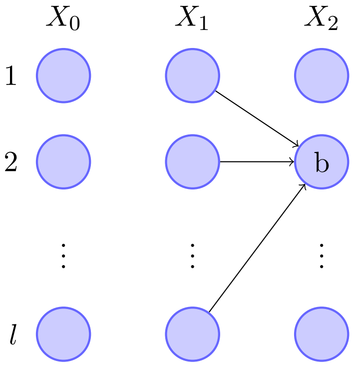
$$X_1 \sim \mu^0 T,\ \ X_2 \sim \mu^0 T^2,\ \ X_3 \sim \mu^0 T^3, \ldots,$$
$$\ldots, X_t \sim \pi,\ \ X_{t+1} \sim \pi,\ \ X_{t+2} \sim \pi \ldots$$



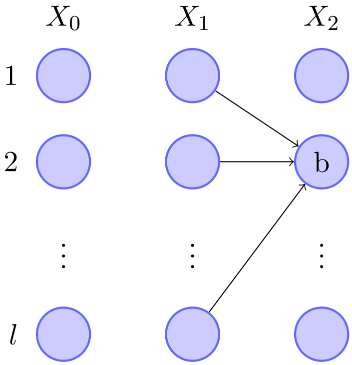
- Is it always easy to draw these samples?



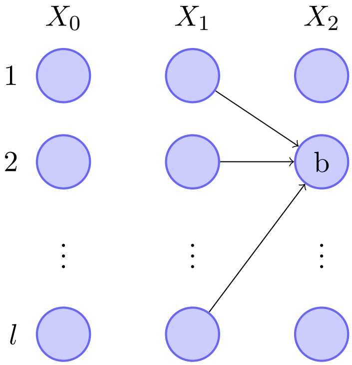
- Is it always easy to draw these samples? No



- Is it always easy to draw these samples? No
- $|\mu^k| = 2^n$ which means that we need to compute the probability of each of the possible 2^n values that X^k can take



- Is it always easy to draw these samples? No
- $|\mu^k| = 2^n$ which means that we need to compute the probability of each of the possible 2^n values that X^k can take
- In other words the joint distribution μ^k has 2^n parameters which is prohibitively large



- Is it always easy to draw these samples? No
- $|\mu^k| = 2^n$ which means that we need to compute the probability of each of the possible 2^n values that X^k can take
- In other words the joint distribution μ^k has 2^n parameters which is prohibitively large
- I wonder what can I do to reduce the number of parameters in a joint distribution (I hope you already know what to do but we will return back to it later)

The story so far...

- We have seen what a discrete space, discrete time, time homogeneous Markov Chain is

The story so far...

- We have seen what a discrete space, discrete time, time homogeneous Markov Chain is
- We have also defined μ^0 (initial distribution), T (transition matrix) and π (stationary distribution)

The story so far...

- We have seen what a discrete space, discrete time, time homogeneous Markov Chain is
- We have also defined μ^0 (initial distribution), T (transition matrix) and π (stationary distribution)
- So far so good! But why do we care about Markov Chains and their properties?

The story so far...

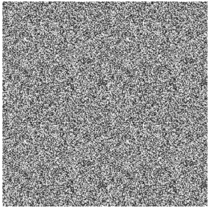
- We have seen what a discrete space, discrete time, time homogeneous Markov Chain is
- We have also defined μ^0 (initial distribution), T (transition matrix) and π (stationary distribution)
- So far so good! But why do we care about Markov Chains and their properties?
- How does this discussion tie back to our goals?

The story so far...

- We have seen what a discrete space, discrete time, time homogeneous Markov Chain is
- We have also defined μ^0 (initial distribution), T (transition matrix) and π (stationary distribution)
- So far so good! But why do we care about Markov Chains and their properties?
- How does this discussion tie back to our goals?
- We will first see an intuitive explanation for how all this ties back to our goals and then get into a more formal discussion

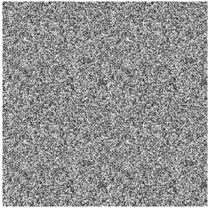
Module 20.2 : Why do we care about Markov Chains?

- Recall our goals



$X \in R^{1024}$

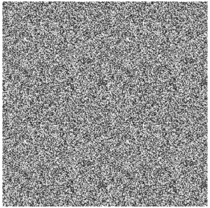
- Recall our goals
- **Goal 1**: Sample from $P(X)$



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

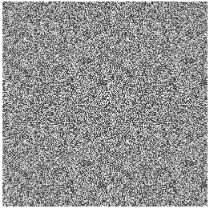
- Recall our goals
- **Goal 1**: Sample from $P(X)$
- **Goal 2**: Compute $\mathbb{E}_{P(X)}f(X)$



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

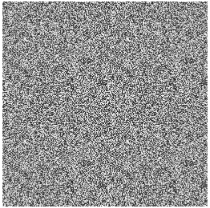
- Recall our goals
- **Goal 1**: Sample from $P(X)$
- **Goal 2**: Compute $\mathbb{E}_{P(X)}f(X)$
- Now suppose we set up a Markov Chain $X_1, X_2, \ldots$ such that



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

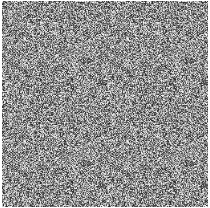
- Recall our goals
- **Goal 1**: Sample from $P(X)$
- **Goal 2**: Compute $\mathbb{E}_{P(X)}f(X)$
- Now suppose we set up a Markov Chain $X_1, X_2, \ldots$ such that
  - It is **easy to draw samples** from this chain and



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

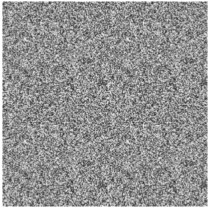
- Recall our goals
- **Goal 1**: Sample from $P(X)$
- **Goal 2**: Compute $\mathbb{E}_{P(X)} f(X)$
- Now suppose we set up a Markov Chain $X_1, X_2, \ldots$ such that
  - It is **easy to draw samples** from this chain and
  - This Markov Chain's **stationary distribution is** $P(X)$



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

- Recall our goals
- **Goal 1**: Sample from $P(X)$
- **Goal 2**: Compute $\mathbb{E}_{P(X)}f(X)$
- Now suppose we set up a Markov Chain $X_1, X_2, \ldots$ such that
  - It is **easy to draw samples** from this chain and
  - This Markov Chain's **stationary distribution is** $P(X)$
- Then it would mean that if we run the Markov Chain for long enough, we will start getting samples from $P(X)$



$X \in R^{1024}$

$\mathbb{E}_{P(X)}[f(X)]$

- Recall our goals
- **Goal 1**: Sample from $P(X)$
- **Goal 2**: Compute $\mathbb{E}_{P(X)} f(X)$
- Now suppose we set up a Markov Chain $X_1, X_2, \ldots$ such that
  - It is **easy to draw samples** from this chain and
  - This Markov Chain's **stationary distribution is $P(X)$**
- Then it would mean that if we run the Markov Chain for long enough, we will start getting samples from $P(X)$
- And once we have a large number of such samples we can empirically estimate $\mathbb{E}_{P(X)} f(X)$ as

$$\frac{1}{n} \sum_{i=l}^{l+n} f(X_i)$$

- We will now get into a formal discussion to concretize the above intuition

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[t\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \quad \text{where } X \in \mathscr{X} \text{ and } X \sim \pi$$

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[t\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \quad \text{where } X \in \mathscr{X} \text{ and } X \sim \pi$$

for any function $f : \mathscr{X} \to R$

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[t\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \quad \text{where } X \in \mathscr{X} \text{ and } X \sim \pi$$

for any function $f : \mathscr{X} \to R$

If, further the Markov Chain is aperiodic then $P(X_t = x_t | X_0 = x_0) \to \pi(X)$ as $t \to \infty \;\; \forall x, x_0 \in \mathscr{X}$

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[t\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \quad \text{where } X \in \mathscr{X} \text{ and } X \sim \pi$$

for any function $f : \mathscr{X} \to R$

If, further the Markov Chain is aperiodic then $P(X_t = x_t | X_0 = x_0) \to \pi(X)$ as $t \to \infty \;\; \forall x, x_0 \in \mathscr{X}$

- So Part A of the theorem essentially tells us that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then using samples from this chain we can compute $E_\pi[f(X)]$ (which we know is otherwise intractable)

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[t\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \quad \text{where } X \in \mathscr{X} \text{ and } X \sim \pi$$

for any function $f : \mathscr{X} \to R$

If, further the Markov Chain is aperiodic then $P(X_t = x_t|X_0 = x_0) \to \pi(X)$ as $t \to \infty \ \forall x, x_0 \in \mathscr{X}$

- So Part A of the theorem essentially tells us that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then using samples from this chain we can compute $E_\pi[f(X)]$ (which we know is otherwise intractable)
- Similarly Part B of the theorem says that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then we can essentially get samples as if they were drawn from $\pi(X)$ (which was otherwise intractable)

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[t\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \quad \text{where } X \in \mathscr{X} \text{ and } X \sim \pi$$

for any function $f : \mathscr{X} \to R$

If, further the Markov Chain is aperiodic then $P(X_t = x_t | X_0 = x_0) \to \pi(X)$ as $t \to \infty \;\; \forall x, x_0 \in \mathscr{X}$

- So Part A of the theorem essentially tells us that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then using samples from this chain we can compute $E_\pi[f(X)]$ (which we know is otherwise intractable)
- Similarly Part B of the theorem says that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then we can essentially get samples as if they were drawn from $\pi(X)$ (which was otherwise intractable)
- Of course Part A and Part B are related!

Theorem: If $X_0, X_1, \ldots, X_t$ is an irreducible time homogeneous discrete Markov Chain with stationary distribution π, then

$$\frac{1}{t}\sum_{i=1}^{t} f(X_i) \xrightarrow[t\to\infty]{\text{converges almost surely}} E_\pi[f(X)], \quad \text{where } X \in \mathscr{X} \text{ and } X \sim \pi$$

for any function $f : \mathscr{X} \to R$

If, further the Markov Chain is aperiodic then $P(X_t = x_t | X_0 = x_0) \to \pi(X)$ as $t \to \infty \;\; \forall x, x_0 \in \mathscr{X}$

- So Part A of the theorem essentially tells us that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then using samples from this chain we can compute $E_\pi[f(X)]$ (which we know is otherwise intractable)
- Similarly Part B of the theorem says that if we can set up the chain $X_0, X_1, \ldots, X_t$ such that it is tractable then we can essentially get samples as if they were drawn from $\pi(X)$ (which was otherwise intractable)
- Of course Part A and Part B are related!
- Further note that it doesn't matter what the initial state was (the theorem holds for $\forall x, x_0 \in \mathscr{X}$)

So our task is cut out now

- Define what our Markov Chain is?

So our task is cut out now

- Define what our Markov Chain is?
- Define the transition matrix T for our Markov Chain

So our task is cut out now

- Define what our Markov Chain is?
- Define the transition matrix T for our Markov Chain
- Show how it is easy to sample from this chain

So our task is cut out now

- Define what our Markov Chain is?
- Define the transition matrix T for our Markov Chain
- Show how it is easy to sample from this chain
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about)

So our task is cut out now

- Define what our Markov Chain is?
- Define the transition matrix T for our Markov Chain
- Show how it is easy to sample from this chain
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about)
- Show that the chain is irreducible and aperiodic (because the theorem only holds for such chains)

So our task is cut out now

- Define what our Markov Chain is?
- Define the transition matrix T for our Markov Chain
- Show how it is easy to sample from this chain
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about)
- Show that the chain is irreducible and aperiodic (because the theorem only holds for such chains)
- For ease of notation instead of $X = V_1, V_2, V_m, \ldots, H_1, H_2, \ldots, H_n$, we will use $X = X_1, X_2, \ldots, X_{n+m}$

Module 20.3 : Setting up a Markov Chain for RBMs

V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}

- We begin by defining our Markov Chain

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1								
2								

- We begin by defining our Markov Chain
- Recall that $X = \{V, H\} \in \{0,1\}^{n+m}$, so at time step 0 we create a random vector $X \in \{0,1\}^{n+m}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2								

- We begin by defining our Markov Chain
- Recall that $X = \{V, H\} \in \{0, 1\}^{n+m}$, so at time step 0 we create a random vector $X \in \{0, 1\}^{n+m}$
- At time-step 1, we transition to a new value of X

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2								

- We begin by defining our Markov Chain
- Recall that $X = \{V, H\} \in \{0, 1\}^{n+m}$, so at time step 0 we create a random vector $X \in \{0, 1\}^{n+m}$
- At time-step 1, we transition to a new value of X
- What does this mean? How do we do this transition? Let us see

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1								
2								
3								
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1								
2								
3								
4								

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1								
2								
3								
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1								
2								
3								
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2								
3								
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2								
3								
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

- Repeat the above process for many many time steps (each time step corresponds to 1 step of the chain)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3								
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

- Repeat the above process for many many time steps (each time step corresponds to 1 step of the chain)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3								
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

- Repeat the above process for many many time steps (each time step corresponds to 1 step of the chain)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

- Repeat the above process for many many time steps (each time step corresponds to 1 step of the chain)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4								

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

- Repeat the above process for many many time steps (each time step corresponds to 1 step of the chain)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

- Repeat the above process for many many time steps (each time step corresponds to 1 step of the chain)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- We need to transition from a state $X = x \in \{0,1\}^{n+m}$ to $y \in \{0,1\}^{n+m}$
- This is how we will do it
- Sample a value $i \in \{1 \text{ to } n+m\}$ using a distribution $q(i)$ (say, uniform distribution)
- Fix the value of all variables except X_i
- Sample a new value for X_i (could be a V or a H) using the following conditional distribution

$$P(X_i = y_i | X_{-i} = x_{-i})$$

- Repeat the above process for many many time steps (each time step corresponds to 1 step of the chain)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- What are we doing here? How is this related to our goals?

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- What are we doing here? How is this related to our goals?
- More specifically, we have defined a Markov Chain, but where is our Transition Matrix T?

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- What are we doing here? How is this related to our goals?
- More specifically, we have defined a Markov Chain, but where is our Transition Matrix T?
- How is it easy to create this chain (or creating samples $x_0, x_1, ...x_N$) ?

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- What are we doing here? How is this related to our goals?
- More specifically, we have defined a Markov Chain, but where is our Transition Matrix T?
- How is it easy to create this chain (or creating samples $x_0, x_1, ...x_N$) ?
- How do we show that the stationary distribution is $P(X)$ (where $X = V, H$) [We haven't even defined T, then how can we talk about the stationary distribution for T] ?

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- What are we doing here? How is this related to our goals?
- More specifically, we have defined a Markov Chain, but where is our Transition Matrix T?
- How is it easy to create this chain (or creating samples $x_0, x_1, ...x_N$) ?
- How do we show that the stationary distribution is $P(X)$ (where $X = V, H$) [We haven't even defined T, then how can we talk about the stationary distribution for T] ?
- Let us answer these questions one by one

- First, let us talk about the transition matrix

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- First, let us talk about the transition matrix
- We have actually defined T although we did not explicitly mention it

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- First, let us talk about the transition matrix
- We have actually defined T although we did not explicitly mention it
- What would T contain ?

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- First, let us talk about the transition matrix
- We have actually defined T although we did not explicitly mention it
- What would T contain ? The probability of transitioning from any state x to any state y

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- First, let us talk about the transition matrix
- We have actually defined T although we did not explicitly mention it
- What would T contain ? The probability of transitioning from any state x to any state y
- So $T \in R^{2^{m+n} \times 2^{m+n}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- First, let us talk about the transition matrix
- We have actually defined T although we did not explicitly mention it
- What would T contain ? The probability of transitioning from any state x to any state y
- So $T \in R^{2^{m+n} \times 2^{m+n}}$ (when did we define such a matrix?)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- First, let us talk about the transition matrix
- We have actually defined T although we did not explicitly mention it
- What would T contain ? The probability of transitioning from any state x to any state y
- So $T \in R^{2^{m+n} \times 2^{m+n}}$ (when did we define such a matrix?)
- Actually, we defined a very simple T which allowed only certain types of transitions

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- First, let us talk about the transition matrix
- We have actually defined T although we did not explicitly mention it
- What would T contain ? The probability of transitioning from any state x to any state y
- So $T \in R^{2^{m+n} \times 2^{m+n}}$ (when did we define such a matrix?)
- Actually, we defined a very simple T which allowed only certain types of transitions
- In particular, under this T, transitioning from a state x to a state y was possible only if x and y differ in the value of only one of the $n + m$ variables

- More formally, we defined T such that

- More formally, we defined T such that

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(y_i|x_{-i}), & \text{if } \exists i \in \mathbf{X} \text{ so that } \forall v \in \mathbf{X} \text{ with } v \neq i, x_v = y_v \\ 0, & \text{otherwise} \end{cases}$$

- More formally, we defined T such that

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(y_i|x_{-i}), & \text{if } \exists i \in \mathbf{X} \text{ so that } \forall v \in \mathbf{X} \text{ with } v \neq i, x_v = y_v \\ 0, & \text{otherwise} \end{cases}$$

- where $q(i)$ is the probability that X_i is the random variable whose value transitions while the value of X_{-i} remains the same

- More formally, we defined T such that

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(y_i|x_{-i}), \text{ if } \exists i \in \mathbf{X} \text{ so that } \forall v \in \mathbf{X} \text{ with } v \neq i, x_v = y_v \\ 0, \text{ otherwise} \end{cases}$$

- where $q(i)$ is the probability that X_i is the random variable whose value transitions while the value of X_{-i} remains the same
- The second term $P(X_i = y_i|\mathbf{X}_{-i})$ essentially tells us that given the value of the remaining random variable what is the probability of X_i taking on a certain value

- More formally, we defined T such that

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(y_i|x_{-i}), & \text{if } \exists i \in \mathbf{X} \text{ so that } \forall v \in \mathbf{X} \text{ with } v \neq i, x_v = y_v \\ 0, & \text{otherwise} \end{cases}$$

- where $q(i)$ is the probability that X_i is the random variable whose value transitions while the value of X_{-i} remains the same
- The second term $P(X_i = y_i|\mathbf{X}_{-i})$ essentially tells us that given the value of the remaining random variable what is the probability of X_i taking on a certain value
- With that we have answered the first question "What is the transition matrix T?" (It is a very sparse matrix allowing only certain transitions)

- We now look at the second question :

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- We now look at the second question : How is it easy to create this chain (or creating samples $x_0, x_1, ...x_l$)?

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- We now look at the second question : How is it easy to create this chain (or creating samples $x_0, x_1, ...x_l$)?
- At each step we are changing only one of the $n + m$ random variables using the following probability

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- We now look at the second question : How is it easy to create this chain (or creating samples $x_0, x_1, ...x_l$)?
- At each step we are changing only one of the $n + m$ random variables using the following probability

$$P(X_i = y_i | X_{-i} = x_{-i}) = \frac{P(X)}{P(X_{-i})}$$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- We now look at the second question : How is it easy to create this chain (or creating samples $x_0, x_1, ...x_l$)?
- At each step we are changing only one of the $n + m$ random variables using the following probability

$$P(X_i = y_i | X_{-i} = x_{-i}) = \frac{P(X)}{P(X_{-i})}$$

- But how is computing this probability easy? Doesn't the joint distribution on LHS also have 2^{n+m} parameters ?

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- We now look at the second question : How is it easy to create this chain (or creating samples $x_0, x_1, ...x_l$)?
- At each step we are changing only one of the $n + m$ random variables using the following probability

$$P(X_i = y_i | X_{-i} = x_{-i}) = \frac{P(X)}{P(X_{-i})}$$

- But how is computing this probability easy? Doesn't the joint distribution on LHS also have 2^{n+m} parameters ?
- Well, not really !

- Consider the case when $i <= m$ (i.e., we have decided to transition the value of one of the visible variables V_1 to V_m)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Consider the case when $i <= m$ (i.e., we have decided to transition the value of one of the visible variables V_1 to V_m)
- Then $P(X_i = y_i | X_{-i} = x_{-i})$ is essentially

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Consider the case when $i <= m$ (i.e., we have decided to transition the value of one of the visible variables V_1 to V_m)
- Then $P(X_i = y_i | X_{-i} = x_{-i})$ is essentially

$$P(V_i = y_i | V_{-i}, H) = P(V_i = y_i | H) \quad = \begin{cases} z, \text{if } y_i = 1 \\ 1 - z, \text{if } y_i = 0 \end{cases}$$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Consider the case when $i <= m$ (i.e., we have decided to transition the value of one of the visible variables V_1 to V_m)
- Then $P(X_i = y_i | X_{-i} = x_{-i})$ is essentially

$$P(V_i = y_i | V_{-i}, H) = P(V_i = y_i | H) \quad = \begin{cases} z, \text{if } y_i = 1 \\ 1 - z, \text{if } y_i = 0 \end{cases}$$

where $z = \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i)$

- The above probability is very easy to compute (just a sigmoid function)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Consider the case when $i <= m$ (i.e., we have decided to transition the value of one of the visible variables V_1 to V_m)
- Then $P(X_i = y_i | X_{-i} = x_{-i})$ is essentially

$$P(V_i = y_i | V_{-i}, H) = P(V_i = y_i | H) \quad = \begin{cases} z, \text{if } y_i = 1 \\ 1 - z, \text{if } y_i = 0 \end{cases}$$

where $z = \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i)$

- The above probability is very easy to compute (just a sigmoid function)
- Once you compute the above probability, with probability z you will set the value of V_i to 1 and with probability $1 - z$ you will set it to 0

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- So essentially at every time step you sample a i from a uniform distribution (q_i)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- So essentially at every time step you sample a i from a uniform distribution (q_i)
- And then sample a value of $V_i \in \{0, 1\}$ using the distribution $Bernoulli(z)$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- So essentially at every time step you sample a i from a uniform distribution (q_i)
- And then sample a value of $V_i \in \{0, 1\}$ using the distribution $Bernoulli(z)$
- Both these computations are easy

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- So essentially at every time step you sample a i from a uniform distribution (q_i)
- And then sample a value of $V_i \in \{0, 1\}$ using the distribution $Bernoulli(z)$
- Both these computations are easy
- Hence it is easy to create this chain starting from any x_0

- Okay, finally let's look at the third question: How do we show that the stationary distribution is $P(X)$ (where $X = V, H$)

- Okay, finally let's look at the third question: How do we show that the stationary distribution is $P(X)$ (where $X = V, H$)
- To prove this we will refer to the following Theorem:

- Okay, finally let's look at the third question: How do we show that the stationary distribution is $P(X)$ (where $X = V, H$)
- To prove this we will refer to the following Theorem:

Detailed Balance Condition

To show that a distribution π is a stationary distribution for a Markov Chain described by the transition probabilities p_{xy}, $x, y \in \Omega$, it is sufficient to show that $\forall x, y \in \Omega$, the following condition holds:

$$\pi(x)p_{xy} = \pi(x)p_{yx}$$

- Okay, finally let's look at the third question: How do we show that the stationary distribution is $P(X)$ (where $X = V, H$)
- To prove this we will refer to the following Theorem:

Detailed Balance Condition

To show that a distribution π is a stationary distribution for a Markov Chain described by the transition probabilities p_{xy}, $x, y \in \Omega$, it is sufficient to show that $\forall x, y \in \Omega$, the following condition holds:

$$\pi(x)p_{xy} = \pi(x)p_{yx}$$

- Let us revisit what p_{xy} is and what π is

- Recall that $p_{\boldsymbol{xy}}$ is given by

- Recall that $p_{\boldsymbol{xy}}$ is given by

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(X_i = y_i|\mathbf{X}_{-i}\boldsymbol{x}_{-i}), & \text{if } \exists i \in \{1, 2, \dots, n+m\} \text{ such that } \forall j \in \{1, 2, \dots, n+m\}i \\ 0, & \text{otherwise} \end{cases}$$

- Recall that $p_{\boldsymbol{xy}}$ is given by

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(X_i = y_i|\mathbf{X}_{-i}\boldsymbol{x}_{-i}), & \text{if } \exists i \in \{1,2,\dots,n+m\} \text{ such that } \forall j \in \{1,2,\dots,n+m\}\text{i} \\ 0, & \text{otherwise} \end{cases}$$

- For consistency of notation we will denote $P(X)$ i.e., $P(V,H)$ as $\pi(X)$

- Recall that $p_{\boldsymbol{xy}}$ is given by

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(X_i = y_i|\mathbf{X}_{-i}\boldsymbol{x}_{-i}), & \text{if } \exists i \in \{1,2,\ldots,n+m\} \text{ such that } \forall j \in \{1,2,\ldots,n+m\}\text{i} \\ 0, & \text{otherwise} \end{cases}$$

- For consistency of notation we will denote $P(X)$ i.e., $P(V,H)$ as $\pi(X)$
- Further, as shorthand we will refer to $\pi(X = \boldsymbol{x})$ as $\pi(\boldsymbol{x})$

- Recall that $p_{\boldsymbol{xy}}$ is given by

$$p_{\mathbf{xy}} = \begin{cases} q(i)P(X_i = y_i|\mathbf{X}_{-i}\boldsymbol{x}_{-i}), & \text{if } \exists i \in \{1,2,\dots,n+m\} \text{ such that } \forall j \in \{1,2,\dots,n+m\}\text{i} \\ 0, & \text{otherwise} \end{cases}$$

- For consistency of notation we will denote $P(X)$ i.e., $P(V,H)$ as $\pi(X)$
- Further, as shorthand we will refer to $\pi(X=\boldsymbol{x})$ as $\pi(\boldsymbol{x})$
- Thus, to prove that $P(X)$, i.e., $\pi(X)$ is the stationary distribution for our Markov Chain we need to prove that

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}} \quad \forall \boldsymbol{x},\boldsymbol{y} \in \{0,1\}^{m+n}$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 1:**

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- In this case, by definition

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- In this case, by definition

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{x}) * 0 = 0$$
$$\pi(\boldsymbol{y})p_{\boldsymbol{yx}} = \pi(\boldsymbol{y}) * 0 = 0$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- In this case, by definition

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{x}) * 0 = 0$$
$$\pi(\boldsymbol{y})p_{\boldsymbol{yx}} = \pi(\boldsymbol{y}) * 0 = 0$$

- Hence the detailed balance condition holds trivially

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 2:**

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)
- In this case, by definition

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)
- In this case, by definition

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{x})p_{\boldsymbol{xx}}$$
$$\pi(\boldsymbol{y})p_{\boldsymbol{yx}} = \pi(\boldsymbol{x})p_{\boldsymbol{xx}}$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)
- In this case, by definition

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{x})p_{\boldsymbol{xx}}$$
$$\pi(\boldsymbol{y})p_{\boldsymbol{yx}} = \pi(\boldsymbol{x})p_{\boldsymbol{xx}}$$

- Hence the detailed balance condition holds trivially

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:**

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{x})q(i)\pi(y_i|\boldsymbol{x}_{-i})$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

$$\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{x})q(i)\pi(y_i|\boldsymbol{x}_{-i})$$
$$= q(i)\pi(x_i, \boldsymbol{x}_{-i})\frac{\pi(y_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})}$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

$$\begin{aligned}
\pi(\boldsymbol{x})p_{\boldsymbol{xy}} &= \pi(\boldsymbol{x})q(i)\pi(y_i|\boldsymbol{x}_{-i}) \\
&= q(i)\pi(x_i, \boldsymbol{x}_{-i})\frac{\pi(y_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\
&= \pi(y_i, \boldsymbol{x}_{-i})q(i)\frac{\pi(x_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})}
\end{aligned}$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

$$\begin{aligned}
\pi(\boldsymbol{x})p_{\boldsymbol{xy}} &= \pi(\boldsymbol{x})q(i)\pi(y_i|\boldsymbol{x}_{-i}) \\
&= q(i)\pi(x_i, \boldsymbol{x}_{-i})\frac{\pi(y_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\
&= \pi(y_i, \boldsymbol{x}_{-i})q(i)\frac{\pi(x_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\
&= \pi(\boldsymbol{y})q(i)\pi(x_i|x_{-i})
\end{aligned}$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

$$
\begin{aligned}
\pi(\boldsymbol{x})p_{\boldsymbol{xy}} &= \pi(\boldsymbol{x})q(i)\pi(y_i|\boldsymbol{x}_{-i}) \\
&= q(i)\pi(x_i, \boldsymbol{x}_{-i})\frac{\pi(y_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\
&= \pi(y_i, \boldsymbol{x}_{-i})q(i)\frac{\pi(x_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\
&= \pi(\boldsymbol{y})q(i)\pi(x_i|x_{-i}) \\
&= \pi(\boldsymbol{y})p_{\boldsymbol{yx}}
\end{aligned}
$$

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- There are 3 cases that we need to consider
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable
- In this case, by definition

$$
\begin{aligned}
\pi(\boldsymbol{x})p_{\boldsymbol{xy}} &= \pi(\boldsymbol{x})q(i)\pi(y_i|\boldsymbol{x}_{-i}) \\
&= q(i)\pi(x_i, \boldsymbol{x}_{-i})\frac{\pi(y_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\
&= \pi(y_i, \boldsymbol{x}_{-i})q(i)\frac{\pi(x_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} \\
&= \pi(\boldsymbol{y})q(i)\pi(x_i|x_{-i}) \\
&= \pi(\boldsymbol{y})p_{\boldsymbol{yx}}
\end{aligned}
$$

- Hence the detailed balance condition holds

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Thus we have proved that the detailed balance condition holds in all the 3 cases

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Thus we have proved that the detailed balance condition holds in all the 3 cases
- **Case 1:**

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Thus we have proved that the detailed balance condition holds in all the 3 cases
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Thus we have proved that the detailed balance condition holds in all the 3 cases
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- **Case 2:**

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Thus we have proved that the detailed balance condition holds in all the 3 cases
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Thus we have proved that the detailed balance condition holds in all the 3 cases
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)
- **Case 3:**

To prove: $\pi(\boldsymbol{x})p_{\boldsymbol{xy}} = \pi(\boldsymbol{y})p_{\boldsymbol{yx}}$

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Thus we have proved that the detailed balance condition holds in all the 3 cases
- **Case 1:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of more than one random variable
- **Case 2:** $\boldsymbol{x}$ and $\boldsymbol{y}$ are equal (i.e., they do not differ in the state of any random variable)
- **Case 3:** $\boldsymbol{x}$ and $\boldsymbol{y}$ differ in the state of only one random variable

So our task is cut out now

- Define what our Markov Chain is?
- Define the transition matrix T for our Markov Chain
- Show how it is easy to sample from this chain
- Show that the stationary distribution of this chain is the distribution $P(X)$ (*i.e.*, the distribution that we care about)
- Show that the chain is irreducible and aperiodic

So our task is cut out now

- Define what our Markov Chain is? **(done)**
- Define the transition matrix T for our Markov Chain
- Show how it is easy to sample from this chain
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about)
- Show that the chain is irreducible and aperiodic

So our task is cut out now

- Define what our Markov Chain is? **(done)**
- Define the transition matrix T for our Markov Chain **(done)**
- Show how it is easy to sample from this chain
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about)
- Show that the chain is irreducible and aperiodic

So our task is cut out now

- Define what our Markov Chain is? **(done)**
- Define the transition matrix T for our Markov Chain **(done)**
- Show how it is easy to sample from this chain **(done)**
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about)
- Show that the chain is irreducible and aperiodic

So our task is cut out now

- Define what our Markov Chain is? **(done)**
- Define the transition matrix T for our Markov Chain **(done)**
- Show how it is easy to sample from this chain **(done)**
- Show that the stationary distribution of this chain is the distribution $P(X)$ (*i.e.*, the distribution that we care about) **(done)**
- Show that the chain is irreducible and aperiodic

So our task is cut out now

- Define what our Markov Chain is? **(done)**
- Define the transition matrix T for our Markov Chain **(done)**
- Show how it is easy to sample from this chain **(done)**
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about) **(done)**
- Show that the chain is irreducible and aperiodic **(let us see)**

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A Markov chain is irreducible if one can get from any state in Ω to any other in a finite number of transitions or more formally

$$\forall i, j \in \Omega \quad \exists k > 0 \quad \text{with}$$
$$P(X^{(k)} = j | X^{(0)} = i) > 0$$

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A Markov chain is irreducible if one can get from any state in Ω to any other in a finite number of transitions or more formally

$$\forall i, j \in \Omega \quad \exists k > 0 \quad \text{with}$$
$$P(X^{(k)} = j | X^{(0)} = i) > 0$$

- Intuitively, we can see that our chain is irreducible

	V_1	V_2	$\ldots$	V_m	H_1	H_2	$\ldots$	H_n
	X_1	X_2	X_3		$\ldots$	$\ldots$		X_{n+m}
0	1	1	0		$\ldots$	$\ldots$		1
1	1	0	0		$\ldots$	$\ldots$		1
2	1	0	1		$\ldots$	$\ldots$		1
3	1	0	1		$\ldots$	$\ldots$		1
4	1	0	1		$\ldots$	$\ldots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A Markov chain is irreducible if one can get from any state in Ω to any other in a finite number of transitions or more formally

$$\forall i, j \in \Omega \quad \exists k > 0 \quad \text{with}$$
$$P(X^{(k)} = j | X^{(0)} = i) > 0$$

- Intuitively, we can see that our chain is irreducible
- For example, notice that we can reach from the state containing all 0's to all 1's after some finite time steps

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A Markov chain is irreducible if one can get from any state in Ω to any other in a finite number of transitions or more formally

$$\forall i, j \in \Omega \quad \exists k > 0 \quad \text{with}$$
$$P(X^{(k)} = j | X^{(0)} = i) > 0$$

- Intuitively, we can see that our chain is irreducible
- For example, notice that we can reach from the state containing all 0's to all 1's after some finite time steps
- We can prove this more formally but for now we will just rely on the intuition

- A chain is called aperiodic if $\forall i \in \Omega$ the greatest common divisor of $\{k | P(X^{(k)} = i | X^{(0)} = i) > 0 \quad \wedge \quad k \in N_0\}$ is 1

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A chain is called aperiodic if $\forall i \in \Omega$ the greatest common divisor of $\{k | P(X^{(k)} = i | X^{(0)} = i) > 0 \ \wedge \ k \in N_0\}$ is 1
- The set we have defined above contains all the timesteps at which we can reach state i starting from step i

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A chain is called aperiodic if $\forall i \in \Omega$ the greatest common divisor of $\{k | P(X^{(k)} = i | X^{(0)} = i) > 0 \quad \wedge \quad k \in N_0\}$ is 1
- The set we have defined above contains all the timesteps at which we can reach state i starting from step i
- Suppose the chain was periodic then this set would contain multiples of a certain number

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A chain is called aperiodic if $\forall i \in \Omega$ the greatest common divisor of $\{k | P(X^{(k)} = i | X^{(0)} = i) > 0 \quad \wedge \; k \in N_0\}$ is 1
- The set we have defined above contains all the timesteps at which we can reach state i starting from step i
- Suppose the chain was periodic then this set would contain multiples of a certain number
- For example, $\{3, 6, 9, 12, \dots\}$ and hence the greater common divisor would be 3 (and the Markov Chain would be periodic with a period of 3)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- A chain is called aperiodic if $\forall i \in \Omega$ the greatest common divisor of $\{k | P(X^{(k)} = i | X^{(0)} = i) > 0 \quad \wedge \quad k \in N_0\}$ is 1
- The set we have defined above contains all the timesteps at which we can reach state i starting from step i
- Suppose the chain was periodic then this set would contain multiples of a certain number
- For example, $\{3, 6, 9, 12, \dots\}$ and hence the greater common divisor would be 3 (and the Markov Chain would be periodic with a period of 3)
- However if the chain is not periodic then the set would contain arbitrary numbers and their GCD would just be 1 (hence the above definition)

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Again intuitively it should be clear that our chain is aperiodic

	V_1	V_2	$\dots$	V_m	H_1	H_2	$\dots$	H_n
	X_1	X_2	X_3		$\dots$	$\dots$		X_{n+m}
0	1	1	0		$\dots$	$\dots$		1
1	1	0	0		$\dots$	$\dots$		1
2	1	0	1		$\dots$	$\dots$		1
3	1	0	1		$\dots$	$\dots$		1
4	1	0	1		$\dots$	$\dots$		0
$\vdots$	$\vdots$							
$\vdots$	$\vdots$							

- Again intuitively it should be clear that our chain is aperiodic
- Once again, we can formally prove this but we will just rely on the intuition for now

So our task is cut out now

- Define what our Markov Chain is? **(done)**
- Define the transition matrix T for our Markov Chain **(done)**
- Show how it is easy to sample from this chain **(done)**
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about) **(done)**
- Show that the chain is irreducible and aperiodic

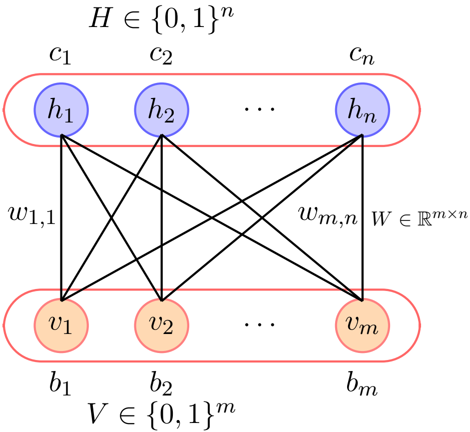
So our task is cut out now

- Define what our Markov Chain is? **(done)**
- Define the transition matrix T for our Markov Chain **(done)**
- Show how it is easy to sample from this chain **(done)**
- Show that the stationary distribution of this chain is the distribution $P(X)$ ($i.e.$, the distribution that we care about) **(done)**
- Show that the chain is irreducible and aperiodic **(done)**

Module 20.4 : Training RBMs using Gibbs Sampling

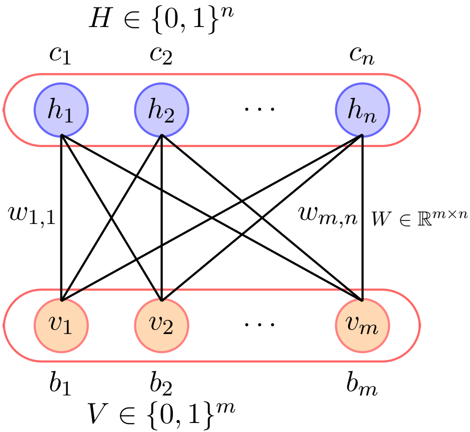
- Okay, so we are now ready to write the full algorithm for training RBMs using Gibbs Sampling

- Okay, so we are now ready to write the full algorithm for training RBMs using Gibbs Sampling
- We will first quickly revisit the expectations that we wanted to compute and write a simplified expression for them



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

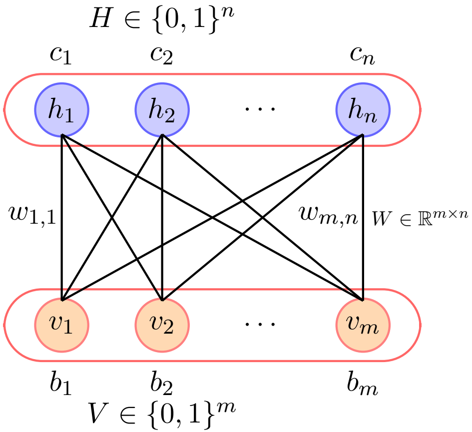
$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$



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

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

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

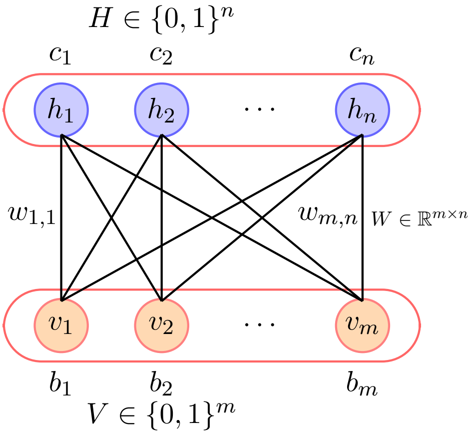


$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

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

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

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



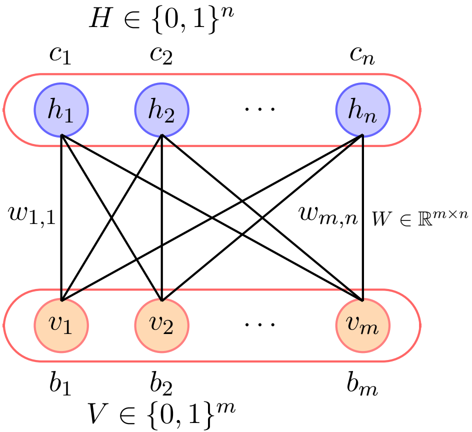
$$E(V,H) = -\sum_i \sum_j w_{ij} v_i h_j - \sum_i b_i v_i - \sum_j c_j h_j$$

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

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

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

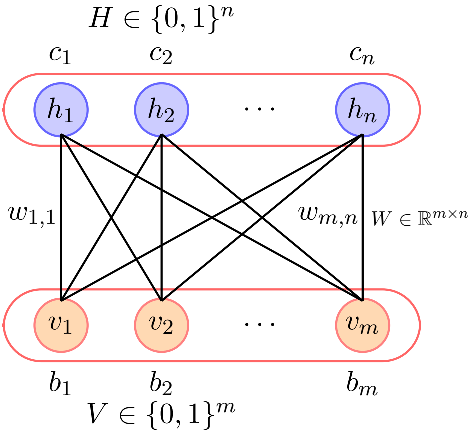
- We were interested in computing the partial derivative of the log likehood w.r.t. one of the parameters (w_{ij})



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

$$\begin{aligned}
&\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} \\
&= -\sum_H p(H|V) \frac{\partial E(V,H)}{\partial w_{ij}} + \sum_{V,H} p(V,H) \frac{\partial E(V,H)}{\partial w_{ij}} \\
&= \sum_H p(H|V) h_i v_j - \sum_{V,H} p(V,H) h_i v_j \\
&= \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]
\end{aligned}$$

- We were interested in computing the partial derivative of the log likehood w.r.t. one of the parameters (w_{ij})
- We saw that this partial derivative is actually the sum of two expectations



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

$$\begin{aligned}
&\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} \\
&= -\sum_{H} p(H|V) \frac{\partial E(V,H)}{\partial w_{ij}} + \sum_{V,H} p(V,H) \frac{\partial E(V,H)}{\partial w_{ij}} \\
&= \sum_{H} p(H|V) h_i v_j - \sum_{V,H} p(V,H) h_i v_j \\
&= \mathbb{E}_{p(H|V)}[v_i h_j] - \mathbb{E}_{p(V,H)}[v_i h_j]
\end{aligned}$$

- We were interested in computing the partial derivative of the log likehood w.r.t. one of the parameters (w_{ij})
- We saw that this partial derivative is actually the sum of two expectations
- We will now simplify the expression for these two expectations

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

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

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j$$

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

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} &= \mathbb{E}_{p(H|V)}[v_j h_i] - \mathbb{E}_{p(V,H)}[v_j h_i] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j
\end{aligned}
$$

We will first focus on $\sum_{h} p(h|v) h_i v_j$

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

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j$$

We will first focus on $\sum_{h} p(h|v) h_i v_j$

$$\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j = \sum_{h_i} \sum_{\boldsymbol{h}_{-i}} p(h_i|\boldsymbol{v}) p(\boldsymbol{h}_{-\boldsymbol{i}}|\boldsymbol{v}) h_i v_j$$

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

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j$$

We will first focus on $\sum_{h} p(h|v) h_i v_j$

$$\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j = \sum_{h_i} \sum_{\boldsymbol{h}_{-i}} p(h_i|\boldsymbol{v}) p(\boldsymbol{h}_{-\boldsymbol{i}}|\boldsymbol{v}) h_i v_j$$

$$= \sum_{h_i} p(h_i|\boldsymbol{v}) h_i v_j \sum_{\boldsymbol{h}_{-i}} p(\boldsymbol{h}_{-i}|\boldsymbol{v})$$

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

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j$$

We will first focus on $\sum_{h} p(h|v) h_i v_j$

$$\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j = \sum_{h_i} \sum_{\boldsymbol{h}_{-i}} p(h_i|\boldsymbol{v}) p(\boldsymbol{h_{-i}}|\boldsymbol{v}) h_i v_j$$

$$= \sum_{h_i} p(h_i|\boldsymbol{v}) h_i v_j \sum_{\boldsymbol{h}_{-i}} p(\boldsymbol{h}_{-i}|\boldsymbol{v})$$

$$= p(H_i = 1|\boldsymbol{v}) v_j$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \mathbb{E}_{p(H|V)}[v_j h_i] - \mathbb{E}_{p(V,H)}[v_j h_i]$$
$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j$$
$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j$$

We will first focus on $\sum_{h} p(h|v) h_i v_j$

$$\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j = \sum_{h_i} \sum_{\boldsymbol{h}_{-i}} p(h_i|\boldsymbol{v}) p(\boldsymbol{h_{-i}}|\boldsymbol{v}) h_i v_j$$
$$= \sum_{h_i} p(h_i|\boldsymbol{v}) h_i v_j \sum_{\boldsymbol{h}_{-i}} p(\boldsymbol{h}_{-i}|\boldsymbol{v})$$
$$= p(H_i = 1|\boldsymbol{v}) v_j$$
$$= \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) v_j$$

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

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i v_j$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j$$

We will first focus on $\sum_{h} p(h|v) h_i v_j$

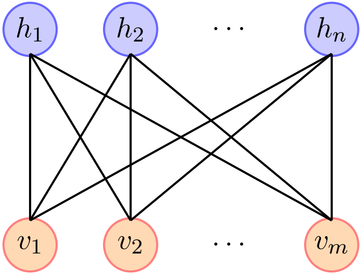
$$\sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i v_j = \sum_{h_i} \sum_{\boldsymbol{h}_{-i}} p(h_i|\boldsymbol{v}) p(\boldsymbol{h}_{-\boldsymbol{i}}|\boldsymbol{v}) h_i v_j$$

$$= \sum_{h_i} p(h_i|\boldsymbol{v}) h_i v_j \sum_{\boldsymbol{h}_{-i}} p(\boldsymbol{h}_{-i}|\boldsymbol{v})$$

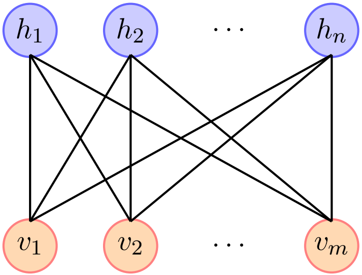
$$= p(H_i = 1|\boldsymbol{v}) v_j$$

$$= \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) v_j$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) v_j$$

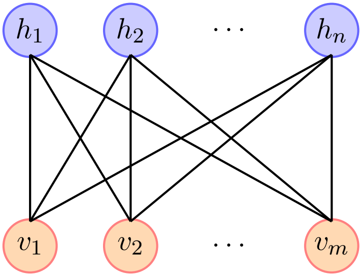


$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)v_j$$



$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)v_j$$

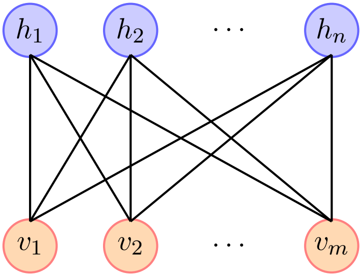
$$= \sigma(\boldsymbol{w}_i\boldsymbol{v} + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\boldsymbol{w}_i\boldsymbol{v} + c_i)v_j$$



$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)v_j$$

$$= \sigma(\boldsymbol{w}_i\boldsymbol{v} + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\boldsymbol{w}_i\boldsymbol{v} + c_i)v_j$$

$$\nabla_{\mathbf{W}}\mathscr{L}(\theta) = \sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T$$



$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) v_j$$

$$= \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

$$\nabla_{\mathbf{W}} \mathscr{L}(\theta) = \sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) \boldsymbol{v}^T - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) \boldsymbol{v}^T$$

$$= \sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) \boldsymbol{v}^T - \mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) \boldsymbol{v}^T]$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_j} = \mathbb{E}_{p(H|V)}[v_j] - \mathbb{E}_{p(V,H)}[v_j]$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_j} = \mathbb{E}_{p(H|V)}[v_j] - \mathbb{E}_{p(V,H)}[v_j]$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h})v_j$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial b_j} = \mathbb{E}_{p(H|V)}[v_j] - \mathbb{E}_{p(V,H)}[v_j]$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) v_j$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j$$

$$\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial b_j} &= \mathbb{E}_{p(H|V)}[v_j] - \mathbb{E}_{p(V,H)}[v_j] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) v_j \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j \\
&= v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial b_j} &= \mathbb{E}_{p(H|V)}[v_j] - \mathbb{E}_{p(V,H)}[v_j] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) v_j \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j \\
&= v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) \\
&= v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) v_j
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial b_j} &= \mathbb{E}_{p(H|V)}[v_j] - \mathbb{E}_{p(V,H)}[v_j] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h})v_j \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})v_j \\
&= v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) \\
&= v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})v_j \\
\nabla_{\boldsymbol{b}} \mathscr{L}(\theta) &= \boldsymbol{v} - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\boldsymbol{v}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial b_j} &= \mathbb{E}_{p(H|V)}[v_j] - \mathbb{E}_{p(V,H)}[v_j] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) v_j \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) v_j \\
&= v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) v_j \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) \\
&= v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) v_j \\
\nabla_{\boldsymbol{b}} \mathscr{L}(\theta) &= \boldsymbol{v} - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \boldsymbol{v} \\
&= \boldsymbol{v} - \mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}]
\end{aligned}
$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial c_i} = \mathbb{E}_{p(H|V)}[h_i] - \mathbb{E}_{p(V,H)}[h_i]$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial c_i} &= \mathbb{E}_{p(H|V)}[h_i] - \mathbb{E}_{p(V,H)}[h_i] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial c_i} &= \mathbb{E}_{p(H|V)}[h_i] - \mathbb{E}_{p(V,H)}[h_i] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h})h_i \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial c_i} &= \mathbb{E}_{p(H|V)}[h_i] - \mathbb{E}_{p(V,H)}[h_i] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h})h_i \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i \\
&= p(H_i = 1|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})p(H_i = 1|\boldsymbol{v})
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial c_i} &= \mathbb{E}_{p(H|V)}[h_i] - \mathbb{E}_{p(V,H)}[h_i] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h}) h_i \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}) h_i \\
&= p(H_i = 1|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) p(H_i = 1|\boldsymbol{v}) \\
&= \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i) - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\sum_{j=1}^{m} w_{ij} v_j + c_i)
\end{aligned}
$$

$$\frac{\partial \mathscr{L}(\theta)}{\partial c_i} = \mathbb{E}_{p(H|V)}[h_i] - \mathbb{E}_{p(V,H)}[h_i]$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h})h_i$$

$$= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i$$

$$= p(H_i = 1|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})p(H_i = 1|\boldsymbol{v})$$

$$= \sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i)$$

$$\nabla_{\boldsymbol{c}}\mathscr{L}(\theta) = \sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})$$

$$
\begin{aligned}
\frac{\partial \mathscr{L}(\theta)}{\partial c_i} &= \mathbb{E}_{p(H|V)}[h_i] - \mathbb{E}_{p(V,H)}[h_i] \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v},\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h})h_i \\
&= \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v})h_i \\
&= p(H_i = 1|\boldsymbol{v}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})p(H_i = 1|\boldsymbol{v}) \\
&= \sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} w_{ij}v_j + c_i) \\
\nabla_{\boldsymbol{c}}\mathscr{L}(\theta) &= \sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) \\
&= \sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) - \mathbb{E}_{\boldsymbol{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})]
\end{aligned}
$$

- Notice that all the 3 gradient expressions have an expectation term

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T]$$

- Notice that all the 3 gradient expressions have an expectation term

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T]$$

$$\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}]$$

- Notice that all the 3 gradient expressions have an expectation term

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T]$$

$$\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}]$$

$$\mathbb{E}_{\boldsymbol{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})]$$

- Notice that all the 3 gradient expressions have an expectation term

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T]$$

$$\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}]$$

$$\mathbb{E}_{\boldsymbol{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})]$$

- Notice that all the 3 gradient expressions have an expectation term
- These expectations are intractable.

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T]$$

$$\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}]$$

$$\mathbb{E}_{\boldsymbol{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})]$$

- Notice that all the 3 gradient expressions have an expectation term
- These expectations are intractable.
- Solution?

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T] \approx \frac{1}{k}\sum_{i=1}^{k}\sigma(\mathbf{W}\boldsymbol{v}^{(k)} + \boldsymbol{c})\boldsymbol{v}^{(k)T}$$

$$\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}]$$

$$\mathbb{E}_{\boldsymbol{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})]$$

- Notice that all the 3 gradient expressions have an expectation term
- These expectations are intractable.
- Solution? Estimation with the help of sampling

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T] \approx \frac{1}{k}\sum_{i=1}^{k} \sigma(\mathbf{W}\boldsymbol{v}^{(k)} + \boldsymbol{c})\boldsymbol{v}^{(k)T}$$

$$\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}] \approx \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{v}^{(k)}$$

$$\mathbb{E}_{\boldsymbol{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})]$$

- Notice that all the 3 gradient expressions have an expectation term
- These expectations are intractable.
- Solution? Estimation with the help of sampling
- Specifically, we will use Gibbs Sampling to estimate the expectation

$$\mathbb{E}_{\mathbf{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})\boldsymbol{v}^T] \approx \frac{1}{k}\sum_{i=1}^{k} \sigma(\mathbf{W}\boldsymbol{v}^{(k)} + \boldsymbol{c})\boldsymbol{v}^{(k)T}$$

$$\mathbb{E}_{\boldsymbol{v}}[\boldsymbol{v}] \approx \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{v}^{(k)}$$

$$\mathbb{E}_{\boldsymbol{v}}[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c})] \approx \frac{1}{k}\sum_{i=1}^{k} \sigma(\mathbf{W}\boldsymbol{v}^{(k)} + \boldsymbol{c})$$

- Notice that all the 3 gradient expressions have an expectation term
- These expectations are intractable.
- Solution? Estimation with the help of sampling
- Specifically, we will use Gibbs Sampling to estimate the expectation

Algorithm 0: RBM Training with Block Gibbs Sampling

Algorithm 0: RBM Training with Block Gibbs Sampling

Input:

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output:

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

  for $t = 0, ..., k, k+1, ..., k+r$ **do**

  end

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

  for $t = 0, ..., k, k+1, ..., k+r$ **do**

    for $i = 1, ..., n$ **do**

    end

  end

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

  for $t = 0, ..., k, k+1, ..., k+r$ **do**

    for $i = 1, ..., n$ **do**

      sample $h_i^{(t)} \sim p(h_i|\boldsymbol{v}^{(t)})$

    end

  end

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D
Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$
forall $\boldsymbol{v} \in D$ **do**
  Randomly initialize $\boldsymbol{v}^{(0)}$
  for $t = 0, ..., k, k+1, ..., k+r$ **do**
    for $i = 1, ..., n$ **do**
      sample $h_i^{(t)} \sim p(h_i|\boldsymbol{v}^{(t)})$
    end
    for $j = 1, ..., m$ **do**
    end
  end
end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

```
init W, b, c
forall v ∈ D do
    Randomly initialize v^(0)
    for t = 0, ..., k, k + 1, ..., k + r do
        for i = 1, ..., n do
            sample h_i^(t) ~ p(h_i | v^(t))
        end
        for j = 1, ..., m do
            sample v_j^(t+1) ~ p(v_j | h^(t))
        end
    end
end
```

Algorithm 0: RBM Training with Block Gibbs Sampling

```
Input: RBM (V_1, ..., V_m, H_1, ..., H_n), training batch D
Output: Learned Parameters W, b, c
init W, b, c
forall v ∈ D do
    Randomly initialize v^(0)
    for t = 0, ..., k, k+1, ..., k+r do
        for i = 1, ..., n do
            sample h_i^(t) ~ p(h_i | v^(t))
        end
        for j = 1, ..., m do
            sample v_j^(t+1) ~ p(v_j | h^(t))
        end
    end
    W ← W + η ∇_W 𝓛(θ)
end
```

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

  for $t = 0, ..., k, k+1, ..., k+r$ **do**

    for $i = 1, ..., n$ **do**

      sample $h_i^{(t)} \sim \; p(h_i|\boldsymbol{v}^{(t)})$

    end

    for $j = 1, ..., m$ **do**

      sample $v_j^{(t+1)} \sim \; p(v_j|\boldsymbol{h}^{(t)})$

    end

  end

  $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \frac{1}{r}\sum_{t=k+1}^{k+r} \sigma(\mathbf{W}\boldsymbol{v}^{(t)} + \boldsymbol{c})\boldsymbol{v}^{(t)T}]$

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

  for $t = 0, ..., k, k+1, ..., k+r$ **do**

    for $i = 1, ..., n$ **do**

      sample $h_i^{(t)} \sim \ p(h_i|\boldsymbol{v}^{(t)})$

    end

    for $j = 1, ..., m$ **do**

      sample $v_j^{(t+1)} \sim \ p(v_j|\boldsymbol{h}^{(t)})$

    end

  end

  $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \frac{1}{r}\sum_{t=k+1}^{k+r} \sigma(\mathbf{W}\boldsymbol{v}^{(t)} + \boldsymbol{c})\boldsymbol{v}^{(t)T}]$

  $\boldsymbol{b} \leftarrow \boldsymbol{b} + \eta \nabla_{\boldsymbol{b}} \mathscr{L}(\theta)$

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

  for $t = 0, ..., k, k+1, ..., k+r$ **do**

    for $i = 1, ..., n$ **do**

      sample $h_i^{(t)} \sim \ p(h_i|\boldsymbol{v}^{(t)})$

    end

    for $j = 1, ..., m$ **do**

      sample $v_j^{(t+1)} \sim \ p(v_j|\boldsymbol{h}^{(t)})$

    end

  end

  $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \frac{1}{r}\sum_{t=k+1}^{k+r} \sigma(\mathbf{W}\boldsymbol{v}^{(t)} + \boldsymbol{c})\boldsymbol{v}^{(t)T}]$

  $\boldsymbol{b} \leftarrow \boldsymbol{b} + \eta[\boldsymbol{v}_d - \frac{1}{r}\sum_{t=k+1}^{k+r} \boldsymbol{v}^{(t)}]$

end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D
Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$
init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$
forall $\boldsymbol{v} \in D$ **do**
  Randomly initialize $\boldsymbol{v}^{(0)}$
  for $t = 0, ..., k, k+1, ..., k+r$ **do**
    for $i = 1, ..., n$ **do**
      sample $h_i^{(t)} \sim \ p(h_i|\boldsymbol{v}^{(t)})$
    end
    for $j = 1, ..., m$ **do**
      sample $v_j^{(t+1)} \sim \ p(v_j|\boldsymbol{h}^{(t)})$
    end
  end
  $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \frac{1}{r}\sum_{t=k+1}^{k+r} \sigma(\mathbf{W}\boldsymbol{v}^{(t)} + \boldsymbol{c})\boldsymbol{v}^{(t)T}]$
  $\boldsymbol{b} \leftarrow \boldsymbol{b} + \eta[\boldsymbol{v}_d - \frac{1}{r}\sum_{t=k+1}^{k+r} \boldsymbol{v}^{(t)}]$
  $\boldsymbol{c} \leftarrow \boldsymbol{c} + \eta\nabla_{\boldsymbol{c}}\mathscr{L}(\theta)$
end

Algorithm 0: RBM Training with Block Gibbs Sampling

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

forall $\boldsymbol{v} \in D$ **do**

  Randomly initialize $\boldsymbol{v}^{(0)}$

  for $t = 0, ..., k, k+1, ..., k+r$ **do**

    for $i = 1, ..., n$ **do**

      sample $h_i^{(t)} \sim \; p(h_i|\boldsymbol{v}^{(t)})$

    end

    for $j = 1, ..., m$ **do**

      sample $v_j^{(t+1)} \sim \; p(v_j|\boldsymbol{h}^{(t)})$

    end

  end

  $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \frac{1}{r}\sum_{t=k+1}^{k+r} \sigma(\mathbf{W}\boldsymbol{v}^{(t)} + \boldsymbol{c})\boldsymbol{v}^{(t)T}]$

  $\boldsymbol{b} \leftarrow \boldsymbol{b} + \eta[\boldsymbol{v}_d - \frac{1}{r}\sum_{t=k+1}^{k+r} \boldsymbol{v}^{(t)}]$

  $\boldsymbol{c} \leftarrow \boldsymbol{c} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c}) - \frac{1}{r}\sum_{t=k+1}^{k+r} \sigma(\mathbf{W}\boldsymbol{v}^{(t)} + \boldsymbol{c})]$

end

Module 20.5 : Training RBMs using Contrastive Divergence

- In practice, Gibbs Sampling can be very inefficient because for every step of stochastic gradient descent we need to run the Markov chain for many many steps and then compute the expectation using the samples drawn from this chain

- In practice, Gibbs Sampling can be very inefficient because for every step of stochastic gradient descent we need to run the Markov chain for many many steps and then compute the expectation using the samples drawn from this chain
- We will now see a more efficient algorithm called k-contrastive divergence which is used in practice for training RBMs

- Just to reiterate, our goal is to compute the two expectations efficiently

$$\mathbb{E}_{p(H|V)}[v_j h_i] = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

$$\mathbb{E}_{p(V,H)}[v_j h_i] = \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

- Just to reiterate, our goal is to compute the two expectations efficiently
- We already have a simplified formula for the first expectation

$$\mathbb{E}_{p(H|V)}[v_j h_i] = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

$$\mathbb{E}_{p(V,H)}[v_j h_i] = \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

$$\mathbb{E}_{p(H|V)}[v_j h_i] = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

$$\mathbb{E}_{p(V,H)}[v_j h_i] = \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

- Just to reiterate, our goal is to compute the two expectations efficiently
- We already have a simplified formula for the first expectation
- Furthermore, note that the first expectation depends only on the seen training example ($\boldsymbol{v}$)

$$\mathbb{E}_{p(H|V)}[v_j h_i] = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

$$\mathbb{E}_{p(V,H)}[v_j h_i] = \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

- Just to reiterate, our goal is to compute the two expectations efficiently
- We already have a simplified formula for the first expectation
- Furthermore, note that the first expectation depends only on the seen training example ($\boldsymbol{v}$)
- The second expectation depends on the samples drawn from the Markov chain $(v_1, v_2, ..., v_n)$

$$\mathbb{E}_{p(H|V)}[v_j h_i] = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i)v_j$$

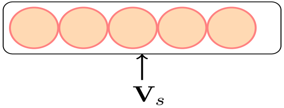
$$\mathbb{E}_{p(V,H)}[v_j h_i] = \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i)v_j$$

- Just to reiterate, our goal is to compute the two expectations efficiently
- We already have a simplified formula for the first expectation
- Furthermore, note that the first expectation depends only on the seen training example ($\boldsymbol{v}$)
- The second expectation depends on the samples drawn from the Markov chain $(v_1, v_2, ..., v_n)$
- The first expectation thus depends on the empirical samples, whereas the second expectation depends on the model samples (because the samples are generated based on $P(V|H)$ and $P(H|V)$ output by the model)

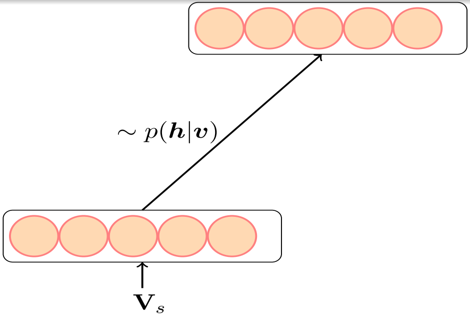
- Contrastive divergence uses the following idea

- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance

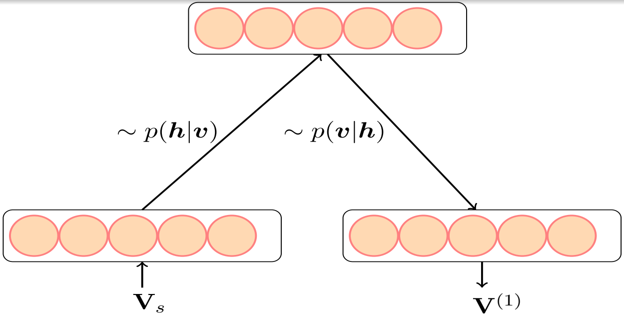
- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$



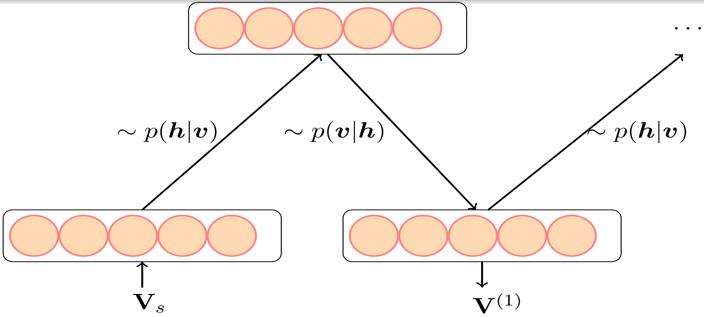
- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate



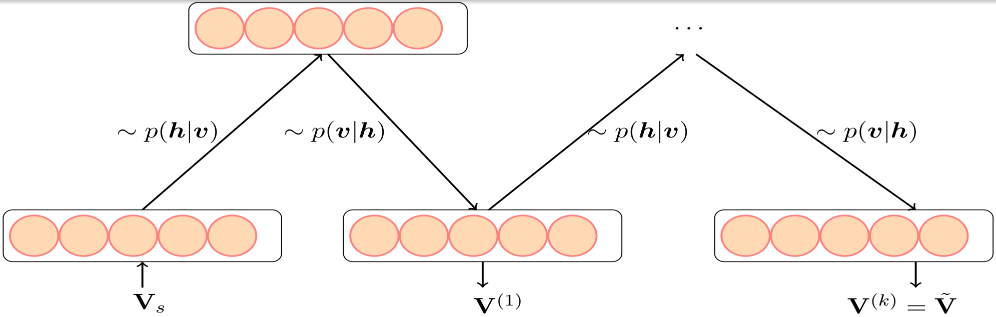
- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate



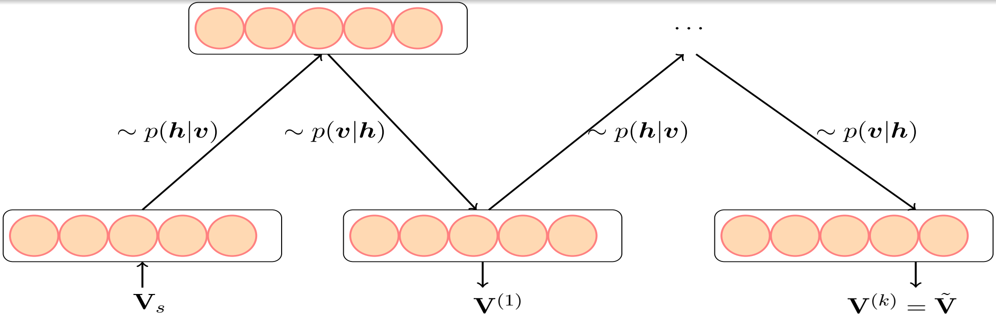
- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate



- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate

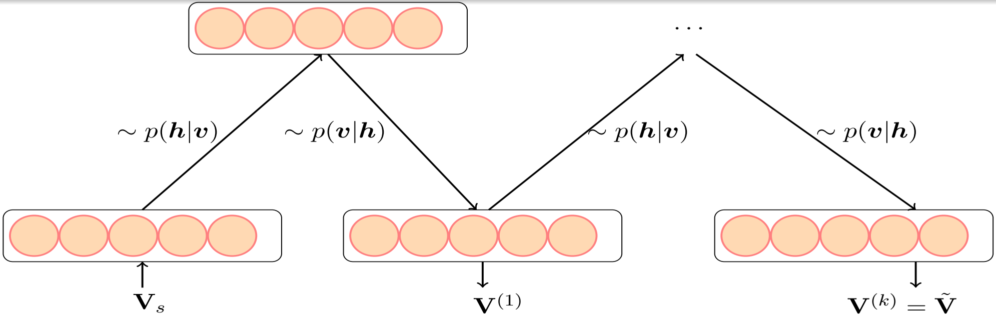


- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate



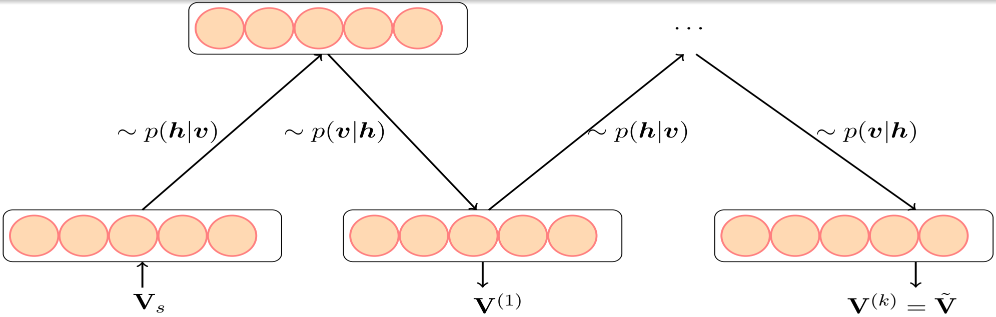
- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate

$$\mathbb{E}_{p(V,H)}[v_j h_i]$$



- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate

$$\mathbb{E}_{p(V,H)}[v_j h_i] = \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\boldsymbol{w}_i\boldsymbol{v} + c_i)v_j$$



- Contrastive divergence uses the following idea
- Instead of starting the Markov Chain at a random point ($V = \boldsymbol{v}^0$), start from $\boldsymbol{v}^{(t)}$ where $\boldsymbol{v}^{(t)}$ is the current training instance
- Run Gibbs Sampling for k steps and denote the sample at the k^{th} step by $\tilde{\boldsymbol{v}}$
- Replace the expectation by a point estimate

$$\mathbb{E}_{p(V,H)}[v_j h_i] = \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\boldsymbol{w}_i\boldsymbol{v} + c_i)v_j \approx \sigma(\boldsymbol{w}_i\tilde{\boldsymbol{v}} + c_i)\tilde{v}_j$$

- Over time as our model becomes better and better $\tilde{\boldsymbol{v}}$ should start looking more and more like our training (empirical) samples

- Over time as our model becomes better and better $\tilde{v}$ should start looking more and more like our training (empirical) samples
- Once that starts happening what will happen to the gradient ?

- Over time as our model becomes better and better $\tilde{\boldsymbol{v}}$ should start looking more and more like our training (empirical) samples
- Once that starts happening what will happen to the gradient ?
- We consider the derivative w.r.t w_{ij} again

- Over time as our model becomes better and better $\tilde{\boldsymbol{v}}$ should start looking more and more like our training (empirical) samples
- Once that starts happening what will happen to the gradient ?
- We consider the derivative w.r.t w_{ij} again

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\sum_{j=1}^{m} \boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

- Over time as our model becomes better and better $\tilde{\boldsymbol{v}}$ should start looking more and more like our training (empirical) samples
- Once that starts happening what will happen to the gradient ?
- We consider the derivative w.r.t w_{ij} again

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} \boldsymbol{w}_i \boldsymbol{v} + c_i)v_j$$

- We have two summations here

- Over time as our model becomes better and better $\tilde{\boldsymbol{v}}$ should start looking more and more like our training (empirical) samples
- Once that starts happening what will happen to the gradient ?
- We consider the derivative w.r.t w_{ij} again

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} \boldsymbol{w}_i \boldsymbol{v} + c_i)v_j$$

- We have two summations here
- The first term can be thought of as summation over a single point v from training example

- Over time as our model becomes better and better $\tilde{\boldsymbol{v}}$ should start looking more and more like our training (empirical) samples
- Once that starts happening what will happen to the gradient ?
- We consider the derivative w.r.t w_{ij} again

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i) v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v}) \sigma(\sum_{j=1}^{m} \boldsymbol{w}_i \boldsymbol{v} + c_i) v_j$$

- We have two summations here
- The first term can be thought of as summation over a single point v from training example
- Similarly, for the second term, the summation over $\tilde{\boldsymbol{v}}$ is being replaced by a point estimate computed from the model sample

- Over time as our model becomes better and better $\tilde{\boldsymbol{v}}$ should start looking more and more like our training (empirical) samples
- Once that starts happening what will happen to the gradient ?
- We consider the derivative w.r.t w_{ij} again

$$\frac{\partial \mathscr{L}(\theta)}{\partial w_{ij}} = \sigma(\boldsymbol{w}_i \boldsymbol{v} + c_i)v_j - \sum_{\boldsymbol{v}} p(\boldsymbol{v})\sigma(\sum_{j=1}^{m} \boldsymbol{w}_i \boldsymbol{v} + c_i)v_j$$

- We have two summations here
- The first term can be thought of as summation over a single point v from training example
- Similarly, for the second term, the summation over $\tilde{\boldsymbol{v}}$ is being replaced by a point estimate computed from the model sample
- As training progresses and $\tilde{\boldsymbol{v}}$ (model sample) starts looking more and more like our training (empirical) samples, the difference between the two terms will be small and the parameters of the model will stabilize (convergence)

Algorithm 0: k-step Contrastive Divergence

Algorithm 0: k-step Contrastive Divergence

Input:

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output:

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

  Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

  Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

  for $t = 0, ..., k$ **do**

  end

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

  Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

  for $t = 0, ..., k$ **do**

    for $i = 1, ..., n$ **do**

    end

  end

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

  Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

  for $t = 0, ..., k$ **do**

    for $i = 1, ..., n$ **do**

      sample $h_i^{(t)} \sim \ p(h_i|\boldsymbol{v}^{(t)})$

    end

  end

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D
Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$
init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$
forall $\boldsymbol{v} \in D$ **do**
    Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$
    for $t = 0, ..., k$ **do**
        for $i = 1, ..., n$ **do**
            sample $h_i^{(t)} \sim p(h_i|\boldsymbol{v}^{(t)})$
        end
        for $j = 1, ..., m$ **do**
        end
    end
end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

    Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

    for $t = 0, ..., k$ **do**

        for $i = 1, ..., n$ **do**

            sample $h_i^{(t)} \sim \ p(h_i|\boldsymbol{v}^{(t)})$

        end

        for $j = 1, ..., m$ **do**

            sample $v_j^{(t+1)} \sim \ p(v_j|\boldsymbol{h}^{(t)})$

        end

    end

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D
Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

```
init W = b = c = 0
forall v ∈ D do
    Initialize v^(0) ← v
    for t = 0, ..., k do
        for i = 1, ..., n do
            sample h_i^(t) ~ p(h_i | v^(t))
        end
        for j = 1, ..., m do
            sample v_j^(t+1) ~ p(v_j | h^(t))
        end
    end
    W ← W + η ∇_W L(θ)
end
```

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \nabla_{\mathbf{W}} \mathscr{L}(\theta)$$

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D
Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

```
init W = b = c = 0
forall v ∈ D do
    Initialize v^(0) ← v
    for t = 0, ..., k do
        for i = 1, ..., n do
            sample h_i^(t) ~ p(h_i | v^(t))
        end
        for j = 1, ..., m do
            sample v_j^(t+1) ~ p(v_j | h^(t))
        end
    end
    W ← W + η[σ(W v_d + c) v_d^T − σ(W ṽ + c) ṽ]
end
```

$$\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \sigma(\mathbf{W}\tilde{\boldsymbol{v}} + \boldsymbol{c})\tilde{\boldsymbol{v}}]$$

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

    Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

    for $t = 0, ..., k$ **do**

        for $i = 1, ..., n$ **do**

            sample $h_i^{(t)} \sim \; p(h_i|\boldsymbol{v}^{(t)})$

        end

        for $j = 1, ..., m$ **do**

            sample $v_j^{(t+1)} \sim \; p(v_j|\boldsymbol{h}^{(t)})$

        end

    end

    $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \sigma(\mathbf{W}\tilde{\boldsymbol{v}} + \boldsymbol{c})\tilde{\boldsymbol{v}}]$

    $\boldsymbol{b} \leftarrow \boldsymbol{b} + \eta\nabla_{\boldsymbol{b}}\mathscr{L}(\theta)$

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

    Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

    for $t = 0, ..., k$ **do**

        for $i = 1, ..., n$ **do**

            sample $h_i^{(t)} \sim \; p(h_i|\boldsymbol{v}^{(t)})$

        end

        for $j = 1, ..., m$ **do**

            sample $v_j^{(t+1)} \sim \; p(v_j|\boldsymbol{h}^{(t)})$

        end

    end

    $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \sigma(\mathbf{W}\tilde{\boldsymbol{v}} + \boldsymbol{c})\tilde{\boldsymbol{v}}]$

    $\boldsymbol{b} \leftarrow \boldsymbol{b} + \eta[\boldsymbol{v} - \tilde{\boldsymbol{v}}]$

end

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D
Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

```
init W = b = c = 0
forall v ∈ D do
    Initialize v^(0) ← v
    for t = 0, ..., k do
        for i = 1, ..., n do
            sample h_i^(t) ~ p(h_i | v^(t))
        end
        for j = 1, ..., m do
            sample v_j^(t+1) ~ p(v_j | h^(t))
        end
    end
    W ← W + η[σ(W v_d + c) v_d^T − σ(W ṽ + c) ṽ]
    b ← b + η[v − ṽ]
    c ← c + η ∇_c ℒ(θ)
end
```

Algorithm 0: k-step Contrastive Divergence

Input: RBM $(V_1, ..., V_m, H_1, ..., H_n)$, training batch D

Output: Learned Parameters $\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}$

init $\mathbf{W} = \boldsymbol{b} = \boldsymbol{c} = 0$

forall $\boldsymbol{v} \in D$ **do**

  Initialize $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{v}$

  for $t = 0, ..., k$ **do**

    for $i = 1, ..., n$ **do**

      sample $h_i^{(t)} \sim \; p(h_i|\boldsymbol{v}^{(t)})$

    end

    for $j = 1, ..., m$ **do**

      sample $v_j^{(t+1)} \sim \; p(v_j|\boldsymbol{h}^{(t)})$

    end

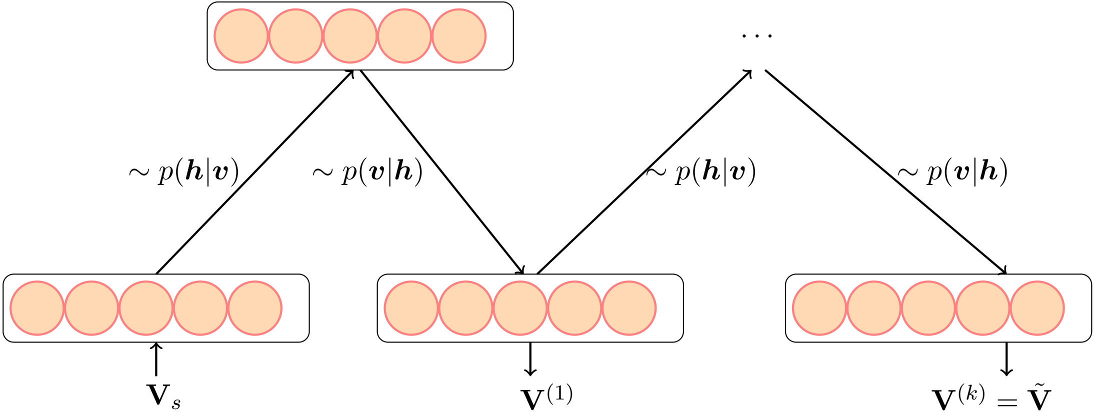
  end

  $\mathbf{W} \leftarrow \mathbf{W} + \eta[\sigma(\mathbf{W}\boldsymbol{v}_d + \boldsymbol{c})\boldsymbol{v}_d^T - \sigma(\mathbf{W}\tilde{\boldsymbol{v}} + \boldsymbol{c})\tilde{\boldsymbol{v}}]$

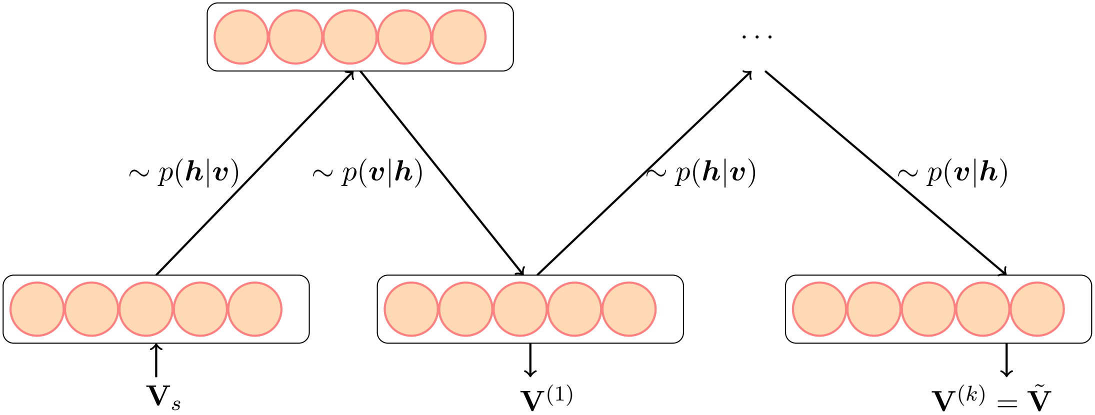
  $\boldsymbol{b} \leftarrow \boldsymbol{b} + \eta[\boldsymbol{v} - \tilde{\boldsymbol{v}}]$

  $\boldsymbol{c} \leftarrow \boldsymbol{c} + \eta[\sigma(\mathbf{W}\boldsymbol{v} + \boldsymbol{c}) - \sigma(\mathbf{W}\tilde{\boldsymbol{v}} + \boldsymbol{c})]$

end



- In practice, $k = 1$ also works well



- In practice, $k = 1$ also works well
- The higher the value of k, the less biased the estimate of the gradient will be.